NOVA IMS Information Management School

Machine Learning – Project 2

Clustering

Group 7

Stefano Sperti 20222246

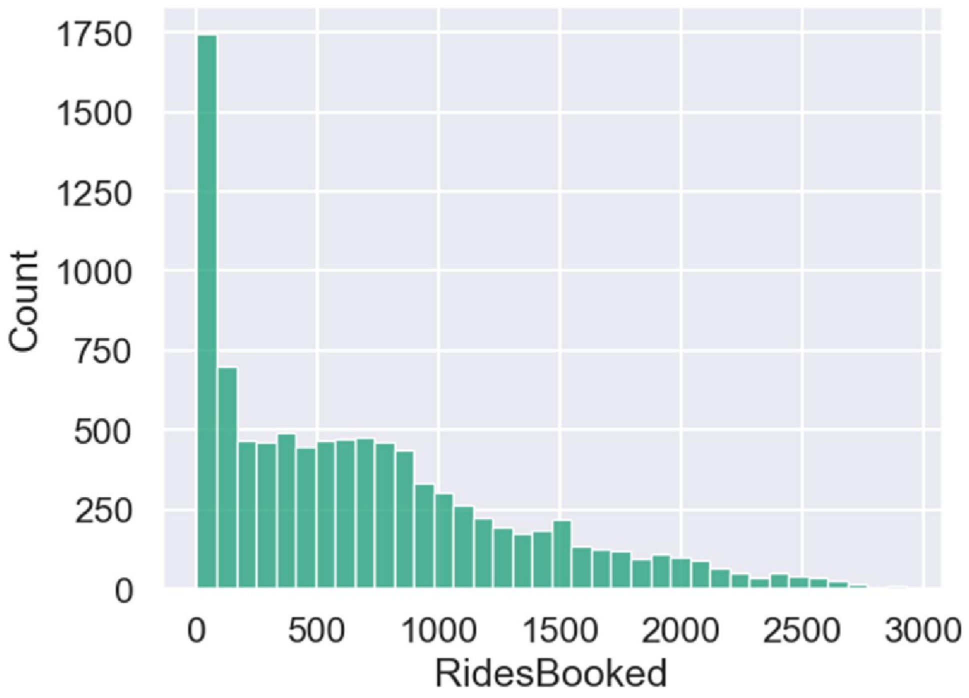
Anna Kwiatkowska 20222216

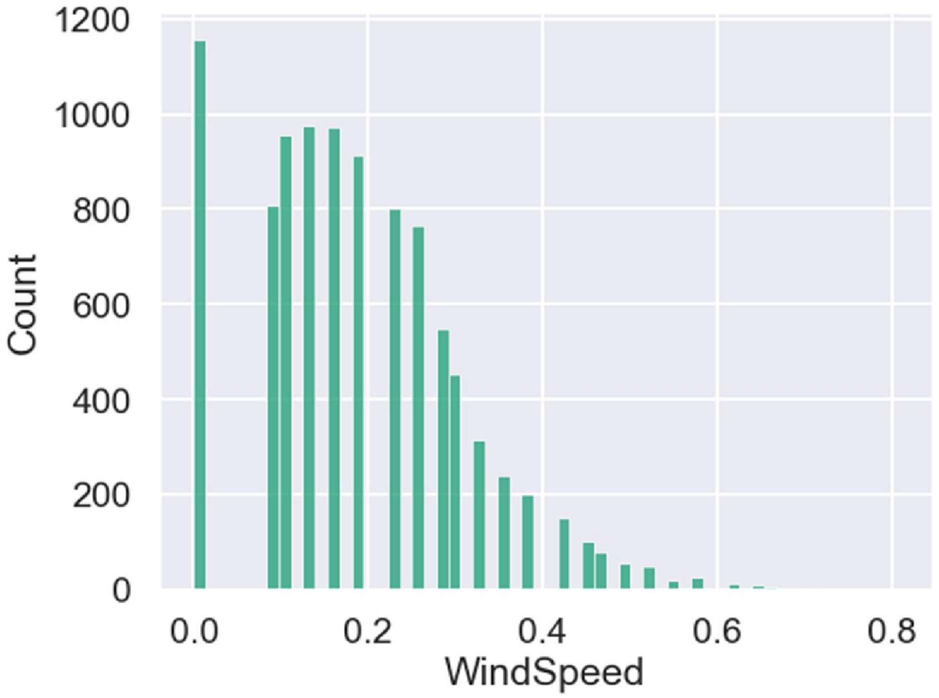
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Data Exploration and Understanding

Important aspects that we found:

- skewness in RidesBooked variable
- ‘blank spaces’ in histograms of Windspeed and Humidity

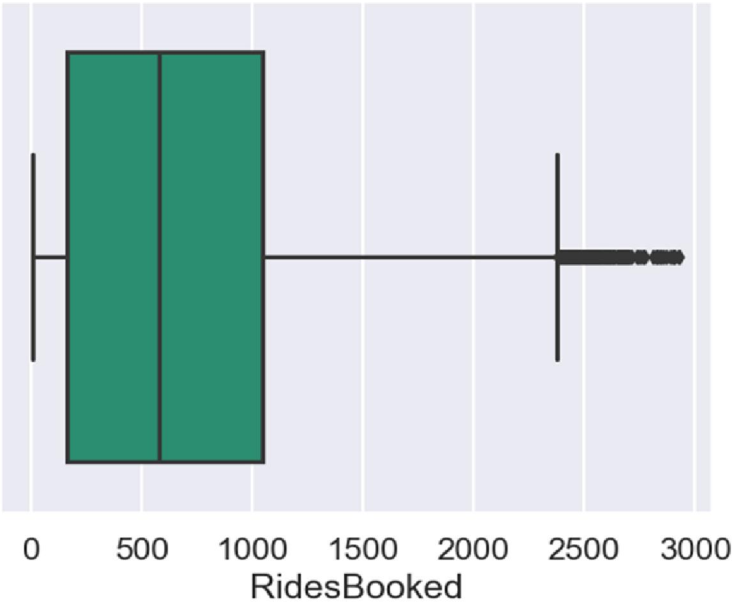


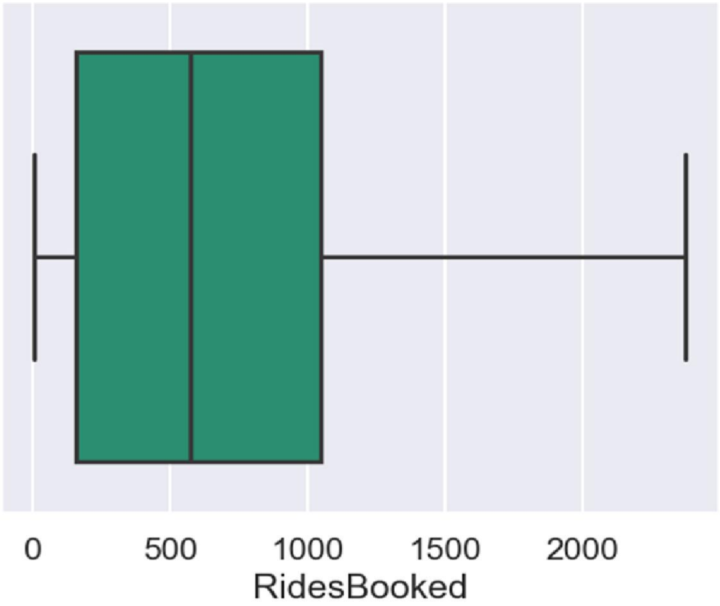


Data Preprocessing

The key processes that were done:

- converting Month and DayofWeek variables into numeric
- fixing outliers - replacing them at the whiskers of the boxplots
- creating a new feature - Date
- grouping the dataset by by Date, DayofWeek, HourofDay and Month

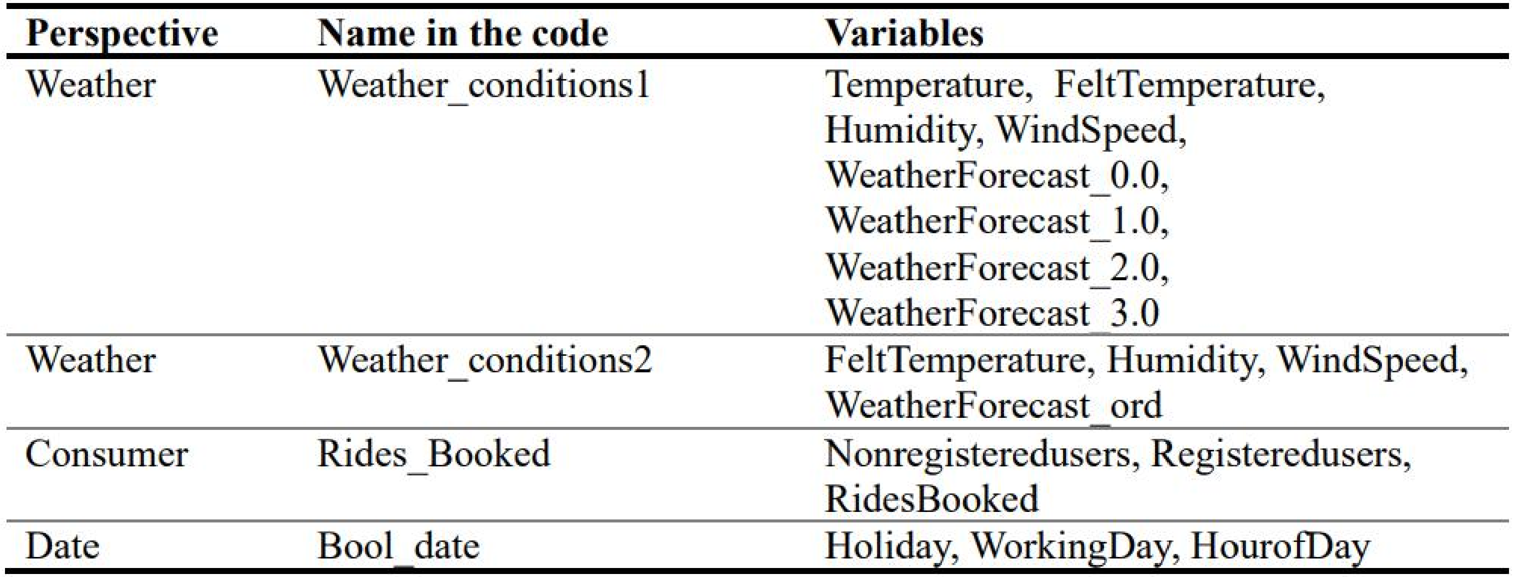




Scaling and feature selection

For scaling and feature selection we performed:

- MinMax Scaling
- Spearman correlation combined with the pairplots
- created 3 different perspectives



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

Models

For modelling we utilized:

- K-means
- K-modes

Additionally, we used three different approaches to define number of clusters “k”:

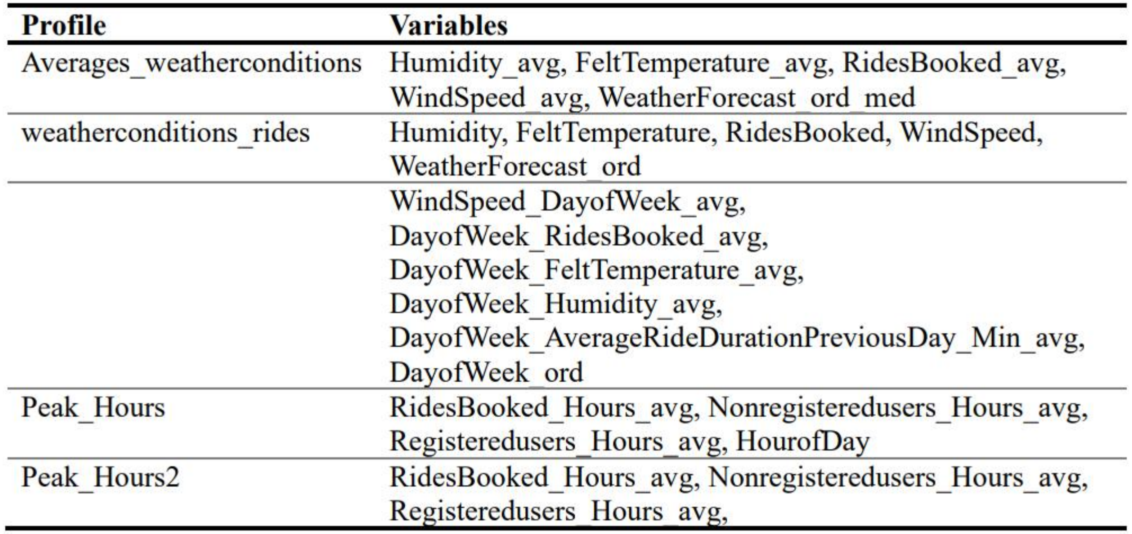
- „elbow” method
- dendograms
- silhouette score

At the end we visualized the results using the pararel plots and histograms.

Model assessment

To improve our models, we:

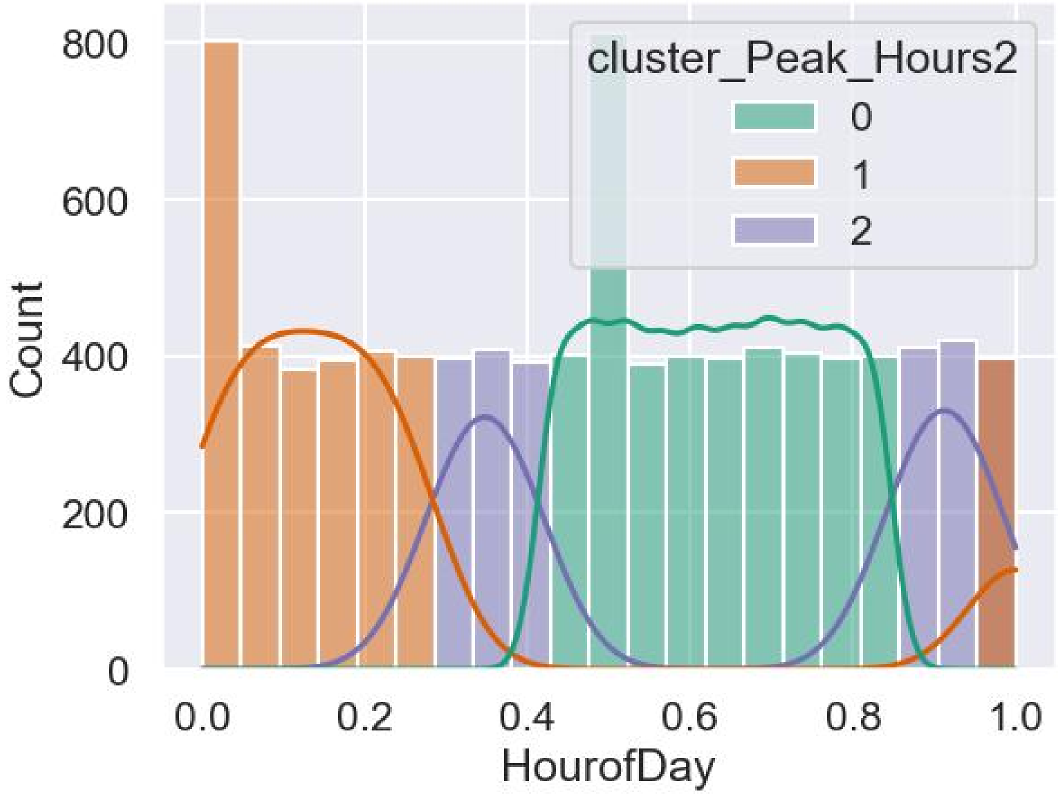
- explored the possibilities to clustering with both the categorical and numerical features
- changed the number of clusters “k”
- created 4 profiles for better understanding of the data
- deleted the Temperature variable
- merged the perspectives



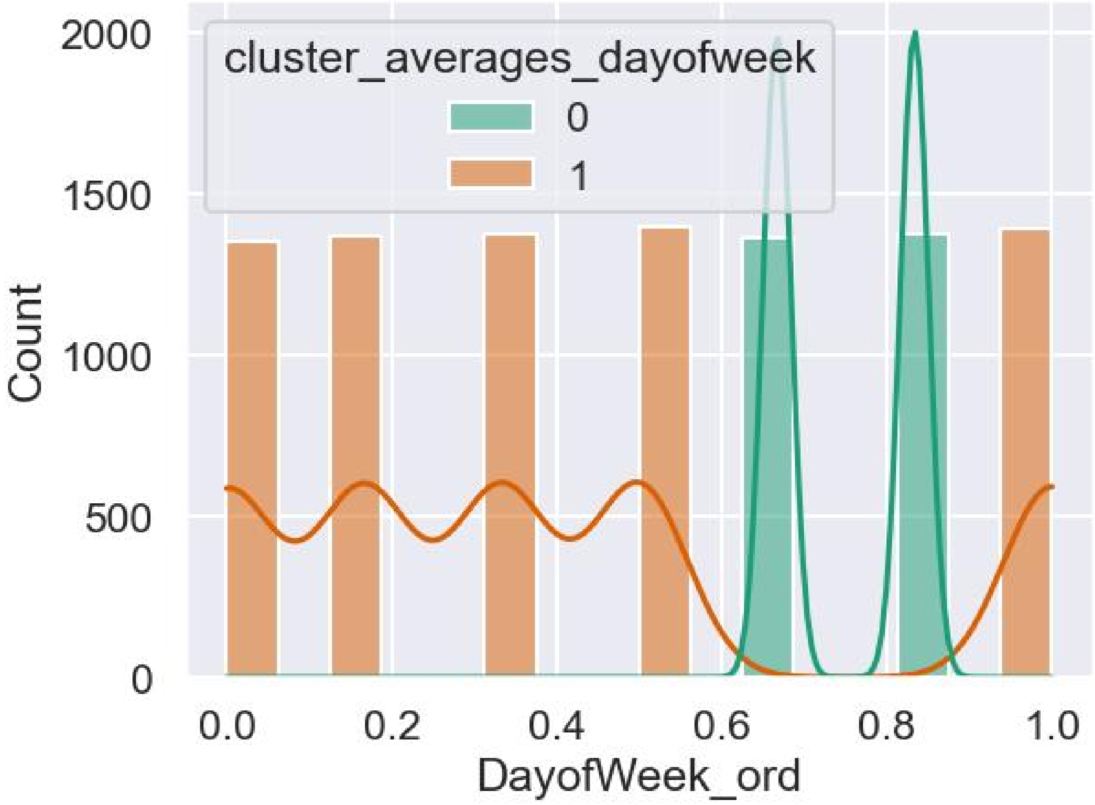
Profile	Variables
Averages_weatherconditions	Humidity_avg, FeltTemperature_avg, RidesBooked_avg, WindSpeed_avg, WeatherForecast_ord_med
weatherconditions_rides	Humidity, FeltTemperature, RidesBooked, WindSpeed, WeatherForecast_ord
	WindSpeed_DayofWeek_avg, DayofWeek_RidesBooked_avg, DayofWeek_FeltTemperature_avg, 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,

Profiles

HourofDay histogram for profile " Peak_hours2"

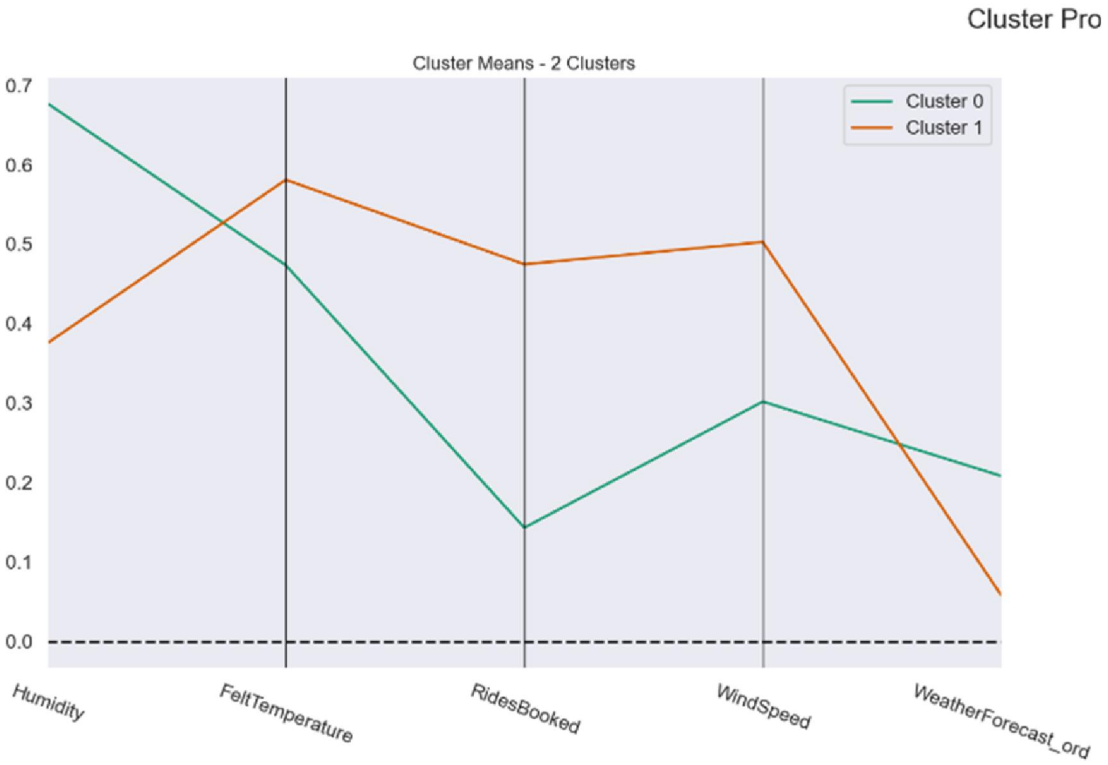


DayofWeek histogram for profile "averages_dayofweek”

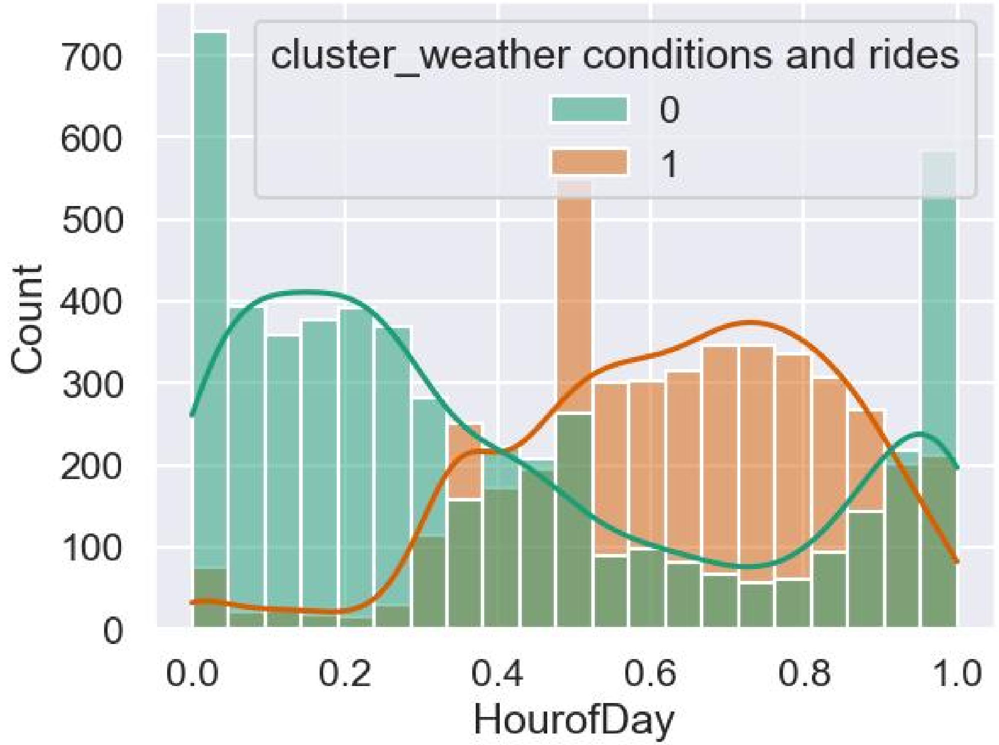


Profiles (2)

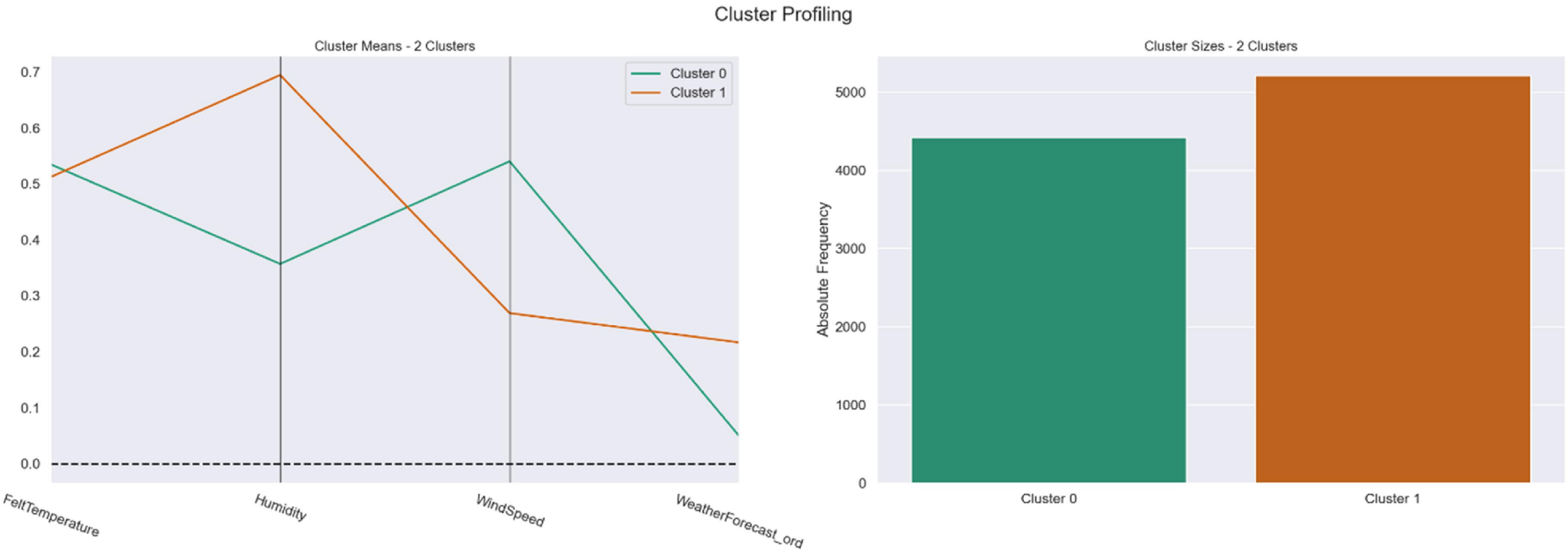
Plot for profile "weatherconditions_rides"



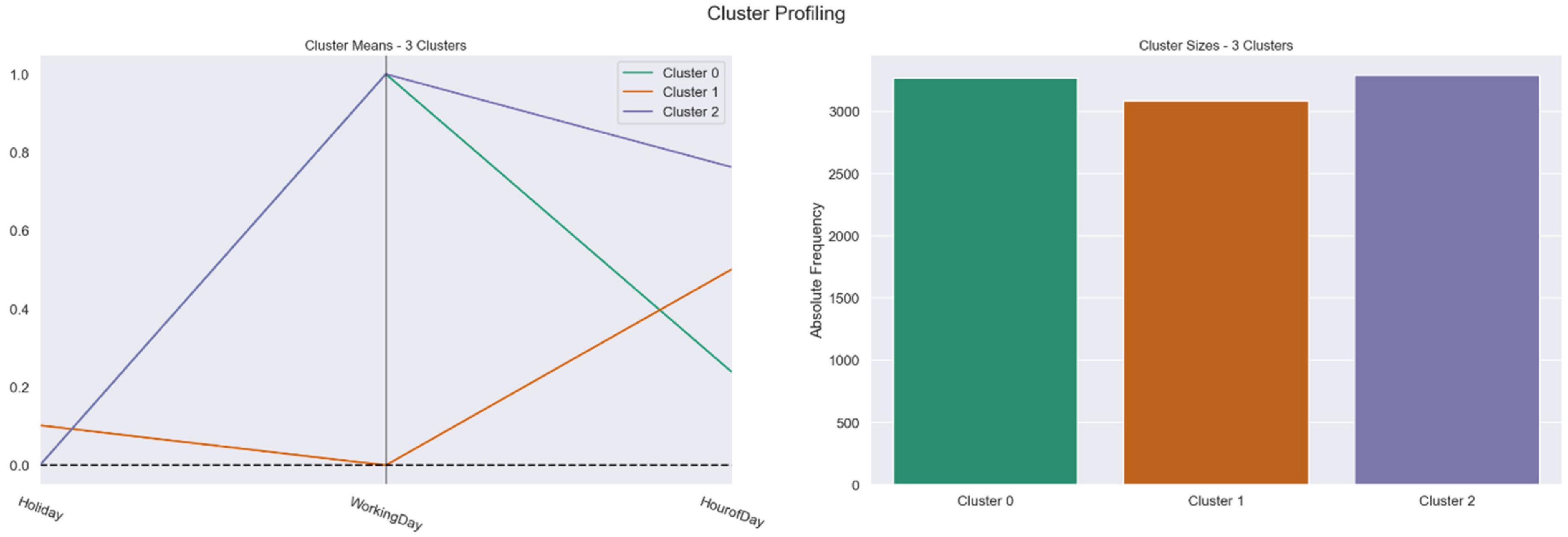
“HourofDay” histogram for profile "weatherconditions_rides"



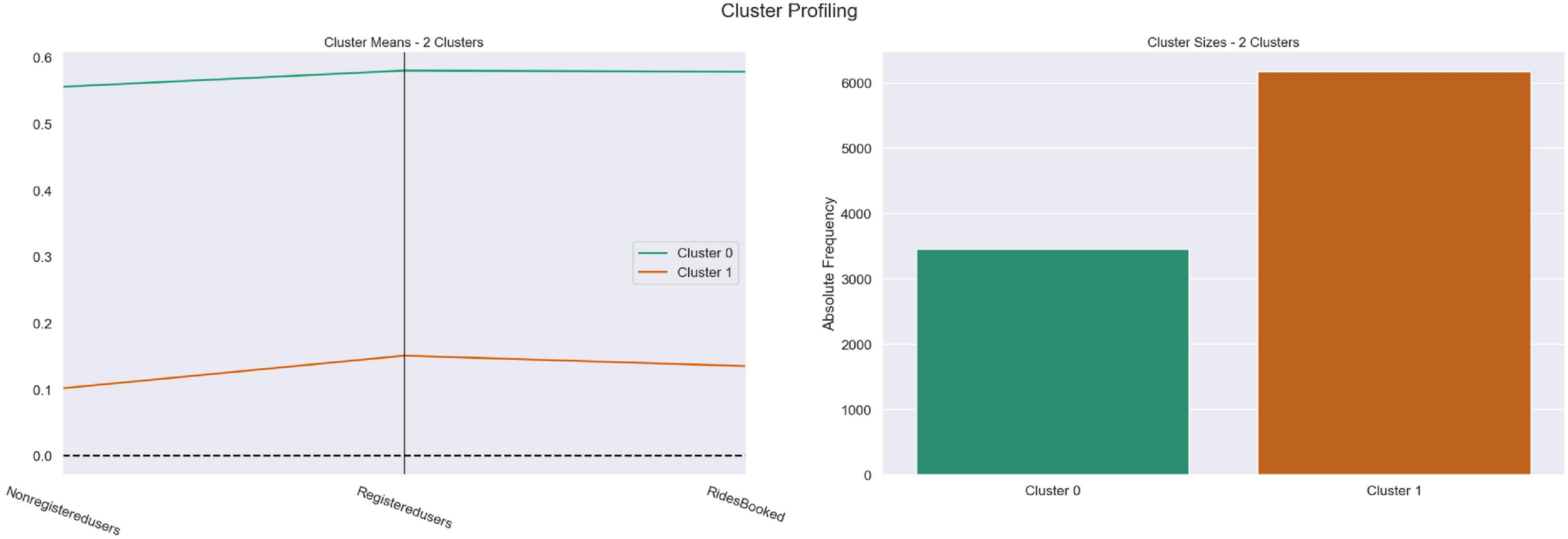
Perspective Weather_condition



Perspective Bool_date

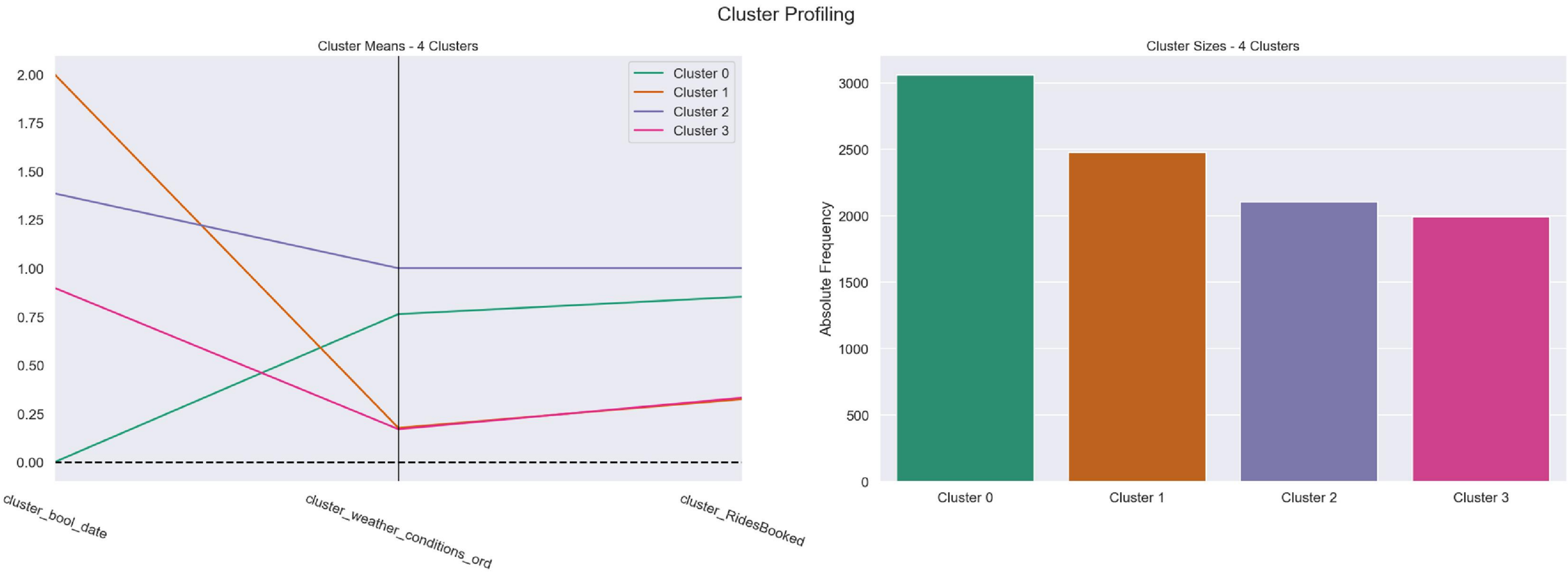


Perspective Rides_booked



Clusters

- From 3 perspectives we obtained 12 “semi” final clusters, that later were merged.
- The final number of clusters is 4.



Results

The main results from our work are:

- 4 different associacion rules, that could explain the customer behavior
- 3 variables, that influence the customer behavior the most

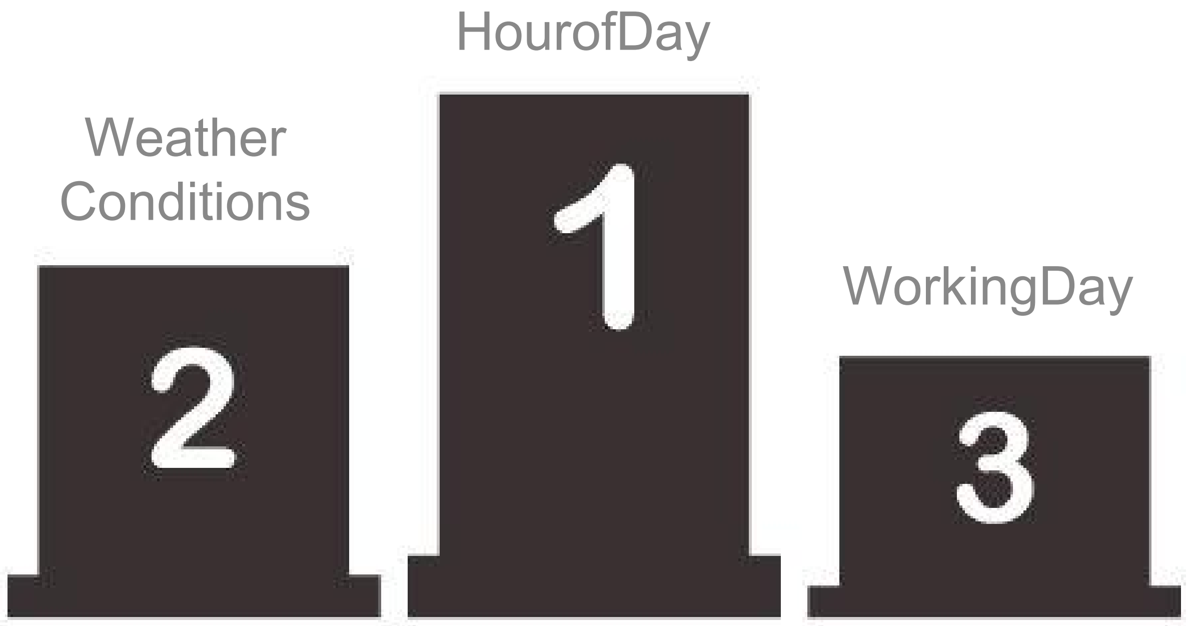
In the mornings of working days with poor weather, there is a low number of rides ordered.

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.

Results



HourofDay

Weather Conditions

WorkingDay

Thank you!

Morada: Campus de Campolide, 1070-312 Lisboa, Portugal
Tel: +351 213 828 610 | Fax: +351 213 828 611

Acreditações e Certificações da NOVA IMS

Cofinanciado por