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, 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.

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?

For this project we used the Mann-Whitney-U-Test to analyze the subway data as a two tailed test. In this case the null hypothesis is that the average ridership on days with rain does not differ significantly from days without rain. The p-critical value in this instance was set at 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.

The Mann-Whitney test does not rely on the data to be normally distributed, which made it a good candidate for our scenario since the data I was working with did not follow 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.

For the test I got the following values:

- *Mean With Rain: 1105.4463767458733*
- *Mean Without Rain: 1090.278780151855*
- *P-value: 0.05 (two tailed)*
- *U Statistic: 1924409167.0*

1.4 What is the significance and interpretation of these results?

Because of the large U statistic and p-value of 0.05 there is evidence to support rejecting the null hypothesis. i.e. There is a statisticaly significant difference between ridership on days with rain vs days without.

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:

a. OLS using Statsmodels or Scikit Learn
b. Gradient descent using Scikit Learn
c. Or something different?

For the regression analysis we used both A and B in different parts of the analysis. OLS on problem set 3.5 and Gradient descent on problem set 3.8.

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

For gradient descent I used the following fields in the model: 'rain', 'precipi', 'Hour', 'meantempi','fog', 'mintempi', 'meanpressurei', 'meanwindspdi'
For OLS I used the following fields in the model: 'rain', 'Hour', 'maxpressurei', 'minpressurei', 'maxdewpti', 'mindewpti', 'maxtempi', 'mintempi'
For both I used the following dummy variable: 'UNIT'

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 R^2 value."

Here are my reasons for including each:

- rain, fog, mintempi, maxtempi - these seems like they could easily drive people to take a more consistent (not impacted by weather) sheltered form of transportation
- percipi, meantempi, maxdewpti, mindewpti, meanpressurei - These seemed like they might contribute to people taking the subway more or less in a less significant way. When adding them they did increase the r^2 value.
-

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

- rain -57.102807
- Hour 65.541167
- maxpressurei 261.128488
- minpressurei -524.342714
- maxdewpti 28.060998
- mindewpti -11.433030
- maxtempi 6.573379
- mintempi -35.082744

2.5 What is your model's R^2 (coefficients of determination) value?
OLS: `0.481647439531`

2.6 What does this R^2 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 R^2 value?

In this case an R^2 value of 0.48 indicates that with our given parameters we can explain about 48% of the variability in the Entries of the NYC subway data.

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.

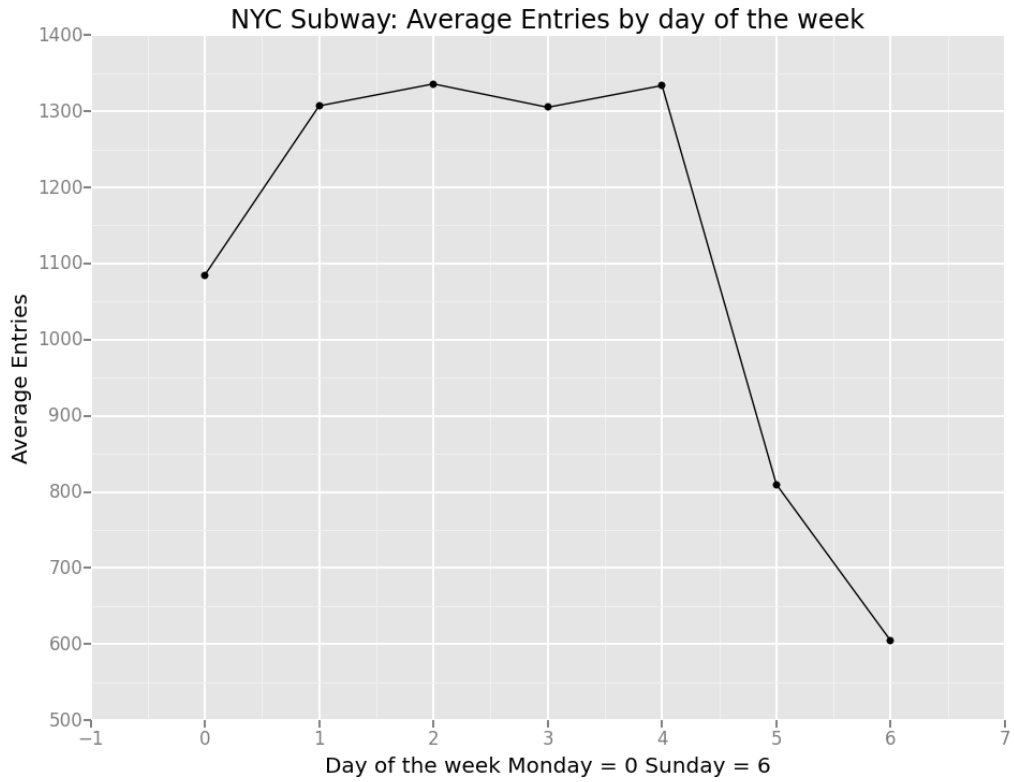

Figure1.

Figure1 shows the average number of entries into the subway by day of the week. This is interesting because it clearly shows that far fewer people ride the subway on the weekend, but also interestingly, fewer people ride on Monday, then the rest of the week .

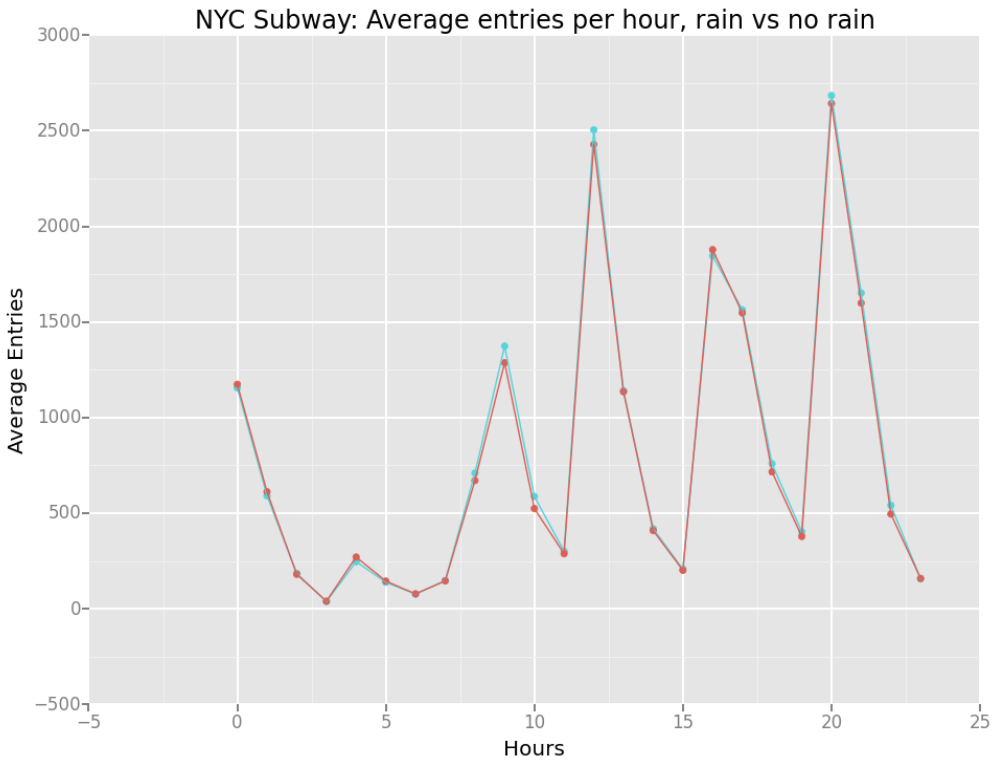

Figure 2.

Figure 2 shows the average entries by hour in the rain and no rain. This is interesting to see the normal fluctuations of the daily boardings are similar for rain vs no rain. It also gives a good visual way to see that though we have found statistical significance between rain and no rain, the values are still very close overall. Note there is an issue with ggplot that will not allow it to show the legend, it is a bug, reported and fixed in a later version, but not the version that is installed via pip. In this case red is no rain and blue is rain

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.

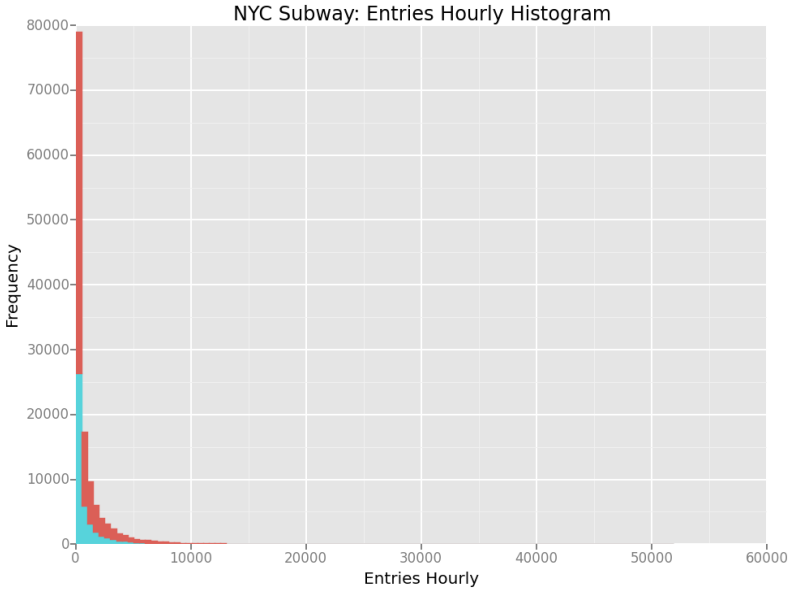

Figure 3.

Figure three shows a stacked histogram of the Hourly Entries into the subway for times when it is raining and when it is not. Note there is an issue with ggplot that will not allow it to show the legend, it is a bug, reported and fixed in a later version, but not the version that is installed via pip. In this case blue is no rain and red is rain

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

See figure 1 and 2

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?

Using the statistical methods and tools I have learned in this class I can say with confidence that more people ride the NYC subway when it is raining, then when it is not raining. This was originally observed after analyzing using the Mann-Whittney-U-Test and later confirmed with a Linear regression using both OLS and Gradient Descent methods.

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.

Initially we performed the Mann Whitney U test on the data given which produced a very large (1924409167.0) U statistic with a p-critical value of 0.05. This demonstrated that the means of the two samples were related.

Later we used a Linear regression analysis to show that rain did contribute to the boarding of passengers in the subway. Using the OLS methods we computed an r^2 value of 0.481647439531 using the parameters: 'rain', 'Hour', 'maxpressurei', 'minpressurei', 'maxdewpti', 'mindewpti', 'maxtempi', 'mintempi'. This indicates that these factors do contribute to the number of boardings but only explain about 48% of the variability in the boarding data.

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 - the data is provided by someone else and I am not able to independently verify the accuracy of the data. We take it on faith that the data is accurate, but there could be issues with data collection or methods of tabulation. Further more there would be good

benefit in seeing if the rain at a time of day would give a more accurate indication of the number of boardings at a particular time of day. It would also be helpful to have a larger number of days to help identify trends from day to day.

2. Analysis, such as the linear regression model or statistical test. The linear regression model gets slower and slower as you add more parameters. it might be helpful to spin up a high computer node on a cloud provider and try to broaden the analysis to include more factors and iterations using the increased computing power.

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