# Section 0 – References


1) https://statistics.laerd.com/premium-sample/mwut/mann-whitney-test-in-spss-2.php
2) http://www.pythonlearn.com/html-009/
3) http://en.wikipedia.org/wiki/Gradient_descent
4) http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html
5) http://docs.ggplot2.org/current/geom_bar.html
6) http://docs.scipy.org/doc/numpy/reference/generated/numpy.sum.html
7) http://docs.scipy.org/doc/numpy/reference/generated/numpy.mean.html
8) http://mathesaurus.sourceforge.net/matlab-numpy.html


# 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?

*Ans – I used Mann-Whitney U test to analyze the NYC subway data. I used a two-tail P value. The null hypothesis in the test was following – "The distribution of number of entries on Non-rainy and rainy days are statistically similar" i.e.*

$$H_0 : \mu(entries/rain) = \mu(entries/noRain)$$

*The p-critical value is 0.05.*

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
.

*Ans – A histogram plot of the number of entries on rainy and non-rainy days showed that the distribution is non-normal. Hence parametric tests like Welch t-test are not applicable. For such non-normal distribution, Mann-Whitney U test is applicable and hence was used to analyze the dataset. Mann-Whitney U test makes the following assumptions*

1) *Assumption1: **There exists 'one dependent variable' that has been measured at the continuous or ordinal level**. In our case it's the 'ENTRIESn_hourly' variable.*
2) *Assumption 2: **There exists one independent variable that has two independent groups**. In our case its 'Rain' variable and the independent groups are 'Rainy days' and 'Non rainy days'*
3) *Assumption 3: **The observations are independent**. We assume that the observations made in the dataset are independent.*

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.

*Ans - I got 4 values. The p-value = 0.025 (one tail value), to convert it to two tail we need to multiply it by 2, so 2-tail p-value = 0.05,*

*U =1924409167.0,*

*mean_ridership_on_rainy_days = 1105.45*

*mean_ridership_on_non_rainy_days = 1090.28 (As implemented in exercise 3.3)*

1.4 What is the significance and interpretation of these results?

*Ans – Since the p-critical value was 0.05 and the value returned by the Mann Whitney U test is 0.05 (p-value), hence p-value=p-critical. It shows that the distribution of entries on rainy days and non rainy days is statistically different. So the null hypothesis is rejected.*

# 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?

*Ans – I used **gradient descent approach** to compute coefficients and produce predictions for ENTRIESn_hourly. (as implemented in exercise 3.5)*

*Additionally I used **Ordinary least squares** using Statsmodels to compute coefficients and produce predictions for ENTRIESn_hourly. (as implemented in exercise 3.8)*

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

Ans – *I used several input variables as features in my model.*

*They are – 'rain' - if rain occurred within the calendar day at the location,*
*'hour'- Hour of the timestamp from TIMEn ,*
*'hour$^2$', 'hour$^3$' – higher powers of 'hour' variable*
*'meantempi' - Daily average of tempi for the location*
*'day_of_week' 0-Monday, 1-Tuesday ---- '6'-Sunday*
*'day_of_week$^2$' – Higher power of 'day_of_week'*

*Yes, I also used dummy variables as part of my features. The remote unit 'UNIT' collecting the turnstile data was the dummy variable I used in my feature set.*

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.

*Ans – First I tried to look for weather phenomena that will make subway travel preferable for travelers.*

1) *One such is 'rain', if it is raining outside, people are more likely to choose subway.*
2) *'meantempi' – gives an indication of the average temperature. Any extremes in temperature will make subway travel more lucrative. If it is very cold or very hot, more people will like to take the subway*

*Besides these indicators, I also tried to include time indicators which influence usage of subway*

1) *'Hour' – It is more likely that subway is used more during certain time of day, say during evening when people are making their way back home from office. I also included $hour^2$ and $hour^3$ to increase $R^2$.*
2) *'day_of_week' – More people are likely to use the subway on weekday than on weekends*

*All these intuitive choices were then validated with experiments. Following are the $R^2$ values with different feature set*

| Feature | $R^2$ value | Comments |
|---|---|---|
| 'UNIT', 'rain' | 0.427464693925 | $R^2$ value when feature set included only 'UNIT' and 'rain' |
| 'UNIT', 'rain','meantempi' | 0.428664209452 | The $R^2$ values increased further shows including temperature led to better prediction |
| 'UNIT', 'rain','meantempi',<br><br>'hour', 'hour$^2$', 'hour$^3$' | 0.4713221298 | The considerable increase in $R^2$ values indicates that time of day does play a role in ridership |
| 'UNIT', 'rain','meantempi',<br><br>'hour', 'hour$^2$', 'hour$^3$'<br><br>'day_of_week',<br>'day_of_week$^2$' | 0.483087572639 | The $R^2$ values increased further shows including weekday led to better prediction |

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

| Feature | Weigths |
|---|---|
| 'rain' | 10.92 |
| 'meantempi' | -11.95 |
| 'hour' | 326.45 |
| 'hour$^2$' | 196.01 |
| 'hour$^3$' | -61.99 |
| 'day_of_week' | 30.76 |
| 'day_of_week$^2$' | -281.76 |
| y-intercept | 1061.09 |

*All the values were generated using the gradient_descent.py program contained in the submission*

2.5 What is your model's $R^2$ (coefficients of determination) value?
*Ans - $R^2$ using gradient descent = 0.48*
*$R^2$ using OLS = 0.50*

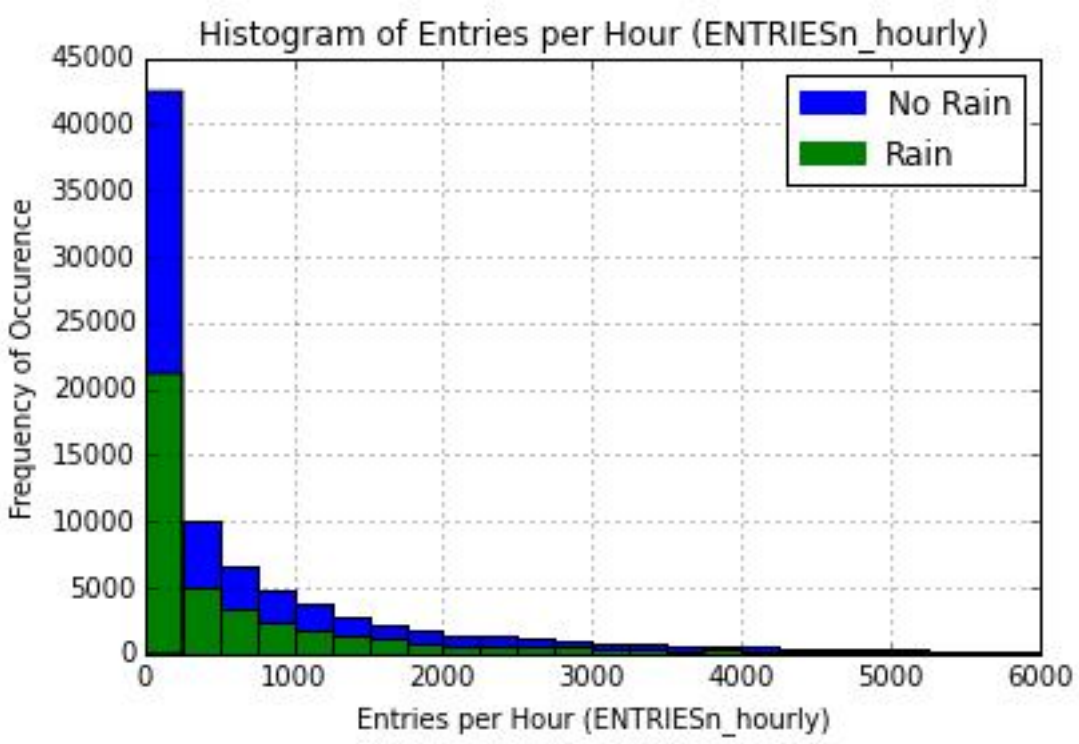
2.6 What does this $R^2$ value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this $R^2$ value?
*Ans - $R^2$ is generally interpreted as the measure of response variation that can be explained using the features used in our model.*

*I achieved an $R^2$ value of 0.50 using the OLS method. This indicates that features used in my model can explain 50% of the variation present in the 'ENTRIESn_hourly' variable. The rest 50% is present in some lurking variables that are not included in my feature set.*

*An $R^2$ value of 0.50 is not that good as a lot of variation (50%) in the 'ENTRIESn_hourly' variable is still not captured. Hence the linear model developed to predict ridership is not appropriate and either needs to be augmented to increase the $R^2$ value or else we can go for non linear fitting of data to get a good predictive model.*

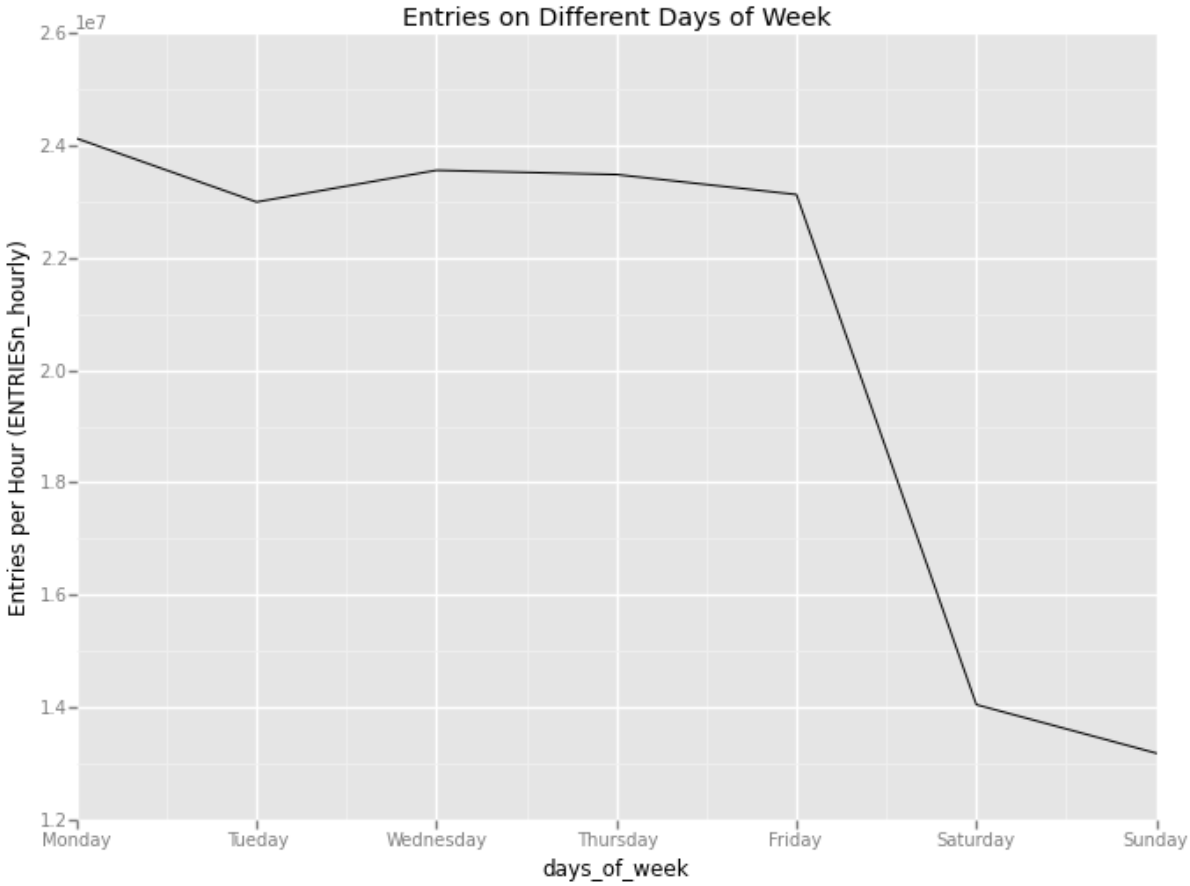
# Section 3 – Visualization

### 3.1 Histogram of entries on non-rainy days and rainy days




**IMP:** x-axis in this example image has been truncated at 6,000 cutting off values in the long tail which extends beyond 30,000

**3.2 a) Ridership by day of week**



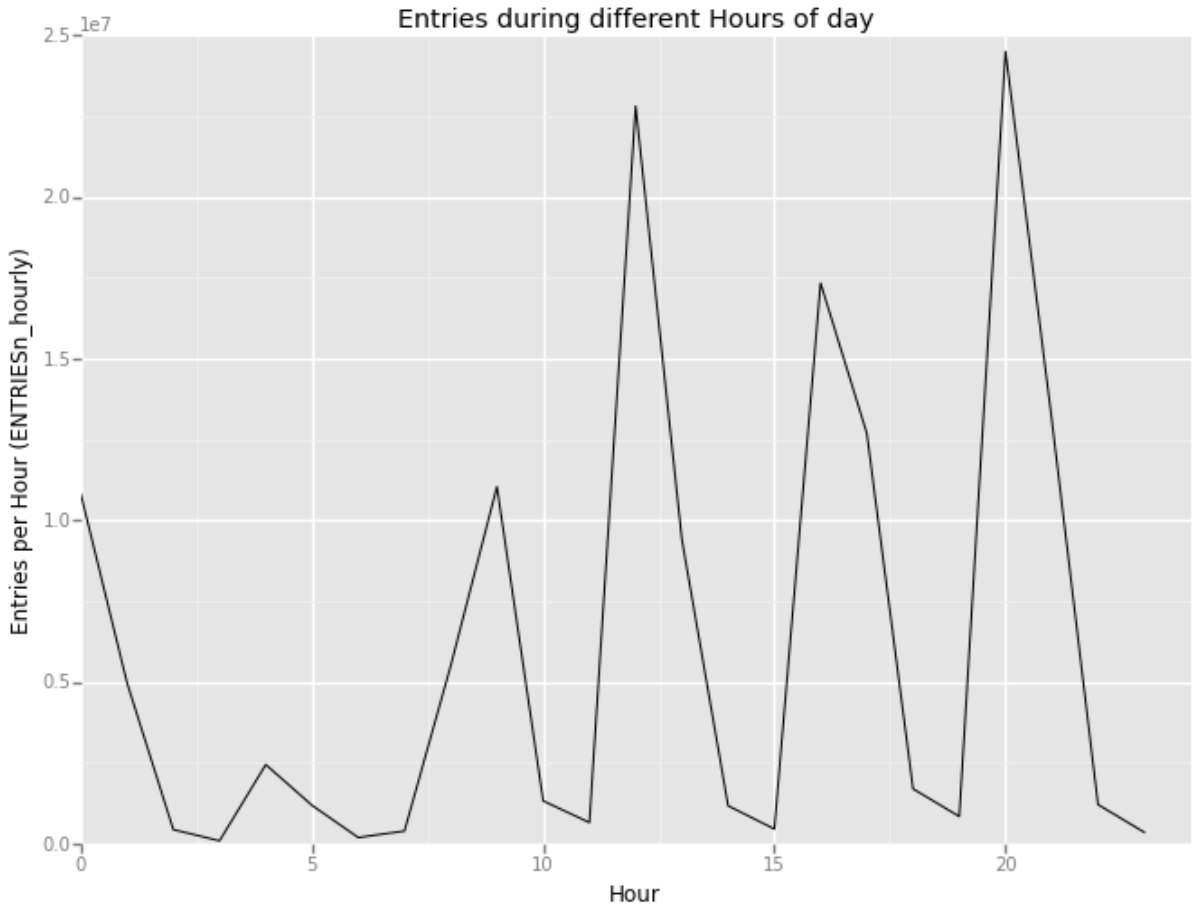

*Few observations:*

1) *Fewer people take subway ride on the weekends than on weekdays. So 'days_of_week' affect the ridership behavior of people. This reaffirms our intuition of using 'days_of_week' as a feature in our model to predict ridership*

2) *The day with highest overall ridership is Monday, with the rest of the weekdays showing almost similar ridership pattern*

3) *Since we are plotting ridership values over a period of time, a line plot appears more appropriate than bar plot.*

**3.2 b) Ridership by time of day**




*Few observations:*

1) *Fewer people take subway ride during 2am-4am period. So 'hour' affects the ridership behavior of people. This reaffirms our intuition of using 'hour' as a feature in our model to predict ridership*

2) *The 'hour' with highest overall ridership is 8pm (20 hrs), with a similar peak observed at 12noon.*

**3)** *Since we are plotting ridership values over a period of time, a line plot appears more appropriate than bar plot***.**

# Section 4 – Conclusion

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?
Ans – *More people ride the NYC subway* ***when it is 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
Ans – *There are numerous evidences which show that more people use the subway when it is raining*

1) *During the Mann-Whitney U test we calculated the mean of entries for rainy and non rainy days.*
   *Mean (rainy days) = 1105.45*
   *Mean (non-rainy days) = 1090.28*

   *These values show that on average, more people took the subway during rainy days*

2) *In the Mann-Whitney U test, the p-critical value was 0.05. The two tail p-value generated during the test was 0.05. So p-value = p-critical. This shows that the distribution of number of entries on rainy days and non-rainy days is statistically different. Hence the difference in mean values observed above is statistically significant and could not have happened because of chance.*
3) *While developing a linear predictor for 'entries' using gradient-descent method*

   *$R^2$ is 0.43013 when using 'day_of_week', 'UNIT' as feature*
   *$R^2$ is 0.43023 when using 'day_of_week', 'UNIT', 'rain' as feature*

   The increase in $R^2$ value indicates that 'rain' feature can explain some variation in the 'entries' data which was not being captured by 'UNIT' and 'day_of_week' alone. This variation is the increase in number of ridership during rainy days.

# Section 5 – Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset – **'turnstile_data_master_with_weather.csv (all discussion is w.r.t to the dataset used in the above file. It was used throughout the course**.),

*Ans – Although the dataset has lot of parameters related to weather phenomena, time and location attributes which could affect ridership patters, following are some of the drawbacks in my opinion*

1) *'rain' and 'precipi' contain almost identical information. Wherever 'rain' is 1, the 'precipi' column contains 0.5. Hence 'precipi' plays no role in increasing $R^2$, if 'rain'*

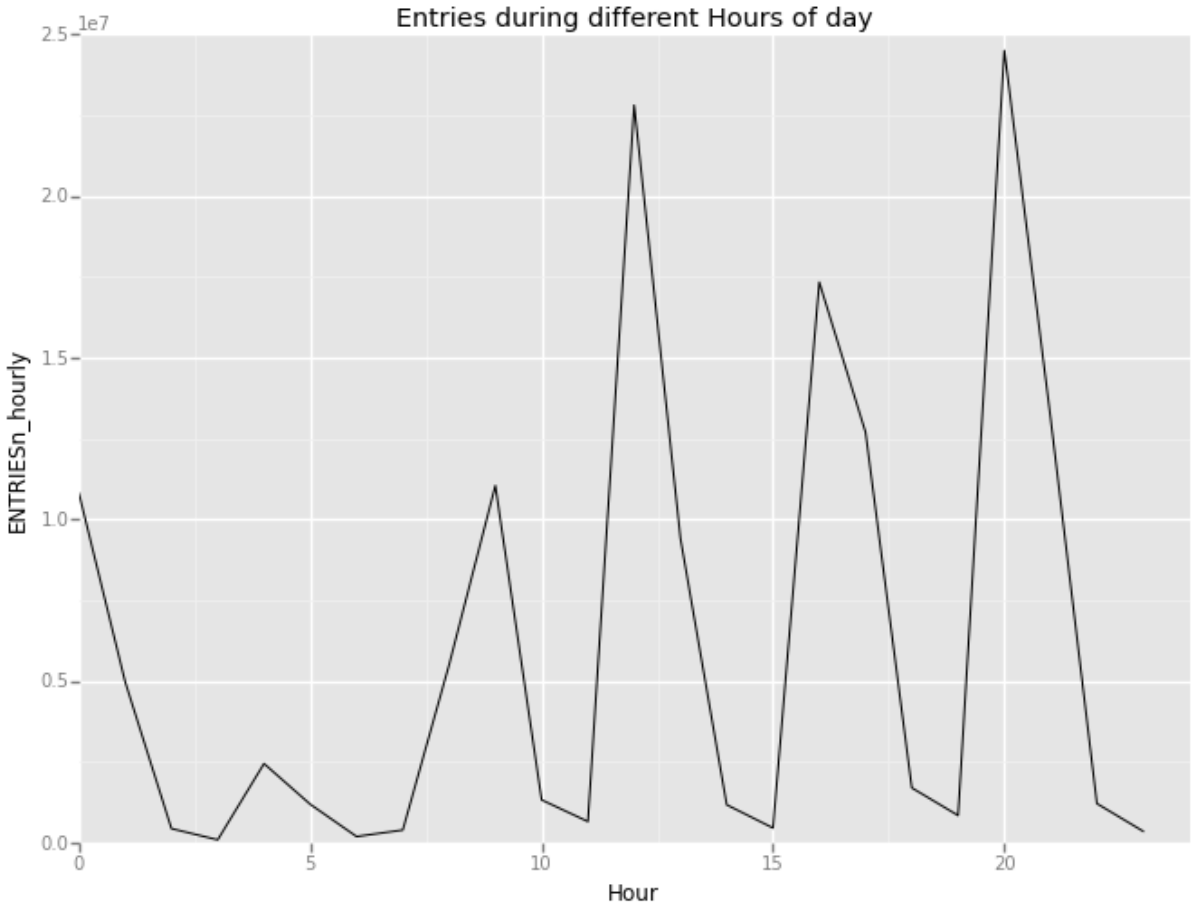
*has already been included in feature set. Intuitively the more it rains the more people should choose subway for travel. Yet the limited information in 'precipi' meant that this relationship could not be established during linear regression modelling.*

2) *The number of days where 'rain' column is 1 is much less compared to the number of non rainy days. Naturally more people would take subway on rainy days, so more data with rainy days would have helped in developing a better predictive model*
3) *Columns like 'DESCn' have just one value – 'REGULAR', hence they provide no discriminatory value with respect to predicting ridership.*

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

*Ans – In order to predict the ridership on subway, I used two techniques, one based on gradient descent and the other based on ordinary least squares. Following are some of the drawbacks in my opinion*

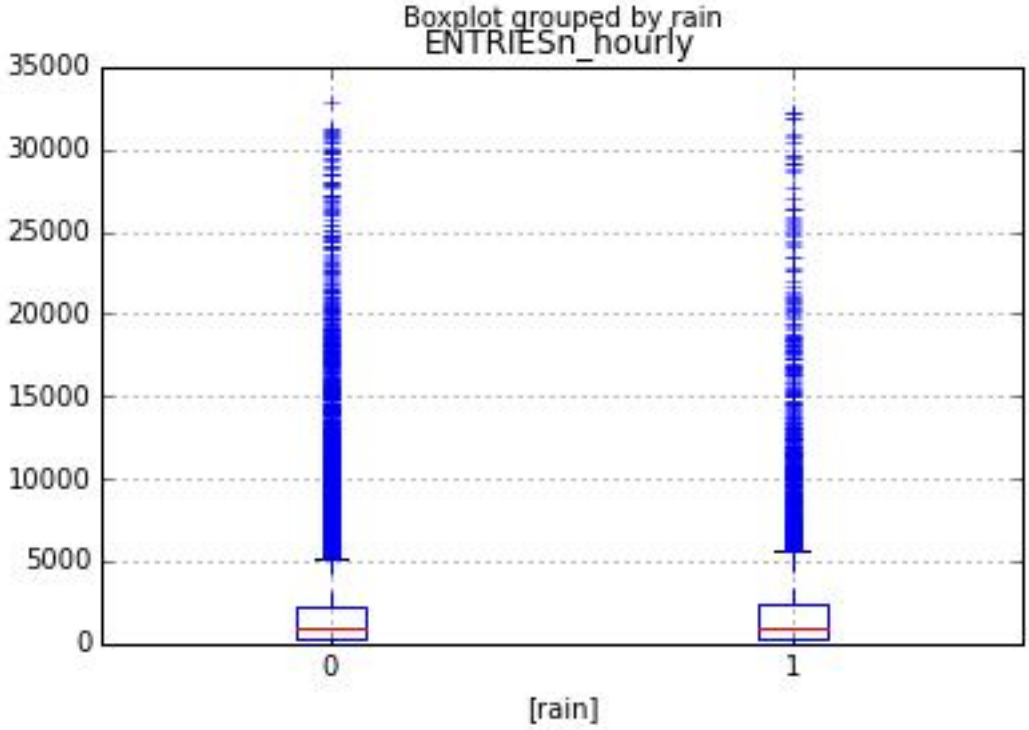
1) *Below is the graph showing ENTRIES w.r.t to the 'hours'. It clearly shows that 'entries' does depend on 'hours' yet a linear relationship between 'entries' and 'hours' does not exist. To overcome this I used higher order features such as hour$^2$ and hour$^3$. However the true relationship appears to be non linear and hence a non linear model would be better able to capture the relationship. Similarly there are other data terms which do not show linear relationship between them and 'ENTRIESn_hourly'. So using linear regression when linear relationship between independent variable and dependent variable is not that strong is a drawback.*




2) $R^2$ *using gradient descent = 0.48*
$R^2$ *using OLS = 0.50 (both gradient descent and OLS used same features)*

*These two data show that the OLS was able to find a better solution than gradient descent. This might happen because gradient descent can get stuck in local minima and never reach globally optimal solution. Getting stuck in a local minima is a major drawback of gradient descent scheme.*

3) *Linear regression is sensitive to outliers and the dataset contains some outliers which would become apparent from the box plot below. Presence of outlier reduces the effectiveness of linear regression and hence constitutes a drawback according to me.*


Boxplot grouped by rain
ENTRIESn_hourly
35000
30000
25000
20000
15000
10000
5000
0
0
1
[rain]