

## 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, meanwindspd, 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.

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 that they would influence the decision of someone to enter the subway system, i.e. if it's too hot/cold (maxtemp), too 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. 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, -0.15113973, -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?

This means the model used is not a good fit for the model the ideal value being close to one, therefore the linear regression used is not appropriate, it may be possible to get a better fit but based on this analysis the linear regression used is not appropriate.

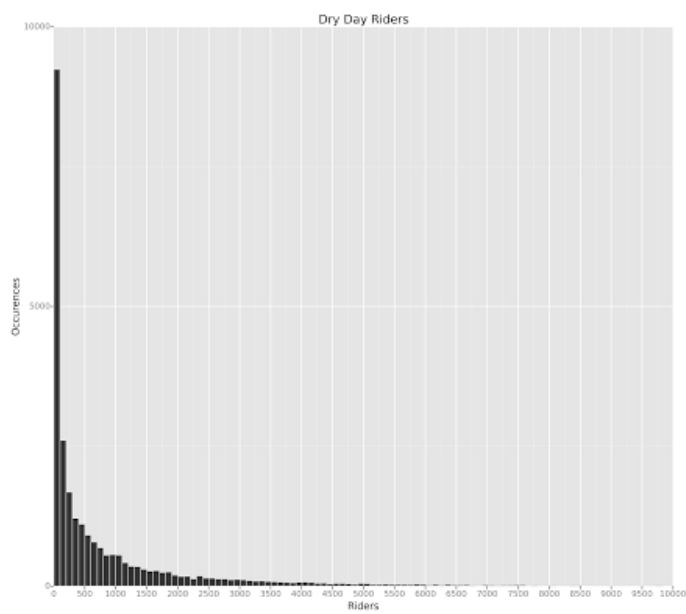
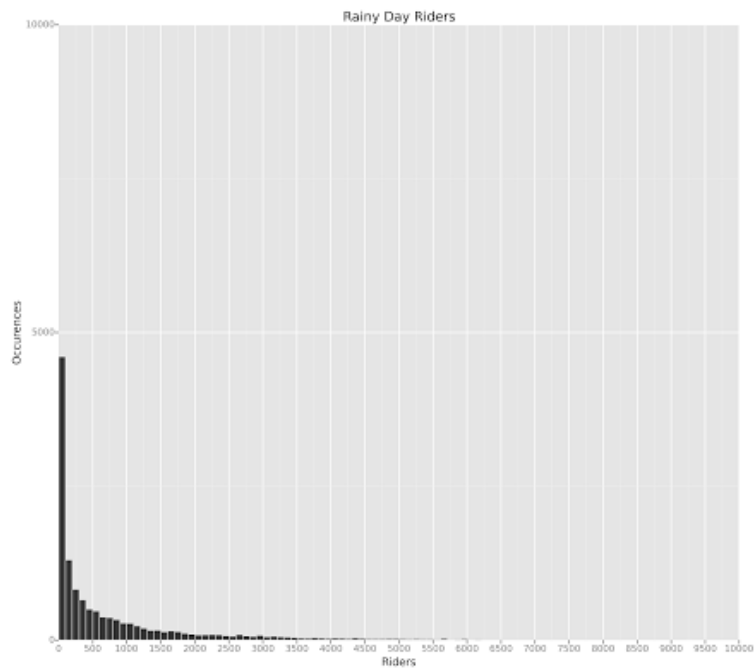
### Section 3. Visualization

Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

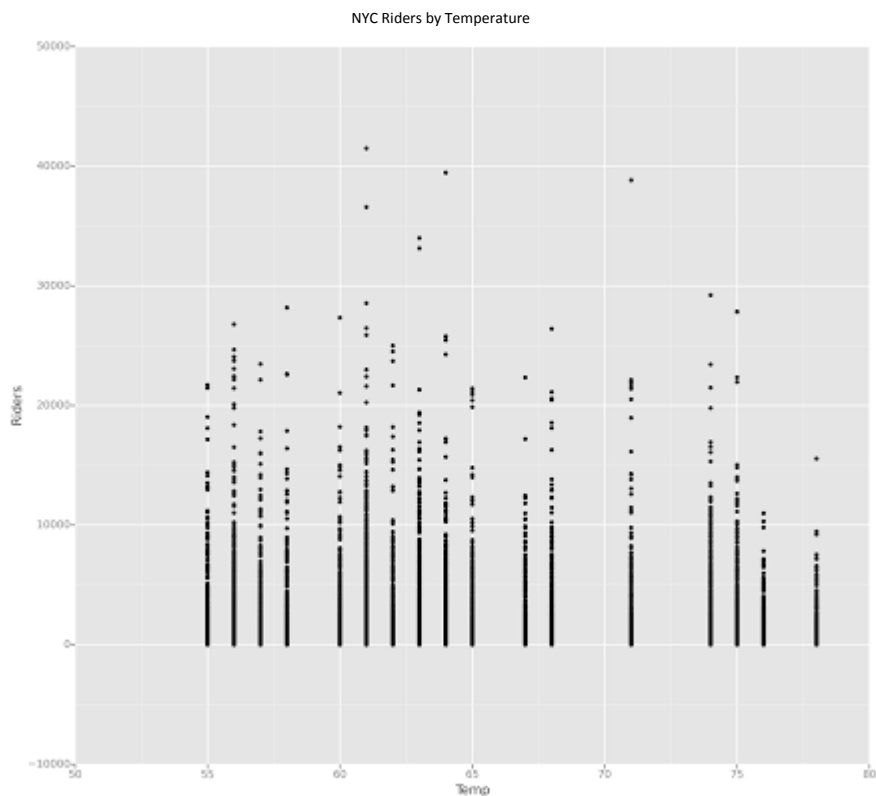
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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 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 occurrences 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.**



- Ridership by temperature
  - It can be seen that the extreme outliers in terms of temperature are not concentrated on a temperature range, making it possible that this is not a key contributor to ridership figures by itself

#### 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 appear to ride the subway when it is raining. This is demonstrated through the fatter tails in the histogram above and the lower occurrence of less than 100 people entering the turnstiles. The difference is also 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 Linear regressions does not 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. This may have had a material impact on the data for example if the data was 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?