Project 1: Analyzing the NYC Subway Data Set

Faiza Ihsan - ihsan_610074

1. References

http://matplotlib.org/1.3.0/examples/pylab_examples/histogram_demo_extended.html

http://stackoverflow.com/questions/6986986/bin-size-in-matplotlib-histogram

http://discussions.udacity.com/t/project-1-section-3-1-creating-2-histograms-for-entries-on-rainy-days-vs-non-rainy-days/18027/3

http://stackoverflow.com/questions/2849286/python-matplotlib-subplot-how-to-set-the-axis-range

https://pypi.python.org/pypi/ggplot/

http://matplotlib.org/examples/statistics/histogram_demo_multihist.html

http://stackoverflow.com/questions/29268134/how-to-merge-rows-in-dataframe-according-to-unique-elements-and-get-averages?rq=1 (to find average entries by hour)

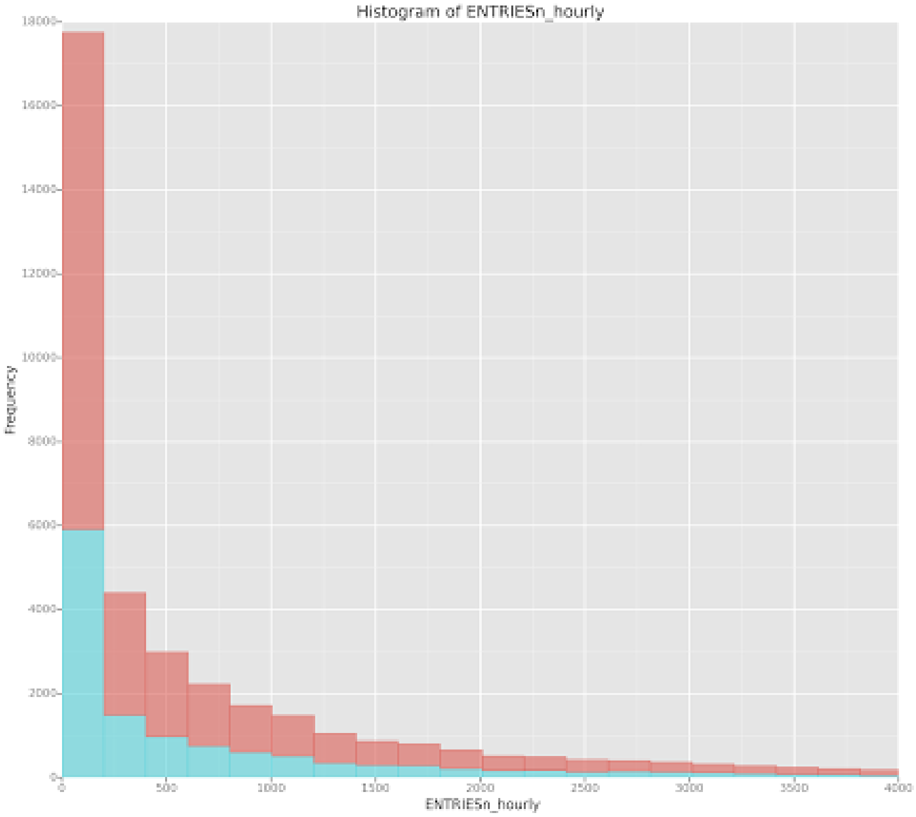
http://discussions.udacity.com/t/written-project-should-be-based-on-results-obtained-in-problem-sets-1-5/17878

http://blog.minitab.com/blog/adventures-in-statistics/regression-analysis-how-do-i-interpret-r-squared-and-assess-the-goodness-of-fit

2. Statistical Test

2.1. I used the Mann Whitney U-test to analyze the NYC subway data on the original data set. I used a two-tail P value as we are only checking for change in ridership caused by rain and cannot assume before-hand that it will increase or decrease. My null hypothesis is that, both the distributions (rainy day ridership and non rainy day ridership) are from the same population. In other words, mean ridership of the NYC subway remains the same on rainy days and non-rainy days. The p-critical value is 0.05.

2.2. When we charted out the ridership data ('ENTRIESn_hourly' for Raining vs Non Rainy days) we saw that the data was not normally distributed hence we cannot use the Welch's t-test which assumes normal distribution.



Code used to generate the above histogram:

```
plot = ggplot(turnstile_weather,
aes(x='ENTRIESn_hourly',colour='rain',fill='rain' )) + \
geom_histogram(binwidth=200, alpha=0.6) + \
ggtitle('Histogram of ENTRIESn_hourly') + \
xlim(0, 4000) + xlab('ENTRIESn_hourly') + ylab('Frequency')
```

It should be noted that just because non-rainy day's frequency of ridership is more in the above histogram, we cannot conclude that people use the subway more on non-rainy days. When we look at the actual data we can see that there are more non-rainy days overall. We will also see later that the mean of rainy days ridership is more than non-rainy day's ridership.

I also did the Shapiro-Wilk test to find out if the two distributions were Gaussian. The null hypothesis was that the data is drawn from a normal distribution. For both the distributions I got a P value of 0.0 and thus rejected the null hypothesis.

The Mann Whitney U-test is a non-parametric test which can be used on distributions that are not normal.

2.3. Using the code in problem set 3.3 we got:

Mean ridership on rainy day	1105.45
Mean ridership on non-rainy day	1090.28
P value (one-tail)	0.024999912793489721 ~ 0.025
P value (two-tail)	0.049999825586979442 ~ 0.05

2.4. If we do not round up the two-tail p value it is less than our threshold of 0.05. So we reject the null hypothesis that the distributions are from the same population. We accept the alternate hypothesis that the distributions are from two different populations. In other words they are statistically different and that rain actually does have an effect on ridership. We can also see that the mean ridership on rainy days is more than on non-rainy days

3. Linear Regression

3.1. I used gradient descent to compute the coefficients theta and produce prediction for ENTRIESn_hourly in my regression model as implemented in exercise 3.5

3.2. I used *'rain', 'precipi', 'Hour', 'meantempi', 'fog', 'meanwindspdi'* as features in my model. I used 'UNIT' as my dummy variables as it's a categorical variable.

3.3. From the Mann Whitney U-test I performed I chose the above features based on intuition and also by trial and error by looking at their effect on R^2

- 'rain' this is the factor which is part of our null hypothesis so it was natural to select this. Also I assumed that on a rainy day people would be more prone to use the subway instead of being stuck in traffic jams and for safety. I know I would!
- 'precipi' was included because if it's just a drizzle then it may not affect people's choice of whether to use the subway but if it's raining heavily that may affect ridership.
- 'Hour' is a feature because there are peak usage hours during the day – morning rush to offices and evening rush back home as we will see in our visualization below.
- 'fog' obviously impacts visibility and as opposed to being in a car, the subway has its fixed tracks and routes and is much safer. Hence this might also influence people's choice.

- ‘meanwindtspi’ gave a slight improvement to the R^2 value. My intuition behind this was that on a very windy day people may also choose not to use cars. Since I saw a positive effect after adding it, I left it as a feature
- ‘UNIT’ also is important because certain stations may have more entries than others because of location. When we remove UNIT we can see that R^2 really goes down. It is used as a dummy variable because we can’t do any numerical calculations on this – it’s a categorical variable.

3.4 Coefficients of non-dummy features

Feature	Coefficient
'rain'	-1.01382395e+01
'precipi'	-2.48062493e+01
'Hour'	4.68435930e+02
'meantempi'	-5.37848030e+01
'fog'	7.31131541e+01
'meanwindspdi’	6.40183862e+01

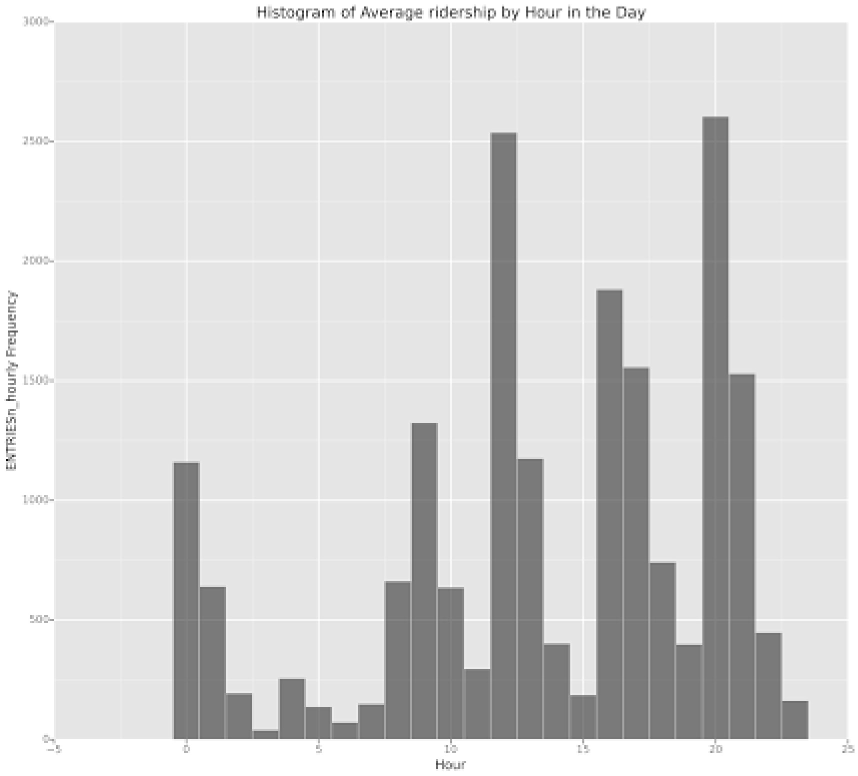
3.5 My R^2 value is 0.465094157151

3.6 This means that 46.5% of the variation in the value of my dependent variable, ENTRIESn_hourly, is explained by the variation in my features. This is a reasonably good fit and is statistically significant considering we are trying to measure a human decision which can be influenced by a LOT of factors.

4 Visualization

4.4 Histogram for ENTRIESn_hourly is shown in Section 2

4.5 Free form plot average ridership by time-of-day



We can see that there are peak hours during noon and evening. It could be people going out for lunch or other personal activities during lunch break (and leaving their cars). We see a peak around 8pm which isn't really the time when people are leaving office so it could be people going out for leisure activities in the evening.

```
df2 =   turnstile_weather.groupby(['Hour'])['ENTRIESn_hourly'].mean().reset_index()
plot = ggplot(df2, aes(x='Hour', y='ENTRIESn_hourly')) + \
geom_bar(alpha=0.60, stat='bar') + \
ggtitle('Histogram of Average ridership by Hour in the Day') + \
xlab('Hour') + ylab('ENTRIESn_hourly Frequency')
```

5 Conclusion

5.4 Statistically the Mann Whitney U-Test result of 0.049 shows that the two distributions are from a different population – that there is more ridership on a rainy day versus on non-rainy days. But the p-value is very close to our p-critical value of 0.05 and our r~2 is only 0.465 (the closer it is to 1 the better fit it is). So we cannot definitely conclude that rainy days have more subway ridership.

5.5 If we look at the difference in mean of ridership on rainy vs non rainy days there is only a difference of about 15 people per hour, which is around a 1% difference. In real world that's really not a lot. In our regression model too keeping all the features I selected above, I played around with 'rain' and 'percipi' and removing them didn't make a significant change to the regression value.

I moved rain to a dummy variable too as its either raining or not and used precipitation as a feature and we can numerically measure the amount of rain through this. The extremely minor change to the r^2 value means we cannot confidently conclude that rain is causing any effect whatsoever at all. The biggest change in r^2 value was obtained by UNIT as a dummy variable which means that certain UNITs have more ridership than others.

r^2	Rain/percipi
0.465083550449	without rain/with percipi as feature
0.464964580824	without rain /without percipi as features
0.465021310938	with rain as a dummy variable / no percipi as feature

6 Reflection

6.4 The features I selected may have correlation example a very rainy day may also be very windy and may have low temperatures. There may be a lot of other variables that influence ridership which weren't available in the data. Weather conditions also very by the hour. It could also be that during a rainy day more people prefer to stay home for safety but the percentage of ridership (based on mean) is higher. Other events like public holidays, parades, election-day, summer holidays, weekends, fuel costs, car taxes etc. may also influence the choice to opt for or against the subway.

We also have data for only one month which isn't a long enough interval to make any serious conclusions.

I should have also looked at the residual plot along with the r^2 value.