Section 0:

http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/
google.com
Stackoverflow.com
http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms
http://docs.ggplot2.org/current/

Section 1.Statistical Test

1.1)The hypothesis test chosen here is Mann-Whitney U-test.Here we choose this test because the samples are not normal as shown in histogram later.
Mann-Whitney U test is non parametric.A non parametric test does not require a particular distribution of shape
Null hypothesis: Probability that a no of people riding the subway on rainy days drawn at random has a greater value than the number of people
Riding the subway on non rainy days drawn at random =0.5.
P(entriesn_hourly(rainy)>entriesn_hourly(non-rainy)) = 0.5

Alternative Hypothesis:Probability that a no of people riding the subway on rainy day drawn at random has a greater value than the number of people
Riding the subway on non rainy day drawn at random !=0.5
P(entriesn_hourly(rainy)>entriesn_hourly(non-rainy)) != 0.5

Test chosen:Two tail test.
P critical =0.05(0.025 on each side)

1.2) Mann-Whitney U-test. This test is applicable since the distribution is not normal and assume that it is non parametric.

1.3) x(rain)=1105.4463767458733 x(without rain)=1090.278780151855
Mann -whitney u statistic=1924409167.0 p- value=0.049999825586 (i.e 0.024999912793489721 on each side)
Pcritical=0.05 (i.e 0.025 on each side).

1.4)
Reject null hypothesis, Probability that a no of people riding the subway on rainy day drawn at random has a greater value than the number of people
Riding the subway on non rainy day drawn at random !=0.5. Since p<p-critical.

Section 2. Linear Regression:

2.1 a)OLS using Statsmodels

2.2) List of features are(const, rain, precipi, Hour, meantempi, unit_R001, unit_R002,unit_R003,unit_R004,unit_R005,unit_R006,.....unit_R552)
The model used a dummy variables

2.3) list of features chosen:rain, precipi, Hour, meantempi
1)rain may influence the number of people riding the subway when it is raining people tend to choose subway
2)Temperature may also influence the number of people riding the subway.If it is too hot or too cold the number of people who use subway might increase
3)the hour might also increase the number of people using the subway.If the road has high traffic at

a particular time of day people might use the subway more at that point of time.

4)The amount of precipitation also might influence the number of people using the subway.

5)unit_R001....unit_R552 are dummy variables

6) const is a constant included to calculate the intercept.

2.4)intercept/const:1539.126

Parameters:

Rain:29.4645

Precipi:28.72

Hour:65.334

Meantempi:-10.531

Unit_R001 :4078.93811

....

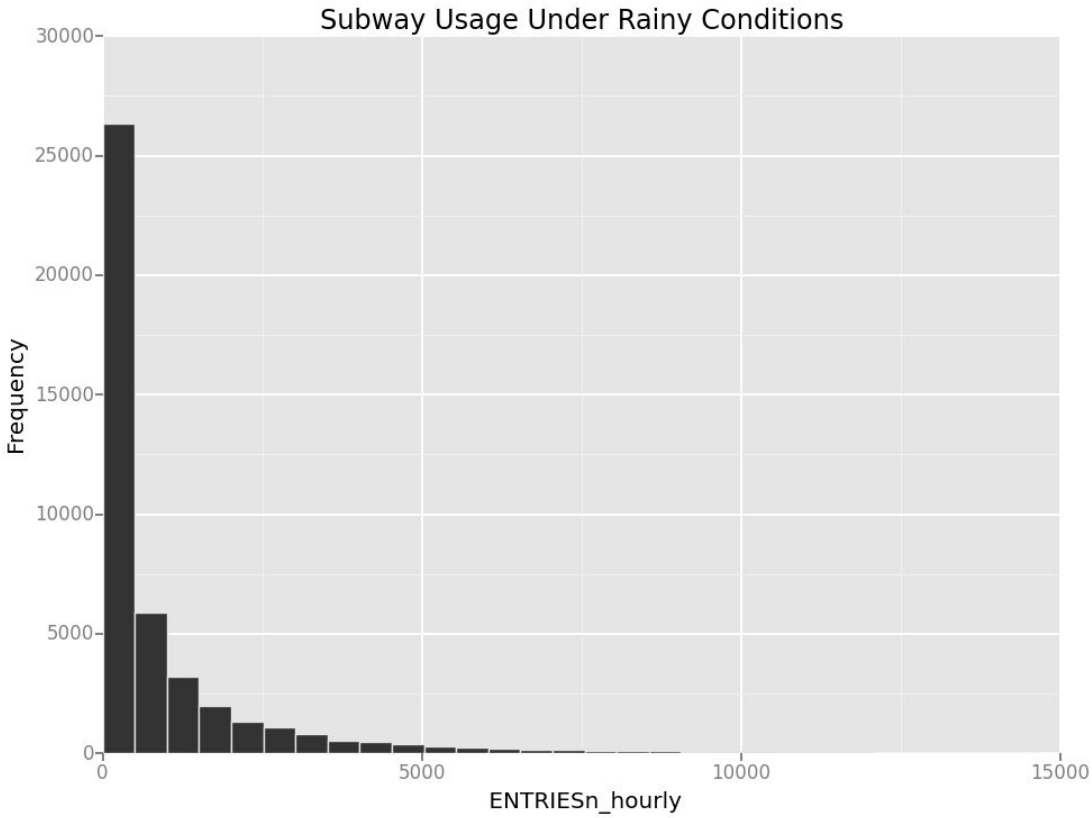
Unit_R552:-1458.5951

2.5) r^2 value =0.47924770782

2.6) R^2 is a statistical measure of how close to data are to the fitted regression line.It is known as coefficient of determination.I think the linear model is not good to predict the ridership given the above value of r^2

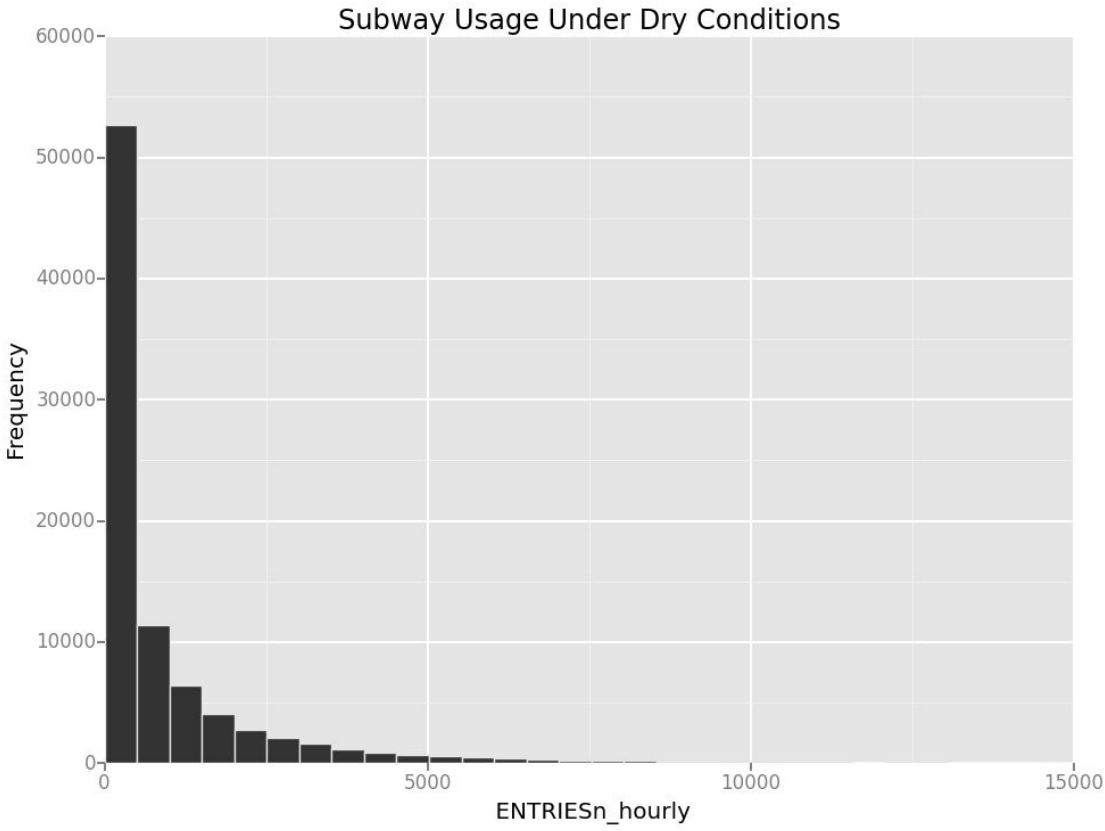
Section 3:

3.1) scale(x axis 1 unit=5000 entries per hour ,y-axis:1 unit=5000 frequnecy)



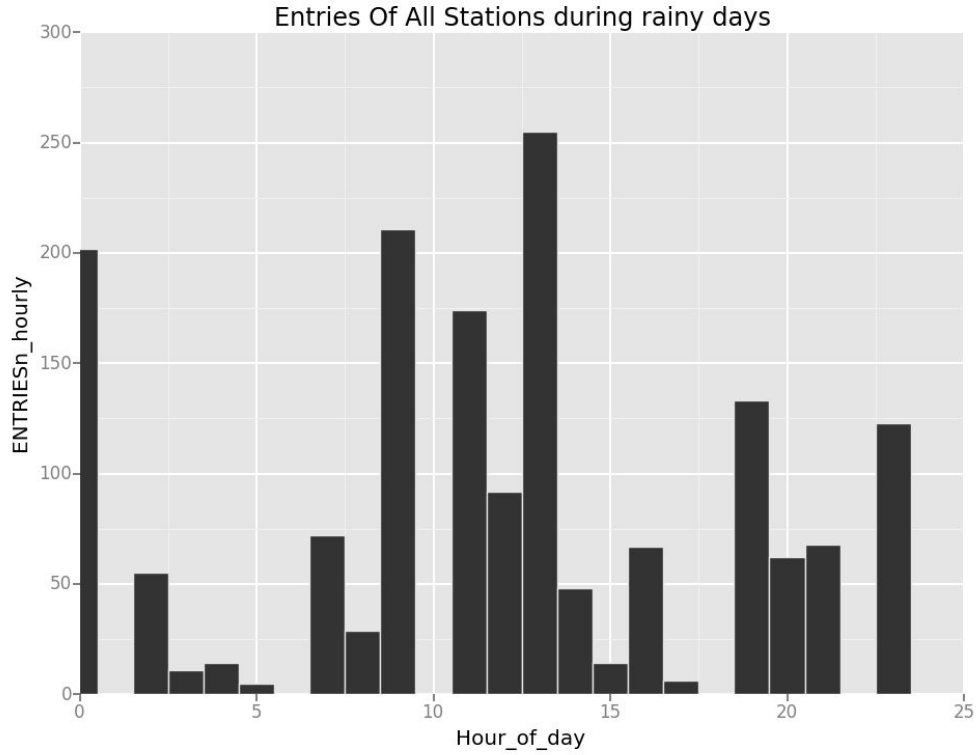
Maximum frequency is about 26000 for (0 -500) entriesn_hourly

scale(x axis 1 unit=5000 entries per hour ,y-axis:1 unit=10,000 frequnecy)

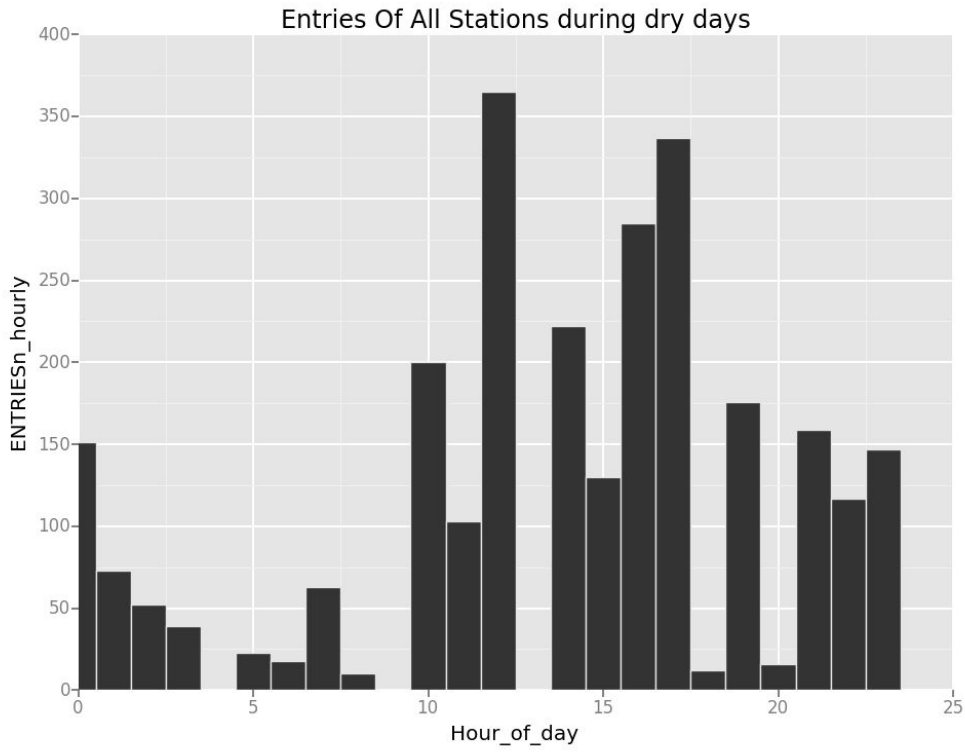


In this case the maximum frequency is about 52000 for (0 -500) entriesn_hourly

3.2)



Maximum number of entries is about 250 at 13:00
There are no entiries at 1:00,6:00,10:00,18:00 and 22:00



Maximum number of entries is about 360 at 12:00
There are no entiries at 4:00,9:00,13:00

Section 4:

1)From the statistical test it implies probability that a number of people riding the subway on rainy day

drawn at random has a greater value than the number of people
Riding the subway on non rainy day drawn at random !=0.5.
2)The co effieicnts obtained in linear regression shows that the ridership of subway depends mostly(i.e directly proportional)
on the hour of the day and also factors such as rain and precipitation.The co efficients of rain and precipitations are almost the same
and hour of day form an important factor for people to ride subway.Subway ridership also depends on the mean temperature which is a
negative co efficient which means the ridership is inversly proportional to mean temperature(meantempi).From the co-efficients obtained
from linear model it is evident that more people ride the subway on rainy days since the co effcient is positive.

Section 5:
5.1

1) Let us suposse we denote a ratio (R)=(No of people using the subway(Ns)/No of people who are out(No))

No: here denotes the total number of people using the subway and the number of people who are out.

Here the data does not represent the number of people who are actually out.I mean when they are weather conditions like rain or more precipitation
people may not go out so the number of people using the subway might decrease but if we compare the ratio of number of people who are
are using subway during rainy days(R rainy-days) may be greater than (R non-rainy days) so we cannot perfectly conclude that the number of
People riding the subway during non-rainy days are greater than the number of people riding the subway in rainy days unless there is some more
data about the number of people who came out(here out means came out of houses and used other means).

2)

If there are too many features then the solution is not unique and the performance decreases and it becomes difficult to compute.

Least squares method cannot handle noise in independent variables and the choice of features also influences the output.