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

1.1 Which statistical test did you use to analyse the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

Answer:
We compared the mean of entries with rain to those without rain asking if the mean of the hourly entries at this two weather conditions are the same or different.
We used the Mann-Whitney U-statistic and p-value comparing the number of entries with rain and the number of entries without rain. We used the two tailed P value. The null hypothesis was that the number of entries is the same for both weather conditions which we rejected given the critical P value of 0.05, since we obtained a P value of 0.025

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.

Answer:
Since the number of Entries is not normally distributed we could not use the Welch t-test.
We used the non-parametric Mann-Whitney which does not makes assumptions about the underlying distribution of the data we compared.

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.

Answer:
The mean of the number of entries with rain = **1105.45**
The mean of the number of entries without rain = **1090.28**

The value of the Mann-Whitney U statistic = **1924409167.0**
and the P Value of the 2 sided test P= **0.025**

1.4 What is the significance and interpretation of these results?
The average number of entries on rainy days is significantly (P=0.025) different,at a significance threshold of P=0.05, than on days without rain.

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?

Answer: I used ordinary least squares (OLS) to compute the coefficients theta of the linear model and I used numpy's dot product to compute the predictions.

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

Answer: As input variables I used 'rain', 'precipi', 'Hour', 'meantempi','fog','meanwindspdi' added a column of 1's to model the intercept and used the the 'UNIT' as indicator (dummy) variables.

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."

Answer:

I choose the variables by first inspecting of the scatterplots of the variables versus the response variable number of entries.
I used on offset since I assumed that there is a base level of the use of the Subway. Than I used the Hour since definitely the number of entries will depend on the hour of the day. Furthermore I included weather conditions into the model (rain, temperature, fog, windspeed)

since I assumed that this will increase or maybe decrease the use of the Subway. Bad weather - people might prefer to use car to get to work or to use subway instead of walking.

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

rain -7.610994
precipi -34.916972
Hour 430.074794
meantempi -67.977471
fog 97.111724
meanwindspdi 67.270255
ones 1065.242100

Answer:

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

Your R^2 value is: 0.485

Answer:

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?

The R square of means that almost 50% variability could be explained by the model. If this is good enough could be examined using diagnostic plots, i.e. plotting the residues of the predicted values and the response variable versus the fitted value.

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.

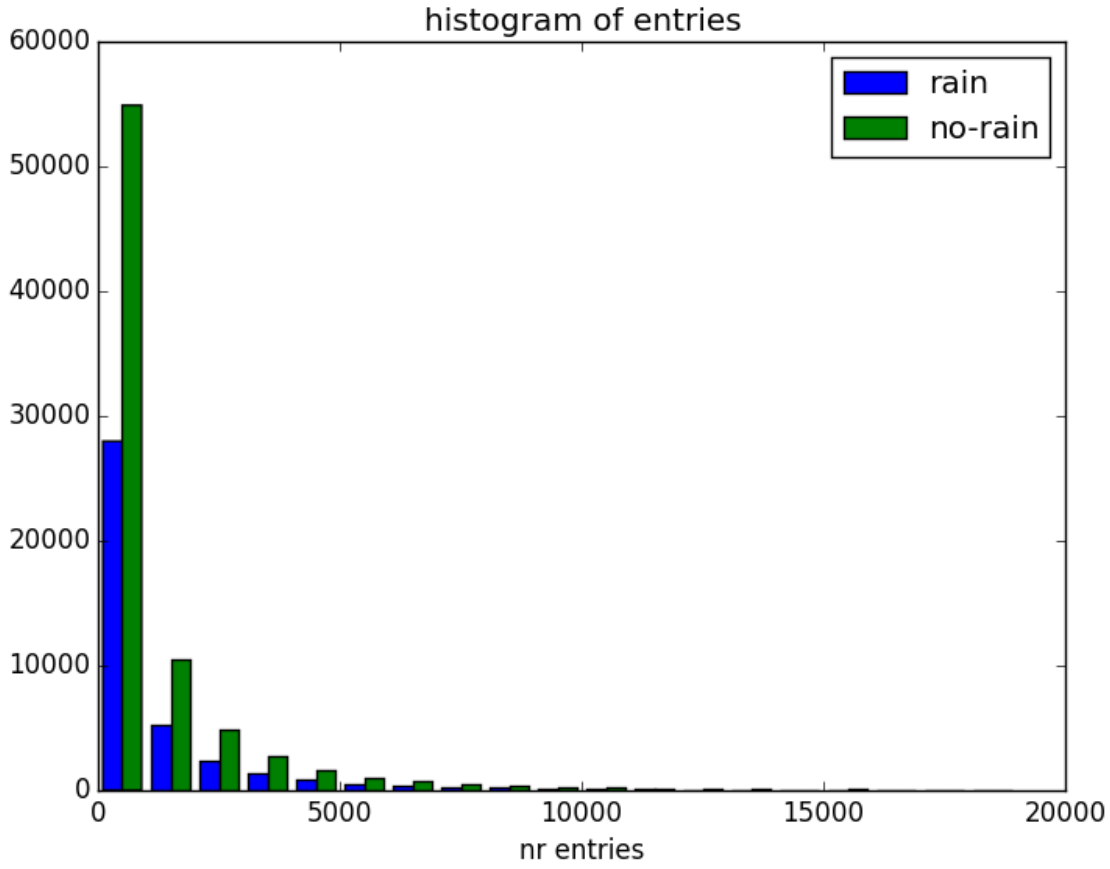
Remember to add appropriate titles and axes labels to your
plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.

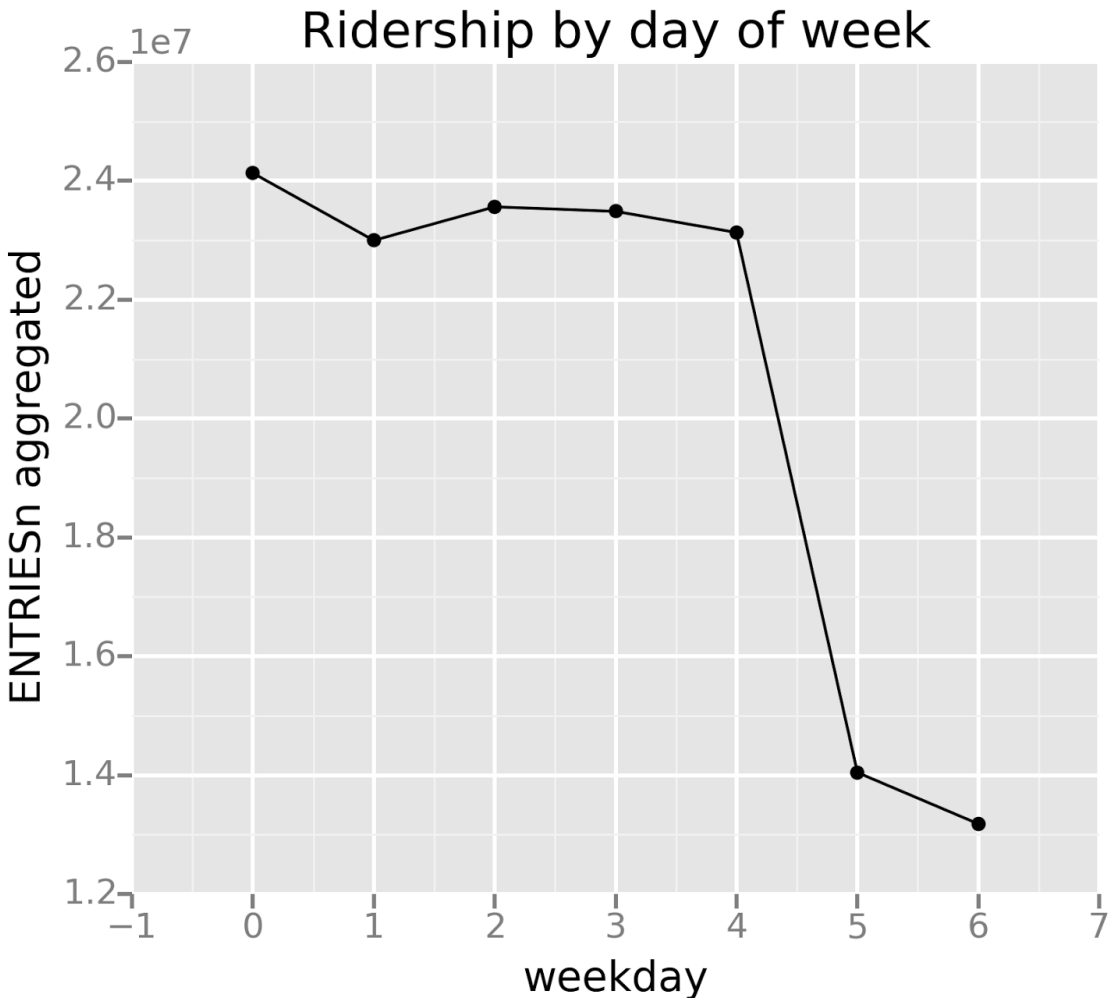
Answer:



Legend: histogram of entries into the subway station at rainy (blue) and days without rain (green).

3.2 One visualization can be more freeform. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



Legend: Ridership by day of week (0 - Monday … 6- Sunday)

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?

Answer:

Looking at the absolute numbers, there are much less days when it is raining. Therefore, the total use of the subway when it is raining is also lower than on not rainy days.

If we however you look at the average of the hourly entries to the subway at a rainy day it is higher than the average hourly entry to the subway at non rainy days (which was asked for in

the problem set). It is also interesting to see that at rainy days there are some hours with very high number of entries.
However, to fully answer this question actually the average total entries per rainy day should be computed and compared with the average usage of the subway at non rainy day.

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.

Answer:

First the the Mann-Whitney test used in Problem Set 3 which indicates that the average number of entries per hour is higher than at non-rainy days. Secondly the linear model fitted indicates that fog and wind speed influences the hourly number of entries into the subway.

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.

Answer: The dataset used is limited to data of the month May. It might therefore be impossible to make predictions about the subway usage in other months such as August or January and how the usage of the subway is influenced by weather conditions.
My analysis is based on problem sets to be solved in the lecture. Normally I would start the data analysis by foremost performing an exploratory analysis. Linear regression and statistical tests as the Mann-Whitney test used are classical statistical tools, and although both have limitations they big strength is that rather large audience can comprehend the performed analysis and assess their viability.

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