Alex P. Kalish

Udacity Project 1: NYC Subway Data

Due Date: 6/1/2015

Short Answers

Section 0 - References:

N/A (I visited a lot of sites but did not save any of them as I didn't realize that was going to be part of the assignment. I will start to save all of the links from now on!)

Section 1 – Statistical Test:

1.1 I used the Mann-Whitney U-Test to analyze the subway data because the features in the data were non-parametric, meaning the data did not conform to any specific probability distribution. I used a two-tail P value to test whether two samples come from the same population or have the same distribution, making my null hypothesis $\mu_0 = \mu_1$. The alternative hypothesis is that the distributions are not equal. The null hypothesis can be rejected if the probability of the sample distributions being equal is low enough to satisfy our specifications. I used the critital p-value of 0.05, which is a 95% confidence level. The p-value of 0.05 represents a threshold in which we can reject the null hypothesis. If the probability that the distributions are equal is less than 0.05 we can reject the null hypothesis.

1.2 The Mann-Whitney U-Test is applicable because the data is non-parametric. A T-test would be applicable if the distribution of the data was normal but our data is not normal based on the histrogram in 3.1. Another advantage of the MW test is that it is not necessary for the two samples to have the same number of observations.

1.3 The mean of hourly entries while raining is 1105.45 and the mean entries when not raining is 1090.28. The p-value is 0.0249. To analyze the hypothesis, we need to multiple the p-value by 2 to examine as a two-tailed test. At the 95% confidence level, we reject the null hypothesis that the distributions are the same because the p-value*2 (0.049) is less than 0.05, which means the distributions are statistically different at this confidence level.

1.4 The result tells us that more people use the subway when it rains, since we reject the null hypothesis that the distributions are similar or the samples come from the same population, which helps validate our hypothesis that subway usage increases during rain.

Section 2 – Linear Regression:

2.1 I used gradient descent and linear regression to predict the theta in my regression model.

2.2. The features used in my model were a dummy variable for rainy days and for the unit, the amount of precipitation, the hour of the day and the average temperature of the day as well as a dummy variable for each station.

2.3 I chose to use the dummy for rain because rain would push travelers who prefer walking into the subway. Because of rush hour, the hour of the day will dramatically affect subway usage and needs to be included in the model to avoid having the dependent variable correlate with error. Precipitation is important as well because it may be able to show how long it rained in a day and if there was a greater chance it was raining during rush hour. An extremely high or extremely low temperature would also push daily walkers into the subway for comfort. A higher average temperature may also bring out seasonal changes in subway usage, such as student vacations. I added the unit dummy variable after seeing how it affects my R^2 value. This suggest that rush hour changes in unit usage are a strong variable in changes to entries and exits in each unit.

2.4 The non-dummy coefficients for my parameters are:

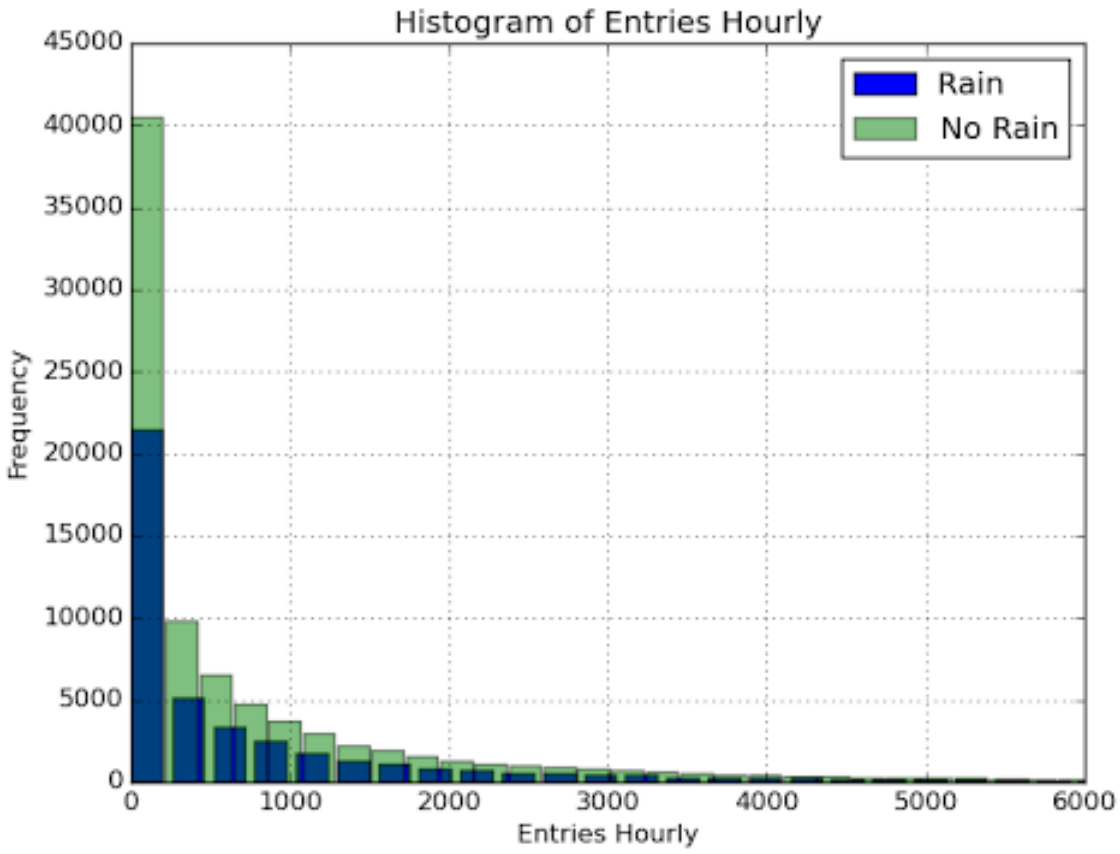
Precipitation	28.7
Hour	65.3
Mean Tempurature	-10.5

2.5 The r^2 value of my model is 0.48.

2.6 The R^2 value measures the good-ness of fit of the model by comparing the change in features (independent variables) with the accuracy of predictions vs. the change in error with accuracy of predictions. This tells you how effective your model is in predicting the dependent variable with the current features.

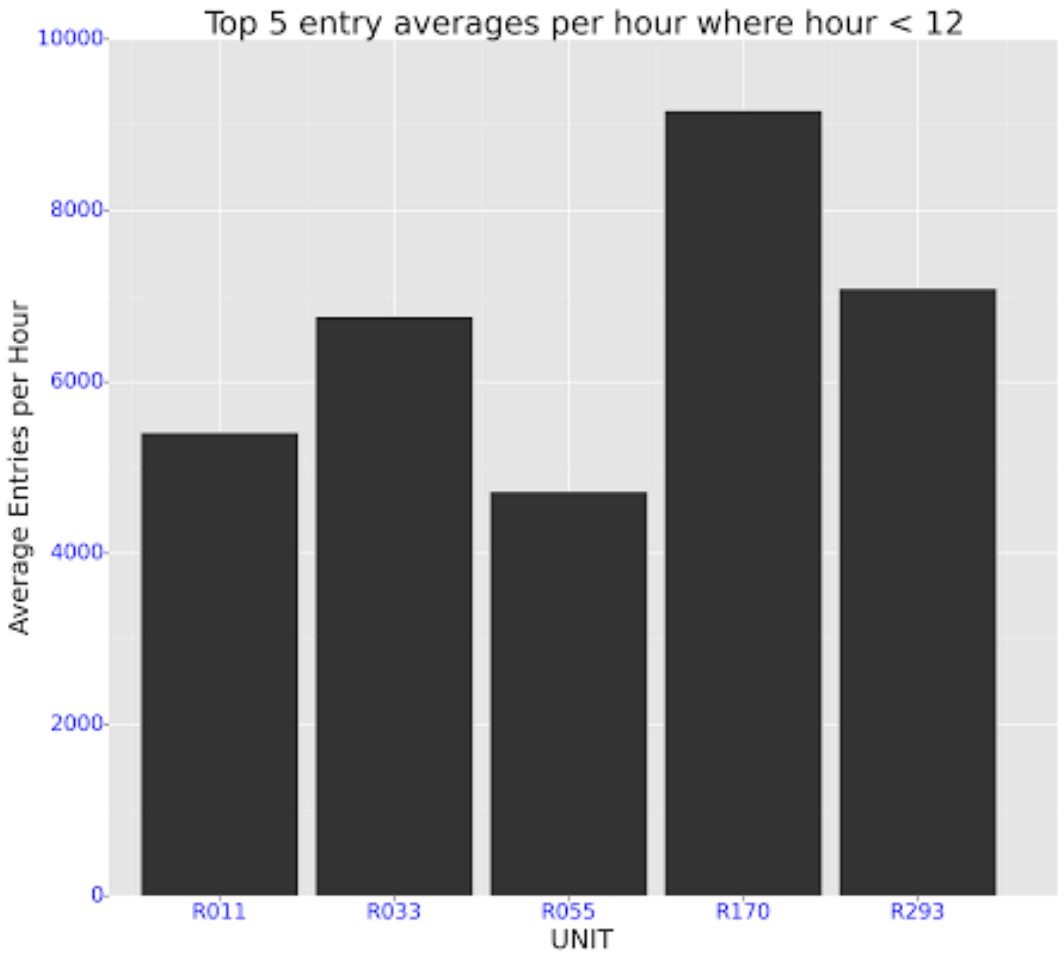
2.62 I think this linear model is appropriate because of the high value of the R^2 value. A 0.48 of change explained by independent variables in the model is high especially for human decision based data.

3.1



In the chart above we see a histogram with bars representing the frequency of a certain range of hourly entries when raining (blue) and when not raining (green). This histogram shows there are many more occurrences of no rain than there are occurrences of rain in our data. It also shows us that our data is not normally distributed but similarly distributed, making the Mann-Whitney U-Test a viable option for statistical testing of mean similarities.

3.2



The chart above shows the top 5 subway stations measured by average entries per hour in the morning, or more specifically before 12pm. It would be put to good use comparing the same stations averages after hours and the top 5 stations after hours to show the changes to individual units from rush hour traffic.

4.1 – 4.2: With a 95% confidence level, I would say that more people ride the NYC subway when it is raining. The mean subway entries during rain is higher than when it is not raining and using the Man-Whitney test, statistically at a 95% confidence level, the two means are significantly different. Therefore, it is plausible to say that more people enter the subway when it is raining.

Supporting this idea, a regression analysis of the data shows that a rain Boolean value 1 shows an increase of entries per hour of about 25 riders. Finally, we can see that there is a positive correlation between the amount it precipitates and subway usage, specifically that ridership increases by 28.7 for every added inch of precipitation. This means that not only will riders enter the subway when it rains but are more likely to enter the subway during heavier periods of rain. Our model includes variables to account for changes in rush hour traffic by including the hour and the station for higher one way traffic

periods in order to reduce the chance of error. These variables increased the strength of our model as we were able to boost the correlation coefficient to about 0.4, an acceptable and generous goodness-of-fit.

5.1.1 – The dataset loses strength by including weather variables by the day while the subway data is recorded at hourly intervals. Having time variables be uniform allows a more accurate analysis of the data. This can be seen in the dominance of the unit bool in affecting the regression. Clearly, different stations have times of the day where they are used most, probably because of rush hour effects. Although the regression shows that precipitation affects total ridership in a day, we could have a more accurate view of its affect if we knew the time of the day it rained and when it rained most.

5.1.2 – It may be possible that the relationships between the dependent and independent variables are non-linear in which case it is not appropriate. It may also be that the relationship is linear for certain values of x and y but the relationship changes with more extreme values. Finally, a regression model can give us the correlation between to data sets but cannot tell us whether a change in one variable is the actual cause of the change in the other variable. For example, the regression shows there is a relationship between the rain and ridership, but it may be some other variable that is directly causing the ridership to rise during periods of rain. Two events taking occurring together do not necessarily have a cause-and-effect relationship. Our stats test of the mean when raining and mean when not raining assumes the data does not conform to any specific probability distribution which may be untrue if our sample data was larger.