Analyzing the NYC Subway Dataset

Short 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 1. Statistical Test

1. Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value?

I used a Mann-Whitney U Test with a two-tail P value

2. Why is this statistical test appropriate or applicable to the dataset?

Mann-Whitney U Test doesn't assume equal sample size and equal variance. On top of that, the data is not normally distributed; hence we had to rule out Welch t-test

3. What results did you get from this statistical test? These should include the numerical values: p-values, as well as the means for each of the two samples under test.

I calculated the p-value through scipy.stats.mannwhitneyu and found 5.48213914249e-06, however this is for a one-sided hypothesis. For a two-sided p-value we need to double it : 1.0964278285e-05.

4. What is the significance of these results?

As the calculated p-value is far below the significance level of 0.05, we can reject the null hypothesis and assume that there's a significant impact of the rain on the number of entries in the NYC Subway and that both samples (rainy and non-rainy) are not part of the same population.

Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction in your regression model:

Gradient Descent as in 3.5

2. What features did you use in your model? Did you use any dummy variables as part of your features?

I used all the below features :

- Precipitation
- Temperature
- Wind Speed
- Fod
- Pressure
- Hour, Day of the Week and Weather Conditions with Dummy Variables

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.

There were some features that I decided to rule out because of potential multicollinearity and I decided to keep the feature that I thought was the most impactful of the correlated features such as :

- Weekday and Day of the Week

- Rain and Precipitation

- Precipitation
    - Precipitation levels can potentially increase/decrease the number of entries as heavy rain or light rain may have an impact on the commuter's decision to take the subway or not.
    - Rain is correlated to Precipitation so I decided to rule it out of the model as how heavy it rains has a bigger impact than just the information whether it rains or not

- Temperature
    - Temperature is also another factor as people may choose between transports depending on the temperature
    - While Subway may be a great transportation while it's cold outside, it may not be the preferred transportation type if it's warm outside.

- Wind Speed
    - People might be annoyed when there's a lot of wind outside (especially women with their hair), hence this also could be an important feature.

- Weekday
    - Weekdays (Mon_Fri or Sat-Sun) are certainly having an impact on the ridership of the tube as a lot of people travel during Sat-Sun for their week-end.

- Fog
    - Fog can potentially lead to an increase in the ridership due to people preferring to take a safe transport than taking their car or a bus which can be dangerous in those conditions.

- Pressure
    - Barometric pressure is a good metric to assess the weather conditions as high/low pressure can indicate an incoming storm and this can influence people in taking the subway by fear of rain coming. Although it is not something that we can measure/see ourselves, people can feel the change by looking at the sky/clouds and assess if there's a raining threat

- Hour
    - The time of the day at which we measure the subway ridership has an important impact as there will be more entries during the day than during the night for example, or during peak-hours (work related)

- Day of the week
    - Week of the day is correlated to weekday, however we can still dive down into specific days as there might be more ridership on Monday than on Fridays for example.
    - I decided to remove Weekday as it was correlated to Day of the week and I believe that day of the week is the feature that will help to build a better model

- Weather Conditions
    - As explained in "Pressure", this is what the people are assessing by looking at the sky and taking their decision depending on the fact that it is cloudy, clear, rainy etc...

4. What is your model's R^2 (coefficients of determination) value?

My model's R2 is equal to 0.5480 (0.547993675572)

5. What does this R^2 value mean for the goodness of fit for your regression model?

It means that 54.8% of the predicted values through the model are fitting to the original data

6. Do you think this linear model is appropriate for this dataset, given this R^2 value?

Considering the limited number of available variables (weather and time), I do believe that this model can be appropriate. The subway ridership can not only be limited to weather/time variables as we need to add many other variables such as social factors/behaviours.

Section 3. Visualization

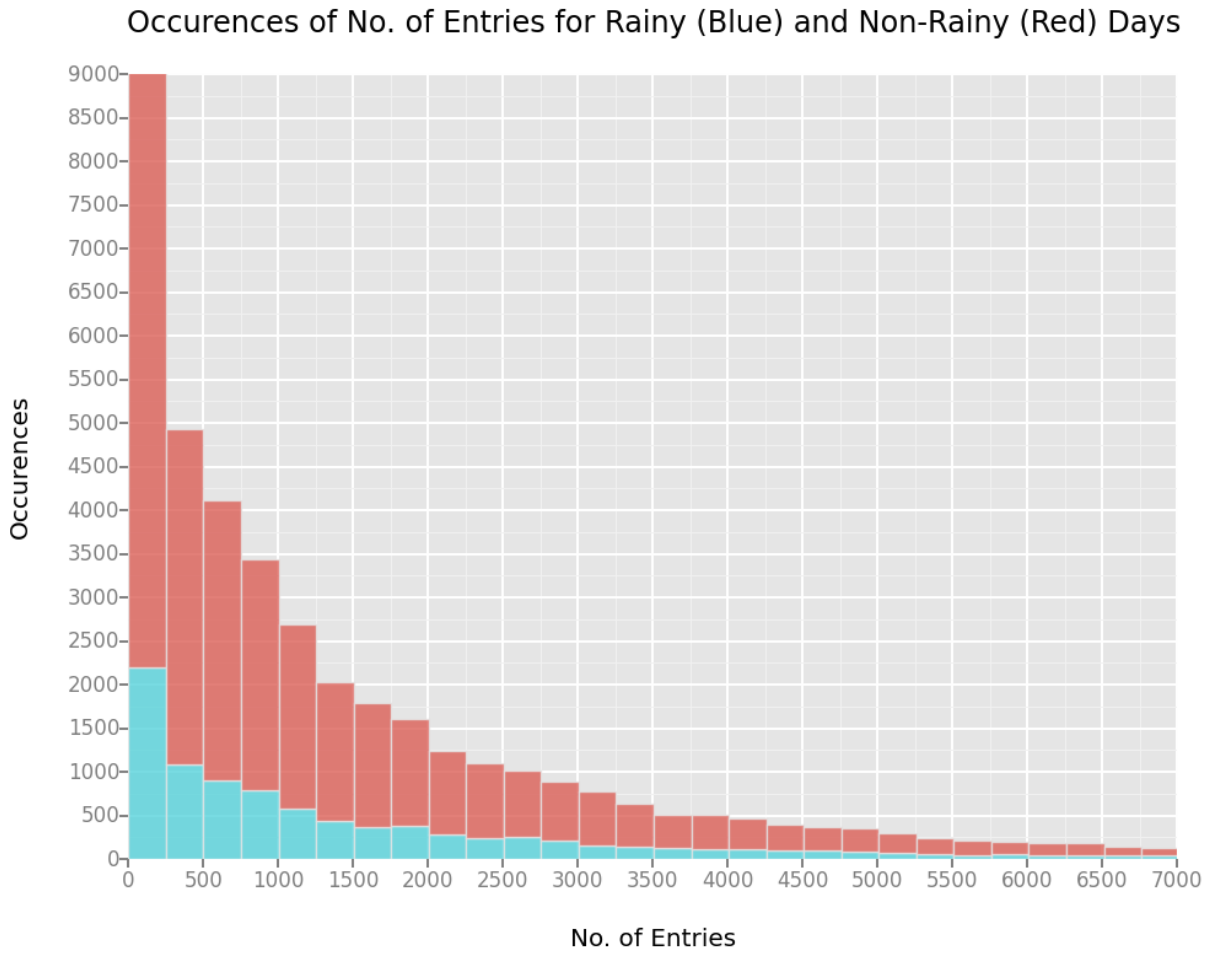
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatterplots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.

1. One visualization should be two histograms of ENTRIESn_hourly for rainy days and non-rainy days. Remember to increase the number of bins on the histogram (by having larger number of bars, each with smaller width). The default bin width is not sufficient to capture the variability in the two samples.

Unfortunately the current version of ggplot doesn't show the colour legend for the variables used; hence I have mentioned the colours in the title of the graph. In order to have a more conclusive graph, I have edited the x and y axis limits in order to focus on the relevant part of the graph. I used a bin size of 250 in order to fit ggplot automatic x axis sequence.

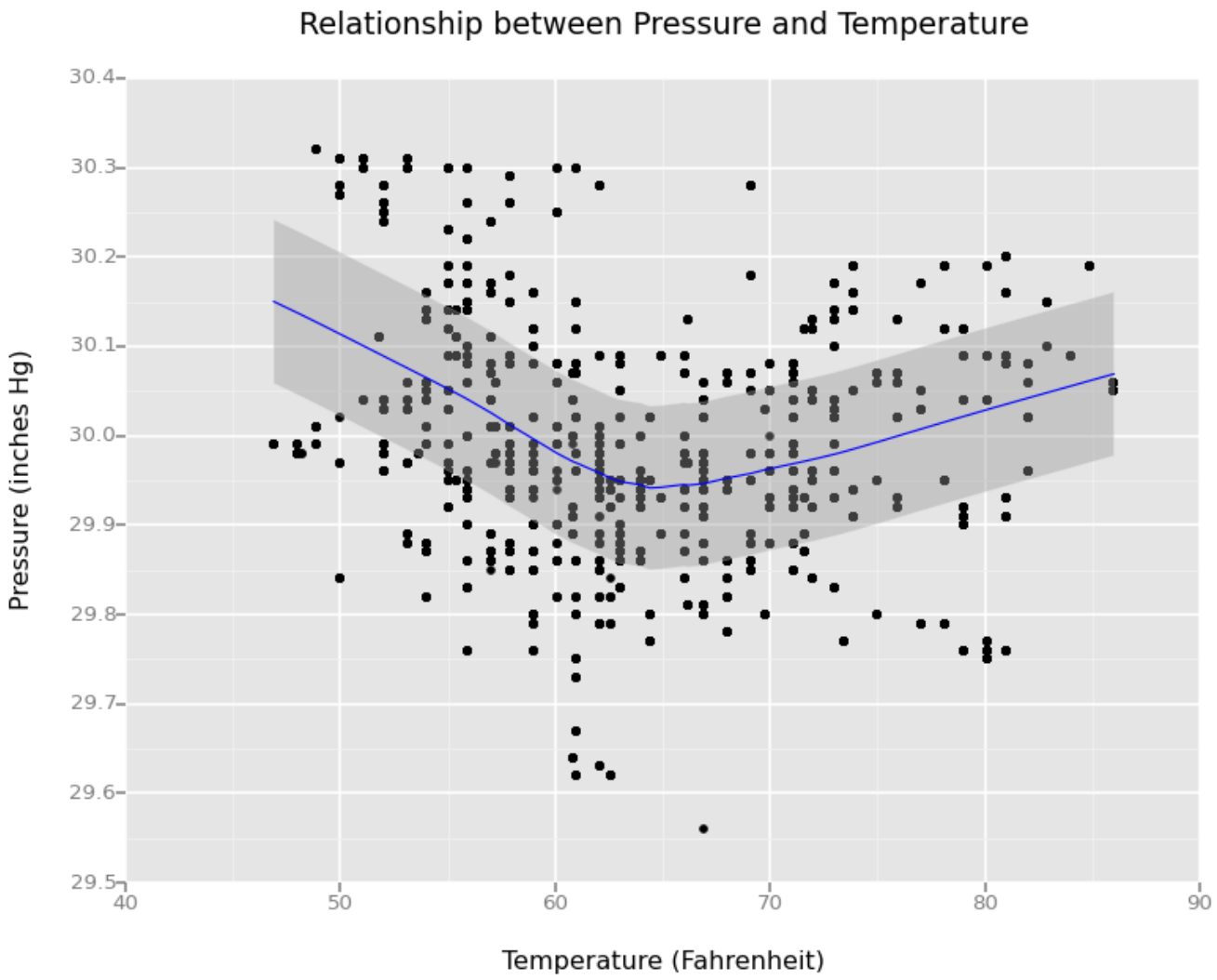


2. One visualization can be more freeform

I decided to investigate a relationship between 2 weather variables by looking at all of the available variables and to find some correlation between some of them. I tried to look at variables such as: temperature, pressure, wind speed, rain etc... By investigating, I found that there was an interesting relation between the temperature and the pressure when it is not raining. This correlation is not as strong as when it is raining so I decided to leave out the rainy days.

The scatter plot seem to identify a trend, however I needed to find a function that would track it as the scatter plot wasn't clear enough by its own. I therefore decided to add a smoother through a generalized linear model.

Note that I made sure to remove all duplicates and that I took only data from each weather station as a single weather station can cover multiple subway station.



Section 4. Conclusion

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

1. From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?
2. What analyses lead you to this conclusion?

The Mann-Whitney U-Test only indicates if both samples are coming from the same population, hence it is not possible to make a conclusion out of this test. In order to make an assumption, I measured the median and mean of both samples in order to identify which one is higher/lower. The mean for non-rainy days is 2028 while its median is 939, for rainy days the mean is 1845 and the median is 893.

This indicates that there's more ridership when it is not raining rather than when it is raining as both median/mean are higher for non-rainy days.

Section 5. Reflection

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

1. Please discuss potential shortcomings of the data set and the methods of your analysis.
2. (Optional) Do you have any other insight about the dataset that you would like to share with us?

First of all, the gathered data only represents 1 month of ridership in the subway and it is quite a small sample compared to the whole population of ridership in the subway which could be measured on several years (or even decades). One of the most defining criterion to influence the ridership, should it be a rainy or non-rainy days, is human/social behaviour, and it is quite difficult to ascertain.

By looking into the free form graph question, I tried to dive deeper into the data and figured out that not all subway stations were included. On top of that, there are a lot of events that can influence heavily the ridership of subway such as: Sports, Concerts, Bank Holiday (the dataset includes one bank holiday) etc..

I could figure out that there's more ridership on a bank holiday's eve compared to a normal day or following a sports event (MLB Games at the Yankee Stadium). Unfortunately the Yankee Stadium subway station is not included in the dataset, therefore I did some python calculation in order to take into account the surrounding subway stations based on their GPS coordinates and managed to find small variances, it would have been easier with the real Yankee Stadium subway station.