For this project, I used the original data set on a local machine.

Section 0. References

http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

http://www.datarobot.com/blog/ordinary-least-squares-in-python/

http://blog.yhathq.com/posts/ggplot-for-python.html

http://people.duke.edu/~rnau/rsquared.htm

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

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I used the Mann-Whitney U-test to compare ridership during rainy and non-rainy days. Since I did not make a prior assumption on whether rain or no-rain increased subway ridership, I used a two-tailed p-value. The set of hypothesis is as follows:

$$H_0: Distribution\ of\ ridership_{rain} = Distribution\ of\ ridership_{no\ rain}$$

$$H_A: Distribution\ of\ ridership_{rain} \neq Distribution\ of\ ridership_{no\ rain}$$

Namely, the Mann-Whitney tests whether ridership of the subway when it is raining comes from the same population as ridership of the subway when it is not raining. Depending on the result of the statistical test, we can conclude whether or not rain is a statistical factor in determining ridership. I will use the commonly accepted p-value of 0.05.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Since I was comparing paired data, I used the Mann-Whitney U-test in favor of the t-test because the Mann-Whitney test does not assume the distribution of the data, whereas the t-test requires that the paired sets of data are normally distributed.

In order to determine what statistical test I could use to analyze the data, I first had to determine the distribution of the ridership data for both rain and no-rain. Using the result of the histogram plot of the two data sets (see figure 3.1), I concluded that the two were similar in nature: they were not normal distributions, and it skewed heavily to the right. This ruled out the t-test, and the Mann-Whitney test is applicable in this scenario.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

P-value = 0.0386

$\mu_{rain} = 1105.446$
$\mu_{no\ rain} = 1090.279$

1.4 What is the significance and interpretation of these results?

The p-value returned by default is for a one-tailed hypothesis test, so for my two-tailed test, I had to multiply it by 2 for a result of 0.0386. The result is less than my p-critical value of 0.05 which allows me to reject the null hypothesis in favor of the alternative hypothesis. The alternative hypothesis says that the distribution of ridership when it is raining is **not** the same as the distribution of ridership when it is not raining. This allows for the conclusion that rain has an effect on ridership. The higher mean of ridership when it is raining would suggest that ridership increases when it rains.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?

I used the statsmodels OLS to obtain the theta coefficients.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

I experimented with a variety of methods in choosing my feature combination of 'Hour', 'mintempi', 'meanwindspdi', and 'meantempi.' The dummy variables were the 'UNITS,' which could be thought of as the individual stations turnstiles.

2.3 Why did you select these features in your model?

In my first approach, I ran the linear regression with each variable on its own and ranked them by their r^2 values, and focused on the top four combinations. I realized that the R2 increased minutely with each additional feature in the model. Since this model did not include a penalty for having too many features, I wanted to focus on just the top four features, and features that did not have coefficient p-values that was above 0.05.

Features	Theta	R2 value
Hour	67.395	0.4575
mintempi	-10.132	0.4191
meanwindspdi	29.886	0.4190
meantempi	-7.598	0.4188
minpressurei	-308.344	0.4187
maxpressurei	-302.384	0.4186
meanpressurei	-271.987	0.4186
maxdewpti	-2.493	0.4184
precipi	40.867	0.4184
rain	14.804	0.4183

I tried other combinations, including 'rain' and 'precipi,' but substituting out the top four only decreased the R2 value. I also tried other methods, such as starting with all features and eliminating the features one by one by the greatest coefficient p-value. In the end, I settled on my first approach.

2.4 What are the parameters of the non-dummy features in your linear regression model?

Feature	Coefficient
constant	1007.930
Hour	67.411
mintempi	-20.599
meanwindspdi	24.022
meantempi	13.525

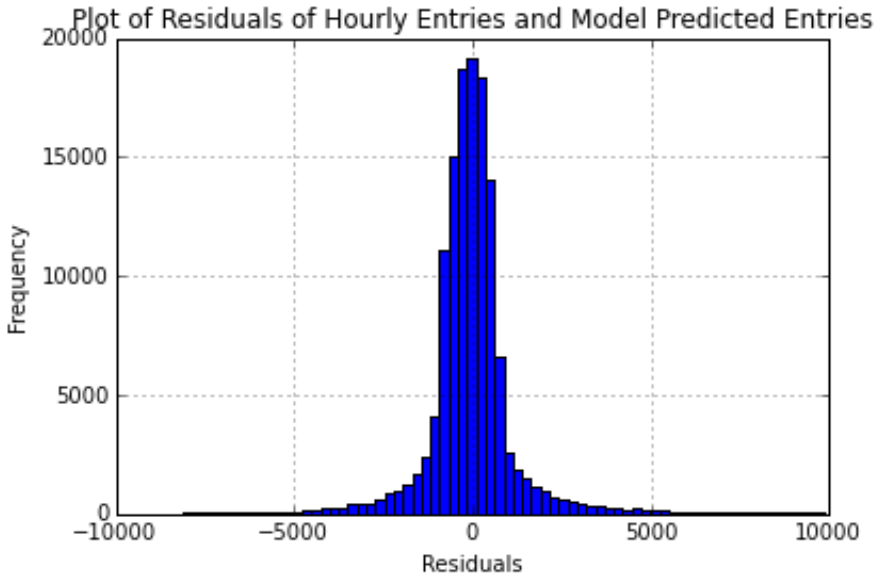
2.5 What is your model's R2 (coefficients of determination) value?

R2 = 0.4588

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

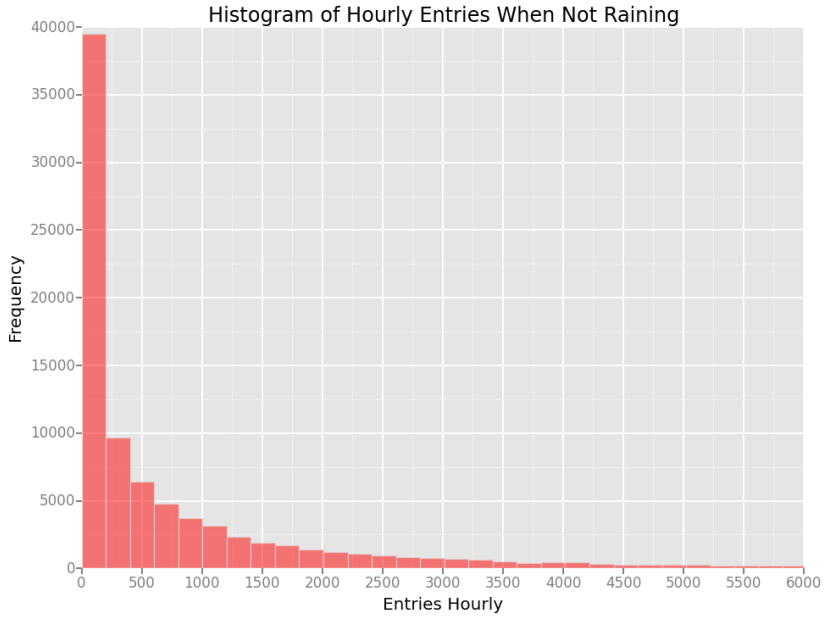
The R2 value of 0.4588 represents the percentage of the response variable variation that is explained by the linear regression model. This means that roughly 46% of the variation between the modeled and actual hourly entries can be explained by the linear regression model, namely the features that I selected in the model. Based on the R2 value alone, it would suggest that the model does not appropriately predict ridership. Although predicting accurate hourly entries may not be feasible using this model, it still serves a purpose in identifying significant features and describing their relationship with ridership.

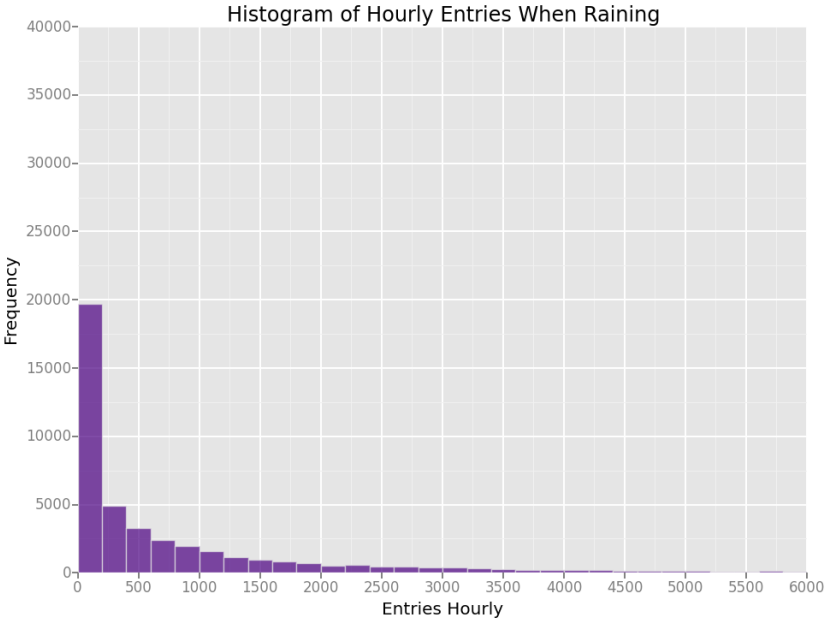
To further assess whether the model is appropriate, it is important to consider other factors in addition to the R2 value. The residual plot of actual and predicted ridership helps in this assessment. A good fit model should have a residual plot that resembles a normal distribution. A close inspection of the residual plot for the model (pictured below) reveals that the residual distribution has long tails on both sides, with residuals extending out to nearly 40,000 in difference. This leads to the conclusion that the residuals are not normally distributed, and that the model is not a good fit for the data.



Section 3. Visualization

3.1 Two histograms: ENTRIESn_hourly for rainy days and ENTRIESn_hourly for non-rainy days

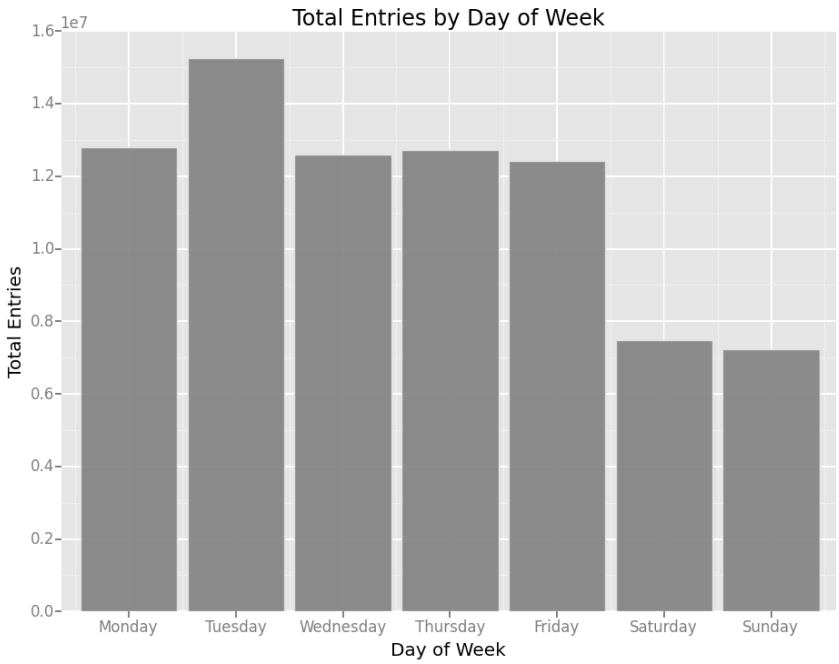




The two histograms that represent the frequency of hourly turnstile entries show the similarity of the two samples in the shape of the distribution. The distribution of both sets of data is heavily right skewed towards less entries per hour. This prevented the use of t-test for normal distributions in favor of the Mann Whitney U-test. The data points for ‘no rain’ outnumber the data points for ‘rain.’

3.2 Freeform visualization

In this visualization, I used the fixed data from 'turnstile_weather_v2.csv'.



In this bar plot, the 'ENTRIESn_hourly' field was summed over all the data by the day of the week. The purpose of the graph is to compare the volume of ridership for each day of the week. This graph shows that in the month of May in 2011, Tuesdays had the most turnstile entries at just over 15 million entries. It is also interesting to note that the entries are somewhat consistent over the weekday, and drops off by nearly half during the weekend. What this graph fails to account for is the number of a particular weekday that occurs in the month of May in 2011. As it happens, there is one more Monday, Tuesday, and Wednesday in this month than the other days.

Section 4. Conclusion

4.1 and 4.2: From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining? What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

From my analysis of the NYC subway data in May of 2011, the data in this time period suggests that more people did ride the NYC subway when it was raining versus when it was not raining.

The result of the Mann Whitney test suggested there was a statistical difference in the distributions between subway ridership when it was raining and not raining. The means of the ridership between these two data sets showed that ridership was greater when it was raining than when it was not. Given that these distributions were statistically different and the mean of ridership when it was raining being greater than when it was not raining, it can be concluded that rain has an effect of increasing ridership.

The linear regression model also gives insight into the relationship of ridership and rain. When I performed the linear regression with rain as the only variable, I got a coefficient of 14.804. Since 'rain' was a Boolean value, with 0 representing no rain and 1 representing rain, the positive coefficient shows that the model associates an increase in ridership when it is raining. The coefficient of 14.804 represents the mean change in hourly entries per each unit of change in the 'rain' variable, and since rain is a categorical variable, it suggests that rain increases ridership by about 15 entries whenever it rains.

These two tests in combination solidifies the conclusion that ridership of the NYC subway in the month of May 2011 increased when it was raining.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1) Dataset

The data consists of NYC ridership in the month of May of 2011. The weather during this month was mild, with a lowest of 46 degrees and a highest of 86 degrees. It would have been interesting to see the effects of extreme weather conditions, such as snow or extreme heat and their effects on ridership. Given a mild weather month, the effect of weather on ridership did not seem amplified, or significant enough for a strong correlation.

The short duration of the sample data also could lead to false conclusions. May of 2011 may not have been a representative sample size for ridership behavior. For example, students at universities or high schools may not take the subway for half the month because the school year is over, or there could have been a subway strike during this time period. A "rainy day in May" may not have the same effect on subway ridership behavior as a "rainy day in January," because the prior may be desirable, whereas the latter is less desirable to walk in. A longer period of data could lead to a more representative model to evaluate the effect of rain on ridership.

2) Analysis, such as the linear regression model or statistical test.

The regression model tried to create a model that could fit all stations (UNITS) at all times of the day. I think a more interesting model would look at one station over a longer period of time, say two years. I think this would have given more insight into the effects of weather conditions on a given station. I surmise that the variance in station entries attributed to the low R2 value. I also noticed that the 'Hour' feature was the best predictive feature when compared to features concerning weather. Given that there are peak ridership times, limiting the dataset to Manhattan stations during normal business hours could have reduced the variance in the model and led to more concrete evidence of weather on ridership.

The linear regression method also has inherent shortcomings. Some of these shortcomings could include overfitting the data. Some of the features seemed redundant, such as precipitation and rain, or mean temperature, max temperature, and min temperature. These features could be dependent in nature, leading to overfitting or bias. Perhaps the data in non-linear, in which case the linear model cannot accurately model the behavior, leading to underfitting.