

Analysing the NYC Subway Dataset

Section 0. References

<https://bespokeblog.wordpress.com/2011/07/11/basic-data-plotting-with-matplotlib-part-3-histograms/>

<http://flowingdata.com/2014/02/27/how-to-read-histograms-and-use-them-in-r/>

<https://plot.ly/histogram/>

<http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm>

http://en.wikipedia.org/wiki/Ordinary_least_squares

http://docs.ggplot2.org/0.9.3.1/geom_bar.html

http://dss.princeton.edu/online_help/analysis/interpreting_regression.htm

<http://people.duke.edu/~rnau/regintro.htm>

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm

http://graphpad.com/guides/prism/6/statistics/index.htm?one-tail_vs_two-tail_p_values.htm

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?

Statistical test: Mann–Whitney U test, using two-tail P value

Null hypothesis: there is no significant difference between the average ridership in rainy & non-rainy hours

P-critical value: 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.

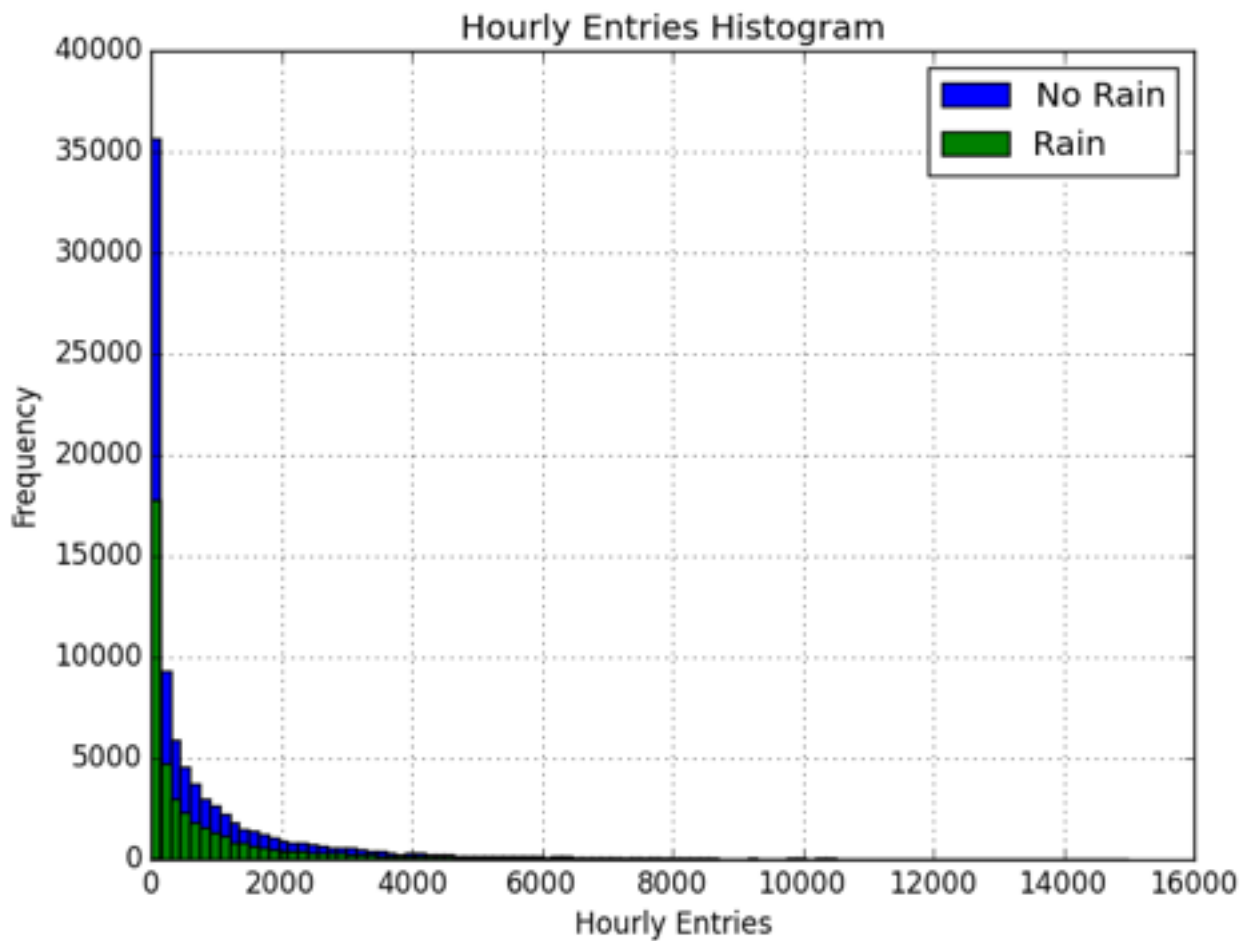


Figure 1: histogram plotting the ridership distribution of rainy and non-rainy hours

As we see in figure 1 the ridership distribution of the two independent samples (rain, no rain) is skewed right (NOT Normal), which is a good case to use Mann–Whitney U test.

Also this dataset meet other assumptions Mann–Whitney *U* test as follows:

- a. The dependent variable here (enries_hourly) is continuous.
- b. The independent variable here (rain status) consists of two categories (rain, no rain)

I have used two-tailed test here because I need to test if there is a significant difference between the mean of the two groups (ridership in rainy & non-rainy hours) regardless of which one is greater than or less than the other group.

Also I have no certainty about the direction of difference, I can't assume one direction and neglect the effect of the other one

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.

U= 1924409167.0

P-value: $0.024 \times 2 = 0.048$

Mean of ridership with rain: 1105

Mean of ridership without rain: 1090

1.4 What is the significance and interpretation of these results?

I can reject null hypothesis, as p-value is less than p-critical value, and assure that there is a significant difference between the average ridership in rainy and non-rainy hours.

Section 2. Linear Regression

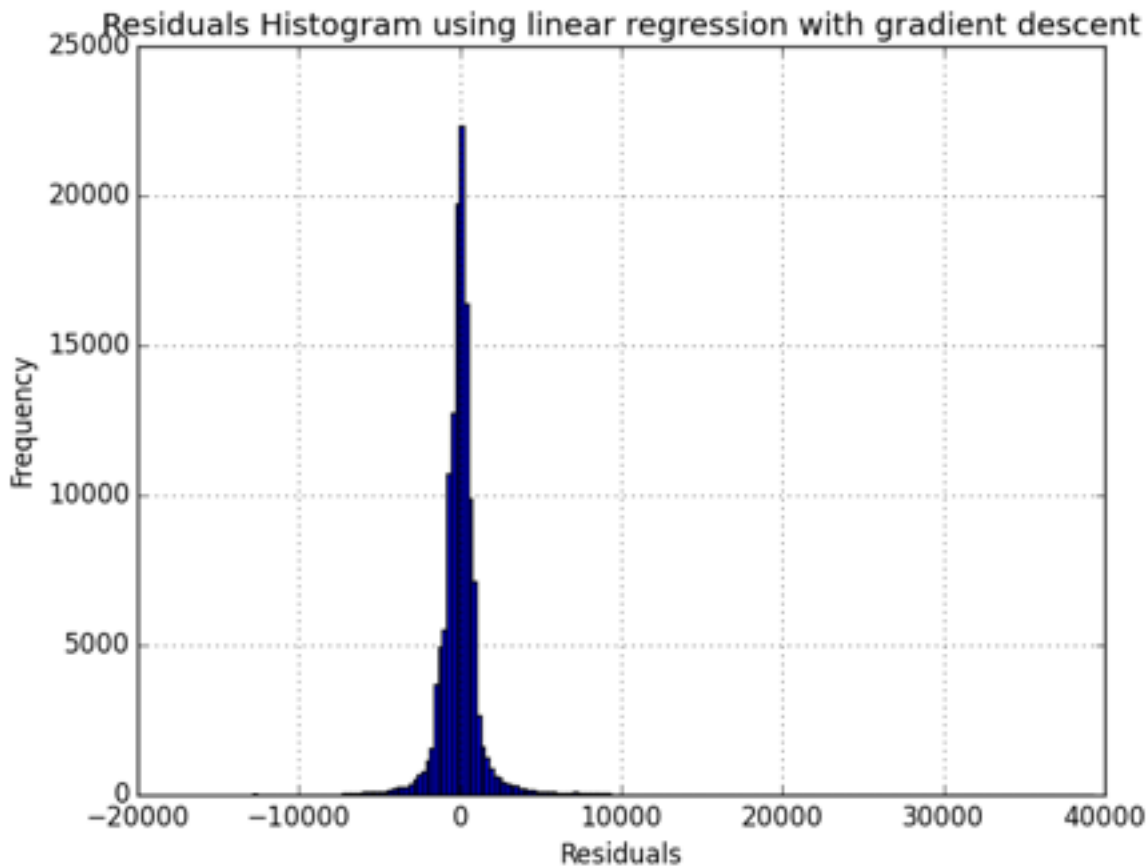


Figure 2: histogram plotting the residual between the observations and the predictions of the entries_hourly for the linear regression with gradient descent model

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model?

*Gradient descent (as implemented in exercise 3.5)
OLS using Statsmodels
Or something different?*

In exercise 3.5 I have used Gradient descent with some modifications on the feature list and the dummy variable, the result r squared was 0.50.

After plotting the residual between the resulted predictions and the observed ones as shown on figure 2, I found that it has a normal distribution which is a good indicator of using the regression model, but I was interested to use another approach of the linear regression. So I decided to use 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?

Features: rain, meantempi

There is also a dummy variables including Hour and 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.

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."

1. rain: is used because I thought that when it is raining people might prefer using the subway
2. meantempi: I thought that the mean temperature may affect the decision of using the subway
3. unit [dummy variable]: is used because I think that there may be stations have more people using the subway than other stations, it has a very powerful effect on the r squared value.
4. hour [dummy variable]: is used because I think that the hour of the day affects the ridership of the subway it may be high during rush hours and very low on the night hours

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?

<u>rain</u>	2.88
<u>meantempi</u>	-8.04

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

0.50

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 resulted r squared value indicates that the model (with its features) accounts for 50 % of the variability of the ridership. I think that this linear model to predict ridership is appropriate for this dataset because this value of r squared means that there is a linear relationship exists between features and the ridership

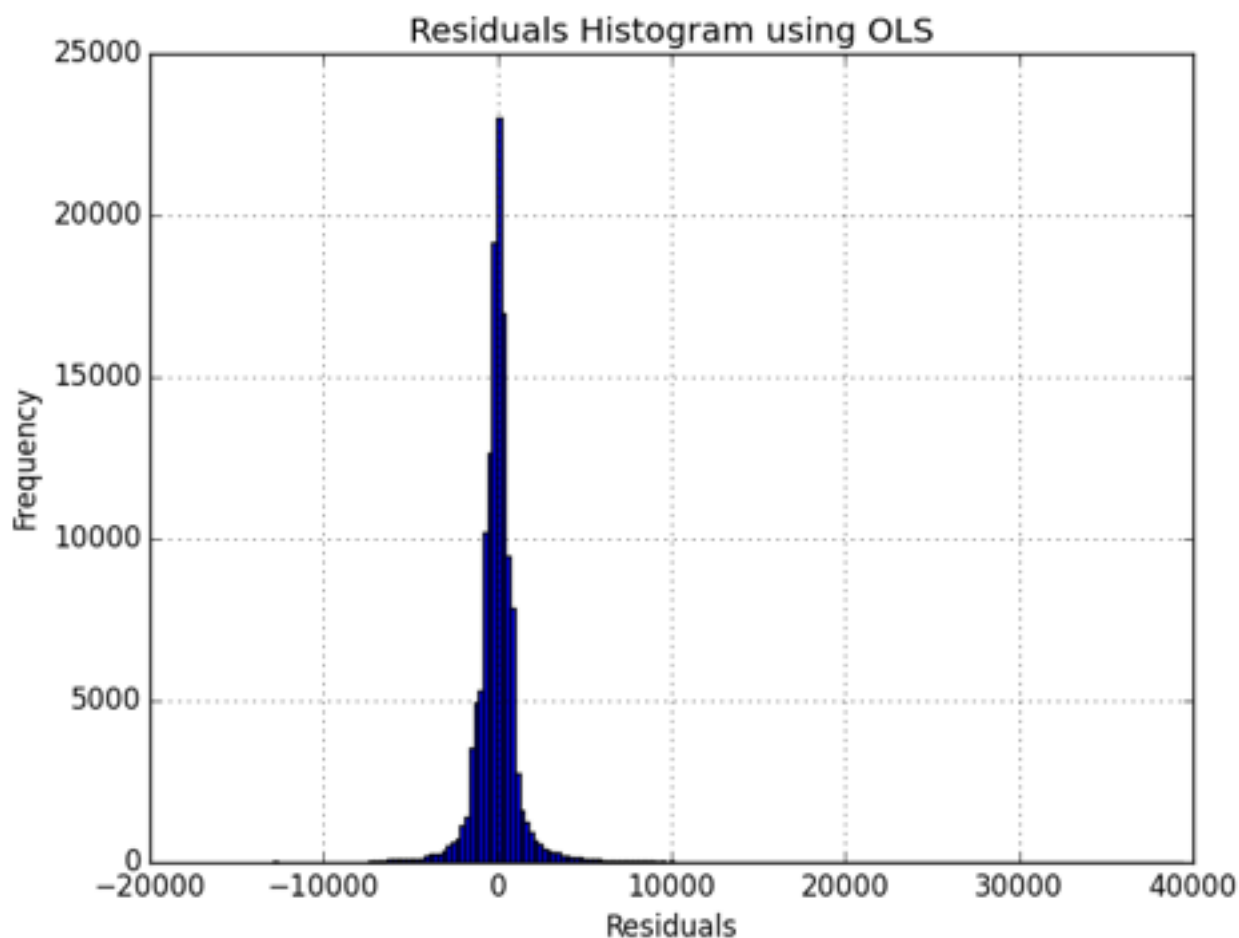


Figure 3: histogram plotting the residual between the observations and the predictions of the entries_hourly for the linear regression model OLS

Section 3. Visualisation

Please include two visualisations 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 visualisation should contain two histograms: one of `ENTRIESn_hourly` for rainy days and one of `ENTRIESn_hourly` for non-rainy days.

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

2. If you decide to use 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.

3. 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.

4. 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.

Figure 1 in section.1 shows the histograms of `ENTRIESn_hourly` for rainy hours and `ENTRIESn_hourly` for non-rainy hours.

We can notice that the frequency of the ridership on non-rainy hours is always greater than the ridership on rainy hours.

There also another thing on the hourly entries, we can see that the frequency is very high on the most left (the smallest number of entries per hour) and skewed to right.

3.2 One visualisation 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

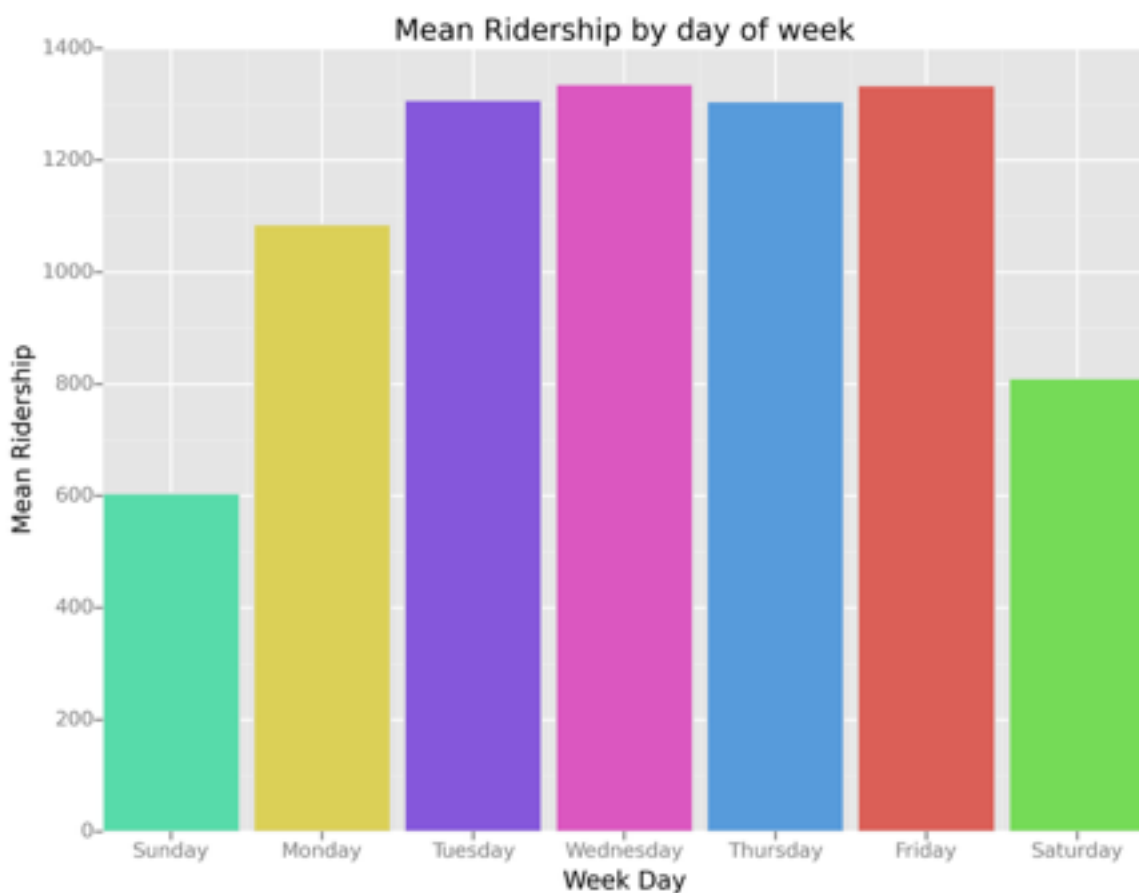


Figure 4 :mean ridership per day of week

Figure 4 shows that the highest mean ridership happened on Tuesday, Wednesday, Thursday, and Friday, in which mean ridership exceeded 1300 entries_hourly.

The lowest mean ridership happens on Sunday, almost 600 entries_hourly

In between Saturday and Monday with mean ridership 800 and 1100 respectively

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.

Using Mann–Whitney U test, with two-tail P value indicated that there is a significant difference between the average ridership on rainy and non-rainy hours, because the p-value (0.048) is less than 0.05.

The results show that there is a difference of 15 entries increase in the mean ridership in rainy hours than non-rainy hours.

Using my regression model OLS, I expect that the ridership is increased by 2.88 when it is raining if the other features are fixed

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?

Dataset shortcomings:

I think that weather condition may have a big effect on the ridership. The problem in this dataset is [DATE] which is all about May 2011. This month might have its own weather characteristics (rain, temperature...), which differs from month to another, so if we train our model on this data, we should apply this model only on 'May' to be fair.

Analysis method shortcomings:

There was also a shortcoming of my analysis, as it is a multiple regression model I was faced by a multicollinearity problem which caused my coefficients to be unstable. It took me time to detect and remove it.