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-test since both data sets are not normally distributed
- Two Tail P value
- Null Hypothysis: There is no statistical evidence that more people ride the subway when it is raining versus when it is not raining.

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.

- Looking at the histogram of the data, we can see that data is non-normal. Since MWUt is a non-parametric test which does not assume any particular distribution, it is the best fit for the NYC subway data set.

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 : 1105
- Without Rain Mean : 1090
- Two tail P value : 0.04988

(Used R for this analysis since I'm getting NaN p value using python)

```
df2 <- read.csv('turnstile_data_master_with_weather.csv')
by(df2$ENTRIESn_hourly, df2$rain, summary)
wilcox.test (df2[df2$rain == "0",]$ENTRIESn_hourly, df2[df2$rain == "1",]$ENTRIESn_hourly)
```

1.4 What is the significance and interpretation of these results?

The MWUt returned a p-value of 0.04988,, so we reject the null hypothesis that both data sets are identical and have the same mean. In other words, both sample means are statistically different.

Conclusion can be drawn with 95% confidence that subway usage increases when it rains, in a statistically significant way.

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?

I used both gradient descent (GD) and OLS models run linear regression on the NYC subway dataset.

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

In the GD model the features used were: rain, presipitation (precipi), hour of the day (Hour), mean temperature (meantempi) and dummy variables for individual station (UNIT).

In the OLS model, the features used where: rain, mean temperature (meantempi) and dummy variables for stations (UNIT) and dummy variables for hours of day (Hour).

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.

Based on intuition and to maximize R ^2.

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

The R squared for the GD model is 0.461. The R squared for the OLS is 0.525.

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?

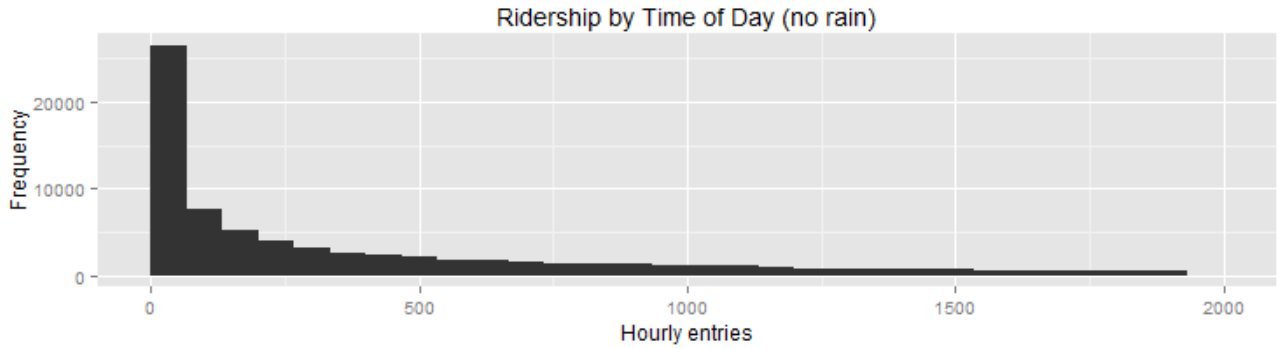
The R squared for the OLS is 0.525 which means we can explain about 52.5% of the data variability with the model.

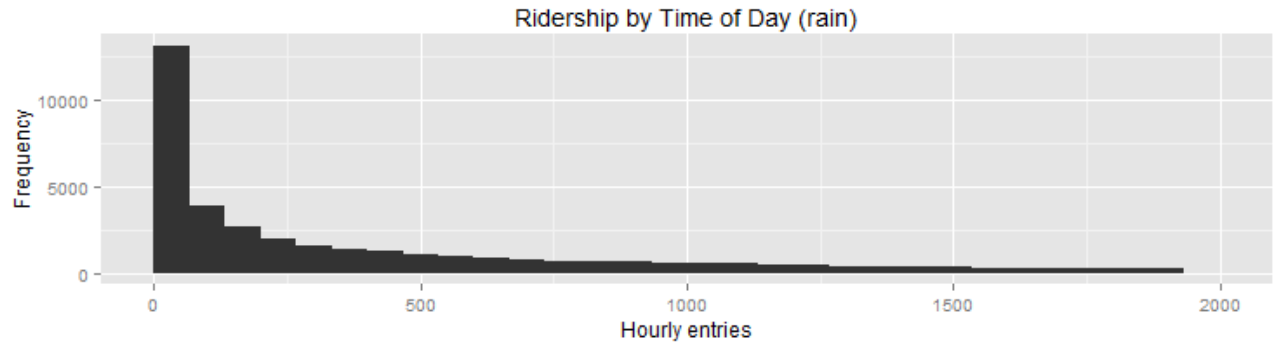
Section 3. Visualization

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.

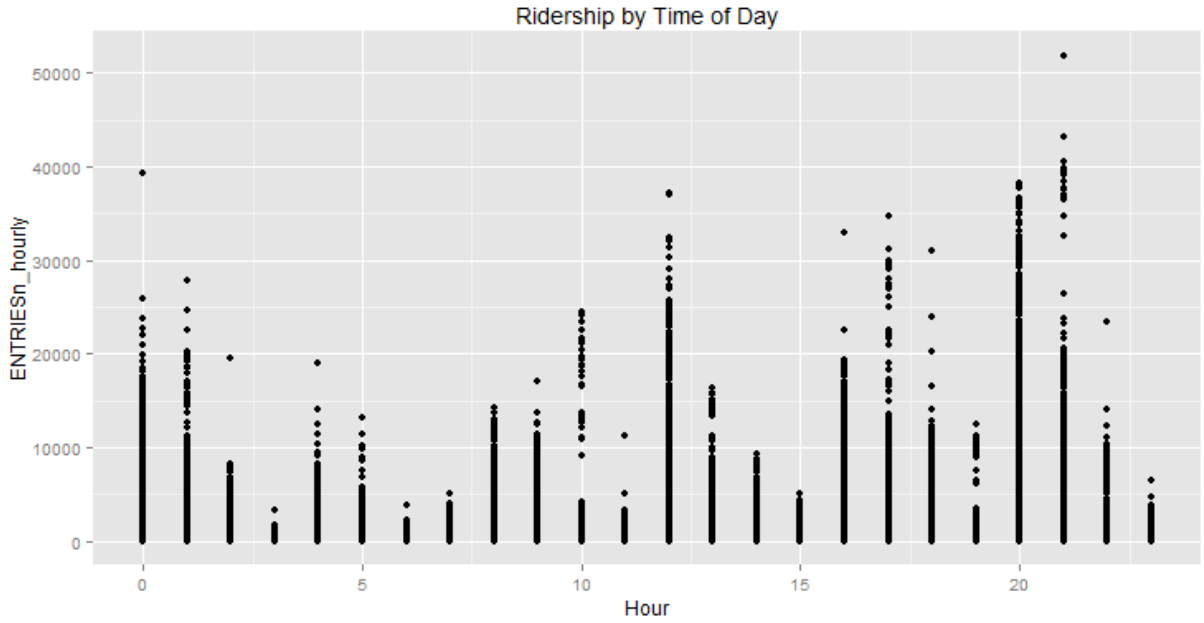
- You can combine the two histograms in a single plot or you can use two separate plots. If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms,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

On average, between 15 to104 more people ride the NYC subway on a rainy day compared to a non-rainy day. In the mean comparison, we see a difference of 15 entries per hour, while in the gradient descent model the theta for the rain variable was 104.5. Given that the rain variable is a boolean the interpretation of the theta is that when it rains (rain = 1), the model predicts on average 104.5 more people will ride the 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 comparison of both means using the Mann-Whitney U-test gives us good reason to believe that there is a statistical significant difference between the two data distributions. Other hints also show up in the Gradient Descent and OLS models, where rain feature had a positive theta of 104.5 and 54.3 respectively.

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.

One of the possible shortcomings of the dataset is that there might be events such as festivals, accidents or maintenance activities etc, impacting the ridership, which are not considered in the analysis.

Reference:

Researched number of sources on net to get right syntaxes and correct approaches. Also, used help in R

Studio.. Below are some of the resources -

- GGPlot (http://ggplot.yhathq.com/docs/index.html)
- statsmodels.regression.linear_model.OLS

  (http://statsmodels.sourceforge.net/devel/generated/statsmodels.regression.linear_model.OLS.html)