POLITECNICO DI TORINO

ICT in Smart Mobility

Labratory Reports - Group 12

Report 1 - Data analysis

Authors

Arman MohammadiGilani	s301000
Sara Gholamhosseinzadeh	s301689
Nikoo Arjang	s300262

Supervisors
Marco Mellia
Luca Vassio

Academic Year 2022/23

Introduction

A MongoDB database that is accessible on a server owned by Politecnico di Torino is used for the data analysis. The Appendix contains all the codes.

1 Task 1 – Preliminary data analysis

1.1 How many documents are present in each collection?

There are four collections for Car2Go named: ActiveBookings, ActiveParkings, PermanentBookings and PermanentParkings. The counts are as follows: *ActiveBookings*: 8743; *ActiveParkings*: 4790; *PermanentBookings*: 28180508; *PermanentParkings*: 28312676.

Also, there are four collections for Enjoy. The counts are as follows: *ActiveBookings*: 0; *ActiveParkings*: 0; *PermanentBookings*: 6653472; *PermanentParkings*: 6689979.

1.2 Why is the number of documents in PermanentParkings and PermanentBooking similar?

The logic is that when a car is booked, it will be counted as active booking; when it is parked, it will be counted as ActiveParking. The permanent booking and parking are the records saved as it means that the booking and parking are already made, so the numbers are similar because parking follows the booking of a car. The parked ones are more because some booked cars were probably unused.

1.3 For which cities is the system collecting data

The cities that are in the *Car2Go* Data Collection are as follows (26 cities): 'Wien', 'Washington DC', 'Vancouver', 'Twin Cities', 'Toronto', 'Torino', 'Stuttgart', 'Seattle', 'San Diego', 'Roma', 'Rheinland', 'Portland', 'New York City', 'Munchen', 'Montreal', 'Milano', 'Madrid', 'Hamburg', 'Frankfurt', 'Firenze', 'Denver', 'Columbus', 'Calgary', 'Berlin', 'Austin', 'Amsterdam'

Additionally, the following is a list of *Enjoy Data* Collection (6 cities): 'Bologna', 'Catania', 'Firenze', 'Milano', 'Roma', 'Torino'

1.4 When the collection started? When the collection ended? Time started and ended?

For *Car2Go* data collection:

First measurement: Stuttgart	Last measurement: Washington DC
Timestamp: 1481650703	Timestamp: 1517404293
GMT: 2016-12-13 18:38:23 at GMT+1.	GMT: 2018-01-31 14:11:33 at GMT+1.
Local timezone: 2016-12-13 18:38:23	Local timezone: 2018-01-31 08:11:33

For *Enjoy* data collection:

First measurement: Catania	Last measurement: Milano
Timestamp: 1493996781	Timestamp: 1560186980
GMT: 2017-05-05 17:06:21 at GMT+1.	GMT: 2019-06-10 19:16:20 at GMT+1.
Local timezone: 2017-05-05 17:06:21	Local timezone: 2019-06-10 19:16:20

1.5 What about the time zone of the timestamps?

The time zone of the time stamp is based on GMT, Greenwich Mean Time. Local time zones are used to represent dates in a human readable format.

1.6 What is the total number of cars seen in the whole period in each city? How does this relate to the fleet size at a given time? How many bookings?

In Berlin, 1871 car2go vehicles have been listed. Only 1188 were running as of the last measurement. In Firenze, 455 car2go vehicles have been listed. Only 225 were running as of the last measurement. In Toronto, 641 car2go vehicles have been listed. Only 460 were running as of the last measurement.

1.7 How many bookings have been recorded on January 2018 in each city?

The recorded number of bookings in January 2018 in each city is as follows:
Berlin: 357217; *Firenze: 41423*; *Toronto: 45261*.

1.8 How many bookings have also the alternative transportation modes recorded in each city?

Our chosen cities don't have the transportation mode field.

2 Task 2 – Analysis of the data

2.1 Derive the Cumulative Distribution Function of booking/parking duration and plot them. Which consideration can you derive from the results?

In the initial phase of the data analysis, the months between the beginning of December 2017 and the 31st of January as the most recent date were considered. The analysis of system utilization over time was undertaken by calculating the number of monthly rentals and dividing it by the number of weekdays. Figure 1 displays the CDFs for both bookings and parking. Due to outliers, the booking's CDF curve appears to be longer. Approximately 80% of parking durations exceed 120 minutes. This is most likely because the car has been parked for an extended period of time, such as at night, or has been serviced.

a. Which of the CDFs is longer? Are there some outliers?

In each of the three cities, the longest CDF is the one associated with the duration of the bookings. It increases to 10^5 minutes in Toronto and 10^4 minutes in Firenze and Berlin. This behavior is unusual, but it cannot be recognized by the system in the event of a car malfunction or system failure.

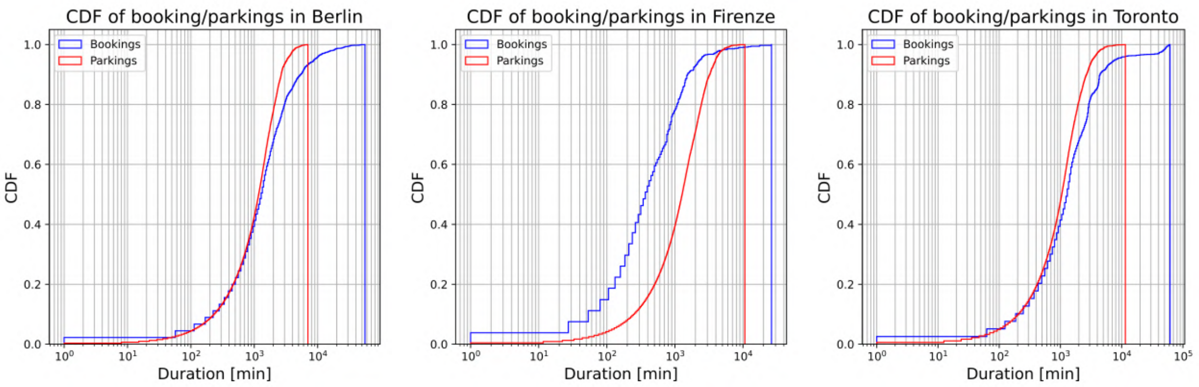


Figure 1: CDF of RAW Bookings/Parkings.

b. Does the CDF change per city? Why?

Parking usually takes up more time than the booking time, although the overall pattern varies mostly as a result of urban density, which causes delays in denser metropolitan areas.

c. Does the CDF change over time (e.g., aggregate per each week of data, or each day or the week?) Why?

For Berlin and Toronto, the CDF of bookings does not dramatically alter over the duration of a day; however, for Firenze, as seen in Figure 2, it fluctuates with time, with the weekend's length, for example, being longer than the weekdays.

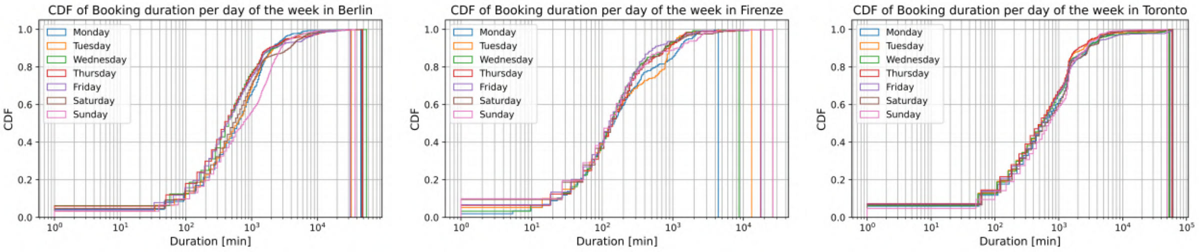


Figure 2: CDF of RAW Bookings data per each day of the week.

2.2 Consider the system utilization over time: aggregate rentals per hour of the day, then plot the number of booked/parked cars (or percentage of booked/parked cars) per hour versus time of day. Do you notice any outliers? Can you explain them?

Figure 3 shows a plot of the total number of bookings that were made throughout each hour of the day. There is a common pattern throughout all the cities. The quantity of booked cars increases in the morning, peaks in the afternoon, and decreases during the night. This helps to highlight several anomalies that were found in the data. To begin, several clusters of outliers may be identified in the data. Because the system was unable to acquire data, it seems that there were no bookings made for the 10th of December in any of the cities.

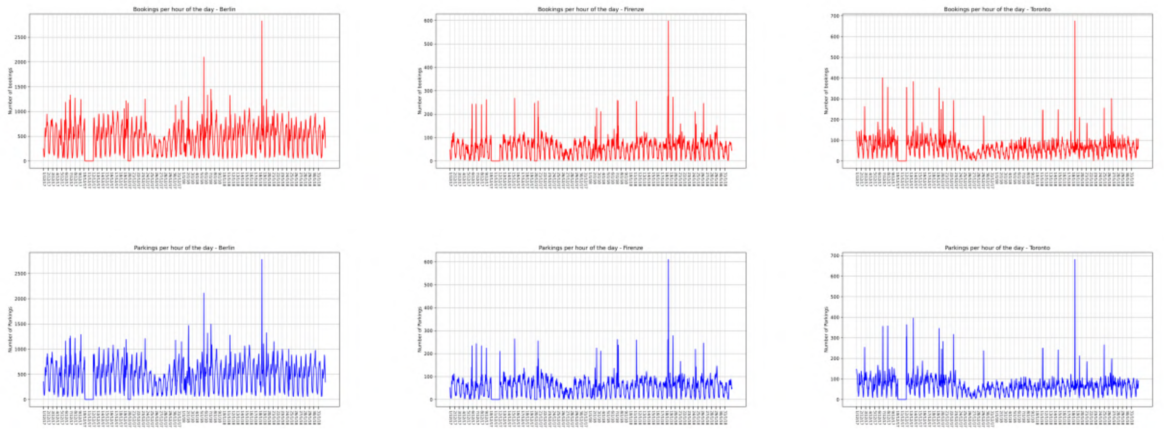


Figure 3: Number of bookings/parkings per hour of day from DEC ”17 through JAN ”18 period.

2.3 Derive a criterion to filter possible outliers (booking periods that are too short/too long) so to obtain rentals from bookings, filtering system issues or problems with the data collection.

It must be noted that data is missing on December 10th; this is most likely the result of a system malfunction at that time, since the exact same issue can be seen in all datasets. Continue working on the analysis; as was previously said, certain outliers in the collection compromise an accurate data analysis. In order to highlight the trends of the filtered statistics of bookings/parkings over days, Figure 4 displays the results of some filtering operations on bookings/parkings of Berlin. Many rentals have been excluded because of their short or long durations or because the coordinates of origin and destination were the same.

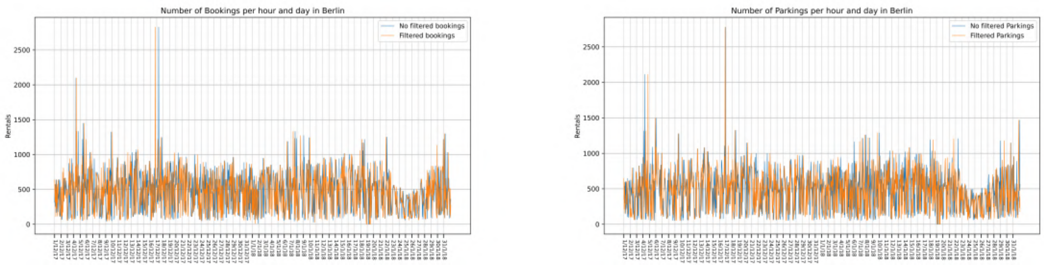


Figure 4: Filtering of bookings and parkings in Berlin.

2.4 Filtering data as above, consider the system utilization over time and the CDF of booking and parking duration. How do this change? Are you able to filter outliers efficiently for both type of events? Consider also to plot the CDF of the filtered events. How do these appear?

Figure 5's filtered representation of the total number of bookings/parking aggregated by hour of the day reveals a more predictable trend. The removal of outliers shows a number of notable events, including not just an overall decrease in system utilization over holidays but also the emergence of a weekly periodic pattern with much fewer bookings throughout the weekends. These patterns make it seem as if the outliers were successfully removed. There are still some peaks in the parking pattern, which is probably the result of a systemic failure.

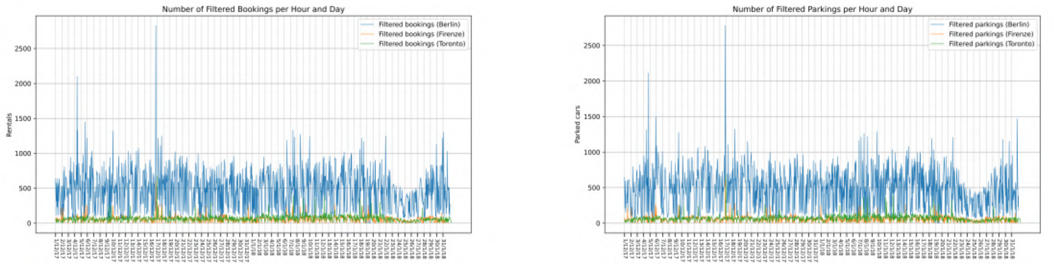


Figure 5: Number of Filtered Booked/Parked cars data.

2.5 Filtering the data as above, compute the average, median, standard deviation, and percentiles of the booking/parking duration over time (e.g., per each day of the collection).

Figure 6 demonstrates once again how filtering might affect various data statistics. With the mean, median, standard deviation, and 90th percentile of the bookings duration each day for the cities, it attempts to emphasize the patterns of the data of bookings across days. The median is usually significantly lower than the mean, which is the first characteristic that stands out. There are occasional increases at the beginning of the weeks (for example, on the 4th and 5th of December), but for instance, New Year's Eve sees one of the the lowest average length of bookings.

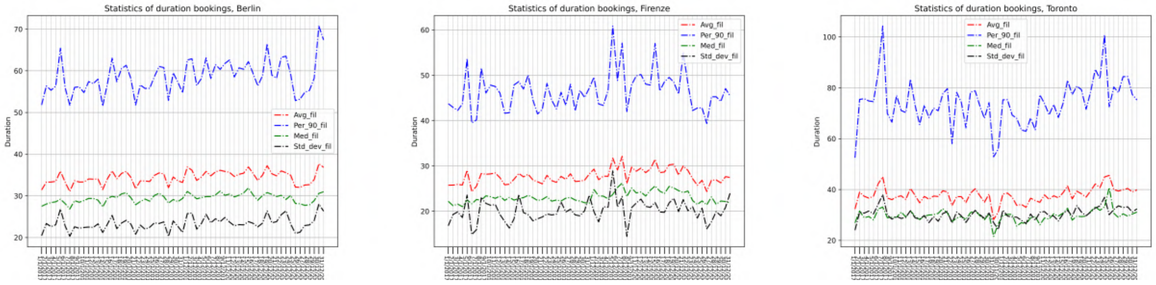
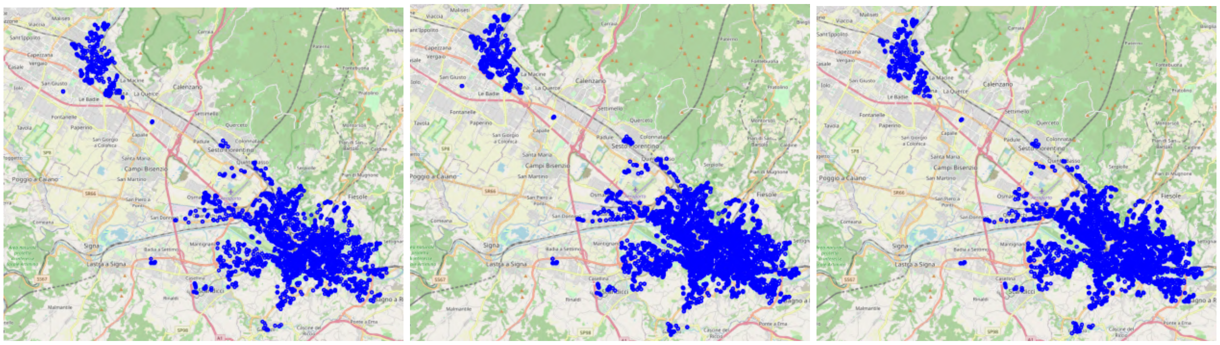


Figure 6: Statistics of filtered data for bookings.

2.6 Consider one city of your collection and check the position of the cars when returned and compute the density of cars at rental ending time (the destination matrix) during different hours of the day.

The purpose of this task is to evaluate the distribution of car parks across the city. Due to its smaller size as compared to Berlin and Toronto, Firenze was chosen for this purpose. Three slots at a time were selected. In November 2017, the chosen timespan ranged from 6 to 12, 12 to 18, and 18 to 24 hours. Figure 7 shows the parked cars throughout the day.



(a) 6 to 12.

(b) 12 to 18.

(c) 18 to 24.

Figure 7: Location of cars in a time interval, Firenze - November 2017.

Another finding is that the distribution of cars is denser in the middle of the day compared to other times. In contrast, it is clear throughout the hours that follow that there are fewer cars overall, especially in the city's center. The spots were divided into zones of 500 meters by 500 meters to examine the density of parked cars around the city over the same time period, as shown in Figure 8, where each color indicates a zone.

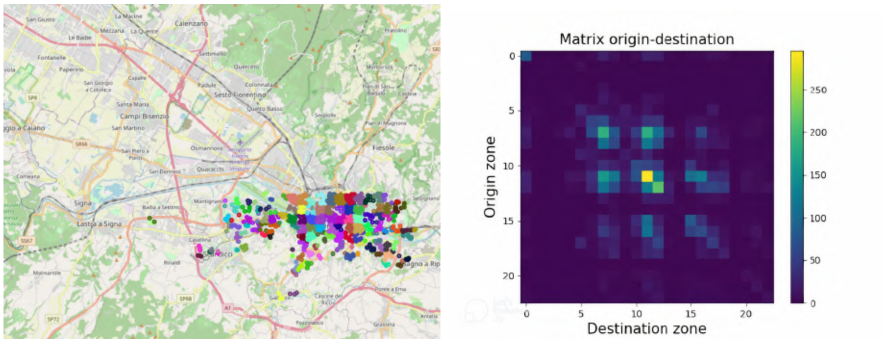


Figure 8: OD matrix regarding Density of cars divided by zones - November 2017.

The origin-destination matrix was built using the zones from the prior experiment. From top to bottom and from left to right, the zones were listed. However, it was challenging to conduct a thorough study since the number of zones was too large with the divide of 500 m. Since the lightest colors are concentrated in the middle of the matrix, it can be observed that the central zones were the ones that were used the most.