Judy Barker Jensen
Jbark1967@gmail.com

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
http://stackoverflow.com/questions/22377539/plotting-one-scatterplot-with-multiple-dataframes-with-ggplot-in-python

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

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

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
Other
https://onlinecourses.science.psu.edu/stat501/node/281

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) Two-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 (critical) = 0.05

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.

a) Mean, with rain: 1105.45
b) Mean, without rain: 1090.28
c) U = 1924409167
d) p = 0.025 (single sided => 0.050 two sided))

```
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, the distribution of the number of entries on a non-rainy day is the same as the number of entries on a rainy day, has a low probability (<0.05 of being incorrectly rejected. The threshold is 0.05 and so the null hypothesis is rejected and the distribution are statistically different.

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

Judy Barker Jensen
Jbark1967@gmail.com

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, Mean temp, fog, Turnstile Unit (dummy variable), Day of week(dummy variable), Hour(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 of day (Hour) and days of the week are busier than others based on rush hour traffic, and also that some stations (Units) will be busier than others (i.e. expect the financial district to be very quiet on the weekends). I confirmed these assumptions by graphing average entries (mean of ENTRIESn_hourly) by hour, day of week, and unit. This exercise also highlighted the need to create a dummy variable for hour as this is not a continuous variable.

I then graphed the various features v. average entries in excel and noted which ones resulted in a change in the trend line on the data, as well as the goodness of fit of the trend line (i.e. slope and R2).

With this information at hand, I tested my model with various parameters one at a time based on the results of my graphing and noted if they had any effect on R2, and also the coefficient and probability of the parameter looking for a high coefficient and a low P.

Intuitively, I expected that temperature, rain, fog and dewpoint would result in co-linearity issues with the model as temperature and dewpoint are directly related to rain and so tested the model with different combinations to maximize R2. Ultimately, I removed rain from the model as the mean dew point had a better effect on the model.

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

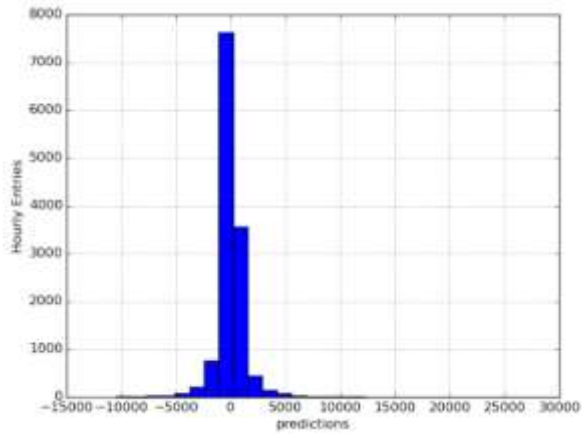
mean temp (meantempi)	-10.4
fog	136
average dew point (meandewpti)	-0.9

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

a. $R^2 = 0.53429$

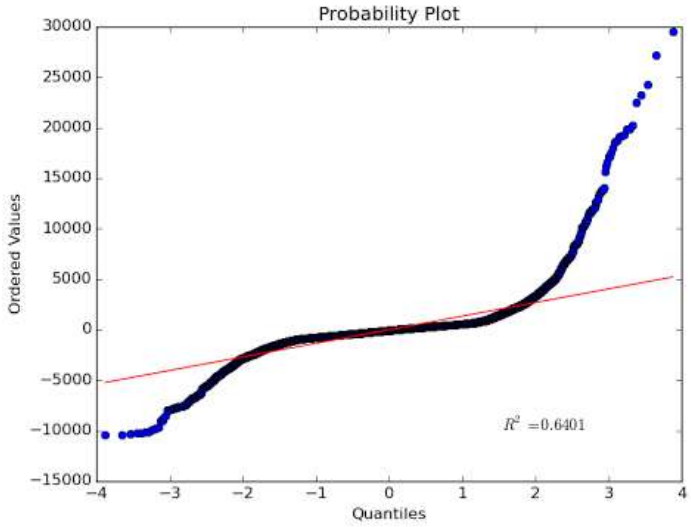
2.6 What does this R^2 value mean for the goodness of fit for your regression model? Do you think this linear model to predict ridership is appropriate for this dataset, given this R^2 value?

a. Fit – the model explains 53% of the variability of entries per hour based on the parameters of mean temp, fog, average dew point, and our dummy variables of hour, day of week, and entry unit. A histogram of the residuals shows there is a wide variance from the predicted value but the shape appears normal (Gaussian)



Given the long tail on both sides, I completed a probability plot which shows that non-linear distribution of the residuals. This plot indicates that our linear model is not a good fit for our data.

Judy Barker Jensen
Jbark1967@gmail.com

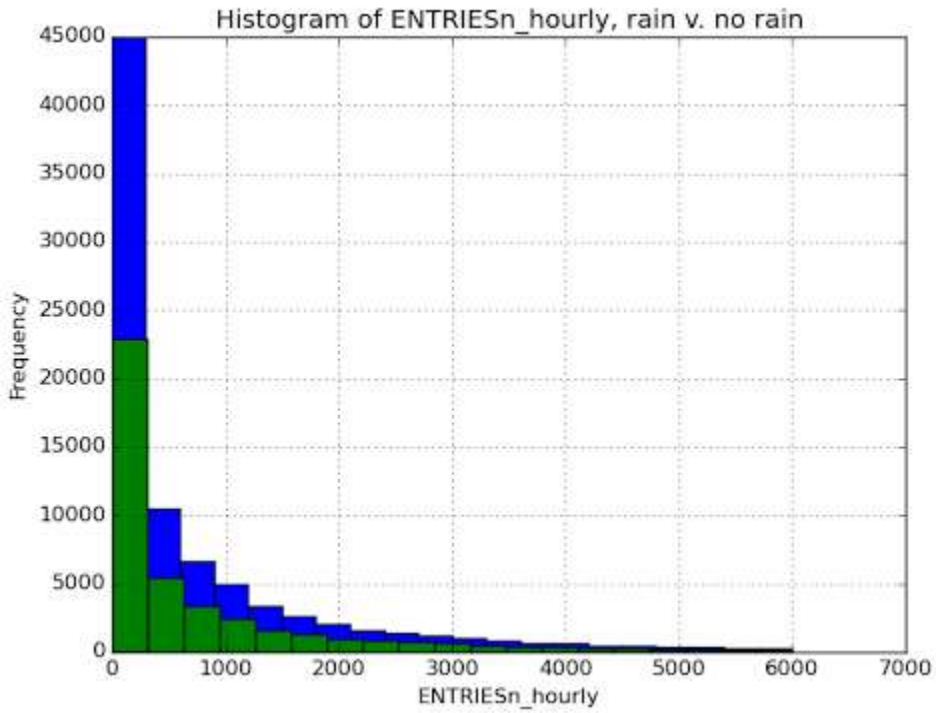


This is also indicated by a scatter plot of predicted v. actual entries where the model fit is poor near 0, and also at higher # of entries (i.e. greater than 4k). This makes sense intuitively – the model will predict negative entries which is not possible, and there are a significant number of actual very high entries which can not be attributed to any of the variables in the model.

Predictive value – the precision of the model may be suitable for making gross assumptions on the expected ridership on the subway, however based on the histogram, it is common to get an errors > +/- 1000 from the prediction, so anything but gross assumptions on the number of riders is not appropriate.

Section 3. Visualization

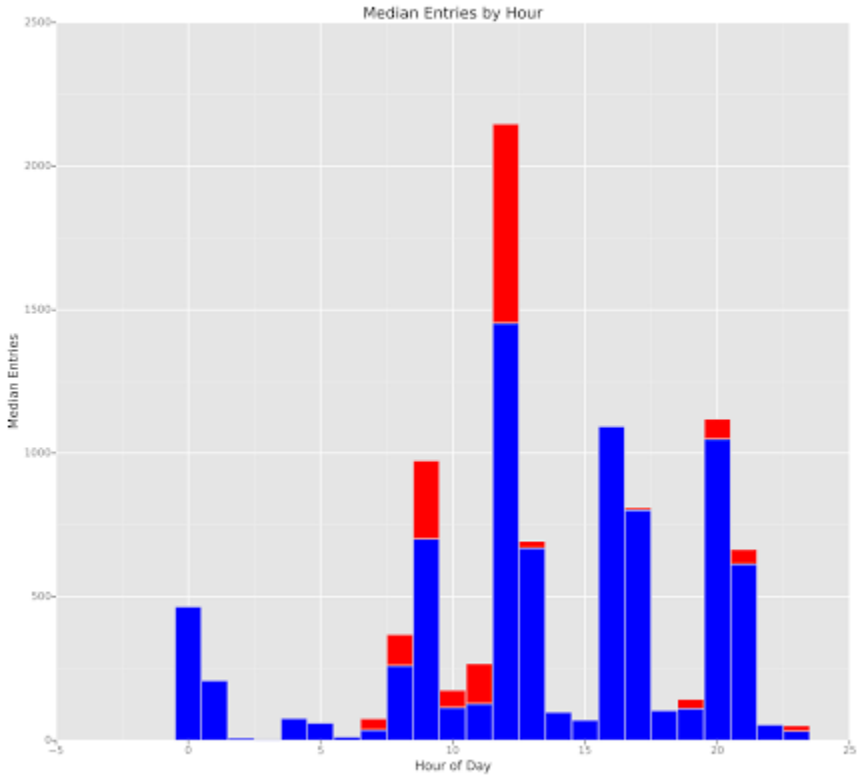
Histogram of hourly Entries:



Judy Barker Jensen
Jbark1967@gmail.com

This histogram graphs hourly entries when it is raining (Green) v. not raining (Blue). There are several outliers >6000, so the data is limited to the range of (0, 6000). (ggplots is not plotting legends. Corrected color coding in description)

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.



This plot of median Entries by hour demonstrates the difference between Entries when it is foggy (red) and not foggy (blue). This chart clearly shows the uneven distribution of entries across the various hours in the day which reinforces the importance of the Hour parameter. It also shows a high variability of the presence of fog. The distribution is unexpected, and I have included the procedure (from PS4:2):

```
def plot_weather_data(turnstile_weather):
    '''
    plot_weather_data is passed a dataframe called turnstile_weather.
    Use turnstile_weather along with ggplot to make another data visualization
    focused on the MTA and weather data we used in Project 3.

    You can check out the link
```

Judy Barker Jensen
Jbark1967@gmail.com

```
https://www.dropbox.com/s/meyki2wl9xfa7yk/turnstile_data_master_with_weather.csv
to see all the columns and data points included in the turnstile_weather
dataframe.
However, due to the limitation of our Amazon EC2 server, we are giving you a random
subset, about 1/3 of the actual data in the turnstile_weather dataframe.
'''
turnstile_weather.is_copy = False
pandas.options.mode.chained_assignment = None
x = 'Hour'
y = 'ENTRIESn_hourly'

Turnstile_fog = turnstile_weather.loc[turnstile_weather.fog == 1,[x, y]]
Turnstile_nofog = turnstile_weather.loc[turnstile_weather.fog == 0,[x, y]]
TS_fog = Turnstile_fog.groupby(x).median().reset_index()
TS_nofog = Turnstile_nofog.groupby(x).median().reset_index()
plot = ggplot(aes(x = x, y = y), data = TS_fog) + \
   geom_bar(stat = 'bar', fill = 'red') + geom_bar(aes(x=x), data = TS_nofog, stat = 'bar', fill = 'blue') +\
   ggtitle('Median Entries by Hour') + xlab('Hour of Day') + ylab('Median Entries')
return plot
```

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, however I did not include rain in my predictive model as other factors had a stronger affect, and were highly correlated to rain (i.e. temp and dew point) which introduced co-linearity issues.

A deeper look at the data using ordinary least squares linear regression highlighted that most of the variability has nothing to do with the weather! 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 0.53309 and adding the weather factors of min temp, fog and dewpoint increases the fit of the model (R2) to 0.53429, or 0.2%!

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 improve the fit of the model. .

Section 5. Reflection

A. Data Shortcomings

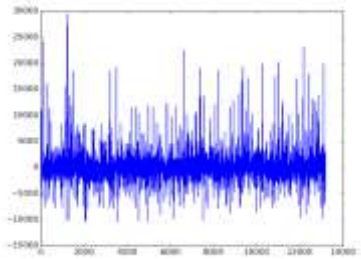
One noticeable data shortcoming is that the data only covers one month. I would expect that temperature extremes, as well as the presence of snow would have a noticeable effect on subway ridership, but this is not significant in the model based on a very comfortable mean temperature range of 55-76F.

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

Judy Barker Jensen
Jbark1967@gmail.com

The analysis resulted in a moderate R2 score. Two methods were utilized to judge the suitability of a linear model for this project. First, plotting the residuals over the data set indicates significant variations of the entries which indicates a poor linear fit.



Second, the probability plot of the residuals is not linear, and coupled with the histogram of the residuals, indicates a non-linear, heavy tailed distribution. This indicates that the relationship between the variables and the entries is not well described by the OLS method (linear regression) used.