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.

N/A. I used Google search for Python syntax questions and read all the forum topics.

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?

Per instructions on problem set #3, I used the Mann-Whitney U-test to test if the rain and no rain sample sets comes from the same population, i.e. if they are statistically from the same data distribution or not. The null hypothesis is that rain and no rain are from different data set. The resulting p-critical value suggest the null hypothesis is true. P-critical value is 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.

The Mann-Whitney test is used if there is no underlying assumption that the arrival times are of normal distribution. Based on the histogram chart, the sample set is not “normal”, so it was appropriate to use Mann-Whitney.

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.

```
Here's your output:
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

Mean of 1st sample: 1105.4463767458733

Mean of 2nd sample: 1090.278780151855

P value: 0.025

1.4 What is the significance and interpretation of these results?

Since p values is less than .05, the hypothesis that the samples are from different data sets is true. Computing the mean of each sample set confirms that hypothesis that with high probability the samples are different.

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?

I used Gradient Descent as instructed in the problems set.

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 ‘maxtempi’, ‘precipi’, ‘Hour’, and ‘meantempi’ as 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 R2 value.”

‘Hour’ and ‘precipi’ are obvious independent variables to predicting ridership, since you would expect higher ridership during rush hour and when it’s raining outside. Also if the average temperature is either too high or too low, that may prompt people to use underground transportation instead of enjoying the walk outdoors.

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 R2 (coefficients of determination) value?

```
You calculated R^2 value correctly!
Your calculated R^2 value is: 0.318137233709
```

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?

Since R2 is a measure of the variance between actual data vs. the model prediction, it describes how well the model approximate the actual data. A higher R2 means that the model explains a greater percent of the variability of the data around the mean. R2 of 31% means that our model explains about 31% of the variability around the mean which is rather low. Given a R2 of 31%, the model is not very good at predicting ridership.

Section 3. Visualization

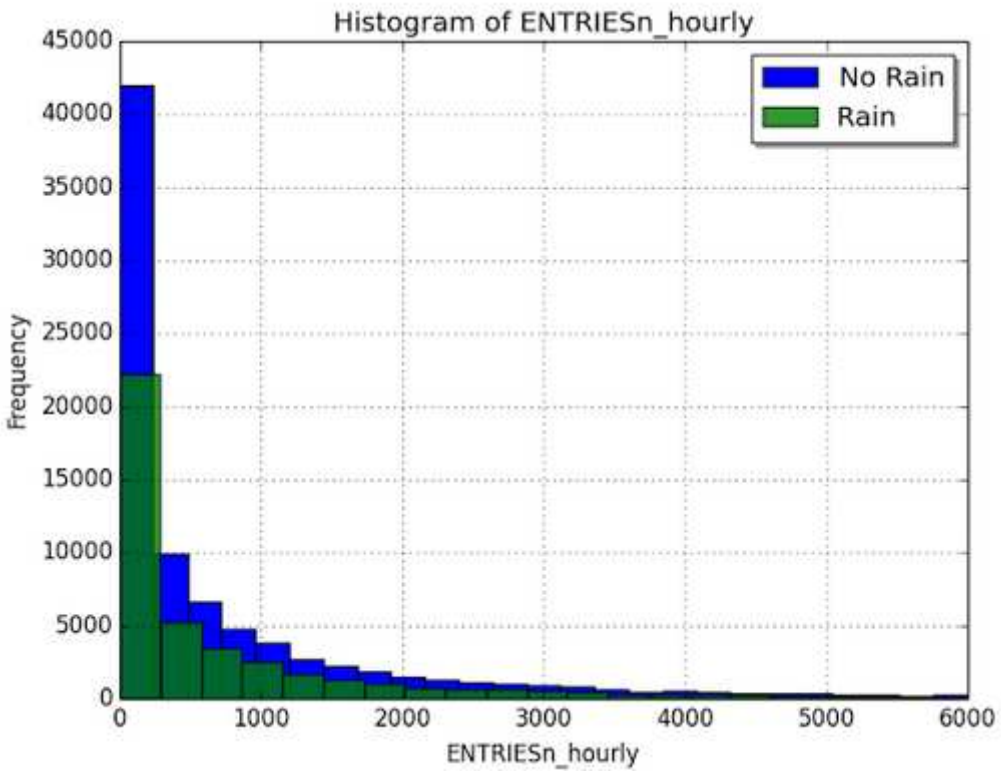
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

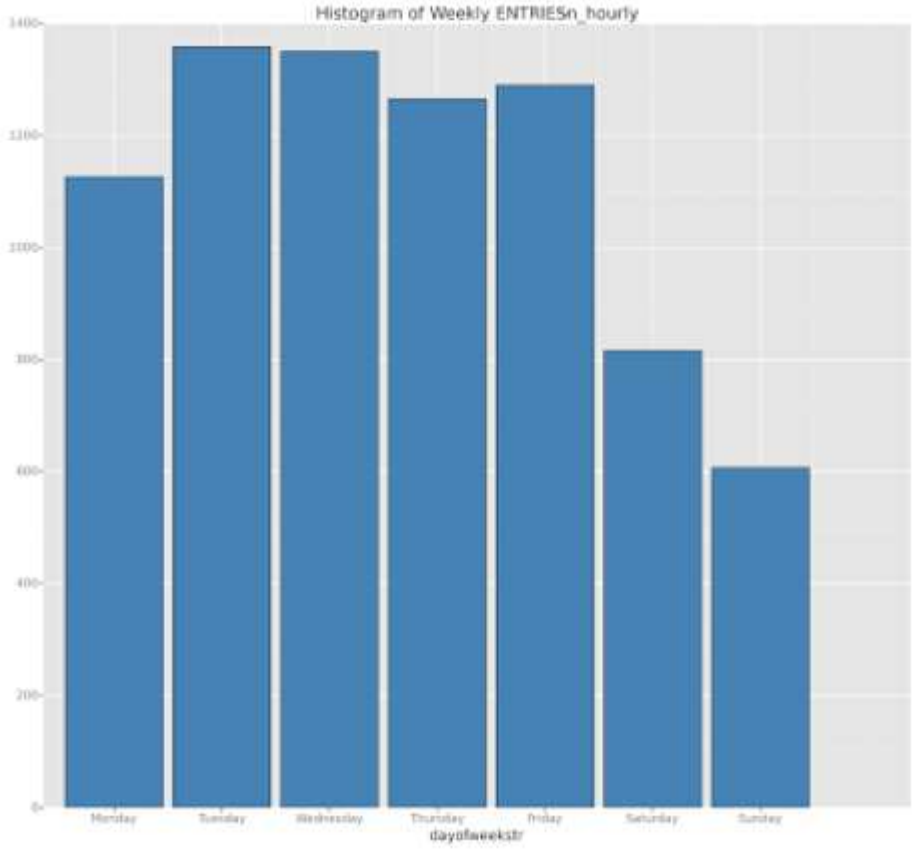
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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



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?

From my analysis and interpretation of the data, it appears that more people ride the NYC 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.

To test the hypothesis that raining day ridership and non-raining day ridership are samples from different population and statistically different from each other, I used the Mann Whitney U test. This test works well on non-normal distributions, which is appropriate in this case. When I plotted a histogram of the

ridership for both raining and non-raining samples, it was not “bell shaped” as a normal distribution would show.

Using the Mann Whitney U test we get the p-values less than .05 which supports the hypothesis that ridership is different between raining and non-raining days. The mean is also different between the 2 samples, which is another indicator that the samples are from different populations and supports the fact that NYC subway ridership are different between raining and non-raining days.

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.

The linear regression model with Rain as an independent variable for ridership was not very strong. I was not able to successfully pick the right features that yielded better than .3 for the R2. R2 of .3 very low in terms of a model fitting the data points. The model I developed was able to only account for 30% of the variability of the actual data around the mean.

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