Section 0. References

http://pandas.pydata.org/pandas-docs/stable/dsintro.html

http://docs.scipy.org/doc/numpy/reference/arrays.ndarray.html

http://docs.scipy.org/doc/scipy/reference/tutorial/linalg.html

http://docs.scipy.org/doc/scipy/reference/tutorial/stats.html

http://synesthesiam.com/posts/an-introduction-to-pandas.html

https://docs.python.org/2/tutorial/controlflow.html#lambda-expressions

http://www.bogotobogo.com/python/python_functions_lambda.php

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

http://stackoverflow.com/questions/419163/what-does-if-name-main-do

http://en.wikipedia.org/wiki/Ordinary_least_squares

http://www.bigdataexaminer.com/14-best-python-pandas-features/

http://bconnelly.net/2013/10/summarizing-data-in-python-with-pandas/

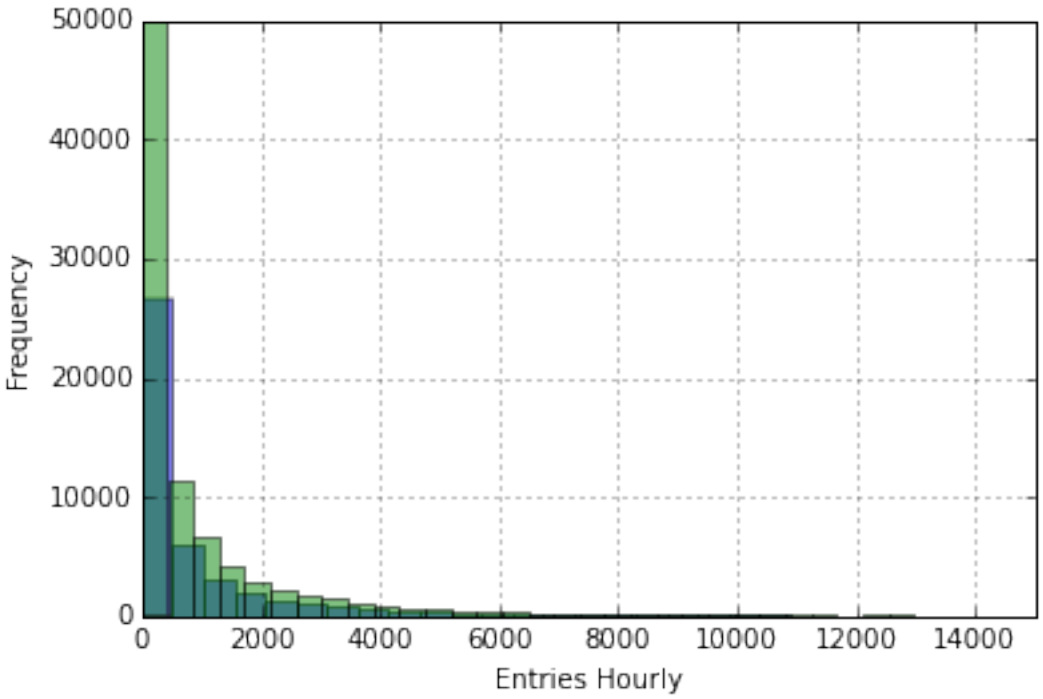
http://blog.yhathq.com/posts/logistic-regression-and-python.html

Section 1. Statistical Test

1.1
For our NYC Subway data set we propose the question does there exist difference in ridership on rainy days as opposed to non-rainy days. To answer this question we implement a Mann-Whitney U statistical test (MWUt). The MWUt method in python's scipy.stats library tests a one-sided hypothesis and thus yields a one-sided p-value by default. In order to receive the correct p-value corresponding to our two-sided hypothesis (that the samples are not equal rather than one greater or less than the other) we simply multiply by two the p-value output by the function. Our null hypothesis is therefore that the two samples are representative of the same population and our p-critical value is 0.05. The alternative hypothesis is that they are not from the same population.
The comparison values used for the two samples (rainy vs. non-rainy) in the MWUt are the hourly entries across all subway stations (which are referred to as ridership) for the month of May, 2011 excluding the 31st.

1.2
The MWUt is a non-parametric test which assumes no underlying probability distribution and is able to discern with some level of confidence whether or not two samples come from the same population. It

is more accurate than the t-test for non-normal distributions. A histogram overlay plot of ridership on rainy vs. non-rainy days clearly shows that ridership does not follow a normal distribution and resembles more an exponential distribution. Thus our MWUt is the appropriate choice of statistical analysis.



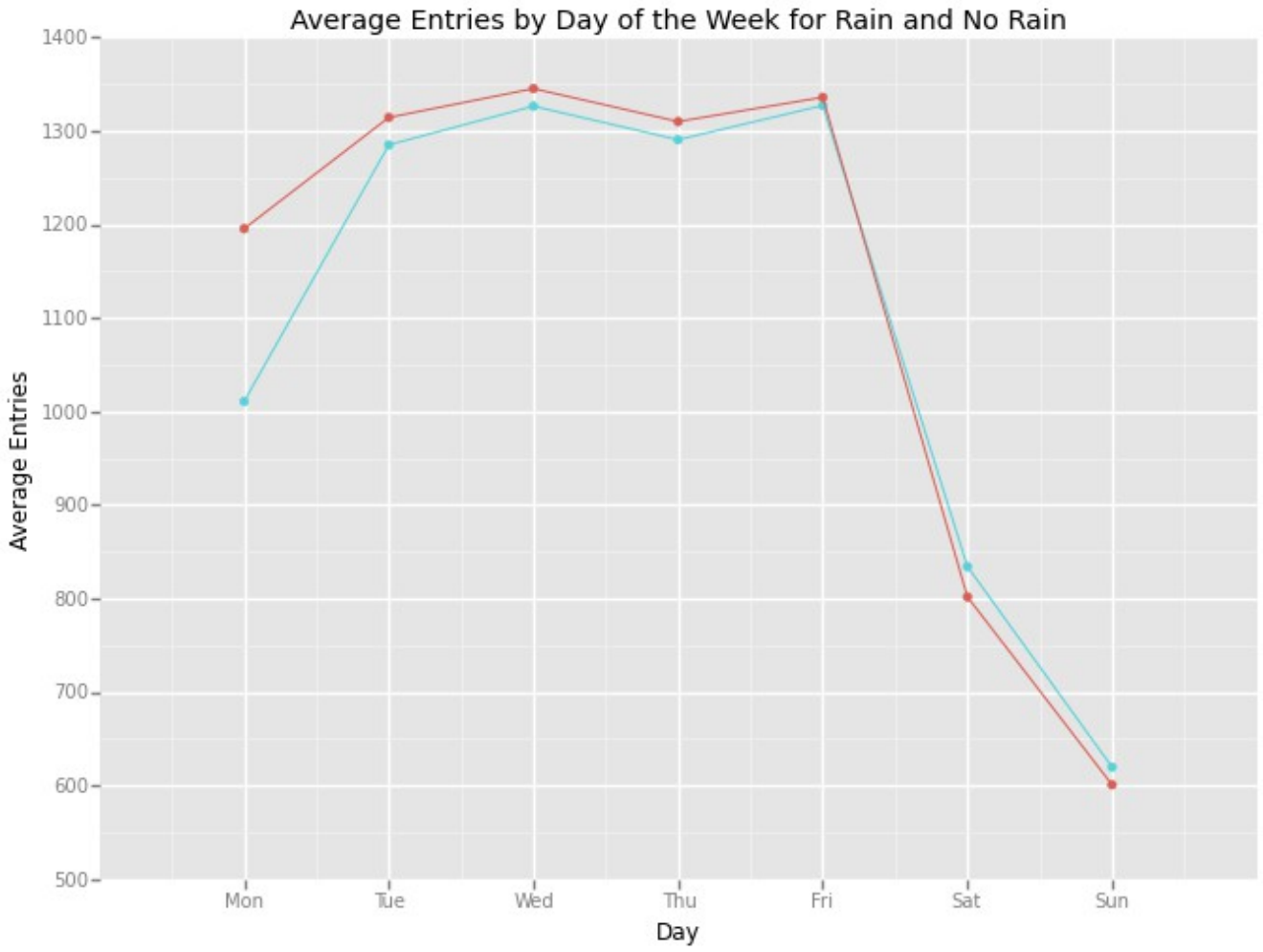
The chart above is a histogram that illustrates the distribution of ridership on rainy and non-rainy days. The number of hourly entries are on the x-axis while the frequency is on the y-axis. The green bars correspond to non-rainy days and the purple bars correspond to rainy days.

1.3

The results of the MWUt are a U test statistic of 1924409167.0, a one-sided p-value of **0.0193,** mean ridership with rain **1105.446**, and mean ridership without rain **1090.279**. In order to get the correct two-sided p-value for our non-directional null hypothesis we multiply the p-value given by the MWUt to get **0.03862**

1.4

From our MWUt test the two-sided p-value is less than 0.05, so we reject the null hypothesis in favor of the alternative hypothesis. These results lead us to conclude that ridership is not the same on rainy days as opposed to non-rainy days. A plot of average ridership on rainy days vs non-rainy days by day of the week helps to illustrate this finding. Although we can now see that there appears to be an anomaly with weekend travel that may allow for another question.



The plot above shows average entries by day of the week for rain and no rain. The red line corresponds to non-rainy days and the blue rain corresponds to rainy days.

Section 2. Linear Regression

2.1
I created my own OLS Regression utilizing numpy arrays and numpy methods. Analysis was performed using matrix linear algebra of which equations are provided in the wikipedia page for OLS under estimation.

2.2
The features used to predict ridership are hour, rain, precipitation, mean temperature, as well as dummy variables which were created in the dataframe that correspond to each subway station. A value of 1 was assigned to the variable if data for that entry was from that particular subway station, otherwise 0 was assigned.

2.3
The choice of features was based upon intuition but also upon previous analysis done with gradient descent. A calculation of the coefficient of determination with and without the dummy variables that

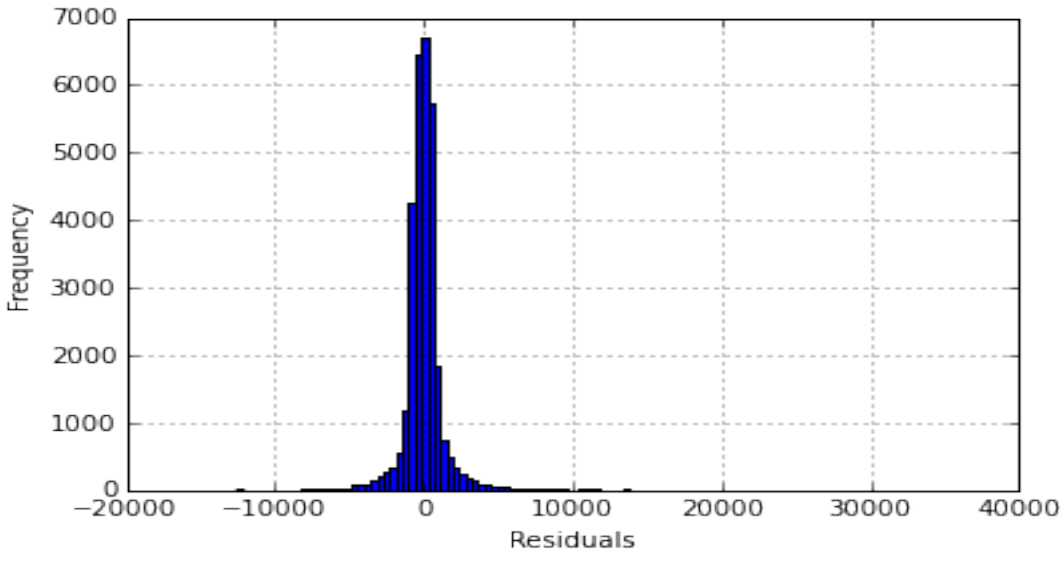
correspond to subway station showed that station was the strongest predictor of ridership. This is also intuitive if we can imagine the higher traffic around places of work or business as compared to neighborhoods. After this, hour seemed to be a big factor as traffic would vary during particular times of the day as opposed to others (rush hour 9am or 5pm vs 4am). Based upon our MWUt we were able to determine that there is significant difference in ridership due to rain and it follows that amount of precipitation could also play a factor. For this reason I included both of these variables. Lastly, I chose temperature because I hypothesized that temperature might be cooler in the subway than above ground (heat rises and below ground is usually cooler) and on hot days people might prefer to ride in the subway rather than walk in the hot May sun.

2.4
The coefficients for both rain and precipitation were **0**. The coefficient for hour was **71.19**, and the coefficient for mean temperature was **-207.19**. The hour was positively correlated with ridership whereas temperature was negatively correlated (against my hypothesis). We might infer that the subway temperature is warmer than the above ground temperature.

2.5
The coefficient of determination (R-squared value) is **0.479**.

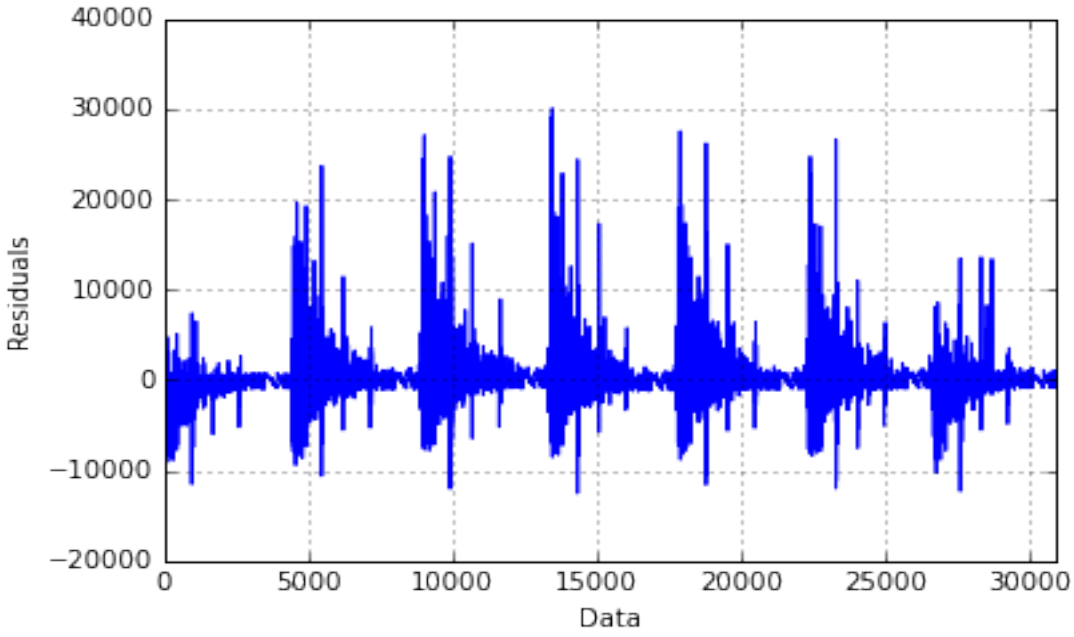
2.6
R-squared values indicate how well our model (and thus chosen features) fit the data. In more precise terminology it is how well our predictors (subway station, hour, rain, precipitation, temperature) explain the outcome variable (hourly entries). When multiplied by 100, the R-squared value is a percentage of how much the variation in our outcome variable is being explained by the variation of our model. Considering our particular R-squared value, you could say that our model roughly accounts for 50% of the variation. Given that we are trying to predict human behavior I would say that our linear model does a fairly decent job of predicting ridership for this data set.

To gain a deeper understanding of our model we take a look at the residuals.



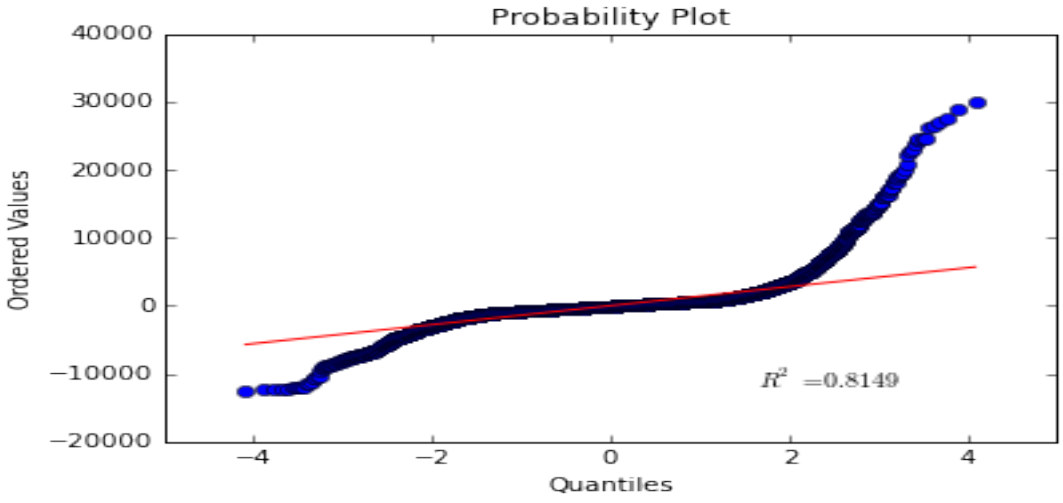
This histogram (with bin size 100) depicts the frequency of residuals on the y-axis and the amount of the residual on the x-axis for a subset of the data.

Glancing at the residuals from our model, it appears that they resemble some sort of skewed normal distribution. There is a zero mean, although we would expect this as OLS regression minimizes the sum of squared errors. The tails of our distribution however are quite long. Another expectation of a good model fit would be the appearance of homoscedasticity (or random noise) if we were to plot each residual in a sequence.



The plot above shows residuals plotted on the y-axis along side each data point on the x-axis for a subset of the data.

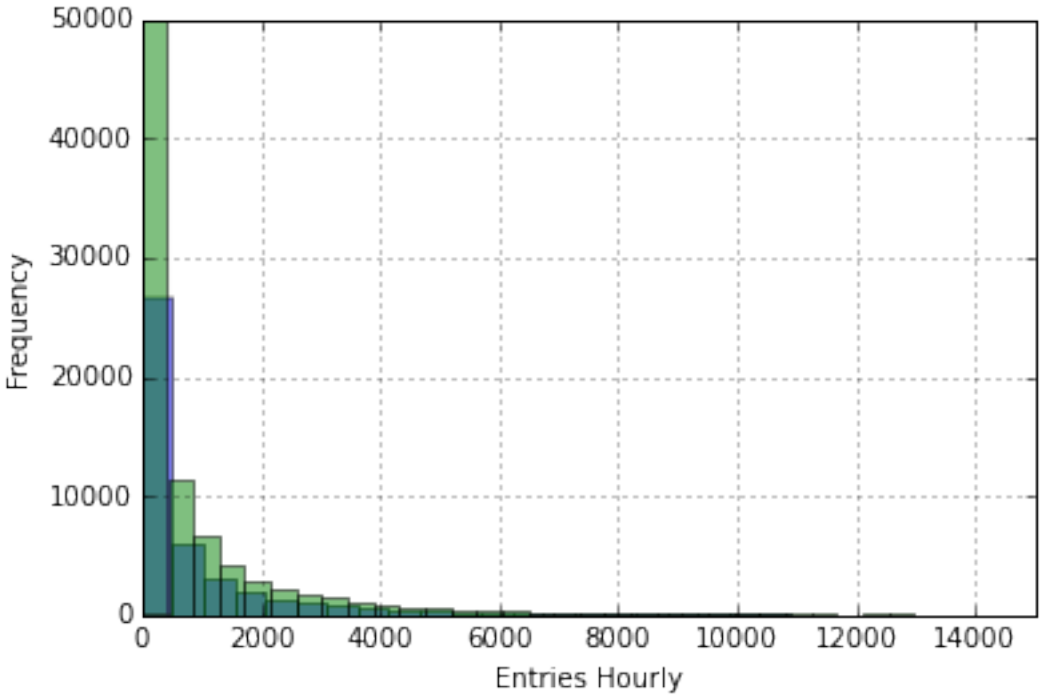
We can see that our residuals do no exhibit randomness but rather a cyclical pattern. This fact indicates a poor model fit and a fundamental flaw in model selection. Further comparison of our residuals with a normal distribution in a probability plot also confirms this.



A probability plot of our residuals (blue) along side a normal distribution (red).

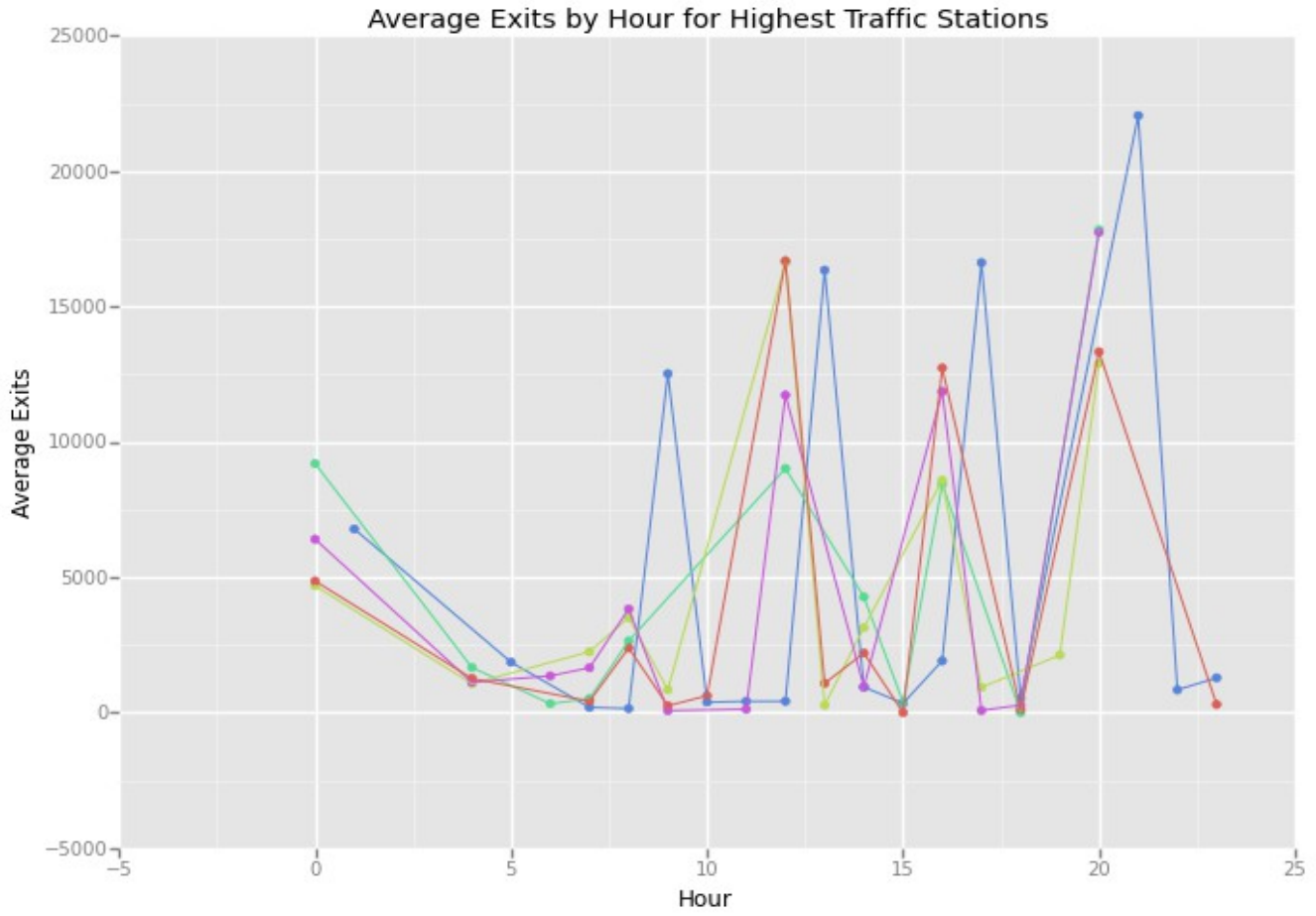
Section 3. Visualization

3.1



This figure shows a histogram of the frequency of entries hourly for rainy (purple) and non-rainy days (green) with 100 bins.

3.2



This plot shows the average hourly exits by time of day for the highest traffic stations (99 percentile), There are 5 stations included in this plot, the interesting thing about this plot is that we can see the different patterns of ridership for each station. The peaks in ridership for the blue station exhibit a typical morning and lunch rush hour and we might infer here that this station is a high traffic exit point for people going to work.

Section 4. Conclusion

4.2

According to the p-value of our statistical analysis by way of a MWUt, the ridership on rainy days for the NYC Subway data set during the month of May, 2011 is significantly different than the ridership on non-rainy days. However, the average difference as seen from our analysis in section 1 is about **15** entries per hour which is a little more than **1%** of the total ridership for either of the two samples. As such, although it is statistically significant, it is not necessarily a good predictor of ridership as a whole for the entire data set.

The results of linear regression outlined in section 2 support this as we were able to see that our weighted coefficients for the predictors of rain and precipitation were calculated to be zero. In fact when these variables were removed as predictors from the features dataframe, our R-squared value remained unchanged supporting the insignificance of these. In conclusion, we admit that there is a statistically significant difference of ridership on rainy days as opposed to non-rainy days however rain and precipitation as factors in determining ridership are insignificant.

Section 5. Reflection

5.1
There are several possible shortcomings with regard to the data set that I noticed. There were some turnstile numbers missing (if there is in fact a complete sequence numbered from 1 to 552) and also several turnstiles with very limited data. These two shortcomings could affect the value of our weighted coefficients by either not weighting enough for some turnstiles or not weighting at all for turnstiles that do in fact exist. Another possible factor in ridership that was not accounted for could be holidays that took place during May such as Cindo de Mayo, Mother's Day, and Memorial Day.

The statistical analysis of rainy vs. non-rainy ridership leaves a further question. Our graph from section 1.4 of mean ridership by day of week for rainy and non-rainy days shows that there is a reversal of trend for the weekend. This poses another question as to why ridership behavior changes on the weekend with rain. The answer could be that there is not enough data for rainy weekend ridership.

As demonstrated in section 2, the residuals of our OLS model show that it is imperfect. The implications of an imperfect model are that some predictions may be relatively accurate while others are entirely off. Further, because the residuals exhibit a pattern, we are consistently inaccurate in a predictable way and this leaves room for improvement with a non-linear model. As we saw in section 3, hour doesn't necessarily follow a linear relationship with ridership. It could be that a polynomial model is more appropriate. Whatever the model choice, we would expect a good model to have residuals that exhibit Gaussian behavior.

5.2
If we wanted to test a hypothesis that rain increases erratic ridership behavior we could have used a parametric test such as the Brown–Forsythe test to see if variances from the median were significantly different between the two samples. Also, using the updated data set with latitude and longitude, there could be room to perform a more complex analysis such as spatial correlation or k-means clustering. With this we could determine if proximity (towards Manhattan for example) is a significant factor in ridership.