1.1 I used the Mann-Whitney U test to analyze the NYC subway data. Since we did not predict which group would have a higher average we use a two-tailed P value where $p < 0.05$. The Null Hypothesis (H_0) is that the number of entries on rainy days and the number of entries on non-rainy days are the same or in other words that rain has no effect upon subway ridership. The p-critical value is equal to 0.05 or 5%.

1.2 The Mann-Whitney U test was necessary due to the fact that our data sets do not follow a normal distribution as shown in the histogram in figure 3.1 below. The Mann-Whitney U test has a greater efficiency than the t-test on non-normal distributions and it also considers the ranks of the values rather than the values itself, which makes it less likely to misconstrue the outcomes in the presence of data outliers.

1.3 From the Mann-Whitney U test we were able to evaluate the H_0 stated above compared to the Alternative Hypothesis (H_1) which is that the number of entries is not the same on rainy versus non-rainy days. The means of our samples are (without_rain_mean) = 1090.27878 and (with_rain_mean) = 1105.44638. The results of the Mann-Whitney U test are U = 1924409167 and p = 0.0193 (running on 64-bit version).

1.4 This U statistic is relatively high and very close to the maximum. A U statistic with half the maximum value would indicate the H_0 hypothesis is true, which is not the case with ours. We can say the probability of rejecting H_0 when this hypothesis is true is smaller than the significance level ($0.0193 < 0.05$) and therefore we reject the null hypothesis with a 95% level of confidence. Also, in comparing the means we realize there are 1.4% more subway entries when it is raining versus not raining, which would support the rejection of the null hypothesis.

2.1 I used linear regression with gradient decent to compute the coefficients theta and produce predictions for ENTRIESn_hourly. Default values of alpha = .1 was used for the learning rate along with a default of 75 iterations.

2.2 My features included rain, precipi, meanwindspdi, meantempi, meanpressurei and meandewpti. I used UNIT and Hour as dummy variables. UNIT was used as a default, but made sense to use as a dummy variable since it would be hard to keep track of quantitatively however it was important to track as there could be a wide variation between subway stations. I also used Hour as a dummy variable since the subway data showed there weren't recordings every hour, but rather every few hours. Therefore we had to categorize data for different periods of the day which included several hours rather than an hour at a time. Once I removed Hour from the non-dummy features and made it a dummy variable the R^2 value jumped by about 5%.

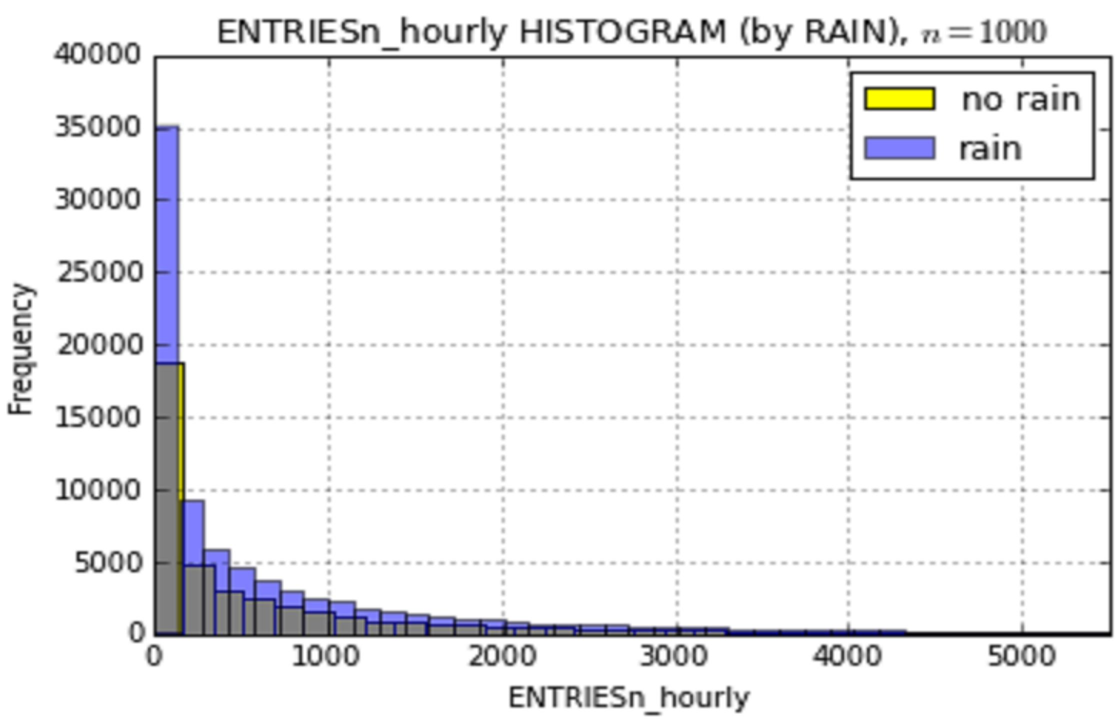
2.3 I believed that the weather variables I used all had a linear relationship to affect ridership and therefore decided to use them. My rationale behind this was typically related to the weather and how one variable has a cause on another, most notably rain. For instance precipitation would be a great indicator of rain and so would pressure and dew point. Temperature had a slight effect on the model so I decided to use that as well; although it can rain when it is warmer and cooler so my initial reasoning would not hold true for this variable.

2.4 The coefficients of the non-dummy features in my model are: -2.80019511e+00, 4.27154725e+00, 4.44581083e+01, -2.97405380e+01, -3.08010753e+01, -1.24151398e+01

2.5 The coefficients of determination or R^2 value is: 0.5019492, which indicates there is about a 50% chance of our model predicting the desired outcome.

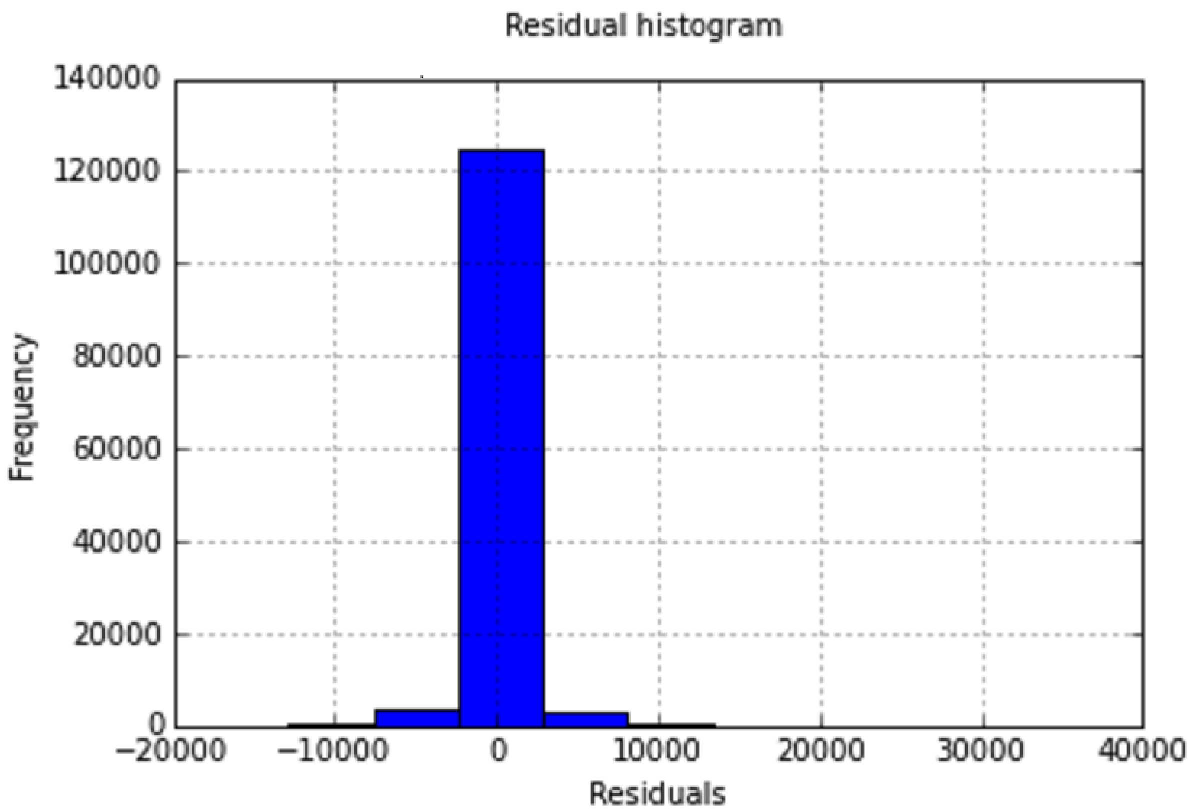
2.6 Based on this R^2 value I do not believe this linear model is appropriate for this dataset because the dataset includes multiple stations and multiple hours, which abate the potential correlation between weather and ridership. However this is more of a data issue than a modeling issue and in analyzing the histogram of the residuals in figure 3.2 we see that most of the residuals were close to 0 (+/- 5,000). So we can determine from the residuals that our model is a close enough fit for our desired outcome; our original hourly entry data and the predicted values are roughly normal and approximately interdependently distributed with a mean of 0 and a constant variance.

3.1 Histogram comparing ENTRIESn_hourly for rainy days versus non-rainy days (Figure 3.1).

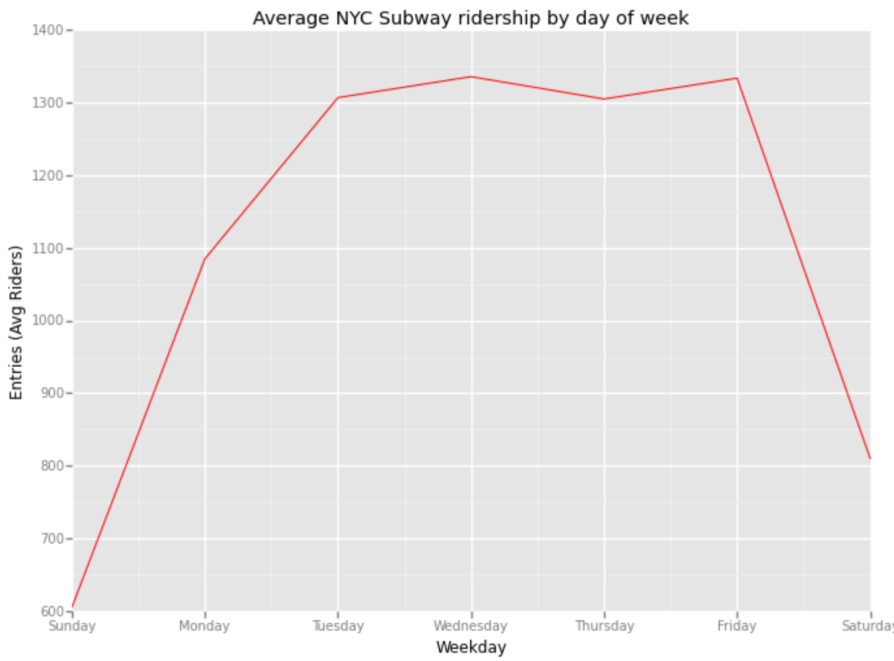


This figure shows a non-normal distribution of the samples sets for number of entries into the subway while raining (blue) and not-raining (yellow).

3.2 Residual histogram of our linear regression with gradient descent model which shows a roughly normal and approximately interdependently distributed data set with a mean of 0.



3.3 Line Chart comparing Ridership by day-of-week:



This graph illustrates a spike in ridership during the weekdays versus the weekends, with Mondays showing less average ridership than the rest of the weekdays. A special note on the data would indicate a holiday fell on a Monday during the time of the data acquired in May 2011 (May 30, 2011 was Memorial Day), which would most likely explain the dip in average Monday riders.

4.1 According to the Mann-Whitney U test we performed there appears to be a relationship between whether or not it's raining on the number of people riding the subway. According to the results from our linear regression model we cannot determine a strong linear relationship exists between rain and ridership on the NYC subway.

4.2 The Mann-Whitney U test indicates there is a significant difference between the numbers of subway riders when it's raining versus not raining. We can say with a 95% confidence level that rain has an impact on the overall subway ridership where rain tends to increase the number of riders. On the other hand the linear regression model, with a coefficient of determination value of just over 50%, seems insufficient to support a correlation. However the fact that it is over 50% would indicate a slight impact of rain on ridership.

5. Using the linear regression modeling I was unable to find a *strong* correlation between the number of riders and rain versus no-rain days. A major reason for this is due to the data consisting of all different stations and the "hours" variable reported for each station consisted of blocks of data over several hours rather than individual hours. If we could select one station and

view the ridership at each hour rather than over the course of a few hours, we would possibly find a stronger correlation. The data provided did not support this type of detailed analysis. In addition we only looked at one month of data, which could certainly have an impact on the results. Had we evaluated the ridership over the course of a year, we would have a larger sample size and therefore more accurate results. Another data anomaly was the greater number of rider entries than exits, which could be the result of subway stations not reporting data. This anomaly would presumably have just as great an impact on non-raining ridership and raining ridership so this should have no effect on our analysis.