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

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 one-tail P value.
- Null Hypothesis: The average hourly entries on the days when it's raining tends to be equal to the average hourly entries when it's not raining.
- 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 0.025. 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 less than the one-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 higher 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 and unit as input variables in my model. I used unit as dummy variable.

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.
- 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 variable.

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

Ans. Rain: 2.31   Hour: 438.96   Fog: 46.48

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

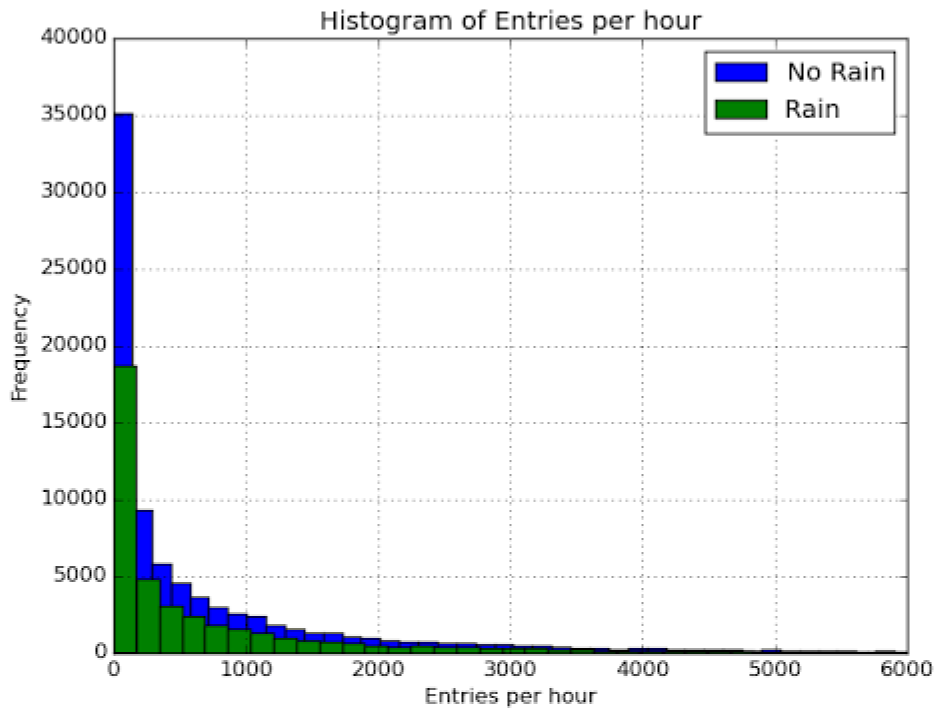
Ans. 0.462275

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.

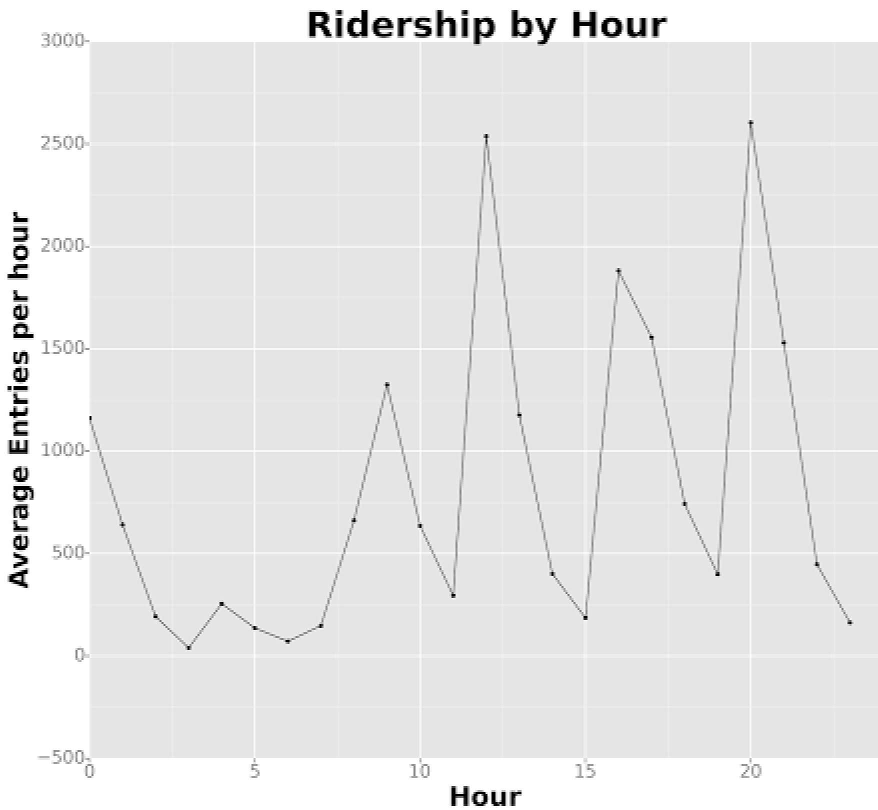
Section 3. Visualization

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 more number of people tend to travel in the subway on the rainy days as opposed to the days when it's 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.
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 average ridership on rainy vs. non rainy

days is almost the same where both the populations share the same mean, 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 one-tail p-value, which says that the average ridership is significantly higher 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 positive theta coefficient for this input variable after performing gradient descent. This indicates that the presence of rain contributes positively to the ENTRIESn_hourly which is an indicator of ridership during that hour.

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.

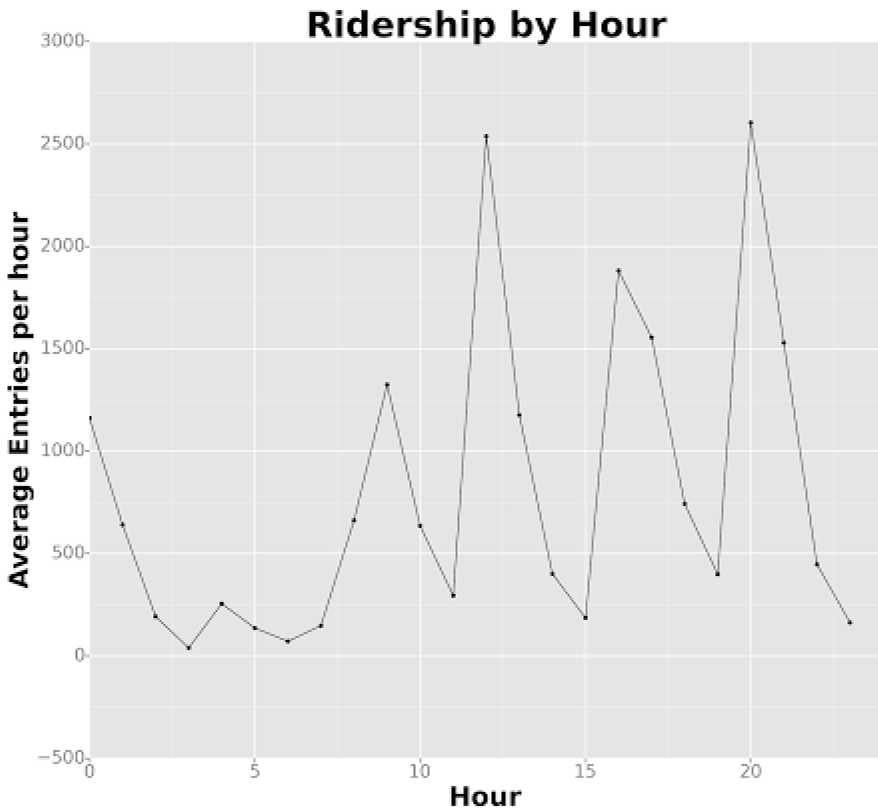
Ans. I believe that the fact that whether a particular day is a public holiday or not in the city would have significantly helped in determining the ridership of the subway per hour. Therefore, the addition of that information is desirable. This can be achieved by looking up the public dataset of public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

The coefficients derived from the linear regression analysis has a low value of R^2. A higher value of R^2 would have increased the reliability of our conclusion that a positive coefficient of the rain as input variable indicates that the presence of rains contributes positively to the ridership in New York Subway.

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