Section 0. References

a) Histograms: http://pandas.pydata.org/pandas-docs/stable/visualization.html#histograms. discussions.udacity.com/t/ggplot-histograms-opening-blank-windows-in-ipython-notebook/19624
b) Mann-Whiteny U test: http://en.wikipedia.org/wiki/Mann%E2%80%93Whitney_U_test http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
c) Shapiro-Wilk test: http://en.wikipedia.org/wiki/Shapiro%E2%80%93Wilk_test
d) Linear Regression: http://statsmodels.sourceforge.net/0.5.0/generated/statsmodels.regression.linear_model.OLS.html http://statsmodels.sourceforge.net/0.6.0/examples/notebooks/generated/predict.html http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.LinearRegression.html http://scikit-learn.org/0.14/modules/generated/sklearn.linear_model.SGDRegressor.html http://stackoverflow.com/questions/30387365/how-to-use-sgdregressor-in-scikit-learn
e) Ggplot: https://pypi.python.org/pypi/ggplot/
f) Python: https://docs.python.org/2/library

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?

ANS: I used non-parametric test; i.e., the Mann Whitney U test to analyze the NYC subway data. This test provides a one-tail p-value.

Null Hypothesis: The 2 samples of data (number of Hourly entries on rainy days and entries on non-rainy days) come from the same population. i.e., Average #Hourly entries are same for rainy and non-rainy days. (That is to say that the probability of #Hourly entries on rainy days > non-rainy days is same as probability of #Hourly entries on non-rainy days > rainy days).

Alternative Hypothesis: The samples of data for number of Hourly entries on rainy days and entries on non-rainy days come from different populations.

The one-tail p-value obtained for the sample NYC subway data for May 2011 is 0.0249. I considered p-critical value of 0.05 (95% significance level).

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.

ANS: Plotting the data on a Histogram shows that the data sets do not follow a normal distribution. This is further supported by using Shapiro Wilks test. With null hypothesis that sample of data is drawn from a population that is normal, probability of getting test statistic

w=0.4716 for rainy days or w=0.4766 for non-rainy days is 0.0. Hence the alternative is true that the samples do not belong to a population that is normal. So t-tests do not apply.

Non-parametric tests like Mann-Whitney U test need to be used that do not assume data to be from any particular distribution.

In addition the sample data is only for the month of May. It cannot be considered a large enough representation of the population. Here we can assume the total population (training data) as ridership since NYC subway was setup. The data over the years the turnstile has been functioning is a subset of the past ridership.

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.

ANS: Using Mann Whitney U test with

Null Hypothesis: The 2 samples of data (number of Hourly entries on rainy days and entries on non-rainy days) come from the same population and the mean entries for both samples are equal.

Alternative Hypothesis: The samples of data for number of Hourly entries on rainy days and entries on non-rainy days come from different populations and the mean entries for both samples are not equal.

- Test statistic (U) = 1924409167.0
- One sided p-value is 0.0249.
- #of Rainy days sample records = 44,104
- Mean for Rainy days sample records =1105.446377
- #of Non Rainy day sample records = 87,847
- Mean for Non Rainy days sample records=1090.278780

1.4 What is the significance and interpretation of these results?

ANS: This means probability of getting U value of 1924409167, if we considered the data sets for rainy day entries and non-rainy day entries from same population is 2.49%. Considering 95% significance level this p-value 0.0249 is less than p-critical of 0.05. Hence mean rainy day entries being greater than mean non-rainy day entries in these samples are statistically significant.

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?

ANS: I used 3 approaches to compute coefficients theta and produce prediction for ENTRIESn_hourly a) Ordinary Least Sqaures using Statsmodels.

b) Stochastic Gradient descent using Scikit Learn, linear_model (with initial learning rate - eta0=0.0004)

c) By writing code to descend the gradient. The program starts with coefficient 0 and programmatically updates coefficients through given number of iterations with given alpha value.

The model that gave me best value for R2 coefficient was a) OLS using Statsmodels

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

ANS: Features I used and decided to not use are below

A) Used following input variables:

a) All dummy variables, except unit_R459 representing the turnstile unit as a part of the features.
    - Units 0-400 contribute significantly to r^2 value.
    - Units 400 onwards contribute ~0.015 to r^2.
    - unit_R459 has minimum coefficient (theta) value – hence dropped it from the feature set.
b) Hour: Among the rest of the input variables this most (~ 0.03pts to to r^2 value).
c) Fog
d) Minimum temperature
e) Mean wind speed
f) Mean pressure

B) Did not use rest of input variables as they contribute very little (~0.002) to increase in r^2 value ('rain', 'maxtempi', 'precipi', 'meandewpti', 'meantempi', 'minpressurei', 'maxpressurei', 'maxdewpti','mindewpti').

C) Thunder cannot be used as it has a single uniform value=0 in the entire dataset and hence will not contribute as an estimator to predict.

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.

ANS: I initially used only input variables like ‘Hour’, ‘rain’, ‘fog’, ‘meantempi’ – but the R2 value was < 0.04. On adding the turnstile units as dummy variables the R2 value increased. On experimenting with different input variables I found that:

- Units 0-400 contribute significantly to r^2 value. 'unit_R170' has maximum coefficient (theta) value.
- ‘Hour’ contributes significantly (~0.03) to r^2 value.
- Units 400 onwards contribute 0.015 to r^2. unit_R459 has minimum coefficient (theta) value, hence decided to drop it from the feature set.
- 'fog' 'mintempi' 'meanwindspdi' 'meanpressurei’ contribute to ~0.003 to r^2
- The rest of input variables like can be added, but contribute very little (~0.002) to increase in r^2 value. ('rain', 'maxtempi', 'precipi', 'meandewpti', 'meantempi', 'minpressurei', 'maxpressurei', 'maxdewpti','mindewpti')
- Thunder cannot be used as it has a single uniform value=0 in the entire dataset and hence will not contribute as an estimator to predict.

Finally I used the following features of a) all turnstile units (as dummy variables) except unit_R459, b) Hour, c)Fog, d) Minimum temperature e) Mean wind speed f) Mean pressure

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

ANS: Coefficients for non-dummy features in my linear regression model (OLS) are:

a) 'Hour': 65.40384071
b) 'fog': 169.95793528
c) 'meanpressurei' : -215.53562424
d) 'meanwindspdi': 24.54976013
e) 'mintempi': -12.73487389

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

ANS: R2 for my Model with OLS using Statsmodel is 0.48087

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?

ANS: Here R2 value is ~0.48. The closer R2 is to 1, the better our model; as it means that the predicted values are closer to actual values. Here, as the value is not close to 1 (i.e. R2 is less than 0.5), I conclude linear model is not a best fit for this dataset.

Section 3. Visualization

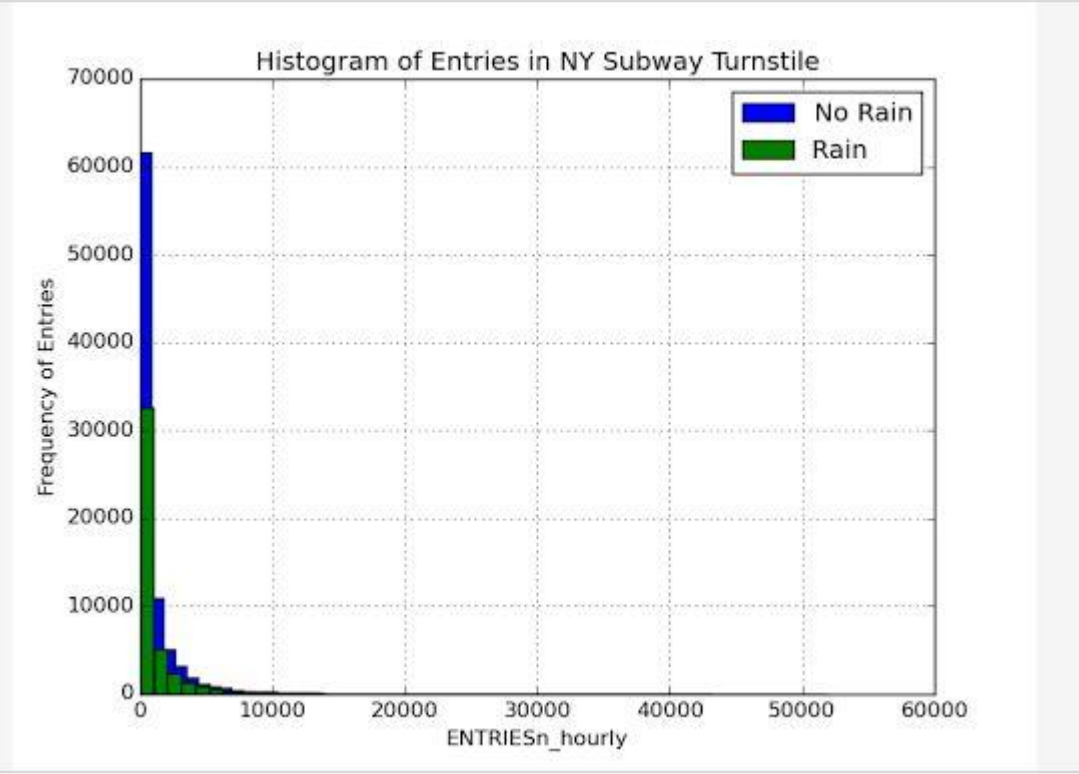
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.

Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.

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

- You can combine the two histograms in a single plot or you can use two separate plots.
- If you decide to use to two separate plots for the two histograms, please ensure that the x-axis limits for both of the plots are identical. It is much easier to compare the two in that case.
- For the histograms, you should have intervals representing the volume of ridership (value of ENTRIESn_hourly) on the x-axis and the frequency of occurrence on the y-axis. For example, each interval (along the x-axis), the height of the bar for this interval will represent the number of records (rows in our data) that have ENTRIESn_hourly that falls in this interval.

- Remember to increase the number of bins in the histogram (by having larger number of bars). The default bin width is not sufficient to capture the variability in the two samples.



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:

- Ridership by time-of-day

- Ridership by day-of-week

Section 4. Conclusion

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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?

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.

Section 5. Reflection

Please address the following questions in detail. Your answers should be 1-2 paragraphs long.

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.

5.2 (Optional) Do you have any other insight about the dataset that you would like to share with us?