**Analyzing the NYC Subway Dataset**

**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.

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.

https://bespokeblog.wordpress.com/2011/07/11/basic-data-plotting-with-matplotlib-part-3-histograms/

http://pandas.pydata.org/pandas-docs/version/0.15.2/basics.html#iteration

http://blog.yhathq.com/posts/ggplot-for-python.html

http://pandas.pydata.org/pandas-docs/version/0.15.2/indexing.html

http://stackoverflow.com/questions/22391433/count-the-frequency-that-a-value-occurs-in-a-dataframe-column

**Section 1. Statistical Test**

* 1. **Which statistical test did you use to analyze the NYC subway data?**   
     Mann–Whitney U-test  
     Shapiro-Wilk Test

Did you use a one-tail or a two-tail P value?   
Two-tail p value   
  
What is the null hypothesis?   
The distributions of both rainy and non-rainy days are equal

What is your p-critical value?   
0.05

* 1. **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.**   
     Our dataset are not equal sample size nor do they have the same variance.   
     Sample Size of Rainy Days = 44104  
     Variance for Rainy Days = 5619401.454

Sample of Non-Rainy Days = 87847  
Variance for Non-Rainy Days = 5382422.911   
  
Mann-Whitney U-test is applicable to the dataset because the dependent variable is either ordinal or continuous but not normally distributed. Using Shapiro-Wilk Test, it is determined the data is not normally distributed. The p-value of the test is less than the alpha level which we would reject the null hypothesis of the data tested and accept the alternate.

* 1. **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.**Mann\_Whitney U Test:  
     p-value = 0.024999912793489721 x 2 = 0.0499998255869794 = 4.99998255869794 %  
       
     Welch’s t-test:  
     p-value = 0.2695064  
       
     rainy day means = 1105.446   
     non-rainy day means = 1090.279
  2. **What is the significance and interpretation of these results?**  
     Using the results from the Mann-Whitney U-test, the p value is ~0.05 which is less than or equal to the p-critical value. This is telling us there is a significance difference in ridership when it rains compare to non-rainy days. We would reject the null hypothesis and accept the alternate hypothesis. The alternate hypothesis would be the distributions of both rainy and non-rainy days are not equal. The difference in mean of ridership between rainy and non-rainy day is ~15 riders. It may be small but it is statistically significance to matter.

**Section 2. Linear Regression**

* 1. **What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn\_hourly in your regression model:Gradient descent (as implemented in exercise 3.5) OLS using Statsmodels Or something different?**  
     gradient descent
  2. **What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?**   
     The features which I used were rain, precipi, meanwindspdi, and Hour.

Yes, rain was used as a dummy variable along with ‘UNIT’.

* 1. **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.”**Rain would deter people from being outside and force them to find a way to get to their destination without getting wet. This is where the subway comes into play. No standing out in the wet rain while waiting for a cab and plus it is cheaper going this route.

The reason for using Precipi pertains to how much it was raining. The idea is if it rained more, people would more likely ride the subway. The more it rain, the more I believe it would deter people to be outside and look to use the subway instead.

Hour of the day was used because subway are on a set schedule. During certain hours of the day, you tend to have more people walking and travelling in NYC in daylight compare to night time. Also, you would have more subways picking up riders compare to other hours say at midnight.

Adding meanwindspdi helped the r2 value increase slightly compare with other features that I added.

* 1. **What are the coefficients (or weights) of the non-dummy features in your linear regression model?**  
     rain = 1.11135485e+01  
     precipi = 7.20205481e+00  
     meanwindspdi = 6.91129475e+01  
     Hour = 4.61187362e+02
  2. **What is your model’s R2 (coefficients of determination) value?**  
     0.464079690026
  3. **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?**  
      R2 value tells you how ‘close’ your data points are to the fitted regression line overall. It is the percentage of the response variable variation that is explained by a linear model. R2 will always be between 0 and 1. 0 indicates the model explains none of the variability of the response data while 1 indicates the model explains all of the variability of the response data around its mean. The higher the R2 the better our model fits the data.   
       
     I think this linear model to predict ridership is appropriate for this dataset. R2 value came in at 42.5% which is a reasonable value to predicting if people are riding the subway or not. The attempt to predict human behavior is very hard to do when compare with for example manufacturing process.

**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.

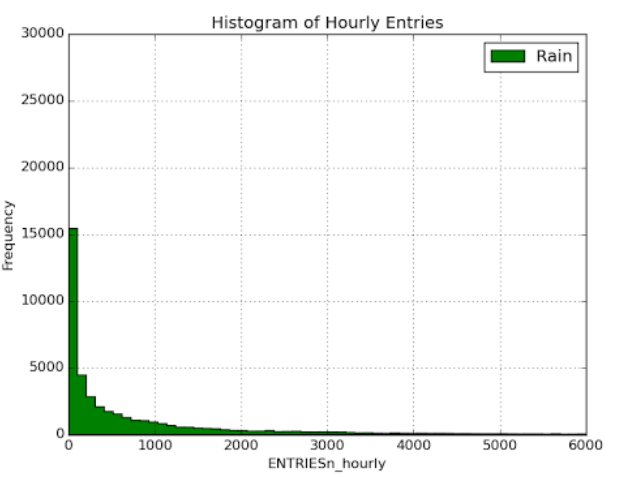
* 1. **One visualization should contain two histograms: one of ENTRIESn\_hourly for rainy days and one of ENTRIESn\_hourly for non-rainy days.**

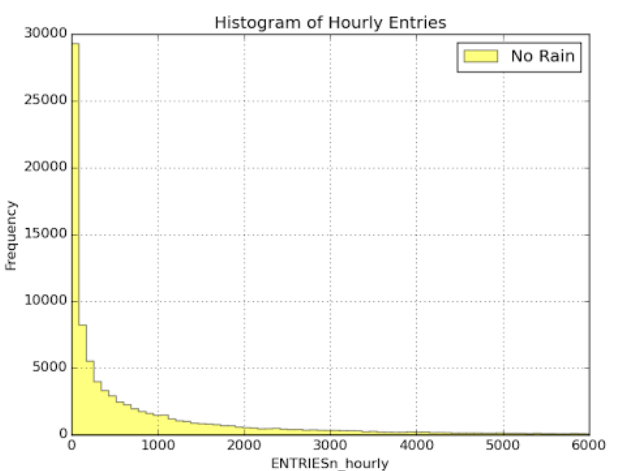
You can combine the two histograms in a single plot or you can use two separate plots.

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.

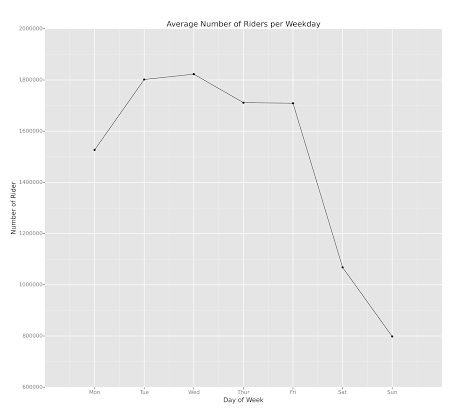
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.





The histogram of hourly entries depicts the frequency of people riding the subway when it is raining and not raining. During non-rainy days, the level of frequency occurs a lot more than rainy days.

3.2 **One visualization can be more freeform**.



The graph above is displaying the daily average ridership for each day of the week. As you can see, more people tend to ride the subway in the mid part of the week peeking on Wednesday before you see a drop off heading into the weekend. The lowest average ridership is on Sunday.

**Section 4. Conclusion**

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

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 conclusion obtain from Mann Whitney U test is the distribution of riders is greater when it is raining compare to non-raining days. Since the p value is less than the p critical value, one would reject the null hypothesis and accept the alternate hypothesis. More people ride the NYC subway when it is raining.
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 results from the statistic tests of Mann Whitney U test says to reject the null hypothesis, distribution of both groups are not equal. The mean of rainy day ridership is 1105 while the mean for non-rain days is 1090. This result would indicate that the rainy days had more people riding the subway when compare with non-rainy days.   
     
   From the linear regression model, the coefficients (weight) for rain had a value of 11.1. The coefficient indicates that for every time it rains you can expect ridership to increase by an average of 11.1 riders.

**Section 5. Reflection**

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

1. **Please discuss potential shortcomings of the methods of your analysis, including: Dataset, Analysis, such as the linear regression model or statistical test.**  
   One might use a longer time frame of dataset to determine the answer for the question. Instead of 30 days increase the dataset to possibly 3 months which would capture the three wettest months for NYC historically. One would also see if any kind of special events that may occur during those days. Those couple of events might skew the results a bit. Another factor is the regular riders, those that ride the subway daily for their job. If it is possible, figure out how many monthly riders there are. These are the people who are going to ride the subway regardless of rain or shine. Another dataset that could help with the model is determine how many subway trains were running during that specific period of hour or track down how many times Uber app was used for a cab ride. One could look into the subway schedule to see what percentage or how many were on-time.   
     
   Using nonparametric tests, we are making fewer assumptions about the distribution of the data. Just because the normality test was performed and showed the data as being not normal doesn’t exactly mean it does. Data can fail the normality test because of presence of outliers. But with our sample data set being so large, it may not have any impact in terms of tests power (misleading).   
     
   R2 cannot determine whether coefficient estimates and predictions are biased. So to combat it, one must assess the residual plots. R2 is just a sample/estimate and not the entire population.
2. (**Optional) Do you have any other insight about the dataset that you would like to share with us?**