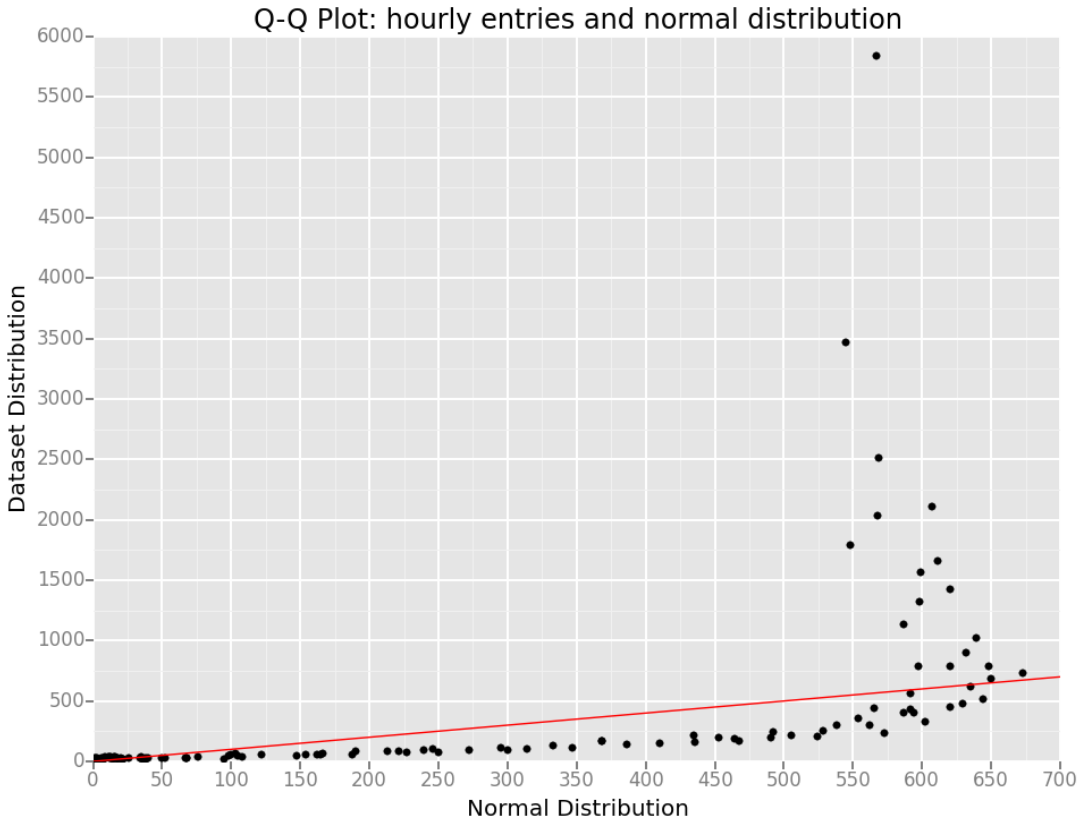
Section 0. References

http://blog.yhathq.com/posts/ggplot-for-python.html
https://ggplot.yhathq.com
https://www2.palomar.edu/users/rmorrissette/Lectures/Stats/ttests/ttests.htm
http://www.datarobot.com/blog/ordinary-least-squares-in-python/
http://en.wikipedia.org/wiki/F-test

Section 1. Statistical Test

1.1 Which statistical test did you use to analyse 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?



As our data distribution is not normal (check the Q-Q plot above), we didn't use Welch's T-Test. Instead, we used Mann-Whitney U-Test as it does not assume any particular distribution.

The test is two-tailed and the NULL hypothesis is: “rain does not affect subway’s ridership”.
The alternative hypothesis is: “rain does affect subway’s ridership”.
The p-critical value is ~5% (4.999%).

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.

This test is applicable because it doesn’t assume any particular distribution of the data-set.

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.

Mean of ridership for rainy hours: 1105.45
Mean of ridership for non-rainy hours: 1090.28
U: 1924409167
p: ~2.5%

1.4 What is the significance and interpretation of these results?

Assuming NULL hypothesis is true, there is ~4.99% chance to have a U-value as extreme as two times 1924409167 (two-tail test).
With a confidence level of 95%, the NULL hypothesis is rejected. The subway ridership depends on rain.

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:

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

I used Gradient descent and OLS.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

For those models, I used the following features: Hour, EXITSn_hourly.
For Gradient Descent, I tried different values of alpha.

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."**

I started with all features. Then I used trial and error on different sets of features trying to maximize R^2 and minimize costs for Gradient Descent only. At each step I removed features with very small or no impact on R^2.
I ended with the following set of features : 'rain', 'Hour', 'EXITSn_hourly', 'meandewpti', 'meanpressurei' and 'meantempi'; with R^2~0.54 for Gradient Descent and R^2~0.55 for OLS.
OLS comes with a metric measuring collinearity. I used this metric in order to have less collinear features and I ended with the following set: 'Hour', 'EXITSn_hourly' and almost the same value (to .001) for R^2.
Trying the same set with Gradient Descent didn't make a difference on R^2.

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

Set 1: 'rain', 'Hour', 'EXITSn_hourly', 'meandewpti', 'meanpressurei' and 'meantempi'

rain	: 6.1438
Hour	: 19.7794
EXITSn_hourly	: 0.855
meandewpti	: -0.4982
meanpressurei	: 9.9403
meantempi	: -2.3683

Set 2: 'Hour', 'EXITSn_hourly'

Hour	: 27.4904
EXITSn_hourly	: 0.8585

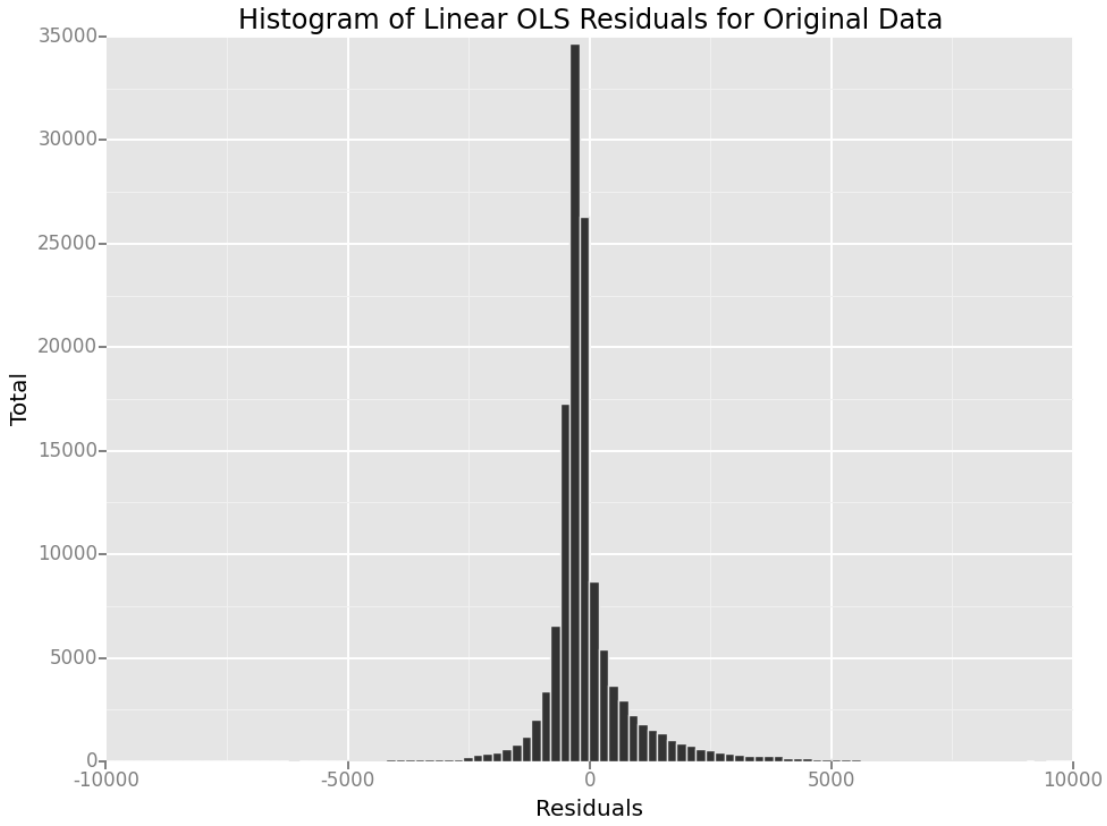
Features are not normalized.

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

0.5547

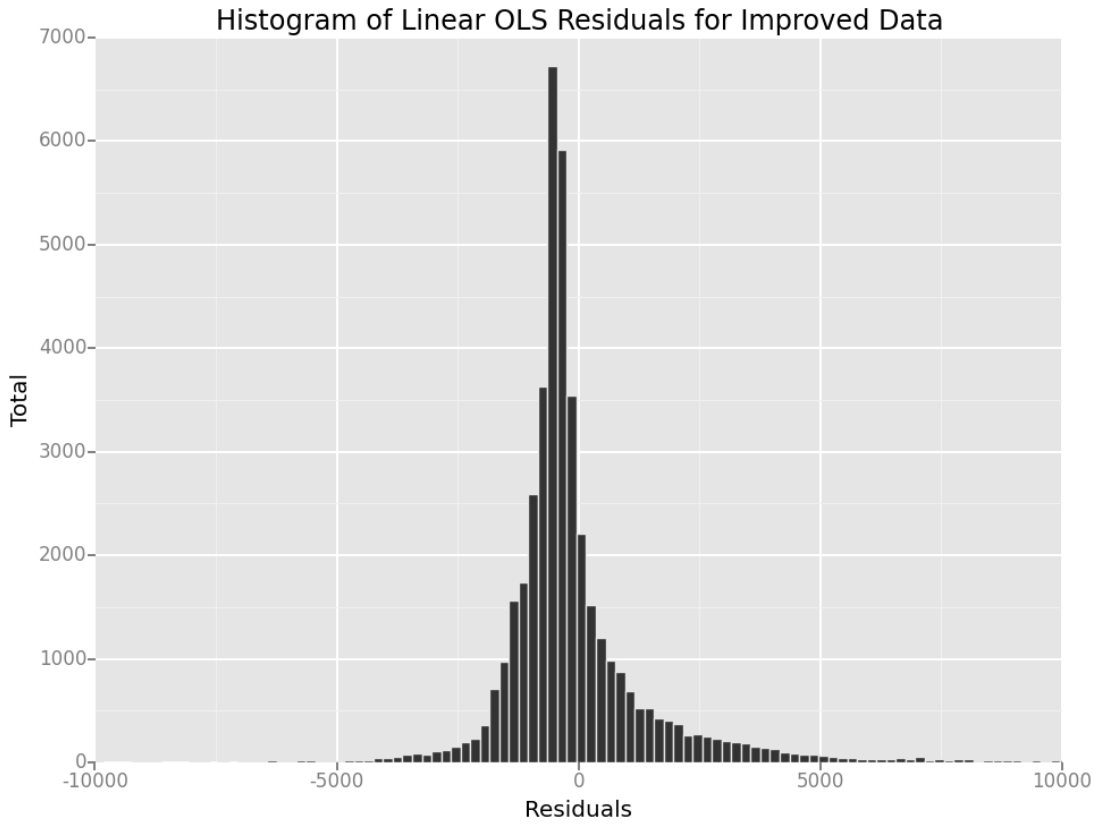
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R^2 value measures how much our model fits data. The best value found is 0.55. In order to verify if this value is acceptable, let's plot the residuals.



Residuals above (raw data minus predictions) seem slightly skewed to the right (fat tail). If our model balances well between less frequent large underpricing and more frequent small overpricing, we can consider that our R^2 value is acceptable, and that our model will predict well subway's ridership.

The plot above is drawn from the complete dataset. With the enhanced dataset, the R^2 metric is 0.42.

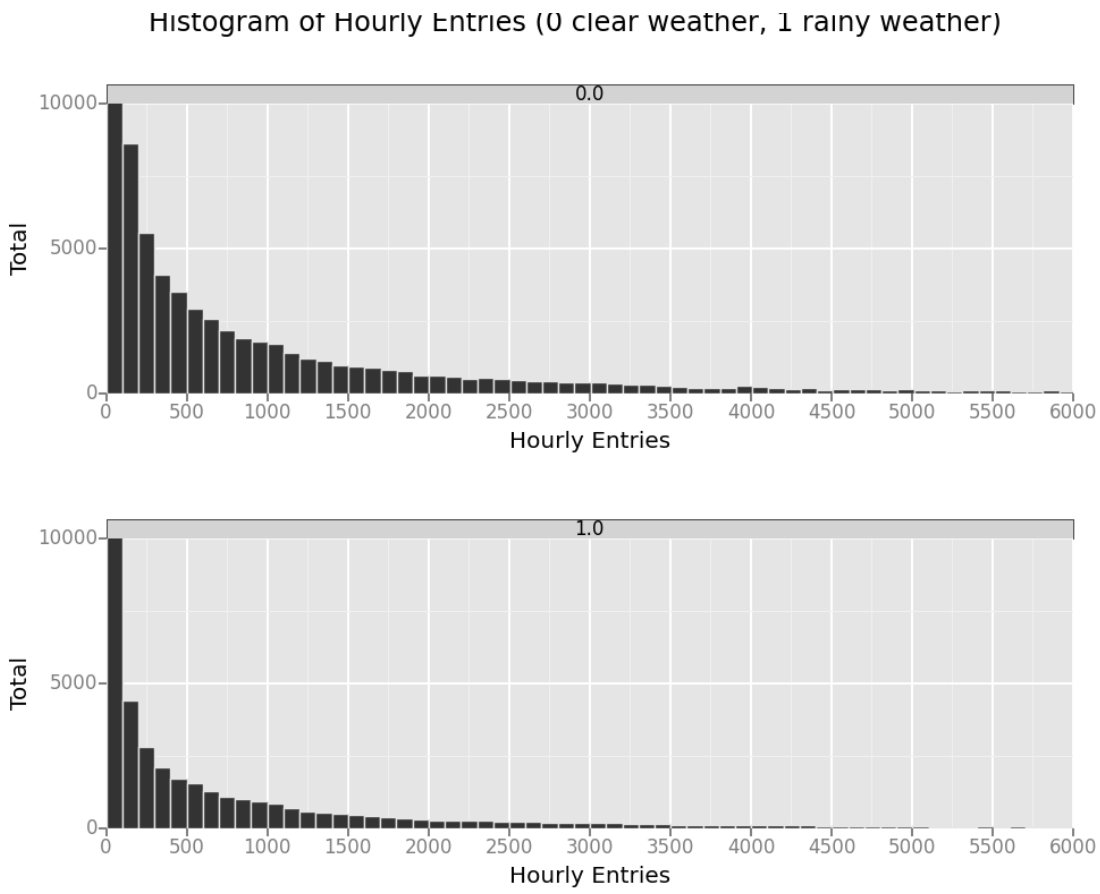


With the new dataset, residuals shape is approximatively the same, but R^2 metric is smaller: 0.42. Also, the fat tail is more important. A parametric or polynomial model may be more adapted.

Section 3. Visualization

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

We cannot infer any comparison from these histograms. However, we can see that the distribution of hourly entries, whether it is raining or not, is skewed to the right with a fat tail. We may use logarithm to approach a normal distribution.

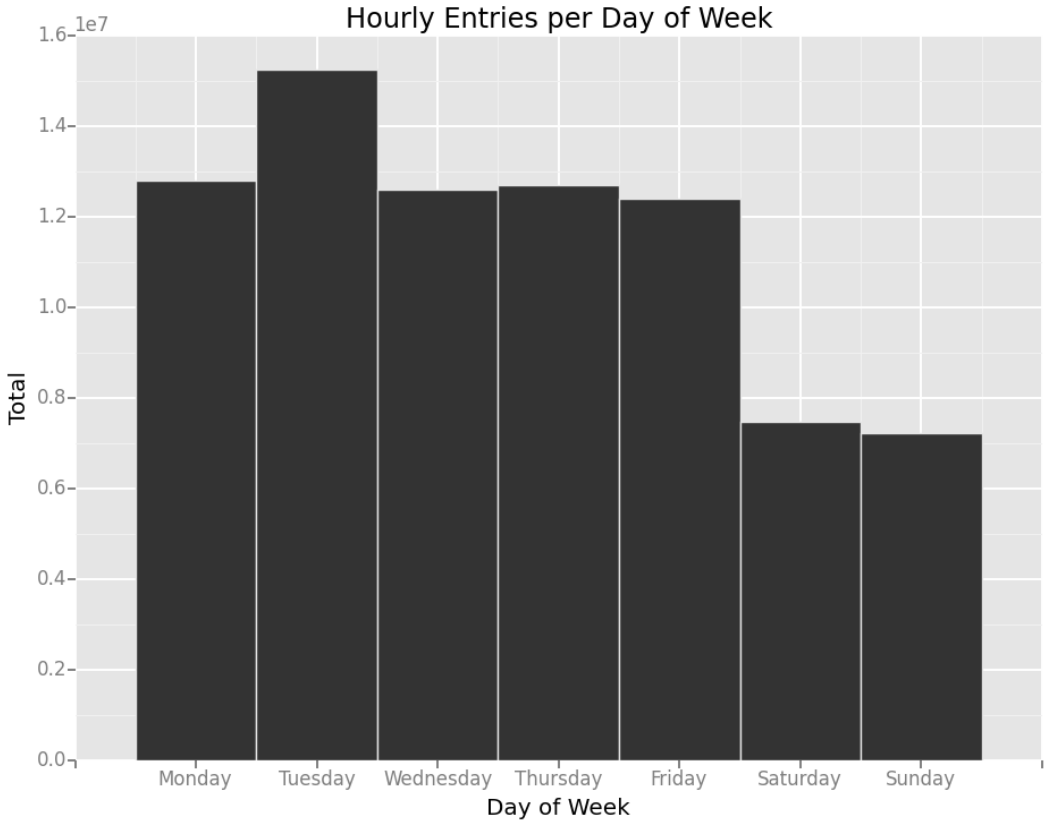


3.2 One visualization can be more freeform. Some suggestions are:

Ridership by time-of-day

Ridership by day-of-week

Ridership by day of week: total number of hourly entries



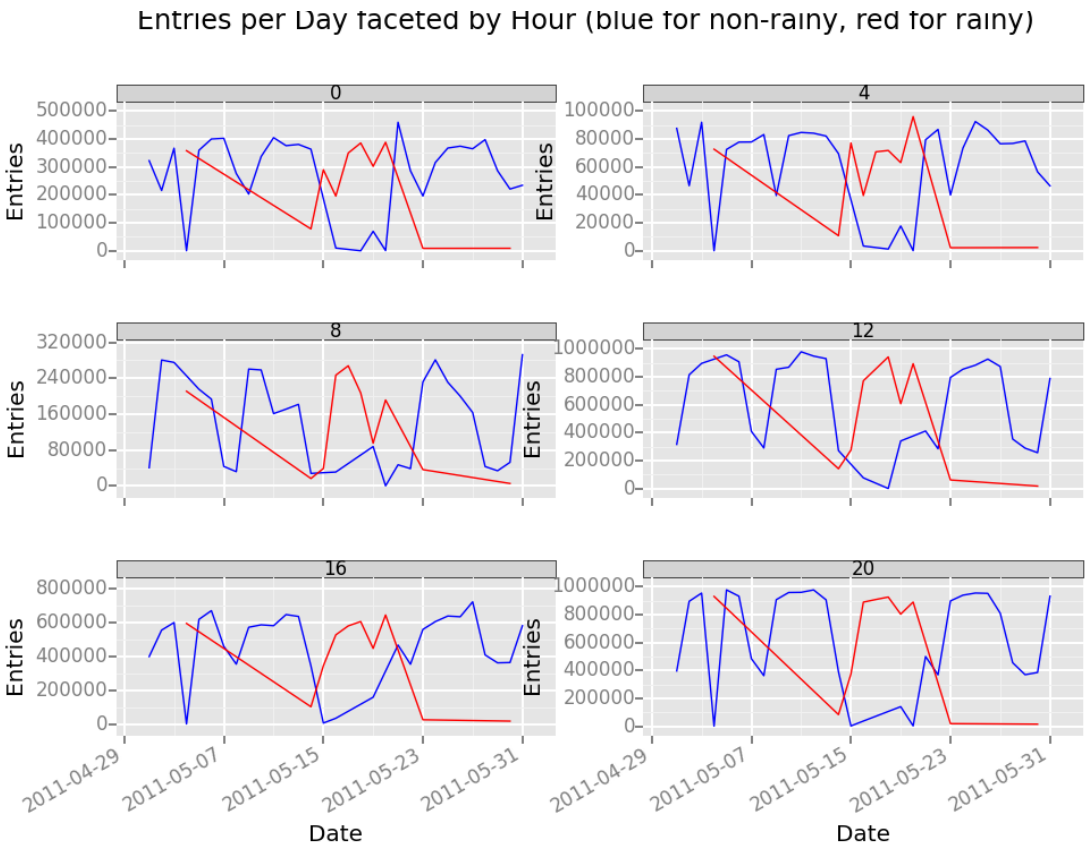
There are more riders on Sunday, Monday and Tuesday than for the other days.

<u>Comparison of entries per day for rainy and clear weather</u>

Legend:

- Rainy: red
- Non-rainy: blue

Graphically, there is no difference between riderships of rainy and non-rainy hours.



Section 4. Conclusion

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?

Yes they do.

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.

With the visualization of number of entries according to the weather (rainy or not), it is not clear whether ridership behavior is different between rainy and non-rainy times. However, the non-parametric test of Mann-Whitney confirms this difference. Though, we don't know in which direction.
On the other hand, when we introduce 'rain' in the regression model, collinearity increases. Which means that 'rain' explains in the same way our target as 'hour' and 'EXITSn_hourly'. Thus, 'rain', explains hourly entries.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:
Dataset,
Analysis, such as the linear regression model or statistical test.

There is no much information about the quality of the dataset: there was certainly some wrangling, and it could be useful to know how much this wrangling did influence results. Following some forum discussions, there was a lot of treatments around precipitation data. Moreover, times have been aggregated, thus diluting the impact of rain on entries (as the norm is clear weather - most of the time it is clear and from time to time it rains unless NYC is in Scotland - using larger time slots hides the real potential increase in ridership).
Regression models don't show any strong correlation between the weather (rainy or not) and the number of riders.
In my analysis, I couldn't rely much on graphs, as the difference in ridership behaviour between rainy and clear hours was not very accentuated, so I couldn't draw conclusions with high certainty. Maybe applying some logarithmic or parametric functions on data may have helped to bring into light some difference. I will give it a try later.

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

The dataset contained data of May, which is not enough to draw conclusions on NYC subway ridership for the whole year. The temperature, the intensity of rain, the wind, all are parameters that seem obvious to have an influence on ridership of NYS subway. However they don't as the regression model shows. Maybe because it is spring season and the weather is not extreme.