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?

The mann-whitney test was used
It was a two sided test
The null hypothesis is that the 2 distributions are identical
The p-critical value is 0.05 for a 2 sided test

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.

The test is applicable because the observations are independent and can be compared / ranked against each other

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.

It had a p-value of 0.049999, mean 1 of 1105.44638, mean 2 of 1090.27878

1.4 What is the significance and interpretation of these results?

The p value is less than the p-critical value, this implies that the null hypothesis is rejected and it cannot be stated that the 2 distributions are identical with a probability greater than 95%. Therefore the 2 populations can be considered to be distinct, ie there is a difference distribution of those riding the subway at rainy and non-rainy times

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

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

precipi, maxtempi, meanwindspdi, I used a number of different variables to test the prediction during the exercise, none with any huge affect on efficiency. I also included the entries in the “UNIT” and “HOUR” columns as dummy variables. These bring the greatest efficiency in terms of an increase in the R-squared value

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.

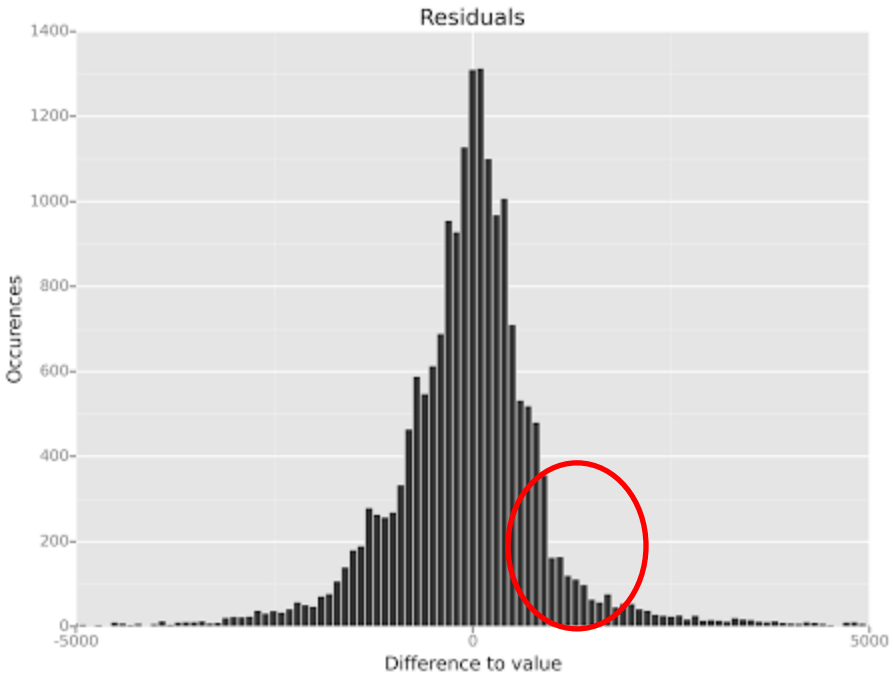
I chose these features because it is reasonable to believe thay they would influence the decision of someone to enter the subway system, i.e if it's too hot/cold (maxtempi), too precipitous, or too windy there are more likely to be people seeking refuge in the subway. At certain hours of the day there is also likely to be different levels of riders to others, e.g. during rush hour commutes and obvisouly this requires dummy variables rather than the actual hour to be used as there would be natural cycles throughout the day (ie rush hour peaks) such that nunbers would not increase or decrease linearly based on hour of day in perfect ascending/descending order. Also it is natural to think that different terminals would be busier than others. Assigning each different hour and terminal as dummy variables brings the greatest gain in efficiency

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?
-0.40121265, 1.59798, -0.77481067

2.5 What is your model's R^2 (coefficients of determination) value?
0.50578

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 value means the regression can be used to explain just over 50% of the variability around the actual ridership and so there remains nearly 50% to be explained through other factors (the ideal value for the R-squared would be 1).

From looking at the residuals below it can be seen that there appears there may be a negative skew in the distribution, with limited data in the highlighted section, such that the residuals seem to be negatively biased and not normally distributed. This is not what would be expected from residuals if the structure of the model was good, implying that there are additional factors which may be needed to be integrated to produce a better fit, or that the data is non-linear or suffers from other limitations.



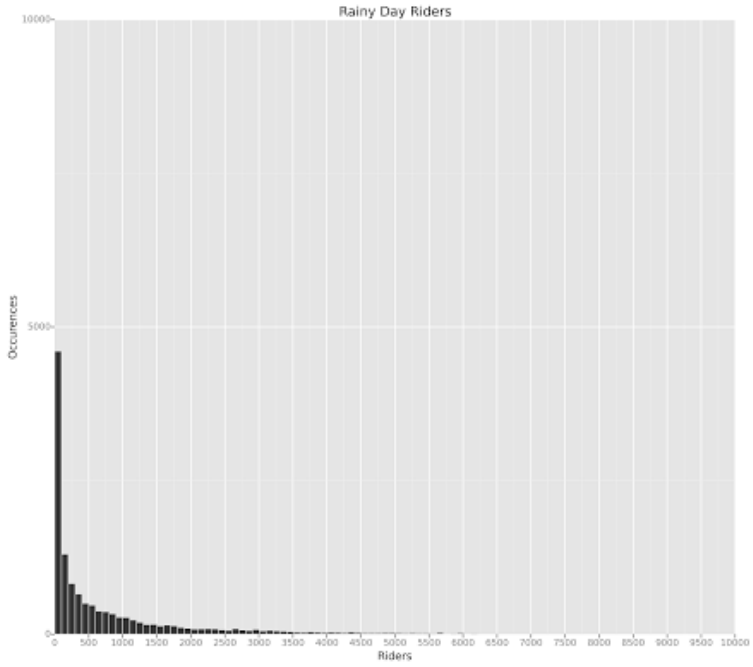
Section 3. Visualization

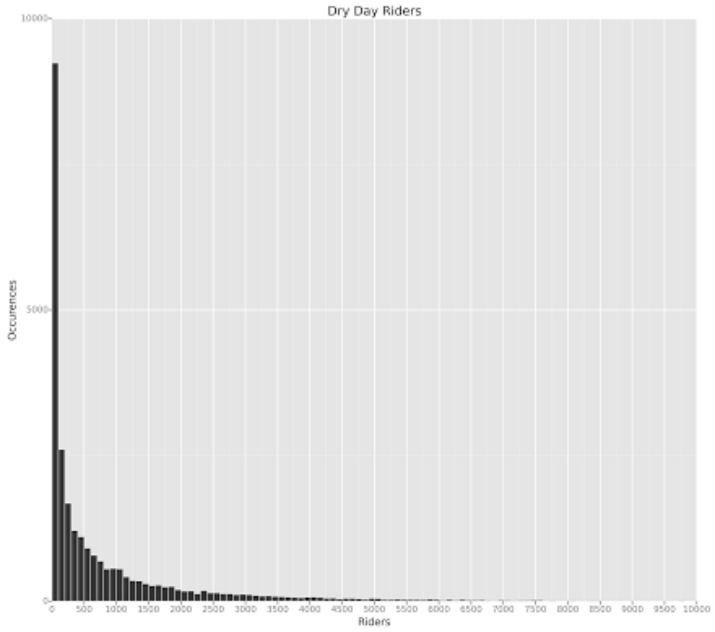
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

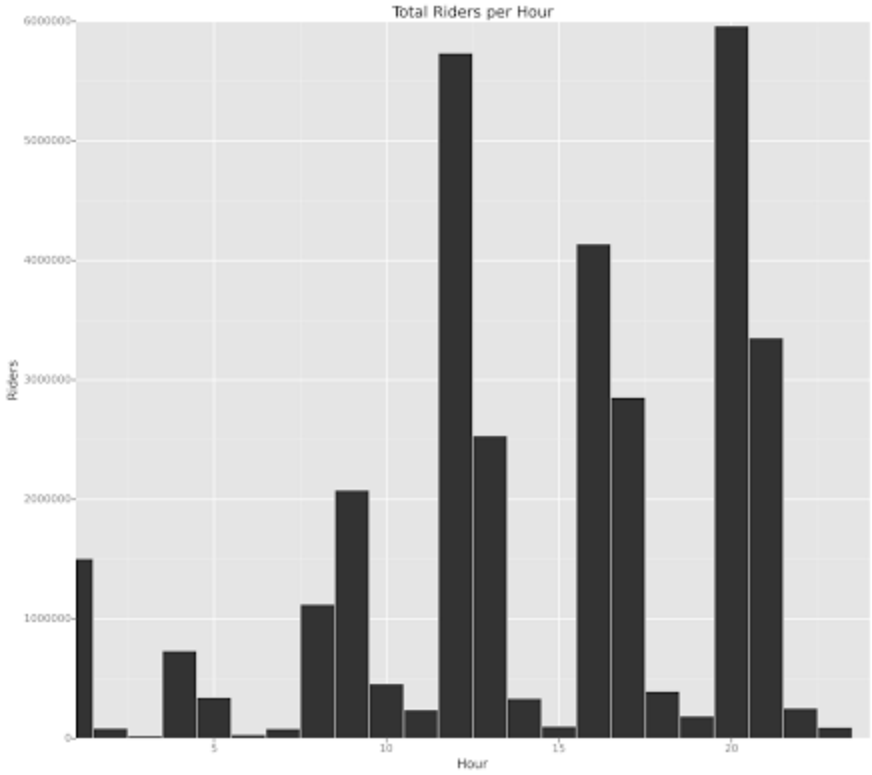
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms,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





It can be seen that the occurences of low number (0 to 100) riders per hour is much higher for dry hours (“Dry day riders”) than for non dry hours (“Rainy day riders”). Also the rainy hour rider diagram is fatter tailed.



- Riders per hour
    - It can be seen from the above that the total number of riders varies substantially at difference hours of the days which is to be expected, making this an important factor in ridership prediction.

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.

More people ride the subway when it is raining. The difference is significant enough to pass the mann-whitney test so it cannot be said with a 95% confidence that the distributions of people riding the subway on rainy vs non-rainy times are different.

The right hand tails in the histogram titled “dry day riders” are also lower than the tails in the histogram entitles “rainy day riders” above, showing a larger number of occurences of high numbers riding the subway on rainy days

The Linear regression however does not in itself contribute to this argument as it considers a number of other factors and also leads to a poor fit of prediction.

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 dataset used has been truncated to allow the server to perform computations. The timespan it covers in particular is quite limited and its reliability has been reduced as a result. This may have had a material impact on the data if for example other elements of the data were also not randomly distributed originally and key biases were incorporated as a result of the truncation, e.g. transportation options may have changed in later data such that there is no longer a reason to ride the subway. Also I have not researched the source of the data and don't know it to be an accurate or complete recording of the actual NYC data.

The linear regression I analysed was a poor fit, however I did experiment with a number of factors, bringing in methods of better incorporating time of day for example could enhance the prediction.

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