1.1. I use Mann-Whitney U test. The p value is two-tail. The null hypothesis is that the distribution of the number of entries statically the same between rainy & non-rainy days. I set the p-critical value 0.05.

1.2. By plotting the histogram, it is clear that the distribution of the number of entries is not Gaussian. So we can use the non-parametric tests, specifically Mann-Whitney U test, to check whether the 2 populations are the same, ignoring the effects of other variables.

1.3. The one-tail p value from python scipy is 0.02499, so the 2-tail p value is 0.02499*2 ≈ 0.05. The mean of entries with rain is 1105.45; the mean of entries without rain is 1090.28. The sample size of entries with rain is 44104; the sample size of entries without rain is 87847.

1.4. The 2-tail p value is 0.05. By applying Mann-Whitney U test, the null hypothesis would be rejected at any level of significance >= 0.05. So the 2 distributions are not the same.

2.1. I use a. OLS using Statsmodels.

2.2. Features: 'day_off' , 'rain', 'Hour', 'meantempi', 'meanwindspdi','fog','meanpressurei'. 'UNIT' is used as dummy variable. Specifically, if the day is weekend or holiday (Memorial Day), then 'day_off' is 1; otherwise it is 0.

2.3. First, when it is raining, or temperature is low, or the wind outside is big, or there is fog, or the pressure outside is not comfortable, I think people are more likely to take the subway; Secondly, there are definitely more people taking subway at rush hours, like 8am and 6pm. Also less people take subway in weekends and holiday (Memorial Day), so 'Hour' and 'day_off' are selected as features; thirdly, location is also important. If a subway station is close to a big company, more people will take subway in this station. So 'Unit' is selected as feature. Moreover, when I included the 'day_off' in my model, the R^2 increased from 0.4792 to 0.4922.

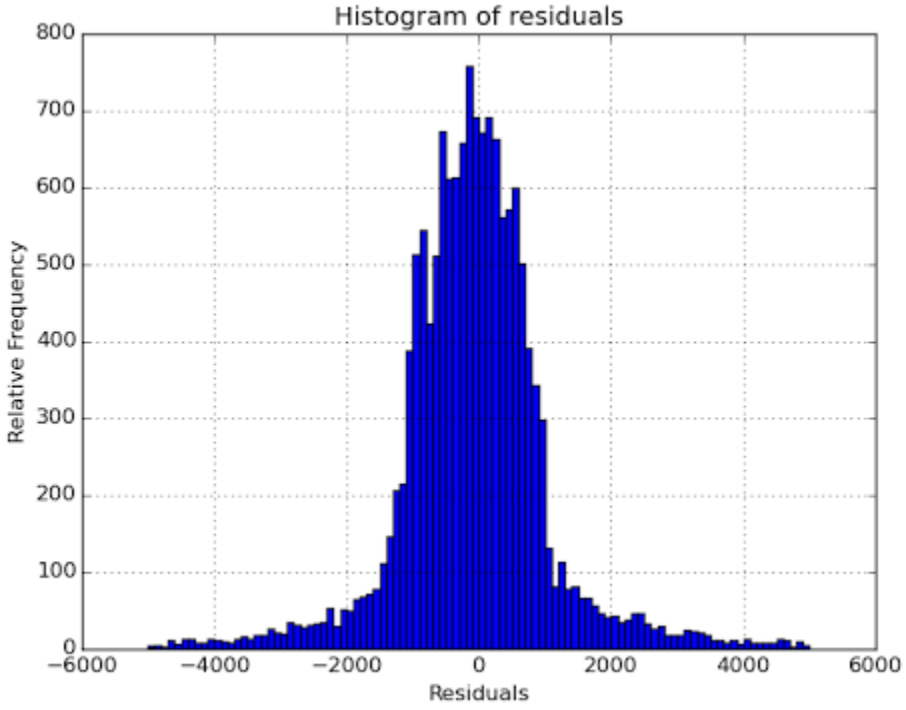
2.4. The parameters are as follows:

'day_off': -581.28; 'rain': 11.56; 'Hour':65.55; 'meantempi': -2.36; 'meanwindspdi': -2.85; 'fog': 28.98; 'meanpressurei': -142.91

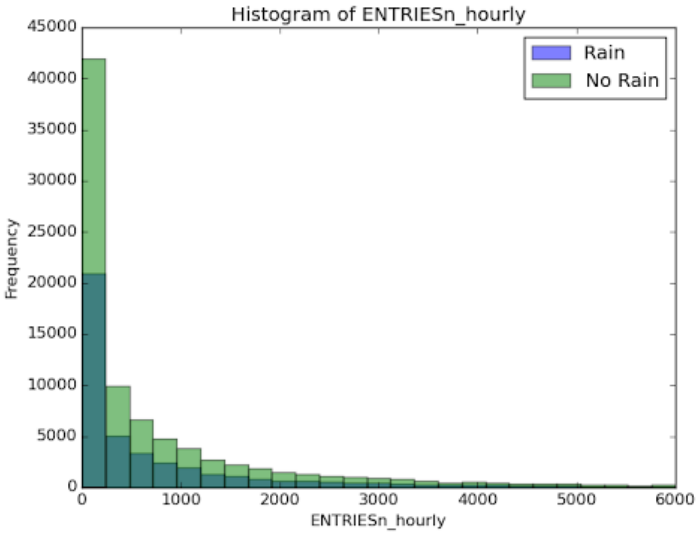
2.5. R^2 is 0.4922 in my model.

2.6. The closer R^2 is to 1, the better our linear regression model is. Given the R^2 equal to 0.4922 in my model, it seems that my linear model to predict ridership is not appropriate for this dataset. About 50% of the original variability has been explained by the linear model; another 50% may be due to nonlinearity between features and the free variables. Another metric for evaluating the model's fit on the training data is to look at residuals plots. We want residuals to be normally distributed around 0. Just as the histogram of residuals in the next page, plots of the residuals show fairly normal distributions with a mean close to 0, which means the level of the error is independent of when the observation occurred in the study, or the size of the observation being predicted, or even the factor settings involved in

making the prediction. From the perspective of the residuals, the linear model is OK for the analysis.

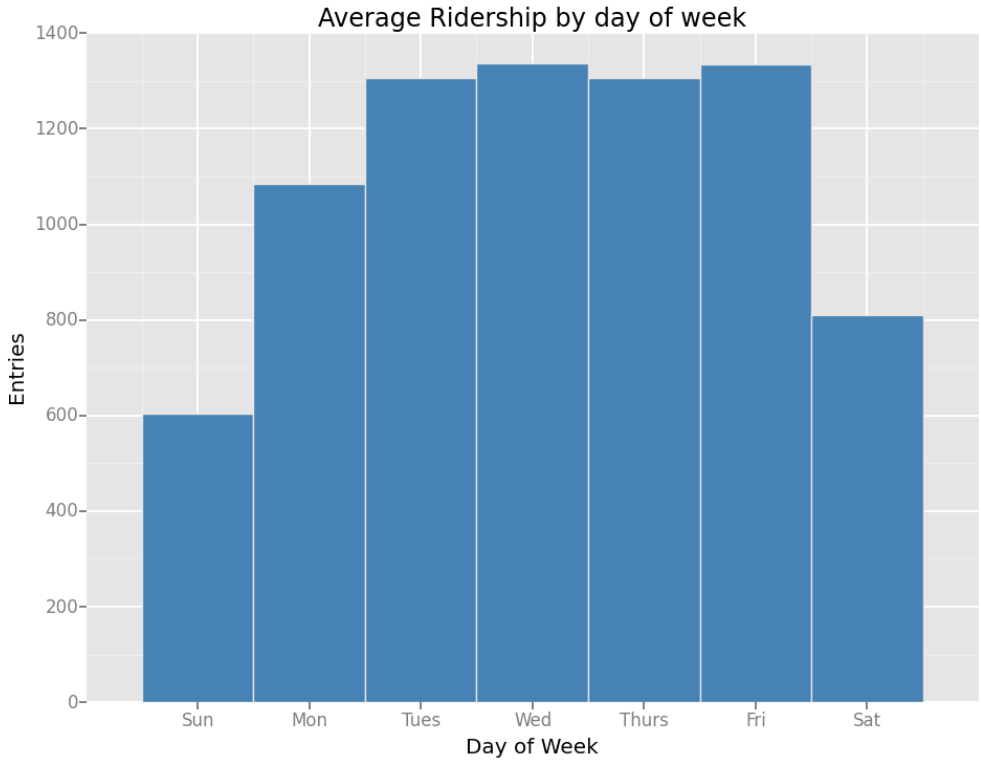


3.1.

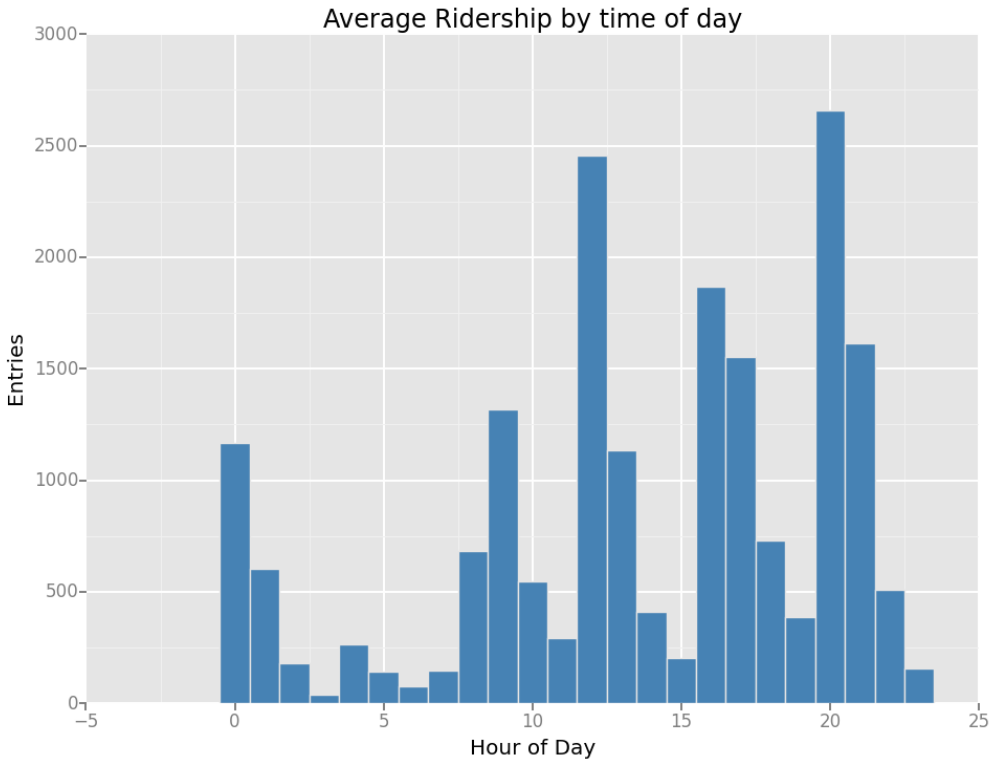


The distributions of ENTRIESn_hourly appear to not be normally distributed and skewed to the left on both rainy and non-rainy days. For both distributions most occurrences are within the smallest bins. There are far fewer observations on rainy days than non-rainy days.

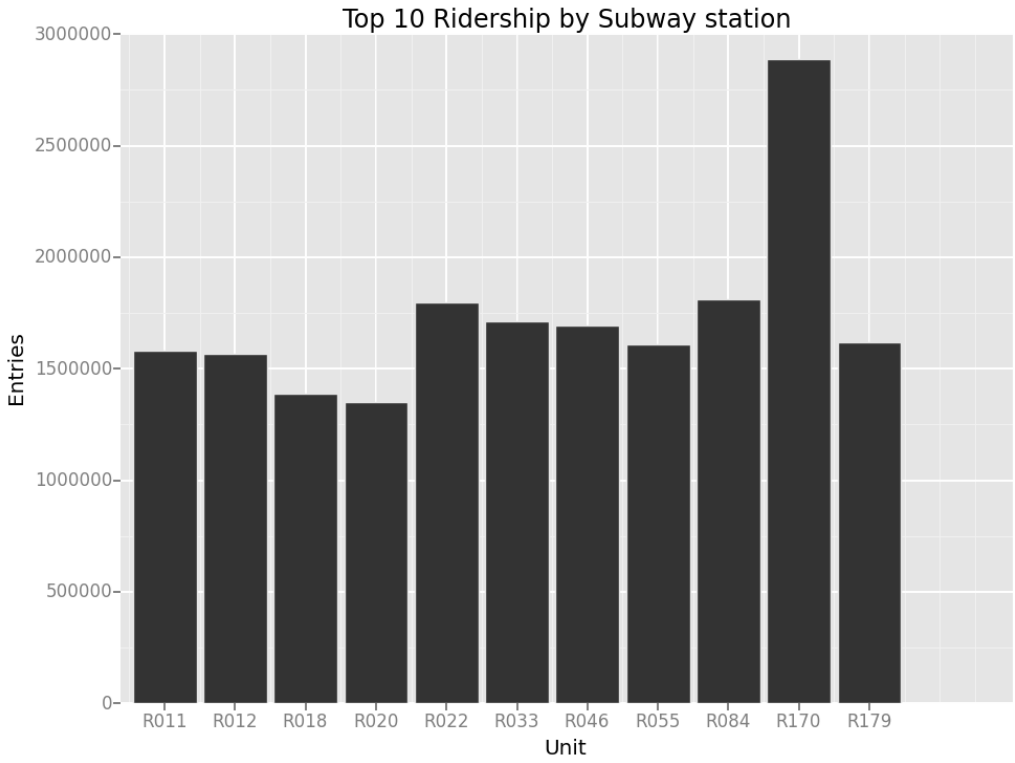
3.2.



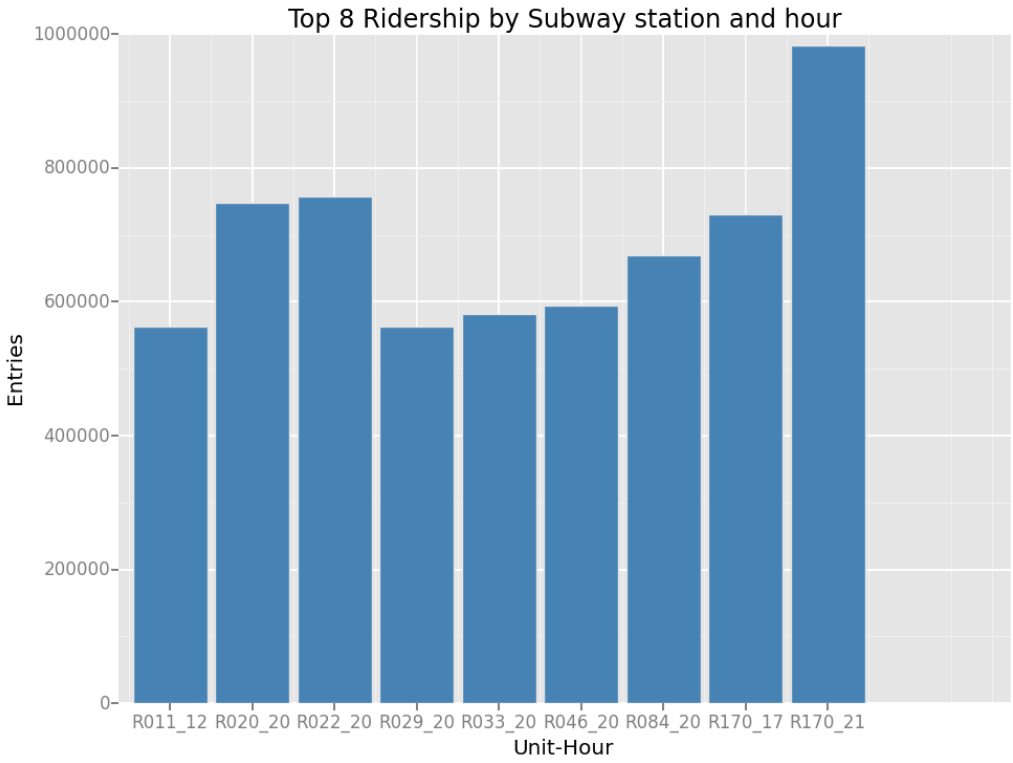
The above bar chart shows the average hourly ridership by day of week. The bar chart shows that the average hourly ridership is higher on weekdays than weekends, with Saturday seeing significantly higher ridership than Sunday. It appears that the average hourly ridership on Monday is significantly lower than the rest of the weekdays. This may be due to a seasonal effect of Monday holidays. The given data set is a sample from May 2011. There is at least one major holiday that falls on Monday in the month of May: Memorial Day.



The above bar chart shows the average hourly ridership by hour of day based on a whole month's data. It is obvious that more people taking subway at rush hours, like 9am and 4pm, when people are going to work or back home from work. Also 12pm is lunch time; 8pm is time for entertainment; 0am is time to back home for sleep after entertainment. So we see peaks at these time slots.



The above bar chart shows the top 10 ridership by subway station. For example, R170 has the highest hourly entries, maybe because the location of R170 is in the business center.



The above bar chart shows the top 8 ridership by station and hour. It is consistent with the 2 figures above: the average hourly ridership by hour of day and the top 10 ridership by subway station. We can see that all units for the top 8 ridership appear in the top 10 ridership by subway station lists, and the hours are also the busy hours. For example, R170 has the highest entries at 9pm.

4.1 & 4.2. I believe that more people take the NYC subway when it is raining. This conclusion is based on two facts: first, as seen in Section 1.3, the mean of ENTRIESn_hourly is greater for hours with rain than without (1,105 vs. 1,090). The comparison of both means using the Mann-Whitney U-test gives us good reason to believe that there is a statistical significant difference between the two data distributions. Secondly, in the OLS model, the rain feature had a positive theta of 11.56, which means that when it raining ('rain' = 1), the predicted entries is increased.

5.1 & 5.2. The data set provided contains only one month of MTA data (May 2011). It cannot reflect how the change of season affects the ridership. The largest shortcoming I see with this data set is that the MTA component of the data is produced on an hourly basis, but it is joined to daily weather data. For example, if it rained at any point in a given day, then in the dataset it rained every hour of that day, which is not true in common sense. And it will definitely affect the accuracy of the prediction. Moreover, the categorical nature of data in the given data set, such as 'rain', 'fog' and 'day_off', and a lot of dummy variables, make a linear model inappropriate.

- **Reference**:

GGPlot (http://ggplot.yhathq.com/docs/index.html)

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

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

http://www.statsoft.com/Textbook/Multiple-Regression#cresidual

http://blog.minitab.com/blog/adventures-in-statistics/how-high-should-r-squared-be-in-regression-analysis