Analyzing the NYC Subway Dataset

Atef SHAAR

Section 0. References

I have used references mentioned in the course videos

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?

Because the two groups are not normally distributed I used Mann-Whitney U test to analyze the NYC subway data, I used two tail P value, and the null hypothesis is that the two distribution are identical, the p-critical value is 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.

The Mann-Whitney U-test is applicable because the two samples are not normally distributed, so we cannot use Welch's t-test to check the difference between the two samples.

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 entries with rain	1105.44637
Mean of entries without rain	1090.27878
U-statistics	1924409167.0
P-value	0.0249999

1.4 What is the significance and interpretation of these results?

Based on the mean of entries of rain days vs. the mean of entries without rain, it appears that people tend to use metro more on rainy days, which makes sense.

Based on Mann-Whitney U-test, the low p-value which almost 0.05 (2 * 0.024999) and the p-critical value is 0.05 , so with 95% confidence I reject the null hypothesis and conclude that ridership with rain is different from the ridership without 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:

I used 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?

To build my model, I used rain, perception, mean of wind speed, hour and mean of temperature and I used UNIT as a 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.

I think that rain, temperature, wind speed, perception is logically a potential variable that may correlate with the ridership predictions power, when I tried several combination of variables I have found the R^2 has been increased using those variables.

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

rain	0.0108199545849
percipi	0.014760467411
meanwindspdi	0.0305520201215
hour	0.177995874366
meantempi	-0.0244638332156

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

R^2 = 0.464422514498

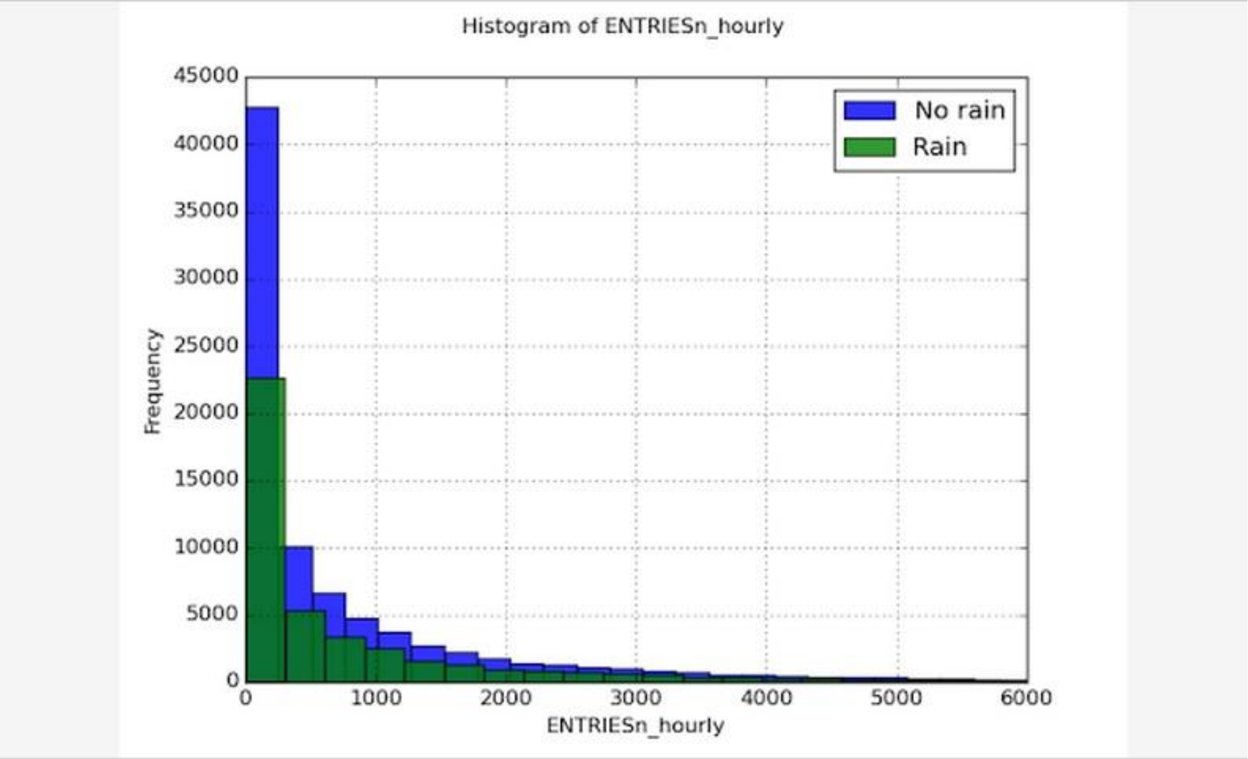
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

R^2 shows the proportion of the variance in the data that is explained by the model, so an R^2 = 0.464422514498 means we can explain about 46.4 % of the data variability with the model.

It is not very good for accuracy, but based on the available variables in the dataset (and without using Exits per hour because it does not make sense to use it (as it reflects the same action of entries per hour)), I think with an accuracy of 46% cannot be reliable for an effective decision.

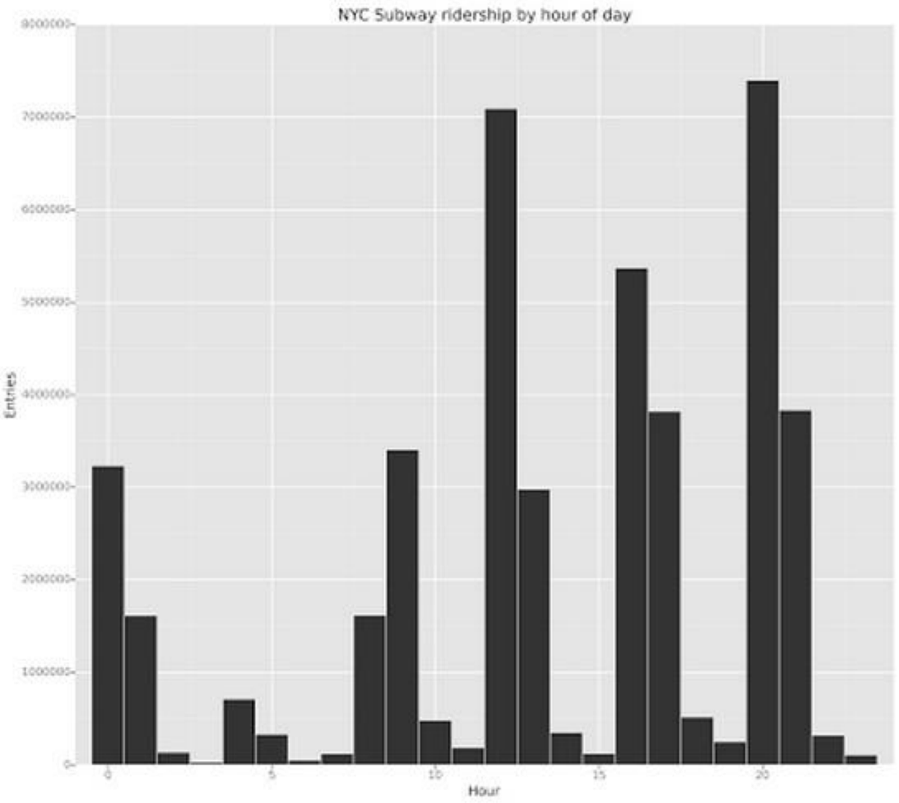
Section 3. Visualization

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



Based on this visualization, we can see that the ridership in rainy days and ridership in non-rainy days are not normally distributed.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class



Based on this visualization, we can see the maximum ridership hours at 13:00 and 20:00 , and we can see the rush hours between 8-10 and between 16-18 .

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?

More people ride the NYC subway when it is raining , but it's a small difference .

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.

Based on Mann-Whitney U test , i concluded there is a difference between the two samples or ridership in rainy vs non-rainy , and looking and the mean and the median of those distribution I can see that the ridership are more in the rainy days even if it is a small difference .

Also the coefficient of rain variable is low, so the effect or the relationship between ridership and rain are questionable.

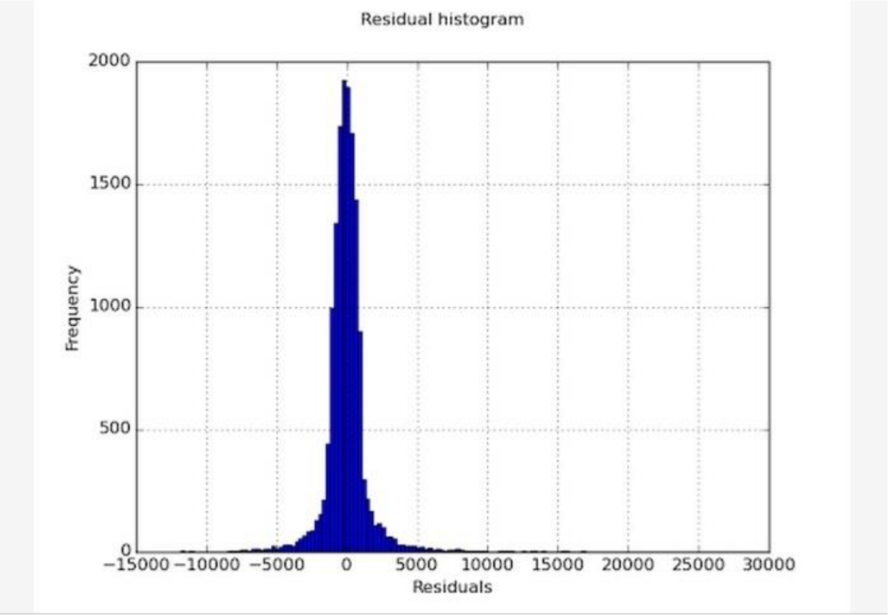
Section 5. Reflection

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

One think that I want to mention is that the methods I used for the analysis are based on the course material only, actually I was applying what I have learned , so I think I need to read more about other methods so I can be more confident and accurate in choosing the suitable method based on the problem .

Also linear regression method used to find out the relation between dependent and independent variables, so we can predict more accurately, like in the case of rainy vs. non-rainy ridership, to generalize it we are studying people using NYC metro and weather, in this case I think variables about people would be helpful and clustering could be the best method to predict the ridership in future.

By visualizing the residuals which is the difference between predictions and the actual values,



We can see from the residuals figure above that the residuals follow a cyclical pattern, and most of the residuals are between (-5000 , + 5000) , that lead us to question the linearity is the data , some non-linearity in the data should be addressed by designing a non-linear model .

Also the data provided in the data set is just for one month , and to build a reliable model that could predicts ridership , we need more data for about others months , a minimum for 6 months so we can define the difference (for example about weather) more accurately .