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

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

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

What is the null hypothesis?
If rain made any significant difference in the number of subway rider ship.

What is your p-critical value?
1.96%

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 applicable to the dataset because the dependent variable is either ordinal or continuous but 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.
p-value = 0.025 x 2 = .0500
rainy day means = 1105.446
non-rainy day means = 1090.279

1.4 What is the significance and interpretation of these results?
Using the results from the Mann-Whitney U-test, the p value is 5% which is greater than our p-critical value telling us that there is no significance in the ridership when it rains or not. The difference in mean of ridership between rainy and non-rainy day is ~15 riders. At first glance, 15 riders is about 1.3 percent of the average total for the day which isn't significance which the mann-whitney u-test is telling us. With that being said, we would reject the null hypothesis

Section 2. Linear Regression

1.1 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

1.2 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 fog, maxtempi, and rain. Yes, dummy variables were used.

1.3 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."

The temperature of the day, I feel, impacts people decision to walk to their destination (if it is within reason). If the mean temp of the day was low say below freezing, more people would want to get off of the street and ride a nice warm subway versus waiting for a cab.
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 cub and plus it is a lot cheaper.

1.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
fog = 5.11970917e+01
maxtempi = -4.85891106e+01
rain = -1.13910731e+01

1.5 What is your model's R2 (coefficients of determination) value?
0.425831938562

1.6 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?
The closer the R2 value is to 1 the better the regression model is to 'fitting' within the line. In this case, ~43% of the data points fall within the results of the line formed by the regression equation. I would believe it is an acceptable model given the value of R2.

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.

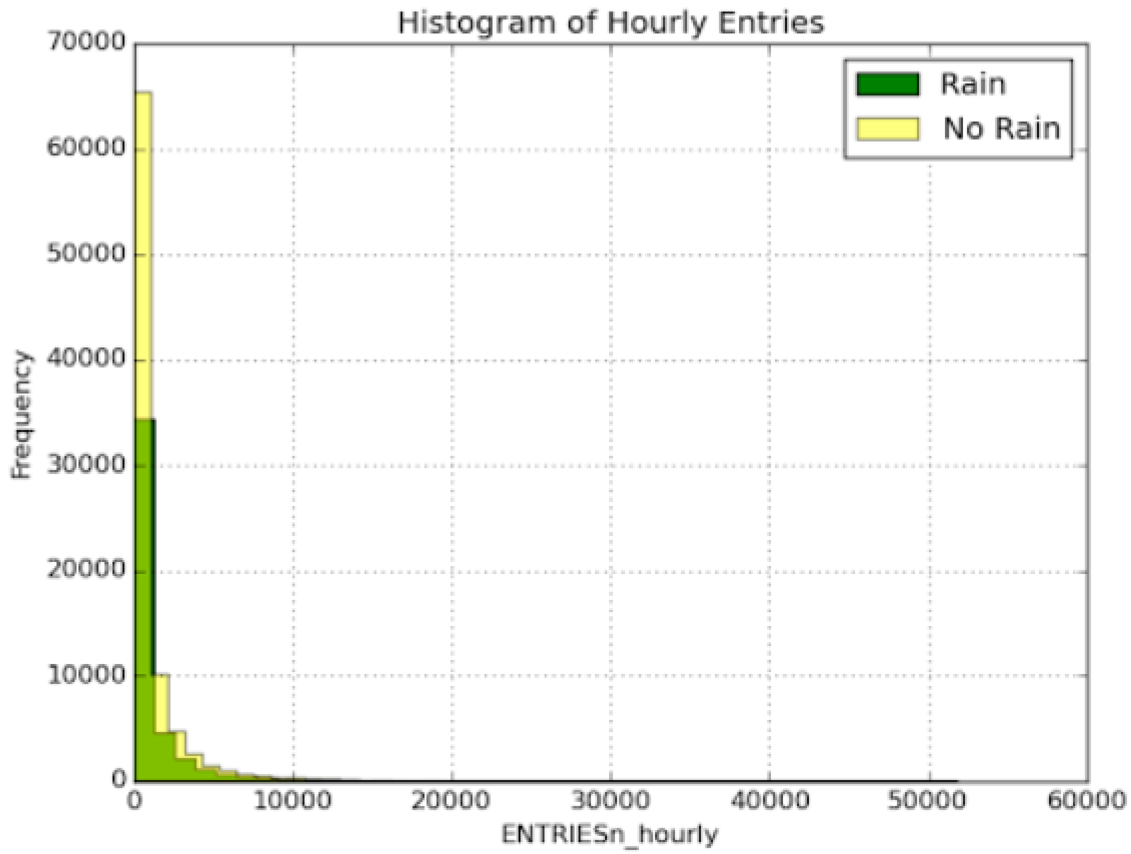
4.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

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.

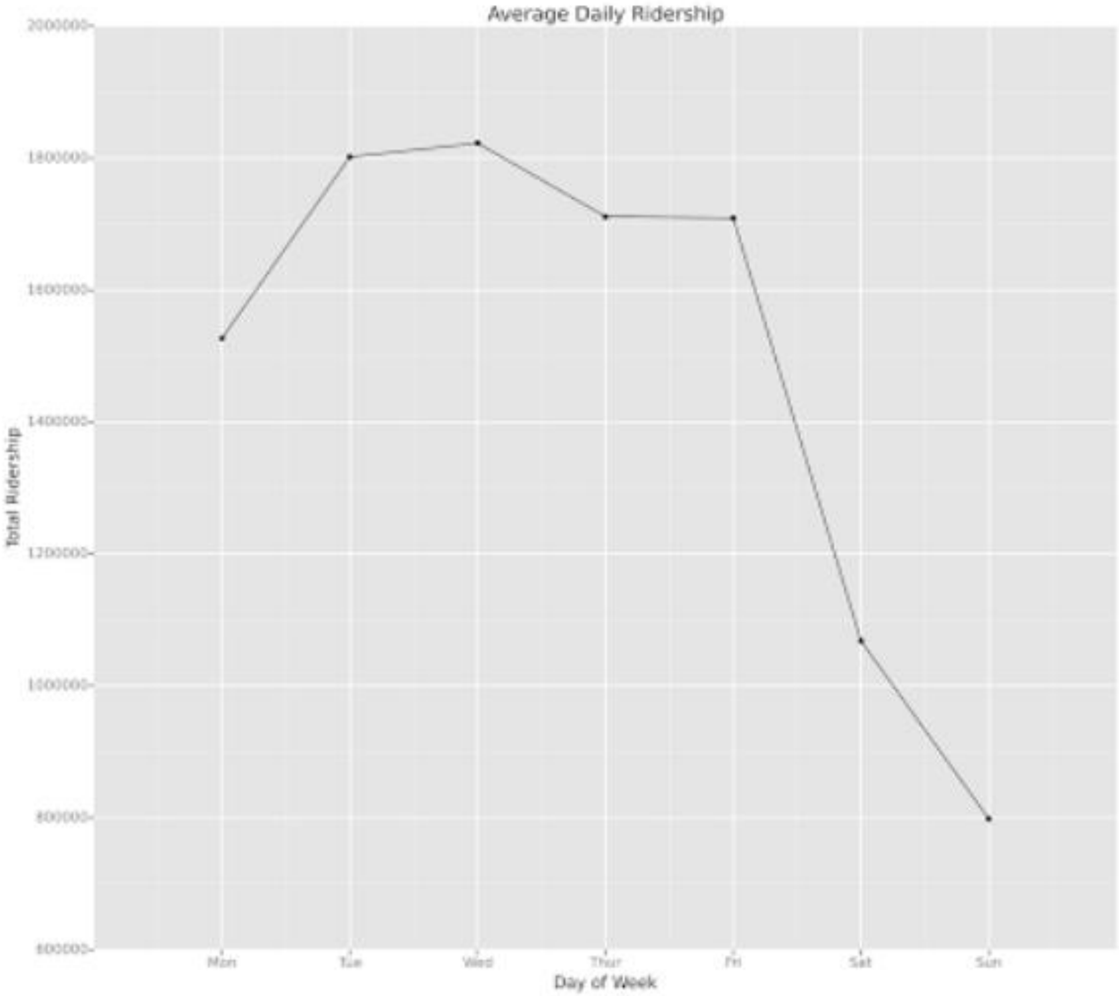
For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows 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.

<u>Section 4. Conclusion</u>

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?
With the tests performed and data collected so far, it isn't possible to determine if more people ride

the NYC subway on rainy or non-rainy days. If the question ask the number of people riding the subway, then the non-rainy days have more people riding. Number of people riding for non-rainy days equal 28,966,813 while rainy days equal 15,112,589.

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 results from the Mann-Whitney U-test says statistically, the rain has no impact on ridership. The p value is greater than the p-critical value which tells me the null hypothesis can be rejected. To support the Mann-Whitney U-test, by looking at the average ridership between rainy and non-rainy days is just ~15 riders but more riders occur during rainy days. The difference of 15 riders says no major impact in terms of ridership as it is only 1% of the average riders. The graph of average daily ridership shows more people ride during the mid-week peaking on Wednesday then on the weekend. This tells me that if it rained on Tuesday and Wednesday more often than it would push the average ridership towards rainy days. R2 coefficients of determination is .426 which is somewhat acceptable but it isn't great in giving us an accurate prediction going forward.

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.
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.

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