Short Questions to Analyzing the NYC Subway Dataset

Short Questions

Section1. Statistical Test

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?*

Statistical test used: Mann-Whitney U test

Type: 1 sided P value

Null hypothesis: There is no difference in hourly entries between rainy and non-rainy days

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.*

Reason for Applicability: The two sample distributions do not appear to be normal, so it is valid to use the Mann-Whitney U test

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.*

P-value: 0.0249999127935

Mean of hourly entries with rainy days: 1105.44637675

Mean of hourly entries with non-rainy days: 1090.27878015

4- *What is the significance and interpretation of these results?*

Interpretation: Using the Mann–Whitney U test, we got a significant p-value (< 5%), which means that we should discard the null hypothesis (which states that there is no difference in entries hourly between rainy and non-rainy days) in favor of the alternative.

Significance: Conclude that it is expected that rainy days have (on average) more hourly entries than non-rainy days.

Section2. Linear Regression

1- What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

a- Gradient descent (as implemented in exercise 3.5)

b- OLS using Statsmodels

c- Or something different?

I used gradient descent and OLS (later in the optional part)

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

Features used (named as the dataframe column names): **rain, Hour, meantempi, meanpressurei, meanwindspdi and UNIT (dummy)**

3- Why did you select these features in your model?

The following points explain why I picked these features in the format:

{Feature: reason for including the feature}

- **Rain:** Well, Rain has been the topic of discussion for many exercises so by nature I included it, also over the course of the lesson we were trying to figure out whether Rain affects the hourly entries column so that's a good reason to include it
- **Hour:** By intuition not all people ride the subway on the same time, so rush hours definitely got to have more entries than others
- **meantempi, meanpressurei, meanwindspdi:** All of these features represent the state of the weather, so in a sense it's a similar feature to Rain. I thought of these features as a complementary feature to Rain, if it's not raining but the wind speed is high maybe people will use the subway more, same thinking goes to pressure and temperature.
- **UNIT:** This one is very important because some units see a lot of people moving through them, some very few.

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

0.46834

5- What does this R2 value mean for the goodness of fit for your regression model?

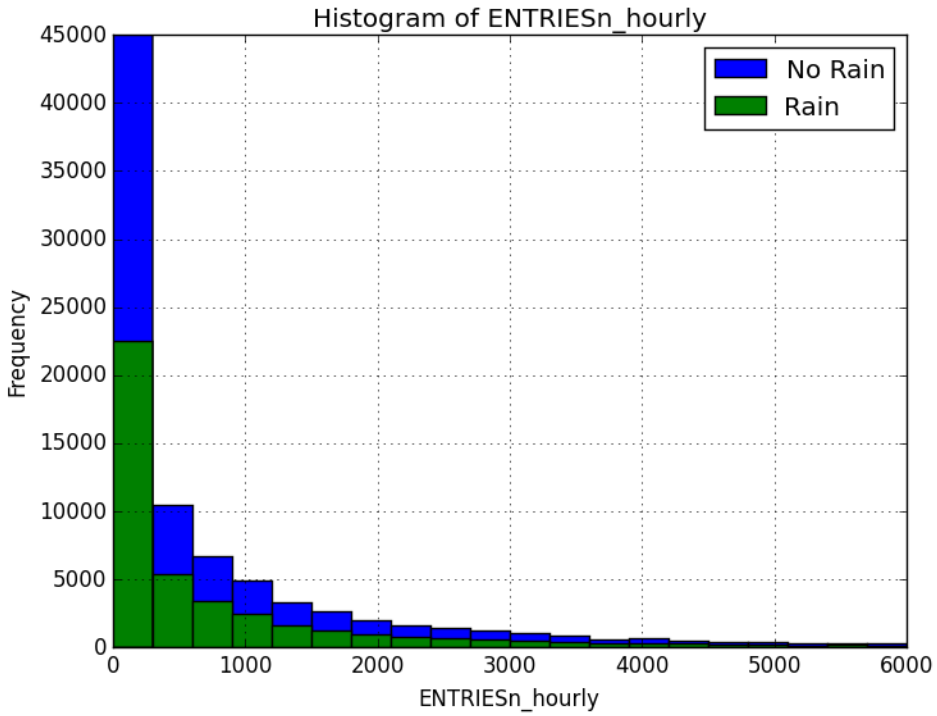
We are only capturing about 46% of the variations in our model or more specifically the explained variation. It's a low value so that probably means we aren't fitting the data well.

6- Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

No

Section3. Visualization

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 different plots.

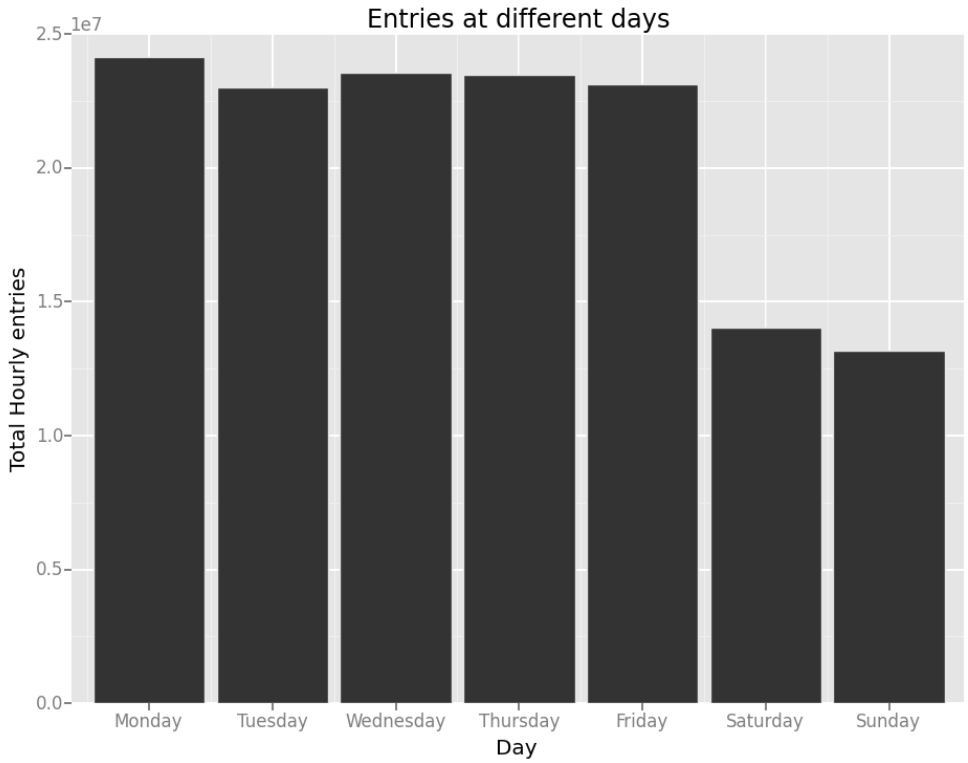


We can see that on average “No Rain” has more hourly entries than “Rain” but this is due to the difference of sample sizes. “Rain” has many fewer samples. This plot is shown just to compare the distributions.

2- One visualization can be more freeform, some suggestions are:

1. Ridership by time-of-day or day-of-week
2. Which stations have more exits or entries at different times of day

The figure below is plotted against the complete data set not just 1/3 of the data (I ran it locally on my machine on the whole dataset).



We can see that most weekdays have roughly the same amount of total hourly entries and in weekends its much less.

Section4. Conclusion

1. *From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining versus when it is not raining?*

   Yes, from the statistical test that we conducted and mentioned earlier in this report we can safely say that on average people ride the subway more when it's raining than when it's not.

2- *What analyses lead you to this conclusion?*

   The non-parametric Mann-Whitney U test led me to this conclusion; I was hoping to strengthen this conclusion even more by the aid of the data visualization (the histogram figure) but the sample sizes of the two distributions differed greatly.

Section5. Reflection

1. *Please discuss potential shortcomings of the data set and the methods of your analysis.*

   Regarding the data set, we needed a bit of cleaning first to make the data tidy and easy to use, and I think it contained just enough variables for us to answer the question of this study(What's the effect of rain on NYC subway ridership?). Another potential shortcoming of the data is that it is not well suited for doing linear regression. (Check the regression diagnostics section for more info)

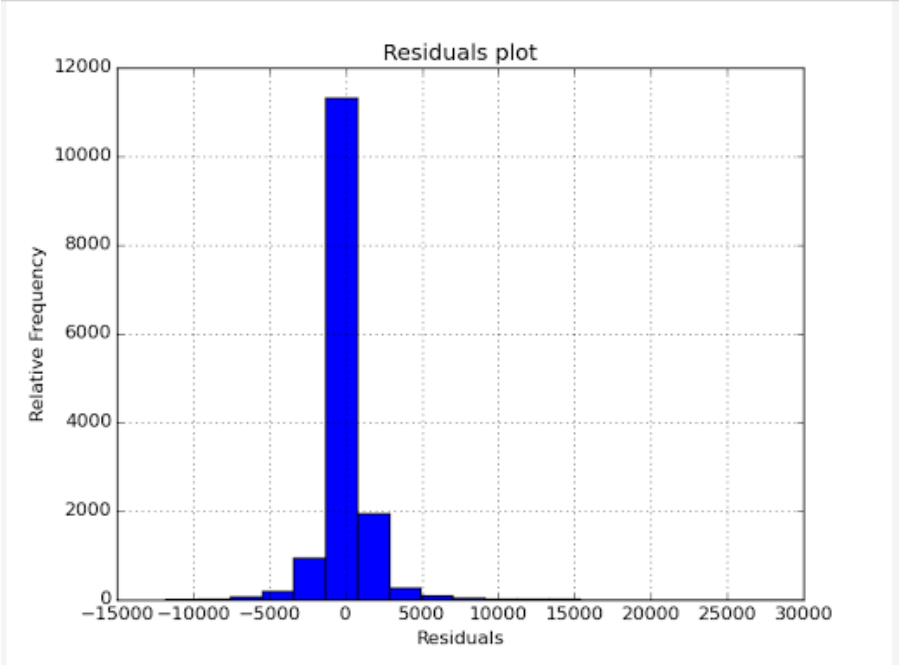
   Regarding the analysis, I guess we succeeded in answering the question we asked but we can definitely extend our analysis to answer a more generalized question like "What are the factors that determine or affect the hourly entries of the NYC subway?", Also linear regression did a rather poor job of predicting entries hourly given the data we had so a more sophisticated technique might be required to do a better job in prediction of future data.

Regression Diagnosis

I mentioned earlier in Section 5 that the data is not perfectly suited for linear regression and here is the reasoning behind this. To verify that our data fits for doing linear regression we need to take a look at 3 plots:

1. Residuals distribution

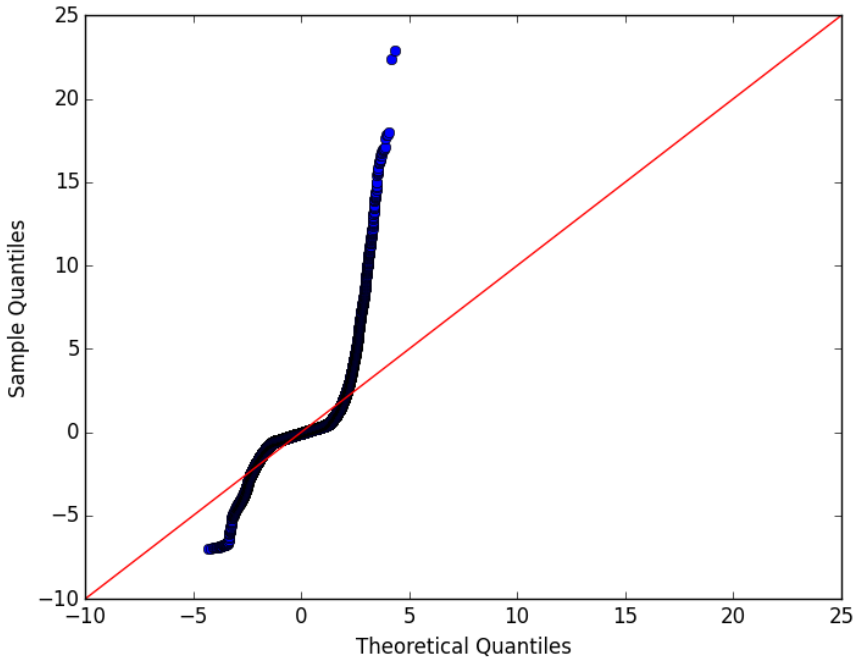
We are looking for normal distribution for residuals, so if the data was perfect for linear regression then we should see an approximately normal distribution of residuals



While it can be argued that the distribution looks roughly like the normal distribution and there is a lot of values around (x = 0) which is a good thing, I say that the shape of the distribution looks much more squished than it should be.

2. Normal Q_Q plot

On the 45 degree line shown in red, we would hope to see that the most of the points lie on the line.

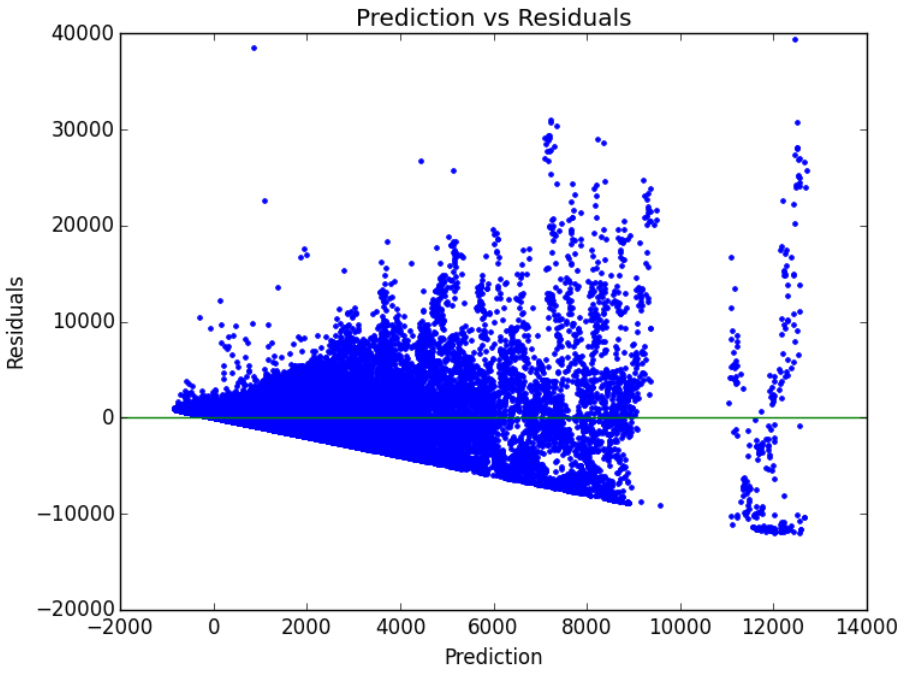


So we see that a lot of the points deviate from the perfect 45 degree line in the upper part of the plot.

I think there is no argument about this one unlike the histogram of the residuals.

3. Scatter plot of the residuals

One important feature that needs to be present in this graph is the constant variability of our error or ‘Homoscedasticity’ of our error.



Around the perfect 0-line in the plot we expect random scatter of the residuals, but instead the residuals show a cone-like scatter, or in other words the variability isn’t constant.