Section 0: References

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm

http://stackoverflow.com/questions/6986986/bin-size-in-matplotlib-histogram

http://stackoverflow.com/questions/6871201/plot-two-histograms-at-the-same-time-with-matplotlib

http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?

r2_ameasureofgoodness_of_fitoflinearregression.htm

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

http://scikit-learn.org/stable/modules/linear_model.html

http://stackoverflow.com/questions/22377539/plotting-one-scatterplot-with-multiple-dataframes-with-ggplot-in-python

http://stackoverflow.com/questions/tagged/python-ggplot

http://web.stanford.edu/~cengel/cgi-bin/anthrospace/ggplot-from-python-with-rpy2

http://stackoverflow.com/questions/1514553/how-to-declare-an-array-in-python

http://pandas.pydata.org/pandas-docs/stable/indexing.html#indexing-view-versus-copy

http://stackoverflow.com/questions/17950374/converting-a-column-within-pandas-dataframe-from-int-to-string

http://stackoverflow.com/questions/15356433/how-to-generate-pandas-dataframe-column-of-categorical-from-string-column

http://stackoverflow.com/questions/3172509/numpy-convert-categorical-string-arrays-to-an-integer-array

http://stackoverflow.com/questions/8238918/ggplot-error-similar-data-graphs-why-not-anymore

http://docs.ggplot2.org/current/geom_histogram.html

http://stackoverflow.com/questions/26680066/python-ggplot-set-axis-range-for-datetime

http://stackoverflow.com/questions/24917700/adding-a-row-to-a-multiindex-dataframe-series

http://stackoverflow.com/questions/6919025/how-to-assign-colors-to-categorical-variables-in-ggplot2-that-have-stable-mappin

http://stackoverflow.com/questions/6986986/bin-size-in-matplotlib-histogram

http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm

http://stackoverflow.com/questions/7837722/what-is-the-most-efficient-way-to-loop-through-dataframes-with-pandas

http://stackoverflow.com/questions/16476924/how-to-iterate-over-rows-in-a-dataframe

http://stackoverflow.com/questions/15315452/selecting-with-complex-criteria-from-pandas-dataframe

http://stackoverflow.com/questions/17432814/filter-rows-based-on-a-true-value-in-a-column-python-pandas-data-frame

http://stackoverflow.com/questions/23934905/pandas-conditionally-select-column-based-on-row-value

http://stackoverflow.com/questions/25597196/selecting-rows-of-a-dataframe-based-on-two-conditions-in-pandas-python

http://stackoverflow.com/questions/22341271/get-list-from-pandas-dataframe-column

http://stackoverflow.com/questions/22611446/perform-2-sample-t-test

http://www.tutorialspoint.com/python/time_strptime.htm

http://stackoverflow.com/questions/10982089/how-to-shift-a-column-in-pandas-dataframe

http://stackoverflow.com/questions/15907200/how-to-add-a-header-to-a-csv-file-in-python

http://stackoverflow.com/questions/23072082/combining-columns-of-multiple-files-in-one-file-python

Section 1: Statistical Test

1.1 The statistical test that was used to analyze the NYC subway data was Mann Whitney U test. I used a one-sided P value. The null hypothesis in this case is that the ridership on rainy days is same as the ridership on non-rainy days. The one-tailed test will test either if the ridership on rainy days is greater than that on non-rainy days or if the ridership on rainy days is less than that on non-rainy days, but not both. The p-critical value used is 0.05.

1.2 First, the distribution of hourly entries was examined when raining versus not raining. A histogram was plotted and it was observed that the distribution is not normal. Hence Welch's T test could not be used. Therefore, Mann Whitney U test was used for analysis.

1.3 The following are the results from the statistical test:

a. p-value: 0.0249
b. mean entries while raining: 1105.44
c. mean entries while not raining: 1090.27

1.4 As the p value is less than 0.05, we can reject the null hypothesis i.e., the distribution of number of entries is statistically different between rainy and non-rainy days.

Section 2: Linear Regression

2.1 Gradient Descent approach was used to compute the coefficients theta and produce prediction for ENTRIESn hourly in the regression model.

2.2 The features that were used in the model were: rain, precipitation, hour and mean temperature. UNIT was used as a dummy feature.

2.3 Rain and precipitation were used because I thought that people would be using subway more when it rains. Hour and mean temperature were included because the squared error improved a lot by including these features.

2.4 The weights of non-dummy features are given by the variable theta. The weights of the features rain, precipi, hour and meantempi respectively are mentioned below:

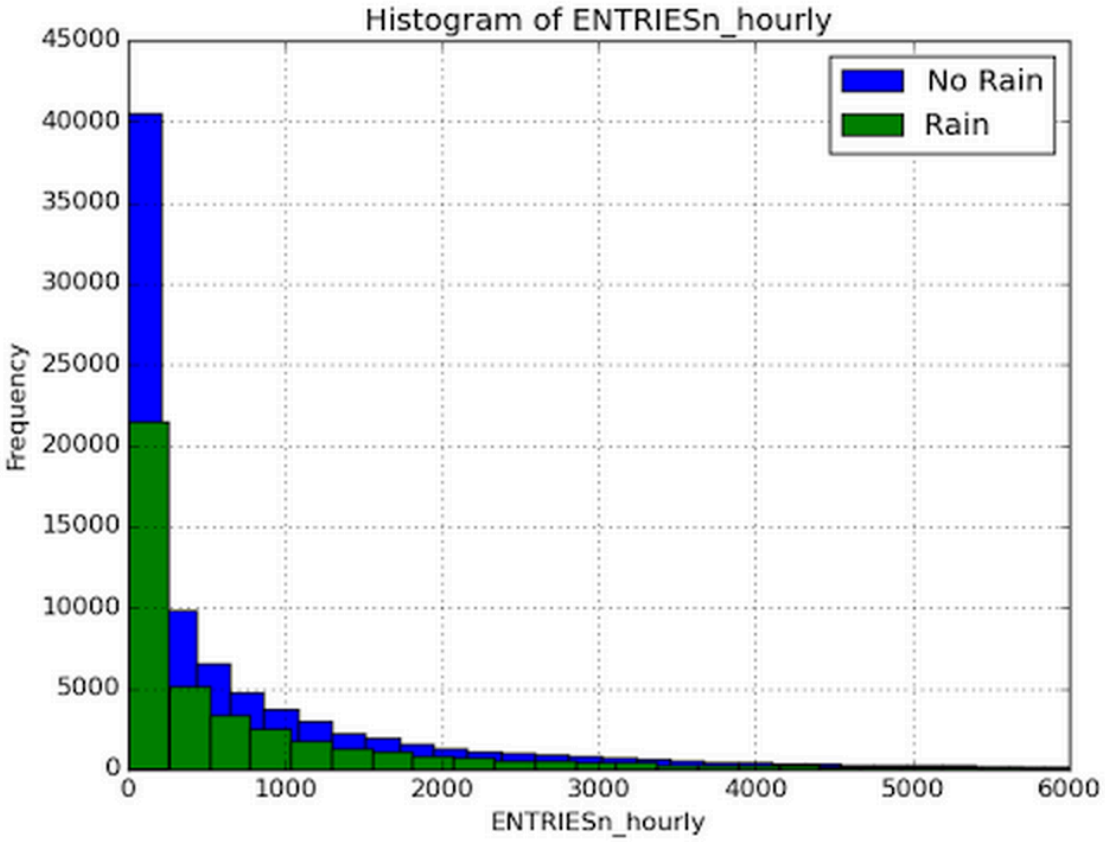
```
2.92398062e+00   1.46526720e+01   4.67708502e+02  -6.22179395e+01
```

2.5 The coefficient of determination is 0.463968815042.

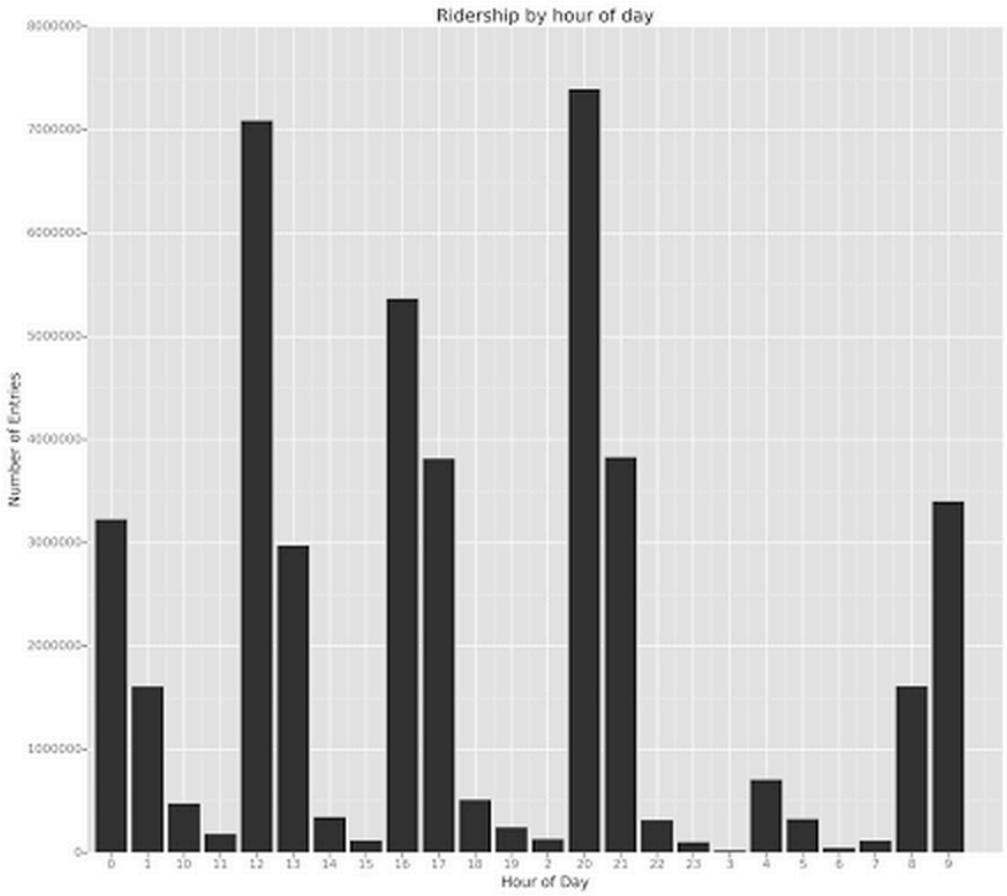
2.6 The coefficient of determination is less than 50%. If I also take a look at the residual plots, I observe that the residuals are mostly centered on zero throughout the range of fitted values. Also, as the R-square value is 0.46 it means that we have explained 46% of the original variability. Ideally, we would like to explain most if not all of the original variability. In other words, this linear model to predict ridership is appropriate and is correct on average for all fitted values.

Section 3: Visualization

3.1



3.2



The above bar graph shows the distribution of number of turnstile entries by hour of day. The values in the x-axis i.e., the hour of day is not ascending order as the values had to be converted to string before plotting.

Section 4: Conclusion

4.1 More people ride subway when it rains. When we look at the histogram that shows total number of entries between rainy and non-rainy days, the number of people who take NYC subway on a rainy day is almost half of the number of people who take subway on a non-rainy day. But this is true because the given dataset has more non-rainy days than rainy days. Hence, it makes sense to look at the average number of entries for rainy and non-rainy days in this case. When we do this, we observe that the average number of entries for rainy day is 1105.44 and that for a

non-rainy day is 1090.27; hence we can say that more people ride subway when it rains.

4.2 First, I started with the null hypothesis stating that the distribution of number of entries is the same for rainy and non-rainy days. After performing Mann Whitney U test, the p-value came to be 0.02. As the p-value is less than 0.05, we can reject the null hypothesis. Hence, the distribution of hourly entries is statistically different between rainy and non-rainy days. After performing linear regression and analyzing the coefficient of determination and the residuals, it was learnt that the model that was used to predict the ridership is appropriate. Most of the residuals were centered on zero. Also by looking at the average hourly entries for rainy versus non-rainy days, it was observed that most people take NYC subway during rainy days.

Section 5: Reflection

5.1 Some of the shortcomings of the analysis and the dataset given are found below:

a. While performing linear regression on the given data to predict ridership on rainy and non-rainy days, we looked only at linear relationship between entries hourly and other features because linear regression only looks at linear relationships between dependent and independent variables. It assumes there is a straight-line relationship between them. However, this might not always be the case.

b. Linear regression is also sensitive to outliers, if present in the data given.

c. Linear regression also assumes that the data are independent i.e., the scores of one subject have nothing to do with those of another. This is often sensible but not always (clustering in space and time).

d. When observing the data values, I observed that the values for "thunder" are all zeros, which does not provide any value to our analysis.

e. There is also a limitation for Mann Whitney U test and a reason why this test was used in the analysis. Mann Whitney u test is non-parametric test. These tests have a lower power than parametric tests. This means that if there is really a difference between two groups, non-parametric tests are less likely to find it. The reason this test was used in the analysis of NYC subway data is because the data set is small. As central limit theorem assumes normality in most cases if the sample size is large, these tests are best used for small data sets.