ro

Machine Learning

Ready, Steady Ride

Group project 2 - Clustering

Group7

28/05/2023

Stefano Sperti 20222246

Anna Kwiatkowska 20222216

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

ABSTRACT

The purpose of this report is to develop a clustering model that identifies the key factors influencing customer behavior and potential customer groups for Ready, Steady Ride, a ride-sharing company. The model aims to adapt the company's services to changes in climate that impact customer behavior. The project utilized a dataset comprising approximately 9,600 records provided by the company. Prior to applying machine learning algorithms such as K-means and K-modes, the data underwent exploration and preprocessing. To enhance our model, various profiles were employed to identify valuable patterns. Furthermore, we explored the possibility of clustering using both numerical and categorical variables. The analysis delivered patterns in different customer behaviors, which can be translated into four association rules. Remarkable features influencing customer behavior include the time of day, weather conditions, and whether it was a working day.

KEYWORDS

Machine Learning, Clustering, Customer Behavior

CONTENTS

INTRODUCTION

The Ready, Steady Ride ride-sharing service has recognized the importance of adapting their services to specific customer behaviors, as well as acknowledging the impact of climate change on traditional approaches to monitoring seasonal effects. Seasonality plays a crucial role in resource allocation, making it essential to incorporate it into the company's services. The dataset we have analyzed consists of approximately 9,600 records collected by the company over a one-year period.

The objective of this report is to utilize the available data to identify patterns that can lead to better decision-making regarding the optimal number of riders required at any given time. Additionally, we aim to develop a more automated approach to determine the suitable equipment for each rider during their shifts. We will employ unsupervised machine learning techniques, to achieve:

- Identifying the key variables that significantly influence the differentiation in customer behaviors.
- Establishing the relationship between weather conditions and customer behavior to determine the number of ride bookings and necessary equipment for the ride.

The following report offers an in-depth overview of the methodology employed to develop an optimal model. It encompasses various stages, beginning with data exploration and preparation, and culminating in the selection of the most suitable model. Moreover, the report presents thorough conclusions derived from the project, providing a comprehensive analysis of the findings.

BACKGROUND

SILHOUETTE SCORE

The silhouette score is a widely used metric for evaluating the quality of clustering results. It provides a measure of how well individual data points fit within their assigned clusters. The silhouette score takes into account both the cohesion and separation of the data points.

To calculate the silhouette score for a data point, several steps are involved. First, the average distance between the data point and all other points within its own cluster is computed. This represents the cohesion, indicating how well the data point is similar to other points in the same cluster. Next, the average distance between the data point and all points in the nearest neighboring cluster is calculated. This represents the separation, indicating how well the data point is dissimilar to points in other clusters.

The silhouette score is then computed as the difference between the average nearest cluster distance and the average intra-cluster distance, divided by the maximum value between these two distances. The silhouette score ranges from -1 to 1, where a higher score indicates better clustering results. A score close to 1 suggests that the data point is well-clustered, with a clear distinction from other clusters. Conversely, a score close to -1 suggests that the data point may be assigned to the wrong cluster.

The overall silhouette score for a clustering solution is the average silhouette score across all data points. A higher average silhouette score indicates better separation and cohesion of clusters, suggesting more reliable and distinct clusters in the data.

The silhouette score is a valuable tool for comparing different clustering algorithms or assessing the optimal number of clusters. It helps practitioners make informed decisions about the quality and appropriateness of a clustering solution based on the inherent structure and relationships within the data.

K - MODES

K-modes clustering is a robust algorithm specifically developed to handle categorical data analysis, offering a versatile solution for various clustering tasks. Unlike traditional algorithms that focus on numerical attributes, K-modes effectively clusters instances based on categorical variables. By considering the unique characteristics of categorical data, it can uncover patterns and relationships that might be hidden in other methods.

One of the key advantages of K-modes is its ability to directly handle categorical variables without requiring complex transformations or encoding schemes. This simplifies the clustering process and ensures that the original nature of the data is preserved. Moreover, K-modes produces easily interpretable results, allowing analysts to understand and explain the cluster assignments and the characteristics of each cluster. This interpretability facilitates decision-making and the development of actionable insights.

Efficiency and scalability are also significant advantages of K-modes. The algorithm can handle large datasets with a high number of categorical attributes and instances. Its computational efficiency ensures that clustering results can be obtained in a reasonable amount of time, even for substantial datasets. This scalability is especially beneficial for real-time and big data applications, where efficient processing is essential.

The algorithm begins by randomly initializing K cluster centroids. Each centroid represents a cluster's prototype or mode. Then, it iteratively assigns data points to the nearest cluster based on a dissimilarity measure called the mode dissimilarity. The mode dissimilarity calculates the dissimilarity between two categorical data points by counting the number of features (variables) that differ between them.

During the assignment step, each data point is assigned to the cluster with the most similar mode. After assigning all data points, the algorithm updates the cluster modes by choosing the most frequent value for each feature within the cluster. This process continues iteratively until convergence is reached, meaning that the cluster assignments no longer change significantly.

In summary, K-modes clustering offers several advantages, including its direct handling of categorical variables, interpretability of results, efficiency, scalability, flexibility in capturing different cluster shapes, and wide applicability. These benefits make it a valuable algorithm for extracting valuable insights from categorical data in various domains.

METHODOLOGY

DATA EXPLORATION AND UNDERSTANDING

After an initial examination of the dataset in order to gather basic information, the variables have been categorized into three groups: metric and categorical variables, and "not a feature" variable which was “Unnamed: 0”. Boxplots were utilized to identify any potential outliers. Furthermore, an assessment of missing values and duplicated rows was conducted. It was discovered during the exploration that the variable "City_Detroit" had a single value, namely "1" meaning the variable had no additional information to our dataset.

Additionally, we came across an important finding aligned with the project description provided to us. It became evident that the RidesBooked variable does not represent the sum of RegisteredUsers and NonRegisteredUsers, by looking at the histograms. It appears that the RegisteredUsers and NonRegisteredUsers variables actually indicate the number of clients rather than the number of rides booked. This explanation effectively captures the left side of the histograms, but fails to accurately describe the right side. The histograms for the variables RidesBooked, RegisteredUsers and NonRegisteredUsers are shown respectively in Graph 1, 2 and 3. Indeed, the sum of the two variable is the same in the value zero and since there are no rides booked it cannot be the case that some costumers buy some rides. Another possible explanation that well explain the histogram is that also RidesBooked doesn’t represent the number of rides but the number of person that has utilize the service. However, from now on we will consider these variables in the context of client count throughout the remainder of the report. This also doesn’t change the reasoning because we can observe that the trend of the histogram is the same and after using the MinMax Scaling the numbers for a sum of RegisteredUsers and NonRegisteredUsers, and RidesBooked are quite similar.

During our exploration of the dataset, we observed that a majority of the variables exhibited skewness, either to the right or left. Particularly intriguing was the variable "Rides Booked," as shown in Graph 1. Notably, we noticed a considerable number of rows where the value was "0," indicating instances where the company failed to attract customers to book a ride. This insight can prove valuable in understanding customer behaviors and preferences.

Furthermore, we observed the presence of “blank spaces” in some of the variable histograms. This occurrence is likely a result of rounding the values to either the minimum or maximum value. Graphs 4 and 5 illustrate examples of these “blank spaces”.

DATA PREPROCESSING

In the preprocessing phase of the project, we initiated by converting the values of the variables "DayofWeek" and "Month" from letters to numeric representations. This transformation was necessary to facilitate further analysis. Additionally, we processed the categorical variable “weather conditions” to ensure it is in a suitable format for next operations that will require both binary and ordinal numeric representation.

The next step involved addressing outliers within the dataset, since there were no missing values or duplicated rows. Due to the considerable presence of outliers, their removal would potentially impact the models and final results. Recognizing that the subsequent use of a k-means model relies on the distance between data points, we determined that relocating all outliers to the whiskers of the boxplots would be an appropriate solution. In the boxplots the whiskers were at a distance of 1.5 times the interquartile range (IQR). Considering the potential negative consequences of completely removing outliers, we believed that relocating them would have a less negative effect on the results. This approach effectively reduces the impact of outliers without sacrificing any valuable information contained within these data points.

Moreover, we have created a new feature known as "Date" by combining a date from the numerical values of the Day of the Month and Month variables. As the variables were not chronological, the values of feature “Date” are not sequential days, but more like a “first day of the first month”. By creating this feature, we aimed to explore potential patterns or trends that may appear.

Subsequently, we grouped the variables by Date, DayofWeek, HourofDay, and Month. This step was taken to enhance our understanding and visualization of the variables, as well as to identify any discernible patterns. Afterwards, the variable "City_Detroit" was removed from the dataset as it contained a single value and did not contribute any valuable information.

SCALING AND FEATURE SELECTION

Considering the non-normal distribution of variables, the application of MinMax Scaling became essential. The objective behind feature scaling is to bring all the variables onto a unified scale, particularly beneficial for algorithms that rely on distance-based metrics as K-means, which we will use as a model for clustering.

Subsequently, we conducted Spearman correlation analysis to identify the variables with the strongest correlations among each other. To enhance the clarity of the associations, the results were visualized using a heatmap. During our analysis, we noticed a correlation between the variables Temperature and FeltTemperature, as well as between the variables RidesBooked, RegisteredUsers, and NonRegisteredUsers. Despite this correlation, we made a decision to retain all variables in the dataset at this stage.

The following step required grouping all the variables into 3 different perspectives. This approach, utilizing perspectives in clustering, enabled improved divisions and resulted in better outcomes. The creation of these groups was guided by logical reasoning and the observed similarities between the variables. A comprehensive overview of the perspectives and their respective components is presented in Table 1. Furthermore, we conducted an analysis to examine other relationships between variables across different perspectives, and no significant correlation between different perspectives was found.

The initial perspective, referred to as "weather," aims to identify differences in the weather conditions. It is important to note that we explored two different approaches within this perspective to determine which one provided better results, considering the presence of both numerical and categorical variables. The second perspective focused on the consumer's standpoint and examined variables such as RidesBooked, RegisteredUsers, and NonRegisteredUsers. Lastly, we created the "Date" perspective, to analyze how the day type and hours affected customer behavior.

MODELS

Considering the two primary clustering approaches, namely hierarchical clustering and K-Means, we have chosen to utilize the latter in this project. K-Means clustering is a widely adopted unsupervised machine learning algorithm employed for clustering analysis. Its objective is to group data points with similar features into clusters. The algorithm operates through iterative steps, where data points are assigned to the nearest centroid and the centroids are subsequently updated to minimize the overall within-cluster variance.

The K-Means algorithm offers efficient computation, which is particularly advantageous for our dataset, which can be classified as medium-sized. Moreover, since the objective of this project is to identify distinct types of customer behavior, the K-Means algorithm performed better. This clustering algorithm is well-suited for scenarios with well-separated and spherical clusters. With 3 perspectives established, the modeling phase remains consistent across each perspective, and a detailed description of the model is presented below.

To diversify our approaches, we made the decision to explore alternative algorithms that could be suitable for our dataset and project objectives. After careful consideration, we chose to explore the K-Modes algorithm, which is well-suited for categorical variables as it utilizes modes. For a more comprehensive understanding of this approach, please refer to the Background section where it is elaborated upon in detail.

Number of clusters

To begin with, it is essential to determine the optimal number of clusters as a requirement for the K-Means as well as for the K-Modes algorithms. To accomplish this, we have chosen to employ three distinct methods to identify the appropriate value for "k" - the number of clusters. By utilizing a majority vote approach, if two or more methods indicate the same value for "k," it is considered the winning choice. However, this initial reasoning only served as a starting point, suggesting a reasonable number of clusters. As the analysis progressed and the initial results were observed, the final number of clusters could be adjusted to better align with the data.

The first approach that was used is known as the "Elbow" method. This technique aids in identifying the "elbow" point within a plot representing the within-cluster sum of squares (WCSS) against the number of clusters. The WCSS measures the sum of squared distances between each data point and its respective centroid within a cluster. The elbow point signifies the number of clusters at which further additions do not significantly enhance the clustering performance.

The second approach involved the dendrogram method, which is a technique utilized to determine the optimal number of clusters. A dendrogram is a tree-like diagram that depicts the hierarchical relationships between data points and clusters. By analyzing the structure of the dendrogram, one can identify the suitable number of clusters.

The third approach included the use of the silhouette score, a widely adopted metric for assessing the quality of clustering outcomes. This score offers a measurement of how well individual data points align with their assigned clusters, considering both the cohesion and separation aspects. For a comprehensive understanding of this method, please refer to the Background section of this report, which provides a detailed description.

Creating clusters

To create clusters using the k-means algorithm, several necessary steps need to be followed. The detailed procedure is outlined below:

1. After selecting the optimal number of clusters (k), initialize the centroids by randomly selecting k data points from the dataset. These centroids will act as representatives for the clusters.
2. Assign data points to clusters by calculating the distance between each data point and the centroids using a distance metric such as Euclidean distance. Assign each data point to the cluster with the nearest centroid.
3. Update the centroids by recalculating the centroids of each cluster. This is done by taking the mean of the data points assigned to that cluster and moving each centroid to the newly calculated mean location.
4. Repeat the steps of assigning data points to clusters and updating the centroids until convergence is achieved. Convergence occurs when the centroids no longer move significantly or when the maximum number of iterations is reached.
5. Finally, assign each data point to a final cluster based on the updated centroids, resulting in the formation of distinct clusters.

Visualization and exploration

Once the clusters have been formed, it is essential to visualize and analyze the results. Firstly, we computed the fundamental statistics for each perspective and cluster, including the mean and standard deviation. These statistics provide insights into the distribution and variations among the clusters. Additionally, we generated subplots to illustrate the standard deviation of specific variables. These subplots provide a visual representation of the variation in the data for each variable, allowing us to examine the dispersion and identify any noteworthy patterns or differences.

Secondly, we generated plots using various variables and dimensions to gain a deeper understanding of the clusters and their characteristics. We utilized parallel plots to display the mean values of each variable and cluster on the Y-axis, while the variables were represented on the X-axis. Additionally, we generated graphical plots displaying the mean and standard deviation of each cluster, providing a visual representation to assess any potential statistical differences among the clusters. Furthermore, we generated scatterplots and histogram to enhance our comprehension of the clusters.

Moreover, we applied the t-SNE algorithm to visualize our findings. T-SNE is a non-linear dimensionality reduction technique utilized for visualizing high-dimensional data in a lower-dimensional space, often 2D or 3D. It aims to preserve the local relationships between data points while mapping them from a high-dimensional space to a lower-dimensional one. By modeling the pairwise similarities between data points in both the original high-dimensional space and the reduced space, t-SNE ensures that similar points are represented as close neighbors, while dissimilar points are positioned further apart.

The primary objective of this phase is to identify and describe the distinctions between clusters. The existence of significant variations between the clusters across various dimensions would signify the successful realization of the project goals.

Models Assessment

Once the clusters were created using the models, it became necessary to evaluate our methodologies and identify opportunities for improvement.

Due to the challenges encountered with one of the first perspectives, namely Weather_conditions1, which comprised both numerical and categorical variables, we made the decision to remove this perspective from the analysis. This step was taken due to the limitation of the K-Means algorithm when dealing with numerous binary variables. Interestingly, we observed that the cluster effectively separates the binary variable, assigning it higher importance compared to the other variables. The models did not perform adequately with this combination of variables. This situation prompted us to investigate the underlying reasons and identify potential areas for improvement as we will show below.

Subsequently, we checked whether varying the number of clusters (k) would have any noticeable impact and improve our outcomes. As outlined in the Models section, we used three distinct methods to determine the optimal value for "k". However, there were instances where these approaches indicated more potential solutions. During the evaluation of the models, we examined more values for “k” and selected the more favorable option. Our selection was based on assessing the dissimilarities between clusters, such as differences in the values of variables. To address this, a graphical approach is employed, involving the creation of graphs to better comprehend the clusters.

Additionally, during the creation of clusters using both the "Temperature" and "FeltTemperature" variables, we noticed signs of potential overfitting in the models. To address this concern, we decided to only use the "FeltTemperature" variable.

Upon analyzing the K-Means algorithm, we discovered that including numerous variables simultaneously could lead to issues and produce undesirable outcomes. Furthermore, combining categorical and numerical variables had the potential to disrupt the algorithm. The detailed explanation of our methods to address this issue is provided below.

Clustering with both numerical and categorical variables

Clustering datasets with a mix of numerical and categorical variables presents challenges due to the differences in their nature and representation. Categorical variables lack a natural numerical scale, making it challenging to calculate meaningful distances or similarities. Numerical variables, on the other hand, may vary widely in scale and distribution, potentially impacting the clustering process. Combining these variables in a single clustering analysis can lead to suboptimal results and obscure valuable patterns within the data. To address this issue, we have investigated four distinct approaches, outlined and described as follows:

Creating Dummy Variables - One common approach involves transforming categorical variables into dummy variables (therefore numerical). This technique converts each category into a binary feature, enabling their inclusion in the clustering analysis. By encoding categorical variables numerically, clustering algorithms designed for numerical data can be utilized effectively. We have realized that this approach does not yield the desired results as shown in perspective Weather_conditions1.

Pespectives - Another strategy involved creating subgroups based on similar meaning within each cluster. This process entails identifying subsets of data with similar categorical variable and numerical and treating them as distinct subclusters. By isolating the impact of categorical variables within clusters, it becomes possible to explore specific patterns or relationships unique to each subgroup.

Separate Treatment of Variable Types - An additional approach focused on treating categorical and numerical variables separately during the clustering process. This strategy allows for a more focused analysis of each variable type, facilitating the identification of clusters driven predominantly by either categorical or numerical features. Examining the resulting clusters individually provides a more comprehensive understanding of the underlying patterns. Also this approach does not yield the desired results.

Groping numerical into categorical - By incorporating one categorical variable for each cluster and computing statistics such as the median or mean for each numerical variable grouping by each value of the categorical variable, we aim to enhance the clustering performance and attain a more profound understanding of the data. The computation of statistics like the median or mean allows us to obtain a representative value that summarizes each numerical variable. This valuable information assists in the interpretation and labeling of the clusters based on their distinct characteristics.

Profiles

In this paragraph, our aim was to explain the concept of "profiles'' and their role in improving our understanding of the dataset and clusters. Table 2 presents all the profiles that we have used and their components. The profiles refers to the “Groping numerical into categorical” in the section above.

The profiles "Average_weatherconditions" and "weatherconditions_rides" show different approaches for clustering using the same information. The first profile employed average values for each feature across days, while the second profile focused on hourly data. The parallel plots for these clusters are shown on Graph 6 and 7. These two clusters underscored the significance of the "HourofDay" variable and revealed that averaging the variable fails to capture consumer behavior throughout the day. The key concept is that individuals make decisions about booking a ride based not solely on the overall weather conditions of the day, but rather on the specific hour they need to order as shown on Graph 8. For example, if someone is going out at night, there is no need to order a ride in the afternoon.

. Moving on, we have used the profile "averages_dayofweek" to understand the difference in behavior on different days of the week. In this profile we used the mean of the numerical variable grouped by the DayoftheWeek. Highlighted a noticeable pattern, as illustrated in Graph 9 and 10, indicating increased ride bookings on two specific days. However, due to the dataset structure where days were represented as not ordered variable, we couldn't definitively specify the exact days exhibiting this trend. It can be hypothesized that weekends were the primary contributors. To support this hypothesis, we examined the histogram of the variable WorkingDay in this profile. The findings revealed a graphically statistically significant difference among the clusters as shown in Graph 11.

The subsequent profile, "Peak_hours," aimed to explore the impact of HourofDay on customer behavior. Unfortunately, due to the presence of categorical variables, the profile delivered unreliable results as the algorithm struggled to accurately measure distances between hours. Indeed, the K-means fails to capture the fact that the distance between 24 and 1 is the same as between 15 and 16. On the other hand, the "Peak_hours2" profile, which was a modified version of the previous profile without categorical variables, clearly showcased variations in customer behavior based on the time of day and user registration. The plots for this profile are presented on Graph 12. For instance, it was observed that in the afternoon, there was a higher number of non-registered users booking rides, potentially tourists, while more registered users tended to book rides in the morning or evening, likely for commuting purposes. Furthermore, there was a decrease in ride bookings during nighttime, indicating a period of reduced activity as shown in graph 13.

Merging the perspectives

We also have significant result using the perspective approach. The initial perspective, "weather," generated two clusters: cluster0, representing good weather conditions characterized by lower humidity and windspeed, and cluster1, depicting unfavorable weather conditions with higher humidity and lower windspeed. The plots for this perspective are presented on Graph 14. Moving on to the "customer" perspective, two relatively straightforward clusters emerged, yet they proved valuable in integrating the perspectives. Cluster0 could be described as positive, indicating a high volume of rides booked and a large number of users, whereas cluster1 indicated a lower count of rides and users. The plots for this perspective are presented on Graph 15. Lastly, the "Date" perspective resulted in three distinct clusters: cluster0, consisting of working days in the morning; cluster1, corresponding to not working days and holidays; and cluster2, representing working days in the afternoon and evening. The plots for this perspective are presented on Graph 16.

Afterwards, we integrated all of the three perspectives to obtain comprehensive outcomes. In pursuit of this objective, we employed three distinct methodologies and subsequently selected the most effective one based on the observed results. Initially, we implemented the K-means algorithm. However, the presence of both categorical and numerical variables led to unsatisfactory outcomes. Consequently, we explored the K-modes algorithm, which provided promising final clusters. Still, we decided to further improve our approach.

The final approach involved integrating the deterministic method with the KNN imputer. Initially, we generated all possible clusters by combining the outcomes from the three perspectives. Subsequently, we selected the four more significant clusters. For the remaining data points, we replaced the column that identify “the number of cluster” with NA. We then applied the KNN imputer to calculate and adjust the remaining data based on these four clusters. The KNN imputer determined which cluster was closest to and assigned the datapoint to this cluster. Finally, we assessed whether this approach was effective by verifying if the division into the four clusters remained accurate using scatterplot and histogram.

Upon evaluating the outcomes, we concluded that the third method would be the most suitable choice. Remarkably, both the K-modes and the deterministic method provided comparable results. However, we opted for the third approach, because it was easier to interpret. Unlike in K-modes, there was no need to convert every variable to binary, making the results more straightforward to understand.

RESULTS

In this section, our aim was to present the ultimate outcomes of our work, namely the clusters that have been formed.

After consolidating the perspectives, we obtained four distinct and conclusive clusters that illustrate the potential impact of customer behavior on the number of booked rides as shown in graph 17. The quantitative analysis for each cluster for variables RidesBooked, WeatherForecast, HourofDay and FeltTemperature are presented in Table 3, 4, 5 and 6 respectively. Additionally, a comprehensive description of each cluster is provided as follows:

- Cluster "0" outlined the morning of a working day characterized by not really good weather conditions. In this cluster, there were only a few rides booked.
- Cluster "1" showcased the afternoon of a working day with favorable weather, where a significant number of rides were ordered. These could be customers who work during the day and want to make the most of the sunny afternoon by engaging in outdoor activities.
- Cluster "2" portrayed days with extremely poor weather conditions, particularly high humidity. The time of day and type of day were not significant factors here, as customers did not book rides.
- Cluster "3" illustrated holidays with good weather. During these days, customers booked a large number of rides, resulting in significant revenue generation.

The analysis of the profiles "Average_weatherconditions", "weatherconditions_rides", "Peak_hours" and "Peak_hours2" provided also some valuable insights into different aspects of customer behavior and clustering. These profiles emphasized the importance of the "HourofDay" variable, revealing that people's decisions regarding booking rides are influenced more by the specific hour they need to go out rather than the overall weather conditions of the day. The analysis of "averages_dayofweek" shows the difference in customer behavior on different days of the week.

In conclusion, customer behavior was primarily influenced by weather conditions and the type of day (working day or holiday). Furthermore, we observed variations in customer behavior throughout different hours of the day. The patterns related to weather conditions can be summarized by noting that on good days (low humidity, higher windspeed, and favorable weather conditions), people tend to book more rides compared to bad days (high humidity, lower windspeed, and unfavorable weather conditions).

CONCLUSION

The primary objective of this report was to determine the key variables that impact customer behavior when booking a ride, as well as the significance of weather conditions in influencing customer behavior. To achieve these objectives, the dataset underwent a series of processing steps, including identifying and resolving inconsistencies, preprocessing the data, eliminating the redundant variables and evaluation of the models.

The outcomes of our analysis present the final four clusters, which can be represented as association rules:

- In the mornings of working days with poor weather, there is a low number of rides ordered.
- Regardless of the time of day, when the weather conditions are unfavorable, there is a minimal number of booked rides.
- During favorable weather, particularly in the afternoons of working days, there is a high volume of rides booked.
- On holidays (non-working days) with favorable weather, there is a significant increase in the number of rides booked.

Moreover, we identified the variables that have the greatest influence on customer behavior. In descending order of relevance, they are: HourofDay, Weather Forecast, and WorkingDay.

However, we have identified potential areas for improving our project. One suggestion involves incorporating statistical tests to evaluate the differences between clusters, which could provide further validation for our cluster decisions. Additionally, exploring alternative clustering methods or algorithms is worth considering.

REFERENCES

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. John Wiley & Sons.

Maulik, U., & Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(12), 1650-1654.

Cao, F., Liang, J., & Bai, L. (2009). A new initialization method for categorical data clustering. Expert Systems with Applications, 36(7), 10220-10226.

Zou, C., Li, Y., Liu, Y., & Li, J. (2013). K-modes clustering algorithm for categorical data sets. Expert Systems with Applications, 40(18), 7236-7241.

Huang, Z. (2011). A fast clustering algorithm to cluster very large categorical data sets in data mining. Data Mining and Knowledge Discovery Handbook, 2nd Edition, 1107-1118.

APPENDIX

Table 1 Perspectives and their components

Perspective	Name in the code	Variables
Weather	Weather_conditions1	Temperature, FeltTemperature, Humidity, WindSpeed, WeatherForecast_0.0, WeatherForecast_1.0, WeatherForecast_2.0, WeatherForecast_3.0
Weather	Weather_conditions2	FeltTemperature, Humidity, WindSpeed, WeatherForecast_ord
Consumer	Rides_Booked	Nonregisteredusers, Registeredusers, RidesBooked
Date	Bool_date	Holiday, WorkingDay, HourofDay

Table 2 Profiles and their components

Profile	Variables
Averages_weatherconditions	Humidity_avg, FeltTemperature_avg, RidesBooked_avg, WindSpeed_avg, WeatherForecast_ord_med
weatherconditions_rides	Humidity, FeltTemperature, RidesBooked, WindSpeed, WeatherForecast_ord
	WindSpeed_DayofWeek_avg, DayofWeek_RidesBooked_avg, DayofWeek_FeltTemperature_avg, DayofWeek_Humidity_avg, DayofWeek_AverageRideDurationPreviousDay_Min_avg, DayofWeek_ord
Peak_Hours	RidesBooked_Hours_avg, Nonregisteredusers_Hours_avg, Registeredusers_Hours_avg, HourofDay
Peak_Hours2	RidesBooked_Hours_avg, Nonregisteredusers_Hours_avg, Registeredusers_Hours_avg,

Table 3 Quantitative analysis of clusters - RidesBooked

RidesBooked	count	mean	std	min	25%	50%	75%	max
Cluster0	1881.0	205.15	249.4	3.0	24.0	84.0	339.0	1095.0
Cluster1	1239.0	1414.34	558.1	690.0	954.0	1251.0	1750.5	2931.0
Cluster2	1293.0	248.32	202.81	3.0	69.0	201.0	372.0	912.0
Cluster3	790.0	1404.22	398.74	546.0	1080.0	1426.5	1704.0	2349.0

Table 4 Quantitative analysis of clusters - WeatherForecast

WeatherForecast	count	unique	top	freq
Cluster0	1881	3	Clear	999
Cluster1	1239	3	Clear	1085
Cluster2	1293	4	Clear	664
Cluster3	790	3	Clear	703

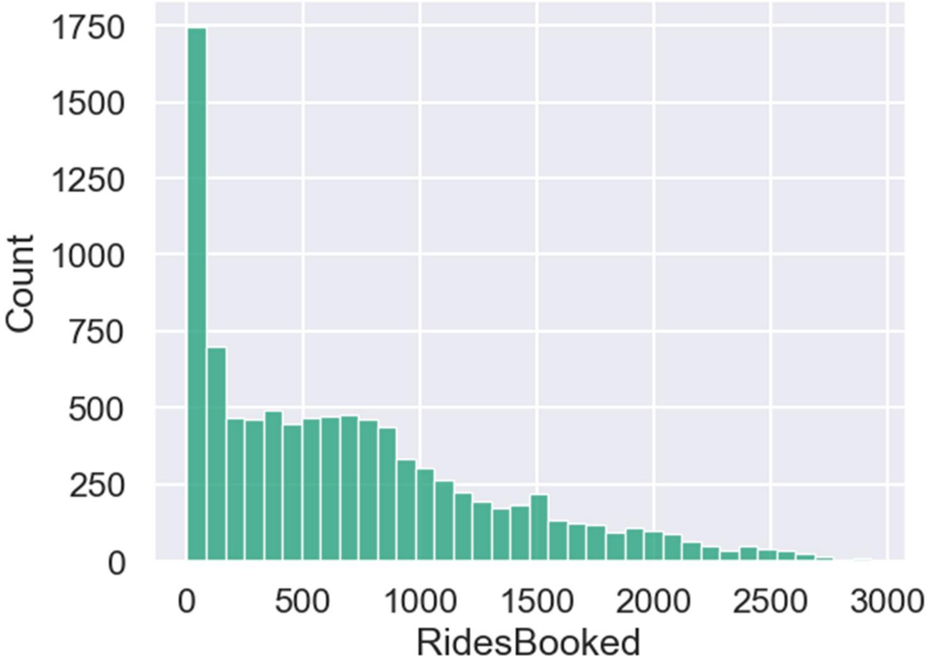
Table 5 Quantitative analysis of clusters - HourofDay

HourofDay	count	mean	std	min	25%	50%	75%	max
Cluster 0	1881.0	4.38	3.17	0.0	2.0	4.0	6.0	11.0
Cluster 1	1239.0	16.48	2.48	12.0	15.0	17.0	18.0	22.0
Cluster 2	1293.0	8.96	7.58	0.0	3.0	6.0	16.0	23.0
Cluster 3	790.0	14.74	3.25	9.0	12.0	15.0	17.0	23.0

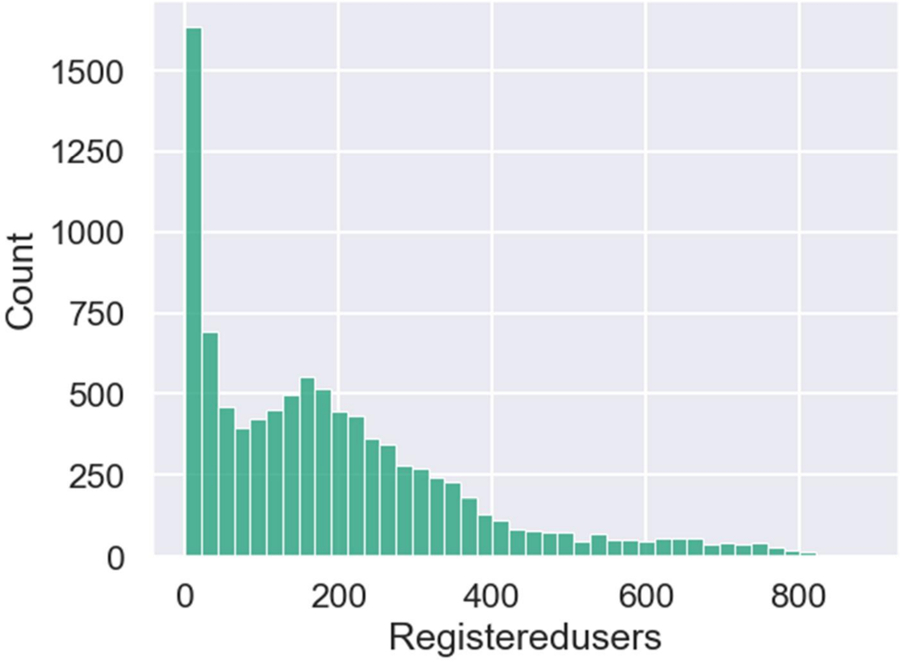
Table 6 Quantitative analysis of clusters - FeltTemperature

FeltTemperature	count	mean	std	min	25%	50%	75%	max
Cluster 0	1881.0	15.4	4.83	3.61	11.33	15.45	19.1	28.85
Cluster 1	1239.0	20.6	4.42	7.73	17.52	21.12	23.7	29.88
Cluster 2	1293.0	14.88	5.12	1.55	10.82	14.94	19.1	27.3
Cluster 3	790.0	20.29	4.7	7.73	17.0	21.12	23.18	30.91

Graph 1 RidesBooked histogram

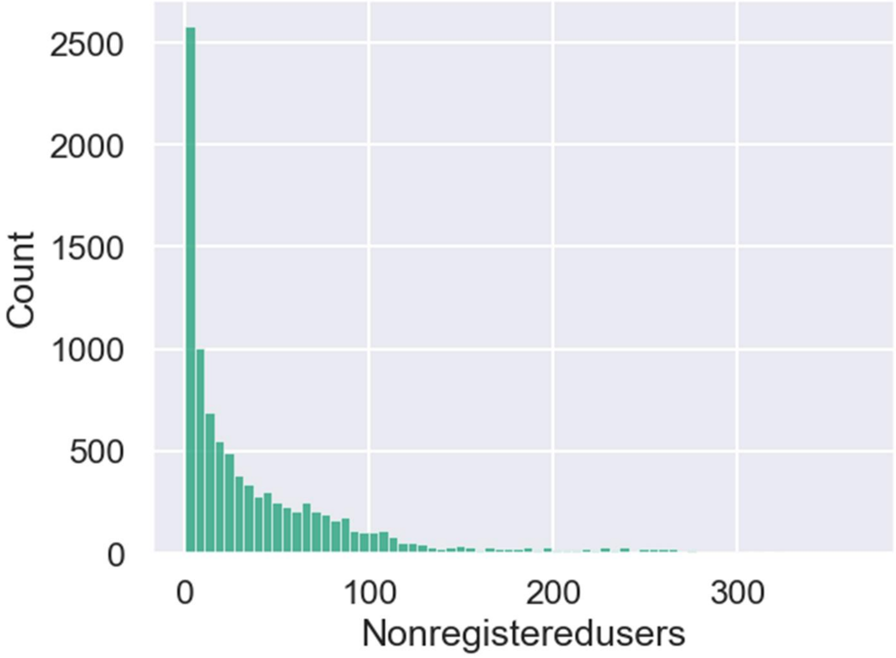


Graph 2 RegisteredUsers histogram

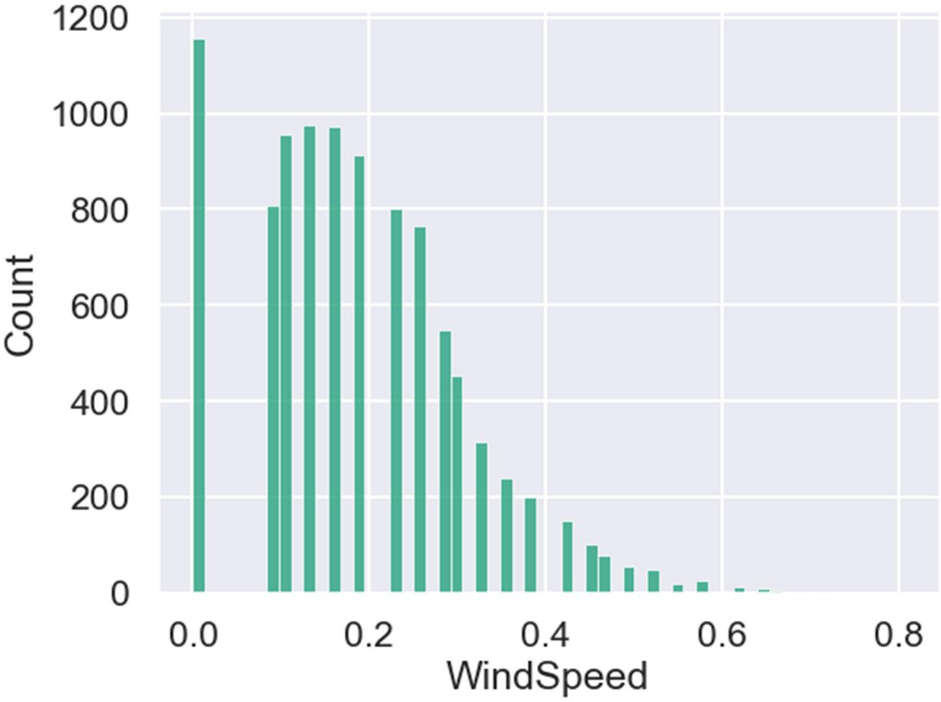


Graph 3 NonRegisteredUsers histogram

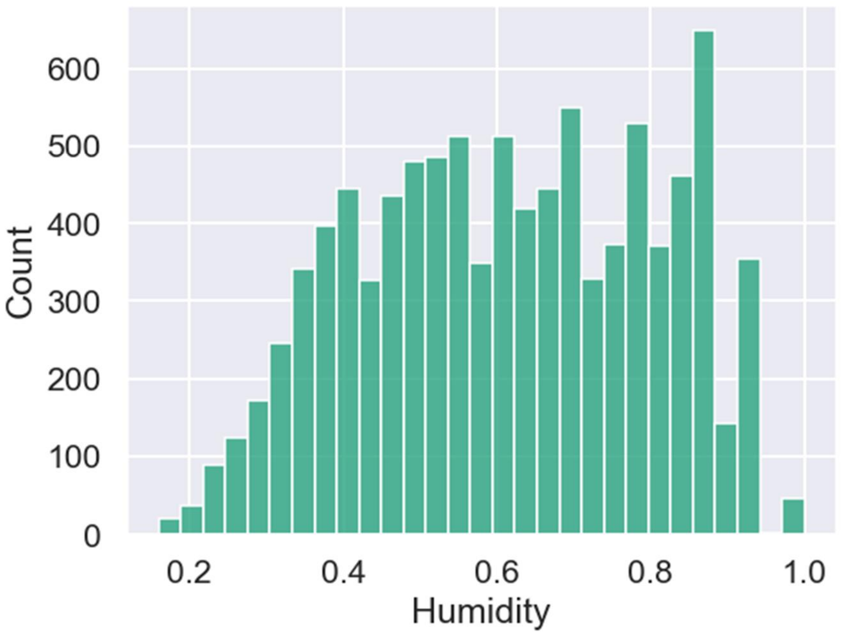
graph



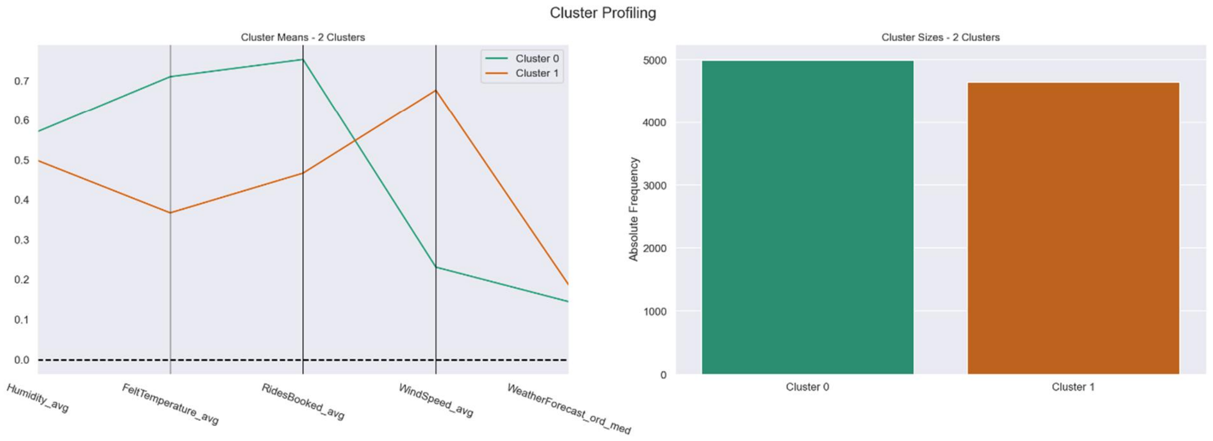
Graph 4 Windspeed histogram



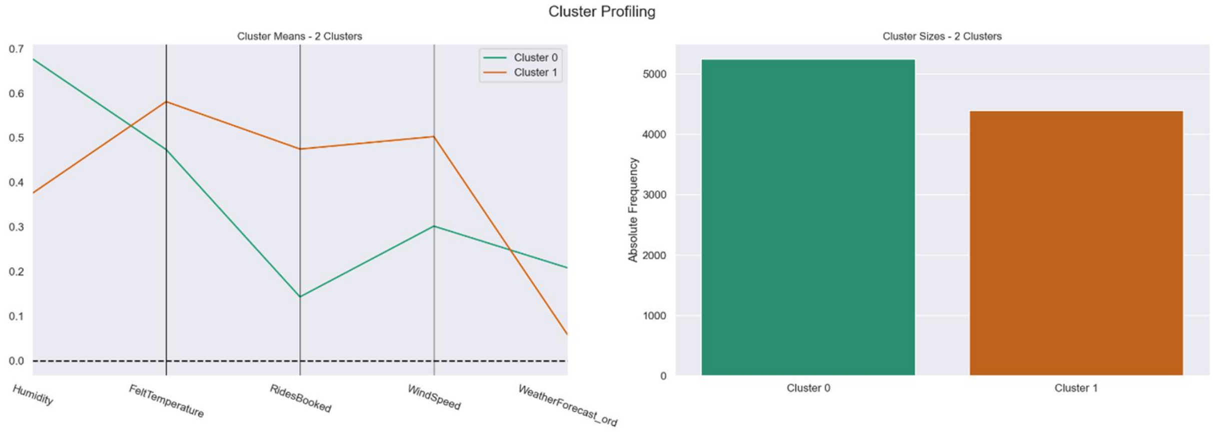
Graph 5 Humidity histogram



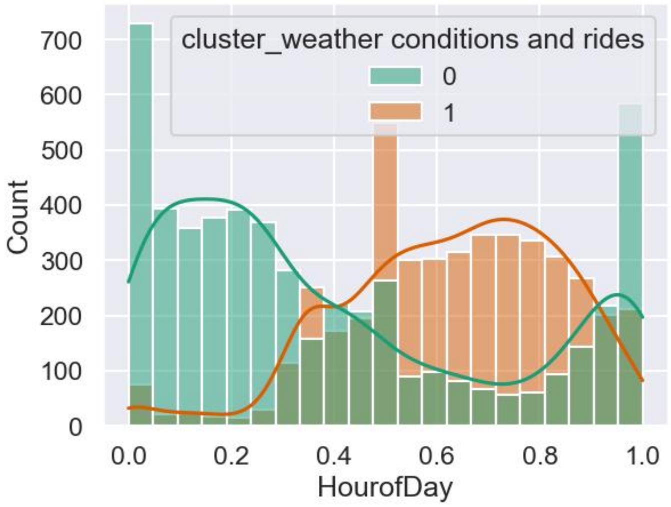
Graph 6 Plots for profile "Average_weatherconditions"



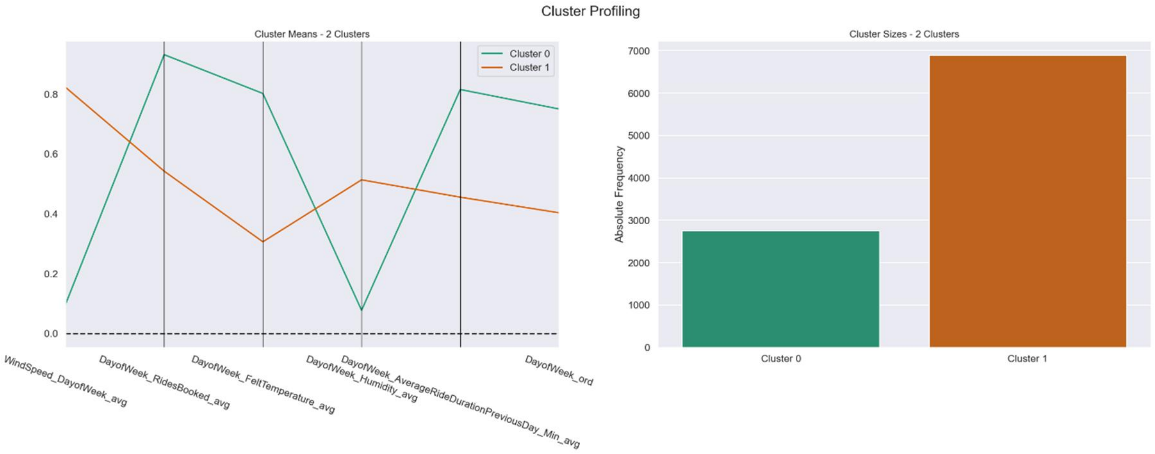
Graph 7 Plots for profile "weatherconditions_rides"



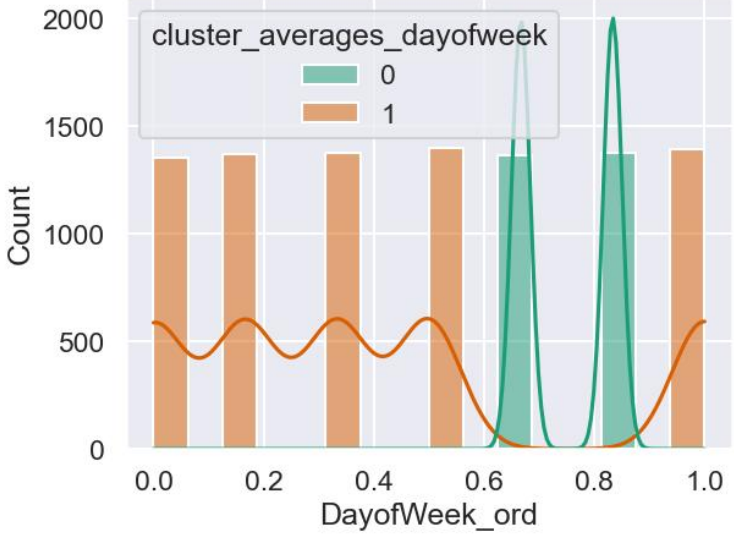
Graph 8 “Hours of the day” histogram for profile "weatherconditions_rides"



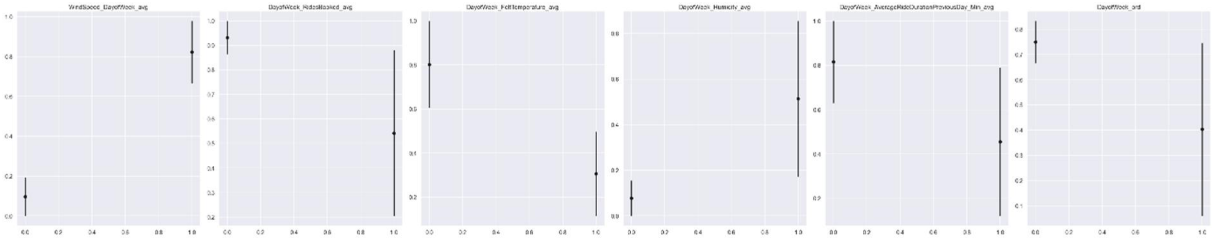
Graph 9 Plots for profile "averages_dayofweek”



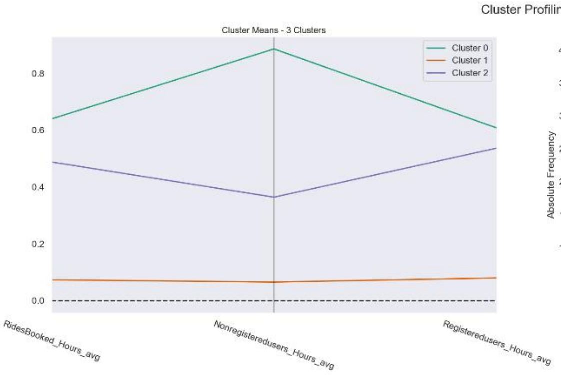
Graph 10 dayofweek histogram for profile "averages_dayofweek”



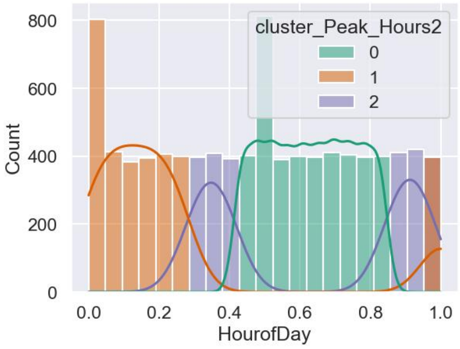
Graph 10 mean and standard deviation for each cluster in profile "averages_dayofweek”



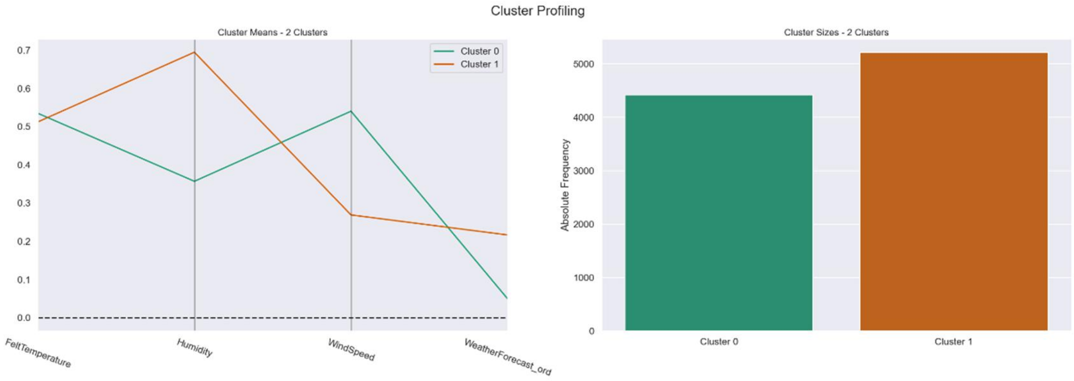
Graph 12 Plots for profile “Peak_hours2”



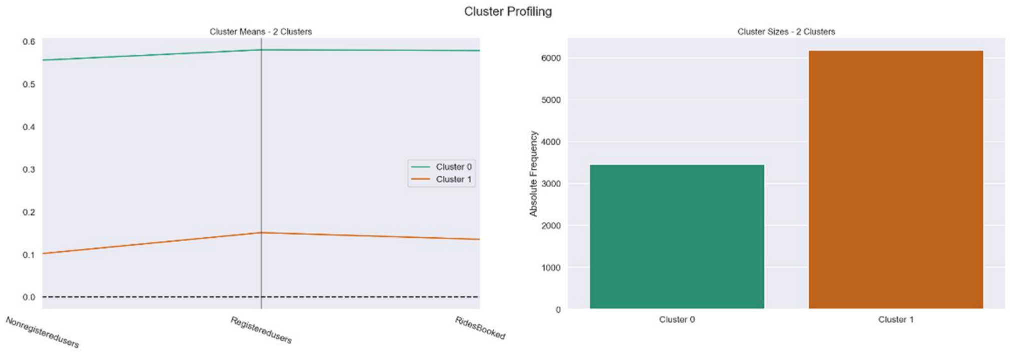
Graph 13 “Hours of the day” histogram for profile " Peak_hours2"



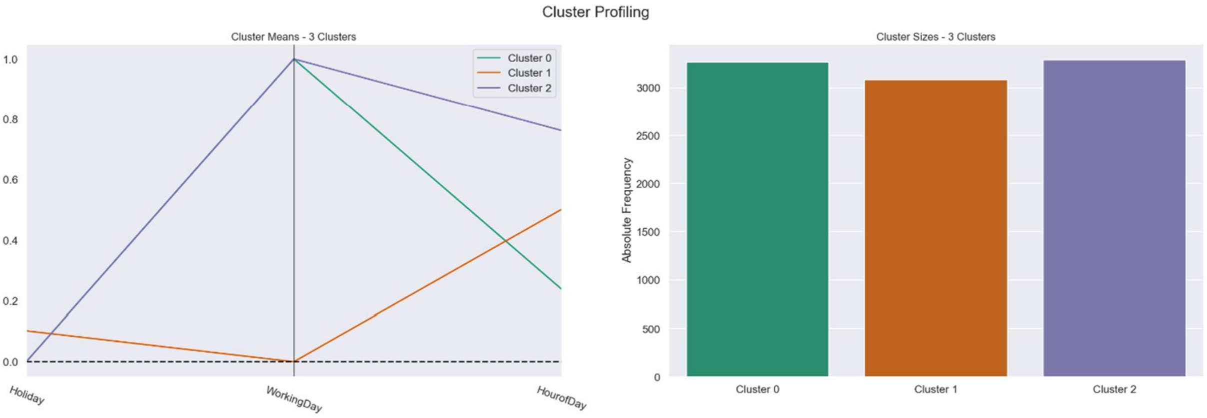
Graph 14 Plots of perspective Weather condition



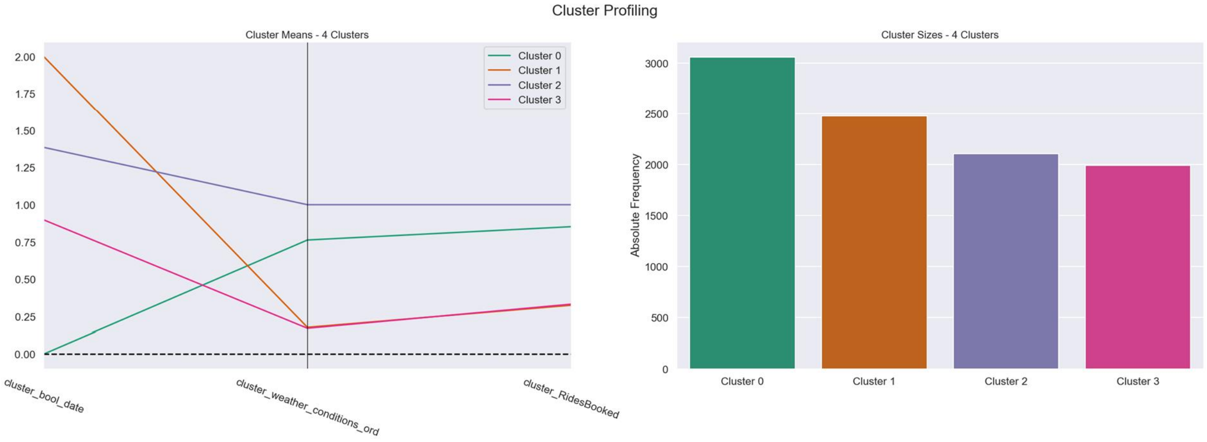
Graph 15 Plots of perspective Rides_booked



Graph 16 Plots of perspective bool_date



Graph 17 Plots for final association rule merging three perspectives



NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa