Section 1

1.1 I have used Mann Whitney U test, since the distribution is not normal probability distribution

And used two-tail P value because the aim is to tell if rider number on rainy and non-rainy days are from the same distribution or not

The hypotheses
Is there a difference between the sample mean of number of entries per hour in rainy days and no rain days
The null hypotheses X : $P(X > Y) = P(Y > X)$ the distribution of both populations are equal

The alternative hypotheses X : $P(X > Y) \neq P(Y > X)$ the distribution of both populations are not equal

P-critical value used is .05 at confidence level of 95%

1.2 The test used is suitable because it does not assume that the data is drawn from any particular probability distribution

1.3

```
        With_rain_mean         = 1105.4463767458733
        Without_rain mean      = 1090.278780151855
        U statistic            = 1924409167.0
        P value                = 0.024999912793489721
        P value * 2            = 0.049999825586979442
P*2 will be used as this is a two-sided t test
```

```
1.4
   The results indicate that the difference between the two means is not by
   chance they are actually from different distributions.
```

Section 2

2.1 OLS is used

2.2 Features used: 'rain', 'precipi', 'Hour', 'meantempi', 'meanwindspdi', 'fog'

Rain and fog are dummy variable, since they're either 1 or 0

2.3

Rain; I think when it's rainy people would decide to use the subway

Percipi; I thought the visibility would matter and people might prefer the subway over driving

Hour; it makes sense the during specific hours, like rush hours, people might want to avoid the traffic

Fog; if it's foggy people might prefer to use the subway instead of driving or using other transportations

All of the features selected had increased R^2 when added, as below

[['rain', 'precipi', 'Hour', 'meantempi', 'meanwindspdi', 'fog']]

0.480456675828

[['rain', 'precipi', 'Hour', 'meantempi', 'meanwindspdi']]

0.479687594832

[['rain', 'precipi', 'Hour', 'meantempi']]

0.47924770782

[['rain', 'precipi', 'Hour']]

0.478412154934

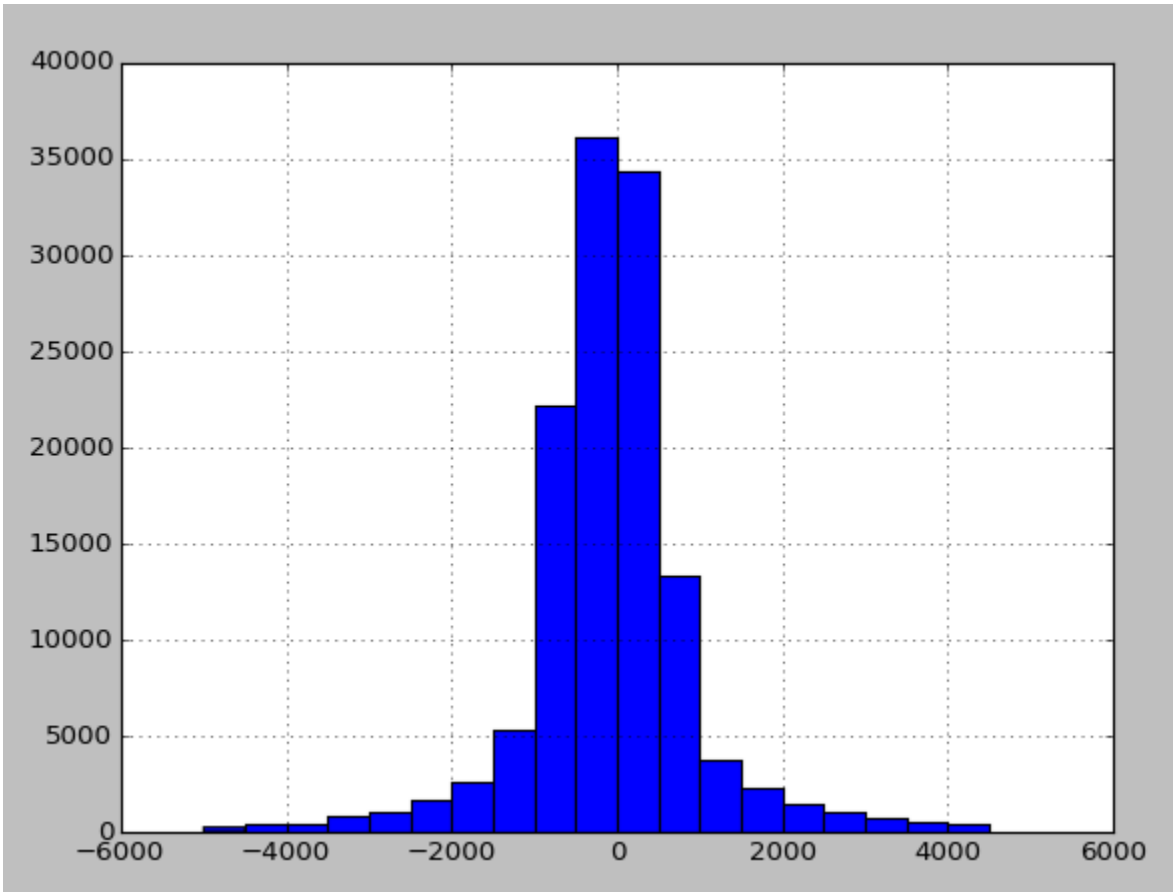
2.4 Params

```
rain            -32.267252
precipi         -22.618057
Hour             67.397395
meantempi        -5.912159
meanwindspdi     26.279924
fog             120.275809
```

```
2.5 R^2 = 0.480456675828
```

2.6 The R^2 value means that the percentage of the response variable variation explained is 48%

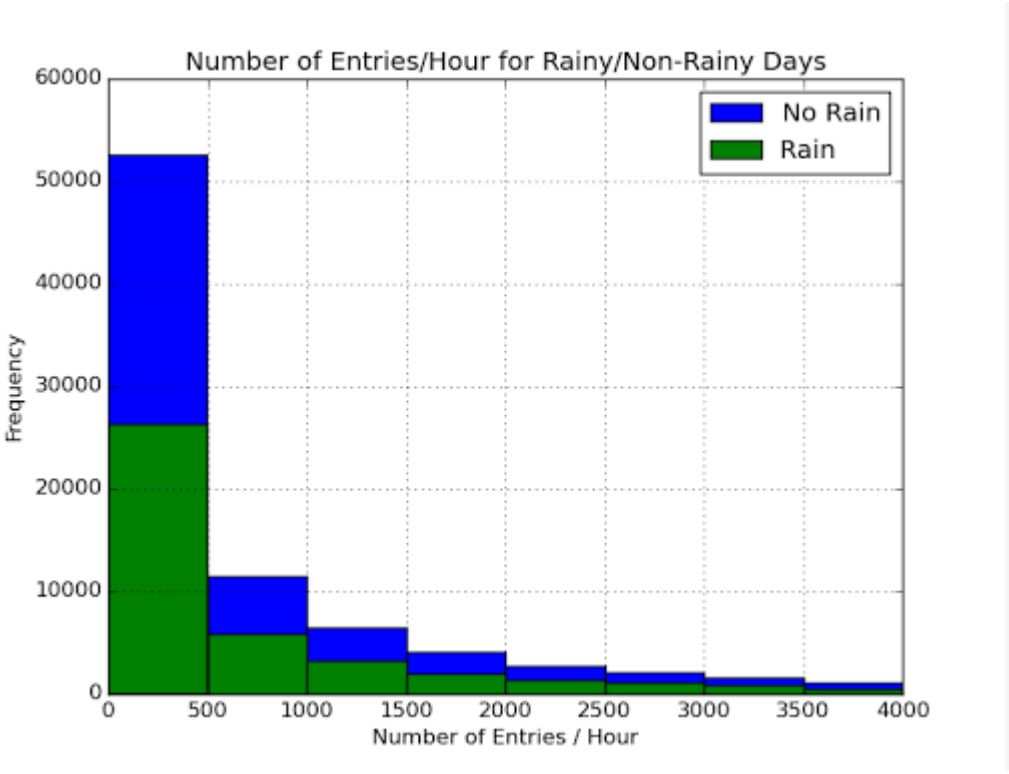
While this value is considered low, it is actually expected to obtain a low value since the model is trying to predict human behavior, in such circumstances it is typical for R^2 to be lower than 50%



Distribution of the residuals

The long tails indicate very large residuals which means that the distribution is not normal, which in turn suggests that the linear model is not appropriate for this dataset

Section 3



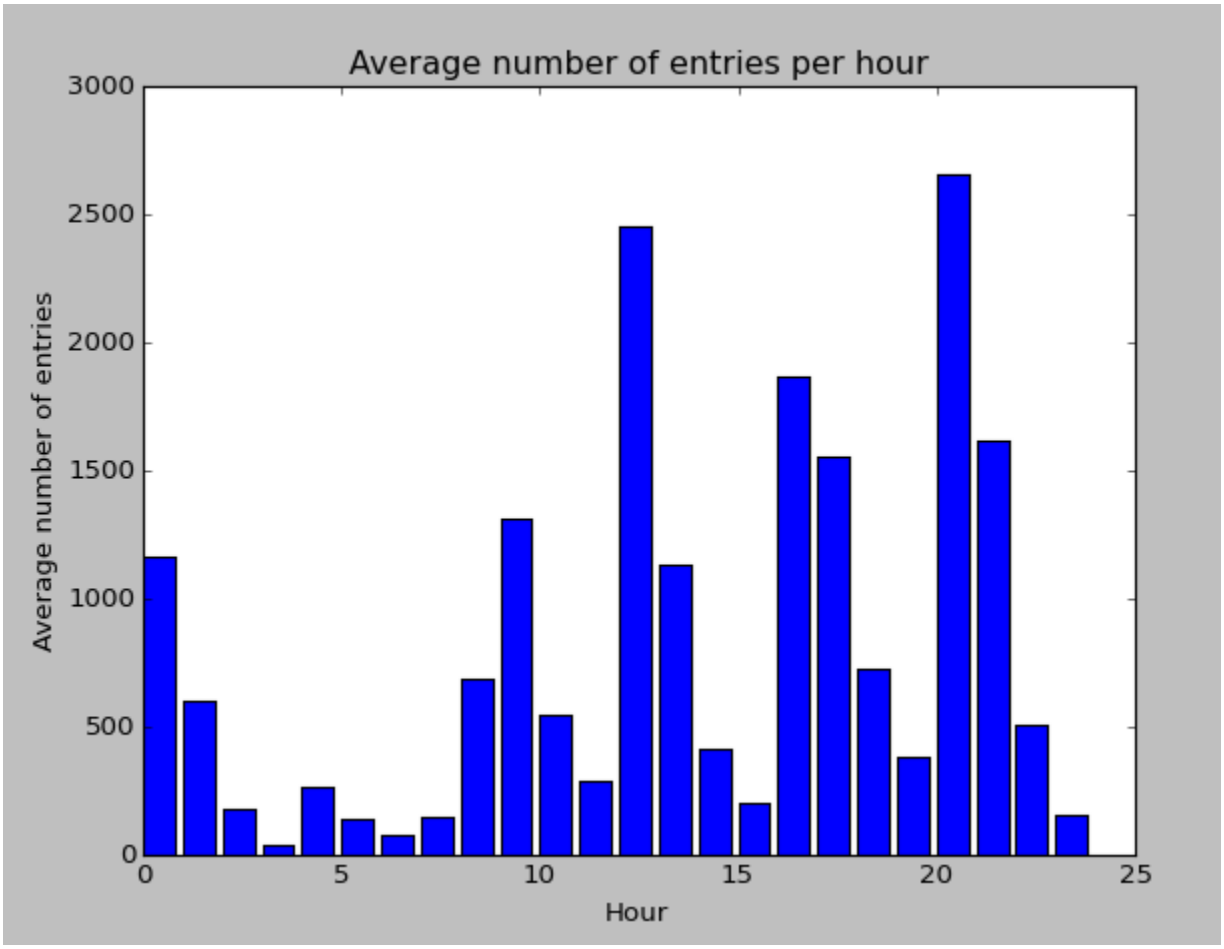
3.1

Distribution of number of entries per Hour

It's noted that the highest frequencies for both rain and no rain days are for number of entries below 500, this means that high numbers of entries occur only on specific times

Another observation is that the number of entries per hour is higher for no rain days across all frequencies

3.2



Average number of entries per hour

The above figure shows that the average number of entries peaks only at specific hours of the day (specifically 0, 9, 12, 13, 16, 17, 20, and 21)

Section 4

4.1 it seems that people ride the subway more on days when it's raining,

4.2 The statistical tests performed indicate more people riding the subway on rainy days, as the mean for entries with rain is larger than the mean of entries without rain, and the Mann Whitney U test confirms the difference, that means that the difference between the means is statistically significant.

The linear model coefficient of rain is negative; this would mean that with rain the number of entries decreases, however this model is questionable since there are residuals with large values as observed on the residuals histogram

Section 5

The linear model performance is questionable, as observed from the residual histogram it have large residuals.

The dataset has observation period covers only one month; this is short period to observe the weather (rain) effect, as maybe it wouldn't rain enough days in one month, also this would little samples to train the linear model.

Some variables in the dataset are related and would cause a collinearity issues, we should include only one of the correlated independent variables (i.e. min, max, avg, we would just use the avg)

Increasing the observation period would provide more data to train and therefor improve the linear model