Analyzing the NYC Subway Dataset: Answers to Short Questions

Section 1. Statistical Test

1.1
I used the Mann Whitney U-test to analyze this data. The P value was two-tailed and the null hypothesis was that the two samples came from the same population. The results of this test (P = .05) along with the two means of the samples (1,105.45 and 1,090.28) indicate that the samples did not come from the same population, and that the mean from the rainy days sample is higher. We know that the samples come from different populations from the Mann Whitney U-test, and we can see that the mean of the rainy days sample is higher, indicating that that particular population tends to have higher values. The p-critical value used was .05.

1.2
This statistical test is applicable to this data set because it is a non-parametric test, meaning it does not assume that our data is normally distributed. Since our data is not normally distributed, as indicated by the histogram, this is an appropriate test to use.

1.3
P-value: .05
Mean of rainy day entries: 1,105.45
Mean of non-rainy day entries: 1,090.28

1.4
The .05 P value satisfies our p-critical value of .05, meaning that we can reject our null hypothesis with 95% certainty, meaning that there is a difference in ridership between rainy and non-rainy days. The rainy day mean is higher than the mean of non-rainy days, indicating ridership is greater on rainy days.

Section 2. Linear Regression

2.1
I used the gradient descent approach to produce the prediction model for ENTRIESn_hourly.

2.2
I used unit, Hour, meantempi, rain, precipi, and what day of the week it is. I used dummy variables for hour, unit, and day of the week.

2.3
I used rain because we had previously proven that whether it's raining or not has an effect on ridership. Intuitively, bad weather should effect ridership as well, and this was proven when adding meantempi, and precipi improved my R^2 value (although they only improved it very slightly). I also expected day of the week to drastically effect the model, and this hypothesis was correct as adding a day of the week feature

improved R^2 drastically. From some external exploration with the data, it did appear that the different units had different types of results, and so I added this feature to the model as well.

2.4
My non dummy theta values are contained in the following matrix:

[11.20690012, 64.78394589, 64.78394589]

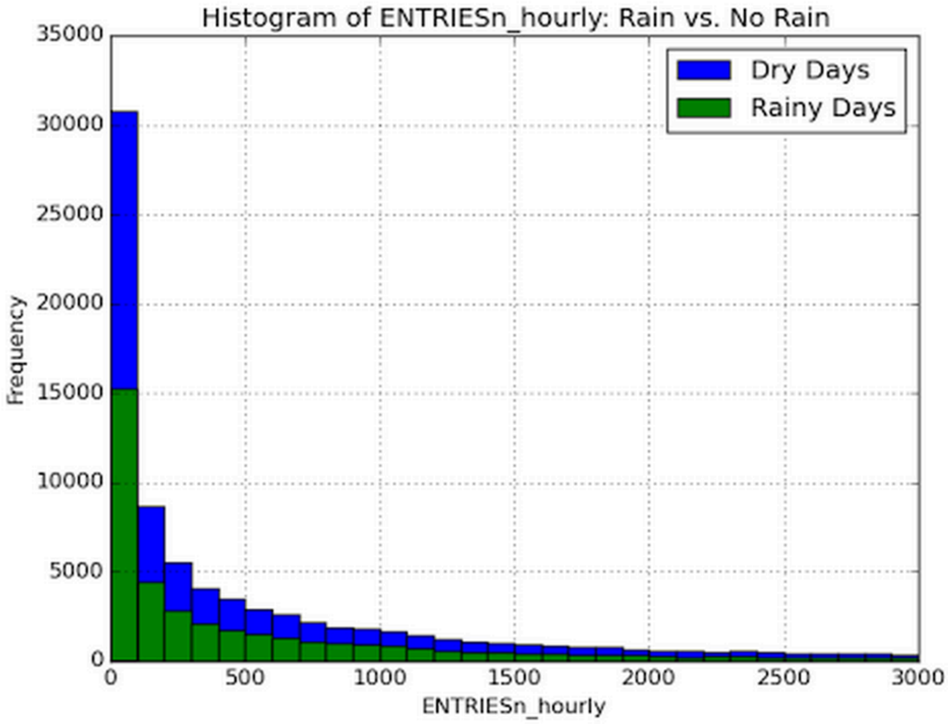
The features represented by this are:
[meantempi, rain, precipi]

2.5
My R^2 value is .52.

2.6
The Coefficient of Determination measures the percentage of the response variable variation that is explained by the linear regression. Since mine is above .5, I would say the regression model is a fairly good fit, but not great. I think that since this data is limited to just the month of May (a fairly mild month, weather-wise), we may be able to do an equally accurate prediction simply using time of day, Unit, and hour. In fact, when I tried using only these features, the R^2 only decreased a tiny bit (~.001). However, I think that if we had all the data from the year, weather would definitely play a more significant role in predicting how many people will ride the metro. Perhaps there is a temperature threshold where below a certain temperature a lot more people suddenly start riding.

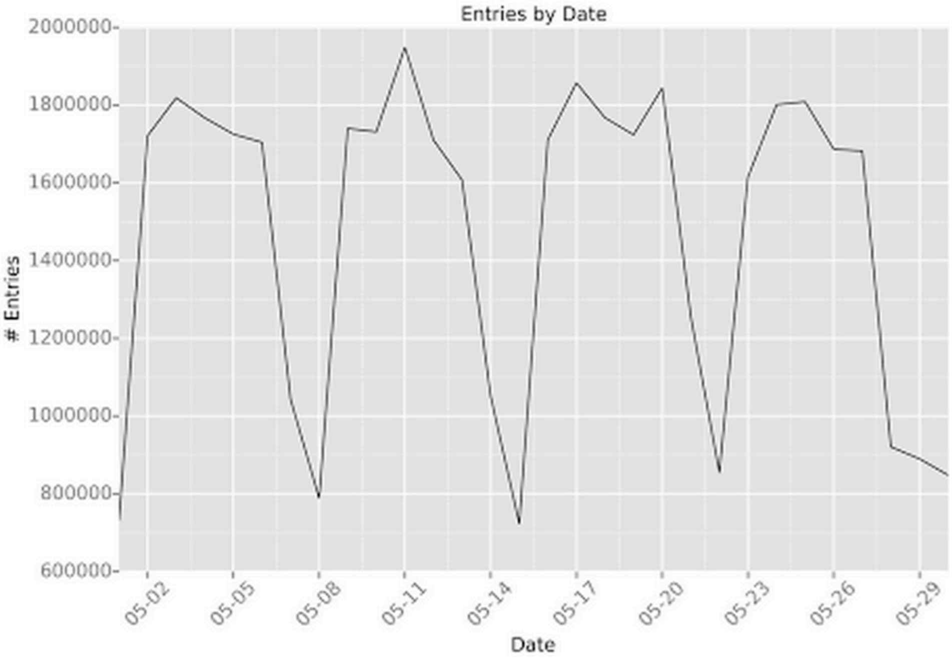
Section 3. Visualization

3.1



This plot represents a Histogram of ridership on rainy and non-rainy days. You can see that the distribution is non-normal, and the distributions of the two groups is fairly similar.

3.2



This graph represents # of riders from day to day during the month of May. It is clear that there is a drastic drop in ridership over the weekend, indicating that day of the week is likely a good predictor of subway ridership.

4.1 and 4.2

Due to the Mann-Whitney U-test, we can state with 95% certainty that ridership on the subway differs between rainy and non-rainy days. This result and the fact that the mean on rainy days is higher indicates that more people ride the NYC subway on rainy days. However, what we have seen based on our linear regression model using the gradient descent method is that rain alone is not a good predictor of how many people will ride the metro on a given day/time. A linear model that did not take rain (or even any weather related feature) into account was more predictive than a model that only took these features into account as it had a significantly higher R^2 value. This would indicate that weather is not a great predictor of ridership, though in combination with other features is does help improve the R^2.

However, I think it would be conceivable that we would see weather as having a higher impact if we were to look at data from over the course of the year rather than just from May. May is a fairly mild month, and therefore there was not a great variance of data within the set. While we found that day of the week and time of day are very predictive of subway ridership, a conclusion that is not likely to change, we may also find that if it gets below freezing, or if its snowing, there would be an upsurge in subway ridership. Therefore, while more people do ride the subway when it’s raining than when it is not, whether or not it is raining is not enough to predict how many people will ride the metro with high accuracy.

Section 5. Reflection

5.1

1.

There are a few potential shortcomings to the dataset. The primary one is that it is only from the month of May. Subway ridership likely changes month to month do to a wide variety of variables (students being home/gone over vacation, weather patterns, holidays, etc.) and therefore we are missing out on a lot of potential features that likely have an impact on ridership. Additionally, the ENTRIESn_hourly field is looking at entry data from every 4 hours. 4 hours is a fairly wide grouping of data, particularly considering that the likely “rush hours” may only be an hour long. By grouping busier hours together with less busy hours, you decrease the extremity of the ridership during those hours. For example, 4 hours where there were 200 riders every hour would look the same as 1 hour where there were 700 riders followed by 3 hours with 33 riders each.

2.

Linear regressions only look at linear relationships by definition. Though we try to correct for this using dummy variables, we are assuming that the relationship between the features and the outcome is always going to be linear, which may not necessarily be the case. Additionally, linear regression assumes that the data are independent. However, this is not always true. Linear regressions are also particularly sensitive to outliers, which could potentially throw off the results. I would imagine there would be a lot of outliers in subway data that would not be related to the features – for example, sporting events, breakdowns, power issues, etc.

3.
The conclusion of a statistical test is only reliable if the underlying data is reliable. We are looking at a very specific data set from one month of one year to draw a large conclusion. We also need to be careful about reporting out the proper P value from the scipy Mann Whitney U-test results, as it assumes a one-sided P value.

Sources Used:

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm
http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit