UTRECHT UNIVERSITY

School of Natural Sciences

Applied Data Science Master Thesis

Timing is Everything: How the Timing of Market Approach Can Influence the Number of Registrations in Procurement Contracts for Maintenance Projects

First examiner:
Arno Siebes

Candidate:
Ayman el Hattache (5021189)

Second examiner:
Thijs van Ommen

In cooperation with:
Rijkswaterstaat

Team mate:
Mohamed Hajji (7450575)

Company supervisor:
Johan Kessels

June 30, 2024

In the name of Allah, The Most Gracious, The Most Merciful

Abstract

Rijkswaterstaat (RWS) is a Dutch governmental organization, responsible for the construction, management and maintenance of the Netherlands' main infrastructure. To do this, RWS offers procurement contracts through TenderNed, a Tender platform, allowing for contractors to bid on these contracts. However, since a few years RWS struggles with low registrations on these contracts, leaving them often with unfavorable terms to offer. In this study, there will be a focus on determining the optimal timing for market approach for RWS to maximize the number of registrations on their contracts, based on historical tender analysis.

In the study, it became clear that many industries have clear seasonal patterns in the timing of their tender publications. However, based on statistical testing, it also became clear that for the majority of the industries, there is no significant difference in number of registrations between certain periods in time. This was also confirmed by literature, which mentioned that market approach time is on the lower end off significant influences for the tendering authority.

In the broader context, this research also showed that there is a lot of data available about the Dutch tendering market, which is not only beneficial for RWS, but for many other (semi-) public entities such as Municipalities or Educational institutions. In practice, the use of data in these tendering procedures is still not matured enough yet, and could be more applied in the future.

1. Preface - Shared

This thesis contains a shared research part and therefore several sections also have shared authorship. The goal of this project was to investigate several scenarios for Rijkswaterstaat Programma's, Projecten en Onderhoud (RWS PPO). Both students, Ayman el Hattache (5021189) and Mohamed Hajji (7450575) have used the same dataset and did a significant amount of work together, resulting in shared content. Each section or subsection is labeled to indicate whether it was written together (shared) or individually (individual).

Contents

2. Introduction

2.1 Context - Shared

Rijkswaterstaat (RWS) is a major Dutch governmental organization responsible for the construction, management, and maintenance of the Netherlands' main infrastructure networks, including roads, waterways, and water systems, ensuring safety, accessibility, and environmental sustainability [1]. RWS runs various new construction and maintenance projects annually to achieve these goals. In 2023, RWS received a €6,6 billion budget from the Ministry of Infrastructure and Water Management, of which €2,9 billion was spent on new construction projects and €2,0 billion on maintenance projects [2]. To perform these projects, RWS offers them as a contract for contractors to bid on. RWS offers these procurement contracts through TenderNed, a platform that is the central place for publishing procurement contracts in the Netherlands. The platform promotes a transparent landscape and ensures a fair marketplace for every bidder and contractor.

However, RWS is not the only entity that offers procurement contracts through TenderNed; other public entities like provincial and municipal authorities also use the platform to publish theirs. Most of the published tenders are performance contracts, with a maximum duration of 4 years. Therefore, they require careful planning to ensure a smooth transition between one contract's end and another's start.

One way to improve tendering is by learning lessons from the past and incorporating data into decision-making [3]. By doing so, the tendering authority can ensure their procurement contracts align well with the current market, attracting more contractors. This study provides a deeper insight into past procurement contract data, specifically focusing on the temporal aspects of the 'moment to publish' and 'contract duration'.

2.2 Literature - Shared

2.2.1 Tender contracts - Shared

Since procurement contracts are the middle point of this study, it is essential to understand what they are. The phenomenon of publishing contracts and having different contractors bidding to do the job is also known as 'tendering' [4], where **contractors** bid on a contract offered by the **tendering authority**. The goal for the tendering authority is to have as many registrations by the contractors, so they get to choose the best price and quality contract available. In the Netherlands, all government agencies have to publish their procurement contract as a tender according to the **procurement law** (Aanbestedingsswet 2012) [5]. In the construction industry, one of the industries that will be investigated in this study, tendering authorities can be identified as public sector entities, such as municipalities, port authorities, and RWS among others [6].

From the literature, various studies have investigated which factors the tendering authority can manipulate to influence the contractors to bid. According to research specifically in the construction industry, there have been 20+ factors identified that could influence the contractor into bidding or not, with a ranking from important to not important [7], which can be seen in Table A.1. Remarkable here, is that factors that have the most influence are on the tendering authorities' side, indicating that they can assert a significant influence on contractors to bid.

According to another study taking into account other industries besides construction, 28 critical factors have been identified for the bidding decision [8]. In line with the article focused on construction, there is significant influence from the tendering authorities' side. However, both articles confirm that **contract duration** is one of the lowest importance. The articles do not mention anything about the publishing period but do mention that the **availability of other projects** is a mediator-influence, which can relate to the publishing date project.

For the offered contracts, there is no specific regulation regarding their duration according to Dutch law. However, there is an exception for **'Raamovereenkomsten'** [9]. These usually have a maximum duration of four years, unless there is a proper justification for an extension.

2.2.2 Missing data - Shared

Missing data is a frequent data quality issue that impacts data pipelines, Machine Learning (ML) applications, and data analyses [10]. It is crucial to approach this issue correctly, as it can cause bias and inaccuracies if not done right [11]. Depending on the underlying reason for the missing data, different solutions can be used. These reasons can be categorized into the following three groups [12]:

- **MCAR**: Missing completely at random, refers to when there is no pattern between the missing data and any other data.
- **MAR**: Missing at random, refers to when the missingness depends on the observed data and there is a possible pattern.
- **MNAR**: Missing not at random, refers to when the missingness of the data is based on several specific missing values.

The amount of missing data should be carefully considered before selecting an imputation method, as the complexity and potential bias increase the more data is missing [13]. Advanced ML is a method that can resolve missing data and also minimally impacts the distribution of the data [14], however in many situations median or average imputation also performs well, especially when the data coverage is > 70% [15]. However, it has to be mentioned that this threshold differs per study and therefore lacks strong empirical agreement among researchers. What also matters are the different data types, which can make imputation complex [16]. In Table A.2 there is an overview of how different imputation methods perform with different data types and types of missing data.

2.2.3 Spearman correlation - Shared

Spearman a correlation coefficient is a metric that is used to assess the strength and direction between two variables. Spearman is highly effective as it can capture linear and non-linear relationships [17].

2.2.4 Machine learning models - Individual

When using ML for imputation, it is important to make sure the model is not over- or underfitting. Especially when a model uses a lot of features, there is an increased risk of overfitting [18]. Cross-validation is an effective method to detect this as it tests the model on different sets of data [19]. When evaluating the performance of ML models after imputation, several statistical metrics are commonly used to evaluate the performance. These are Mean Squared Error (MSE), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R^2 [20].

Furthermore, outliers can influence the performance of an ML model as they introduce noise. It is important to remove this from the dataset. The interquartile range (IQR) approach is a robust way to solve this. The IQR approach defines outliers as everything that falls below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR [21]. By removing these from the dataset, an ML model will be less likely to be disturbed by the outliers and capture the patterns more effectively.

2.2.5 Seasonality - Individual

Temporal data series can be processed by seasonal trend decomposition, which reveals hidden trends or seasonal trends in the data [22]. With this method, the original data is split into four components: observed data, trend, seasonal patterns, and residual components. While this decomposition theory has been in use since the last century, it is still a powerful and useful technique for identifying seasonal patterns. To quantify the seasonality, one can use ANOVA testing to determine the difference between data points [23].

2.2.6 Statistical test - Individual

T-tests are often used to compare the means of two groups and to confirm whether there is a statistically significant difference between them [24]. However, to perform a T-test, the two data groups need to validate a few assumptions [25]:

1. The observations must be independent of each other
2. The data should be normally distributed (or approximately)
3. The variances of the two groups should be similar (homogeneity of variance)

If one of these validations is violated, one might prefer to perform the Mann–Whitney U test. The Mann-Whitney U test is nonparametric and is particularly useful when the datasets are not normally distributed and variances are not similar [26].

2.3 Research Question and Hypothesis - Individual

This study will focus on the manipulation of the timing market approach and its effect on the number of registrations on the respective tenders. Since RWS is curious to know how to maximize its number of registrations, the following research question can be formulated:

What is the optimal timing for market approach in project tenders for Rijkswaterstaat PPO to maximize the number of submissions and ensure a fair competitive field, based on historical tender data analysis?

This research question can be structured around 3 key topics, each also has a sub-question that can be answered as support to the main research question:

- **Seasonality**: Are there specific periods in the year when the tender market is under higher pressure than the rest of the year? And if so, how does it correlate with RWS's historical pattern?
- **Statistical significance**: What are the differences in the mean number of tender registrations across these periods?
- **Industries**: Are there any differences in patterns between different industries? RWS is active in multiple industries and therefore attracts contractors from different markets, does this differ?

Following the subquestions, the following hypothesis can be formulated after initial consultation with RWS, which will either be confirmed or rejected in this study:

- *H1: There are seasonal patterns available in the tendering, likely falling together with major holidays (Christmas and 'bouwvak') and the ending of financial book years.*
- *H2: The average number of registrations is significantly higher in off-peak periods compared to peak periods.*
- *H3: Because certain industries have more contractors than others, there will likely not be a uniform answer for all industries but there will be differences between them.*

3. Data

3.1 TenderNed dataset - Shared

To answer the research question mentioned in 2.3, it is essential to obtain a dataset that contains all procurement publications of the industries that are related to those that RWS active in. The dataset used in this study is the complete dataset from TenderNed. This data is publicly available and can be accessed through the TenderNed webiste. The website publishes the dataset in both JSON and XLSX formats. For this research, the dataset with XLSX is used for the analysis, since the XLSX files are significantly smaller in size, and for its compatibility with Microsoft Excel, which has been used in the data preparation. TenderNed publishes the XLSX dataset in two sets: one covering the period from 2010 to 2015, and the other covering the period from 2016 to 2023.

In addition to the TenderNed dataset, an internal dataset from RWS is also used, which complements the TenderNed dataset. This internal RWS dataset contains additional information on the procurement contract publications published by RWS. More information on the utilization of this dataset can be seen in 3.3.

3.2 Description of the data - Shared

3.2.1 Data processing - Shared

In both XLSX datasets, each row represents the announcement of a procurement contract through the TenderNed platform. The 2010-2015 dataset contains 68 columns, while the 2016-2023 dataset compasses 85 columns in total, where all of the 68 previous columns are available, but 17 extra columns have been added to add new information related to the contracts. When combined, the dataset covers 85 columns and in total 301.252 rows. This contained data on all procurement contracts in the Netherlands, ranging from those in the construction industry to healthcare projects. Among these 85 columns, several were particularly important:

- **Publicatiedatum**: Publishing date of the contract. This variable is relevant as it will be a key factor in the analysis.
- **Naam aanbestedende dienst**: Name of the entity that offers the contract.

- **Aantal inschrijvingen**: The number of registrations for the respective publication. This variable is especially relevant as it is the focus of this research.
- **Hoofd CPV definitie code**: The Common Procurement Value (CPV) code relating to the industry in which the contract is offered.

Additionally, new columns have been added to make the dataset more useful for the analysis. These columns were:

- **Publicatiemaand**: Transformation of the publishing date to publishing month.
- **Cleaned name AB**: The cleaned name of column 'Naam aanbestedende dienst'.
- **Category AD**: The respective category of the contracting authority. This indicated what type of authority was offering the contract (e.g.: Municipality, Province, etc.)
- **Hoofd CPV code**: Strip of the column 'Hoofd CPV definitie code' to only 3 digits, since that was more relevant for the scope of this analysis.

The original **Naam aanbestedende dienst** column contained several inconsistencies with the names, like because of:

1. Typos and mistakes in writing.
2. The use of different names referring to the same entity.

As a result, for example, RWS PPO is labeled by 11 different names in the dataset, as can be seen in Table A.3. By doing a manual check on all the unique names in Excel (1,636 names), the unique names were then mapped manually to their correct 'cleaned name' (1,350 names). By using the INDEX-MATCH function combination, the new **Cleaned name AB** column could easily display the correct name.

With the significant number of distinct procurers, it was considered useful to categorize them into a category, as many small organizations published a tender. An experiment was conducted with a clustering algorithm, attempting to group the 'Cleaned names' into similar groups. To do this, the column **Category AB** was added, which is similar to the original column **'Soort aanbestedende dienst'**. This column had missing ± 10% missing data and was also inconsistent with categorizing; therefore it made sense to re-categorize them. The unique value from **Cleaned name AB** were categorized in one of the 11 groups that can be seen in Table A.4. These 11 categories were chosen in consultation with RWS.

To categorize the names into logical groups, few-shot prompting in a Large Language Model (LLM) was used. LLMs are used because they excel at understanding natural language nuances, crucial for categorizing names accurately. They are highly effective with large amounts of textual data, making them a suitable solution for the initial categorization task [27]. Additionally, their capacity for few-shot learning makes them particularly useful in scenarios where comprehensive training datasets are limited [28]. The prompt can be seen in A.4, and the LLM utilized was the **'gpt-35-turbo-16k'**. After the preliminary categorization of the LLM, there was a manual check and the categorization where needed.

3.2.2 Data exploration - Shared

Initial data exploration showed that over the years, more procurement tenders were published on the platform, which was a development confirmed by RWS. This can also be supported by a polynomial trend in Figure 3.1.

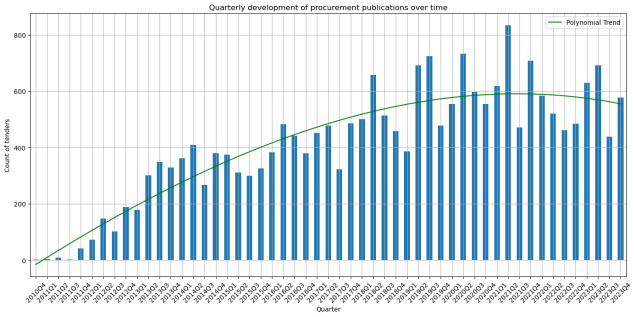

Figure 3.1: Quarterly development of publications of procurement contracts on TenderNed

Furthermore, in Figure 3.2 the data also reflects that most procurement contracts were offered by municipalities, but that RWS also is a significant large player in the pool.

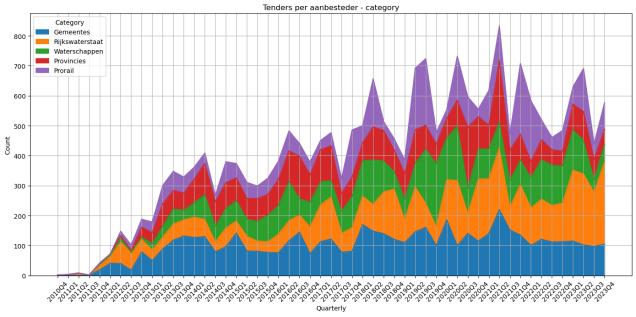

Figure 3.2: Quarterly development of offered procurement contracts by the 5 biggest players

This is relevant since municipalities usually offer maintenance projects that are similar to those of RWS, but smaller in scale. Based on these initial insights, and after consultation with RWS, the choice was made to exclude all categories that were not offered by either Provinces, Regional Water Authorities, RWS, or ProRail. In addition, from the municipalities, only the top 5 cities (Amsterdam, Rotterdam, Den Haag, Utrecht & Eindhoven) were still kept in the dataset, as these were still relevant to RWS.

3.2.3 Data filtering - Shared

Based on these first pre-processing steps, a significant part of the dataset could be removed. By using the **Hoofd CPV code** column, all irrelevant industries could be removed from the dataset. The relevant codes were provided by RWS, which were in this case: '**45-**', '**50-**', '**713-**', '**906-**' and '**34927100**'. This first data slice in Excel resulted in the dataset decreasing from **301.252** to **55.959** rows.

Next, by filtering out all irrelevant categories, the dataset was further narrowed down from **55.959** rows to **21.735**. With this dataset, initial data pre-processing in Excel was finished and now is ready for further pre-processing with Python.

3.3 Missing data pre-processing - Individual & Shared

3.3.1 Temporal filtering - Shared

With the current dataset, there seem to be a total of **4.621** missing values in the **Aantal inschrijvingen** column, which corresponds to **21,3%** of the dataset. At first, the complementary dataset from RWS was used to fill the empty values in rows of RWS, however, this only resulted in the filling of dozens of rows. In Figure 3.3, it becomes clear that the data coverage is in particular low during the the period. Figure 3.4 also shows that around 2016 the data coverage stayed consistent above 50%.

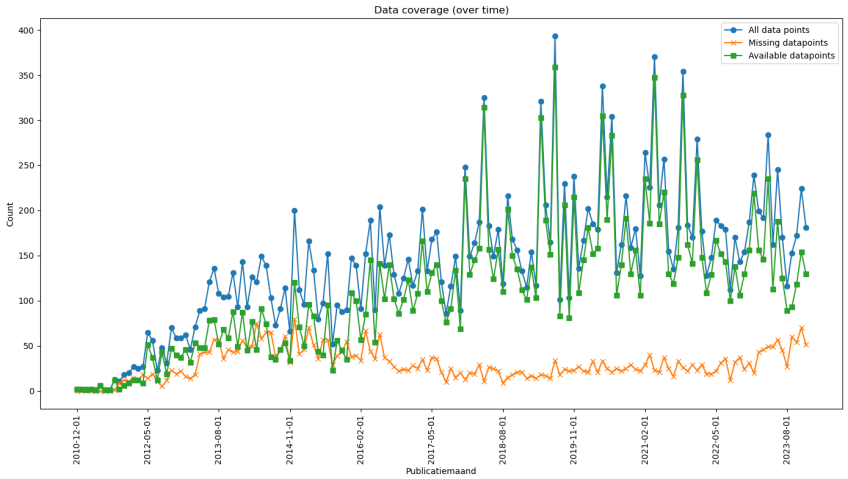

Figure 3.3: Data coverage of the Aantal inschrijvingen column over time.

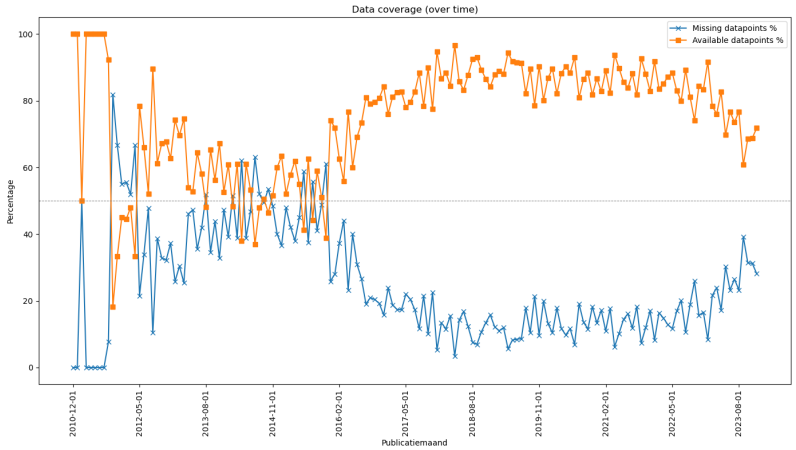

Figure 3.4: Data coverage of the Aantal inschrijvingen column over time (percentages.

As a result, only data after November 2015 (last month with less than 50% data coverage) is being used in this analysis. As a result, the dataset narrows down again from **21.735** to **17.430** rows, with now **2.763** missing values in the **Aantal inschrijvingen** column. This means that **15,9%** of the data is missing, a decrease of **5,4**-percent points.

3.3.2 CPV-codes - Shared

As the end goal of this research is not only to provide analysis across the whole dataset but also to segment by unique CPV code, there is a deeper dive into this to filter out the CPV codes that have insufficient data points. Since the time covered by the dataset from December 2015 to December 2023, the aim was to have at least 96 data points per CPV code, corresponding to one per month. In Table A.5, it becomes clear that 5 CPV codes contained insufficient data, and will be removed from the dataset as they cannot be analyzed.

In the same Table A.5, it also became evident that the amount of missing data in some CPV-codes exceeded **20.0%**. This is troublesome, as imputation generally works less stably when more than **20.0%** of the data is missing [14]. However, since removing the CPV-codes that have more than **20.0%** missing data would significantly reduce the size of the dataset, the choice was made to split the dataset into two sets: one 'low risk' dataset, containing all CPV-codes with missing values $\leq$ **20.0%**, and another dataset named 'high risk', containing the other CPV-codes. Both datasets will be prepared and analyzed with the same techniques; however, this aids in consulting RWS with the interpretation of the results.

3.3.3 Missing data patterns - Shared

Before resolving the missing data, it is essential to know what kind of missing data it is, as was mentioned in 2.2.2 (MAR, MCAR, MNAR). First, this can be done by trying to see if there are any potential patterns in the datasets by looking at correlations. Since most of the columns in the dataset were categorical columns, one-hot encoding was used to get numeric variables of the categorical ones. The correlations for the low risk dataset can be seen in Table A.6. In this table, it becomes evident that there is some strong correlation with other columns, which makes it hints at MAR over MCAR and MNAR, as the missingness is strongly related to other columns as well. Visualization of a heatmap can be seen in A.1.

For the high risk dataset, similar patterns can be observed, with several columns, such as **'Datum gunning'**, **'Perceel id '**, and **'Opdracht'**, showing strong correlations with the missing data. These correlations are presented in Table A.7. Therefore, the data appears to be MAR. This indicates that simply deleting or ignoring the missing data is not a viable approach, as it would introduce bias. Therefore, imputation emerges as the most appropriate option to resolve this issue. Corresponding heatmaps of the high-risk dataset are illustrated in Figure A.2.

3.3.4 Imputation with Machine Learning - Individual

To resolve missing data on a MAR basis, advanced Machine Learning (ML) options emerge as the preferred option, as this minimally impacts the distribution of the dataset [14]. In the dataset, columns had both numerical and temporal datatypes. This is important to consider for model selection, as this already rules out several modes like decision trees and Knn, which do not handle mixed data types as efficiently as other models like Random Forest or Gradient Boosting. Ensemble methods tend to handle diverse data types better for imputation tasks, so those are preferred [29]. Due to significant missing data in other columns as well, the **'HistGradientBoostingRegressor'** model was chosen, as it possesses the native ability to handle NaN values [30]. This is also supported by literature, and therefore we have opted for this model [31].

The **HistGradientBoostingRegressor** model showed promising results after initial tests on the low risk dataset. The model was tested on the dataset, with the outlier removed according to the IQR method. Initially, the model was tested on all features in the dataset. With the use of permutation importance, the most effective features could be extracted from the dataset and were only trained on those. After running a cross-validation test with different features, the low risk imputation model was trained on the top 75 features, as this showed the lowest MSE, as can be seen in Figure **??**.

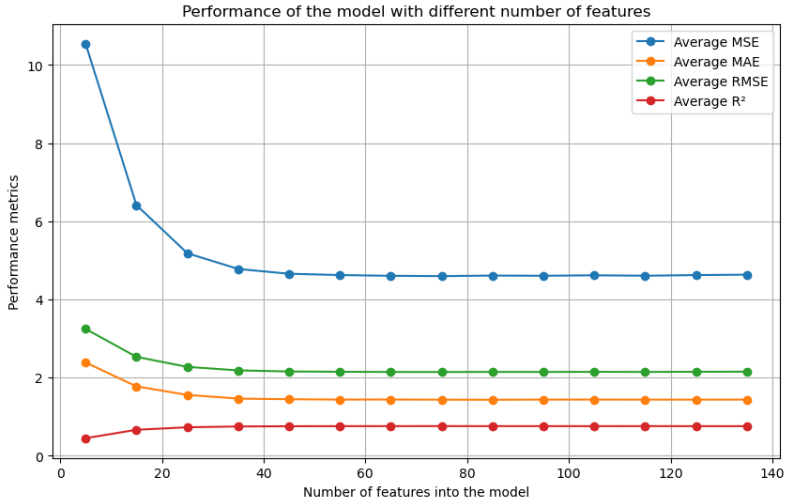

Figure 3.5: Performance of the model taking into account different numbers of features for the low risk dataset

After doing 10-fold cross-validation, the average performance scores can be seen in Figure 3.6.

Average Mean Squared Error (MSE):	4,7
Average Mean Absolute Error (MAE):	1,4
Average Root Mean Squared Error (RMSE):	2,2
Average R-squared (R^2):	0,76

Figure 3.6: Model performance on low risk data after 10 fold cross validation

After imputation, the distribution over time didn't change significantly, as can be seen in Figure 3.7. In the figure, it becomes clear that the pattern after the imputation is following the pattern before the imputation, and therefore it seems not to be bothered much by the imputation. Therefore, the imputation seems to be successful.

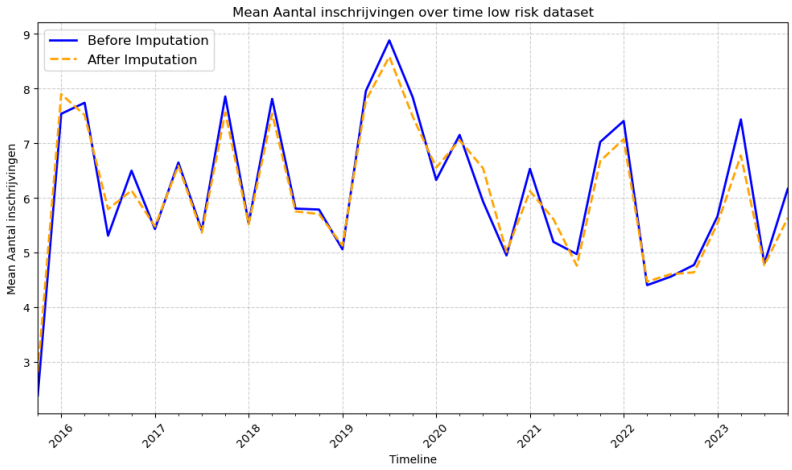

Figure 3.7: Aantal inschrijvingen mean over time difference before and after imputation

After testing the **'HistGradientBoostingRegressor'** model on the high risk dataset, the model needed fewer features than with the low risk dataset. The ideal amount of features to take into the model is in this case 55. How this develops can be seen in Figure 3.8.

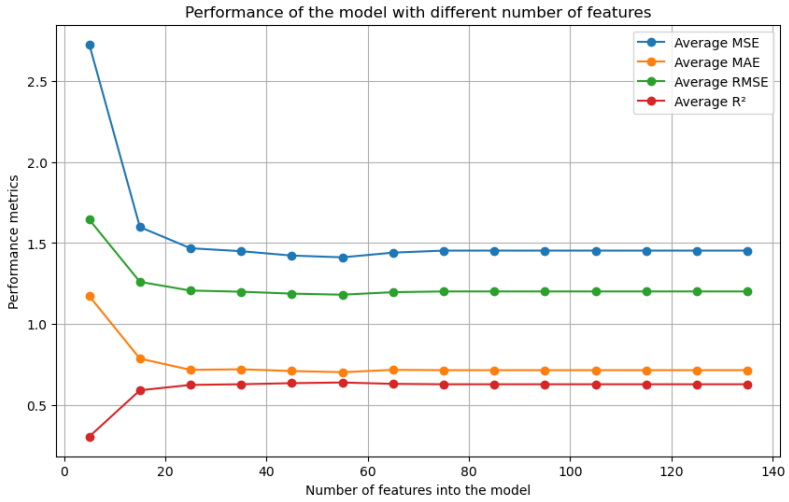

Figure 3.8: Performance of the model taking into account different numbers of features for the high risk dataset

After doing 10-fold cross-validation, the average performance scores can be seen in Figure 3.9. Although the MSE, MAE, and RMSE are lower than with the low risk dataset, indicating a better performance, the model is not able to explain why it is predicting better and therefore not effectively capturing the pattern in the data well enough, which leads to a lower (R^2). This can be explained due to the simple fact that there is more missing data, which can negatively influence the performance and the ability to capture patterns [32].

Average Mean Squared Error (MSE):	1,4
Average Mean Absolute Error (MAE):	0,8
Average Root Mean Squared Error (RMSE):	1,2
Average R-squared (R^2):	0,64

Figure 3.9: Model performance on high risk data

After imputation, the distribution didn't change here significantly as well, as can be seen in Figure 3.10. In this figure too, it becomes clear that the new pattern follows the old one closely, although with more discrepancy than with the low risk dataset. However, since overall the pattern seems to be the same, the imputation here too seems to be successful.

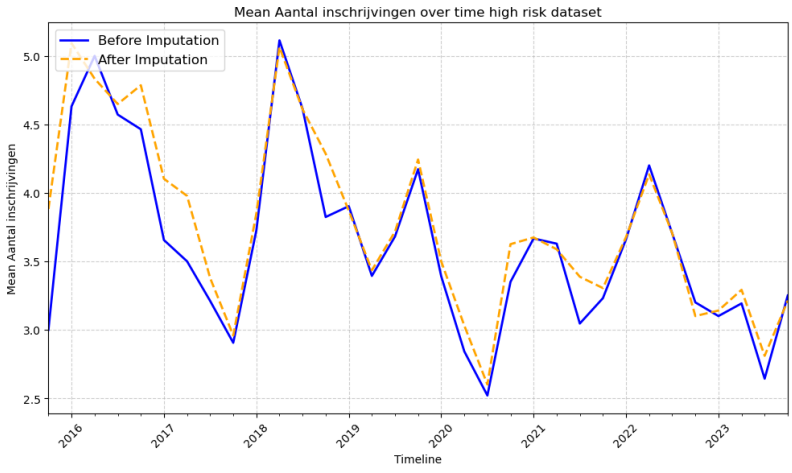

Figure 3.10: Aantal inschrijvingen count over time difference before and after imputation

After these imputations, both the low risk and high risk datasets are ready for analysis.

4. Method - Individual & Shared

4.1 Industry differences - Shared

To investigate whether there are differences in the industries, there will be an analysis of the distribution of the values in **Aantal inschrijvingen** compared to publishing date and contract duration. This is done by correlation analysis using both Pearson and Spearman correlation coefficients on the values of both columns. The absolute values of all one-hot encoded columns will then be aggregated and be taken as the mean, to see whether CPV code does have an influence.

4.2 Seasonality - Individual

Seasonality decomposition was conducted to examine trends, seasonality cycles, and residual values. This was done on a 12-month basis. This decomposition allows for a visual inspection of patterns, enabling the identification of peak and off-peak periods by comparing the publication counts to the mean count. Periods with higher counts than the mean are classified as peak periods, while lower counts are classified as off-peak periods.

4.3 Statistical testing - Individual

Once seasonality and peak/off-peak periods were identified, this was tested by statistical tests. This can be either T-tests or Mann-Whitney-U, depending on the data distribution of the samples. The important aspect here is whether the assumptions for the T-tests are being violated. These tests determine whether an increased mean is likely significant or caused by random variations. As t-tests were performed for each unique CPV value, the Bonferroni correction method was employed to avoid inflated Type I errors. This adjustment serves to hold the overall alpha level, which helps make results more robust.

5. Results

5.1 Industry differences - Shared

In Figure 5.1 there is an overview of the distribution of the values for '**Aantal inschrijvingen**' and those of '**contract duration**' per different CPV codes. As can be seen in the figure, is that there are different distributions of the values for the **Aantal inschrijvingen** for different CPV codes. This can be seen because of the distribution across the Y-dimension. CPV code '**90-**' shows a narrower distribution than the code '**71-**' for example.

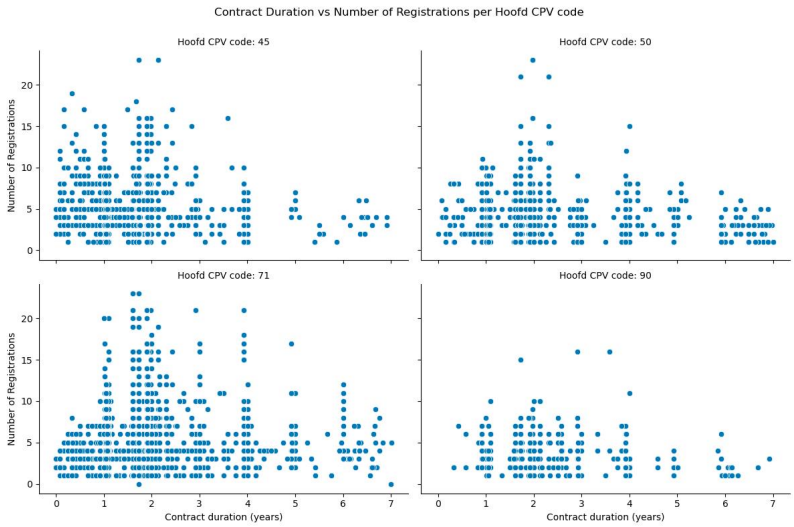

Figure 5.1: Plot of several CPV codes with time distribution

In Figure 5.2 there is an overview of the top 5 most influential features for the column '**Aantal inschrijvingen**'. These features are absolute aggregated. As can be seen from the figure, is here indeed that the CPV code is the most influential one, which directly responds to the sub-question regarding the differences between the industries.

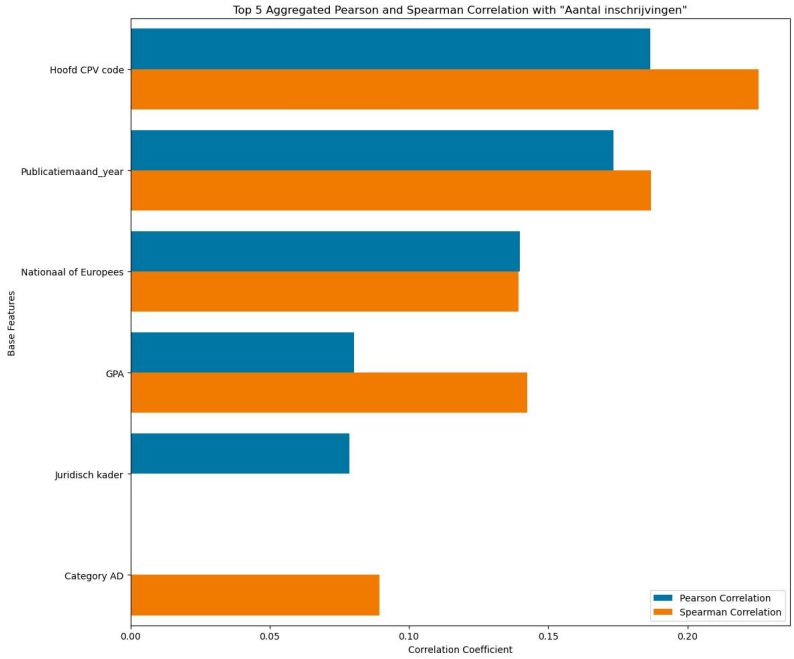

Figure 5.2: The top 5 most influential features for the Aantal Inschrijvingen column

5.2 Analysis results - Individual

5.2.1 Low risk dataset

After the analysis, it became evident that all CPV codes have a seasonality component. This can be seen in Table 5.1, where the seasonal strength is shown and the outcome of the ANOVA test. What is notable in the table is that the seasonal strength does differ from industry to industry. In section A.7.1 there is an overview of all STL visualizations of the different CPV codes.

CPV	Industry	Seasonal Strength	Seasonality
452	Construction Works	0,21	✓
500	Repair and Maintenance	0,06	✓
450	Construction Works	0,13	✓
454	Building Completion	0,11	✓
508	Miscellaneous Services	0,13	✓
713	Architectural Services	0,09	✓
906	Environmental Services	0,14	✓
501	Repair and Maintenance	0,12	✓
349	Transport Equipment	0,26	✓

Table 5.1: Summary of Seasonal Decomposition per CPV Code (low risk)

In Table 5.2 there is an overview of the results of the further analysis. First of all, it becomes clear that none of the two data groups validate the assumptions for the statistical T-test, so for all of them, the Mann-Whitney U test was needed to check for significant differences between the periods. The visualizations of the distributions can be seen in section A.7.2. Furthermore, after the Bonferroni correction, **none** of the CPV codes show a significant difference in the mean amount of registrations, whether in peak or off-peak season. Therefore it seems like the differences between the two groups are more likely to be related to variation than a real difference.

CPV	Industry	Shapiro-Wilk	T-test?	Stat. Test	P-val	Bonf. P-val	Signif.?
452	Construction	0,71 / 0,84	×	M-W U	0,02	0,22	×
500	Repair	0,38 / 0,91	×	M-W U	0,03	0,29	×
450	Construction	0,58 / 0,38	×	M-W U	0,01	0,05	×
454	Building	0,51 / 0,55	×	M-W U	0,48	1,00	×
508	Miscellaneous	0,68 / 0,38	×	M-W U	0,04	0,35	×
713	Architectural	0,71 / 0,68	×	M-W U	0,06	0,50	×
906	Environmental	0,52 / 0,74	×	M-W U	0,12	1,00	×
501	Repair	0,46 / 0,73	×	M-W U	0,61	1,00	×
349	Transport	0,85 / 0,74	×	M-W U	0,01	0,07	×

Table 5.2: Hypothesis Testing and Validation per CPV Code (low risk)

5.2.2 High risk dataset

In Table 5.3 there is an overview of the analysis of the high risk dataset. Here too, all CPV codes do have a significantly measurable seasonality pattern. In section A.8.1 there is an overview of all STL visualizations of the different CPV codes for the high risk dataset.

CPV	Industry	Seasonal Strength	Seasonality
502	Repair	0.12	✓
451	Construction	0.16	✓
453	Construction	0.18	✓
507	Miscellaneous Services	0.21	✓

Table 5.3: Summary of Seasonal Decomposition per CPV Code (high risk)

In Table 5.4 it becomes clear that in this case there are two CPV codes in which there is a difference between the two groups. For CPV codes '**502-**' and '**453-**' there is a notable difference. Furthermore, for all of these tests, the Mann-Whitney U test was used, since the assumptions for the T-test were not validated as well. The visualizations of the distributions can be seen in section A.8.2.

CPV	Industry	Shapiro-Wilk	T-test?	Stat. Test	P-val	Bonf. P-val	Signif.?
502	Repair	0.85 / 0.79	×	M-W U	0.00	0.01	✓
451	Construction	0.88 / 0.81	×	M-W U	0.20	0.82	×
453	Construction	0.69 / 0.86	×	M-W U	0.00	0.02	✓
507	Miscellaneous	0.77 / 0.81	×	M-W U	0.02	0.07	×

Table 5.4: Hypothesis Testing and Validation per CPV Code (high risk)

6. Conclusion & Discussion

6.1 Conclusions - Individual

Looking back at the sub questions of this study, the following answers can be given:

- **Seasonality**: *According to the ANOVA testing, for all CPV codes there is a significant difference between different times of the year, and therefore all CPV codes do have a seasonality pattern. The seasonality strength however differs per CPV code. After looking at the visualizations, it looks like the highs and lows are primarily surrounding the major breaks in the industries.*
- **Statistical significance**: *According to the statistical test, the number of registrations for the majority of the CPV values does not matter on the publication date. For most of them, there is no statistical difference based on specific periods in the year, which also confirms the literature.*
- **Industries**: *Yes, this does differ. In the distribution plots, it already became clear that there are different numbers of registrations per CPV code, and also the Spearman correlation showed that the CPV code was the most influential.*

Therefore, the answer to the main Research question is **it depends**. For the majority of the CPV values, according to the statistical tests, there is no real difference according to several periods in the year. For the two CPV codes **502** and **453**, it tends to be *after* the major breaks, so in January and September.

6.2 Discussion - Individual

In this dataset, the column **Aantal inschrijvingen** missed a lot of data, which influenced this study significantly. A lot of time went into pre-processing the data and making it ready for analysis. Since the study was limited to 10 weeks, it left not much room for further analysis besides what has been done.

However, if this study had to be done over, one specific topic for further research is to investigate the effects of multiple variables combined. In this study, only the effect of publishing date has been analyzed, while in reality (also supported by literature), there are many factors influencing the decision to bid for contractors. It would be particularly interesting to see how the publishing date in combination with macro-economic events influences the number of registrations per tender.

Another subject to look into is to take into account external events. During the analyzed period, several major events have shaken up the world, such as increased global tensions, inflation, COVID-19 pandemic among others. These could have significantly influenced the tender market but were not taken into account in this study.

Lastly, for the seasonality patterns, in this study, only 12 months were analyzed, as for the simplicity for 1 year. However, this could be manipulated and altered to other periods.

A. Appendix

A.1 Literature review

A.1.1 Factors to bid/no-bid

Groups	Critical factors according to their nature	Average RII
Tendering authority	Previous experience of contractor with employer. Type of employer. Financial stability of employer. Identity and reputation of employer in the industry. Employer special requirements. Promptness of employer in payments process. Employer efficiency in decision making. Strength and position of employer in the industry. Qualifications and quality of employer staff.	75.87
Contractor	Availability of required cash. Facilities available to contractors from financial institutions. Quality of available contractor's staff. Previous experience in similar projects. Need for public exposure. Availability of labor. Current workload. Need for work. Previous profit in similar projects. Availability of subcontractors.	67.87
Bidding situation	Tendering method. Tendering duration. Number of bidders. Identity of bidders. Availability of other projects. Tender documents purchasing price. Bid bond size and validity.	63.42
Contract	Contract type. Contract durations. Contract conditions. Payment scheme. Tender documents quality level. Value of liquidated damages. Size and validity of security bonds required. Insurance premium required.	60.52
Project	Type of Project. Size of project. Location of project. Safety level required. Complexity level. Identity of designer. Identity of construction supervisor. Type of equipment required.	57.74

Table A.1: Critical factors according to their nature and average Relative Importance Index (RII) [7]

A.1.2 Imputation methods for missing data

Imputation Method	Numerical	Categorical	MCAR	MAR	MNAR
Listwise Deletion	✓	✓	✓	×	×
Pairwise Deletion	✓	✓	✓	×	×
Mean/Median	✓	×	✓	×	×
Mode	×	✓	✓	×	×
LOCF & NOCB	×	✓	✓	✓	×
Maximum Likelihood	✓	✓	✓	✓	×
Matrix Completion	✓	×	✓	✓	×
Bayesian Approach	✓	✓	✓	✓	×
Regression	✓	✓	✓	✓	×
K-Nearest Neighbour	✓	✓	✓	✓	×
Tree Based	✓	✓	✓	✓	×
SVM Based	✓	✓	✓	✓	×
Clustering Based	✓	✓	✓	✓	×
Hist Gradient Booster	✓	×	✓	✓	×

Table A.2: Summary of various imputation methods with different data types and type of missing data [16]

A.2 Data discrepancy

Different entries of RWS PPO
Rijkswaterstaat Programma's Projecten en Onderhoud
Rijkswaterstaat Programma's, Projecten en Onderhoud
Rijkswaterstaat PPO
Rijkswaterstaat Programma's, Projecten en Onderhoud (PPO)
Rijkswaterstaat Programma's, Projecten en Onderhoud Zee en Delta
Rijkswaterstaat mede namens ProRail en de gemeente Amsterdam RWS Grote Projecten en Onderhoud t.a.v. Aanbestedingsteam GPO-PPO
Rijkswaterstaat Programma's, projecten en Onderhoud - Uitvoerend Secretariaat VNSC
Rijkswaterstaat Programma's Projecten en Onderhoud (PPO)
Rijkswaterstaat Projecten Programma's en Onderhoud
Rijkswaterstaat Programma Projecten en Onderhoud (PPO)
Rijkswaterstaat Progamma's, Projecten en Onderhoud

Table A.3: Different entries for RWS PPO

A.3 Unique categories

Categories	Number of players in each category
Educational institutions	355
Municipalities	480
Commercial organisations	125
Ministries	68
Healthcare	19
Other	209
Provinces	19
Infrastructure	33
Rijkswaterstaat	14
Regional Water Authorities	27
Prorail	1
Total	**1350**

Table A.4: Number of players across the categories

A.4 Prompt for Initial Output

Objective: *The goal of this task is to analyze a given list of organizational names and group them into logical categories based on the type of organization.*

Data: *The data includes names of various organizations such as government agencies, schools, municipalities, etc., examples being 'Rijkswaterstaat Noord-Brabant', 'Gemeente Rotterdam', and 'Universiteit van Amsterdam'.*

Task: *Category Mapping - Group all organizations logically based on their nature and descriptive naming. Categories that can be used are 'Other', 'Educational institutions', 'Ministries and agencies', 'Municipality', 'Healthcare', 'Infrastructure and transportation', 'Provinces', 'Regional Water Authorities', 'Rijkswaterstaat' and 'Commercial companies'.*

Output: *Provide an Excel table with two columns: the first column for the original names of the organizations and the second column for the corresponding categories. Ensure each organization is accurately categorized alongside the original name.*

A.5 Missing data

A.5.1 Data points per CPV code

Hoofd CPV code	All data points	Available datapoints	Missing datapoints	Missing ratio
506	2	2	0	0.000000
455	4	4	0	0.000000
503	38	24	14	0.368421
504	77	63	14	0.181818
505	106	93	13	0.122642
508	140	129	11	0.078571
349	147	137	10	0.068027
507	206	153	53	0.257282
501	192	174	18	0.093750
451	292	214	78	0.267123
502	362	262	100	0.276243
453	658	511	147	0.223404
906	541	515	26	0.048059
454	591	549	42	0.071066
500	805	724	81	0.100621
450	2120	1714	406	0.191509
713	3932	3508	424	0.107833
452	7217	5891	1326	0.183733

Table A.5: Descriptive statistics of available data per CPV code

A.5.2 Missing data patterns

A.5.2.1 Low risk dataset

Column	Correlation
Perceel id	0.490629
Datum gunning	0.429638
Opdracht	0.281320
Hoofdgunningscriterium	0.215622
Procedure	0.189537
Plaats opening kluis	0.170408
Voltooiing opdracht	0.115627
Sluitingsdatum aanbesteding	0.098406
Trefwoorden	0.088918
Termijn verkrijgen documenten	0.084508

Table A.6: 10 columns with most correlation with missingness of column 'Aantal inschrijvingen' in the low risk dataset

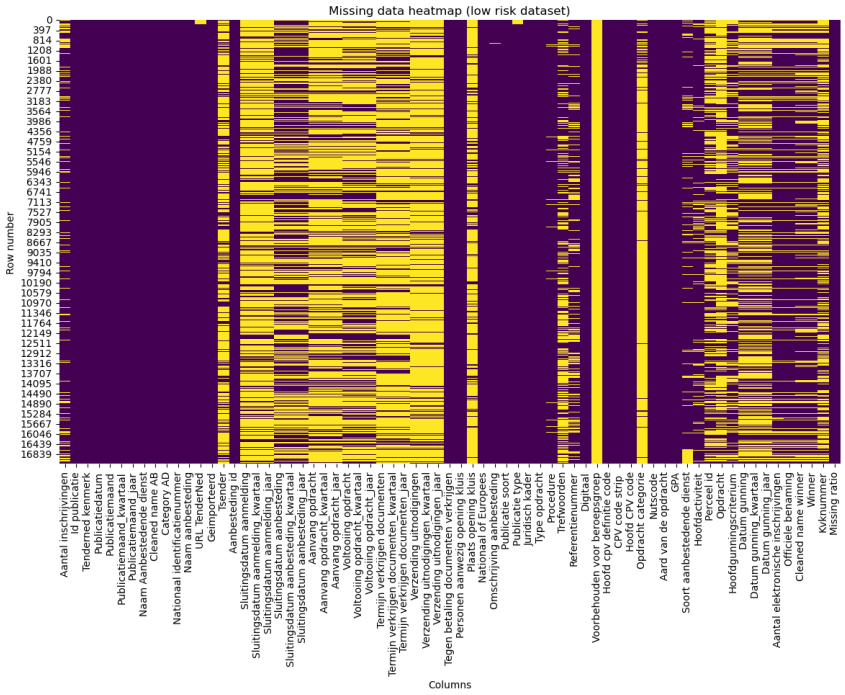

Figure A.1: Heatmap on missingness of the data (low risk dataset)

A.5.2.2 High risk dataset

Column	Correlation
Datum gunning	0.600640
Perceel id	0.540495
Opdracht	0.319894
Hoofdgunningscriterium	0.224969
Procedure	0.220109
Plaats opening kluis	0.209320
Voltooiing opdracht	0.184098
Verzending uitnodigingen	0.165406
Aanvang opdracht	0.120690
URL TenderNed	0.112435

Table A.7: 10 columns with highest correlation with missingness of column 'Aantal inschrijvingen' in the high risk dataset

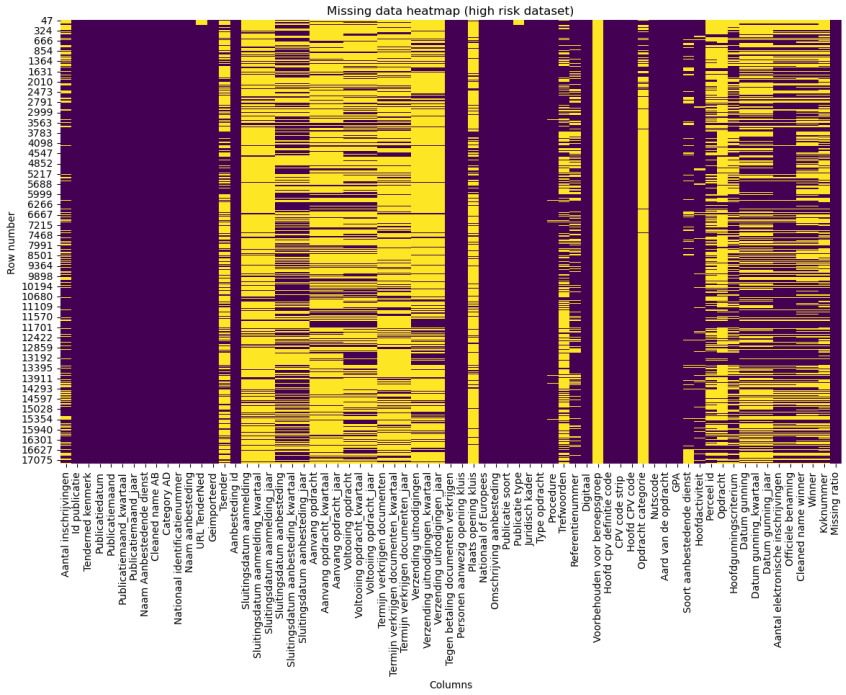

Figure A.2: Heatmap on missingness of the data (High risk dataset)

A.6 Summary statistics and visualizations imputation

A.6.1 Low risk dataset

A.6.1.1 Summary statistics

Statistic	Before Imputation	After Imputation
Count	12.308	15.685
Mean	6,37	6,24
Std	4,38	4,05
Min	0,0	0,0
25%	3,0	3,0
50% (Median)	5,0	5,0
75%	9,0	8,0
Max	20,0	20,0

Table A.8: Summary statistics before and after imputation on the low risk dataset

A.6.1.2 Distribution visualisation

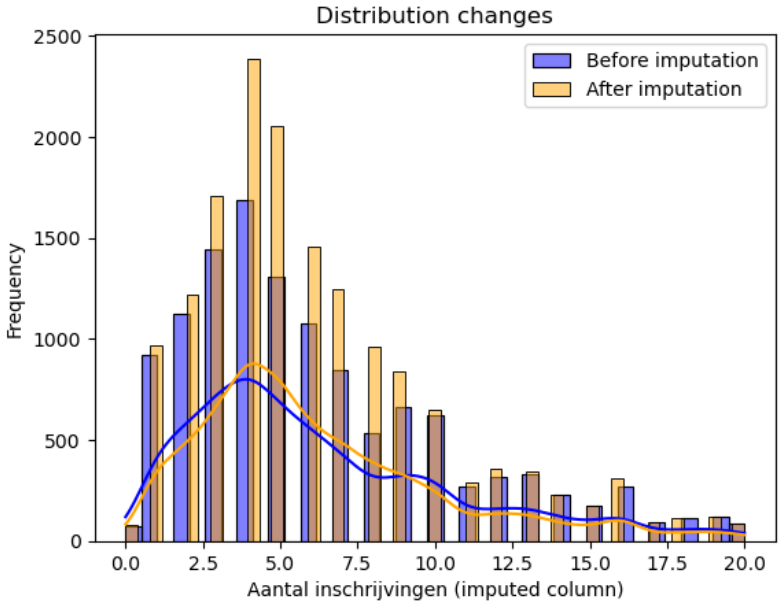

Figure A.3: Distribution difference before and after imputation

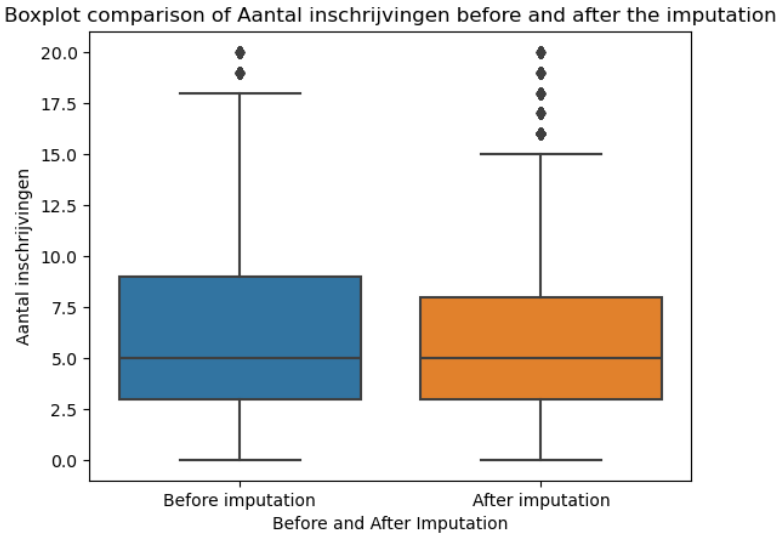

Figure A.4: Difference in distribution before and after imputation for the low risk dataset

A.6.2 High risk dataset

A.6.2.1 Summary statistics

Statistic	Before Imputation	After Imputation
Count	1.110	1.518
Mean	3,68	3,78
Std	1,99	1,85
Min	0,0	0,0
25%	2,0	2,0
50% (Median)	3,0	4,0
75%	5,0	5,0
Max	9,0	9,0

Table A.9: Summary statistics before and after imputation on the high risk dataset

A.6.2.2 Distribution visualisation

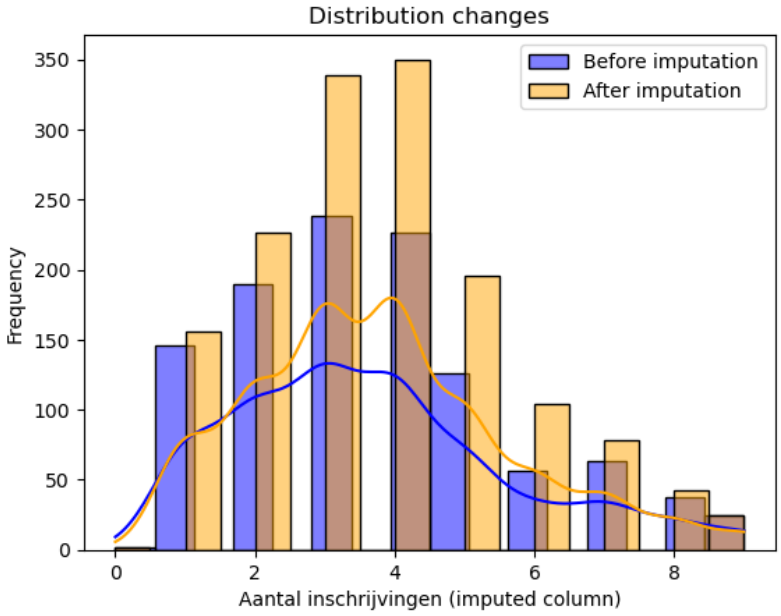

Figure A.5: Distribution difference before and after imputation

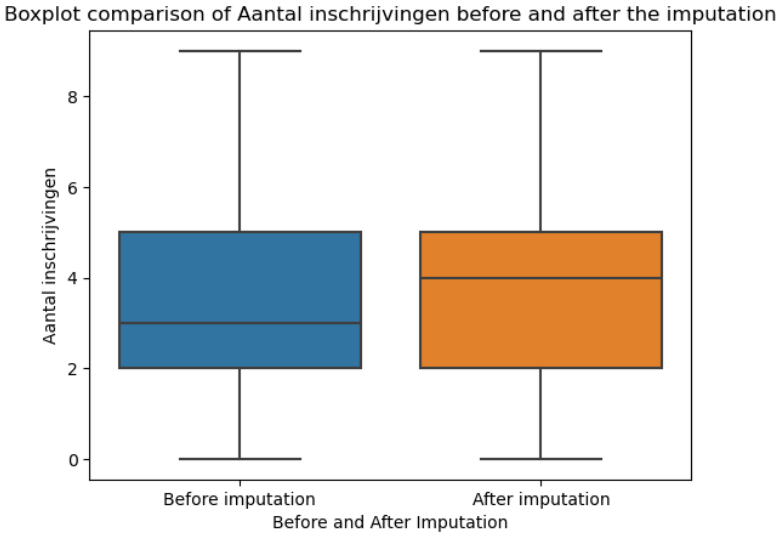

Figure A.6: Difference in distribution before and after imputation for the high risk dataset

A.7 Results low risk

A.7.1 Seasonal decomposition

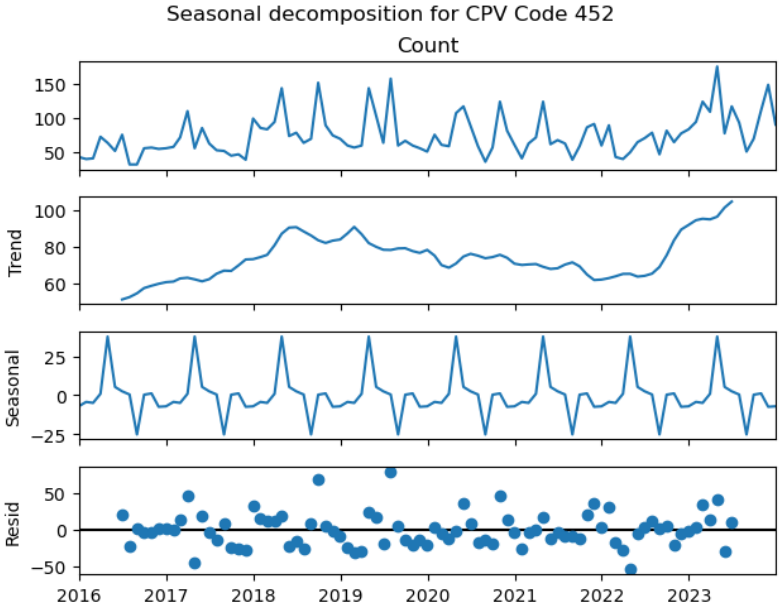

Figure A.7: Seasonality decomposition visualization of CPV code 452

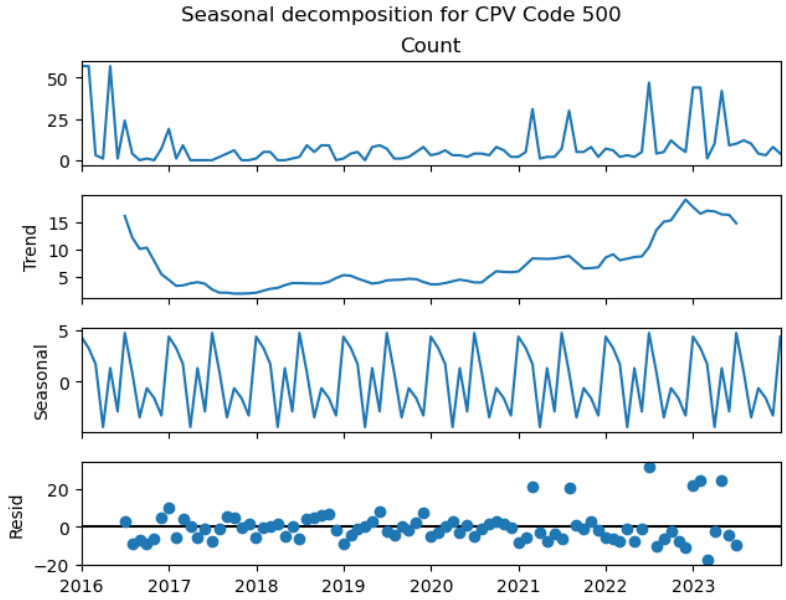

Figure A.8: Seasonality decomposition visualization of CPV code 500

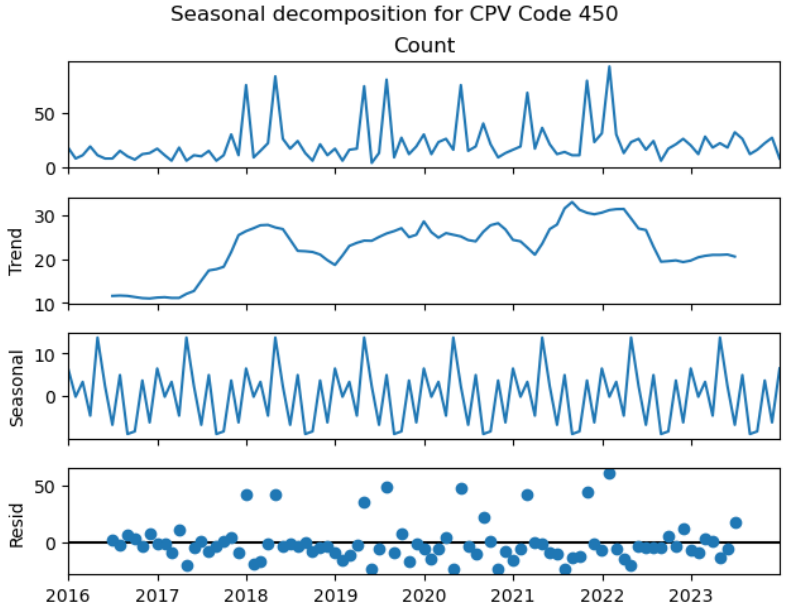

Figure A.9: Seasonality decomposition visualization of CPV code 450

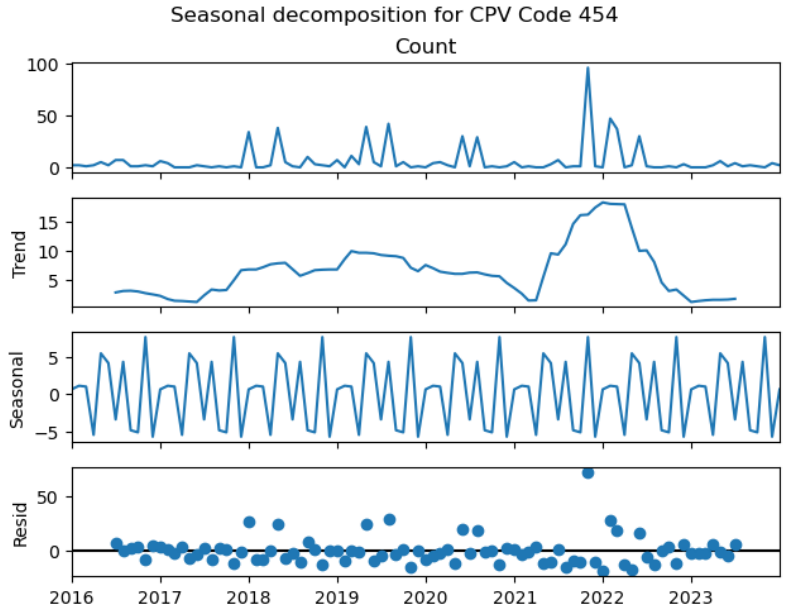

Figure A.10: Seasonality decomposition visualization of CPV code 454

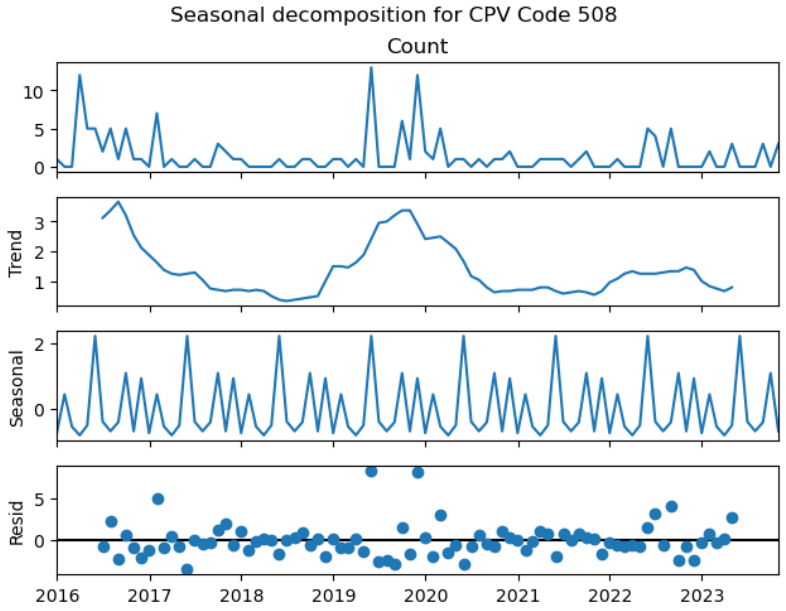

Figure A.11: Seasonality decomposition visualization of CPV code 508

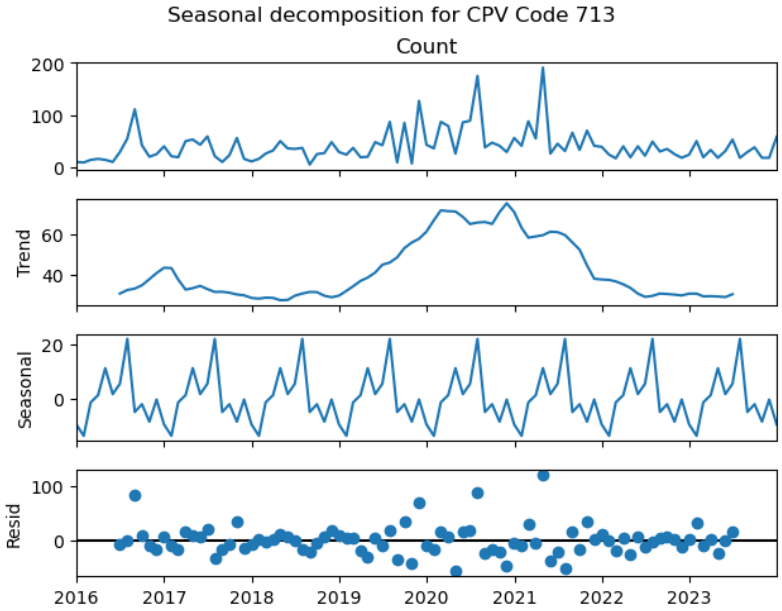

Figure A.12: Seasonality decomposition visualization of CPV code 713

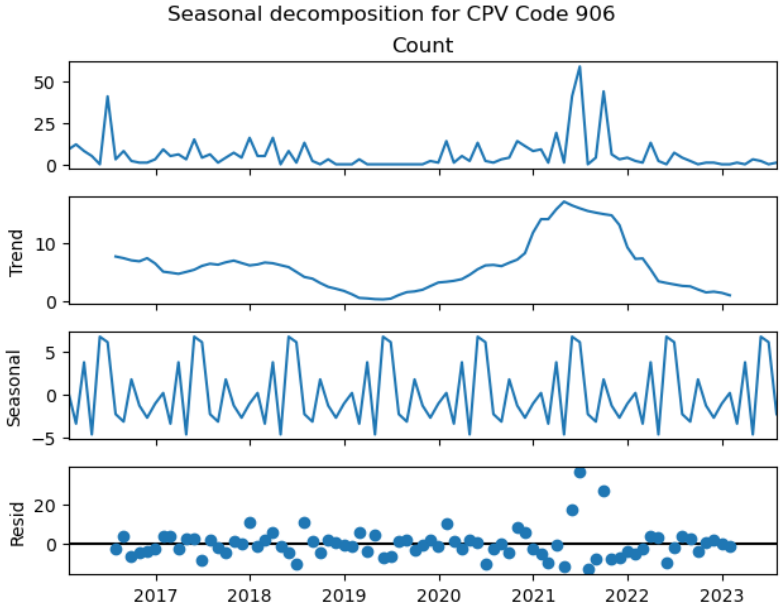

Figure A.13: Seasonality decomposition visualization of CPV code 906

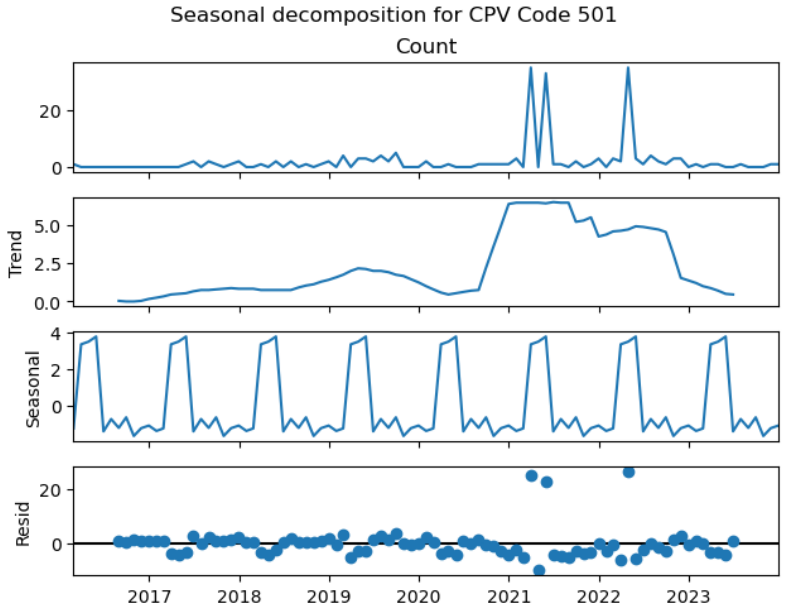

Figure A.14: Seasonality decomposition visualization of CPV code 501

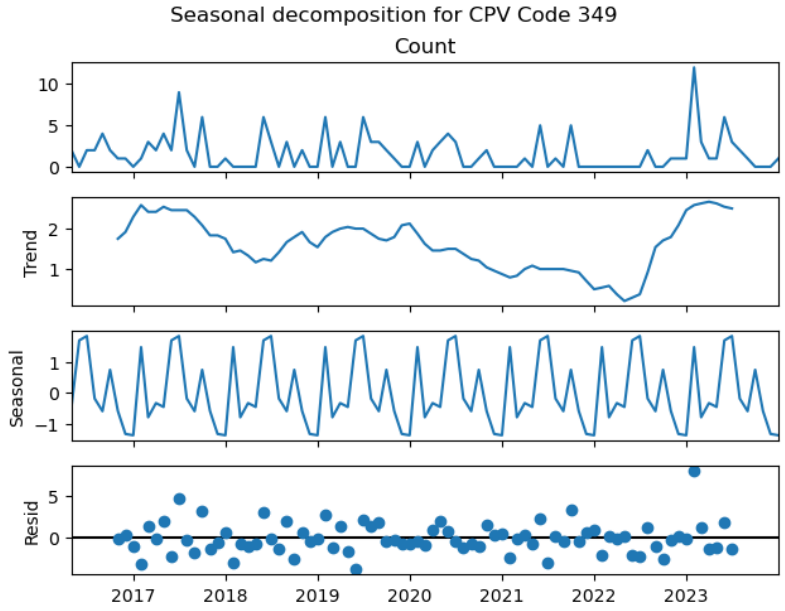

Figure A.15: Seasonality decomposition visualization of CPV code 349

A.7.2 Distribution groups

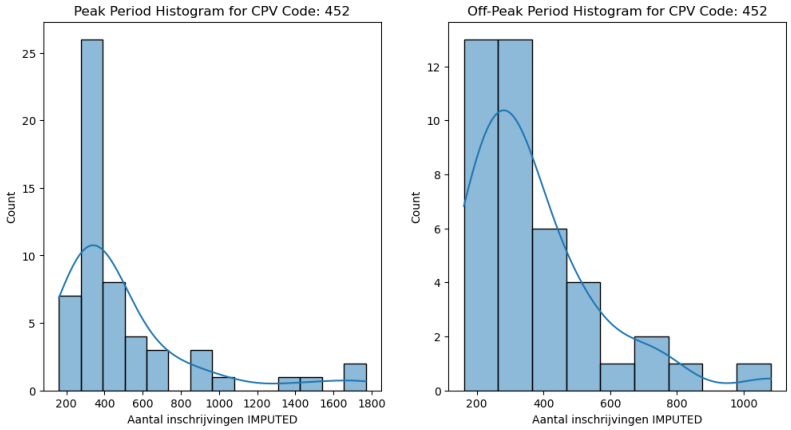

Figure A.16: Data distribution of the peak and off peak groups for CPV code 452

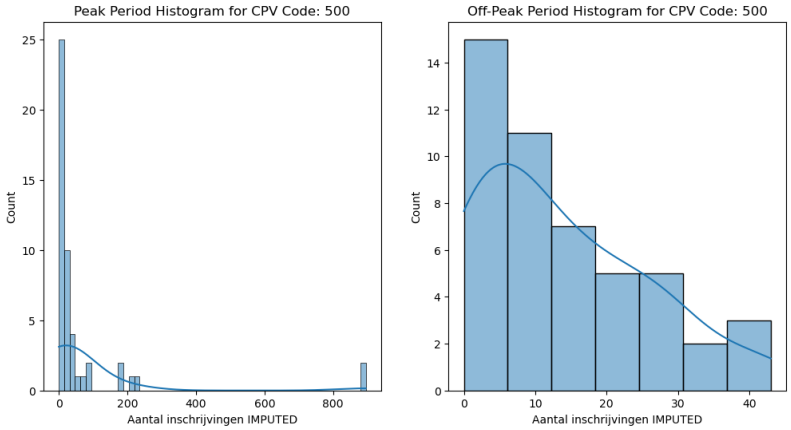

Figure A.17: Data distribution of the peak and off peak groups for CPV code 500

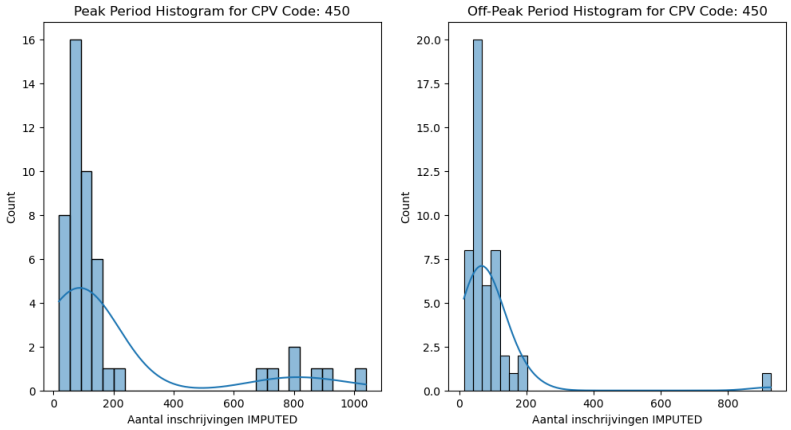

Figure A.18: Data distribution of the peak and off peak groups for CPV code 450

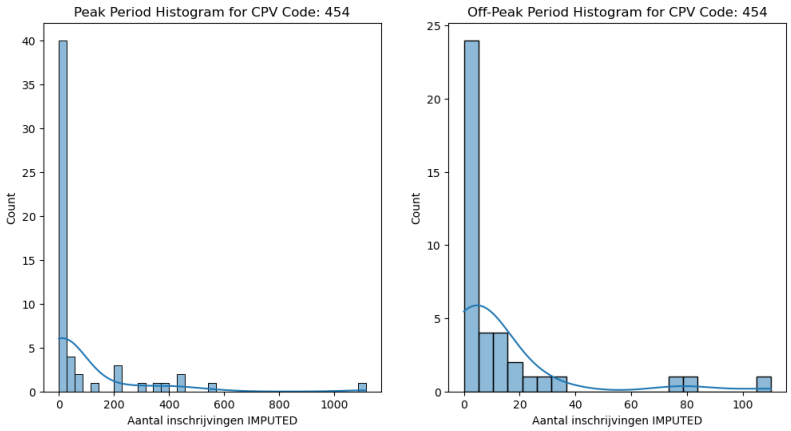

Figure A.19: Data distribution of the peak and off peak groups for CPV code 454

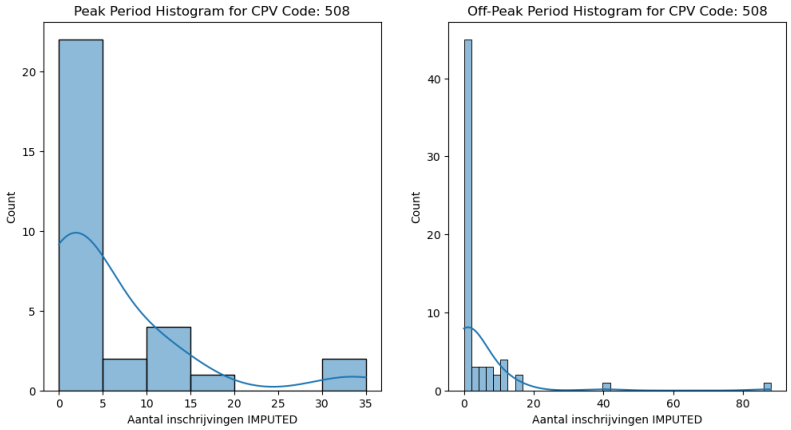

Figure A.20: Data distribution of the peak and off peak groups for CPV code 508

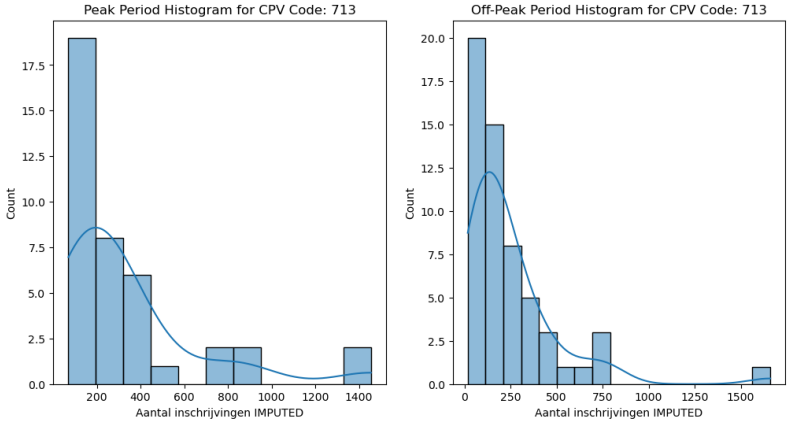

Figure A.21: Data distribution of the peak and off peak groups for CPV code 713

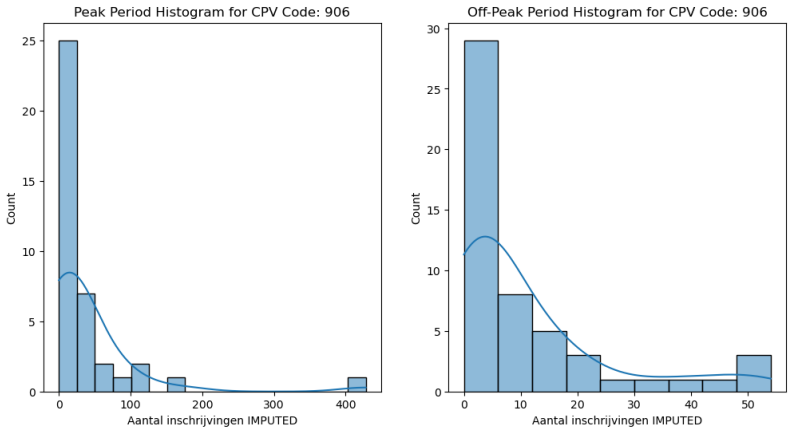

Figure A.22: Data distribution of the peak and off peak groups for CPV code 906

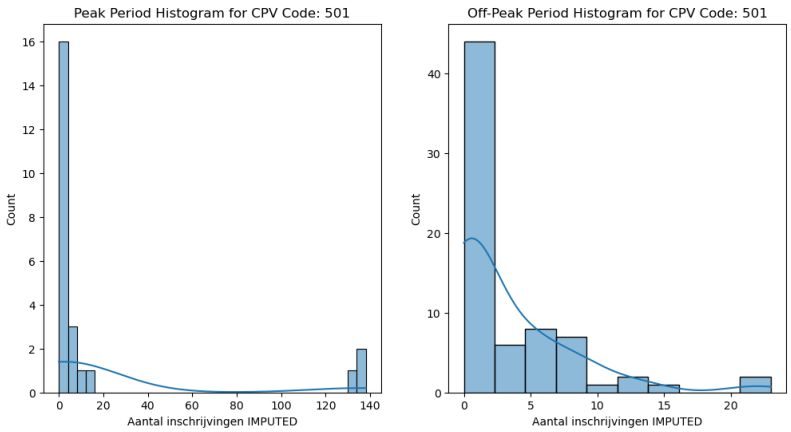

Figure A.23: Data distribution of the peak and off peak groups for CPV code 501

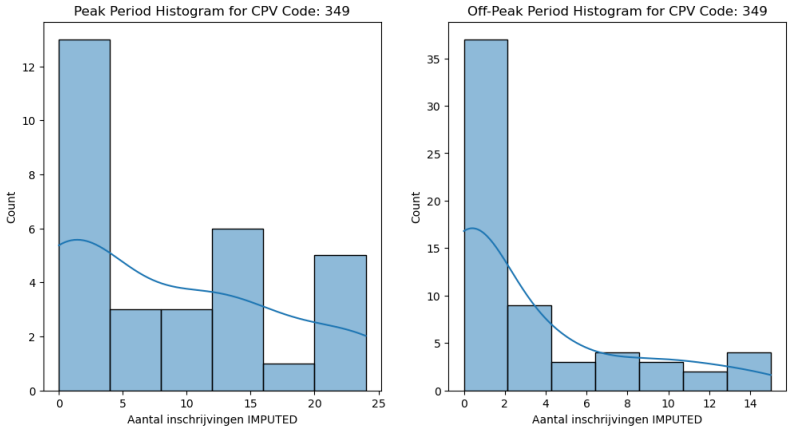

Figure A.24: Data distribution of the peak and off peak groups for CPV code 349

A.8 Results high risk

A.8.1 Seasonal decomposition

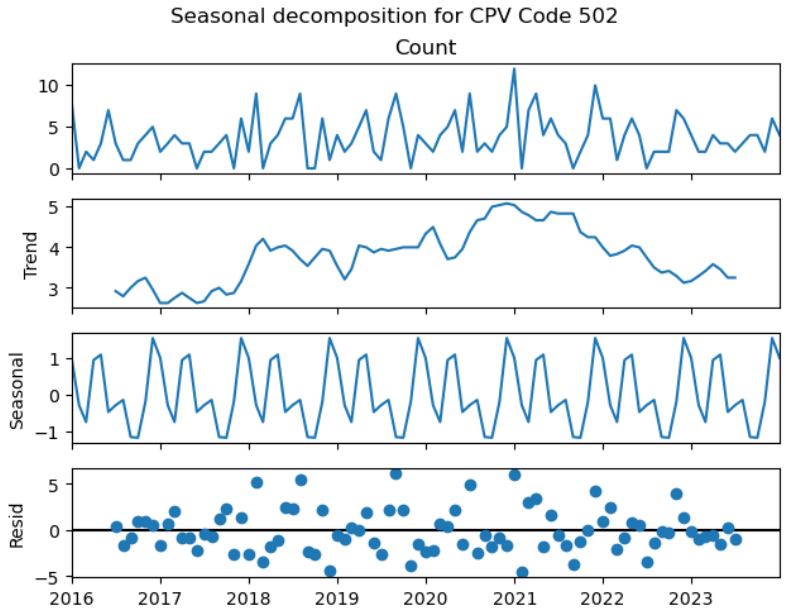

Figure A.25: Seasonality decomposition visualization of CPV code 502

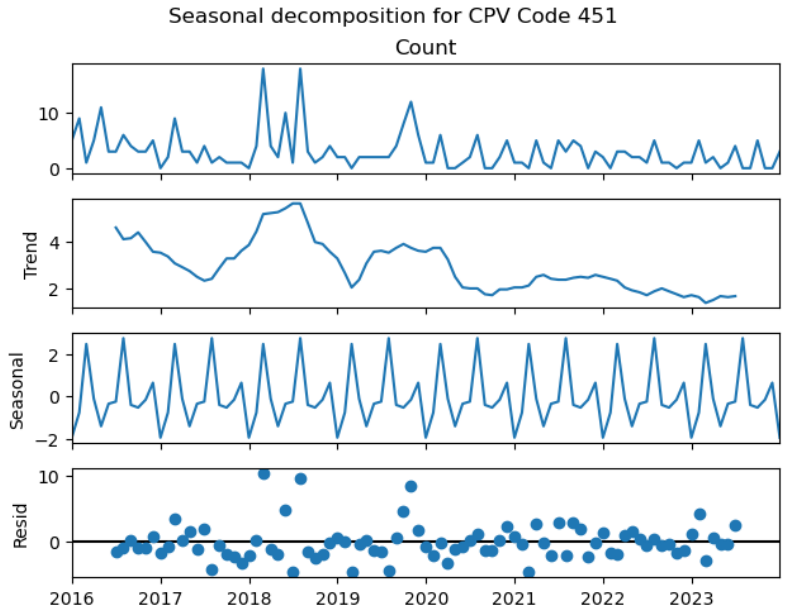

Figure A.26: Seasonality decomposition visualization of CPV code 451

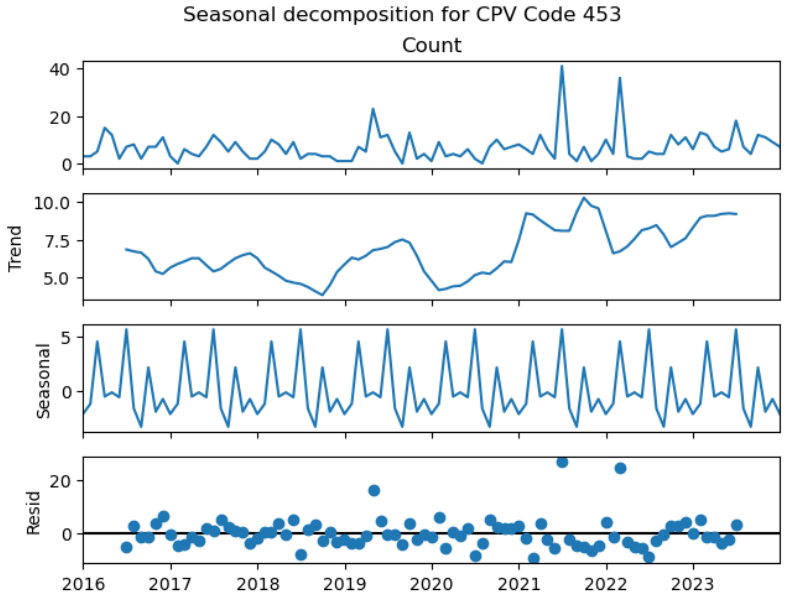

Figure A.27: Seasonality decomposition visualization of CPV code 453

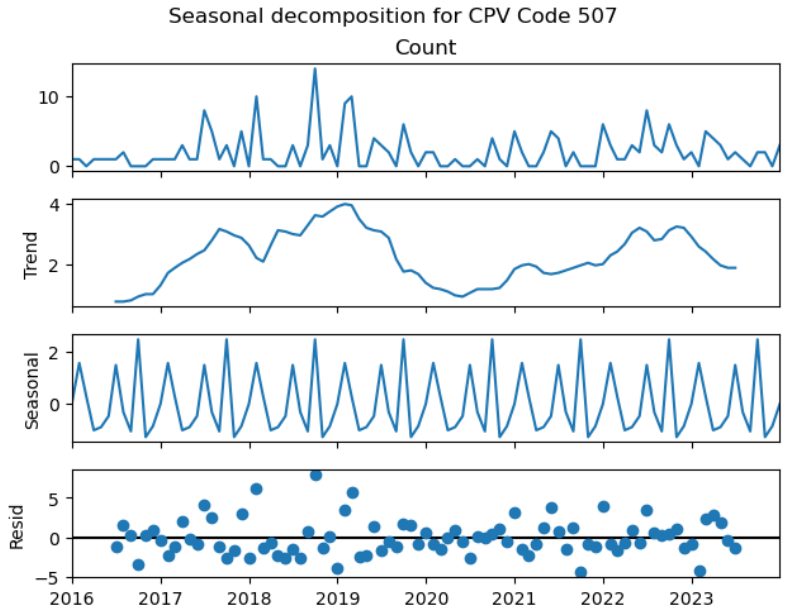

Figure A.28: Seasonality decomposition visualization of CPV code 507

A.8.2 Distribution assumptions t test

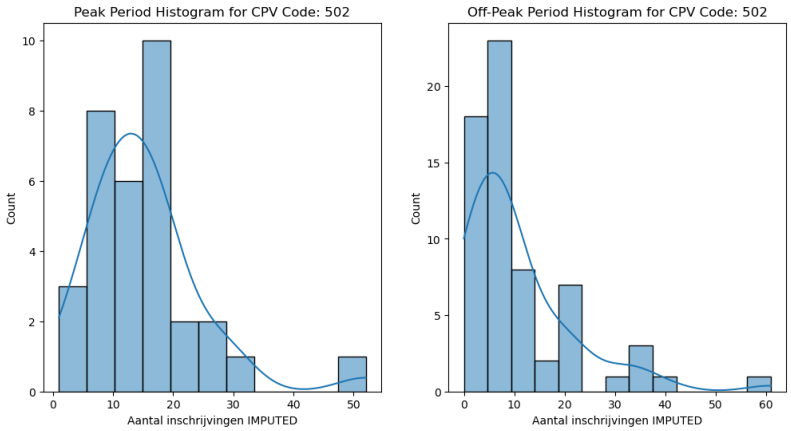

Figure A.29: Data distribution of the peak and off peak groups for CPV code 502

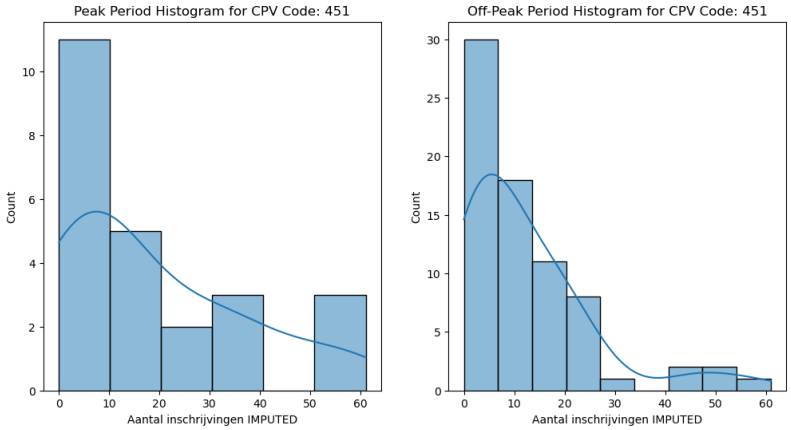

Figure A.30: Data distribution of the peak and off peak groups for CPV code 451

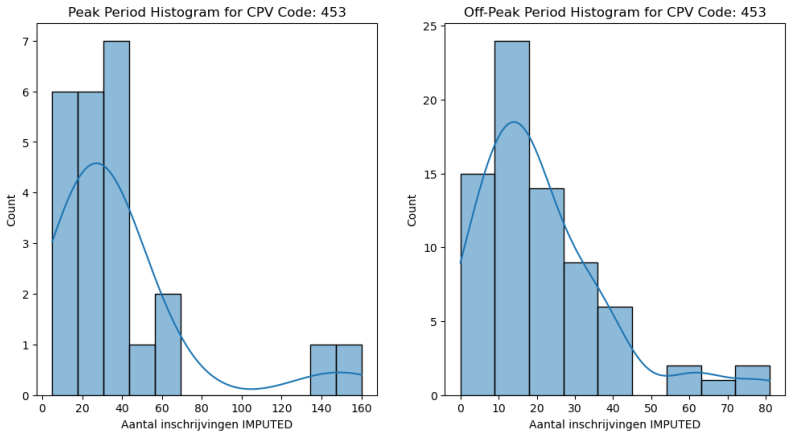

Figure A.31: Data distribution of the peak and off peak groups for CPV code 453

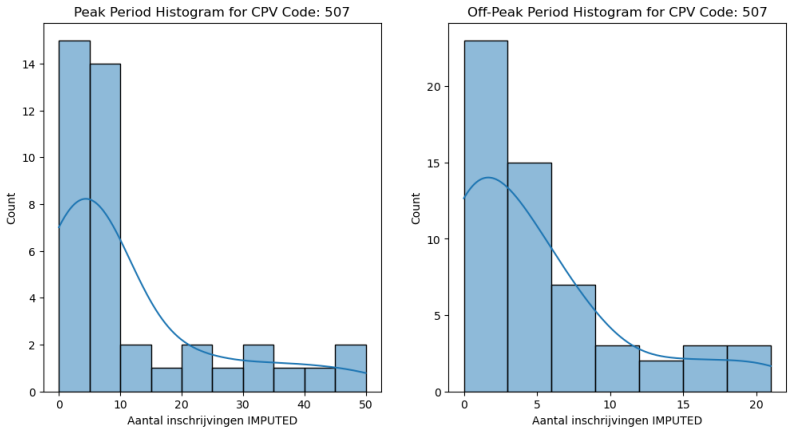

Figure A.32: Data distribution of the peak and off peak groups for CPV code 507

A.9 Notebook and Excel

The used notebook can be found in this Github respository. The Excel used in this study can be found on this website. Please request access through email: a.elhattache@students.uu.nl

Bibliography

[1] G. of the Netherlands, *Ministry of infrastructure and water management - organisation*, Accessed: 2024-06-22, 2024. [Online]. Available: https://www.government.nl/ministries/ministry-of-infrastructure-and-water-management/organisation.

[2] Rijkswaterstaat, *Uitgaven en kosten 2022-2023*, Accessed: 2024-06-23, 2024. [Online]. Available: https://www.magazinesrijkswaterstaat.nl/rwsjaarbericht/2023/01/uitgaven-en-kosten_2.

[3] PIANOo, *3 voordelen van datagedreven aanbesteden*, Accessed: 2024-06-23, 2024. [Online]. Available: https://www.pianoo.nl/nl/actueel/nieuws/3-voordelen-van-datagedreven-aanbesteden.

[4] M.-B. Dumo, "Bidding/tendering strategies: As success determinants of construction tenders," *World Journal of Innovation and Modern Technology*, vol. 6, no. 1, pp. 1–18, 2022. DOI: 10.56201/wjimt.v6.no1.2022.pg1.18. [Online]. Available: https://doi.org/10.56201/wjimt.v6.no1.2022.pg1.18.

[5] Rijksoverheid, *Aanbestedingsregels*, Retrieved June 22, 2024, from https://www.rijksoverheid.nl/onderwerpen/aanbesteden/aanbestedingsregels, n.d.

[6] S. Hardeman, "De feiten rond aanbesteden clustering, proportionaliteit en gebruik emvi bij aanbesteding van werk," Economic Institute for Construction and Housing (EIB), Report, 2013. [Online]. Available: https://www.publicspaceinfo.nl/media/uploads/files/EIB_2013_0002.pdf.

[7] G. Martínez Montes, F. J. Alegre Bayo, and B. Moreno Escobar, *Title of the paper*, 21st International Congress on Project Management and Engineering, Cádiz, 12th - 14th July 2017. Retrieved June 22, 2024, from http://dspace.aeipro.com/xmlui/bitstream/handle/123456789/349/AT02-001.pdf?sequence=1&isAllowed=y, 2017.

[8] B. L. Oo, T. H. B. Lim, and G. Runeson, "Critical factors affecting contractors' decision to bid: A global perspective," *Buildings*, vol. 12, no. 3, 2022, ISSN: 2075-5309. DOI: 10.3390/buildings12030379. [Online]. Available: https://www.mdpi.com/2075-5309/12/3/379.

[9] M. van Binnenlandse Zaken en Koninkrijksrelaties, *Aanbestedingswet 2012*, Accessed: 2024-06-22, 2022. [Online]. Available: https://wetten.overheid.nl/BWBR0032203/2022-03-02/0.

[10] S. Jäger, A. Allhorn, and F. Bießmann, "A benchmark for data imputation methods," *Frontiers in Big Data*, vol. 4, 2021. DOI: 10.3389/fdata.2021.693674. [Online]. Available: https://doi.org/10.3389/fdata.2021.693674.

[11] V. Österlund, "Missing data - a gentle introduction," Ph.D. dissertation, 2020. [Online]. Available: https://www.diva-portal.org/smash/get/diva2:1444918/FULLTEXT01.pdf.

[12] H. Kang, "The prevention and handling of the missing data," English, *Korean Journal of Anesthesiology*, vol. 64, no. 5, pp. 402–406, 2013, Published online: May 24, 2013, ISSN: 2005-7563. DOI: 10.4097/kjae.2013.64.5.402. [Online]. Available: https://doi.org/10.4097/kjae.2013.64.5.402.

[13] M. K. Hasan, M. A. Alam, S. Roy, A. Dutta, M. T. Jawad, and S. Das, "Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021)," *Informatics in Medicine Unlocked*, vol. 27, p. 100 799, 2021. DOI: 10.1016/j.imu.2021.100799. [Online]. Available: https://doi.org/10.1016/j.imu.2021.100799.

[14] T. Emmanuel, T. M. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, pp. 1–21, 2021. DOI: 10.1186/s40537-021-00516-9.

[15] F. O. Odhiambo, "Comparative study of various methods of handling missing data," *Mathematical Modeling And Applications*, vol. 5, no. 2, p. 87, 2020. DOI: 10.11648/j.mma.20200502.14. [Online]. Available: https://doi.org/10.11648/j.mma.20200502.14.

[16] Y. Zhou, S. Aryal, and M. R. Bouadjenek, "Review for handling missing data with special missing mechanism," *arXiv*, 2024. arXiv: 2404.04905 [stat.ME]. [Online]. Available: https://doi.org/10.48550/arxiv.2404.04905.

[17] E. A. Zúñiga, S. C. Chambi, C. C. Carbajal, *et al.*, "La correlación de pearson o de spearman en caracteres físicos y textiles de la fibra de alpacas," *Revista de Investigaciones Veterinarias del Perú*, vol. 33, no. 3, e22908–e22908, 2022.

[18] D. L. Banks and S. E. Fienberg, "Data mining, statistics," in *Elsevier eBooks*, Elsevier, 2003, pp. 247–261. [Online]. Available: https://doi.org/10.1016/b0-12-227410-5/00164-2.

[19] M. De Rooij and W. Weeda, "Cross-validation: A method every psychologist should know," *Advances in Methods and Practices in Psychological Science*, vol. 3, no. 2, pp. 248–263, 2020. DOI: 10.1177/2515245919898466.

[20] S. P. M. J, "A comparison of six methods for missing data imputation," *Journal of Biometrics & Biostatistics*, vol. 6, no. 1, 2015. DOI: 10.4172/2155-6180.1000224.

[21] V. Magar, D. Ruikar, S. Bhoite, and R. Mente, "Innovative inter quartile range-based outlier detection and removal technique for teaching staff performance feedback analysis," *Journal of Engineering Education Transformations*, vol. 37, no. 3, pp. 176–184, 2023. DOI: 10.16920/jeet/2024/v37i3/24013.

[22] R. B. Cleveland, W. S. Cleveland, and I. Terpenning, "Stl: A seasonal-trend decomposition procedure based on loess," English, *Statistics Sweden (SCB)*, p. 3, Mar. 1990, ISSN: 0282423X.

[23] H. Chen, K.-T. Shao, and H. Kishino, "Bayesian hierarchical anova model of stochastic seasonality for diodon holocanthus in northern taiwan," *Journal of Marine Science and Technology*, vol. 24, no. 2, pp. 303–310, 2016. DOI: 10.6119/JMST-015-0428-1.

[24] D. L. Greene, "Comparison of t-tests for differences in sexual dimorphism between populations," *American Journal of Physical Anthropology*, vol. 79, no. 1, pp. 121–125, 1989. DOI: 10.1002/ajpa.1330790113. [Online]. Available: https://doi.org/10.1002/ajpa.1330790113.

[25] M. W. Fagerland, "T-tests, non-parametric tests, and large studies—a paradox of statistical practice?" *BMC Medical Research Methodology*, vol. 12, p. 78, 2012. DOI: 10.1186/1471-2288-12-78. [Online]. Available: https://doi.org/10.1186/1471-2288-12-78.

[26] R. Wall Emerson, "Mann-whitney u test and t-test," *Journal of Visual Impairment & Blindness*, vol. 117, no. 1, pp. 99–100, 2023. DOI: 10.1177/0145482X221150592. [Online]. Available: https://doi.org/10.1177/0145482X221150592.

[27] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017. DOI: 10.48550/arxiv.1706.03762.

[28] T. B. Brown, B. F. Mann, N. C. Ryder, *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020. DOI: 10.48550/arxiv.2005.14165.

[29] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *Journal of Machine Learning Research*, vol. 15, no. 90, pp. 3133–3181, 2014. [Online]. Available: http://jmlr.org/papers/v15/delgado14a.html.

[30] scikit-learn, *Histgradientboostingregressor*, Accessed: 2024-06-20, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.HistGradientBoostingRegressor.html.

[31] A. Perez-Lebel, G. Varoquaux, M. L. Morvan, J. Josse, and J. Poline, *Benchmarking missing-values approaches for predictive models on health databases v2*, 2022. [Online]. Available: https://doi.org/10.17504/protocols.io.b3nfqmbn.

[32] F. Santore, E. C. de Almeida, W. H. Bonat, E. H. M. Pena, and L. E. S. de Oliveira, "A framework for analyzing the impact of missing data in predictive models," *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20)*, p. 4, 2020, October 19–23, 2020, Virtual Event, Ireland. DOI: 10.1145/3340531.3412129. [Online]. Available: https://doi.org/10.1145/3340531.3412129.