



Data 180 Recap

Cole Jennings, Mai Nguyen, and Whitney Finney

Tuesday, February 6th, 2024



AGENDA

1. Major concepts in DATA 180
2. R Application in student's project
3. References to *Ace the Data Science Interview*
4. Question and Discussion

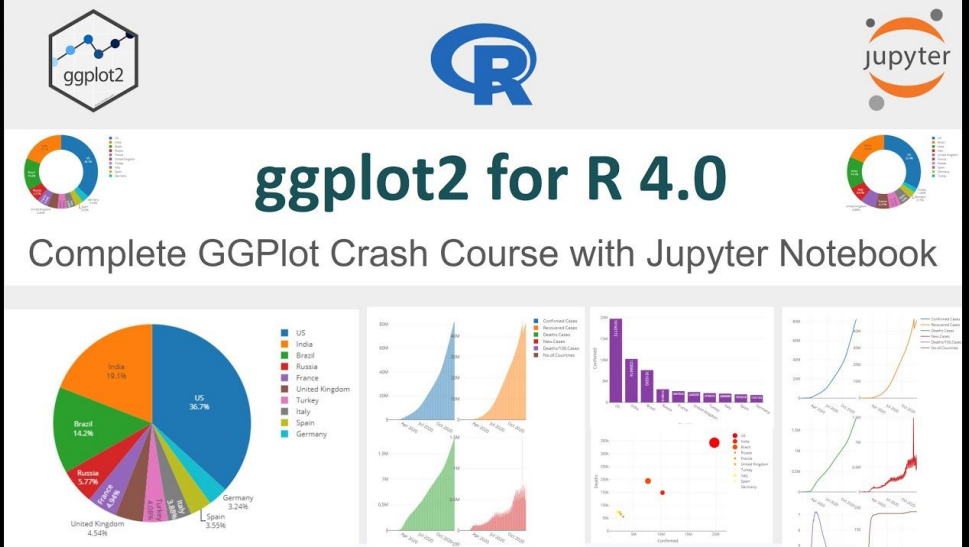


Major Concepts in DATA 180

- Visualization
- Unsupervised vs. Supervised Learning
- Clustering
- Regression
- Classification

Data Visualization

- ggplot2
- Used for most visualizations throughout DATA 180
- Allowed for scatter plots, bar graphs, histograms, etc
- Can use package for visualizations for DATA 400 project
 - Can use a Python package such as Matplotlib to do similar things



The banner for the 'ggplot2 for R 4.0' course features the R logo, the ggplot2 logo, and the Jupyter logo. It includes a small donut chart in the top left and right corners. The main title 'ggplot2 for R 4.0' is in a large, bold, dark green font. Below the title is the subtitle 'Complete GGPlot Crash Course with Jupyter Notebook'. The bottom section displays a collage of various plots: a large pie chart showing the percentage of cases by country (US: 38.7%, India: 13.1%, Brazil: 14.2%, Russia: 5.77%, France: 4.54%, United Kingdom: 4.54%, Germany: 3.24%, Spain: 3.55%), a bar chart, a scatter plot, and several line plots.

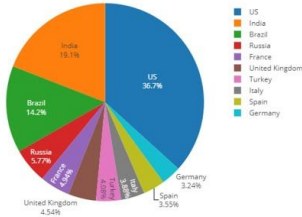
ggplot2

R

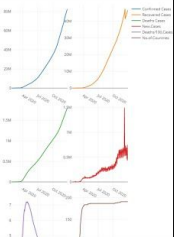
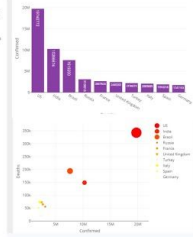
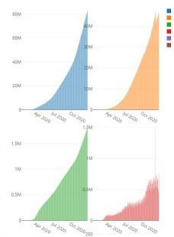
jupyter

ggplot2 for R 4.0

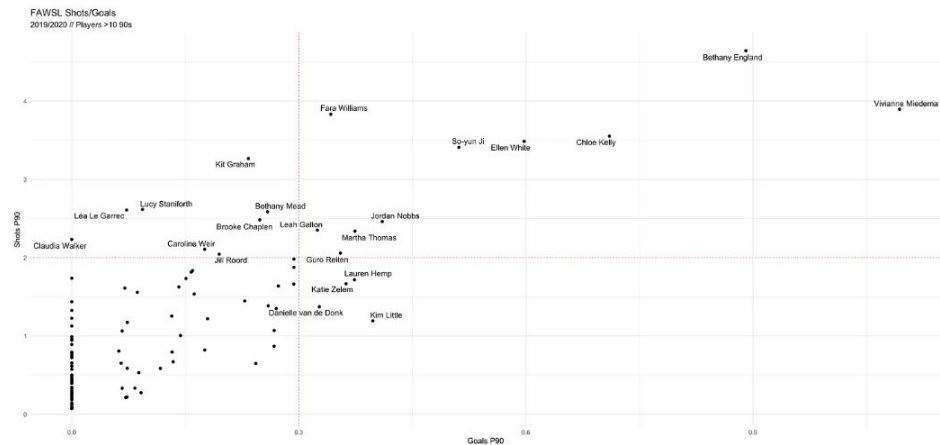
Complete GGPlot Crash Course with Jupyter Notebook



Country	Percentage
US	38.7%
India	13.1%
Brazil	14.2%
Russia	5.77%
France	4.54%
United Kingdom	4.54%
Germany	3.24%
Spain	3.55%



Data Visualization



- Example of use in “Shot Maps in R with Statsbomb data”

```
##shots p90/goals p90 scatter
ggplot(data = summary,
       aes(x = goals_p90,
           y = shots_p90))+
  geom_point() +

  geom_hline(yintercept = 2,
             colour = "red",
             alpha = 0.7,
             linetype = "dotted")+

  geom_vline(xintercept = 0.3,
             colour = "red",
             alpha = 0.7,
             linetype = "dotted")+

  geom_text_repel(data = summary %>%
                 filter(shots_p90 >= 2 | goals_p90>= 0.3),
                 aes(label = player.name))+

  theme_minimal() +

  labs(title = "FAWSL Shots/Goals",
       subtitle = "2019/2020 // Players >10 90s",
       x = "Goals P90",
       y = "Shots P90")
```

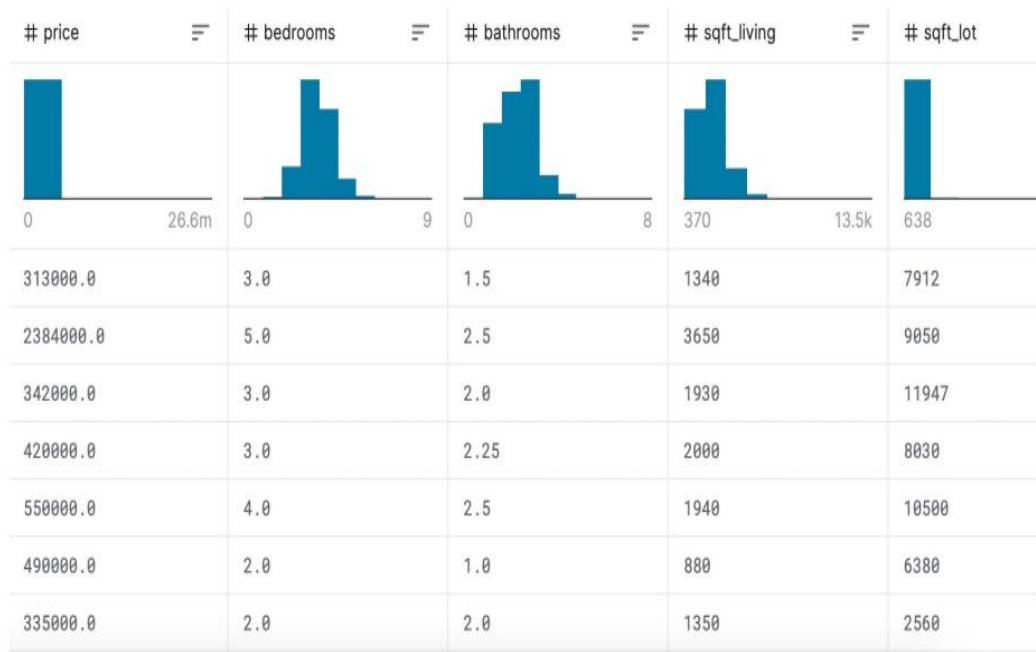
Supervised vs. Unsupervised Learning



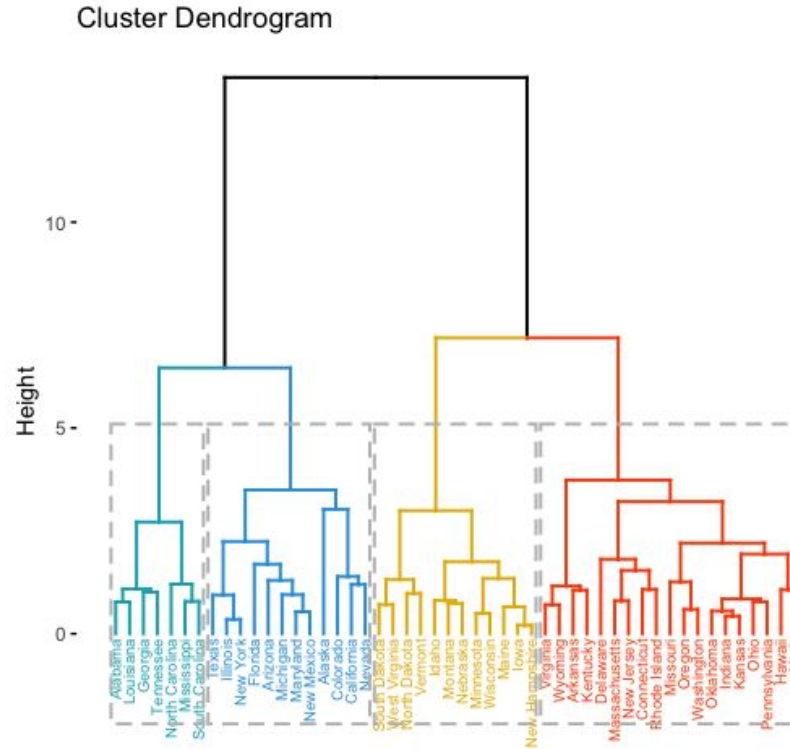
- **Supervised Learning:** modelize relationships between the predictor and and response variables given a dataset with both types of variables. The model relationships can be used for prediction or inference. Common and important methods:
 - Regressions
 - Support Vector Machine
 - Decision Trees
 - K-nearest neighborhood (KNN)
- **Unsupervised Learning (aka Exploratory Data Analysis):** explore relationships between the variables and look for structure in the dataset given a dataset without an output variable -> more challenging. Common and important method: Cluster Analysis

Supervised vs. Unsupervised Learning

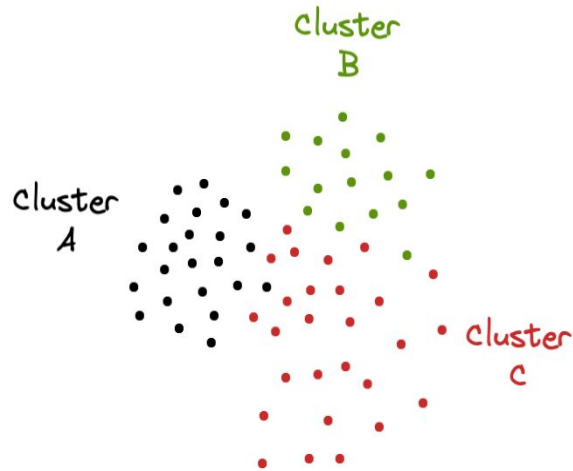
- **Supervised learning:** what are the variables that might impact housing price? How are they impacting housing prices?
- **Unsupervised learning:** find subgroups in the dataset:
 - houses that look similar in terms of ALL variables are in the same group, and
 - houses that look different are in different groups.



Clustering

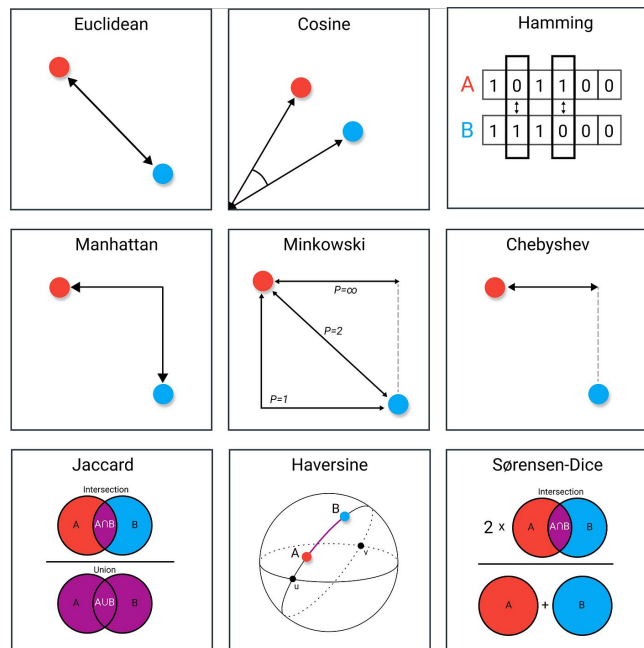


Clustering



- Cluster is a number of similar things that occur together:
 - Individual cluster: should display some level of homogeneity (specific attribute)
 - Different clusters should display some level of separation
- Steps to perform cluster analysis:
 - Selection of cluster variate, which is the set of characteristics we will use to cluster the objects under study (require domain-specific knowledge)
 - Selection of a measure of proximity: a method for combining the measurements of the traits into a "distance" (single number) that hopefully measures how similar (or dissimilar) the two objects are:
 - Euclidean distance
 - Manhattan distance (aka city block distance)
 - Minkowski distance

Clustering



→ Clustering then becomes:

- Find a group of sample units so that these units are close to each other, i.e., small in terms of distance, and
- sample units from different groups are far from each other, i.e., large in terms of distance.

Or we can say:

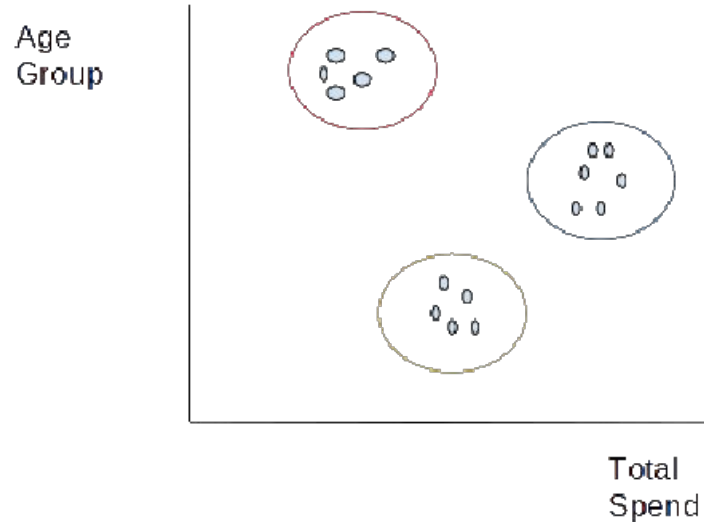
- Minimizing within-group distance, and
- Maximize between-group distance.

K-Means Clustering

- K-means is a centroid-based clustering algorithm, where we calculate the distance between each data point and a centroid (mean) to assign it to a cluster. The goal is to identify the K number of groups in the dataset.

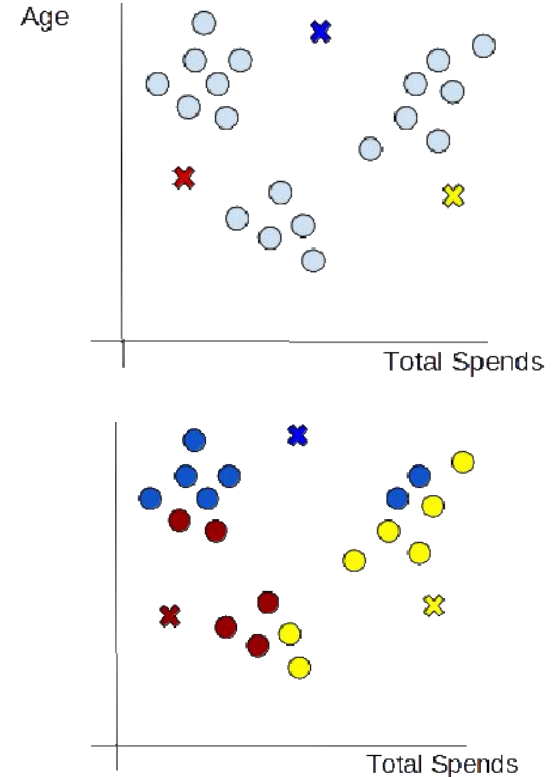
Example:

- You work at Walmart and want to improve marketing strategies.
- It's difficult to segment your customers manually.
- BUT - you have data that can group customers based on their spending -> through clustering!
- Once the customers are segmented, you can devise appropriate strategies for each group.



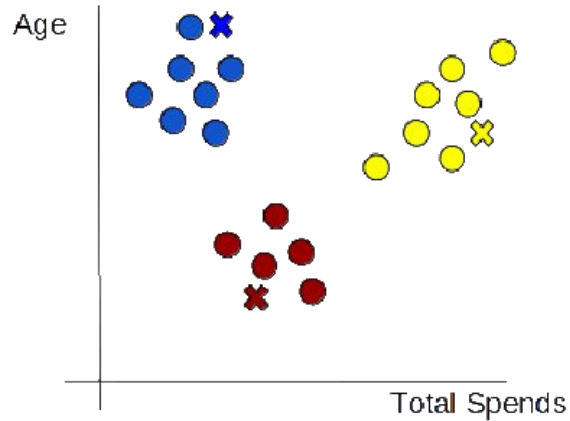
K-Means Clustering

1. **Choosing the number of clusters:** define the K number of clusters in which we will group the data. Let's select K=3.
2. **Initializing centroids:** Centroid is the center of a cluster but initially, the exact center of data points will be unknown so, we select random data points for each group and define them as centroids for each cluster. We will initialize 3 centroids in the dataset.
3. **Assign data points to the nearest cluster:** first calculate the distance between data point X and centroid C using Euclidean Distance metric, then choose the cluster for data points where the distance between the data point and the centroid is minimum.
4. **Re-initialize centroids:** calculate the average of all data points of that cluster.



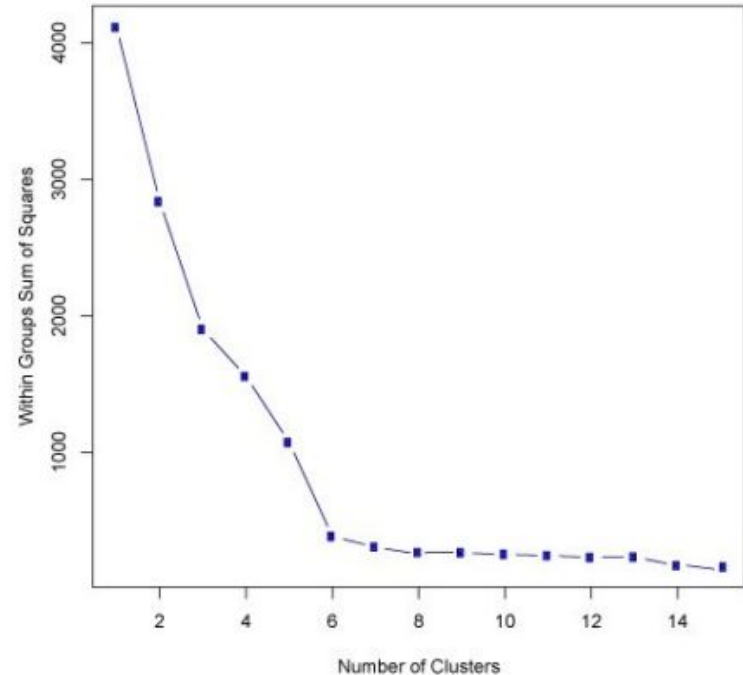
K-Means Clustering

5. **Repeat steps 3 and 4** until have optimal centroids and the assignments of data points to correct clusters are not changing anymore.



K-Means Clustering

- **Disadvantage:** You have to specify the number of clusters.
→ Use “Elbow Method” to optimize k value.
- **Elbow Method** is a plot of number of clusters against the Within-Groups Sum of Squares (WGSS), the sum of the distances for each of the points within a cluster.
- The goal is to have each cluster contain points that are the closest to it and to minimize the distance. Generally, as the number of clusters increases, the WGSS decreases.
- This right graph shows when $k = 6$, the decrease in WGSS is minimal.



Plant Derived Epigenetic Weapons

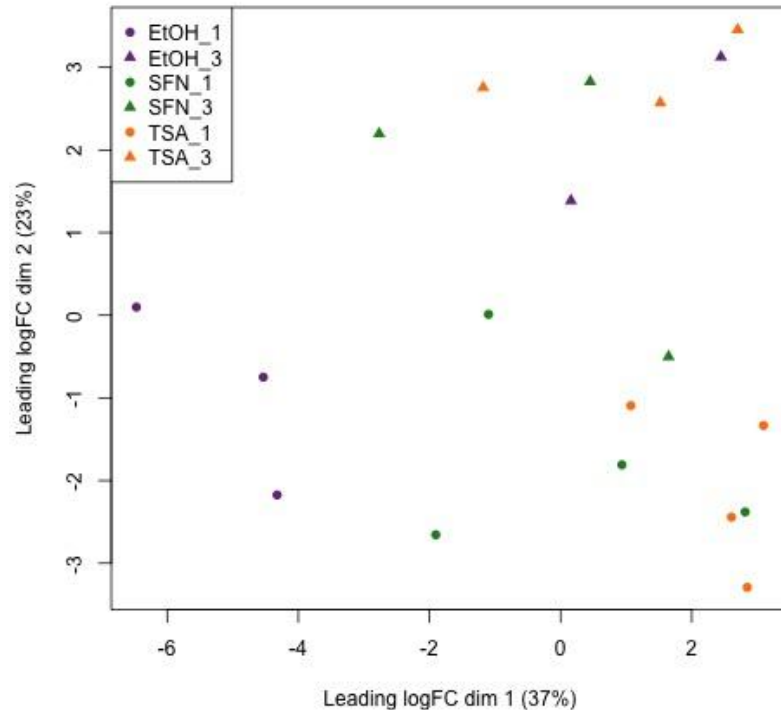
featureCounts in R

Genomic features

Output: A table of the
significant genes

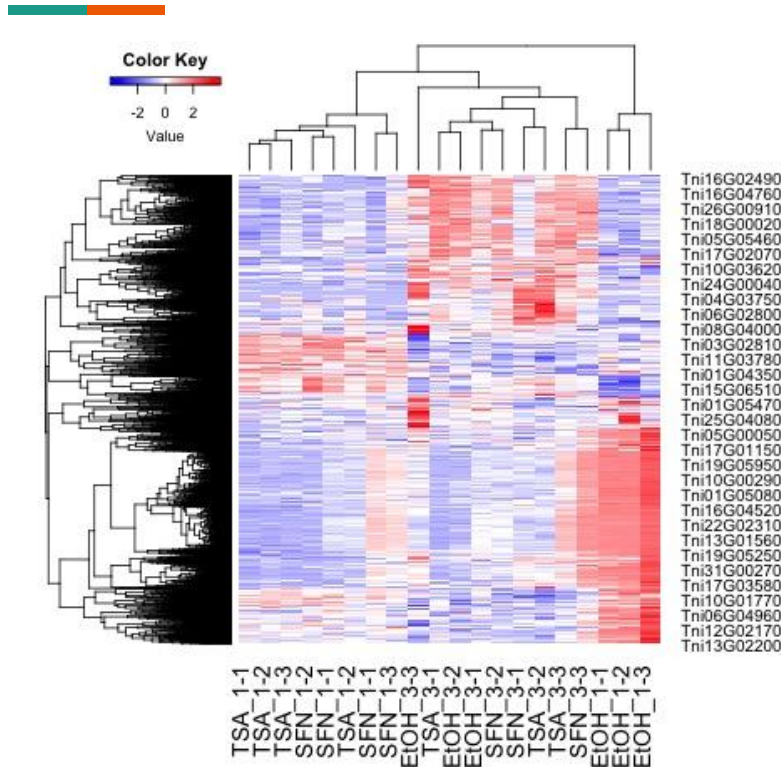
	A	B	C	D	E	F	G	H	I	J	K
1		22044R-0	22044R-01-0	22044R-01-0	22044R-01-0	22044R-01-0	22044R-01-0	22044R-01-0	22044R-01-0	22044R-02-0	22044R-02-0
2	Tni00G00010	477	633	478	1137	3042	760	1094	1090	2316	2223
3	Tni00G00020	900	896	492	804	883	709	963	1173	846	1151
4	Tni00G00030	2183	468	2323	1351	3527	1432	2722	3461	1824	3859
5	Tni00G00040	12	12	45	37	11	41	16	49	20	31
6	Tni00G00050	172	28	129	102	156	108	106	99	341	94
7	Tni00G00060	152	566	46	137	634	132	125	145	795	61
8	Tni00G00070	0	0	0	0	0	0	0	0	0	0
9	Tni00G00080	646	326	448	546	244	361	482	543	1285	590
10	Tni00G00090	0	0	0	0	0	0	0	0	2	1
11	Tni00G00100	17	8	59	2	653	56	147	22	10	147
12	Tni00G00110	772	1237	49	129	1	0	1	1	4314	11
13	Tni00G00120	462	657	21	44	0	0	0	0	2123	7
14	Tni00G00130	0	0	0	0	0	0	0	0	0	0
15	Tni00G00140	13	17	3	2	0	5	3	6	47	4
16	Tni00G00150	60	175	45	54	61	25	39	55	206	41
17	Tni00G00160	0	0	1	2	0	0	0	0	1	2
18	Tni00G00170	4	1	0	0	0	0	0	0	8	0
19	Tni00G00180	0	1	0	0	0	5	1	5	1	0
20	Tni00G00190	2	2	0	0	1	7	7	13	1	1
21	Tni00G00200	152	544	125	169	191	105	96	137	337	113
22	Tni00G00210	0	0	0	4	0	0	0	0	0	0
23	Tni00G00220	2	12	0	0	0	0	0	0	0	0

Plant Derived Epigenetic Weapons



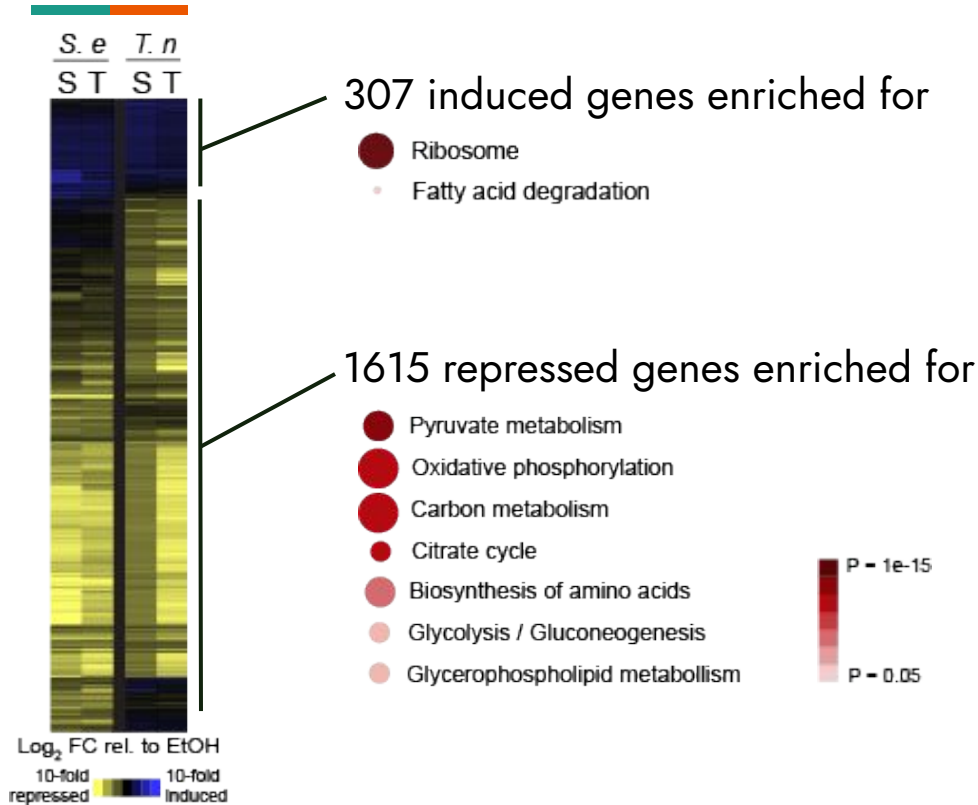
- Differential expression
- Examining the difference between samples
- ***Growth data is notoriously non-normal***
- Can you see any obvious clusters?

Clustering (heat maps)



- Gene expression for a large number of genes.
- Samples, hierarchical clustering, differentially expressed genes.
- Blue is repressed, red is induced.

Combining information:



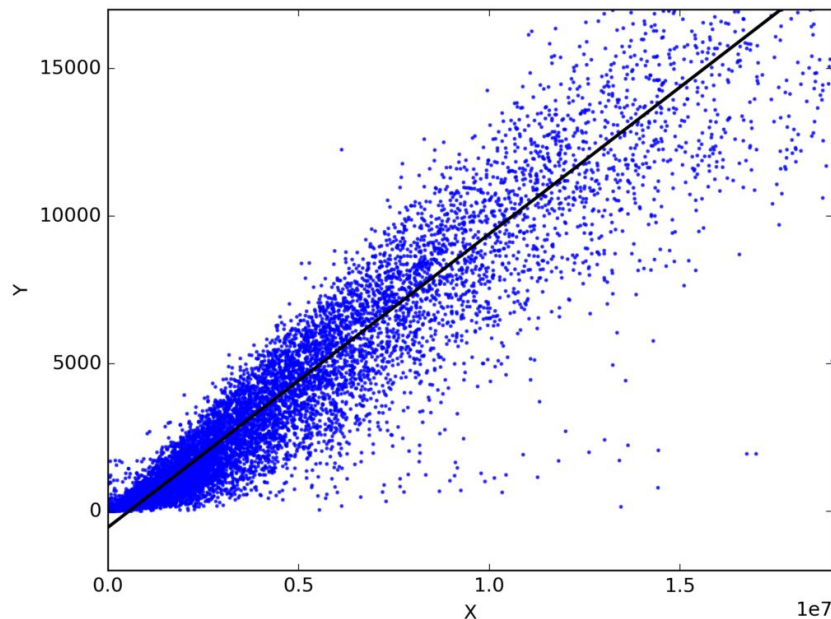
- Genes associated with ribosome production and fatty acid degradation are upregulated.
- Genes associated with energy production are downregulated.
- SFN and TSA generate similar patterns in gene expression.

Regression

- Linear regression is to assume a linear relationship between the predictors and the response.

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_Dx_D$$

- The goal of regression model is to calculate the coefficients so that the model best fits the training set (R-squared), then the model can also be used to predict the response variable.
- Advanced regression models:
 - Non-linear regression model
 - Ridge regression
 - Lasso regression



Classification



- Classification refers to the type of supervised learning models with a binary response variable, for example:
 - Is this email a spam or not?
 - Is this patient diagnosed with cancer or not?
 - Is this picture a cat or not?
- Common methods:
 - Logistic regression
 - K-nearest neighborhood (KNN)
 - Support Vector Machines
 - Naïve Bayes
 - Decision Trees (and their extensions)
 - Neural Network

Classification



Intuition behind KNN:

- Suppose I interview everyone in your friend circle and figured out they are all senior, which year are you most likely to be in?

→ Senior

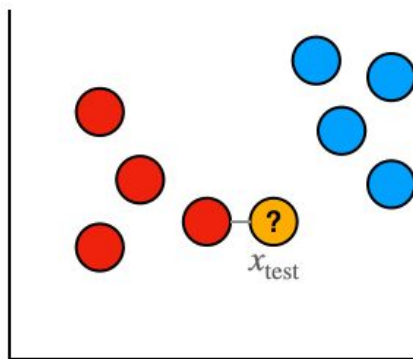
- Suppose I interview everyone in your friend circle and figured out 80% of them are senior and the rest of the 20% are junior, which year are you most likely to be in?

→ Senior

KNN Classification Technique:

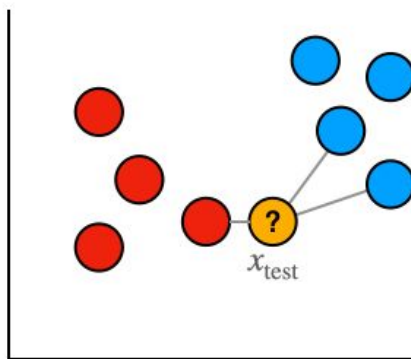
- For any new sample unit, the algorithm first finds its closest K sample units.
- It then finds the “class” (given by the response variable) of these K units.
- Their mode becomes the prediction for this new sample unit.

Classification



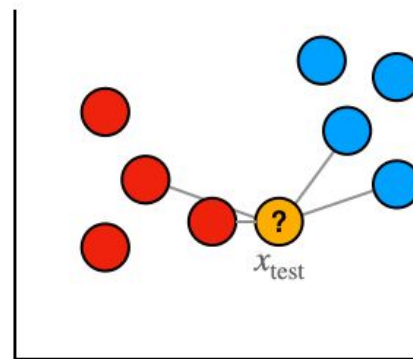
$k = 1$

Nearest point is **red**, so x_{test} classified as **red**



$k = 3$

Nearest points are {**red**, **blue**, **blue**} so x_{test} classified as **blue**



$k = 4$

Nearest points are {**red**, **red**, **blue**, **blue**} so classification of x_{test} is not properly defined

Classification

- Quality of Classification Model: confusion matrix -> a lot of ways to measure the model quality so need to critically think about the implication behind choosing the metric. Sometimes accuracy is not always the most suitable metric.

		Predicted condition	
		Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Positive (P)	True positive (TP), hit	False negative (FN), type II error, miss, underestimation
	Negative (N)	False positive (FP), type I error, false alarm, overestimation	True negative (TN), correct rejection

Classification

- Once we determined the quality measure to use, we could then try different K and pick the K that gives us the best result.
- If you chose accuracy/recall as the quality measure, then higher accuracy/recall means better model.

```
library(class)
movie <- read.csv("~/downloads/movie_regression.csv")

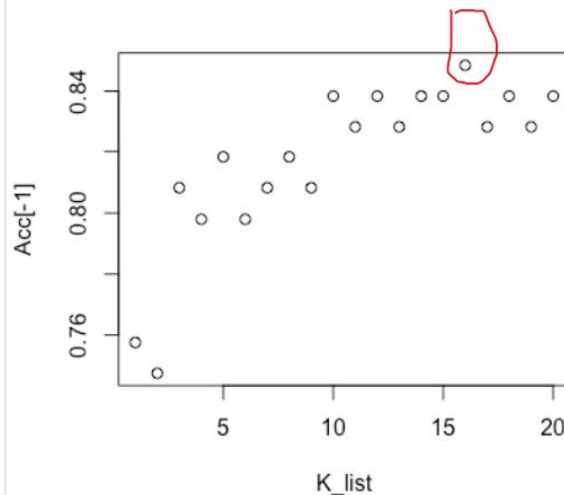
movie <- na.omit(movie)
movie <- dplyr::mutate(movie, class = ifelse(movie$Critic_rating > 7, 1, 0))
movie <- dplyr::select(movie, -Critic_rating, -X3D_available, -Genre)

training <- sample(1:nrow(movie), 0.8*nrow(movie))
trainingset <- movie[training, ]

validation <- setdiff(1:nrow(movie), training)
validationset <- movie[validation, ]

K_list <- 1:20
Acc <- 0

for (k in K_list){
  knn_pred <- knn(trainingset,
                  validationset,
                  trainingset$class,
                  k = k)
  conf_mat <- table(validationset$class, knn_pred)
  Acc <- c(Acc, sum(diag(conf_mat))/sum(conf_mat))
}
plot(K_list, Acc[-1], ylab = "Accuracy")
```



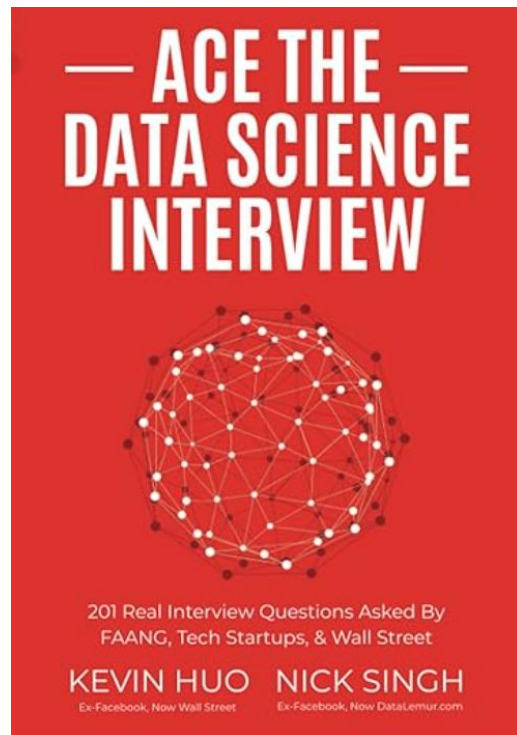
Suggestions for DATA 180

- Include a crash-course on Tableau and/or Power BI during data visualization component of course
 - Can be very helpful for students trying to land internships
 - Relevant to data visualization part of course
- Include a crash-course on Excel and how to use scripts to communicate with Excel
- Offering a Python section of DATA 180/option to submit work in Python
 - Gives students flexibility



ADSI Connections

- Statistics
 - Entire chapter of ADSI
 - One of the course materials for data 180: An Introduction to Statistical Learning: with applications in R
 - Introduce elements of statistics relevant to modeling
 - P-value, R-squared, bias, variance
- Coding
 - Entire chapter of ADSI
 - R is a commonly used scripting language
 - Class focuses on building proficiency with data analysis packages
 - dplyr





Questions and Discussion

- What was the hardest part of DATA 180?
- What have you used for internships/research/other classes?

Helpful Resources:

- <https://www.reddit.com/r/rprogramming/>
- <https://www.datacamp.com/>