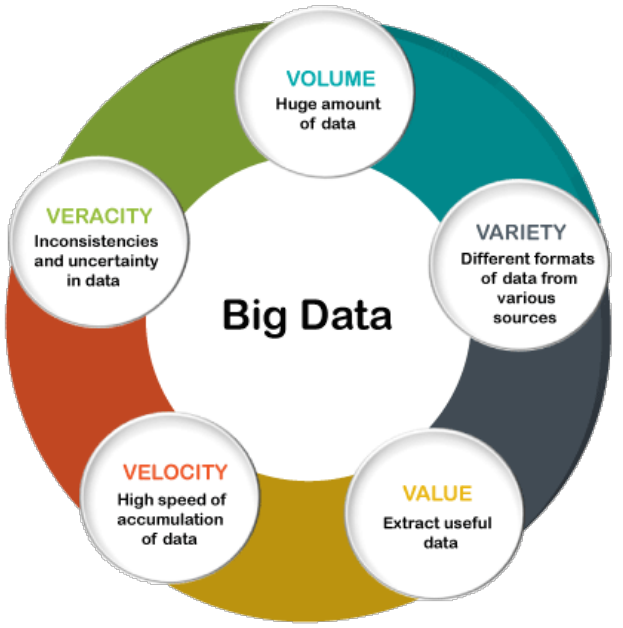
Common Uses for Data

Justin Post

Big Picture

- 5 V's of Big Data
  - Volume
  - Variety
  - Velocity
  - Veracity (Variability)
  - Value
- Will look at the Big Data pipeline later
  - Databases/Data Lakes/Data Warehouses/etc.
  - SQL basics
  - Hadoop & Spark

- **What to do with the data?**



Statistical Learning

Statistical learning - Inference, prediction/classification, and pattern finding

- Supervised learning - a variable (or variables) represents an **output** or **response** of interest
    - May model response and
        - Make **inference** on the model parameters
        - **predict** a value or **classify** an observation
- Unsupervised learning - **No output or response variable** to shoot for
    - Goal - learn about patterns and relationships in the data

Standard Rectangular Data

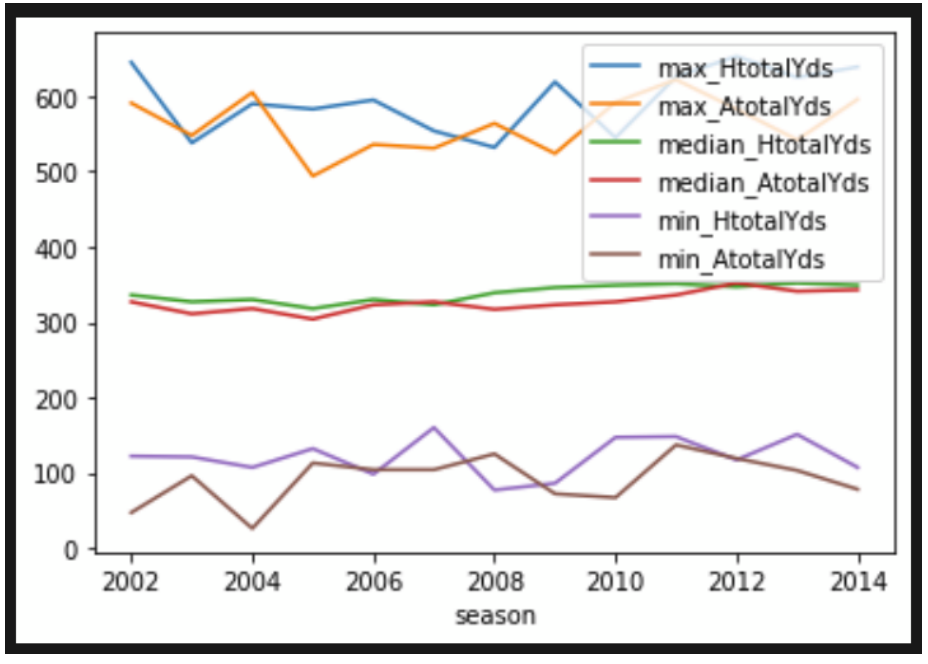
	A	B	C	D	E	F	G	H	I	J	K	L	M
1	X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
2	7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
3	7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
4	7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
5	8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
6	8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
7	8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
9	8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0
10	8	6	sep	tue	91	129.5	692.6	7	13.1	63	5.4	0	0
11	7	5	sep	sat	92.5	88	698.6	7.1	22.8	40	4	0	0
12	7	5	sep	sat	92.5	88	698.6	7.1	17.8	51	7.2	0	0
13	7	5	sep	sat	92.8	73.2	713	22.6	19.3	38	4	0	0

Data Driven Goals

Four major goals when using data:

1. Description

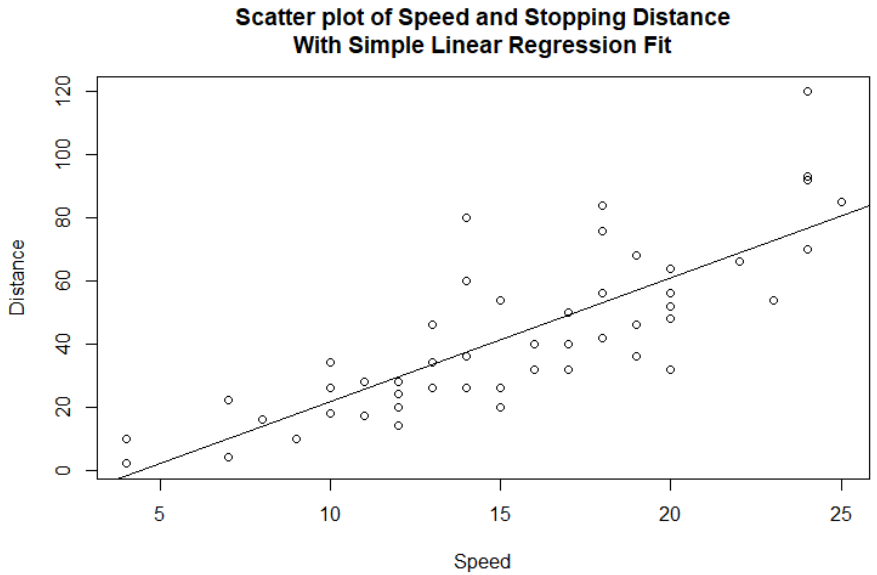
	mean		median	
	AtotalYds	HtotalYds	AtotalYds	HtotalYds
season				
2002	324.161049	333.782772	327	336
2003	308.247191	331.340824	311	327
2004	321.161049	335.659176	318	330
2005	308.378277	322.722846	304	318
2006	316.449438	328.745318	323	330
2007	321.906367	328.112360	327	323
2008	318.760300	334.932584	317	339
2009	323.977528	347.640449	323	346
2010	329.734082	341.441948	327	349
2011	341.655431	354.734082	336	351
2012	347.089888	351.573034	352	347
2013	340.685393	357.524345	341	352
2014	343.310861	352.838951	343	349

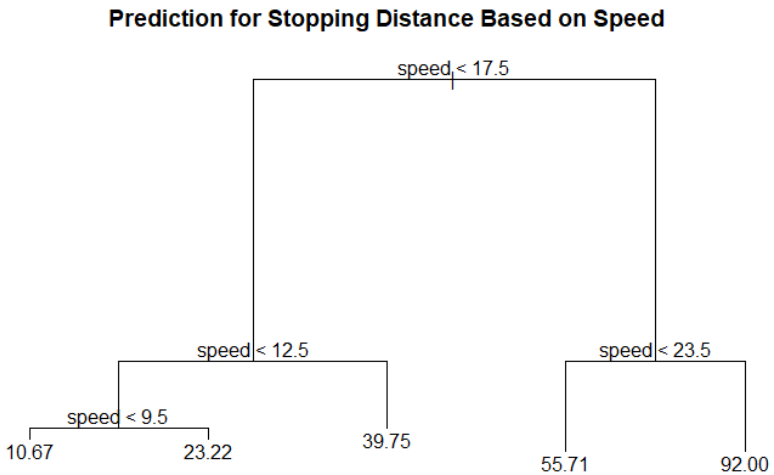


Data Driven Goals

Four major goals when using data:

2. Prediction/Classification





Data Driven Goals

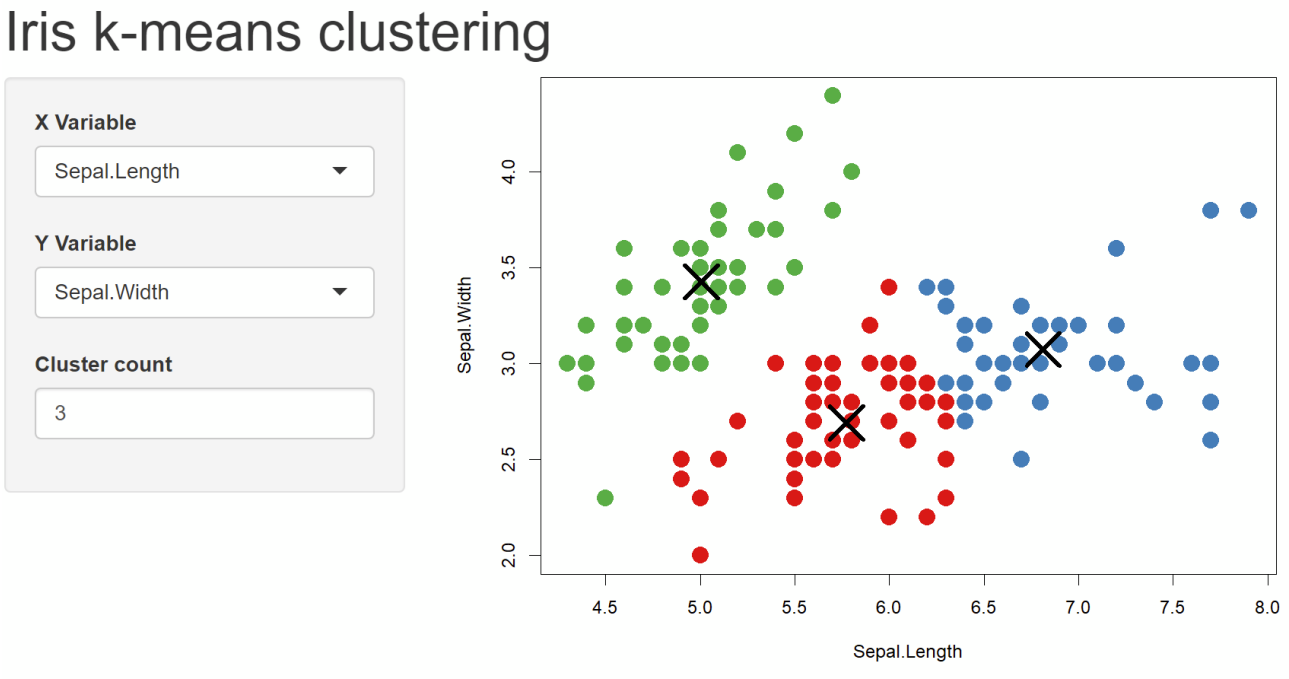
Four major goals when using data:

3. Inference
    - Confidence Intervals
    - Hypothesis Testing

Data Driven Goals

Four major goals when using data:

4. Pattern Finding



1. Describing Data

Goal: Describe the **distribution** of the variable

- Distribution = pattern and frequency with which you observe a variable
- Numeric variable - entries are a numerical value where math can be performed

For a single numeric variable,

- Shape: Histogram, Density plot, ...
- Measures of center: Mean, Median, ...
- Measures of spread: Variance, Standard Deviation, Quartiles, IQR, ...

For two numeric variables,

- Shape: Scatter plot
- Measures of Dependence: Correlation

Quick Example

Read in some data

```
import pandas as pd
wine_data = pd.read_csv("https://www4.stat.ncsu.edu/~online/datasets/winequality-full.csv")
wine_data.head()
```

```
##    fixed acidity  volatile acidity  citric acid  ...  alcohol  quality  type
## 0            7.4              0.70         0.00  ...      9.4        5   Red
## 1            7.8              0.88         0.00  ...      9.8        5   Red
## 2            7.8              0.76         0.04  ...      9.8        5   Red
## 3           11.2              0.28         0.56  ...      9.8        6   Red
## 4            7.4              0.70         0.00  ...      9.4        5   Red
##
## [5 rows x 13 columns]
```

Lots of Summaries!

- Use the `describe()` method on a `pandas` data frame

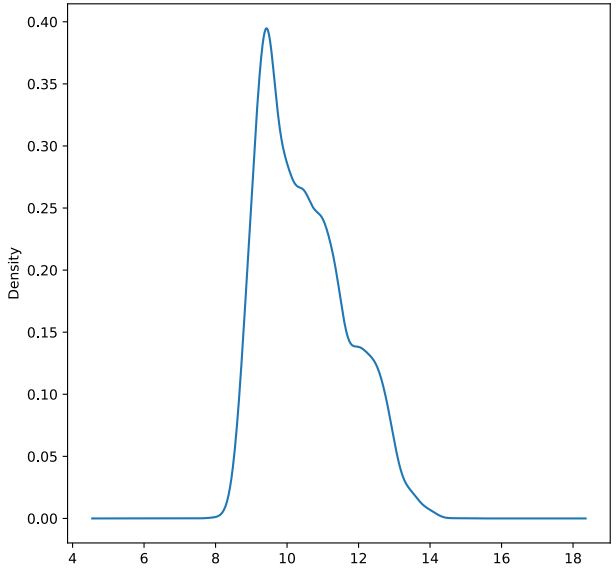
```
wine_data.describe()
```

```
##        fixed acidity  volatile acidity  ...      alcohol      quality
## count    6497.000000       6497.000000  ...  6497.000000  6497.000000
## mean        7.215307          0.339666  ...    10.491801     5.818378
## std         1.296434          0.164636  ...     1.192712     0.873255
## min         3.800000          0.080000  ...     8.000000     3.000000
## 25%         6.400000          0.230000  ...     9.500000     5.000000
## 50%         7.000000          0.290000  ...    10.300000     6.000000
## 75%         7.700000          0.400000  ...    11.300000     6.000000
## max        15.900000          1.580000  ...    14.900000     9.000000
##
## [8 rows x 12 columns]
```

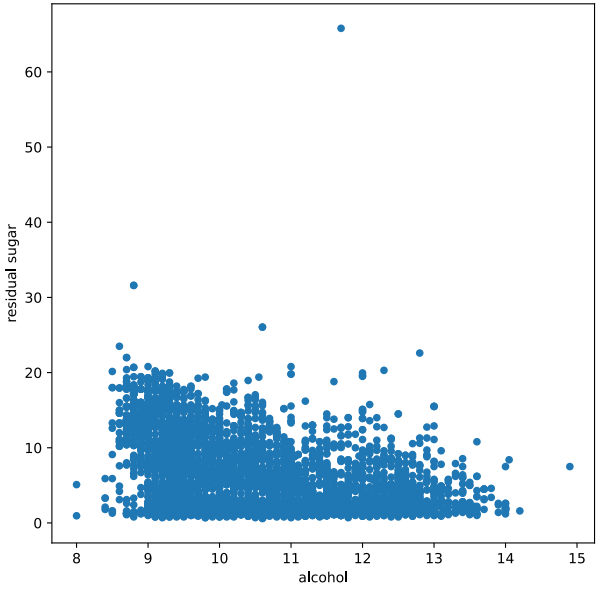
Graphs

- Many standard graphs to summarize with as well

```
wine_data.alcohol.plot.density()
```




```
wine_data.plot.scatter(x = "alcohol", y = "residual sugar")
```



2. Prediction/Classification

- A mathematical representation of some phenomenon on which you've observed data
- Form of the model can vary greatly!

Simple Linear Regression Model

$$\text{response} = \text{intercept} + \text{slope*predictor} + \text{Error}$$

$$Y_i = \beta_0 + \beta_1 x_i + E_i$$

- Assumptions often made about the data generating process to make inference (not required)

Simple Linear Regression Model

- We'll learn how to 'fit' this model later

```
from sklearn import linear_model
reg = linear_model.LinearRegression() #Create a reg object
reg.fit(X = wine_data['alcohol'].values.reshape(-1,1), y = wine_data['residual sugar'].values)
```

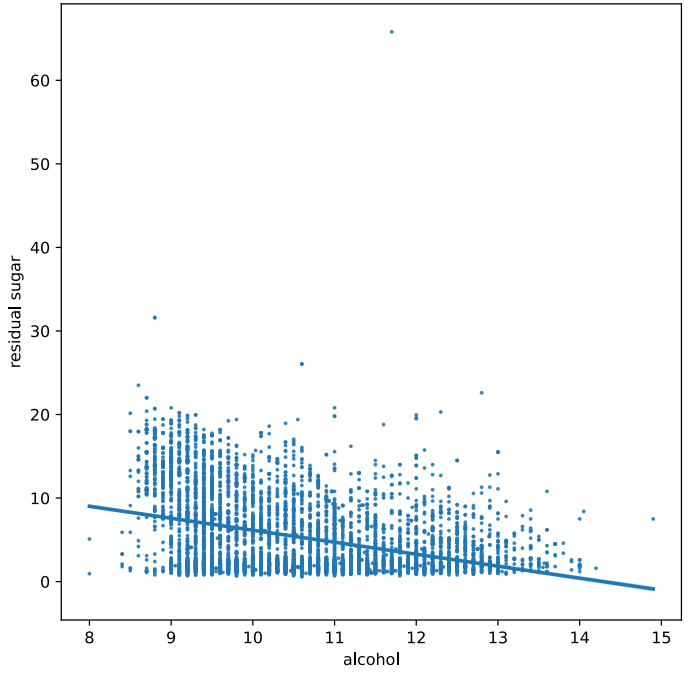
```
<style>#sk-container-id-1 {color: black;background-color: white;}#sk-container-id-1 pre{padding: 0;}#sk-containe
```

```
print(round(reg.intercept_, 3), round(reg.coef_[0], 3))
```

```
## 20.486 -1.434
```

Simple Linear Regression Model

```
import seaborn as sns
sns.regplot(x = wine_data["alcohol"], y = wine_data["residual sugar"], scatter_kws={'s':2})
```



2. Prediction/Classification

- A mathematical representation of some phenomenon on which you've observed data
- Form of the model can vary greatly!

Regression Tree

- This model can be used for prediction

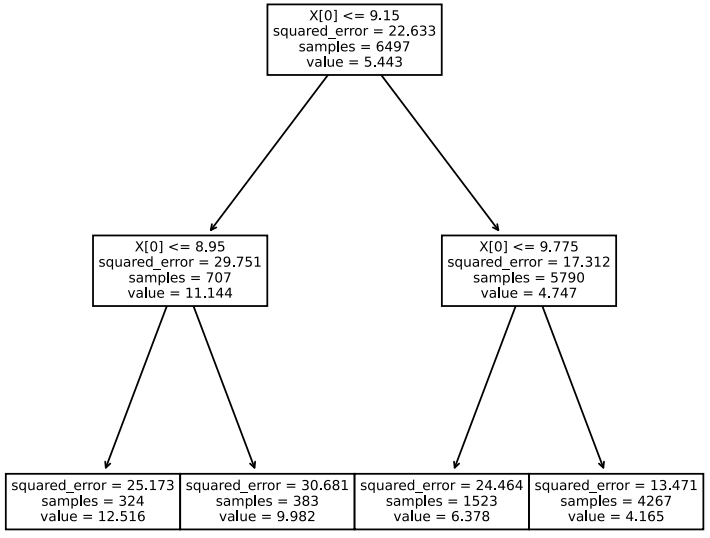
```
from sklearn.tree import DecisionTreeRegressor
reg_tree = DecisionTreeRegressor(max_depth=2)
reg_tree.fit(X = wine_data['alcohol'].values.reshape(-1,1), y = wine_data['residual sugar'].values)
```

```
<style>#sk-container-id-2 {color: black;background-color: white;}#sk-container-id-2 pre{padding: 0;}#sk-containe
```

Regression Tree

- This model can be used for prediction

```
from sklearn.tree import plot_tree
plot_tree(reg_tree)
```



2. Prediction/Classification

- A mathematical representation of some phenomenon on which you've observed data
- Form of the model can vary greatly!

Logistic Regression

- Consider binary response and classification as the task

$$P(\text{success}|\text{predictor}) = \frac{e^{\text{intercept}+\text{slope}*\text{predictor}}}{1+e^{\text{intercept}+\text{slope}*\text{predictor}}}$$

$$P(\text{success}|\text{predictor}) = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

We'll investigate a number of different models later in the course!

Recap

Four major goals with data:

1. Description
2. Prediction/Classification
3. Inference
4. Pattern Finding

- Descriptive Statistics try to summarize the distribution of the variable
- Supervised Learning methods try to relate predictors to a response variable through a model
    - Some models used for inference and prediction/classification
    - Some used just for prediction/classification