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, 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.
http://pandas-docs.github.io/pandas-docs-travis/groupby.html#dataframe-column-selection-in-groupby
http://ggplot.yhathq.com/docs/index.html

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?
Mann-Whitney U Test. I used the two-tail P value. Null hypothesis was that the ridership is no different on a rainy day then a non rainy day. 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.
Mann Whitney U Test applies can be applied to unknown distribution while t-test can be applied on to normal distribution only.
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 – 1105.446
Without Rain Mean –1090.278
p-value – 0.025
1.4 What is the significance and interpretation of these results?
From the test we can deduce that both the distribution of entries is statistically 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. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

I used OLS using Statsmodels

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 Hour, Precipi and meanwindspdi
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."

Time at which wind is high with rain as well should result in people deciding to use subway for that time for it being safer option.

2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?
Hour 65.386178
precipi 40.466678
meanwindspdi 35.792983

2.5 What is your model's R^2 (coefficients of determination) value?
0.479217847323

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?
The R value predicts around 50% of the variance only, so this might not be a good fit for our regression model

Section 3. Visualization

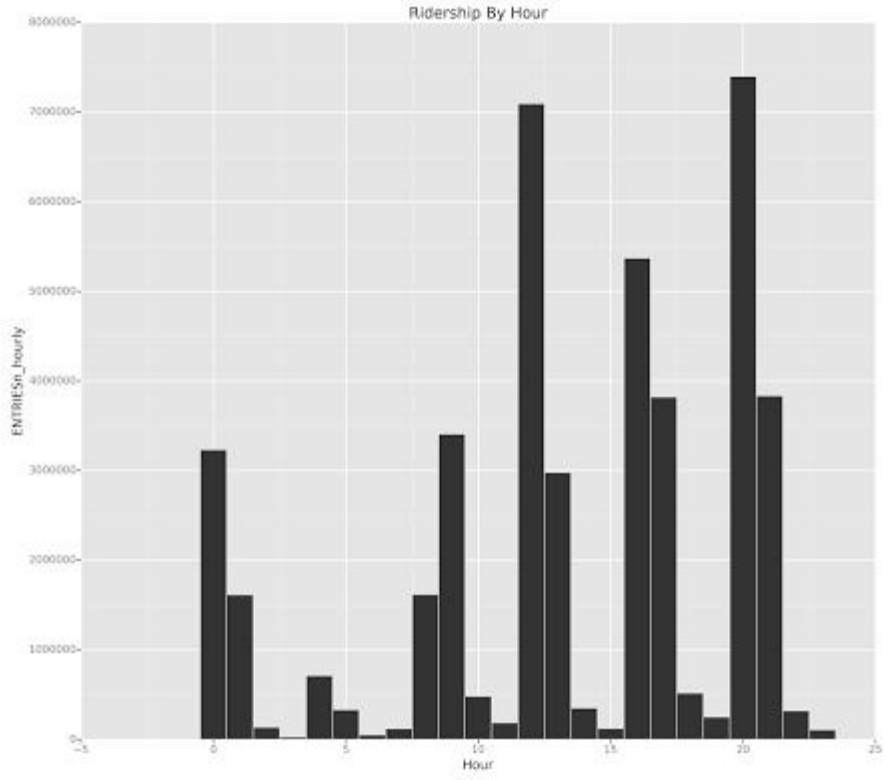
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

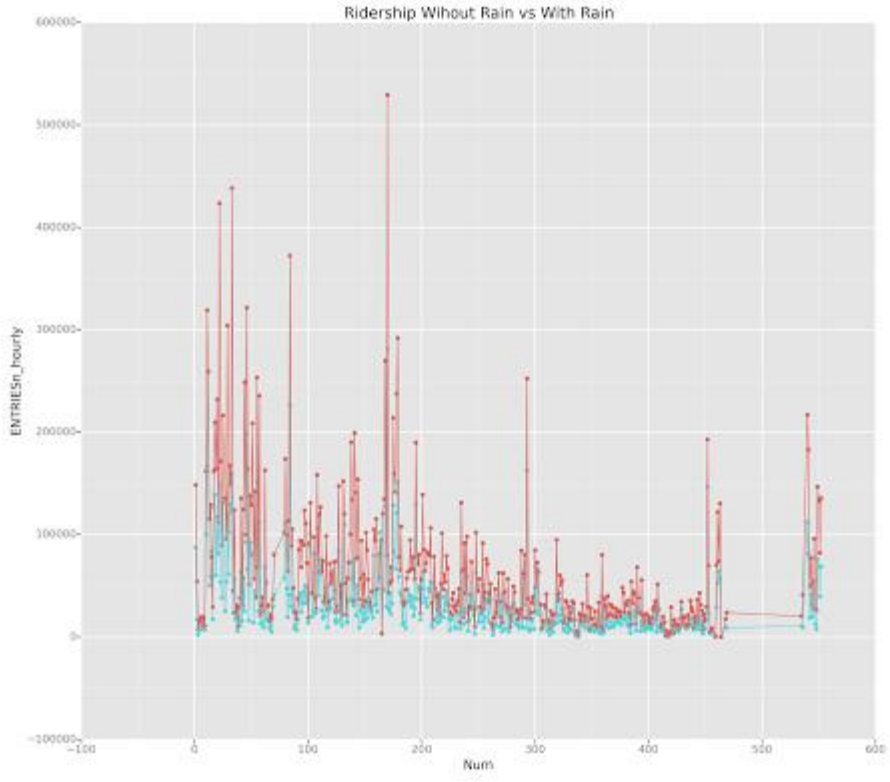
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms,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


Ridership By Hour
ENTRIESn_hourly
8000000
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
−5
0
5
10
15
20
25
Hour



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?
More people ride subway when it is raining, this can be concluded from the Mann Whitney U test and the R^2 value test.
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.
Statistical Mann Whitney U test stated that there both entries were statically different which is because the p-value 0.25 is at the critical value for the mean comparison which is out of the 95% confidence interval which makes it different from the mean, and R^2 value is also shows around 50% variance from the model.

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.

We are only covering few parameters here like rain, fog or wind but many other factors can also contribute to the analysis, for e.g. if there is some road blockage near a certain station people living around that area would tend to take subway that day to get away from the traffic jam, also the ridership can also be divided between categories, some of the riders have only subway as an option so if we can somehow get some statistics on how many of these people do have some other options and they choose subway then we can deduce better analysis over our statistics. If R^2 value is computed over the categorical ridership data then our value could give better result or higher values.

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