Udacity

Project 2: Analyzing the NYC Subway Dataset

Name: Anup Dudani

Section 0: References

1] Statistics Course https://www.udacity.com/course/viewer#!/c-ud134-nd
2] Python Course https://teamtreehouse.com/tracks/learn-python
3] For some statistic topics www.khanacademy.com
4] Naked Statistics: Stripping the Dread from the Data by Charles Wheelan
5] Welch's T-test: https://en.wikipedia.org/wiki/Welch%27s_t_test
6] Histogram with Pandas: https://bespokeblog.wordpress.com/2011/07/11/basic-data-plotting-with-matplotlib-part-3-histograms/
7] mannwhitneyu http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.mannwhitneyu.html
8] ManWhitney U test: https://statistics.laerd.com/spss-tutorials/mann-whitney-u-test-using-spss-statistics.php
9] ManWhitney U test: Udacity Notes
10] Pandas dummy variable http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html
11]

Section 1: Statistical Tests

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?

H_0 : There is no effect of rain on number of entries or there is decrease in number of of people taking subway when its raining.
H_a : There is increase in number of people.

I used ManWhitney U test to compare the effect of rain on subway ridership. By default the ManWhitney U test is one sided test, it requires to be multiplied by 2 when we want to do two sided test.

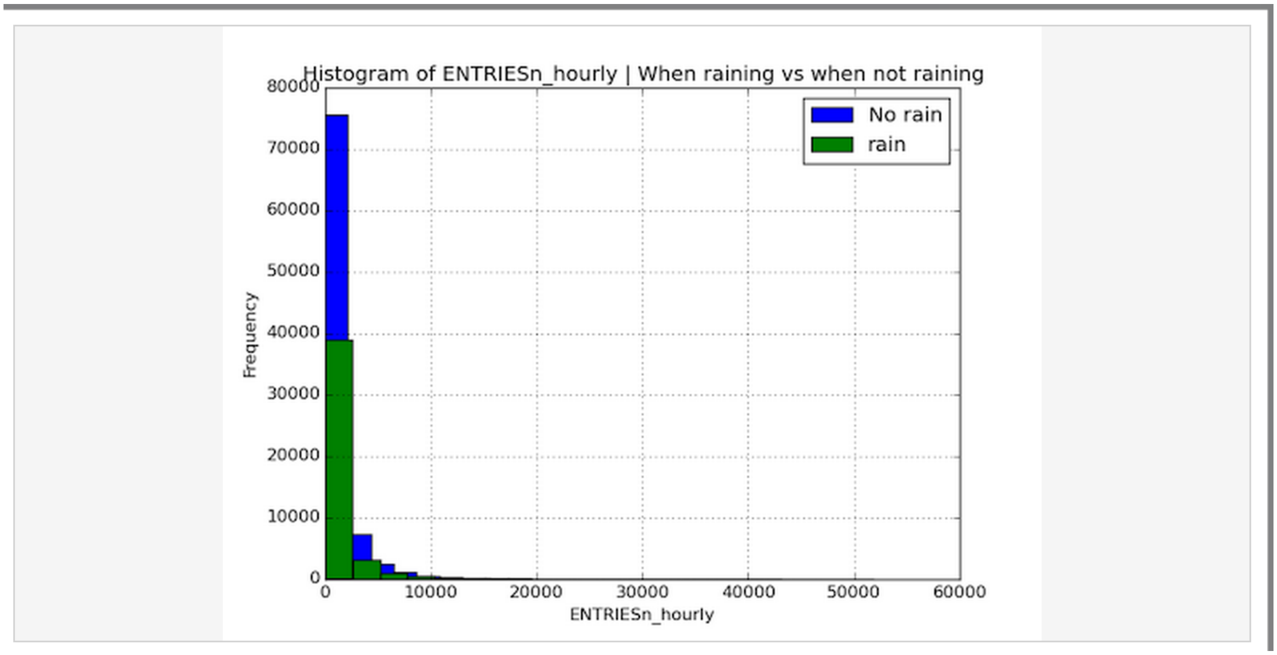
1.2. Why is this statistical test applicable to the dataset? In particular, consider the assumptions that the test is making about the distribution of ridership in the two samples.

Ans:

ManwhitneyU test (Wilcoxon ranksum test) is used to compare the independent groups when the dependent variables are either ordinal or continuous but not normally distributed. In graph we plotted between the Entries_hourly vs. frequency, we can see the data is continuous, but normal:

Assumption:
ManWhitney U test is good when the number of observation is >20 and you have 2 independent samples of ranks.

The reported p-value is for a one-sided hypothesis, to get the two-sided p-value multiply the returned p-value by 2.



So I applied ManwhitneyU test. We also considered Welch's t-test but as it is only applicable to normally distributed dependent variables, where two dependent variables may be in equal or unequal numbers.

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: The results are as follows:

```
p value: 0.024999912793489721
with_rain_mean: 1105.4463767458733
without_rain_mean: 1090.278780151855
U statistics: 1924409167.0
```

1.4 What is the significance and interpretation of these results?

- The number of entries when its raining is statistically different than when its not raining.
-

Section 2:
2.1 What approach did you use to compute the coefficients theta and produce prediction for ENTRIESn_hourly in your regression model:
Ans:

OLS using Statsmodels is used

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

I used the following features(input variables) :
1.rain
2. precipi
3. Hour
4. meantempi
5. meandewpti
6. mintempi

Yes, I used dummy variable as “dummy_units”. which is python pandas dummy variable.

More info: http://pandas.pydata.org/pandas-docs/stable/generated/pandas.get_dummies.html

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.
Reasoning behind selecting the inputs features for predictive model

- Your reasons might be based on intuition. For example, response for fog might be: “I decided to use fog because I thought that when it is very foggy outside people might decide to use the subway more often.”
- Your reasons might also be based on data exploration and experimentation, for example: “I used feature X because as soon as I included it in my model, it drastically improved my R^2 value.”

I selected these features, by experimenting which feature increase R^2 value and I found all of them increased the R^2 value.

I also though when its raining, temp, dew, and temp there will be more effect on ridership of people.

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

2.5 What is your model's R^2 (coefficients of determination) value?
Ans:
R^2 (coefficients of determination) value= 0.480025674116

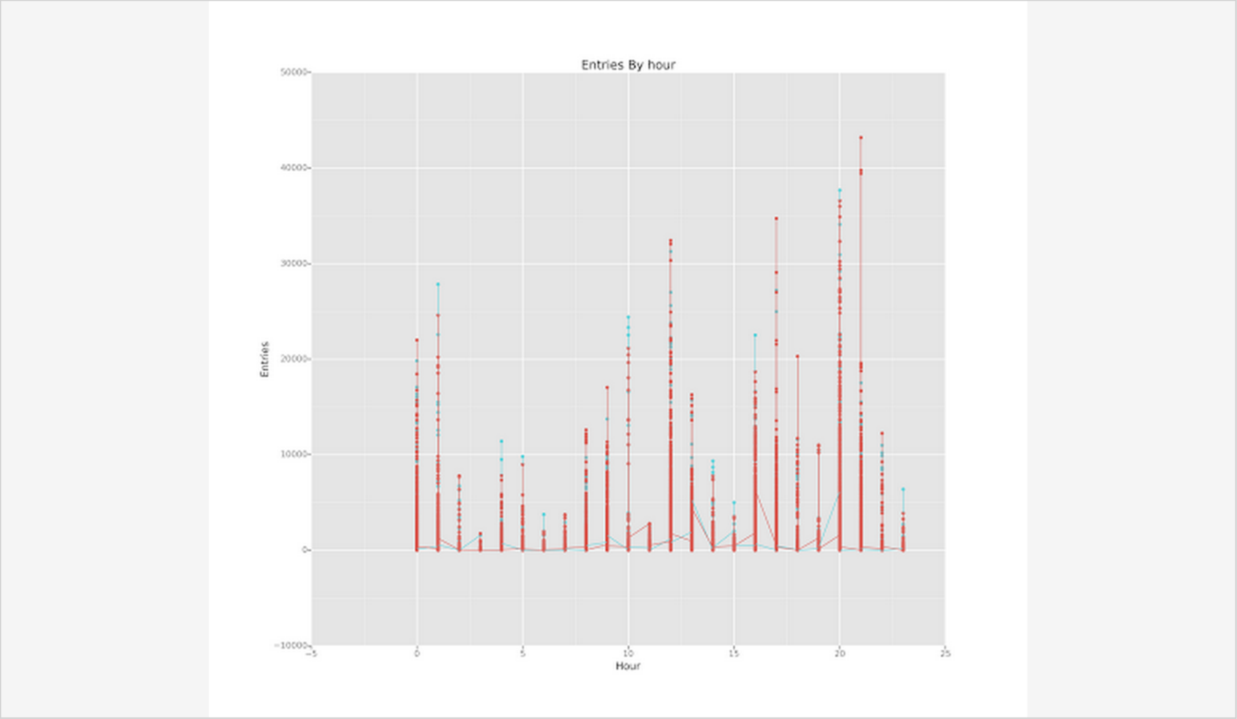
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

The more the R^2 (coefficients of determination) value, the better the prediction. So I think, this linear model to predict ridership is appropriate for this dataset, given this R^2 value.

Section 3:

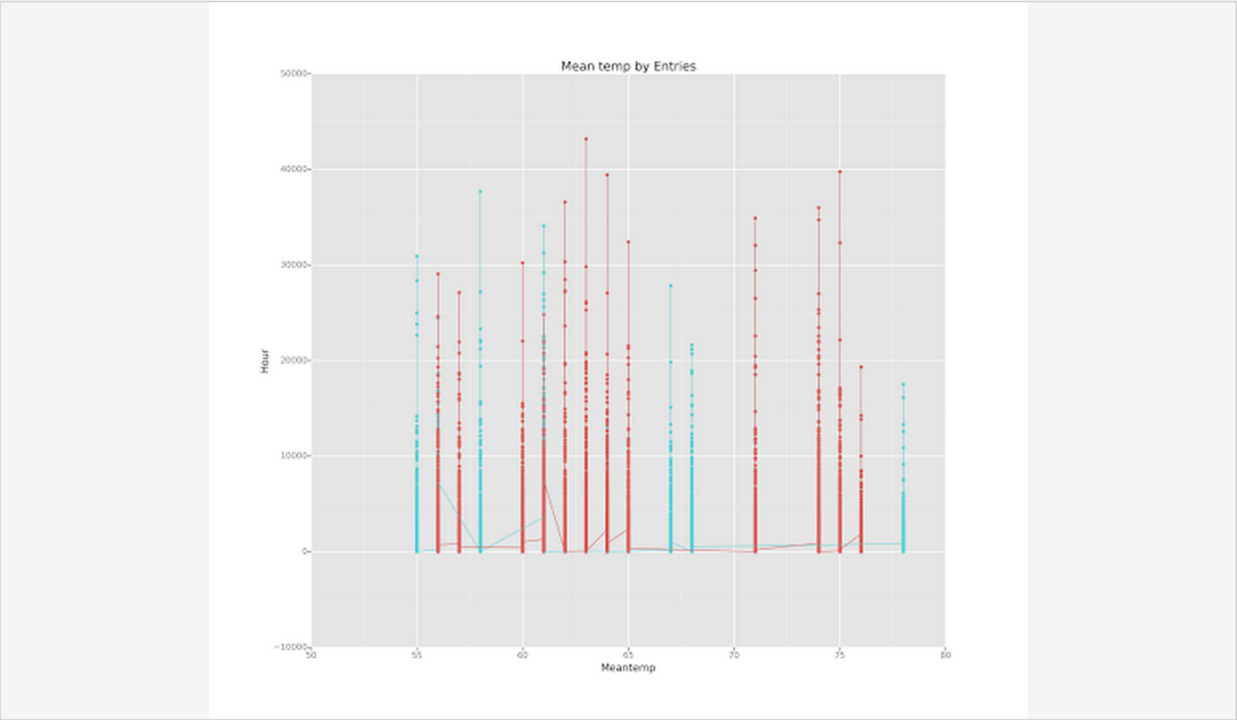
Please include two visualizations that show the relationships between two or more variables in the NYC subway data.
Remember to add appropriate titles and axes labels to your plots. Also, please add a short description below each figure commenting on the key insights depicted in the figure.
3.1 One visualization 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:

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:

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?