P2: Analyzing the NYC Subway Dataset

Questions and Answers

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?

- Mann-Whitney U test was used.
- We are interested in the difference in the ridership with rain and without rain. So, the null hypothesis is
  H0: P(x > y) = 0.5, where x is random draws from the population of the ridership with rain, and x is random draws from the population of the ridership without rain.
  Alternative hypothesis is
  H1: P(x > y) =/= 0.5.
- A two-tail P value is used.
- My p-critical value is 0.05 (95% confidence level)

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 a non-parametric test, which is used for two populations with unknown distributions. We don’t know the distribution of ridership in rainy days and in non-rainy days. I am interested in whether one of the two distributions is more likely to generate a higher value than the other. This is the reason the non-parametric test was used for the dataset.

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 returns the Mann-Whitney statistics and one-sided p-value. To get the two-sided p-value multiply the returned p-value by 2.
- The Mann-Whitney statistics is 153635120.5.
- The p value is approximately 5.482xe-06.
- Significance level alpha is 0.05.
- the means with rain and without rain are approximately 2028.1960 and 1845.5394, respectively.

1.4 What is the significance and interpretation of these results?

- Since the p value is less than the significance level (0.05), it is statistically significant and the hypothesis is rejected. In other words, the average number of people who use the subway when it rains is not the same as that of people who do when it doesn’t 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:

OLS using Statsmodels or Scikit Learn
Gradient descent using Scikit Learn
Or something different?

- OLS using Statsmodels is used to compute the coefficients theta and produce prediction for ENTRIESn_hourly.

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

- Features include ['rain', 'fog', 'hour', 'weekday', ‘UNIT’], where ‘UNIT’ is 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.
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.”

- ‘rain’ is a key feature obviously. ‘fog’ is an indicator if there was fog at the time and location. I think it helps people to decide whether to use the subway or not.
- I used features ‘hour’ and ‘weekday’ because when I included them in my model, they drastically improved my R^2 value.
- Features ‘precipi’ and ‘meantempi’, and others were merely contributed to the improvement, while increasing more computational complexity relatively. So, I decided not to include them.

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

- [‘rain’, 'fog', 'hour', 'weekday'] = [57.936517, -596.515322, 123.712526, 973.276640]
- Intercept is -103.779608371

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

- It is 0.481783370977

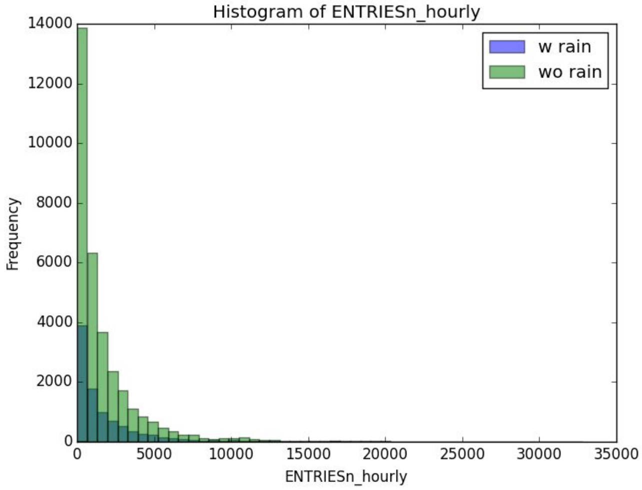
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

- R^2 is a statistical measure of how close the data are to the fitted regression line. It is the percentage of the response variable variation that is explained by a linear model. That is,
  R^2 = Explained variation / Total variation
- It is always between 0 and 1; 0 indicates the model explains none of the variability of the response data around its mean, whereas 1 indicates that the model explains all the variability of the response data around the mean.
- The higher the R^2, the better the model fits the data. In this model, it is approximately 0.4818. I think the linear model to predict ridership is appropriate for the dataset.

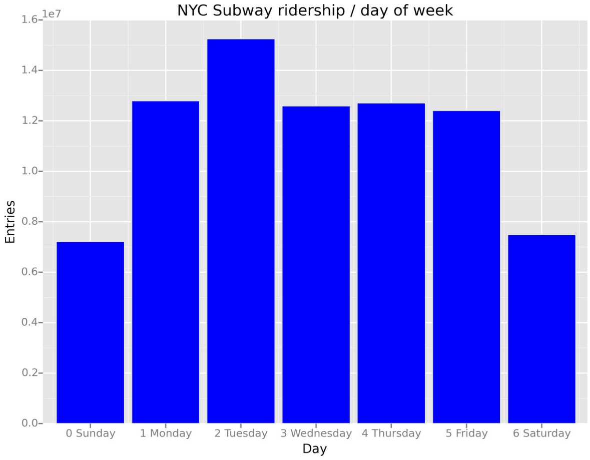
Section 3. Visualization

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 falls 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.

This figure shows the relationships between riderships with rain and without rain. Intervals on the x-axis represent the volume of ridership (value of ENTRIESn_hourly), and the height of the bar for each interval on the y-axis represents the number of records that have ENTRIESn_hourly that falls in this interval.
The distribution of ENTRIESn_hourly seems not to be normally distributed and skewed to the right on both cases. there are far fewer observations on rainy days than non-rainy days.



3.2 One visualization can be more freeform.
The figure shows the NYC subway ridership by day-of-week. As can be seen, people use the subway less often in weekend than in week days.



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?

- More people ride the NYC subway when it's raining than 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.

- The Mann-Whitney U-Test shows that the population means of ridership with rain and ridership without rain are different, and the average population of ridership with rain is greater. Also, my linear regression shows that the coefficient of rain is positive, which means that when it's raining it contributes to more people riding the subway.

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:

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

- There is the discrepancy between the two datasets; the MTA turnstile entry data is produced on an hourly basis, while the weather data is produced daily. It does not reflect the change in the number of riders when the weather changes within a day.
- We have only a single month of MTA data, which cannot be represented on effects of seasonality, as ridership is affected.
- Since the Mann-Whitney U test used for this project is a non-parametric test, it is less powerful than parametric tests. In other words, it doesn't take the shape of distributions into account, and if there is really a difference between two groups, it is less likely to find it.