AN ANALYSIS OF REASONS FOR USING MULTIPLE VETERINARIANS IN THE EQUINE INDUSTRY

Universiteit Utrecht

Abstract:

This study provides an in-depth exploration of the reasons horse owners utilize multiple veterinarians and the attributes they most value in their primary veterinarians. By leveraging text analytics, the study dissected the responses of two client groups in the equine industry with a total of 1389 respondents. The research highlighted the shared and distinct preferences among these groups. Veterinarians demonstrated a preference for quality of care and professional expertise, valuing aspects such as dental care, emergency services, and localized service availability. However, they showed differences in their preferences. One group mainly placing emphasis on continuity and personalized care, while the other group focused more on resolving specific issues, such as lameness. The results also revealed universally valued attributes in a primary vet across all groups, including honesty, knowledge and competency. The study offers a comprehensive understanding of client needs and preferences in equine veterinary care and highlights potential areas for service improvement to meet diverse client requirements. The findings contribute to the growing body of research on consumer behaviour in the equine veterinary industry and provide actionable insights for veterinary practitioners to tailor their services and improve client satisfaction.

Contents

1. Introduction

Horses play an integral role in society, contributing to industries such as sports, recreation and bringing companionship to many. To ensure the horses well-being, regular veterinary care, which ranges from routine check-ups to vaccinations and treatments for various illnesses and injuries is essential. A growing trend in the equine industry of horse owners is to engage with multiple veterinarians for their horses' care, this study seeks to understand the underlying reasons and implications of this practice.

The complex and varied health needs of horses often necessitate the engagement of multiple veterinary practitioners by horse owners (Batchelor & McKeegan, 2012). While the multiplicity of care can offer a wider spectrum of services and expertise, it also poses unique challenges in ensuring consistency, continuity, and quality of care (Coe, Adams, & Bonnett, 2007). Understanding the perspectives of horse owners on why they engage multiple veterinarians and what they value in their primary veterinarian, can offer significant insights for improving the design and delivery of veterinary services.

The role of veterinarians extends beyond the provision of veterinary services; they are also instrumental in informing owners' decisions regarding their horses' health and welfare (Elte et al., 2021). As such, a more nuanced understanding of the client-veterinarian relationship and the factors influencing veterinary service utilization is crucial to better meet client needs and ensure optimal horse care.

Research into pet owners' perceptions and behaviours regarding their pets' veterinary care has been previously conducted (Coe, Adams, & Bonnett, 2007; Kogan, et al., 2018). However, there is a scarcity of studies specifically investigating the viewpoints of horse owners, particularly in the context of using multiple veterinary services. Equine veterinary care usually demands more specialized and complex attention compared to smaller animals, creating a distinctive area of study (Mcbride, Hemmings 2004).

This research paper aims to fill this gap by employing various data science techniques to analyse textual responses from horse owners regarding their reasons for using multiple veterinarians and their most appreciated aspects of the primary vet. These methods include Latent Dirichlet Allocation (LDA), Term Frequency-Inverse Document Frequency (TF-IDF), Jaccard similarity, n-gram analysis, and decision tree classifier. Through these techniques, we seek to uncover patterns and themes in the attitudes and perceptions of horse owners, which could guide the improvement of equine veterinary services and enhance client satisfaction.

The research paper is structured as follows: First, we present the background and data source. Then the methodology of the data processing and analysis. This is followed by the chapters exploratory data analysis and in-depth analysis. These chapters include the key themes identified in the horse owners' responses followed by a discussion on the implications for veterinary practice. The final part of this paper is a summary of the key findings and their potential impact on equine veterinary practices, we acknowledge the limitations of our study and suggest potential possibilities for future research.

By adopting these innovative (data analysis) methods, we strive to offer a more thorough and insightful analysis than traditional methods might permit. This approach seeks to effectively link the

diverse background of equine veterinary practice to our research tools, ensuring a comprehensive, innovative, and nuanced exploration of why horse owners use multiple veterinarians.

2. Background

Client satisfaction in equine veterinary practice is a complex, multifaceted construct that extends beyond giving only veterinary services. It encompasses a diverse range of elements that together contribute to the overall experience of the horse owner (Elte et al., 2021). As such, understanding the factors that influence client satisfaction is crucial for enhancing service delivery, improving client retention, and ultimately, promoting the welfare of the horse.

This study particularly zooms in on a rather unexplored aspect, the reasons horse owners have for engaging multiple veterinarians, and what they value in their primary veterinarian. This distinctive focus aims to deepen our understanding of the quality of care from the clients' perspective.

Quality of care is a key part of client satisfaction and goes beyond the basic requirement of possessing the knowledge and skills for diagnosing and treating physical health conditions. It involves demonstrating a high level of competence in a wide range of surgical procedures, as well as a deep understanding of animal behaviour and welfare. This includes continuous learning of the latest developments in veterinary medicine (Coe et al., 2007). Given the variety of health issues that horses can present with, the expertise of the equine veterinarian in these areas is directly tied to horse owner satisfaction.

Quality of service is not solely about the medical aspects of veterinary care but also encompasses the logistical and administrative aspects of the practice. Horse owners appreciate practices that offer flexible appointment scheduling, uphold high standards of hygiene, and are easily accessible in terms of location and contact ability (Coe et al., 2007).

Horsemanship, which entails a nuanced understanding of equine behaviour and a gentle, confident approach to handling horses, is another aspect of veterinary practice that can greatly influence horse owner satisfaction. Good horsemanship can minimize stress for the horse during veterinary procedures, thereby enhancing the horse's welfare and increasing client satisfaction. Veterinarians who demonstrate good horsemanship not only show respect for the horse as a patient but also exhibit a level of industry knowledge that horse owners appreciate (Elte et al., 2021).

The cost of service has a role to play. Although this is not the primary factor in client satisfaction, it is important for horse owners to have the idea that they are receiving value for their money. This perception can be improved by providing detailed estimates of treatment costs upfront, offering a clear explanation of the bill, as well as ensuring that clients understand the potential cost of different treatment options (Elte et al., 2021).

Interpersonal skills are integral to effective veterinary practice. Beyond the transmission of information, they involve empathy, the ability to understand and respond to client concerns, and the capacity to manage communication in stressful or emotionally charged situations (like someone's favorite horse is possibly dying). Veterinarians who outperform in these areas can build strong relationships with clients, leading to higher levels of client satisfaction (Elte et al., 2021).

Transfer of knowledge is another key factor. The ability to educate clients about the horse condition in a way that is understandable and useful is a discipline on its own. When clients feel informed and involved in their horse's care, they are likely to be more satisfied with the service they receive from the veterinary. More important, an educated client can contribute to better health outcomes for the horse by making informed decisions about ongoing care and following through on home care instructions (Elte et al., 2021).

Professionalism extends beyond maintaining up-to-date knowledge and skills. It involves maintaining the reputation of the veterinary profession by treating clients and their animals with respect, being honest and being transparent in communications. Also, they should effectively manage any conflicts of interest that may arise. The ability to balance the welfare of the horse, the needs and expectations of the client, and the ethical responsibility of the profession is a key aspect of professionalism in veterinary practice (Elte et al., 2021; Batchelor & McKeegan, 2012).

Equine veterinary practice can significantly benefit from the insights provided by Elte et al. (2021). The emphasis on high-quality care and service, interpersonal skills, and horsemanship, among other factors, underscores the multifaceted nature of horse owner satisfaction and highlights the need for a holistic approach to service delivery.

Excellence in medical skills is a basic expectation, but alone, it is insufficient. Veterinarians also need to focus on the service aspects of their practice. This might involve offering flexible scheduling options, maintaining a well-kept and hygienic practice, and providing clear transparent information about treatment costs. Additionally, they should strive to establish and maintain effective communication channels to keep clients informed about their horse's progress and promptly address any concerns risen by horse owners.

Horsemanship and interpersonal skills are other areas that equine veterinarians should pay attention to. Regular training to improve horsemanship skills can help minimize stress for the horse and improve horse owner satisfaction. Similarly, developing communication and interpersonal skills can go a long way in building strong relationships with horse owners. This might involve attending workshops or courses on client communication or seeking feedback from clients and colleagues to identify areas for improvement.

Professionalism is another area that veterinarians should focus on. This includes keeping their knowledge and skills up to date through ongoing professional development, behaving in a manner that upholds the reputation of the profession, treating clients and their animals with respect and effectively managing the complex ethical challenges that arise in veterinary practice.

Building on the diverse aspects of equine veterinary practice and client satisfaction, it becomes clear that assessing horse owner perspectives is a multi-layered and complex task. Traditional methods of qualitative analysis may fall short in covering the breadth and depth of these perspectives. This is where the research stands out. Leveraging advanced data science techniques to delve into the multi-dimensional opinions and attitudes of horse owners.

The multifaceted nature of client satisfaction – encompassing quality of care, quality of service, horsemanship, cost considerations, interpersonal skills, knowledge transfer, and professionalism – mirrors the complexity of the data we aim to analyze. Considering this, the use of advanced data science techniques, such as Latent Dirichlet Allocation (LDA), Term Frequency-Inverse Document Frequency (TF-IDF), n-gram analysis, and decision tree classifier are of value. These methods allow to not only analyze the textual responses of horse owners on a large scale but also capture the nuanced themes and patterns that underpin their attitudes and decisions.

In using these methods, we can provide a more detailed and insightful analysis than traditional methods might afford. This approach serves to bridge the connection between the multifaceted background of equine veterinary practice and the research tools employed, ensuring a robust, innovative, and nuanced exploration of horse owner reasons to use multiple veterinarians.

In conclusion, this study aims to offer an understanding of the factors that contribute to client satisfaction in equine veterinary practice by specifically analyzing horse owners' reasons for using multiple veterinarians and their most appreciated aspects of their primary vet. By focusing on these areas, equine veterinarians can enhance client satisfaction, improve the welfare of the horses under their care, and uphold the reputation of the veterinary profession. Future research could delve deeper into these themes, investigating specific actions and behaviors that equine veterinarians can adopt to further improve client satisfaction and animal welfare.

3. Data

This study is based upon an extensive survey with multiple questions in different areas of 1389 horse owners. This survey ensured we obtained a broad representation of the horse owning community and captured a comprehensive account of their experiences and preferences (Bryman, 2016). The responses were recorded in a mixture of formats including open-ended, multiple-choice, and numerical scaled questions to encourage comprehensive and diverse feedback from the participants.

The data for this study was collected using Qualtrics, an online survey platform recognized for its robust and diverse data collection capabilities (Qualtrics, Provo, UT) (Sax, Gilmartin, & Bryant, 2003). This survey tool allowed horse owners to provide responses conveniently and anonymously to the questionnaire, which were then compiled and exported from Qualtrics for further analysis.

Post extraction, all subsequent stages of exploratory data analysis (EDA), data cleaning, transformation, and integration were carried out using Python, a popular programming language for data analysis and machine learning applications (VanderPlas, 2016).

EDA, the first step of this analysis, involved the creation of visualizations and summary statistics for key variables, such as the number of veterinarians utilized, the frequency of consultations, and the reasons for using multiple veterinarians. This process helped in finding patterns and connections in the data relevant to the research questions (Pandas Development Team, 2020).

The data cleaning process, conducted using Python's pandas and NumPy libraries, included eliminating test responses and checking for missing values (McKinney, 2010; Walt, Colbert, & Varoquaux, 2011). Fortunately, the dataset was complete, and no imputation process was required for the relevant columns/attributes.

During the transformation stage, categorical variables were converted into frequency distributions using Python's data manipulation capabilities. These distributions highlighted the percentage of respondents ranking each aspect of care. Subsequently, entropy values were calculated from these frequency distributions, enabling the measurement of consensus among respondents. (Cover & Thomas, 2006).

Data integration involved collating the frequency distribution for each sub-scenario into a unified dataset. This dataset facilitated comparison of preferences across different aspects of care and different sub-scenarios. Using this integrated dataset, a heatmap and a dendrogram were generated using the seaborn and SciPy libraries in Python, respectively, providing visual insights into the relationships between the sub-scenarios and the aspects of care (Waskom, 2021; Virtanen et al., 2020).

Following the initial stages, the next stage involved an in-depth analysis of the dataset, further modeling and interpretating the results.

This in-depth analysis was implemented using Python. Key features of Python, such as its easy syntax, robust library support, and advanced data manipulation capabilities made it the ideal choice for the in-depth research. All the data wrangling steps were repeated and adjusted were relevant for the in-depth analysis.

There has been ethical approval for the usage and analysis of the data.

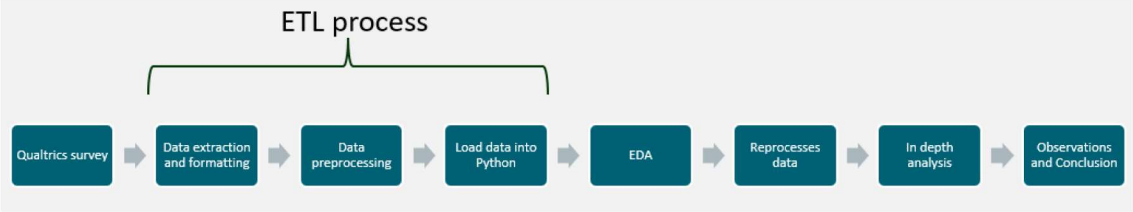


Figure 1 Extract, transform, load and analysis.

4. Method

The main research aim - understanding the reasons behind horse owners employing multiple veterinarians - made the focus on a subset of the survey. This subset incorporated the seven aspects of veterinary care as described in the background. Besides the aspects of care, some particular columns/attributes are of interest for the research question:

- The number of horses owners have.
- The main reasons for using multiple veterinarians.
- The most appreciated aspects about the primary veterinary.
- The duration of service (in years) of the primary veterinary.

4.1 Extract, transform, load process.

The data collection process utilized the Qualtrics platform for its robust features and wide usage in academic research (Sue & Ritter, 2012). This data was then exported into a spreadsheet format, which allowed for compatibility with multiple data analysis tools and methodologies (Hair et al., 2017).

In the context of the research, the preprocessing steps undertaken, including lowercasing, punctuation removal, tokenization, and stop words removal, were essential in preparing the data for subsequent analysis (Jurafsky & Martin, 2019). Considering that the survey was administered in both English and Dutch and contained domain-specific jargon, the removal of stop words was adapted to this context. Lowercasing and punctuation removal are fundamental techniques that standardize the data, eliminating any discrepancies in interpretation due to varying case usage or punctuation inconsistencies (Bird, Klein, & Loper, 2009). Tokenization, a process that breaks down the text into individual words or tokens, is necessary for the accurate application of text analysis methods (Manning, Raghavan, & Schütze, 2008). With the removal of stop words (English, Dutch, and several domain-specific terms added manually (Jurafsky & Martin, 2019)), the frequently used words with minimal semantic meaning, it further refines the data, allowing to focus on contextually relevant words in the analysis (Rajaraman & Ullman, 2011).

4.2 EDA methods

Post preprocessing, we conducted an exploratory data analysis (EDA) to gain initial insights into the patterns and trends within the data. The insights derived from the EDA, such as potential correlations and preliminary trends, informed the more nuanced and targeted analysis that followed.

To examine the variability in preferences across different aspects of veterinary care, we undertook an entropy analysis. Entropy, fundamentally, is a measure of uncertainty or randomness in a set of data. In the context of this study, a high entropy value implies a greater level of disagreement or variation in preferences among horse owners (Cover & Thomas, 1991). When calculated from the frequency distributions of each aspect of veterinary care, entropy provides us with a useful metric to pinpoint areas with high variability. These high-variability areas signal a lack of consensus among horse owners about the importance or quality of various care aspects. It serves as a powerful tool to

identify elements of veterinary care that might require greater standardization or areas where more communication between the veterinarian and client might be beneficial to align expectations and perceptions.

For visualization purposes and to provide an intuitive and accessible representation of these patterns, these entropy values are also presented in a heatmap. Through this visualization, we can readily see and compare the level of consensus or disagreement for each aspect, providing a clearer picture of where efforts to improve client understanding and satisfaction might be most needed.

Hierarchical clustering with Python's Scikit-learn library (Pedregosa et al., 2011) is used in the EDA. Hierarchical clustering, an unsupervised machine learning technique, clusters together similar items. In the context of this study, the aim was to unearth potential patterns among horse owners, particularly focusing on their reasons for consulting multiple veterinarians and the qualities they value in their primary veterinarian.

Unfortunately, the initial trials with hierarchical clustering failed to produce distinct or significant clusters that could provide additional insights into the data. To validate these findings, the Elbow method was used, this is a widely recognized technique in clustering analysis. This method plots the explained variation against the number of clusters, identifying the 'elbow' point where further clusters don't substantially contribute to the explained variation (MacQueen, 1967).

For the dataset, the elbow method affirmed the initial findings: adding more clusters did not significantly enhance our understanding of the data. Therefore, we decided to exclude hierarchical clustering from further in-depth analysis.

However, a cross-tabulation was performed to examine the preferences among professional and recreational users. This form of analysis can shed light on trends and patterns in how these two distinct groups of horse owners perceive and engage with veterinary services, thereby providing useful insights for service improvement and customization.

4.3 In-depth analysis methods

While EDA provided valuable first insights, it overlooked the contextual relevance of words and phrases. Thus, a more in-depth and comprehensive analysis was conducted, involving several advanced techniques:

The decision to use topic modeling via Latent Dirichlet Allocation (LDA) and the Term Frequency-Inverse Document Frequency (TF-IDF) metric, came from the nature of the research question. These analytical methods have a track record of revealing hidden semantic structures and distinguishing significant features within a textual dataset, aligning with the objective to understand the motivations behind horse owners using multiple veterinarians (Blei, Ng, & Jordan, 2003; Ramos, 2003).

Topic modeling is an unsupervised machine learning method utilized to discover hidden thematic structures in a document corpus. This technique identifies words that commonly co-occur in the same context, suggesting a common topic or theme (Blei, Ng & Jordan, 2003). By applying this method, significant themes within the text data are revealed, thereby enhancing the understanding of major factors impacting client satisfaction in equine veterinary care.

The LDA model, a type of topic modeling framework, was used to uncover latent topics hidden within the data. By identifying patterns of word co-occurrences, it unveils topics and their associations within the text corpus. This offers deep insights into the reasons for horse owners using multiple veterinarians (Blei et al., 2003). To optimize the model, an iterative process was used where we compared models with a different number of topics, ranging from 5 to 20. We settled on the optimal number of topics (10) based on a balance of automated metrics and manual examination for interpretability.

TF-IDF: To underscore distinctive textual features, the Term Frequency-Inverse Document Frequency (TF-IDF) analysis was used. TF-IDF is a weighting scheme which calculates the relative importance of a word within a specific document and their presence in the entire corpus of documents.

The formula for TF-IDF: $$TF - IDF(t, d, D) = TF(t, d) * IDF(t, D)$$

where:

- t represents a term (word)
- d is a specific document
- D represents the entire corpus of documents

TF(t, d) is the Term Frequency - the number of times term t appears in document d.
IDF(t, D) is the Inverse Document Frequency, calculated as log(Total number of documents / Number of documents with term t in it)

The resulting TF-IDF score is high for terms that are more prevalent in a specific document but less frequent across other documents, thereby enabling the identification of terms that are uniquely significant to each document (Ramos, 2003).

In the context of the research, TF-IDF analysis was used to highlight unique reasons cited by horse owners for engaging multiple veterinarians and the specific attributes they appreciate in their primary veterinarians. This method was particularly useful in distinguishing the unique semantic features within the responses of different participants, providing a nuanced understanding of their individual perspectives.

The combination of LDA and TF-IDF proved particularly effective for our research question. While LDA helped uncover underlying topics in the corpus, TF-IDF pinpointed key terms that differentiated these topics. This combination enabled a more comprehensive and nuanced understanding of the complex semantic relationships within the data, allowing us to address the research question effectively.

Similarity Assessment: In the study, we aimed to understand if there were common themes among different respondent groups. To facilitate this, the Jaccard index was used. This is a measure commonly used for assessing the similarity and diversity between two sets (Real & Vargas, 1996).

By comparing sets of unique words, we were able to quantify the overlap or shared vocabulary among the groups of 2 and 3(+) veterinary users, providing insights into potential similarity patterns and associations in their reasons for using multiple veterinarians. The Jaccard index formula is as follows:

$$Jaccard\ Index = (the\ number\ of\ shared\ words\ between\ Set\ A\ and\ Set\ B) \\ / (the\ total\ number\ of\ unique\ words\ across\ both\ Set\ A\ and\ Set\ B)$$

The Jaccard index ranges from 0 (no shared words) to 1 (all words are shared), thereby enabling a precise understanding of the overlap between the groups' expressed reasons and sentiments.

N-grams: When analyzing textual data, understanding the context in which words appear and how they combine can be as important as the words themselves. For this reason, a n-gram analysis is used (Manning & Schütze, 1999).

An n-gram is a contiguous sequence of 'n' items from a given sample of text. A unigram (one-word), a bigram (two-word sequences) and a trigram (three-word sequences) analyses were used. This approach captures the relationship between consecutive words and identify phrases of significance that may not have been apparent in a single-word analysis like TF-IDF top features.

For instance, while analyzing single words might reveal frequent mentions of 'expertise' and 'availability', bigram and trigram analyses could potentially uncover significant phrases like 'specialized expertise' or 'always available'. This provides a deeper understanding of the horse owners' preferences and reasons for using multiple vets.

It is important to clarify that the n-gram analysis was a separate part of the textual data analysis and was not incorporated into the LDA model. Instead, it complemented the LDA and TF-IDF analyses by providing additional context and uncovering more nuanced patterns within the data.

Decision Tree Classifier: A decision tree classifier was used to organize and categorize the respondents' comments (Quinlan, 1986). By recognizing patterns and establishing rules within the responses, this tool guided us to the core of the research question.

The classifier's settings, including the depth of the tree, were optimized through cross-validation and parameter tuning to enhance the model's performance (James, et al., 2013). This iterative process helped to prevent overfitting and underfitting, ensuring that the model captured meaningful patterns without being overly sensitive to noise in the data.

The decision tree classifier worked in tandem with the LDA topic model. In the approach, the LDA model identified underlying topics in the 'reasons' for using multiple veterinarians, based on word distribution across the documents. By selecting the most prominent topic for each document, they were essentially categorizing each 'reason' into one of the topics, thus creating the classes for the decision tree.

The classifier then organized these topics or classes according to the features of the data. the features: the number of horses owned, the duration of service, and the number of vets involved were used. Each rule in the decision tree classifier was built upon these features, effectively forming a decision-making guide that helped us predict likely topics from the LDA model given specific conditions.

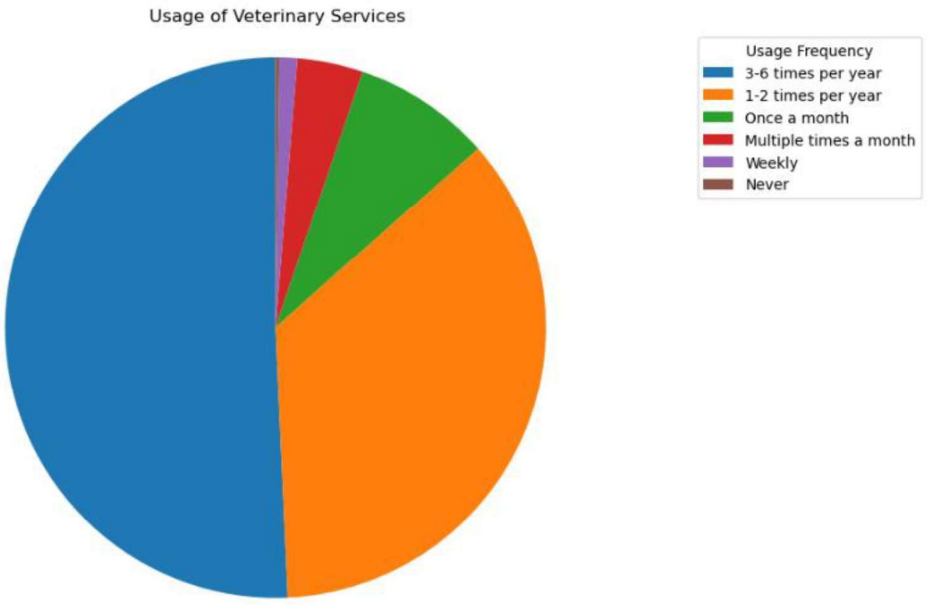
The Gini impurity measure, a measure of disorder within a node of the tree, played an important role in building the decision tree. It determined which feature to split at each node, aiming to enhance node homogeneity and improve the model's predictive accuracy (Raileanu & Stoffel, 2004).

Ultimately, this combination of topic modeling and decision tree classification allowed it to dive deeper into the motivations and concerns of horse owners, uncovering nuanced patterns that would have been difficult to identify through manual analysis alone.

The multiple data analysis strategies that were used to ensure the reliability and validity of the findings in this paper (Creswell & Creswell, 2017), are a mix of analysis techniques for text mining, providing a holistic view of the factors influencing the use of multiple veterinarians (Hair et al., 2017). An in-depth literature review was used to align the findings with existing literature and research. Together, these varied methods offered a comprehensive perspective on the multifaceted reasons influencing horse owners' reasons for using multiple veterinarians, underscoring the depth and complexity of the data at hand.

5. Results (exploratory)

The exploratory results began by assessing the frequency at which veterinary services were utilized by horse owners. The distribution of respondents' reported usage was depicted using a pie chart, supported by a detailed table (Figure 2). This step aimed to lay the foundation for understanding the interaction frequency between horse owners and veterinary services.



Usage Frequency	Count
3-6 times per year	670
1-2 times per year	473
Once a month	109
Multiple times a month	52
Weekly	14
Never	3

Figure 2 usage of veterinary services

Proceeding to the next phase of the results section, we sought to determine if a correlation existed between the number of horses (including ponies, foals, etc.) under the respondents' care and the number of veterinarians or practices they engage. In this analysis, an outlier value of 140 horses was excluded to prevent skewed results (Osborne & Overbay, 2004).

A statistical correlation analysis was conducted and found that the Pearson correlation was 0.09, indicating a very weak positive relationship. The spearman correlation was 0.12, also signifying a weak positive correlation. These results suggest that the number of horses a respondent cares for does not have a strong correlation with the number of veterinarians they use.

To visually represent this relationship, a violin plot was used. This specific type of plot was chosen as it combines the benefits of a boxplot and a kernel density plot, providing a detailed density estimation of the data at different values (Figure 3). The violin plot's width at any given point represents the proportion of data at that value, thus giving a clear view of data density.

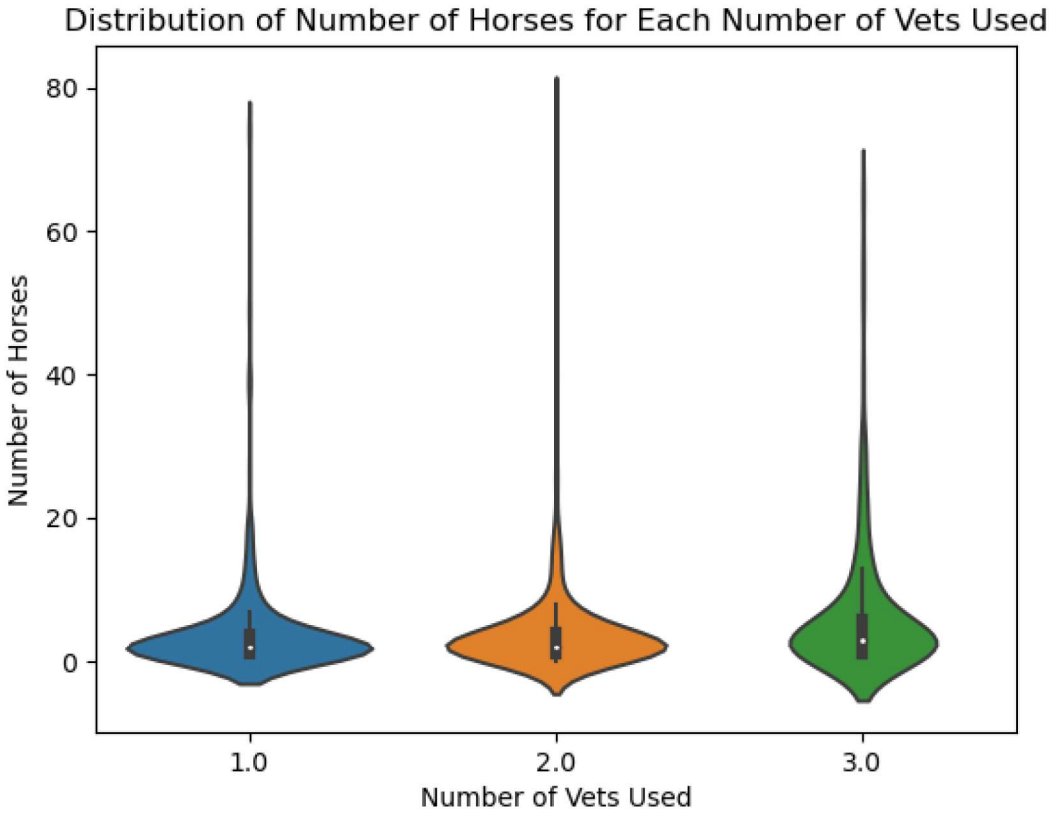


Figure 3 distribution of horses per number of vets

Following the investigation of horse numbers and veterinary engagements, the focus shifted towards the horse owners' preferences regarding aspects of their interaction with vets. The respondents were asked to rank the importance of various aspects during their vet's visit. The question posed was as follows: "Which aspect do you find most important around and during that visit from your vet? Put the aspect most important to you at number 1 and the least important at number 7." The frequency distributions of these rankings across the aspects were then analysed and visualized.

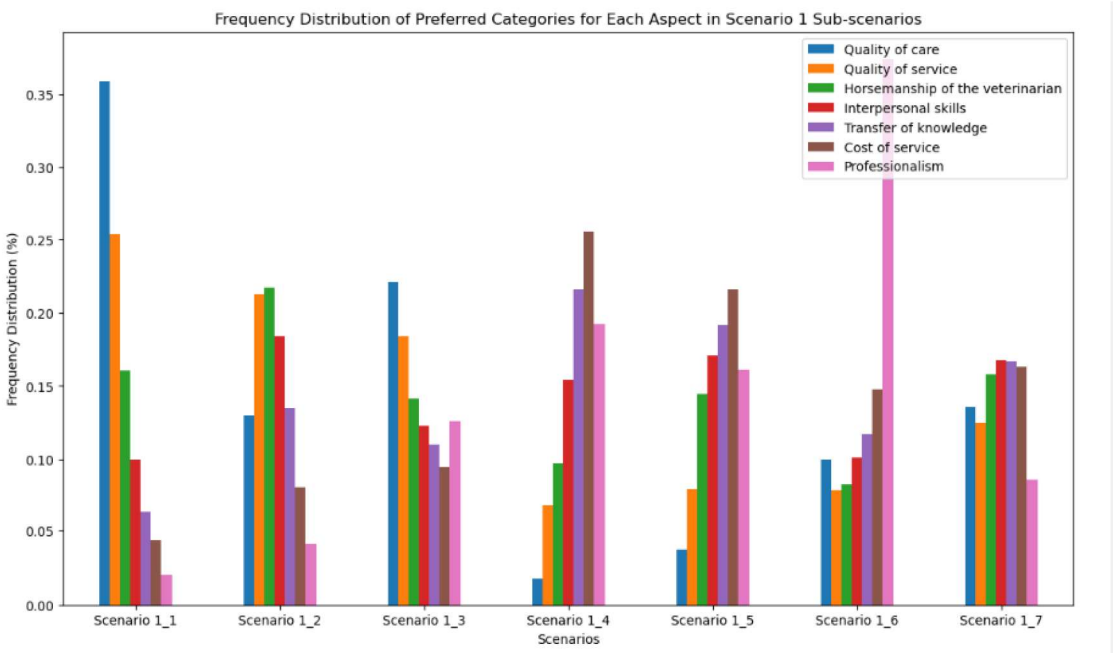


Figure 4 frequency distribution of most preferred aspects

The calculated entropy value of 1.75 indicates a moderate level of variability in the preferences. This is because the maximum possible entropy value, when there are seven equally likely outcomes, is log2(7) ≈ 2.81. An entropy of 1.75, therefore, suggests that while there is some variation in preferences, there is also a degree of consistency among the respondents.

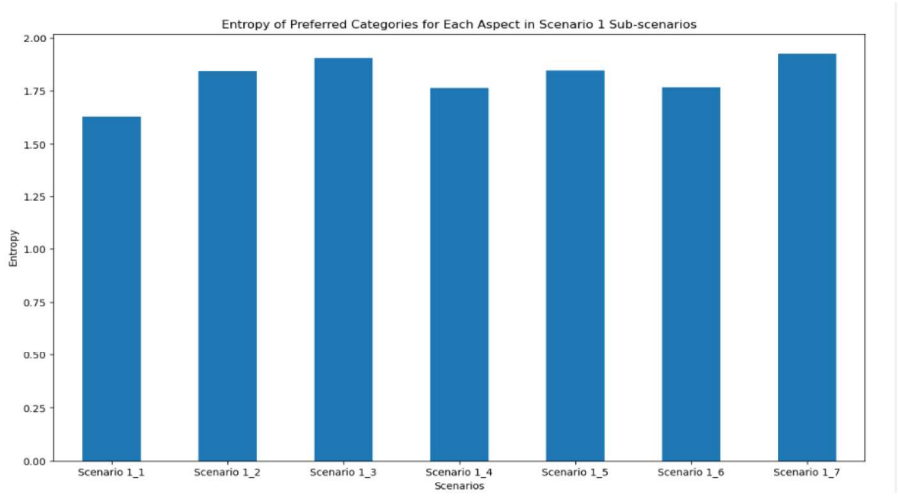


Figure 5 entropy of preferred aspects

Therefore, the conclusion drawn from this is that while different respondents do have varying preferences, these preferences are not completely random or scattered. The moderate entropy suggests that targeted strategies addressing the most common preferences could be effective, while

also acknowledging the need to maintain a level of flexibility to cater to the diversity of preferences among respondents.

To complement the entropy calculations, a hierarchical clustering of the sub-scenarios based on their preferences for aspects of care was conducted. The dendrogram generated from this analysis visually represented these clusters (Figure 6). The clustering analysis was pivotal in uncovering underlying factors driving the patterns in preferences, adding depth to our understanding of why owners use multiple vets.

As a result, some conclusions can be drawn:

- The owners using veterinarians are not a homogenous group but rather can be segmented into distinct clusters based on their preferences for different aspects of care.
- Understanding these clusters can provide valuable insights for veterinary practices. It can inform more targeted service offerings, communication strategies, and overall customer management approaches to cater to the needs of different owner clusters effectively.
- The clustering also underscores the complex and multifaceted nature of decision-making in veterinary care, further emphasizing the importance of personalized, flexible service offerings in this sector.

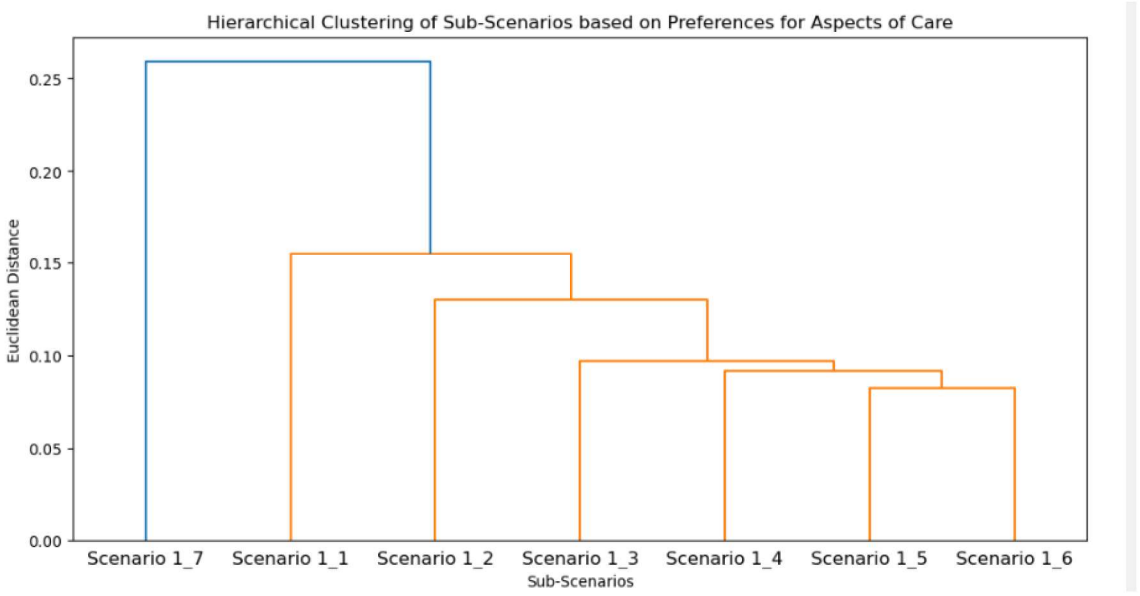


Figure 6 hierarchical clustering (dendrogram)

While we considered partitioning the data into additional clusters using the elbow method, the analysis indicated a linear structure in the sum of squared errors (SSE) reduction as the number of clusters increased. This suggested that additional clusters did not substantially improve our understanding of the variability within the dataset (Figure 7). This piece of information is important

to the analysis methodology, as it guided us to stick with the initial clustering.

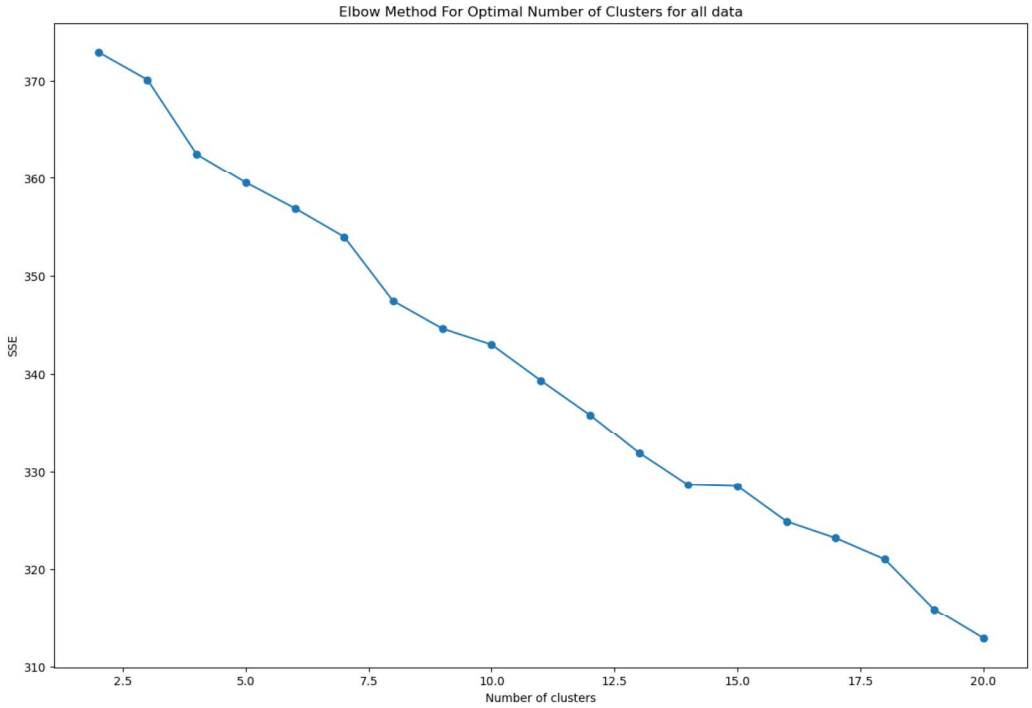


Figure 7 elbow method optimal clusters

For easy visual analysis a heatmap is created (figure 8). It is apparent that respondents placed a higher importance on the quality of care compared to the cost of service. For clarification, in this heatmap, 1_1 denotes the most important aspect of care according to the respondents, whereas 1_7 indicates the least important aspect of care.

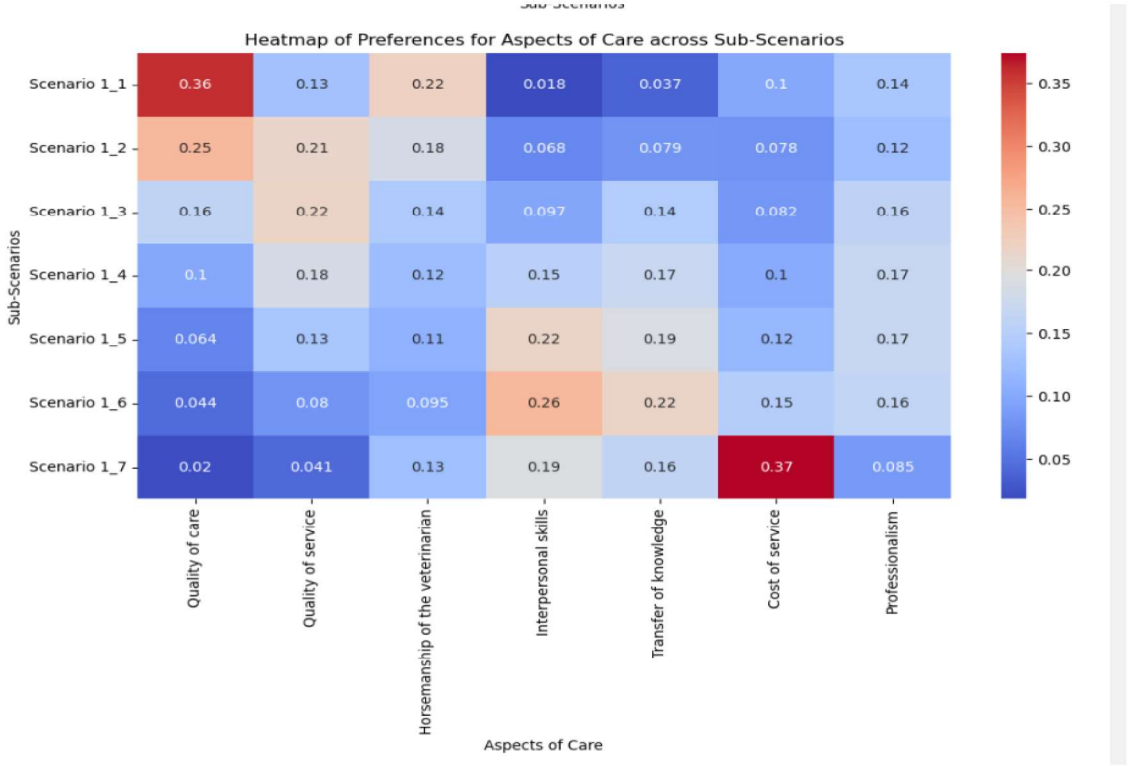


Figure 8 heatmap of preferences for aspects of care

For further analysis and to dive deeper into understanding why horse owners use multiple vets, the purpose of keeping horses and whether the respondents' kept horses for professional reasons was analyzed. Cross-tabulation of these data points, represented visually via a heatmap, allowed us to see any connection between these variables (Figure 9).

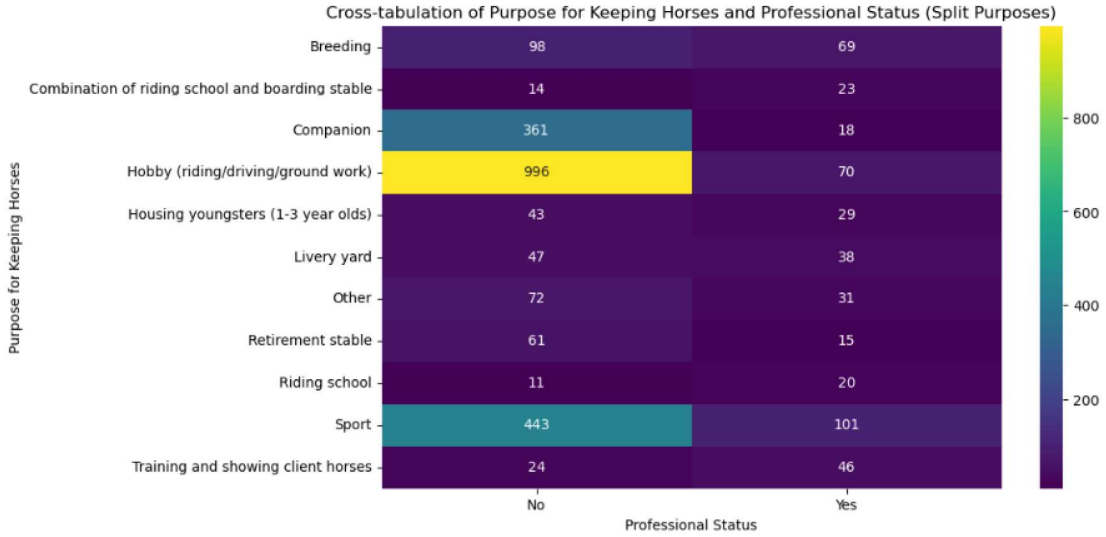


Figure 9 cross tabulation heatmap

Through this heatmap, the relationship between the respondents' purpose for keeping horses and their professional status is shown. The colour variations demonstrated that most respondents are non-professional horse keepers, maintaining horses for companionship, hobbies, and sports.

In summary the exploratory analysis revealed distinct variations in horse owners' preferences for different aspects of equine care. For instance, owners consistently ranked 'quality of care' highly, reflecting the importance of veterinarians' expertise and proficiency. In contrast, 'cost of service' showed more variability, suggesting differing levels of financial considerations among owners.

Additionally, we noticed a clear link between the purpose of horse ownership and the use of multiple veterinary services. Professional horse owners or trainers, who often manage a more significant number of horses with varying needs, were more likely to engage multiple vets. This pattern could be attributed to the necessity of specialized care for distinct horse breeds or competitive requirements.

6. Results (in depth analysis)

To conduct a more robust analysis, selected columns are kept for a in depth investigation. This predominantly focuses on the number of veterinarians used. Instances where only one vet is utilized are excluded as they don't contribute to the in-depth analysis due to the lack of filled reasons.

The dataset is subsequently divided into subgroups with two veterinarians along with their associated reasons, and those with three or more vets and the corresponding reasons. A column combining both these categories is also maintained for benchmarking. Much of the thorough investigation is concentrated on these columns. In addition, the following columns are kept for additional analysis:

'Number of horses (including ponies, foals, etc) under your care? (Numeric response only)' - referred to as 'number of horses'.

'What is the main thing you appreciate in your veterinarian?' - referred to as 'appreciation'.

'How long have you been using the services of your current, primary, veterinary practice?' - referred to as 'service duration'.

In this analysis, a variety of methods including topic modeling, TF-IDF/count vectorizer analysis, text similarity measures, N-grams, and decision tree classifier are utilized to gain insights into client perspectives and priorities in equine veterinary care. The investigation involves comparisons between groups based on different categorizations, thereby revealing both similarities and disparities that can provide useful guidance for tailored communication and service strategies in veterinary practices.

6.1 Topic modeling (LDA), count vectorizer and TF-IDF

The analysis began by examining the key aspects of care emerging from the discussions on veterinary services. We used two term weighting methods, count vectorizer and TF-IDF, to process the dataset before conducting topic modeling. It's essential to note that count vectorizer and TF-IDF are not themselves topic modeling algorithms but are used to prepare the text data, helping to highlight the significance of certain words or terms within the documents.

Once the data was appropriately weighted and processed, we applied Latent Dirichlet Allocation (LDA) as our topic modeling technique. This enabled us to group the discussions into various topics based on word occurrence patterns, helping to uncover the underlying themes.

Summary of Findings:

Top 5 words per topic for the group all TF-IDF method	
Topic:	Top 5 words in the topic:
0	Letterlijk, verspreid, eigen, routine, practice
1	Nodig, care, probleem, holistische, maakt
2	Medicine, home, tanden, wolvega, vriendin
3	Enten, basis, dental, come, uithof
4	Chiro, specialistische, weg, service, buurt
5	Specialistische, algemene, stal, zorg, procedure
6	Weekenddiensten, specifiek, koliek, reproductive, work

7	Issue, sportpaarden, regular, entingen, kreupelheden
8	Large, 2e, others, tandarts, avond
9	Familie, vaccinatie, meerdere, enten, onderzoek

Table 1 TF-IDF top 5 words per topic

For a complete breakdown of the top words across all topics (TF-IDF and count vectorizer) and groups, see appendix A.

The count vectorizer model identified ten primary topics, including emergency veterinary care, costs, orthopedic issues, dental care, and local veterinary services. On the other hand, the TF-IDF model brought forward themes such as holistic care, routine practices, cost-efficient services, and availability of services during off-hours.

Certain themes were consistent across both models, underlining their significance. These shared aspects of care included emergency services, specialized services, routine and basic care, dental care, orthopedic care, local or community-based services, and referral services. Additionally, some topics were unique to each model. For example, the count vectorizer model emphasized complex cases, while the TF-IDF model underscored the importance of holistic care and off-hours availability.

As the groups 2 and 3(+) were not that different for both models we are focusing on the combined group for further analysis. The table below interpreters the topics manually, which could mean the interpretations are biased and subjective.

Group all	
Count Vectorizer:	**TF-IDF:**
Topic 0: This topic seems to focus on the idea of emergency veterinary care. Words such as "spoed" (urgency), "hospital", "vaccination" and "clinic" are prevalent. **Topic 1**: Appears to address costs, orthopedic issues, and holistic care ("kosten", "orthopedie", "holistisch"). **Topic 2**: This topic could be about dental care and rural veterinary services ("tanden", "rural", "spoedeisende"). **Topic 3**: It looks like this topic is about specialized services, possible issues related to horse colic and expertise ("specialist", "koliek", "expertise"). **Topic 4**: The focus here seems to be on specialized and mobile services, and x-ray examinations ("specialistische", "mobile", "rontgenfotos"). **Topic 5**: The topic discusses the role of specialists in basic care and potential referrals ("specialist", "basis", "doorverwijzing"). **Topic 6**: Appears to deal with larger, complex cases, and the idea of having a local veterinarian ("grotere", "ingewikkelde", "lokale"). **Topic 7**: This topic seems to be about routine care for sport horses and dental care ("sportpaarden", "routine", "gebit"). **Topic 8**: The emphasis here is on routine vaccinations, different types of care, and	**Topic 0**: This topic seems to discuss routine practices and specialist mobile services ("routine", "praktijken", "mobile"). **Topic 1**: Seems to emphasize holistic care, orthopedics, and community-based services ("holistische", "orthopedie", "buurt"). **Topic 2**: This topic might revolve around cost-efficient services and emergency first-line and second-line care ("goedkoper", "eerstelijns", "tweedelijns"). **Topic 3**: Looks at dental care, specialist services, and choice of services ("enten", "specialistische", "keuze"). **Topic 4**: It discusses chiropractic care and emergency services ("chiro", "emergency"). **Topic 5**: The topic revolves around general care, specialized work, and referral services ("algemene", "gespecialiseerd", "doorverwijzing"). **Topic 6**: The emphasis here is on specialized weekend services, orthopedic care, and home-based services ("weekenddiensten", "orthopeed", "thuis"). **Topic 7**: This topic addresses issues with sports horses, orthopedic care, and general lameness ("sportpaarden", "orthopedisch", "lameness").

emergency situations ("vaccination", "different", "emergency"). **Topic 9**: Seems to be about multiple specialists, vaccinations, and normal services offered by local clinics ("meerdere", "inentingen", "normale").	**Topic 8**: Focuses on large-scale care, dental services, and availability during off-hours ("large", "gebit", "bereikbaarheid"). **Topic 9**: This topic seems to be about vaccinations, dental care, and consultation services ("vaccinatie", "raadplegen").
Shared Aspects (Topics that appear in both models): - Emergency care - Specialized services - Routine and basic care - Dental care - Orthopedic care - Local/community-based services - Referral services	
Distinct Aspects (Topics that are unique to each model): Count Vectorizer: Emphasis on costs, larger complex cases TF-IDF: Holistic care, availability during off-hours, family care	

Table 2 comparisons count vectorizer and TF-IDF

Further analysis of these themes allowed to categorize them into seven key aspects that impact client satisfaction in equine veterinary care. These were quality of care, quality of service, horsemanship, cost of service, interpersonal skills, transfer of knowledge, and professionalism.

Topic modeling vs different aspects of care:

Quality of care:

This relates to possessing the necessary knowledge and skills for diagnosing and treating equine health conditions. The use of terms such as 'specialist', 'specialized', 'orthopedics', 'dental', 'lameness', 'reproductive', 'emergency', 'vaccination', 'routine', and 'regular' across various topics indicates the range of health issues that horses can present with and the diversity of services that equine vets offer. The repeated occurrence of these terms suggests the importance of specialized care in areas like dental health, lameness, and emergency response, which directly affects client satisfaction.

Quality of service:

This covers the logistical and administrative aspects of the practice. Terms like 'availability', 'location', 'emergency', 'weekend service', and 'practice' highlight the importance clients place on aspects such as flexible scheduling, accessibility, and quick response times.

Horsemanship:

A good understanding of equine behavior is important in handling horses. However, this factor isn't directly highlighted in the topic modeling results. There might be implicit connections in the way terms related to care and specialized service are discussed.

Cost of service:

The term 'goedkoper' (cheaper) occurs in the topics, indicating a consideration for cost-effective alternatives. Terms related to costs are not so common which is also the least relevant in aspects of care stated by the respondents.

Interpersonal skills:

These are crucial for effective veterinary practice. Words like 'communication', 'care', 'consult', and 'question' are suggestive of the emphasis clients put on good communication with their vets.

Transfer of knowledge:

The ability to educate clients about their horse's condition is a key factor. While not explicitly stated in the topic modeling results, the occurrence of terms related to specific treatments suggests that clients are informed and understand the various aspects of care their horses receive.

Professionalism:

Words such as 'honesty', 'respect', and 'transparency' are not directly mentioned, but concepts related to good communication and specialized care hint at the importance of professionalism in veterinary practice.

Each of these elements, while distinct, contribute to a holistic understanding of the respondents' experiences and expectations of equine veterinary care. They offer insights into the wide range of services that equine vets offer and the diversity of health issues that horses can present with, revealing a demand for specialized care, flexible scheduling, and strong interpersonal skills among veterinarians.

Summary of topic modellng analysis:

The results highlight the multifaceted nature of client satisfaction in the context of equine veterinary care. High-quality care and service, cost considerations, and the veterinarian's expertise, interpersonal skills, and communication abilities are all critical elements.

The topic modeling results, though they offer a simplified view of complex human attitudes and perceptions, provide invaluable insights that can guide the improvement of veterinary services to better meet client needs and expectations. They also reveal areas that may require more specific exploration in future studies, such as the cost of service, horsemanship, and interpersonal skills.

However, it's important to bear in mind that topic modeling is an exploratory technique and not a definitive representation of these factors. The insights derived from this analysis should be complemented with other methods of research for a more comprehensive understanding.

6.2 TF-IDF top features, reasons for usage of multiple vets, and most appreciated aspects of the primary vet

This analysis uses term frequency-inverse document frequency (TF-IDF) for single words instead of topics. This is a reliable analytical tool to investigate the reasons clients utilize multiple veterinarians and what qualities they appreciate most in their primary vets. The data is categorized into three groups again - group '2 vets', group '3(+) vets', and an 'all' group. Each group exhibits both unique and shared characteristics, providing a nuanced understanding of client preferences.

The TF-IDF score of a word in a document is usually proportional to its frequency in the document and inversely proportional to its frequency in the entire corpus. A high TF-IDF score for a word means that the word is significant in that document, which might not be the case in other documents. The score can range between 0 and 1, where a higher score means that the word is more important in distinguishing the document from others.

Top features reasons for usage of multiple vets for Group 2 vets	**Top features most appreciated at primary vet for group 2 vets**	**Top features reasons for usage of multiple vets for Group 3(+) vets**	**Top features most appreciated at primary vet for group 3(+) vets.**	**Top features reasons for usage of multiple vets for group all**	**Top features most appreciated at primary vet for group all.**
Tfidf score:	Tfidf score:	Tfidf score:	Tfidf score:	Tfidf score:	Tfidf score:
0 tandarts 0.028991	0 eerlijkheid 0.048428	0 afhankelijk 0.031051	0 eerlijkheid 0.036532	0 tandarts 0.027067	0 eerlijkheid 0.043677
1 emergency 0.021407	1 kennis 0.029897	1 spoed 0.026770	1 kennis 0.032717	1 emergency 0.020623	1 kennis 0.029475
2 eigen 0.020431	2 eerlijk 0.026177	2 tandarts 0.025334	2 eerlijk 0.024341	2 eigen 0.018288	2 eerlijk 0.024709
3 specialist 0.020061	3 meedenken 0.021557	3 lameness 0.024476	3 snel 0.021241	3 specialist 0.017014	3 meedenken 0.020810
4 care 0.018138	4 kundig 0.020348	4 emergency 0.021704	4 duidelijkheid 0.021209	4 lameness 0.016997	4 kundig 0.018116
5 zorg 0.017598	5 kunde 0.020047	5 practice 0.021300	5 meedenken 0.021137	5 vaste 0.016289	5 kunde 0.017875
6 vaste 0.016465	6 vriendelijk 0.017134	6 local 0.020227	6 communicatie 0.021120	6 afhankelijk 0.015919	6 duidelijkheid 0.017521
7 gespecialiseerde 0.016246	7 duidelijkheid 0.016336	7 different 0.019303	7 bereikbaar 0.020050	7 spoed 0.015772	7 snel 0.017287
8 reguliere 0.015724	8 snel 0.016319	8 vaste 0.018185	8 bereikbaarheid 0.019613	8 zorg 0.015206	8 communicatie 0.016588
9 nodig 0.015386	9 overleg 0.015598	9 buurt 0.018173	9 duidelijk 0.018724	9 gespecialiseerde 0.015188	9 open 0.014295
10 enten 0.015232	10 communicatie 0.015416	10 general 0.016792	10 good 0.018453	10 enten 0.015187	10 care 0.014103
11 huis 0.015002	11 open 0.014524	11 entingen 0.016751	11 care 0.018270	11 different 0.015066	11 vriendelijk 0.014089
12 lameness 0.014856	12 mee 0.013959	12 enten 0.016390	12 uitleg 0.017347	12 care 0.014967	12 overleg 0.012843
13 different 0.013904	13 care 0.013325	13 issue 0.016299	13 nodig 0.016888	13 nodig 0.014457	13 kundigheid 0.012819
14 stal 0.012641	14 kundigheid 0.013195	14 eigen 0.015578	14 kundig 0.015385	14 practice 0.014177	14 mee 0.012694
15 zaken 0.012569	15 tijd 0.012009	15 gespecialiseerde 0.014588	15 open 0.015042	15 huis 0.013649	15 bereikbaar 0.012228
16 practice 0.011997	16 duidelijk 0.011346	16 expertise 0.013929	16 communication 0.013762	16 reguliere 0.013308	16 uitleg 0.011992
17 spoed 0.011516	17 knowledge 0.011326	17 specialisatie 0.013888	17 expertise 0.013610	17 buurt 0.012245	17 good 0.011752
18 regular 0.011066		18 nodig 0.013725	18 kunde 0.013442	18 stal 0.012110	18 duidelijke 0.011669
19 eerste 0.010656		19 probleem 0.013684		19 local 0.011793	19 duidelijk 0.011441
20 kleine 0.010498				20 entingen 0.011698	

21 entingen 0.010353 22 available 0.010294 23 buurt 0.010286 24 routine 0.010060	18 knowledgeable 0.011230 19 uitleg 0.009936 20 contact 0.009823 21 luisteren 0.009782 22 cost 0.009532 23 good 0.009530 24 betrokkenheid 0.009519	20 klinieken 0.013621 21 location 0.013354 22 klacht 0.012307 23 stal 0.012051 24 tanden 0.011888	19 duidelijke 0.013081 20 kundigheid 0.012866 21 ervaring 0.011876 22 skill 0.011810 23 extremely 0.011752 24 contact 0.011548	21 general 0.010795 22 kleine 0.009808 23 onderzoek 0.009702 24 specialistische 0.009528	20 tijd 0.011271 21 bereikbaarheid 0.011198 22 knowledge 0.010663 23 nodig 0.010418 24 communication 0.010140

Table 3 top features TF-IDF

Considering the reasons for using multiple vets, the analysis shows both group '2' and '3+' clients value the quality of care and professional expertise. Common features across these groups emphasize the importance of dental care, the availability of emergency services, and specialized veterinary services. The term 'buurt' (neighborhood) emerges as significant, suggesting that both groups appreciate having veterinary services close to their residence.

A deeper look into each group reveals distinct preferences. Group '2' clients seem to value the continuity and personalization of care. This is evident from the high TF-IDF scores for 'eigen' (own) and 'vaste' (regular), signifying the desire for dedicated and consistent veterinary attention. In contrast, group '3+' appears more issue-focused, emphasizing the resolution of specific problems such as 'lameness'.

Analyzing the 'all' group can provide a benchmark or reference point for evaluating the performance and characteristics of individual groups. It allows us to assess whether certain features or topics are more prominent or distinctive within specific groups compared to the overall data. The table below presents the most prominent shared and distinctive features for the groups.

Shared features:	Distinct features:
Tandarts (Dentist): Both groups prioritize dental care for their animals, indicated by the high TF-IDF scores for "tandarts" and "tanden" in group2 and group 3, respectively. **Emergency:** The term "emergency" is important for both groups, highlighting the need for immediate veterinary care in critical situations. **Gespecialiseerde (Specialized)**: Both groups value specialized veterinary services, as indicated by the presence of "gespecialiseerde" and "specialisatie" in group 2 and group 3, respectively. **Buurt (Neighbourhood):** The concept of proximity and having a veterinary practice nearby is important for both groups, as seen in the high TF-IDF scores for "buurt/local" in both groups.	**Eigen (Own):** Group 2 emphasizes the importance of personalized care and having their own dedicated veterinarian, as indicated by the high TF-IDF score for "eigen." **Zorg (Care):** Group 2 values comprehensive care and attention for their animals, as indicated by the high TF-IDF score for "zorg." **Vaste (Regular)**: Group 2 expresses a preference for having a regular veterinarian or veterinary practice, as suggested by the high TF-IDF score for "vaste." **Lameness:** Group 3 shows a specific concern for lameness issues in animals, as indicated by the high TF-IDF score for "lameness." **Dependent (Afhankelijk):** Group 3 has as a top TF-IDF score “afhankelijk” which could mean the availability of certain veterinaries is not always optimal. (Which could be due to living in rural areas)

Table 4 shared and distinct features

The table below presents a comparison between group 2, group 3(+), and the overall group regarding the seven aspects of veterinary service: quality of care, quality of service, professionalism, interpersonal skills, horsemanship, cost of service, and transfer of knowledge. It examines these aspects based on the reasons horse owners use multiple vets and what they appreciate most in their primary vet according to the TF-IDF top features analysis.

Table looking at the seven aspects in relation to the features:			
	Group 2 vets:	Group 3(+) vets:	Group all:
Reasons:	Reasons for using multiple vets for group 2 include the need for specific care like 'tandarts' (dentist), the importance of availability during emergencies, and the consistency of 'vaste' (regular) veterinary care. This indicates the value group 2 clients put on the quality of care, quality of service, and professionalism of their vets.	Reasons for using multiple vets for group 3 again include the need for 'tandarts' (dentist) and urgency ('spoed'), along with the term 'afhankelijk' (dependent), indicating that group 3 clients highly value quality of care and quality of service	Across all groups, the reasons for using multiple vets centre around quality of care ('tandarts', 'specialist', 'gespecialiseerde') and quality of service ('emergency', 'zorg').
Most appreciated:	For what they appreciate in their primary vet, group 2 highlights 'eerlijkheid' (honesty), 'kennis' (knowledge), and 'kundig' (competency), demonstrating appreciation for quality of care, the transfer of knowledge, and professionalism. The term 'meedenken' (thinking along) suggests appreciation for interpersonal skills as well.	When it comes to what is appreciated about their primary vet, group 3 shares group 2's appreciation for 'eerlijkheid' (honesty) and 'kennis' (knowledge). The term 'snel' (quick) suggests group 3 clients also highly value the quality of service they receive.	For what is appreciated about their primary vet across all groups, the top features again align with quality of care ('eerlijkheid', 'kennis', 'kundig'), transfer of knowledge ('meedenken'), and interpersonal skills ('communicatie').

Table 5 relation to aspects of care

The aspects of horsemanship of the veterinarian and cost of service are not directly captured by the TF-IDF results, though 'afhankelijk' (dependent) from group 3 might relate to cost considerations. It's also possible that aspects of horsemanship are embedded in terms related to quality of care, as an understanding of equine behaviour would be necessary to provide effective treatments.

These results could suggest areas of focus for improving veterinary services based on what horse owners value most. By emphasizing aspects such as quality of care, quality of service, professionalism, and interpersonal skills, veterinary practices can better meet the needs and expectations of their clients.

This data would capture what is analyzed in the EDA where horse owners stated in scenario 1 that cost of service is not important to the and quality of care the most important.

Summary of TF-IDF top features analysis:

The text uses TF-IDF, a robust analytical tool, to delve into the reasons why multiple vets are used and what aspects are most appreciated in a primary vet, from both a data science and veterinary perspective. The results are split across three groups, labeled '2', '3(+)', and 'all', with each group having distinct and shared features.

In terms of reasons for using multiple vets, both groups '2' and '3(+)' appreciate the quality of care and expertise. Shared characteristics across both groups underscore the importance of dental care, emergency services, and specialized veterinary services. The term 'buurt' (neighborhood), which highlights the importance of service proximity, also stands out. However, while group '2' values personalized care and continuity, group '3' is more inclined towards addressing specific issues like lameness and resolving problems.

On the other hand, what is most appreciated about a primary veterinary seems to center on qualities such as honesty, knowledge, and competency across all groups, suggesting that these traits are universally valued in a primary veterinarian. The term 'meedenken' (thinking along) in groups '2' and 'all' hints at the importance of interpersonal skills, while the term 'snel' (quick) in group '3' underlines the significance of quality of service.

In conclusion, the results reveal a layered understanding of client needs and preferences in equine veterinary care. The findings can help veterinarians tailor their services to meet the specific requirements of different client groups while maintaining universally valued qualities such as honesty, knowledge, and competency.

6.3 Similarity scores

The Jaccard index is a metric used to measure the similarity between two sets, in this case, the vocabulary used by the 'all' group, group 2, and group 3(+). The Jaccard index between the 'all' group and group 2 is approximately 0.70, indicating that roughly 70% of the total unique words in the 'all' group and group 2 overlap. Similarly, the index is approximately 0.54 between the 'all' group and group 3, suggesting that about 54% of their total unique words are common. These significant indices highlight a considerable vocabulary similarity between the 'all' group and the other two groups.

The Jaccard index between group 2 and group 3 is approximately 0.24, which implies that only about 24% of their unique words are shared. This comparatively lower score suggests that there might be differences in the language usage or discussed topics between group 2 and group 3.

Jaccard index between 'all' group and group2:	0.7017364657814096
Jaccard index between 'all' group and group3(+):	0.5444330949948928
Jaccard index between group2 and group3(+):	0.24616956077630234

Table 6 Jaccard similarities

While the Jaccard index provides a helpful measure of overall vocabulary overlap, it doesn't account for the frequency of words. This could potentially influence the interpretation of these similarity scores. A word frequency-based analysis could further clarify the similarity and differences in language use between the groups.

6.4 N-grams

In the following analysis, three key text mining techniques are applied - unigrams, bigrams, and trigrams - to extract patterns and themes from the textual responses of two distinct groups. These techniques help break down the text into individual words (unigrams), two-word phrases (bigrams), and three-word phrases (trigrams), which then allows us to explore the frequency of these entities within the dataset. All the n-grams are based upon the column “reasons” for using multiple veterinarians.

Unigram analysis:
The unigram analysis reveals the most frequent individual words across the responses. This gives us a broad sense of the recurring themes or topics that respondents frequently discuss.

Bigram analysis:
In the context of natural language processing, a bigram is a sequence of two adjacent elements from a string of tokens, typically letters, syllables, or words. Bigrams can provide important contextual information by capturing the dependencies between successive words.
Moving beyond individual words, bigrams help us understand the context better by looking at pairs of words that commonly occur together. This can reveal important relationships between concepts and highlight key phrases that often appear in the discussion.

Trigram analysis:
Trigrams offer an even more nuanced view by examining three-word phrases. Although they can be less frequent than unigrams and bigrams, trigrams can provide valuable insights into more complex ideas or specific themes discussed by respondents.

However, the effectiveness of bigram and trigram analyses depends on the amount and diversity of text data. In cases where we have shorter text responses or the diversity of the responses is limited, these analyses may yield less informative results, as phrases tend to appear with low frequency.

Through these techniques, the analysis tries to derive meaningful patterns and insights that can help us understand the opinions, experiences, and concerns of the two groups.

6.4.1 Unigram analysis

Top 20 unigrams for group 2 vets:			unigrams for group 3(+) vets:		
	unigram	frequency		unigram	frequency
0	emergency	22	0	spoed	15
1	specialist	19	1	local	14
2	care	19	2	lameness	13
3	zorg	17	3	practice	12
4	eigen	17	4	emergency	11
5	tandarts	16	5	tandarts	11
6	gespecialiseerde	15	6	general	9
7	enten	14	7	vaste	8
8	zaken	14	8	issue	8
9	vaste	14	9	entingen	8
10	lameness	14	10	afhankelijk	8
11	practice	13	11	buurt	8
12	nodig	12	12	different	8
13	spoed	11	13	routine	6
14	stal	11	14	gespecialiseerde	6
15	huis	10	15	needed	6
16	onderzoek	10	16	care	6
17	kleine	10	17	enten	6
18	reguliere	10	18	eigen	6
19	different	10	19	dental	5

Table 7 unigram analysis

The unigram analysis underscored a significant focus on themes related to veterinary practices, emergency services, and specific medical conditions among both groups. Terms like 'emergency', 'specialist', 'zorg' (care), 'lameness', 'spoed' (urgency), 'tandarts' (dentist), and 'enten' (vaccination) frequently appeared, shedding light on the primary areas of concern. Notably, group 2 showed a higher frequency of the term 'emergency' than group 3, indicating a possible greater emphasis on emergency services. 'Zorg', or care, also appeared more frequently in group 2, signifying the group's stronger focus on overall animal healthcare.

While the exact interpretation can depend on the context in which these terms are used, the unigram analysis gives a broad sense of the key concerns and topics discussed in both groups.

6.4.2 Bigram analysis

Top 20 bigrams for group 2 vets:			bigrams for group 3(+) vets:		
	bigram	frequency		bigram	frequency
0	routine care	4	0	mile away	4
1	maak gebruik	4	1	orthopedische problemen	3
2	sport medicine	4	2	local practice	3
3	specialistische zorg	3	3	sport medicine	2
4	general care	3	4	local basic	2
5	lameness specialist	3	5	issue lameness	2
6	care emergency	3	6	spoed kleine	2
7	emergency regular	2	7	afhankelijk klacht	2
8	lokale enten	2	8	general practice	2
9	enten simpele	2	9	depends issue	2
10	kleine inenten	2	10	xray machine	2
11	vaste veearts	2	11	veterinary practice	2
12	simpele zaken	2	12	routine care	2
13	emergency call	2	13	holistische beeld	1
14	diagnose behandeling	2	14	beeld diagnostiek	1
15	vaste tandarts	2	15	diagnostiek operatief	1
16	huis sportpaarden	2	16	operatief ingrijpen	1
17	rontgen echo	2	17	ingrijpen reguliere	1
18	eerste tweedelijns	2	18	chiropractic dental	1
19	eerste lijn	2	19	dental eastern	1

Table 8 bigram analysis

The top bigrams for group 2 underline the recurrent themes of care, particularly 'routine care', 'general care', and 'emergency care'. The presence of 'specialistische zorg' (specialized care) and 'lameness specialist' reinforces the importance of specialized service in the satisfaction of group 2 clients. Additionally, terms like 'maak gebruik' (make use), 'lokale enten' (local vaccinations), 'kleine inenten' (small vaccinations), 'simpele zaken' (simple matters) emphasize a preference for localized and simple veterinary services.

The top bigrams for group 3 echo some of these themes, specifically with 'routine care', 'local practice', and 'general practice'. A notable mention is 'mile away', indicating a potential concern about the location or accessibility of the veterinary service. Another theme that emerges is the focus on orthopedics, as illustrated by 'orthopedische problemen' (orthopedic problems). 'Holistische beeld' (holistic image) and 'chiropractic dental' further highlight the interest in holistic and specialized care.

6.4.3 Trigrams analysis

Top 20 trigrams for group 2 vets:		
	trigram	frequency
0	praktische redenen vraag	2
1	buurt huis spoed	1
2	huis spoed minder	1
3	spoed minder spannende	1
4	minder spannende specialisten	1
5	tandverzorging enten zorg	1
6	locatie regulier werk	1
7	regulier werk enten	1
8	werk enten kloris	1
9	enten kloris spoed	1
10	kloris spoed orthopedisch	1
11	spoed orthopedisch onderzoek	1
12	zorg nauwverwant samenwerkt	1
13	nauwverwant samenwerkt tandarts	1
14	samenwerkt tandarts sederen	1
15	dierenartspratijk spoed weekenddienst	1
16	spoed weekenddienst praktijken	1
17	hoofddierenarts tandartsdierenarts reserve	1
18	huisdierenarts grotere huisdierenarts	1
19	grotere huisdierenarts apparatuurruimte	1

Table 9 trigram analysis group 2

Top 20 trigrams for group 3(+) vets:		
	trigram	frequency
0	holistische beeld diagnostiek	1
1	beeld diagnostiek operatief	1
2	diagnostiek operatief ingrijpen	1
3	operatief ingrijpen reguliere	1
4	chiropractic dental eastern	1
5	dental eastern medicine	1
6	eastern medicine general	1
7	medicine general service	1
8	general service emergency	1
9	service emergency call	1
10	emergency call sport	1
11	call sport medicine	1
12	huis klinieken nodig	1
13	gebit enting vaste	1
14	regio routine werk	1
15	routine werk vaccinatiesmestonderzoek	1
16	werk vaccinatiesmestonderzoek wondverzorgimg	1
17	vaccinatiesmestonderzoek wondverzorgimg kreupe...	1
18	wondverzorgimg kreupelsomderzoek gebitsbehande...	1
19	drachtigheid ziekten keuringen	1

Table 10 trigram analysis group 3

Nearly all the trigrams appear only once, which makes this analysis less useful for identifying common themes or topics. The value of trigram analysis typically comes from finding recurring patterns in the text. It helps to identify specific phrases or language patterns that are commonly used in each context, providing more detail than individual words or bigrams.

In this case, the trigrams are unique and don't recur. This could be due to several reasons:

Specificity: The specificity of trigrams makes them less likely to appear frequently. While single words or bigrams might recur often, it's less common for the exact same three-word sequence to appear multiple times, unless there are standard phrases or language patterns in the text.

Data volume: If the volume of data is small, the chances of trigrams reappearing decrease significantly. The volume of data of the data is small compared to “big” datasets.

Language variation: The text is written by different authors and the language used might be varied, leading to a wide range of different trigrams.

Inter-group comparison:

When comparing the top unigrams and bigrams between the two groups, some shared themes appear, but the frequency and order of their appearance suggest different areas of emphasis. For instance, the term 'emergency' shows up more frequently in group 2, suggesting that this group may place a greater emphasis on emergency veterinary services. The phrase 'mile away' being a top bigram for group 3 suggests that accessibility or proximity of veterinary services may be a more significant concern for this group. The comparative analysis suggests that while both groups share some common concerns and preferences, the importance they attach to each issue may vary.

Trigram analysis, despite being less informative due to data limitations, provides a more detailed context in which words are being used. However, the comparison of trigrams between the two groups doesn't yield substantial patterns due to their uniqueness.

Summary:

The n-gram analysis illuminated key themes, areas of concern, and preferences among the two groups, contributing valuable insights for understanding their perspectives. The comparative analysis between group 2 and group 3 allowed us to identify not only the common concerns but also the distinct priorities and areas of emphasis for each group.

While the trigram analysis did not yield substantial recurring patterns due to the uniqueness of most trigrams in the data, the unigram and bigram analyses revealed important insights into the concerns and preferences of both groups.

From these findings, veterinary practices could derive actionable insights: the importance of emergency services for group 2 indicates the need for readily available and efficient emergency care, while the emphasis on proximity for group 3 implies the importance of accessible veterinary services. Tailoring services and communication to address these preferences and concerns could enhance client satisfaction and loyalty among these groups.

The specificity and diversity in our dataset limited the insights gained from the trigram analysis, suggesting that unigram and bigram analyses may be more effective for this dataset. Future work could focus on gathering more data to conduct a more robust trigram analysis and continue to refine these insights.

6.5 Decision tree classifier

The decision tree classifier is a potent classification tool that utilizes a tree-like structure to make decisions (Loh, 2011). Each node symbolizes a feature, each branch signifies a decision rule, and each leaf denotes an outcome, be it categorical or continuous (Hastie, Tibshirani, & Friedman, 2001). The rules that govern the classifier are determined by the data, which can be visualized. The conditions on number of horses, number of vets and service duration (in years) set the rules. Each rule leads to a class represented by an array of samples, where the array's length indicates the number of topics. The values within the array represent the count of samples under that topic.

This tree-based method offers a hierarchical means to understand data, enabling visualization of the decision-making process and identifying key variables that impact predicted topics. Adjusting the tree's depth (3 for this model) can help avoid overfitting (capturing too much noise) or underfitting (not capturing enough patterns).

Per the LDA (TF-IDF) model, each document is deemed a blend of various topics. When an LDA model is fit to the data, it identifies the underlying topics based on word distribution across the documents and provides the distribution of topics across each document (Blei, Ng, & Jordan, 2003). By selecting the topic with the highest proportion for each document, you are essentially classifying each 'reason' into one of the topics, thus forming the classes for the decision tree (Ramage, Dumais, & Liebling, 2010).

The term "Gini" in the decision tree refers to the Gini impurity. It's a measure of impurity or disorder, used in the decision tree to decide which feature to split at each step in the tree (Raileanu & Stoffel, 2004). A Gini impurity of 0 indicates a perfectly pure node where all data points belong to a single class.

The Gini impurity of a node is calculated as the probability that a randomly selected element from the set would be incorrectly labeled if it was randomly labeled according to the distribution of labels in the subset. It is determined using the formula:

$$Gini = 1 - \Sigma(p\_i)\text{^}2$$

Where p_i is the probability of an element being classified to a specific class.

In the context of decision trees, the Gini impurity is used as a cost function to evaluate splits in the data. The tree algorithm chooses the split that results in the lowest Gini impurity. By choosing splits that minimize Gini impurity, the decision tree aims to enhance node homogeneity and improve its predictive accuracy. In this specific context, each split in the decision tree reduces the Gini impurity, creating more homogeneous nodes with respect to the distribution of topics derived from the 'reasons' text data. Each rule's conditions in the decision tree represent the condition that resulted in the most substantial reduction of Gini impurity at that decision tree building step.

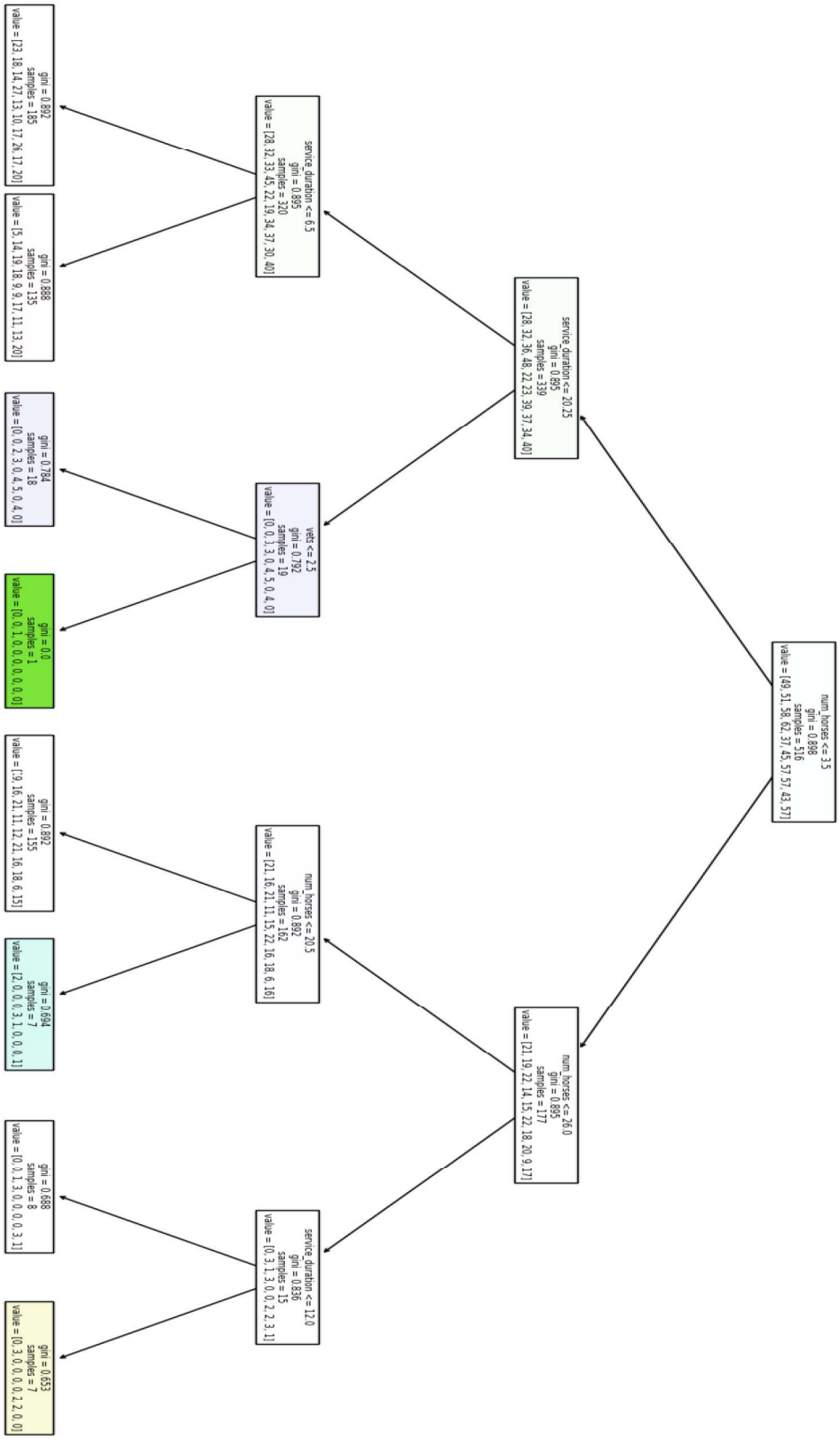


Figure 10 decision tree classifier

	Rule	Class	Samples
0	num_horses <= 3.5 & service_duration <= 20.25 & service_duration <= 6.5	[[23.0, 18.0, 14.0, 27.0, 13.0, 10.0, 17.0, 26.0, 17.0, 20.0]]	185
1	num_horses > 3.5 & num_horses <= 26.0 & num_horses <= 20.5	[[19.0, 16.0, 21.0, 11.0, 12.0, 21.0, 16.0, 18.0, 6.0, 15.0]]	155
2	num_horses <= 3.5 & service_duration <= 20.25 & service_duration > 6.5	[[5.0, 14.0, 19.0, 18.0, 9.0, 9.0, 17.0, 11.0, 13.0, 20.0]]	135
3	num_horses <= 3.5 & service_duration > 20.25 & vets <= 2.5	[[0.0, 0.0, 2.0, 3.0, 0.0, 4.0, 5.0, 0.0, 4.0, 0.0]]	18
4	num_horses > 3.5 & num_horses > 26.0 & service_duration <= 12.0	[[0.0, 0.0, 1.0, 3.0, 0.0, 0.0, 0.0, 0.0, 3.0, 1.0]]	8
5	num_horses > 3.5 & num_horses <= 26.0 & num_horses > 20.5	[[2.0, 0.0, 0.0, 0.0, 3.0, 1.0, 0.0, 0.0, 0.0, 1.0]]	7
6	num_horses > 3.5 & num_horses > 26.0 & service_duration > 12.0	[[0.0, 3.0, 0.0, 0.0, 0.0, 0.0, 2.0, 2.0, 0.0, 0.0]]	7
7	num_horses <= 3.5 & service_duration > 20.25 & vets > 2.5	[[0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]	1

Table 11 output decision tree classifier

To interpret the results, we can analyse the rules created by the decision tree as follows:

- **Rule 0:** Clients with fewer horses (≤ 3.5) and shorter service duration (≤ 6.5 years) are likely to discuss dental care, specialist services, sports horses' issues, and orthopedic care (T3 & T7 being prevalent topics).
- **Rule 1**: Clients with a moderate number of horses (>3.5 & ≤ 20.5) tend to prioritize cost-efficient services, emergency care, general care, specialized work, and referral services (T2 & T5 being the dominant topics).
- **Rule 2**: Clients with fewer horses (≤ 3.5) but longer service duration (> 6.5 years & ≤ 20.25 years) may focus on cost-effective services, emergency care, family vaccinations, and dental care (T2 & T9 being the main topics).
- **Rule 3**: With fewer horses (≤ 3.5), a longer service duration (> 20.25 years), and fewer vets (≤ 2.5), clients are likely to discuss general care, specialized work, referral services, large-scale care, dental services, and availability during off-hours (T5 & T8 being the prevalent topics).
- **Less representative rules, i.e., Rules 4-7**, cater to fewer samples and may not reflect broader trends but provide niche insights. For instance, with more horses involved but a shorter service duration, dental care, specialist services, large-scale care, and off-hours availability (T3 & T8) become focal points. Conversely, with more horses and longer service durations, holistic care, orthopedics, community-based services, specialized weekend services, orthopedic care, and home-based services (T1 & T6) come to the forefront.

Summary:

The combined use of decision tree and LDA model has yielded significant insights into the motivations and concerns of customers utilizing veterinary services. These machine learning models are particularly useful at handling the complexity and variability of natural language data, transforming unstructured customer 'reasons' into structured data points and topics.

Based on conditions: the number of horses owned, service duration, and the number of vets involved, the decision tree predicts likely topics derived from the LDA model. The distilled insights from the decision tree model include:

- Customers with fewer horses and shorter service durations often express concerns about dental care, specialist services, sports horses' issues, and orthopedic care.

- When dealing with a moderate number of horses, discussions around cost-efficient services, emergency care, general care, specialized work, and referral services become prominent.
- Clients with fewer horses but requiring a longer service duration seem to prioritize cost-effective services, emergency care, family vaccinations, and dental care.
- Customers with fewer horses, a longer service duration, and fewer vets are more likely to discuss general care, specialized work, referral services, large-scale care, dental services, and availability during off-hours.
- When more horses are involved, the customers may bring up topics such as dental care, specialist services, large-scale care, availability during off-hours, chiropractic care, emergency services, holistic care, orthopedics, community-based services, specialized weekend services, orthopedic care, and home-based services.

While these patterns provide a solid starting point, they are based on the specific data used for this analysis. It's always good practice to periodically re-evaluate the models with fresh data to confirm their continued validity. Even though the limited dataset size led to the use of the entire dataset for model creation, limiting its generalizability, future studies intending to shift from exploration to prediction should consider a train/test split to validate the model's performance. Nonetheless, this analysis underlines the usefulness of machine learning models in distilling meaningful insights from customer data, facilitating informed decision-making for business strategy and customer engagement.

7. Discussion

The research study delved into understanding the perspectives of horse owners about their utilization of multiple veterinarians and their appreciation for their primary vet. These findings offer considerable insights that might inform the design and delivery of veterinary services, potentially leading to increased satisfaction among horse owners and improved welfare outcomes for the horses.

The horse owners appreciated the quality of care, quality of service, professionalism, and interpersonal skills of their veterinarians, in alignment with earlier studies (Coe, Adams, & Bonnett, 2007; Elte et al., 2021). This emphasis reiterates the crucial role these aspects play in attracting and retaining clients in a veterinary practice.

Quality of care was especially significant for horse owners, possibly due to the complex and often specialized needs of horse care (Mcbride, Hemmings 2004). Veterinarians might want to consider continuous professional development and training in various aspects of equine care to enhance their service offerings and meet the needs of horse owners. Training in softer skills such as communication could also be beneficial, as our findings highlight the value horse owners place on 'honesty', 'knowledge', and 'meedenken' (thinking along) which involves collaborative communication and problem-solving.

The utilization of multiple veterinarians by horse owners may present challenges for maintaining continuity and consistency in care. However, it seems to be driven by practical needs such as availability during emergencies and requirement of specific care. It might be helpful for veterinary practices to explore collaborative models of care where veterinarians can work together to ensure comprehensive care for the horses. Also, strategies to ensure quick response in emergencies, such as on-call services, could enhance client satisfaction.

The results indicate that while there are shared features across all groups of horse owners, there are also distinct preferences and needs. This could be due to various factors such as the specific health needs of their horses, their past experiences, and their individual expectations from veterinary care. This highlights the importance of personalized care that considers the unique needs and preferences of each client (Batchelor & McKeegan, 2012). It might be useful for veterinary practices to adopt client-centered approaches that allow for the customization of services tailored to the clients' needs.

However, while client-centered care is crucial, it's also essential to balance it with the welfare needs of the animals. As Kidd and Kuznetsova (2020) discuss, veterinarian-client communication in challenging contexts like euthanasia decisions is a delicate balance of aligning with the client's needs and ensuring the best interests of the animal. This also aligns with the ethical considerations raised in the findings, where client demand should not lead to over-treatment or unnecessary procedures.

The analysis techniques used in this study, such as LDA, TF-IDF, n-grams and decision tree classifiers, allowed for in-depth analysis of textual data. However, they also come with limitations. The LDA, for example, assumes a Dirichlet prior over the topics which might not always hold true (Blei, Ng, & Jordan, 2003). TF-IDF, while effective in highlighting relevant words in a document, doesn't consider the context or semantics of the words (Ramos, 2003). Future research might want to consider the application of more advanced techniques like word embeddings and neural network models that can capture the semantics of words.

Lastly, while the research used anonymized data, ensuring data privacy and confidentiality remains a critical ethical responsibility. The increasing capabilities of machine learning techniques to extract detailed information underscore the importance of vigilant ethical practices to protect sensitive data (Pedregosa et al., 2011).

In conclusion, this research has provided valuable insights into horse owners' perspectives about their veterinarians, which could guide improvements in veterinary practice. However, it also highlights the need for balancing client needs with animal welfare considerations, and the importance of ethical research practices in a rapidly evolving data science landscape. Future research in this area could further explore the factors influencing client satisfaction and the strategies that veterinary practices can adopt to enhance their service delivery.

8. Conclusion

This research aimed to provide a deeper understanding of equine veterinary practices from the perspective of horse owners. The results provide insights into the decision-making process of horse owners when it comes to the selection of their (primary) veterinarians, and the evaluation of the veterinary services received.

Various data science techniques were used, such as LDA, TF-IDF, Jaccard similarity, n-gram analysis, and the decision tree classifier. These methods were used to analyze the textual responses of horse owners regarding their reasons for using multiple veterinarians and their most appreciated aspects of the primary vet.

The analysis revealed that the quality of care, expertise, and availability during emergencies were significant factors for horse owners in choosing multiple vets. These aspects were consistent across different groups of owners who use two vets, three or more vets and the group as a whole.

Moreover, when it comes to the most appreciated aspects of their primary vet, honesty, knowledge, and competency stood out across all groups, with specific emphasis on interpersonal skills and the quality of service. These findings suggest the universal importance of these attributes in a primary veterinarian from the perspective of horse owners.

The Jaccard similarity indices indicated significant overlap in terms used by different groups of horse owners, suggesting that despite the number of veterinarians they use, many of their concerns and appreciations are common.

However, it is also important to note that certain unique differences exist among the groups. The distinction, although not very large, might be due to specific needs or experiences that vary depending on the complexity of care required by their horses.

Ultimately, this research highlights the significance of quality care, expertise, honesty, and effective communication in equine veterinary practice. Understanding the aspects of care is important to improving veterinary services and aligning them with client expectations. Further qualitative research is needed to explore the nuances of these findings and how they can effectively be used in the veterinary healthcare field.

This study is not without limitations and constraints. The reliance on self-reported data, which could be subject to respondent bias and inaccuracies related to a lack of knowledge, is one of the major constraints of the dataset (MacKenzie, & Podsakoff, 2012). Moreover, the use of text analysis methods, while robust, may miss nuances in communication that a direct interview or qualitative analysis might capture (Abbasi, Chen, & Salem, 2008). Furthermore, the findings may not be objective for all horse owners as the study participants are potentially not a representative sample of the whole horse owner's population and might represent a group of individuals with particular interests and motivations.

Despite the limitations acknowledged, the insights gained from this study offer a valuable perspective that can improve equine veterinary services in the future. It also underscores the need for more in-depth research to understand the motivations and behaviors of horse owners in engaging veterinary services. Future research should consider diversifying data collection methods and widening the respondent pool to enhance the robustness and generalizability of findings.

References:

American Psychological Association. (2010). Publication Manual of the American Psychological Association (6th ed.). Washington, DC: APA.

Batchelor, C. E. M., & McKeegan, D. E. F. (2012). Survey of the frequency and perceived stressfulness of ethical dilemmas encountered in UK veterinary practice. Veterinary Record, 170(1), 19.

Bird, Steven & Klein, Ewan & Loper, Edward. (2009). Natural Language Processing with Python.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.

Coe, J. B., Adams, C. L., & Bonnett, B. N. (2007). A focus group study of veterinarians' and pet owners' perceptions of the monetary aspects of veterinary care. Journal of the American Veterinary Medical Association, 231(10), 1510–1518.

Cover, T.M. & Thomas, J.A. (2006). Elements of Information Theory (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Creswell, J. W., & Creswell, J. D. (2017). Research design: Qualitative, quantitative, and mixed methods approaches.

Elte, Y., Wolframm, I., Nielen, M., & van Weeren, R. (2021). Client satisfaction in equine veterinary practice: A structured review and qualitative synthesis. Veterinary Record, 187

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2017). Multivariate data analysis. Cengage Learning.

Hastie, T., Tibshirani, R., & Friedman, J. (2001). The Elements of Statistical Learning. Springer series in statistics.

Jaccard, P. (1912). The distribution of the flora in the alpine zone. New Phytologist, 11(2), 37-50.

Jurafsky, D., & Martin, J. H. (2019). Speech and language processing. Pearson.

Kidd, L., & Kuznetsova, D. (2020). Veterinarian communication in the context of euthanasia decisions. Animals, 10(2), 228.

Kogan, L., Oxley, J. A., Hellyer, P., Schoenfeld, R., & Rishniw, M. (2018). UK pet owners' use of the internet for online pet health information. Veterinary Record, 182(21), 601.

Kucher, K., & Kerren, A. (2015). Text visualization techniques: Taxonomy, visual survey, and community insights. In 2015 IEEE Pacific Visualization Symposium (PacificVis) (pp. 117-121). IEEE.

Loh, W.-Y. (2011). Classification and Regression Trees. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 14–23.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.

Mcbride, S., Hemmings, J. A. (2004). A preliminary study of the effect of massage on arterial blood flow in healthy horses. Veterinary Therapeutics, 4(3), 218–226.

McKinney, W. (2010). Data structures for statistical computing in python. In S. van der Walt & J. Millman (Eds.), Proceedings of the 9th Python in Science Conference (pp. 51 - 56).

Pandas Development Team. (2020). Pandas: Powerful data structures for data analysis, time series, and statistics. Retrieved from https://pandas.pydata.org

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.

Ramage, D., Dumais, S., & Liebling, D. (2010). Characterizing Microblogs with Topic Models. Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

Ramos, J. (2003). Using tf-idf to determine word relevance in document queries. Proceedings of the First Instructional Conference on Machine Learning, 242, 133–142.

Raileanu, L. E., & Stoffel, K. (2004). Theoretical Comparison Between the Gini Index and Information Gain Criteria. Annals of Mathematics and Artificial Intelligence, 41(1), 77–93.

Rajaraman, A. and Ullman, J.D. (2011). Mining of Massive Datasets.

Sax, L. J., Gilmartin, S. K., & Bryant, A. N. (2003). Assessing response rates and nonresponse bias in web and paper surveys. Research in Higher Education, 44(4), 409–432.

Shannon, C. E. (1948). A mathematical theory of communication. Bell System Technical Journal, 27(3), 379-423.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and Statistical Modeling with python. Proceedings of the 9th Python in Science Conference.

Sibson, R. (1973). SLINK: An optimally efficient algorithm for the single-link cluster method. The Computer Journal, 16(1), 30-34.

Sue, V. M., & Ritter, L. A. (2012). Conducting online surveys. SAGE Publications, Inc.

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media, Inc.

Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., ... & SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in python. Nature methods, 17(3), 261-272.

Walt, S. v. d., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering, 13(2), 22-30.

Waskom, M. (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021.

Appendix

A count vectorizer and TF-IDF LDA topic modeling:

Group 2 vets	Group 3(+) vets	Group all (combined)
CountVectorizer - Top Words for Each Topic: Topic 0: rontgenfotos expertise call dienst backup work emergency service standaard practice primary zaken different paardentandarts location lijn 2e eigen specialist eerste Topic 1: plaatselijke inentingen licht vaccinatie large dichterbij tegelijk eigen problemen utrecht gespecialiseerde onderzoek pensionstal tandarts kleine nodig spoed rest stal vaste Topic 2: wolvega schapen elder work veterinary standaard gespecialiseerde vaccination needed lameness pony regular specialistische nodig zaken art buurt specialist practice tandarts Topic 3: close peesblessure spoed praktijken main gebruik mobile specialty specialist eigen needed eerstelijns clinic plaatselijke eerste tweedelijns availability home behandeling emergency Topic 4: require ingewikkelde spoed scan issue bereiken eigen enten specialisatie ligt inenten huisdierenarts grotere onderzoek zaken barn available klein e different dichtbij Topic 5: access main meerdere gespecialiseer	**CountVectorizer - Top Words for Each Topic:** Topic 0: navicular advanced eerste availability basic chiropractic zorg call sport vaccinaties reguliere location stal emergency lijn spoed issue medicine general lameness Topic 1: live case handelingen good work regio pensionstal zoekende care rural area stal routine oude expertise emergency specialty lameness practice dental Topic 2: depending specialised onderzoeken reguliere basis verschillende afstand keuren specialistische gebit stal holistische nodig speciality service maken praktijken vaste afhankelijk gebruik Topic 3: gebit ky spor t osteopathie diagnostic dentistry special nodig oude beschikbaarheid woonplaats mile beter away zorg probleme n local care clinic different Topic 4: mile away mogelijkheden problem injury problemen orthopedische procedure behandelingen plaatselijke regulier specialistische huis beter surgery kleine gespecialiseerde general practice spoed Topic 5: spoed expertise specialisatie kreupelheid aangesloten hour gespecialiseerd xray	**CountVectorizer - Top Words for Each Topic:** Topic 0: basic behandeling beter away work spoed reguliere lamenes s vaccination praktijken dichtstbijzijnde hospital needed entingen veterinary diagnose specialist clinic local practice Topic 1: mestonderzoek kosten geneeskunde eerste maak zorg orthopedie gespecialiseerde specifieke holistisch problemen enting stal buurt home rest regulier e tanden vaste afhankelijk Topic 2: tanden goedkoper denk location vraag echo away plaatselijke primary nodig spoed eisende rural tweedelijns entingen paardentandarts procedure eerste emergency zorg availability Topic 3: lingehoeve licht zaken onderzoek specialist gespecialiseerde hangt surgery koliek problemen specialisatie veearts jaar expertise eentje specialisme vaste location nodig eigen Topic 4: specialistische regio standaard emergency call service rontgenfotos onderzoeken specialised mobile issue reguliere specialty local veterinary holistische lameness general practice different

de werkt gespecialiseerd faculteit tandverzorging appointment sportpaarden buurt tanden basis local entingen simpele specialistische lokale enten zorg Topic 6: geneeskunde gevallen utrecht acute werk orthopedisch onderzoek spoed jaar weg doorverwijzing grote orthopedie locatie gespecialiseerd tandarts eentje holistische regulier gespecialiseerde Topic 7: fokkerij art tijd x1 closer simpele normale afhankelijk dental gespecialiseerde blessures zaken emergency veearts weg sportpaarden bel vaste reguliere huis Topic 8: honden spoedgevallen katten nodig enting hospital specialisten procedure weg mobile zorg bereikbaarheid goedkoper keuze farm spoedeisende huis specialisme specialist eigen Topic 9: nodig specialty medicine sport rest specialized reguliere clinic afhankelijk available tanden stal university regular general local routine emergency lameness care	routine tanden available meerdere klinieken depends tandarts enten local eigen needed entingen Topic 6: kreupelheidsonderzoek vragen ken vaccineren vraag gezamenlijk vaccinaties afhankelijk weekend mestonderzoek lingehoeve come klacht beste lameness vaccinatie probleem onderzoek buurt tandarts Topic 7: university general art werk basic spoed basis sport clinic breeding kreupelheid dichtstbijzijnde thuis weg utrecht buurt local ingewikkelde veterinary bel Topic 8: tijd orthopedisch meerdere simpele weg zorg enting huis onderzoeken location specialisatie enten ky b uurt licht different echte spoed afhankelijk vaste Topic 9: expert state specific koliek andere n travel mestonderzoek lameness specialisme farm diensten hangt dichtstbijzijnde hospital depends practice specialist local issue emergency	Topic 5: specialist basis depends gebit door verwijzing praktijken enten lijn veearts gespecialiseerd werk gespecialiseerde simpele specialistische zaken spoed zorg onderzoek stal tandarts Topic 6: rest gevallen grotere faculteit zorg utrecht lokale ingewikkelde huisdierenarts buurt tandarts thuis weg plaatselijke vaste spoed art dichtbij kleine enten Topic 7: specialized maak regulier weg sport paarden general gebruik local behandeling entingen gespecialiseerde needed tandarts bel routine spoed lameness care huis specialist Topic 8: vaccination clinic different basic call hospital routine sport specialty main case local available medicine regular lameness dental issue care emergency Topic 9: werkt meerdere reproductive specialisatie pensionstal specialisten specialistische grotere plaatselijke lameness practice afhankelijk barn inentingen kleine spoed normale gespecialiseerde klacht buurt
TF-IDF - Top Words for Each Topic: Topic 0: huisarts convenient inherent hand primary practice barn handelingen doorverwijzing specialisaties backup 2e zaken service lijn specialist eerste different location paardentandarts Topic 1: problemen gespecialiseerde doel specialisten stal experie	**TF-IDF - Top Words for Each Topic:** Topic 0: issue diensten location medicine available veulen geboorte nacht echter kreupelheden call rural lameness spoed orthopedie voortplanting availability emergency come klacht Topic 1: gespecialiseerde klinieken genus practitioner reproductiv	**TF-IDF - Top Words for Each Topic:** Topic 0: letterlijk verspreid eigen routine practice ara alternatief veterinary praktijken tandarts entingen specialist mobile land vaccination reguliere verschillende spoed clinic available Topic 1: nodig care probleem holistische maakt gespecialiseerde st

nce zaken denk reproductive apart large door verwezen specialize in entingen vaccinatie rest nodig kleine licht pensionstal Topic 2: work pony vaccination specialty regular weg sap eerstelijns basic practice vragen sport specialist medicine standaard gespecialiseerde buurt needed wolvega tandarts Topic 3: imms equipmen texpertise dracht regulier praktijken vaccination specialisatie tweedelijns verschillende onderzoek referred emergency werk behandeling locatie enten specialist main plaatselijke eigen Topic 4: went imaging available wormenkuren plaatse avond barn behandeld inschakelen huisdierenarts kleine emergency home eigen grotere inenten different ligt enten dichtbij Topic 5: appointment someren orthopedie ziektebehandeling buurt entingen gespecialiseerde basis gebit afstand maakt vaste faculteit specialistische bel werkt tandverzorging enten art zorg Topic 6: needed florida general sportbegeleiding orthopedische holistisch regulier vaste eentje grote ara dentist huistuinkeuken team gespecialiseerde practice gespecialiseerd tandarts uithof availability Topic 7: bijzonderheden shot emergency gespecialiseerde lameness x1 afhankelijk dental vriendin ambulant beenproblemen deskundigheid duur dienst goedkoper	e holistische dierensartsen dag weinig others convenient favorite eigen dental specialty specialisatie holistisch eerste lijn buurt Topic 2: dental vaccination clinic prepurchase exam yearly specialised gebruik diagnosticstreatment normally weekend maak vervanger needed beschikbaarheid mestonderzoek enting doel special afhankelijk Topic 3: veearts away clinic eerstelijnshulpingewikkeldespecialistische ky mile location apparatuur passend specialty holistisch tanden reguliere home trainer fijnst aandoening zorg different Topic 4: orthopedische enten tandarts licht creatief behandelingen specialismen veterinary vinden moeilijk ervaring paardentandarts beter multiple mogelijkheden practice regulier nodig spoed huis Topic 5: care tandarts smdc toepassing klachten vage locale specialistenkliniek opinion lingehoeve universiteitskliniek routine enten pensionstal stal gespecialiseerd eigen osteopaat entingen expertise Topic 6: whats wrongneeded physical orthopeed alledrie osteopathie gespecialiseerde diensten land verspreid repro general lameness gelang vaccinatie onderzoek beste buurt probleem tandarts Topic 7: carevaccinations beschikbaarheid drachtigheid ziekten keuringen doorverwijzingen area werkzaam oude s	al main specifieke reguliere vaste buurt enting orthopedie kliniek en rest tanden zorg holistisch afhankelijk Topic 2: medicine home tanden wolvega vriendin locatie schapen goed koper lameness general vragen try probleem entingen eerste regulier e eerstelijns tweedelijns emergency availability Topic 3: enten basis dental come uithof specialistische jaar home problemen licht vaste gespecialiseerde eentje keuze different specialisatie expertise nodig location eigen Topic 4: chiro specialistische weg service buurt enten gebruik maak zaken nodig tandverzorging rontgenfotos zorg holistische regular lameness general emergency different practice Topic 5: specialistische algemene stal zorg procedure eerste veearts afhankelijk doel onderzoek gespecialiseerde spoed werk gespecialiseerd simpele zaken doorverwijzing specialist lijn tandarts Topic 6: weekenddiensten specifiek koliek reproductive work specialty kleine ingewikkelde vaste basiszorg orthopeed spoed thuis faculteit needed dichtbij art regulier plaatselijke enten Topic 7: issue sportpaarden regular entingen kreupelheden enten beenproblemen general lameness hospital practice orthopedisch behandeling routine university gespecialiseerde spe

normale sportpaarden v aste reguliere huis Topic 8: wegens vier f requently travel huis holiday vaccinaties fa rm zu 1hospital eigen zorg keuze spoed speci alisme reguliere holis tische bereikbaarheid gebruik maak Topic 9: specialized b ack honden simpele loc al vaste clinic genera l afhankelijk speciali st mobile tanden avail able regular lameness stal routine nodig eme rgency care	pecialismeklacht tanda rts specialtiesreprodu ction familie emergenc y local bel aangeslote n lameness kreupelheid entingen Topic 8: zoekende bewe gingsapparaat algemene travel enting pensioen afstand echte buitenla nd binnen speciality p robleem spoed speciali st practice general sp ecialisaties different vaste afhankelijk Topic 9: woonplaats be ter speciality art ent en vaste letterlijk ho ek verkoop training fa rm dichtstbijzijnde ta nden nodig emergency d epends thuis hangt spe cialisme issue	cialist care paardenta ndarts huis Topic 8: large 2e othe rs tandarts avond hosp ital local different g ebit issue tijd case r outine care specialisa ties bel bereikbaarhei d stal emergency vaste Topic 9: familie vacci natie meerdere enten o nderzoek raadplegen ta ndheelkundige x1 inent ingen specialistische sap 1hospital grotere specialisten pensionst al werkt mogelijkheden normale buurt klacht

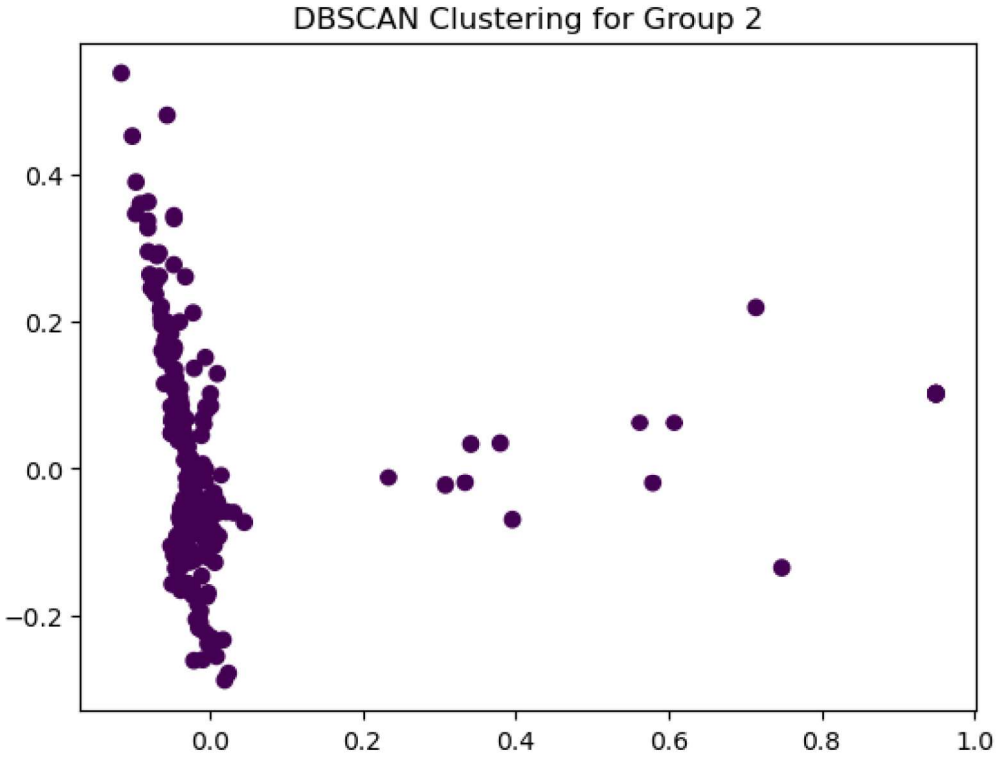
B Decision tree:

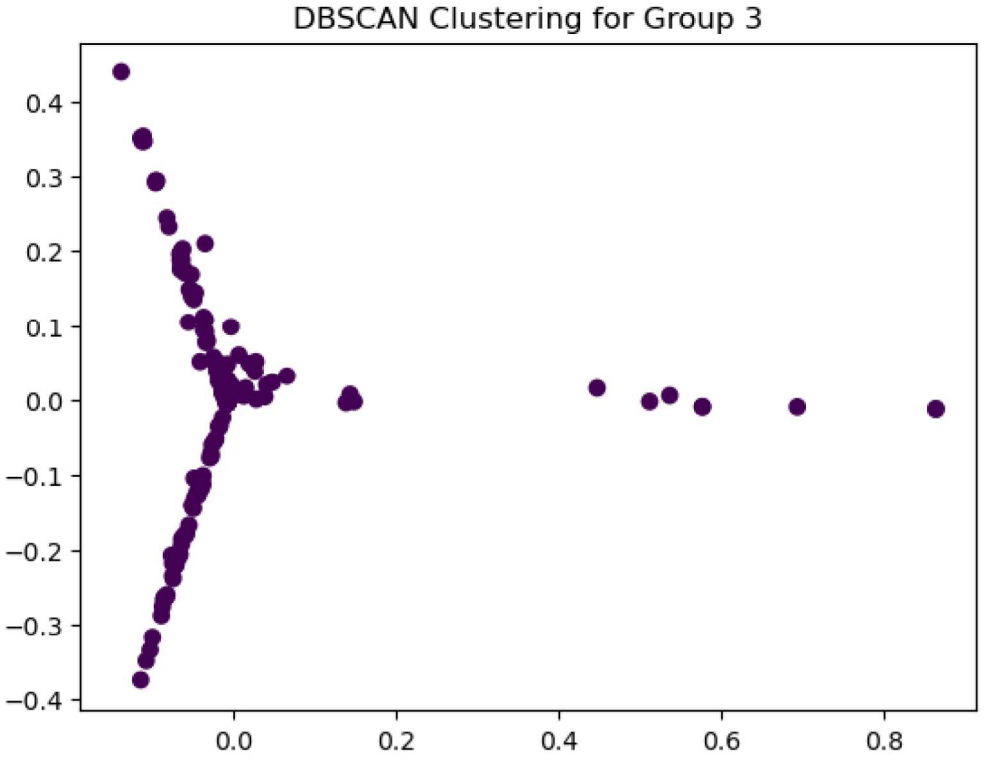
How to read decision tree output table:

	Rule	Class	Samples
0	num_horses <= 3.5 & service_duration <= 20.25 & service_duration <= 6.5	[[23.0, 18.0, 14.0, 27.0, 13.0, 10.0, 17.0, 26.0, 17.0, 20.0]]	185
1	num_horses > 3.5 & num_horses <= 26.0 & num_horses <= 20.5	[[19.0, 16.0, 21.0, 11.0, 12.0, 21.0, 16.0, 18.0, 6.0, 15.0]]	155
2	num_horses <= 3.5 & service_duration <= 20.25 & service_duration > 6.5	[[5.0, 14.0, 19.0, 18.0, 9.0, 9.0, 17.0, 11.0, 13.0, 20.0]]	135
3	num_horses <= 3.5 & service_duration > 20.25 & vets <= 2.5	[[0.0, 0.0, 2.0, 3.0, 0.0, 4.0, 5.0, 0.0, 4.0, 0.0]]	18
4	num_horses > 3.5 & num_horses > 26.0 & service_duration <= 12.0	[[0.0, 0.0, 1.0, 3.0, 0.0, 0.0, 0.0, 0.0, 3.0, 1.0]]	8
5	num_horses > 3.5 & num_horses <= 26.0 & num_horses > 20.5	[[2.0, 0.0, 0.0, 0.0, 3.0, 1.0, 0.0, 0.0, 0.0, 1.0]]	7
6	num_horses > 3.5 & num_horses > 26.0 & service_duration > 12.0	[[0.0, 3.0, 0.0, 0.0, 0.0, 0.0, 2.0, 2.0, 0.0, 0.0]]	7
7	num_horses <= 3.5 & service_duration > 20.25 & vets > 2.5	[[0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0]]	1

" Rule 0: If a customer owns less than or equal to 3.5 horses and uses a vet service for less than or equal to 20.25 years, further classified by those who use the service for less than or equal to 6.5 years, then a certain topic is predicted as the reason for using a vets services. The vector [[23.0, 18.0, 14.0, 27.0, 13.0, 10.0, 17.0, 26.0, 17.0, 20.0]] associated with Rule 0, represents the count of samples that fall under each topic. In this case, we have 23 instances of topic 0, 18 of topic 1, and so forth. The topics are the topic from the TF-IDF topic modeling results.

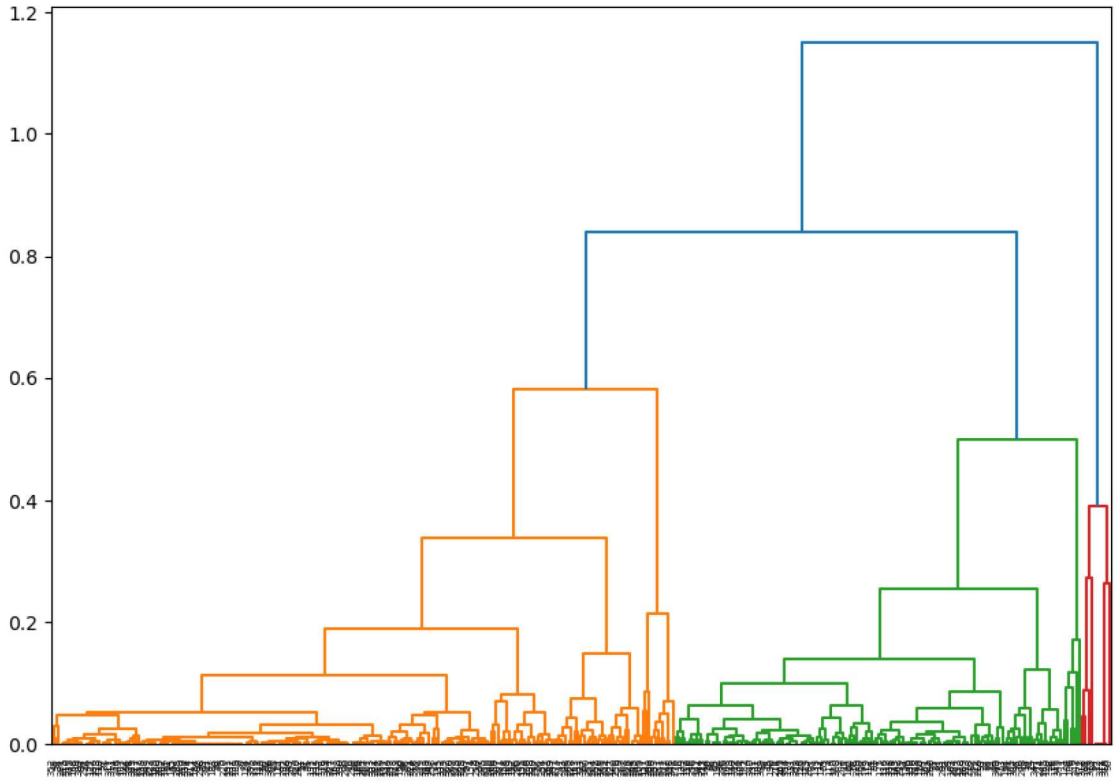
C DBSCAN clustering:



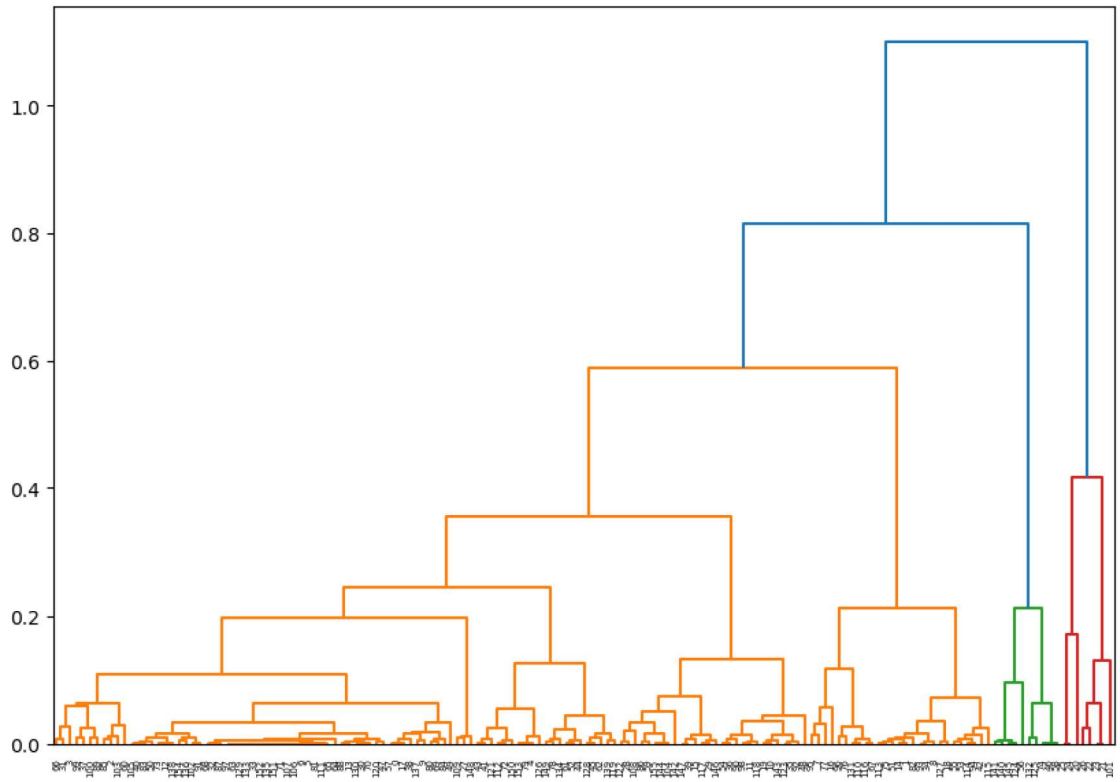


D dendrogram complete linkage reasons column:

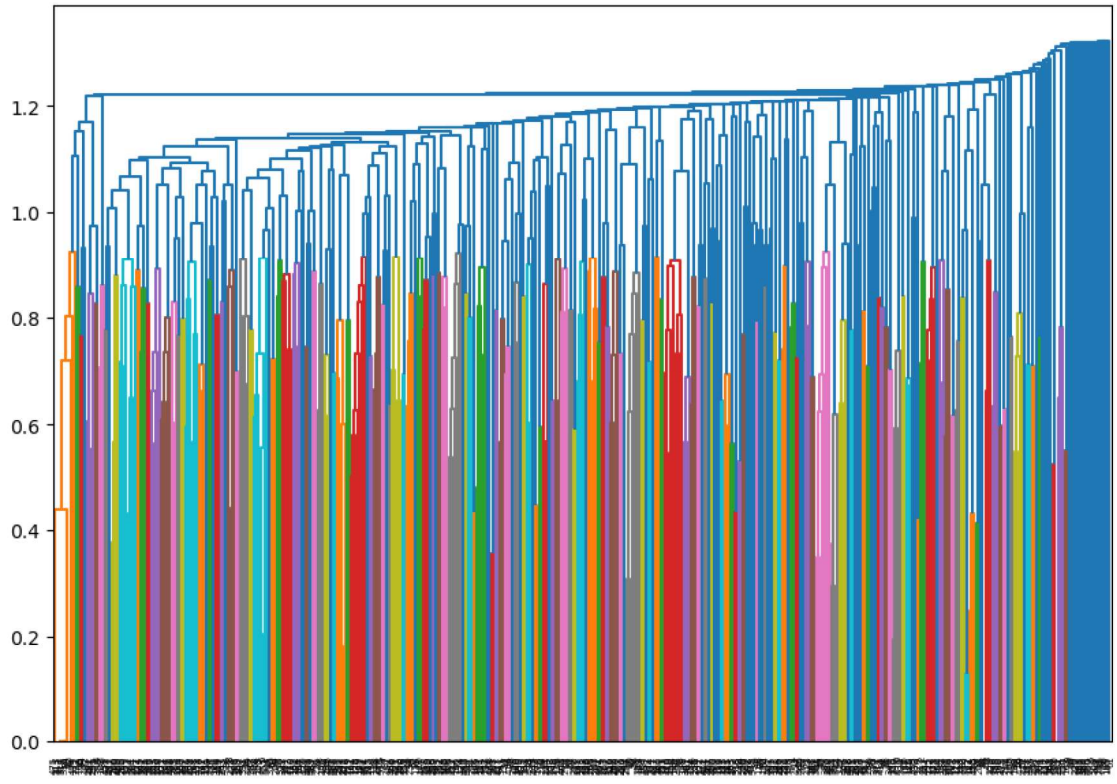
Group 2



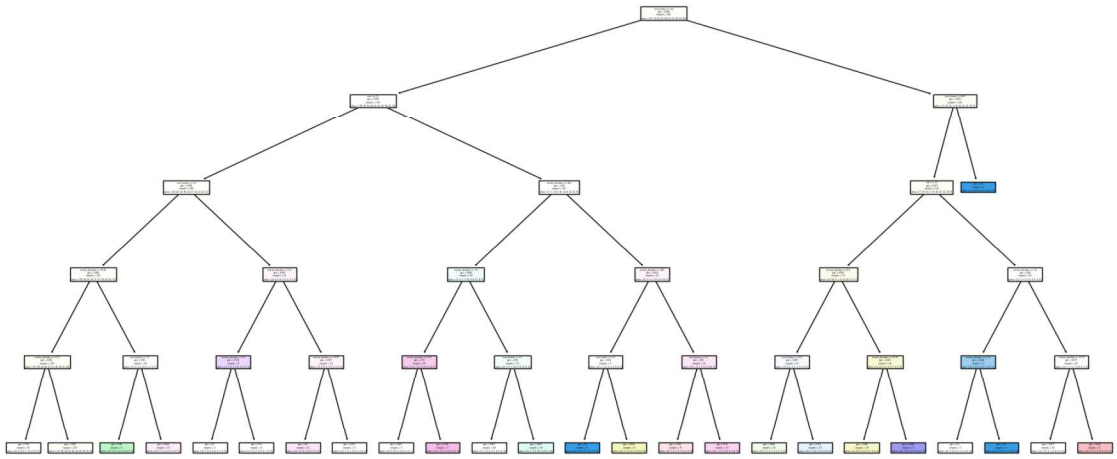
Group 3



Group all

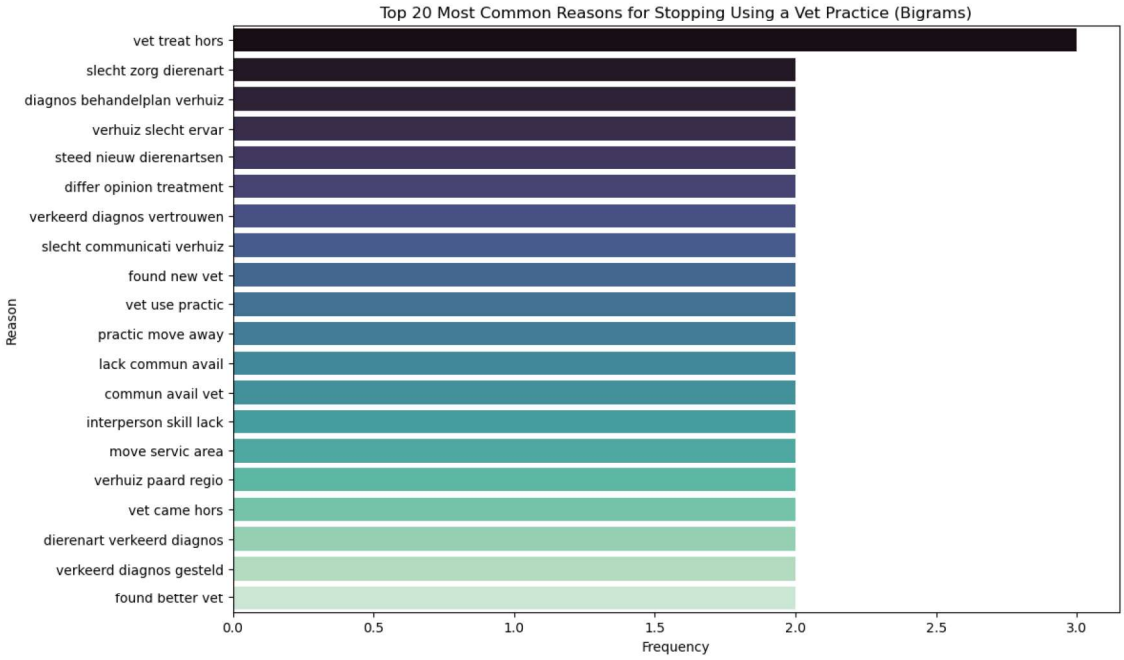


E decision tree depth 5:

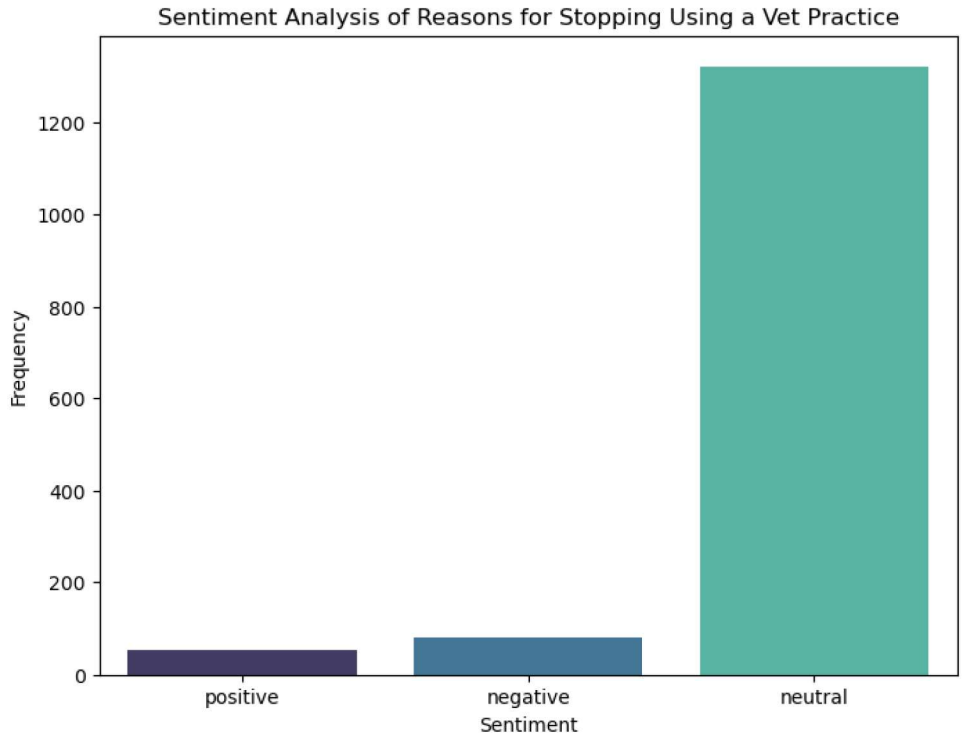


	Rule	Class	Samples
0	num_horses <= 4.5 & vets <= 2.5 & num_horses <= 3.5 & service_duration <= 20.25 & service_duration > 0.75	[[17.0, 36.0, 24.0, 23.0, 26.0, 24.0, 29.0, 18.0, 21.0, 9.0]]	227
1	num_horses <= 4.5 & vets > 2.5 & service_duration <= 14.5 & service_duration > 1.5 & num_horses > 1.5	[[5.0, 1.0, 3.0, 3.0, 12.0, 4.0, 4.0, 6.0, 2.0, 4.0]]	44
2	num_horses > 4.5 & num_horses <= 66.0 & vets > 2.5 & service_duration > 1.5 & service_duration <= 23.75	[[2.0, 7.0, 6.0, 1.0, 7.0, 4.0, 4.0, 6.0, 1.0, 4.0]]	42
3	num_horses > 4.5 & num_horses <= 66.0 & vets <= 2.5 & service_duration > 8.5 & service_duration <= 23.75	[[2.0, 14.0, 2.0, 0.0, 2.0, 5.0, 5.0, 3.0, 4.0, 1.0]]	38
4	num_horses <= 4.5 & vets > 2.5 & service_duration <= 14.5 & service_duration > 1.5 & num_horses <= 1.5	[[1.0, 2.0, 3.0, 4.0, 2.0, 5.0, 4.0, 5.0, 2.0, 0.0]]	28
5	num_horses > 4.5 & num_horses <= 66.0 & vets <= 2.5 & service_duration <= 8.5 & num_horses <= 11.5	[[3.0, 1.0, 5.0, 1.0, 1.0, 3.0, 3.0, 2.0, 1.0, 2.0]]	22
6	num_horses <= 4.5 & vets <= 2.5 & num_horses > 3.5 & service_duration > 3.5 & service_duration <= 23.75	[[3.0, 3.0, 0.0, 4.0, 3.0, 0.0, 0.0, 1.0, 6.0, 0.0]]	20
7	num_horses <= 4.5 & vets <= 2.5 & num_horses <= 3.5 & service_duration > 20.25 & num_horses > 1.5	[[2.0, 0.0, 1.0, 1.0, 2.0, 0.0, 1.0, 2.0, 3.0, 1.0]]	13
8	num_horses <= 4.5 & vets > 2.5 & service_duration > 14.5 & service_duration > 16.5 & num_horses > 2.5	[[0.0, 1.0, 2.0, 0.0, 0.0, 2.0, 0.0, 2.0, 4.0, 0.0]]	11
9	num_horses > 4.5 & num_horses <= 66.0 & vets <= 2.5 & service_duration <= 8.5 & num_horses > 11.5	[[0.0, 3.0, 1.0, 0.0, 0.0, 4.0, 0.0, 0.0, 3.0, 0.0]]	11
10	num_horses <= 4.5 & vets <= 2.5 & num_horses <= 3.5 & service_duration <= 20.25 & service_duration <= 0.75	[[0.0, 2.0, 0.0, 3.0, 1.0, 3.0, 0.0, 0.0, 0.0, 1.0]]	10
11	num_horses <= 4.5 & vets > 2.5 & service_duration > 14.5 & service_duration > 16.5 & num_horses <= 2.5	[[0.0, 0.0, 0.0, 1.0, 0.0, 1.0, 0.0, 2.0, 2.0, 3.0]]	9
12	num_horses <= 4.5 & vets > 2.5 & service_duration <= 14.5 & service_duration <= 1.5 & service_duration > 0.75	[[1.0, 1.0, 1.0, 0.0, 1.0, 0.0, 0.0, 0.0, 3.0, 0.0]]	7
13	num_horses <= 4.5 & vets <= 2.5 & num_horses <= 3.5 & service_duration > 20.25 & num_horses <= 1.5	[[0.0, 0.0, 0.0, 3.0, 2.0, 0.0, 0.0, 0.0, 0.0, 0.0]]	5
14	num_horses <= 4.5 & vets <= 2.5 & num_horses > 3.5 & service_duration <= 3.5 & service_duration > 2.5	[[1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0]]	4
15	num_horses <= 4.5 & vets > 2.5 & service_duration > 14.5 & service_duration <= 16.5 & num_horses > 1.5	[[0.0, 2.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0]]	4
16	num_horses > 4.5 & num_horses <= 66.0 & vets > 2.5 & service_duration > 1.5 & service_duration > 23.75	[[0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.0]]	4
17	num_horses <= 4.5 & vets <= 2.5 & num_horses > 3.5 & service_duration > 3.5 & service_duration > 23.75	[[0.0, 1.0, 0.0, 1.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0]]	3
18	num_horses <= 4.5 & vets > 2.5 & service_duration <= 14.5 & service_duration <= 1.5 & service_duration <= 0.75	[[0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 1.0, 1.0]]	3
19	num_horses > 4.5 & num_horses <= 66.0 & vets <= 2.5 & service_duration > 8.5 & service_duration > 23.75	[[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.0, 0.0, 1.0, 0.0]]	3
20	num_horses <= 4.5 & vets <= 2.5 & num_horses > 3.5 & service_duration <= 3.5 & service_duration <= 2.5	[[0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0]]	2
21	num_horses > 4.5 & num_horses <= 66.0 & vets > 2.5 & service_duration <= 1.5 & service_duration <= 0.75	[[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 1.0, 0.0, 0.0, 0.0]]	2
22	num_horses > 4.5 & num_horses > 66.0	[[0.0, 0.0, 0.0, 0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 0.0]]	2
23	num_horses <= 4.5 & vets > 2.5 & service_duration > 14.5 & service_duration <= 16.5 & num_horses <= 1.5	[[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0]]	1
24	num_horses > 4.5 & num_horses <= 66.0 & vets > 2.5 & service_duration <= 1.5 & service_duration > 0.75	[[0.0, 0.0, 0.0, 0.0, 0.0, 1.0, 0.0, 0.0, 0.0, 0.0]]	1

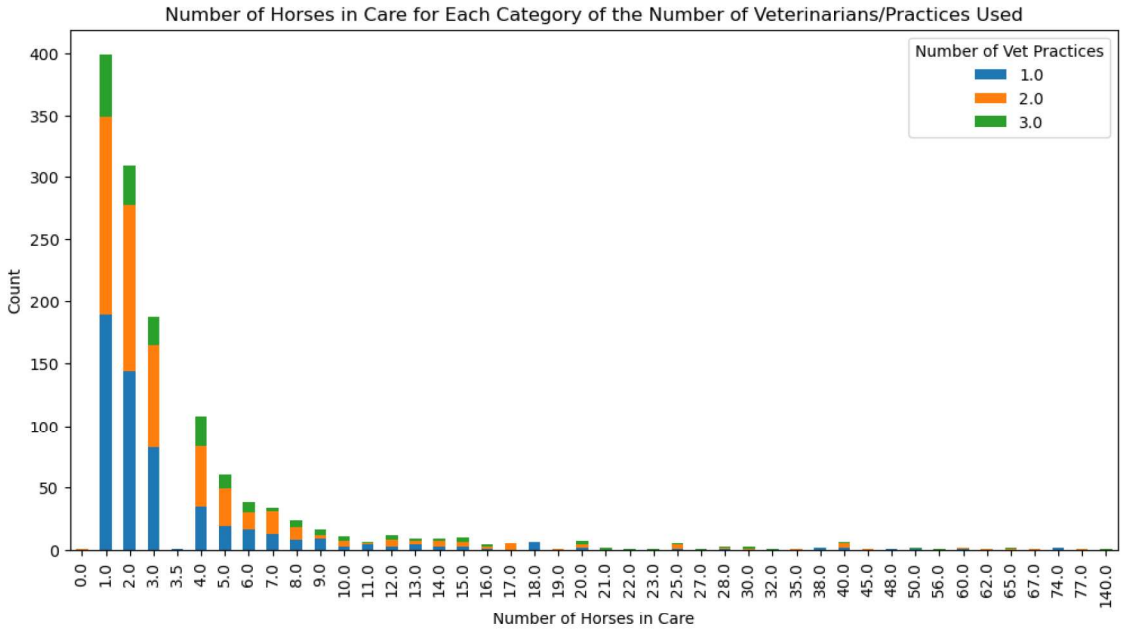
F bigram most common reasons stopping at a vets practice:



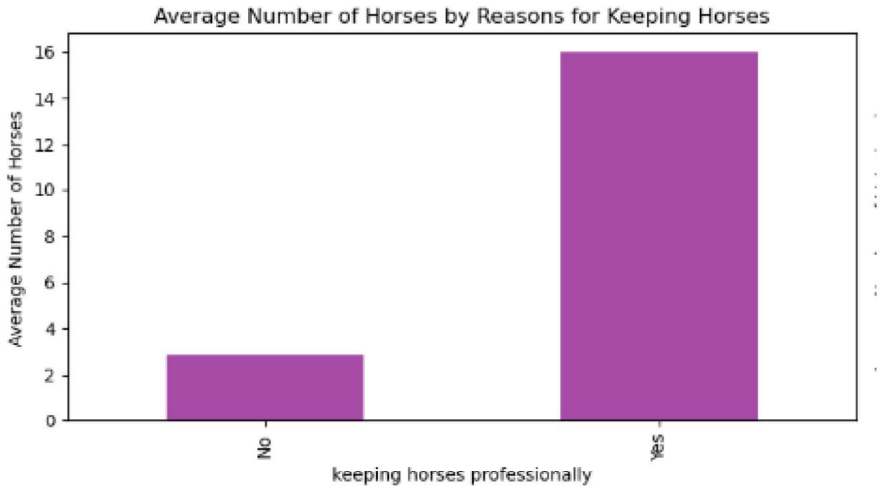
G sentiment of reasons for stopping at a vets practice:



H number of horses in care per number of vets used:



I average number of horses kept, professional or not:



J count of each purpose keeping horses: