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, and 4 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.

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?
A Mann Whitney U-test was used to analyze the NYC subway data. I used a two-tail P-value. My null hypothesis was the following: "The distribution of the hourly ridership in rainy and non rainy day populations are equal." My p-critical value was $\alpha = 0.05$.

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.
The Mann Whitney U-test is applicable to this dataset because it assumes the sample distribution of NYC subway entries is non-normal, which is true for this 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.
With_rain_mean =1105,
Without_rain_mean = 1090,
Two tailed p-value = 0.04999982558698
(scipy returns one tailed p-value = 0.024999912793489721)

1.4 What is the significance and interpretation of these results?
Mean entries on the NYC subway on days with and without rain were 1105 and 1090 per hour; the distributions in the two groups differed significantly (Mann–Whitney U = 1,924,409,167, n_rain = 39,895, n_no-rain = 78,715, p-value < 0.05 two-tailed).

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. **OLS using Statsmodels** or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

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

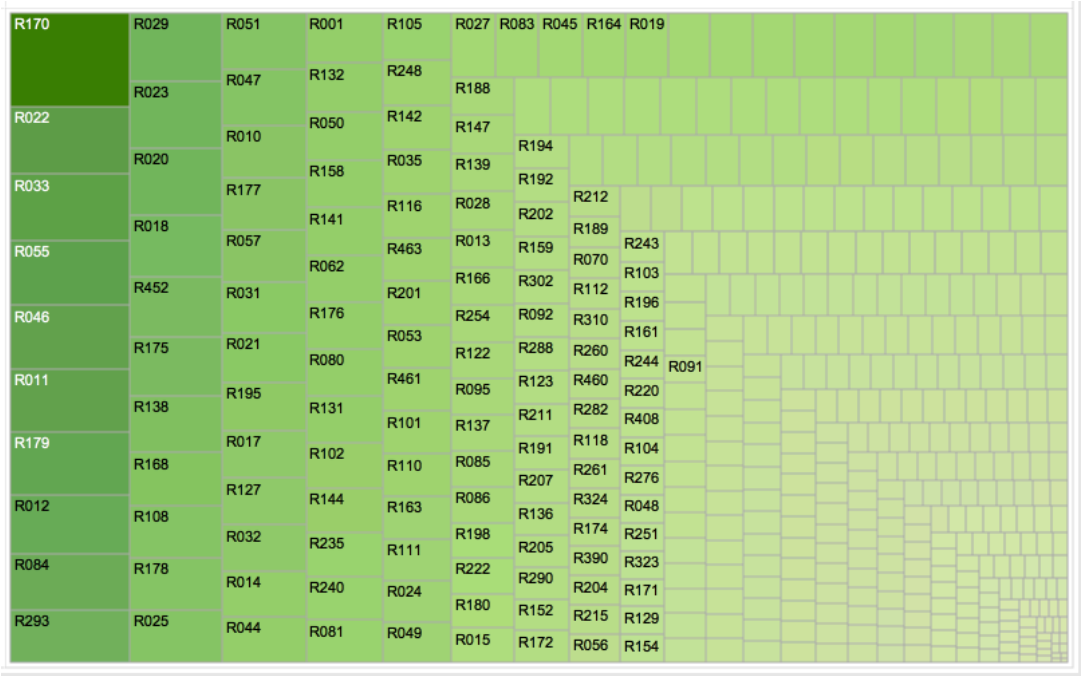
I used rain, weekend and holiday (I identified weekends and Memorial Day)
I used UNIT and Hour as dummy variables

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.

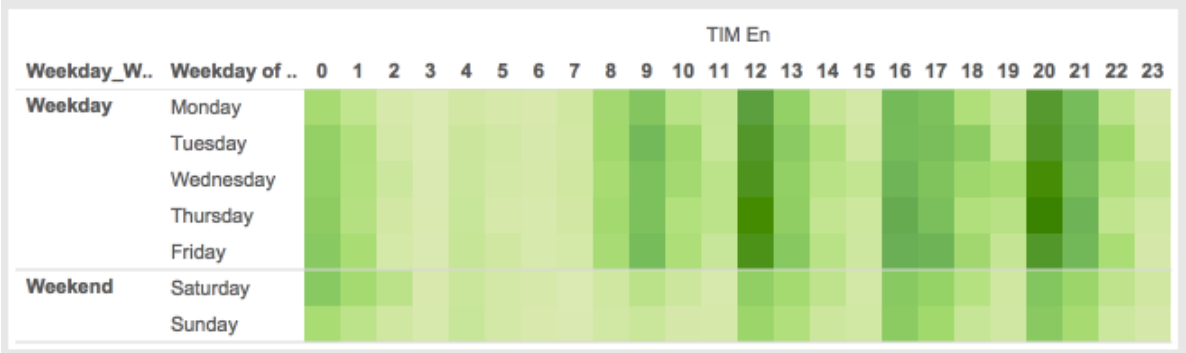
Rain was used because I thought it would weigh heavily in a person's decision to take the subway on a given day. I left out other measures of weather as I thought they could be correlated with rain.

Weekend was used because I noticed a substantial difference when plotting a chart of the avg. entries per hour by days of the week. I assumed trains run less frequently on the weekend and people choose to use alternative forms of transportation when they do not have to go to work.

Units were used as a dummy variable as it became clear when plotting a tree map that some stations experience a lot more traffic than others.



I decided to use Hour as a categorical variable after plotting average entries on a heat graph by day of the week and time. I noticed that there were pockets of time where travel peaked.



- 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.”

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

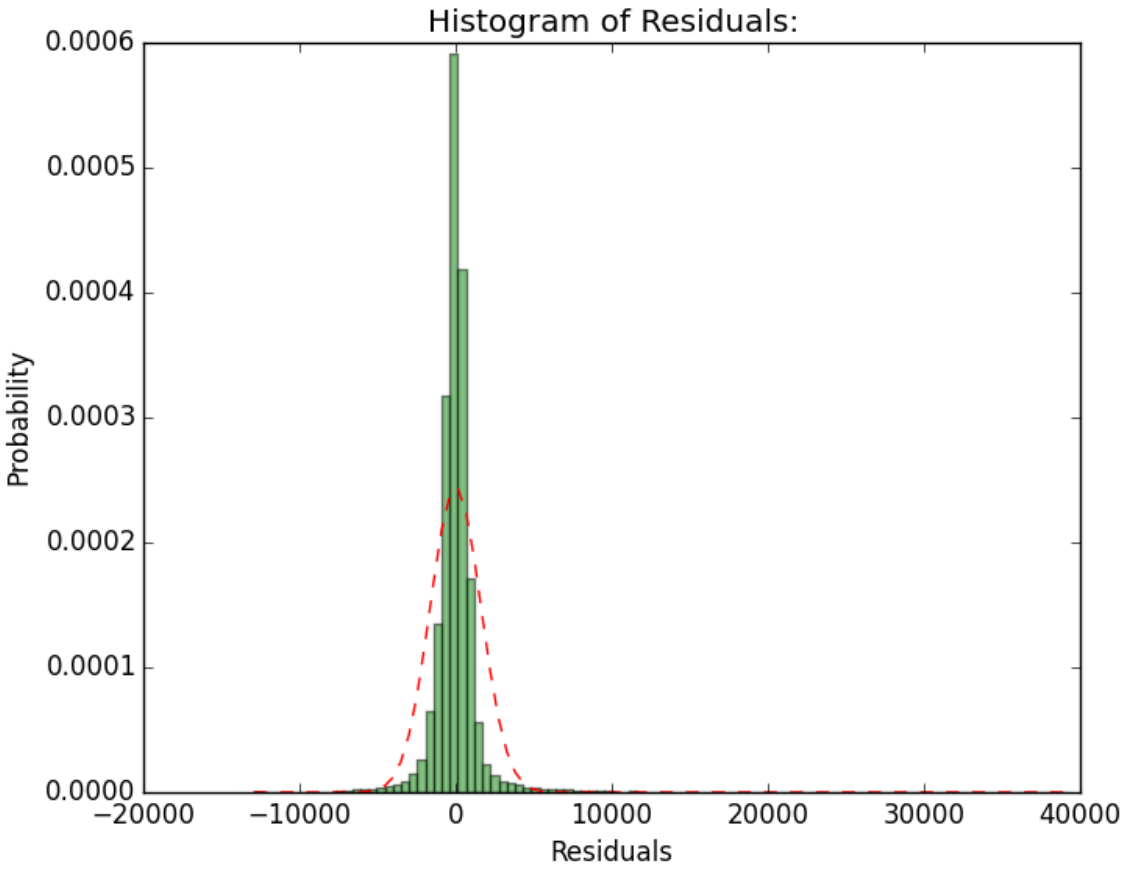
rain	**86.3911**
weekend	**-555.1255**
holiday	**-649.7585**

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

r^2 value is 0.529449747356

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^2 indicates that the model explains ~53% of the variability of the response data around its mean. Taking a closer look at the residuals we notice that the distribution is NOT normal, but has a higher peak and longer tails with a kurtosis = 48.68 (Fisher, normal = 0). The higher the kurtosis coefficient is above the "normal level", the more likely that predictions will be either extremely large or extremely small. This result serves as a good reason for further exploration of the data set in order to improve the regression model.

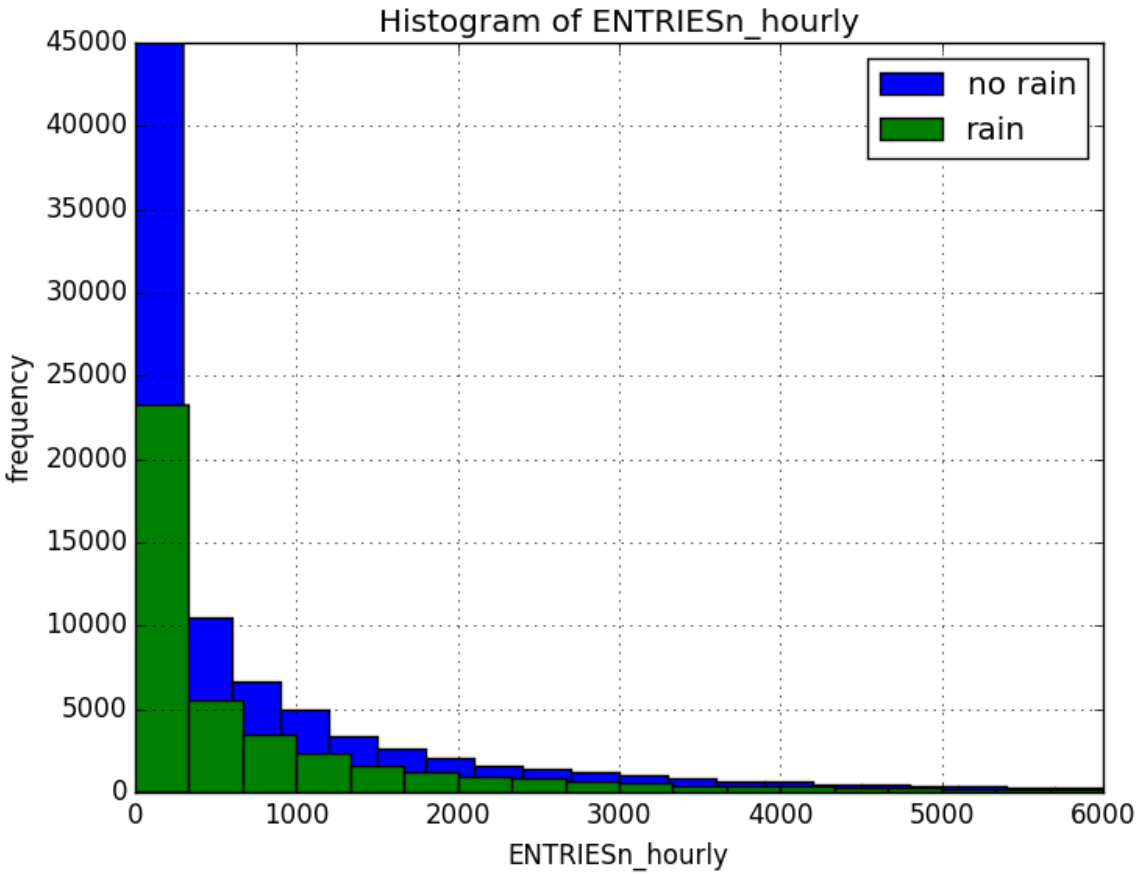


Section 3. Visualization

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.

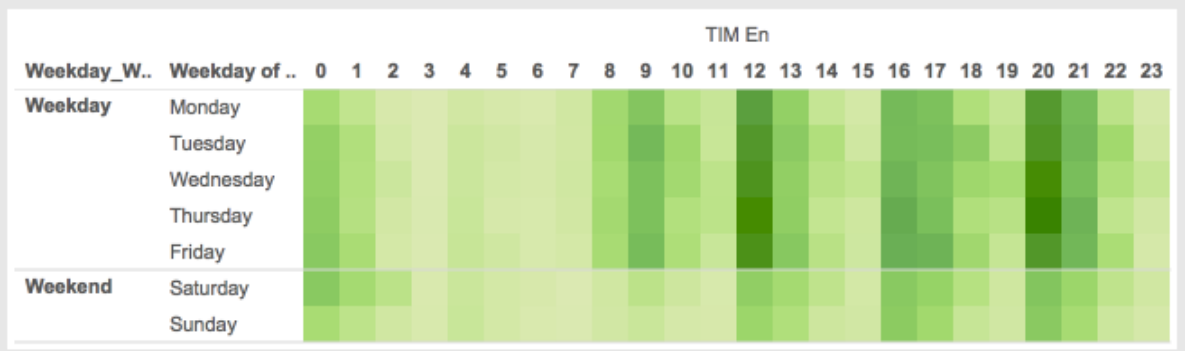


The histograms above illustrate that hourly entries are lower when it rains than when it does not rain.

- 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.

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



In the heatmap above it can be observed that the weekends have lower entries than the weekdays. In addition, there are periods of heavy ridership around 12p, 4p, and 8pm.

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?
My analysis and interpretation of the data lead me to conclude that more people ride the subway 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.
A number things led me to this conclusion. The sample mean Entries per hour is 1105 when it rains vs 1090 when it is not raining. The histogram comparing raining and not raining shows a clear difference in the distribution of ridership between the two. The Mann Whitney U-test showed that the distributions in the two groups differed significantly (Mann–Whitney U = 1,924,409,167, n_rain = 39,895, n_no-rain = 78,715, p-value < 0.05 two-tailed). Lastly, the sample mean Entries per hour is 1105 for the coefficient for rain in the OLS model showed that ridership increased by ~25 riders per hour when it rains.

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,

   The dataset only reflects a one-month sample. We may get more insight into ridership on rainy vs. non rainy days if we have a larger sample covering a longer time frame. If we look at a longer time period it would also be helpful to capture economic data that could influence ridership such as employment, gas prices, subway fares, etc.

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

   Some of the shortcomings of using OLS are as follows:

   (1) Outliers – I didn't scrub the data for outliers. This could adversely impact the results.
   (2) Non-linearity – I didn't have a strong understanding of how a transportation system works. Perhaps there are some transformations that I can apply to the data with better knowledge of the data I am attempting to model.
   (3) Dependency – I did not test for correlation among the independent variables. If it exists, there may be alternative regressions that adjust for this.

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

I thought the data on rain could be more granular. For example, it may be raining at one station and not at another. Or it may be raining at a specific time at a given station. It appeared the dataset assumed that if it rained in a given day it rained all day at that location.