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 Mann Whitney U test with a two-tailed p-value. The null hypothesis is that the two populations, ridership on rainy days verses ridership on non-rainy days are the same. The p-critical value is 5% or 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.
Histograms of the data show that it is skewed to the lower end of the spectrum as opposed to a normal, bell curve shape. This indicates that the data is a non-normal distribution, so a Mann Whitney test is a better choice than a t-test which is applicable only to a normal distribution. The Mann Whitney does not care about the type of 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.
The one-sided p-value is .039. Rainy days have a mean of 1105.45 riders per hour while non-rainy days have a mean of 1090.28 riders per hour.

1.4 What is the significance and interpretation of these results?
Since the P-value is so small, much lower than our p-critical value, we can reject the null hypotheses in favor of the alternative that the ridership on rainy days verses ridership on non-rainy days are not the same.

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 gradient descent.

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?
Rain, fog and Hour were the features that I used. Dummy variables for the individual units were also 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.
Rain and fog are intuitive in that you would expect more people to ride the subway on days that are rainy or foggy. The Hour feature was just a random experiment that seemed to improve my R^2 value.

- 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."

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
Rain: -2.54
Fog: 26.13
Hour: 453.70

2.5 What is your model's R^2 (coefficients of determination) value?
.46

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?

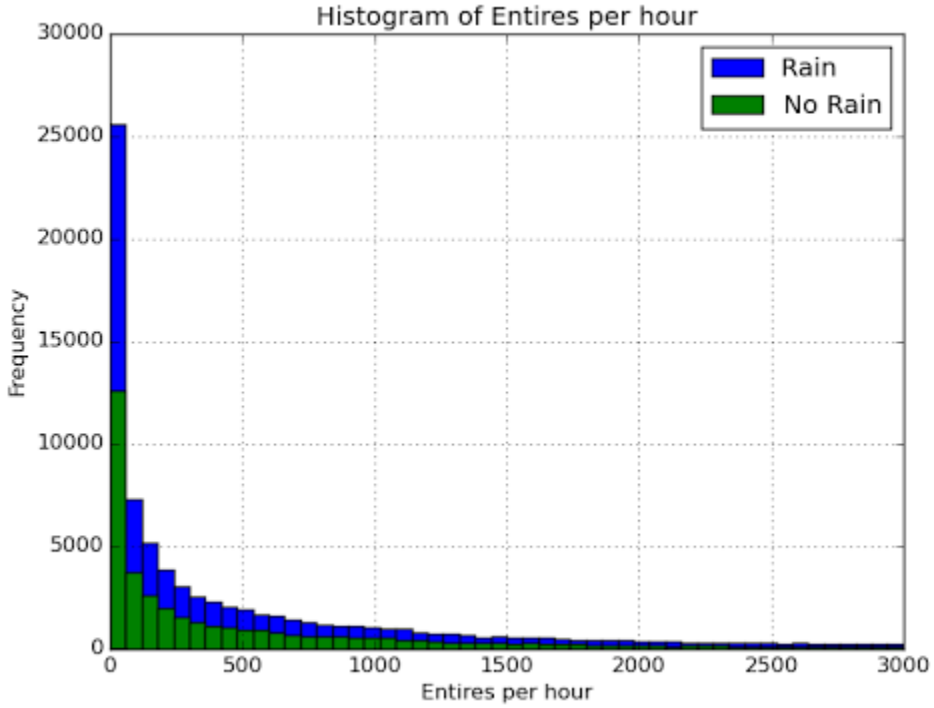
The range of and R^2 value can be from 0 to 1, zero being a poor model and one being a perfect model. With that in mind, my .46 is good, but there is room for improvement. Perhaps a different regression technique would have produced a better result. Plotting the residuals, we find that they are not randomly distributed, lead me further to believe that the model needs some improvement.

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.

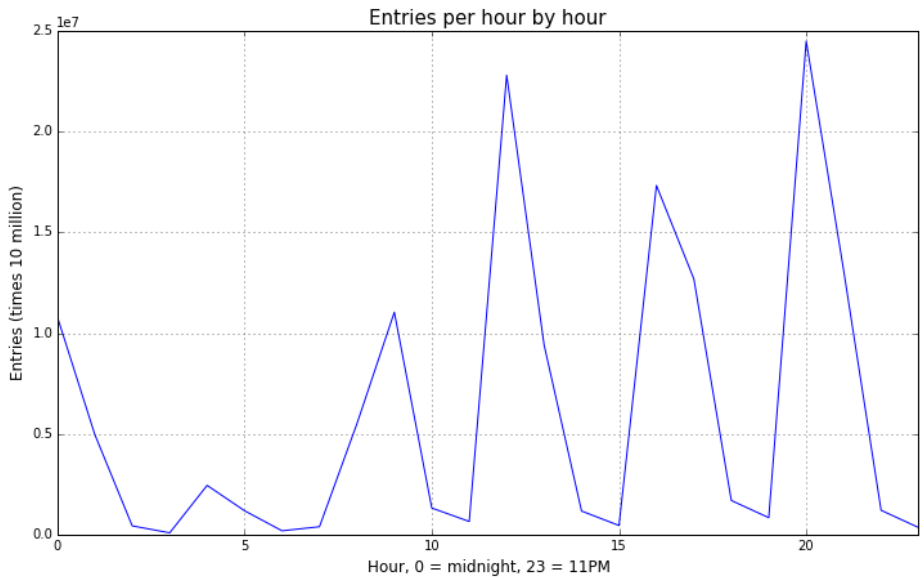


-

The histogram above makes it clear that the data set is not a normal distribution. It also shows that the sizes of the samples between rain and no rain differ.

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
-



- This chart shows the total entries for each hour in the day. You can clearly see spikes at certain periods during the day. On the surface, these likely correlate to morning, lunch and afternoon rushes. However a closer inspection of the time data shows that the only data points given are for hours 0, 4, 8, 12, 16, and 20. This is the only reason we see spikes. Higher fidelity data would be needed to show daily rush hours.

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.

My analysis shows that more people tend to ride the NYC subway on days that it is raining. While analyzing the means of the populations, I found them to be statistically different according to the Mann Whitney U test. The test gives us a p value of .039 or 3.9%. This shows us that there is only a 3.9% chance that we would find two means in the population that are the same. Since this is lower than our p-critical value of 5%, we can reject this null hypotheses in favor of the alternate hypotheses of the means are different. My gradient descent model does a fair job of predicting ridership, but rain is not the most significant input variable. Hour and fog show a higher weight in predicting ridership.

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.

The month of May might not be a good representation of the weather in NYC. Perhaps a year's worth of data might yield different results. Weather is very local. It may be raining at one subway station and not at another. The weather can also change throughout the day. The data is presented as once it rains, it rains for that entire day at that station. A more accurate set of data that shows the weather local to the individual stations and whether it is raining at a specific hour would likely show different results and perhaps yield more accurate predictions.
In regards to my analysis, gradient decent is a fairly basic and simple regression technique. It's seems that even with my decent R^2 value, the model is not a good fit for the data. After all, predicting human behavior is a very complex endeavor. Another type of regression technique may have produced a better model.