Exploratory Data Analysis of Taxi Demand Data

Zilong Wang[a,1] **and Athanasios Lolos**[a,1]

[a]Georgia Institute of Technology

This manuscript was compiled on December 10, 2020

We analyse the Taxi Service Trajectory dataset (1) in the UCI Machine Learning Repository, which was used in a prediction challenge in ECML KDD 2015, and attempt a more in depth exploratory data analysis (EDA) stage. Chief amongst these methods used are the L2 Wasserstein metric along with k-means to cluster the taxi stands by their distribution of service times and service inter-arrival times, along with other classical dimensionality reduction tools such as principal component analysis (PCA). We hope that we will get a better feel of the exact nature of the dataset before attempting to replicate a portion of the paper by (2) in order to verify their assumptions with regards to the data generating process.

Histograms | Principal Component Analysis | Wasserstein Distance | k-means | Stratification Analysis | Linear Regression

This is the technical report for Fall 2020 ISyE 7405's final project using the style of PNAS's template. The following sections are organized as follows: Section 2 will contain a description of the dataset features, Section 3 contains details on our Data-Preprocessing, Section 4 and 5 contains details on our findings during the exploratory analysis and explanatory findings using linear regression. Finally section 6 and 7 attempts to strengthen the findings above with clustering analysis, before running in clustering regression. We conclude with sections 8 and 9 with some final remarks regarding our endeavor so far and some future work that can be done towards this direction.

Description of Dataset

For this project we will use an accurate dataset that describes the busy trajectories (the trajectories when a customer uses the taxi) performed by all the 448 taxis running in the city of Porto, in Portugal for a whole year (from 07/01/2013 to 06/30/2014). These taxis operate through a taxi dispatch central, using mobile data terminals installed in the vehicles. In total we have 1710670 Data Points and 63 Taxi Stand Locations.

Dataset Headers. Each data sample corresponds to one completed trip and contains a total of nine features that are described below:

1 **TRIP_ID**:(String) It contains a unique identifier for each trip.

2 **CALL_TYPE**: (Char) It identifies the way used to demand this service. It may contain one of three possible values.

   i **'A'** if this trip was dispatched from the central.

   ii **'B'** if this trip was demanded directly to a taxi driver at a specific stand

   iii **'C'** otherwise (e.g. a trip demanded on a random street).

3 **ORIGIN_CALL**: (integer) It contains a unique identifier for each phone number which was used to demand at least one service. It identifies the trip's customer if CALL_TYPE= 'A'. Otherwise, it assumes a NULL value.

4 **ORIGIN_STAND**: (integer) It contains a unique identifier for the taxi stand. It identifies the starting point of the trip if CALL_TYPE= 'B'. Otherwise, it assumes a NULL value.

5 **TAXI_ID**: (integer) It contains a unique identifier for the taxi driver that performed each trip.

6 **TIMESTAMP**: (integer) Unix Timestamp (in seconds). It identifies the trip's start.

7 **DAYTYPE**: (char) It identifies the daytype of the trip's start. It assumes one of three possible values.

   i **'B'** if the trip started on a holiday or any other special day (i.e. extending holidays, floating holidays, etc.)

   ii **'C'** if the trip started on a day before a type-B day

   iii **'A'** otherwise (i.e. a normal day, workday or weekend)

8 **MISSING_DATA**: (Boolean) It is FALSE when the GPS data stream is complete and TRUE whenever one (or more) locations are missing.

9 **POLYLINE**: (string) It contains a list of GPS coordinates (i.e. WGS84 format) mapped as a string. The beginning and the end of the string are identified with brackets. In addition, each pair of coordinates is also identified by brackets [LONGITUDE, LATITUDE]. The list contains one pair of coordinates for each 15 seconds of trip. The last list item corresponds to the trip's destination, while the first one represents its start.

Other Details. From the CALL_TYPE feature we see that essentially there are three main ways to pick-up a passenger. The first way is that a passenger goes to a taxi stand stand

Significance Statement

Public datasets are now commonly reused for different works once introduced, it is therefore important to curate and properly document them with extensive Exploratory Data Analysis for future referencing and reproducibility.

Please provide details of author contributions here.

No conflicts of interest

[1]A.O.(Zilong Wang) contributed equally to this work with A.T. (Athanasios Lolos)

[2]To whom correspondence should be addressed. E-mail: zwang937gatech.edu

and picks-up a taxi (CALL_TYPE= ‘B’). The second way is that the passenger calls the taxi network central and demands a taxi for specific location/time (CALL_TYPE= ‘A’) and finally a passenger may pick a vacant taxi on any street at random (CALL_TYPE= ‘C’).

Data Pre-Processing

From the nine features of each data point we can extract important information for the trips (e.g. Total Trip Time), for the Taxi Stand Locations and the interarrival times of the customers of each Taxi Stand Location.

Calculation of Total Trip Time. In order to better understand our data set, we used the feature "POLYLINE" of each data point, in order to calculate the Total Trip Time (i.e. the duration of each trip). Since, in the feature "POLYLINE" we have a list of GPS coordinates, that contains one pair of coordinates for each 15 seconds of trip, we can calculate the duration of the trip in minutes by multiplying the total number of pairs by 15 and dividing what we find by 60. In Figure 1, we provide the Frequency Distribution of the Total Trip Times. NOTE: All the trips with duration ≥ 60 are binned together, which can be clearly seen in Figure 1.

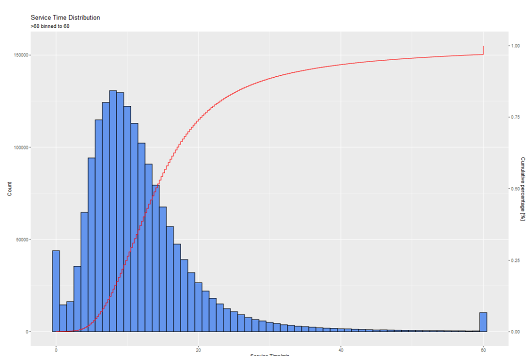


Fig. 1. Frequency Distribution of Total Trip Times (in minutes)

Data Selection. We excluded from our analysis all the data points, whose "MISSING_DATA" feature is TRUE. Additionally from Figure 1, we observe that there is a significant number of trips, whose duration is less that 3 minutes, which seems odd. After careful examination of the data we concluded that in most of these cases, probably there was a problem with the tracking. For example, it is not reasonable to have many trips with duration less than 30 seconds. In order to avoid including data, whose features may be misleading, we decided to discard all trips with Total Trip Time ≤ 3 minutes. After following this procedure, we excluded 4.8% of the data, i.e. we excluded 81259 out of the 1710670 data points, and we continued our Data Analysis with the remaining 1629411 data points. Another matter that we considered was whether or not we should discard trips with duration ≥ 60 minutes. However, we concluded that we should not discard these trips as after careful examination of the data points, it does not appear to be erroneous tracking. It is important to note here that when we study duration of trips, we expect to have some trips with very large duration. Our choice to not discard these trips is strengthened from Figure 2, which shows that we have a long gentle slopped tail in the histogram.

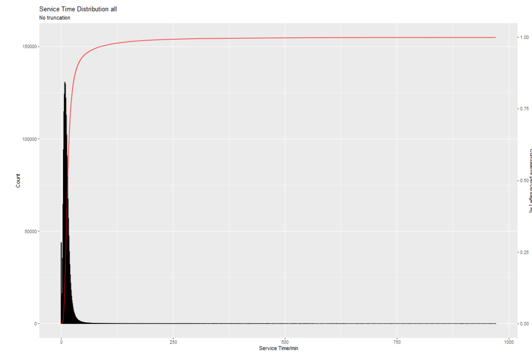


Fig. 2. Frequency Distribution of Total Trip Times-not binned (in minutes)

Starting Time of the Trip. The feature "TIMESTAMP" contains all the information that we need for the starting time of the trip, however, this information is given in Unix Timestamp (in seconds). Thus, in order to use this information, we converted them to date and time, considering the time zone of the city of Porto.

Calculation of Additional Features. Here we describe besides the Total Trip Time, what other features we calculated for our analysis from the initial nine features of each data point. We found the origin of each trip (for the CALL_TYPE="B" trips this is equivalent to finding the initial Taxi Stand Location) by using the first pair of coordinates of the feature "POLYLINE". In addition, from the feature "POLYLINE" we calculated the Euclidean Distance for each trip by using the first and the last pair of coordinates. Moreover, we were able to stratify the trips in different groups according to the analysis we wanted to perform, with respect to their starting time. Specifically, we created the four boolean vectors "EARLYSHIFT" (0am-8am), "MIDSHIFT" (8am-4pm), "LATESHIFT" (4pm-0am) and "WEEKEND", and for our final regression models, where we did further stratification, we considered 24 boolean vectors, one for each hour of the day. Finally, we calculated the interarrival times of the different Taxi Stand Locations, by subtracting the starting times of successive trips in each Taxi Stand Location.

Exploratory Data Analysis

Summary Tables. Once the data has been pre-processed, we proceeded with stratification analysis. As shown in Table 1, we

		AVERAGE		
	TotalServices	EARLY	MID	LATE
Workdays	1206390	557	1270	929
Weekends	423021	387	289	319
AllDaytypes	1629411	929	1556	1235

Table 1. TAXI SERVICES VOLUME (AVERAGED PER DRIVER)

stratified the number of services (trips) by Early $[0000, 0800)$, Mid $[0800, 1600)$, and Late $[1600, 0000)$ shifts according to the time of the day and also by whether it was on a weekday or weekend.
Likewise, we also present the stratification results for the services and time taken per service (cruise times) per driver. As shown in Table 2, there is also considerable variance between

	ServicesPerDriver	TotalCruiseTime
Max	7468	118656
Min	2	20
Mean	3695	47101
SD	1462	18035

Table 2. TAXI SERVICES VOLUME(PER DRIVER/CRUISE TIME)

individual drivers with regards to the number of services and the duration of each service. Both of these are in the ballpark range similar to those obtained by (2)

Geo-spatial Distribution of Service Times and Interarrival Times. We now proceed to examine the distributions of the service times at each station. In the field of operations and revenue management, it is very common in vehicle routing and server applications to assume that both the rate of incoming customers and service times are modelled as Poisson processes. This allows practitioners to use common modeling techniques such as queuing analysis and Poisson regression. Therefore, the next step is to empirically examine and verify that such assumptions hold. Figure 3 shows the locations of the taxi stands across the city of Porto, Portugal and from initial observations, we can see that it is highly concentrated in the south-eastern region. This might indicate that the majority of the customers originate there, and as we will see in a later section, the data shows that this is indeed the case.

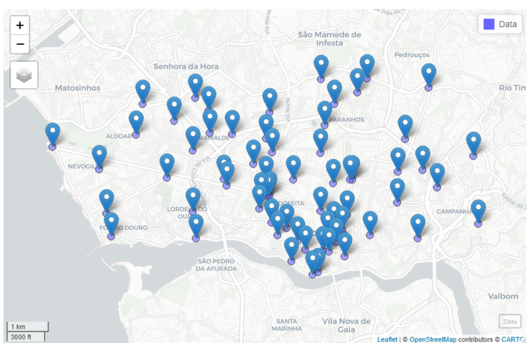


Fig. 3. Location of taxi stands in Porto, Portugal

Service Time Distributions. As shown in Figure 4, we can see that the average trip times at each station are rather similar. As such, it warrants further investigation as to the exact distribution. For the sake of brevity, we will display only a subsection of the histograms that show the most common distributions as seen in Figure 5 and they indeed look like they take on Exponential Distributions.

Inter-arrival Time Distributions. We now examine the distributions of inter-arrival times between services. This requires more careful explanation and examination of the mechanics. It must be noted that a **trip** begins if and only if there is a passenger and taxi present at the same stand. This may not be exponentially distributed as **there are now potential endogeneity concerns**. For example, the inter-arrival times of taxis to a station might depend on the number of customers present (more customers waiting implies abundance of service opportunities, leading to possibly shortened taxi inter-arrival times). As such, we should examine the inter-arrival times carefully at each station.

To do this, we first examine the total demand per station.

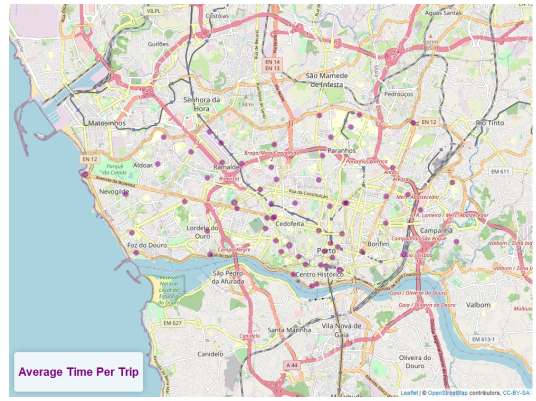


Fig. 4. Average Trip Times (Radii)

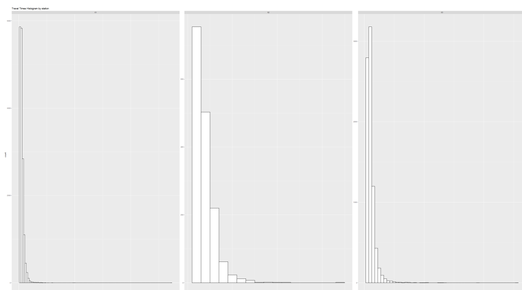


Fig. 5. Representative Histograms of Service Time Distributions

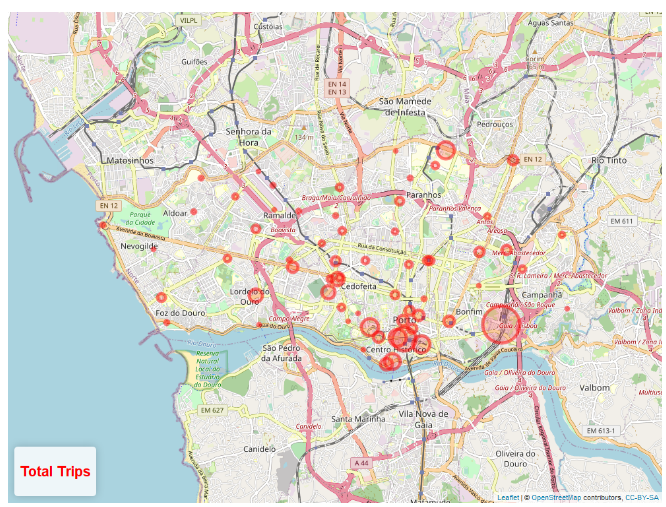


Fig. 6. Total Trips of each station (Radii)

As we have guessed, as shown in the radii of the circles in Figure 6 the south eastern taxi stands have a larger proportion of trips as indicated by their larger radii. This therefore implies the converse in terms of interarrival time between trips: we should expect to see larger average gaps between trips for stations with fewer trips. This is confirmed in Figure 7.

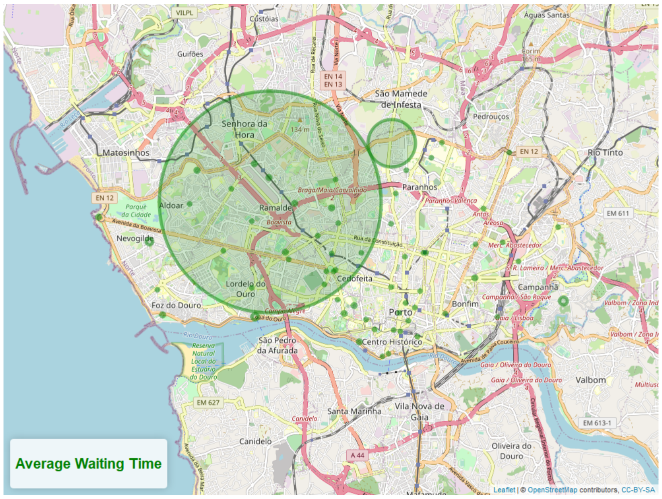


Fig. 7. Average Inter-arrival times of services for each station (Radii). Note the massive outliers station 5 and station 48 (2nd largest and largest circles)

The incredibly large outliers of station 5 and station 48 (the 2nd largest and largest circles respectively) only had 47 and 6 trips respectively. In Figure 8, we exhibit the most representative histograms

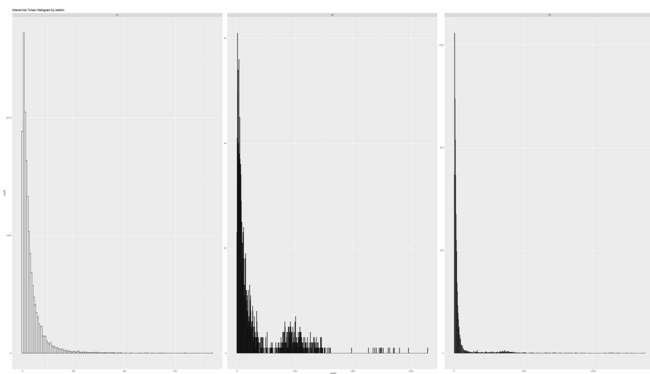


Fig. 8. Representative Histogram of Inter-arrival Times, bimodality may be symptomatic of endogeneous self-selection of taxis to stands that are busier

In general, most of the station's inter-arrival times appear to be exponentially distributed with some looking bimodal, which may indicate difference service periods (maybe some stations, while busy overall, are way less busy at some times of the day) or endogeneity in terms of taxi self-selection towards busier stations as previously suspected. This therefore raises very worrying concerns with regards to modeling taxi services as Poisson processes and using the typical accompanying methods such as queuing analysis and Poisson regression.

Principal Component Analysis. Before we move on to using linear regression to examine the explanatory power of the dataset, it is good to see how much of the variance can be explained. While we note that most of the features are ordinal and as such, would benefit more from Multi-factor Analysis, due to hardware and time restrictions, we were unable to perform it. Nevertheless, PCA can expose some potential issues that may arise when working with the dataset, and it can be clearly seen in the following exhibit

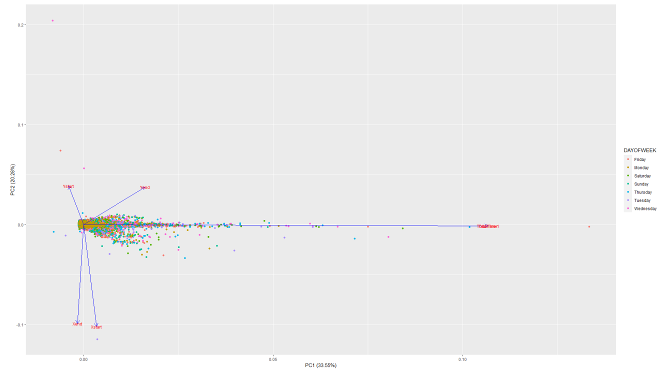


Fig. 9. Principal Component Analysis of all continuous features

Even after standardization of features, the variance of the dataset is dominated by the service time, with the starting locations serving little to explain the variability of the dataset. Indeed, while knowing where the trip originated from might provide some information as to the nature of the trip, there is just too little information available, indicating that there are a lot of other unobserved explanatory variables not present in the dataset. This problem will later be shown in more statistical detail in the analysis of the dataset's variability using standard Ordinary Least Squares.

Linear Regression Analysis

In this section are presented some of the linear models that we used for our Analysis. It is important to note here, that we used linear Regression and ANOVA in order to see how well the data set that we have, explains the Total Trip Time and the Interarrival Waiting Time, i.e. using these models for predictions is not our main purpose. We will provide more information for each model separately. For all the regression models that we will describe below, we consider only the data for which CALL_TYPE="B", i.e. they start from a Taxi Stand Location. Moreover, we standardized the data in the columns "Distance","Xorigin","Yorigin","Xdestination","Ydestination". For our Analysis we use 80% of the data for finding the regression model and 20% for testing. Finally, for each model you can also refer to the R code for the ANOVA table and the RMSE, MAPE and Average Absolute Difference values (the absolute difference is the difference between the real and the predicted values). Due to the length of the tables that we got from the Regression Analysis, we present the Tables for each one of the Regression Models in the Appendix. It is also important to note that we converted the Total Trip Time in minutes for our analysis, but we kept the interarrival waiting times (interarrival times of customers in the same Taxi Stand Location) in seconds, since some waiting times were very small, thus, keeping them in seconds instead of minutes was helpful for our analysis.

Linear Regression Model I. For the first Regression Model, we consider the Response to be the Total Trip Time (T) and we have six Explanatory variables, which are the Distance (D), the Xorigin (X), the Yorigin (Y), the Earlyshift (E), the Midshift (M) and the Weekend (W). The linear Regression

Model that we find is:

$$Y = 12.464 + 3.779D + 0.018X - 0.550Y \\ - 2.513E - 0.297M - 1.145W, \quad [1]$$

where all the variables are considered to be significant at a significance level $\alpha = 0.05$. We include all the important information for this model in the first column of Table 9. For this model, we have $R^2 = 0.229$, $RMSE = 7.4465$ and $MAPE = 0.3714$. We see that R^2 is relatively small, but this is something that we expect and we will try to improve beginning with the next regression model that we consider.

Linear Regression Model II. For this Regression Model, we consider the Response to be again the Total Trip Time and we add to the Explanatory variables that we consider above, the Waiting Time (WT) (interarrival time for the next customer to arrive considering the same Taxi Stand Location). The results for this model can be seen in the second column of Table 9. However, the addition of the Waiting Time to the explanatory variables does not improve Regression Model I and specifically, we found that the Waiting Time is considered insignificant with respect to the Total Trip Time. What we find here is reasonable, since the duration of a trip should be independent from the interarrival time of the next customer. Therefore, this assumption is partially verified here.

Linear Regression Model III. For this Regression Model, we consider the Response to be the Waiting Time (WT) and we have the same six Explanatory variables as in Regression Model I, which are the Distance (D), the Xorigin (X), the Yorigin (Y), the Earlyshift (E), the Midshift (M) and the Weekend (W). The linear Regression Model that we find is:

$$WT = 31.327 - 1.525D + 14.153X + 13.047Y \\ +33.450E + 10.634M + 0.589W. \quad [2]$$

However, at a significance level $\alpha = 0.05$ we find the D (Distance) and W (Weekend) variables to be insignificant (refer also to the ANOVA table in the R code). An analysis can be found in our R code where we refit our model considering only the significant variables. We include all the important information for this model in Table 10. For this model, we have $R^2 = 0.006855$, which despite the fact that it is very small, it is something that we expected and we are looking for ways to improve this model.

Idea for Improvement. In order to improve the Linear Regression Models that we have we will further stratify our dataset based on the exact hour of the day that the trip starts for more sophisticated modelling.

Linear Regression Model IV. For the this Regression Model, we consider the Response to be the Total Trip Time (T) and we have 28 Explanatory variables, which are the Distance (D), the Xorigin (X), the Yorigin (Y), the Xdestination (Xd), the Ydestination (Yd) and the 23 Explanatory Variables that represent the different hour intervals for each day (e.g. 0am-1am,1am-2am,... etc.). Note that we do not need 24 variables for the hours of the day, since if the trip has zero in the first 23 variables then for sure the last variable will be one and if it has an one in any of the first 23 variables, then the last variable will certainly be zero, this is also the reason why we do not consider the "LATESHIFT" as an explanatory variable in the regression models above. The linear Regression Model that we find can be seen in Table 11. We avoid writing down this model analytically due to the space restriction that we have. For this model, we have $R^2 = 0.238$, We see that our model has slightly improved, but the R^2 is still relatively small.

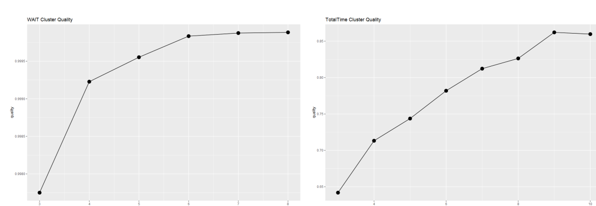
Linear Regression Model V. For the this Regression Model, we consider the Response to be the Waiting Time (WT) and we have the same 28 Explanatory variables as in the Regression Model IV. The linear Regression Model that we find can be seen in Table 12. We avoid writing down this model analytically due to the space restriction that we have. For this model, we have $R^2 = 0.009702$, We see that our model has slightly improved, but the R^2 is still very small.

Remark for Regression Analysis. Despite the fact that with the further stratification of the dataset, our models were slightly improved, we see that the R^2 values are still small, especially for the Waiting Time. We can understand why this is happening by observing the Tables 5, 6, 7 in the Appendix. It is clear that there isn't that much variation between stratas (this is surprising), which explains why the stratification idea did not significantly improve our models. However, this is not a problem in our case as we wanted to use the Regression Analysis in order to see how well our data describe the Total Trip Time and the Waiting Time. However, our analysis implies that if we wanted to consider a linear regression model for predictions we should definitely augment our data with external datasets that may provide more information.

Clustering Analysis

After looking at the histograms of the trip lengths and inter-arrival times of each station, a natural next step is to cluster these stations according to these characteristics, i.e. cluster these stations based on their histograms by the inter-arrival times and travel times separately.

Since histograms are empirical proxies of probability distributions a very natural distance metric to use would be the $L2$ Wasserstein Distance, which would be used with k-means clustering. In order to tune the hyperparameter k (the number of clusters) we opted to observe the quality (sum of square deviations explained by the model) and choose a good cutoff point where the gain in quality is marginal to avoid "over-fitting". We utilized the *HistDAWass* package (3). As seen



(a) Quality Increase per number of k-means clusters using the L2 Wasserstein metric over the distribution of Inter-arrival times of stations. We opted for $k = 6$

(b) Quality Increase per number of k-means clusters using the L2 Wasserstein metric over the distribution of service times of stations. We opted for $k = 7$

in Figure 10a there seems to be a clear distinction between clusters with respect to the inter-arrival times, which is in favour of our hypothesis of some form of self-selection by taxis, and at $k = 6$, the marginal improvement in quality tapers off. As for the clustering based on the distribution of service lengths, Figure 10b's increase in quality is more haphazard, which is to be expected as the service length distributions are less distinct.

With these clusters at the ready, we proceed to rerun the regression within each subset of clusters for the inter-arrival times and service times respectively, with the hope that this will demonstrate that simple stratification and clustering of data can lead to massive improvements even on simple elementary models. We present the exact cluster assignments in the two tables 13, 14 in the Appendix.

Within Cluster Regression Analysis

In this section we present our results from running Regression Models separately within clusters in order to improve the explanatory power of our models. For our Analysis we run Regression Models within each cluster according to the Tables 13 and 14 in the Appendix. First, we consider the Response to be the Total Trip Time and the Explanatory Variables to be those used for Regression Model IV (28 in total). We run this regression model seven times one for each cluster, considering the respective data points every time. The results for the R^2 and the total number of observations in each cluster are shown in Table 3. For further information on these models you can refer to the R code.

Cluster	R^2	Total Number of Observations
Cluster 1	0.2463	31524
Cluster 2	0.1787	345976
Cluster 3	0.02434	2803
Cluster 4	0.2702	62439
Cluster 5	0.3024	128282
Cluster 6	0.02131	840
Cluster 7	0.1674	211506

Table 3. Regression Models within Clusters for Total Trip Time

In Table 4 are shown the results when we consider the Response to be the Interarrival Waiting Time and the Explanatory Variables to be those used for Regression Model V (28 in total, same as Regression Model IV). We run this regression model six times one for each cluster, considering the respective data points every time. The results presented here are again the R^2 and the total number of observations in each cluster. For further information on these models you can refer to the R code. It is important to be noted that we discard Cluster 5 for this Analysis as it contains only a Taxi Stand Location with 6 observations, which is considered to be an outlier.

It is clear from the values of the R^2 that the clustering analysis helped us to improve our models in some cases, whereas in other cases provided worse models. In general the R^2 even for the improved models are still relatively small. However, this method showed us that if we are interested in a specific Taxi Stand Location or Group of Taxi Stand Locations that belong to the same cluster, using the clustering analysis could significantly improve our model (see for example Cluster 5 for the Total Trip Time). However, our analysis still implies that

Cluster	R^2	Total Number of Observations
Cluster 1	0.1088	19633
Cluster 2	0.2275	46
Cluster 3	0.05968	759995
Cluster 4	0.03716	1320
Cluster 5	Discarded	5
Cluster 6	0.1413	2371

Table 4. Regression Models within Clusters for Interarrival Waiting Time

if we want to consider a linear regression model for predictions we should augment our data with external datasets that may provide more information.

Concluding Remarks

Using rather elementary, but thorough exploratory data analysis techniques, we managed to thoroughly sift through the data and expose potential issues that is symptomatic amongst literature which would be likely to use this data. Chief amongst these are questionable distributional assumptions of parameters such as inter-arrival times between taxi services and the features of the dataset itself being rather insufficient as a whole in terms of explanatory power (and thus would benefit from being augmented with an external dataset).

Nevertheless, very significant insights were also obtained, such as the surprisingly low variation across hours and the nature of demand and service patterns across stations distributed across the city.

This goes to show that simple but rigorous analysis of the dataset itself can serve as a strong robustness check for model specifications and assumptions, and should never be neglected in any case. The results of this report can be replicated via the code hosted here:

https://github.com/Runespear/ISYE7405FinalProject

Future Work

For more rigorous analyses of the distribution of inter-arrival times and service times, non-parametric methods such as the Kolmogorov-Smirnov test or Anderson-Darling test could be used to ascertain the parameters of the distributions. With regards to dimensionality reduction and factor analysis, Multi Factor Analysis could be used if sufficient hardware is available, due to the large number of categorical variables generated by our data processing. Finally, more thorough treatment of the data by taking the geo-spatial information of the stations into account would probably lead to even more insightful findings and robust results.

ACKNOWLEDGMENTS. We would like to thank the Professor Shihao Yang and the TA Tianyi Liu for their guidance in this project

1. Taxi Service Trajectory prediction challenge, ecml pkdd 2015 data set (https://archive.ics.uci.edu/ml/datasets/Taxi+Service+Trajectory+-+Prediction+Challenge%2C+ECML+PKDD+2015#) (2015) Accessed: December 10, 2020.
2. L Moreira-Matias, J Gama, M Ferreira, J Moreira, L Damas, Predicting taxi-passenger demand using streaming data. *IEEE Transactions on Intell. Transp. Syst.* **14**, 1393–1402 (2013).
3. DCF Irpino A., Rosanna V., Dynamic clustering of histogram data based on adaptive squared wasserstein distances. *EXPERT SYSTEMS WITH APPLICATIONS, vol. 41, p. 3351-3366* **41**, 3351–3366 (2014).

Appendix

Due to page limits and the large amount of results we have to display, we move some of the more verbose details to the appendix here.

Stratification by Hour. We present more detailed stratification of the amount of services on average a driver can expect to get in the whole year for each hour on each type of day in Tables 5, 6, 7. Surprisingly, there isn't that much variation between stratas.

Daytype	Total Services	E1	E2	E3	E4	E5	E6	E7	E8
Workdays	1206390	93	92	88	95	78	58	68	116
Weekends	423021	50	52	51	56	73	72	55	39
AllDaytypes	1629411	136	137	134	144	146	124	119	152

Table 5. EARLYSHIFT TAXI SERVICES VOLUME (AVERAGED PER DRIVER) STRATIFIED BY HOUR

Daytype Group	Total Services	M1	M2	M3	M4	M5	M6	M7	M8
Workdays	1206390	170	180	165	151	145	156	165	162
Weekends	423021	36	38	38	39	39	39	39	39
AllDaytypes	1629411	204	216	201	189	182	193	202	198

Table 6. MIDSHIFT TAXI SERVICES VOLUME (AVERAGED PER DRIVER) STRATIFIED BY HOUR

Daytype Group	Total Services	L1	L2	L3	L4	L5	L6	L7	L8
Workdays	1206390	156	144	130	115	111	102	97	94
Weekends	423021	38	40	44	46	51	49	46	46
AllDaytypes	1629411	192	182	171	157	156	144	137	135

Table 7. LATESHIFT TAXI SERVICES VOLUME (AVERAGED PER DRIVER) STRATIFIED BY HOUR

For completeness, we also include the following Table 8, which shows in which time intervals the variables shown in the first column take value 1. Recall that the vectors that we get for the variables shown in the first column are boolean.

Variables	Time Interval
E1	[0AM-1AM)
E2	[1AM-2AM)
E3	[2AM-3AM)
E4	[3AM-4AM)
E5	[4AM-5AM)
E6	[5AM-6AM)
E7	[6AM-7AM)
E8	[7AM-8AM)
M1	[8AM-9AM)
M2	[9AM-10AM)
M3	[10AM-11AM)
M4	[11AM-12PM)
M5	[12PM-1PM)
M6	[1PM-2PM)
M7	[2PM-3PM)
M8	[3PM-4PM)
L1	[4PM-5PM)
L2	[5PM-6PM)
L3	[6PM-7PM)
L4	[7PM-8PM)
L5	[8PM-9PM)
L6	[9PM-10PM)
L7	[10PM-11PM)
L8	[11PM-12AM)

Table 8. Time Intervals For Value 1

Regression Table Model I and Model II. The regression output for Model I and Model II is displayed below 9. The first column corresponds to Regression Model I and the second column corresponds to Regression Model II.

	Dependent variable:	
	TotalTime	
	(1)	(2)
WaitingTime		-0.0001^{*}
		(0.00003)
Distance	3.779^{***}	3.779^{***}
	(0.009)	(0.009)
Xstart	0.018^{**}	0.017^{*}
	(0.009)	(0.009)
Ystart	-0.550^{***}	-0.549^{***}
	(0.009)	(0.009)
EARLYSHIFT	-2.513^{***}	-2.512^{***}
	(0.026)	(0.026)
MIDSHIFT	-0.297^{***}	-0.297^{***}
	(0.020)	(0.020)
WEEKEND	-1.145^{***}	-1.145^{***}
	(0.021)	(0.021)
Constant	12.465^{***}	12.466^{***}
	(0.016)	(0.016)
Observations	626,696	626,696
R^2	0.229	0.229
Adjusted R^2	0.228	0.228
Residual Std. Error	6.961 (df = 626689)	6.961 (df = 626688)
F Statistic	$30{,}935.800^{***}$ (df = 6; 626689)	$26{,}516.900^{***}$ (df = 7; 626688)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 9. Model I and II Regression Output for Total Time

Regression Table Model III. The regression output for Model III is displayed below in Table 10.

Regression Table Model IV. The regression output for Model IV is displayed below in Table 11.

Regression Table Model V. The regression output for Model V is displayed below in Table 12.

	Dependent variable:
	WaitingTime
Distance	-1.525^{***}
	(0.370)
Xstart	-14.153^{***}
	(0.366)
Ystart	13.047^{***}
	(0.365)
EARLYSHIFT	33.450^{***}
	(1.057)
MIDSHIFT	10.634^{***}
	(0.799)
WEEKEND	0.589
	(0.855)
Constant	31.327^{***}
	(0.634)
Observations	626,696
R^2	0.007
Adjusted R^2	0.007
Residual Std. Error	284.238 (df = 626689)
F Statistic	721.897*** (df = 6; 626689)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10. Regression output for Model III

Dependent variable:TotalTime	
Distance	3.775***
Xstart	-0.001
Xend	6.063***
Ystart	-0.571***
Yend	2.362***
E1	-0.260***
E2	-0.338***
E3	-0.508***
E4	-0.625***
E5	-0.696***
E6	-0.568***
E7	-0.257***
E8	1.408***
M1	1.957***
M2	1.966***
M3	2.081***
M4	2.124***
M5	1.722***
M6	2.040***
M7	2.508***
M8	2.992***
L1	3.494***
L2	4.011***
L3	3.454***
L4	1.727***
L5	0.707***
L6	0.471***
L7	0.248***
Constant	-35.120***
Observations	626,696
R^2	0.238
Adjusted R^2	0.238
Residual Std. Error	6.919 (df = 626667)
F Statistic	6,982.163*** (df = 28; 626667)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 11. Regression output for Model IV

Dependent variable:Waiting Time	
Distance	-1.933***
Xstart	-14.763***
Xend	37.844***
Ystart	12.455***
Yend	32.805***
E1	6.092**
E2	-0.940
E3	1.915
E4	23.765***
E5	14.283***
E6	40.707***
E7	64.425***
E8	55.736***
M1	20.781***
M2	5.637**
M3	-2.489
M4	-7.986***
M5	-9.856***
M6	-12.200***
M7	-13.128***
M8	-12.727***
L1	-14.801***
L2	-13.143***
L3	-14.261***
L4	-12.543***
L5	-9.207***
L6	-6.190**
L7	-2.060
Constant	-981.661*
Observations	626,696
R^2	0.010
Adjusted R^2	0.010
Residual Std. Error	283.831 (df = 626667)
F Statistic	220.268*** (df = 28; 626667)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 12. Regression output for Model V

Histograms of Inter-arrival Times of Each Station. The detailed histograms of the inter-arrival times between each trip for each station is exhibited below

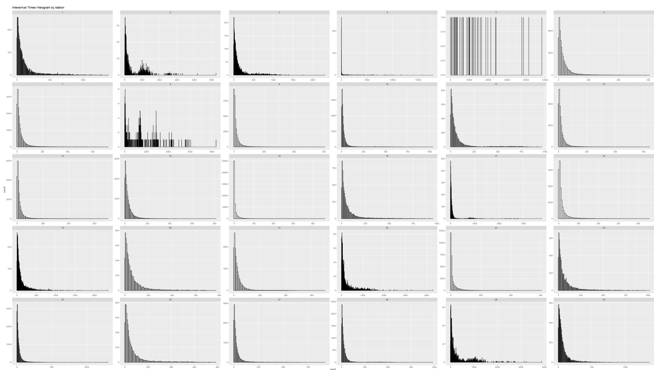


Fig. 11. Histograms of Inter-arrival Times of Station 1 to 30

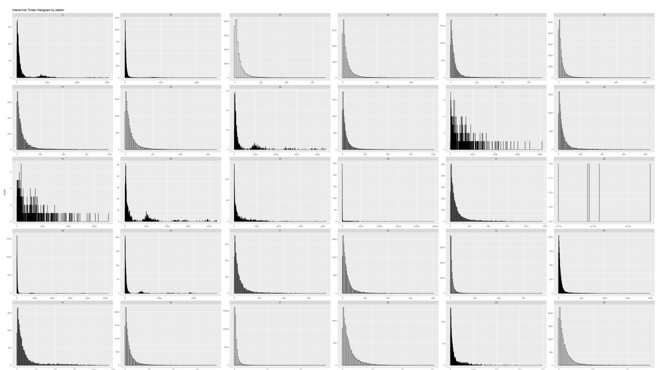


Fig. 12. Histograms of Inter-arrival Times of Station 31 to 60

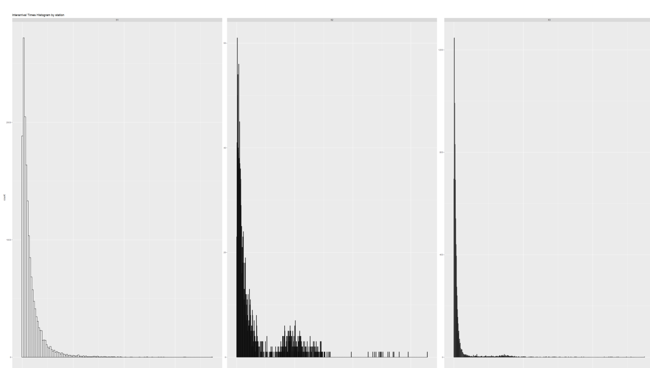


Fig. 13. Histograms of Inter-arrival Times of Station 61 to 63

Histograms of Trip Durations of Each Station. The detailed histograms of the trip (service) durations for each station is exhibited below

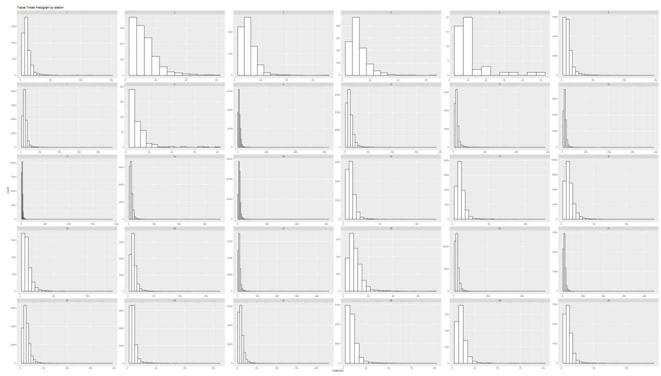


Fig. 14. Histograms of Trip Durations of Station 1 to 30

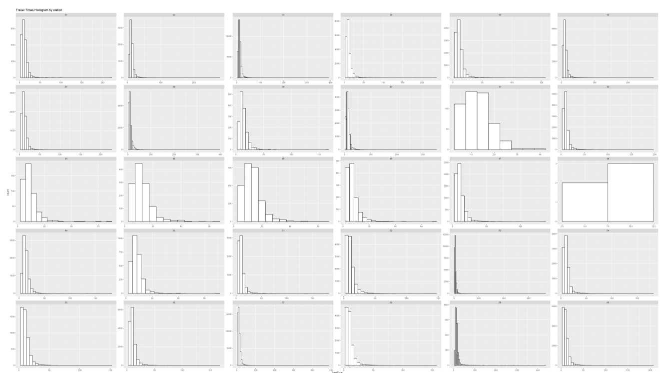


Fig. 15. Histograms of Trip Durations of Station 31 to 60

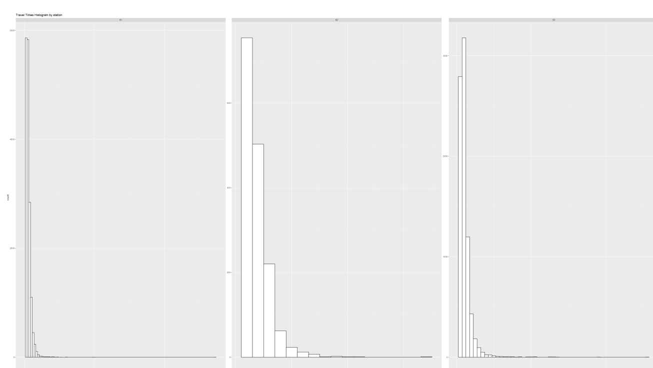


Fig. 16. Histograms of Trip Durations of Station 61 to 63

Cluster Assignments. We display the exact cluster assignments, which were used for the Regression Models within clusters below in Tables 13 and 14.

ORIGIN_STAND	TotalTime.clusters	WAIT.clusters
1	2	3
2	4	1
3	1	3
4	4	6
5	4	2
6	2	3
7	2	3
8	6	4
9	5	3
10	2	3
11	5	3
12	7	3
13	7	3
14	7	3
15	5	3
16	1	3
17	2	3
18	4	3
19	2	3
20	2	3
21	7	3
22	4	1
23	2	3
24	7	3
25	4	3
26	2	3
27	2	3
28	1	3
29	2	1
30	2	3
31	2	1
32	2	3

Table 13. Cluster Assignment of stations 1 to 32. Metric used was L2-Wasserstein distance, clustering technique used was k-means

ORIGIN_STAND	TotalTime.clusters	WAIT.clusters
33	2	3
34	2	3
35	2	3
36	2	3
37	2	3
38	7	3
39	4	1
40	7	3
41	4	4
42	2	3
43	6	4
44	4	1
45	4	1
46	2	6
47	2	3
48	1	5
49	2	3
50	4	1
51	2	3
52	2	3
53	7	3
54	2	3
55	1	3
56	2	3
57	7	3
58	2	3
59	3	1
60	2	3
61	7	3
62	1	1
63	5	3

Table 14. Cluster Assignment of stations 33 to 63. Metric used was L2-Wasserstein distance, clustering technique used was k-means