FORZADATA

# Udacity Project
# NYC Subway Data Analysis

Prepared for: Udacity Data Science
Prepared by: Charles Upjohn

March 31, 2014

**Charles Upjohn** 2025 Woodmont Blvd Apt 317 Nashville, TN 37215 615-412-9152 upjohnc@gmail.com

# Subway Operational Analysis

## Summary

At the end of Udacity's data science course, I was given a set of data on NYC subway usage.  From this data set, I performed an analysis of usage for operational planning.  Similar business scenarios would be planning order fulfillment, server usage, or call center planning.

In this analysis, I asked two questions.  One, are there consistent patterns in subway ridership by time or by day of the week?  Two, are there data elements that can predict fluctuations in those patterns such as rain or temperature?

### Data Set

The data set of this project is similar to most operational data.  It contains the entries and exits by hour and by unit.  This set is for one month, May 2011.  It also has weather elements for that hour and unit, such as temperature, dew point, pressure, fog, rain, etc.

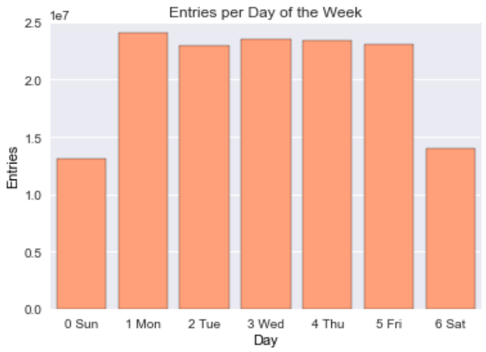
### Analysis Steps

The first step I took was to create a visual of the entries by day of the week and then by time of the day.  I aggregated the data in order to develop the barplots and then created the barplots.  The second step was to determine which weather elements correlate to entries.  I developed scatterplots with one weather element on an axis and entries on the other.  This view led to choosing the elements to test in an ordinary least squares regression analysis.

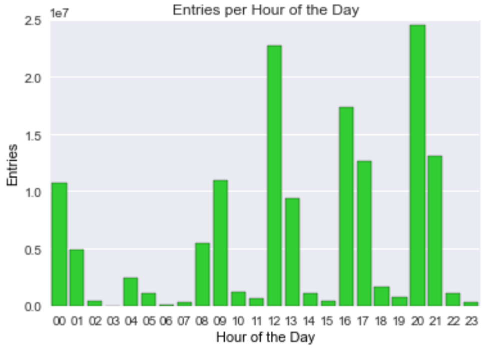
## Usage by Days and Time

In the two histograms, there are distinct patterns of usage by days and time.  Weekday usage is around 1.5 to 2 times higher than on the weekends.  Also, there is high usage during rush hour, lunchtime, evening and then midnight.

#### Entries per Day of the Week



### Entries per Hour of the Day



## Statistical Significance

While the graphs show difference between weekdays and weekends and also between different hours of the day, a statistical significance test adds a level of understanding to the data.  A Mann Whitney U test, statistical significance of difference in the means, shows the following results.  The Mann Whitney test was performed because some of the data did not have a normal distribution and the test does not require a normal distribution.

## Weekday to Weekend Comparison

| Comparison | Significant Difference at 95% |
|---|---|
| Monday to Sunday | Yes |
| Tuesday to Sunday | Yes |
| Wednesday to Sunday | Yes |
| Thursday to Sunday | Yes |
| Friday to Sunday | Yes |
| Monday to Saturday | No |
| Tuesday to Saturday | Yes |
| Wednesday to Saturday | Yes |
| Thursday to Saturday | Yes |
| Friday to Saturday | Yes |

### Peek Hour to Non-Peek Hour Comparison

| Comparison | Significant Difference at 95% |
|---|---|
| Noon to 10am | Yes |
| Noon to 3pm | No |
| 8pm to 10am | No |
| 8pm to 3pm | Yes |

Weekend and weekday show significant difference. Even Monday compared to Saturday returns a p-value of 0.07 which is close to 0.05 level required to be significant at 95%. The conclusion is that decisions can be made to allocate resources differently for weekend and weekday.

On the other hand, peek hours compared to non-peek hours do not show significant difference. This is most likely due to variance of usage at the separate units and subway stations. It would be recommended to analyze the usage at the subway unit level to gain greater insight.

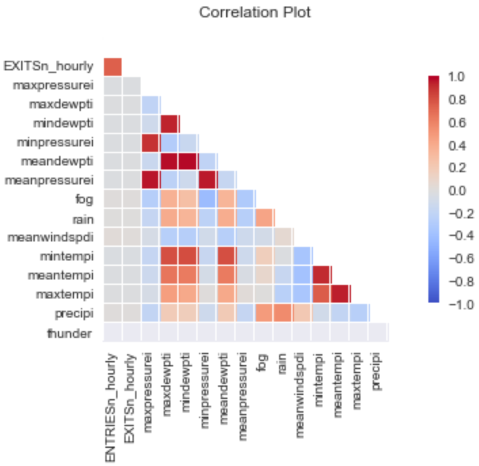
## Predicting Usage

As a first step to understanding the data, the correlation plot is very useful for initial review of all the elements in the data set. This graph shows a high correlation between exits and entries with the other elements not very strongly correlated. However, exits is not an element useful to predicting usage. We cannot project out the exits. Whereas we can project out the probability of rain which we could leverage in a formula produced from regression analysis. As an example, if there is an 80 percent chance of rain and the act of rain increases entries by 0.1, then it can be projected that entries will increase by 0.08.

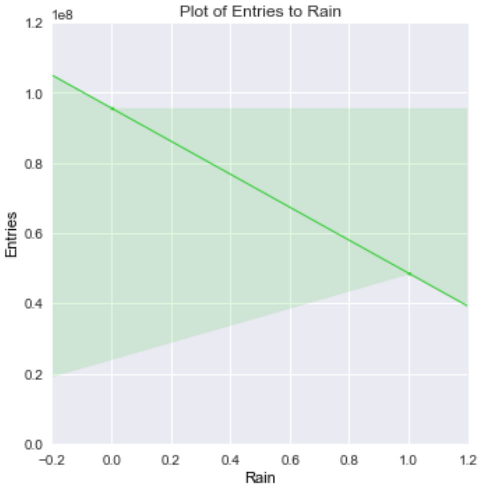
After review of the correlation plot, I created scatterplots of entries against the elements in the data set. This step was to see the potential elements to test in a regression analysis. The elements that showed potential were rain, minimum temperature, and maximum dew point. While the plots did not show high potential, it is worth the effort to test.

I then ran the regression formula on combinations of the three elements. The results were poor with the best result having a 0.176 r-squared (calculation of fit with 1 being perfect fit).

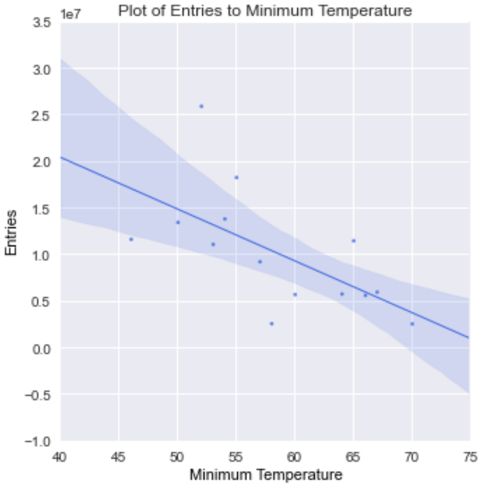
### Correlation Plot



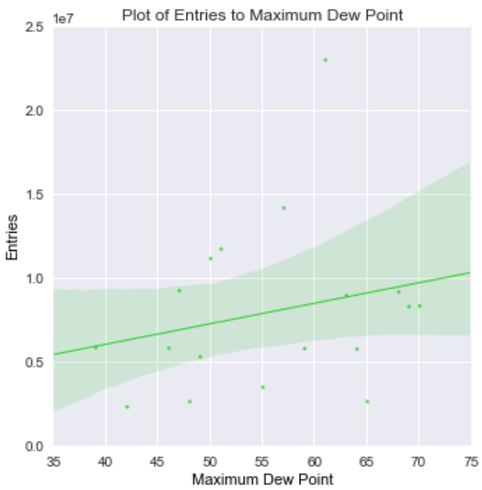
## Scatterplot of Entries to Rain



## Scatterplot of Entries to Minimum Temperature



### Scatterplot of Entries to Maximum Dew Point



### Results of Regression Analysis

| Combination | R-Squared |
|---|---|
| rain, min temp, max dew | 0.176 |
| min temp, max dew | 0.176 |
| rain | 0.161 |
| min temp | 0.175 |
| max dew | 0.175 |

## Results

For the first question, the graphs resulted in a good answer. NYC Subway authorities can plan its operations around time and day. For instance, the operational needs can be planned accordingly.

The regression analysis did not produce the desired results of augmenting the patterns found in the answer to the first question. In other words, the elements in the data set did not support predicting usage changes.

## Further Analysis Recommendations

This analysis was a first pass at understanding the NYC operations based on this data set. A deeper analysis on some of the findings is recommended. For instance, the t-test comparing peek hours to non-peek hours showed a lack of statistical significance. The graph reveal a difference, which leads to further analysis. The recommendation is to run a t-test on the means at the unit level rather than an aggregation of units.

Collecting more data would also be beneficial. This would be additional months and years of the data elements in the data set and also more data elements. Having more data can both confirm the patterns found and also show additional patterns by month and season. More data elements can possibly correlate to usage and thereby augment the knowledge of the patters by predicting fluctuations.

A mapreduce solution should also be considered as the data sets grow. The data set in the exercise had approximately 132K rows and if five years worth of this data were collected it would be nearly 8MM rows. Mapreduce solutions help to speed the reading of the data from its source. Examples of map and reduce code from the Udacity's exercises are here:link to the code from the Udacity Exercises

# Code from Analysis

Github of Project

NbViewer: Graph Analysis

NbViewer: Regression Analysis

NbViewer: Mapreduce