**Analyzing the NYC Subway Dataset**

**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 a two tailed Mann-Whitney U test to perform an analysis on NYC subway data. Intuitively, one would think that rain would increase subway riders but we are no evidence to make this hypothesis. Therefore, I used a two-tailed test with the null hypothesis that the two populations are the same (rain does not effect how many people use the subway) and alternative hypothesis that the populations are different (rain does effect how many people use the subway). This test was run using a 95% confidence or p-critical of 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 histogram of the hourly entries when it was raining and not raining was non-normal (see Figure X). Therefore, Welch’s t-test is not appropriate. I decided to use a non-parametric test (Mann-Whitney U test) because it does not assume the data was drawn from any underlying probability 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.**

Output of code I wrote is below:

Mean ridership when it is raining is and 1105.44637675

Mean ridership when it is not raining is and 1090.27878015

U stat is and 1924409167.0

p-value is and 0.0193096344138

**1.4 What is the significance and interpretation of these results?**

Based on the p-value of [0.0193], the null hypothesis is rejected and we can say there is a significant difference between subway ridership when it is raining and when it is 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:**

Using the SciKit Learn package in Python, I used a Linear Regression model with Gradient Descent to converge on the parameters that minimized the cost function and create the prediction model for hourly entries (‘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?**

**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.”**

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

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

**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?**

The R-squared value represents how much of the variation is explained by the model or more concretely the percentage of variation explained by the model. I think the model is appropriate for a quick estimate of subway ridership but I don’t think it’s the best model. It is likely that more advanced algorithms could develop a better prediction model (in terms of R-squared) but one would have to be careful to avoid over fitting the data when developing such a model.

**Section 3. Visualization**

**Section 4. Conclusion**

**Section 5. Reflection**

The main issue I noticed from the start was that the entries and exits did not lineup. This is concerning from a data integrity standpoint as you could be missing a lot of information.

The linear regression model was sufficient for the purposes of this class and being able to make a decent prediction of subway ridership. It is certainly not anywhere close to a model that could actually be used for a production type environment. By introducing a linear model, you are inherently introducing bias – meaning the model lacks the ability to account for a lot of the variation (hence the difficulty of raising the R-squared). Additionally, it is usually better to split up the available data into a training set, cross-validation set, and test set in order to create an environment where you can test the model out of sample. This would become more important when using the more complicated models.

The Mann-Whitney U test was used to understand if there was a significant difference between ridership when it’s raining versus not raining. By using this test, we are limiting ourselves in terms of what we can quantitatively say about the actual difference between the two poplulations. We provide the mean and median but we don’t have any information regarding the variance