Universiteit Utrecht

Utrecht University
Faculty of Science Artifical Intelligence

Contrastive explanation of the output of machine learning models

By Remco Leen (6009557)

Supervisors: Dr. A.J. Feelders, Utrecht University
Prof. dr. mr. H. Prakken, Utrecht University

November 11, 2019

Abstract

In this research we aim to automatically generate an explanation of decisions made by machine learning models. To be able to do this we adapted the explanation based model by Feelders and Daniels. We describe the building blocks of the model and consider different options for determining the required reference object. In this thesis we calculate the required distances with either Gower's distance or the simplex method. For the type of object reference we use either the closest object with the desired classification or the medoid object of the desired classification. We test the proposed algorithm with a questionnaire that tested the quality of the explanation and parts there of. We found that the medoid reference type was significantly better received by respondents than the closest reference type. In the ordinal logistic regression model we found a significant negative effect of the number of errors people made in the subject knowledge questions on the perceived explanation quality. We were unable to find any significant results for the other factors, and we found no significant effect on which distance function performs better, or if adding the counter-acting to the contributing causes had an effect on the overall perception of the explanation. As the number of participants to the empirical study was rather small we opted to go for a more exploratory approach to find factors that could be interesting to investigate in further studies. Due to this reason future research with this method is recommended as we were unable to get a definitive conclusion for our developed model.

Acknowledgement

I would like to extend my sincere thanks to Dr. A.J. Feelders for the to me amazing guidance through the process of making this master thesis. I really appreciate the small sparring on ideas and the guidance through the writing process which is not my greatest skill.

Also a major thanks to Prof. dr. ir. L.C. van der Gaag for providing a writing class for master thesis students in which I learned a lot.

A major thanks to my friend MSc. M. Burghoorn for keeping me sane throughout this master thesis process.

I would also like to thank my sister MSc. N.A. Leen for talking me through the wonderful world of statistics.

And a final thanks to my parents for keeping up with my most abnormal sleeping schedule turning me into a night animal.

Contents

1 Introduction

With the wide spread usage of Artificial Intelligence (AI), in the current day and age people see more and more decisions being made for them by predictive models. For these people an explanation as to why a particular decision was made is expected. Today's predictive models almost all suffer from the issue of being unable to explain why a certain prediction is made over the other. Other fields (medicine, Law, etc.) which insist on having a decision made which can be explained are currently limited in their wide spread use of predictive models due to this shortcoming. The ability to see how decisions are made by the predictive models might also give us a better understanding as to how they work. There has been a substantial amount of research on the subject of explaining decisions made by prediction models. Most of these solutions work with a model-agnostic method of trying to provide the explanation. This means that it takes into account the input and output of the model and tries to come up with an explanation without having access to a description of the model itself. The idea of these explanations is that depending on the most important variables people can make their own decision about the correctness of the explanation. We approach this issue from a different side, instead of looking purely at the features side of the model we first want to find out how humans provide explanations to other humans. Philosopher Miller has looked into how humans explain to one another (Miller, 2018). According to Miller a good way of explaining why something happened or why something is, is to provide an example of a similar situation with a different result to provide a comparison for a person to deduce with. For example if we want to know why student A didn't get accepted into college we could look at an other student B who did get accepted and compare their grades to see that student A has lower grades compared to student B. While methods like these exist they tend to not involve the feature importance and only provide a comparison (Molnar, 2019). What we strive for is an approach which sits somewhere between the model-agnostic methods and example-based explanations. The proposed method falls between these two, as we use the existing datapoints to find a reference object (example-based explanation) while the explanation given is based on the features and how to get the desired result (model-agnostic).

While research to the specific approach is limited a great number of tools are available to make this possible. We use multiple building blocks, first we make use of Feelders and Daniels explanation framework as the starting point (Feelders and Daniels, 2001). Because we are using this we require to search for a fitting reference object which we can use to provide the comparison for the explanation. To be able to compare objects we will have to look at different distance measures. As we want the algorithm to work with mixed numeric and categorical data we also need to make sure we are able to use categorical data in these distance functions.

For the algorithm to function we require a reference object, which we use to compare the input with. The main purpose of this reference object is to provide an object with the desired classification different from the one the explainee got assigned to. To acquire the reference object we looked at multiple solutions, which primarily came from the field of Data Clustering (van de Velden et al., 2018). All of these solutions make use of a distance functions, which we also require to choose.

1.1 Research questions

Summarizing we aim to answer the following question:

> Does the explanation framework which we will refer to as NEXT (Norm-based EXplanation Technique) presented in (Feelders and Daniels, 2001) provide a good basis for explaining decisions of machine learning models?

To make the explanation framework operational, we specifically consider the following choices:

1. Which of the considered reference objects produces the best explanations?
2. Which of the considered distance functions provides us with the best reference objects, going by the quality of the explanation produced?
3. How does the inclusion or exclusion of counteracting causes influence the quality of the provided explanations?

In chapter 2 we will describe NEXT. In chapter 3 we will explain which tools we are going to use to get the distance functions and reference object we need to use in NEXT. In chapter 4 we will go into how we plan to test the models performance in the form of an empirical study. In chapter 5 we will explain the different analyses we ran on the empirical study results. In chapter 6 we will end with the end results of our algorithm and the accompanied empirical study to see it's effectiveness. We finish in chapter 7 where we take a look at what we could improve for future studies and what might need to change.

2 A formal model of explanation

In this chapter we provide a summary of the research done in the paper by Feelders and Daniels (Feelders and Daniels, 2001). We will explore the proposed explanation model and explain how we implemented it in the model. We will discuss the overall ideas of the explanation model by Feelders and Daniels, the requirements it has, what has been used to meet those requirements and why we made certain choices.

2.1 An explanation based model

Feelders and Daniels developed a model for diagnosis and explanation in financial models. Their model is largely based on Humphreys' notion of aleatory explanations. In short an event E occurred because of C^+, despite C^-. Where C^+ are the *contributing causes*, which help explain why E happened, and C^- are the (possibly non existing) *counteracting causes*, which could prevent E from happening (Feelders and Daniels, 2001; Humphreys, 1989). The event E in this regard is divided into two kind of possible events:

- variable y has some particular value at time t
- variable y *changes* value from time t to t'

Feelders and Daniels change the meaning of the event E to make it more descriptive in their usecase. They evolve it from E to the usage of Hesslow's $\langle a, F, r\rangle$ explanandum of an event (Hesslow, 1983). Here a is an object, F being a property of that object in reference to a reference class r. An example would be $\langle$student A, has a relatively low grade, other students in his class$\rangle$. In these explanandum r can take up different forms e.g. we could also compare "student A" with "student A" at a later date to change the reference type to a temporal object, multiple kinds of options exist we only need to make sure that the reference objects we are using are viable in the comparison.

As this is a very abstract way of being able to compare two events, we are able to use it in a great many cases. Two issues arise to make us able to use this way of explaining to automate an explanation in the model:

- A way to find the contributing (C^+) and counteracting causes (C^-).
- A viable reference object used for comparison

Feelders and Daniels already developed a way to calculate the C^+ and C^- which will be explained in chapter 2.2.2.

As we want to be able to also use this on mixed models, this brings up the issue of converting categorical values to real numbers when we need to calculate distances of objects from each other. This is also used to find a reference object from any dataset as we will have to deal with categorical data in these cases. As our goal is to find an explanation for decisions/classifications made by a algorithm we need to be able to automatically generate a viable reference object from the training data that was used on that model.

2.2 Finding the explanation

Having a distance function and reference object available, we are now able to use them in the algorithm to calculate causes by using the method provided by Feelders and Daniels (Feelders and Daniels, 2001). If we take the previous explanandum $\langle a, F, r\rangle$ we now always know the variable a, the object we want an explanation for and variable r the object we are comparing it to. From the context of our situation we know we want to find why a didn't get the value that r got, this is the F variable. In our case these are the causes we want to compute to find out why. For example: why was a's loan application rejected, whereas b got accepted.

2.2.1 Finding the set of causes

To begin formulating an explanation we want to find the difference in value between the two objects the object a with value $y^{(a)}$ and the object of reference c with value $y^{(c)}$. To get the degree of difference we need to calculate the difference in the outcome of their values which we call the Δy.

$$\Delta y = y^{(c)} - y^{(a)} \tag{1}$$

Generating an explanation we compare and take into account the influence of the different variables in the current classifier. This can be achieved by looking at the influence on the prediction the different input variables have. We can calculate this influence using the variables of both the applicant $x^{(a)}$ and reference object $x^{(c)}$. To get the influence of a single variable i of $x^{(a)}$ denoted as $\inf(x_i)$ we replace the value of $x^{(a)}$ with the variable i of $x^{(c)}$ (denoted as $f(x^{(a)}_{-i}, x^{(c)}_i)$ and calculating the difference it made on the value we got previously. This answers what y value would a have had, if it had c's value for x_i. Resulting in the formula:

$$\inf(x_i) = f(x^{(a)}_{-i}, x^{(c)}_i) - y^{(a)} \tag{2}$$

This gives us a score for a single variable i. We can extend this equation from the influence of a single variable to that of a set of variables X with index set $I \subseteq \{1, 2, ..., n\}$ on y. This changes the previous equation 2 to:

$$\inf(X_I) = f(x^{(a)}_{-I}, x^{(c)}_I) - y^{(a)}$$

Instead of now calculating for only a single variable we calculate the influence of every variable. This collection of influences we then use to define the set of causes C. This is simply adding any variable which influence does not equal 0 to the set of causes C (Feelders and Daniels, 2001).

2.2.2 Counteracting and contributing causes

Now that the algorithm has a way of finding the causes of the difference between two objects we want to be able to find which causes contribute to the current value and which causes counteract the current value. To separate the causes C we look at the previously calculated Δy in equation 1. If C_i has the same sign as the Δy has it gets assigned to the set of contributing causes C^+ if it has the opposite sign it gets assigned to the counter-acting causes set C^-.

A problem arises if big objects are used to compare to each other and these objects also differ in a lot of variables. We can end up in a *data overload* for the person needing to interpret the causes. To prevent even the tiniest influences to be shown to the person receiving the causes we filter out the less important ones based on a threshold. This is done by using the notion definition of a parsimonious set:

Definition 1. *(Parsimonious set of contributing causes).* The parsimonious set of contributing causes C^+_p, is the smallest subset of C^+ such that:

$$\frac{\inf(C^+_p)}{\inf(C^+)} \geq T^+$$

The parsimonious set of counteracting causes is defined analogously. The parsimonious set of contributing causes is the smallest subset of the set of contributing causes, such that its influence on y exceeds a particular fraction T^+ of the influence of the complete set. In case there are several sets of equal cardinality that explain a fraction larger than T^+, the one with the highest inf-value is called the parsimonious set. The definition with respect to counteracting causes is clearly analogous. The fraction T^+ and T^- are numbers between 0 and 1, and will typically be close to 1. We can now use this parsimonious set to explain to the user why a certain event/classification happened.

Coefficients:	value	student A (A)	student B (B)
grade1 (g_1)	0.05	10	9
grade2 (g_2)	0.01	5	6
bad behaviour (b)	-0.25	1	0

Table 1: Example coefficients and student score

2.2.3 Example

If we would have a simple linear model to calculate if a students got a admitted to a master programm or not which includes three variables: 'grade1' ([1-10]), 'grade2' ([1-10]) and 'bad behaviour' ([0-1]) with the coefficients shown in table 1 (with an intercept of 0 for simplicity sake) this results into the linear equation 3. In this example we consider a student to have been admitted if $y \geq 0.50$. If we would want to see why student A didn't get admitted with his score while student B did as shown in the coefficient table 1.

$$y = 0.05 \times g_1 + 0.01 \times g_2 - 0.25 \times b \tag{3}$$

We now use the previous described method to find out why student A failed to gete admitted compared to student B. We first want to calculate the difference in y value student A has from our reference object student B:

$$y^{(B)} - y^{(A)} = 0.51 - 0.3 = 0.21 = \Delta y \tag{4}$$

We now want to find the influences of that each separate variable has we do this by replacing one value of student A with that of student B and see what changes in the y value. For each feature we now want to swap their value for the one in the reference object student B and calculate their influence:

$$\begin{aligned} \inf(g1) = y(g_1^{(B)}, g_2^{(A)}, b^A) &= (0.05 \times \ \ 9 + 0.01 \times 5 - 0.25 \times 1) - 0.3 = \ \ -0.05 \\ \inf(g2) = y(g_1^{(A)}, g_2^{(B)}, b^A) &= (0.05 \times 10 + 0.01 \times 6 - 0.25 \times 1) - 0.3 = \ \ \ 0.01 \\ \inf(b) = y(g_1^{(A)}, g_2^{(A)}, b^B) &= (0.05 \times 10 + 0.01 \times 5 - 0.25 \times 0) - 0.3 = \ \ \ 0.25 \end{aligned} \tag{5}$$

Grade1 has a different sign than Δy we can say this is a counter-acting cause, grade2 has the same sign as Δy we add this to the set of contributing causes. For bad behaviour just like grade 2 the sign of coincides with Δy also making this a contributing cause.

We now have a set of causes C and we know which causes are contributing causes C^+ and counter-acting causes C^-. The final part is to find the parsimonious set of contributing causes. In the example we pick a threshold $T = 0.95$ to filter the contributing causes. We now want to calculate the parsimonious set of contributing causes, in our case this consists only of the bad behaviour feature since if we add the grade2 feature as well we go above our threshold $T = 0.95$:

$$\frac{inf(g_2) = (0.25)}{inf(C^+) = (0.25 + 0.01)} = 0.96 \geq T \tag{6}$$

Since our function is linear we end up having to add both our influences together, in the case of different relations between features the influence of two features might differ when taken together. We can see that in this case only the bad behaviour would be shown as its effect is higher than the threshold. In the end this tells us the main reason that student A didn't get admitted was his bad behaviour and additionally can show us that he did score better in grade1 than student B but not enough to outweigh his bad behaviour.

3 Adaption of the explanation based model

In chapter 2, we presented the NEXT explanation model. In this model we require a reference object (object of reference) to generate explanations. A reference object could for example be the most similar object that received the opposite classification or an object that is in some sense representative for the group of objects that received the opposite classification. In either case, we need a similarity measure (or distance measure) that quantifies the similarity between objects. Since objects can have both numerical and categorical features, we need a similarity measure that can handle such mixed type data.

3.1 Calculating distances between mixed type datapoints

A lot of research has been done on the topic of converting categorical variables to continuous ones (Robnik-Šikonja and Kononenko, 2008; Drezner and Turel, 2011), but even with all this research a perfect solution has not yet been found. For our research and for the purpose of having a comparison that is easier to grasp we chose to go with two different distance functions which incorporate the changing of categorical data to continuous. The first one is the more naive Gower's distance (Gower, 1971), the second one is the more complex Simplex method (McCane and Albert, 2008).

3.1.1 Gower's distance

Gower's distance is an easy to grasp distance measure to calculate the difference between two mixed-type datapoints. It's used to compare two objects A and B to get a distance measure $D_g(A, B)$. To calculate this distance measure we need to first find the distance between each individual variable i in the two objects expressed as $d_{(A,B)i}$. the calculation differs depending on the type of variable that is being evaluated. When the variable is categorical in nature and A_i and B_i are of the same category we set the value to 0; if they differ we set it to 1. If the variable is continuous we end up with a different formula which calculates a number between 0 and 1, where 0 means the variables are exactly the same and 1 means they are at opposite ends of their respective range. The formula used to calculate this score is as follows:

$$d_{(A,B)_i} = \frac{|A_i - B_i|}{R_i}$$

Here R_i is the range of variable i i.e. $R_i = max(X_i) - min(X_i)$ where the maximum and minimum are taken over the collection of all datapoints. Gower's then sums up the scores and divides them by their availability $\delta_{(A)(B)_i}$, which gets set depending if the two compared variables are present in the database, and not missing. Our algorithm assumes that the database is always complete, so availability always equals 1. This results in the main part of the Gower's distance:

$$D_{(A,B)} = \frac{\sum_{i=1}^{p} d_{(A,B)_i}}{\sum_{i=1}^{p} \delta_{(A,B)_i}}$$

It is possible to add weights to each variable if desired. In that case the formula becomes:

$$D_{(A,B)} = \frac{\sum_{i=1}^{p} d_{(A,B)_i} \times W_i}{\sum_{i=1}^{p} \delta_{(A,B)_i} \times W_i}$$

If we look at how Gower's distance treats the categorical values we can see that it does not care for the impact a categorical variable actually has. Rather it assumes the worst case and says that for one change in categorical variable it equals a max range difference even if two categorical variables could very well be close to each other in practice. We picked this distance measure as it is simple to grasp for a person how this function works and so can be used as a good point of comparison for the more complex Simplex method.

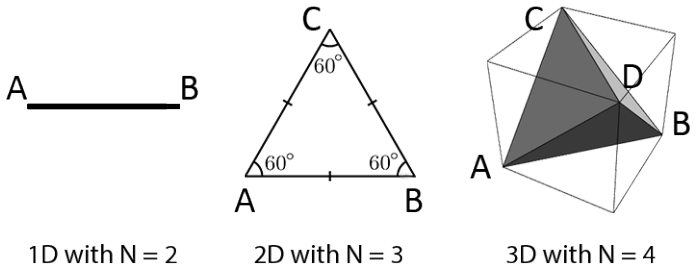


Figure 1: Simplexes for n = 2,3,4.

3.1.2 Simplex method

The Simplex method aims to find the impacts that different values in a categorical feature has in comparison to its other values. For example, we want to know the difference between the values A, B, C from a categorical variable X it is possible for the values A and B to be closer than the values B and C. This comes from the fact that after getting the coordinates we calculate the distance with the inverted covariance matrix. Which results in a value between 0 and 1 for each different column. unlike what the Gower's distance does where these distances are always either 0 or 1. The Simplex method as its name states uses a simplex, a $(n-1)d$ shape where all point are the same distance away from each other, to replace the categorical values of a variable with the coordinates of a n sized simplex within a $n-1$ space. The examples of the simplexes generated when $n = 2, 3, 4$ are shown in figure 1, as can be seen in the figure all of the points displayed are within the same distance of each other if we were to measure them. If we were to convert our categorical variable X into simplex coordinates we would end up with the following coordinates:

$$A = (0, 0)$$
$$B = (1, 0)$$
$$C = (0.5, \frac{\sqrt{3}}{2})$$

Which in our eventual table would look as followed:

Value	v_1	v_2
A	0	0
B	1	0
C	0.5	$\frac{\sqrt{3}}{2}$

The use of the Simplex is preferred due to this property as all the different categorical values get the same starting point in importance. From here on we replace the original categorical values with the coordinates of the simplex' points in their respective space as a vector. To summarize, each categorical value is replaced by $(n-1)$ numerical values, where n is the size of the domain of the variable.

From here the Simplex method calculates a standard covariance matrix for all the datarows in the dataset. It then uses the Mahalanobis distance to calculate the distance for both the continuous and former categorical variables separately. For the Mahalanobis distance we use the inverse covariance matrix Σ^{-1}. When we use this in the algorithm and we want to compare person A to person B for example, we use the inverted covariance matrix we calculated and use the following formula twice. Once for continuous variables D_{con} and once for categorical D_{cat}:

$$D(A, B)_{con/cat} = (A_{con/cat} - B_{con/cat})^T \Sigma^{-1} (A_{con/cat} - B_{con/cat})$$

When both distance D_{con} and D_{cat} have been calculated we combine them using the euclidian distance measure, we can do this as the Mahalanobis distance inherently scales the variance of each column down to 1:

$$D(A, B) = \sqrt{(D_{con})^2 + (D_{cat})^2}$$

The differences between Gower's and the Simplex method consist of the different way of calculating the distance between categorical values. Another difference is the fact that the Simplex method makes use of the Mahalanobis distance which takes into account the correlation between variables (through Σ^{-1}) whereas Gower's ignores the correlations entirely.

3.1.3 Correlation example Mahalanobis and Gower's

In table 2 we show an example dataset where we have the age of people combined with their income. The trend in the dataset is that with age the income also increases except for one outlier being the person aged 23 having an income of 3000. The last two columns show the total distance to all other datapoints combined for both the Gower's distance and the Mahalanobis distance. If we look at our outlier here we can see that in the case of using Gower's distance this one is quite average compared to the others. In comparison the Mahalanobis distance shows the outlier which has the biggest distance from the other datapoints by a large amount. This indicates that something seems to be different than what we would expect in the dataset (in our case the positive correlation of age and income).

	age	income	Gower's	Mahalanobis
1	23	3000	2.40	31.28
2	26	2400	2.54	18.37
3	32	2600	2.23	21.81
4	54	4400	2.50	15.59
5	44	4000	2.07	11.73
6	60	5000	3.28	21.18

Table 2: Example dataset and the total distances to all others on distance function.

3.2 Reference object(s)

For reference object choices we looked at how data-clustering algorithms cluster their different datapoints around a singular existing point, finding the center this way. We ended up with K-medoid for it's ability to find a prototypical object of our classification classes with usage of an existing datapoint. This way we can keep the realistic reference without averaging all separate variables in a dataset. The other approach we went with is the closest datapoint with the desired classification (Wilson and Martinez, 1996). This gives us a nice comparison to see if people will prefer a person being as close as possible or the more prototypical approach.

3.2.1 Closest object

The closest point type of reference object is quite simple. We calculate the shortest distance we can get to another datapoint with the desired classification. To do this we simply use one of the distance functions explained in the previous subsection. Due to having two different options for the distance function it's possible we get different results using a different distance function.

3.2.2 Prototypical object

The prototypical object or medoid approach we take uses the aforementioned approach of finding the datapoint that is the least distance away from all other datapoints of the desired classification. Same as with the closest object approach the actually output object can differ depending on which distance function we use. Let X be a set of data points, then x_m is a medoid of X (with respect to the distance measure d). If and only if:

$$x_m = \underset{x \in X}{arg\,min} \left\{ \sum_{x' \in X} d(x, x') \right\}.$$

Notice that a medoid is always an observed data point. An alternative would be to compose a new data point by taking the average (numerical) or mode (categorical) of each single attribute. Such an object may however be highly unrealistic and therefore not suited as a reference object.

For the data in table 2, observation nr. 5 is the medoid according to both Gower's distance and Mahalanobis distance.

Variable Name	Range	Description
Ratio of payment to income (dir)	percentage	This is the total monthly mortgage payment divided by total monthly income
Ratio of housing to property value (hir)	percentage	This is the inhouse expense to total income ratio
Ratio of loan size to property value (lvr)	percentage	This is the total size of the requested loan divided by the properties total value
Consumer credit score (ccs)	[1-6]	1: if no "slow" payments or delinquencies. 2: if one or two slow payments or delinquencies. 3: if more than two slow payments. 4: if insufficient credit history for determination. 5: if delinquent credit history with payments 60 days overdue. 6: if delinquent credit history with payments 90 days overdue.
Mortgage credit score (mcs)	[1-4]	1: if no late mortgage payments. 2: if no mortgage payment history. 3: if one or two late mortgage payments. 4: if more than two late mortgage payments.
Bad credit record (pbcr)	Yes/No	If any public record of credit problems (bankruptcy, charge-offs, collection actions) exist
Unemployment rate in industry (uria)	[0%-100%]	This is the percentage of people that are unemployed in your line of work
deny	Yes/No	If the applicant was denied their mortgage application (1 = yes; 0 = no)

Table 3: Description of variables

4 Experimental design

In this chapter we show how we selected the cases for explanation and we will explain how we tested the quality of our generated explanations. To achieve this we use a questionnaire as explanations can be seen as adequate for one person but unhelpful to another. This section covers the cases we used to put in our questionnaire, which questions were asked with each case and the presentation of the questionnaire itself. We also explain why we chose our specific questions and what we expected to see in our return.

4.1 The cases

For our questionnaire we used the Boston HMDA 1997-1998 dataset included in the ECdat library of R. From this dataset we created a logistic regression prediction model and used this prediction model to classify our objects for the eventual use in our algorithm. Then we run the algorithm on different datarows to create our cases.

4.1.1 The logistic regression model

For the logistic regression model we don't use the full Home Mortgage Disclosure Act(HMDA) dataset. The columns used for the predictions variables are visible in table 3. The classification values are the deny column (Munnell et al., 1996). With this data we create a logistic regression model with deny as the Y variable and the variables in the table as the prediction variables. After calculating the model we take all the results and transform the classification predictions from a continuous number to either a 1 if the probability of denial was 50% or higher else it got a 0. This way we ensure our algorithm only has access to the classification gotten instead of the chance that get calculated by the model. If we look at the coefficients in table 4 we see that most variables behaved as we expected, the only exception to this is the coefficient that 'ratio of housing to property value' has. Our expectation was that if a person would spend more of his total income on housing expenses this would give a higher chance of denial not a lesser chance. If we look at the actually significance of the different variables we see that four of them are highly significant and two (mortgage credit score and employment rate) are significant. We see that the 'ratio of housing to property value'(hir) is not significant due to this reason and the reason that with our current threshold it never was part of the set of causes and it's sign is counter-intuitive, we decided to remove it from the questionnaire.

Coefficients:	Estimate	Std	z value	Pr(>|z|)
(Intercept)	-7.43649	0.53381	-13.931	<2e-16
ratio payment to income	5.30583	0.99184	5.349	8.82e-08
ratio of housing to property value	-0.56651	1.15599	-0.490	0.6241
ratio loan size to property value	2.64364	0.45908	5.759	8.48e-09
consumer credit score	0.28924	0.03647	7.930	2.19e-15
mortgage credit score	0.30316	0.13110	2.312	0.0208
bad credit record 'yes'	1.29967	0.19056	6.820	9.10e-12
unemployment rate in industry	0.06324	0.03144	2.012	0.0443

Table 4: Summary of the logistic regression model.

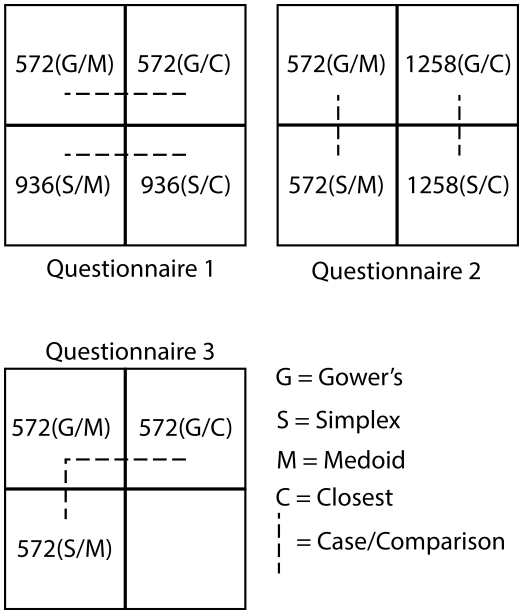


Figure 2: The distribution of cases over the Questionnaires.

4.1.2 Selecting the cases

To select the cases we set a threshold T for the counteracting and contributing causes of $T = 0.88$. We then proceeded to calculate all the causes for the different combinations of setups for distance functions and reference objects, this results in each datarow which had a rejected mortgage in the model having 4 different cause sets to compare with each other (Simplex/Closest, Simplex/Medoid, Gower's/Closest, Gower's/Medoid). We exclude all the datarows that contained no differences between all four options. From all the remaining cases we chose a number of cases that would represent the non-outliers (those of which had no strong single cause). To minimize the number of different questionnaires to be filled we arranged it so we could use both within and in-between subject testing. This was done as shown in Figure 2.

In Figure 2 the numbers represent the ID's of the data rows and the dotted lines the comparison in the within subject analyses. The comparisons always test for just one difference in the possible distance functions or reference objects with the exception of the single case in questionnaire 3 which is meant to be a single case. The main reason for this version is the addition of counteracting causes being displayed and wanted to see the difference in-between subject groups. This means Gower's/Medoid vs Gower's/Closest gets compared. But Gower's/Medoid vs Simplex/Medoid doesn't in the same questionnaire (except for the one case in questionnaire 3). The ID's overlap in the different questionnaires as this makes it possible to use the results we got in questionnaire 1 and 2 to be compared to those in questionnaire 3 which has the

counter-acting causes added to it. Case 572 was chosen as it provided one of the few situations where there was a difference in causes for each possible combination of distance function and reference object.

4.2 Experimental setup

As we want to minimize the number of people needed to participate in the questionnaire we tried to set it up in such a way that with a minimal number of cases we could test our different questions within and in-between subject groups. To achieve this we split the questionnaire in three different versions, to prevent the questionnaire from becoming too long for a single person to fill in comfortably.

4.2.1 Background knowledge

At the start of the questionnaire we ask participants to read a small explanation of the cases they will be exploring. This section explains what each variable means. After the participant reads this explanation he or she is asked to answer a number of questions, each question relating to what they would expect to happen if we increase or decrease a single variable. This has been done to see if the participant filling in the questionnaire has an understanding of what they are being shown in the cases.

4.2.2 Applicant overview

We split each comparison in to a case giving the participant an applicant profile, which shows them their variables from $x^{(a)}$ with a short explanation of what each variable in this table means for them, an example of this can be seen in figure 3.

4.2.3 Depiction of the explanation

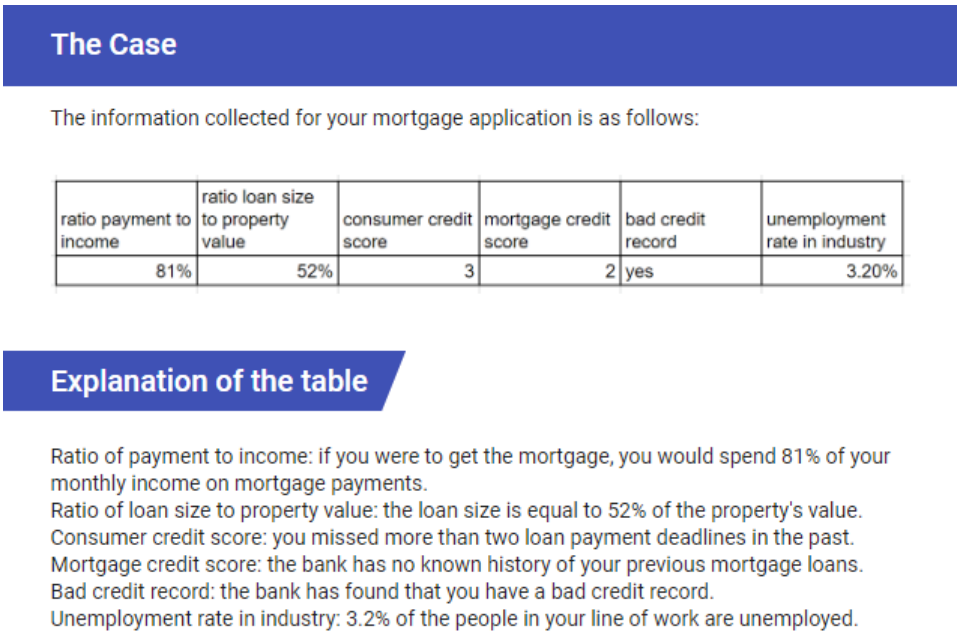
After reading the case the participant reads that his mortgage application was denied (everyone is denied). After reading the decision, the explanation is presented to the participant. This explanation consists of a small paragraph explaining what the participant is shown, followed by a table where the participant is shown his case compared to that of the person he is being compared to. The contributing causes are shown in red. The variables which counteracted the decision are colored in green (in case the participant is filling in questionnaire 3). This is followed by the explanation itself which summarizes on which parts the applicant got rejected, naming the most important ones first. An example of this can be seen in figure 4.

4.2.4 Questions

Finally the participant is asked to rate several aspects of the explanation. Our aim for the questionnaire was to find out the following:

1. How well received are the explanations?
2. Which of the distance functions is preferred when it comes to the reference object.
3. Which of the reference object types is better received when it comes to the reference object.
4. Are some variables more important to people than others in the generated explanation.

Firstly he/she is asked "How satisfied are you with explanation (A/B/C)?" on a scale of 1-5, with the lower end labeled as "Completely unsatisfied" and the higher end labeled "Completely satisfied". Secondly the participant is asked to rate the reference object: "Do you think the



The Case

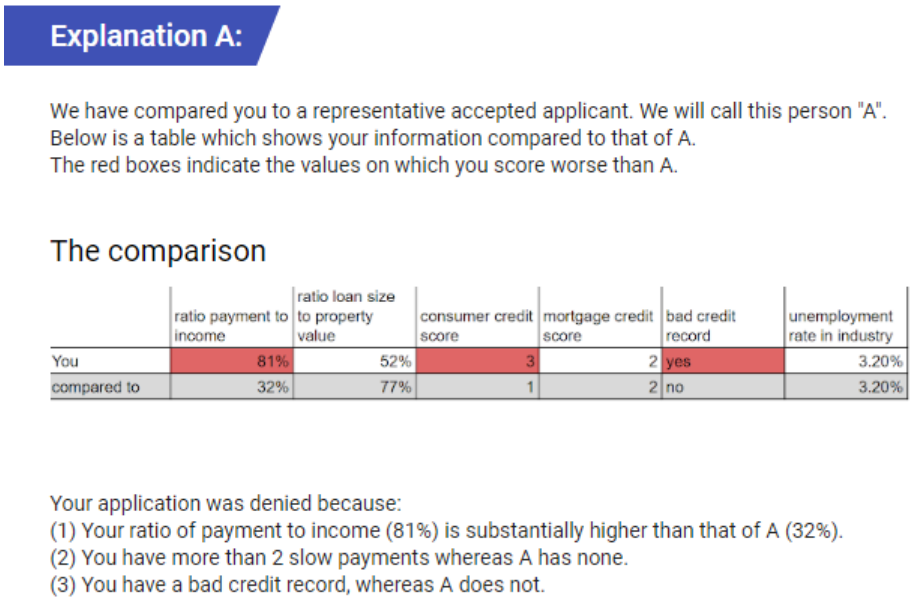
The information collected for your mortgage application is as follows:

ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
81%	52%	3	2	yes	3.20%

Explanation of the table

Ratio of payment to income: if you were to get the mortgage, you would spend 81% of your monthly income on mortgage payments.
Ratio of loan size to property value: the loan size is equal to 52% of the property's value.
Consumer credit score: you missed more than two loan payment deadlines in the past.
Mortgage credit score: the bank has no known history of your previous mortgage loans.
Bad credit record: the bank has found that you have a bad credit record.
Unemployment rate in industry: 3.2% of the people in your line of work are unemployed.

Figure 3: Example of a case as presented to a questionnaire participant



Explanation A:

We have compared you to a representative accepted applicant. We will call this person "A".
Below is a table which shows your information compared to that of A.
The red boxes indicate the values on which you score worse than A.

The comparison

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	32%	77%	1	2	no	3.20%

Your application was denied because:
(1) Your ratio of payment to income (81%) is substantially higher than that of A (32%).
(2) You have more than 2 slow payments whereas A has none.
(3) You have a bad credit record, whereas A does not.

Figure 4: Example of a case comparison and explanation as presented to a questionnaire participant.

comparison to person (A/B/C) is relevant?". The scale is again 1-5 with the labels changed for the lower end to "Completely irrelevant" and the higher end labeled "Completely relevant". Thirdly we ask the participant to rate each individual cause and if they find it to be a valid reason for denial worded as:'In your opinion, is your score on "ratio payment to income"(e.g.) a valid reason to deny your application'. The participant get a scale of 1-5 with the lower end labeled "Not at all" and the higher end labeled "Very much so" (Coleman, 2018). Finally the participant is asked if they have any suggestions, which they can freely fill in if they so desire (this is the only optional question in the questionnaire). An example of all types of questions can be seen in figure 5.

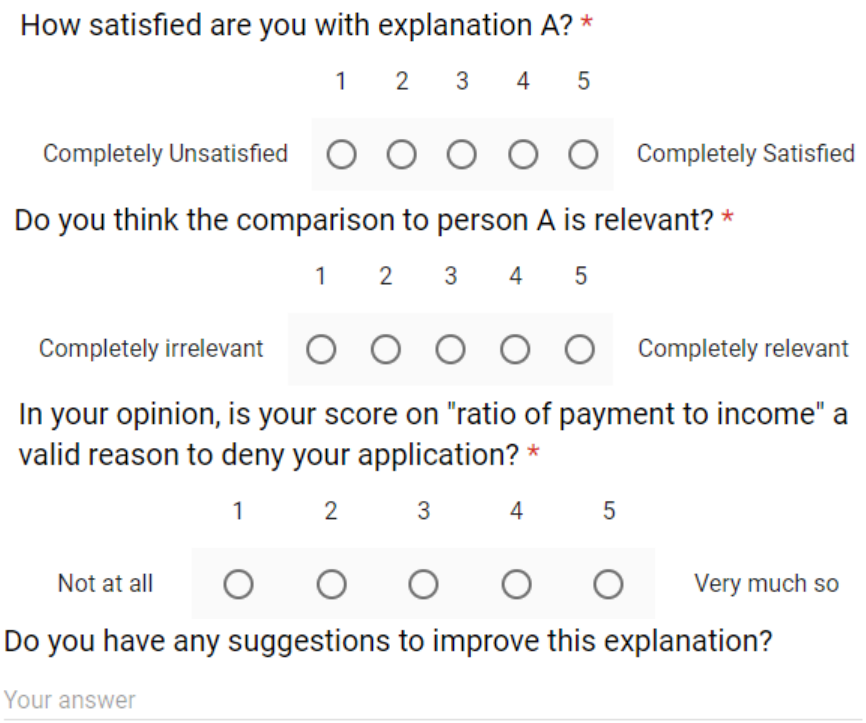


Figure 5: Example of all types of question for the participant.

5 Methods of analysis

In this section we will go through the analyses methods used to find the results shown in section 6. As we have three different questionnaires and not all have the same number of responses our data for the majority of the comparisons is based on the percentage of people answered. All required questions were made as a Likert scale type question, it is argued that this gives us the ability to also use the mean as a measurement but is not universally agreed on for this reason we ran four analysis. Two depicting the answers as ordinal and the other two depicting the answers as numeric.

5.1 Methods

Due to the small number of participants (40) we chose to more in depth with our analysis and try to find out where possible effects are visible. Due to this we consider multiple ways to analyse the questionnaire results.

5.1.1 Cumulative distribution

Our first analysis looks at the cumulative distribution of the responses. This analysis is used to see if we can see if one of the compared distance functions or reference objects is scoring consistently lower than the other. This is done the following way: If we take the question "How satisfied are you with explanation A?" comparing on Gowers vs Simplex as an example, we count the number of responses received for each distance function. We then proceed to go from 1 towards 5 for each separate distance function with the count for rating 1 being the number of people that rated this explanation as a 1. Rating 2 then gets shown as the count of both rating 1 and rating 2 combined, then continue doing this until we reach rating 5 where we should end up with the total number of answers as the total count. Since we might have a different number of participants for different distance functions or reference objects we use percentage of total answers. In the mathematical sense we are trying to find if the cumulative conditional probability given condition A is always lower than given condition B. So we can state that if:

$$\begin{aligned} P(y \leq j|A) \leq P(y \leq j|B) \\ \forall j = 1, \ldots, 5 \end{aligned} \tag{7}$$

then $P(y|A)$ is "stochastically larger" than $P(y|B)$ and in that sense we can state that A tends to produce better explanations than B (if in this case y represents the quality of the explanation).

Visually, this means that the cumulative distribution of y given A lies completely below the cumulative distribution of y given B.

5.1.2 Within subject dependent t-test

We use the dependent t-test (also called a paired sample t-test) to compare the mean difference between our samples to the difference we would expect to find between population means, and then takes into account the standard error of the differences. This means we compare the means of our two different conditions and the standard error to determine if the differences in means is likely or not.

5.1.3 Pearson chi-square

The Pearson chi-square is used to check for individual effects of variables on our predictions, this way we can see if any single effect has a strong correlation with the rating provided by the participant. In this case we try to find the most parsimonious model which is not significant in our Pearson and likelihood ratio test. The eventual model can be viewed as a mosaic plot

and through this we can see the effect sizes in form of the standard residual if there are any significant differences from our null hypothesis. If any of these differences occur we can state that there is an effect in that variable. We use the same outcome variables y for this method.

5.1.4 One-way independent ANOVA

The one-way independent ANOVA is used for finding effects for single independent variables. We use it for the same reason we used Independent factorial ANOVA for the logistic regression only in this case to compare it with the Pearson chi-square. Same as the Pearson chi-square we use this ANOVA to test for single independent variables and if they have a significant influence on their own. The main difference is the output is processed as a continuous variable instead of a factor making it possible to see smaller changes more easily. Before building the regression model we first use Levene's test to verify if the assumptions are satisfied. Levene's test assesses the equality of variances for two or more groups, which is an assumption that the one-way independent ANOVA has. If the Levene's test is non significant and therefore the assumption holds, we can continue building the linear model which will be used for analysis. In the analysis we use as y values the explanation rating and reference object rating, and a single predictor variable (reference type, distance function or number of contributing causes). If the Levene's test is significant and therefore the assumption is not met we can use Welch's F test to adjust for the differences in group variances. If in Welch's F test we do find a significance we can state that there was a significant difference between the two different variables.

5.1.5 Ordinal logistic regression

For looking at the combined effects of both the reference object type and the distance function used when considering the replies as an ordinal variable and not a continuous we use a logistic regression model. In this case our y value consists of either the rating given on the question "How satisfied are you with explanation (A/B/C)?" or "Do you think the comparison to person (A/B/C) is relevant?" as a factor, depending on what we are testing for. As the predictor variables we use the distance function used (factor), the reference object type(factor), the number of errors a person made in the test questions(numeric) and the number of contributing causes that were given(numeric). The result will be a model which consists of multiple separating lines (in our case 4 lines as we got 5 different factor levels). This gives us a total of 5 areas which are coinciding with the possible answers, this way we can calculate with the input in which area a certain combination will most likely fall. With the ordinal logistic regression model built we look at the coefficients and if any are significant. Another thing we look for is the signs of the coefficients if these coincide with what we would expect. After making the model we do not automatically get the p-values, to acquire this we compared the t-value against the standard normal distribution. This calculation is not completely accurate but gives us a good estimate to the p-value. The formula for the ordinal logistic regression is show in equation 8. In the formula we calculate the proportional odds j for the class y given x, we calculate this using the intercept t for j using the coefficients $\beta^\top$ for variable x (Long and Freese, 1997). We use the polr function in the R package MASS to do this for us.

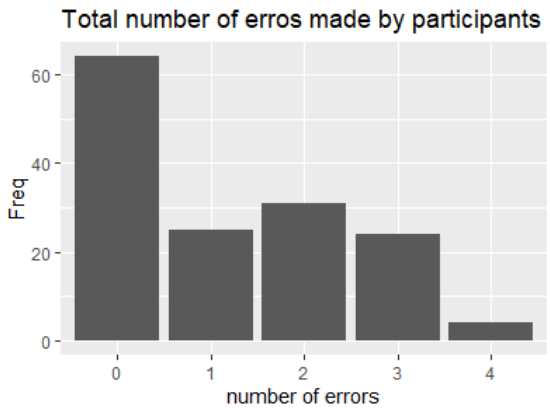
$$P(y \leq j|x) = \frac{exp(t_j - \beta^\top x)}{1 + exp(t_j - \beta^\top x)} \tag{8}$$

5.1.6 Independent factorial ANOVA/multiple regression

A factorial ANOVA is an ANOVA which uses more than one independent variable and which uses a prediction output which is continuous. We use this analysis to see if we have different results compared to the ordinal logistic regression when we see the output as a continuous variable instead of a factor as the participants have answered. Reasoning behind this is the fact

that we might find differences in outputs which we can miss if we use the factors due to our low number of participants. Instead of the logistic model used in the ordinal logistic regression we use a linear model for the ANOVA to determine if there are any significant variables. We take the same outcome types for y (explanation rating/comparison rating), we also use the same predictors as with the logistic regression (distance function, reference type, number of contributing causes and errors made on test questions). The standard model for the factorial ANOVA is shown in equation 9. The y value given variable x_i and z_i is calculated with the intercept β_0 and the coefficients for single variables are β_1 and β_2 and the interaction effects coefficient uses both variables calculated with β_3. The coefficients are the sum of squares of the variable x_i, z_i and the interaction effect $(x_i \times z_i)$ (Field et al., 2012, p. 403).

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \times z_i) \tag{9}$$



number of errors	Freq
0	64
1	25
2	31
3	24
4	4

Figure 6: Total Number of errors made by participants

6 Results

In this chapter we present the results we found by performing the analyses discussed in chapter 5.

6.1 Questionnaire results

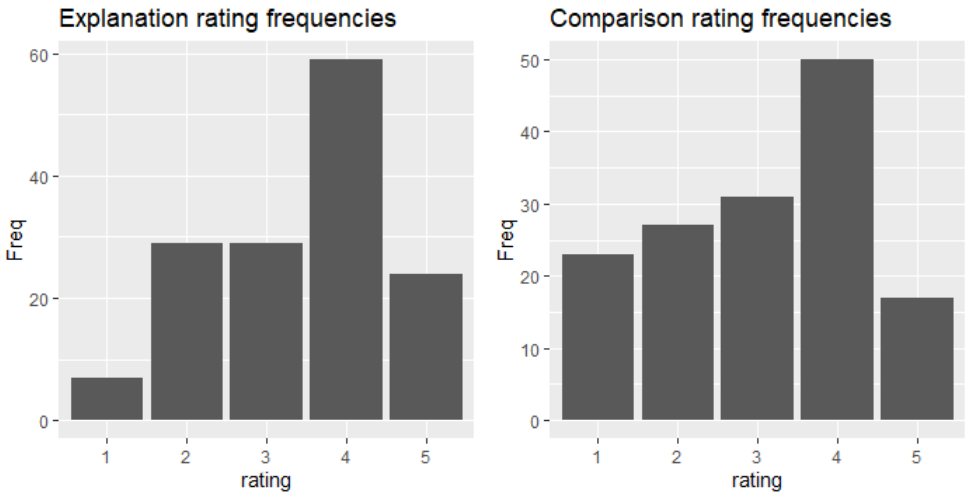
We had a total of 40 respondents on our questionnaire divided over the different questionnaires as: questionnaire 1 (14), questionnaire 2 (14) and questionnaire 3 (12). Most participants were between the age of 24-36 and were either currently studying or started their career. We summed the total number of errors made by participants as shown in figure 6. In figure 7 the ratings given by the participants to the question "How satisfied are you with explanation (A/B/C)" and the rating given to the question "Do you think the comparison to person (A/B/C) is relevant". In both cases the mode is at 4.

6.2 Cumulative distribution

We used the cumulative distribution to find a stochastically larger effect in single independent variables.

6.2.1 Explanation rating with reference object type

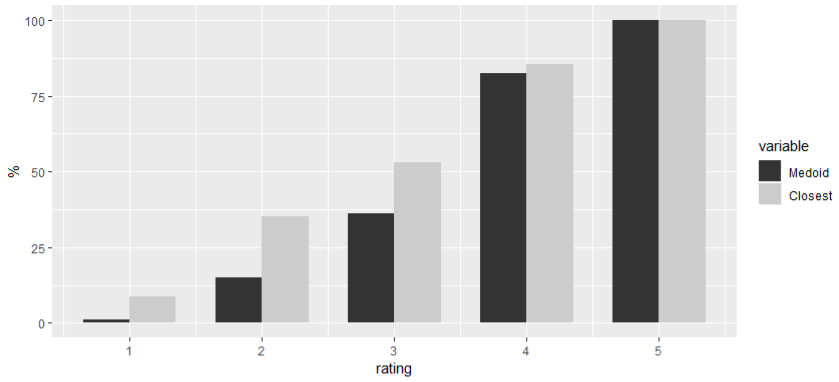
If we look at figure 8 we can clearly see for closest type keeps it's cumulative percentage above that of the medoid type reference object. Medoid in our case seems to perform better over closest in the overall cases when it comes to the quality of the explanation as rated by participants.



rating	Freq
1	7
2	29
3	29
4	59
5	24

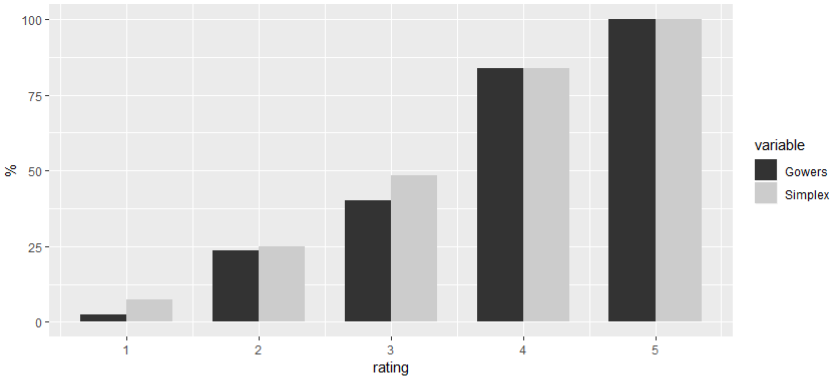
rating	Freq
1	23
2	27
3	31
4	50
5	17

Figure 7: The frequencies of ratings given to the explanation and reference object type



rating	variable	%	variable	%
1	Medoid	1.25	Closest	8.82
2	Medoid	15.00	Closest	35.29
3	Medoid	36.25	Closest	52.94
4	Medoid	82.50	Closest	85.29
5	Medoid	100.00	Closest	100.00

Figure 8: Cumulative distribution for quality of explanation depending on the reference object type (closest or medoid)



rating	variable	%	variable	%
1	Gowers	2.50	Simplex	7.35
2	Gowers	23.75	Simplex	25.00
3	Gowers	40.00	Simplex	48.53
4	Gowers	83.75	Simplex	83.82
5	Gowers	100.00	Simplex	100.00

Figure 9: Cumulative distribution for quality of explanation depending on distance function

6.2.2 Explanation rating with distance function

If we look at figure 9, it is visible that the Simplex method just about stays above the cumulative percentage of Gowers' percentage. The effect seems to be less than that of the previous graph comparing reference object types. It still holds that Gowers is stochastically larger than Simplex, this would mean that Gowers is the preferred option of finding the reference object.

6.2.3 reference object rating

As can be seen in figure 10 and figure 11 in both cases we do not achieve a stochastically larger variable in the case of the reference object rating.

6.2.4 Contributing causes with and without counter-acting causes

In figure 12 we see that including counteracting causes in the explanation doesn't seem to improve its quality. Except for rating 1, the cumulative distribution of quality of explanations mentioning just contributing causes is below that of explanations including counteracting causes. The exception for rating 1 is due to a single respondent giving this rating for 1 explanation.

6.3 Within subject dependent t-test

We use the dependent t-test (in R function "t.test" with paired set to TRUE) to see if our subjects within the questionnaire displayed a preference for one of the reference type or distance function based on their explanation rating. The results of the t-tests can be seen in table 5. The comparisons that were done are shown in the earlier discussed figure 2 the dashed lines here are the comparisons we do within subject combining them per questionnaire. As seen none of the results test significant ($p < 0.05$). What we can take from the t-test is that when testing

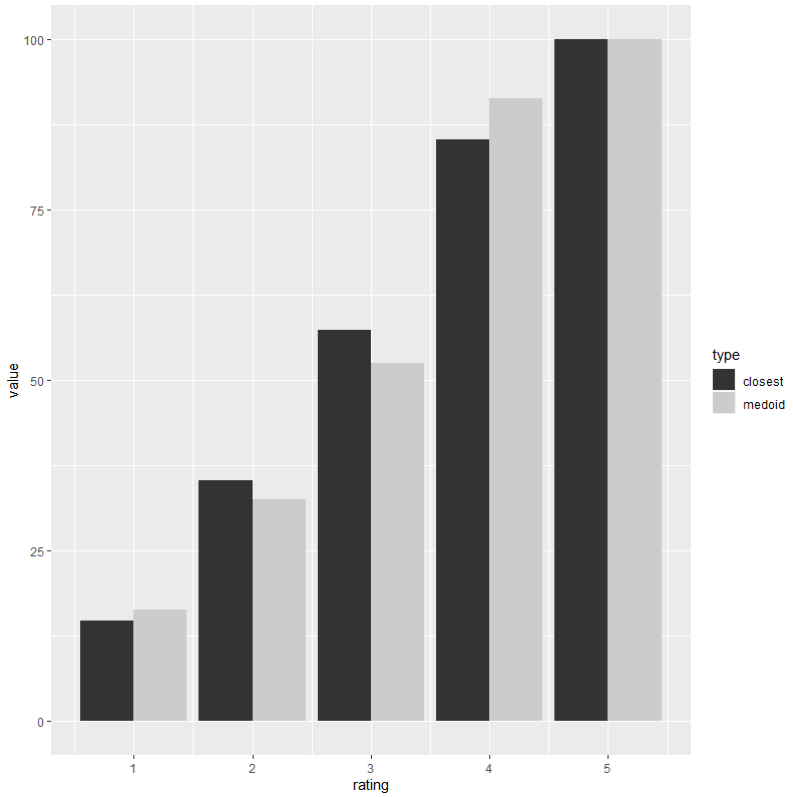


Figure 10: Cumulative distribution for quality of reference object depending on the reference object type (closest or medoid)

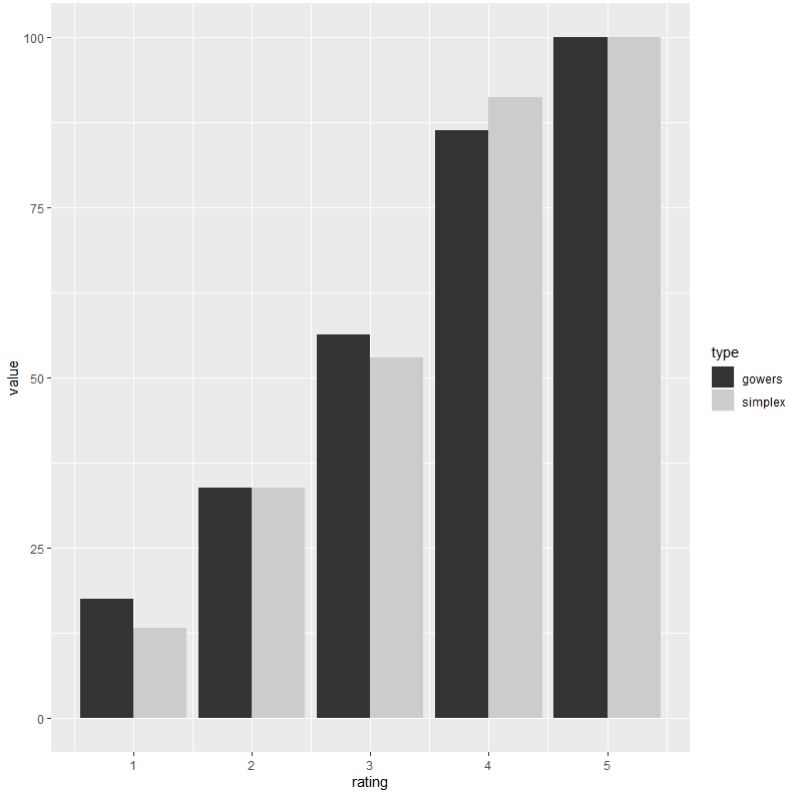
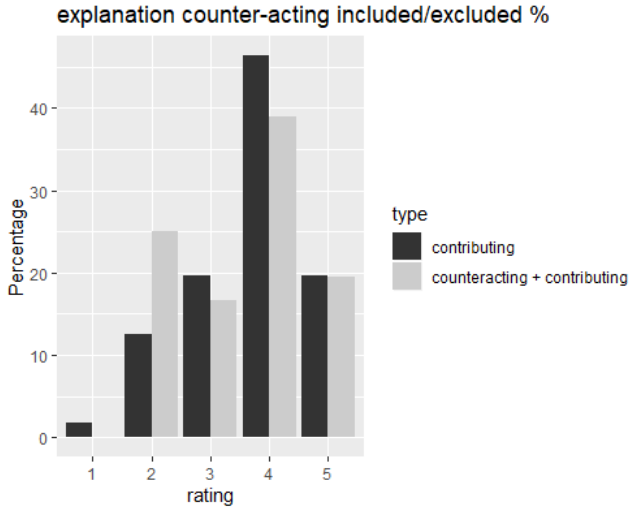
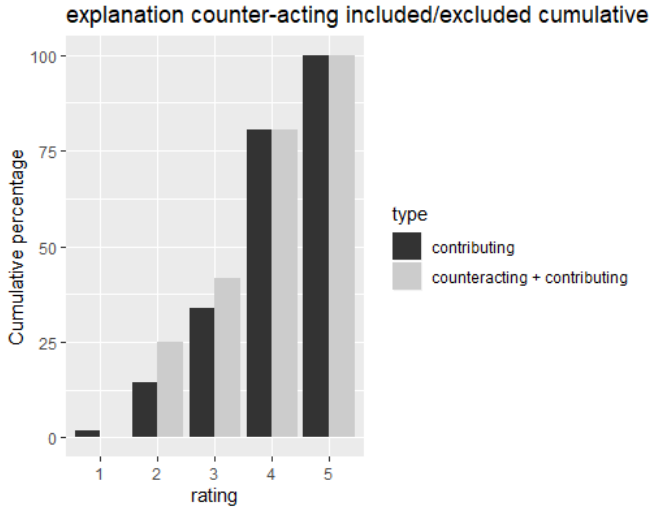


Figure 11: Cumulative distribution for quality of reference object depending on distance function (gowers or simplex)



rating	%	type
1	0.00	counteracting + contributing
2	25.00	counteracting + contributing
3	16.67	counteracting + contributing
4	38.89	counteracting + contributing
5	19.44	counteracting + contributing
1	1.79	contributing
2	12.50	contributing
3	19.64	contributing
4	46.43	contributing
5	19.64	contributing



rating	%	type
1	0.00	counteracting + contributing
2	25.00	counteracting + contributing
3	41.67	counteracting + contributing
4	80.56	counteracting + contributing
5	100.00	counteracting + contributing
1	1.79	contributing
2	14.29	contributing
3	33.93	contributing
4	80.36	contributing
5	100.00	contributing

Figure 12: explanation ratings comparison counter-acting and contributing causes

Questionnaire	Comparison	t-value	df	p-value
1	medoid/closest	0.66749	27	0.5101
2	gowers/simplex	0.44073	27	0.6629
3	medoid/closest	1.4832	11	0.1661
3	gowers/simplex	0.32063	11	0.7545

Table 5: Results of all dependent t-tests

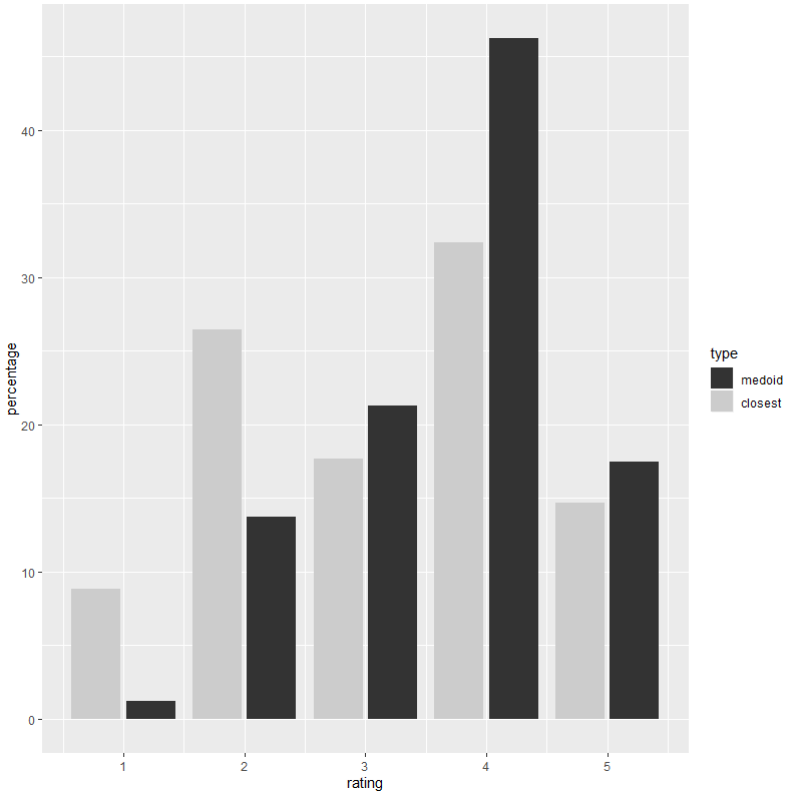


Figure 13: Percentage distribution of explanation rating for medoid and closest

within subject the t-value indicates the same tendency towards medoid and Gowers as analysis later in this section will show.

6.4 Pearson chi-square

For the Pearson chi-square analysis we look at the influence of reference type and distance function separately on the explanation rating and comparison rating. The complete test can be found in the appendix.

6.4.1 reference object influence on explanation rating

The chi-square test for reference object type against explanation rating also shows in this test to have a significant influence for when a person would have the medoid type for a good rating with $X^2(1) = 9.694136, p < .05$. If we then look at our percentage graph in figure 13 we can see that medoid scores better than the closest type in the higher ratings. The odds of scoring a rating 4 or higher are 1.98 (1.758621,0.888889) times higher when using the medoid reference object.

Test	$X^2(1)$	df	p-value
explanation | reference object	9.694136	4	0.04590745
explanation | distance type	3.727172	4	0.4441862
comparison | reference object	2.91621	4	0.571944
comparison | distance type	2.524658	4	0.6402259

Table 6: All chi-square test results

(center = median)	Df	F value	Pr(>F)
group	1	9.522	0.002429
	146		

Table 7: Levene's Test for Homogeneity of Variance (center = median)

6.4.2 Distance type influence on explanation rating

As shown in the table of all chi-square (table 6) results we can see that the influence that the different distance types on the explanation rating were non significant $p = 0.4441862 \neq p < 0.05$.

6.4.3 reference object influence on comparison rating

Just like with the previous test we do not see any significant effect for the reference type on the comparison rating. As shown in table 6 we have a non significant p-value $p = 0.571944 \neq p < 0.05$.

6.4.4 Distance type influence on comparison rating

The influence that the different distance types had on the comparison rating also tested non significant, $p = 0.6402259 \neq p < 0.05$. This also indicates that there was no relation found between the distance type used and the comparison rating.

6.5 One-way independent ANOVA

For the One way ANOVA we take a look at the separate independent variables to see if they have any influence. Same as with the Pearson chi-square analysis we take a look at the influence of distance function and reference type separately on the explanation or comparison rating. For the one-way independent ANOVA we do the Levene's test to see if the predictor is significant. If it is we apply Welch's F test, else we can build a linear model to see if the predictor is significant.

6.5.1 Explanation rating one-way independent ANOVA

For our explanation rating predicted by the distance function Levene's test was insignificant with $p = 0.3184$. And showed in the linear model to be insignificant with $p = 0.428$. The explanation rating predicted by reference type had a significant Levene's test shown in table 7. This tells us that when we use the reference type as a predictor changing the reference type has a big influence on changing the outcome variable (explanation rating). If we do the follow up Welch's F we get another significant return with $F = (1, 126.3) = 6.5751, p = .01151$. This indicates that even after adjustments that reference type still has a significant influence on the explanation rating, which is in line with our earlier analysis.

6.5.2 Comparison rating one-way independent ANOVA

For the comparison rating predicted by distance function, Levene's test came back insignificant $p = 0.6126$. As with previous results the generated linear model shown in table 8 shows that

	Df	Sum Sq	Mean sq	F value	Pr(>F)
Distance function	1	0.0	0.0243	0.015	0.903
Residuals	146	236.16	1.6175		

Table 8: Levene's test One-way ANOVA comparison rating/distance function

	Df	Sum Sq	Mean sq	F value	Pr(>F)
Reference type	1	0.0	0.0001	0	0.994
Residuals	146	236.2	1.6177		

Table 9: Levene's test One-way ANOVA comparison rating/reference type

the distance function has no significant effect on the comparison rating with $p = 0.903$. The F-value in the levene's test also indicates that the means are almost identical to each other. The comparison rating predicted by reference type also shows an insignificant Levene's test $p = 0.8376$. From this linear model just as the previous result shows a non significant effect in the prediction of the comparison rating by reference type as shown in table 9 with $p = 0.994$. We can also see from the F-value that the means are almost identical just as the previous result.

6.6 Ordinal logistic Regression

We use the ordinal logistic regression to find significant effects when we have multiple independent variables which need to predict an ordinal outcome.

6.6.1 Explanation rating logistic model

In the ordinal logistic model in table 10 we see that the reference type is significant with $p = 0.0018$, where the coefficient indicates that the reference type medoid scores better than the closest. Another significant variable in this test is the number of contributing causes in the explanation ($p = 0.0422$) with the coefficient being negative indicates that when participants were shown an explanation with less contributing causes they are more likely to rate the explanation positively. Another variable of note is the errors made, it's close to being significant and according to the coefficient people that made more errors at the start are more likely to give the explanation a lower rating.

Coefficients:	value	Std. Error	t value	p-value
Distance type.simplex	-0.2204	0.3062	-0.7199	0.4715992
Reference type.medoid	1.1911	0.3807	3.1288	0.0017553
Number of contributing causes	-0.5059	0.2490	-2.0315	0.0422030
Number of errors	-0.2219	0.1236	-1.7953	0.0726136

Intercepts:	value	Std. Error	t value
1—2	-4.0073	0.6555	-6.1138
2—3	-2.0762	0.5564	-3.7317
3—4	-1.1258	0.5423	-2.0761
4—5	0.8757	0.5406	1.6200

Table 10: The summary of the logistic regression model for explanation rating

Coefficients:	value	Std. Error	t value	p-value
Distance type.simplex	0.06706095	0.2988123	0.2244250	0.8224266418
Reference type.medoid	0.19435939	0.3554711	0.5467656	0.5845397668
Number of contributing causes	0.21910064	0.2412739	-0.9080991	0.3638258464
Number of errors	0.07728198	0.1172310	-0.6592285	0.5097490473

Intercepts:	value	Std. Error	t value
1—2	-2.16547305	0.5583698	-3.8782058
2—3	-1.14283200	0.5360185	-2.1320756
3—4	-0.27555674	0.5305947	-0.5193356
4—5	1.58763960	0.5570865	2.8498976

Table 11: The summary of the logistic regression model for comparison rating

(center = median)	Df	F value	Pr(>F)
group	3	1.5612	0.2014
	144		

Table 12: Levene's Test for Homogeneity of Variance (center = median)

6.6.2 Comparison rating logistic model

Compared to the explanation rating the model for the comparison rating does not include any significant variables (see table 11).

6.7 Independent factorial ANOVA/multiple regression

The factorial ANOVA is used to make a linear model of both the explanation and comparison ratings.

6.7.1 Explanation factorial ANOVA

Our ANOVA is fit for the explanation rating predicted by the reference type, distance function, number of contributing causes and the number of errors made at the background questions test. If we look at the Levene's test in table 12 we can conclude that due to $F(3, 144) = 1.56, p = 0.2014$ is not significant, that the assumptions of the ANOVA are met. The ANOVA's type III test is shown in table 13. In the table we can see that the main effect with a significance of ($p < 0.01$) is the reference type for its influence on the explanation rating and a minor significance of ($p < 0.05$) in the number of contributing causes. Just like the previous results other predictors are non significant. In comparison to the ordinal logistic regression we see the same significant results.

6.7.2 Comparison factorial ANOVA

Same as with the previous ANOVA we first need to see if our assumptions hold by looking if our Levene's test in table 14 is significant or not. Our assumption holds again as $F(3, 144) = 0.0882, p = 0.9664$ is not significant. This means we can now look at the type III test again to determine the important predictors. In the type III test we can see that none of the predictors are significant so we can not state that any is important enough as a predictor, this also coincides with the results found earlier.

Response: explanation	Sum Sq	Df	F value	Pr(>F)
(Intercept)	186.261	1	159.9888	<2.2e-16
Reference type.medoid	14.415	1	12.3818	0.0005817
Distance type.simplex	0.735	1	0.6314	0.4281761
Number of contributing causes	5.545	1	4.7630	0.0307110
Number of errors	3.065	1	2.6325	0.1069015
Residuals	166.483	143		

Table 13: ANOVA table (Type III tests) explanation

(center = median)	Df	F value	Pr(>F)
group	3	0.0882	0.9664
	144		

Table 14: Levene's Test for Homogeneity of Variance (center = median)

Response: comparison	Sum Sq	Df	F value	Pr(>F)
(Intercept)	136.820	1	83.4467	5.778e-16
Reference type.medoid	0.432	1	0.2634	0.6086
Distance type.simplex	0.088	1	0.0535	0.8175
Number of contributing causes	1.331	1	0.8119	0.3691
Number of errors	0.422	1	0.2576	0.6126
Residuals	234.464	143		

Table 15: ANOVA table (Type III tests) comparison

6.8 Conclusion

Explanation based on medoid objects of reference tend to be rated higher than explanation based on "most similar" objects of reference. Something to note is that participants did not show the same effect when asked directly if the reference object was relevant to their case. This could be due to the fact that the participants do not link the explanation rating to that of the comparison used and only look at the explanation separate from the comparison person. Another possibility is that the participants overall found the reference objects too close to each other and might not have been able to significantly differentiate them from each other.

The ordinal logistic regression gave us a significant result on the number of errors people made on the the test. This indicated that people were giving worse ratings if they made more errors on the test. For the non significant results of the distance function we do see the coefficients that we expected to see from our participants where they find the easier comparison of Gower's distance to be better rated than the more complex Simplex method. Another observation can be made for the number of causes presented to the user. If we look at the coefficients given to the number of contributing causes we can see that the rating tends to be higher if less causes are presented. This is in line with Feelders, as the parsimoneous set is used to prevent the less crucial causes to be filtered out preventing info overload for the user. We can see this effect in the results.

Participants were also able to fill in an open field on the questionnaire which asked how they think we could improve the explanation. Eight (20%) people noted that in the case of our mortgage question they wouldn't want to be compared to another person as they see their application as a singular case. A positive note is that when the medoid was used as a reference type most people voted for a rating of 3 or higher indicating to us that the explanation might not be passable right now as we can consider a rating of 3 to also indicate "no opinion" or "neutral", but does not get completely rejected. Looking at our research questions we can see if we are able to answer them now:

1. Which of the considered objects of reference produces the best explanations?
2. Which of the considered distance functions provides us with the best reference objects, going by the provided end result?
3. How does the inclusion or exclusion of counteracting causes influence the quality of the provided explanations?

1. If we look at the results of the analysis we see that the medoid type scores significantly better than the closest type when it comes to the explanation rating. The results of the question if people were satisfied with the comparison person did not give us significant results.

2. If we look to the analysis we can not exclude either of the options, we have not been able to find a significant difference in any of the analyses performed. Hence, it doesn't seem to be worthwhile to use the more complicated Simplex method in combination with the Mahalanobis distance. The simpler Gower's distance appears to work just as well

3. There is a slight indicator in the cumulative distribution of the counter-acting included and excluded analysis that only showing the contributing causes has a better effect on the ratings. This is not a significant indicator and might only be taken as an indicator for future research.

7 Discussion

In this section we will describe the implications of the results and discuss the (possible)shortcomings of the research done in this paper. From our conclusion we see that one of our three research questions has gotten an answer (which reference type is better closest/medoid). If we go back to the other two questions we have no certainties to indicate our different options are better than the other, with only a possible indicator in the counter-acting excluded way of explaining. An interesting point we did pick up in one of our analysis is that the number of contributing causes the explanation uses was influencing the rating with significance in the ordinal logistic model. One of our biggest issues in this research was the lack of participants in our questionnaire, for any research further on this subject which uses a questionnaire to be viable for a good analysis should consist around 100+ participants as the lack of data can be really shown when we take a look at our analysis. Due to this lack of data a lot of the analysis is looking for variables that could be significant and our overall analysis could very well be wrong. With a bigger set we would have been able to run a more precise analysis and go into greater detail as to what is important when it comes to the explanation.

Another point is the setup for the questionnaire for the setup we should have had a strict divide of within and between tests, our current implementation was chosen to minimise the number of participants needed for maximum number of comparisons, this in the end conflicted with the strict divide of within and between subject testing. It would have improved our analysis if we would have done only one independent variable being tested instead of multiple at the same time, this being the comparison of the non counteracting causes questionnaires (1/2) and the added counteracting causes questionnaire (3). The reason being that we combined the same cases with and without counteracting causes to at the same time compare them to the distance function or reference object types. Resulting in some single variable analysis not having an single independent variable but two.

The database that was used to create the questions for the questionnaire might have been to complex for the differences in explanations that we wanted to achieve. The main issue with this is that we tried to find a database which is understandable for the participants but might have ended up with a too complex explanation for the low level testing that was ran. We limited our research to only two options in each category (distance function/reference object type) this was mostly done for ease of comparison in future research a more broad approach would be needed to find what distance function or reference object type would work best.

A major point that didn't get looked into was how to present such an explanation when generated. We chose to go with the more raw type where we give the actually contributing cause with a very short clear description. If this were to change to a more natural type of explanation, which in the real world could be done by a subject expert, people might also increase their liking of the explanation provided due to the more natural way of it being presented to them.

References

Coleman, R. (2018). *Designing Experiments for the Social Sciences*, volume 1. Sage Publications Inc.

Drezner, Z. and Turel, O. (2011). Normalizing variables with too-frequent values using a Kolmogorov–Smirnov test: A practical approach. *Computers & Industrial Engineering*, 61(4):1240–1244.

Feelders, A. and Daniels, H. A. (2001). A general model for automated business diagnosis. *European Journal of Operational Research*, 130(3):623–637.

Field, A., Miles, J., and Field, Z. (2012). *Discovering Statistics Using R*. SAGE Publications.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, pages 857–871.

Hesslow, G. (1983). Explaining differences and weighting causes. *Theoria*, 49(2):87–111.

Humphreys, P. (1989). *The chances of explanation: Causal explanation in the social, medical, and physical sciences*, volume 1051. Princeton University Press.

Long, J. S. and Freese, J. (1997). *Regression Models for Categorical and Limited Dependent Variables*. Advanced Quantitative Techniques in the Social Sciences. SAGE Publications.

McCane, B. and Albert, M. (2008). Distance functions for categorical and mixed variables. *Pattern Recognition Letters*, 29(7):986–993.

Miller, T. (2018). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*.

Molnar, C. (2019). *Interpretable Machine Learning*. github. https://christophm.github.io/interpretable-ml-book/.

Munnell, A. H., Tootell, G. M., Browne, L. E., and McEneaney, J. (1996). Mortgage lending in Boston: Interpreting HMDA data. *The American Economic Review*, pages 25–53.

Robnik-Šikonja, M. and Kononenko, I. (2008). Explaining classifications for individual instances. *IEEE Transactions on Knowledge and Data Engineering*, 20(5):589–600.

van de Velden, M., Iodice D'Enza, A., and Markos, A. (2018). Distance-based clustering of mixed data. *Wiley Interdisciplinary Reviews: Computational Statistics*, page e1456.

Wilson, D. R. and Martinez, T. R. (1996). Value difference metrics for continuously valued attributes. In *Proceedings of the International Conference on Artificial Intelligence, Expert Systems and Neural Networks*, pages 11–14. Citeseer.

Appendix

Cases used

going from Questionnaire 1-3, comparison in blocks of 2 except for the 3 final ones which get compared together.

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	32%	77%	1	2	no	3.20%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	30%	80%	3	2	yes	3.20%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	40%	88%	6	1	yes	3.20%
compared to	35%	75%	2	2	no	3.90%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	40%	88%	6	1	yes	3.20%
compared to	32%	90%	5	1	yes	3.20%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	35%	75%	2	2	no	3.90%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	32%	77%	1	2	no	3.20%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	34%	100%	2	4	yes	4.30%
compared to	40%	90%	2	3	yes	3.60%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	34%	100%	2	4	yes	4.30%
compared to	34%	71%	3	4	yes	3.20%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	35%	75%	2	2	no	3.90%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	32%	77%	1	2	no	3.20%

	ratio payment to income	ratio loan size to property value	consumer credit score	mortgage credit score	bad credit record	unemployment rate in industry
You	81%	52%	3	2	yes	3.20%
compared to	30%	80%	3	2	yes	3.20%

Pearson chi-square tables

full tables of the Pearson chi-square test done in the analysis

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  148

                     | Survey.dat$Ctype
Survey.dat$explanation |  closest  |   medoid  | Row Total |
-----------------------|-----------|-----------|-----------|
                     1 |         6 |         1 |         7 |
                       |     3.216 |     3.784 |           |
                       |     2.409 |     2.048 |           |
                       |    85.714% |    14.286% |     4.730% |
                       |     8.824% |     1.250% |           |
                       |     4.054% |     0.676% |           |
                       |     1.552 |    -1.431 |           |
-----------------------|-----------|-----------|-----------|
                     2 |        18 |        11 |        29 |
                       |    13.324 |    15.676 |           |
                       |     1.641 |     1.395 |           |
                       |    62.069% |    37.931% |    19.595% |
                       |    26.471% |    13.750% |           |
                       |    12.162% |     7.432% |           |
                       |     1.281 |    -1.181 |           |
-----------------------|-----------|-----------|-----------|
                     3 |        12 |        17 |        29 |
                       |    13.324 |    15.676 |           |
                       |     0.132 |     0.112 |           |
                       |    41.379% |    58.621% |    19.595% |
                       |    17.647% |    21.250% |           |
                       |     8.108% |    11.486% |           |
                       |    -0.363 |     0.334 |           |
-----------------------|-----------|-----------|-----------|
                     4 |        22 |        37 |        59 |
                       |    27.108 |    31.892 |           |
                       |     0.963 |     0.818 |           |
                       |    37.288% |    62.712% |    39.865% |
                       |    32.353% |    46.250% |           |
                       |    14.865% |    25.000% |           |
                       |    -0.981 |     0.905 |           |
-----------------------|-----------|-----------|-----------|
                     5 |        10 |        14 |        24 |
                       |    11.027 |    12.973 |           |
                       |     0.096 |     0.081 |           |
                       |    41.667% |    58.333% |    16.216% |
                       |    14.706% |    17.500% |           |
                       |     6.757% |     9.459% |           |
                       |    -0.309 |     0.285 |           |
-----------------------|-----------|-----------|-----------|
          Column Total |        68 |        80 |       148 |
                       |    45.946% |    54.054% |           |
-----------------------|-----------|-----------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  9.694136     d.f. =  4     p =  0.04590745


Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.04753064


       Minimum expected frequency: 3.216216
Cells with Expected Frequency < 5: 2 of 10 (20%)
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  148

                     | Survey.dat$Dtype
Survey.dat$explanation |   gowers  |   simplex  | Row Total |
-----------------------|-----------|------------|-----------|
                     1 |         2 |          5 |         7 |
                       |     3.784 |      3.216 |           |
                       |     0.841 |      0.989 |           |
                       |    28.571% |    71.429% |     4.730% |
                       |     2.500% |      7.353% |           |
                       |     1.351% |      3.378% |           |
                       |    -0.917 |      0.995 |           |
-----------------------|-----------|------------|-----------|
                     2 |        17 |         12 |        29 |
                       |    15.676 |     13.324 |           |
                       |     0.112 |      0.132 |           |
                       |    58.621% |    41.379% |    19.595% |
                       |    21.250% |    17.647% |           |
                       |    11.486% |     8.108% |           |
                       |     0.334 |     -0.363 |           |
-----------------------|-----------|------------|-----------|
                     3 |        13 |         16 |        29 |
                       |    15.676 |     13.324 |           |
                       |     0.457 |      0.537 |           |
                       |    44.828% |    55.172% |    19.595% |
                       |    16.250% |    23.529% |           |
                       |     8.784% |    10.811% |           |
                       |    -0.676 |      0.733 |           |
-----------------------|-----------|------------|-----------|
                     4 |        35 |         24 |        59 |
                       |    31.892 |     27.108 |           |
                       |     0.303 |      0.356 |           |
                       |    59.322% |    40.678% |    39.865% |
                       |    43.750% |    35.294% |           |
                       |    23.649% |    16.216% |           |
                       |     0.550 |     -0.597 |           |
-----------------------|-----------|------------|-----------|
                     5 |        13 |         11 |        24 |
                       |    12.973 |     11.027 |           |
                       |     0.000 |      0.000 |           |
                       |    54.167% |    45.833% |    16.216% |
                       |    16.250% |    16.176% |           |
                       |     8.784% |     7.432% |           |
                       |     0.008 |     -0.008 |           |
-----------------------|-----------|------------|-----------|
          Column Total |        80 |         68 |       148 |
                       |    54.054% |    45.946% |           |
-----------------------|-----------|------------|-----------|


Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  3.727172     d.f. =  4     p =  0.4441862



Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.4616081


       Minimum expected frequency: 3.216216
Cells with Expected Frequency < 5: 2 of 10 (20%)
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  148
```

Survey.dat$comparison	Survey.dat$Ctype: closest	medoid	Row Total
1	10	13	23
	10.568	12.432	
	0.030	0.026	
	43.478%	56.522%	15.541%
	14.706%	16.250%	
	6.757%	8.784%	
	-0.175	0.161	
2	14	13	27
	12.405	14.595	
	0.205	0.174	
	51.852%	48.148%	18.243%
	20.588%	16.250%	
	9.459%	8.784%	
	0.453	-0.417	
3	15	16	31
	14.243	16.757	
	0.040	0.034	
	48.387%	51.613%	20.946%
	22.059%	20.000%	
	10.135%	10.811%	
	0.201	-0.185	
4	19	31	50
	22.973	27.027	
	0.687	0.584	
	38.000%	62.000%	33.784%
	27.941%	38.750%	
	12.838%	20.946%	
	-0.829	0.764	
5	10	7	17
	7.811	9.189	
	0.614	0.522	
	58.824%	41.176%	11.486%
	14.706%	8.750%	
	6.757%	4.730%	
	0.783	-0.722	
Column Total	68	80	148
	45.946%	54.054%	

```
Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2.91621     d.f. =  4     p =  0.571944


Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.5741826

      Minimum expected frequency: 7.810811
```

```
   Cell Contents
|-------------------------|
|                   Count |
|         Expected Values |
| Chi-square contribution |
|             Row Percent |
|          Column Percent |
|           Total Percent |
|            Std Residual |
|-------------------------|

Total Observations in Table:  148
```

Survey.dat$comparison	Survey.dat$Dtype: gowers	simplex	Row Total
1	14	9	23
	12.432	10.568	
	0.198	0.233	
	60.870%	39.130%	15.541%
	17.500%	13.235%	
	9.459%	6.081%	
	0.445	-0.482	
2	13	14	27
	14.595	12.405	
	0.174	0.205	
	48.148%	51.852%	18.243%
	16.250%	20.588%	
	8.784%	9.459%	
	-0.417	0.453	
3	18	13	31
	16.757	14.243	
	0.092	0.109	
	58.065%	41.935%	20.946%
	22.500%	19.118%	
	12.162%	8.784%	
	0.304	-0.329	
4	24	26	50
	27.027	22.973	
	0.339	0.399	
	48.000%	52.000%	33.784%
	30.000%	38.235%	
	16.216%	17.568%	
	-0.582	0.632	
5	11	6	17
	9.189	7.811	
	0.357	0.420	
	64.706%	35.294%	11.486%
	13.750%	8.824%	
	7.432%	4.054%	
	0.597	-0.648	
Column Total	80	68	148
	54.054%	45.946%	

```
Statistics for All Table Factors


Pearson's Chi-squared test
------------------------------------------------------------
Chi^2 =  2.524658     d.f. =  4     p =  0.6402259


Fisher's Exact Test for Count Data
------------------------------------------------------------
Alternative hypothesis: two.sided
p =  0.6550356


       Minimum expected frequency: 7.810811
```