NTNU

Kunnskap for en bedre verden

TDT4225 - Very Large, Distributed Data Volumes

Assignment 1

Group 99

Author:
Ola Munthe Vassbotn
Sasha Elisabeth Landell-Mills
Sepanta Jamshid Ganjei

Date

Table of Contents

1 Introduction

In this exercise, the dataset `porto.csv` was explored using EDA, cleaned based on the findings, and then used to answer various questions about the data using SQL and python. The `porto.csv` file contains information about taxi trips taken between July 2013 and June 2014 in Porto, Portugal.

2 Results

2.1 Analysis and Cleaning

The main focus of this section is the analysis and cleaning of the porto.csv dataset for further use.

2.1.1 Dataset before cleaning

Number of rows: 1 710 670.
Number of columns: 9.
Date range: 2013-07-01 to 2014-06-30

In Table 1 the different columns names and their type is shown. CALL_TYPE can either be A, B or C. A means the taxi is dispatch from central, B means the taxi is requested directly from a taxi stand, and C means the taxi is randomly hailed down in the street.

Table 1: Overview of columns and their data types before cleaning

Column name	**Data type**
TRIP_ID	int64
CALL_TYPE	object
ORIGIN_CALL	float64
ORIGIN_STAND	float64
TAXI_ID	int64
TIMESTAMP	datetime64[ns]
DAY_TYPE	object
MISSING_DATA	bool
POLYLINE	object

2.1.2 EDA of porto.csv dataset

We first examined how many rows had the attribute `MISSING_DATA` equal to true. We then discovered that this was the case for 10 rows.

Further, we wanted to visualize the data to see how it was distributed across the different columns and to identify any potential outliers. All the code related to these graphs can be found in `visualize_porto.py`.

The distribution of the different call types in the column `CALL_TYPE` was examined through the graph in Figure 1.

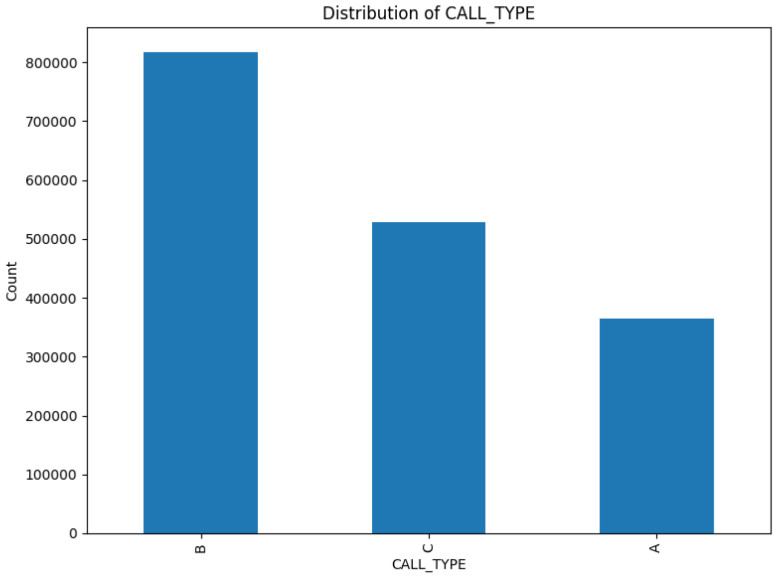

Figure 1: Distribution of the different call types, before cleaning

The number of trips per day were found to be vary from day-to day, as shown in Figure 2. With the two extremal days having 2,185 and 7,493 trips each. The overall average was 4,686.77 trips per day.

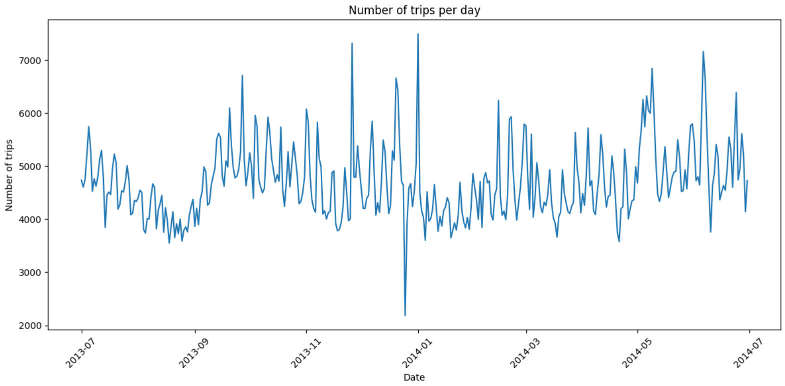

Figure 2: Number of trips registered per day modeled from dataset before cleaning

The distribution of number of trips per hour of the day was examined as well, as seen in Figure 3. The two extremal hours were found to be at 9:00 AM with 98,564 trips, and 2:00 AM with 48,417 trips.

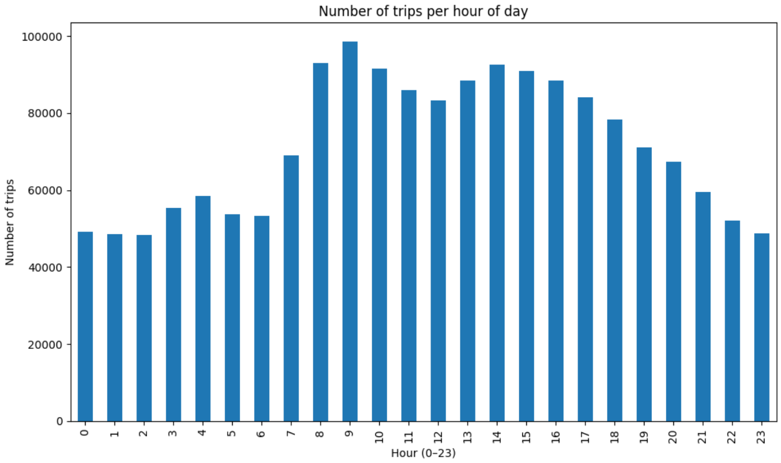

Figure 3: Trips registered per hour of the day, before cleaning

The number of trips were plotted against trip duration using the polyline attribute to visualize the distribution of trip duration. We initially plotted this including all the data, as seen in Figure 4. However, we re-plotted it with a cap at the 95th percentile, as seen in Figure 5. The number of bins in the histogram where chosen to be 100 to divide the trips into 15,3 s intervals. The reason we chose to re-plot the data was because the graph gave little value initially.

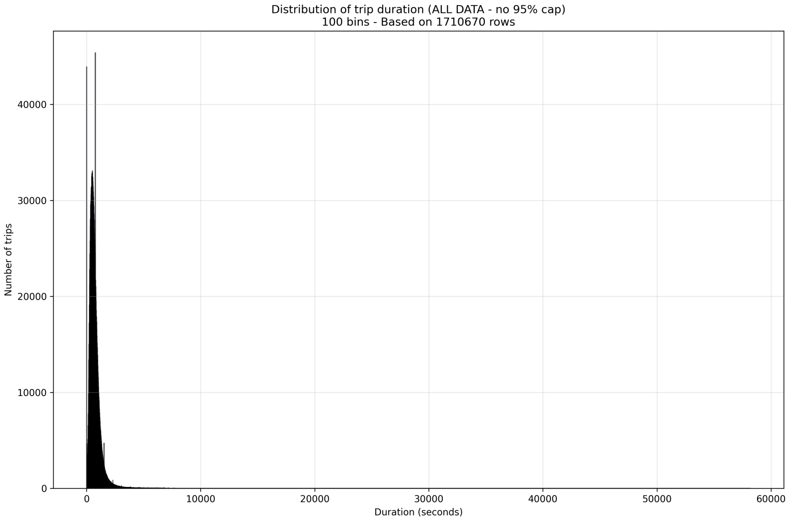

Figure 4: The number of taxi trips plotted against their duration, before cleaning

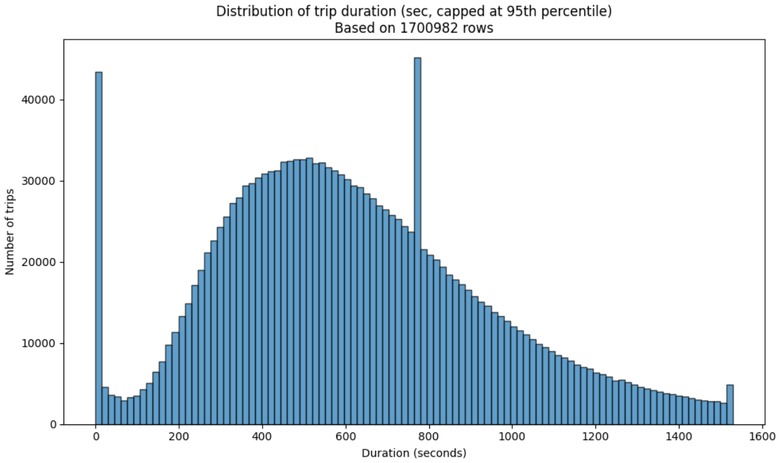

Figure 5: The number of taxi trips plotted against their duration, capped at 95th percentile, before cleaning

The two extremal bins in Figure 5 were found to be:

1. Bin 51:
Duration range: 765.0 - 780.3 seconds
Duration range: 12.8 - 13.0 minutes
Bin center: 772.7 seconds (12.9 minutes)
Number of trips: 45,167
Percentage of total: 2.79Actual trips in range: 45,167
Average duration in bin: 772.4 seconds (12.9 minutes)
Min duration in bin: 765.0 seconds
Max duration in bin: 780.0 seconds

2. Bin 1:
Duration range: 0.0 - 15.3 seconds
Duration range: 0.0 - 0.3 minutes
Bin center: 7.7 seconds (0.1 minutes)
Number of trips: 43,413
Percentage of total: 2.68Actual trips in range: 43,413
Average duration in bin: 2.5 seconds (0.0 minutes)
Min duration in bin: 0.0 seconds
Max duration in bin: 15.0 seconds

2.1.3 Data cleaning results

The dataset was cleaned based on the findings from the EDA, as discussed earlier in Section 2.1.2. Cleaning involved: removing the 10 rows with missing data, removing all rows in the data recorded on the days that were the global extremes for trips registered per day (2,185 and 7,493), and removing trips with duration less than 30.6 seconds. In this section, the results for each cleaning operation are shown in chronological order. This means that the cleaning steps were performed in the order they are listed, and the results build on top of each other.

After removing the rows with missing data:
Number of rows removed: 10 (∼0%)
Columns after cleaning: 8 (MISSING_DATA removed)

After removing the rows with that landed on the extremal days:
Number of rows removed: 9,688 rows (0.57%)
Columns after cleaning: 8

After removing the rows with trip duration of ≤ 30.6s
Number of rows removed: 47,995 (2,82%)
Columns after cleaning: 8

Total number of rows after cleaning:: 1,652,987
Total reduction: 57,683 rows (3.37%)
Total number of columns after cleaning: 8

2.1.4 Dataset after cleaning

Figure 6 displays the updated graph for the number of trips per day, in the cleaned dataset.

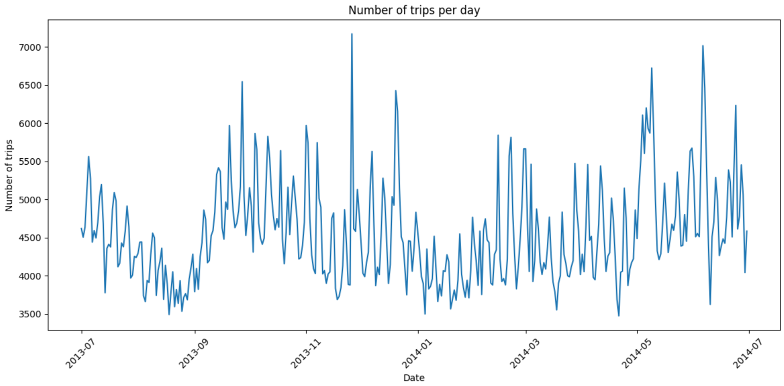

Figure 6: The number of taxi trips registered for each day, after cleaning

Since we removed all short trips (≤30.6s), we plotted a new histogram for the distribution of trip duration. This resulted in a much cleaner and simple curve. See Figure 7.

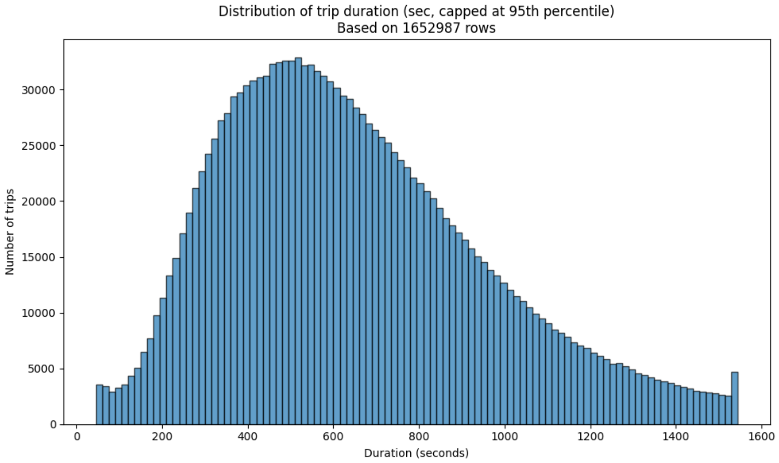

Figure 7: The number of taxi trips plotted against their duration, capped at 95th percentile, after cleaning

Figure 8 however does not display as much of a drastic change after cleaning. Neither does the call type distribution as seen in Figure 9.

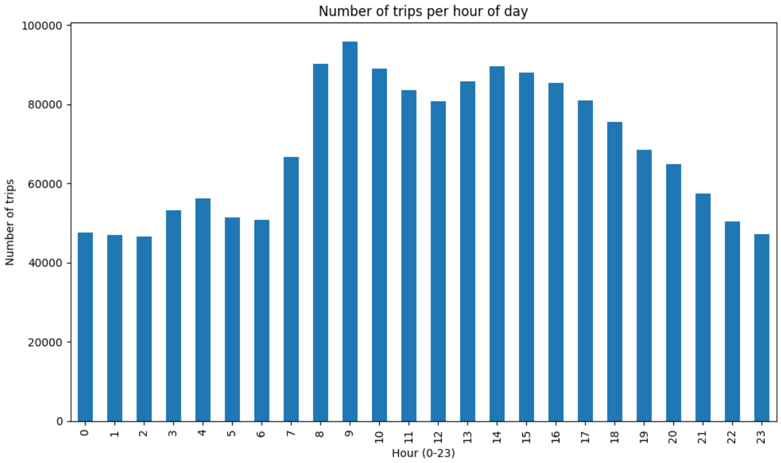

Figure 8: Trips registered per hour of the day, after cleaning

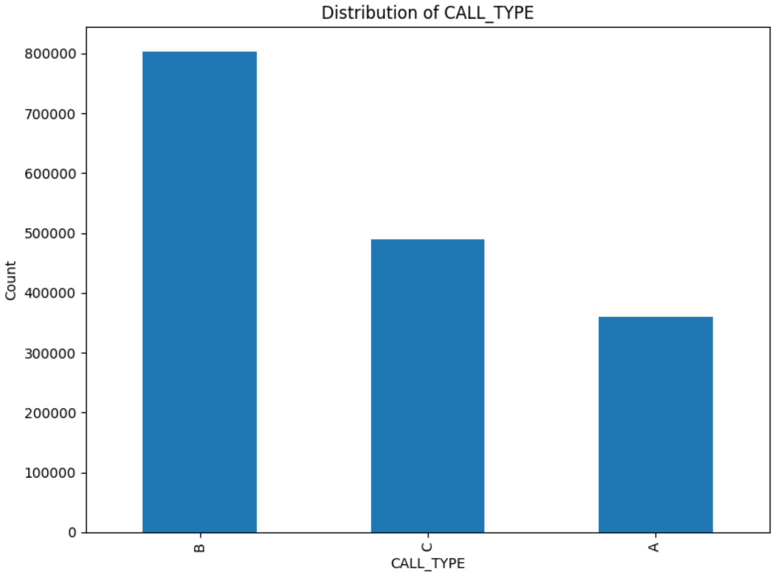

Figure 9: Distribution of the different call types, after cleaning

2.2 Queries

In this section, queries were executed on the dataset to answer the questions presented. Each file containing the corresponding query found within the files delivered with this report, within the `part2` directory. All of them are in the format of `query[NUMBER].py` or `query[NUMBER].sql`.

Some of the results were too large to include in text format in this report. For the queries this applies to, there is a corresponding file within the `part2/results/` directory, named `query[NUMBER]_final_results.json`. It is worth noting that we used the cleaned dataset to answer the questions below, which contains fewer rows than the original dataset.

1. How many taxis, trips, and total GPS points are there?
Taxis: 441 Number of trips: 1652981 Number of GPS points: 82 919 325

2. What is the average number of trips per taxi?

Average trips per taxi: 3748.2698

3. List the top 20 taxis with the most trips.

Table 2: Top 20 taxis with the most trips from cleaned data

Taxi id	Trip count
20000483	7554
20000403	7274
20000307	7209
20000621	7205
20000364	7133
20000492	7048
20000129	7025
20000424	6961
20000089	6879
20000529	6864
20000042	6414
20000678	6409
20000616	6367
20000235	6336
20000304	6335
20000179	6316
20000263	6296
20000325	6283
20000140	6236
20000233	6208

4. a) What is the most used call type per taxi?

A sample of the results are shown in the table below. The full results are available in the delivered files.

Table 3: Most used call type per taxi

taxi_id	most_used_call_type	call_count
20000001	B	1303
20000002	B	1528
20000003	C	818
20000004	B	2654
20000005	B	3004
20000006	B	1687
20000007	B	2733
20000008	B	2617
20000009	B	2066
20000010	B	2799

b) For each call type, compute the average trip duration and distance, and also report the share of trips starting in four time bands: 00–06, 06–12, 12–18, and 18–24.

Call Type A:

- Average Duration: 772.1 seconds

- Average Distance: 5437.6 meters
- Time Band Shares:
  - 00-06: 0.118 (11.8
  - 06-12: 0.335 (33.5
  - 12-18: 0.322 (32.2
  - 18-24: 0.225 (22.5

Call Type B:

- Average Duration: 689.7 seconds
- Average Distance: 5120.9 meters
- Time Band Shares:
  - 00-06: 0.122 (12.2
  - 06-12: 0.301 (30.1
  - 12-18: 0.345 (34.5
  - 18-24: 0.231 (23.1

Call Type C:

- Average Duration: 841.1 seconds
- Average Distance: 6717.9 meters
- Time Band Shares:
  - 00-06: 0.330 (33.0
  - 06-12: 0.232 (23.2
  - 12-18: 0.240 (24.0
  - 18-24: 0.199 (19.9

5. Find the taxis with the most total hours driven as well as total distance driven. List them in order of total hours.

Since this question doesn't specify a defined limit but still asks for the taxis with the 'most' total hours, we have decided to only keep the top 20 results. The results can be seen in Table 4.

Table 4: Top taxis by total hours driven

Taxi ID	Total Hours	Total Distance (km)
20000904	1975.22	62510.04
20000129	1668.06	36268.71
20000307	1606.63	39693.56
20000529	1528.59	42154.97
20000276	1438.61	46845.59
20000436	1422.46	43973.58
20000483	1419.53	35935.60
20000372	1400.79	40348.94
20000616	1366.45	35135.08
20000179	1343.59	39629.70
20000574	1314.55	37211.22
20000235	1305.35	36456.08
20000621	1304.93	31772.40
20000364	1299.00	41259.54
20000435	1294.72	40774.32
20000446	1284.30	35055.88
20000199	1283.95	39179.54
20000011	1281.95	38229.14
20000492	1280.66	32842.81
20000395	1279.24	33986.07

6. Find the trips that passed within 100 m of Porto City Hall. (longitude, latitude) = (-8.62911, 41.15794)

Number of trips that passed within 100m of Porto City Hall: 125 605. A sample of the results are shown below. The full results are available in the delivered files.

Trip IDs:

- 1372637610620000497
- 1372638361620000154
- 1372641991620000231
- 1372636956620000167
- 1372641197620000653

7. Identify the number of invalid trips. An invalid trip is defined as a trip with fewer than 3 GPS points.

We found zero invalid trips, but this is (probably) because we have already removed trips that are below 30s in length.

8. Find pairs of different taxis that were within 5m and within 5 seconds of each other at least once.

Total amount of close pairs: 94158.

The full results are available in the delivered files.

9. Find the trips that started on one calendar day and ended on the next (midnight crossers).

8228 trips were midnight crossers.

The full results are available in the delivered files.

10.Find the trips whose start and end points are within 50 m of each other (circular trips).

A total of 19,527 round trips were found. A sample of the results are shown in the table below. The full results are available in the delivered files.

Table 5: Some of the circular trips (start and end points within 50 m) in the cleaned dataset.

Trip ID	**Distance (m)**	**Start Point**	**End Point**
1372639092620000233	8.77	(41.168295, -8.632737)	(41.168232, -8.632800)
1372638303620000112	2.47	(41.162427, -8.587116)	(41.162436, -8.587143)
1372644642620000305	40.94	(41.144490, -8.605926)	(41.144706, -8.606322)
1372650711620000403	6.73	(41.147271, -8.615736)	(41.147298, -8.615808)
1372661911620000496	10.98	(41.236668, -8.669259)	(41.236578, -8.669205)
1372664679620000186	13.77	(41.162499, -8.615457)	(41.162382, -8.615511)
1372664787620000397	17.79	(41.150979, -8.620074)	(41.151015, -8.619867)
1372662261620000031	2.51	(41.157585, -8.682174)	(41.157603, -8.682156)
1372667871620000152	18.52	(41.160339, -8.609616)	(41.160303, -8.609400)
1372668421620000093	34.94	(41.148585, -8.585631)	(41.148783, -8.585955)
1372669656620000184	13.79	(41.148585, -8.585730)	(41.148693, -8.585811)
1372667532620000443	4.27	(41.161392, -8.598087)	(41.161410, -8.598132)
1372670071620000395	16.17	(41.154255, -8.649261)	(41.154399, -8.649288)
1372671524620000349	18.03	(41.149917, -8.621001)	(41.150079, -8.620992)
1372669341620000388	0.75	(41.161995, -8.587854)	(41.161995, -8.587845)
1372670376620000579	2.14	(41.146758, -8.620011)	(41.146740, -8.620002)
1372669944620000398	5.50	(41.156262, -8.653149)	(41.156298, -8.653194)
1372663272620000206	12.97	(41.150295, -8.602101)	(41.150313, -8.602254)
1372669439620000460	21.14	(41.148630, -8.585658)	(41.148756, -8.585847)
1372673017620000432	2.14	(41.152275, -8.665074)	(41.152293, -8.665083)

11.For each taxi, compute the average idle time between consecutive trips. List the top 20 taxis with the highest average idle time

Table 6: Top 20 taxis with the highest average idle time between trips based on the clean dataset.

Taxi ID	Avg Idle Time (hours)	Idle Periods	Total Trips
20000941	1746.61	3	4
20000969	307.71	22	23
20000510	13.22	621	629
20000312	13.20	577	580
20000609	12.09	550	553
20000579	11.47	740	743
20000079	9.90	95	96
20000072	9.53	896	899
20000185	9.07	663	667
20000449	8.88	958	961
20000902	7.71	1019	1030
20000170	6.73	6	7
20000535	6.65	1273	1274
20000315	6.50	1297	1306
20000225	6.32	1339	1344
20000071	6.26	1352	1355
20000545	6.00	1226	1230
20000407	5.65	1437	1440
20000205	5.58	1315	1319
20000443	5.44	1544	1547

3 Discussion

3.1 Analysis and Cleaning

This section discusses the data cleaning and preprocessing steps performed prior to the queries in Section 2.2. The focus is on identifying and handling missing data, outliers, and irregular trip durations to ensure the reliability and consistency of the dataset. Furthermore, the section examines how these operations affected the overall data distribution.

Based on the row name `MISSING_DATA` and the fact that there were only 10 rows where this value was equal to "true", these rows were removed. This was done to remove incomplete data points. One common problem with doing this is that the rest of the data might be skewed, but since these data points were $\sim 0.0\%$ of the data, this was not seen as an issue (Kang, 2013).

The data was further normalized in the cleaning by removing outliers. Outliers are values that deviate significantly from the rest of the dataset, and they can skew the median and average in the set. There are two main techniques for handling this; either filling in an average value or deleting the data in question (Syam, 2024). In the Figure 2 showing the number of registered trips each day, two clear extremes were identified. The highest at 7,493 and the smallest at 2,185 trips per day. These were seen as outliers by the team due to them being so far away from the average at 4,686.77 trips per day. Therefore, they were removed so as not to distort the data findings in the queries performed later. The team considered filling in these values with the average value but choose not to, due to them only representing 0.57% of the dataset. This is something the team would reconsider redoing if the results are unexpected.

The number of trips was sorted per hour in Figure 3. The extremes were found to occur at 9 a.m. and 2 a.m. This was not surprising to the team, as we imagined that people are typically on their way to work around 9 a.m., and at 2 a.m. there are few people awake and therefore fewer people in need of taxis. Since these values did not deviate significantly from the surrounding data, the team did not classify them as outliers or apply further normalization.

When modeling the distribution of trip duration, the team first analyzed all data without applying a 95% cap. This resulted in Figure 4, where several data points exceeded 1,600 s. As this graph provided little insight into the overall distribution of trip durations, an additional visualization was created with a 95% cap, shown in Figure 5. The 95th percentile was used to exclude extreme outliers and obtain a more representative view of typical trip durations, allowing clearer visualization of general patterns (SigNoz, 2024). Here, a clear pattern resembling a normal distribution can be observed.

One of the largest bin in the histogram (Figure 5) was found to be Bin 1, containing all trips between 0.0 s and 15.3 s. The team interpreted this as an error in the dataset or in the measuring equipment, as it is unlikely that a taxi trip would last that short. For the same reason, Bin 2, containing trips between 15.3 s and 30.6 s, was also removed.

Removing these data points entailed discarding approximately 2.82% of the dataset, which is substantial. However, the team considered these points to be measurement errors and therefore found their removal justified. After removing these values, the distribution of trip duration, capped at the 95th percentile (Figure 7), exhibited a clear normal distribution curve. This was not surprising, as most trip durations are expected to follow a near-normal distribution when extreme values are excluded, since the majority of taxi rides tend to occur within similar time frames determined by typical urban traffic conditions, average travel distances, and consistent service demand patterns. Since a clear normal distribution pattern appeared after this first bin removal, the team did not see any reason to address the other vertex in the graph, as a vertex between 765 s and 780.3 s seems reasonable.

When comparing the call type distribution before cleaning (Figure 1) and after cleaning (Figure 9), the overall distribution appears to remain largely unchanged. This indicates that the cleaning process did not significantly alter the proportions of the different call types, suggesting that data preprocessing preserved the original structure of the dataset.

3.2 Queries

For this section, we have chosen to discuss only the results that stood out to us, as many of the results from the queries were as expected.

The results of the queries gave us insight into the data. For example, the average number of trips per taxi was found to be 3,748 in total, which is about an average of 10.5 trips per day. This is in line with the finding that the taxi with the most driving hours had 1,975.22 in total, which averages out to about 5.5 hours per day. None of this information is surprising, as we imagine that many taxis might only operate part of the day.

Call type B was the most common type, as shown in Figure 9. However, based on Query 4b), there is no clear time period during which this call type dominates. This is surprising, but not impossible.

In Query 7, no invalid trips were found. This was because all invalid trips had already been removed beforehand during the cleaning process.

In Query 8, we used GitHub Copilot to optimize its execution in order to run the query within a reasonable timeframe. We asked it to implement concurrency through parallelized threads to reduce computation time, as well as to create functions for saving progress during execution to prevent data loss in case of a power outage. These optimizations made it possible to complete the query and obtain results after approximately five hours. We also explored potential database optimizations but found no obvious critical flaws in our design that could be improved upon to gain a significant performance increase.

In Query 8, the total number of pairs of cars within 5 meters and 5 seconds of each other was found to be 94,158, and all pairs are listed in the `query8_final_results.json` file. In reality, this number might actually be higher due to the removal of two days from the dataset during the cleaning process when analyzing the number of trips per day. Furthermore, taxi trips shorter than 30 seconds were also removed. The latter cleaning step may be particularly relevant, as many taxi trips could originate from the same taxi stand (for trips of type B) or be dispatched simultaneously from the central dispatch (for trips of type A). In future work, the team should consider the potential impact of data cleaning steps on subsequent queries more carefully, as the removed days might represent meaningful patterns in the data rather than mere outliers.

For Query 9, the number of midnight crossers was found to be 8228 trips. However, similar to Query 8, this number might in reality be higher due to the cleaning process performed beforehand. In this case, the removal of the two days is particularly important, as many trips could have started or ended during those days. This represents an unintended side effect of cleaning the data in this manner. The team should have been more conscious of the fact that the dataset was being prepared for queries rather than for analytical purposes.

In Query 10, the total number of round trips was found to be 19,527. Out of a total of 1,652,987 trips (after cleaning), this corresponds to approximately 1.2% of all trips. It is therefore unlikely, though not impossible, that the data cleaning process had a significant impact on this result. In the absence of prior information about how taxis in Porto operate — particularly regarding taxi sharing practices (similar to Uber) and other operational factors, this proportion appears reasonable. The team thus considers these findings to be both plausible and consistent with expectations.

In Query 11, illustrated in Table 6, the two taxis at the top of the ranking exhibit long average idle times between trips. However, the idle time appears to level out further down the list, suggesting a pattern of decreasing variation among the remaining taxis. This observation cannot be confirmed from this table alone, but it may indicate a smoothing trend that could be further investigated through additional analysis. The differences in idle time could also be explained by varying work patterns, with some drivers working part-time and others full-time, though this remains only a hypothesis.

Bibliography

Kang, H. (2013). The prevention and handling of missing data. *NIH: National Library of Medicine*, *64*, 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

SigNoz. (2024). *How to calculate 95th percentile of an average in prometheus* [Accessed: 2025-10-09]. https://signoz.io/guides/how-to-get-the-95th-percentile-of-an-average-in-prometheus/

Syam, S. S. (2024). *Understanding and handling outliers in data analysis*. https://medium.com/@heysan/understanding-and-handling-outliers-in-data-analysis-727a768650fe