Project 2: Analyzing the NYC Subway Dataset

Submitted by: Bharteesh Kulkarni

Course: Udacity Data Analyst Nanodegree

Date: 10/19/2015

Section 0. References

1. http://www.statisticssolutions.com/mann-whitney-u-test/
2. https://www.linkedin.com/pulse/regression-analysis-how-do-i-interpret-r-squared-assess-gaurhari-dass
3. http://blog.minitab.com/blog/adventures-in-statistics/why-you-need-to-check-your-residual-plots-for-regression-analysis
4. http://docs.ggplot2.org/current/scale_continuous.html
5. http://stattrek.com/statistics/dictionary.aspx?definition=coefficient_of_determination
6. http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
7. https://storage.googleapis.com/supplemental_media/udacityu/649959144/MannWhitneyUTest.pdf

Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

a) To analyze the NYC subway data, I used Mann-Whitney U-test to compare two population means (mean of entries with rain, mean of entries without rain).
b) Two-tailed P value is used to evaluate whether or not there's a statistically significant difference between the populations.
c) Null Hypothesis (H_0): Probability of the mean ranks of a randomly selected rainy sample from population X being greater than the mean ranks of a randomly selected non-rainy sample from population Y is 0.5.
   H_0: $P(x > y) = 0.5$
   x – random draw from population X (ridership entries on rainy days)
   y – random draw from population Y (ridership entries on non-rainy days)
d) p-critical value used : $p \leq 0.05$

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Mann-Whitney U test is a non-parametric test, which is used to compare two population means, and this test is appropriate to analyze NYC subway dataset because both rainy and non-rainy datasets are not normally distributed.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

Sample Means:

Mean of entries with rain: 1105.4463

Mean of entries without rain: 1090.2788

Test results:

U-statistic: 1924409167

One-tailed p-value: 0.025

Two-tailed p-value: 0.50

1.4 What is the significance and interpretation of these results?

Results show that subway ridership is higher when it rains. Two-tailed p-value satisfies p-critical value of 0.05, which indicates that the mean value is statistically different for rainy and non-rainy days. Hence, null hypothesis could be rejected.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

a) **OLS using Statsmodels or Scikit Learn**
b) **Gradient descent using Scikit Learn**
c) **Or something different?**

I used OLS using Statsmodel module to run linear regression on the NYC Subway dataset.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

I used these features: 'rain', 'precipi', 'Hour', 'meantempi','fog','meanwindspdi'

I experimented with only one Dummy variable (‘UNIT’).

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

After adding “fog” and “meanwindspdi” features to the 4 provided features (rain, precipi, Hour and meantempi), r^2 value increased slightly in the regression model. Typically, people use subway during inclement weather; so, I tried to include other weather related features such as meandewpti, meanpressurei but they did not seem to improve the r^2 value significantly.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

The following parameters are from the OLS model:

rain	0.040560
precipi	-73.976935
Hour	65.364525
meantempi	-9.491420
fog	214.086883
meanwindspdi	32.568316

2.5 What is your model’s R^2 (coefficients of determination) value?

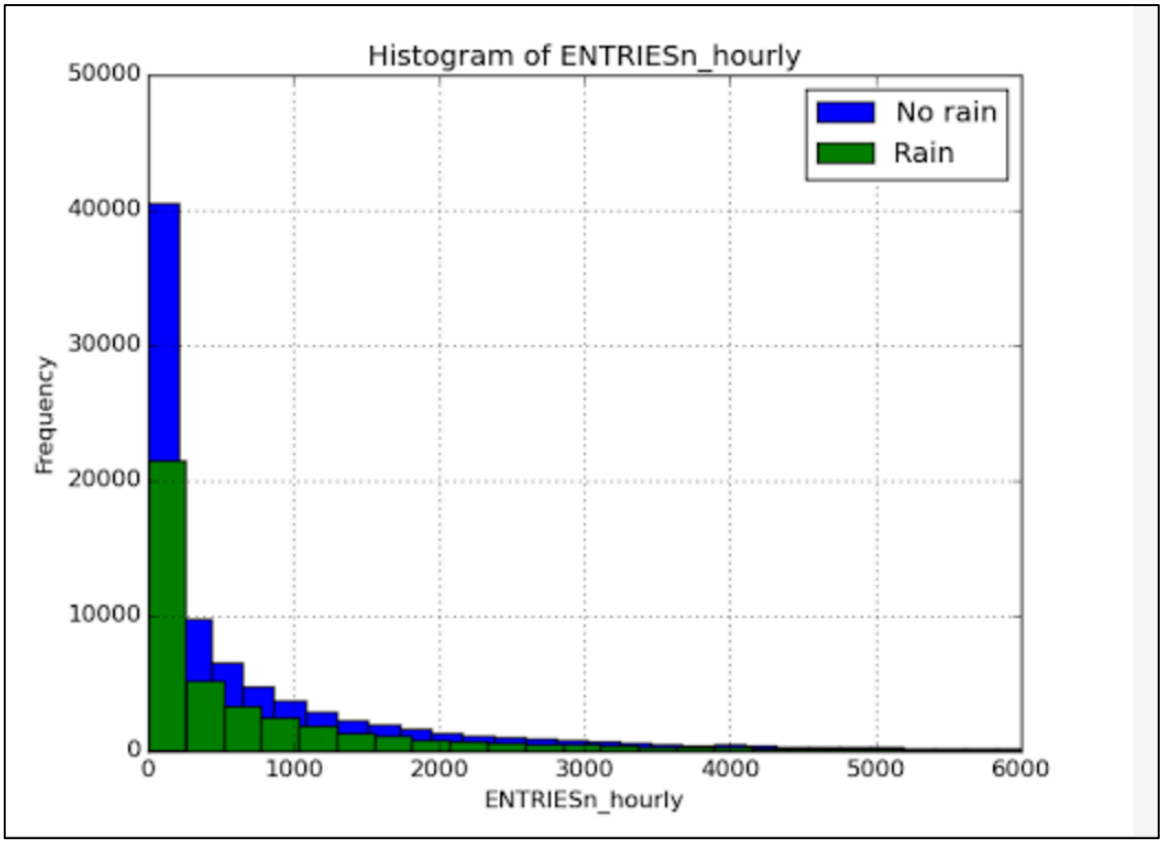
OLS R^2 value - `0.480456675828`

2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2value?

R^2 is the percentage of the response variable variation that is explained by a linear model, which is always between 0% and 100%. For the NYC subway dataset, R^2 value for OLS model explains 48% of the variance.
It might not be possible to conclude whether this model was good fit just by knowing the R^2 value. Residual plots can reveal unwanted residual patterns and we will be able to assess whether observed error (residuals) is consistent with stochastic (random/unpredictable) error.

Section 3. Visualization

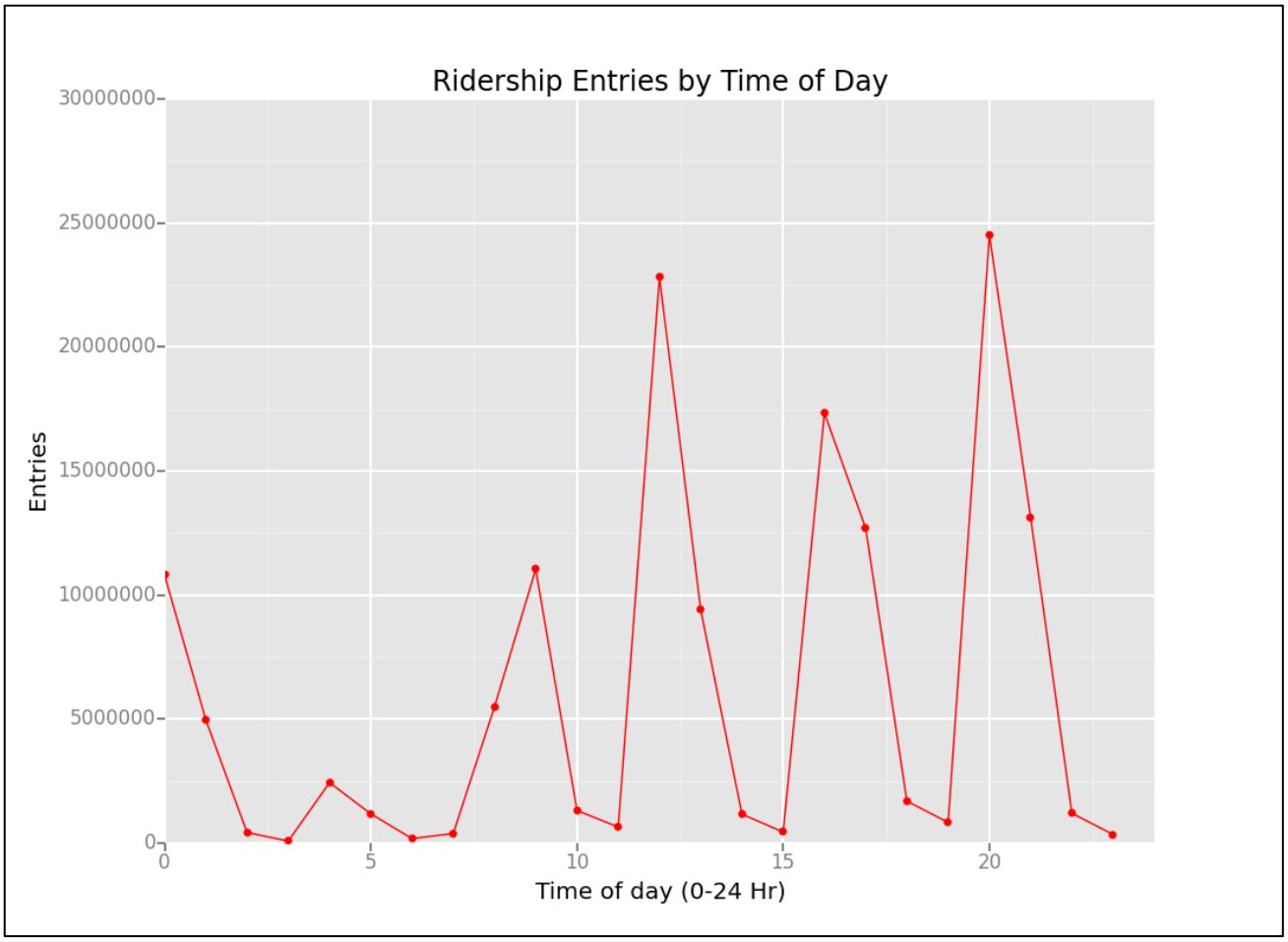
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.



Histograms illustrate that both distributions (rain and no rain) are not normally distributed.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like.

Ridership by time of day:



This line plot (also from ggplot) shows that average ridership is higher during morning work hours (peaks at noon, lunch time may be?), increases between 3-4PM and finally reaching peak at 8PM.

Section 4. Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

Based on the analysis and interpretation of the data by using the statistical methods, more people ride the NYC subway when it is raining.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

Mann-Whitney U Test results tell us that on an average, 15 more riders take the subway when it is raining as compared to when it is not raining. Also, the test confirmed that the two-tailed p-value satisfied the p-critical value of 0.05, hence rejecting the null hypothesis.

In the OLS regression model, R^2 value for OLS model explains only 48% of the variance, which I think is not a pretty good value and could be enhanced to improve the explanatory power of the model. When looking at the coefficients of non-dummy features such as **rain** which has a positive value indicates that there exists a positive linear relationship between the ridership and rain.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. **Dataset,**
2. **Analysis, such as the linear regression model or statistical test.**

Regarding the Dataset provided, it included the ridership data for only one month – May ’11. Having additional data for other months, or may be for few years could have significantly improved the test results and predictions.

Mann Whitney U-Test only used Entries data for rainy (or) non-rainy days. Other important parameters such as UNIT or Time or Hour have not been accounted for.

In the regression model, R-squared value could not be significantly improved with the features provided in the dataset. This is likely due to some limitations in the subway dataset (or) may be an advanced regression model would have produced better results.