Short Questions to Analyzing the NYC Subway Dataset

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?
   Ans: I used Mann Whitney U test to analyze the data.

2. Why is this statistical test applicable to the dataset?
   Ans: Based on the Histograms we created from the data available in Exercise 3.1 we saw that the data is non-normal; so we should be using a non-parametric test with this data. Hence I used MWU test.

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.
   Ans: Two tailed p value was approximately .049999
   Mean for data with rain = 1105.4463767458733
   Mean for data without rain = 1090.278780151855

4. What is the significance and interpretation of these results?
   Ans: As we can see from above that two tailed p <.05 and U < Ucritical which denotes that this cannot happen by chance and there is some statistical difference between the two data. Hence we can conclude that people are more likely to take subway when it's raining.

Resources: I read the following to understand MWU test in detail
http://www.sussex.ac.uk/Users/grahamh/RM1web/MannWhitneyHandout%202011.pdf

Section 2. Linear Regression

1. What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:
   Ans: Gradient descent (as implemented in exercise 3.5)
2. What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
   Ans: I used 'rain', 'Hour', ‘meantempi','maxpressurei','precipi','meandewpti'.
   Yes, We used UNIT as the dummy variable.

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.”
   Ans: Based on the findings of Statistics MWU test. I observed statistical difference between entries per hour with and without rain, so rain becomes an obvious choice. Moreover the R^2 without rain is .454 and with rain it was .46 hence rain . Remaining input variables are purely on exploratory analysis, so as to improve my R^2.

4. What is your model’s R2 (coefficients of determination) value?
   Ans: My R2 is 0.474163985591

5. 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?
   Ans: One way to determine the goodness of our model is by looking at R^2, closer it is to one, better our model is. As we can see its value is 0.47 which is closer to 0 than 1. Hence we can say that the model we had is not a good fit.

Section 3. Visualization

1. 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., scatter plots, 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.

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 different plots.

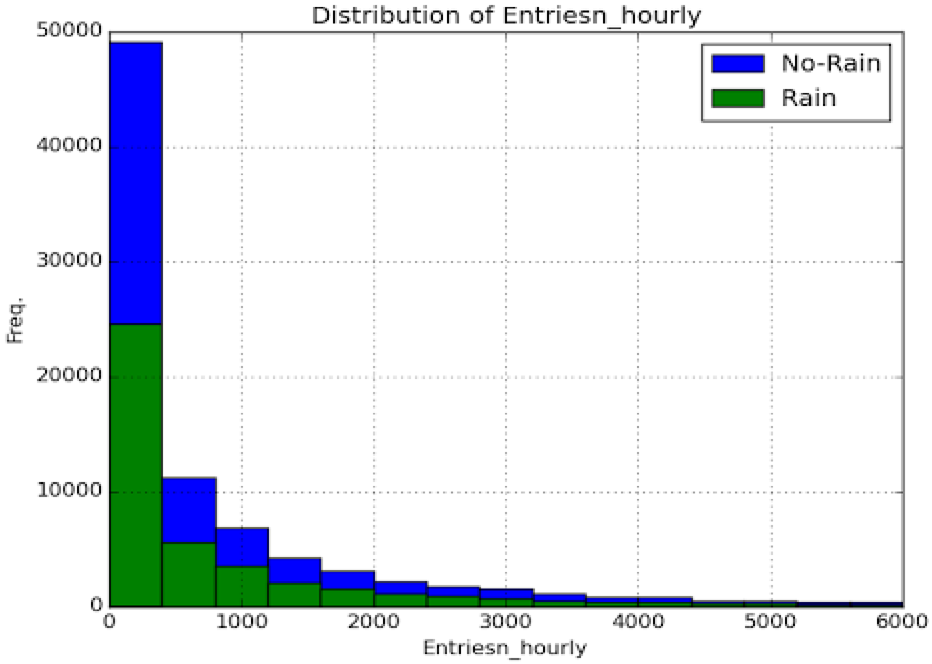
For the histogram, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, you might have one interval (along the x-axis) with values from 0 to 1000. The height of the bar for this interval will then represent the number of records (rows in our data) that have ENTRIESn_hourly that fall into 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.

One visualization can be more freeform, some suggestions are:

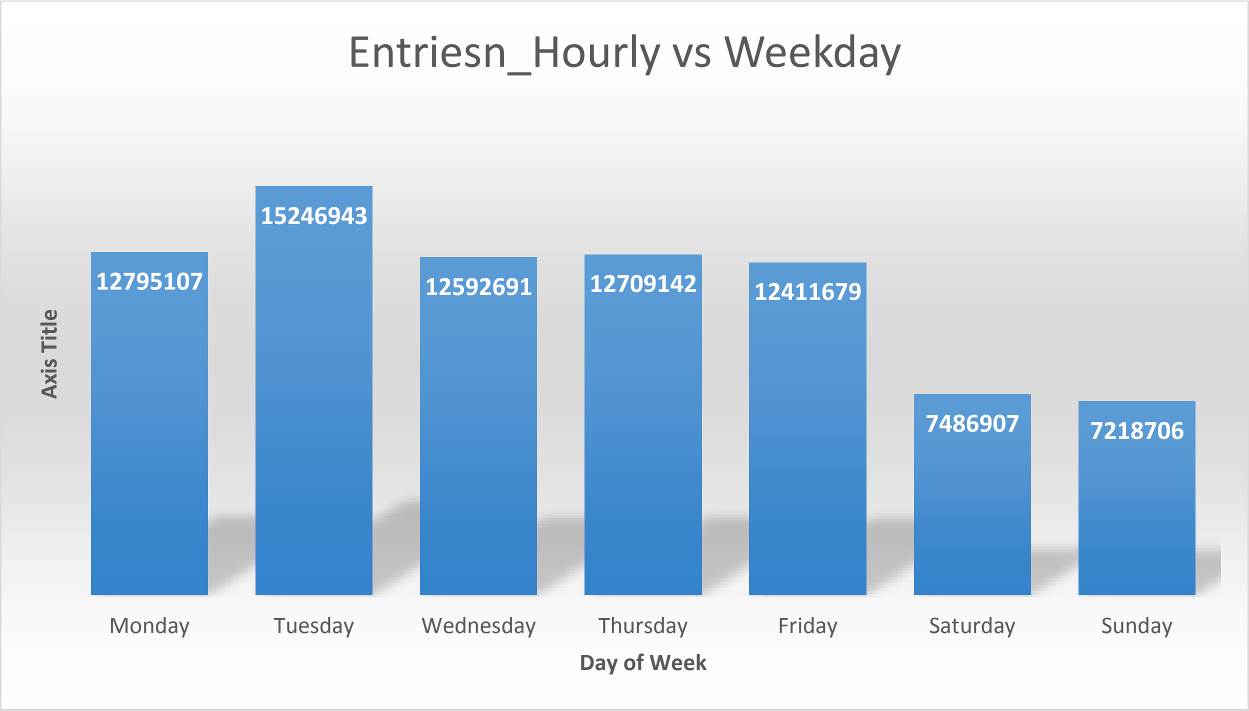
- Ridership by time-of-day or day-of-week
- How ridership varies by subway station
- Which stations have more exits or entries at different times of day

Ans: Visualization for Histograms on Rainy and Non Rainy days:



Visualization 2: Ridership(ENTRIESn_hourly) vs day_of_week

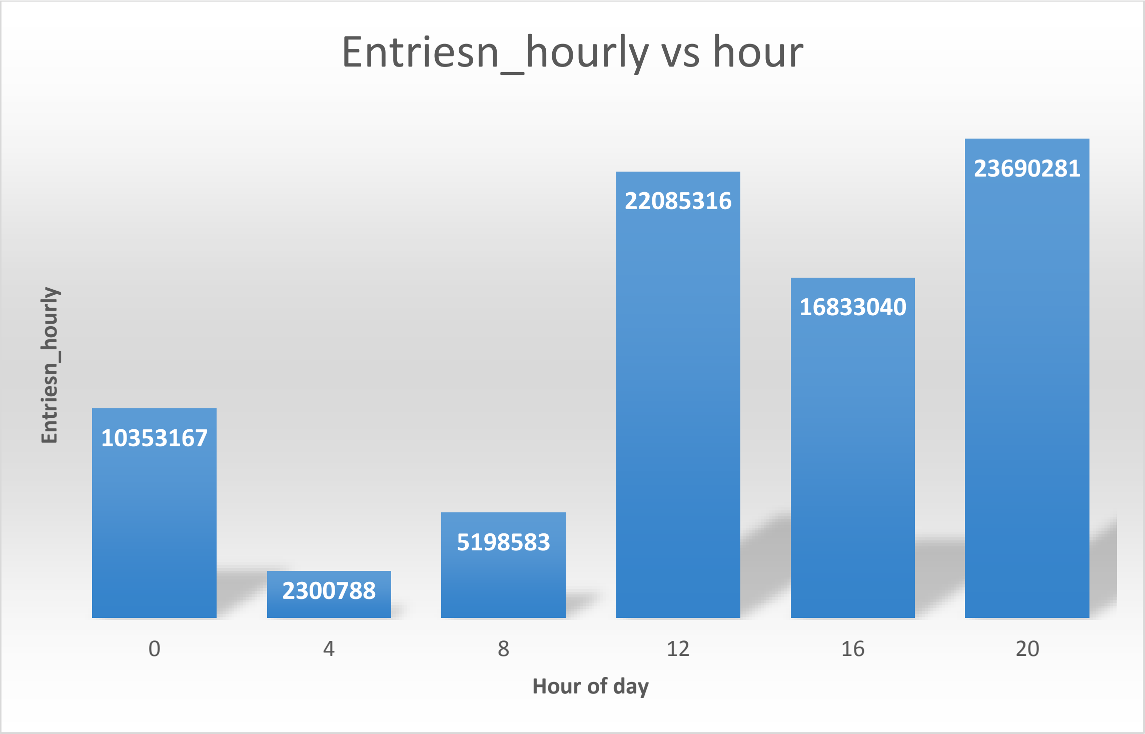
In below visualization the value Entriesn_hourly is summed based on the day it belongs.



Visualization 3: Ridership(ENTRIESn_hourly) vs hour

In below visualization the value Entriesn_hourly is summed based on the hour.

NOTE: the data provided in the dataset is in 4 hours buckets. Hence the below graph is also using 4 hours intervals.



Section 4. Conclusion

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

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?

What analyses lead you to this conclusion?

Ans: Based on the statistical analysis done in this project we saw that mean of Entriesn_hourly was higher on rainy days than non-rainy days. Upon doing Mann Whitney test we also saw that two tailed p<.05 which lead us to believe that this cannot happen by chance and there is some statistical significance to this data, hence more people ride subway on rainy days in comparison to non-rainy days.

Section 5. Reflection

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

Please discuss potential shortcomings of the data set and the methods of your analysis.

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

Ans: I have been working with the original data set and realized that there were various shortcomings in the data set, for e.g. the DATEn column had missing values, additionally it was hard to figure out what multiple ENTRIESn_hourly meant for same UNIT and HOUR.

The Linear Regression model we created using Gradient Descent did satisfy our class assignment condition(R^2 >.20) but it was no way close to being optimal (close to 1). Which lead us to believe that our model isn't correct one.

I was also surprised to see the effect of meantempi on the R^2; without it the R^2 dropped to .40