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.

http://stackoverflow.com/
https://www.youtube.com/watch?v=SqcxYnNlI3Y
http://www.statisticshowto.com/what-is-an-alternate-hypothesis/

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 is used to analyze the NYC subway data.
The two-tailed test is used.

The null hypothesis is that two populations (the ridership with rain vs without rain) are the same.

My p-critical value is 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.

Because both rainy and non-rainy histogram are non-normally distributed, as such the non-parametric Mann-Whitney U test is good fit.

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.

rain_mean: 1105
no_rain_mean: 1090
one-sided p value: 0.0245

1.4 What is the significance and interpretation of these results?

Since I am doing 2-tailed test, the p value is 0.0245*2=0.049 satisfies(less than) the p critical value 0.05, so we reject the null hypothesis that ridership is the same with vs without rain.

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?

Gradient descent

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

I use 'precipi', 'meantempi','hour' and 'meanwindspdi' features.

I take 'UNIT' as dummy variables.

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 believe ‘precipi’, ‘rain’,‘meantempi’ and ‘meanwindspdi’ could be major factors to affect people make decision to stay indoor or outdoor. The ‘hour’ feature is observed how ridership varies with time of day

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

rain: Coefficients is 8.35086061e+00
percipi: Coefficients is 2.95851077e+00
meantempi: Coefficients is -4.34275560e+01
hour: Coefficients is 4.68191545e+02
meanwindspdi: Coefficients is 5.55959004e+01

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

R2 =0.464

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?

The closer the R2 is to 1, the better our model is to describe the observed phenomenon. Due to our goal is to predict the ridership based on the predictors (features), given this low R2 value, this model is not very good fit. It tells me only 46% of variation in the ridership can be explained by the variation in the predictors of module.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

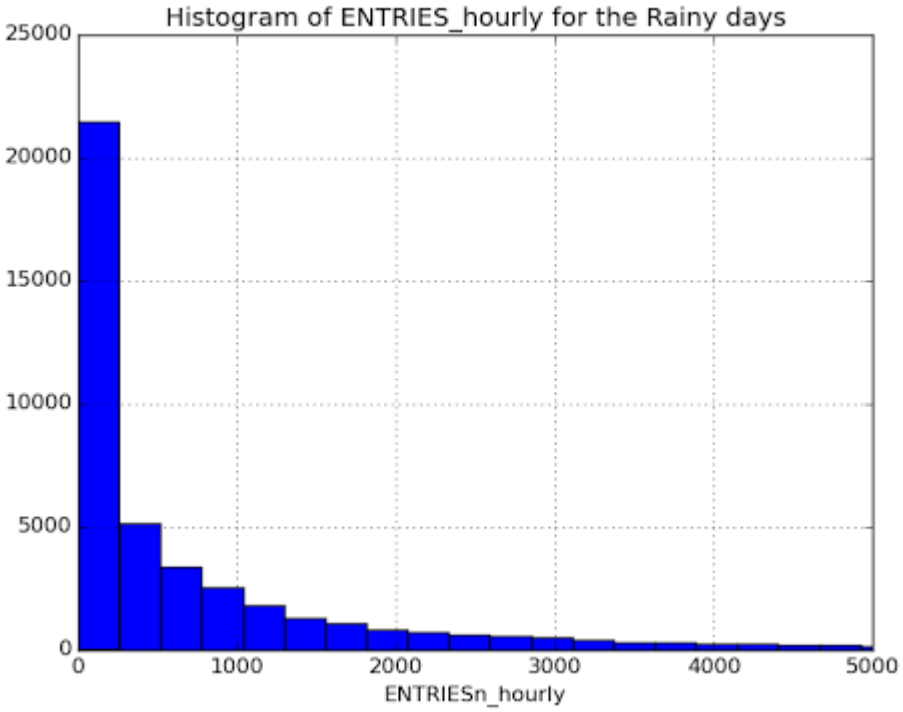
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

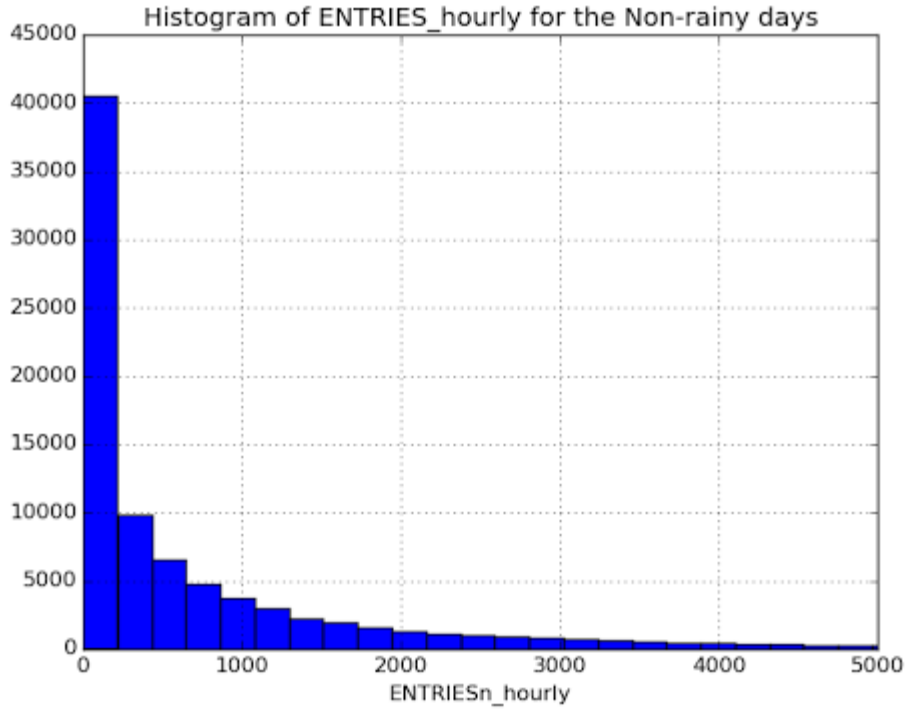
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

- 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.

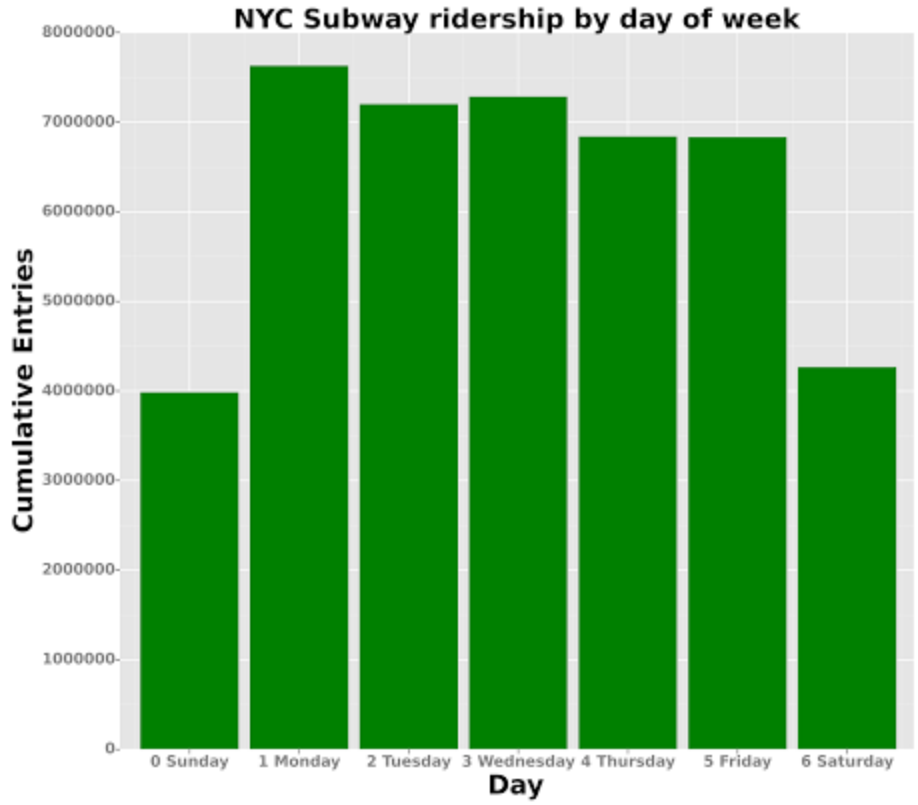




The graph of histograms of subway entries for both rain and no-rain show both distribution are not normally distributed. Also the frequency of ENTRIESn_hourly for non-rainy days is higher than rainy days.

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



The bar chart shows that the cumulative hourly ridership is higher on weekdays than weekends.

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?

It seems more people ride the subway when it is raining vs 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.

The Mann-Whitney U test result: the double-sided p value is 0.0245*2=0.049 and is less than p critical value 0.05. So there's a statistical significant difference between the ridership of rainy and non-rainy days. We reject the null hypothesis. Considering the mean of ENTRIESn_hourly on rainy days was greater than the mean of ENTRIESn_hourly on non-rainy days, conclude more people ride the subway when it is raining vs not raining. In the liner regression module where 'rain' and 'percipi' feature had a positive theta of 8.35 and 2.96 respectively.

Section 5. Reflection

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

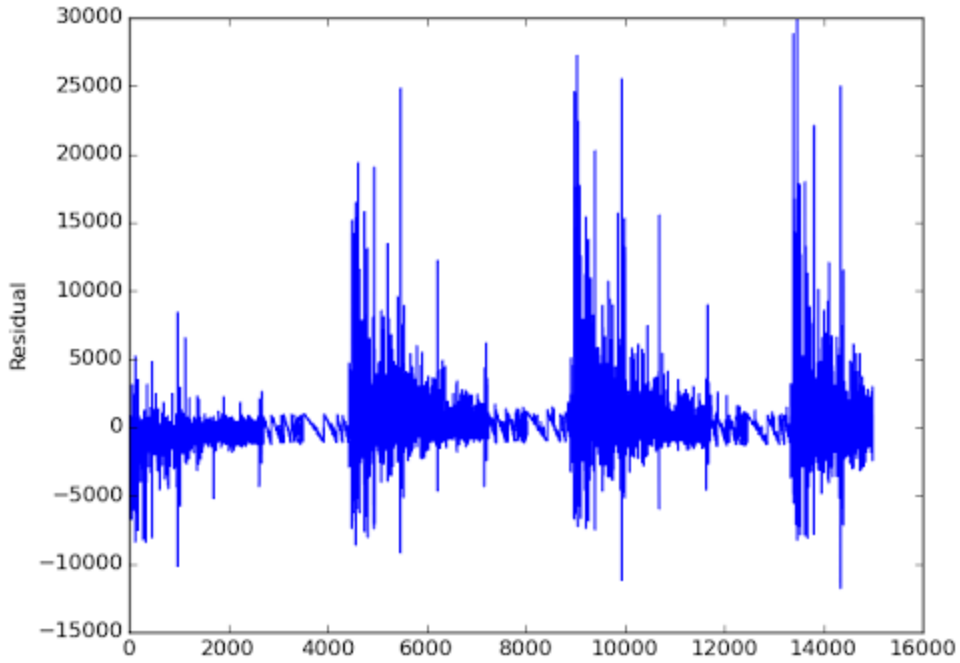
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset:

   Increasing the sample size may change the results of analysis and like increase time interval. Also I would like to see season features in the dataset.

2. Analysis, such as the linear regression model or statistical test.

   On the other hand, based on R2 result, it seems insufficient to prove the weather has correlation with the ridership. It is unclear ridership has a liner relationship with rain. Also the residual plot looks like does not randomly dispersed around the horizontal axis, so it seems a linear regression model is not appropriate for the data;



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