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?

Answer: To analyze the NYC subway data we used the Mann-Whitney U-Test. For implementation purposes we used scipy.stats.mannwhitneyu.

The reported p-value (0.024999912793489721) is for a one-sided hypothesis. To get the two-sided p-value the returned p-value shall be multiplied by 2. The p-critical value is .05

My null hypothesis is that there is no difference in ridership between rainy and non-rainy days.

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

Answer: This statistical test is applicable for the dataset because it does not assume our data is drawn from any particular underlying probability distribution. The Mann-Whitney U-test tests null hypothesis that 2 populations are the same. It tests whether or not 2 samples (in our case the distribution of the number of riders on rainy days and non-rainy days) come from the same population.

It is useful to report the results of this test along with other information such as the 2 sample means.

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.

Answer:

The mean of entries with rain is 1105.4463767458733

The mean of entries without rain is 1090.278780151855

The Mann-Whitney U-statistic value is 1924409167.0

The p-value is 0.024999912793489721

The 2-tailed p value is obtained by multiplying the critical p-value by 2 = 0.05

4. What is the significance and interpretation of these results?

Answer: The mean of number of entries between the rainy and non-rainy day is different, the mean of entries with rain being higher than the mean of non-rainy days.

The p-critical value is 5%, and if the ridership on rainy days and non-rainy days follow the same distribution, the differences we will see are as big as the ones we see p=5% of the time.

Given these results:

- **we reject the null hypothesis, that there is no difference in ridership between rainy and non-rainy days**
- **we accept the alternative hypothesis as true. The distribution of the number of entries between rainy and non-rainy days is statistically different.**

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:
    a) Gradient descent (as implemented in exercise 3.5)
    b) OLS using Statsmodels
    c) Or something different?

Answer: to predict the ENTRIESn_hourly we used the Gradient Descent.

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

Answer: to predict the ENTRIESn_hourly I took a data exploration and experimentation approach, aiming to improve the R2 value as much as possible. I sequentially ran the code using various subset of features and observed how R2 has improved with every run as follows:

A. **features = dataframe[['rain', 'precipi', 'Hour', 'meantempi']]**
   no dummy variable
   R2 value is 0.0962311786182
B. **features = dataframe[['rain', 'precipi', 'Hour', 'meantempi']]**
   Add UNIT to features using dummy variables
   R2 value is 0.461129068126
C. **features = dataframe[['rain', 'precipi', 'Hour', 'meantempi','meanwindspdi']]**
   Add UNIT to features using dummy variables
   R2 value is 0.470049747281
D. **features = dataframe[['rain', 'precipi', 'Hour', 'meantempi','meanwindspdi', 'maxtempi']]**
   Add UNIT to features using dummy variables
   R2 value is 0.472682297649
E. **features = dataframe[['rain', 'precipi', 'Hour', 'meantempi','meanwindspdi', 'maxtempi', 'mintempi']]**
   Add UNIT to features using dummy variables
   R2 value is 0.473552297101
F. **features = dataframe[['rain','Hour','meanwindspdi','maxtempi','mintempi','maxpressurei', 'minpressurei','precipi','mindewpti','maxdewpti']]**
   Add UNIT to features using dummy variables
   R2 value is 0.474345151066
G. **features = dataframe[['rain','Hour','meanwindspdi','maxtempi','mintempi','maxpressurei', 'minpressurei','precipi','mindewpti','maxdewpti']]**
   Add UNIT and DATEn to features using dummy variables
   R2 value is 0.474349836743

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."

Answer: to predict the ENTRIESn_hourly I took a data exploration and experimentation approach, aiming to improve the R2 value as much as possible

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

Answer: We used rain with positive theta of 104.5

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

Answer: See values listed in my answer above

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?

Answer: R2 is a statistical measure of how close the data are to the fitted regression line. The R2 value is between 0 and 1; the higher the R2, the better the model fits the data.

Based on the R2 values obtained, this is not a very good regression model.

I do not think that a linear model is the best way to predict ridership. Ridership volume for each subway unit is highly asymmetric. We have a lot of stations with low ridership and a few stations with high ridership. A different model might lead to better results.

Section 3. Visualization

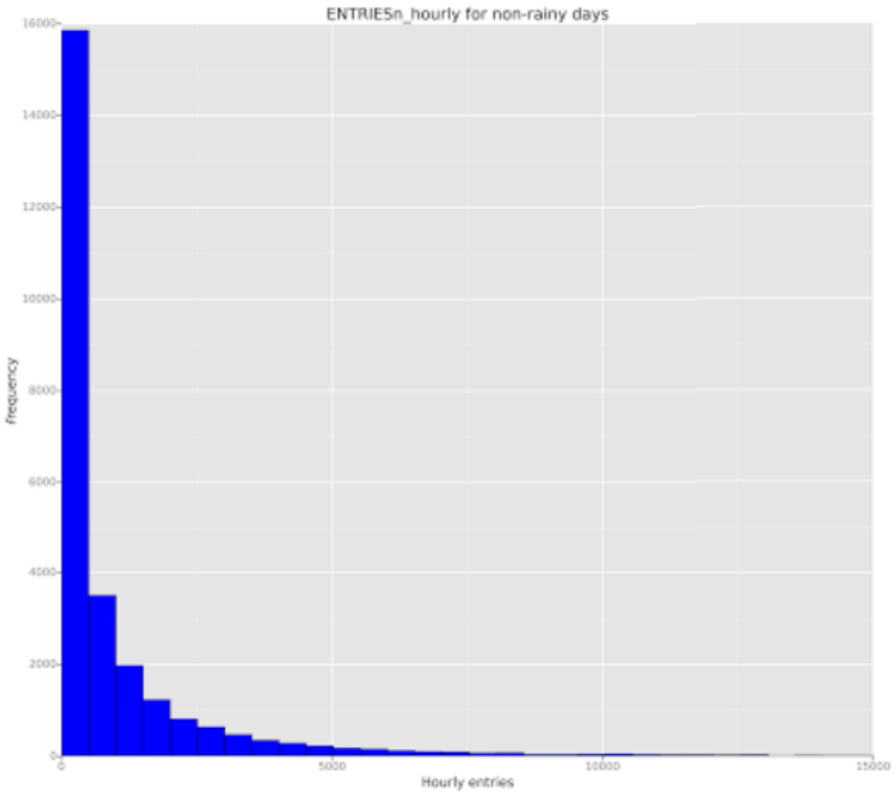
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

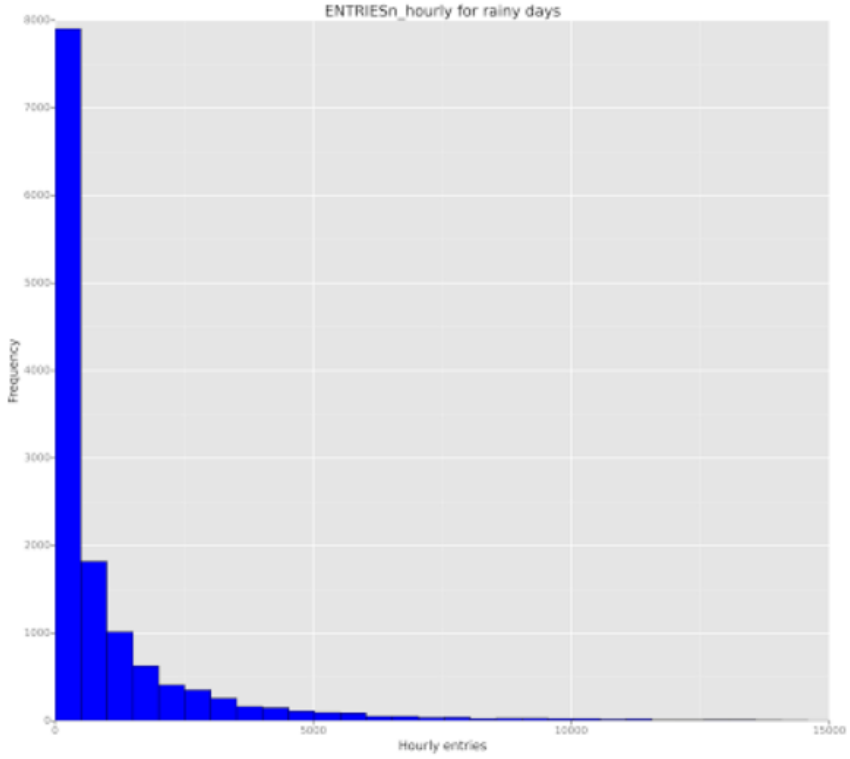
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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 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.
   Answer:

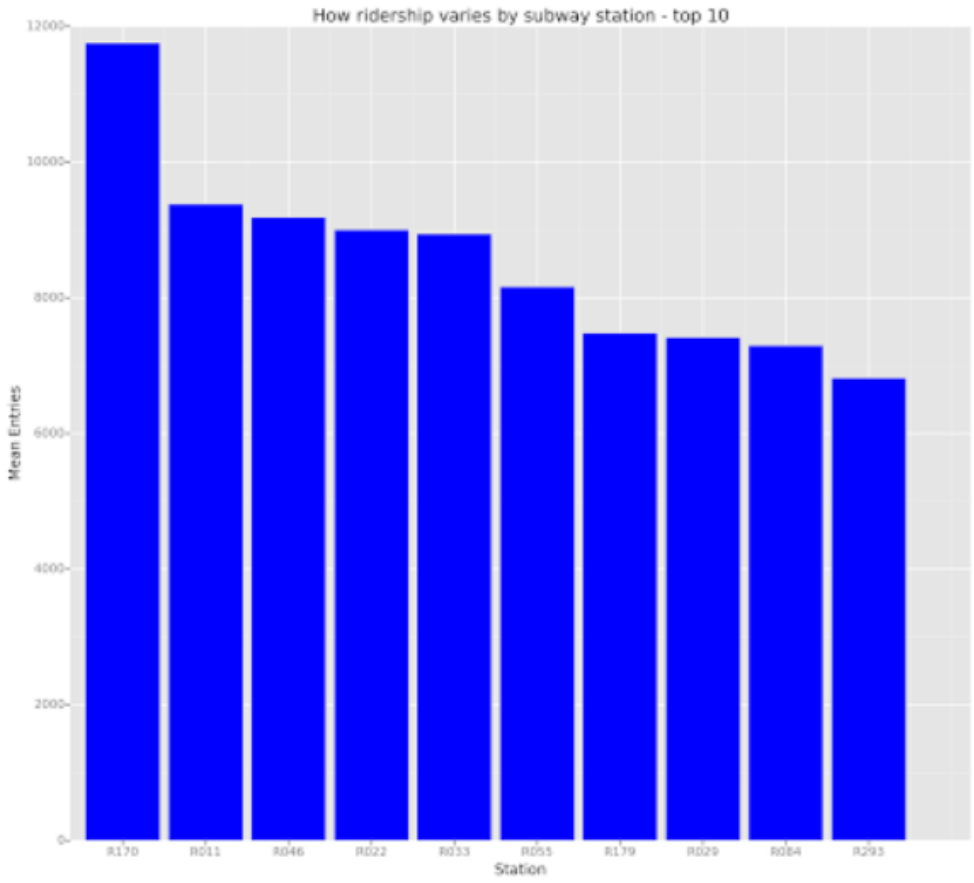

ENTRIESn_hourly for non-rainy days
Frequency
Hourly entries
0
2000
4000
6000
8000
10000
12000
14000
16000
0
5000
10000
15000


ENTRIESn_hourly for rainy days
Frequency
Hourly entries
0
1000
2000
3000
4000
5000
6000
7000
8000
0
5000
10000
15000

The above histograms represent the distribution of the hourly entries for non-rainy and rainy days respectively. The intervals on the x axis represent the volume of ridership (value of ENTRIESn_hourly); The values on the y axis represent the frequency of occurrence.

2. One visualization can be more freeform, some suggestions are:
    a) Ridership by time-of-day or day-of-week
    b) How ridership varies by subway station
    c) Which stations have more exits or entries at different times of day

Answer:



The above bar chart represents the top 10 most "busy" stations in terms of ridership (ENTRIESn_hourly). The x axis represents the station code; the y axis represents the mean number of ENTRIESn_hourly.

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?

Answer: Based on my analysis and interpretation of the data, the average number of people riding the NYC subway on rainy days is greater than those days when it is not raining.

2. What analyses lead you to this conclusion?

Answer: The Mann-Whitney test results and the mean of entries with rain being higher than the mean of non-rainy days suggest that more people ride the subway in 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.
   Answer:
   Shortcomings of the data:
   - **Not all turnstiles are reporting at the same hours**
   - **The quantity of rain per location is relevant to a certain day, but it is not relevant to specific time of the day when data is reported. We are not able to compare ridership between rainy and non-rainy hours of the days; we are treating the rainy days in the same manner, no matter when the rain occurred during the day.**
   - **The data is only for a month; we need data for more months to be able to make a better prediction**
   - **A linear regression model is not the best way to predict ridership. Ridership volume for each subway unit is highly asymmetric. We have a lot of stations with low ridership and a few stations with high ridership. A different model might lead to better results**