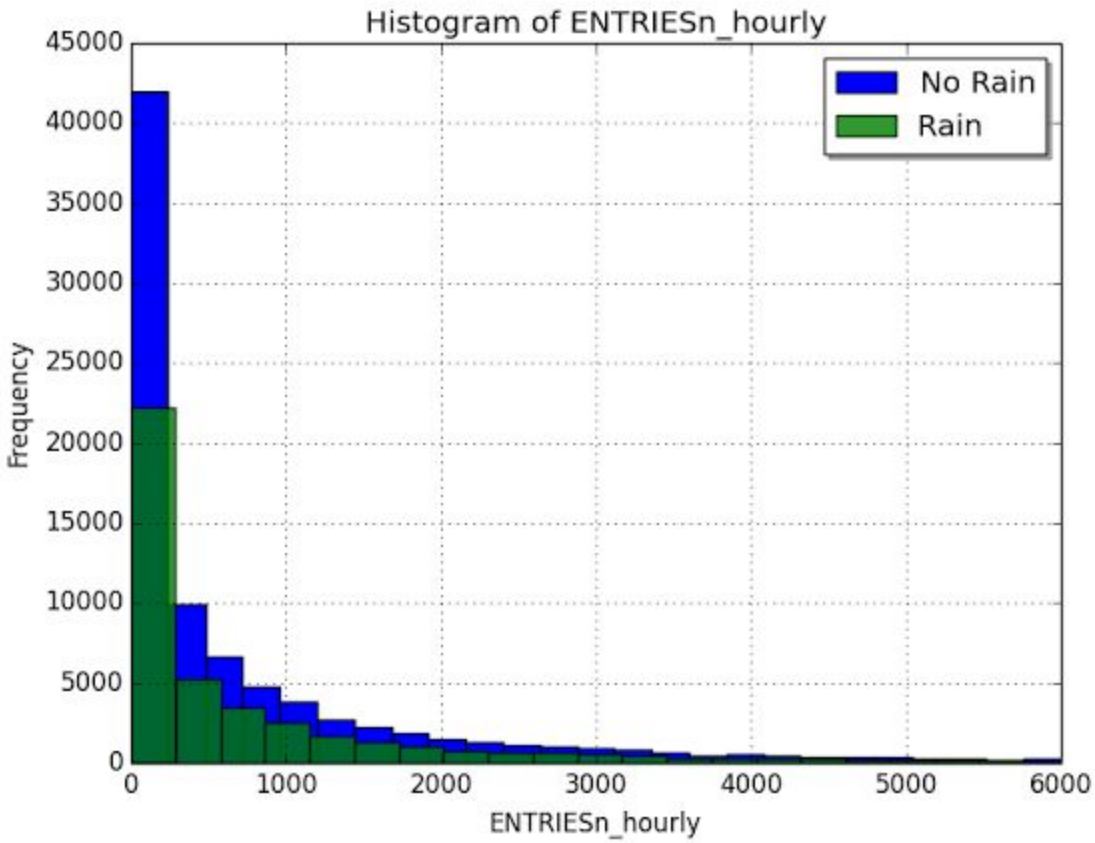
# Analyzing the NYC Subway Data

## 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?

Since the distribution of hourly entries on rainy days/non-rainy days is not normal a nonparametric test is used.




Mann Whitney U test was performed to check the significance of differences between subway ridership on rainy days compared to non-rainy days.

Since we do not make assumption about directionality, a two-tailed test is chosen.

Null-Hypothesis: hourly ridership distribution on Rainy-Days is identical to hourly ridership distribution on Non-Rainy-Days.
Alternative Hypothesis : hourly ridership on Rainy-Days is not identical to hourly ridership distribution on Non-Rainy-Days

p-critical : .05

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.

Mann Whitney U does not assume normality in the sample.
The sample size is > 20 for Mann Whitney U test to be performed.
The two samples are independent, so Mann Whitney is applicable[1].

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.

| Mean Hourly Entries on Rainy Days | Mean Hourly Entries on non-Rainy Days | The Mann-Whitney statistics | Two-sided p value |
|---|---|---|---|
| 1105.4463767458733 | 1090.278780151855 | 1924409167.0 | 0.04999982558 |

1.4 What is the significance and interpretation of these results?

From our statistical test ( Mann Whitney U ) we reject the null hypothesis at p < .05
We can conclude the ridership of NYC subway on rainy days is significantly different compared to non-rainy days at p < 0.5

| | Rainy days | Non-rainy days |
|---|---|---|
| mean ENTRIESn_hourly | 1105.45 | 1090.28 |
| median ENTRIESn_hourly | 282.0 | 278.0 |
| Num Data points | 44104 | 87847 |

From the above results and the Mann Whitney U test, we can conclude that more people ride the subway when it is raining.

# 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:

- OLS using Statsmodels or Scikit Learn

[1] http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.mannwhitneyu.html

- Gradient descent using Scikit Learn
- Or something different?

A Orinary Least Squares (OLS) from Statsmodles was used.

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

The following were the input variables used in the model :

'rain', 'fog', 'meantempi' , 'precipi', 'meanpressurei', ‘weekday’, ‘Hour’, ‘UNIT’ (proxy for location)
Since, ‘weekday’, ‘Hour’, UNIT are infact multiple variables in themselves they were represented as “DUMMY” variables by using ‘pandas.get_dummies’ function.

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.”

'rain', 'fog', 'meantempi' , 'precipi', 'meanpressurei' were used because they are all related to weather data. A bad weather day might influence subway ridership. Example, on a very hot day perhaps, less people might take the subway.

At the same time, some of the other weather indicatores like 'meanwindspdi', 'thunder' , 'meandewpti' do not seem to strong indicators of ridership in this model.

Adding ‘Hour’ and ‘UNIT’ data drastically increased the $R^2$ value, so these were kept. Also it does make sense, since subway ridership is very dependent on time of the day and location of the subway station.

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

```
            coefficients

rain             -79.9124

fog              106.4658

precipi          -60.1037
```

```
meanpressurei  -192.9981

meantempi        -8.6407
```

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

0.514467278677

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?

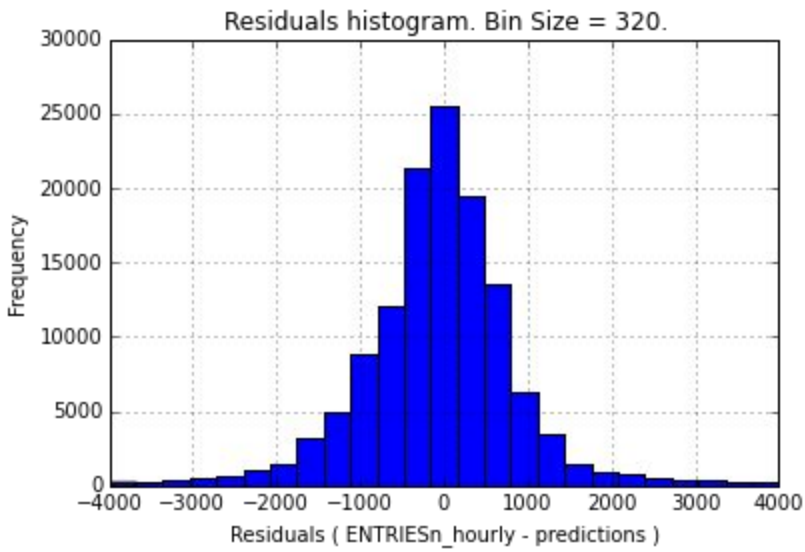
$R^2$ value of ~0.51 means that the regression model can explain 51% of the variances of our dependent variable ( Hourly Subway ridership ). The rest of the 49% variance in Hourly Subway ridership is unexplained by our model.

$R^2$ value in itself cannot be a good indicator of the model. Residual plots need to considered. Our residual plot looks good ( See figure, explanation below ). Given the number of variables (factors) and residual plots, $R^2$ value of 0.51 would be acceptable. There seem to be other issues regarding multicollinearity in the regression model. This is evident from the summary of statsmodels.OLS api[2] . However, Multicollinearity does not affect $R^2$ of the model, hence the prediction strength of the model remains the same.[3]

Following is the histogram of residuals. since the residual graph is normal, this is an indicator that our linear regression model is not overfitted (or biased in other words).

[2] Display from the results section of OLS : " The smallest eigenvalue is 1.32e-22. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
0.514467278677"

[3] http://blog.minitab.com/blog/adventures-in-statistics/what-are-the-effects-of-multicollinearity-and-when-can-i-ignore-them

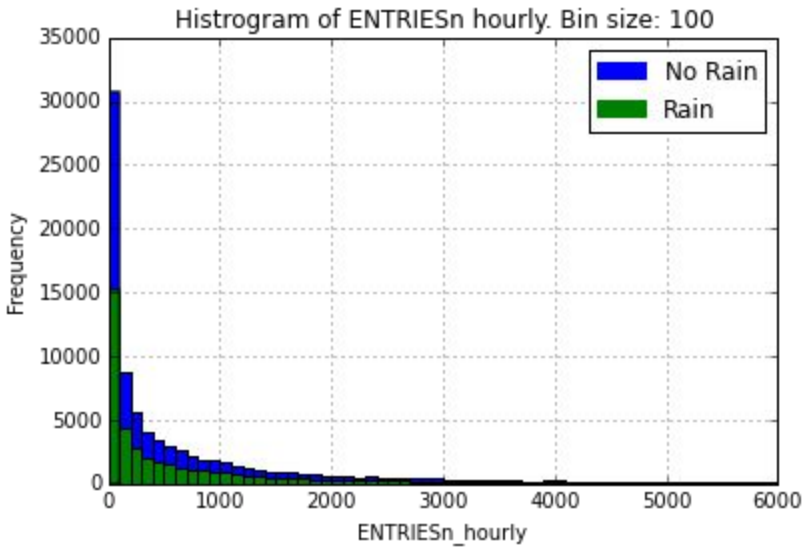



# Section 3. Visualization

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.

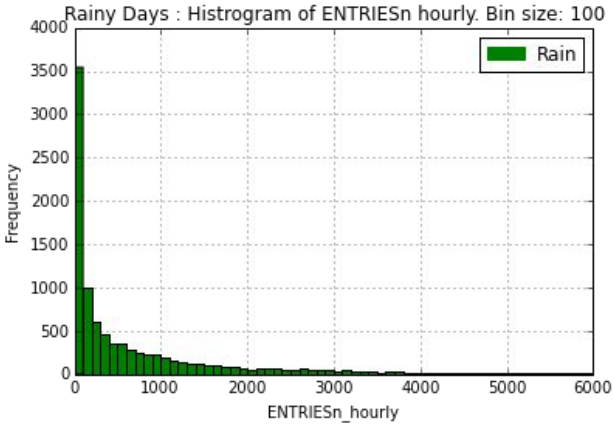
- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms,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

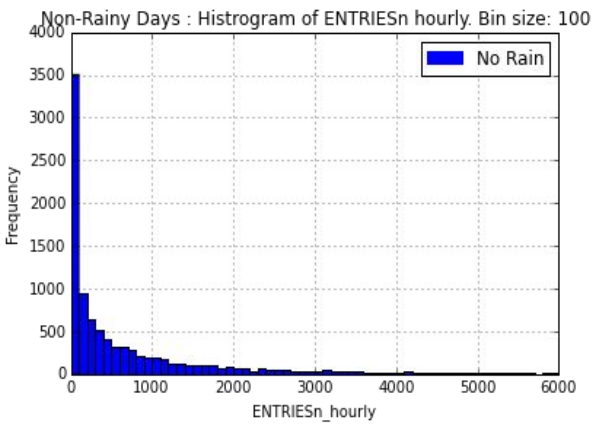



| | Rainy days | Non-rainy days |
|---|---|---|
| mean ENTRIESn_hourly | 1105.45 | 1090.28 |
| median ENTRIESn_hourly | 282.0 | 278.0 |
| Num Data points | 44104 | 87847 |

Since the data points for rainy days is much less than non-rainy days. The above graph might not be good to make conclusions/inference.

We collected a random sample of 10000 from both samples to compare them in histograms:

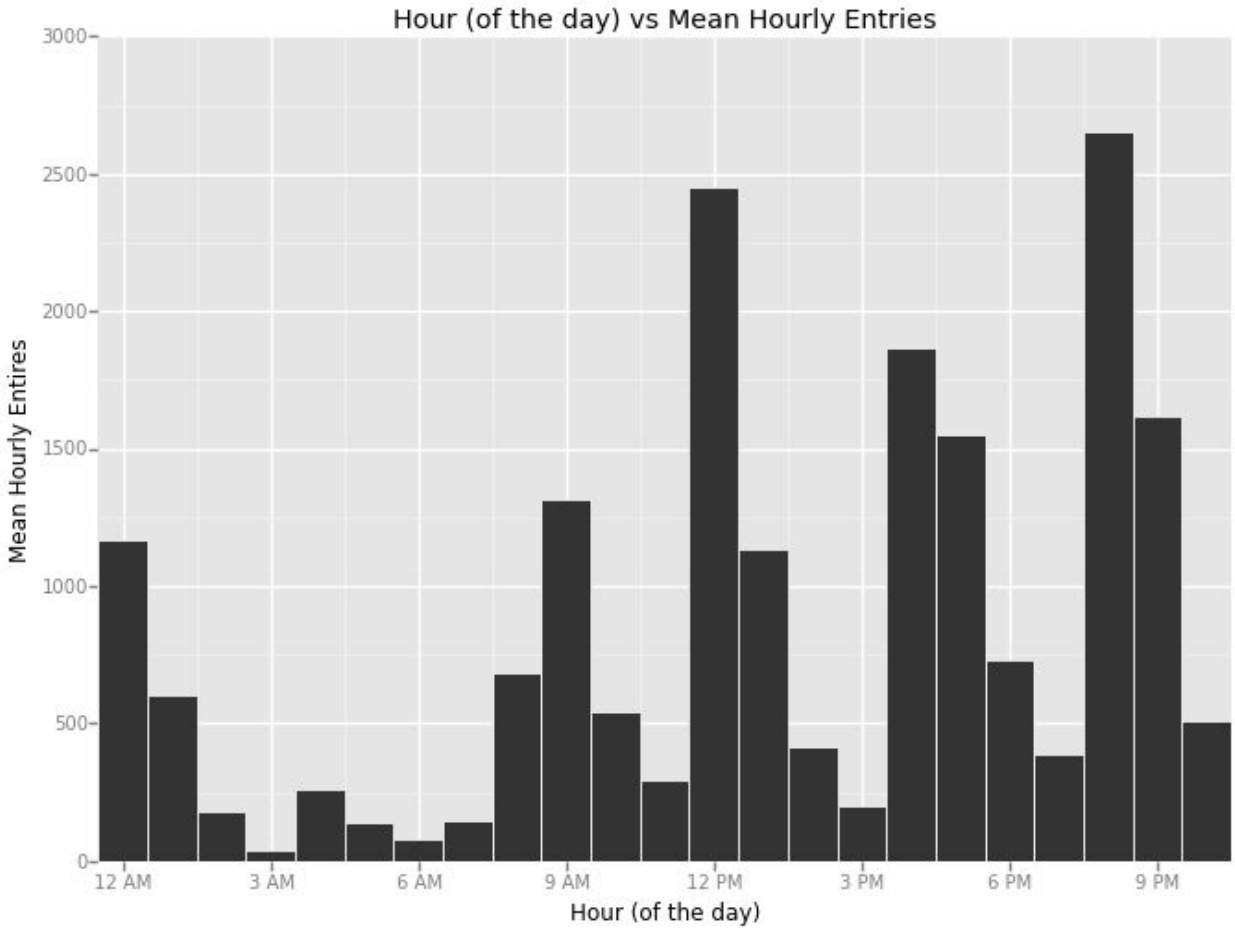


There does not seem to much significant difference to the naked eye in these two randomly drawn samples. However, as seen from the mean of Hourly entry data between Rainy vs non-rainy data is different, rain does seem to affect ridership.

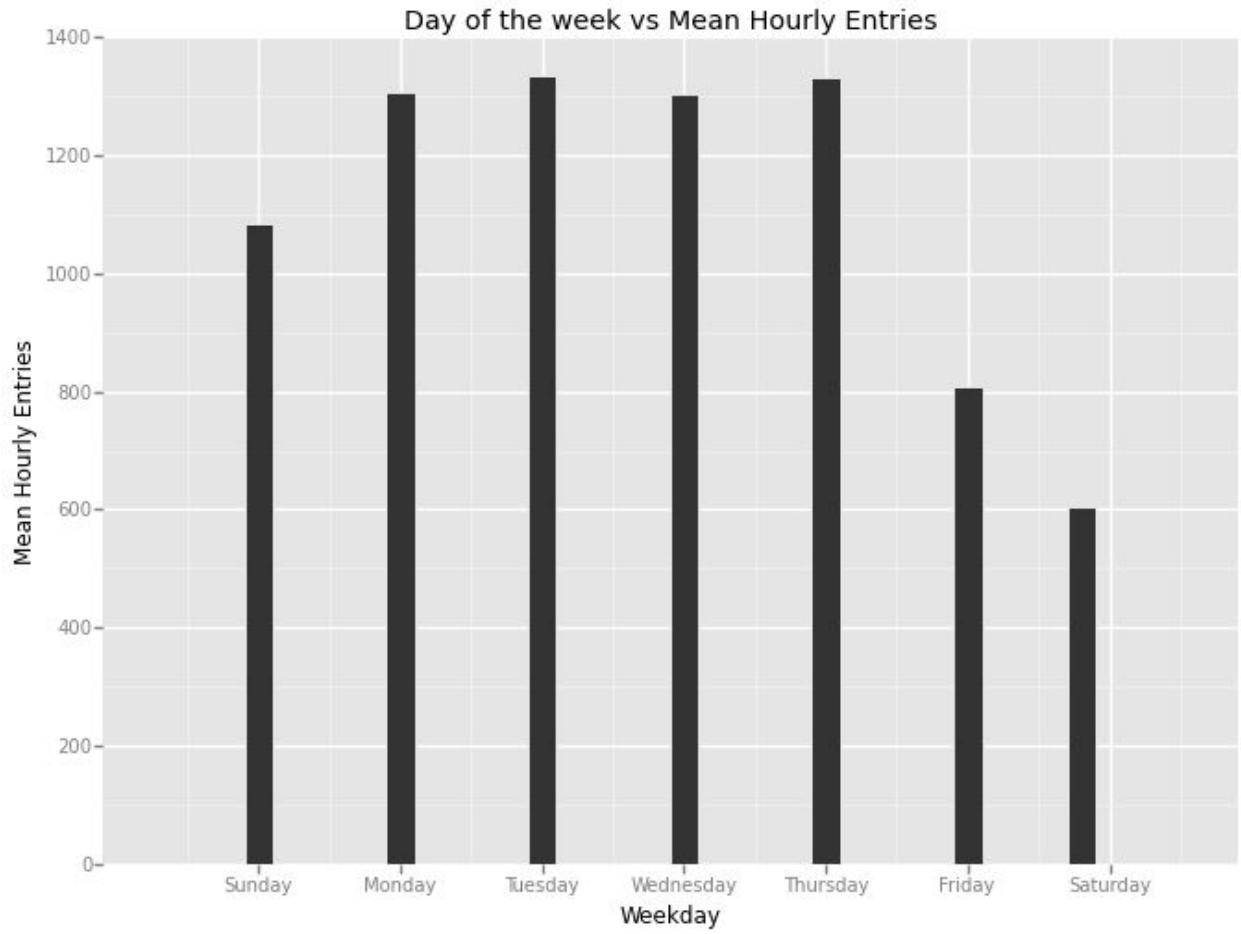
3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

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




Noon time, 8-9 PM seems to be busiest hours of the day.

Morning commute time does not seem to be all that busy when compared to noon time and 8-9PM, which is a bit of a surprise.




Ridership seems to be evenly distributed across the week with the exception of perhaps Saturday when there does not seem to be very high entries compared to other days.

# 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 subway when it is raining. From a preliminary look at the data by plotting, it is not very evident regarding the relationship of rain and ridership.

From looking at the mean, median of ridership between Rainy days vs non-Rainy days, it appeared that subway ridership is more on rainy days.

To bolster our observation from mean, median a statistical test was performed. Since the samples do not follow a normal distribution, a nonparametric test, Mann Whitney U test was performed. From the outcome

of Mann Whitney U test we reject the null hypothesis ( that subway ridership on rainy days vs non-rainy days is equal).

From the significance level of Mann Whitney U test and the respective means for rainy days and non-rainy days, we can confirm that ridership on rainy days is higher than non-rainy days.

A regression model was built to predict the subway ridership based on a number of factors. The factors themselves were carefully selected and a model was built. Of course, we are interested in effect of rain, so rain is one of the factors considered in our model.

Rain does show an effect on the ridership in our model. However, the coefficient of 'rain' factor being negative, we observe that the our model would indicate that ridership would decrease when it is raining. This is contrary to our statistical analysis. But it appears multicollinearity may be affecting our results and the coefficients themselves cannot be trusted, but the validity of the model itself to predict the ridership is good enough.

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.

| | Rainy days | Non-rainy days |
|---|---|---|
| mean ENTRIESn_hourly | 1105.45 | 1090.28 |
| median ENTRIESn_hourly | 282.0 | 278.0 |

The mean and median statistics between rainy and non-rainy days indicate that subway ridership is more on rainy days.

From Mann whitney results we conclude that the ridership on rainy days is significantly different on rainy days compared to non-rainy days.

U = 1924409167.0 , p = 0.04999982558 (two-tailed)

However, our regression model does not predict a positive coefficient for 'rain' as a factor. It appears multicollinearity might be affecting our coefficients. The coefficient for rain is -79.9124. As noted before, this is contrary to our statistical tests.

# 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.

It appears Scipy's Mann Whitney U test does not tell us which of the U Score it chooses while reporting its output[4]. This makes it difficult to tell which of the groups has higher U score. Of course, comparing mean, median of two groups gives us this information but seemingly scipy's interface could be better here(?).

Too many parameters in the dataset increase the difficulty in picking the variables for linear regression. I suspect that there might be a better way to pick variables to be considered in the regression. The value of R-squared was hard to guess/predict based on the factors we pick. Also, looks like having many factors which are related (multicollinear) affect the linear regression model. For example, rain and fog are related, it is more likely to be foggy on a rainy day. This effects might be hurting our coefficients in our regression model.

The Dataset provided had data only for the month of May. Hence the regression model built might not be useful for other times of the year.

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

[4] http://docs.scipy.org/doc/scipy-0.16.0/reference/generated/scipy.stats.mannwhitneyu.html