Judy Barker Jensen
Jbark1967@gmail.com
831-818-4471

Project 1 – Data Analyst Nanodegree

Section 0. References

Stackoverflow.com:

c# - Combine date and time when date is a DateTime and time is a string - Stack Overflow
http://**stackoverflow**.com/questions/4371231/removing-punctuation-from-python-list-items
python - How to change uppercase to lowercase in a string? - Stack Overflow
Best way to strip punctuation from a string in Python - Stack Overflow
python - Convert Date String to Day of Week - Stack Overflow
python - Run an OLS regression with Pandas Data Frame - Stack Overflow
gradient descent using python and numpy - Stack Overflow
python - Add item to pandas.Series? - Stack Overflow
python - How to add a single item to a Pandas Series - Stack Overflow
python - convert datetime format - Stack Overflow
combine multiple text files into one text file using python - Stack Overflow
Python TypeError: coercing to Unicode: need string or buffer, file found - Stack Overflow
python - TypeError: expected a character buffer object - Stack Overflow
python - TypeError: expected a character buffer object - while trying to save integer to textfile - Stack Overflow
Python concatenate text files - Stack Overflow
http://**stackoverflow**.com/questions/252703/python-append-vs-extend/252711#252711
http://**stackoverflow**.com/questions/22727291/duplicate-number-in-list-remove-in-python/22727510#22727510
hash - Python dictionary : TypeError: unhashable type: 'list' - Stack Overflow
python - Getting key with maximum value in dictionary? - Stack Overflow
http://**stackoverflow**.com/questions/8693116/performance-for-finding-the-maximum-value-in-a-dictionary-versus-numpy-array
python - Getting key with maximum value in dictionary? - Stack Overflow
Python take a list of numbers and return the average - Stack Overflow

Udacity forums:

discussions.udacity.com/t/deep-reverse-practice-problem-question/16600
discussions.udacity.com/t/lesson-2-exercise-5/16877
discussions.udacity.com/t/lesson-2-exercise-6/17107/7
discussions.udacity.com/t/lesson-4-plotting-syntax-error/18411
discussions.udacity.com/t/lesson-5-mapper-and-reducer-with-aadhaar-data/17840
discussions.udacity.com/t/lesson-5-quiz-2-cant-get-solution-code-to-work/16341
discussions.udacity.com/t/no-more-help-for-free/16461
discussions.udacity.com/t/problem-set-2-8-get-hourly-entries/15406
discussions.udacity.com/t/problem-set-2-part-11-includes-answers/15205
discussions.udacity.com/t/problem-set-2-q-11-date-format-how-to-truncate-the-time/17340
discussions.udacity.com/t/problem-set-3-8-more-linear-regression-optional/15337
discussions.udacity.com/t/problem-set-3-q1-spoiler-alret-cant-get-code-to-run-on-my-local/15811
discussions.udacity.com/t/problem-set-4-how-to-get-day-of-week/15617
discussions.udacity.com/t/problem-set-4-is-there-a-bug-of-ggplot/16729
discussions.udacity.com/t/problem-set-5-3-spoilers/15652

Judy Barker Jensen
Jbark1967@gmail.com
831-818-4471

discussions.udacity.com/t/problem-set-5-discussions-no-spoilers-please/15195
discussions.udacity.com/t/problem-set-5-question/17751/2
discussions.udacity.com/t/problem-set-5-question-2-problem-not-averaging/17325
discussions.udacity.com/t/spoiler-three-questions-about-the-subway-dataset-and-project-1/17174
discussions.udacity.com/t/use-of-checking-and-continue-in-lesson-5-quiz-3-aadhaar-data/18760
discussions.udacity.com/t/use-of-checking-and-continue-in-lesson-5-quiz-3-aadhaar-data/18760/2
http://discussions.udacity.com/t/deep-reverse-practice-problem-question/16600/2
http://discussions.udacity.com/t/lesson-2-exercise-5/16877/3
http://discussions.udacity.com/t/lesson-2-exercise-6/17107/8
http://discussions.udacity.com/t/lesson-4-plotting-syntax-error/18411/2
http://discussions.udacity.com/t/lesson-5-mapper-and-reducer-with-aadhaar-data/17840/2
http://discussions.udacity.com/t/lesson-5-quiz-2-cant-get-solution-code-to-work/16341/10
http://discussions.udacity.com/t/lesson-5-quiz-3-aadhaar-data/18760/1
http://discussions.udacity.com/t/no-more-help-for-free/16461/6
http://discussions.udacity.com/t/problem-set-2-8-get-hourly-entries/15406/2
http://discussions.udacity.com/t/problem-set-2-part-11-includes-answers/15205/2
http://discussions.udacity.com/t/problem-set-3-8-more-linear-regression-optional/15337/9
http://discussions.udacity.com/t/problem-set-4-how-to-get-day-of-week/15617/6
http://discussions.udacity.com/t/problem-set-4-is-there-a-bug-of-ggplot/16729/5
http://discussions.udacity.com/t/problem-set-5-3-spoilers/15652/2
http://discussions.udacity.com/t/problem-set-5-3-spoilers/15652/9
http://discussions.udacity.com/t/problem-set-5-question/17751
http://discussions.udacity.com/t/spoiler-three-questions-about-the-subway-dataset-and-project-1/17174/5
http://discussions.udacity.com/t/use-of-checking-and-continue-in-lesson-5-quiz-3-aadhaar-data/18760/3
Problem set 4: Is there a bug of ggplot? - Data Analyst ND - Apr '15 / P1: Intro to Data Science - Udacity Discussion Forum
Problem Set 5 Discussions **NO SPOILERS PLEASE*** - Data Analyst ND - Apr '15 / P1: Intro to Data Science - Udacity Discussion Forum
Use of checking and 'continue' in Lesson 5 quiz 3 (aadhaar data) - Udacity Discussion Forum
discussions.udacity.com/c/nd002-2015-04-01

Python.org documentation:

https://docs.**python.org**/2/library/datetime.html#
8.1. datetime — Basic date and time types — Python 2.7.10rc0 documentation
8.2. calendar — General calendar-related functions — Python 2.7.10rc0 documentation
pandas: powerful Python data analysis toolkit — pandas 0.16.0 documentation
Python Data Analysis Library — pandas: Python Data Analysis Library
Full Text Search: "hashtable" - Python Wiki
Documentation/Ref/TheStandardTypeHierarchy - Python Wiki
KeyError - Python Wiki
9.2. math — Mathematical functions — Python 2.7.10rc0 documentation
https://docs.**python.org**/3/library/statistics.html#averages-and-measures-of-central-location
9.7. statistics — Mathematical statistics functions — Python 3.4.3 documentation
15.7. logging — Logging facility for Python — Python 2.7.10rc1 documentation
https://docs.**python.org**/2/py-modindex.html#cap-m

Judy Barker Jensen
Jbark1967@gmail.com
831-818-4471

Python Module Index — Python 2.7.10rc1 documentation
Index — Python 2.7.10rc1 documentation
https://docs.**python.org**/2/library/datetime.html#
8.1. datetime — Basic date and time types — Python 2.7.10rc0 documentation
https://docs.**python.org**/2/library/datetime.html#datetime-objects
8.1. datetime — Basic date and time types — Python 2.7.10rc1 documentation
ggplot 0.6.5 : Python Package Index
Python Data Analysis Library — pandas: Python Data Analysis Library
https://docs.**python.org**/2/library/datetime.html#date-objects
7.1. string — Common string operations — Python 2.7.10rc1 documentation
https://docs.**python.org**/2/library/string.html#string-functions
https://docs.**python.org**/2/library/string.html#string-formatting
https://docs.**python.org**/2/library/string.html#string-constants
4. Built-in Types — Python 3.4.3 documentation

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?

a) The Mann-Whitney U test
b) One-tail P value
c) Null hypothesis – the distribution of the number of entries on a non-rainy day is the same as the number of entries on a rainy day.
d) p = 0.025

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.

a) The Mann-Whitney U test is a non-parametric statistical test that is useful for testing if two samples are the same (null hypothesis) when no assumptions can be made about the probability distribution.
b) Non-normal distribution as determined by plotting a histogram of the rain and non-rain hourly entries
c) One dependent variable (entries)
d) One or more independent variables (rain)

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, without_rain_mean, U, p = (1105.4463767458733, 1090.278780151855, 1924409167.0, 0.02499
9912793489721
```

1.4 What is the significance and interpretation of these results?
The null hypothesis, that there is no difference between the means of the rainy day entries and the non-rainy day entries, has a very low probability (<2.5%) of being true and should be rejected.

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:
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?
a. Rain, Precipitation, Hour, Mean temp, fog, Turnstile Unit (dummy variable), Day of week(dummy variable)
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.
a. Intuitively, it made sense that certain times and days of the week are busier than others based on rush hour traffic, and also that some stations will be busier than others (i.e. expect the financial district to be very quiet on

Judy Barker Jensen
Jbark1967@gmail.com
831-818-4471

the weekends). I also used the results.summary() feature in statsmodels to see which features had the highest co-efficients and then added and subtracted features to optimize R^2.

2.4 What are the coefficients (or weights) of the non-dummy features in your linear regression model? (run without the dummy variables)

rain	-4.7007
precipi	-30.5838
Hour	57.9852
Meantempi	-12.3691
fog	225.8609

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

a. $R^2 = 0.49167$

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?

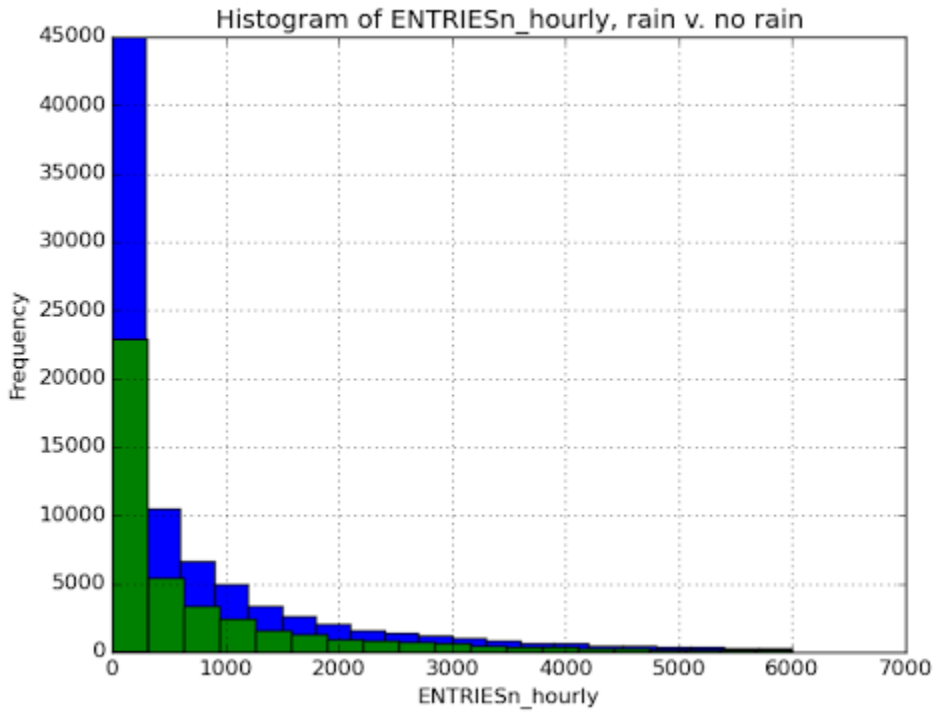
a. I would like to see a higher R^2 value – It may be possible to improve the R^2 value by removing outliers (i.e. exceptionally high numbers because there was a sporting event, or low numbers because of maintenance / delays), or perhaps there is interdependencies among the datapoints.

However, the residuals plot out in a normal distribution which would indicate differences between the observed and the predicted values are random.

I would use this model for rough estimates, but would look for a better model if more precision is involved. For example, it is probably suitable to indicate if more or fewer cars are required based on the weather forecast, time of day, and day of week.

Section 3. Visualization

Histogram of hourly Entries:

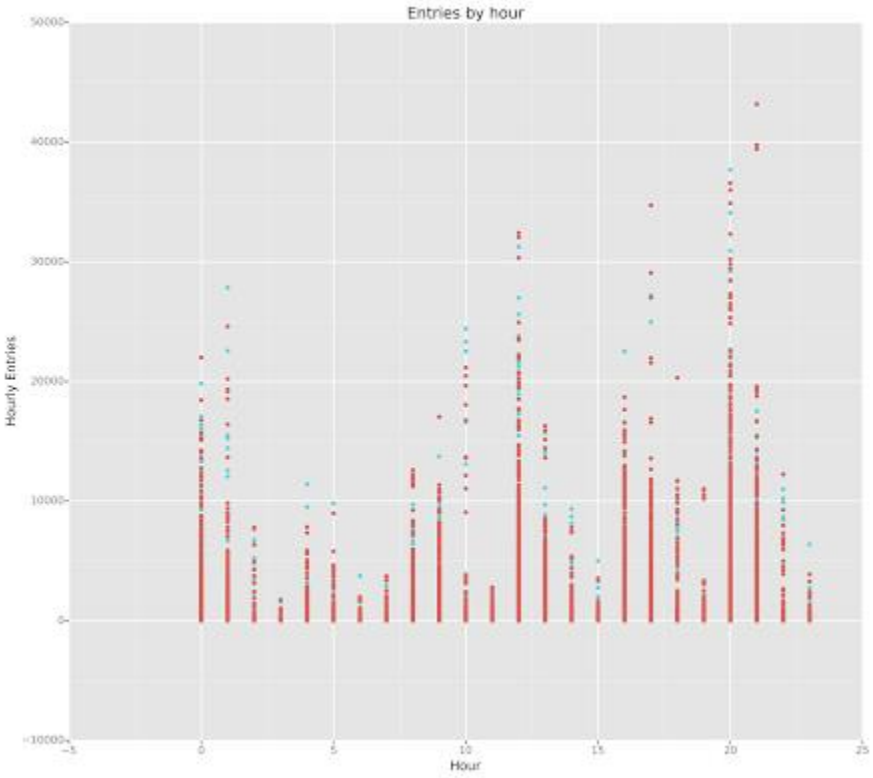


Judy Barker Jensen
Jbark1967@gmail.com
831-818-4471

This histogram graphs hourly entries when it is raining (blue) v. not raining (green). There are several outliers >6000, so the data is limited to the range of (0, 6000).

The key information on this histogram is that the distribution of entries is not normal (indicating Whitney Mann U test v. T-test for statistical analysis). It also appears that the ratio of entries w. rain is higher than with no rain. Finally, the graph shows the samples sizes are noticeably different.

Needs second graph



This plot of Hourly Entries by hour shows significant variation in ridership based on what hour of the day it is. It appears from this chart that when it rains, there is a spike in ridership as indicated by the blue points.

Section 4. Conclusion

I used the Whitney-Mann U test to conclude that the difference between the average number of entries when it was raining compared to not raining is statistically significant.

A deeper look at the data using ordinary least squares linear regression highlighted that the day of the week, the turnstile used (Unit) and the hour are the strongest predictors of ridership (these three factors result in an R2 of 49.0299, and adding rain as a factor increases the R2 to 49.03!

Also, digging into the coefficients, the negative value of rain would indicate that the presence of rain actual decreases ridership, despite the higher mean from the Whitney-Mann U test.

My conclusion is that the Unit, day of the week, and hour are the most significant predictors of subway ridership. Weather factors are small, but still statistically significant effect.

Judy Barker Jensen
Jbark1967@gmail.com
831-818-4471

Section 5. Reflection

A. Data Shortcomings

One noticeable data shortcoming is that the data only covers one month. I would expect that temperature extremes, as well as the presence of snow would have a noticeable effect on subway ridership, but this is not significant in the model based on a very comfortable mean temperature range of 55-76F.

Other useful information that might improve the predictive value of the model is special event information (baseball season), and subway line / destination (i.e. airport, business district, shopping, or event center).

B. Analysis Shortcomings

There is a clear difference between the Whitney-Mann results (strong probability the populations are not the same) and the R2 results from the OLS linear regression tests. My stats background is not sufficient to postulate if this is due to the independent variables having dependencies on each other, or perhaps the correlation is non-linear.

The histogram of the residuals shows a normal distribution, however a scatter plot of the predicted v. the observed values is not linear – which would indicate a different regression model might yield a tighter fit.