1. **Statistical Test**
   1.1. The Mann Whitney U-test has been used. The 2-tailed test is used since we are not making a statement about ridership being less or more but simply checking if there is a difference in ridership during rain vs no rain. The NULL Hypothesis is: There is no difference in ridership trends when there is no rain vs when there is rain. The p-critical value is 0.05
   1.2. The test is applicable since the data has no known distribution and the number of observations are greater than 20. It is assumed that the rain vs no-rain series are independent of each other as well as are independent of each other. The test assumes that under the null hypothesis, the probability of drawing a value from sample X (rain) larger than one from sample Y (no rain) is the same as drawing a value from sample Y (no rain) larger than one from sample X (rain). The alternative hypothesis refutes this assumption.
   1.3. Test Results
      1.3.1. The p-value is 2.74106957124e-06 (this was obtained by multiplying the mann-whitney U Test p-value by 2 to get the 2-sided p value)
      1.3.2. The U value of the Mann Whitney test is 153635120.5
      1.3.3. The means and medians of the data sets are:
         1.3.3.1. Rain → Mean: 2028 (rounded), Median: 939 riders
      1.3.4. No Rain → Mean: 1,846 riders, Median: 893 riders
   1.4. The miniscule p-value suggests that the null hypothesis cannot be accepted implying that the rain and no rain ridership are not similar. Hence, we can safely conclude that the populations for rain vs no-rain are distinct.

2. **Regression -** ***The same "Raw" turnstile data as in Section 1 has been used in this section as well.***
   2.1. I used OLS using statsmodel for my to compute Coefficients
   2.2. I used the following features:
      2.2.1. rain
      2.2.2. meantempi
      2.2.3. precipi
      2.2.4. fog
      2.2.5. UNITS as dummies
      2.2.6. Hours as dummies

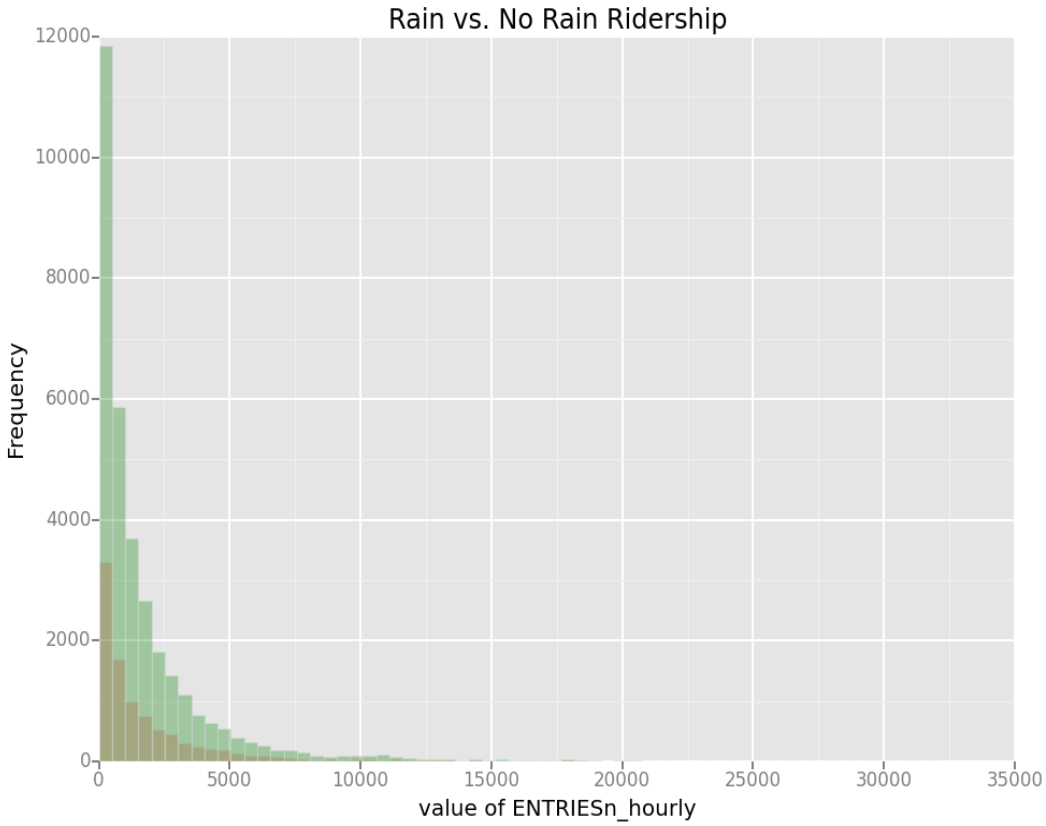
   2.3. The reasons for selecting these features were intuitive:
      2.3.1. **rain** must have an impact on ridership as people are less likely to go out during rain. Also they are more likely to take cabs as the train station may not be easily accessible to everyone in the rain. **meantempi** was chosen as people also are less likely to travel if the temperatures are

extreme such as in the dead of the winter. **Precipi** was chosen for a similar reason since higher precipitation may imply impending rain hence decreasing the chances of people wanting to travel. Fog was chosen as it may increase ridership as people would prefer not to drive during a fog and would feel safer taking the subway. Units were chosen as dummies since each unit may be in a different location and ridership would be tied to that location (close to offices vs. isolated areas). Similarly hours were chosen as dummies since there would be more traffic during certain hours (such as rush hour).

2.4. Coefficients for non-dummy features:
    2.4.1. rain → -4.02
    2.4.2. meantempi → -88.39
    2.4.3. precipi → -19.90
    2.4.4. fog → 87.47
2.5. The R^2 value is 0.526
2.6. R^2 interpretation:
    2.6.1. The higher the R^2 value the better the model can be considered a fit
    2.6.2. An R^2 value of 0.526 is not very high, even though it's not very low. I would **not** be comfortable using this model to predict ridership and would aim for a higher R^2 value. Something higher than 0.7 (70 %) would be more convincing.

3. Visualization - The same data as Sections 1 and 2 is used to show a histogram of rain vs no-rain ridershp

3.1. The image below shows the Rain data in Red and the No Rain ridership data in Green.



We can make the following observations based on this data:

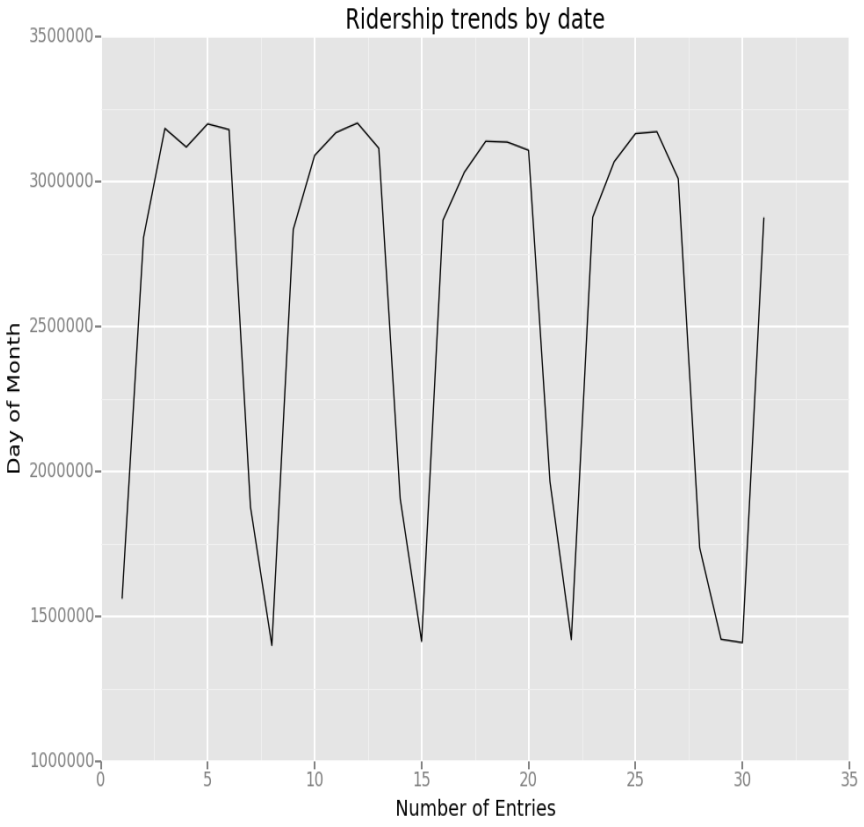
3.1.1. Clearly the ridership is impacted by rain - it is much lower for rainy days

3.1.2. Ridership for no-rain days has a much fatter tail than that for rainy days

3.1.3. Ridership in non-rainy days is significantly higher than non-rainy days

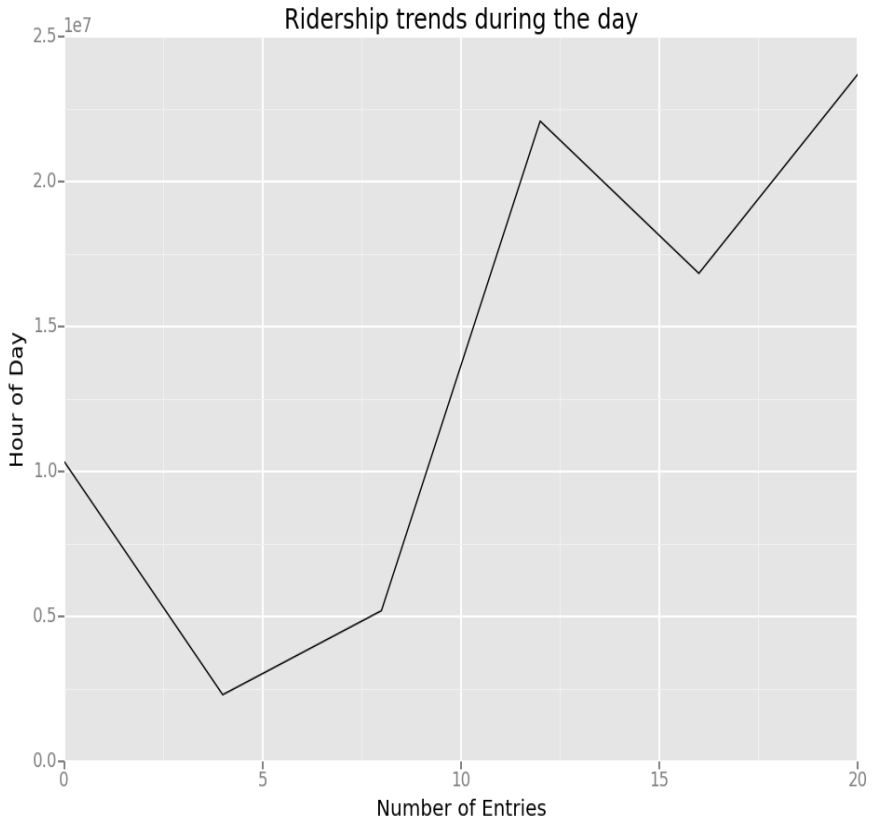
3.2.

3.2.1. The following chart shows **aggregate ridership by date** for the month of May 2011.



This clearly indicates that the trend is cyclical over the week. For example, May 1 2011, which was a Sunday has the lowest points. So ridership is lowest on Sundays and second lowest on Saturdays. This makes sense as ridership must be significantly higher on weekdays due to commuters.

3.2.2. The following chart shows **aggregate ridership by hour** for the month of May 2011



The chart shows ridership going up in the mornings starting around 7am and linearly increasing until 10:30 am at which point it peaks and starts decreasing and then starts rising again in the evening around 4 pm and linearly increasing until 8pm. Since the readings are in 4 hour intervals, the conclusion is not precise in terms of time intervals but more of a broad conclusion showing subway traffic **increasing during rush hours**.

4. Conclusion

4.1. The conclusion is that subway ridership differs on rainy vs. non-rainy days. Specifically, more people ride the subway when it is **not** raining.

4.2. On the basis of the statistical test we can see that ridership is definitely different when comparing rainy vs. non-rainy days. We can also see that the mean and median of the non-rainy days is much higher than that of rainy days.
The regression analysis shows a coefficient of -4 associated with the 'rain' feature. Hence ridership decreases by a factor of ~4 during rainy days.
Both these observations (along with the visualization of the histogram in 3.1) lead us to the conclusion that ridership is lower on rainy days.

5. Shortcomings

5.1. Data Shortcomings

5.1.1. 4 hour interval readings - The subway readings are being taken at 4 hour intervals, hence the data is not precise from the point of view of ridership during a specific hour

5.1.2. Difference in ridership patterns across units: In this analysis ridership has mostly been looked at as a single time-series across days. In many cases it has been summed over different subway stations or units. This may not present an accurate picture of ridership patterns since ridership patterns may vary greatly across subway units. Intuitively, ridership on key hubs such as near Penn Station or Times Square may be very different from lesser used units. Hence, it would have been more accurate to divide ridership by clusters of units such as:
- Very busy
- Somewhat busy
- Not busy

And then examine ridership among units of similar categories.
(See histogram in section 5.4)

5.2. Linear Regression Model Shortcomings

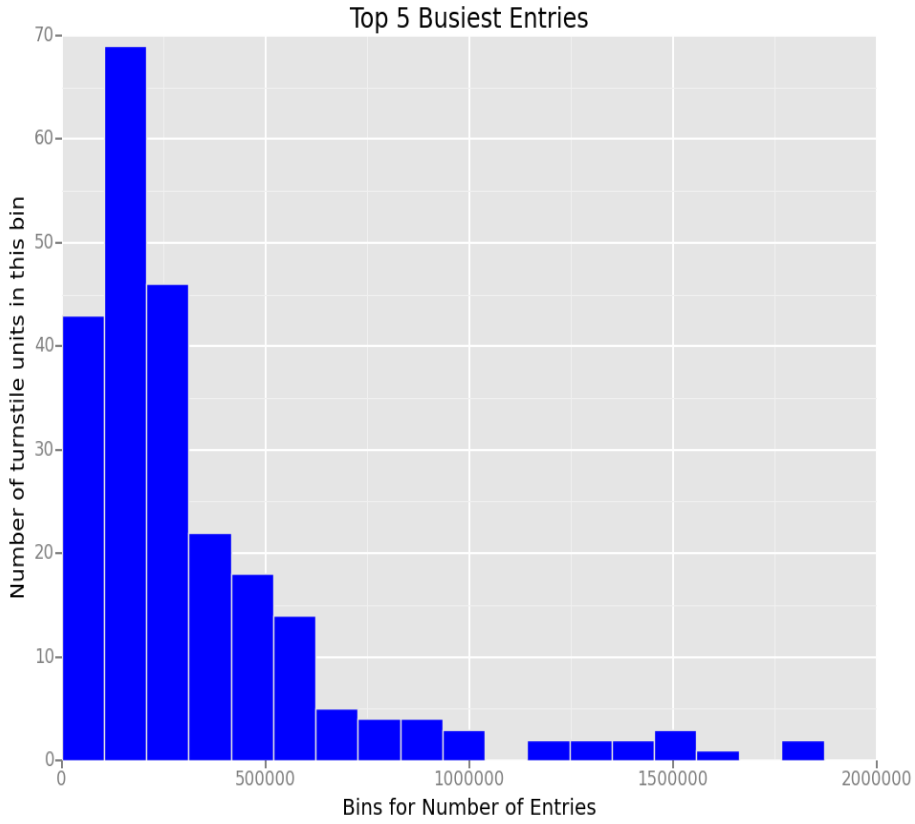
5.2.1. Linear regression in my case has an R^2 of 52.6% which is not very high and hence the results are not very reliable

5.3. Statistical Test Shortcomings

5.3.1. The Mann-Whitney tests is not very precise since it is a test that accepts data for which a distribution is not known.

5.4. Other Insights

5.4.1. To illustrate the point (mentioned in section 5.1.2) about uneven distribution of ridership over units, please see the histogram depicting ridership distribution across units. The x axis shows the sum of entries across units and the y axis shows how many units in that bin.



It is clear that this histogram has a very long tail, that is most units have ridership per day well below 1 Million while a small number of units have ridership over 1 Million. Hence the distribution is not even. It looks more like a exponentially decreasing distribution.