Analyzing the NYC subway dataset

By Aarthy Vallur in fulfillment of Project 2

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?

I used the 2-tailed Mann-Whitney U test. p- critical value = 0.05

Assuming X is the distribution of the "Entries per hour of the Rain group" and Y is the distribution of the "Entries per hour of the No Rain group", the Null hypothesis for this test is that, the probability of X or Y exceeding the other is the same.

Question	Verbal null hypothesis	Symbolic null hypothesis
Does the NYC subway ridership change on days with no rain compared to rainy days	The probability of the distribution of entries per hour on rainy days exceeding that of non- rainy days is the same as that of the probability of the distribution on non- rainy days exceeding that on rainy days.	$H_0 = P(X) = P(Y)$

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 distribution of this data set cannot be categorized as normal or otherwise. The Mann Whitney U test is a non-parametric test that makes no assumptions about the nature of the distribution. It is more robust than a t- test on large sample sizes, of which we cannot assume normality. Also, it is less likely to be affected by outliers. Hence it is applicable to this question and 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.

Mann Whitney U test:

Output: 1105.44637675 1090.27878015
MannwhitneyuResult(statistic=1924409167.0, pvalue=0.024999912793489721)

$u_1 = 1105.446$ and $u_2 = 1090.279$.
Mann-Whitney U statistic = 1924409167.0
One-tailed p- value = 0.025

1.4 What is the significance and interpretation of these results?

The calculated 2-tailed p value of 0.025*2 (0.05) is no different than the critical p- value= 0.05. Hence the difference in the distribution of the 2 groups is not statistically significant. We should keep the null hypothesis and conclude that the ridership of the NYC subway is not significantly affected by 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 OLS using statsmodels

1. OLS using Statsmodels or Scikit Learn
2. Gradient descent using Scikit Learn
3. Or something different?

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
I used rain, minimum temperature, hour and fog. Unit was used as the 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.

- 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 used rain based on intuition and the fact that though the difference is not significant, when considered along with outher features, it may prove important. I used hour based on intuition that, ridership is not a constant on all days of the year or at all hours of the day. It could be more during weekdays or during holidays and will be more before and after typical office hours. I used "fog" and minimum temperature based on both prediction that on days that are foggy or cold, people with prefer to use the subway as well as data exploration. "Fog" and "mintempi" improved my R^2 value compared with my earlier variable, mean temp as well as maximum temperature.

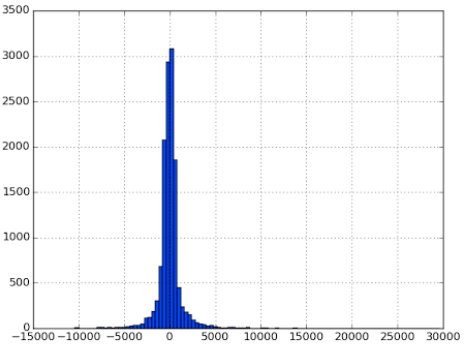
2.4 What are the parameters (also known as "coefficients" or "weights") of the non-dummy features in your linear regression model?

```
rain          -4.261471
mintempi      -15.011470
Hour           65.336815
fog           191.695634
```

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

```
r^2 value is 0.480299592504
```

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?
This R^2 value is greater than 0.4 and is therefore a pretty good fit for the regression model chosen. This is a good basic model to predict subway ridership. However I will prefer to also do a gradient descent model to develop a model to look at the same features. Plotting the residuals gave a normally distributed histogram

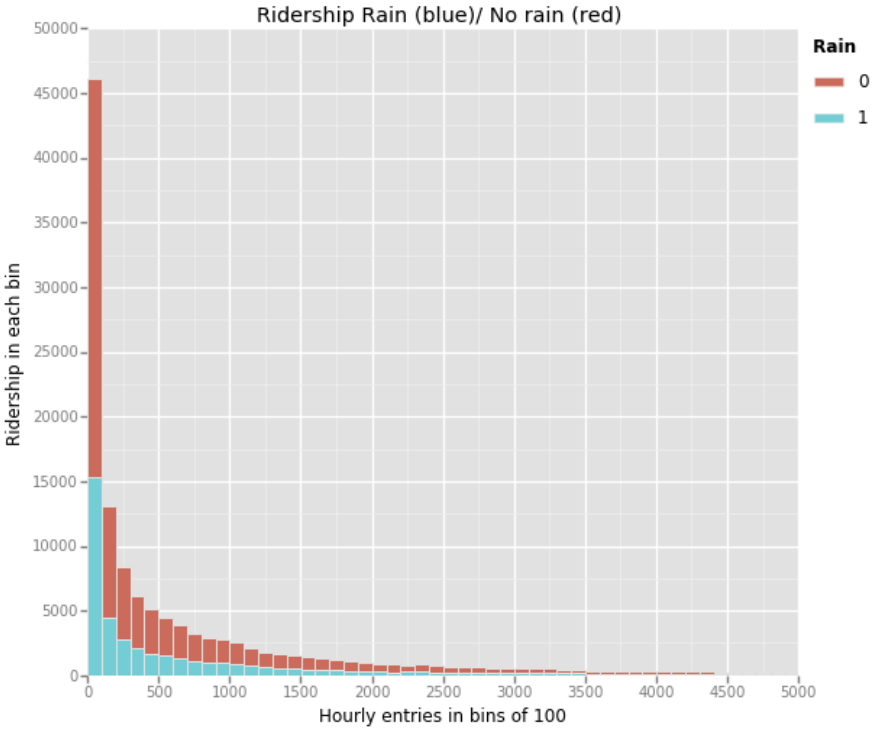


A normal distribution of variance is indicated by the symmetric histogram, attesting to the validity of the models' underlying assumptions.

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

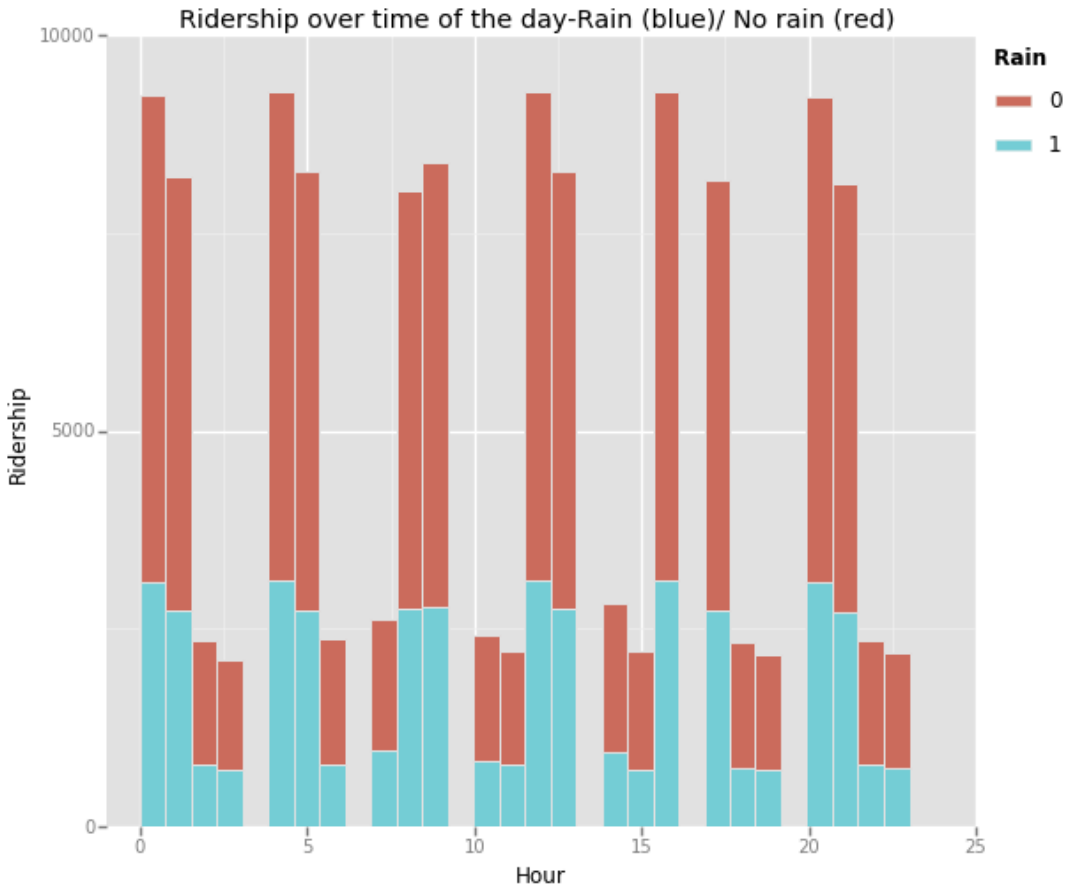
- *You can combine the two histograms in a single plot or you can use two separate plots.*
- *If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.*
- *For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that fls in this interval. Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.*



The above visualization is a stacked histogram showing the number of riders in each bin of 100 hourly entries. Red bars represent the “No rain” days and blue, the “rainy” days. The height of the bars represents the number of riders recorded in each bin. From the graph, entries in each bin on the “No rain” days are more than the “Rain” days. This is true across the X axis in any bin.

3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

- Ridership by time-of-day
- Ridership by day-of-week



The above visualization is a stacked histogram that represents ridership over the hour of the day during days with No rain (Red bars) and Rain (Blue bars). The graph shows higher ridership during certain hours of the day. Irrespective of the hours, ridership is higher when there is no rain.

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
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?
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.
From my analysis, ridership in the NYC subway on days with rain compared to days with out rain is not different. The Mann Whitney U test did not show a statistically significant mean hourly ridership for the "rain" group compared to the "no rain" group. But the visualization seems to suggest that when binned by hour of the day, ridership is higher when counted as entries per hour on days with no rain. It is possible that rain is not a direct determinant of ridership, but when considered with other factors, could play a role.
In the linear regression model, the R^2 value did not change with the exclusion of rain, as detailed below
When features = dataframe[['rain', 'mintempi', 'Hour', 'fog']], r^2 value is 0.480299592504
When features = dataframe[['mintempi', 'Hour', 'fog']], r^2 value is 0.480299017023
In fact, in the absence of 'Hour', the R^2 value drops to 0.444333719541. So, it is likely that hour of the day is a more predictive feature for NYC subway ridership than rain, fog or min temp.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.
5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. *Dataset,*
2. *Analysis, such as the linear regression model or statistical test.*

Visualization does not represent the conclusion of the statistical analysis satisfactorily. For instance, when binned by number of hourly entries, ridership for the days without rain seems to be higher than on those with rain (histograms). This may mean that our reliance on one feature alone for the statistical test may not be fully relevant. Either way, more detailed and diverse analysis that does not just use statistics is needed to really have meaningful conclusions. That may come from more data depth, exploring analysis tools, using more modeling/ machine learning in addition to just data analysis using a statistical test.

The dataset is extensive yet not detailed. For example e, the relationship between rain and ridership maybe more dependent on amount and timing of the rain (hour of the day when it rains) rather than rain or no rain. It is possible to presume from both the linear regression and the visualization that time of the day could influence ridership. We do not have those details. In the absence of details, just assuming that rain affects ridership is not very conclusive. It is possible that the significance observed by the statistical analysis is exaggerated by the absence of more discerning features.

A gradient descent model for linear regression would be a great addition to the analysis, it will potentially let us add more features and use the data set more fully. More features and details will also improve the regression model and give a more accurate means to predict the effects of rain on subway ridership.

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

As is clear from the linear regression model and the R^2 values, 'Hour' could be the most reliable predictor of ridership. Weather related features such as rain, fog or temperature are secondary influencers. It makes more sense to consider them based on the 'Hour' parameter rather than by themselves to evolve a trustable model for NYC subway ridership.

References

1. http://www.statisticssolutions.com/
2. https://georgemdallas.wordpress.com/2013/10/30/principal-component-analysis-4-dummies-eigenvectors-eigenvalues-and-dimension-reduction/
3. http://www.itl.nist.gov/div898/handbook/pri/section2/pri24.htm
4. http://www.originlab.com/doc/Origin-Help/Residual-Plot-Analysis
5. http://www.cookbook-r.com/Graphs/
6. https://docs.scipy.org
7. https://www.unistat.com/guide/nonparametric-tests-unpaired-samples/
8. http://stackoverflow.com/
9. http://projecteuclid.org/euclid.ss/1009213726