Section 1. Statistical Test

1.1 Which statistical test did you use to analyse 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?

Based on the histogram plot of the hourly entries on both rainy and non-rainy day it was clear that the dataset did not seem to have a normal distribution. Hence, a Mann-Whitney test was chosen to validate my null hypothesis. The null hypothesis was that data on rainy days was statistically different from data on non-rainy days. A two-tail P value was used with a p-critical value of 5% (95% certainty). This is a two-tailed since we are only testing for if the two data-sets (rain or non-rainy) are different and not whether one's mean will be higher than the other.

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.

Our data set contains larger than 30 data points and is also not following a normal distribution. hence a Mann-Whitney U test is 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.

The mean of hourly entries on rainy days was 1105.45 (rounded to 2 decimal places)
The mean of hourly entries on non-rainy days was 1090.28 (rounded to 2 decimal places)
p-value was 0.0249. For a 2-tailed test this needs to be multiplied by 2, hence the p-value was 0.049

1.4 What is the significance and interpretation of these results?

The results of this test show that we can say with ~95% certainty that the mean of the samples from rainy days is different from the mean of samples from non-rainy days.

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)
- OLS using Statsmodels
- Or something different?

Gradient descent and OLS both were tried.

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', 'Hour', 'meantempi' were used as the 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 R^2 value.”

I started with 'rain', ‘fog’, 'Hour', 'meantempi' as the initial features selected. These features seemed to be most likely to be causes when people would ride the subway instead of say taking their own car. The ‘Hour’ of the day would really determine how many people are out and about and that in turn would also impact ridership. After starting with this features, I tried replacing one or more of these features with others and none seemed to improve the R2 value. Taking away mean temp also does not make a large difference, but taking out any one of rain, fog or Hour does seem to make the model poorer (lower R2 value).

Hence, the final feature set chosen was 'rain', ‘fog’, 'Hour', 'meantempi'

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

>> Not sure how to get this.

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

R2 Value is 0.4644

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 value of R2 means that 46% of the time the model is able to predict the variability in the training data. The closer the value of R2 is to 1, the better the prediction of the model is. So, this value says this probably is an average model. Now, depending on the application of the predictions made by a model and the cost of collecting data for additional features or creating a model with more complexity, this may be good enough.

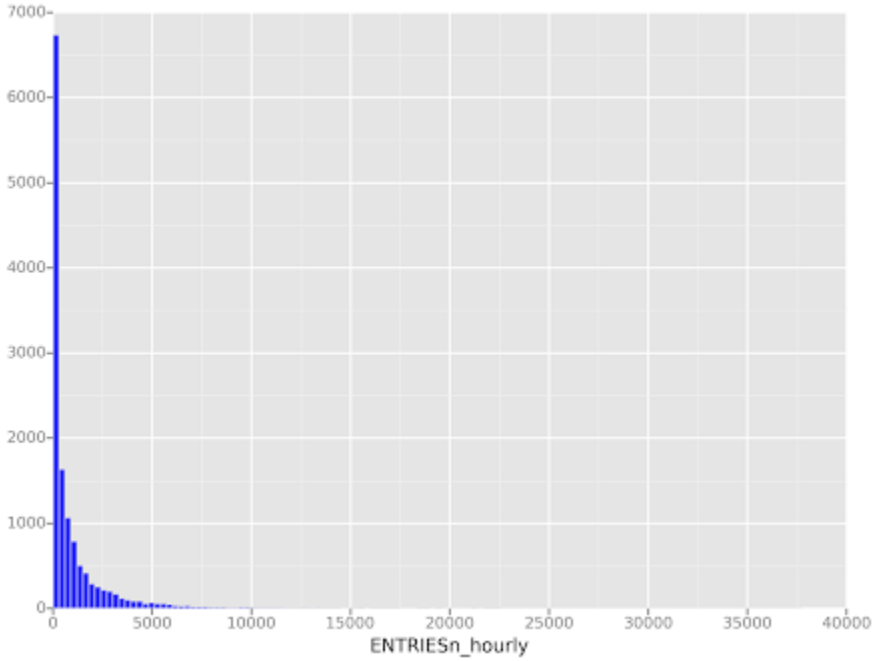
Section 3. Visualization

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.

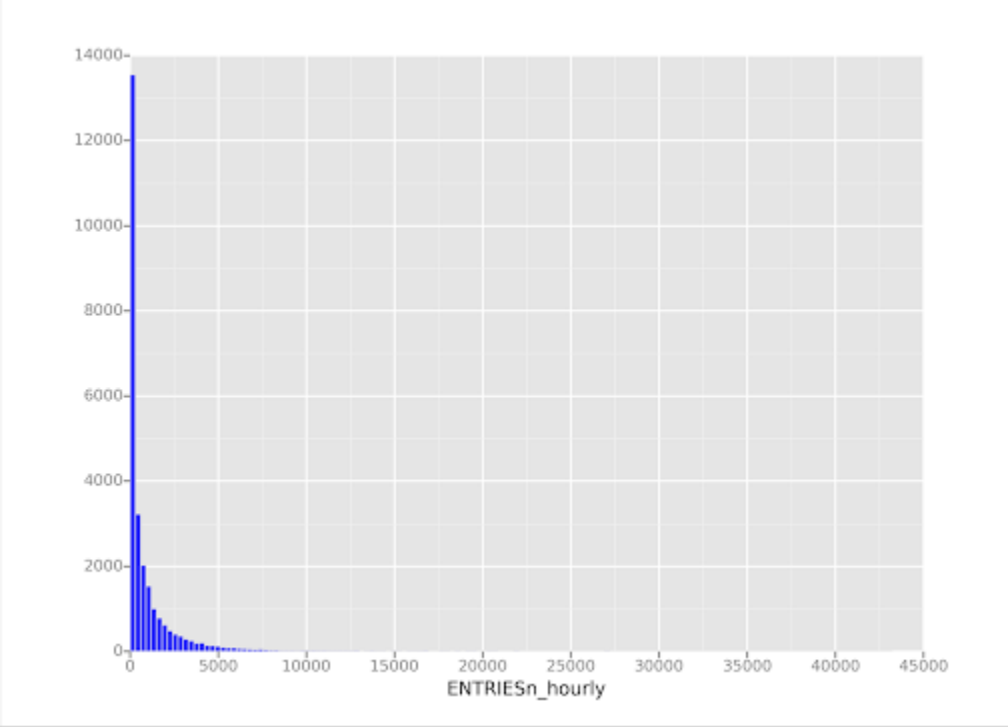
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.
You can combine the two histograms in a single plot or you can use two separate plots.

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 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.

Rainy Day Histogram:



Non-Rainy Day Histogram

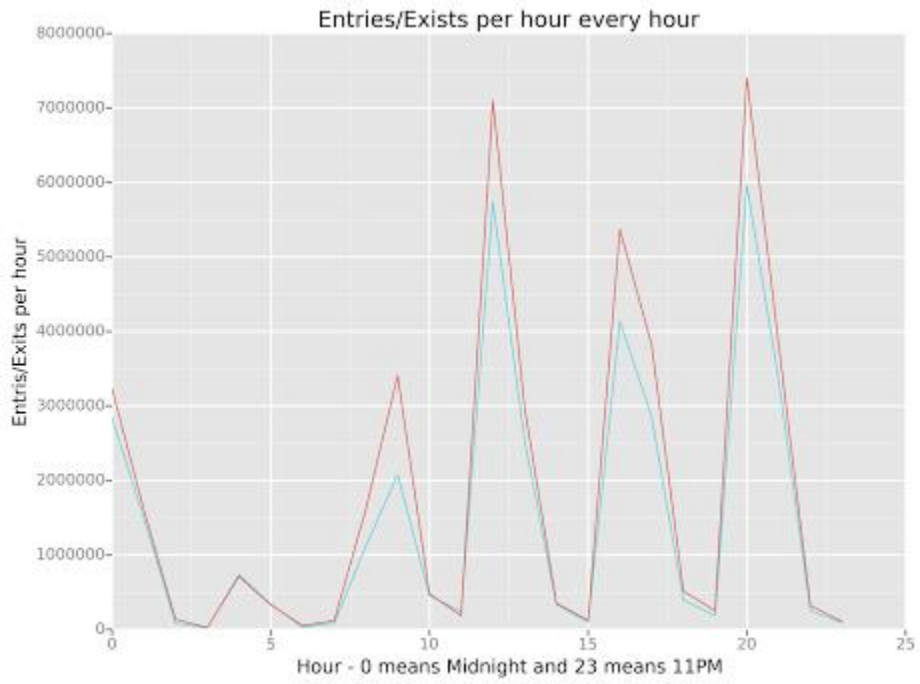


Both the Rainy day and the Non-Rainy day histograms show that the distribution of data is not normal. This information was used to determine which statistical test ought to be used for testing difference in the data.

3.2 One visualization can be more freeform. Some suggestions are:

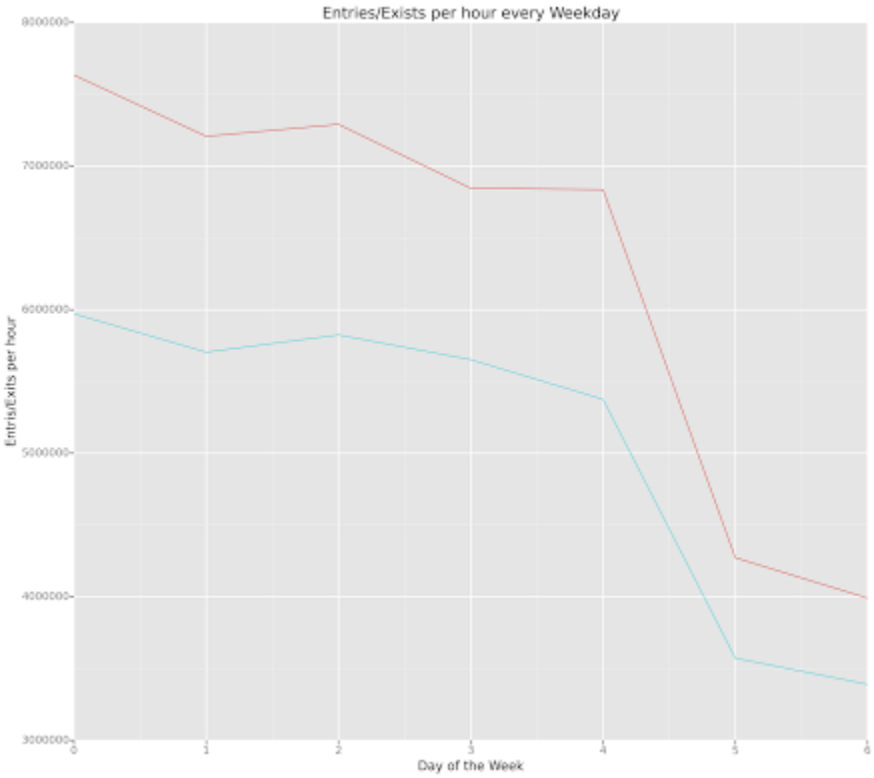
- Ridership by time-of-day
- Ridership by day-of-week

Ridership by time of day



While the normal office hours seem to have the expected peaks of increased ridership, it's surprising to see the highest ridership close to 8PM and 12Noon. Do a lot more people return home from work at 8PM in New York? Do people go out for lunch using the subway - a comparatively large distance from work?

Ridership by day-of-week:



This shows data as expected. The ridership is less over the weekend, when people are more likely to use their own vehicles and during the weekdays a lot more people use the Subway. This shows (what is commonly known) that a lot of people use the NYC subway system to go to work.

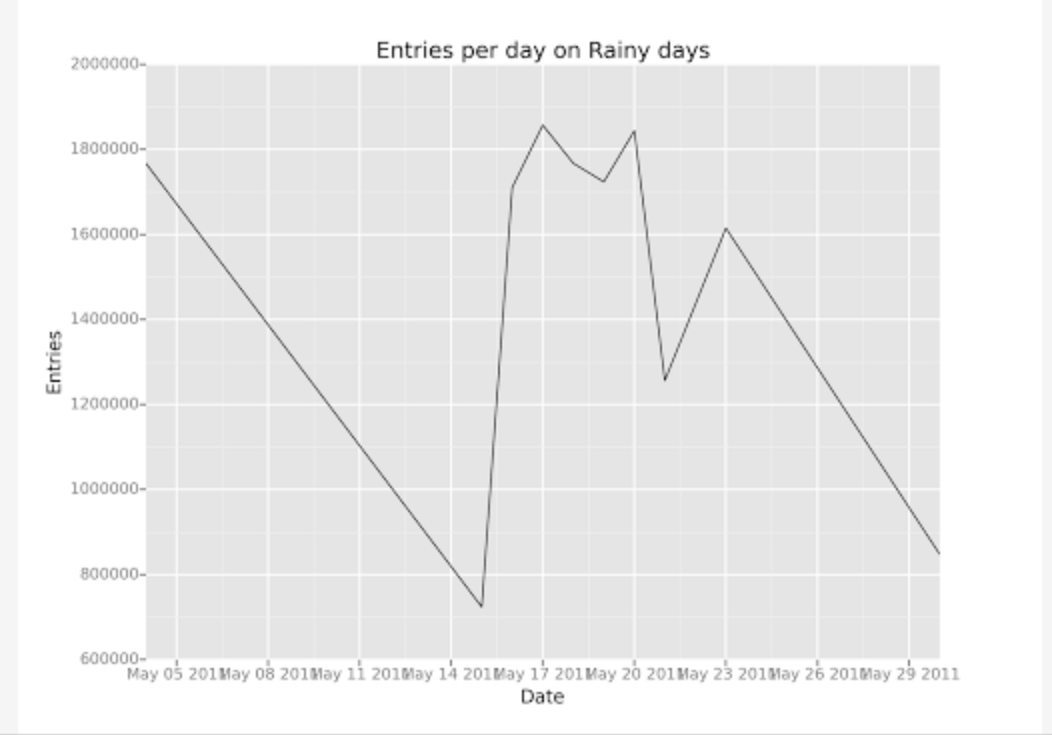
Section 4. Conclusion

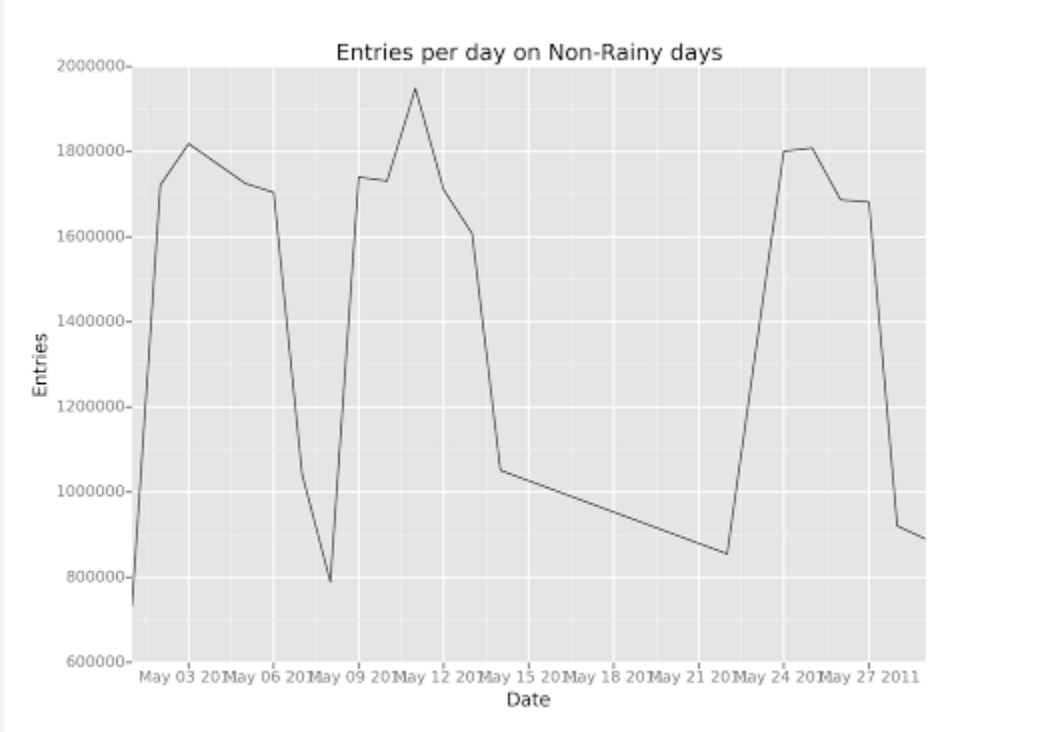
Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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 NYC subway when it is raining than 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.





The data visualization for the ridership on non-rainy and rainy days clearly the shows an increase in ridership on rainy days. The averages of ridership for the rainy and non-rainy days

are statistically different as proven by the Mann-Whitney test and also the averages values show that the ridership on rainy days is higher.

Also, during the regression modelling, we saw that rain, fog and mean temp have significant co-relationship with ridership numbers (along with hour of the day). This tends to confirm the hypothesis from the data above that rainy days tend to increase ridership in the NYC subways.

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,
- Analysis, such as the linear regression model or statistical test.

Since, the number of rainy days is comparatively much lower than the non-rainy days in the data set, the number of samples being analyzed for rainy days are much lesser than non-rainy days. This may cause a bias based on this particular set of data. A larger dataset (let's say the month of May for each of last 5 years) may indicate a different pattern that has been hidden here.

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

References:

http://onlinestatsbook.com

http://en.wikipedia.org/wiki/Ordinary_least_squares (also Wikipedia entries for Co-efficient of Determination and Gradient Descent)

Book: Statistics Without Tears by Derek Rowntree – Very good explanations of Hypothesis testing

Book: Think Python by Allen Downey (For Python outside of Pandas)

Book: Python for Data Analysis by Wes Mckinney (For Pandas)