0. References for project:

http://synesthesiam.com/posts/an-introduction-to-pandas.html#disqus_thread: Introduction to the pandas library, and useful as a reference on how to manipulate data frames. In addition, the website helped clear doubts on how to properly use the strptime function, and apply/lambda
http://www.stat.purdue.edu/~tqin/system101/method/method_two_t_sas.htm : Was used as a

http://stats.stackexchange.com/questions/164542/is-time-in-linear-regression-a-categorical-or-continuous-variable : Wanted to determine whether 'hour' should have been used as a dummy variable or not.
http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm : Used to review correlation coefficient and coefficient of determination.

Weiss, Neil. *Introductory Statistics*. 9th ed. Boston: Pearson, 2012. Print. : Used as a reference for the inferential statistics portion of the project

1. Introduction and Explanation of Statistical Methods

1.1 The goal of this project was to determine whether there is a statistically significant difference between the average number of subway riders on days when it rains versus days that it does not. This was accomplished by making use of ridership data from May of 2011, and applying the Mann-Whitney test. The reason for this is that as noted by Weiss in Introductory Statistics, The Mann-Whitney test should be used when trying to determine whether the value of a population parameter is different for two populations, in cases where the two population parameters have the same distribution. The reason the test is justified in this case is that upon plotting, both samples exhibit a strong right skew. Due to the fact that our samples are extremely large, with data available for a large number of turnstiles in the NYC region, we can infer that the populations of ridership in no rain and when it rains follow a similar distribution.

In addition, because we are interested in deviation in either direction, a two tailed test is appropriate. Average subway ridership for the year on days when it does not rain is represented using μ_1, while μ_2 is used to represent mean subway ridership on days when it does. The null (H_0) and alternative (H_1) hypotheses are as follows

H_0: Mean subway ridership is the same on days when it is dry versus days when it rains; $\mu_1 = \mu_2$
H_1: Mean subway ridership is different on days when it is dry versus days when it rains; $\mu_1 \neq \mu_2$

Because we are dealing with a two tailed test, a p critical-value of 0.025 is established for each tail, corresponding to a 5% significance level.

1.2 In order to be able to apply the Mann-Whitney test, we must assume that the two populations share a similar distribution. Looking at the histogram of a sample of ridership data for the two cases, we can see this is in fact true. Both samples having a strong right skew. From our samples, we are inferring that both populations also have a strong right skew for their Average Hourly Entries.

1.3 When the Mann-Whitney U test was implemented, the U statistic was calculated to be 153635120. Additionally, the mean for hourly entries when it was not raining was calculated to be 1845, while the mean for hourly entries when it was raining was 2028. Upon attempting to calculate a p-value, the program returned “nan”. Upon researching the problem, I discovered that the issue was discussed in the following thread on Udacity:

https://discussions.udacity.com/t/mann-whitney-u-test-on-improved-dataset-yields-p-nan/4470

Based on the suggestion by an Udacity coach, the p value from the Udacity IDE was used in the analysis, which was calculated to be approximately 0.024999912. Because we are interested in conducting a two tailed test, the p-value must be multiplied by two, resulting in the p value being 0.049999824.

1.4 Because the p-value is less than the p-critical value of 0.05, we can reject the null hypothesis in favor of the alternative hypothesis at the 5% significance level. We conclude that at the 5% significance level, there is evidence to suggest that there is in fact a significant difference in the hourly entries for the two populations.

2. Linear Regression

2.1 Because the data set for subway ridership is not too large, linear regression was implemented using OLS in Statsmodels.

2.2 The features that were selected in running the regression were “rain”, “hour”, and “meantempi”. The dummy variables that were included were ‘Unit’ and ‘weekday’.

2.3 Each of the factors helped increase the adjusted R^2 value, with the majority of the increase coming from when the ‘hour’ and ‘unit’ columns were included in the analysis. However, the ‘weekday’ column did help increase the R^2 value a bit, and ‘meantempi’ and ‘rain’ did so marginally.

2.4 The coefficients for the non-dummy features in the regression were as follows:

Hour	122.39
Precipi	-38.034
Meantempi	-14.988
Weekday	991.6411

Constant	860.7791

2.5 The adjusted R^2 value is 0.480, and the R^2 value is 0.483.

2.6 The coefficient of determination means that 48%, of the variance in Hourly entries is determined by the features and dummy variables. This coefficient of determination value is indicative that a linear regression model is not a highly accurate mode of representation for the data.

3. Visualization

3.1 Histograms for the frequency of Hourly Entries – No Rain and Rain:

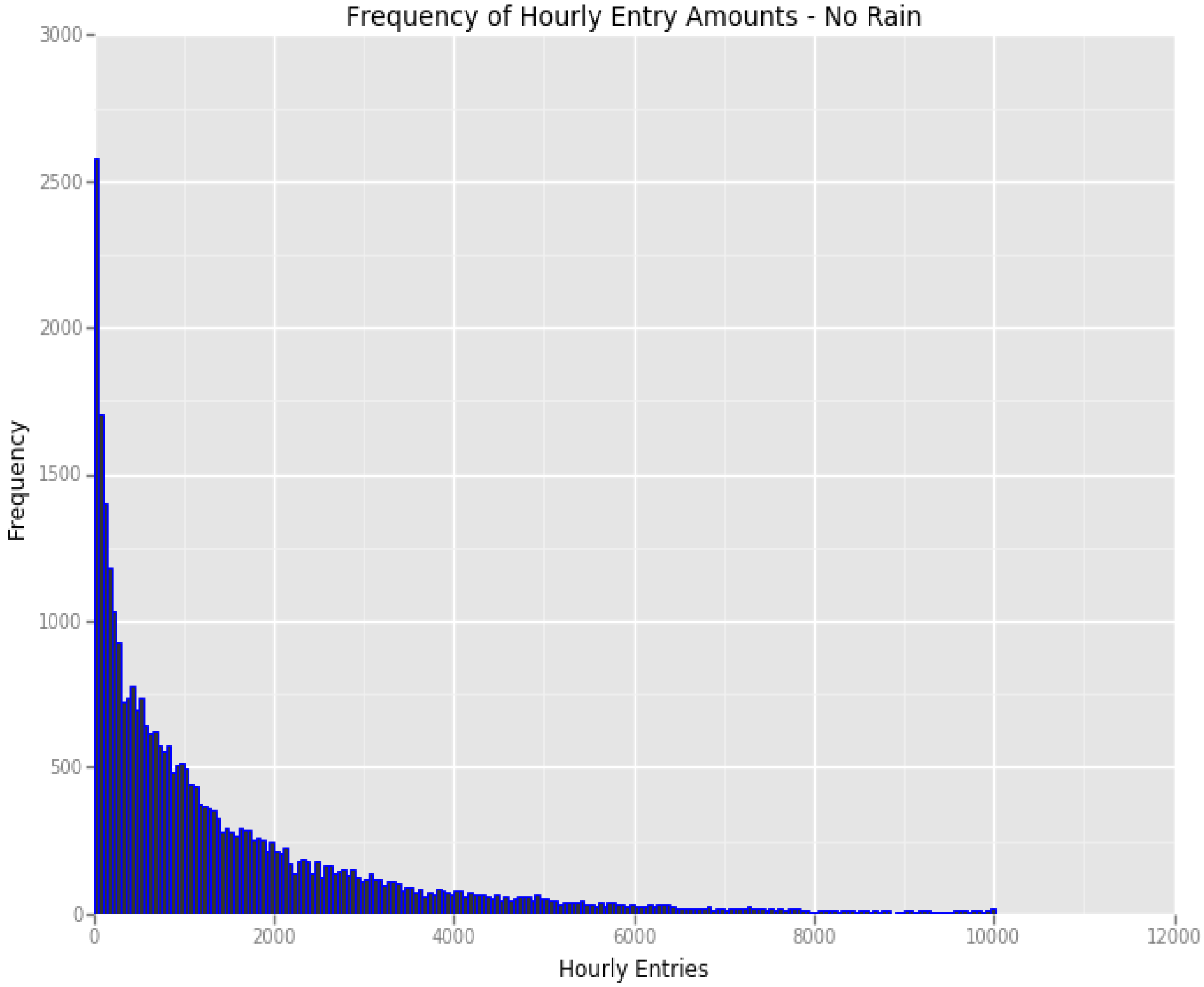

Figure 1. Histogram of the frequency of hourly entries on days when there is no rain. Bin width is set to be 50, and the first bin includes days when there were 0 entries. The data is limited to hourly entries up to 10,000.

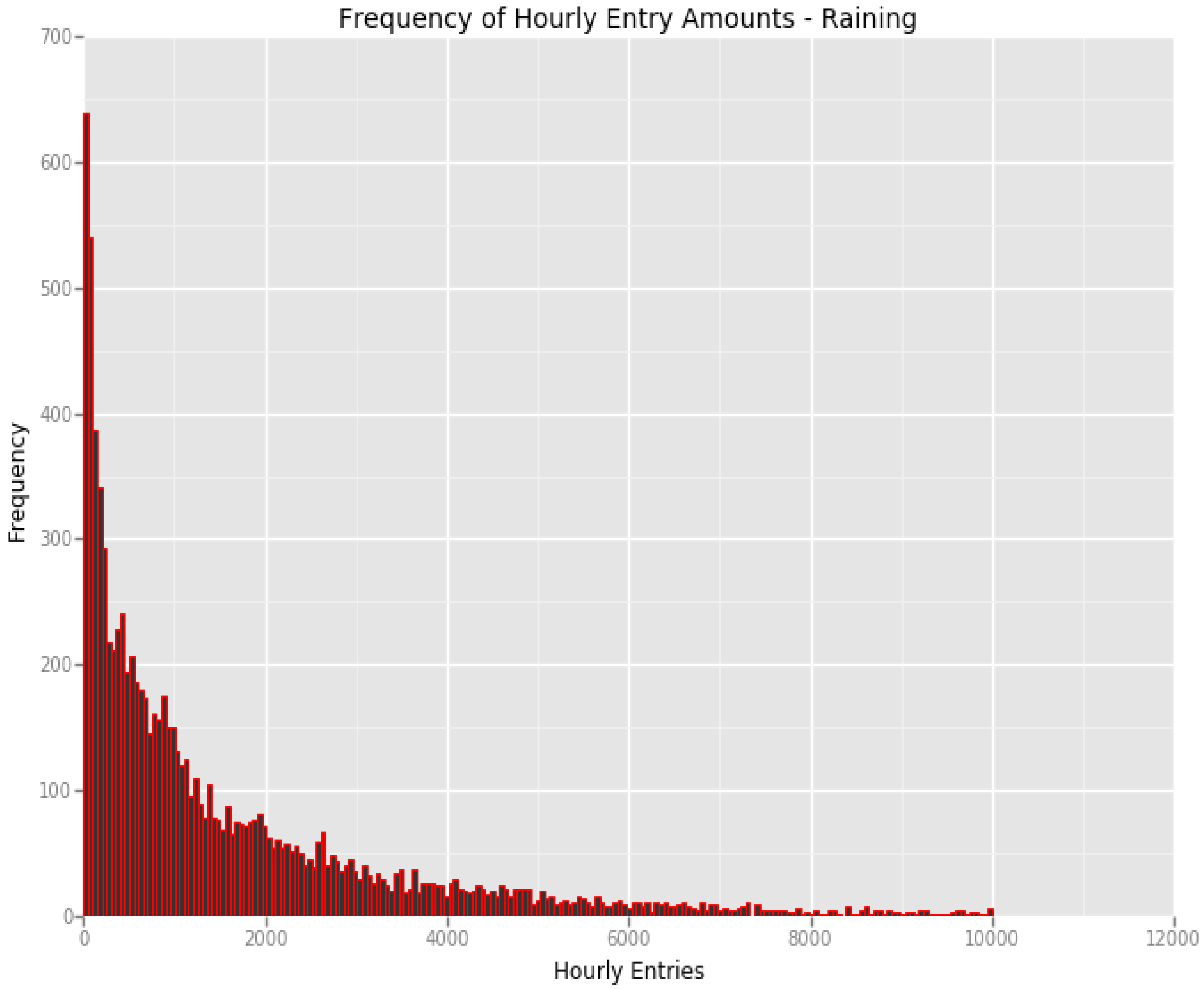

Figure 2. Histogram of the frequency of hourly entries on days when it rains. Bin width is set to 50, and the first bin includes days when there were 0 entries. The data is limited to hourly entries up to 10,000.

3.2 Subway ridership by day of the week:

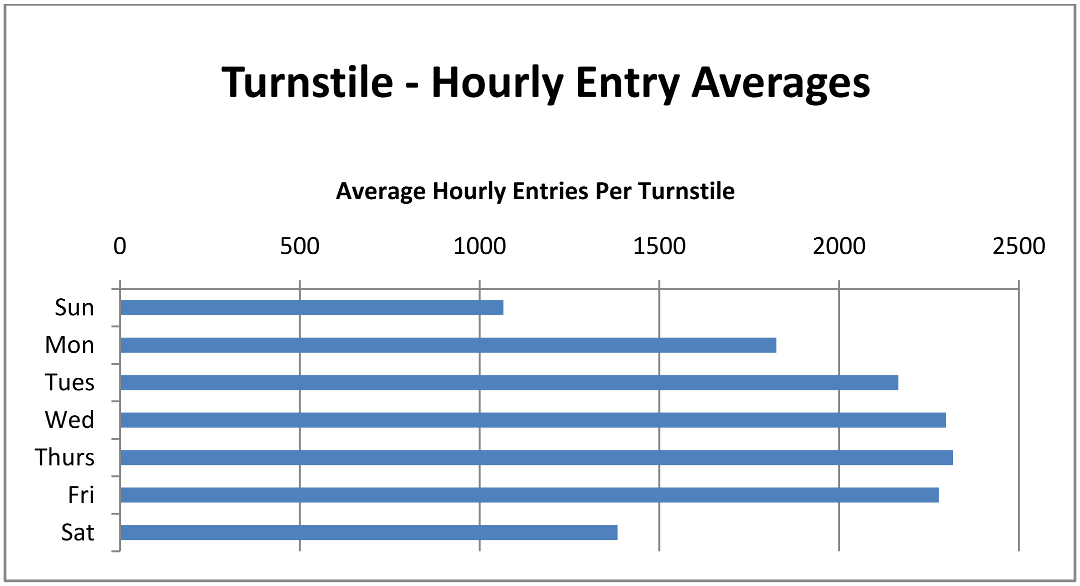

Figure 3. Bar graph for the average hourly entries by day of the week. All analysis of data was conducted in Python, but once the values were obtained, the bar graph was created in Excel, due to some trouble using ggplot in Python

4. Conclusion

4.1 It's possible to reject the null hypothesis in favor of the alternative, that average ridership of the subway differs on days when it rains compared to when it does not. This was shown by use of the Mann-Whitney U Test. Upon making histograms of ridership level for the types of days, it initially appears that ridership is greater when it does not rain, mainly due to the high value of the initial bars. The first bar, however, represents days when there were no entries, and very small number of entries. This distortion would affect our analysis.

4.2 The primary analysis that led to this conclusion was the Mann-Whitney U test. Because the p-value 0.049998024 is less than the p-critical value of 0.05, we can conclude there is in fact a significant difference in the two population means. Although the Mann Whitney U test allowed us to state with certainty that subway ridership differed significantly, the linear regression model did not. The constant for the rain factor in the linear regression model was -38.03, indicating that in the presence of rain, ridership actually went down. However, our analysis also returned a confidence interval of (-89.698, 13.630), meaning that we could be 95% confident that the constant was in that interval. Therefore, it is possible for the constant to be positive.

5. Conclusion

5.1 A potential shortcoming that exists in the dataset is the classification of days when it rains for short periods of time as compared to the entire day. Although a column exists for precipitation amount, there is no way to tell how long it actually rained. Consequently, it is possible that rainfall during one time of the day had no effect on subway ridership during another. Our current classification system does not account for this.

An analytical problem that exists in the linear regression model is the constant -38.03 for rain. Although the Mann-Whitney U test verified that ridership increases when it rains, the regression model implies the opposite. A reason for this could be that a linear model is not the correct form of regression. Other models would have to be implemented to figure which is the appropriate one.