Section 1. Statistical Test

1.1
Which statistical test did you use to analyze the NYC subway data?

Mann-Whitney U-statistic

Did you use a one-tail or a two-tail P value?

Two tailed P Value: because I did no prior assumptions about the difference in the distributions of ridership on rainy and non-rainy days.

What is the null hypothesis?

Rain does not statistically affect Ridership.

What is your p-critical value?

P <=0.05

1.2
Why is this statistical test applicable to the dataset?

Both populations are non-normally distributed, and the mentioned test does not assume any particular 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.

P-Value= 0.025 (Since the result of the test is one tailed it has to be multiplied by two)

U_value= 1924409167.0

with_rain_mean = 1105.45

without_rain_mean = 1090.29

1.4
What is the significance and interpretation of these results?

Since p*2 is smaller than 0.05 the null H can be rejected, and we can conclude the distribution of the number of entries is different between days with and without 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:

Gradient descent (as implemented in exercise 3.5)

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

I used 'weekday', “rain”, "Hour" as features in my model.

And “UNIT” as categorical feature, using 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.

I decided to use the mentioned features because it seemed intuitive that weather conditions affect the ridership of the subway. For example, if its raining then people who normally walk or use the bike will use the metro. I also think that hour and day make a difference, because at certain hours or certain days people start or finish working, causing rush hours.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
Weekday= 85.5306372245
Rain= 27.6707226451
Hour= 467.19832231

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

R^2=0.46

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?

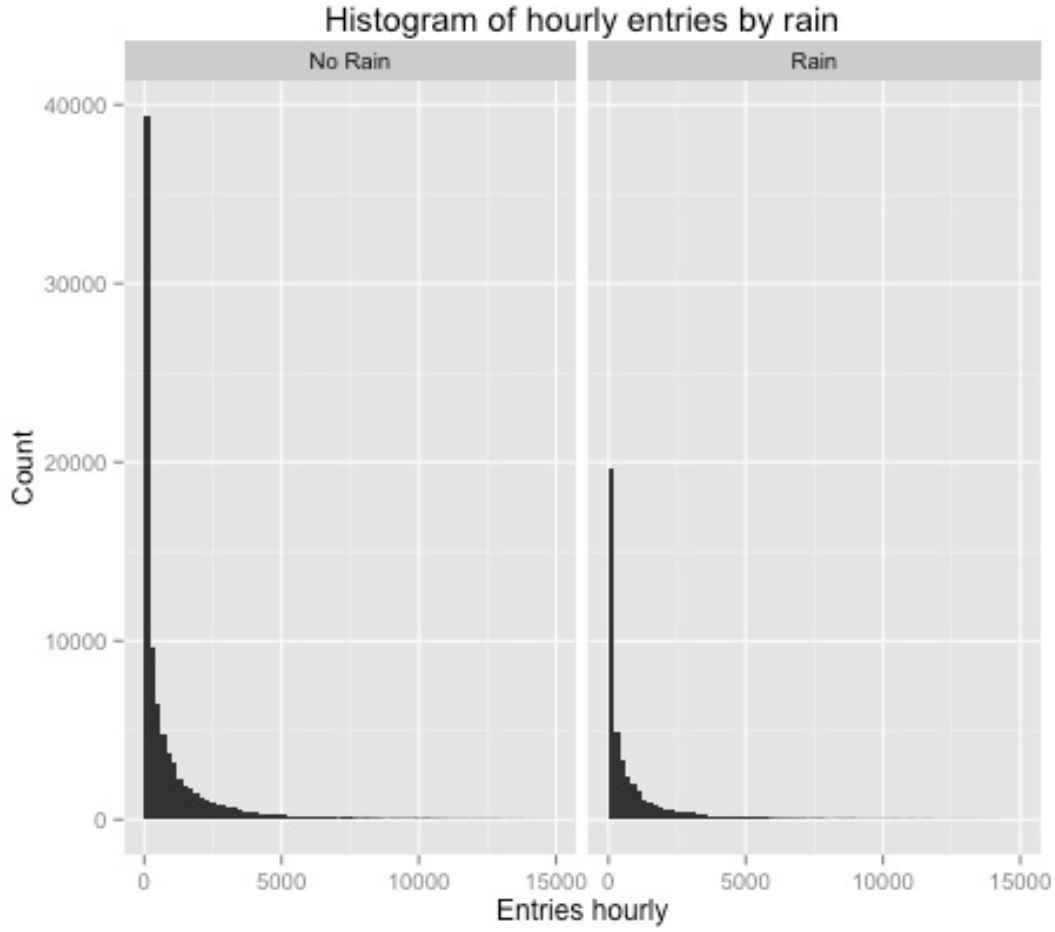
It means that our GD model identifies around 46% of the variation present in the data it was trained on. The model is not appropriate since the R2 values is too small, explaining less than 50% of the variation seems to little.

Section 3. Visualization

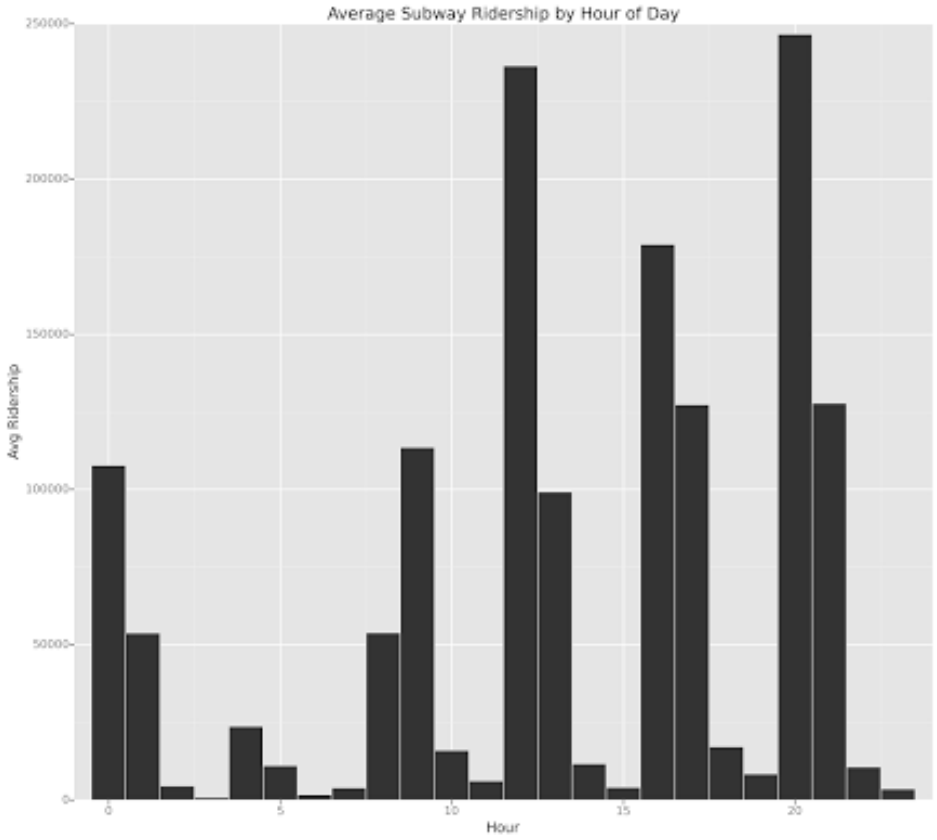
Please include two visualizations that show the relationships between two or more variables in the NYC subway data. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots, or histograms) or attempt to implement something more advanced if you'd like.

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.



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



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 subway when it is 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 mean of entries is bigger when it is raining. Although it must be said that the linear regression shows that the coefficient of determination for rain is small.

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: the data was to small to make a good analysis. It should have included a longer period of time in order to compare different seasons of the year.

Analysis: the regression we used only looks at linear relationships, however there might be different relations.

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