Matt Munsch

User – M.Munsch@yahoo.com

Analyzing the NYC Subway Dataset

Section 0. References

http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs__two-tail_p_values.htm

http://www.itl.nist.gov/div898/handbook/prc/section1/prc131.htm

http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

https://www.moresteam.com/whitepapers/download/dummy-variables.pdf

http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

http://dss.princeton.edu/online_help/analysis/interpreting_regression.htm

Section 1. Statistical Test

1.1

Which statistical test did you use to analyze the NYC subway data? **Mann Whitney U**

Did you use a one-tail or a two-tail P value? **Two-tail**

What is the null hypothesis? **"...the distributions of both groups are identical, so that there is a 50% probability that an observation from a value randomly selected from one population exceeds an observation randomly selected from the other population."**

What is your p-critical value? **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.

The statistical test is applicable to the dataset due to the fact that the data is not normally distributed. A simple histogram comparing the "ENTRIESn_hourly" rain and no rain values supports this non-

normal distribution. Other statistical tests, such as Welch's t-test, require that the data is 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.

with_rain_mean: 1105.4463767458733

without_rain_mean: 1090.278780151855

p: 0.0499998255869794

1.4

What is the significance and interpretation of these results?

The p value is quite small which suggests one should reject the null hypothesis. Thus, the populations are distinct. When it rains, ridership increases by ~15 riders per hour.

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? **'rain', 'precipi', 'Hour', and, 'meantempi'**

Did you use any dummy variables as part of your features? **Yes, for the 'UNIT' feature**

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.

When using the 'hour' feature my r^2 value increased from ~.42 (with other features such as 'rain','meantempi','precipi') to ~.46. Adding in these other features further increased my r^2 value from 0.463247669262 to 0.463968815042.

2.4

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

'rain' = 2.92398062e+00 'Hour' = 4.67708502e+02

'precipi' = 1.46526720e+01 'meantempi' =-6.22179395e+01

2.5

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

0.463968815042

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?

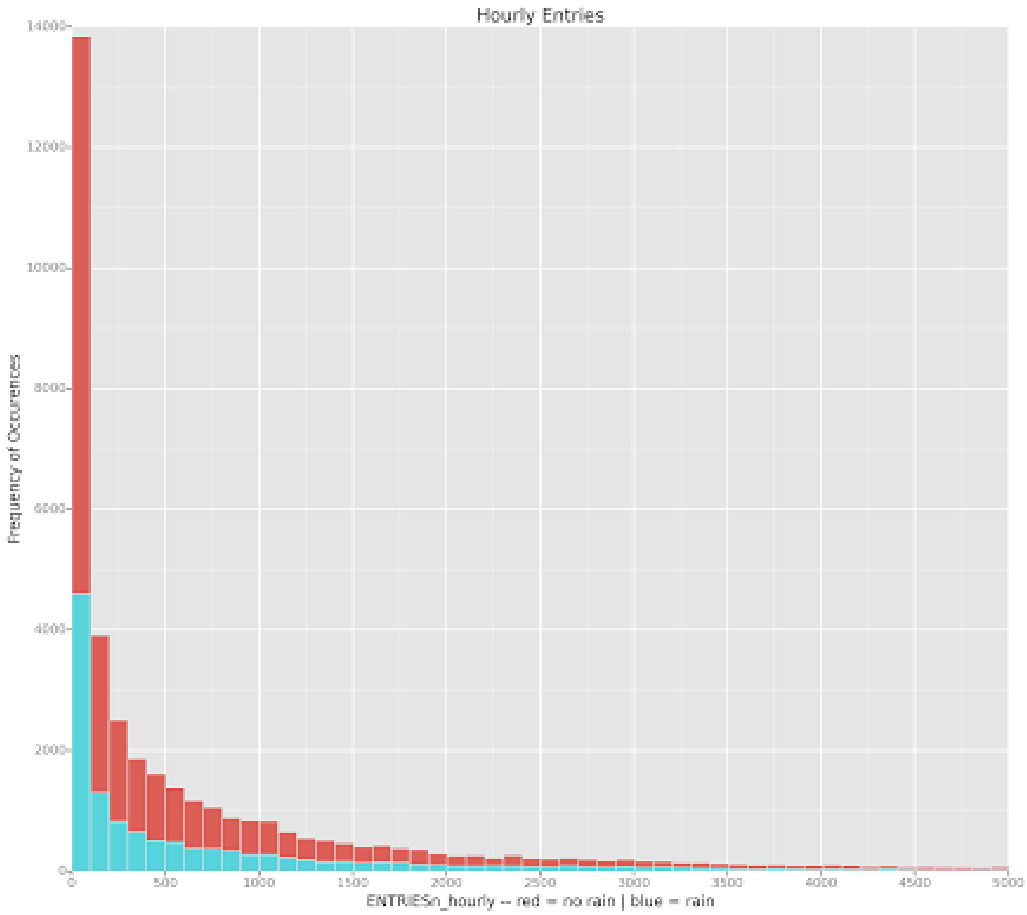
This value means that ~46% of the total variance is explained. Given the R2 value, it is hard to say whether or not this linear model is appropriate for the dataset. Further research has shown me that low or high R2 values are not inherently good or bad.

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.



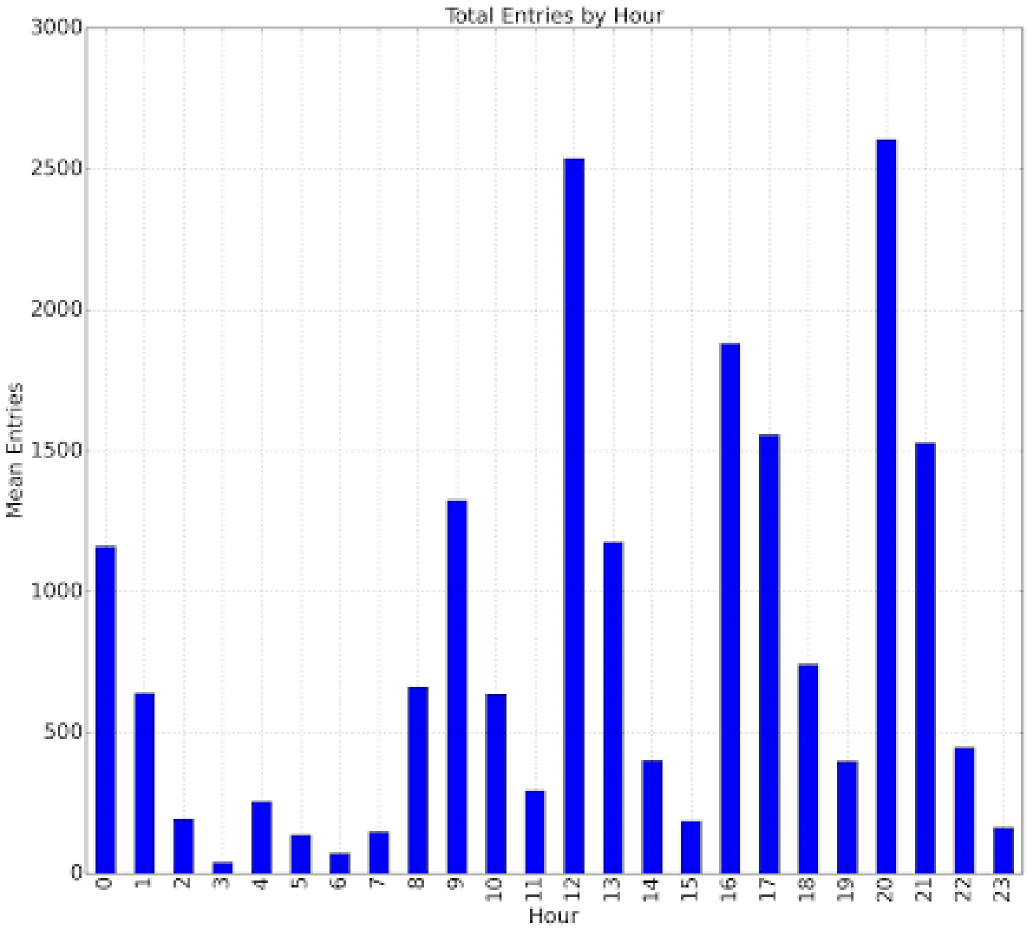
Legend issue with ggplot: http://discussions.udacity.com/t/ggplot-pyplot-trouble-producing-a-legend/13272

Because of this I added the "legend" in the x-axis label

This histogram shows that a large number of "ENTRIESn_hourly" rows had a very low amount of total entries – this suggests that there may be some bias in the data in terms of certain "UNITS" (stations) being closed during certain hours or were possibly closed for maintenance. This could also suggest that the data may or may not have been corrupted to produce "0" entries. Additional analysis with subject matter experts of the logging system should be done to confirm or deny these possibilities.

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.



This graphic was done using pandas ".plot". This graphic answers a fundamental question: On average, how many entries are occurring during each hour of the day? From this graphic we discover that there are "hotspots" of activity during hours: 0, 9, 12, 16, and 20. This pattern is expected given normal commuting/travel patterns.

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?

From my analysis of the data presented, it is clear that there is an increase in the average ridership of the NYC subway during periods when it is raining. However, it should be noted that this

statement can only be confirmed true for this specific dataset. In order to validate this statement further, a much larger sample size is needed. Before analyzing the data, my hypothesis was that individuals would tend to take the subway more often during periods of rainfall over other modes of urban transportation (walking, biking, etc.).

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.

Both the statistical tests and linear regression model proved that, on average, more individuals ride the subway when it is raining. The Mann Whitney U test showed that the p value was 0.04999 and thus, the two populations are distinct. It also showed us that on average, ridership increases by around 15 riders per hour during periods of rain. The linear regression model used also supported this after plotting the residuals as a histogram. The majority of the residuals plot near 0. In addition, the coefficient of 'rain' in the linear regression test amounted to 2.92398062e+00, showing that rain does indeed have a positive effect on ridership.

Section 5. Reflection

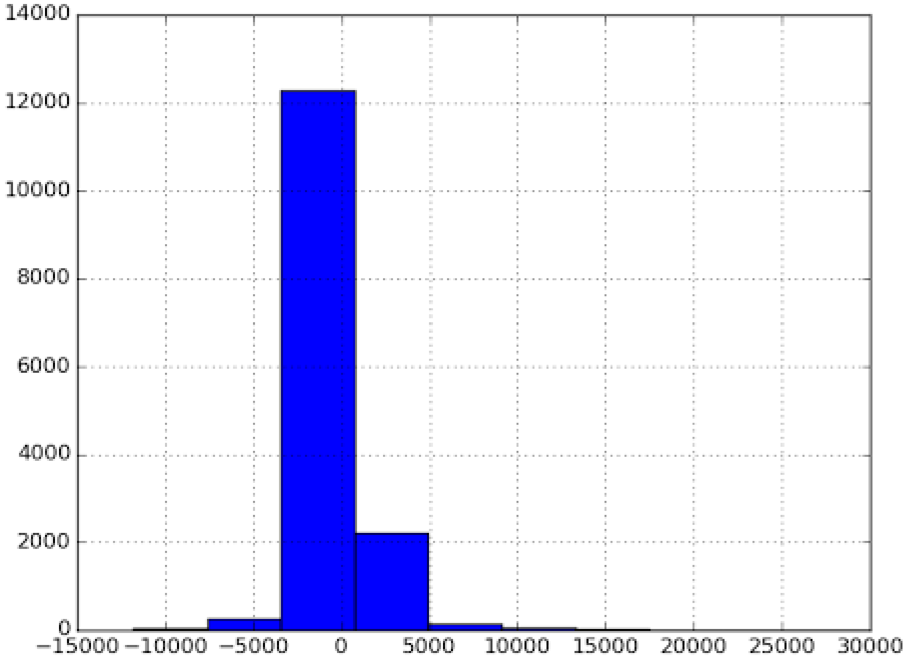
Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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

As with all statistical methods, there are several shortcomings that must be taken into account. After reflecting on this project my biggest worry (and this would be my focus if this was a real-life task) would be the validation of the data coming from the subway system. In order to properly understand the data, it is incredibly important to understand the methods behind HOW the data is collected and stored. There could be certain collection biases that should be taken into account. For example, if the subway station is closed (due to open/close hours or due to things like construction) a data scientist would need to understand if that data is still being collected (i.e. Are we receiving a number of rows with "0" riders for hours when the subway station in in-op). Additionally, the dataset that we received was only for one month. In order to provide a better understanding of the patterns a much larger dataset is required which covers multiple months (and seasons) as well as previous years.

Another potential shortcoming involves the statistical methods and the lack of relevant data. Although we have been given a somewhat large log of ridership data, more research would need to go into discovering other variables that could enhance the statistical methods. For example, during federal holidays or a natural disaster, the average ridership might change drastically. If more potentially relevant shreds of data could be enriched to our dataset, we might find that there are other factors that are better suited to model.

In addition, upon viewing the residuals in a histogram (pictured below) it is clear that their some level of variation that the model is not accounting for. This variance could lie in the fact that we did not have *all* the data necessary to successfully model the data.



5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?