Section 1

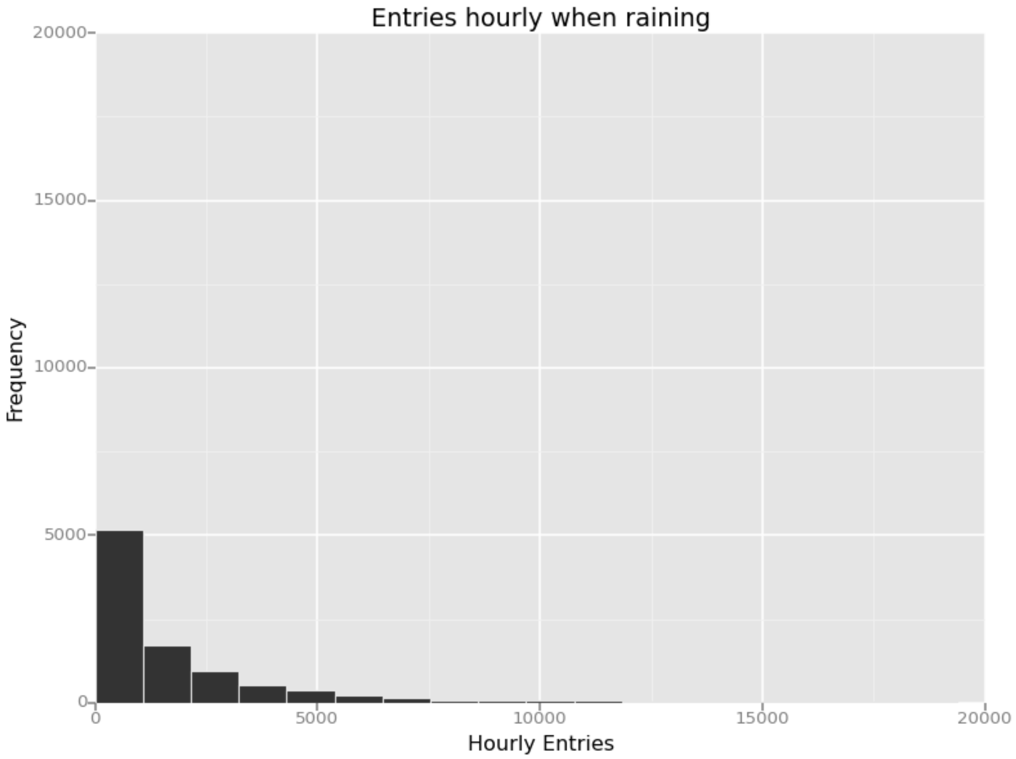
1. **Mann-Whitney's U-Test** was used to analyse the NYC subway data. We wanted to know if raining would cause number of hourly entries to decrease or even increase, thus **two-tailed P-value testing** was used. The null hypothesis here is **"Raining does not have an effect to number of hourly entries."** while alternate hypothesis is **"Raining does increase / decrease the number of hourly entries."**. P-critical value used here is **0.05**.
2. Mann-Whitney's U-Test is applicable to the test since, unlike Welch's t-test, **it can be used for non-normalised data**, in other words it does not assume the distribution of ridership is normal.
3. The result I got from this statistical test was:
    a. p-value = **0.025**
    b. Mean of ridership when raining: **1105.45**
    c. Mean of ridership when not raining: **1090.28**
4. Since the p-value is **positive 0.025**, which is **smaller than p-critical value 0.05**, we can conclude that raining does cause a change in subway ridership. The change is of positive direction, as suggested by **mean of ridership when raining > when not raining**.

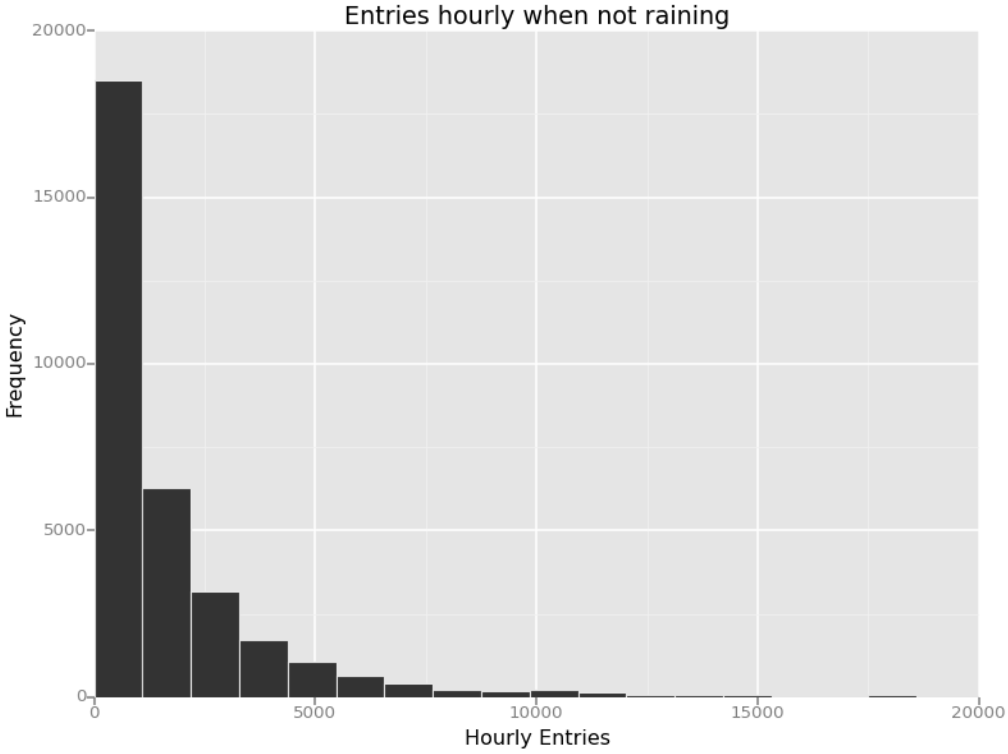
Section 2

1. **Linear regression with gradient descent** and in optional training 3.8 **OLS with stat models** were used to compute the coefficient theta and produce prediction.
2. After trying out adding all available features, I decided to go back on using the four original features: **'rain', 'precipi', 'Hour', 'meantempi'**. Dummy variables are used as part of the features; They are stored in **UNIT** field of given data.
3. By adding all features, r-squared increased from **0.464** to **0.466**. As they did not drastically improve the prediction, I reverted to using the original four features.
4. 2.92398062e+00, 1.46526720e+01, 4.67708502e+02, -6.22179395e+01
5. 0.463968815042
6. It means the model will not be able to predict the future values well. Linear model is **not appropriate** for this dataset.

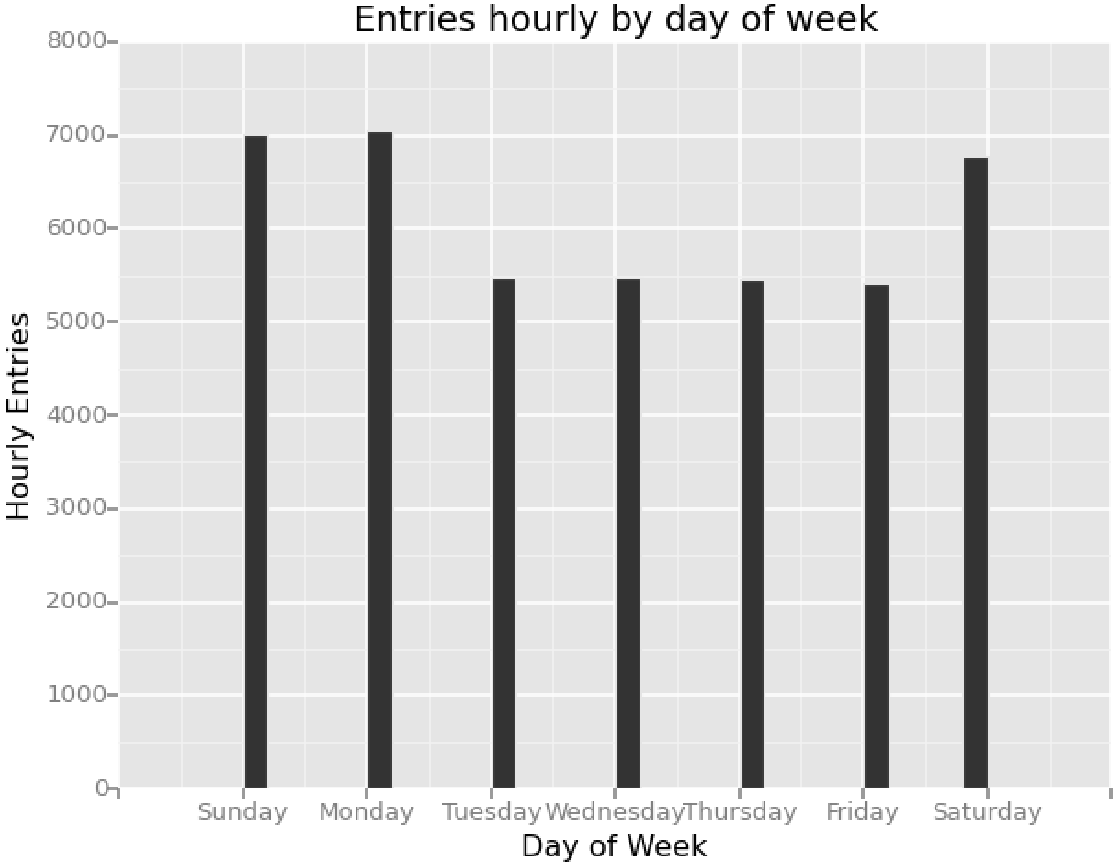
Section 3

1. Following are the histograms:





2. Entries by days of week:



Section 4

1. People seemed to ride NYC subway when it is raining, as shown in the analysis provided in section 1, where mean of ridership when raining is **1105.45**, larger than the mean of ridership when not raining (**1090.28**). But even with that information we still could not conclude the hypothesis that people do ride NYC subway more when it is raining is correct. We then conducted a Mann-Whitney's U-test to ensure if that hypothesis is indeed likely to be true.

   From the test we get that the p-value is **positive 0.025**, which is **smaller than p-critical value 0.05**, and thus we can conclude that **null hypothesis can be rejected**, in other words raining does cause a change in subway ridership. The change is of positive direction, as suggested by **mean of ridership when raining > when not raining**.

2. From statistical test the explanation is as described above. From linear regression, we found out that the weight for **rain** feature was **2.92398062e+00**. Since this is a positive value, we can be sure that rain does positively affect ridership i.e. all other things being equal, when it is raining we can expect ridership to be larger.

Section 5

1.1. Unfortunately the dataset does not account for more demographic information like how many people are there around the area of each UNIT, and the age of that population. That demographic information could be useful since one can always expect that number of ridership within a certain area is larger when that area contains more population of people old enough to ride trains.

    Should this be a more serious project, latitude and longitude can be used to find out the location and we should be able to find number of population in the radius of several hundred meters around each UNIT position. We could do that by combining the data of this dataset with another dataset containing population per area.

1.2. Linear regression is not a good tool to model this dataset, as shown by the low Coefficient of determination value. Polynomial regression would likely yield a better fit to the data.

    Personally, in addition to demographic information as stated above I would enter the features into a Support Vector Machine or Artificial Neural Networks to predict whether a combination of features may result in a ridership or no ridership, and maybe use Bayesian Network to get more understanding of the features used in this dataset. Although I am not quite familiar with the latter concept.

2. I hope we'd get to learn more about Bayesian Network later in this Nanodegree program!