Analyzing the NYC Subway Dataset – Bator Sutton

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, and 4 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.

I had some difficulties with plotting (mainly in making the axis in both graphs of question 2.1 in the same scale) and found help in Stackoverflow discussions as well as the ggplot website. I've also found many good tips in the Udacity forums.

Section 1. Statistical Test

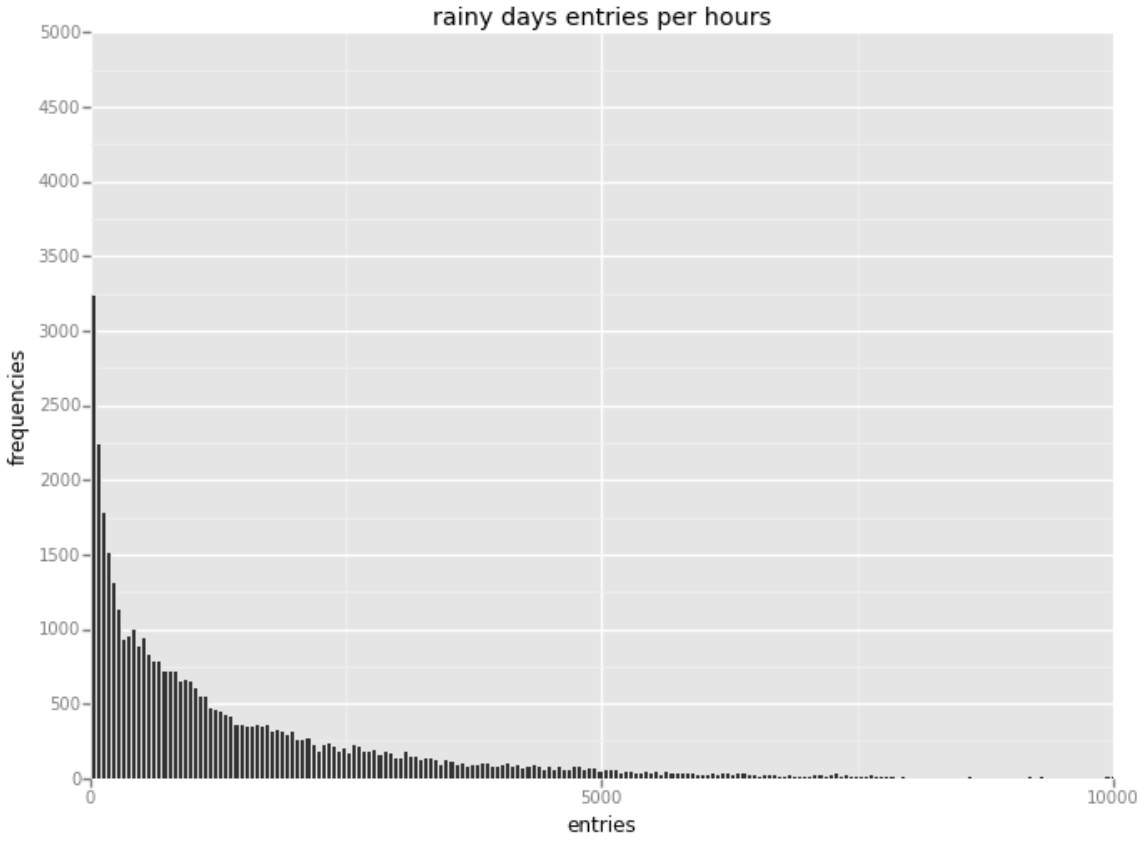
1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

I've used the Mahn Whitney test . The Null hypothesis is that the two groups (in our case – daily entries to subways in non rainy days and rainy days) are not different, IE rain doesn't influence the number of NYC subway riders. The P critical result is: 0.019309634413792565*. Since we're looking for a one sided result (IE we claim that there are more riders when it rains) then we need to multiply the P * 2 = ~0.38 which is still smaller than 0.05, the standard critical P.

* I would like to note that when I used the exact same query on the extended DS, I got a "nan" P value in the Mahn Whitney test. Therefore, just for calculating the Mahn Whitney test, I used the "data master" file.

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 t test is relevant in cases where there's a normal distribution of the data. Since we have N > 5000, the best way to see if our data is distributed normally is through visualization of the data. I used ggplot to plot the ENTRIES_HOURLY column and saw that the data isn't distributed normally (see below), and there's a clear skew towards lower values of Entries_Hour. That meant that I need to use the Mahn Whitney test that doesn't assume a normal distribution.



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 : 0.038619268827585
Mean on rain days (I used the V2 sample) : 2028.19603547
Mean of no rain days: 1845.53943866

1.4 What is the significance and interpretation of these results?

The significance and interpretation of these results is:

- The null hypothesis is rejected: since p critical is below the standard P critical (<0.05), it means that there IS a difference between the two groups
- The means of the 2 groups suggested that on average, there are ~180 more passengers on rainy days compared to non-rainy days

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:

OLS using Statsmodels or Scikit Learn

Gradient descent using Scikit Learn

Or something different?

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

Rain, fog, hour, weekday. I also used as dummy variables the station and condition

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 added rain as we saw that it has influence on the number of passengers using the T test.
- I added hour as I thought the hour itself (rush hour vs/ nights for example) will have an influence on the number of passengers.
- I added weekday as I thought during the week there'll be more use of public transportation (due to rider to work)
- I added the station as I assumed there stations that some are busier than others and can indicate a central location.

- I added fog as saw it has a lot of influence on the R2 value
-

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

- rain 135.565006
- hour 123.504048
- fog -329.577191
- weekday 932.905223

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

R2 = 0.433465

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 R2 value means that the features I selected can "explain" ~43.3% from ridership. It means that more than 50% remains un explained and their for predicting ridership based on these features only will not lead to very accurate results. Personally I think this R2 value isn't good enough so it may be possible that this is not appropriate model. However, it's also possible that weather condition and data such as time of day and day of the week can only account for that much in predicting ridership. At this point I would like to mention that I did try to add more features the model (such as min temp etc.) but I didn't find they added something significate to the R2 result.

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.

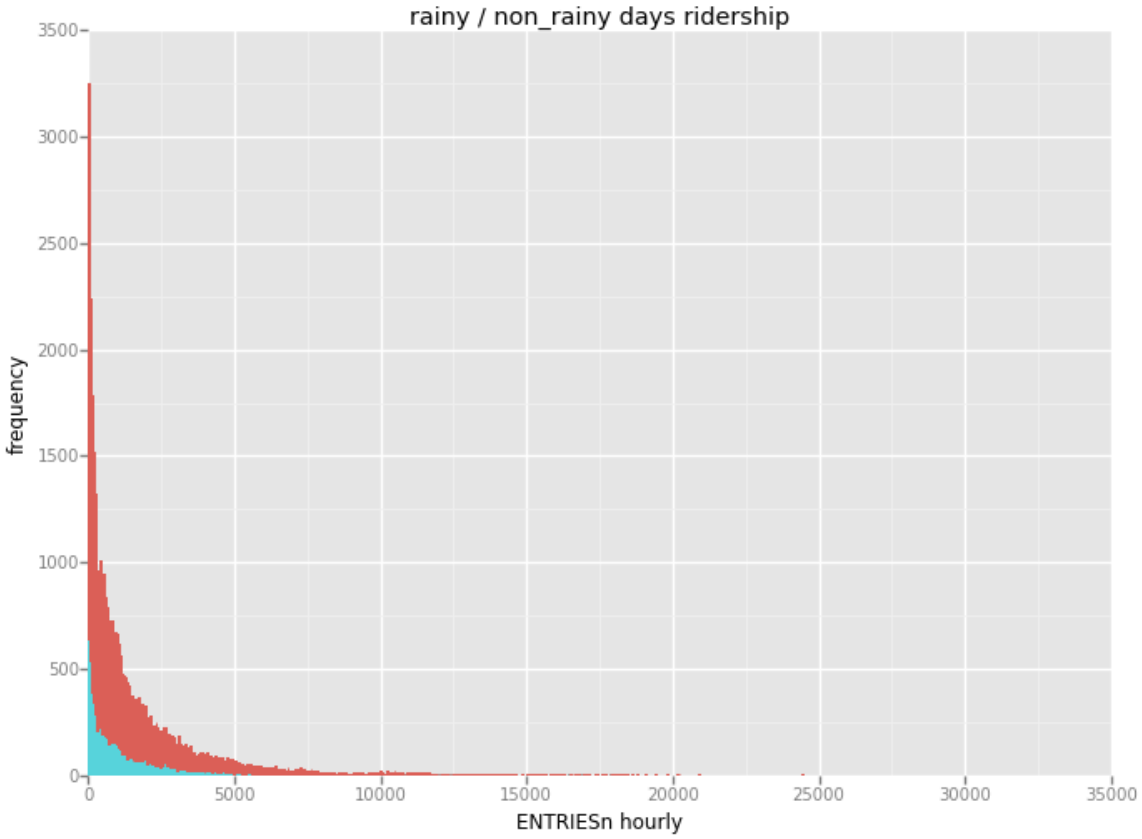
You can combine the two histograms in a single plot or you can use two separate plots.

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.

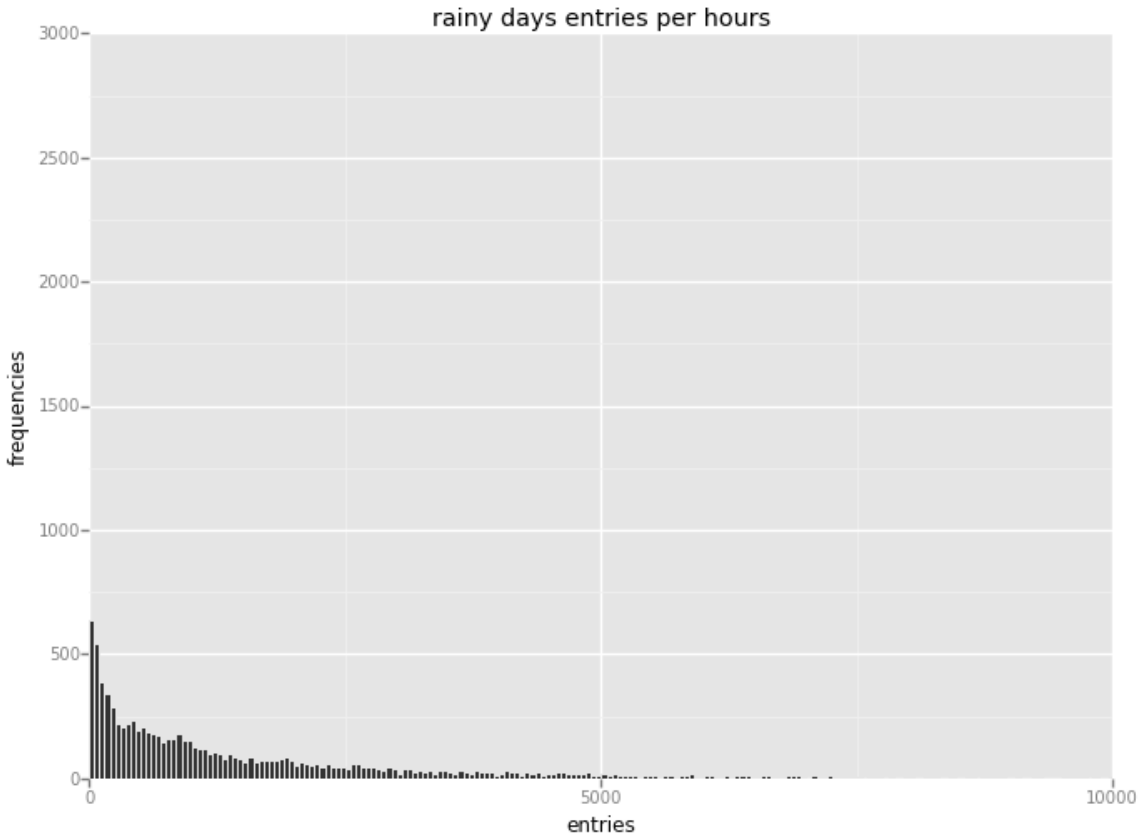
For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows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

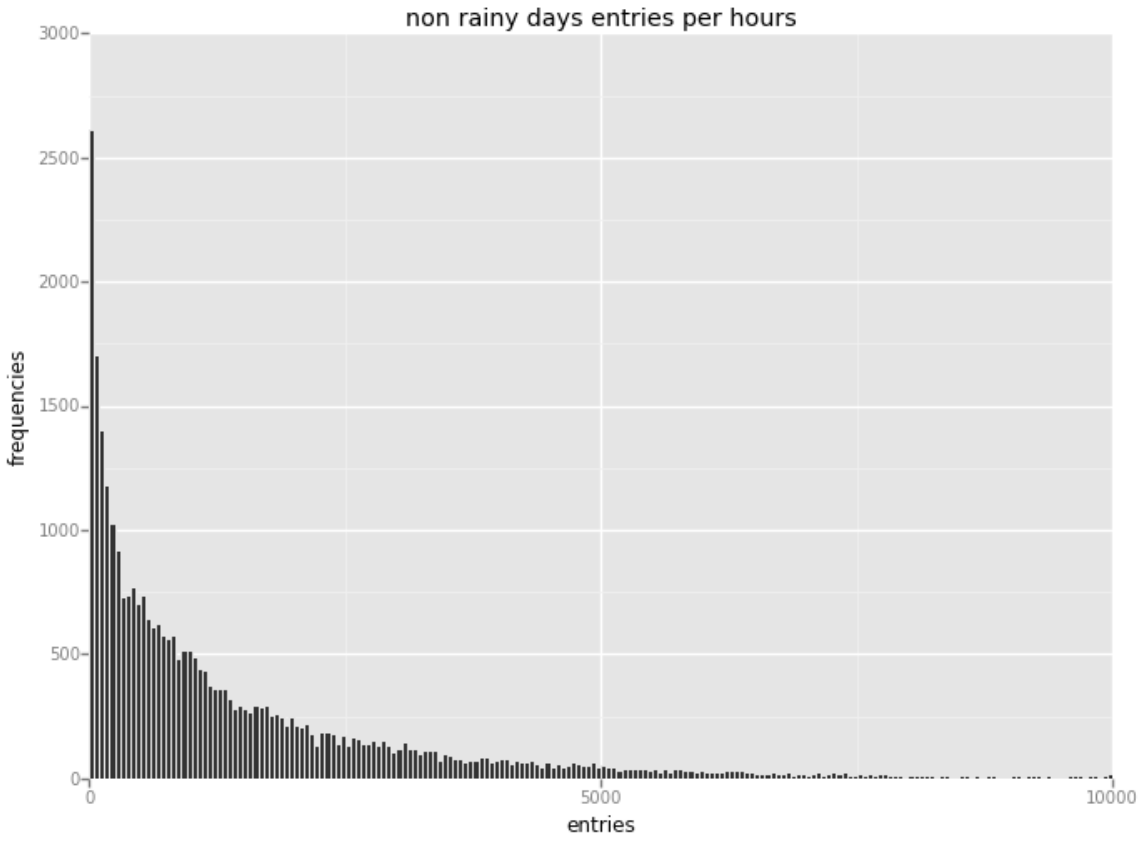
Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.

In one histogram (stacked)



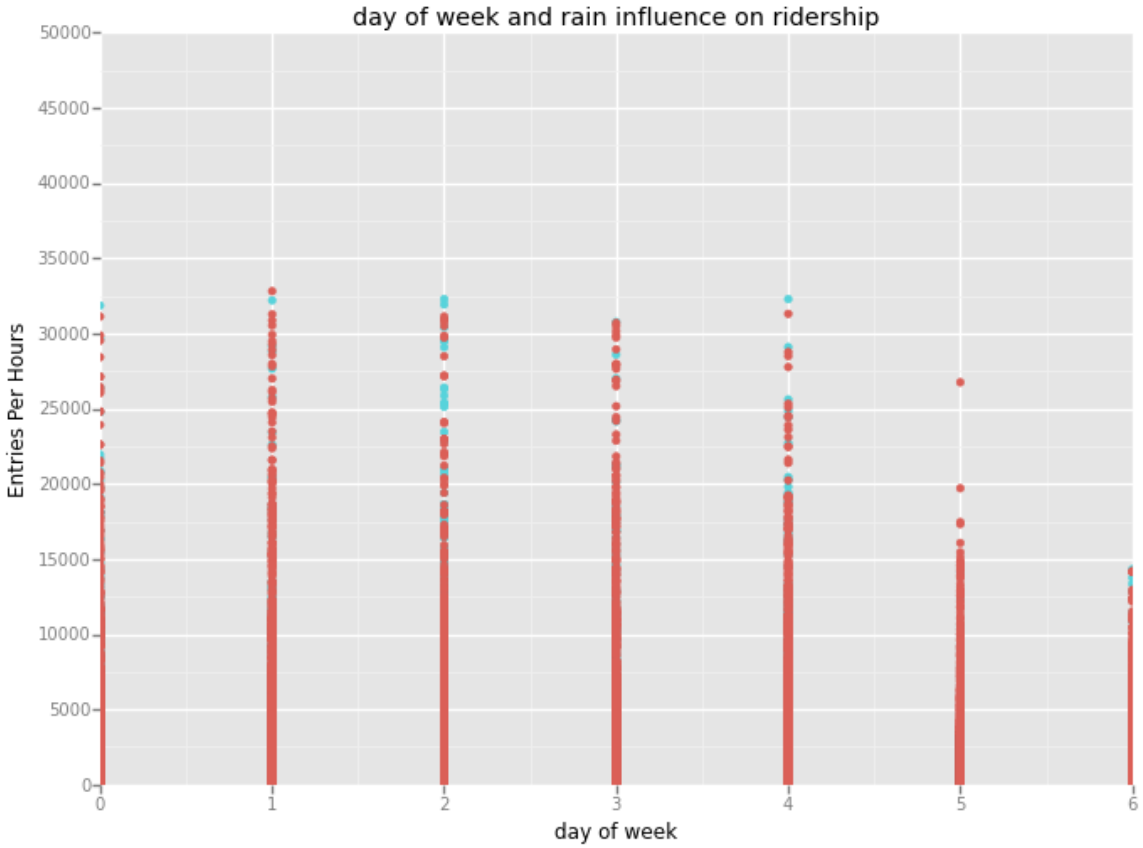
Separate:


rainy days entries per hours
3000
2500
2000
1500
1000
500
0
frequencies
0
5000
10000
entries



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



(rain is indicated by color)

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 analysis of the data leads me to believe that there are more people that rides the NYC subway when it rains. This however doesn't necessarily mean that rain is the *cause* of increase of ridership. When analyzing the data I tried to check if rain on its own isn't the cause for increase in ridership, but rather the low temperature (that often comes with rain). I couldn't reach such conclusion. However, it's possible that since we're only basing our research on the DS that was given to us, we're missing other features that may influence the ridership or be correlated to rain - for example, in rainy days the price of a subway ride is lower. This could mean that the reason for more ridership in rainy days isn't the rain itself, but the the lower price per ride. However, since I didn't have such data at my disposal – I

do reach the conclusion that more people ride the train when it's raining vs when it's not raining.

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.

I was led to this conclusion both by the T Test analysis (P value was very low), meaning that the 2 groups of data – ridership with rain and ridership without rain, are significantly different, in a way that can't occur due to common variance.

In addition, when building the linear regression I could see that the rain factor on its own contributes to the R2 value, meaning that not only does rain have an effect on ridership, the fact that it's raining (or not) can improve the prediction of number of Ridership that time .

This also shows in the parameter that was given to the "rain" feature in the linear regression, which is quite high.

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.

As mentioned above, I think that the DS didn't necessarily include all relevant information to predict the Ridership. For example, it lacked data on special occasions (such as big concerts) or holidays, Drop off stations (I would assume that long train rides won't be influenced greatly by rain, whereas short one will be more common on rainy days), Etc.

With regards to the linear regression model, I felt that in some point I'm just trying to add features without any good explanation as to why they should contribute to the result. Finally I preferred not to include them in the model even though they did improve the R2 result.

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