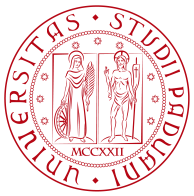

University of Padova

Department of Mathematics

Master Thesis in Data Science

RideMovi Data Analysis: Users Profiling

Supervisor
Francesco Silvestri
University of Padova

Master Candidate
Yelnur Shauketbek

Academic Year
2023-2024

To my family and other loved ones

Abstract

The 21st century brings an era of Big Data, where information belongs to every aspect of our lives. It presents an unprecedented opportunity for companies to leverage their data to enhance business performance. Various methods and techniques are available, tailored to a company's needs, expectations, and goals. One such innovative method is customer profiling, a robust data analysis approach that can significantly enhance the quality and quantity of services and goods offered by companies.

In this Thesis, I explore the application of customer profiling using the RideMovi bike-sharing service dataset. The dataset is thoroughly analyzed and focused on segmenting users into distinct profiles. These profiles are designed based on the identification through selection of characteristic Points of Interest (PoIs). Utilizing these PoIs, I measure the distances between each point and the starting or ending locations of individual rides.

Furthermore, I conduct a comparative analysis across different user profiles, examining metrics such as ride frequency, distance traveled, and duration. These metrics are evaluated on both a monthly and weekly basis. Additionally, an investigation is undertaken to uncover potential correlations between obtained results and prevailing weather conditions.

Through this study, I aim to shed light on the effectiveness of customer profiling as a strategic tool for businesses, offering insights into how they can optimize their services based on user behaviors and preferences. The RideMovi dataset serves as a valuable case study, illustrating the practical applications and benefits of this approach in the context of a bike-sharing service.

Contents

Listing of figures

Listing of tables

Listing of acronyms

PoI . . . . . Point of Interest

GMM . . . . . Gaussian Mixture Model

DBSCAN . . . . . Density-Based Spatial Clustering of Applications with Noise

SOM . . . . . Self-Organizing Map

1 Introduction

RideMovi is a popular bike-sharing service in multiple countries across Europe. RideMovi is a popular service because of many advantages, such as an easy and intuitively understandable smartphone app, a variety of subscriptions, a relatively low price per ride, and the possibility to choose an electric bike (e-bike) or a common bike. Particularly, RideMovi is highly popular among students and Padua citizens because of its easy and fast ability to reach anywhere inside Padua. Any user in Padua can find and ride a bike within several minutes. Another huge advantage of using bikes is making zero pollution and also decreasing the number of traffic jams. People usually prefer to ride a small and compact bike through narrow streets in Padova rather than use a car or bus, especially if the price of bus tickets is increasing. Therefore, if the riding distances are not large and health allows riding a bike, it's better for the environment and for people's healthiness to ride bikes.

Also, one of the most important reasons why people start using RideMovi and other bike services is COVID-19, specifically post COVID-19[1]. It was safer to ride a bike rather than use public transport.

In this thesis, I'm analyzing the RideMovi service dataset that was collected from Padua City in the period from July 2022 to November 2023, totaling 17 months. So, the dataset is fresh and contains currently active users. I'm going to analyze the data, profile users, and compare different profiles. I looked for answers to the following question: How should I efficiently and accurately profile users?

I expected bad weather conditions would reduce the number of rides, but analytical results

refuted the following statement. However, general data analysis shows some interesting trends and actions, such as ride numbers decreasing during winter and summer holidays, and increasing with the start of a new study year. Also, how ride number changes during the week and day.

To profile users, I created and used 7D vectors for each user which includes all selected Points of Interest (PoIs) categories. 7D vectors were obtained using the geodesic distance calculation between Points of Interest (PoIs) collected by hand and all given starting and ending points in the dataset using the Haversine formula. Later these 7D vectors fitted to the clustering algorithms, that cluster users into desired profiles.

I expected to find that the majority class of users are students, as Padua is one of the largest student cities in Italy and this is confirmed. I tried to explain and find differences between different profiles, tried to capture patterns and behaviours of each profile. Additionally, compared these profiles, which gave me quite interesting results.

The thesis is organized as follows:

- Chapter 2 contains information about the dataset all preprocessing done to the dataset, and general analysis;
- Chapter 3 describes analytical results, methods and the decision-making procedure used in this thesis;
- Chapter 4 focuses on the results obtained after user profiling;
- And the last Chapter 5 presents conclusions and possible future research that could complement this thesis.

2

Dataset

As mentioned in Chapter 2, this chapter is dedicated to the Dataset and the preprocessing made to it.

2.1 Dataset

In this Thesis, I'm using the RideMovi bike service rides dataset in the city of Padua. This dataset, provided by the Municipality of the city, contains fully anonymized information. I have collected bike ride information from July 2022 to November 2023, totaling 17 months, which is quite sufficient for analysis. The dataset consists of around 1.3 million rows and 28 columns, but most of the columns are not useful for analysis due to the incompleteness of the information they contain. In total, I have selected 11 columns containing user ID, ride starting time, ride ending time, starting and ending longitudes, starting and ending latitudes, riding time, riding distance, vehicle type, and pass group. Now let's take a closer look at these columns. The User ID needs to be included to profile users by their IDs. Ride starting and ending times contain the date and time of the ride, which will be helpful in future decisions and analysis. Also, for each record in the Dataset, I have decided to calculate the starting and ending time difference and the geodesic distance between starting and ending points using the Haversine formula 2.1, where R is the radius of the Earth and θ is the central angle.

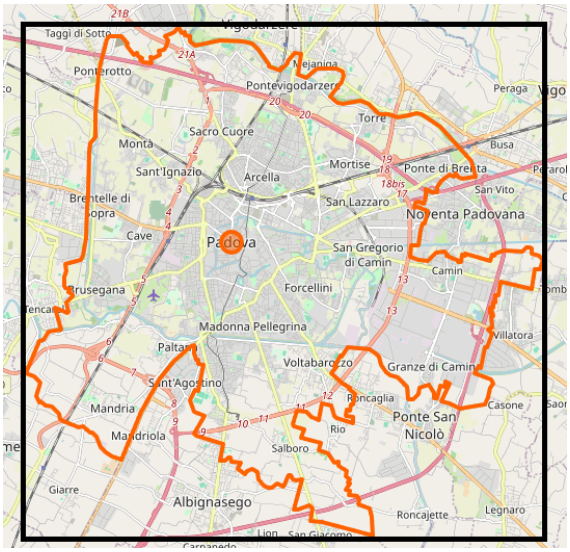

Figure 2.1: Padua city border in orange and square boundary in black

$$d = 2R \arcsin(\sqrt{\sin^2(\frac{\theta}{2})}) \tag{2.1}$$

I need additional calculated time differences and distances in this Dataset because mostly given ride durations and distances are incorrect. To avoid future problems with analyzing, I added these two columns to the Dataset.

2.2 Preprocessing

The Dataset I'm using in this Thesis is a real-world dataset; therefore, it could have typos, mistakes, or unrelated records. The entire Dataset has records in the 11 chosen columns, as was mentioned in Section 2.1. There is no need to remove empty rows. However, the Dataset has multiple issues with starting and ending coordinates; some records show that the user started or ended the ride trip in another city or even country. To avoid this problem, I'm going to create a square on the map that fully includes the city of Padua and nothing else. All records lying outside this boundary will be removed. The city border and the black square that I'm using for checking the starting and ending points of the records are depicted in Figure 2.1. The square boundary is an easy and elegant way to prevent complex calculations of the real city borders. The picture with the city border was downloaded from OpenStreetMap (Map data © OpenStreetMap contributors, 2015)[2].

Another problem encountered is unusually long ride durations. Users wouldn't ride for 4 hours. There are three possible reasons why this is happening:

1. Some users simply forget to end the ride and lock the bike;
2. Users ride bikes for 4 hours, but the riding distance should be long;
3. There is a bug in the recording system, probably short-term issues with servers, and this amount of time is just a filler for missing record indicators.

The third option seems to be the most probable one because additional analysis of such records shows that one user rode only 32 meters in 4 hours. Most other users also rode short distances for the same 4 hours.

One of the crucial indicators is the distance of the ride, but the ride distance column given in the dataset contains too many errors, such as zero distances in places where the real distance is not zero. One example is the record where the real distance is around 1.8 km, but in the dataset, we have zero. That's why I decided to create a new column with the correct distances and replace the initial ride distances column. After calculating the correct distances for all records, I don't have incorrect record distances.

Why is the distance a crucial indicator? Let's imagine a situation where a user sits on a bike and rides 50, 100, or 150 meters and feels something wrong with the bike. Of course, the user will stop and try to find another working bike. As a result, the ending point is not the same as the desired destination point, and the profiling system suffers from such unpredictable and unwanted experiences. Therefore, the best way to handle this problem is to choose records with riding distances of more than 200 meters. I believe that 200 meters is fully enough to understand and feel if there are any problems with the bike, whereas 50 or 100 meters could be insufficient.

2.3 General Dataset Analysis

After preprocessing I can do some general analysis on the obtained filtered dataset. Totally in the new dataset, I have a bit more than 1.2 million rows, which includes 510 days. By simple calculations, the average number of rides per day is almost 2400. Let's look at Figure 2.2, in this Figure I made a histogram of rides that took place in 510 days. Now I want to share the changes that happened with the number of rides during the research period. Changes through time are shown in Figure 2.3. The graph is quite complex with a specific pattern. From July

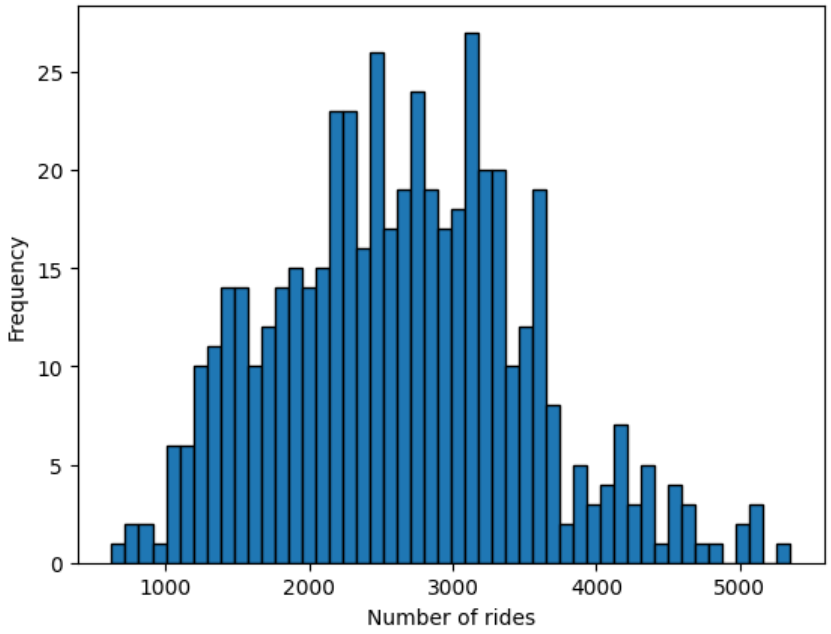

Figure 2.2: Rides histogram

to August for both 2022 and 2023 years I observe a sharp jump down, possibly because of the main summer holidays for students. The main peculiarity can be found in the period from September to November when both years have the most rides. It's probably because most of the students coming or returning to the city to prepare and start their new Academic year. From the winter starts the graph line is dropping down, and starts growing again after February, it's obvious that the reason for these moves depends on the weather conditions.

For better understanding, I made a new graph Figure 2.4 that is responsible for the changes in the number of rides for each month. So from this plot, we can truly see that everything I wrote before is true. New students or returning students are responsible for more rides, bad weather conditions bring fewer rides, better weather conditions from February result in more rides, and starting from July line is dropping down because most of the students leave the city. I want to highlight the following, starting from November 2023 RideMovi has changed its terms and conditions. Before this month Premium subscription allows one to ride a bike for free every 60 minutes, but after changing its policy to a Premium subscription users are paying 0.5 euro per every 20 minutes.

In Table 2.1 it's seen that most users prefer to ride electronic bikes rather than standard bikes. The reason for that probably lies in the fact that e-bike requires less effort and are much faster

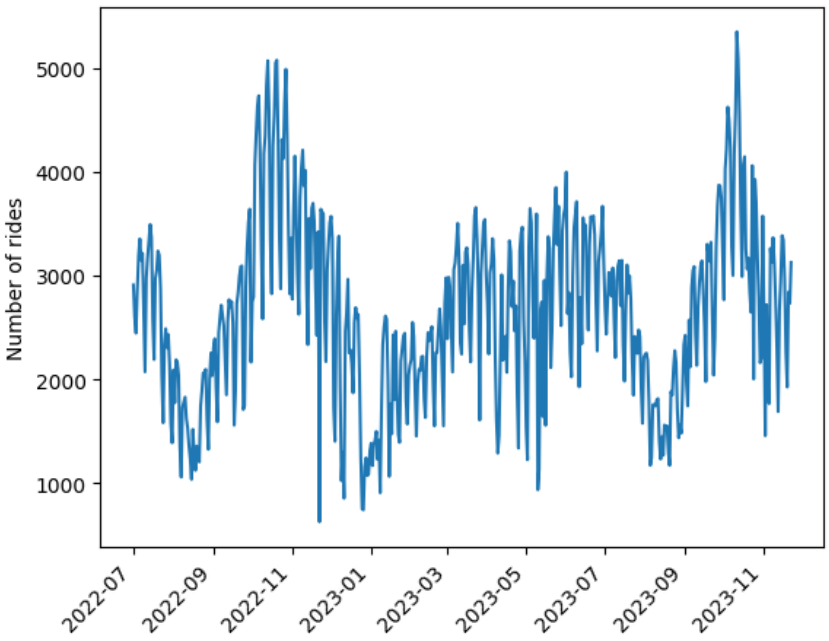

Figure 2.3: Changes in the number of rides per each day

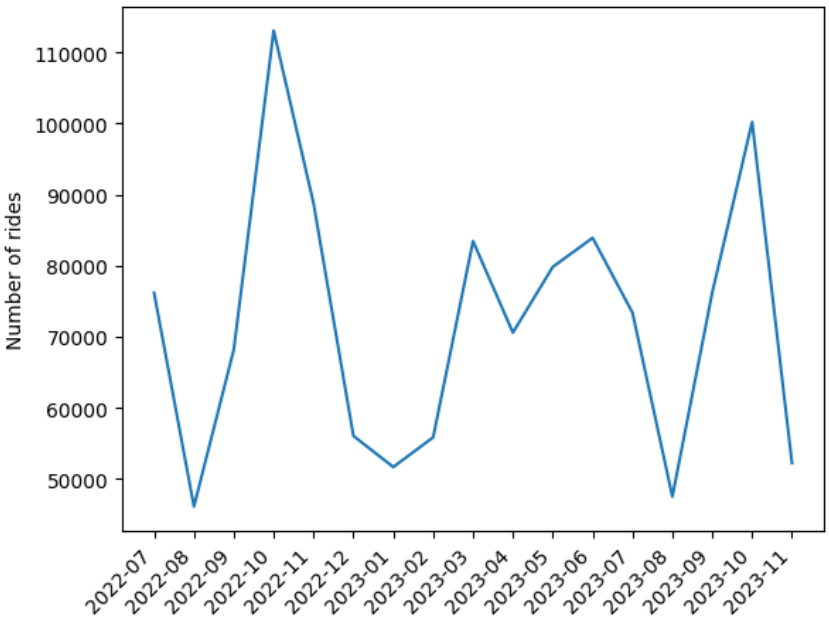

Figure 2.4: Changes in the number of rides for each month

Bike Type	Distibution
e-bike	55.74%
bike	44.26%

Table 2.1: E-bike and bike distribution

Pass Group	Distibution
Paying group	36.61%
Premium Pass	30.13%
Month Card & Daily Pass	18.29%
Times Pass	14.78%
Partner	0.12%
Coupon	0.07%

Table 2.2: Pass groups distribution

than traditional bikes.

What about pass group distribution? In the dataset, I found 6 groups: paying for a ride, premium pass, month card, daily pass, times pass, partner, and coupon. The distribution is shown in Table 2.2, where we can see that most users prefer to pay for their rides, it's economically beneficial when users do not ride bicycles regularly. Premium pass allows users to pay less for their rides, current rates are the following: bike trip fee: € 0.50 / 20 min and e-bike trip fee: € 1.50 / 15 min, while the price of the Premium Pass is 14.99 euro per month. Without Premium Pass rates are much higher. Month card (30, 90, 365 days passes) and Daily pass subscriptions apply only for standard bikes and allow users to use a bike for free every 60 minutes. Times passes including 25, 45, and 90-minute passes, and valid only for e-bikes, which allows free riding during the pass validity period. Partner and coupon passes are not available for the common users, which is why the percentage of these kinds of passes is too small.

In Figure 2.5 I collected average distances for each day and built the histogram. This histogram shows us that usually, the average distance per day is equal to 1.5 km or in other words 1.5 km average distance per ride in the dataset.

In Figure 2.6 we can observe daily average distance changes that happened during the observation period. From March of 2023, we have a linearly growing graph, probably because of the weather conditions suitable for larger distances.

For a better understanding of what happening with the average distances, I prepared an additional, more obvious graph. If we look at Figure 2.7 we can notice that now it's more clearly seen, that from February 2022 average distance per ride is increasing.

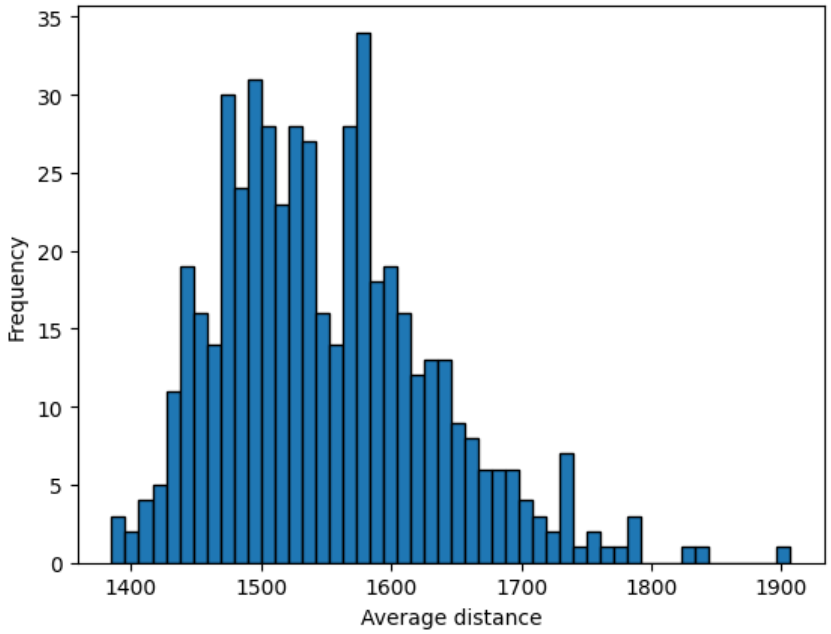

Figure 2.5: Average distances histogram

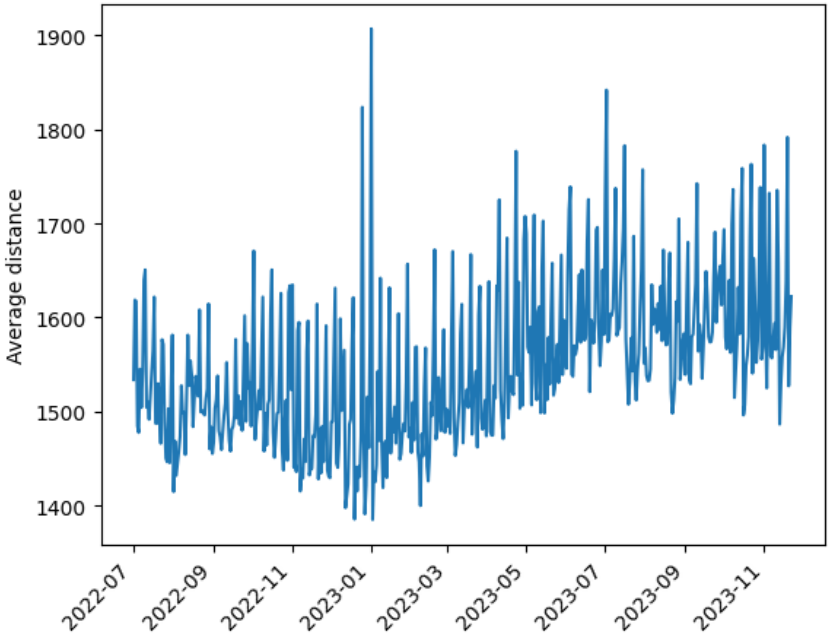

Figure 2.6: Changes in daily average distance through the time

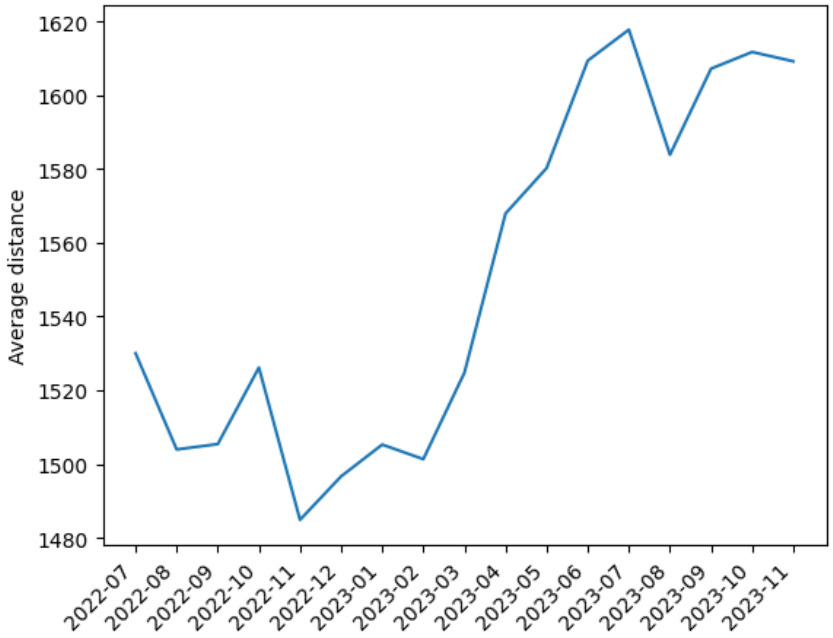

Figure 2.7: Changes in monthly average distance through the time

Let's try to understand how the number of rides changes during the week and also during the day. In Figure 2.8 we can see that most rides occur on working days from Monday to Friday and sharply drop down on the weekends. Figure 2.9 demonstrates hourly changes in the number of rides, in other words, rides distribution during the whole day. During the night hours from 0 to 5 o'clock almost no rides, but starting from 5 to 8 the line rapidly went up, and from 15 to 18 o'clock, I observed an increasing number of rides. After 18 o'clock the line is linearly going down. Therefore I can conclude that most rides happen during the morning when users go to their works or study, and during the afternoon when users return from their works or study.

To better understand sudden droppings in Figure 2.3 I have collected weather history and prepared a dataset using Weather API [3]. Weather dataset including date, time, rain (mm), temperature (°C), and wind speed (m/s). I want to check the correlation between the number of rides and rain and temperature, the same for the average ride distances. In other words, I want to check if the number of rides or average ride distance depends on the weather conditions. I have obtained the following results available in Table 2.3. I have the following results:

1. Number of rides mostly correlated with rain. Correlation is negative which means is there more rain, then fewer rides. However, the correlation is not significant. Wind is less correlated because usually rain accompanied by wind.

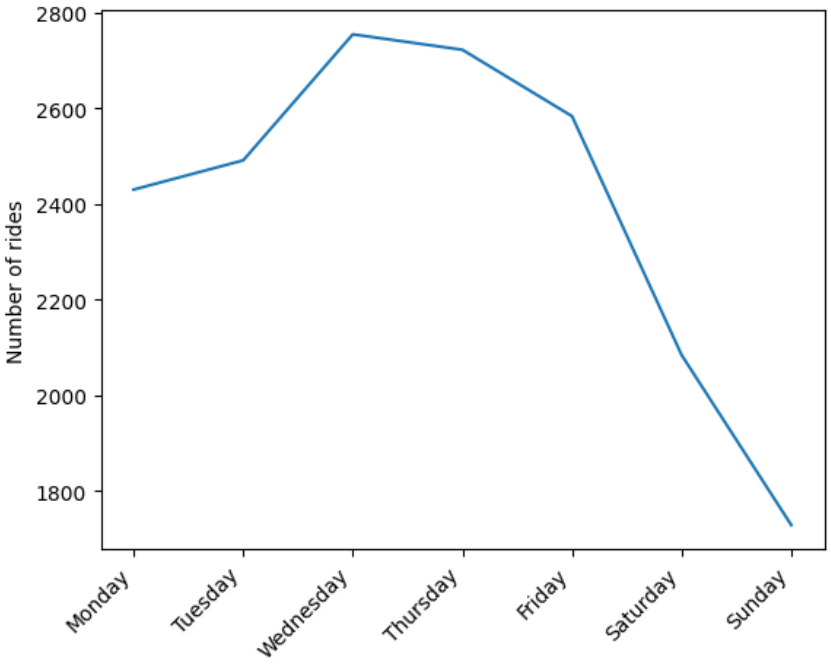

Figure 2.8: Changes in the number of rides during the week

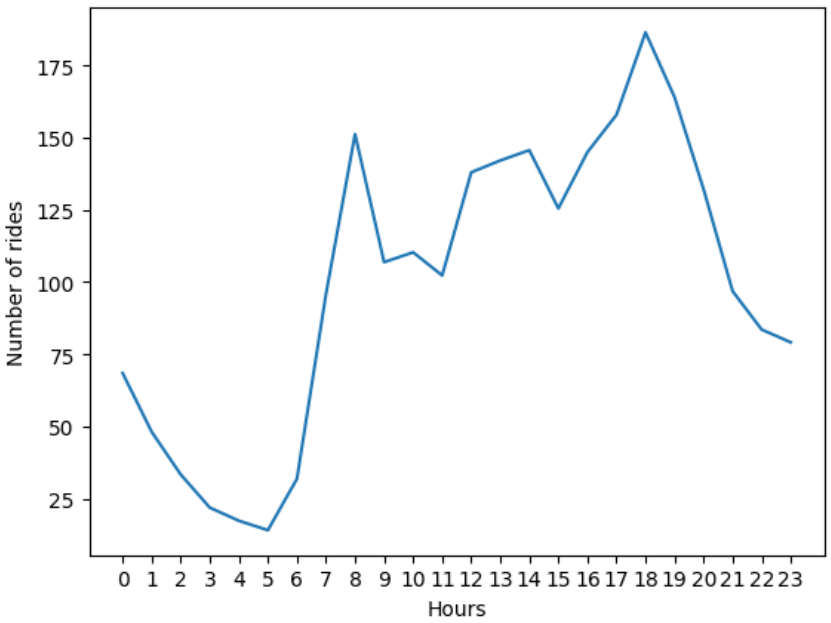

Figure 2.9: Changes in the number of rides during the day

Weather condition	Correlation between number of rides	Correlation between average ride distance
Rain	-0.2338	0.0203
Temperature	0.0521	0.2674
Wind	-0.1149	-0.0312

Table 2.3: Weather condition correlations

Strike Date	Number of rides
23-07-2022	1834
29-10-2022	3099
02-12-2022	3086
18-09-2023	2942
20-10-2023	2842
11-11-2023	2046

Table 2.4: Rides number on strike dates

2. Average ride distance mostly correlated with temperature. This correlation is positive, which indicates that higher temperatures are larger ride distances. As before the correlation is not significant. Other weather conditions almost do not correlate with average distance.

   I couldn't say that bike rides highly depend on the weather conditions, of course, we have some dependency but not significant at all. Only once when the rain was the heaviest one during all 510 days, number of rides was record low.

But what about public transport strikes? I found and took all official BusItalia strikes from 01-07-2022 to 22-11-2023 that have been published in the BusItalia NEWS section. Padua city public transport strikes: 23-07-2022, 29-10-2022, 02-12-2022, 18-09-2023, 20-10-2023, 11-11-2023. After finding the following dates in the dataset, I can share the results. Results are available in Table 2.4, I want to remind you that the average ride number is almost 2400. As a result, we see that on 4 of 6 strike dates ride number is higher than the average number for the whole period of observed time. I can conclude that public transport strikes lead to rides number increasing.

The general analysis of the RideMovi Dataset ends here, other analysis and profiling are available in the further chapters.

3 Analytical Results and Methods

In this Chapter, I try to explain the decision-making processes used for further analyzing the dataset and profiling users. User profiling is the main part of this Thesis.

3.1 Users flow

Before I start user profiling I want to share interesting observations on users riding flows inside the city. It's not a secret that users ride bikes from somewhere to their desired destination places in Padua. I'm going to try to cluster starting and ending points, analyze clusters, and get users flow.

The main idea is to collect all starting and ending points available in the dataset. After I have collected them, I can use the K-means clustering method. The method tries to give a class label to each point in the given clustering problem, in my case it tries to label all given starting or ending points within a given number of labels. Mathematically it's written in the Formula 3.1, where $S = \{S_1, S_2, ..., S_k\}$, $|S_i|$ is the size of S_i, $\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} x$ and the $|| \cdot ||$ is standard L^2 norm.

$$arg \min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||x - \mu_i||^2 = arg \min_{S} \sum_{i=1}^{k} |S_i| Var S_i \tag{3.1}$$

Initial cluster centers were selected fully randomly. After initializing the first k points, I can

use k-means or Lloyd's algorithm to label other points. The whole algorithm works in two steps:

1. Assignment step: Assign each point to the closest cluster or in other words to the cluster with the nearest mean. To find the closest cluster usually uses squared Euclidean distance. Mathematically formulated as:
   $S_i^{(t)} = \{x_p : ||x_p - m_i^{(t)}||^2 \leq |x_p - m_j^{(t)}||^2, \forall j \in \{1, ..., k\}\}$
2. Update step: recalculate and move the centers of each cluster. Mathematically formulated as:
   $m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$

The algorithm works until convergence or in other words, works until the assigning step won't change any point label. For my task, I decided to use 40 clusters, which is good enough to have a general concept of the flow. This number is sufficient to capture the diversity of patterns and behaviors without resulting in an overly granular segmentation that may be difficult to interpret. Also, using 40 clusters strikes a good balance between capturing detailed patterns and maintaining computational efficiency.

Before clustering, I want to plot all endpoints on a map. The ending points of each ride in the Dataset are clearly shown in Figure 3.1 that was plotted and downloaded from the Kepler.gl [4]. The same plot could be done for starting points, but it's unnecessary because it will be the same plot as for the endpoints. The reason for that is quite simple and intuitive: users end their rides somewhere and after that other users take the same bikes to ride somewhere else. Therefore, ending and starting points eventually lie in the same spots.

Clustering available in Figures 3.2 and 3.3. We can notice that clustering graphs are quite identical, which means that most of the starting and ending points lay close to each other.

Now I choose 20 clusters with the most points belonging to these clusters. By selecting the clusters with the most points, I ensure that the majority of the data is represented. This approach highlights the most significant patterns and behaviors within the dataset, making the analysis more focused and impactful. Furthermore, reducing the number of clusters from 40 to 20 helps simplify the analysis. This reduction makes it easier to interpret the results and draw meaningful conclusions without being overwhelmed by too many segments. I'm going to split the map into big areas, such as Stanga, Sacra Famiglia, and so on. All big districts, regions, and areas are shown in Figure 3.4 that was downloaded and available on Hoodmaps.com [5].

20 clusters from the start points clustering and 20 clusters from the endpoints clustering, then I need to obtain their centers. The center of each cluster is given as a set of latitude and

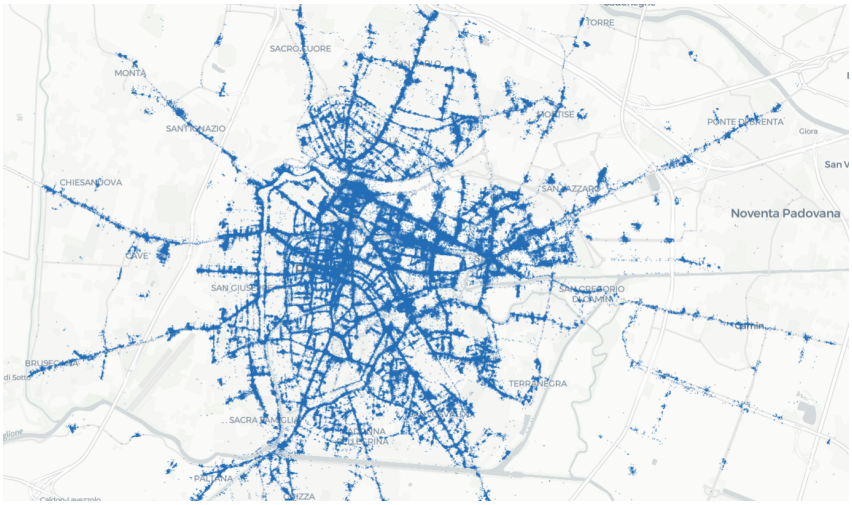

Figure 3.1: End points visualization

longitude. After collecting all centers, I can compare start points and end points centers. If the geodesic distance between centers is less than 100 meters, it could mean only that the centers lay too close to each other and belong to the same big area. After computing the distances between centers I got that 13 of 20 centers for both clustering are the same or located close (less than 100 meters). In Figures 3.5 and 3.6 we can see that the most common point to start or end the ride is the Train Station. It's probably because of the users traveling by train: students or workers living in other neighboring cities. Another bar for both of the figures is the sum of all rides not included in any other areas inside the city.

Figure 3.7 is a more intuitive version of Figure 3.5. But I can represent this diagram even better. The main idea is to connect some areas into one big area, for example, all University areas could be connected into one big University area, the same for the city center. And now from the better diagram represented in Figure 3.8, I can notice the top three areas (not including Others):

1. University area
2. City Center area
3. Train Station

I prepared the same diagrams for ending areas, they are available in Figures 3.9 and 3.10. And now by these figures, I can observe the same three main ending areas as were for the starting areas.

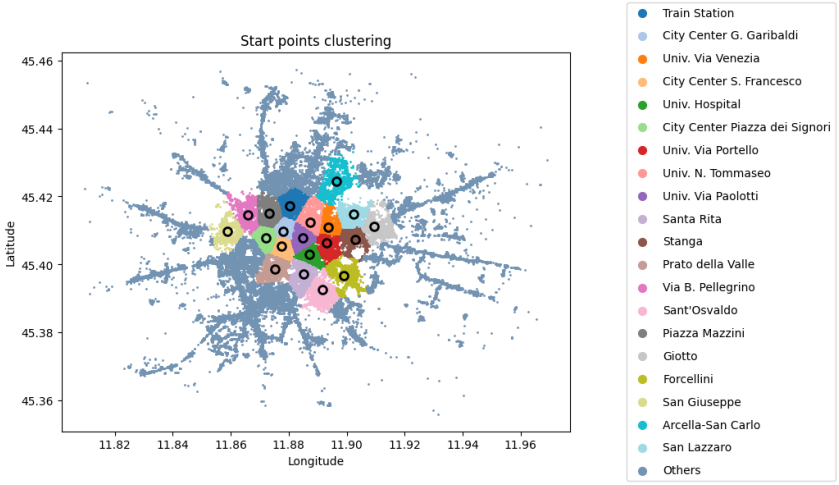

Figure 3.2: Start points clustering

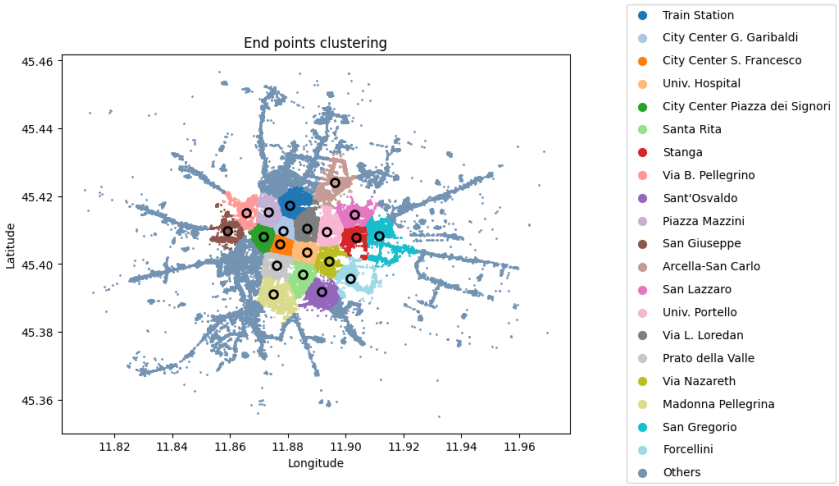

Figure 3.3: End points clustering

Figure 3.4: Padua areas

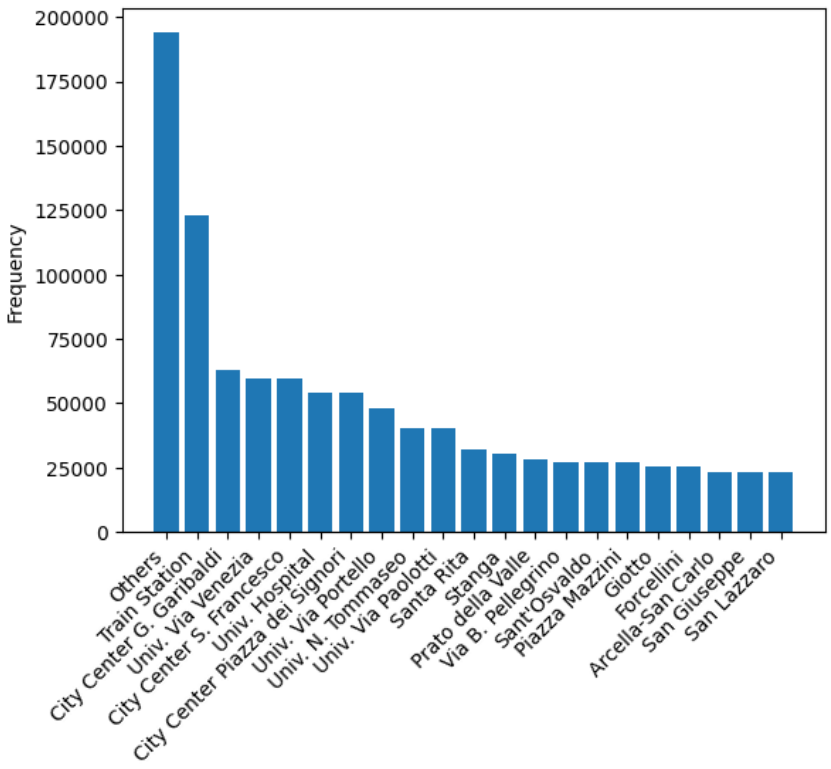

Figure 3.5: Mostly used starting areas

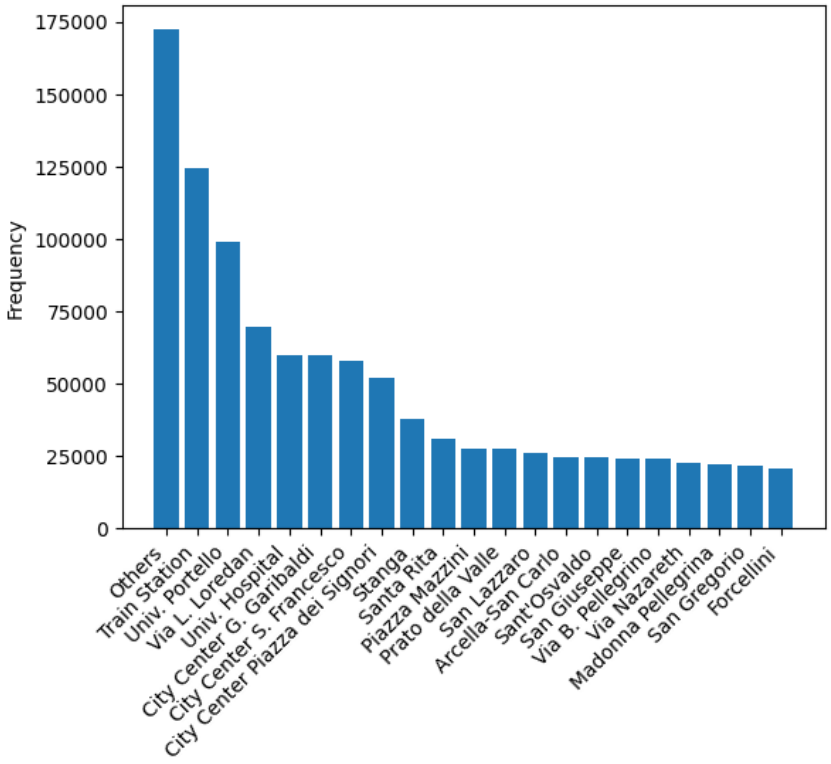

Figure 3.6: Mostly used ending areas

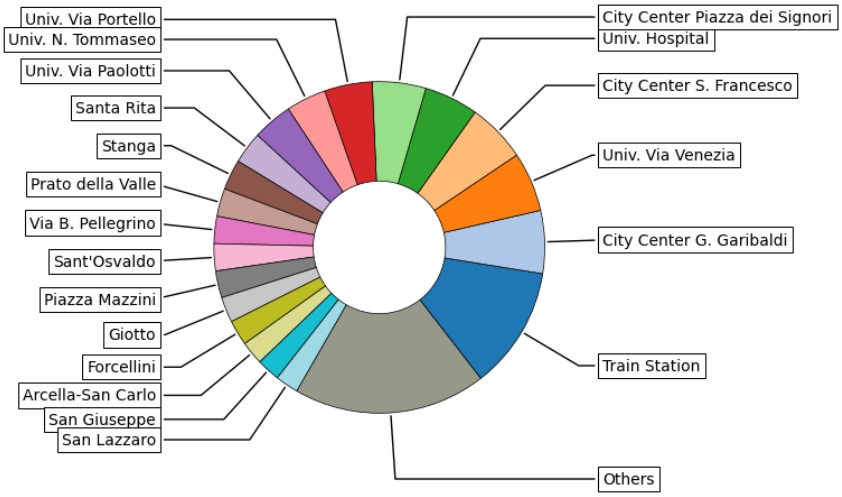

Figure 3.7: Starting areas diagram

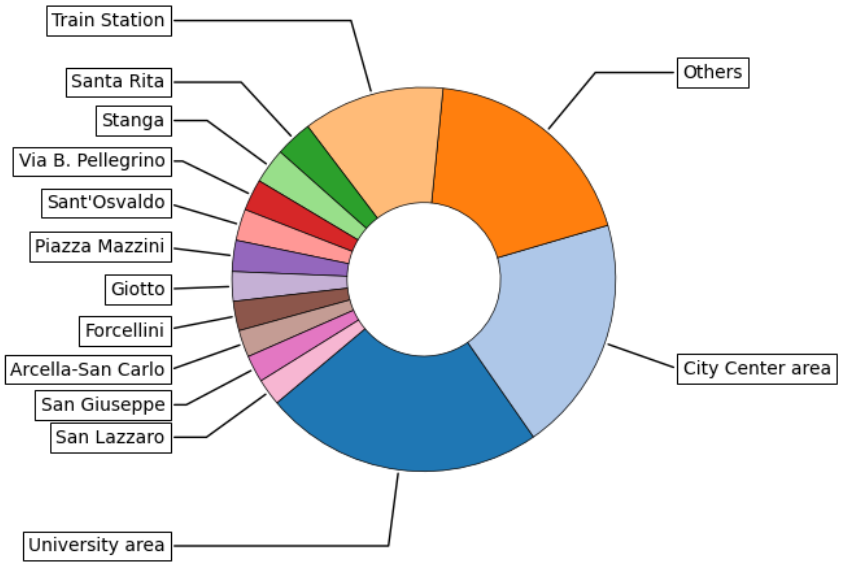

Figure 3.8: Joint starting areas diagram

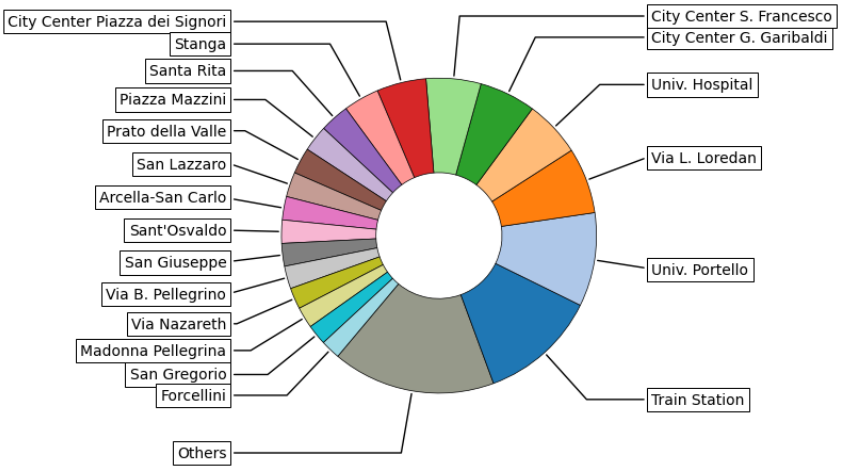

Figure 3.9: Ending areas diagram

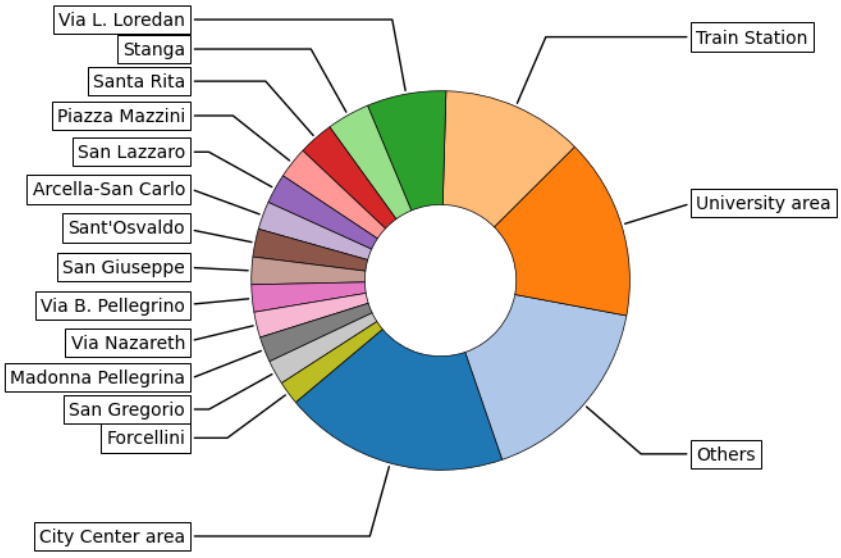

Figure 3.10: Joint ending areas diagram

3.2 Points of Interest

To make users profiling possible I should collect points of interest. It's quite obvious that people usually ride to their desired points of interest. Such PoIs for students are University buildings, medical workers are Hospitals and other Medical buildings, for bank workers are Banks, and so on. A reasonable question is why non-student users couldn't ride to some of the PoIs for students. Of course, they can, but non-student users won't ride to such places periodically. Therefore, the ratio of visited University buildings becomes too small. Firstly, I want to collect all important places for students, such as residences, canteens, study rooms, libraries, departments, and other University buildings. Totally I have collected 102 PoIs for students. All places were collected from the Google Maps (Google, 2024) [6]. Google Maps offers much more valid and actual information about places rather than OpenStreetMap, this is the main reason to use Google Maps. Other PoIs as banks, hospitals and other medical buildings, industrial zone buildings, tourist places, and train stations could be collected in the same way as before for University buildings. The total number of collected places is 235 PoIs for all profiles. All plotted PoIs using Kepler.gl [4] are shown in the Figure 3.11.

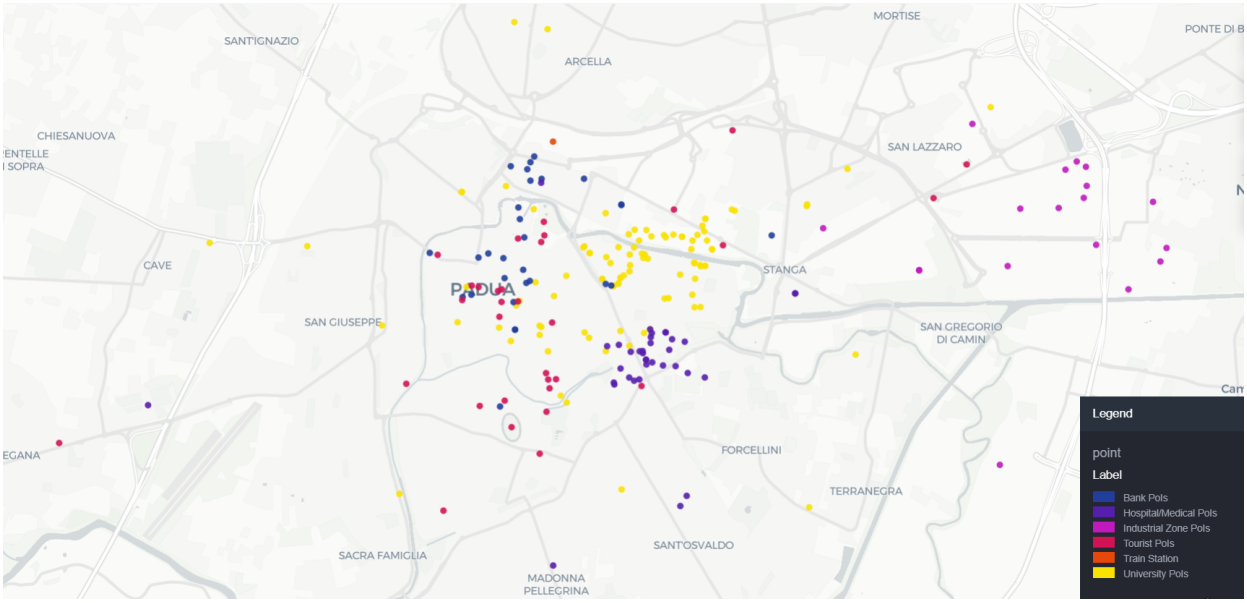

Figure 3.11: Selected Pols on a map

3.3 Users classification

Users classification plays a crucial role in profiling. I decided to make two classes of users:

1. Active user: user who has at least 20 rides during the whole observational period
2. All users: the whole dataset users

By the Figure 3.12 we can see that the number of rides histogram is really spread. Even I can mention at least one user with 1600 rides. If we look closer at the scaled histogram available in Figure 3.13 we can see that most of the users are occasional, mostly users have only one ride. That such users wouldn't clearly show their behavior and it's not possible to correctly profile them. Therefore, I'm choosing active users that have at least 20 rides. 20 rides is the average number of rides per user in the initial Dataset and also, it's fully enough to make a decision about such a user. While the all-users class contains all users that have at least one ride in the considered period. I will use these two classes in different comparisons.

After the classes definition, I have the following users in each class:

1. About 11 thousand active users
2. About 55 thousand all-users

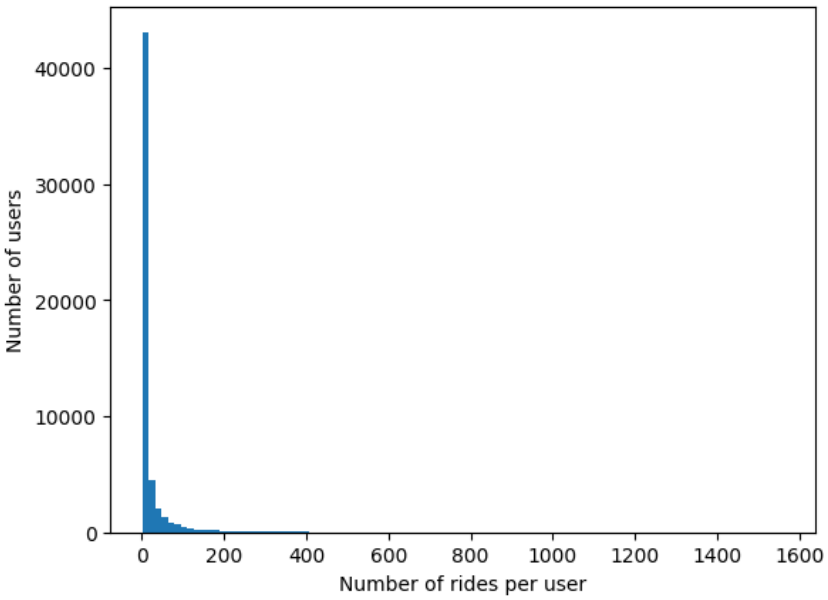

Figure 3.12: Number of rides per user histogram

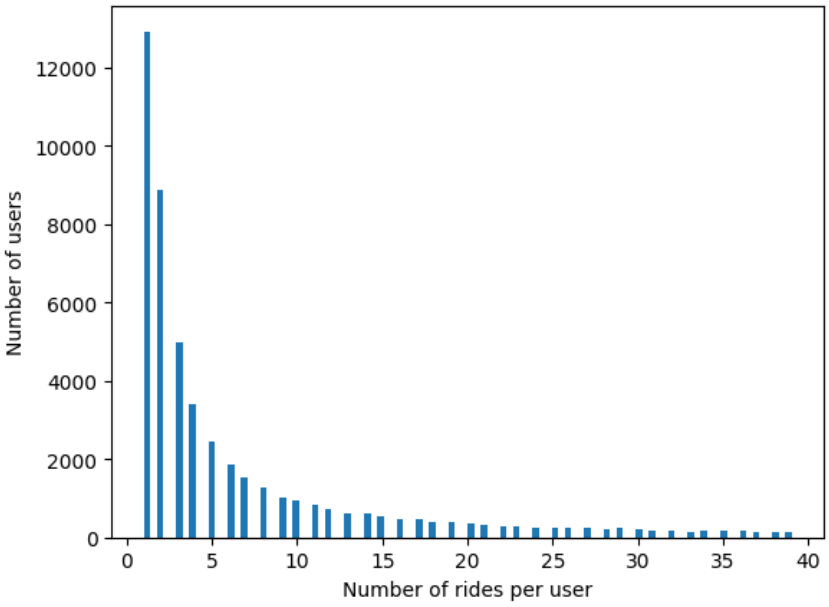

Figure 3.13: Scaled number of rides per user histogram

It's a huge drop, but what about the number of records? It dropped from 1.2 million to 1 million, therefore the majority of ride records remain.

3.4 Decision making procedure

Now I want to explain the main part of the profiling method. The main question here is how I could profile users into certain profile categories. To answer this question I will use multiple approaches and compare them. They are:

1. Simplest approach
2. K-means clustering approach
3. GMM (Gaussian mixture model) clustering approach

I have data related to all the starting and ending points of each ride. I can use it to be able to profile them. Also, I have collected PoIs which will help me to understand which category mostly suits each user. Therefore, I have 7 PoI categories: university, banks, hospitals, industrial, train station, tourists, and others. Here I have added a train station category only for the possibility to distinguish residents and commuters. If a user starts or ends its ride near any PoI it will add +1 to the following PoI category. If starting or ending points lie near two or more PoIs, let's say near the bank and university building, it will add +1 to both of them. It should be done to avoid the superiority of one over the other because two PoIs in this case are equal.

To check if the ride's start or end points are near any PoIs, I'm going to use the 100-meter far criterion. If the distance between the start or end points and any PoIs far than 100 meters, then it means that this point does not belong to any profile's PoI, and goes to the Others category. As a result, I need to check 1 million records of active users and 1.2 million records of all users and profile them. The exact number of distance calculations of each record equal to 235 total PoIs * 2 start and end points or 470 calculations per each record in the dataset. For two classes I have the following number of calculations:

1. Active users records: 470 million distance calculations
2. All users records: 564 million distance calculations

The number of distance calculations is huge even for only active user records. To handle such a large number of calculations, I'm going to use the multiprocessing technique available

user$_id$	University	Banks	Hospitals	Industrial	Train station	Tourists	Others
115	187	58	95	0	126	45	10

Table 3.1: Example of a 7D vector

in Python. This technique allows me to greatly reduce the calculations' total execution time. To understand the scale of the reduction, I can share the following information: without multiprocessing, 1.2 million distance calculations require around 5 minutes, by simple calculations 564 million calculations require 2350 minutes. But with multiprocessing (14 cores and 20 threads), this amount of calculations requires only 117 minutes, so it's almost 20 times faster.

In the end, I have 7D vectors for each user in datasets. 7D vector contains how many times each user has visited desired PoIs. As an example, I can share the 7D vector for one user. Example can be seen in the Table 3.1. As we can see it's a user with the most rides to the University buildings.

Now it's time for selected approaches. Let's start with the simplest approach, I called it this because of the simple and intuitively understandable implementation. After receiving the 7D vector for each user, the simplest approach is profiling each user by the most visited category in this 7D vector. As an example let's take the vector that we saw in Table 3.1. By the simplest approach, this user becomes a student, because the highest number in this vector is related to the University buildings, this means that this user mostly visited University buildings. Also, the simplest approach can divide users into commuters and residents. If the user has mostly visited a train station, then the algorithm will search for the second most visited category. Therefore, any of the 6 profiles could be also divided into commuters and residents. As an example, it could be student-resident and student-commuter.

K-means approach was explained in the 3.1, therefore no need to explain again how this approach works. As before, it's taking 7D vectors and clustering users into 6 profiles. K-means couldn't distinguish users as commuters or residents, therefore users will be reviewed as a whole thing.

The Gaussian mixture model is one of the popular clustering algorithms which is a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters. The main part of GMM is the Expectation-Maximization (EM) algorithm. The algorithm works through steps:

1. Randomly assigning cluster centers;
2. Computes for each center point a probability of being generated by each component of

the model;

3. Tweaks the parameters to maximize the likelihood of the data given those assignments;
4. Repeat 2-3 steps until convergence to a local optimum.

As before with other approaches 7D vectors are given to the GMM algorithm, and it tries to cluster them into 6 profiles. As for K-means, GMM couldn't divide users into commuters and residents.

Additionally, I have tried other clustering algorithms: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and SOM (Self-Organizing Map). Unfortunately, DBSCAN and SOM couldn't solve this problem correctly. Therefore, results of the simplest, K-means and GMM approaches I will include in the results section.

4

Results and Discussion

This chapter is fully dedicated to the results obtained during and after users profiling. As I mentioned before I profile users into 6 profiles: students, bank workers, hospital and other medical workers, tourists, industrial zone companies workers, and others. For better convenience I will split results for active and all user classes.

4.1 Active users results

After collecting all active user records, I can start profiling them. As I mentioned in the previous chapter, to profile a user into one of the desired profiles I'm using three different approaches. As expected different approaches are giving different results. Profiling results using the simplest approach available in Table 4.1. This table contains residents and commuters columns. We can easily see that residents number is larger than the commuters' number. Now it's a good idea to compare residents and commuters:

1. Residents' average ride distance is 1516 meters, while for commuters is 1561, it's quite identical results;

2. Residents: 45% bike and 55% electric bike, commuters: 48% bike and 52% electric bike;

3. Differences in Pass groups: residents mostly buy Premium Passes, while commuters mostly Pay for their rides.

Profile	Residents	Commuters
Bank workers	897	201
Hospital workers	373	66
Industrial workers	97	19
Tourists	-	526
Students	4403	1104
Others	2676	645

Table 4.1: Simplest approach results divided into residents and commuters for active users

Profile	Number of users
Bank workers	1098
Hospital workers	439
Industrial workers	116
Tourists	526
Students	5507
Others	3321

Table 4.2: Simplest approach joint results for active users

I couldn't say that these differences are significant and should be used in future analysis, therefore I decided to merge commuters and residents. Table 4.2 shows us results, but now residents and commuters are joint. As we can see students are the majority category.

Now I want to do some comparisons between student and non-student users, by non-students I mean users except students. Average ride distance comparison:

1. Students: 1447 meters;
2. Non-students: 1599 meters.

So we can see that non-student users have a larger average ride distance than student users. So we have a tendency for non-student users to travel larger distances, the average difference is equal to 152 meters.

Now I want to share the comparison in an average number of rides per user. Results:

1. Students: 93 rides per user;
2. Non-students: 94 rides per user.

As we can see there is no difference between them. But what about the ratio of using bikes and e-bikes? The answer is the following:

Pass Group	Students	Non-Students
Paying	27%	30%
Premium Pass	34%	35%
MonthCard & Daily Pass	26%	16%
Times Pass	13%	19%

Table 4.3: Students and non-students pass group distribution for active users

	K-means	GMM	Simplest
Bank workers	516	1122	1098
Hospital workers	969	1534	439
Industrial workers	276	344	116
Tourists	839	667	526
Students	5876	4289	5507
Others	2531	3051	3321

Table 4.4: K-means and GMM profiling results for active users

1. Students: 54% bike, 46% e-bike;
2. Not Students: 37% bike, 63% e-bike.

I expected these results when students mostly prefer to common bikes and non-student users prefer e-bikes. The most probable reason for that is that common bike price rates and subscriptions are cheaper and economically beneficial rather than for e-bikes. To better understand let's look at Table 4.3. It's easy to see that non-student users usually pay more for their rides, while students purchase more MonthCard and Daily Passes.

After these comparisons, I could say that the difference between student and non-student users is tangible, but not critical.

Now I want to show the results of the other two approaches. Table 4.4 shows K-means and GMM profiling results, but also I have added results of the simplest approach for better comparison. I can mention that every approach did some overcounts and undercounts in comparison with the simplest approach, but in total all results seem to be plausible and valid. Padova city firstly is a student city, with over 70k students, therefore it's obvious expect that most of the users are students. I will use the simplest approach results for future post-profiling analysis.

As we can see profile with the least number of users is industrial area workers. And it's fully understandable because the industrial area is located quite far from the city center and it's problematic to reach these places.

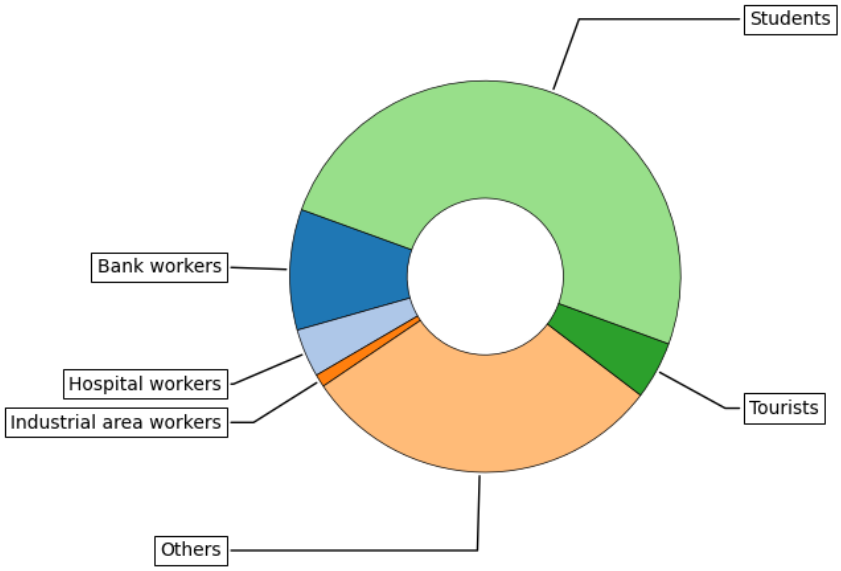

Figure 4.1: Active users profile distribution by the simplest approach

A Bank worker's profile probably contains many users that live in the same or neighborhood buildings where these profile PoIs located. Usually, banks rent commercial premises inside residential buildings. The others column means other users who have no significant number of visits to any other of the selected PoIs. This is fully normal because besides these profiles there exist an enormous number of other professions and possibilities to profile them, but the main problem with these possibly "new" profiles is that there is no specific pattern to recognize and profile them.

In Figure 4.1 it's clearly shown that the most active users are students. It confirms my expectations about students as the biggest profile in the dataset. As I mentioned before the reason for that is quite simple and obvious: Padua is one of the largest student cities in Italy.

4.2 All users results

Here I want to use the same approaches I used for the active users dataset, but now for the whole dataset. Quick reminder the number of users in the whole dataset or in other words the number of all users is equal to 55 thousand. As before, I start with the simplest approach. The results can be seen in the Table 4.5. As before there are more residents rather than commuters. Table 4.6 is showing the merged residents and commuters results.

For now, I want to compare students with non-student users. Average distance comparison

Profile	Residents	Commuters
Bank workers	5572	1452
Hospital workers	1099	290
Industrial workers	435	101
Tourists	-	4783
Students	20039	6035
Others	13147	2359

Table 4.5: Simplest approach results divided into residents and commuters for all users

Profile	Number of users
Bank workers	7024
Hospital workers	1389
Industrial workers	536
Tourists	4783
Students	26074
Others	15506

Table 4.6: Simplest approach joint results for all users

between students and non-student users gives me the following results:

1. Students: 1465 meters;
2. Not students: 1623 meters.

As before non-student users have a larger average ride distance than student users. So we have confirmed the tendency that non-student users travel larger distances, but now the average difference is a bit higher 158 meters.

Now the comparison is in an average number of rides per user. Results:

1. Students: 23 rides per user;
2. Non-students: 21 rides per user.

As we can understand from these results, the number of rides per user highly decreased rather than it was for the active users. It happened because most users were occasional. But even that students have more rides than non-students.

The ratio between using bikes and e-bikes remains almost the same as it was before:

1. Students: 52% bike, 48% e-bike;

Pass Group	Students	Non-Students
Paying	35%	38%
Premium Pass	30%	30%
MonthCard & Daily Pass	23%	14%
Times Pass	12%	17%

Table 4.7: Students and non-students pass group distribution for all users

	K-means	GMM	Simplest
Bank workers	431	3152	7024
Hospital workers	1319	6610	1389
Industrial workers	354	469	536
Tourists	4153	85	4783
Students	43293	37343	26074
Others	5762	7 653	15506

Table 4.8: K-means and GMM profiling results for all users

2. Not Students: 37% bike, 63% e-bike.

I observe some small changes in pass group distribution. Now I have a bit of changed results available in Table 4.7. I can mention that the percentage of users paying for their rides increased for both profiles. The reason for that is the fact that the whole dataset contains many one-time rides, of course, it's cheaper to pay once rather than buy a subscription for a one-time ride.

Now I can use other approaches. Profiling results can be found in the Table 4.8. In the whole dataset, there are more users, hence more profiled users. As we can see results for every approach vary greatly. In comparison with the simplest approach results of K-means and GMM overcounts and undercounts for multiple profile categories. Therefore, K-means and GMM do not seem to be a good idea to use for all-users dataset. Only the simplest approach results seem to be plausible and valid. Let's check the chart pie of all profiles. Chart pie in Figure 4.2 shows almost identical results as it were for the active users, with some changes in profile proportions.

4.3 Profiles analysis

In this section, I decided to analyze and compare different profiles. Also, I want to highlight the fact that I use only active users profiling results, because all-users profiling is identical, and

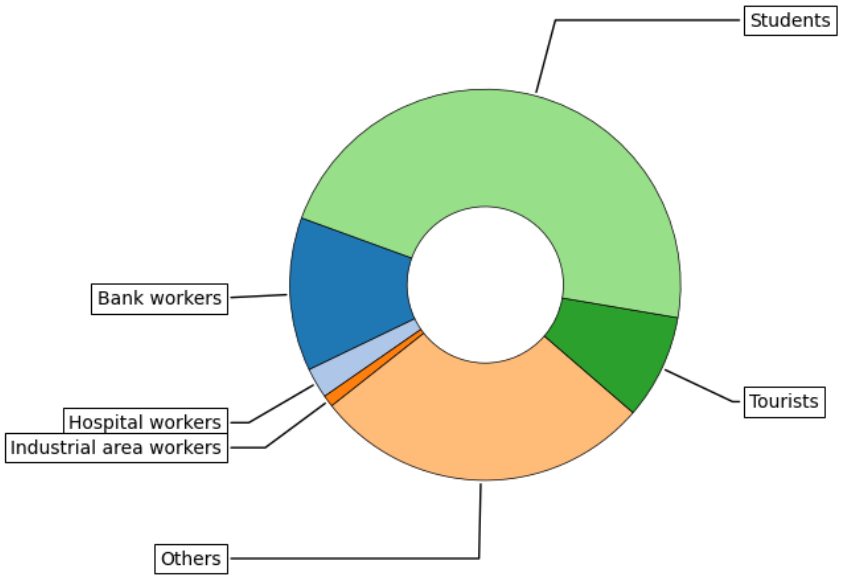

Figure 4.2: All users profile distribution

the main difference is only in the number of users in each profile, but other options like tendency, pattern, and others remain the same as for active users. Therefore, active users results are generalized versions of all-users results.

Let's start with the average ride distance plot available in Figure 4.3. As we can see industrial workers have the highest average distance among other profiles. As I mentioned in previous chapters the reason for that is quite intuitive: the industrial zone is located far from the city center. This was the average ride distance for the whole period of observational time.

What about the changes that happened per month during the period? To answer this question I prepared the graph in Figure 4.4. Only the industrial workers' line has significant ups and downs, while other profile lines seem to be more stable. The main reason for such a line with industrial workers is an insufficient number of people in this profile and, accordingly, a small number of records.

The last comparison in average ride distance is a comparison per day of the week. Results are shown in the Figure 4.5. Almost all profiles tend to ride more distances on the weekends. While the line on working days seems to be straight.

The next thing to compare is the average number of rides. As before I want to start with the comparison per month for different profiles. As we can observe in Figure 4.6 students are making most of the rides during the whole period. This students line is proving my previous words about increasing the number of rides because of the new academic year start and decreas-

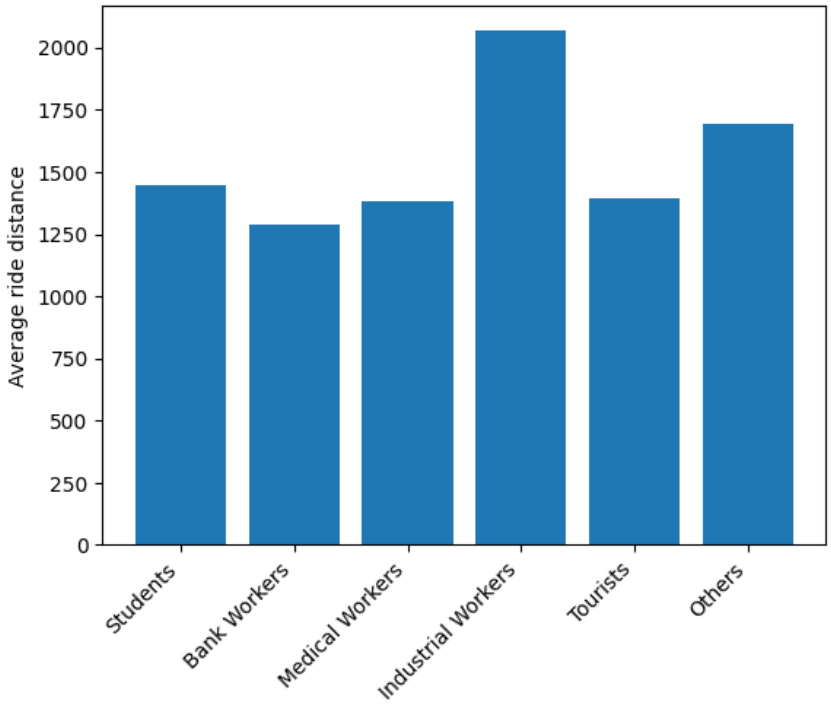

Figure 4.3: Average ride distance for different profiles

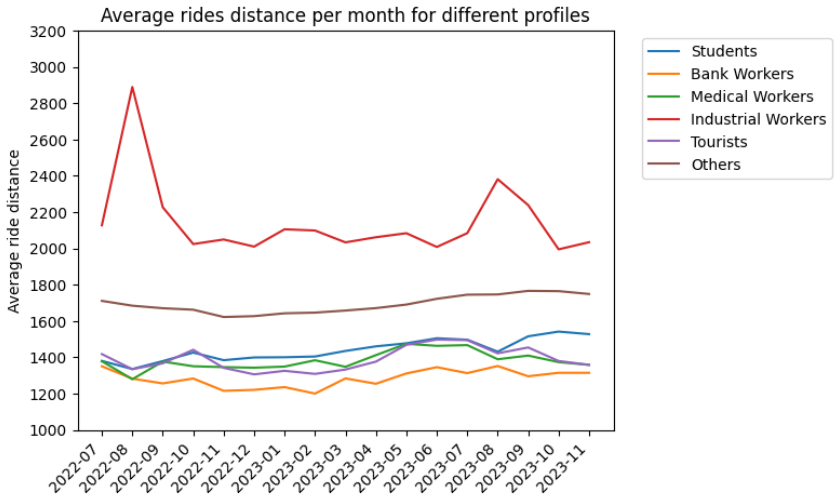

Figure 4.4: Average rides distance per month for different profiles

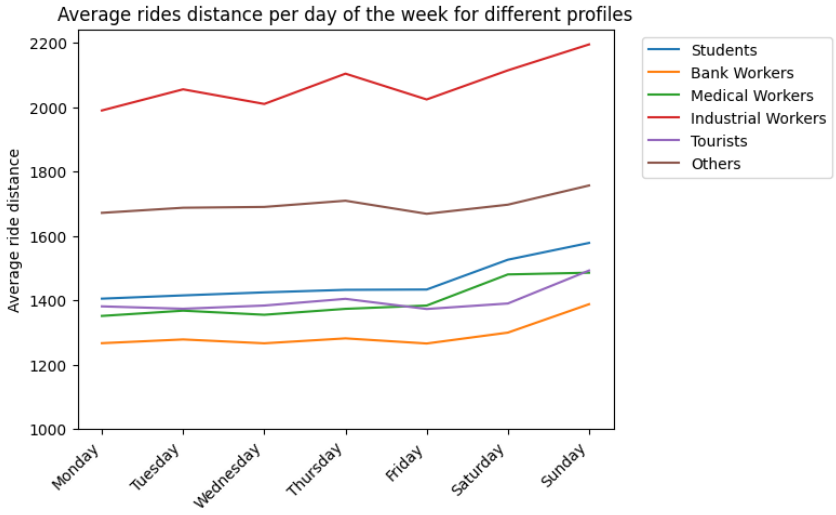

Figure 4.5: Average rides distance per day of the week for different profiles

ing the number of rides during the summer holidays. While in other profiles the number of rides lines is quite straight and doesn't change significantly. I can mention that students mostly depend on the weather conditions, while other profiles' number of rides remains the same even in winter months.

Let's check the rides number per day of the week. The graph can be found in the Figure 4.7. All profiles tend to ride less number of times during the weekends, especially since it easy to see in students' line, after Friday it's significantly drops down.

Let's see how the average rides duration changes per month for different profiles. Figure 4.8 can be found in rides duration in seconds. We can see that most rides duration doesn't change much through months.

As usual average rides duration per day of the week for different profiles is available in Figure 4.9. Duration graphs are fully connected with the average rides distance graphs because distance and time are interdependent measures. That's why I have similar graphs for duration and distance.

To better understand rides duration time I prepared Figure 4.10, where we can see that the average ride duration for all profiles lies in the range between 6 and 9 minutes.

This histogram was for all profiles, but what about different profiles then? To answer I made two histograms: Figure 4.11 and Figure 4.12. From these histograms, I can obtain the usual riding durations for these profiles:

1. Students: 6-7 minutes;

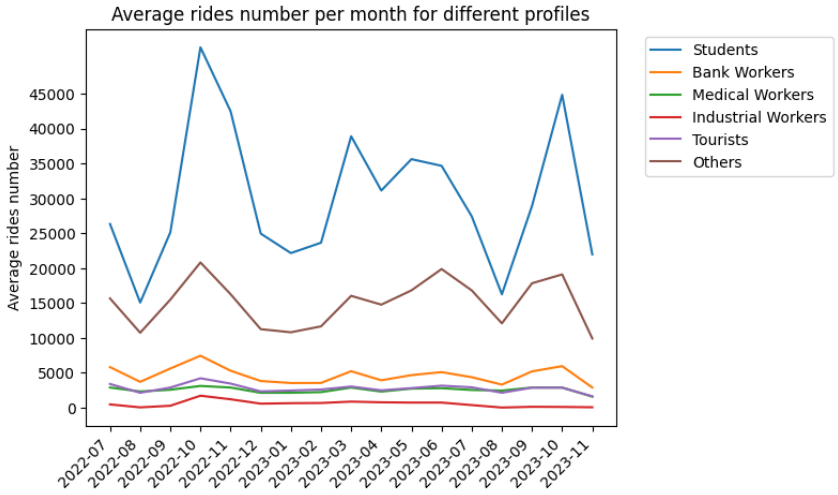

Figure 4.6: Average rides number per month for different profiles

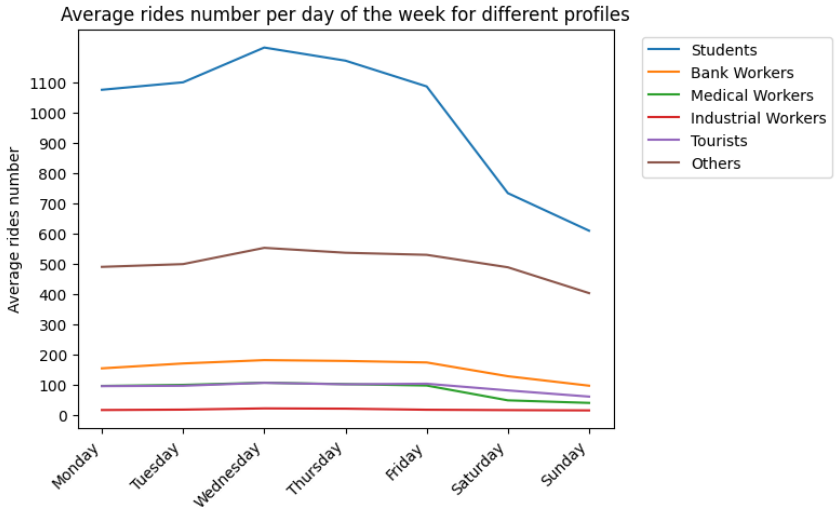

Figure 4.7: Average rides number per day of the week for different profiles

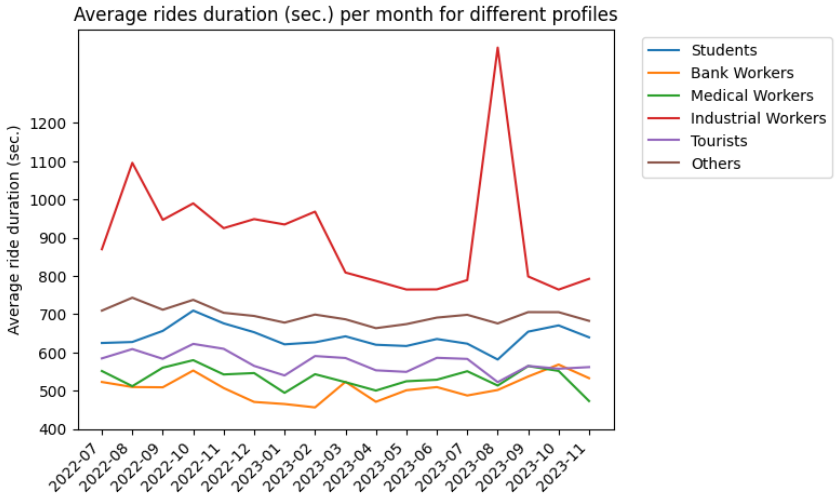

Figure 4.8: Average rides duration (sec.) per month for different profiles

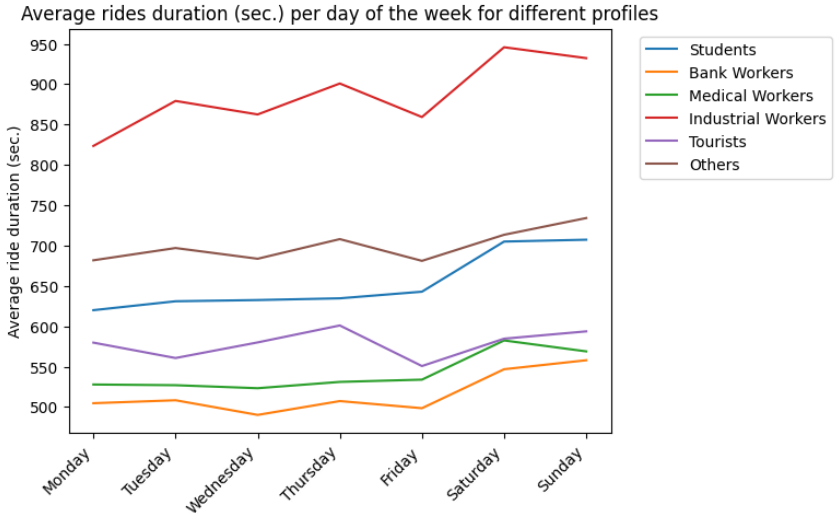

Figure 4.9: Average rides duration (sec.) per day of the week for different profiles

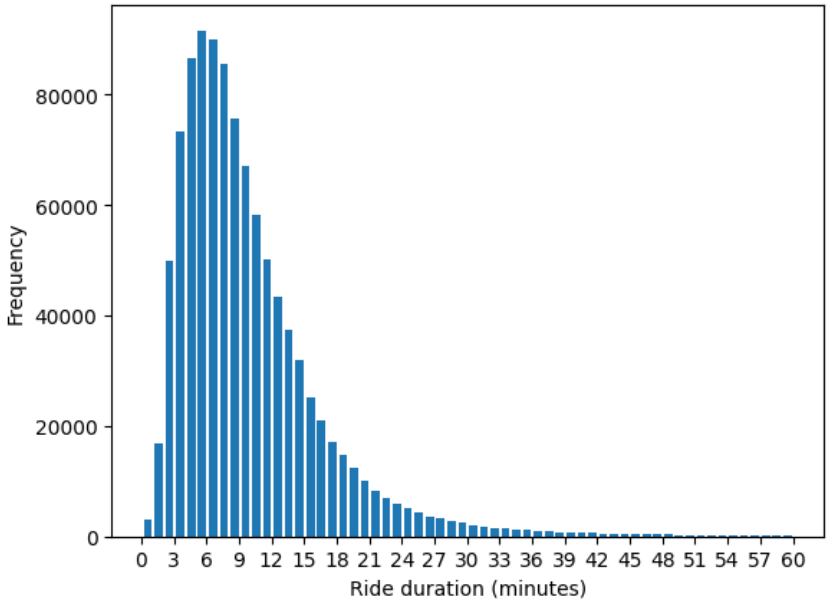

Figure 4.10: Rides duration histogram

2. Bank workers: 4-5 minutes;
3. Hospital and Medical workers: 5-6 minutes;
4. Others: 6-9 minutes;
5. Tourists: 4-5 minutes;
6. Industrial zone workers: 8-12 minutes.

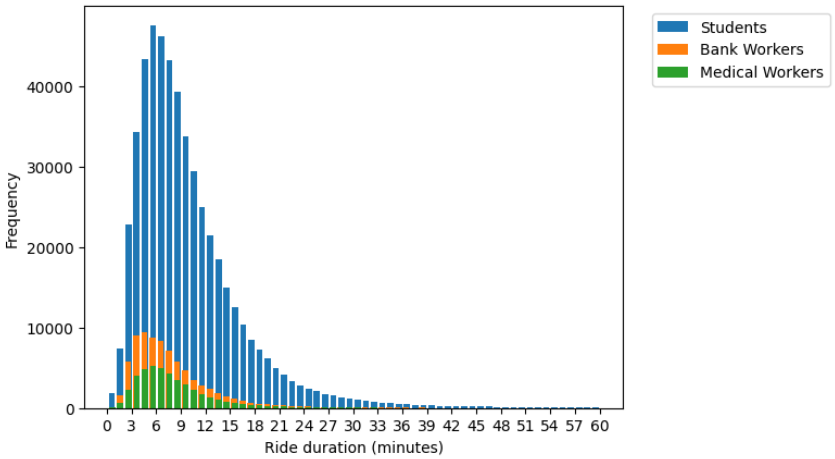

Figure 4.11: Duration histogram: students, bank and hospital workers

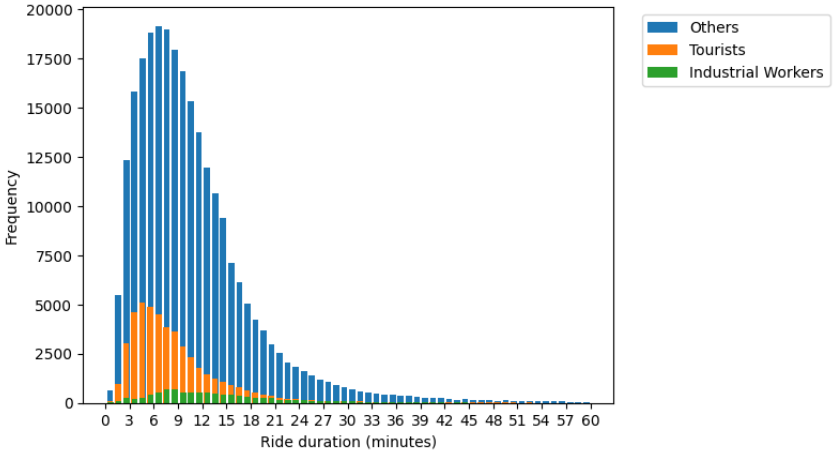

Figure 4.12: Duration histogram: others, tourists and industrial workers

5 Conclusion

In my thesis, I explored the use of user profiling for data analysis. I used various profiling methods based on the starting and ending points of bike rides. Some of my initial assumptions, like the majority of users being students, were confirmed by the analysis. However, other expectations, such as the relationship between ride frequency and weather conditions, were not confirmed.

Through the analysis, I identified different profiles for active users with at least 20 rides, as well as for all-users who had at least one ride during the observation period. These profiles revealed tendencies in average ride distance, duration, and frequency per month and weekdays, as well as more general usage patterns. It became evident that each user has unique behaviors. While I was able to identify and observe six distinct profiles in my thesis, many other profiles and user professions did not exhibit specific patterns.

The most easily identifiable profile was that of students, who displayed distinct behaviors such as frequent rides to university buildings, canteens, study rooms, and student residential areas.

These user profiles can be leveraged in various business strategies. For instance, RideMovi and other bike service companies could offer special subscriptions tailored to different user profiles. For example, a ”for students” subscription could allow students to ride to university areas at no additional cost, while rides to other locations would be priced at standard rates.

Another potential area for user profiling is categorizing users based on ride frequency and distance. For example, users who ride frequently but for short distances could be offered a

special subscription allowing them to ride for free or low cost for short distances, with no restrictions on the number of rides per day. On the other hand, users who ride long distances could be offered a different subscription that allows longer rides but with a limited number of uses per day. These specialized subscriptions could be priced attractively and are likely to find customers.

References

[1] M. Cavattoni, M. Comin, and F. Silvestri. Long term effects of the pandemic on urban mobility: The case of free-floating bike sharing in padova. [Online]. Available: https://ssrn.com/abstract=4272467

[2] Openstreetmap contributors (2015). [Online]. Available: https://www.openstreetmap.org

[3] P. Zippenfenig, "Open-meteo.com weather api," 2023. [Online]. Available: https://open-meteo.com/

[4] Kepler.gl (2024). kepler.gl is a powerful open source geospatial analysis tool for large-scale data sets. [Online]. Available: https://kepler.gl

[5] Hoodmaps: different demographics and neighbourhoods in cities around the world. [Online]. Available: https://hoodmaps.com/padova-neighborhood-map

[6] Google (2024). google maps. [Online]. Available: https://www.google.com/maps/place/Padua

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Professor Francesco Silvestri, for his unwavering support and guidance throughout my thesis. His expertise, patience, and insightful feedback have been invaluable to the development and completion of this work. I am deeply thankful for the time and effort Professor Silvestri invested in supervising this thesis, and for the meetings and discussions that helped shape my ideas and improve my work. I am truly grateful for the opportunity to learn and grow under your supervision.

To my family and loved ones, thank you for your patience, understanding, and endless encouragement. Your support has been invaluable, providing me with the comfort and reassurance needed to persevere through the challenges of my studies. I would also like to extend my heartfelt thanks to anyone who has shown me friendship and kindness during my time as a master's student.

A special thanks to the University of Padua for providing me with the opportunity to pursue my master's degree and for offering a supportive and enriching academic environment. I am particularly grateful to my professors and advisors for their guidance, knowledge, and encouragement throughout this journey. Your dedication and expertise have been instrumental in my academic and personal growth.

I thank the Municipality of Padova for providing the RideMovi rides dataset.