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 statistics to test the null hypothesis that there is no relationship between entries recorded during rain and no rain. One-tail p-critical value was 0.0249.

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.

As the data exhibited a non-normal distribution, Mann-Whitney U test was applicable since it doesn't assume any inherent distribution. Else Welch's T Test would have been appropriate.

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.

with_rain_mean, without_rain_mean, U, p
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.024999912793489721)

1.4 What is the significance and interpretation of these results?

The low p-value helps reject the null hypothesis, or simply that there is a relationship between the two datasets.

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?

Gradient Descent Algorithm

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

Rain, Hour, Peak Hours, Weekday, Holiday, maxtempi, mintempi, UNITs as dummy variables

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.”

UNIT as a dummy var was the single biggest contributor to the r_squared value. Rain because people may not want to walk out in the rain. Hour because subway usage varies by hour. Peak Hours is an

added feature to identify particular hours that caused high traffic (after looking at summary of the data) like at 9pm at night – most likely people returning home from a night out. From some data visualization, it was also clear Sat and Sun had the lowest traffic so weekday became another added feature. Public holidays also meant people might travel more like on Mother's day. Max and Min temperatures simply because depending on the weather type, people may prefer to use the subway more – escape the scorching sun or the blizzard.

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

[-3.77872129e+01 2.82302188e+02 4.42195553e+02 3.00943057e+02
  6.88943979e+01 -2.04211582e+01 -2.79525457e+01]

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

~0.5

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?

I am not entirely sure if 0.5 is a proper benchmark for this dataset. Obviously the higher the better because we are comparing the predicted values against actual sample but I learned in a supply chain course that overfitting is the most common mistake made by new comers. The real goal is to predict the future and not merely fit the available dataset. Having said that, perhaps a simple line produced by the linear model isn't exactly the correct model and a better approach would be to use polynomial model as in $\Theta_0 + \Theta_1 x + \Theta_2 x^2 + \Theta_3 x^3$ or $\Theta_0 + \Theta_1 x + \Theta_2 \sqrt{x}$

Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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 I nterval (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.

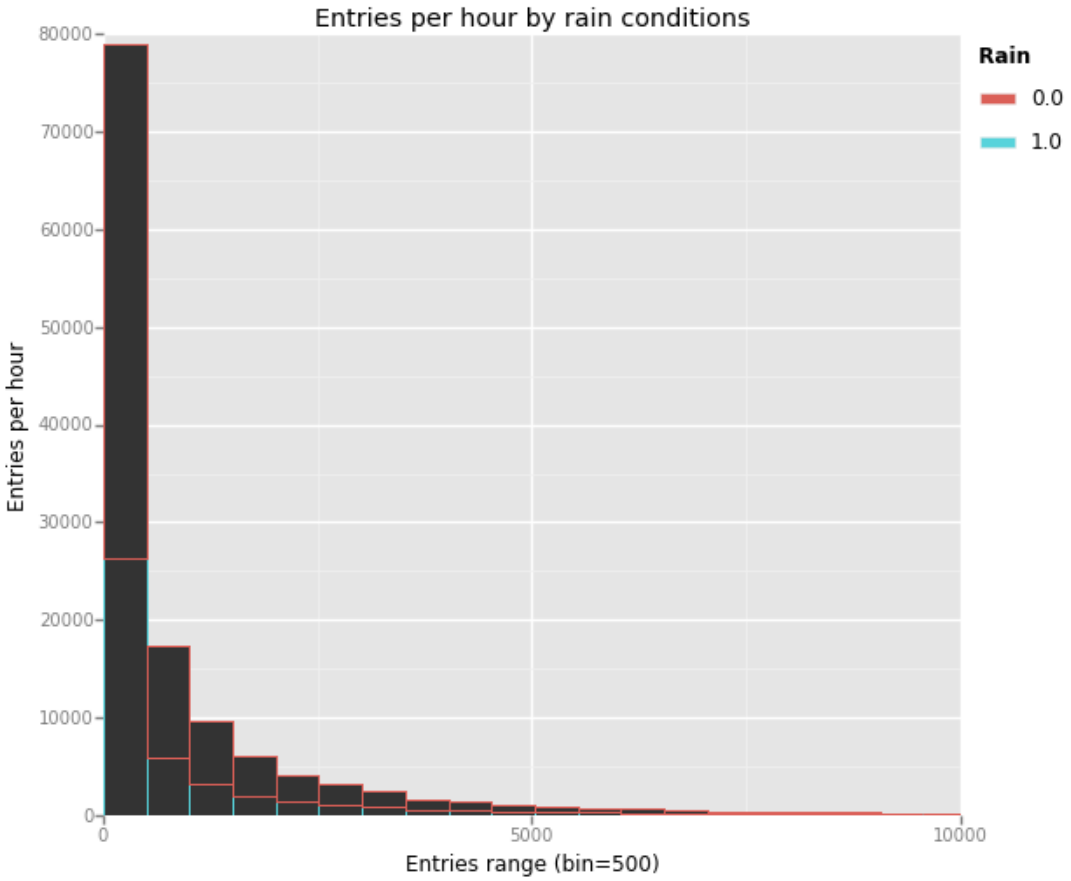

Illustration 1: Rain causes a substantial decrease in Entries

3.2 One visualization can be more freeform. Some suggestions are:
Ridership by time-of-day
Ridership by day-of-week

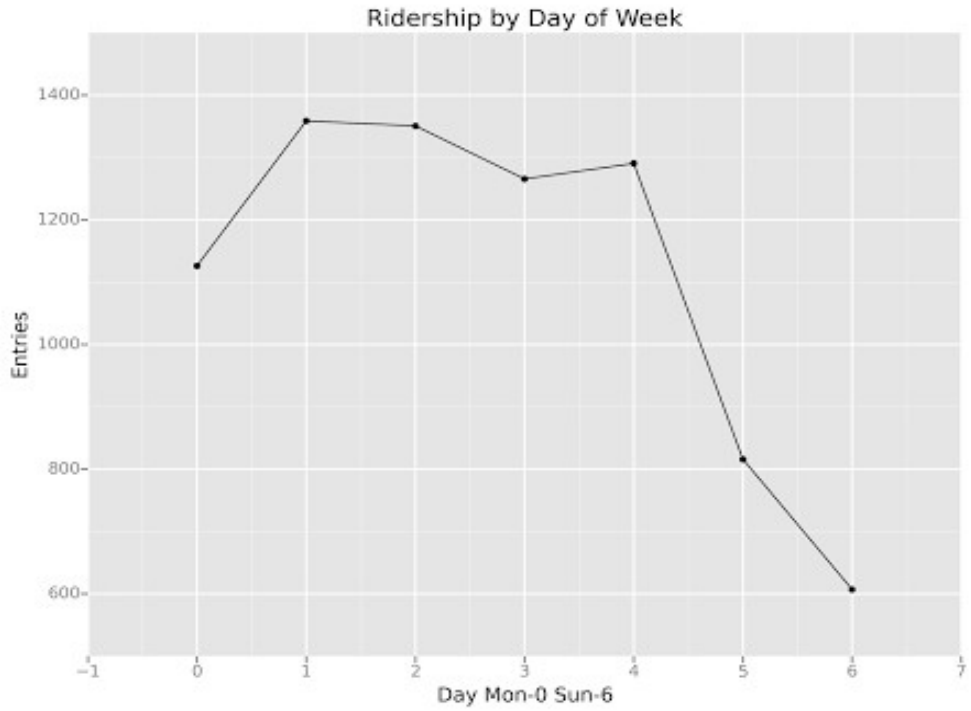

Illustration 2: Shows the average Entries for all stations by the day of the week; Sat-Sun being the lowest

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?

It is very clear from both the visualization and linear regression that rain has an impact on how many people choose to use the NYC subway. Simply stated, people use the NYC subway less when it is r aining.

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 null hypothesis was rejected because of a low p-critical value from the Mann-Whitney U test which means there is a relationship between the two datasets. The histogram of the datasets also revealed this early on but it wasn't until the actual linear regression the fact became most clear. In combination with all the selected features and valid minimization of the cost function, the theta value for 'rain' was -37.8; this would bring down the predicted entries.

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. Linear regression model,
3. Statistical test.

The dataset was not big enough and not fully representative of the NYC traffic. The feature set may not be enough either as there are many other things that could potentially affect traffic – for example planned/unplanned maintenance or breakdowns, tourist seasons, holidays, major events and many others. Also the data we currently have can be more continuous than binary – e.g measuring rain in inches instead of 0 and 1. There might be possible dependent feature sets – e.g relationship between subway stations and how if a person travels via one subway station, they will more than likely travel through some related subways with a higher likelihood. Getting a r_squared value of approximately 0.5 maybe due to the use of linear regression model and a more appropriate model could involve some polynomial regression and weighted analysis. The gradient descent algorithm might have been an overkill for the dataset provided; normal equation could have been more efficient using matrix operations which rely on highly tuned libraries. In short, better data, faster method and some subject matter (with available data) would definitely make the model much more accurate.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?
It doesn't thunder in May in NYC even though there are instances of fog and rain, so using that as a feature set would harm the analysis.