Section 1. Statistical Test

1.1 Which statistical test did you use to analyze the NYC subway data? Did you use a one-tail or a two-tail P value? What is the null hypothesis? What is your p-critical value?

Answer 1.1

Initially I used Mann-Whitney U test to Analyze New York city subway data set. We specifically did not used Welch’s T-test; as the dataset is not normally distributed; or to simply put it’s not yet know which data set is higher or lower.

We used a One-tail P value returned by scipy.stats.mannwhitneyu; counter validate our null hypothesis.

The p-critical values used would be 0.5 or 5%.

1.2 Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

I used this test as the data for rain and no-rain is not normally distributed.

1.3 What results did you get from this statistical test? These should include the following numerical values: p-values, as well as the means for each of the two samples under test.

with_rain_mean = 1105.4463767458733,
without_rain_mean = 1090.278780151855,
U = 1924409167.0,
P = 0.024999912793489721

1.4 What is the significance and interpretation of these results?

Increase of about 1.4% in ridership in New-York subway; But it’s not sufficient to prove our null-hypothesis.

Very High value of U statistics 1924409167.0 and very low value for P 0.024; helps us draw conclusion that the proposed null hypothesis is false; and ridership is different for rainey and non rainy days.

Section 2. Linear Regression

2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:

1. **OLS using Statsmodels or Scikit Learn**
2. **Gradient descent using Scikit Learn**
3. **Or something different?**

1. Initially I use OLS using Statsmodels; without forgetting to add a constant in the features
2. Gradient descent using Scikit Learn - experimented with this as well

2.2 What features (input variables) did you use in your model? Did you use any dummy variables as part of your features?

Answer 2.2

I used features mentioned below; from subway data set:
'rain', 'precipi', 'Hour', 'meandewpti', 'meanwindspdi'
No additional dummy features were used; except default dummy_units which added 'Units' features later to the data frame.

2.3 Why did you select these features in your model? We are looking for specific reasons that lead you to believe that the selected features will contribute to the predictive power of your model.

I went with the features:
'rain', 'precipi', 'Hour', 'meantempi' - as they indicated rain and precip at certain duration of time at a certain temp; as well as they are provided default parameters with R2 0.42

'meandewpti', 'meanwindspdi' – Both of these features dew and wind speed are essential in predicting weather conditions; and improvement in value of R2 indicates the same.

‘fog’ - Just went with a hunch that it might be also useful to predict; increase in value of R2 proves the same,

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model?

Coefficient (or weights) of non-dummy features or params in my linear regression model is:

```
[  2.51573725e+01  -7.10132850e+01   6.53621362e+01  -6.66431278e+00
  -2.72639810e+00   3.20729898e+01   2.18896549e+02   4.07920639e+03
  -9.51978056e+02  -1.28094252e+03  -1.09303580e+03  -4.37826853e+02
  -9.46909732e+02  -1.31592707e+03  -1.16280143e+03  -1.16439172e+03
   4.52193989e+03   7.05527511e+03   7.23063269e+03   1.08442978e+03
   1.92637201e+03  -3.33729911e+02  -6.14548600e+02   2.94535251e+03
   3.43947249e+03   2.62736172e+03   1.98144721e+03   3.12318976e+03
   8.50831438e+03   5.92778172e+03   8.42205821e+02   3.13474843e+03
   1.56412979e+03   1.53194658e+03   5.95092976e+03   5.65619616e+02
   2.51697811e+03   1.92838379e+03   4.52834607e+03  -5.06237506e+02
   1.74960499e+03  -7.50107205e+02  -7.71424745e+02  -1.45036005e+03
  -9.66849298e+02  -1.05394388e+02   1.03101215e+03  -5.47494750e+02
  -9.04444307e+02   2.14689675e+03   2.67173308e+02   6.08069614e+03
   6.11747877e+03   9.29614710e+02   8.15716519e+02   2.92070631e+03
   3.00712383e+03  -4.08604427e+02   1.69863449e+03  -5.45658406e+02
   6.78469176e+03   5.62376766e+02   3.35236893e+03  -7.90229781e+02
  -1.80812697e+02  -8.16657534e+02  -8.18435899e+02   1.38939783e+03
  -4.75839076e+02  -6.53394087e+02  -9.00910393e+02  -1.19921312e+03
  -5.12678463e+02  -1.02844938e+03  -8.18268741e+02  -4.09125269e+02
   1.25986006e+03   3.20924582e+03   2.02184828e+03  -1.88945569e+01
   1.66684836e+03   4.80218075e+03   3.03957826e+02   8.71321365e+02
  -2.68442164e+02  -1.15555880e+03  -1.03829777e+03  -1.18057381e+03
  -6.62454041e+01   3.89588573e+02   2.14000431e+02   1.18416656e+02
   6.49788354e+02   4.45455889e+02   1.86019322e+03   8.64572392e+02
   5.28448637e+02  -8.85064704e+02   2.08337273e+03   3.39340493e+03
  -1.61371186e+02   5.89283228e+02   1.67299568e+03  -4.44633230e+02
  -5.27341920e+02   4.23913942e+03  -8.40376525e+01   1.55162536e+03
   9.06365516e+02   1.67425810e+02   1.77601434e+03  -3.94717487e+02
  -4.49684190e+02   1.31373093e+03  -7.04423273e+02   2.63636952e+01
  -3.65258841e+01   9.48256890e+01   2.36148999e+01   1.09131901e+03
   7.48072048e+02  -9.48680713e+02  -8.42536337e+02   1.63101601e+02
   2.64922946e+03  -1.08514018e+03  -1.67725706e+01  -8.76088351e+02
```

```
 1.61700647e+03  1.80447127e+03 -2.36063174e+02 -1.21980964e+03
-2.19360865e+02 -2.45067267e+02  6.60468361e+02  3.46941003e+03
-1.74385008e+02 -1.81808495e+02  2.44825215e+03  1.59814477e+03
 7.41429332e+02  2.35283782e+03 -9.51336394e+02 -1.10037107e+02
 8.74159848e+02 -8.38713914e+02 -1.59450083e+02 -6.37388838e+02
 2.33469477e+02  1.20342181e+02  8.30994116e+02 -7.19270712e+02
-3.59474055e+02 -3.98527091e+02 -2.06855683e+02  1.72616041e+03
 5.42556823e+01  1.10517003e+03 -5.71618789e+00  2.10970847e+03
 1.48381278e+03 -1.67317574e+03  8.42620212e+02  1.56902869e+03
 3.38071465e+03  2.59066061e+02  8.93569677e+03 -5.01352045e+02
-5.60428543e+02 -6.55300220e+02  2.99222412e+02  3.46413967e+03
 1.56193050e+03  3.46417772e+03  4.84601903e+03  7.34119250e+03
 7.80818941e+02  2.19479160e+02  9.04588431e+02 -4.45442428e+02
-5.75383351e+02 -5.13749388e+02 -9.19274455e+01  3.77884954e+01
 1.07880679e+03 -6.97838857e+00  4.09177414e+02  4.38971690e+02
 1.74553886e+02  3.23877624e+01  4.13697089e+02  3.34550909e+03
-1.40899981e+02  1.72528674e+02  8.69677492e+02 -7.61206275e+02
 3.64620059e+02  1.17339877e+03  1.02233598e+03 -1.30088746e+02
 2.41939178e+02  2.45695554e+02  1.90190027e+02  1.98598243e+02
 1.84410462e+03 -3.37508771e+02 -9.32994372e+02  2.16810608e+02
 6.71320921e+01  1.46825407e+01 -7.45101074e+02 -1.36936089e+02
-8.34152929e+02  2.88619641e+02  6.29789648e+00 -9.75930441e+02
-4.51106894e+02 -2.50889111e+02  1.79798172e+03  1.32845827e+03
-6.30448314e+02 -8.89913767e+02 -7.88824235e+02 -4.57178701e+02
-3.33255379e+01 -1.01928036e+03 -8.30480046e+02 -3.74126013e+02
-3.46746831e+02 -5.07464492e+02 -1.22505526e+03  1.89281425e+03
 4.03817620e+02 -8.32629534e+02  9.83745602e+02 -7.06293479e+02
 1.87193331e+03 -1.09120478e+03 -9.41157973e+02  1.55679656e+02
-3.58315268e+02 -1.14000725e+03 -1.13265422e+03  2.40191255e+03
 9.55498291e+01 -3.71670391e+02 -4.06040608e+02 -7.92848873e+02
-1.00633708e+03  5.27649132e+02 -6.97899402e+02 -6.73708844e+02
 2.98075530e+02  3.01988193e+02 -6.57711593e+02  6.56360411e+01
-1.40337298e+02 -5.53738309e+02 -1.42565976e+03 -1.01007290e+03
-5.64592209e+02 -2.93419149e+02 -2.89717077e+02  1.06809051e+03
-6.95633803e+02 -9.16159164e+02 -1.14187533e+03  6.19751802e+01
-3.10124463e+02 -6.54969162e+02 -7.12345049e+02 -9.03242992e+01
-6.36666044e+02 -1.06125332e+03 -6.93036872e+02 -8.90494785e+02
-1.89295533e+02  6.26558217e-01 -8.28609661e+02 -5.75746226e+02
-9.81480068e+02 -9.37136085e+02 -7.87521767e+02  3.48617760e+02
-9.80724713e+02 -2.53713816e+01  1.39979666e+02 -1.03479766e+03
 5.99424342e+03 -5.83371556e+02 -9.85999048e+02 -1.07918792e+03
-7.03162636e+02 -6.41035242e+02 -1.15446436e+03  1.31815148e+03
-4.88684220e+02  6.96126518e+02  2.44661028e+02 -4.76946979e+02
```

```
-1.41521712e+03 -9.43534151e+02 -5.76047470e+02 -5.64555949e+02
 1.95067199e+02 -1.05884193e+03 -1.12847020e+03 -1.22044474e+03
-1.04810383e+03 -9.47661791e+02 -1.33793009e+03 -8.17747081e+02
-1.18390029e+03  5.47197403e+02 -8.31215974e+02 -4.97101473e+02
 4.17171823e+02 -1.15167865e+02  1.69765616e+02 -1.13426721e+03
-1.11800205e+03 -3.70279103e+02 -1.17423133e+03 -1.44183705e+03
-2.79445653e+02 -1.31864725e+03 -8.96767418e+02 -7.64721423e+02
-5.94469767e+02 -9.94889979e+02 -1.34373003e+03 -1.44790925e+03
-1.56445268e+03 -1.20579497e+03 -1.07687829e+03 -9.92316800e+02
-9.63504563e+02 -9.43522701e+02 -1.11320305e+03 -1.13763818e+03
-4.55895491e+02 -6.57995298e+02 -1.29191610e+03 -6.01284701e+02
-1.18890740e+03 -1.10812138e+03 -9.00731038e+02 -1.16288678e+03
-1.25856086e+03 -8.42345884e+02 -1.48471006e+03 -1.28648944e+03
 7.47757799e+02 -1.45351238e+03 -5.91311940e+02 -9.05072603e+02
-8.69350079e+02 -5.11529367e+02 -1.06076406e+03 -1.11611685e+03
-9.25459256e+02 -1.33898546e+03 -8.61042669e+02 -8.46394871e+02
-8.05881953e+02 -9.80498146e+02 -6.45131813e+02 -7.53358649e+02
-1.10100509e+03 -1.07032586e+03 -2.47126262e+02 -1.70422127e+02
-1.13443724e+03 -7.46062963e+02 -6.36252122e+02 -5.30358588e+02
-1.02487709e+03 -1.25892491e+03 -5.78997034e+02 -7.57652364e+02
-4.05104763e+02 -5.19802861e+02 -1.14564592e+03  1.03624978e+02
-1.95978127e+02 -5.50188944e+02 -8.98660169e+02 -4.93450773e+02
-7.83511408e+02 -9.66353959e+02 -1.22203692e+03 -8.58919655e+02
-1.12595674e+03 -1.20287860e+03 -1.38540698e+03 -1.08803034e+03
-1.14778370e+03 -6.40383139e+02 -1.05241435e+03 -4.14724702e+02
-1.20382109e+03 -1.50872383e+02 -9.63723369e+02 -1.48728552e+03
-1.30042838e+03 -1.11732288e+03 -1.40879509e+03 -1.68354826e+03
-1.42953989e+03 -1.47594676e+03 -1.47490350e+03 -1.61113921e+03
-1.00688143e+03 -1.12608007e+03 -1.32830454e+03 -1.12074954e+03
-1.24578223e+03 -1.09413770e+03 -1.44944617e+03 -1.31836154e+03
-1.44824449e+03 -7.81722013e+02 -9.48955126e+02 -1.13750718e+03
-1.10937626e+03 -1.20839110e+03 -8.85750050e+02 -1.13590324e+03
-1.23947622e+03 -8.69966497e+02 -9.22029341e+02 -7.76373386e+02
-1.23777016e+03 -1.02276979e+03 -1.37120074e+03 -5.56797290e+02
-1.34072985e+03 -5.85311475e+02 -7.36394264e+02 -1.05999858e+03
-1.53028130e+03 -9.13435983e+02 -1.14207019e+03 -6.31099816e+02
 4.89463483e+03 -5.39856189e+01 -1.23667730e+03 -1.32482315e+03
-1.06402649e+03 -1.99048667e+03 -1.51539644e+02  2.22523880e+03
 2.68204390e+01  1.53292285e+03 -1.47039610e+03 -1.04516086e+03
-1.14992095e+03 -1.08751274e+03 -9.97028281e+02 -1.36173325e+03
-1.44817312e+03 -1.49229891e+03 -1.51018903e+03 -1.57227538e+03
-1.50619447e+03 -1.42310392e+03 -1.44012418e+03 -1.43951803e+03
-1.57121620e+03 -1.56606440e+03 -1.47955869e+03 -1.46138335e+03]
```

2.5 What is your model’s R2 (coefficients of determination) value?

My model’s R2 (coefficients of determination) value = 0.480502743303

2.6 What does this R2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R2 value?

“*R-Square*, also known as the *Coefficient of determination* is a commonly used statistic to evaluate model fit. *R-square* is 1 minus the *ratio of residual variability*. “
Simplyput coefficient of determination R2 is quantitative measure of ‘goodness to fit’ for the applied regression model.

Here value of R2 to explains only 48.05% variations. Meaning my model is fit 48.05% of times with the linear regression model is good fit.

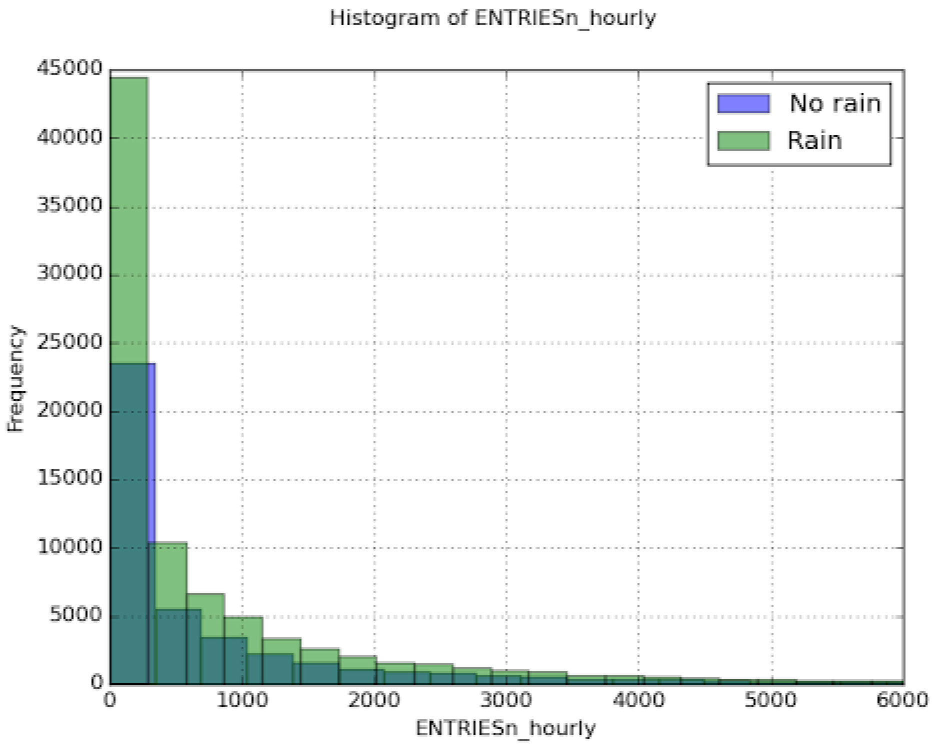
So I think Linear regression model is appropriate for this dataset and for the use case worked upon; where I am to work on linear relation of rain with subway ridership. But were I to add some more features like subway riding safety and security in situation of rain; it would not be wise to use Linear regression model.

But the current variation for R2 can still be worked upon can can still be improvised with more features that could help understand subway ridership surge.

Section 3. Visualization

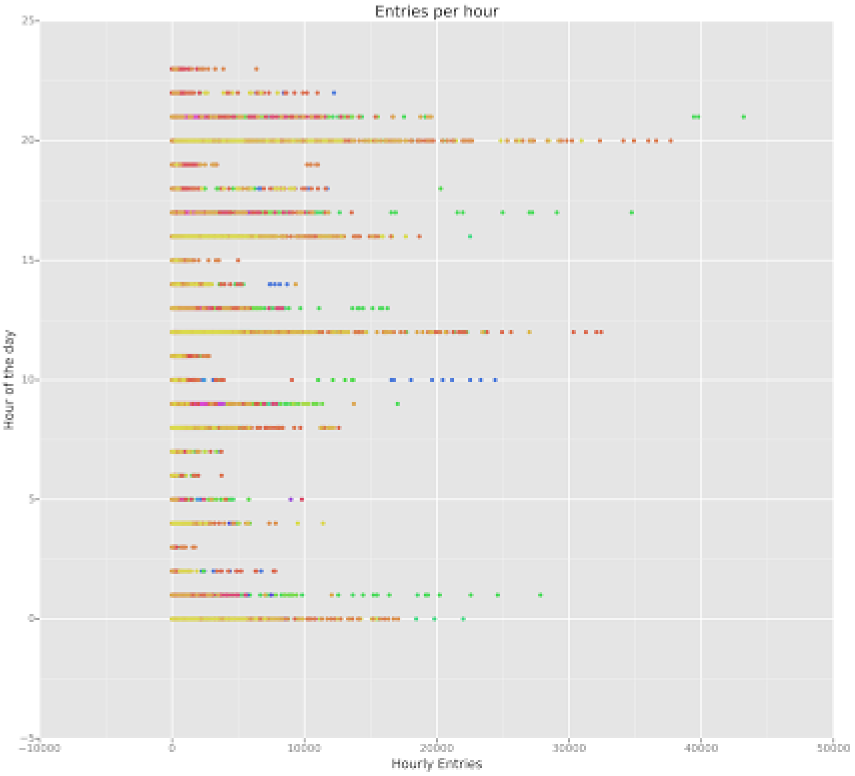
3.1 One visualization should contain two histograms: one of ENTRIESn_hourly for rainy days and one of ENTRIESn_hourly for non-rainy days.

This histogram is mapped for ENTRIESn_hourly for NYC subway data set; for rainy and not rainy days. The Histogram very shows that the people ride more on rainy day compared to non rainy day.



3.2 One visualization can be more freeform. You should feel free to implement something that we discussed in class (e.g., scatter plots, line plots) or attempt to implement something more advanced if you'd like. Some suggestions are:

The plot is the plot for entries per hour vs the hours features; every entries per hour are color coded by Units. The diagram represents entries in units color coded; indicating that certain units are more utilized at certain hours.



Section 4. Conclusion

4.1 From your analysis and interpretation of the data, do more people ride the NYC subway when it is raining or when it is not raining?

As per my analysis and interpretation of data, people do ride more on NYC subway when it’s raining compared to when it’s not; and the first and more than 95% conclusive test I did was Mann-Whitney U test; which had critical P-value of 0.25.Though the test was not helpful for predictions.

4.2 What analyses lead you to this conclusion? You should use results from both your statistical tests and your linear regression to support your analysis.

First what I did was i came to conclusion that people do ride more on NYC subway while it's raining; for that I calculated critical P-value (=0.25 for this case) and U (1924409167.0) giving us sufficient doubt to believe the null hypothesis; and used mean values of riders on rainy and non rainy days. Concluding that more people do ride on subway of NYC on rainy days.

Now using this using linear regression I tried to build a basic prediction system based on the data; and was able to understand that rain (0 or 1) is not the only feature on based of which subway ridership is depends but there are other feature that indicate such as meanwind, meantemp, meandew, fog that indiacete bad weather. And adding those other features helps me to increase accuracy of my model.

This I understood once I tested the model with the four base features first and then kept on adding features that would indicate bad weather and motivate someone to not to be on the street; and my reasoning was justified with the increase of R2 value as I kept adding features.

Hence I conclude that rain definitely increases subway ridership but there as other features that affect the ridership that are not as direct as rain.

Section 5. Reflection

5.1 Please discuss potential shortcomings of the methods of your analysis, including:

1. Dataset,
2. Analysis, such as the linear regression model or statistical test.

Potential shortcomings of analysis:
The linear regression models is certainly the appropriate choice as the study appears to be of linear relationship for the subway ridership on rainy vs non-rain days.

Potential shortcomings of Dataset:
One, The size of the dataset can be increased it would have helped more in linear regression; which compared to MannWhitney's U test does not just based on riders per rainy day for all subway stations of NYC.

Just by increasing data points for other subway units in not enough though; we need more data for each units as well. This would allow us to draw better insight per unit of subway.

Also we can may be have a look at exits data more clearly and understand why some units have more exits than entry.

Bibliography:

http://en.wikipedia.org/wiki/Ordinary_least_squares
http://en.wikipedia.org/wiki/Gradient_descent
http://www.ats.ucla.edu/stat/mult_pkg/faq/general/tail_tests.htm
http://en.wikipedia.org/wiki/Null_hypothesis
sklearn.linear_model.LinearRegression — scikit-learn 0.14.1 documentation
yhat/ggplot
http://pandas.pydata.org/pandas-docs/stable/10min.html
http://www.statsoft.com/Textbook/Multiple-Regression#cresidual