Section 0. References

Please include a list of references you have used for this project. Please be specific - for example, instead of including a general website such as stackoverflow.com, try to include a specific topic from Stackoverflow that you have found useful.

https://github.com/yhat/ggplot/issues/33

https://chrisjmccormick.wordpress.com/2014/03/04/gradient-descent-derivation/

http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm

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?

Answer: Mann-Whitney U-Test was used to analyze the NYC subway data. I used a two-tail P value. The null hypothesis is that the distributions from both groups are identical. The p-critical value used was 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.

Answer: This statistical test is applicable because the two populations do not follow a normal distribution and have the same shape. The independent variable should consist of two independent groups for Mann Whitney U-Test to apply, which it does in this case. i.e rain and no-rain.

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.

Answer:
With Rain Mean :`1105.4463767458733`
`Without Rain Mean :1090.278780151855`
`p-value = 0.024999912793489721`
`U = 1924409167.0`

1.4 What is the significance and interpretation of these results?
The Mann Whitney test provides a p value less than 0.05, which leads to the conclusion that the populations are distinct. The distribution of number of entries is statistically different between rainy and non-rainy days.

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:

1. Gradient descent (as implemented in exercise 3.5)
2. OLS using Statsmodels
3. Or something different?

Answer: The Gradient descent approach was used.

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

Answer: rain, precipi, Hour and meantempi column data were used as features for the model. One dummy variable i.e the UNIT column data was used.

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 R2 value."

Answer: Rain and precipi was used in the model because ridership might increase during rains/snow. Hour was used in the model because ridership changes bases on the time of the day. Mean Temperature was used in the model because the ridership may increase when temperatures are low or high

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

```
Answer: [-2.41395539e+02  -1.50927305e+02   -1.51524518e+02
1.10060866e+03]
```

2.5 What is your model's R2 (coefficients of determination) value?

```
Answer : 0.463968815042
```

2.6 What does this R2 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 R2 value?

This means that only 46% of the variation in ridership can be explained by the linear relationship between ridership and rain. This is model is not a good fit and can be improved.

Section 3. Visualization

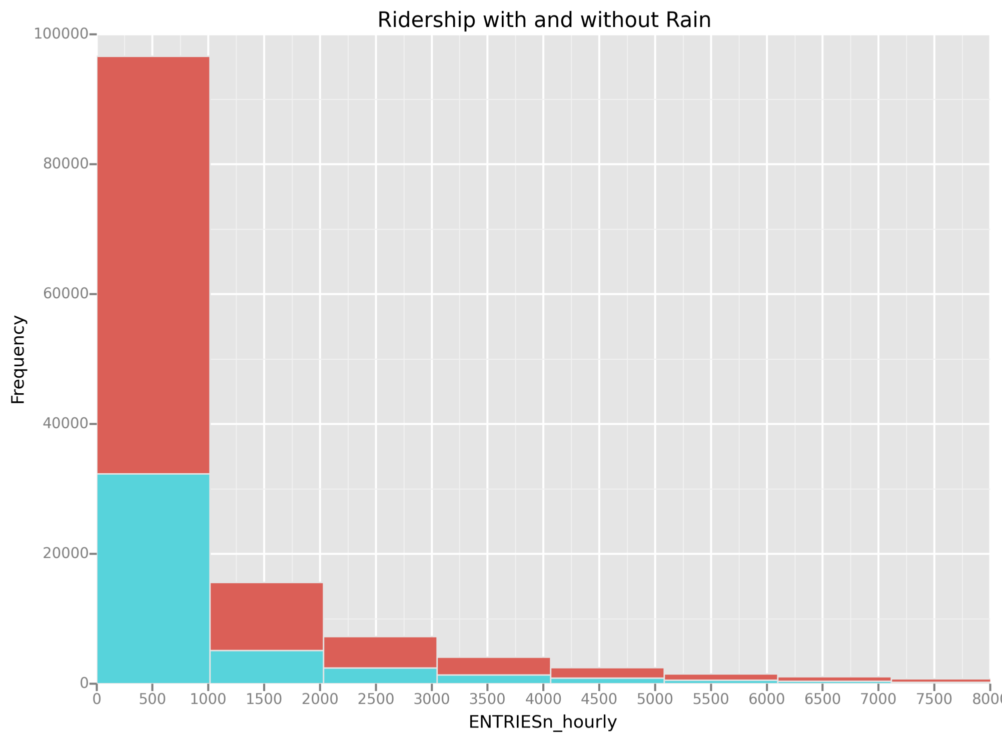
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

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.

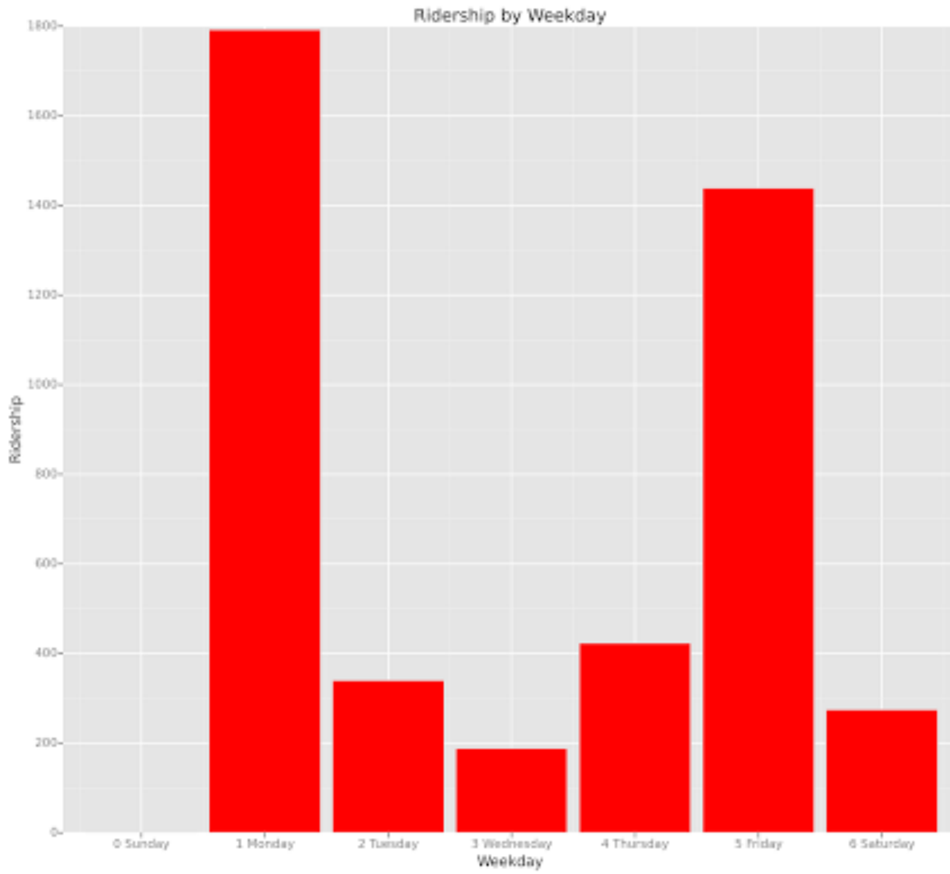
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, 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.



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



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?

The Mann Whitney test provides a p value less than 0.05, which leads to the conclusion that the populations are distinct. The distribution of number of entries is statistically different between rainy and non-rainy days. The linear regression analysis using gradient descent also shows that the rain variable contributes 46% towards the linear relation between rain and ridership. Based on this analysis we can conclude that rain does increase the ridership on the NYC subway.

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 test provides a p value less than 0.05, which leads to the conclusion that the populations are distinct. The distribution of number of entries is statistically different between rainy and non-rainy days. The linear regression analysis using gradient descent also shows that the rain variable contributes 46% towards the linear relation between rain and ridership. Based on this analysis we can conclude that rain does increase the ridership on the NYC 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:

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

Answer: R2 value of 0.46 indicates that the regression model is not a fit that is desired. The model could be improved to obtain a higher R2 value, by experimenting with other feature variables and/or using polynomial regression.

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