Section 1. Statistical Test

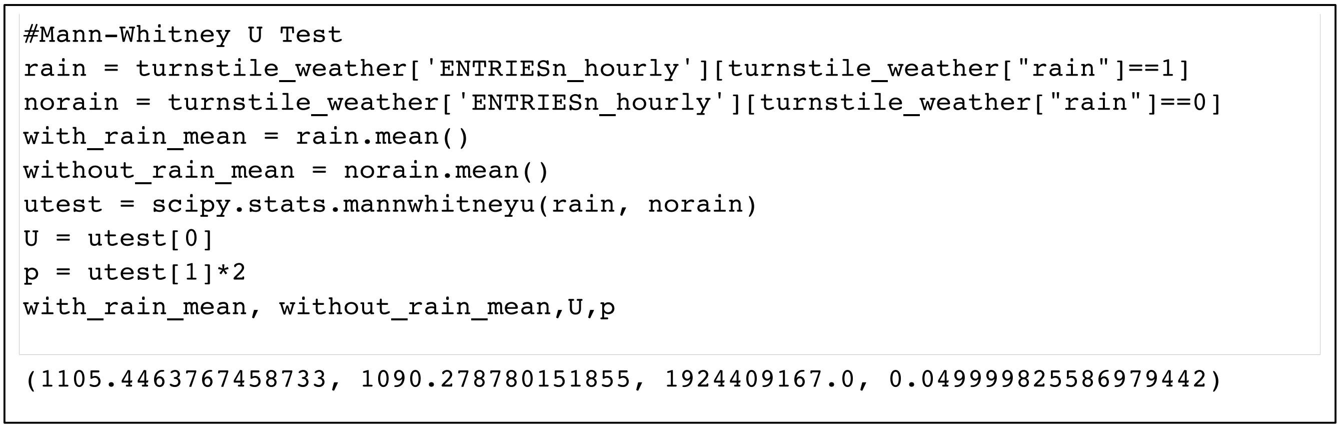
The NYC subway ridership for rain and non-rainy days are not normally distributed (as graphed in section 2), therefore, we can use a non-parametric test, Mann-Whitney U Test, to analyze the NYC subway data.

In this test, we are performing a two-tailed U-test, with α of 0.05. The hypothesis of the test is: The two ridership samples of rainy and non-rainy days are derived form the same population, in the sense that if we randomly draw the ridership from non-rainy days, it will not be more likely to generate higher value than rainy days.

Our Hypothesis testing parameters are:

x = Ridership on Rainy Days
y = Ridership on non-Rainy Days

$H_0: P(\, y > \; x) = 0.5$
$H_A: P(\, y > x \,) \neq 0.5$



```
#Mann-Whitney U Test
rain = turnstile_weather['ENTRIESn_hourly'][turnstile_weather["rain"]==1]
norain = turnstile_weather['ENTRIESn_hourly'][turnstile_weather["rain"]==0]
with_rain_mean = rain.mean()
without_rain_mean = norain.mean()
utest = scipy.stats.mannwhitneyu(rain, norain)
U = utest[0]
p = utest[1]*2
with_rain_mean, without_rain_mean,U,p
```

```
(1105.4463767458733, 1090.278780151855, 1924409167.0, 0.049999825586979442)
```

From the test, we know that the avg ridership on rainy days is 1105, and on non-rainy days is 1090. We also obtained a U-statistic of 1924409167 and p-value of 0.05.

Since our p-value is equal to α (0.05), we failed to reject the null hypothesis, and the ridership of NYC subway between rainy and non-rainy days are not different – ridership on non-rainy days is not always

Section 2. Linear Regression

I used OLS regression model using Statsmodels module from Python to produce prediction for ENTRIESn_hourly.

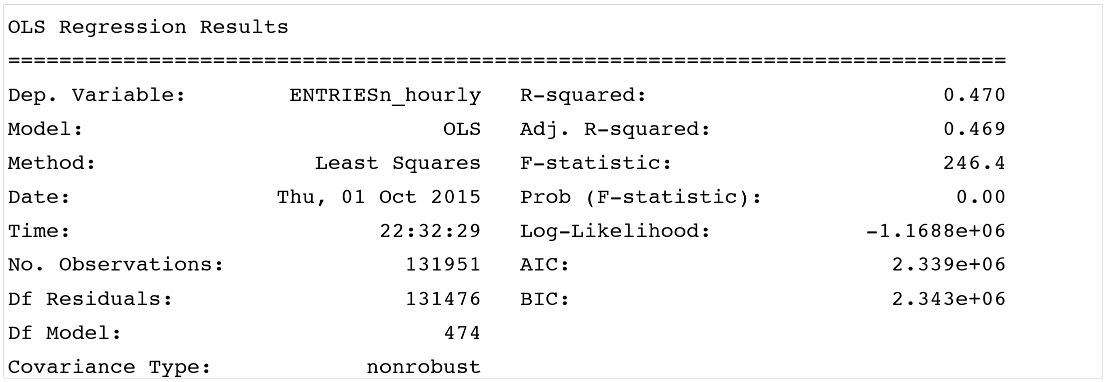
The features I used are “rain”, “fog”, “hour”, “meantempi”, “meanwindspdi”, “UNIT” and “day_of_week”

The reason for choosing these variable is based on intuition and also experimentation – if they do increase R^2:

- **rain**: More people may take subway when it’s raining outside.
- **fog**: Used Fog feature as if it’s foggy outside, more people may take the subway
- **hour**: Hour of the day also matters as there are certain hours (peak hours) when more users would take the Subway
- **meantempi**: Average temperature may be related to ridership, as if it’s too hot or too cold outside may influence people wanting to take subway.
- **meanwindspdi**: Closely related to the above, if it’s too windy outside, more people may want to take subway instead.
- **UNIT**: Unit is a dummy variable for stations. Different stations would have different ridership, and it may also have interaction with time of day (eg. Stations close to work locations during commuting peak hours).
- **day_of_week**: I created this variable using Dates to parse out the day of the week for each Date. The day of the week may have strong influence on Subway ridership, especially during weekdays vs weekends.

```
import numpy as np
import pandas as pd
from statsmodels.formula.api import ols

mod = ols(formula = "ENTRIESn_hourly ~ rain + Hour + meantempi + fog  + UNIT +
day_of_week", data = turnstile_weather )
res = mod.fit()
print res.summary()
```



```
OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.470
Model:                            OLS   Adj. R-squared:                  0.469
Method:                 Least Squares   F-statistic:                     246.4
Date:                Thu, 01 Oct 2015   Prob (F-statistic):               0.00
Time:                        22:32:29   Log-Likelihood:            -1.1688e+06
No. Observations:              131951   AIC:                         2.339e+06
Df Residuals:                  131476   BIC:                         2.343e+06
Df Model:                         474
Covariance Type:            nonrobust
```

The parameters of the non-dummy features in my linear regression models are

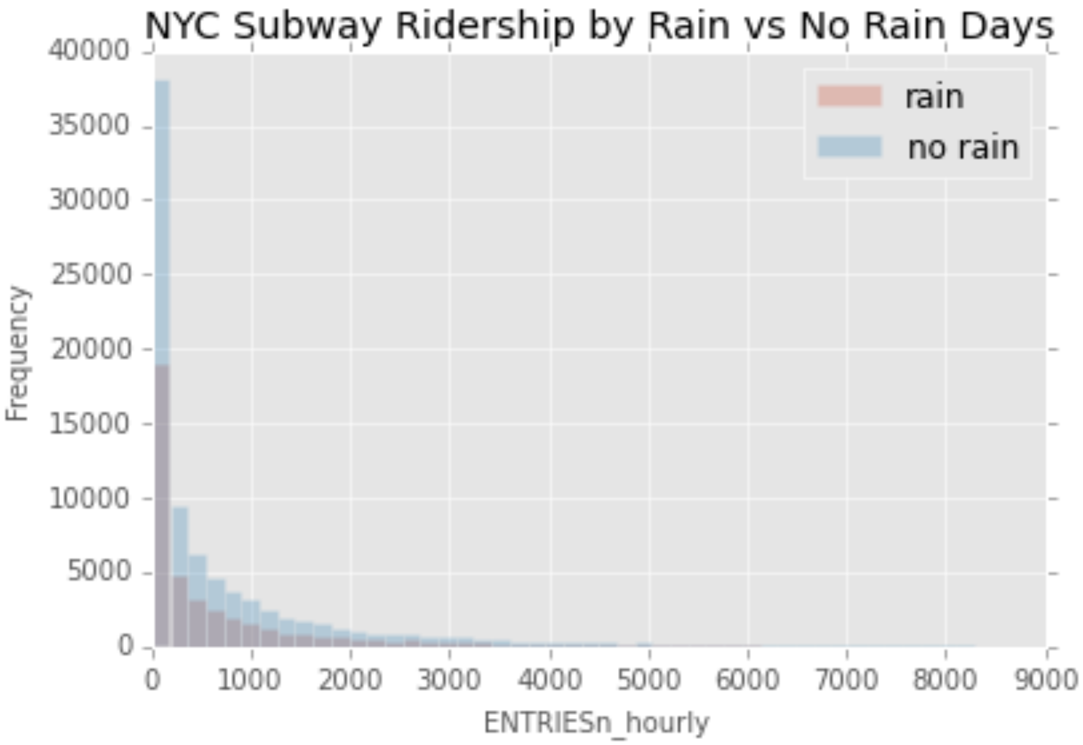
- `hour: 67.3572`
- `meantempi: -6.6495`
- `fog:46.6283`
- `meanwindspdif: 0.15`

The R^2 of the model is 0.47, this means the model is able to explain 47% of the subway ridership using the variables. This model may be somewhat appropriate to predict the ridership given the features available as this model has improved R^2 comparing to the base R^2 0.40.

3. Visualization

a. Histogram of Ridership on Rainy Days and Non-Rainy Days

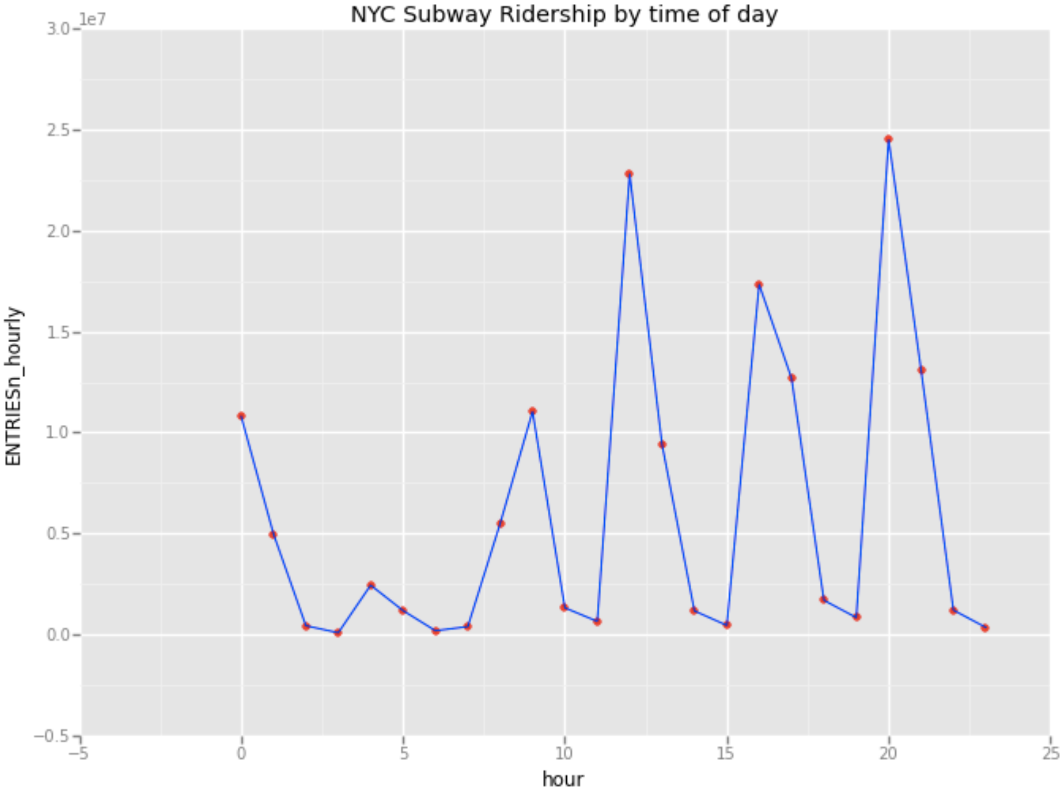
```
#plot ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.
plt.figure()
rain = turnstile_weather["ENTRIESn_hourly"][turnstile_weather["rain"]==1]
no_rain = turnstile_weather["ENTRIESn_hourly"][turnstile_weather["rain"]==0]
df = pd.concat([rain, no_rain], axis =1)
df.columns = ["rain", 'no rain']
plt.figure()
plot = df.plot(kind='hist', alpha=0.3, bins=50, range=[0, 9000]).set_title('NYC Subway Ridership by Rain vs No Rain Days')
plt.xlabel('ENTRIESn_hourly', fontsize=10)
plt.ylabel('Frequency', fontsize=10)
```



The distribution of rain and non rainy days seems to have the same distribution: Heavily skew towards the right tail and with most of the frequencies concentrating on the left tail.

b. Ridership by time-of-day

```
#plot by hour of day
by_hour = turnstile_weather["ENTRIESn_hourly"].groupby(turnstile_weather["Hour"]).sum()
by_hour.index.name = 'hour'
by_hour = by_hour.reset_index()
p = ggplot( by_hour, aes("hour", "ENTRIESn_hourly"))
p + geom_point(color = "red") + geom_line(color = "blue") + ggtitle("NYC Subway Ridership by
time of day")
```



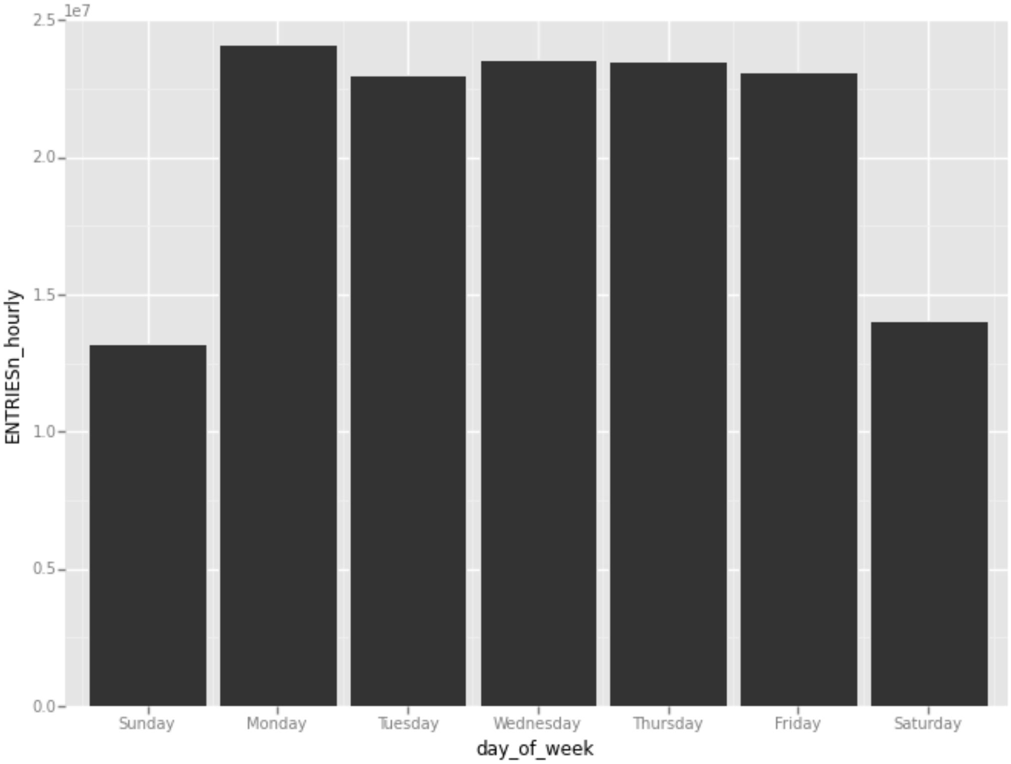
Time of day shows an interesting pattern: We see peaks falls around work time rush hours (9AM, 12PM, 4PM and 8PM)

c. Ridership by day-of-week

```
turnstile_weather["day_of_week"] = map(lambda x: datetime.strptime(x, "%Y-%m-%d").strftime("%A"),
turnstile_weather["DATEn"])

turnstile_weather["day_of_week_num"] = map(lambda x: datetime.strptime(x, "%Y-%m-%d").strftime("%w"),
turnstile_weather["DATEn"])

by_day = turnstile_weather.groupby(["day_of_week_num","day_of_week"]).ENTRIESn_hourly.sum().reset_index()
p = ggplot(by_day, aes("day_of_week", "ENTRIESn_hourly"))
p + geom_bar(stat="identity") + ggtitle("NYC Subway Ridership by Day of Week")
```

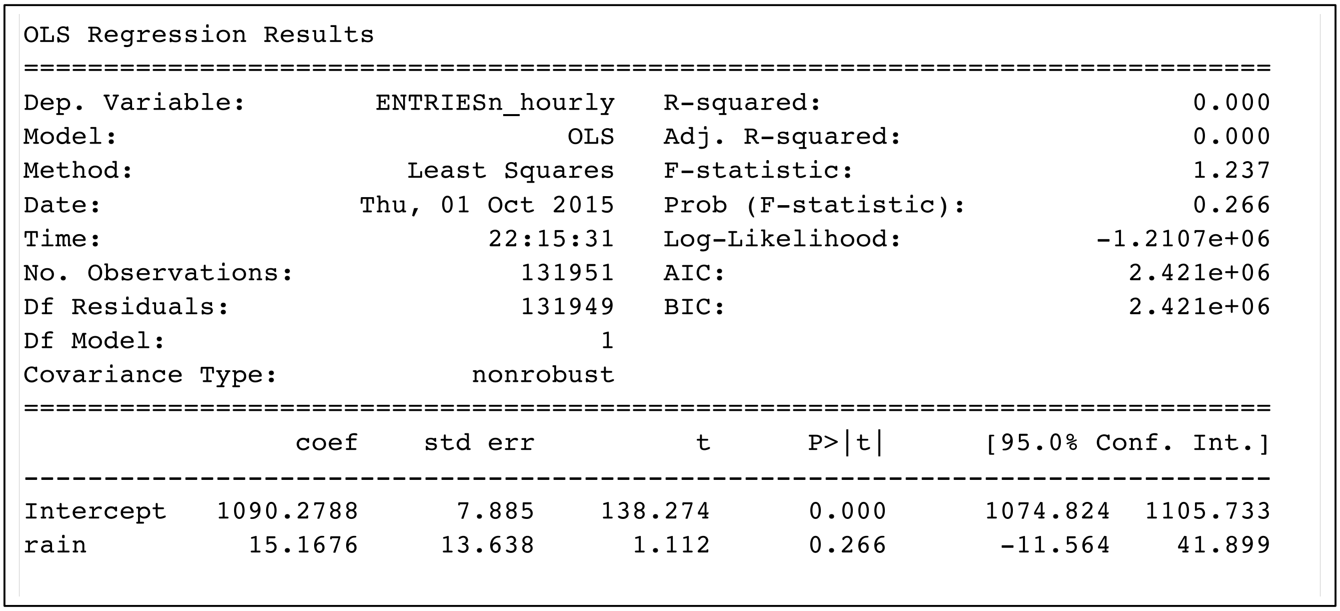


Day of week also shows that ridership during weekdays are higher than weekends.

4. Conclusion

The average ridership of NYC subway on rainy days are 1105 entries/hour, while during non-rainy days, it’s 1090 per hour. Since the distribution of the ridership is non-normal, I conducted Mann-Whiteny U test to test if the ridership is the same between rainy and non-rainy days. The results of the test (p = 0.05) shows that there does not seem to be a difference between ridership on rainy and non-rainy days.

And indeed, if using only rain as a factor to predict ridership, the R^2 is equal to 0, and the p value of the coefficient is greater than 0.05. This means the regression model using only rain as the independent variable does not explain the ridership, and the coefficient it’s not significantly different than 0.



```
OLS Regression Results
==============================================================================
Dep. Variable:        ENTRIESn_hourly   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     1.237
Date:                Thu, 01 Oct 2015   Prob (F-statistic):              0.266
Time:                        22:15:31   Log-Likelihood:            -1.2107e+06
No. Observations:              131951   AIC:                         2.421e+06
Df Residuals:                  131949   BIC:                         2.421e+06
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept   1090.2788      7.885    138.274      0.000      1074.824  1105.733
rain          15.1676     13.638      1.112      0.266       -11.564    41.899
```

Therefore, we have to also use other variables like fog, wind, temperature, time of day, day of week and stations to predict the ridership.

In addition, when adding rain as another explanatory variable to the regression model, we do not see an improvement to R^2. Therefore, it is very possible that ridership is not responding to rainy or non-rainy days.

5. Reflection

Intuitively, we would think that rain would affect subway ridership, but our statistical test does not show there is a difference in ridership between rain and non-rainy days. The reason rain does not have predictive power over ridership may also due to the simplicity of the variable: We don't know how much it rained, and how long it rained. If we do have this data, then we can further fine tune the model, using more "rain" information to improve the model.

In addition, my OLS regression has an R^2 of 0.47, which explains only less than 50% of the subway ridership in NYC, plus, there may be interaction between the variables: Rain, temperature, precipitation and wind could have interaction with each other. More investigation of this interaction is needed to improve feature selection and the regression model.

Resources:

1. Do Rainy Days Impact NYC Subway Ridership? http://rainydaysny.blogspot.com
2. Mann-Whitney U Test: https://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test
3. MTA Data Description: http://web.mta.info/developers/resources/nyct/turnstile/ts_Field%20Description.txt