Data Analyst Nanodegree

Project 1 - Analyzing the NYC Subway Dataset – Part 2

By
Tom Dzorevski
tomdzorevski@gmail.com
647 405 4752

Section 0. References

Mann–Whitney U test (26 May 2015) Retrieved (2015, May 28) from
http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

scipy.stats.mannwhitneyu (Jan 18, 2015) Retrieved (2015, May 28) from
http://docs.scipy.org/doc/scipy-0.15.1/reference/generated/scipy.stats.mannwhitneyu.html

GraphPad Software (n.d.) Analysis checklist: Mann-Whitney test Retrieved (2015, May 28) from
http://www.graphpad.com/guides/prism/6/statistics/index.htm?stat_checklist_mannwhitney.htm

Laerd Statistics (n.d.) Mann-Whitney U test in SPSS Retrieved (2015, May 28) from
https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php

(2015) Problem with Mann-Whitney U test in scipy [Web log post] Retrieved (2015, May 28) from
http://stats.stackexchange.com/questions/116315/problem-with-mann-whitney-u-test-in-scipy

DataRobot Inc.(2015) Ordinary Least Squares in Python [Web blog post] Retrieved (2015, May 28) from
http://www.datarobot.com/blog/ordinary-least-squares-in-python

Engineering Statistics Handbook (n.d.) Normal Probability Plot [Web post] Retrieved (2015, June 4) from
http://itl.nist.gov/div898/handbook/eda/section3/eda33l.htm#examples

Engineering Statistics Handbook (n.d.) How can I test whether or not the random errors are distributed normally? [Web post] Retrieved (2015, June 4) from
http://itl.nist.gov/div898/handbook/pmd/section4/pmd445.htm

Minitab (n.d.) Patterns in residual plots [Web post] Retrieved (2015, June 4) from
http://support.minitab.com/en-us/minitab/17/topic-library/modeling-statistics/regression-and-correlation/residuals-and-residual-plots/patterns-in-residual-plots/

Section 1. Statistical Test

1.1 Used a Mann-Whitney U-Test to analyze the NYC Subway Dataset to determine if the increase in ridership is statistically significant when it is raining. This is a two-tail test with a 95% confidence level, a with $\alpha = 0.05$ and a $p_{critical} = \alpha/2 = 0.25$. The null hypothesis is that the mean of ridership when it is raining and not raining are the same.

Null Hypothesis H_0: $\mu_{rain} = \mu_{no\text{-}rain}$

There is no relationship between ridership when it is raining and not raining.

Alternative Hypothesis H_A: $\mu_{rain} \neq \mu_{no\text{-}rain}$

There a difference in ridership when it is raining either.

$\alpha = 0.05$

$p_{critical} = \alpha/2 = 0.025$ (for two-tail test)

1.2 The Mann-Whitney U- Test is applicable in the case because the test assumptions are met by the samples.

Mann-Whitney U-Test Assumption	Applicability
All the observation are independent of each other from both groups	For this test, the turnstile hourly entries are independent from each other.
One dependent variable that is measured as continuous or ordinal values so the values can be ranked	The dependent variable Entries Hourly is district which satisfies the assumption
There is one independent variable that consists of two categorical independent groups.	The independent variable is rain which is divided into two groups rain and no rain
If the distributions of the two groups have the same shape, then the Mann-Whitney U test determines whether there are differences in the medians.	The distributions of both the rain and no-rain samples have similar shapes with a positive skew as illustrated in figure 1.

1.3 The original data sample (turnstile_data_master_with_weather.csv) was analyzed using Python 2.2.9 Anaconda 2.2.0 (32-bit) Windows 7 version. As shown in figure 4 the ridership is significantly impacted by the type of day. Thus, the analysis if rain impacted ridership was done on 3 sets. First set was the complete original dataset. The second set was only workdays (Monday through Firday with holidays excludes). The third was weekend days (Saturday and Sunday) from the original dataset. The results are as follows:

	Complete Original Data Sample	**Workday sample**	**Weekend sample**
No-Rain Sample mean variance standard deviation	$\mu_{no\text{-}rain} =$ 1090.3 $\sigma^2_{no\text{-}rain} =$ 5382422.9 $\sigma_{no\text{-}rain} =$ 2320.0	$\mu_{no\text{-}rain} =$ 1304.5 $\sigma^2_{no\text{-}rain} =$ 6989105.7 $\sigma_{no\text{-}rain} =$ 2643.7	$\mu_{no\text{-}rain} =$ 686.6 $\sigma^2_{no\text{-}rain} =$ 2105458.9 $\sigma_{no\text{-}rain} =$ 1451.0
Rain Sample mean variance standard deviation	$\mu_{rain} =$ 1105.4 $\sigma^2_{rain} =$ 5619274.0 $\sigma_{rain} =$ 2370.5	$\mu_{rain} =$ 1280.2 $\sigma^2_{rain} =$ 6994438.1 $\sigma_{rain} =$ 2644.7	$\mu_{rain} =$ 727.2 $\sigma^2_{rain} =$ 2386670.3 $\sigma_{rain} =$ 1544.8
Mean Boolean Comparison	$\mu_{no\text{-}rain} < \mu_{rain}$	$\mu_{no\text{-}rain} > \mu_{rain}$	$\mu_{no\text{-}rain} < \mu_{rain}$
Mann-Whitney U-Test U $p_{one\text{-}tail}$ $p_{two\text{-}tail} = 2 * p_{one\text{-}tail}$	U = 1924409167.0 $p_{one\text{-}tail} =$ 0.0193[1] $p_{two\text{-}tail} =$ 0.0386	U = 882759886.0 $p_{one\text{-}tail} =$ 0.00273 $p_{two\text{-}tail} =$ 0.00546	U = 129024195.5 $p_{one\text{-}tail} =$ 0.01 $p_{two\text{-}tail} =$ 2.26e-7
$p_{critical}$ Comparison $p_{two\text{-}tail}$ vs $p_{critical}$	$p_{two\text{-}tail} > 0.025$	$p_{two\text{-}tail} < 0.025$	$p_{two\text{-}tail} < 0.025$
Null Hypothesis	Not Rejected	Rejected	Rejected

1.4 Applied the Mann-Whitney U-Test to the complete original dataset and the 2 subsets. The complete original dataset had a $p_{two\text{-}tail}$= 0.0386 for the two- tail test (scipy.stats.mannwhitneyu). Since $p > p_{critical}$, the null hypothesis ($H_{0:} \mu_{rain} = \mu_{no\text{-}rain}$) is not rejected. There is a 3.86% chance to randomly get a sample with at least the ENTRIES_hourly mean for the rain sample for the complete original sample.. Thus, the alternative hypothesis (H_A: $\mu_{rain} \neq \mu_{no\text{-}rain}$) is not accepted. However, the subsets had a different result. For the workday and weekend subset samples had $p_{two\text{-}tail}$= 0.00546 and respectively. Since $p < p_{critical}$ for both subset, the null hypothesis ($H_{0:} \mu_{rain} = \mu_{no\text{-}rain}$) in each case is reject. The alternative hypothesis (H_A: $\mu_{rain} \neq \mu_{no\text{-}rain}$) for these sample subsets is accepted. For the workday sample had $\mu_{no\text{-}rain} > \mu_{rain}$ with a $p_{two\text{-}tail}$= 0.0054. There is a 0.5% chance of randomly getting a rain sample as small or smaller as the workday sample . On the other hand, the weekend sample had $\mu_{no\text{-}rain} < \mu_{rain}$ with $p_{two\text{-}tail}$ = 2.26e-7 . Essentially, there is almost no chance (2.26e-5 %) of getting a random sample at least as large as the rain sample. Therefore, based on the weekend and work day samples, rain does have statistically significant effect on the ridership in the NY subway.

[1] The p value is different than the value obtained in class Problem Set 3: Analyzing Subway Data > 3 –Mann-Whitney U-Test which had a $p_{one\text{-}tail}$ = 0.0250 with the same data. This value difference seems to be attributed to the different environments used to compute the p value.

Section 2. Linear Regression

2.1 Used the Ordinary Least-Squares (OLS) linear regression from the statsmodels Python package. The input data was the improved data set (turnstile_weather_v2.csv). Note the original data set could not be used because when processing the data, the Anaconda python window 32-bit version, encountered a memory error.

2.2 The features in the used by the model are 'rain', 'precipi', 'hour', 'meantempi', 'holiday', and 'fog' with the additional dummy variables on the variables 'UNIT' and 'day_week'.

2.3 The features selected were select for the following reasons:

'rain': It seems weather would affect ridership. Based on the sample, the rain variable had a statistical impact on ridership as described in section 1.4.

'precipi': Ridership might be impacted by the amount of precipitation. Suspect a drizzle would have less impact on ridership than a down pour.

'hour': The hour of the day would logically impact ridership. During work days ridership would be higher when people are traveling to and from work. Also, ridership would be less when most people are sleeping. Also, adding it greatly increased the R^2 value.

'meatempi': For extreme temperatures, it is likely that more people would take the subway to avoid the extremes. For example, in extremely cold temperatures, individuals maybe more inclined to take the subway to avoid the cold.

'holiday': In the data sample, the ridership on Memorial Day May 30, 2011 was significantly lower than other Mondays as illustrated in figure 2. Thus, the 'holiday' variable was added to take into account the effects of holidays.

'UNIT': This variable was converted to a dummy variable that greatly increased the R^2 value.

'day_week': This variable was converted to a dummy variable which resulted in a significantly increased the R^2 value.

2.4 These are the coefficients of the non-dummy variables.

Variable	Coefficient	[95.0% Conf. Int.]	
rain	55.558	0.211,	110.906
precipi	-3289.0988	-4176.005,	-2402.193
hour	122.5949	119.682,	125.508
meantempi	-5.8304	-9.381,	-2.280
holiday	-908.3306	-1048.762,	-767.899
fog	-908.3306	-598.133,	-165.903

2.5 The R^2 (coefficient of determination) is 0.489.

2.6 The R^2 indicates that 48.9% of variance in the response variable 'Entries_hourly' can be explained by the explanatory variables. The remaining 51.1% can be attributed to lurking variables or inherent variability. In figure 5, the histogram of residual show long tails in both directions in that the OLS regression model has large errors. The Probability Plot of OLS Residuals in figure 6, show an S shape curve that deviates at the ends from the normal deviation line. This indicates that residuals are not normal distributed errors. In figure 7, Residuals vs Predictions, it illustrates that there are large negative predictions which are impossible. Also larger predictions have larger residuals. I feel that this linear model is not appropriate to predict ridership for this dataset, because the R^2 value is only around 50%, the prediction has impossible values, and larger predictions have larger residuals.

Section 3. Visualization

3.1 Histogram of Entries Hourly

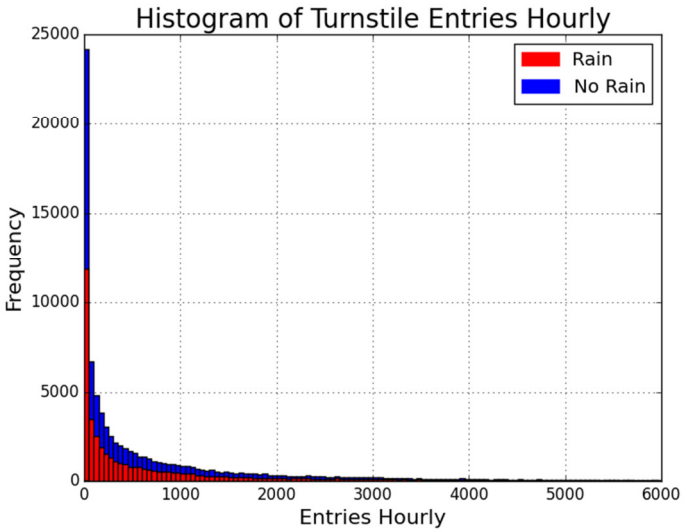

Figure 1 - - Histogram of Entries Hourly using the original data set (turnstile_data_master_with_weather.csv)

In figure 1 the Rain sample size is 44,104 which much smaller than the No Rain sample size of 87,847. Also, the x-axis has been truncated at 6,000, which cuts off the outliers in the long tails that extend past 50,000. Figure 1 illustrates both distributions are positively skewed, and they have similar shapes.

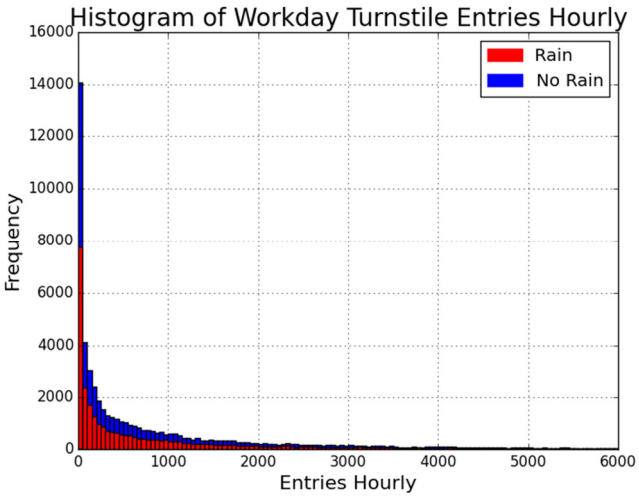

Figure 2 -Histogram of Workdays from original data

In figure 2, the original data was filtered for workdays which are Monday through Friday with holidays were excluded. The rain sample size was 31073, and the no-rain sample size was 57389. Figure 2

illustrates the distributions of the two samples are positively skew and have similar shapes. Also, the x-axis was truncated at 6,000 which cut-off long running tails past 50,000.

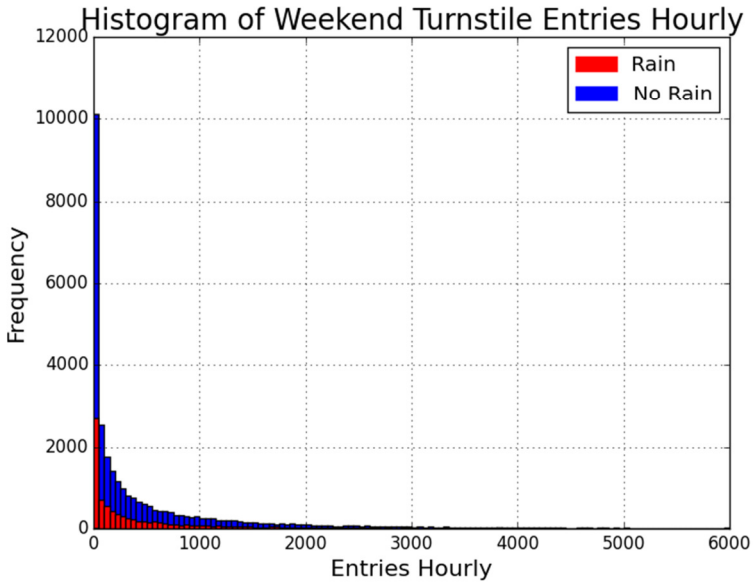

Figure 3 - Histogram of Weekends from original data

In figure 3, the original data was filtered for weekend dates Saturday and Sunday. The Rain sample size is 8,681which much smaller than the No Rain sample size of 30,458. Also, the x-axis has been truncated at 6,000, which cuts off the outliers in the long tails that extend past 39,000. Both distributions are positively skewed, and they have similar shapes.

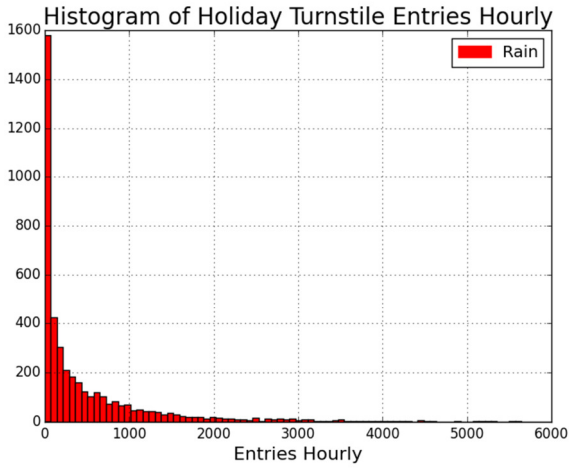

Figure 4 - Histogram of Holiday Entries Hourly

Figure 4 illustrates the histogram of holiday ridership. In the original sample there was only one holiday Memorial Day with 4350 data points, and it rained all day at all the data collecting station. As a result,

there are no no-rain data points for holidays. The x-axis was cut-off at 6,000 clipping the long running tail that extended past 18,000.

3.2 Visualization of Ridership by day-of-week

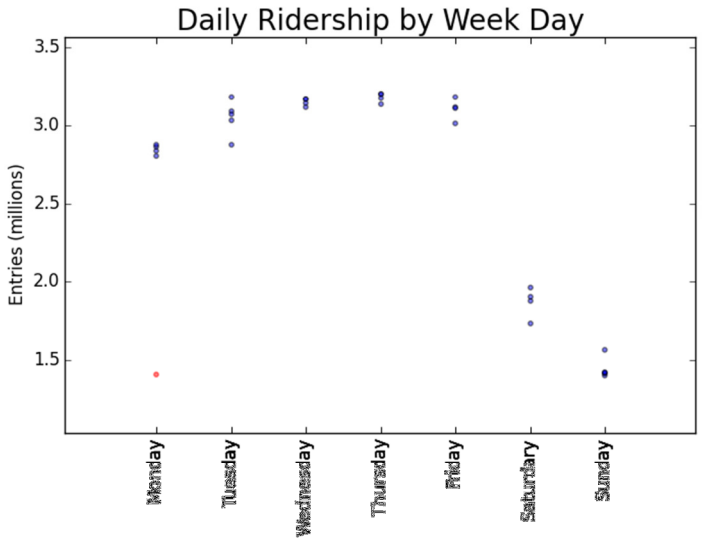

Figure 5 - Ridership by Week Day[2]

Figure 5 illustrates the ridership by day. The red point indicates a holiday (Memorial Day). The holiday has significantly smaller ridership than a normal Monday workday. The ridership is significantly higher on workdays than on weekends or holidays.

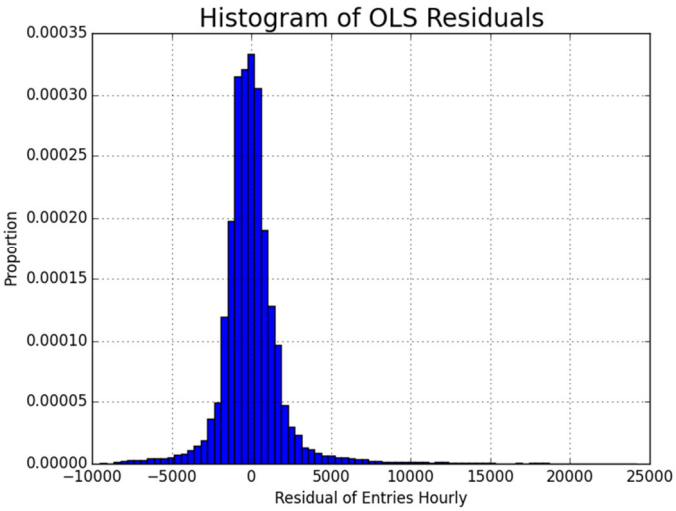

Figure 6 - OLS Residuals Histogram

[2] When the legend was added, it obscured the Friday data points; as a result, the legend was removed.

In Figure 6, the histogram of the OLS model residuals normal distribution shape but the tails in both directions are very long.

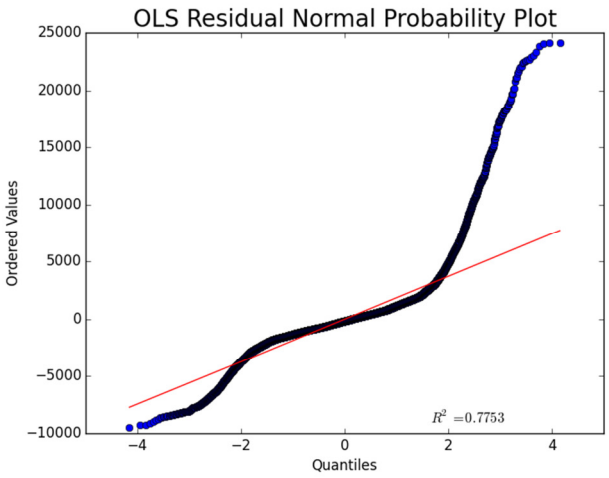

Figure 7 - Probability Plot of OLS Residuals

In figure 7, the normal probability of the OLS residuals has a S shape curve deviating for the normal distribution line. Indicating the residual errors at the end is not due to random error.

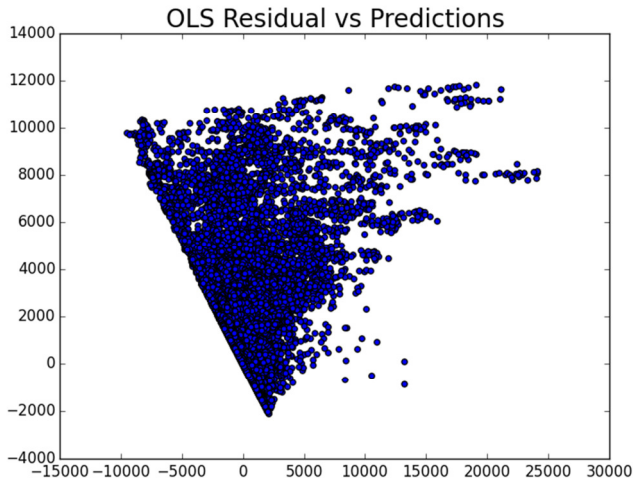

Figure 8 - OLS Residuals vs Predictions

In figure 8, the plot of residuals vs predictions for the OLS regression model shows that large positive predictions we have larger residuals, and there are negative ridership predictions which are not possible.

Section 4. Conclusion

The ridership of the NY Subway is affected when it is raining. The impact on ridership depends on the type of day. For workdays ridership decreases when is raining. Conversely, for weekends ridership increases when it is raining. The analysis of the data showed that the mean of turnstile hourly entries decreased from $\mu_{no\text{-}rain} = 1304.5$ to $\mu_{rain} = 1280.2$ for workdays and for weekends it increased from $\mu_{no\text{-}rain} = 686.6$ to $\mu_{rain} = 727.2$. When the complete data sample was analyzed, the ridership increased from $\mu_{no\text{-}rain} = 1090.3$ to $\mu_{rain} = 1105.4$. A two-tail statistical test with a null hypothesis of H_0: $\mu_{rain} = \mu_{no\text{-}rain,}$ an alternative hypothesis of H_A: $\mu_{rain} \neq \mu_{no\text{-}rain}$ and alpha level of $\alpha = 0.05$ was used to analyze the samples. A Mann-Whitney U-Test was applied on the samples, and it produced the following p values: complete dataset $p_{two\text{-}tail} = 0.0193$, workday subset $p_{two\text{-}tail} = 0.00546$, and weekend subset $p_{two\text{-}tail} = 2.26e\text{-}7$ Since $p_{two\text{-}tail} < p_{critical}$ for the workday and weekend samples, the null hypothesis was reject and the alternative hypothesis was accepted for the weekend and workday samples. Thus for the workday and weekends samples the difference in the means was statistically significant. However, for the complete dataset sample the analysis showed $p_{two\text{-}tail}=0.0386$; thus, $p_{two\text{-}tail} > p_{critical}$ so the null hypothesis was not rejected for the complete sample. This is not unexpected for the complete sample because the effect of rain on ridership is in opposite direction for workdays and weekends.

OLS linear regression was applied to the data with the rain attribute as a feature. The rain feature had coefficient of 55.558 and 95% confidence interval of 0.211 to 110.906. The lower limit of the confidence range is near zero which further supports the conflicting effect of rain in ridership between workdays and weekend.

Section 5. Reflection

The OLS linear regression only had a R^2 of 0.489 which means the 48.9% of the variance can be attributed to the explanatory variables, with 51.1% attributed to lurking variables or inherent variance. As outlined in figure 5, holidays like Memorial Day significantly impact ridership. The residual figures 5, 6, 7, and 8 indicate a poorly fitted linear model.. In the residual normality plot in figure 7, shows a S curve deviating from the normality line which indicates non-normal random errors with long tails in the positive and negative direction. Also, the S curve is an indication of lurking variables, and of missing higher order terms (Minitab). The negative predictions shown in figure 8 are not possible, and the residuals are larger for larger predictions. This further supports the poor quality of the model and that lurking variables are missing from the model. Also the plots show the fanning and spreading of residuals which is an indicator of non-constant variance. Most likely lurking variables are not part of the data set. Some potential lurking variables are sporting events, concerts, parades, celebrations, tourism, and season.

Furthermore, additional tests can be performed on other variables to determine if there is a statistical difference by other variables. Also a different statistical test could be used to further investigate the Null Hypothesis. For this analysis, sample data was only for a single month. Data for additional months may provide additional incites regarding the data relationship and other variables that will affect ridership.