Analyzing the NYC Subway Dataset

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 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.

---

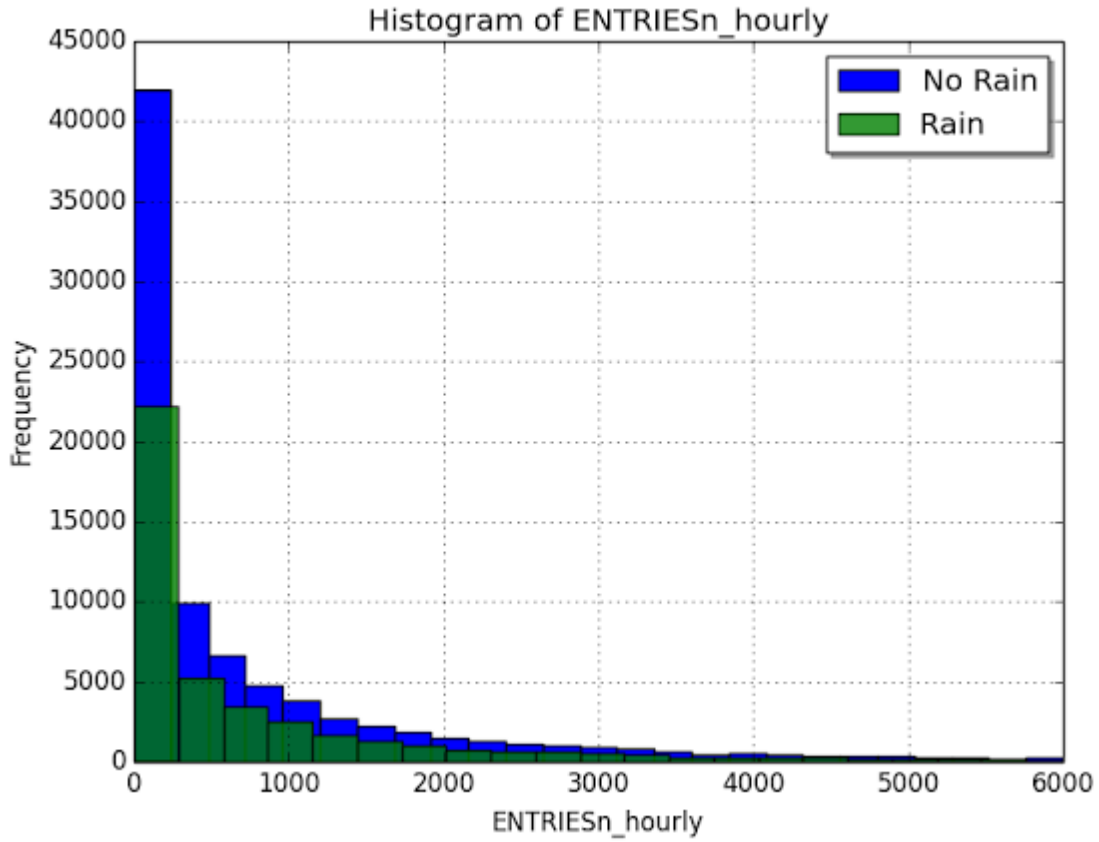
https://docs.python.org/2/
http://stackoverflow.com/
http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
http://discussions.udacity.com/t/unable-to-print-correctly-the-individual-coefficients-of-features-problem-set-3-5/13656/3

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

1.1 I used the Mann-Whitney U-Test to determine if there could be a difference between the distribution of turnstile entries with and without rain. I used a two-tail P value. The null hypothesis was that no distribution was more likely than the other to generate a higher value if values are randomly drawn from it. The p-critical value is 0.05.

1.2 The Mann-Whitney U-Test does not require that the two population distributions be normally-distributed. After creating a histogram of the frequency of hourly entries, it was obvious that the two distributions were not normal.



1.3 Mean with rain: 1105.4463767458733
Mean without rain: 1090.278780151855
P-value: 0.024999912793489721

1.4 The null hypothesis was rejected because the resulting p-value of the test was less than 0.05. This means that the alternative hypothesis that the distribution of entries without rain tends to produce higher values than the distribution of entries with rain holds true.

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?

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

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
2.5 What is your model's R^2 (coefficients of determination) value?
2.6 What does this R^2 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 R^2 value?

2.1 Gradient descent

2.2 I used rain, precipi, Hour, meantempi, meanwindspdi, which were already features in the data, dummy variables for the turnstile units, and created a new column called WEEKDAYn which calculated which day of the week the data point fell on using the DATEn column. This feature was also turned into a set of dummy variables.

2.3 I decided to include rain and exclude precipi because the addition of precipi didn't much improve my R^2 value. I decided to include the Hour variable because the subway will tend to be used more often during rush hour periods. I decided to use meantempi because I thought that people might use the subway more often when they want to avoid a walk in cold weather. I included meanwindspdi because I thought people might want to avoid walking outside during very windy weather. I included WEEKDAYn because I thought it would be likely that people take the subway less often on weekends compared to weekdays (work rush hour). Dummy variables for the turnstile units were included because some locations might have heavier usage in general than others.

2.4

rain:	-8.33176037e+00
Hour:	4.56794502e+02
meantempi:	-5.07689526e+01
meanwindspdi:	1.30671437e+01

2.5 The model's R^2 is 0.476

2.6 This R^2 means that 47.6% of the variation in the model can be explained by the chosen features. This means that this linear model only explains about half of the variation in subway ridership. Further exploration would have to be undertaken. Perhaps there are other variables that can be added that weren't included as features in the base dataset.

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.

- You can combine the two histograms in a single plot or you can use two separate plots.

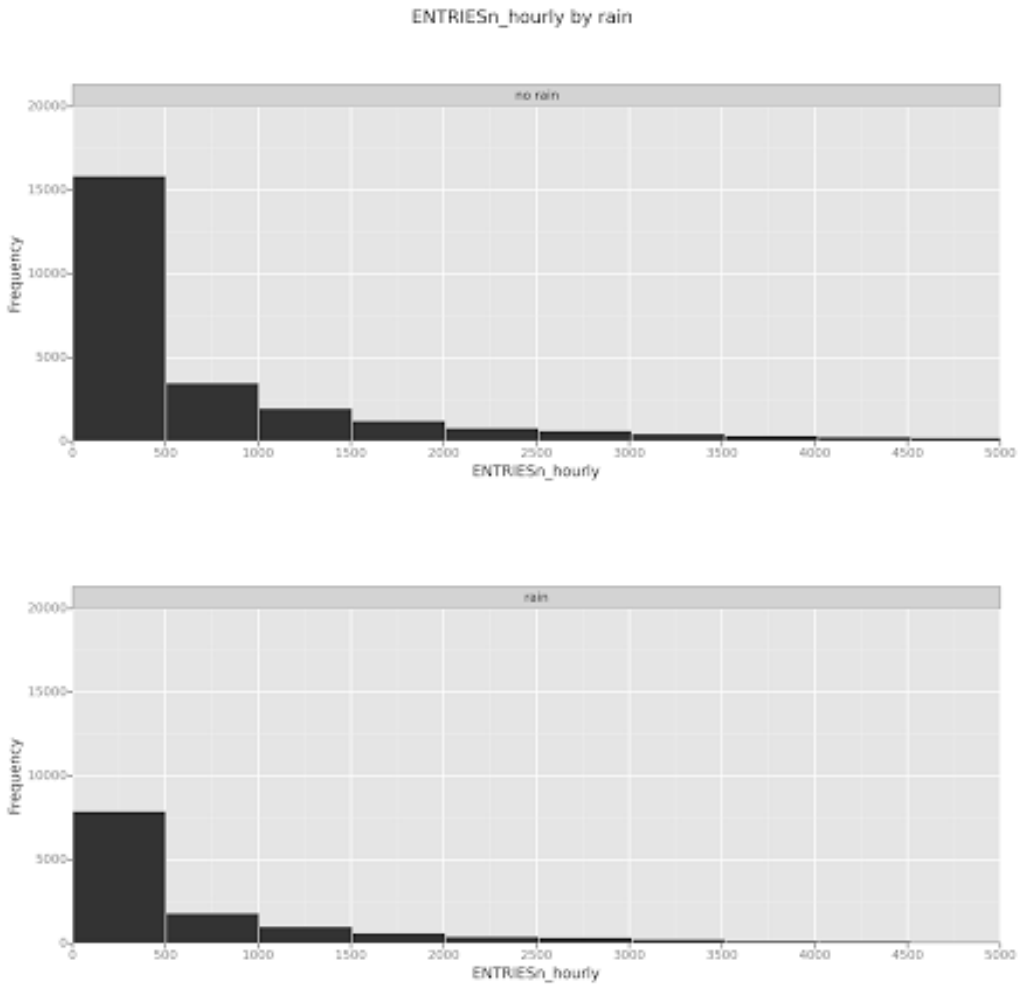
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) 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:

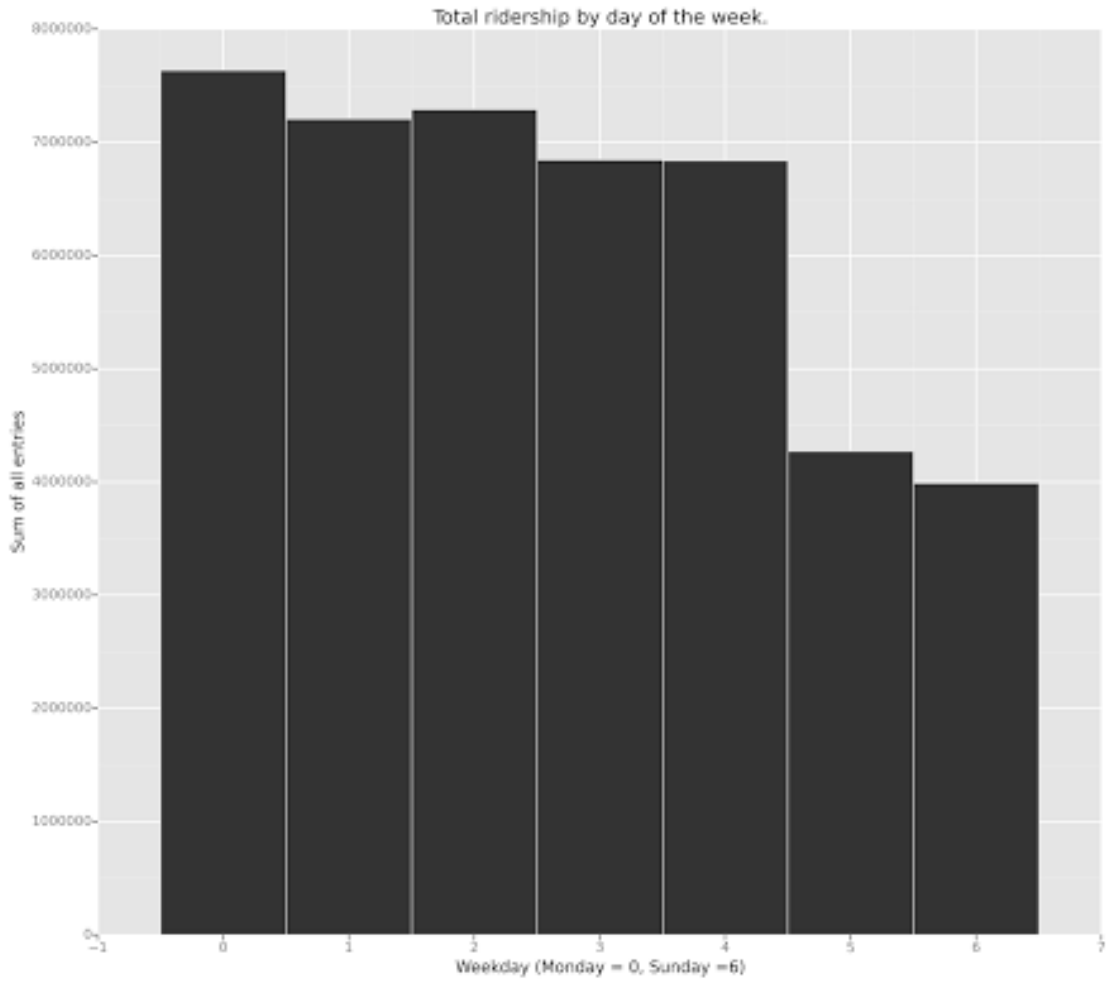
- Ridership by time-of-day
- Ridership by day-of-week

3.1



It appears as if the frequency of ridership with rain is substantially lower than that during periods with no rain.

3.2



This chart shows that total ridership during the study time period is greater on weekdays than on weekends.

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?
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.

4.1 Based on my analysis and interpretation of the data, it appears that more people ride the NYC subway when it's not raining than when it's raining.

4.2 The Mann-Whitney U-Test for differences in the distribution of entries with and without rain concluded that the distribution of entries without rain is more likely to produce larger values if one were to randomly sample from it. This suggests that more ridership occurs during periods without rain.

When a linear regression with gradient descent was run regressing entries on a feature array of rain, hour, meantempi, meandwindspdi, unit, and weekday, the coefficient for rain was negative and large compared to the other coefficients. This suggests that ridership is heavily discounted when it rains compared to when it doesn't rain.

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.

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

5.1 Shortcomings with the methods of analysis used include missing or imperfect variables in the dataset and not testing assumptions that are key to linear regression.

Ideally, the UNIT variable should be grouped by station instead of each individual turnstile unit. Knowing which stations are statistically more likely to be used under varying conditions would be a good transportation planning exercise.

Other variables missing from the dataset include whether or not a significant special event occurred that day (a marathon for example) which could disrupt normal traffic and increase ridership.

Also, linear regression has some key assumptions including linear relationship, normality, little or no multicollinearity, no auto-correlation and homoscedasticity. Ideally, tests should be run for each of these before concluding that the resulting model is a good predictor of ridership. For example, pairwise comparisons should be conducted to see correlation between each variable.