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] http://www.clockbackward.com/2009/06/18/ordinary-least-squares-linear-regression-flaws-problems-and-pitfalls/
12] Gradient Descent Linear Regression Intuition https://www.youtube.com/watch?v=aJTLVBzUw9M
13] Seaborn, https://stanford.edu/~mwaskom/software/seaborn/index.html

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?

ANS==>

- Hypothesis Statements

  H_0 : There is no effect of rain on number of hourly entries

  H_a : The number of hourly entries is different when its raining than when its not raining

- I used ManWhitney U test to compare the effect of rain on subway ridership. I used two tailed test with alpha level 0.05 (p- critical value).

  NOTE: By default the ManWhitney U test is one sided test, it requires to be multiplied by 2 when we want to do two tailed 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.

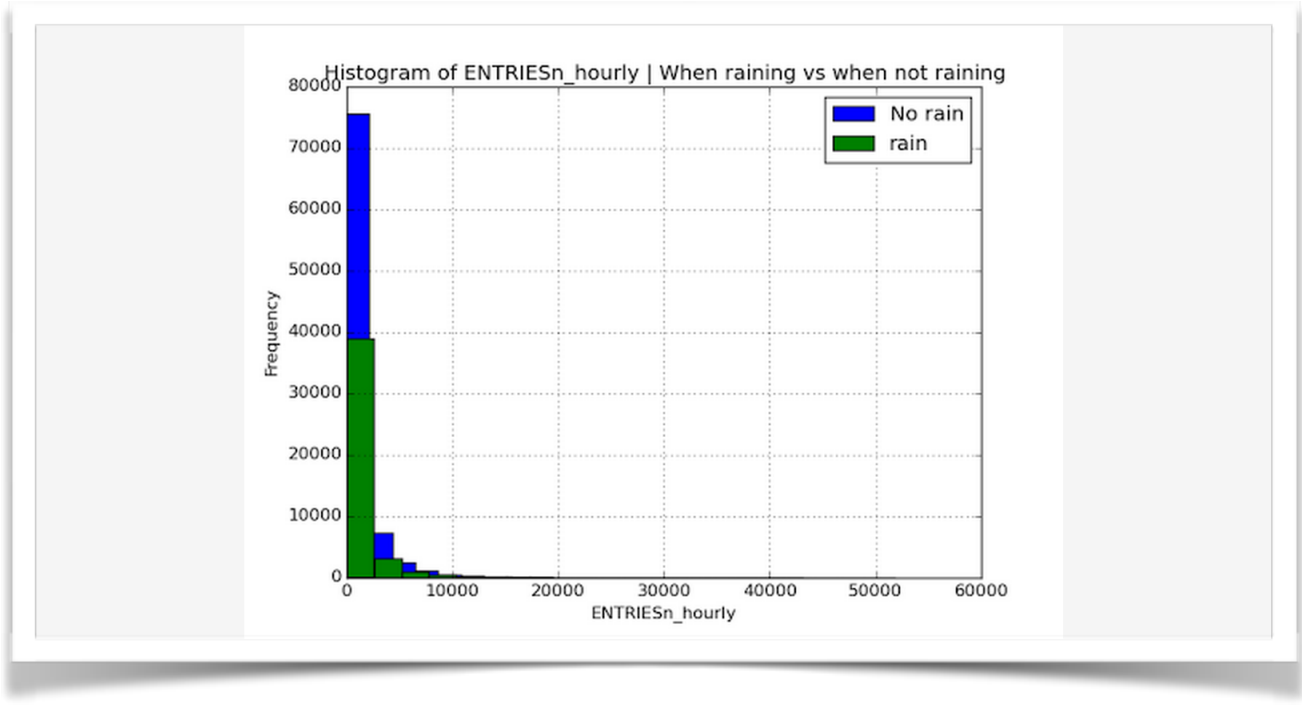


Fig: Distribution of Entries when its rating and when its not raining

So I applied ManwhitneyU test. I 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*2= 0.0499 (p value is multiplied by 2 since its a two tailed test)
- 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?

ANS==>

- Since p value(0.0499) is less than alpha level 0.05, the entries when its raining is significantly different than when its not raining. i.e. hourly entries when its raining is not equal to when its not raining
- So we can reject the null hypothesis i.e. (there is no effect of rain on number of entries)
- The mean number of entries when its raining is around 1105 and when its not raining its 1090.
- Difference between the two means is 15 entries

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:
ANS:

- OLS using Statsmodels is used which is a python module
- We also used gradient descent in lesson 3 exercise

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

ANS==> I used the following features(input variables) :

- rain
- precept
- Hour
- meantempi
- meandewpti
- mintempi

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

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.**

ANS ==>

- 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 thought 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:

Weights of the coefficients are as follows:

Non dummy Features	Parameters
rain	-11.31
precept	25.84
Hour	65.37
meantempi	19.24
meandewpi	11.16
mintempi	-45.73

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?
Ans==>

- R^2 is the coefficient of determination and is the square of correlation(r) between predicted response scores and actual response.
- R^2 value of 0 means the response cannot be predicted, while R^2 value of 1 means response can be predicted without error.
- R^2 between 0 to 1 indicates the extent to which the dependent variable is predictable which is in this case is ENTRIESn_hourly. An R^2 of 0.10 means that 10 percent of the variance in is ENTRIESn_hourly, predictable from features; an R^2 of 0.20 means that 20 percent is predictable; and so on.

- So R^2 of 0.48 has 48% predictability which I think is fairly good to predict ridership.

Section 3: Visualisation

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**

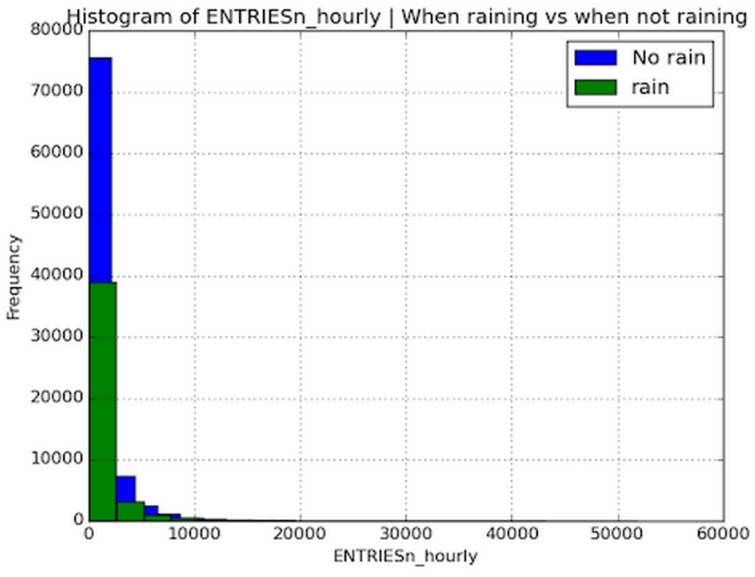
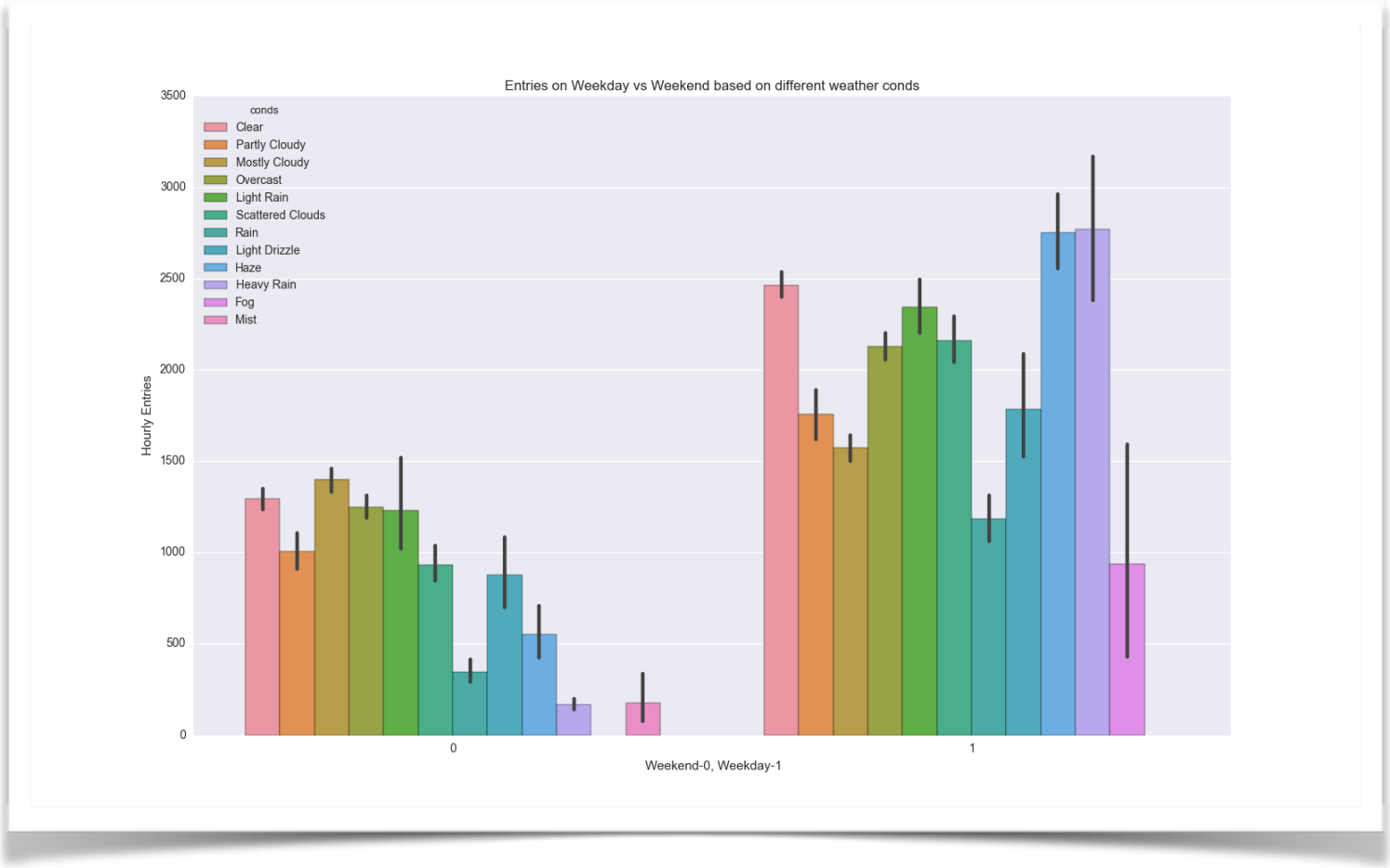


Fig: Entries per hour when its rain vs when its not raining

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**

Fig: Hourly Entries on Weekday Vs Non Weekday with different weather conditions



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?

The mean hourly entries when its raining is:
with_rain_mean: 1105.4463767458733
The mean hourly entries when its not raining is:
without_rain_mean: 1090.278780151855
I also did Mann Whitney U test which has p-value(0.0499) is less than p critical (0.05).
So people take NYC subway, more when its 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.

Ans==>

Mann Whitney U Test:

I did hypothesis testing on two independent samples i.e. hourly entries when its raining and hourly entries when its not raining. I choose mann whitney U test over welch's t-test, since welch's t test require the data to be normally distributed.

For the Mann Whitney U test, I got p value of 0.0499 which is less than the p critical i.e 0.05 so we rejected the null and accepted the alternate hypothesis which is "there is change in number of hourly entries when its raining than when its not raining"

Also we calculated mean entries when its raining and when its not raining which are 1105 and 1090 respectively.

Linear Regression:

When I did linear regression test without rain as a feature, I got R square value of 0.480023 and when I added 0.480025. So there is a increase in predictability when we add rain as feature.

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**

ANS==>

1) Dataset:

- During exercise 2.8 and 2.9 we make a cumulative of entries for particular hours, but that caused the some missing hours. Though that helped in easier analysis and prediction but that lead to misinterpretation as well.
- Also I see there is only the logged entries are for month of may in 2011. If we need to predict the ridership we would need more data spread across more months or years for better prediction

UNIT	DATEn	TIMEn	ENTRIESn	EXITSn	ENTRIESn_hc	EXITSn_hour	datetime	hour
R003	5/1/11	0:00:00	4388333	2911002	0	0	5/1/11 0:00	0
R003	5/1/11	4:00:00	4388333	2911002	0	0	5/1/11 4:00	4
R003	5/1/11	12:00:00	4388333	2911002	0	0	##########	12
R003	5/1/11	16:00:00	4388333	2911002	0	0	##########	16
R003	5/1/11	20:00:00	4388333	2911002	0	0	##########	20

Fig. Hour cumulative log problem

2) Analysis:

- Use of Gradient Descent:

  Gradient descent has a problem of local minima.

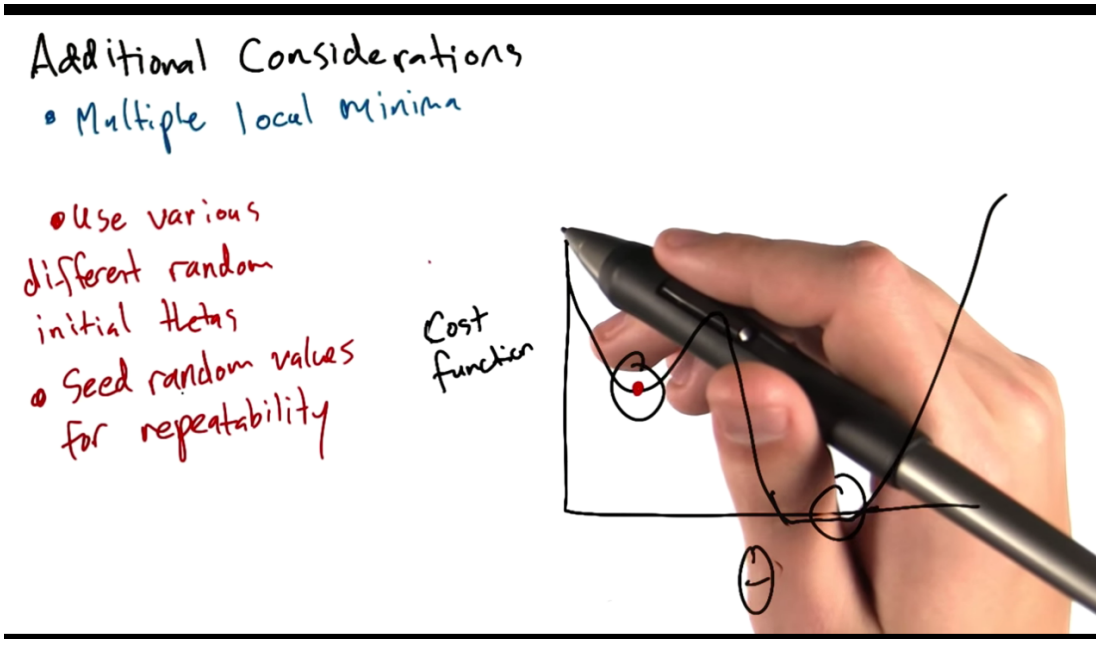


Fig. Local Minima problem.

Note: the figure is used for reference number 12 in reference list.

- Use of Mann Whitney U test:

  Mann Whitney U test works better when the size of the sample is bigger. So, it will be better to use bigger size of the sample

- Use of OLS Model

  Nonlinearity of datasets:

  Most linear regression systems suffer from a problem that most system in reality are not linear. As linear models try to fit the one dimensional data into line, two dimensional data into plane and higher dimensional data in to generalize plane (hyperplane). In practice most of the real world systems cannot be fit into generalize line, plane or hyperplane.

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