Terry Milyard – tm4237@att.com

PROJECT 1 – PART II: Analyzing the NYC Subway Dataset

Section 0. References

I generally used the references listed in the problem sets for additional detail and help. However, I did use the following sites for additional help:

- http://www.michael-noll.com/tutorials/writing-an-hadoop-mapreduce-program-in-python/)
  - This site had a similar example to the ones shared in class
- http://stackoverflow.com/questions/10911057/adding-a-simple-lm-trend-line-to-a-ggplot-boxplot
  - I used this to find the geom_smooth function to add a trend line to my point/scatter data.
- http://docs.ggplot2.org/current/scale_continuous.html
  - This site helped me with chart (ggplot) y-axis scale options. I used the log scale.

Other links from the problem set (and topic):

- https://dev.mysql.com/doc/refman/5.1/en/counting-rows.html
  - How to aggregate using SQL
- http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
  - Format of Mann U function and important note re: one v two tailed test use
- http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html
  - Format of mean function with example
- http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
  - Visual examples illustrating the inverse relationship of residual size to model accuracy
- http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html
  - Format of sum function with example
- https://pypi.python.org/pypi/ggplot/
  - Site provided examples of ways to visually represent data with supporting code

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?

- I used the Mann-Whitney U Test when evaluating the NYC dataset.
- One tailed P-Value
- The null hypothesis stated that there was no difference in the mean ridership of rainy and non-rainy days
- At a 95% confidence level, the critical p-value threshold is .025 (.05/2). I calculated a p-value of .024999 from my data concluding the distributions are different

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.

- Based on our data set we could not assume that the data was drawn from any particular underlying probability distribution. Therefore a non-parametric test was applicable.

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.

- With Rain Mean Ridership = 1105.4463767458733

- Without Rain Mean Ridership = 1090.278780151855
- P-Value = 0.024999912793489721

1.4 What is the significance and interpretation of these results?

- While the distribution and means of the two populations are very close, there was enough of a variance to reject the null hypothesis.

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:. Gradient Descent

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

- Features: MAXTEMP, MINTEMP, PRECIP, WINDSPEED
- Dummy variables: UNIT

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that?

- My assumption was that ridership would change with adverse weather conditions and that I would observe statistically relevant ridership patterns under similar weather conditions. If it's raining, very warm, very cold or very windy, I would expect that people would prefer riding the subway instead of walking.

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

- I do not recall explicitly setting or calculating individual coefficients for each feature, however, the feature data was normalized prior to being used in the model.

2.5 What is your model's R2 (coefficients of determination) value?

- I calculated an R^2 value of 0.42655 when using the features I selected.

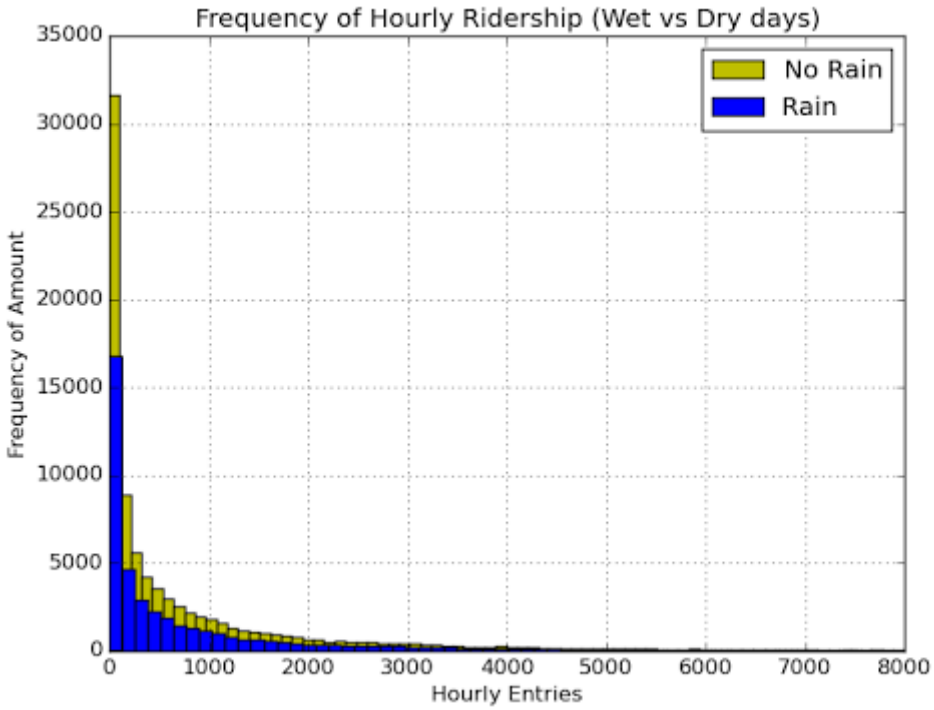
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

- This is a highly subjective question. Because the problem required an R^2 value of >0.2 I feel comfortable in assuming that 0.42655 is a strong indicator of goodness of fit. However, I would also assume that there would be another combination of features that would yield a higher R^2.

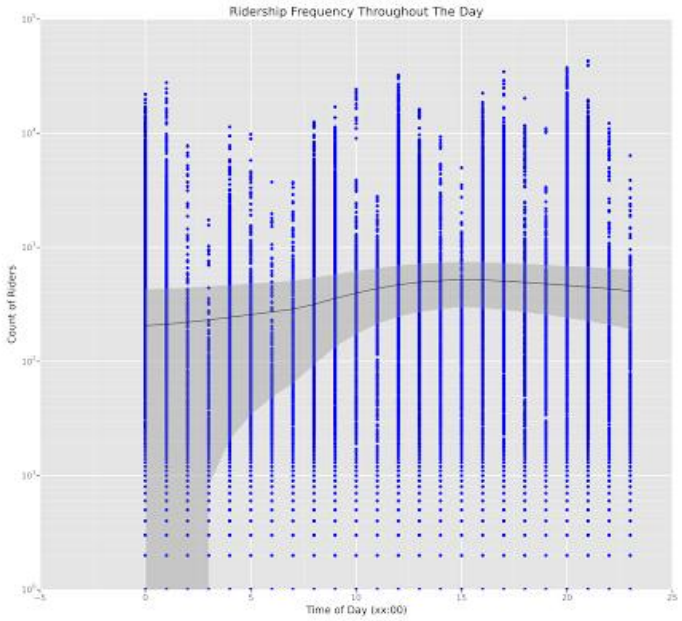
Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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



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.



Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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? 4.2 What analyses lead you to this conclusion?

I believe that more people ride the subway when it is raining. While the means for with and without rain were 1105 and 1090 respectively. While that only represents a difference of a little less than 1.5%, the Mann-U test had a p-value of 0.02499 for these two populations. At a 95% confidence level, one would have to reject the null hypothesis that the mean of the two populations are the same because it is a one-tailed test and therefore we need to divide 5% by 2 giving us a threshold of 2.5%. Furthermore, I saw an increase in the R^2 value of my regression model when bringing features which contain adverse weather conditions. I think the histogram helps to illustrate that the two populations have similar distributions in terms of frequency breakdown and help to show that they do behave very similarly, but the variance was just high enough to conclude a difference; that rain increases ridership.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

5.1 Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test. 5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

I found there was a consistent theme throughout the problem sets that the code compiler/server had its limitations and therefore sometimes we were working with an abbreviated data set. A larger, more complete data set would have been better. I felt that I would have liked to have been able to build an algorithm to combine various sets of features to help increase R^2. There were simply too many iterations to try them all, but finding a way to automate the search would have helped significantly. (This would also have been helpful when applied to the titanic problem in part one). It would have been an advantage to help set the scope of the dataset itself. I understand that in many cases, the data is what it is and by the time it is received by those doing the analysis, there is no chance to go back upstream to change what is captured, but all the same, it would have been helpful to see if any other fields/values were available within the NYC system and/or if the scope of UNITS was limited in any way.