Analyzing the NYC Subway Dataset

Section 0:

http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures/

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

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

http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients

http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis

http://connor-johnson.com/2014/02/18/linear-regression-with-python/

Section 1:

1.1. Mann-Whitney U test is used to analyze the NYC subway data. Two-tail p-value is used. The Null Hypothesis is that the distributions of both groups (i.e. hourly entries with rain and hourly entries without rain) are identical. P-critical value is: 0.05.

1.2. The distribution of ridership in the two samples is not normal. Mann-Whitney U test is a non-parametric test which does not make any assumptions related to the distribution unlike Welch's t test.

1.3. p-value: 2*0.02499 = 0.0499
mean with rain: 1105.44
mean without rain: 1090.27

1.4. This test tells us that, given that the groups are sampled from population with identical distribution, the probability of sampling the data that we sampled is 4.9%. With the small p-value, we can reject the null hypothesis that the difference is due to random sampling, and conclude instead that the populations are distinct.

Section 2:

2.1. OLS using statsmodels was used to compute the coefficients theta and produce prediction for ENTRIESn_hourly.

2.2. The following features were used in the model: 'maxtempi', 'precipi', 'Hour','mintempi','fog'. The name of the 465 turnstile units were used as dummy variables.

2.3. The selection of the features was primarily based on intuition. Maximum temperature, minimum temperature, precipitation and fog are components of weather that may affect people's decision making on whether or not to ride subway. And Hour was chosen because subway ridership varies with time of the day.

2.4. The coefficients of the non-dummy features in the linear regression model are:

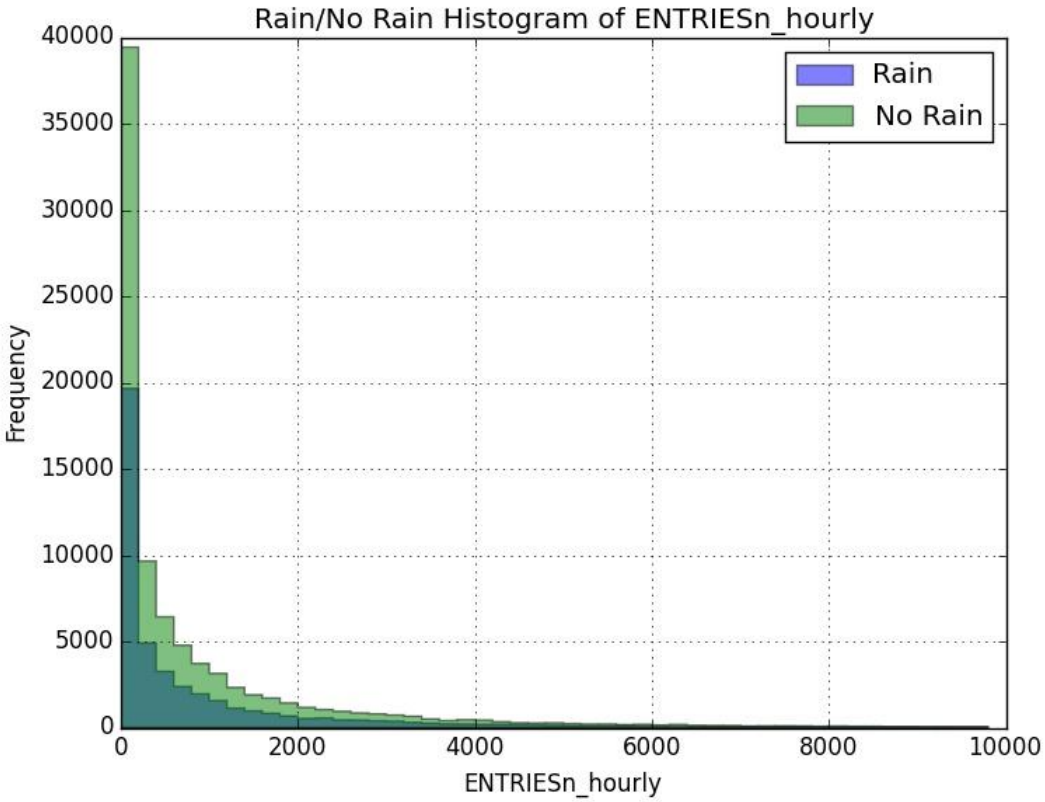
precipi -15.6815 , Hour 429.6614 , maxtempi 19.7553 , fog 92.4650 , mintempi -112.8211

2.5. The models R2 (coefficients of determination) value is 0.48473.

2.6. The R2 value indicates that the model explains around half of the variability of the response data around its mean. This is relatively low and linear model is not appropriate to predict the ridership for this dataset.

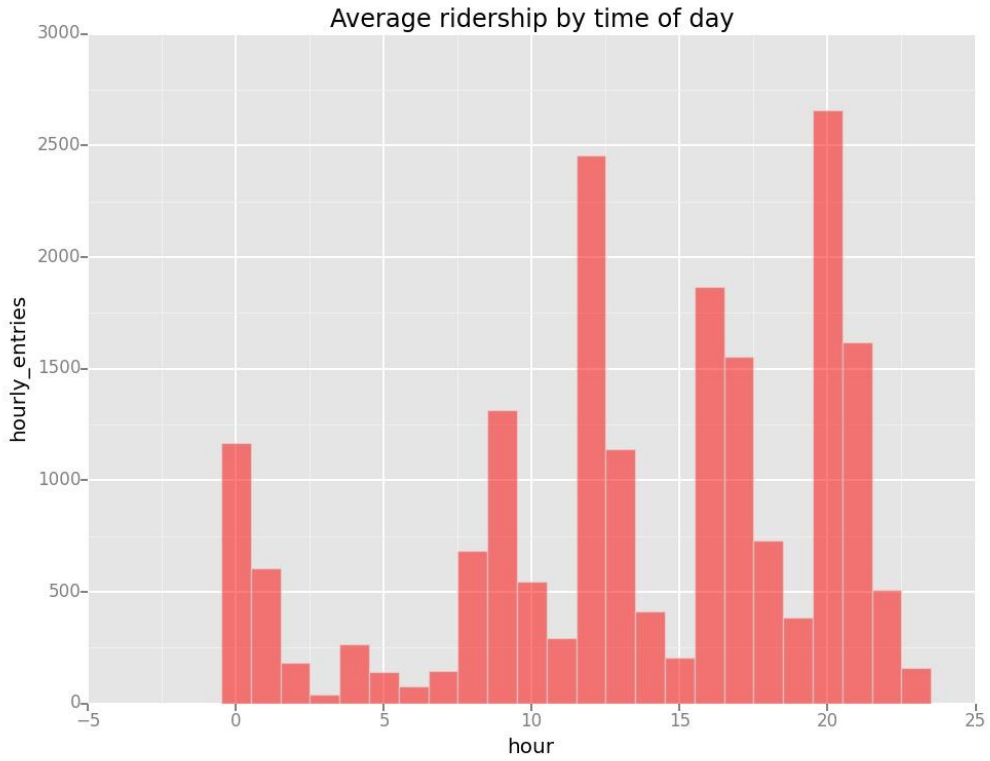
Section 3:

3.1



There is less number of entries for rain than that for without rain in the dataset. The plot shows that the distribution of ridership during rain and without rain is not normal.

3.2.



The ridership in NYC subway varies according to the time of day. The rush hours in the subway are at 9 in the morning, 12 in the noon, 4 in the afternoon and 8 at night.

Section 4:

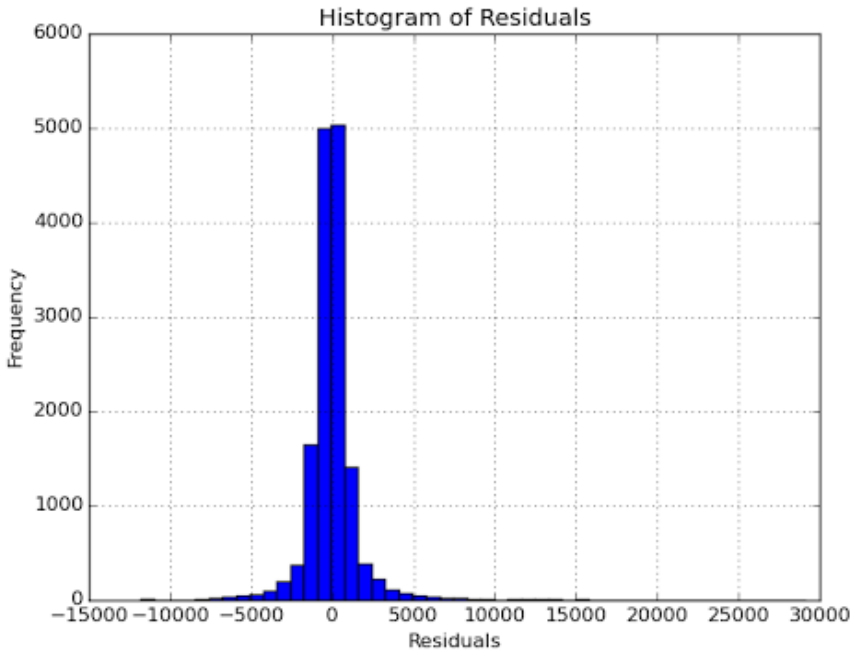
4.1. More people ride subway when it is raining than when it is not raining.

4.2. From the Mann-Whitney U test, we reject the null hypothesis that the groups are sampled from population with identical distribution, and conclude that the populations are distinct. We also observed (from 1.3) that the average ridership during rain is higher than the average ridership without rain. So, I believe that more people ride subway when it is raining than when it is not raining.

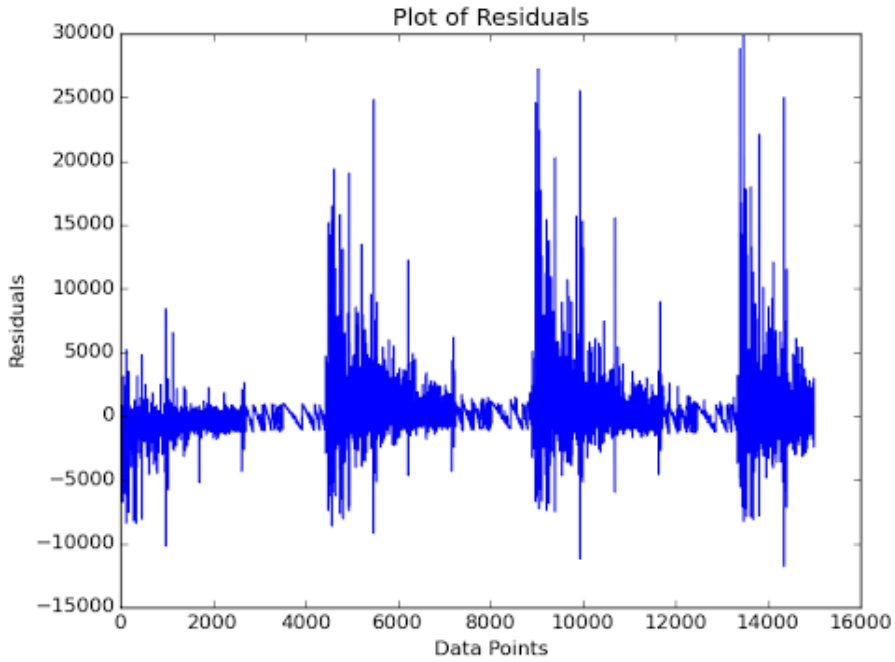
Section 5:

5.1. The dataset contains ridership data only for the month of May and the number of rainy days in the given dataset is less than the number of days without rain. A more balanced data would probably be better for the analysis. Moreover the precipitation data is given for a day and not on hourly basis.

The linear regression model is not appropriate to predict the ridership in this dataset.



From the histogram of residuals, we see that it has a long tail, which shows that some hourly_entries are predicted incorrectly by large margin.



Furthermore, from the plot of residuals per data-point, we see that the residuals are not random and follow a cyclic pattern.