Porto Taxi Trajectory

MIS381N

December 2019

Gaurav Choudhary

LaShay Fontenot

Chris Henson

Shivank Sood

Vikrant Vaidya

Data

The data provides information about taxi trips throughout the city of interest

1 Year

July 2013 - June 2014

442 Taxis

in Porto, Portugal

1.7M Trips

throughout the city

9 Features

including Polyline, timestamp, call type

This Kaggle dataset provides nearly 1.7M observations of taxi trips in Porto, Portugal.

Primary feature is latitude/longitude given every 15 seconds **for a total of 83 million coordinate pairs**

Additional features include taxi and trip id, timestamp, call type

Engineered features will include trip duration, network analysis, etc.

Overview

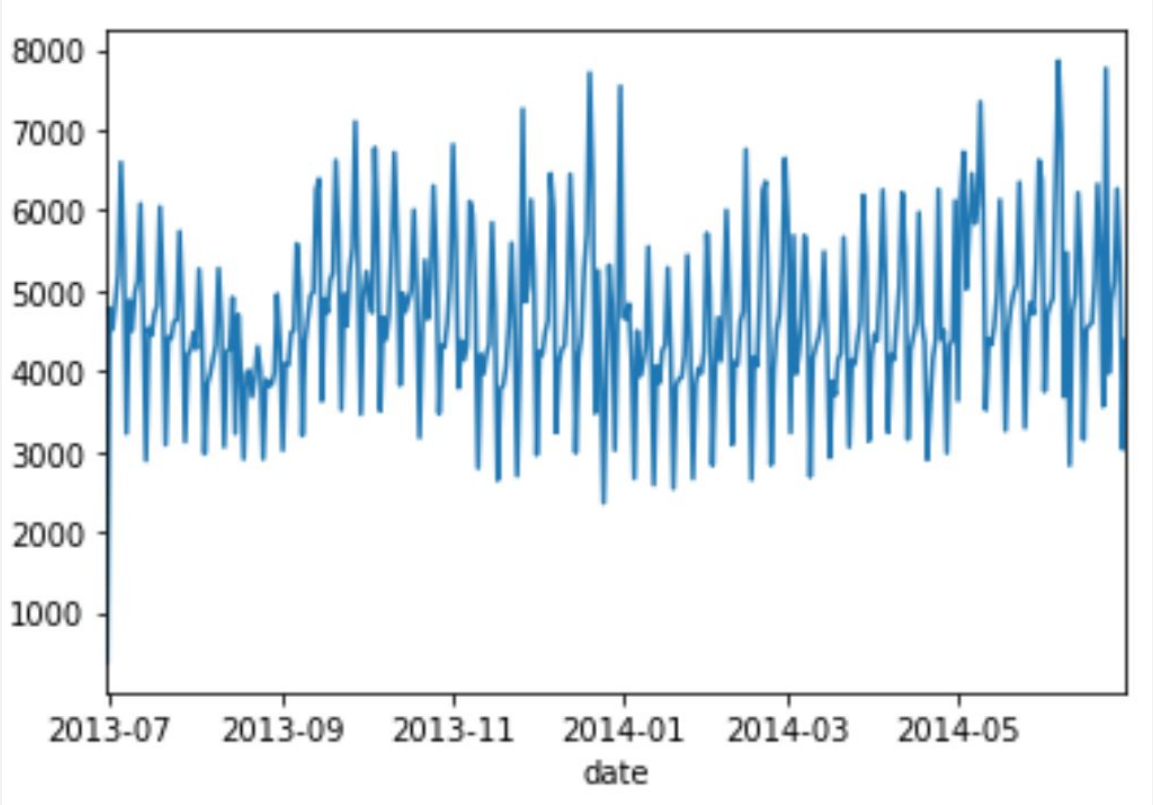
Modeling Approach & Results

Heat & Interactive Mapping Demos

Exploratory Analysis

Popular Days (No. of Trips)

Date	Number of Trips
2014-06-06	7879
2014-06-23	7775
2013-12-20	7744
2013-12-31	7582
2014-05-09	7366
2013-11-26	7277
2013-09-27	7123
2014-06-07	6971
2013-11-01	6840
2013-10-04	6789
2014-02-14	6774
2014-05-03	6751
2013-10-11	6738
2014-02-28	6665

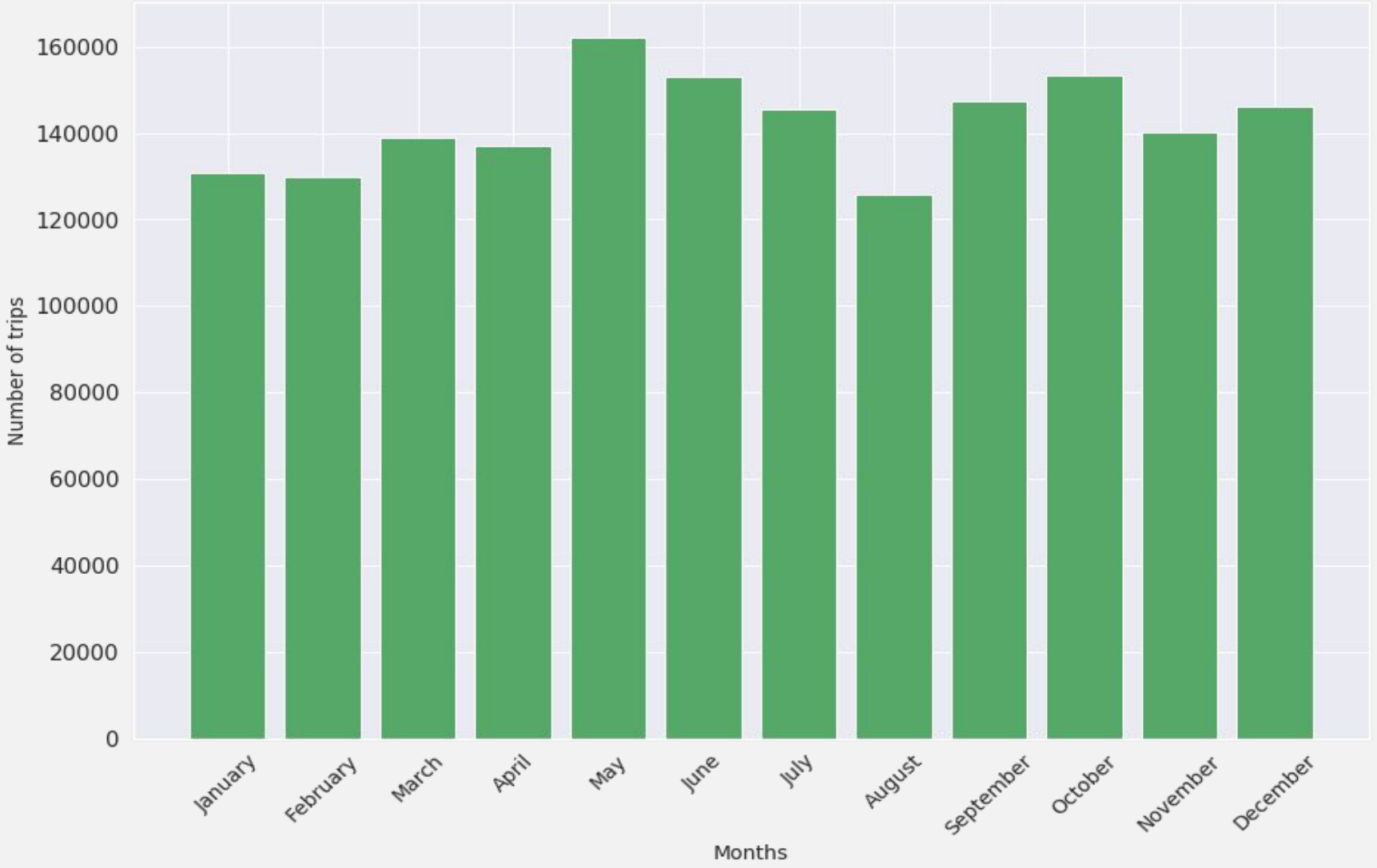


All Saints Day

New Year's Eve

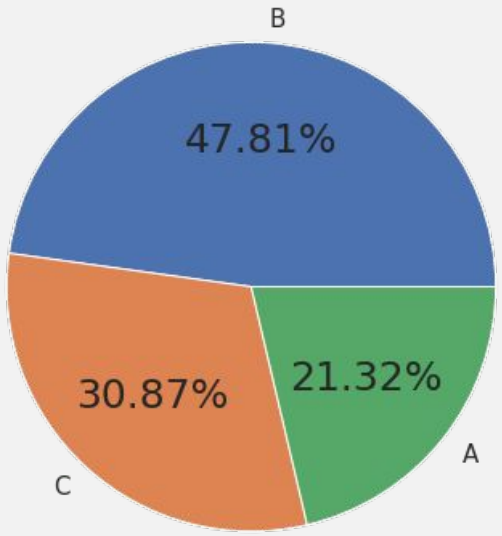
Summer months

Trips by Month

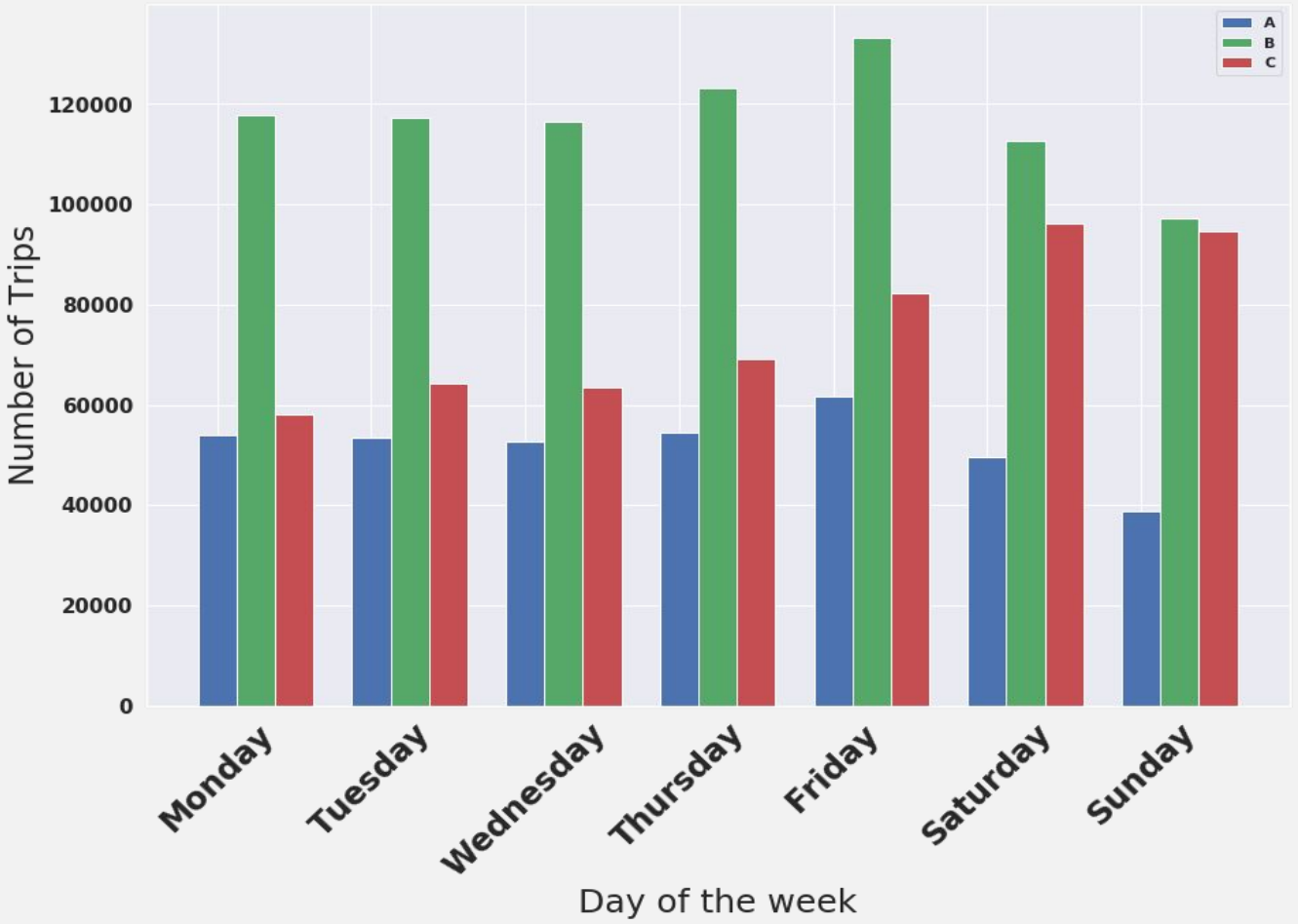


- We see spikes in trips during periods of high tourism
- Colder months see drop off in trips
- August : typical vacation month for locals

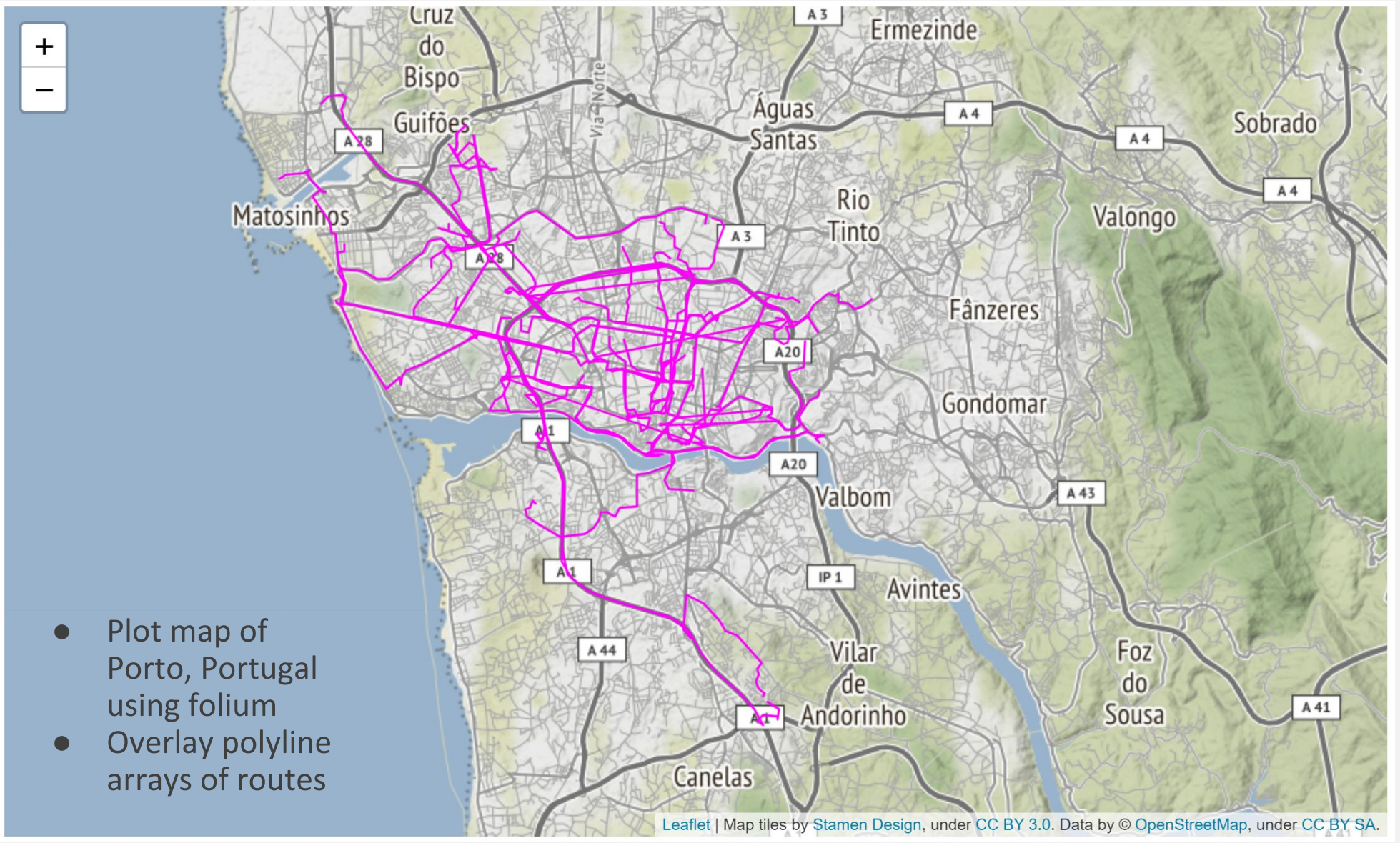
Trip Types



A. Trips dispatched from the central
B. Trips booked directly to a taxi driver on a specific stand
C. Trips booked on a random street



Interactive Route Mapping



- Plot map of Porto, Portugal using folium
- Overlay polyline arrays of routes

Route Mapping : Driver Drill Down

Using an interactive map and aggregated data, we can take a closer look at ‘performance’ of particular taxi drivers

Points of Interest

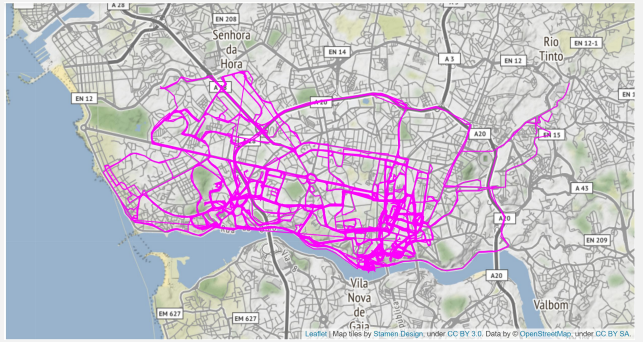
- What routes is the driver taking during his/her longest trips (in terms of duration)?
- Where is the driver making most of his/her pickups and or drop offs?

- What drivers are spending the most time on the road?
  - Are these drivers going to similar locations? i.e. the Airport, famous landmarks, wineries
- What day of the week proves problematic for our drivers? Does this vary among drivers?

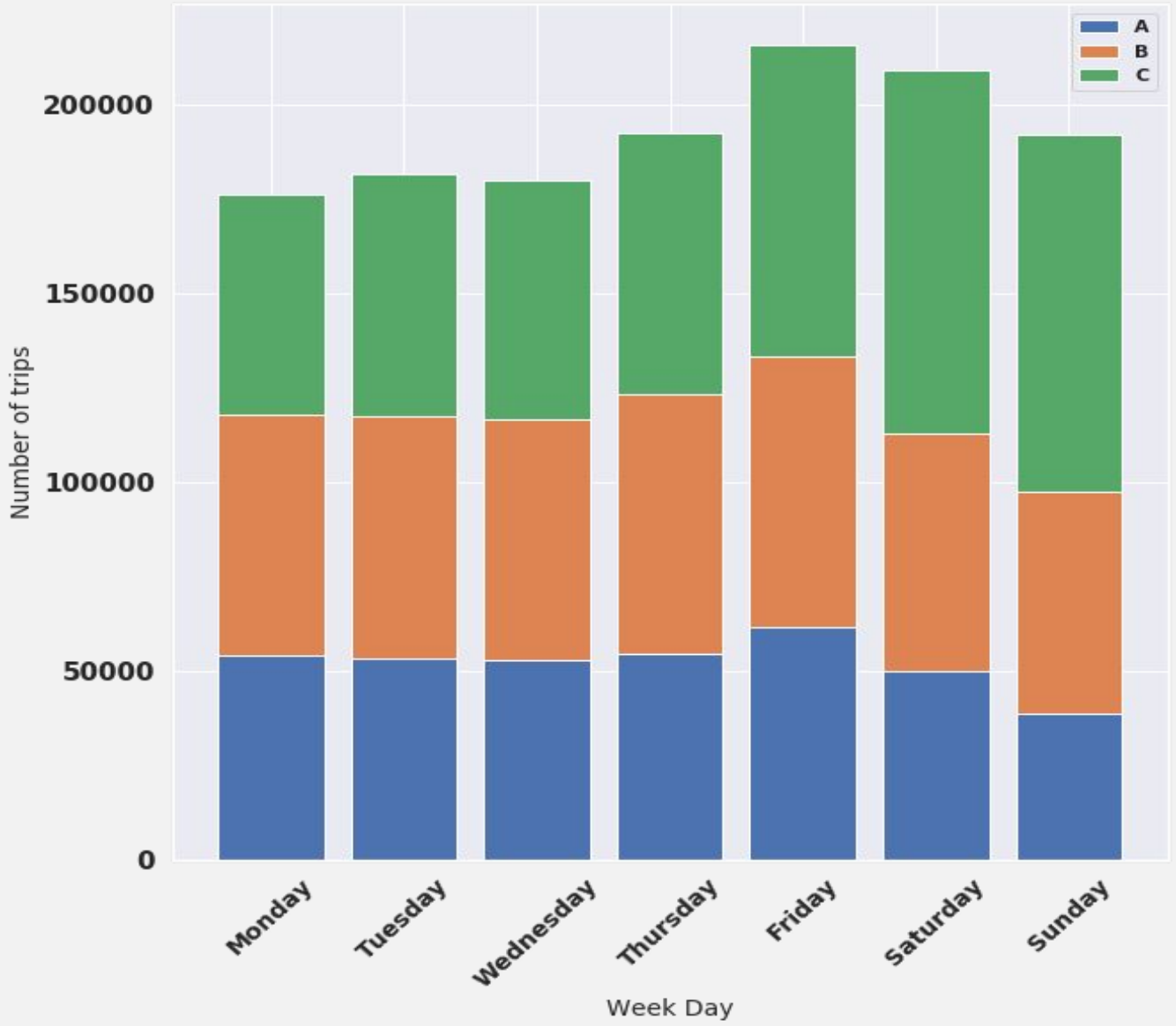
Driver Profile

Taxi ID
20000205

- Average trip Duration : 53 minutes
- Routes concentrated near the water and through the city
- A few trips out to Rio Tinto area
- Busiest on Fridays (292 trips) ; highest avg. duration
- Only 5 Saturday trips

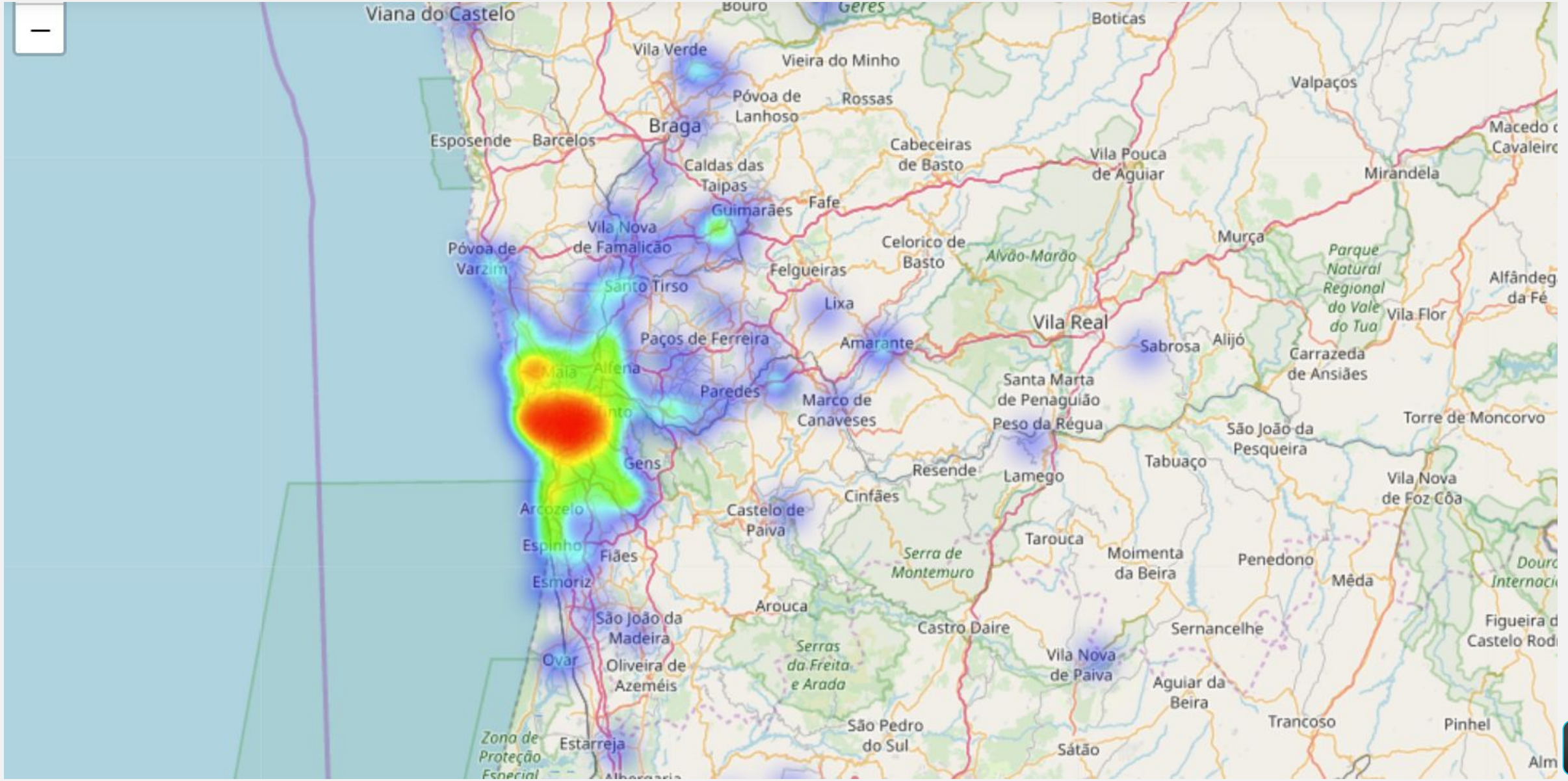


Trips by Day of the Week

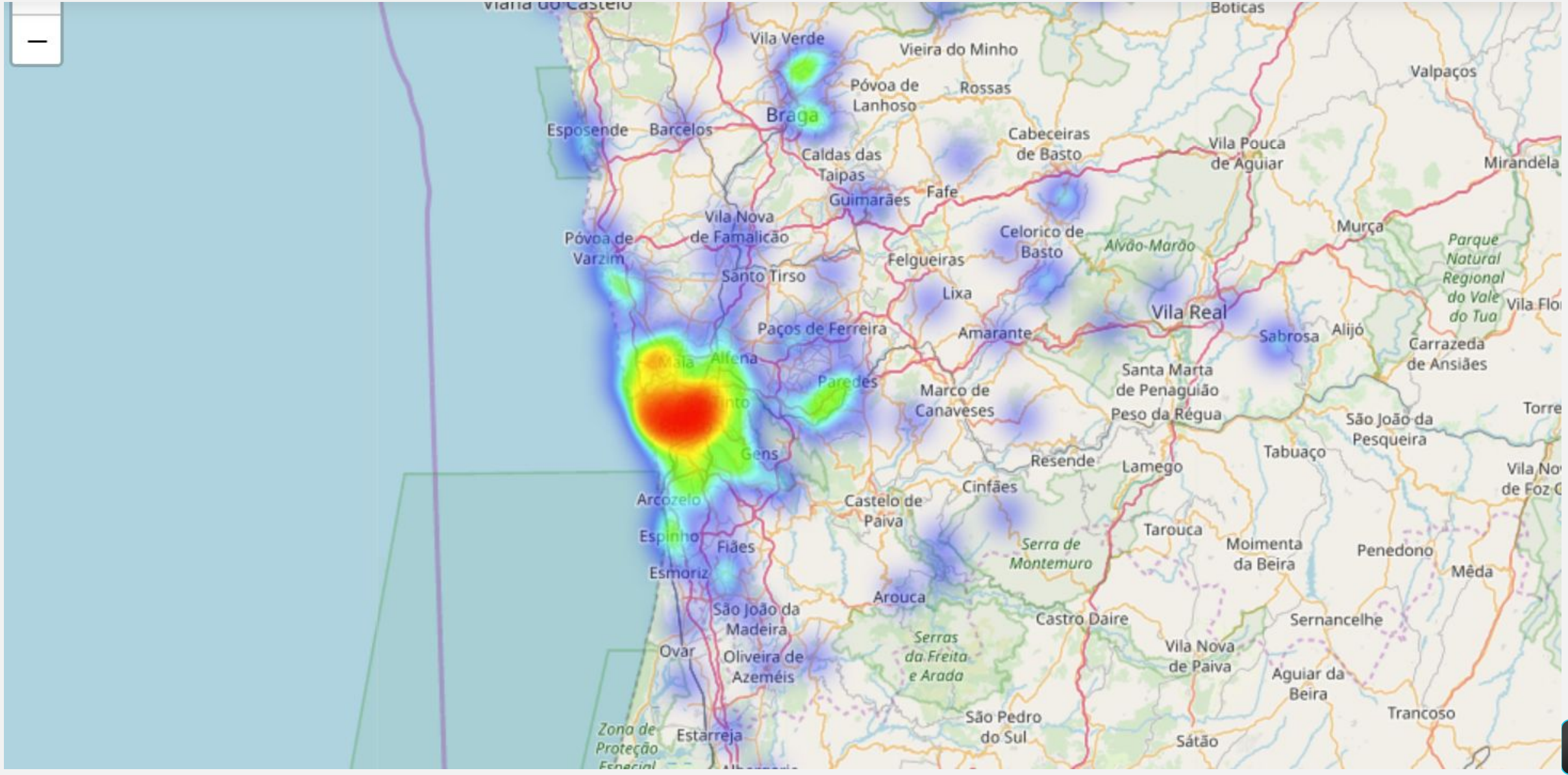


- As expected, trips spike during the weekend (Friday and Saturday)
- Monday sees the lowest number of trips

Heatmap of Trips : Monday



Heatmap of Trips : Sunday



Modeling : Causes of Trip Duration

- Sensor data is provided, every fifteen seconds, so given the entire route we can derive duration
- At the point the taxi is hailed however, we have only a rough estimate of what areas they will travel through.
- The question is, how much of duration can be predicted solely from the geographic information, independent of time of day, holidays, etc.?

Approach : Causes of Trip Duration

- Divide the city of Porto into a rectangular grid of latitude/longitude, using a large bin size (one million)

- Restrict our bin consideration to those that are frequently traveled (threshold of 30,000 taxis traveling the area in a year timeframe), reducing our considered bins to approximately 1000

- For each of these 1000 bins, create a binary variable that marks if that route passed through the corresponding bin

	duration	x_480648	x_480993	x_480709	x_480734	x_480743	x_480984	x_480752	x_480726	x_481079	...	y_314985
0	5.5	0	0	0	0	0	0	0	0	0	...	0
1	4.5	0	0	0	0	0	0	0	0	0	...	0
2	16.0	0	0	0	0	0	1	1	0	1	...	0
3	10.5	0	0	0	0	0	0	0	0	0	...	0
4	7.0	0	0	0	0	0	0	0	0	0	...	0

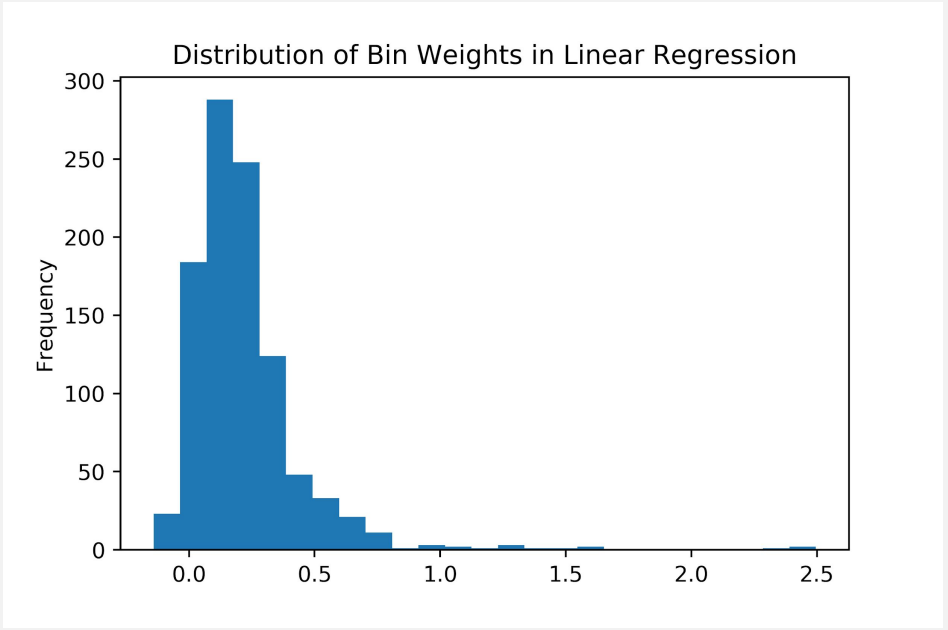
Results : Causes of Trip Duration

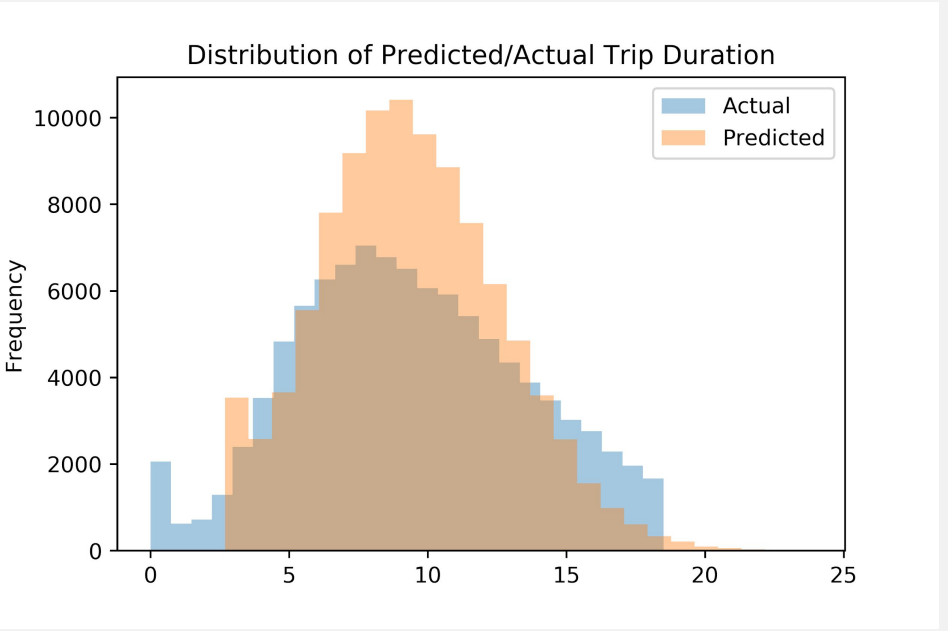
- Simple linear regression on these binary bins accounts for 61.3% of the variance in trip duration
- Comparing actual/predicted duration, we see that our model has a much lower standard deviation. We can (very carefully!) use this to try to attribute variance in individual trip duration to outside influences
- We can use the coefficients of this regression to identify areas of the city that are most likely to add to trip duration (independent of any other conditions)

```
1  coef_df[coef_df.coef >  .5]
```

	coef	bin
220	0.640921	x_481630
241	0.647587	x_481622
303	0.700338	x_482101
317	0.654878	x_481639
333	0.609584	x_481578
...	...	...
984	0.640267	y_315324
986	0.681915	y_314784
989	0.554545	y_314802
994	0.648178	y_314713
995	0.721629	y_314760

75 rows × 2 columns





Conclusion & Applications

City Planning

The model can be used in city / infrastructure planning by identifying areas that are most likely to add to trip duration

Taxi Scheduling

The data used in this analysis can help transportation companies plan driver schedules throughout the year and choosing hub focal points

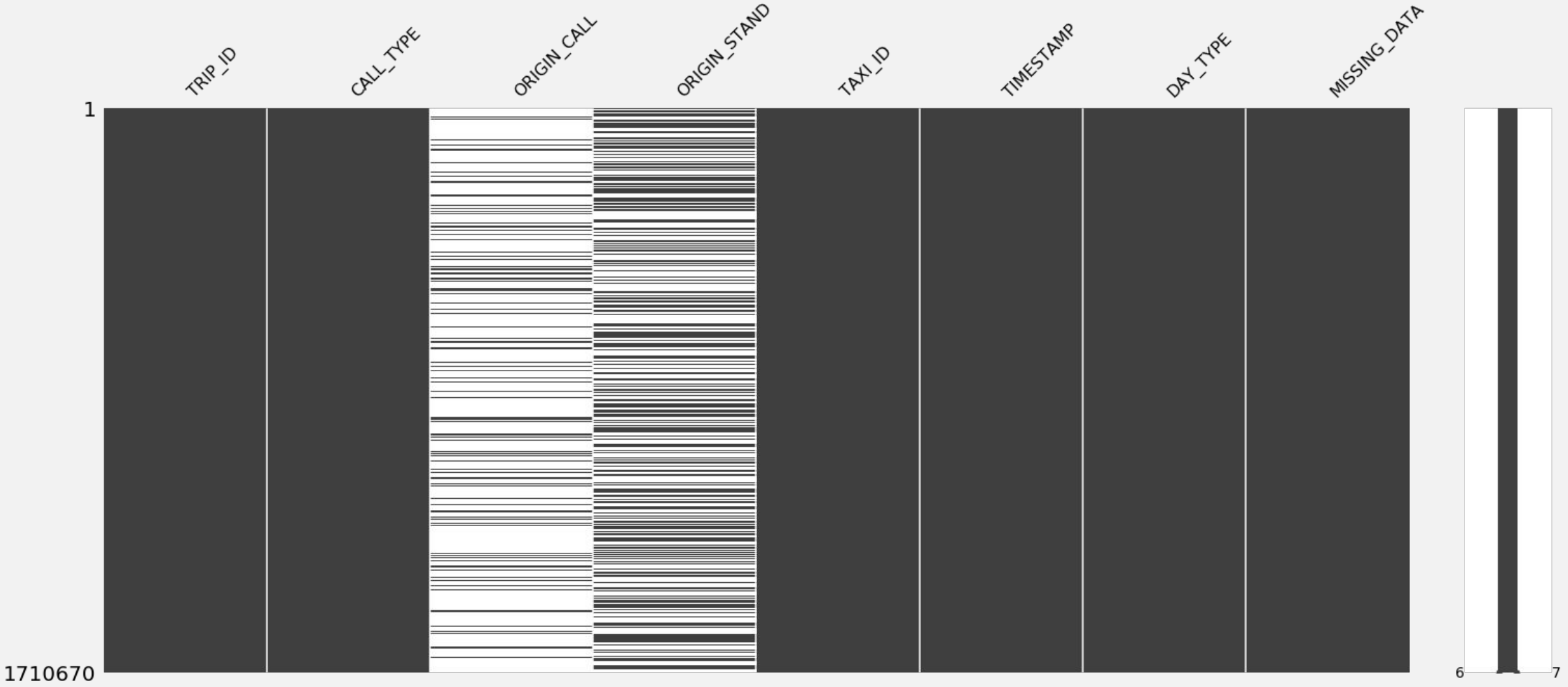
Tourism Insights

Paired with other economic / tourism data, this data can be used to generate insights for the Porto tourism board

Questions?

(link to Github)

MISSING VALUES



Popular Days (No. of Trips)

All Saints Day

New Year's Eve

Summer months

Date	Number of Trips
2014-06-06	7879
2014-06-23	7775
2013-12-20	7744
2013-12-31	7582
2014-05-09	7366
2013-11-26	7277
2013-09-27	7123
2014-06-07	6971
2013-11-01	6840
2013-10-04	6789
2014-02-14	6774
2014-05-03	6751
2013-10-11	6738
2014-02-28	6665

Trip Types by Day of the Week