INTRODUCTION TO DATA SCIENCE - PROJECT

SECTION 0 - REFERENCES

Below is a summary of the references that I used:

I used this to get a formula using numpy for squared differences:

http://stackoverflow.com/questions/2284611/sum-of-square-differences-ssd-in-numpy-scipy

I used the following to improve my understanding of the Mann-Whitney U test and how to interpret the results:

http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test

http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html

http://www.tqmp.org/RegularArticles/vol04-1/p013/p013.pdf

I used the following to help resolve some issues with my graphs (index not starting at 0):

https://github.com/yhat/ggplot/issues/382

SECTION 1 – STATISTICAL TEST

1.1 I performed the Mann-Whitney U test – two tailed. The null hypothesis was that both samples (number of entries with rain and the number of entries without rain) are drawn from the same population and so have the same distribution. The critical p value was just less that 0.05 (two-tailed).

1.2 The statistical test is appropriate because the data in non-normal – we now this from the plot done in Problem set 3.1.

1.3 The results from the statistical test are

Mean – with rain	1105.4463767458733
Mean – without rain	1090.278780151855
U	1924409167.0
p-value	0.024999912793489721 (one-tailed) 0.049999825586979442 (two-tailed)

1.4 As the p-value is less than 0.05 we reject the null hypothesis. If they 2 sample were drawn from the same population the probability of seeing a difference as significant as here is less than 0.05.

SECTION 2 – LINEAR REGRESSION

2.1 Gradient descent

2.2 Input variables used: ['rain','fog','precipi','mintempi','meanwindspdi','maxpressurei']

2.3 It makes sense to me that ridership would be heavily influenced by the weather. I started to build up my features by looking at the obvious ones first (rain, fog) and then adding more weather features (temp, pressure, wind speed) to see if they improved the r^2.

2.4 Thetas:

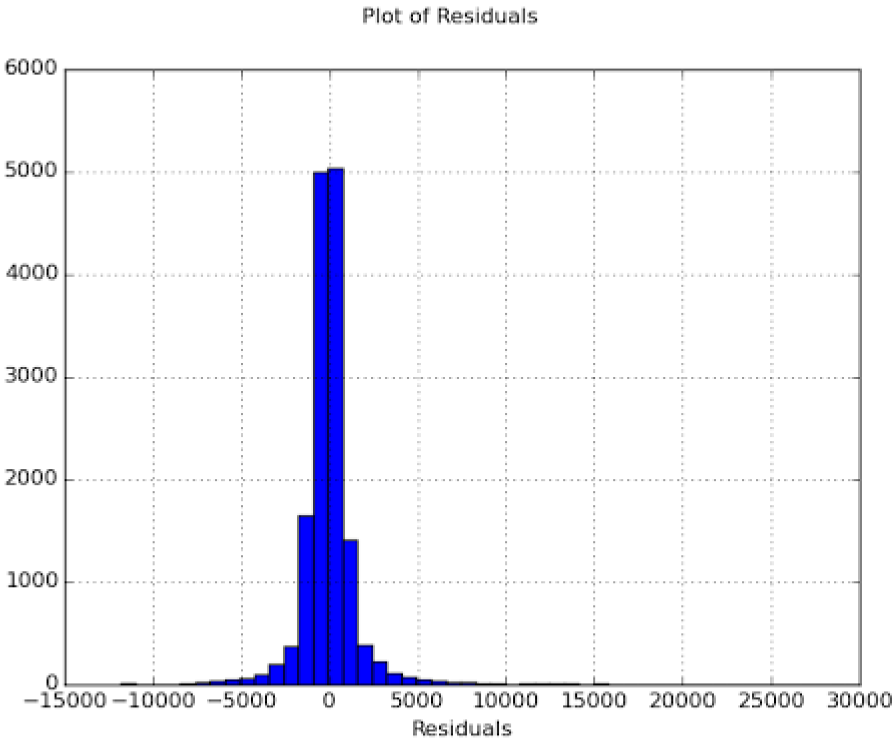
```
 -4.63642909e+00   5.17466459e+01  -1.03653650e+01  -6.80806385e+01
  4.76840828e+01  -4.13157757e+01
```

2.5 The coefficient of determination (r^2) is 0.42416963146.

2.6 The coefficient of determination is a measure of the goodness of fit of the model. It tries to provide a quantitative a measure of the how much of the variance in the underlying data is explained by the model.

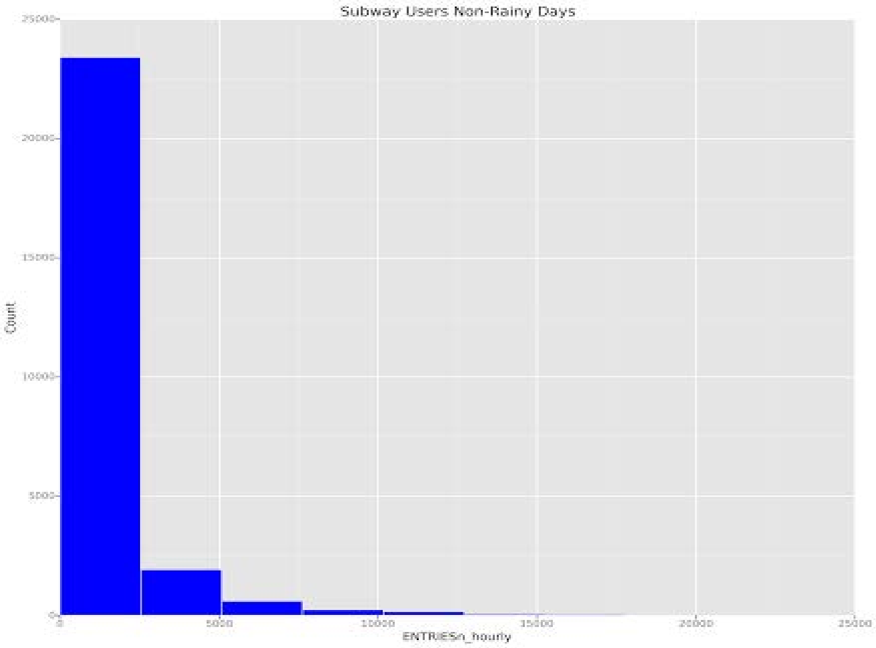
In theory the closer the r^2 is to 1, the better the model fits that data. If the r^2 was 1, for example, the model would be a perfect fit. In this case the r^2 is approximately 0.42. This means that only 42% of the variance in the underlying data in explained by the model - this is not as high as we might like so highlights that perhaps the model is not a good fit.

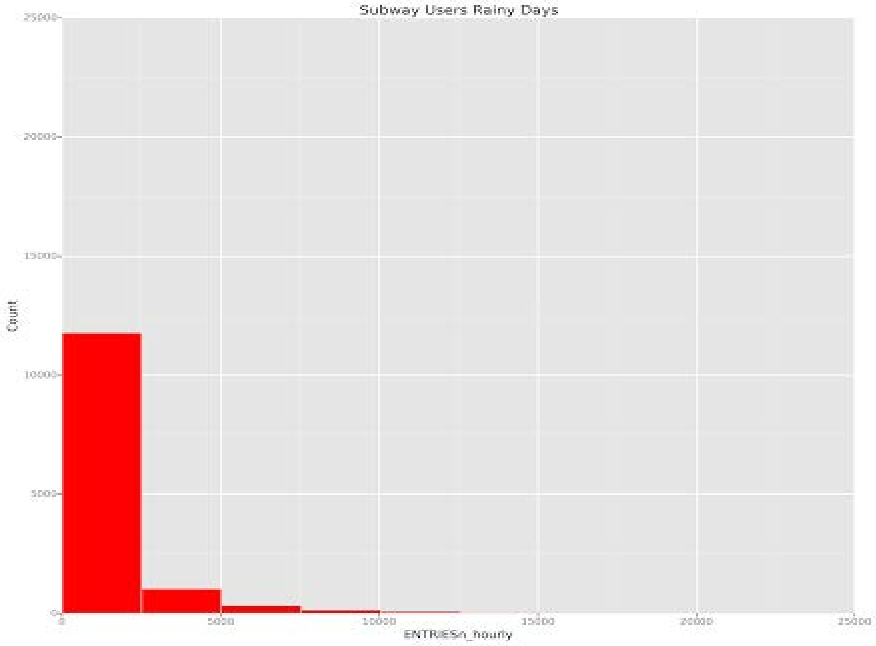
A key assumption of linear regression is that the residual are normally distributed and so it is useful to plot the residual in order to assess this. From looking at the plot below the assumption of a normal distribution seems quite reasonable. However the tails of the distribution are quite long which means there are large residuals (i.e. observations that our model over or under estimates considerably) which suggests that our model is not a good fit.


Plot of Residuals
6000
5000
4000
3000
2000
1000
0
−15000
−10000
−5000
0
5000
10000
15000
20000
25000
30000
Residuals

SECTION 3 – VISUALIZATION

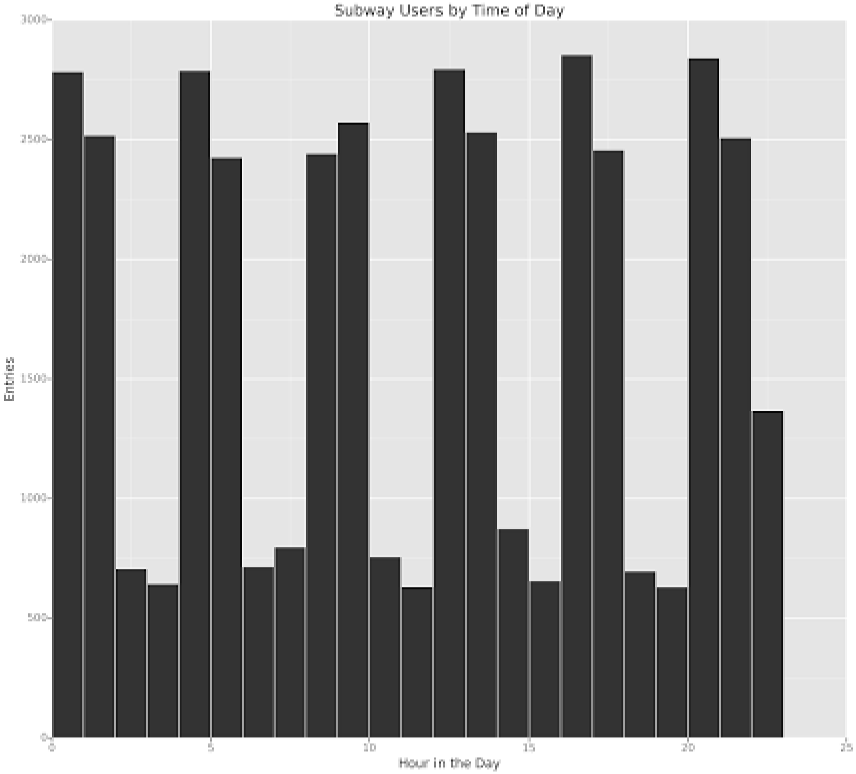
3.1





The graphs plot ENTRIESn_hourly for non-rainy days and ENTRIESn_hourly for rainy days respectively. From this we can see that rain does have an impact on whether or not people use the subway. For example there is a much higher proportion of observations in the 0-2,500 category in the first graph than in the second (i.e. proportionately more rows we less than or equal to 2,500 entries).

3.2



This graph looks at how subway ridership varies over the course of the day. As you would expect there is a lot of variability over the course of a day. In the early hours (2/3am) of the morning we can see that the numbers are quite low. We can see spikes in the figures at particular times as the number of commuters increases (8am-10am and 4pm to 6pm). There are further high points around 9pm and 12/1am which may be people using the subway to attend social events.

SECTION 4 – CONCLUSION

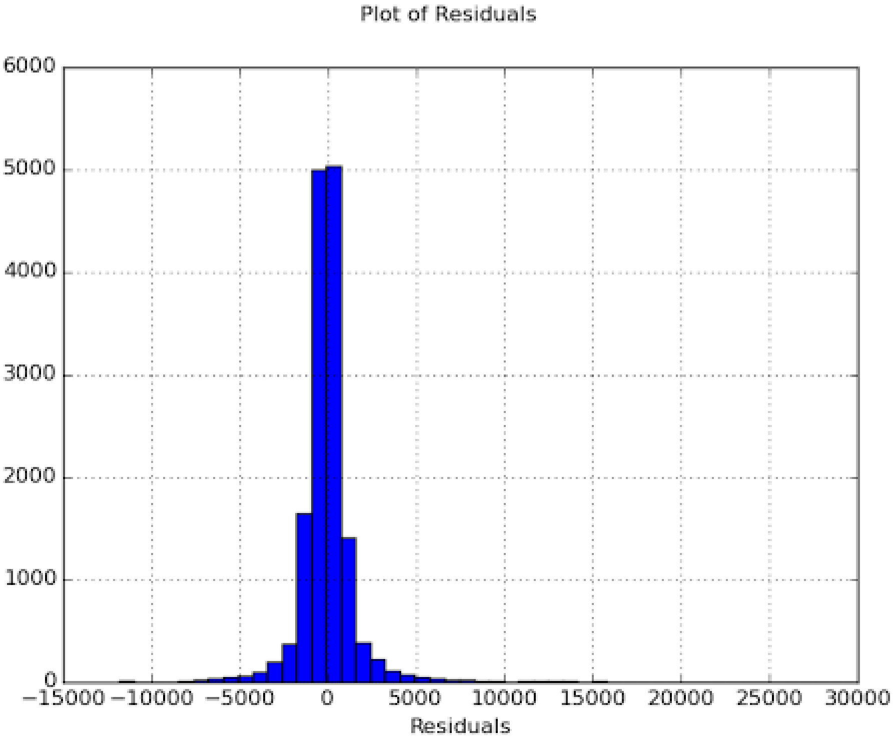
Looking at the result of the Mann-Whitney U test we can see that:

- Average ridership increases when it is raining
- The 2 samples (days with rain vs day without rain) are significantly different

This leads me to think that more people use the subway when it is raining as opposed to when it is not raining. This is supported by the data visualization (histogram) which shows a greater number of people using the subway when it is raining.

However we saw from the linear regression that our model was not a very good fit (r^2 of approx. 0.42). This suggests perhaps that while rain does influence the number of people using the subway, the relationship may be non-linear.

On key assumption of linear regression is that the residuals are normally distributed. From looking at the plot below the assumption of a normal distribution seems quite reasonable. However the tails of the distribution are quite long which means there are large residuals (i.e. observations that our model over or under estimates considerably) which suggests that our model is not a good fit.



SECTION 5 – REFLECTION

From the analysis above I think that the use of linear regression is probably not appropriate. The model could not be used to sufficiently predict subway ridership. The analysis focuses on the weather without considering other factors that might provide insight:

- There may be differences due to individual stations. For example we might expect there to be a difference between stations in touristy areas versus stations located in the business district.
- The analysis did not take account of the impact of the time of day on ridership. As shown in the 2nd visualization above we can see that there are definite peaks and troughs over the course of the day.

In addition there are issues with the data:

- For example the number of observations per station varies greatly (from 83 to 12,198). This means that if there are factors specific individual stations, our analysis is impacting by the weight given to each station based on the number of observations.
- The period covered is just one calendar month (the month of May). It may not be appropriate to use this to try to predict future ridership. For example in May people may not expect rain and so may not plan/dress for it whereas in other months when rain might be expected they may plan/dress and so may be less likely to take the subway if it rains.
- There is no information on the individual stations e.g. where they are based, number of lines that they served etc.