Section 1

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?

Mann-Whitney U-test was used. Null hypothesis was that both the means (for distributions with and without rain are equal). I would like to keep alternative hypothesis as: mean with rain is greater than mean without rain.
Since alternative hypothesis is one sided, we would use one tailed p-test. Using scipy function, will give desired one sided value.
(Reference: http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html)
Typical p-critical value is used: 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.

This statistical test is valid for non-normal distributions, so it is valid for our data.
It assumes number of samples is more than 20. Also assumes that samples should be independent of each other. These are true in our case.

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.

Here are the results of the test:
Mean with rain = `1105.4463767458733`
Mean w/o rain = `1090.278780151855`
p-value of the test = `0.024999912793489721`

1.4 What is the significance and interpretation of these results?

It can be concluded that if the two distributions came from same population, there is 2.4% probability of getting a set that has mean with rain this much larger than mean without rain. Since the probability is less than critical value of 5%, it is safe to reject the null hypothesis. That is - it is safe to conclude that ridership during rain is more than ridership without rain.

Section 2

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?
4.

I used gradient descent same as taught in lesson 3 exercise.

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

Input variables used were: rain, fog, precipi, hour
Yes dummy variable is also used as a feature.

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.”

Reasons for choosing these parameters were mostly logical. I started with the four parameters are indicative of bad weather "rain, fog, precipi, thunder". Being a car driver I know that is any of these conditions are worse, people would prefer to not go by themselves but prefer a public transport like Subway. "thunder" was removed since script reported that it's value does not change. Also I added "hour", since ridership was very much dependent on hour of the days i.e. rush hours of the day should logically have more passengers than non-busy hours.
R^2 value that came out of putting these four parameters was quite above the minimum value expected in the exercise (0.2), so I assumed it was a good combination of features.

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

'rain' = -10.765284733091921
'precipi' = 16.995444308152365
'Hour' = 463.69446810289855
'fog' = 22.138191793522655

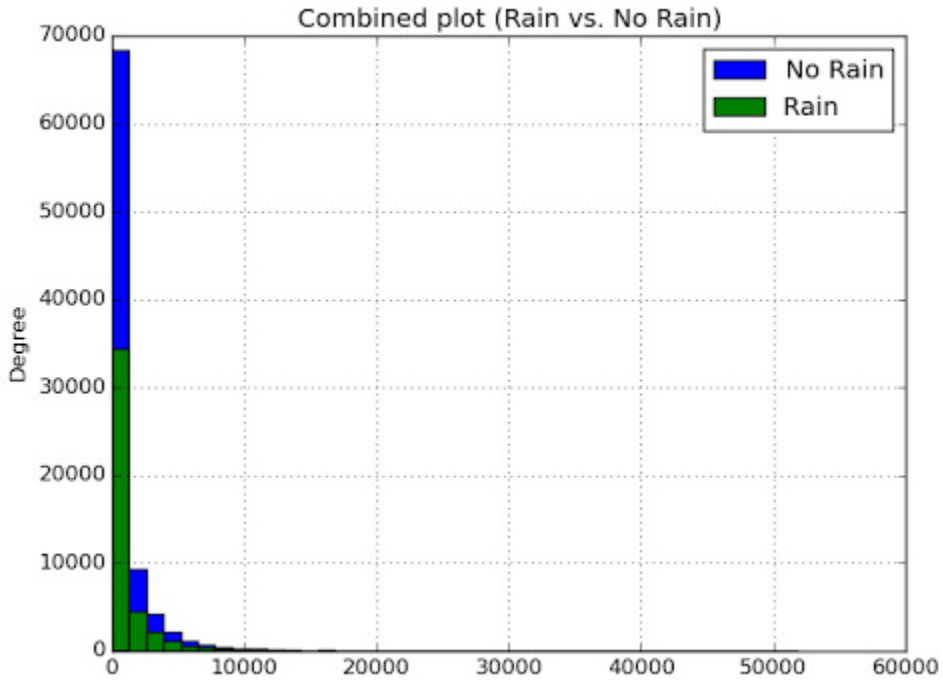
2.5 What is your model’s R^2 (coefficients of determination) value?

R squared = 0.457707

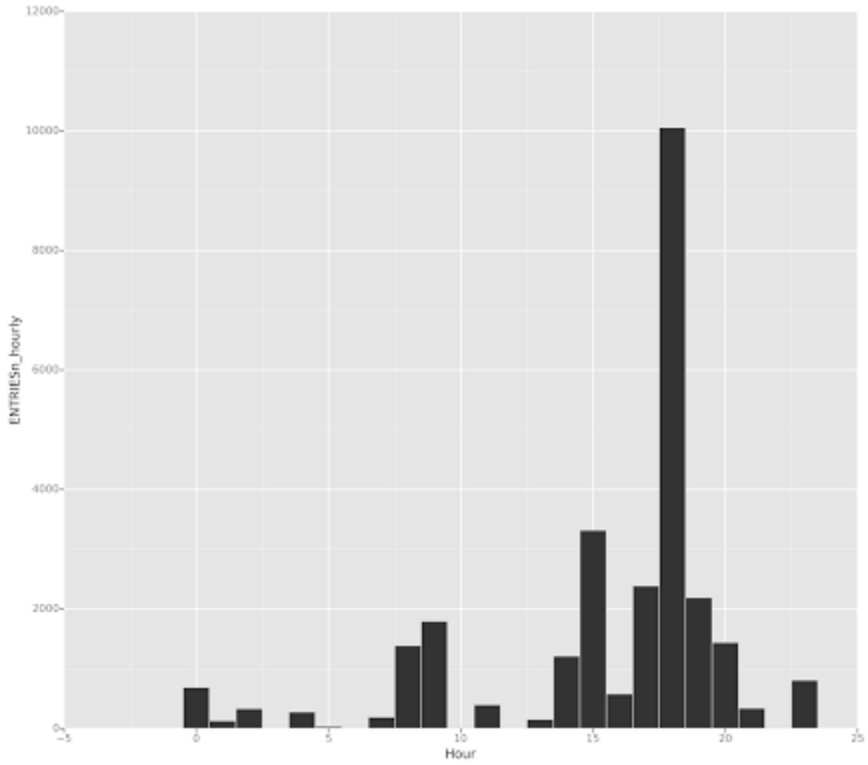
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

Closer the R^2 value to 1, better the model. This is definitely not close to 1 and there must be room for other prediction methods that bring the value closer to one. However, the minimum value condition of 0.2 is met by this particular model. So, I believe, it is a good enough model to start analysis with.

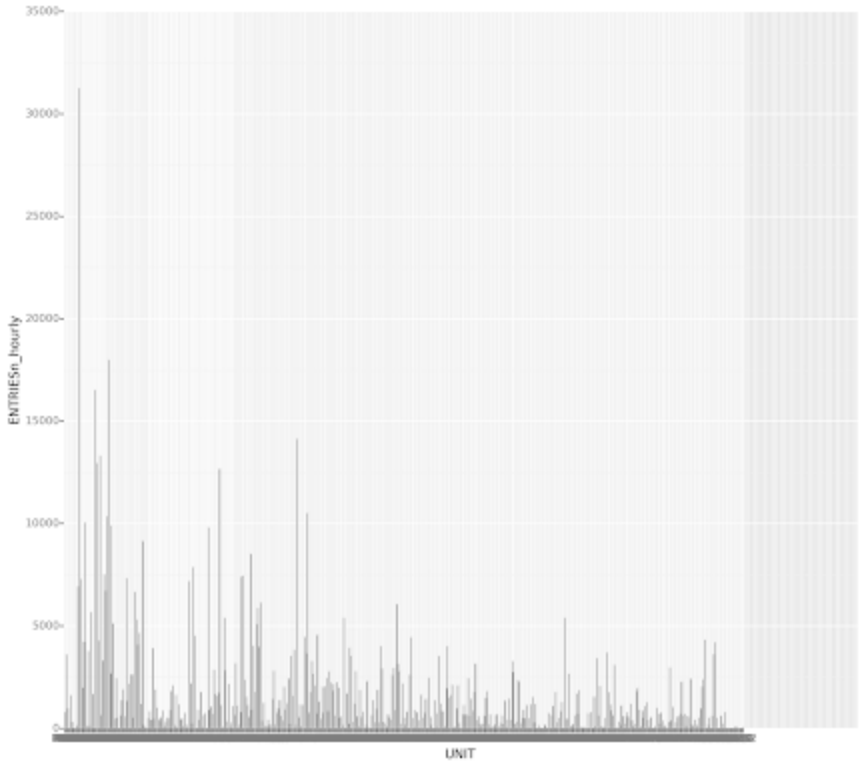
Section 3



3.1 This graphs indicates that the rider data distribution is non-normal. This justified the use of statistical test. It also indicates that the mean of 'No Rain' data is lower than the 'Rain' data



3.2 a Ridership is heavily dependent on hour of the day. Ridership is appears to be high in morning office hours and reaches maximum values in evening hours, peaking at 6 pm.



3.2 b The important/possibly downtown or high population areas have more ridership than others. This accounts for a scope of whole lot of more detailed analysis on the reason behind more ridership in the busy stations. (by collecting population stats or by analyzing their location using data given)

Section 4

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 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.

Mean ridership with rain is more than that without rain. We also applied statistical test with which we could conclude that it is very less likely that such a large difference in mean could occur from two data sets coming from same kind of population.
Also linear regression coefficient of 'precipi' is positive indicating that increased rainy condition leads to more riders.

Section 5

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

Dataset - I used original data set for simplicity and to keep it same for all the results. So it was slow to get the results. I read in forums that improved data set has some rows removed. That could have made getting results fasters, esp. linear regression experiments.
Also the added columns in improved dataset might have given more options of analysis with graphs. But I am still somewhat assured with linear regression results, from data set point of view.

Analysis - For linear regression, I tried to remove 'rain' but still got a pretty good R-square value. 'rain' has a negative coefficient. If we increase the value of rain from 0 to 1, I would expected ridership prediction to go up. 'precipi' column's coefficient indicates that as well. I am not sure but I think this might be an indication that dependence of values on one or more features might not be linear. We could use better models that linear regression and gradient descent (possibly!)