ISyE 7405 Final Project

Cluster Analysis and EDA of Taxi Demand Data

Presenters: Zilong Wang, Athanasios Lolos

Instructor: Prof. Shihao Yang

Contents Page

Overview

- Intro:
  - We conduct Exploratory Data Analysis (EDA) and Cluster Analysis on Taxi Dataset from ECM KDD 2015
  - We want to properly understand what the dataset is about and verify some assumptions that some papers made about the data generating process
- Motivation:
  - It is important to properly curate datasets that will be potentially reused across different studies
  - Verify assumptions commonly used in papers that use such data (are inter arrival times really exponential?)
  - Demonstrate how simple models can have remarkable improvements with careful stratification
- Methods:
  - Principal Component Analysis
  - Linear Regression | ANOVA
  - Dynamic Time Warping | Hierarchical Clustering

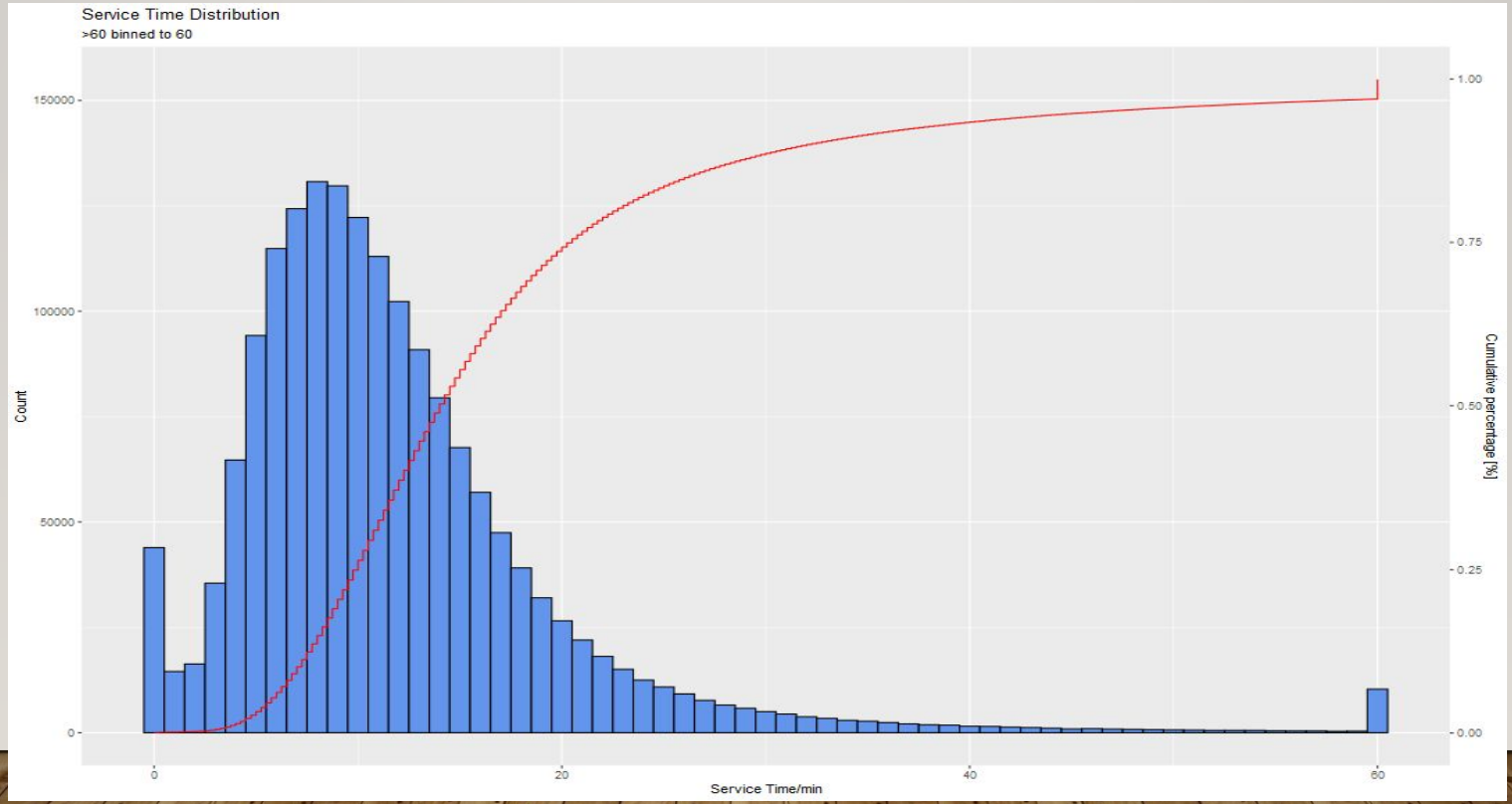
Dataset Description: Data

Taxi Demand Data from Porto, Portugal

- 1710670 Data Points
- 63 Taxi Stand Locations
- 448 Taxi Drivers
- Time Period: 2013-07-01 to 2014-06-30 (1 year)

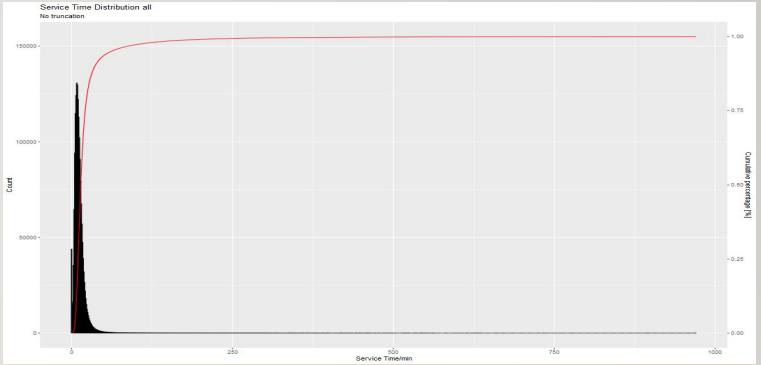
- Each data sample corresponds to one completed trip and it contains a total of nine features:
  - TRIP_ID: Identifier for each trip
  - CALL_TYPE: Way used to demand service
    - Three possible values: A, B, C
  - ORIGIN_CALL: Identifier for phone number
  - ORIGIN_STAND: Identifier for Taxi Stand
  - TAXI_ID: Identifier for Taxi Driver
  - TIMESTAMP: Identifier for Trip Start
  - DAYTYPE: Daytype of Trip's Start
    - Three possible values: A, B, C
  - MISSING_DATA: Checks for Missing Values
  - POLYLINE: List of GPS coordinates (for each 15 seconds of trip)

Dataset Description: Frequency Distribution of Taxi Cruise Times (>= 60 min binned)



Data Pre-Processing

- We discard trips with time taken <= 3 min
  - Probably erroneous tracking
  - 4.8% of the data (81259 out of 1710670 data points)
- We cannot discard trips with time taken > 60 min
  - Does not appear to be erroneous tracking
  - Long gentle sloped tail in histogram

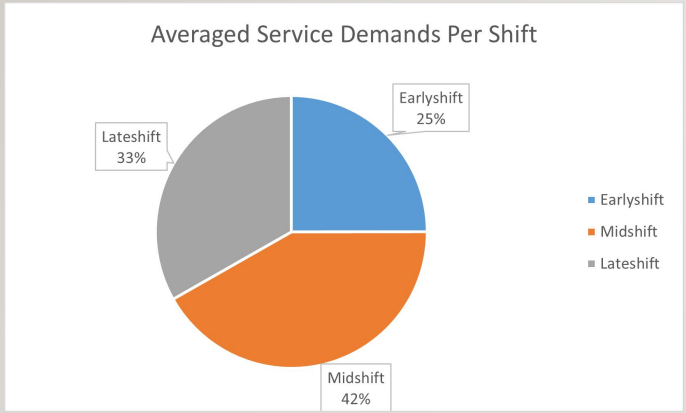


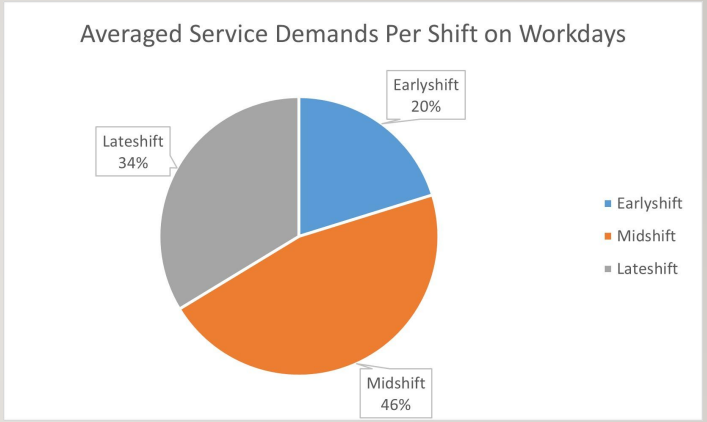
- We found the origin and we calculated the Euclidean Distance for each trip by using the first and the last pair of coordinates of the feature “POLYLINE”
- We created four boolean vectors “EARLYSHIFT” (0am-8am), “MIDSHIFT” (8am-4pm), “LATESHIFT”(4pm-0am) and “WEEKEND” according to the starting time of the trip
- We calculated the interarrival times of the customers of the different Taxi Stand Locations

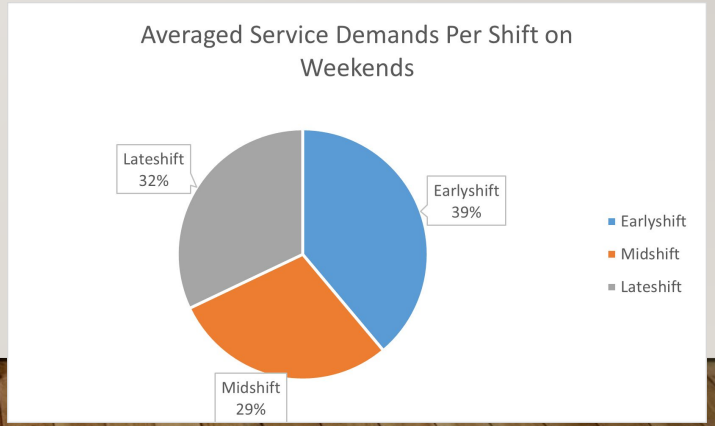
Exploratory Data Analysis: Table I

Daytype Group	Total Services	Averaged Service Demands Per Shift		
		EARLY [0 AM , 8 AM)	MID [8 AM , 4 PM)	LATE [4 PM , 0 AM)
Workdays	1206390	556.7143	1270.0204	928.6227
Weekends	423021	386.6731	288.9725	318.9063
All	1629411	929.0746	1555.7166	1235.2993

Exploratory Data Analysis: Pie Charts of Table I



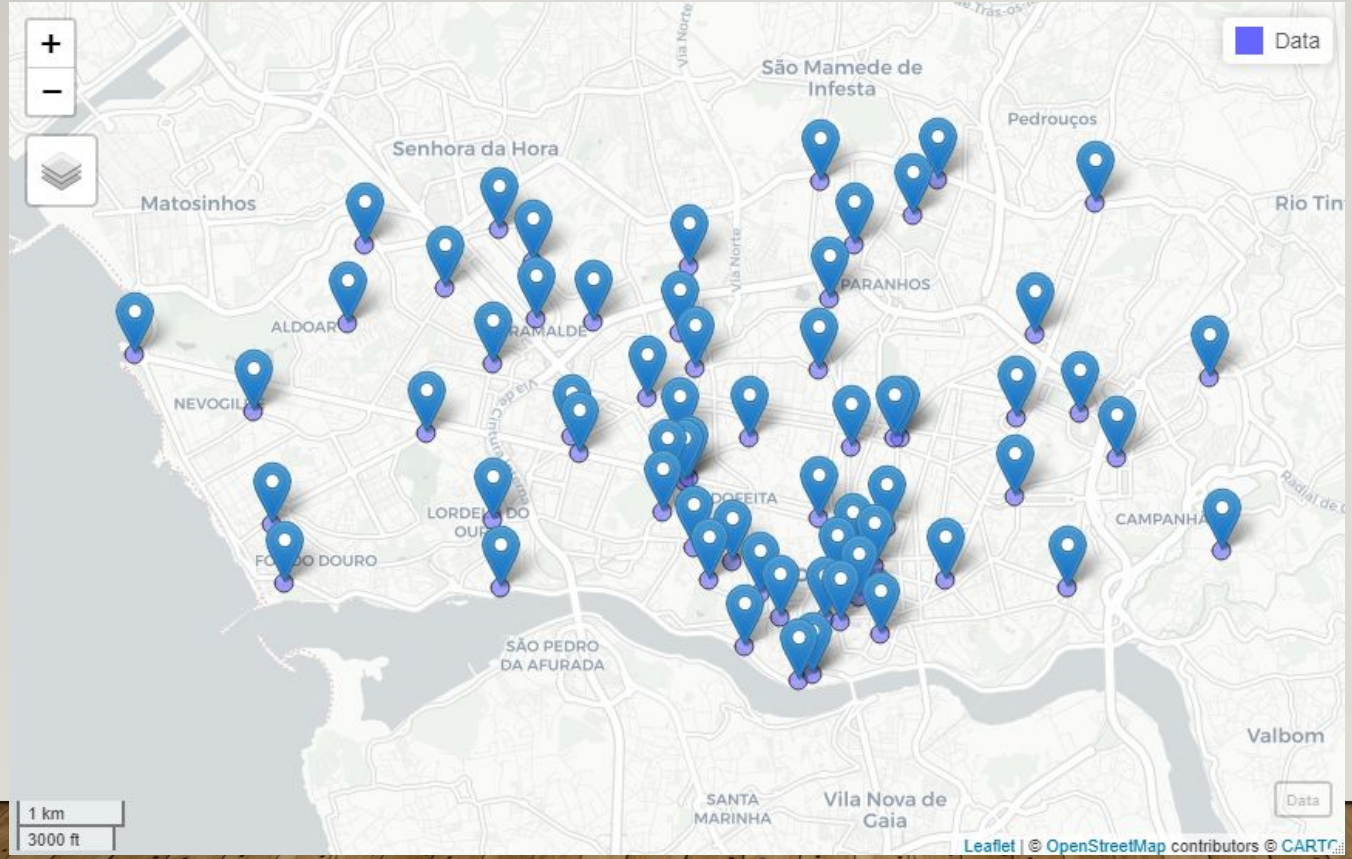




Exploratory Data Analysis: Table II

Taxi Services Volume / Driver		
	Services	Total Cruise Time / min
Max	7468.00	118847.75
Min	2	20.25
Mean	3694.81	47101.42
SD	1462.19	18035.34

Exploratory Data Analysis: Taxi Stand Locations in Porto, Portugal

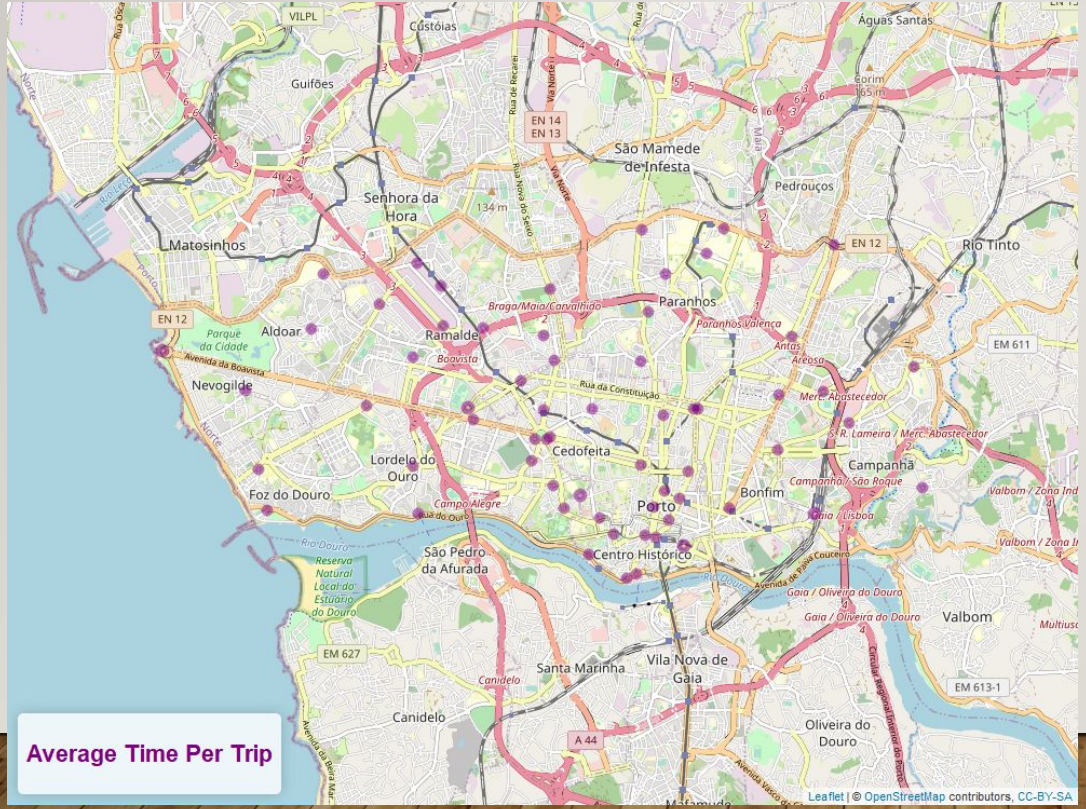


63 Stations distributed rather evenly throughout the city

Exploratory Data Analysis: Service Time Distribution at Each Station

- Cruise Time of trips originating at each station can be thought of as service times at a server
- In this case, time taken depends more on customer than server (driver)
- Check if distribution of service times look exponential (population and drivers relatively homogeneous)

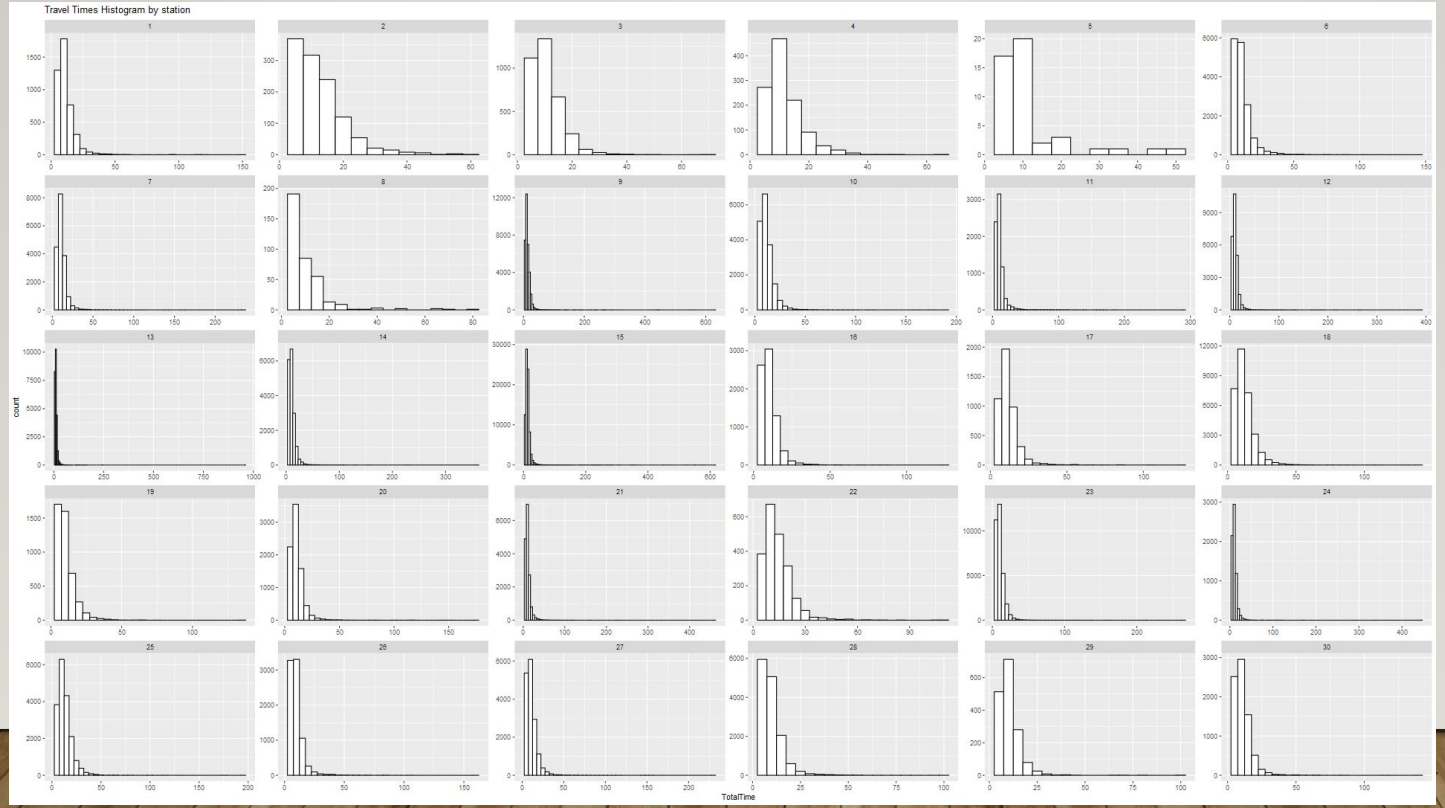
Exploratory Data Analysis: Stratified Average Trip (Service) Length per Station



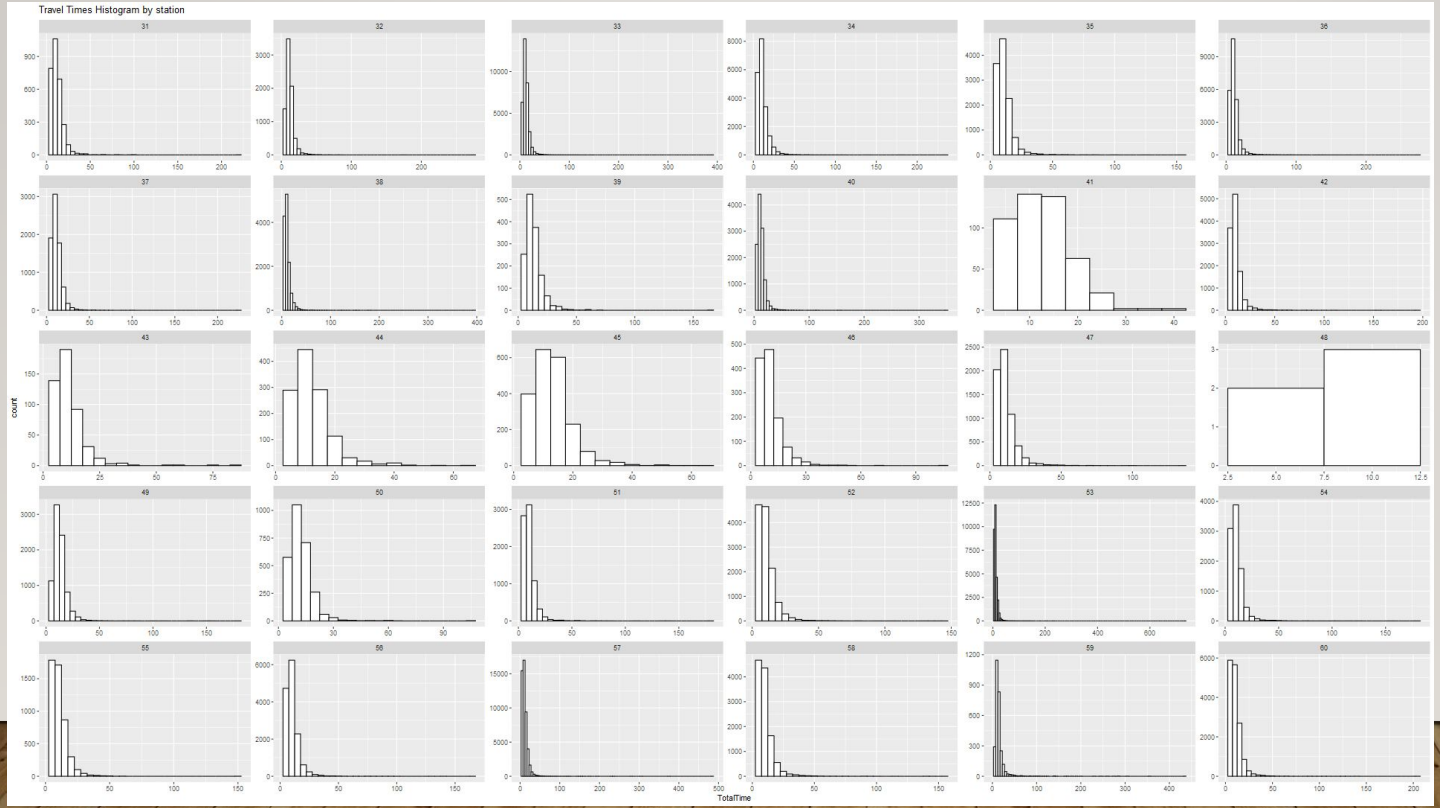
Service Length across all station data points:

Mean: 11.61 min
SD: 7.94

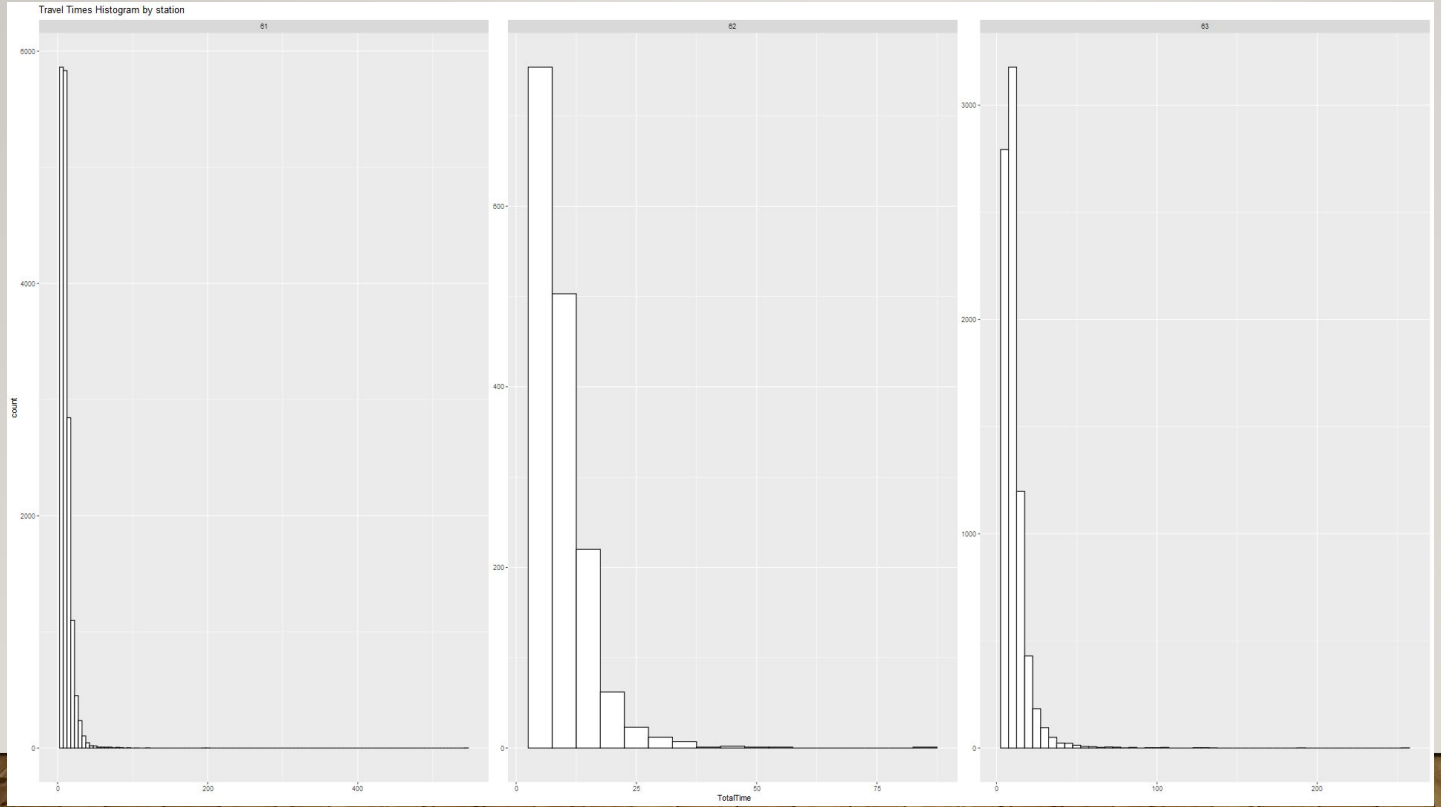
Exploratory Data Analysis: Histogram of Service Times 1-30



Exploratory Data Analysis: Histogram of Service Times 31-60



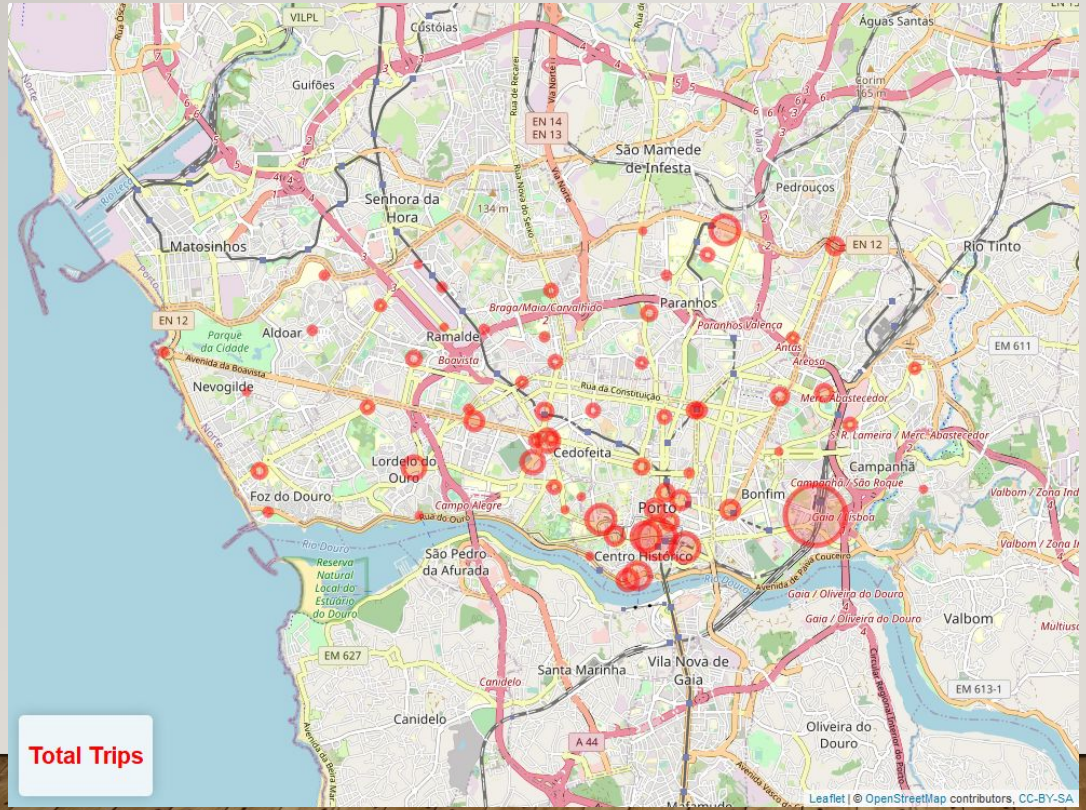
Exploratory Data Analysis: Histogram of Service Times 61-63



Exploratory Data Analysis: Interarrival Between Services at Each Station

- Commonly modelled as a Poisson process
- But is a mix of inter arrivals of customers and empty taxis
- Inter arrivals may not be independent!
  - E.g. More customers at a station may induce more taxis to prowl the location
- Exponential Looking Interarrival Times may indicate saturation of taxis or customers
  - Reduces to case of waiting for taxis (if a lot of customers) and vice versa
- Investigate the characteristics of each of these 63 stations via histograms

Exploratory Data Analysis: Stratified Total Demand per Station

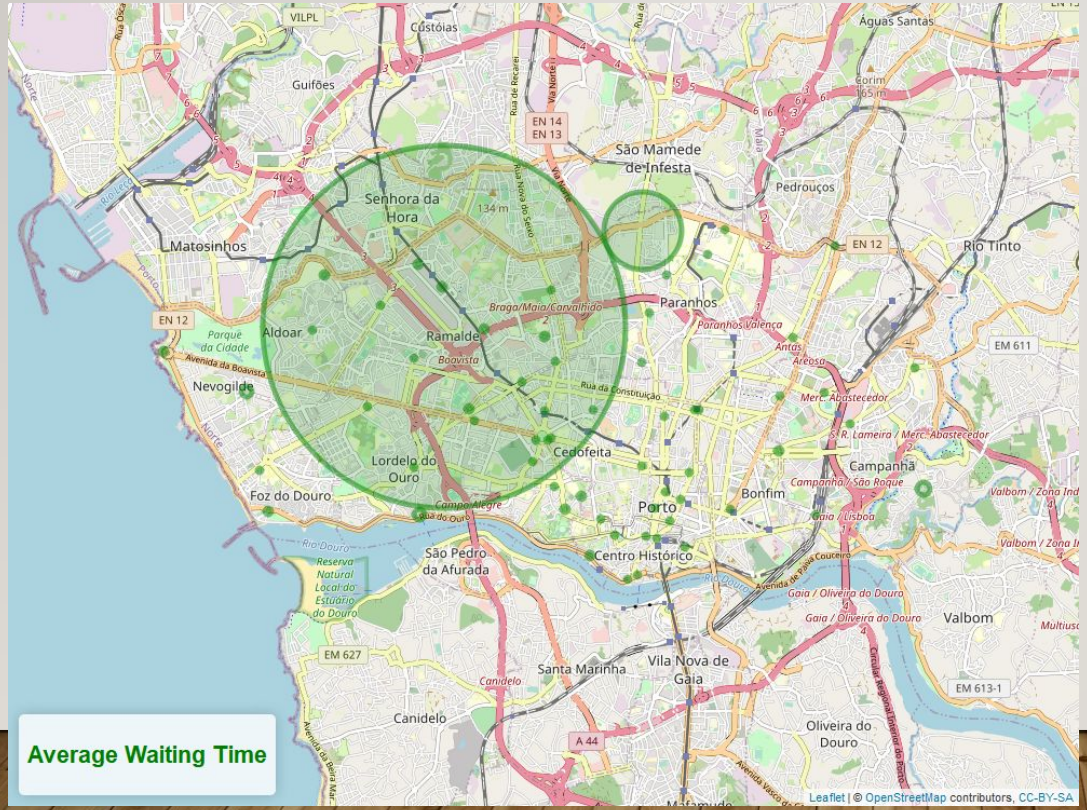


Demand across all station data points:

Mean: 12435 trips

SD: 13432.11

Exploratory Data Analysis: Stratified Average Interarrival Times



2nd Largest circle:
Station 5 only had 47 trips in total

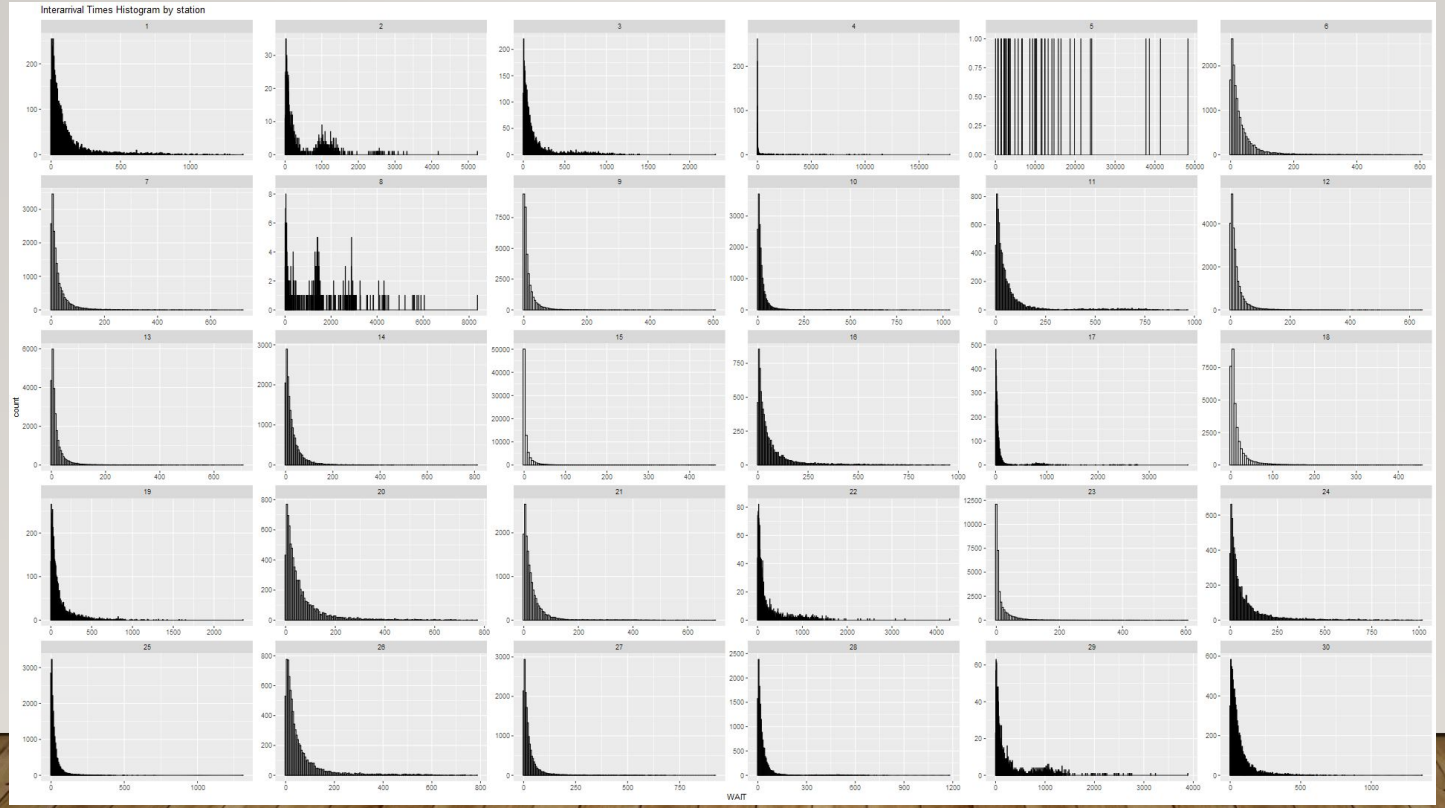
Largest circle:
Station 48 only had 6 trips in total

Look inversely proportional to previous figure

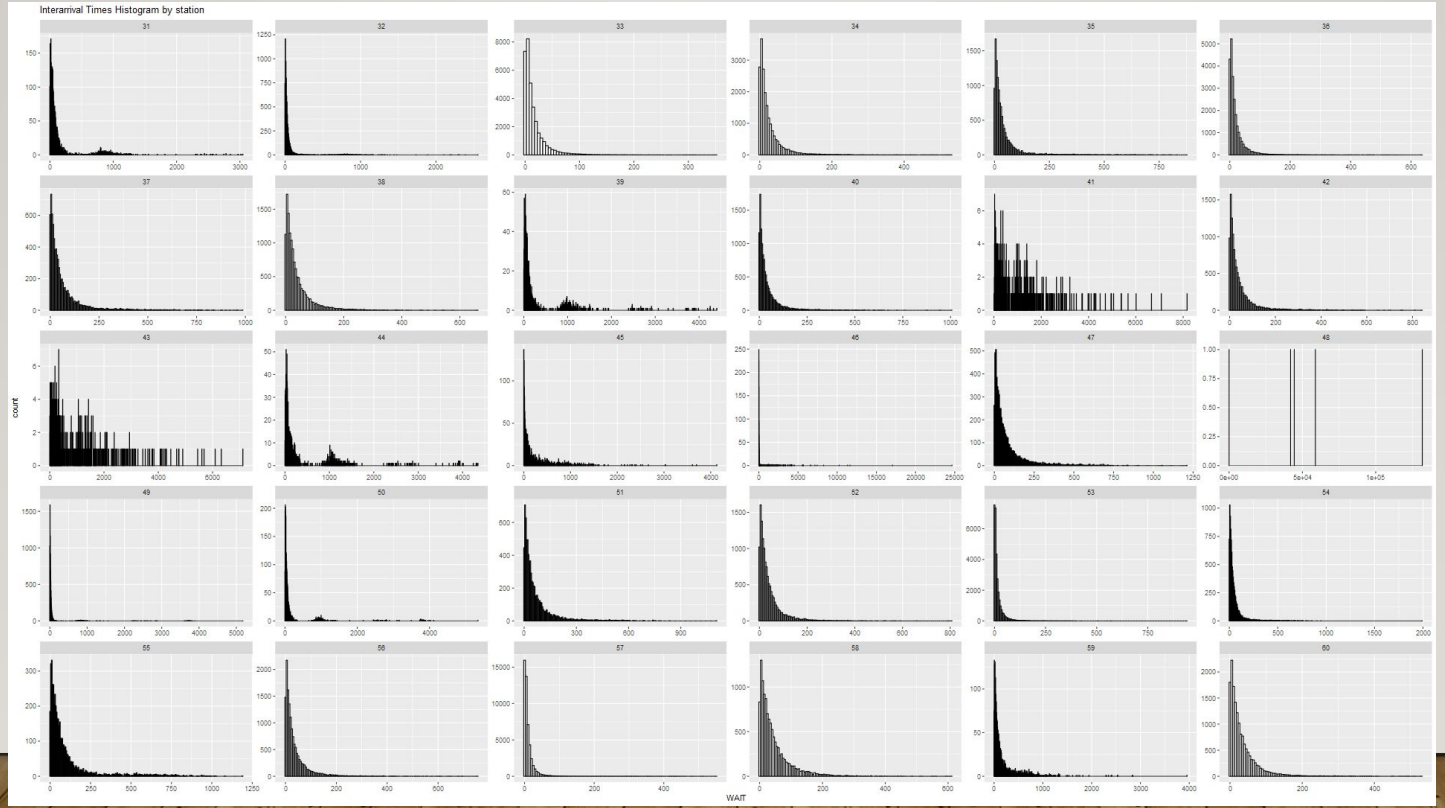
Summary across all station data points

Mean: 41.8 min
SD: 267.8

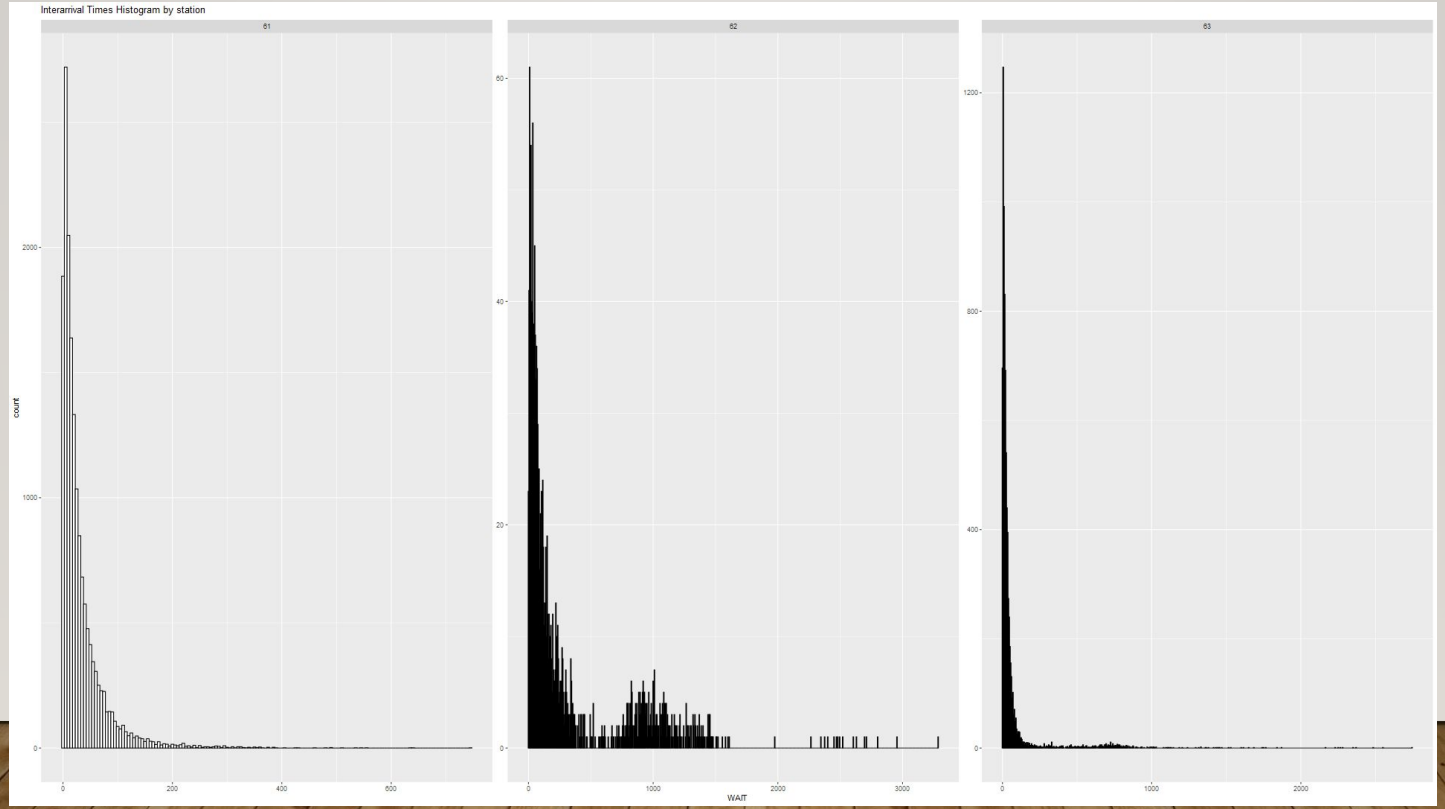
Exploratory Data Analysis: Histogram of Interarrival Times 1-30



Exploratory Data Analysis: Histogram of Interarrival Times 31-60



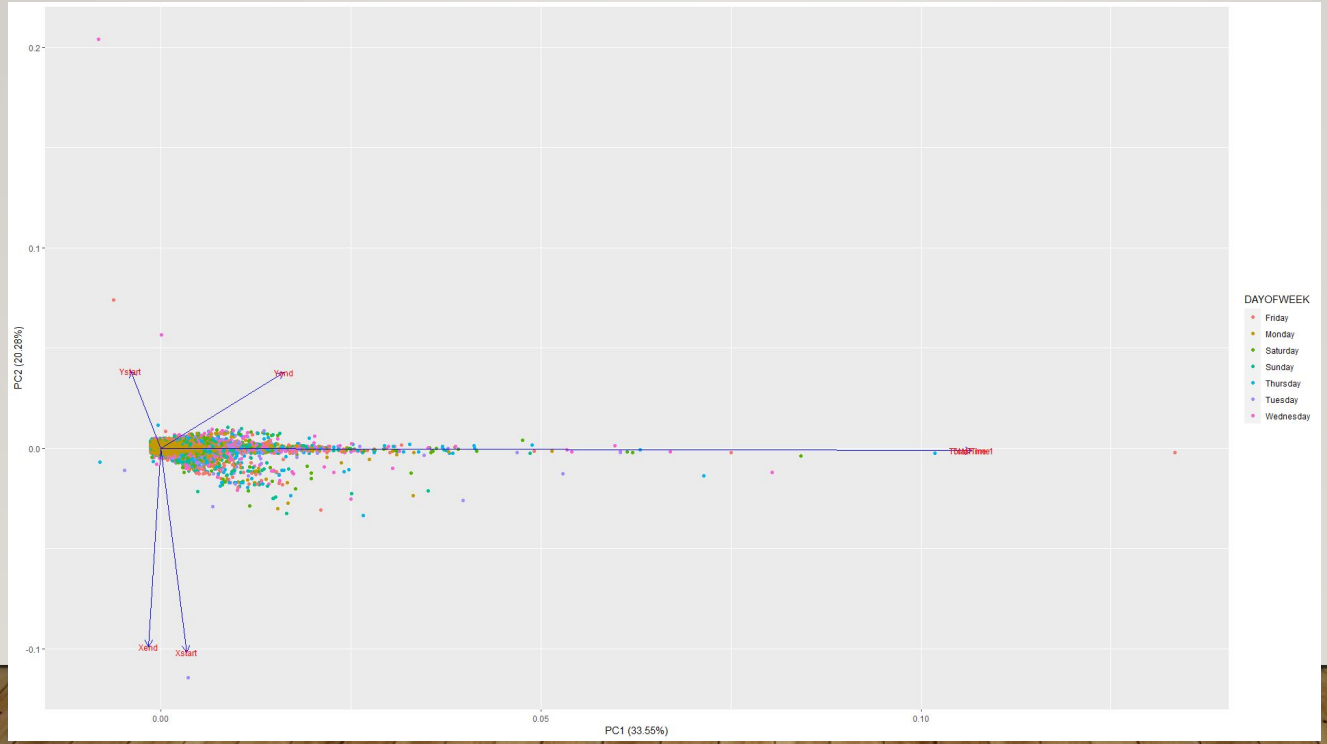
Exploratory Data Analysis: Histogram of Interarrival Times 61-63



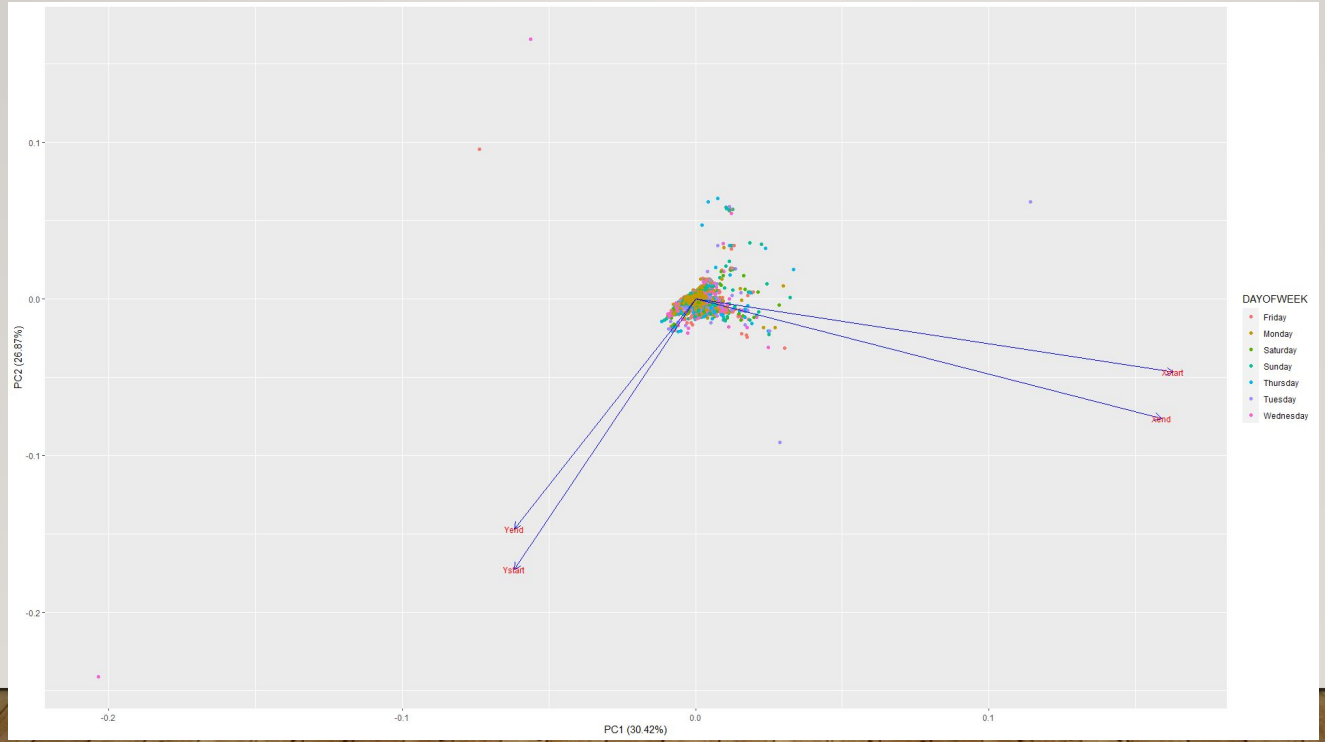
Exploratory Data Analysis: Summary of Interarrival Times Histograms

- Most of the interarrival times look exponentially distributed
- Some bimodal looking ones (7,22,29,41,43,44,50,62,63)
  - Periods where there is paucity of customers or taxis
- The outliers 5 and 48 look like barcodes due to low demand

Exploratory Data Analysis: Principal Components Analysis



Exploratory Data Analysis: Principal Components Analysis (No Total Time)



Linear Regression and ANOVA: Linear Regression Model I

- Response: Total Trip Time (T)
- Explanatory Variables: Distance (D), Xorigin (X), Yorigin (Y), EarlyShift (E), MidShift (M), Weekend (W)
- We consider only the data for which CALL_TYPE= "B" (i.e. start from Taxi Stand Location)
- We standardized the data in the columns: Distance, Xorigin, Yorigin
- We used 80% of the data for finding the regression model and 20% for testing
- Linear Regression Model:
  - Y=12.446+3.472*D+0.020*X-0.517*Y-2.485*E-0.292*M-1.133*W
  - R^2=0.2148, RMSE=7.5157, MAPE=0.3727, Average Absolute Difference=3.6748 mins

Linear Regression and ANOVA: Linear Regression Model I (cont'd)

```
Analysis of Variance Table

Response: TotalTime
               Df   Sum Sq Mean Sq   F value    Pr(>F)
Distance        1  7395278 7395278 153721.69 < 2.2e-16 ***
Xstart          1     7161    7161    148.86 < 2.2e-16 ***
Ystart          1   132943  132943   2763.42 < 2.2e-16 ***
EARLYSHIFT      1   565958  565958  11764.27 < 2.2e-16 ***
MIDSHIFT        1     5757    5757    119.66 < 2.2e-16 ***
WEEKEND         1   141713  141713   2945.72 < 2.2e-16 ***
Residuals  626689 30148898      48
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.446431   0.015471 804.493   <2e-16 ***
Distance     3.470709   0.008689 399.443   <2e-16 ***
Xstart       0.019533   0.008932   2.187   0.0287 *
Ystart      -0.517173   0.008909 -58.050   <2e-16 ***
EARLYSHIFT  -2.485293   0.025790 -96.366   <2e-16 ***
MIDSHIFT    -0.292376   0.019506 -14.989   <2e-16 ***
WEEKEND     -1.132677   0.020869 -54.274   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.936 on 626689 degrees of freedom
Multiple R-squared:  0.2148,    Adjusted R-squared:  0.2148
F-statistic: 2.858e+04 on 6 and 626689 DF,  p-value: < 2.2e-16
```

Linear Regression and ANOVA: Linear Regression Model II

- Response: Waiting Time (WT)
- Explanatory Variables: Distance (D), Xorigin (X), Yorigin (Y), EarlyShift (E), MidShift (M), Weekend (W)
- We consider only the data for which CALL_TYPE= "B" (i.e. start from Taxi Stand Location)
- We standardized the data in the columns: Distance, Xorigin, Yorigin
- We used 80% of the data for finding the regression model and 20% for testing
- Linear Regression Model (All Variables):
  - WT=31.163-1.385*D-13.674*X+12.747*Y+32.280*E+10.506*M+0.891*W (All Variables)
  - R^2=0.007609, Average Absolute Difference=41.777 mins
  - WT=31.402-13.691*X+12.610*Y+32.277*E+10.448*M (Significant Variables)
  - R^2=0.007583, Average Absolute Difference=41.744 mins

Linear Regression and ANOVA: Linear Regression Model II (cont'd)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.1627     0.5837  53.388  < 2e-16 ***
Distance     -1.3854     0.3278  -4.226 2.38e-05 ***
Xstart      -13.6740     0.3370 -40.577  < 2e-16 ***
Ystart       12.7473     0.3361  37.924  < 2e-16 ***
EARLYSHIFT   32.2796     0.9730  33.174  < 2e-16 ***
MIDSHIFT     10.5057     0.7359  14.275  < 2e-16 ***
WEEKEND       0.8914     0.7874   1.132    0.258
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261.7 on 626689 degrees of freedom
Multiple R-squared:  0.007619,  Adjusted R-squared:  0.007609
F-statistic: 801.9 on 6 and 626689 DF,  p-value: < 2.2e-16
```

```
Analysis of Variance Table

Response: WaitingTime
               Df     Sum Sq   Mean Sq   F value Pr(>F)
Distance        1 9.5000e+01        95    0.0014 0.9703
Xstart          1 1.5450e+08 154496667 2256.0836 <2e-16 ***
Ystart          1 9.8079e+07  98079255 1432.2316 <2e-16 ***
EARLYSHIFT      1 6.2951e+07  62950848  919.2585 <2e-16 ***
MIDSHIFT        1 1.3867e+07  13867067  202.4980 <2e-16 ***
WEEKEND         1 8.7767e+04     87767    1.2816 0.2576
Residuals  626689 4.2916e+10     68480
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  31.4023     0.5497   57.13   <2e-16 ***
Xstart      -13.6905     0.3370  -40.63   <2e-16 ***
Ystart       12.6097     0.3345   37.69   <2e-16 ***
EARLYSHIFT   32.2768     0.9669   33.38   <2e-16 ***
MIDSHIFT     10.4479     0.7339   14.24   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 261.7 on 626691 degrees of freedom
Multiple R-squared:  0.007589,  Adjusted R-squared:  0.007583
F-statistic:  1198 on 4 and 626691 DF,  p-value: < 2.2e-16
```

Linear Regression and ANOVA: Linear Regression Model III

- Response: Total Trip Time (T)
- Explanatory Variables: Waiting Time (WT), Distance (D), Xorigin (X), Yorigin (Y), EarlyShift (E), MidShift (M), Weekend (W)
- We used 80% of the data for finding the regression model and 20% for testing
- We concluded that the Waiting Time does not improve Regression Model I

```
Analysis of Variance Table

Response: TotalTime
               Df    Sum Sq Mean Sq    F value    Pr(>F)
WaitingTime     1      4343    4343      90.28 < 2.2e-16 ***
Distance        1   7395295 7395295  153722.67 < 2.2e-16 ***
Xstart          1      6532    6532     135.79 < 2.2e-16 ***
Ystart          1    131125  131125    2725.64 < 2.2e-16 ***
EARLYSHIFT      1    564269  564269   11729.21 < 2.2e-16 ***
MIDSHIFT        1      5718    5718     118.85 < 2.2e-16 ***
WEEKEND         1    141699  141699    2945.43 < 2.2e-16 ***
Residuals  626688 30148727      48
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.245e+01  1.551e-02 802.798   <2e-16 ***
WaitingTime -6.304e-05  3.348e-05  -1.883   0.0597 .
Distance     3.471e+00  8.689e-03 399.428   <2e-16 ***
Xstart       1.867e-02  8.944e-03   2.088   0.0368 *
Ystart      -5.164e-01  8.919e-03 -57.894   <2e-16 ***
EARLYSHIFT  -2.483e+00  2.581e-02 -96.203   <2e-16 ***
MIDSHIFT    -2.917e-01  1.951e-02 -14.952   <2e-16 ***
WEEKEND     -1.133e+00  2.087e-02 -54.272   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.936 on 626688 degrees of freedom
Multiple R-squared:  0.2148,    Adjusted R-squared:  0.2148
F-statistic: 2.45e+04 on 7 and 626688 DF,  p-value: < 2.2e-16
```

Proposed Next Steps

1. Further stratify dataset based on exact hour of day for more sophisticated modelling
2. Cluster Stations by their histograms / historical demand profile
    a. Use Dynamic Time Warping over their time series and get the dissimilarity matrix for Hierarchical clustering
    b. Or use DBSCAN over their Histograms
3. Run regression models separately within clusters to improve explanatory power

References

- 1. Taxi Service Trajectory prediction challenge, ecml pkdd 2015 data set (https: 100 //archive.ics.uci.edu/ml/datasets/Taxi+Service+Trajectory+-+Prediction+Challenge%2C+ 101 ECML+PKDD+2015#) (2015) Accessed: October 28, 2020. 102
- 2. L Moreira-Matias, J Gama, M Ferreira, J Moreira, L Damas, Predicting taxi-passenger demand 103 using streaming data. IEEE Transactions on Intell. Transp. Syst. 14, 1393–1402 (2013). 104
- 3. TK Vintsyuk, Speech discrimination by dynamic programming. Cybernetics 4, 52–57 (1968) 105 Russian Kibernetika 4(1):81-88 (1968).

Thank you!

Questions?