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.

[Toru Yamashige]

- Tokyo Daigaku Kyoyo Gakubu Toukeigaku kyoshitsu, Toukeigaku Nyumon (Introduction of Statistics), Tokyo Daigaku Shuppankai, 1991
- (How to write scatter plot with ggplot), http://motw.mods.jp/R/ggplot_geom_point.html
- (Histogram), http://docs.ggplot2.org/current/geom_histogram.html
- (Linear regression with statsmodel), http://qiita.com/yubais/items/24f49df99b487fdc374d

- (What’s good value for R-squared?), http://people.duke.edu/~rnau/rsquared.htm
- (What is gradient descent? Let’s figure it out with python), http://qiita.com/kenmatsu4/items/d282054ddedbd68fecb0

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?

[Toru Yamashige]

Mann-Whitney U-test with two-tail. The null hypothesis is: “The ridership between rainy days and non-rainy days has no significant difference”. Since this test is two-tailed, p-critical value is 5%/2 = 2.5% = 0.025.

[Toru Yamashige _ 2nd]

The null hypothesis would be: “Both samples for rainy and non-rainy days are extracted from same population”

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.

[Toru Yamashige]

Since the dataset is not distributed as normal distribution and in this case Mann-Whitne U-test is more reliable than Welch’s t-test.

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.

[Toru Yamashige]

The test has got mean ridership in rainy days, non-rainy days, U-value, and p-value as following result respectively.

```
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)
```

1.4 What is the significance and interpretation of these results?

[Toru Yamashige]

Since the p-value is less than p-critical value, we could say that there is a difference between the ridership in rainy days and non-rainy days.

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:

[Toru Yamashige]

Both Gradient descent and OLS.

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?

[Toru Yamashige]

Hour/meanpressurei/meantempi/meanwindspdi/unit(as a dummy)

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.

[Toru Yamashige]

I chose them since p-value of these features are reliable enough according to OLS result(less than 0.05).

[Toru Yamashige _ 2nd]

My choice of these features is based on the question “Is this feature reliable enough on the model? To be more precise, is p-value of the feature less than 5%?”. After experimenting each feature, I have found that those selected features’ p-value are less than 5%. For example, the p-value ofr “rain” turn out to be 71.3% thus I removed this feature from the 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?

[Toru Yamashige]

<<OLS>>

Hour:464.5, meanpressurei:-43.75, meantempi:-28.12, meanwindspdi:42.56

<<Gradient Descent>>

Hour:464.2, meanpressurei:-34.66, meantempi:-26.99, meanwindspdi:42.60

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

[Toru Yamashige]

OLS	: 0.457
Gradient Descent	: 0.459

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?

[Toru Yamashige]

We could say that the model is around 50% reliable to predict the ridership, which means "so-so". However, rather than the prediction reliability, we could interpret from the outcome that 'Hour' value apparently is the most significant to the ridership.

[Toru Yamashige _ 2nd]

According to the R^2 value itself, we can say that the model has decreased the differences between the actual value approximately 50% better than what of the mean value. However, if we look at the plots of the residuals as described on Figure1 below, it can be noticed that the residuals are scattered in cyclic form. This is because that the ridership is highly reliable on the station: UNIT value, and the data has been regularly taken from the station each day.

Since the ridership is not in the linear form but the cyclic form, “linear” formed model is not appropriate to predict the value. Thus I would say the model I have reached is NOT appropriate for the prediction.

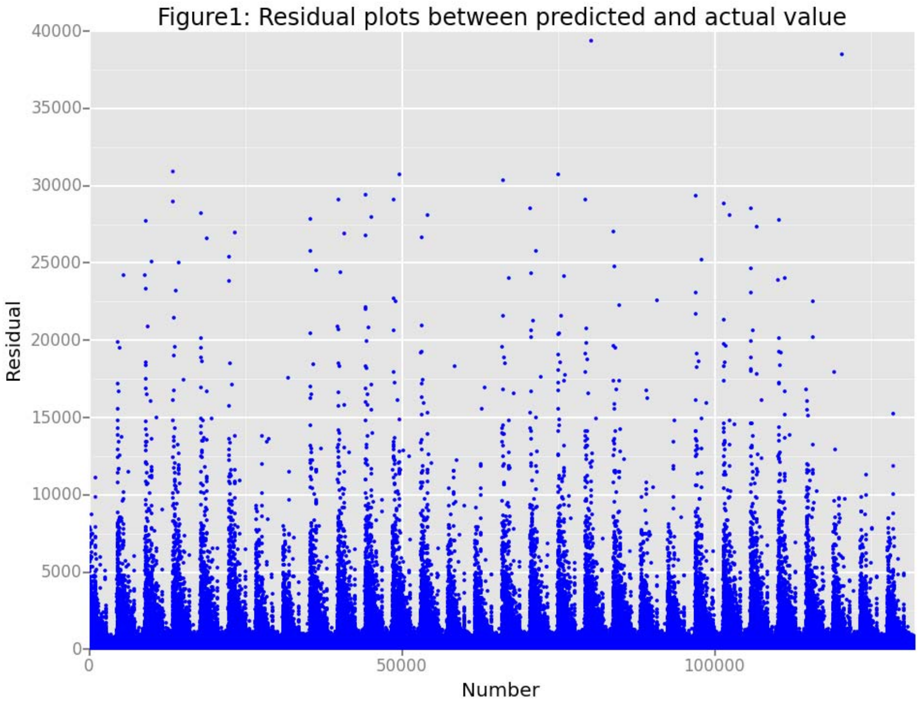


Figure1: Residual plots between predicted and actual value

Section 3. Visualization

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

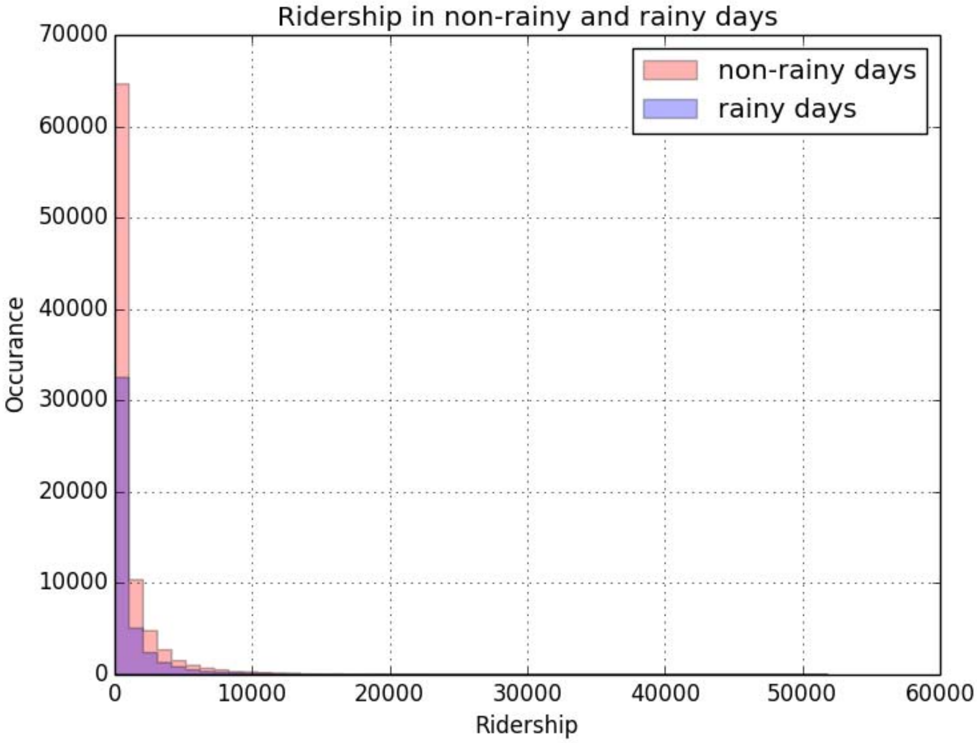
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

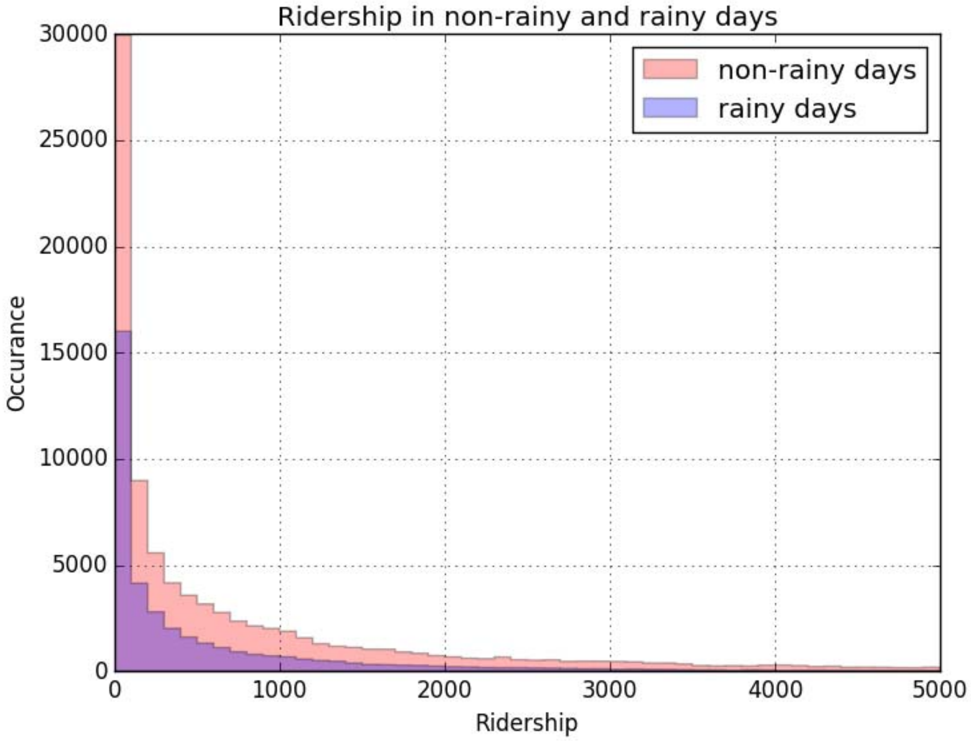
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms,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

[Toru Yamashige]



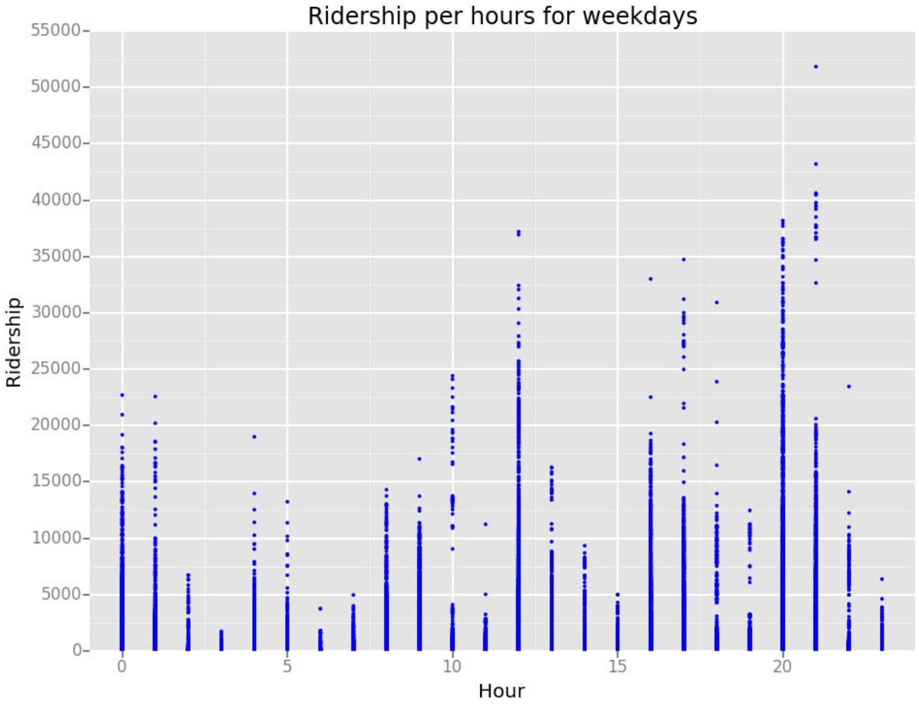
[Toru Yamashige _ 2nd]

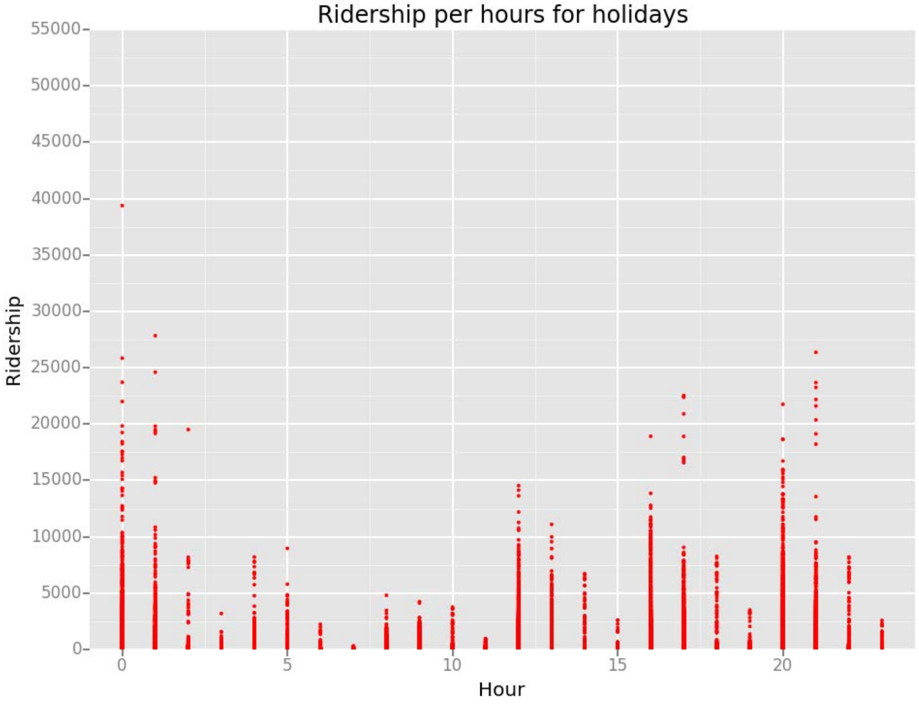


3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

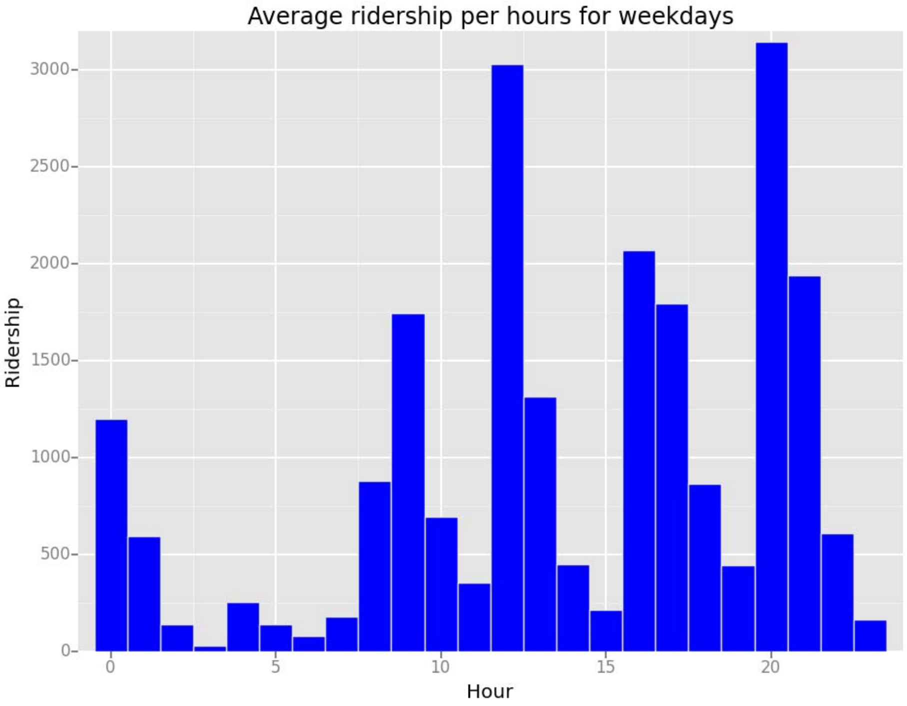
- Ridership by time-of-day
- Ridership by day-of-week

[Toru Yamashige]

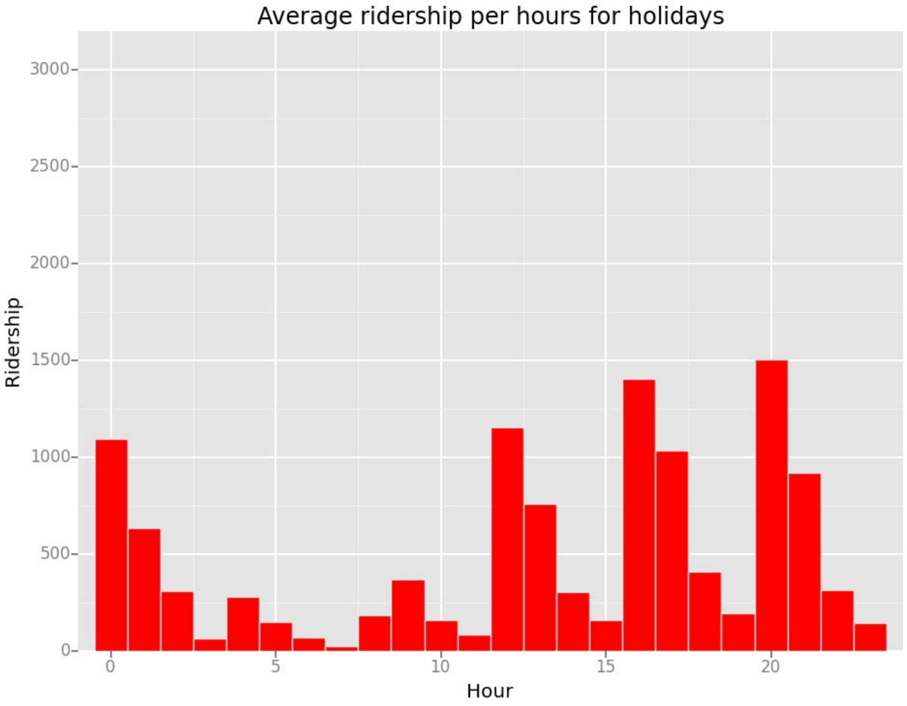

Ridership per hours for weekdays
Ridership
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
5
10
15
20
Hour


Ridership per hours for holidays
Ridership
55000
50000
45000
40000
35000
30000
25000
20000
15000
10000
5000
0
0
5
10
15
20
Hour

[Toru Yamashige _ 2nd]



Ridership on working hours(8am – 21pm) shows bigger than what of on holidays.



Ridership shows less than what of on weekdays, however, the ridership on the midnight is almost same as weekdays.

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?

[Toru Yamashige]

People ride on the subway more 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.

[Toru Yamashige]

By Mann-Whitney U test, p-value was calculated to be less than 0.25 with rainy days mean value bigger than what of non-rainy days, which means there is a significant difference on ridership between rainy and non-rainy days. However, we need to be noted that rainy or non-rainy conditions does NOT reliable on predicting the ridership, since the p-value of the feature is around 0.7, much bigger than 0.05. Also we can assume this from the coefficients of the feature is negative value which does not make sense on the previous U test.

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.

[Toru Yamashige]

From Dataset point of view, if it includes some abnormal values such as outlier obtained by measurement error for example, the model would be drastically impacted as unreliable. In order to prevent this, it'd be better to seek/eliminate those outliers first and try how the reliablity would change.

From analysis point of view, the linear regression; gradient descent will not take effect when there are more than two minimum values on the cost function. Also we need to be careful on determining the alpha value; learning rate otherwise the theta value will be unreliable.

[Toru Yamashige _ 2nd]

In addition to the above comment, I would like to state that the linear regression model will not take effect when the original data has non-linear form.

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

[Toru Yamashige]

As I stated in 3.2, it seems there is a difference on ridership between weekdays and holidays. I have tested Mann-Whitney Utest and figured out that p-value is almost 0, which means there is a significant difference between those two conditions(I chose Mann-Whitney test since the dataset are scattered unlikely the normal distribution). I have included this feature: weekdays/holidays in OLS, then R^2 square was improved to be 0.461, with the coefficient: -242.3 and p-value: 0. Thus, I conclude as including this feature is very reliable on improving the model.

```
====================================================================
```