Sections 0. References

1. scipy.stats.mannwhitneyu
2. Mann–Whitney U test
3. What's a good value for R-squared?
4. Dummy variable (statistics)
5. Size of labels for x-axis and y-axis ggplot
6. Are the model residuals well-behaved?

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?
Ans.

- I used Mann-Whitney U-test to compare the samples on hourly entries for days with rain vs. without rain.
- I used two-tail P value.
- Null Hypothesis: The presence of rain have no effect on the average ridership in the subway.
- p-critical 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.
Ans. Mann Whitney's U-test is a non-parametric test. It's especially useful for testing the null hypothesis of two samples come from the same population against an alternative hypothesis, that one population tends to have larger value than the other. It doesn't assume that the data is drawn from any particular underlying distribution.

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.
Ans. My results are based on the values derived from the Mann Whitney U-Test in the problem set 3 NYC subway data. I got the p-value to be around 0.05. The mean of entries per hour when it's raining is around 1105.45 and mean of entries per hour when it's not raining is around 1090.28

1.4 What is the significance and interpretation of these results?

Ans. As the p-value is equal to the two-tail p-critical value 0.05, I reject the null hypothesis. I conclude that the average ridership on the days when it's raining is statistically significantly different than the average ridership on the days when it's not raining.

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:

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

Ans. Gradient descent.

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

Ans. I used rain, hour, fog, day of the week and unit as input variables in my model. I used unit and day of the week as dummy variables.

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.
Ans.

- Rain - I chose rain because, I could see from the Mann Whitney's U-test comparing ridership with and without rain that the rain has significant impact on the average ridership in the subway.
- Hour - I created a graph for average ridership per hour across all the units and I could see a clear pattern across all of the units that during the peak hours, the ridership was pretty high. So I believe that the hour of day should play a significant role in deciding the average ridership at that time on all the stations.
- Fog - a severe fog can cause the visibility issues while driving on road. So people might prefer to go in subway in such case. Also adding the Fog as predictor increased my R^2 a little bit. So I chose to include it as a minor improvement in my prediction model.
- Day of week - from exploratory analysis by plotting bar graph for average ridership by day of week, I could clearly see a strong decrease in the ridership on weekends. From that I created an intuition that the day of the week must be contributing significantly in deciding the average ridership, so I chose to include as input variable. Because it's of categorical nature, I added it as dummy variables.
- Unit - intuitively different stations will have different amount of ridership based on the locality surrounding that station. So it made sense to include it as a variable. Because it is categorical variable, I included it as a dummy variables.

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

Ans. Rain: -13.09   Hour: 439.35   Precipitation: -12.08   Fog: 43.82

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

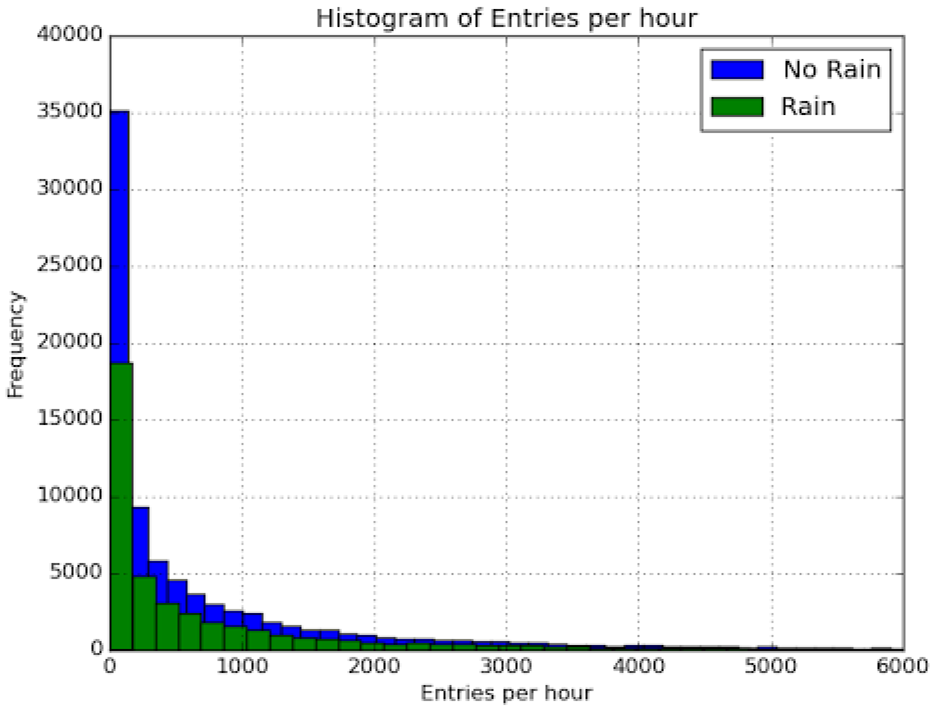
Ans. 0.47487885462

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?

Ans. R^2 represents the proportion of variance in outcomes explained by model. The value closer to 1 represents that model is able to explain most of the variance in outcome. The decision that whether this predictive model is useful depends on the context and threshold value that we set for R^2. So for e.g. if we set the 0.5 as threshold value of R^2 for model to be considered as useful then our model fails to meet that criteria. But if the threshold is 0.4 for R^2, then this model can be seen as appropriate.

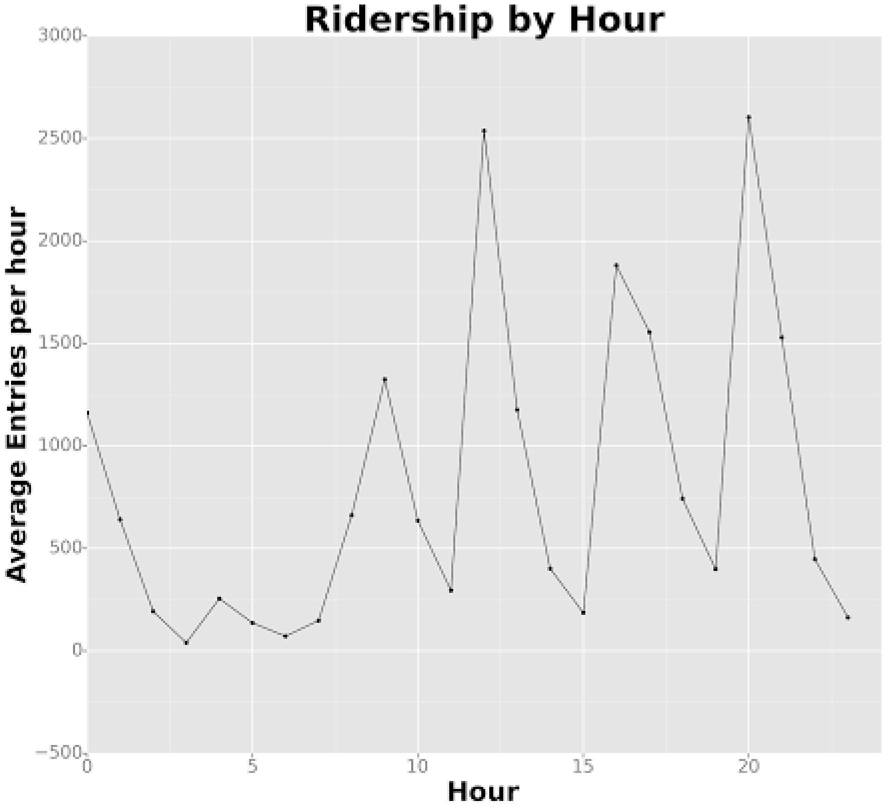
Section 3. Visualization

3.1



From the above histogram, it’s evident that there are fewer hours on rainy days when few people are entering in the subways as compared to the days when it’s not. That means on rainy days, it’s less probable that the subways are going empty.

3.2 See answer on next page



Here we can see a trend in the average ridership based on the hour of the day. We can see that 12pm and 8pm are the hours when we see the highest average ridership across all the units of subway. 9am and 4pm also see a great increase in the ridership. During the night hours there is very low ridership.

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?

Ans. From my analysis I conclude that there are significantly different number of people tend to travel in the subway on the rainy days as opposed to the days when it's not raining based on the Mann Whitney's U-Test. The negative coefficient of rain in my linear regression analysis suggests that while taking important factors into consideration like rain, hour, fog, precipitation, unit and day of the week; the rain has a negative impact on the ridership. It means that based on the model derived from this dataset: when everything else is held constant, the presence of rain will reduce the ridership a little bit.

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.
Ans. The Mann Whitney’s U-Test that I carried out on two populations: one on rainy days and the other on non-rainy days, shows that the difference in the means of both of these populations is statistically significant. The small p value 0.025 suggests that if we assume the null hypothesis to be true i.e. the presence of rain has no impact on the ridership, it’s very unlikely that we would have come up with our sample of observations by chance.

So we must reject the null hypothesis in the favor of our alternative hypothesis with two-tail p-value, which says that the average ridership is significantly different on rainy days than on non-rainy days.

My linear regression analysis of ENTRIESn_hourly includes ‘rain’ as one of the input variable and I received a negative theta coefficient for this input variable after performing gradient descent. This indicates that the presence of rain contributes negatively to the ENTRIESn_hourly which is an indicator of ridership during that hour. But considering the fact that the R^2 of my linear regression is less than 0.5 and the sequence dependent patterns that I could observe in the error residuals of model, I would take any conclusions from the linear regression analysis with a pinch of salt.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

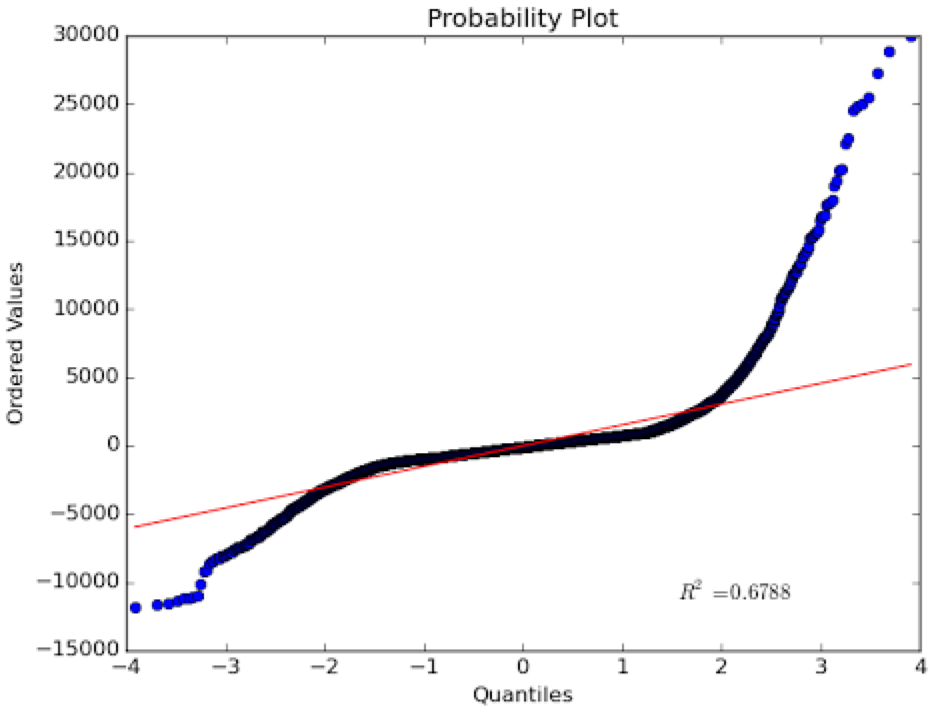
Dataset,

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

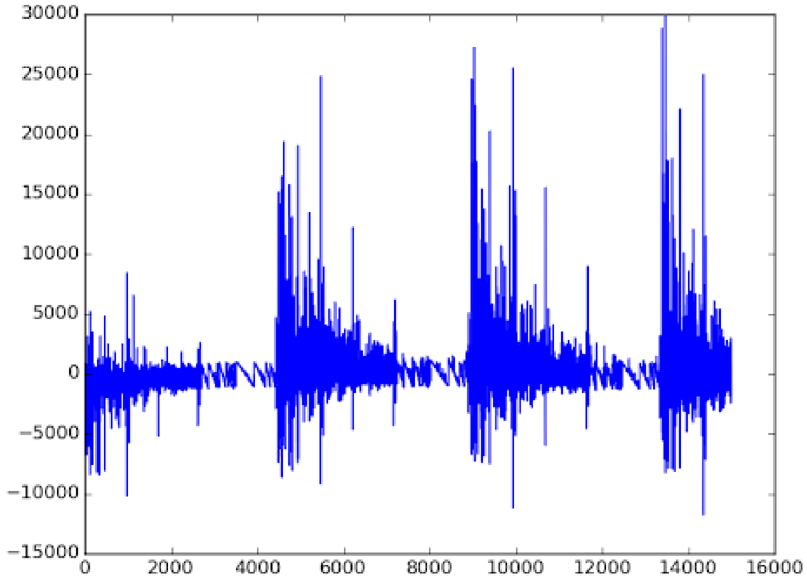
Ans. I believe that the fact that whether a particular day is a public holiday or not in the city would have significantly helped in determining the ridership of the subway per hour. Therefore, the addition of that information is desirable. This can be achieved by looking up the public dataset of public holidays in the New York city and adding a column for that information in our dataset.

My gradient descent based linear regression model has R^2 of 0.47487. I consider it little bit low to rely confidently on any conclusion made from this regression analysis. I would have preferred R^2 to be at least greater than 0.5.

The probability graph of the residuals from linear regression model is shown in the graph below, which has a clear ‘S’ pattern on the right end, which indicates that the residuals are not distributed normally in the extreme quantiles.



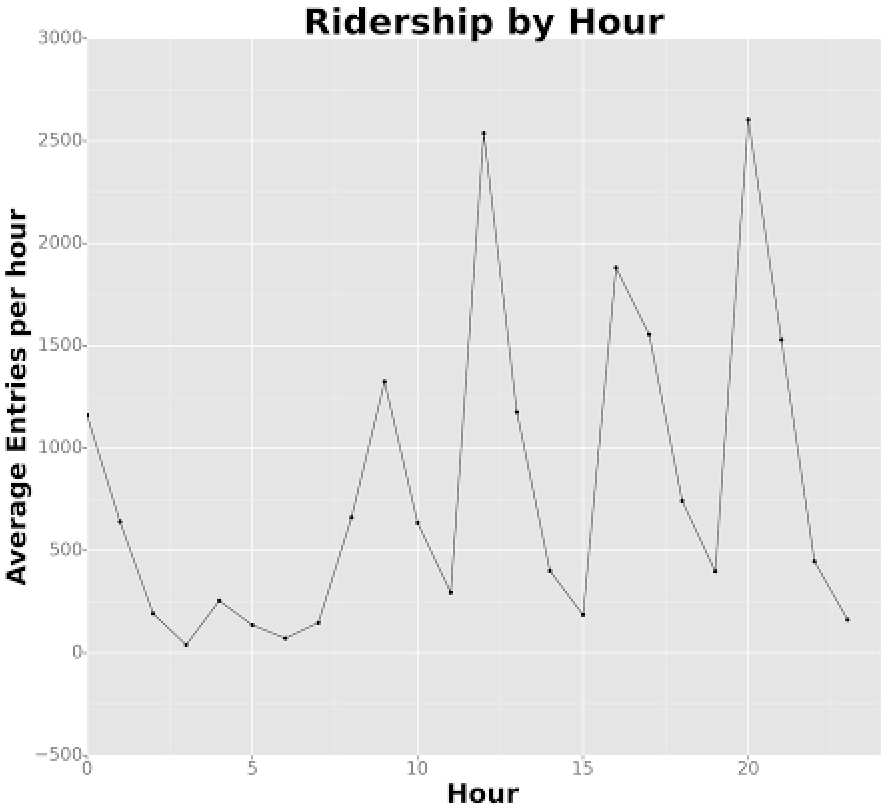
I also plotted the residuals by data points and observed some patterns as shown below, which indicates that there is a dependence of the error size on the order in which the samples were collected in our dataset. This makes our model less appropriate to derive conclusions from. We may need to add some additional variable in our model to eliminate this pattern from the residuals.



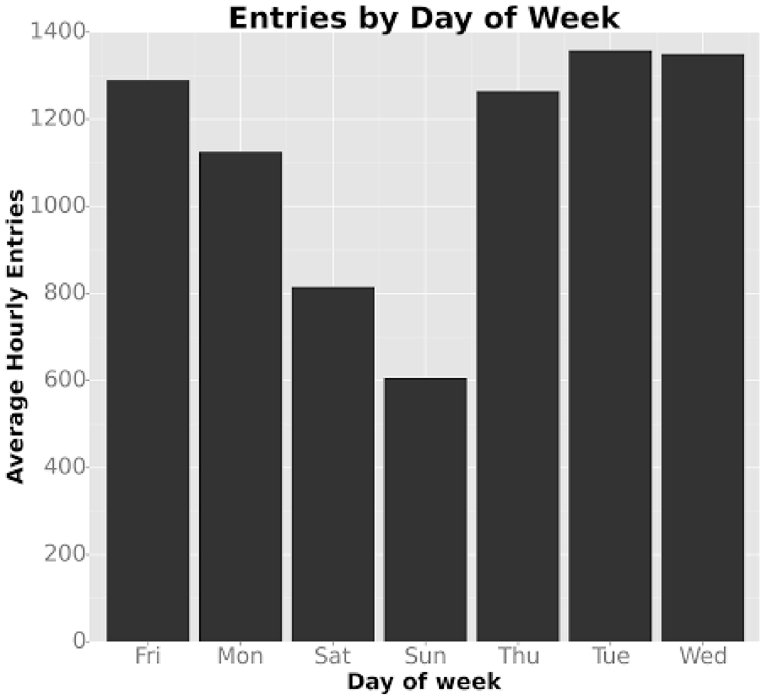
5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?

Ans. I could observe the pattern in average ridership based on the hour of day. One can observe high amount of average hourly entries during the peak hours of the day in the New York Subway.

The graph below plots this observation.



One more observation is that average ridership is rather low on weekends as compared to the workdays as evidenced in this graph below


Entries by Day of Week
Average Hourly Entries
1400
1200
1000
800
600
400
200
0
Fri
Mon
Sat
Sun
Thu
Tue
Wed
Day of week