Analyzing the NYC Subway Dataset

vct-6-3-2015 v4 – updated

Submission # 2 to cover items..

1. Some shortcomings of the model or of the statistical test and the dataset need to be discussed a bit further in 2.6 and 5.1 (the answers are connected when it comes to the linear model).
2. The interpretation and usage of the two tailed P value needs to be reviewed.
3. Some minor plotting issues need to be addressed.

Submission # 3 to cover items..

4. I think I got it this time. The point of linear relationship is to locate a model with the data given to predict ridership. And the means the data should cluster tightly near linear line.
5. So with an R of 0.46 and with the plotting of the residuals that show data points away from the "an actual model line" it proves that the linear regression model is NOT a good model.
   a. The reason for this is that the data is scattered/plotted "far" from the actuals and therefore you can not predict.
6. So R2 of .46 tells you it is not a good model
7. The Plot of high level of residuals tells you it is not a good model.
8. So it would be difficult to draw a conclusion or find a linear model to predict the ridership with the given variables/data.
   a. I believe that if you limit or "tighten" the data down such as to limit the number of stations and time of day, you would reduce the variance and come up with a 'better" linear model.

Vince Tierney – ATT

Opening

Overall I enjoyed this exercise and class. I dusted off my Stats books, ordered new and read a lot on-line to refresh. What I most enjoyed was the Python language. The modules truly enable programmer efficiency from my days of Assembler, PL1, APL and REXX.. (All programming languages that where used probability before you were born! ☺). I added comments to all my code in the exercises (as I was taught to do) to explain the code and explain (most to me) what I 'wanted" it to do..

I did a lot of research online. I listed the majority of the References/URLs in a list at the beginning of this sheet and did not note the specific reference in the general answers. I used, copies, took liberties, etc with various information, documentation, code examples, definitions, and anything that would help me refresh my brain and skills.

Questions

Overview

This project consists of two parts. In Part 1 of the project, you should have completed the questions in Problem Sets 2, 3, 4, and 5 in the Introduction to Data Science course.

Completed

This document addresses part 2 of the project. Please use this document as a template and answer the following questions to explain your reasoning and conclusion behind your work in the problem sets. You will attach a document with your answers to these questions as part of your final project submission.

Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

Books:

- Python for Dummies,
- Python in a Day-Wagstaff

Web sites: (besides ones listed in the classes)

1. https://wiki.python.org/moin/SimplePrograms
2. https://pypi.python.org/pypi/ggplot/
3. https://www.python.org/
4. http://www.statsoft.com/Textbook/Multiple_regression#cresidual
5. http://www.pythonforbeginners.com/code-snippets-source-code/python-code-examples
6. http://en.wikipedia.org/wiki/Python_(programming_language)
7. https://www.udacity.com/course/programming-foundations-with-python--ud036
8. http://learnpythonthehardway.org/book/
9. http://www.tutorialspoint.com/python/
10. http://www.pythonlearn.com/
11. http://spotofdata.com/subway-weather-udacity/
12. http://stackoverflow.com/questions/15427692/perform-a-shapiro-wilk-normality-test
13. http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Nonparametric/BS704_Nonparametric4.html
14. http://hamelg.blogspot.com/2014/02/udacity-is-kicking-off-2014-with-new.html
15. http://www.graphpad.com/guides/prism/6/curve-fitting/index.htm?r2_ameasureofgoodness_of_fitoflinearregression.htm
16. http://java.dzone.com/articles/r-ggplot---plotting-multiple
17. https://github.com/sebasibarguen/udacity-nanodegree-nyc-subway/blob/master/code/advanced_linear_regressions.py
18. https://docs.python.org/2/tutorial/datastructures.html
19. http://ubuntuforums.org/showthread.php?t=522688
20. http://ggplot.yhathq.com/
21. https://pypi.python.org/pypi/ggplot
22. http://docs.ggplot2.org/current/geom_histogram.html
23. http://en.wikipedia.org/wiki/Welch's_t_test
24. http://en.wikipedia.org/wiki/P-value
25. http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm
26. http://www.statisticallysignificantconsulting.com/RegressionAnalysis.htm

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?

- The Mann-Whitney U test was used for the subway analysis because we did not know if the data was normally distributed. When using the M-W test the null hypothesis says that both data sets are the same and that *rain* has no impact on the ridership. (The Mann-Whitney test is a nonparametric test that allows two groups or conditions or treatments to be compared without making the assumption that values are normally distributed). A two tail test, checking the area under the curve at both sides of a normal distribution, seemed appropriate as the data/outcome was not fully known at the time. P was 0.025 (single and 0.05 two sides) (P = probability of observing a result)
- Not being a fully trained Statistician I did a lot of research on the internet and in books as to why you would use the MWU test. Most recommendations point to using the WMU test.
  - "If you're in doubt, by all means use the Mann-Whitney; it has good power properties at the normal and if there's a tendency toward skewness or heavy-tailedness will tend to outperform the *t* on shift-alternatives, possibly quite heavily. And people 'recognize it', which is sometimes useful (in terms of requiring less explanation or justification)."

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.

- This non-parametric Mann-Whitney test would be proper where the Welch's two sample T test would not because it assumes equal distribution. The resulting histogram shows an unequal distribution. The *t*-test is the most commonly used method to evaluate the differences in means between two groups.

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.

- Mean entries with rain: 1105.4463767458733,
- Mean entries without rain: 1090.278780151855,
- U-statistic: 1924409167.0, = the "U" in "U-statistics" stands for "unbiased".
- P-value:
  - 0.024999912793489721 for a **one sided test**
  - **0.05 for a two sided test. (2 x one sided test)**
    - P = the statistical significance of a result is an estimated measure of the degree to which it is "true" (in the sense of "representative of the population").

1.4 What is the significance and interpretation of these results?

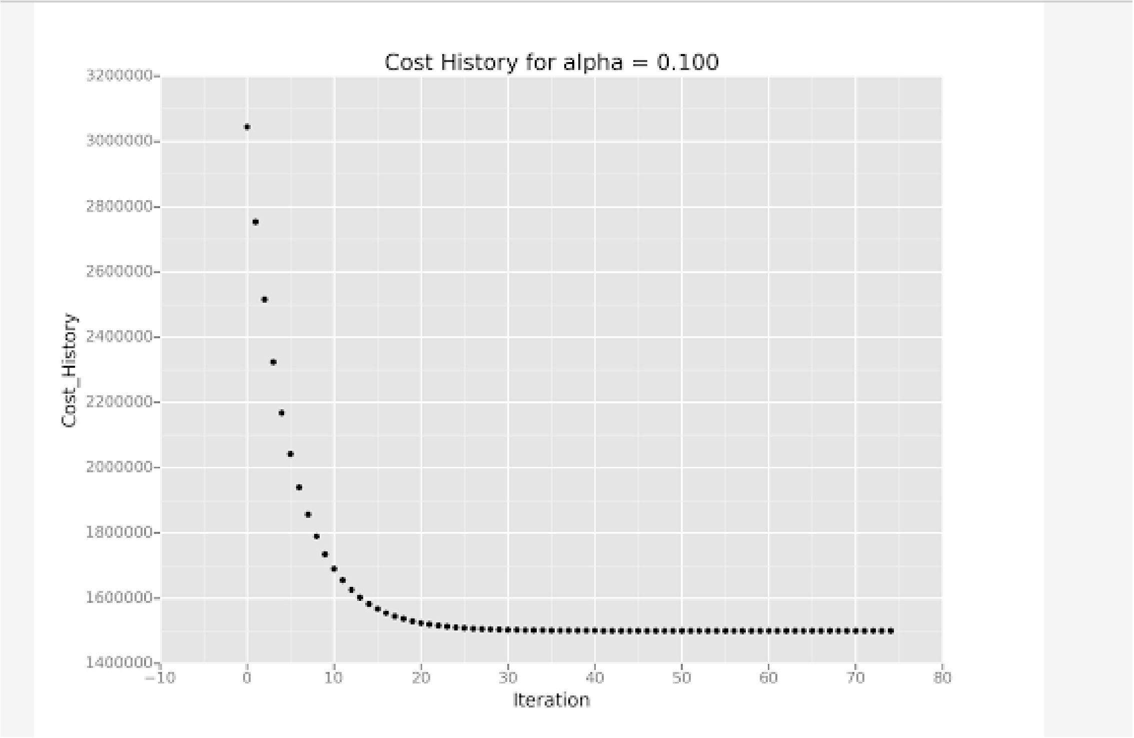
- The mean comparisons say that there are ~ 1.4% (1.357%) more riders per hour when it rains. This seems to draw the conclusion that the ridership is different between rain and no rain. The U-statistic (biases) value is high and that also seems to point to a false hypothesis. The p-value 0.05 shows with confidence that the null hypothesis is false and that ridership is different with rain vs. without rain. (The *p*-value is the probability of observing an effect given that the null hypothesis true).
- In other words, when p < 0.05 we say that the results are statistically significant, meaning we have strong evidence to suggest the null hypothesis is false.

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

- I used gradient descent to test the linear regression coefficients. I kept the default values of learning rate (alpha) 0.1 and 75 iterations. The values were sufficient in converging on a local minimum, as confirmed by plotting the cost history vs. number of iterations. (see plot this section)
- There are additional plots that show linear regression.   (plot is in section 2.6)



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

- Gradient descent (GD) and OLS models(best fix model) where used. These would show linear regression and look for relationships between the features and the predicted values.
- I used
    - 'rain',
    - 'precipi',
    - 'Hour',
    - 'meantempi

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.

- Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
- Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R2 value.”

- I selected rain, precipitation, hour, and mean temperature as those seem to be the most influencing aspects on riders and R2 with this dataset. Other aspects moved R2 closer to 0 which means less linear relationship. (UNIT)
- Used Time of day because riders typical travel at the same time each day (rush hours, theater times, etc)
- Used precipitation because the amount of rain gives you a weight (light, med or heavy) which would influence people wanting to ride and not walk. This expanded the r2 slightly more.

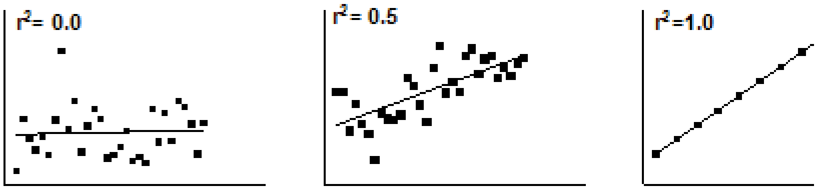
2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?

I used the code in the test to give these results.

- 2.92398062e+00 1.46526720e+01 4.67708502e+02 -6.22179395e+01

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

- Your r^2 value is 0.463968815042
    - “The value r^2 is a fraction between 0.0 and 1.0, and has no units. An r^2 value of 0.0 means that knowing X does not help you predict Y. There is no linear relationship between X and Y, and the best-fit line is a horizontal line going through the mean of all Y values. When r^2 equals 1.0, all points lie exactly on a straight line with no scatter. Knowing X lets you predict Y perfectly.”
    - Examples representing r values



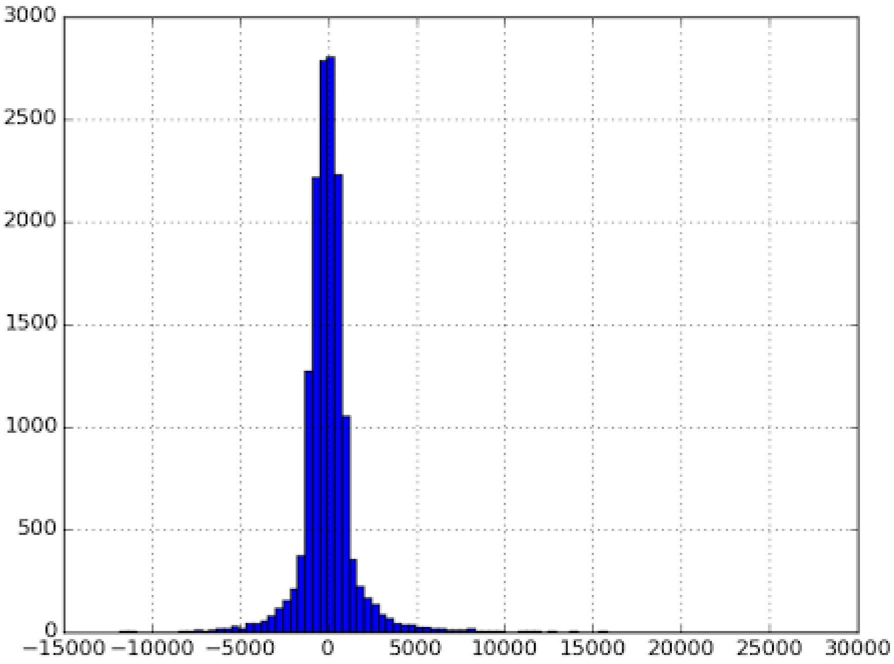
2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

- I think I got it this time. The point of linear relationship is that the data should cluster tightly near or on the function line. So an R of 0.46 and the plotting of the residuals that show data points away from the 0 or “line” proves that the linear regression model is NOT a good model.
- The reason for this is that the data is scattered “far” from the optimum and therefore you can not predict.
- So R2 of .46 tells you it is not a good model
- The Plot of high level of residuals tells you it is not a good model.

- Some had in my thought process that a 46% variance was a GOOD value that would consider linear regression model good where is actually say there is a 64% chance the data will NOT fit the model.

- R2 of 1.0 tells you there is a solid linear relationship, R2 of 0.0 means knowing x you can not predict y. Our R2 of 0.46 tells you there is no way to predict a 100% relationship. We can predict about 46% confidence of the ridership**.** 46% of the total variance in Y is "explained" by the linear regression model R2 value shows you the proportion of the variance in the test data as explained by the model.
- The following plot shows that most residuals fall +-5,000 and to the extent reach out to +- 10,000. I believe that this shows the linear model does NOT show a relationship between rain and non-rain and can predict ridership.
- The residuals fall on each side of "0" .
- The residuals could be explained by general human nature were we are repetitive and creatures of habit. So the weather will not be a factor until it reaches a threshold to overcome human nature.
- Also the residuals could also be explained by errors in data gathering, overall ridership for that day (how do you know how many riders just cancelled their trip due to weather).

- Plot shows the residuals with data provided
- Plot with bins changed to 100 – better representation of the data



- The above plot show the residuals with respect to the linear model

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.

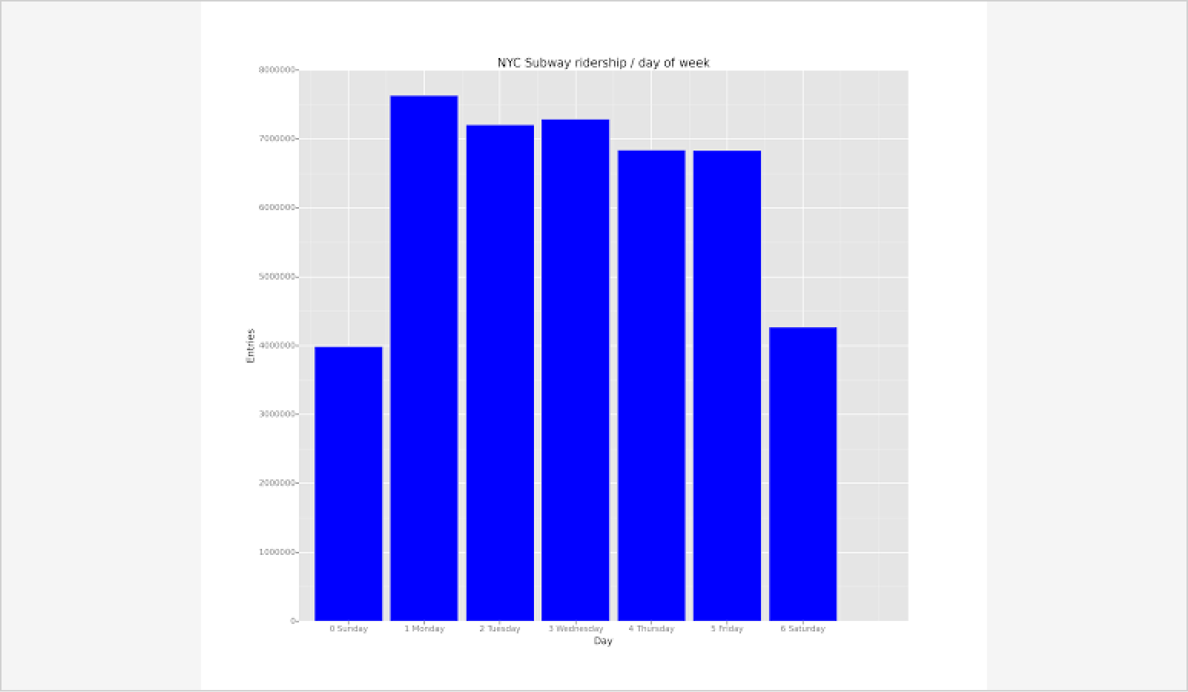
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.

- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

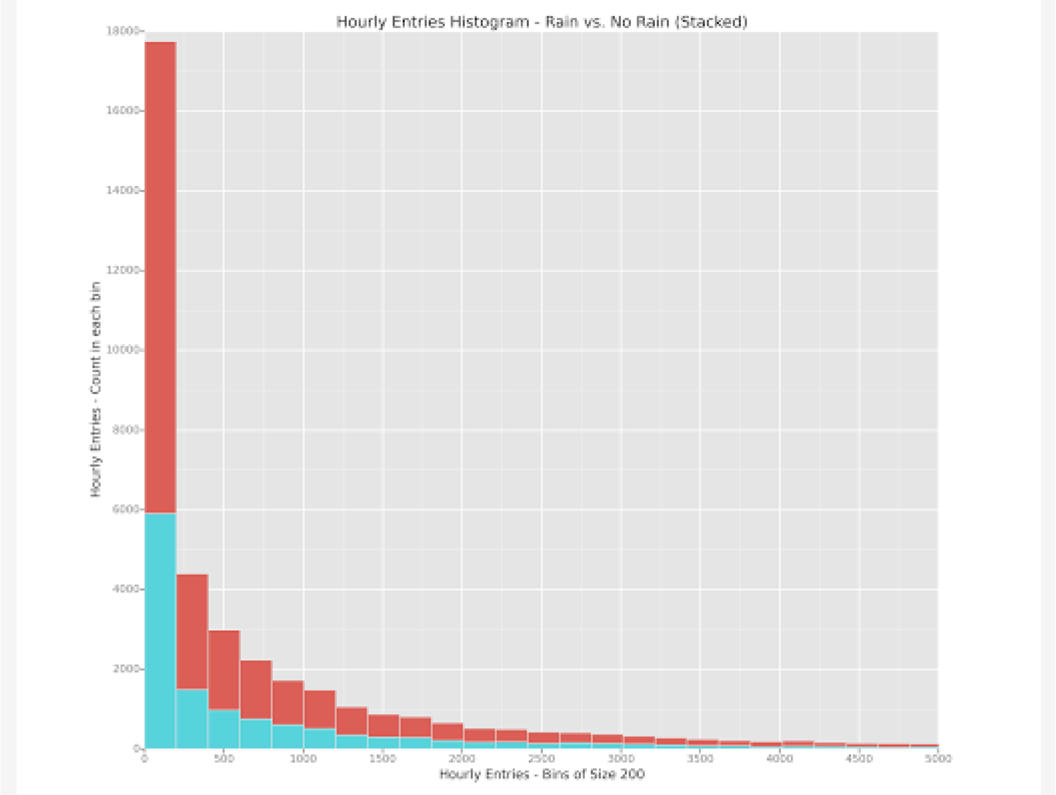
- Plots titles and axis re labeled

This is a plot of Riders by Day ..   Monday's are most busy



- The above plot show the entries of riders by day or week across the data set

Plot of riders on rainy vs non rain days

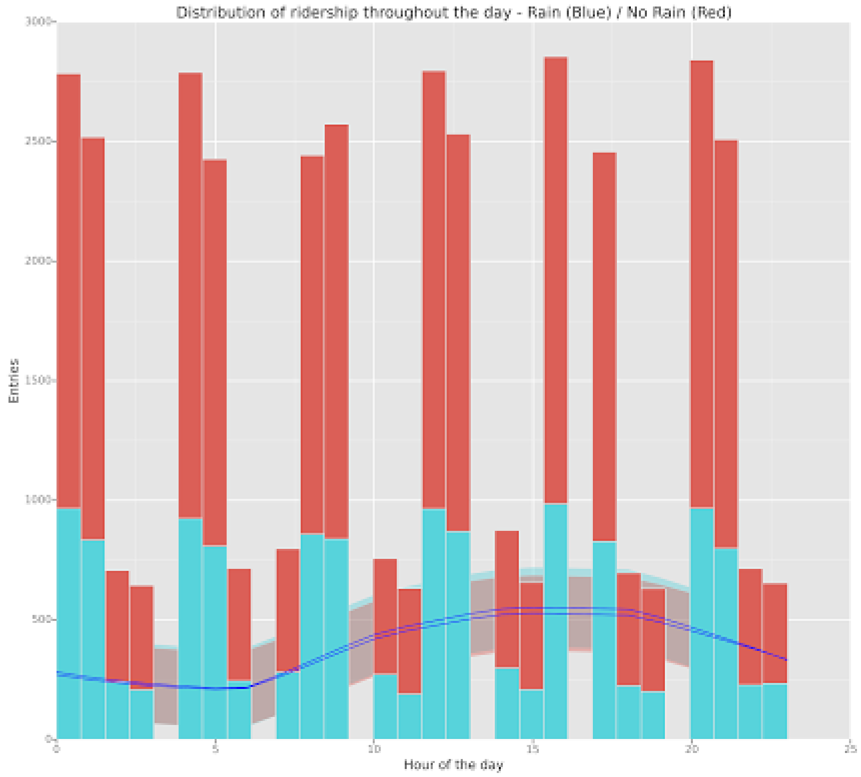


- The above plot show the entries of riders by hour of the day if is it raining or not raining

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. Some suggestions are:

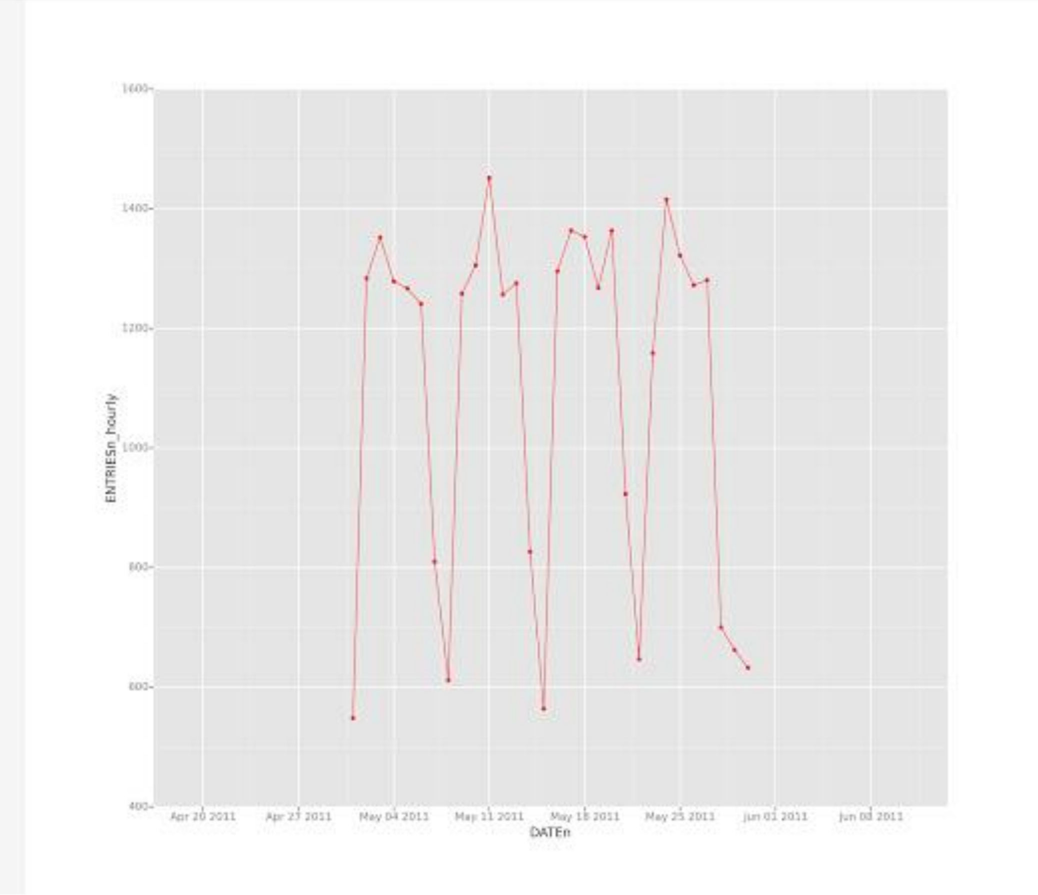
- Ridership by time-of-day
- Ridership by day-of-week

Ridership time of day and a smoothing interval of 0.5. Smoothing is an interesting compliment to the normal ggplot capabilities.



- The above plot shows the rider ship for rain and no rain by hour of the day categorized by day of the week.
- I played with the smoothing graphic so see what additional information I could gain and it appears to show that there is more traffic (riders) starting to ramp up from Mondays to Weds and peaking on Friday. (a busy day for NYC)..

Plot based on Entries/ridership per hour over days – Line format



- The above plot shows a individual day trends. This shows a ramp up in the morning, peak in the afternoon and another spike at ~ 8 pm (pre/post theater/dinner hour or other events).

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?

- The Mann-Whitney U test (p-value: 0.05), shows that with a high level of certainty that more people ride the NYC subway when it is raining. The Mann-Whitney U test was used to confirm that the two data sets are statistically different. The null hypothesis is false. The two distributions are different.
- The results show that that data for rain vs non-rain is not the same and that more customers ride the subway on rainy days.
- The R2 of ~.46 says with 46% confidence that you can not fully predict that value of “Y”. R2 is not the primary stat you need to use to reach this conclusion the MWU-P, coefficients, and R2 all are used to make this conclusion. With all these different statistical tools you can draw th aforementioned conclusion

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.

- The positive coefficient indicates that the presence of rain contributes to increased ridership. With the R^2 being approximately 46% could show some relationship. The means of both data sets are not that different from each other, the Mann-Whitney U test did show there was a statistically significant change in ridership for rain vs. no-rain. Therefor we can claim that rain increases subway 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:

1. Dataset,
    - Shortcomings of the dataset are
        1. The data is representative of a fairly short time frame – 1 month
        2. There looks to be more entries riders than exit riders
        3. The time collected varies by day and is not standardized
    - In my estimation the data set was efficient to a good degree. Said differently, I think there are some dimensions that could have been added to give a truer picture of rider travel preferences. Some additional data points could be:
        1. How far did the rider travel? Short travel or longer travel. I would expect that some riders would not want to pay for a shorter trip vs just walking with an umbrella. Would it be revenue generating to lower the cost of a ticket on inclement days?
        2. Where there extra train cars added? If there was less crowding would more ride the trains?
        3. There were 27m more entries than exits? Did that many riders exit outside of the subway system? Transfer? Or is their an issue with the collection of data?
2. Analysis, such as the linear regression model or statistical test.
    - The linear regression model proved to not be the proper model to use to support the exercise.
    - The subway does have a maximum limit do to the number of cars and riders per car. If you ever rode the subway, there are times the trains just are full and no more riders can get on.
    - I would expect that our dataset, limited to ~40k rows and sometimes a sample of 10k rows, had influence on the overall statistical test. I general the MWU test and others should reflex the associated predictable correctness of a larger set. For such a massive system it would be interesting to use a larger set to validate a larger sample.

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

- Overall the dataset did supply the necessary data for the programs and exercises. I did wonder about the influence from special events such as holidays, sporting events, conventions, etc. Were any of these mixed into the entries and exit numbers? Also were the trains all in time and were there any mechanical malfunctions? The linear model was single variable, Rain or No rain. A multi variable function would be interesting to stretch. You could add local trains (shorter stops) vs express to see if there was a relationship or influence.

- Also I would like to process a dataset that “predicted” poor weather the day before to see if riders planned to ride the trains based on the next day weather report.
- The above could be used to have the MTA add additional trains and plan for additional ridership.