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.

Basic examples for ggplot2
http://www.plob.org/2012/09/20/3553.html

How to use datetime in python
http://www.cnblogs.com/rollenholt/archive/2012/04/11/2441699.html

geom_line
http://docs.ggplot2.org/current/geom_line.html

statsmodels.regression.linear_model.GLS
http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.GLS.html?highlight=gls

Python floating point Numbers keep two decimal places the save as type String
http://blog.chinaunix.net/uid-36569-id-2414373.html

Working with missing data
http://pandas-docs.github.io/pandas-docs-travis/missing_data.html

How to shift a column in Pandas DataFrame
http://stackoverflow.com/questions/10982089/how-to-shift-a-column-in-pandas-dataframe

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 Pvalue? What is the null hypothesis? p-critical value?

Two sample T-test. Two-tail.

The null hypothesis is the **mean** of ridership in rainy day is equal to those in no rain day. The p-critical value is 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.

Because the size of sample is large enough to see its population as a normalized distribution, so I used T-test. And because the test is consider whether there is relation between the rain and ridership, I would estimate it have no difference. In other words, the ridership in two sample should have the same distribution.

1.3 What results include the following numerical values: p-values, as well as the means for each of the two samples under test.

The p-value is 0.050

The mean of two samples are:

Rainy day 1105.45

No rain day 1090.28

1.4 What is the significance and interpretation of these results?

The significance is 0.05 for a two-tailed test

P-value is higher than P-critical.

The result is in favor of the null hypothesis.

The mean of ridership in rainy day is equal to that in no rain day, which suggest the rain did not influence the ridership.

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:

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?

Precipi ,rain,Hour,ENTRIESn_hourly,meantempi ,and dummy unit.

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.

I decided to use fog, rainy, meantempi, because I thought that when it is more foggy, not rainy or lower temperature outside, people might decide to use the subway more often.

I decided to use Unit because different unit may have different ridership, which affext Existsn_hourly.

I used feature EXITSn_hourly 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?

fog	93.07682089359774
rain	2.5169874025780996
EXITSn_hourly	0.6467806449873095
Meantempi	-7.678808067415405

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

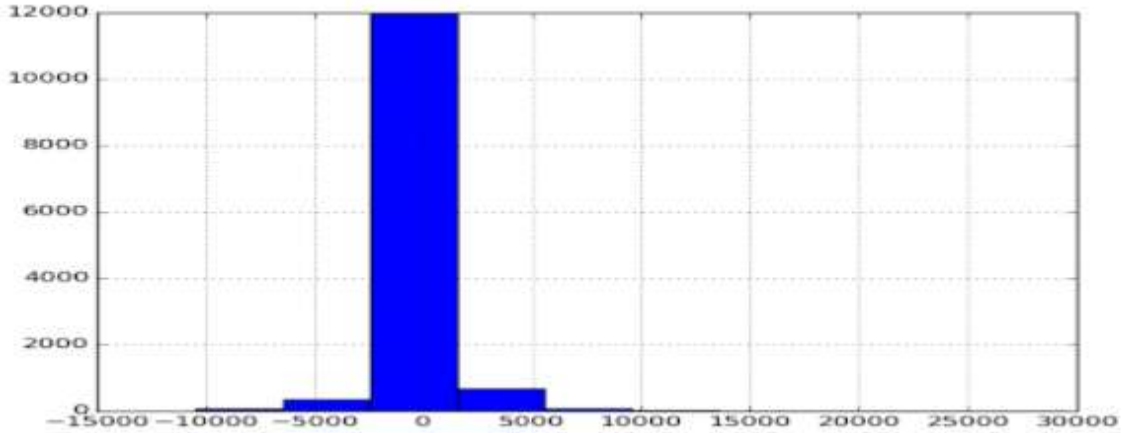
0.628736266033

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 R2 closer to 1, the fit better. The R2 closer to 0, the fit worse.

I think 0.629 is not enough to prove this is the appropriate. *R-square* of 0.629 suggest that the variability of the ridership per hour around the regression line is 0.371 times the original variance; in other words, we have explained 62.9% of the original variability, and are left with 37.1% residual variability.

According to the residual plot as follow:



We can find the residual is roughly normal distribute and a mean of 0 and constant variance of 5000, which means there exists structure that did not contain in the model.

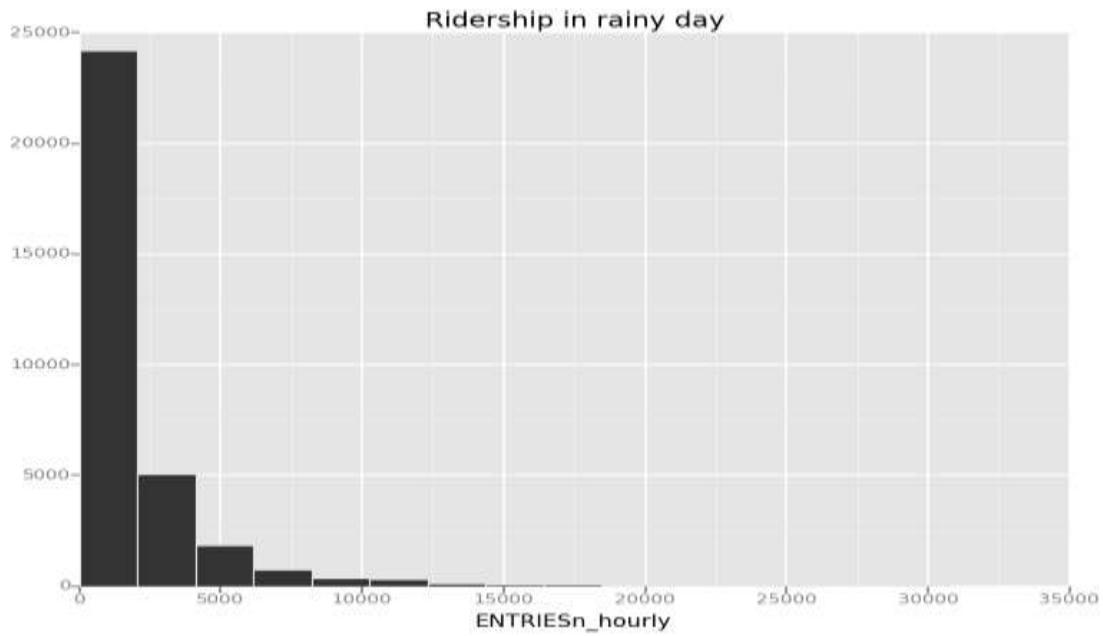
Section 3. Visualization

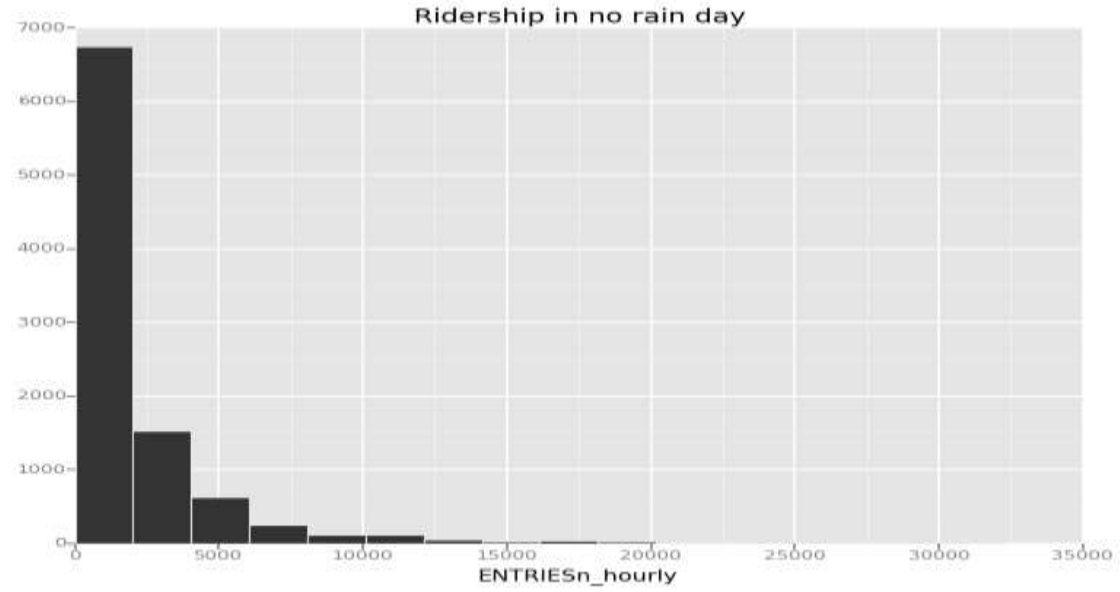
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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.



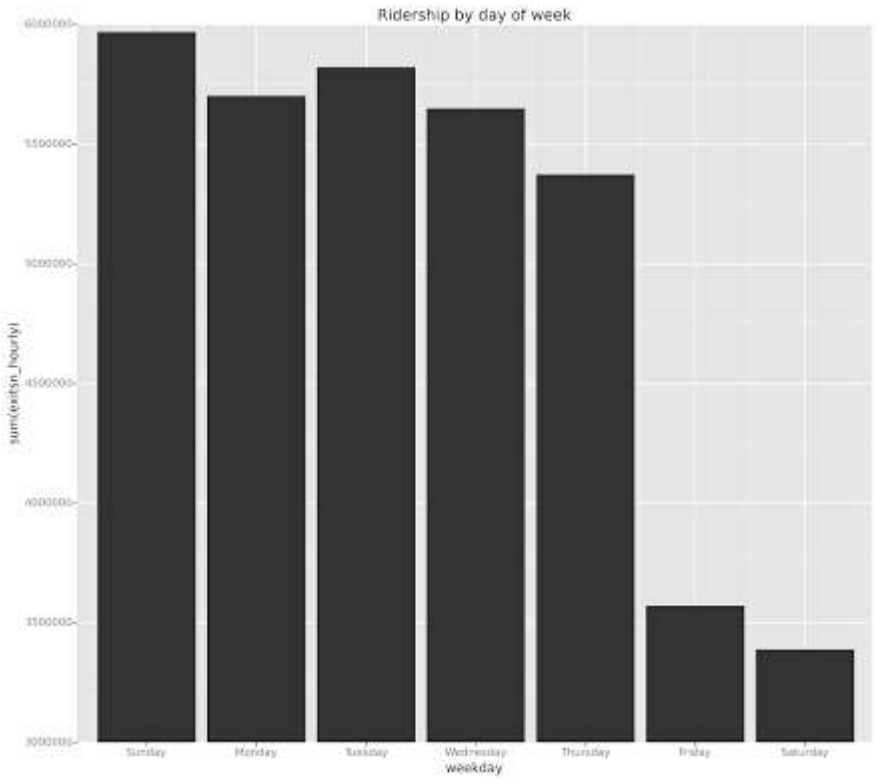


The x-axis is the number of ridership per hour on the subway, have a max number of 35000 and class interval of 2000.

The y-axis is the frequency of ridership per hour in different class.

The distribution in rainy day(upper plot) is similar to those have no rain(down plot), which suggest there is rain cannot influence the ridership.

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:



The x-axis is day of a week from Sunday to Saturday.

The y-axis is the sum of ridership per hour in different day of a week.

The ridership in Friday and Saturday is apparently smaller than other day, and the ridership peak occurs in Sunday.

The result may suggest people take subway more for work other than entertainment.

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?

People ride subway have no relation with rainy

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.

The statistical tests' p value is 0.050 higher than p critical 0.05, the null hypothesis that ridership between the rainy day and not rainy day is equal cannot be reject.

The weight of parameters can also prove the result. We can find the rain have the weight of 2.51 while that of fog is 93.1. The weight of rain parameter was negligible in the prediction 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

   The dataset is wrangled to fill the blank value. Take the example of rainy, if all the blank value is rainy but I choose to wrangle the blank into 0, the answer would be different between the truth and my result.

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

   Using linear regression can fit all the possible result if I take too much parameters in the model. For example, when I calculate the R^2 of the prediction and true value, if I take all the parameters into account, the R^2 will merge to 1.

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