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://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

http://www.graphpad.com/guides/prism/6/statistics/index.htm?how_the_mann-whitney_test_works.htm

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 used Mann-Whitney U test to analyze the NYC Subway data. It had a one-sided p-value of 0.024999. The null hypothesis is that there is 50% probability that a random observation from the ‘Rain’ sample had more entries into the subway than a random observation from the ‘No Rain’ sample.

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.

This test is applicable due to the size of each groups. Both groups are over 100 values. The assumption the Mann-whitney U test assumes the distribution of the ridership of the two samples has the same shape. The graph of each group shows that this is true of the dataset.

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.

The results from the Mann-Whitney U test were p value = 0.024999, with_rain mean = 1105.4464 and without_rain mean = 1090.27878.

1.4 What is the significance and interpretation of these results?

The significance of these results is that the data in each group is not identical. There is a likelihood that the values in the with_rain dataset will be larger than any of the value in the without_rain dataset.

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?

I used Gradient Descent for my regression model.

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

I used rain, fog, minimum temperature, travel hour, minimum pressure and precipi. I did not have any dummy variables in my model.

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 substituted in many of the feature that were included into the data to see if I could improve my model over my initial selection of rain, fog, hour and temperature. My original selection was based on why I would choose to take the subway if I normally would walk or take some other mode of transport. Minimum pressure and precipi seemed to have the biggest impact on my model.

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

1445.43229243, 118.81370118, 644.9446506, 2330.7121009, 244.48226869, 1945.09122323

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

0.465391474466

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 indicates that the linear regression model fits at almost 50%. I think this linear model to predict ridership is appropriate as it is attempting to predict human behavior which can encompass more variables than are available in this data set. Human behavior is difficult to predict do to internal variables as well as external such as weather.

Section 3. Visualization

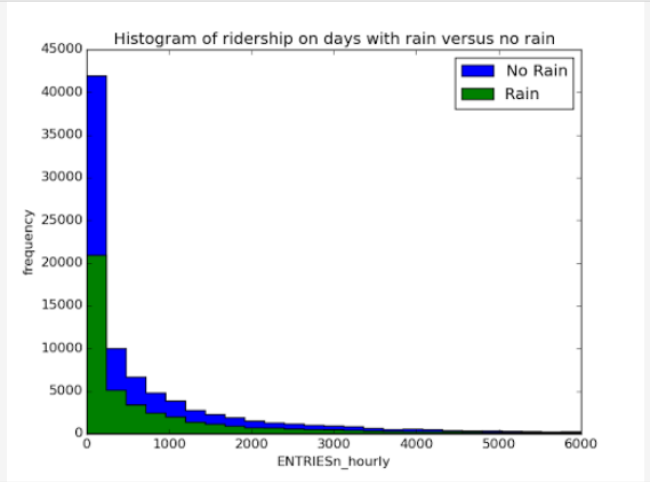
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

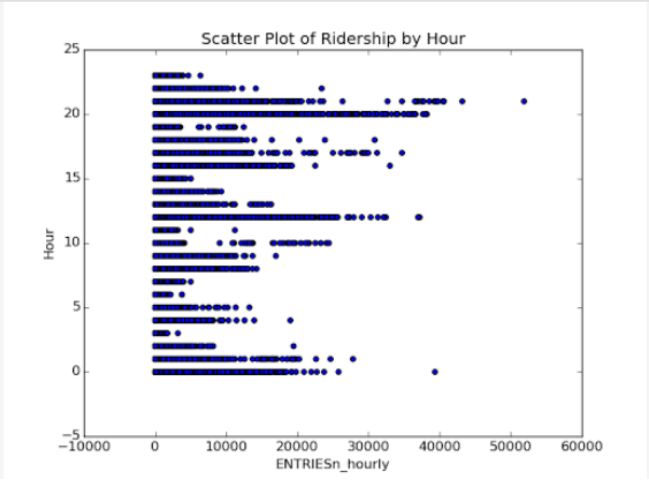
- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



This histogram shows the relationship of ridership between rainy day and non-rainy days. The key insight shown is that the ridership is lowest for non-rainy days as the frequency of ridership below 120 entries per hour is highest for non-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



This scatterplot graph show the relationship between ridership and the time of day. Peak ridership is in the evening between 8:00pm and 10:00pm followed closely by ridership between 12:00pm and 1:00pm.

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?

Yes, slightly more people do ride the subway more when it is 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.

When analyzing the data, there are a few statistics that seem to tell the story with the remaining statistics that back up that analysis. There are approximately twice as many non-rainy days as rainy days in New York as captured by the data presented and the total ridership on rainy days has the same proportion to total ridership on non-rainy days. The mean ridership per observed hour on rainy days is 1105 and the mean ridership per observed hour on non-rainy days is 1090, only a difference of 15 riders per hour.

The Mann-Whitney U test does show that the datasets are not identical and the linear regression with gradient descent does indicate that the features I used in my model has a reasonable predictor of whether someone would ride the subway in NYC. There seems to be more to the story as to whether someone would ride the subway rather than just based on rain or no rain. Other factors that influence ridership are hour of the day, temperature and actual precipitation.

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

A potential shortcoming of the methods of my analysis is the lack of different perspectives. I only used basic statistics, Mann-Whitney and linear regression to analyze the data. There are so many different ways to look at the dataset and analyze it, most of which are not included in this analysis. This analysis is a very narrow look at the dataset.

Another shortcoming is that the dataset cannot possibly incorporate all the reasons why people may take the subway when it is raining. Does it matter where they are going or what they are coming from? The location of the subway station would affect how many people are using the subway when it is raining. Also, with the data set only capturing one hour of four, there is data missing in the analysis as the weather can change on a dime. The original dataset also reported it as raining if it rained at all that day. This can skew the data as well.

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