STAT 177, CLASS 0

Richard Waterman

July 2020

WELCOME TO THE COURSE

OBJECTIVES

- Learn how to code Python.
- Focus on applications for data science.
- Familiarize with popular data science libraries.
- Learn about the statistical underpinnings of the analyses.
- Have some fun!

SYLLABUS REVIEW

- You read it.
- I'll review it.
- Resources (including Canvas and Piazza).
- Deliverables.
- Schedule.
- Grading.

PYTHON AND "DATA SCIENCE"

WHAT IS DATA SCIENCE ANYWAY?

- A Data Scientist is usually involved in various parts of the analytics pipeline:
  - Data capture.
  - Data maintenance.
  - Data exploration and analysis.
  - Reporting and presentation.
- What's the difference between a data scientist and a statistician?

TOP TEN REASONS FOR LIKING PYTHON

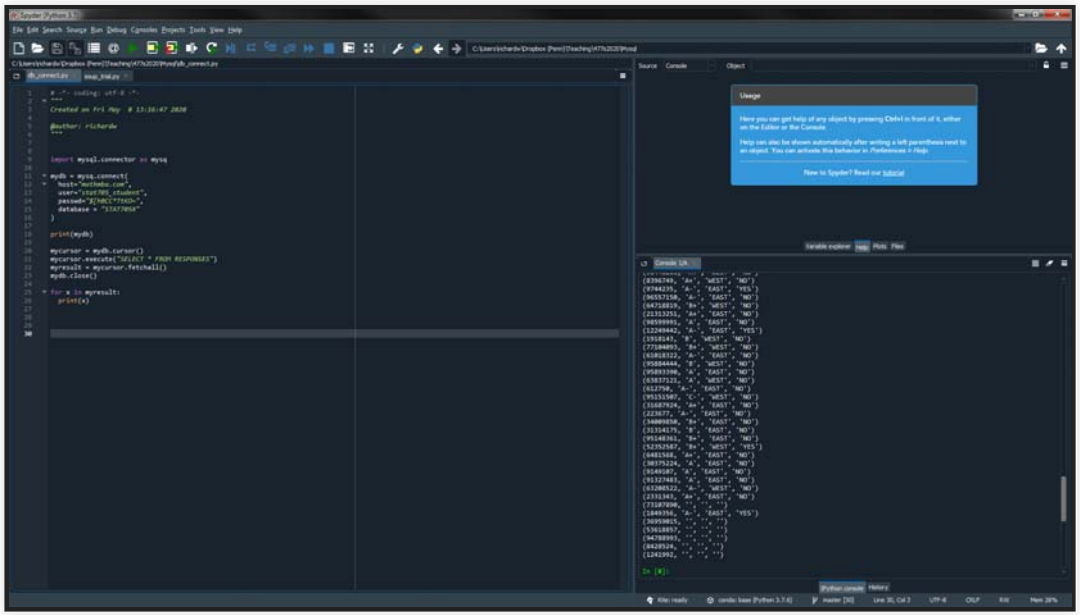
1. The most popular language for data science (but R is great too).
2. In demand/preferred skill.
3. All purpose software engineering framework (good at many different things).
4. Great for scientific computing (e.g. machine learning).
5. Scripting (formalizing a set of repetitive/boring tasks).

TOP TEN REASONS FOR LIKING PYTHON (CTD.)

6. Open source.
7. Extensible (add-on packages for specific activities, e.g visualization).
8. Large support community.
9. Generous learning curve.
10. At least 10 more reasons!

BY THE END OF THE COURSE ...

CODE SOME PYTHON



Python coding

IN 5 WEEKS, THE CODE BELOW SHOULD MAKE SENSE!

```
# Recode the levels of a categorical variable
print(carTable['Transmission'].value_counts())
recodes = {'A':'Automatic','AS':'Automatic','AV':'Automatic', 'M':'Manual'}
carTable['Transmission'] = carTable['Transmission'].map(recodes)
print(carTable.Transmission.value_counts())
```

RETRIEVE DATA (FROM VARIOUS SOURCES)

```
import pandas as pd
import os
from sklearn import tree

# This is where the data lives:
os.chdir('C:\\Users\\richardw\\Dropbox (Penn)\\Teaching\\477s2020\\DataSets')
print(os.getcwd())

carTable = pd.read_csv("Car08_just_499.csv")
print(carTable)
```

```
C:\Users\richardw\Dropbox (Penn)\Teaching\477s2020\DataSets
     Make/Model  MPG_City  MPG_Hwy  Weight(lb)  Seating  Horsepower  HP/Pound  \
0       Acura_RL        16       24        4014        5         290  0.072247
1       Acura_TL        17       26        3674        5         286  0.077844
2       Acura_TL        18       27        3559        5         286  0.080360
3      Acura_TSX        20       28        3345        5         205  0.061285
4      Acura_TSX        19       28        3257        5         205  0.062941
..           ...       ...      ...         ...      ...         ...       ...
494    Volvo_V50        20       28        3321        5         168  0.050587
495    Volvo_V70        16       24        3527        5         235  0.066629
496   Volvo_XC90        14       20        4356        5         235  0.053949
497   Volvo_XC70        15       22        4092        5         235  0.057429
498   Volvo_XC90        13       19        4826        7         311  0.064443

     Displacement  Cylinders  Origin Transmission     EPA_Class  Length Fuel  \
```

RETRIEVE FROM A DATA BASE

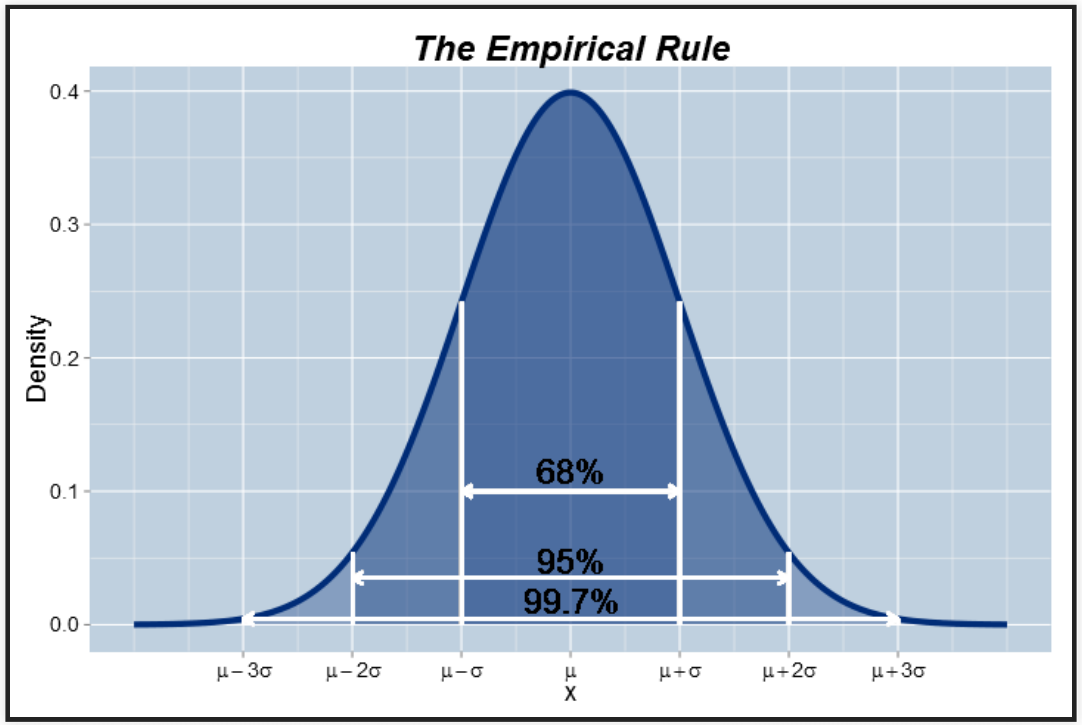
```
import mysql.connector as mysq

mydb = mysq.connect(
  host="mathmba.com",
  user="stat705_student",
  passwd="$[h0CC*TtKO~",
  database = "STAT705X"
)

mycursor = mydb.cursor()
mycursor.execute("SELECT * FROM RESPONSES")
myresult = mycursor.fetchall()
mydb.close()
```

```
(98616981, 'B', 'EAST', 'YES')
(31004395, 'A', 'EAST', 'NO')
(42889147, 'A-', 'EAST', 'NO')
(3215038, 'A', 'EAST', 'NO')
(21698690, 'B+', 'EAST', 'NO')
(22549780, 'A-', 'EAST', 'NO')
(26222822, 'A-', 'WEST', 'NO')
(23218140, 'B', 'EAST', 'NO')
(98114241, 'A+', 'EAST', 'NO')
(55528599, 'A-', 'EAST', 'NO')
(42039091, 'B+', 'EAST', 'NO')
(91812638, 'A+', 'EAST', 'NO')
```

```
(90446266, 'A+', 'WEST', 'NO')
(8396749, 'A+', 'WEST', 'NO')
(9744235, 'A-', 'EAST', 'YES')
```

LEARN SOME CORE STATS IDEAS

Making inferences:



The Empirical Rule

CLEAN AND WRANGLE DATA

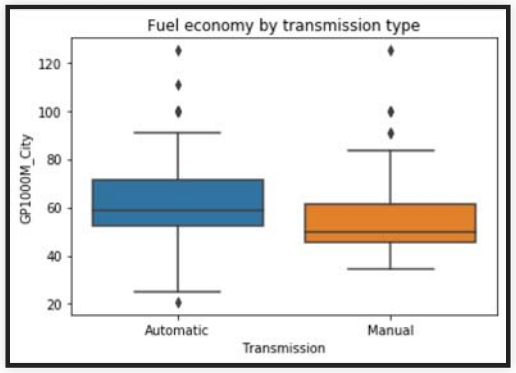
```
# Recode the levels of a categorical variable
print(carTable['Transmission'].value_counts())
recodes = {'A':'Automatic','AS':'Automatic','AV':'Automatic', 'M':'Manual'}
carTable['Transmission'] = carTable['Transmission'].map(recodes)
print(carTable.Transmission.value_counts())
```

```
A      217
AS     129
M      122
AV      31
Name: Transmission, dtype: int64
Automatic    377
Manual       122
Name: Transmission, dtype: int64
```

GRAPHICAL EXPLORATION OF DATA

```
import matplotlib.pyplot as plt
import seaborn as sns

sns.boxplot(x = 'Transmission', y = 'GP1000M_City', data = carTable)
plt.title('Fuel economy by transmission type')
```

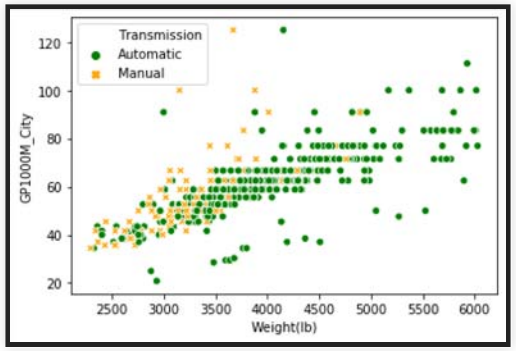
```
Text(0.5, 1.0, 'Fuel economy by transmission type')
```



BIVARIATE RELATIONSHIPS

```
sns.scatterplot(x = 'Weight(lb)', y = 'GP1000M_City', hue='Transmission', data=carTable,
  style='Transmission',
                    palette=['green','orange'], legend='full')
```

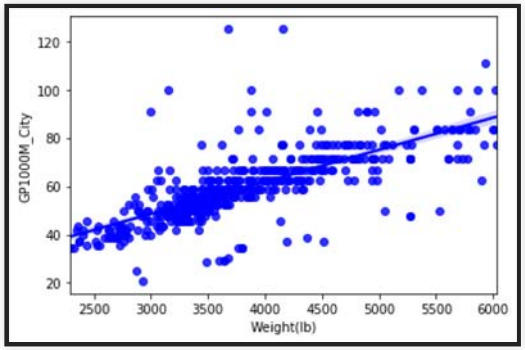
```
<matplotlib.axes._subplots.AxesSubplot at 0xc06cd88>
```



CREATE PREDICTIVE MODELS

```
# A scatterplot with the linear regression line and confidence bands
sns.regplot(carTable['Weight(lb)'], carTable['GP1000M_City'], color ='blue')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0xc116988>
```



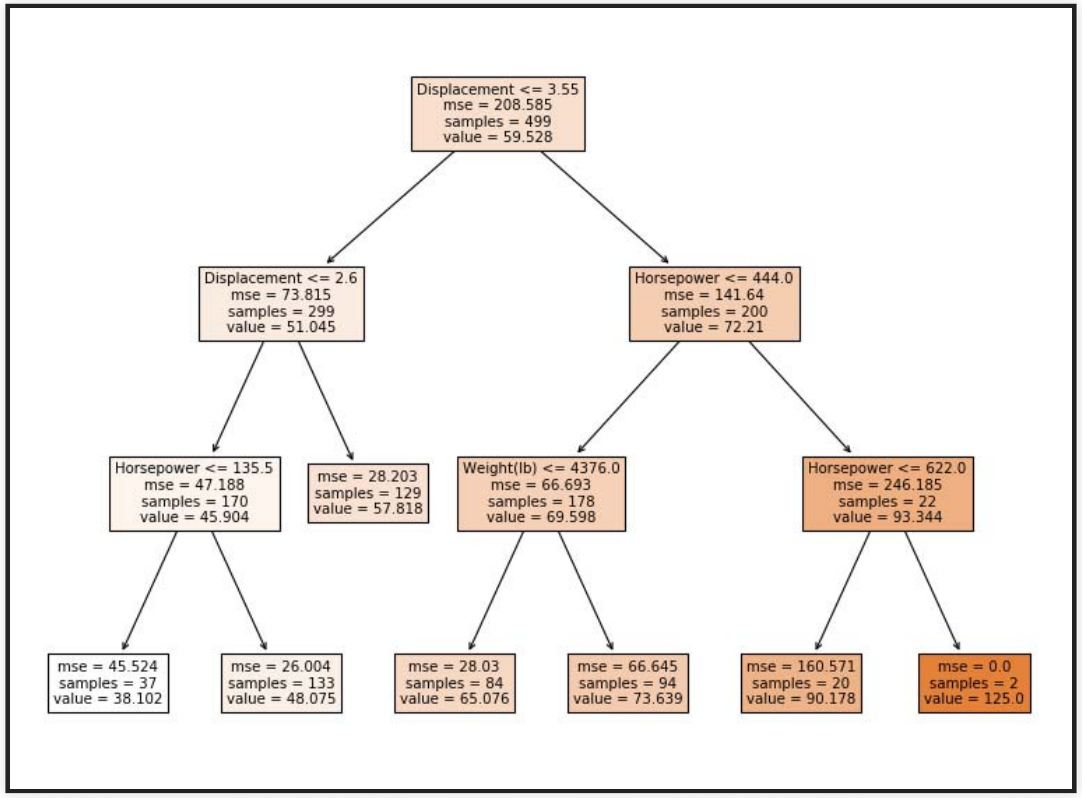
FIT A REGRESSION TREE

```
from sklearn.tree import DecisionTreeClassifier, plot_tree

X = carTable[['Weight(lb)', 'Displacement', 'Horsepower']]
y = carTable['GP1000M_City']
regtree = tree.DecisionTreeRegressor(max_leaf_nodes = 7)
regtree = regtree.fit(X, y,)
```

VISUALIZE MODELS

```
os.chdir('C:\\Users\\richardw\\Dropbox (Penn)\\Teaching\\477s2020\\Notes')
fig = plt.figure(num=None, figsize=(12, 9), dpi=80, facecolor='w', edgecolor='k')
plot_tree(regtree, filled=True, feature_names=X.columns)
plt.savefig('images/tree_01.png',  bbox_inches='tight')
plt.close(fig)
```

VISUALIZE MODELS



Decision tree

CLASS SUMMARY

SUMMARY

- Objectives.
- Syllabus review.
- Deliverables.
- Resources.
- Reasons why Python is so popular.
- Got a sense of the types of data activities we will learn to do.

NEXT TIME

NEXT TIME

- Opening the Spyder IDE
  - Introduction to interactive Python
  - Getting help
- Types of data
- Storing data in variables
- Data structures