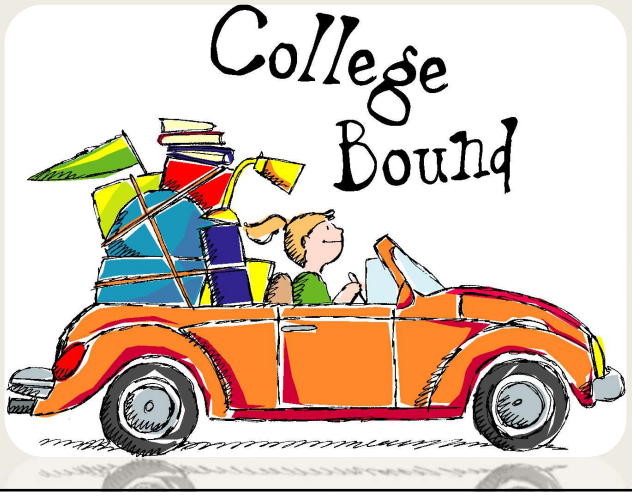
Cluster Analysis and Predictive Modeling on Transfer Students' Success rate.



TAIR 2021
Presenter: Daniel Le
Research Analyst
Contact info: dle@dcccd.edu

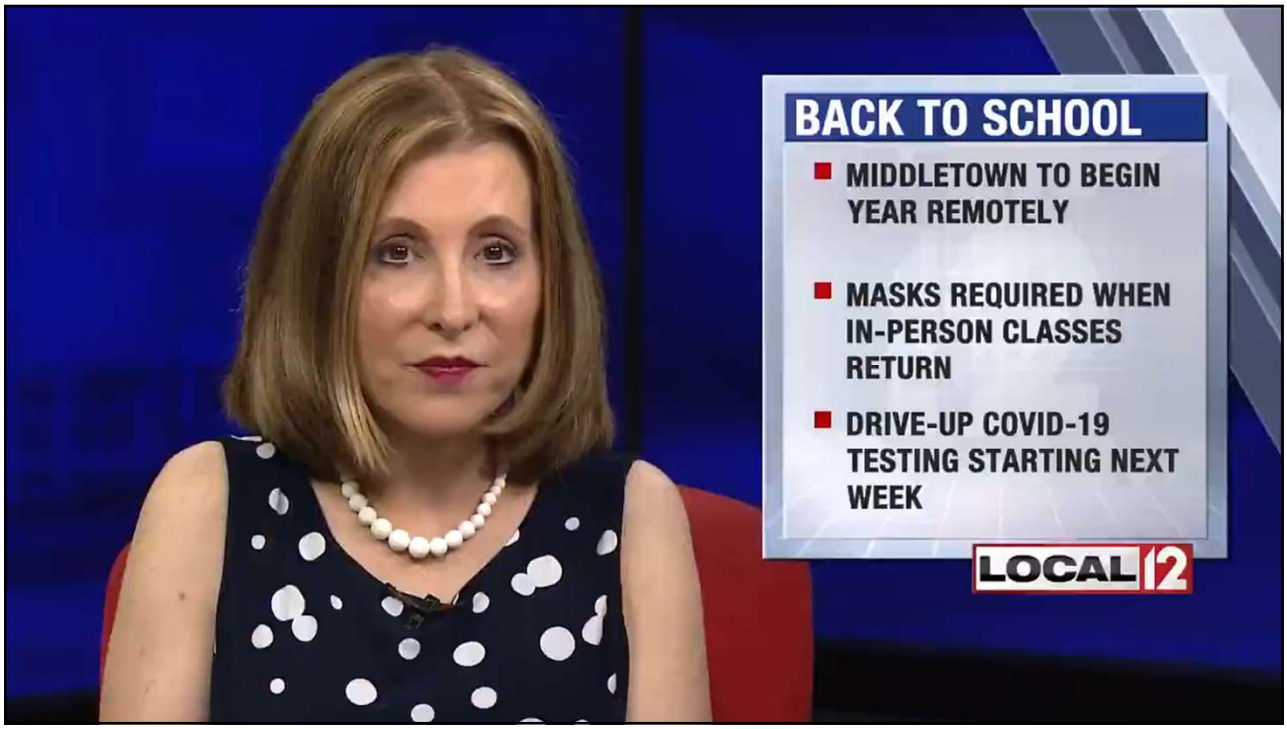
DALLAS COLLEGE

5

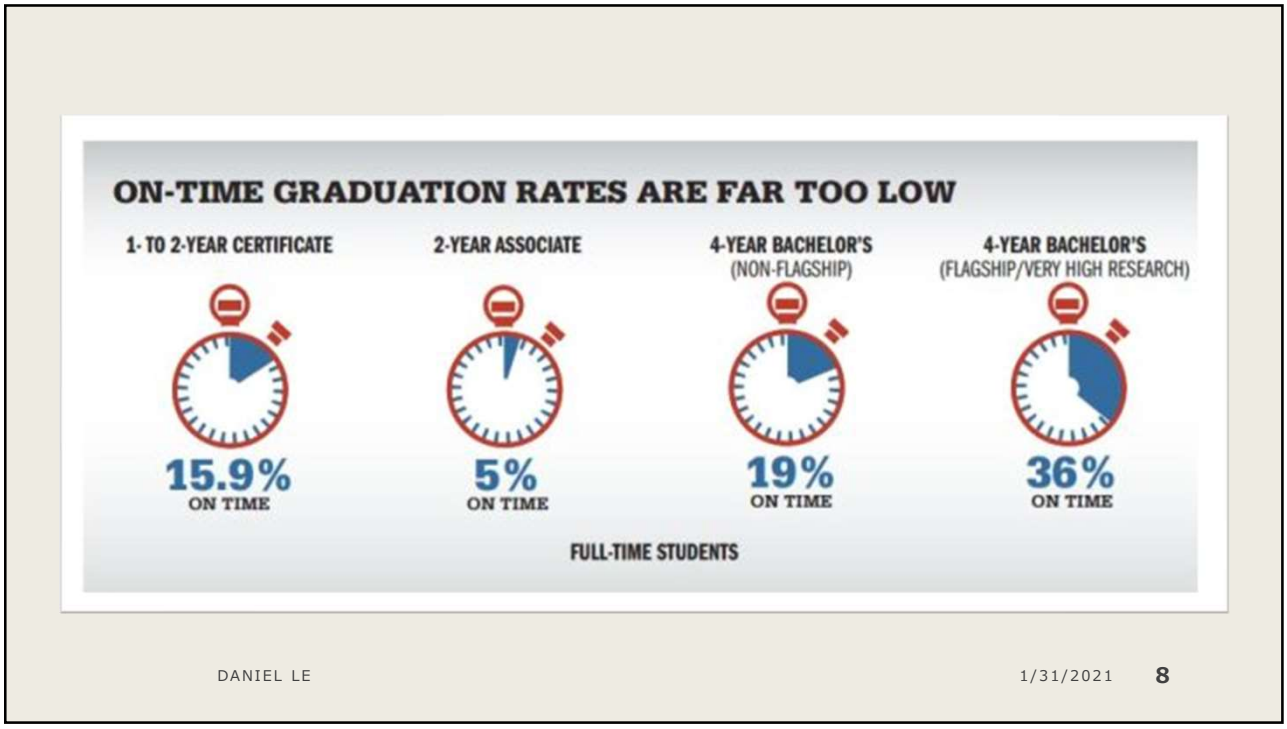
PART I:

INTRODUCTION

6

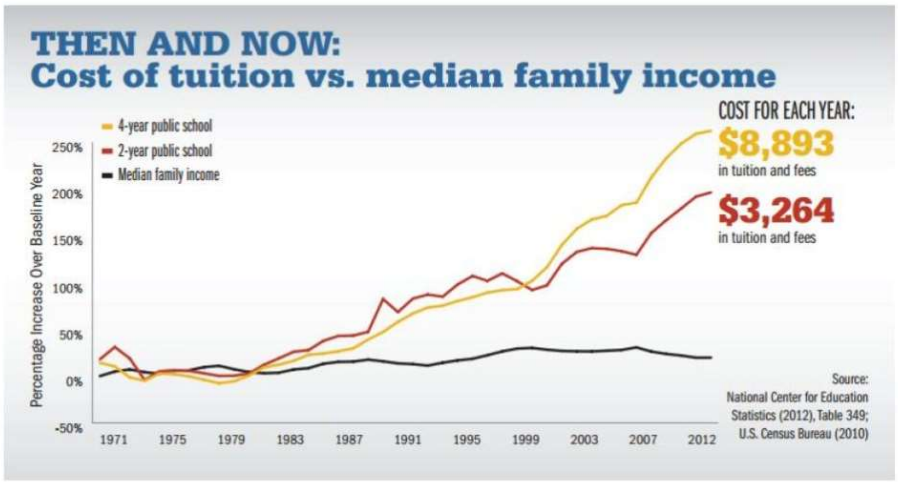

BACK TO SCHOOL
MIDDLETOWN TO BEGIN YEAR REMOTELY
MASKS REQUIRED WHEN IN-PERSON CLASSES RETURN
DRIVE-UP COVID-19 TESTING STARTING NEXT WEEK
LOCAL 12

7


ON-TIME GRADUATION RATES ARE FAR TOO LOW
1- TO 2-YEAR CERTIFICATE
15.9%
ON TIME
2-YEAR ASSOCIATE
5%
ON TIME
4-YEAR BACHELOR'S (NON-FLAGSHIP)
19%
ON TIME
4-YEAR BACHELOR'S (FLAGSHIP/VERY HIGH RESEARCH)
36%
ON TIME
FULL-TIME STUDENTS
DANIEL LE
1/31/2021
8

8

The cost of higher education has drastically outpaced increases in median family income. As a result, obtaining the education necessary for success has become far more difficult and costly, and students have been forced to pile on even more debt in the process.



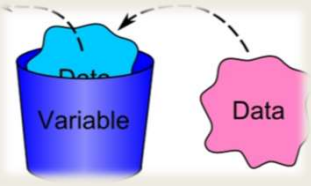
9

10

- Following IRB (Institutional review board) approval, a prospectively maintained database of sample of Fall 2020 transfer students to Dallas College is reviewed.
- **Inclusion criteria:** Fall 2020 transfer – in students who enrolled in Fall 2020 in Dallas College.
- **Exclusion criteria:** any students who do not meet the above inclusion criteria will be removed from the dataset.

Variables:

- Term GPA, Term credit hours, Cumulative GPA, Cumulative hours (Cluster analysis).
- Student info: Enrollment status, Classification, Age, Gender, Race/Ethnicity, Number of dependents, Employment status, Income range, Mother's education level, Father's education level (predictive model).
- Response (outcome) variable (predictive model): Success (1: cumulative GPA ≥ 2.0, otherwise it's 0).



PART II: STATISTICAL ANALYSIS

13

14

15

Section I: Clustering

- A task of dividing up data into groups (clusters), so that points in any one group are more similar to each other than to points outside the group.
- Main uses:
  - ✓ Summary: deriving a reduced representation of the full data set.
  - ✓ Discovery: looking for new insights into the structure of the data.
  - ✓ Investigating the validity of pre-existing group assignments.
  - ✓ Helping with prediction (classification or regression).

16

17

Clustering algorithms

- Suppose the number of clusters K < N is pre-specified. C(i) = k ∈ {1,2,…,K} is an encoder that assigns the *i*th observation to the k*th* cluster.
- We seeks the particular encoder that minimizes the within-point scatter (i.e. sum dissimilarities with clusters)

$$W(C) = \frac{1}{2}\sum_{k=1}^{K}\sum_{C(i)=k}\sum_{C(i')=k} d(x_i, x_{i'})$$

- This is computationally infeasible as there are many possible cluster assignments.
- Feasible strategies are based on iterative greedy descent (examining a small fraction of all possible assignments)
  - At each step, the cluster assignments are changed in such a way that the value of the criterion is improved from its previous value.
  - When the prescription is unable to provide an improvement, the algorithm terminates with the current assignments as its solution.

18

K-means Clustering

- Starting with guessing those cluster centers, and it repeats the following steps:
  - ✓ For each data point, the closest cluster center (in Euclidean distance) is identified.
  - ✓ Each cluster center is replaced by the average of all data points that are closest to it.
  - ✓ Stop when the cluster assignment does not change.

19

Algorithm 14.1 *K-means Clustering.*

1. For a given cluster assignment C, the total cluster variance (14.33) is minimized with respect to $\{m_1, \ldots, m_K\}$ yielding the means of the currently assigned clusters (14.32).

2. Given a current set of means $\{m_1, \ldots, m_K\}$, (14.33) is minimized by assigning each observation to the closest (current) cluster mean. That is,

$$C(i) = \underset{1 \le k \le K}{\operatorname{argmin}} ||x_i - m_k||^2. \qquad (14.34)$$

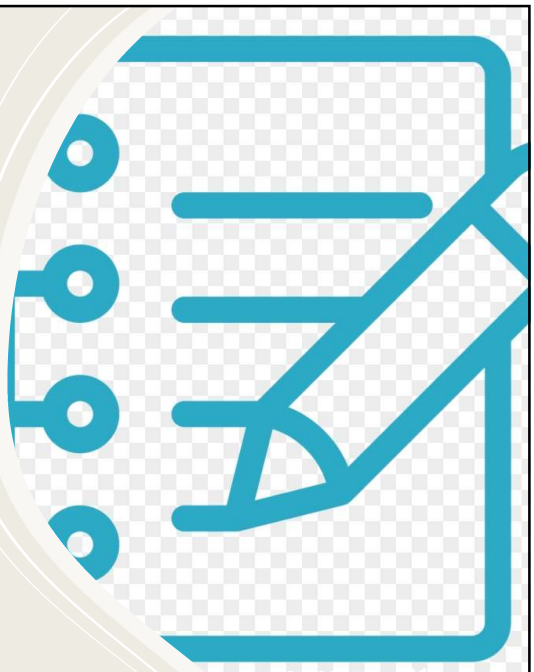
3. Steps 1 and 2 are iterated until the assignments do not change.

$$\bar{x}_S = \underset{m}{\operatorname{argmin}} \sum_{i \in S} ||x_i - m||^2. \qquad (14.32)$$

20

K-means Remarks

- Within-point scatter decreases with each iteration of the algorithm (the sum of squared distance of each observation from the cluster mean decreases).
- The final clustering depends on the initial cluster centers. We typically run K-means multiple times with random guesses, then choose among from collection of centers based on which one gives the smallest within-point scatter.
- The algorithm is not guaranteed to deliver the clustering that globally minimizes within-cluster variation.



```
Classes 'tbl_df', 'tbl' and 'data.frame':       2537 obs. of  4 variables:
 $ CUMM_CREDS: num  3 71 22 3 73 1 4 8 3 5 ...
 $ CUMM_GPA  : num  3 3.86 2.86 4 2.97 4 4 4 4 4 ...
 $ TERM_CREDS: num  3 3 3 3 3 1 4 8 3 5 ...
 $ TERM_GPA  : num  3 4 4 4 4 4 4 4 4 4 ...
```


23

The Elbow Method

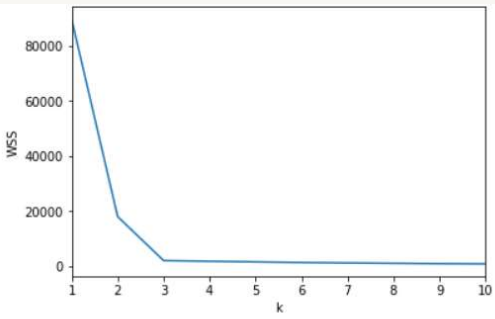
This is probably the most well-known method for determining the optimal number of clusters. *It is also a bit naive in its approach.*

*Calculate the **Within-Cluster-Sum of Squared** Errors (WSS) for **different values of k**, and choose the k for which WSS becomes first starts to diminish. In the plot of WSS-versus-k, this is visible as an **elbow.***

Within-Cluster-Sum of Squared Errors sounds a bit complex. Let's break it down:

- The Squared Error for each point is the square of the distance of the point from its representation i.e. its predicted cluster center.
- The WSS score is the sum of these Squared Errors for all the points.
- Any distance metric like the Euclidean Distance or the Manhattan Distance can be used.

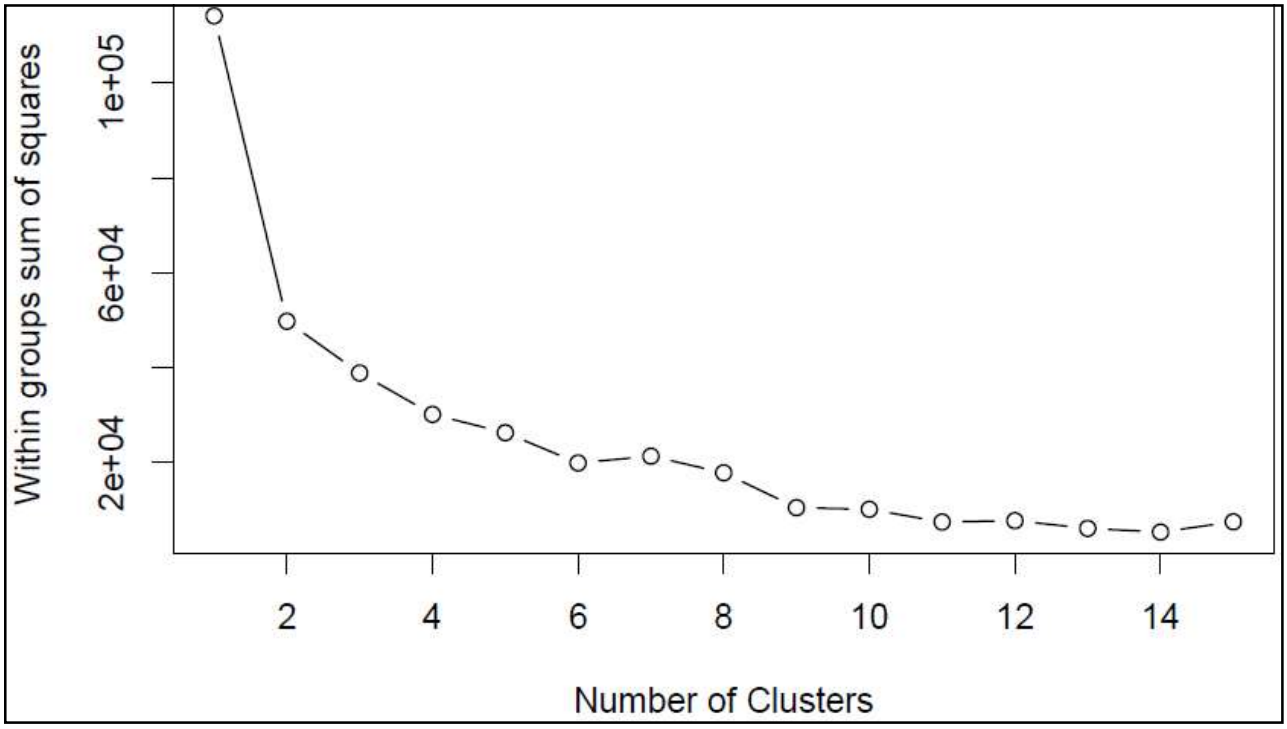
24



As expected, the **plot looks like an arm with a clear elbow at k = 3.**

Unfortunately, we do not always have such clearly clustered data. This means that the elbow may not be clear and sharp.

25



26

NbClust

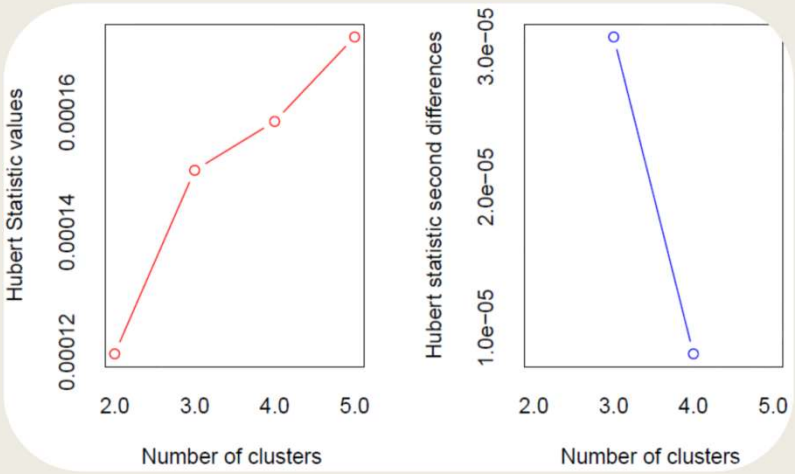
From NbClust v3.0 by Malika Charrad — 99.99th Percentile

NbClust Package For Determining The Best Number Of Clusters

`NbClust` package provides 30 indices for determining the number of clusters and proposes to user the best clustering scheme from the different results obtained by varying all combinations of number of clusters, distance measures, and clustering methods.

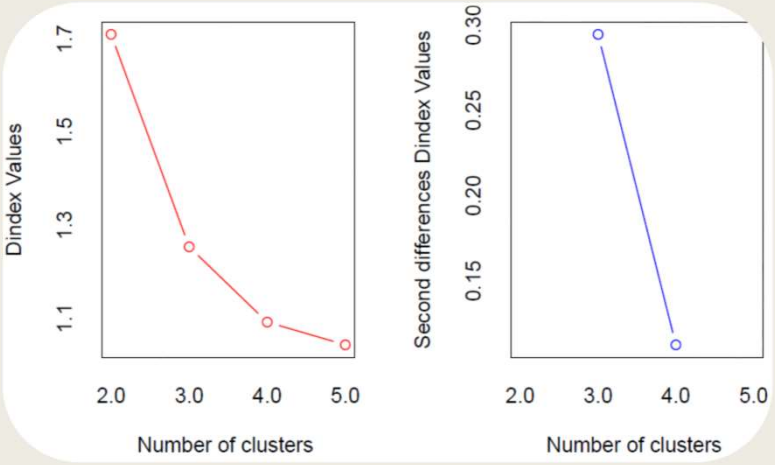
27

The Hubert index is a graphical method of determining the number of clusters. In the plot of Hubert index, we seek a significant knee that corresponds to a significant increase of the value of the measure i.e the significant peak in Hubert index second differences plot.

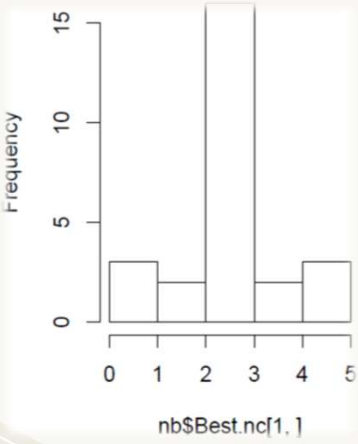


28

The D index is a graphical method of determining the number of clusters. In the plot of D index, we seek a significant knee (the significant peak in D index second differences plot) that corresponds to a significant increase of the value of the measure.



29




```
*****************************************************************
* Among all indices:
* 2 proposed 2 as the best number of clusters
* 16 proposed 3 as the best number of clusters
* 2 proposed 4 as the best number of clusters
* 3 proposed 5 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3


*****************************************************************
```


30

Distances between Final Cluster Centers

Cluster	1	2	3
1		65.583	10.207
2	65.583		59.020
3	10.207	59.020	

Final Cluster Centers

	Cluster 1	Cluster 2	Cluster 3
CUMM_GPA2_CREDS	4	70	11
CUMM_GPA2	2.410926766	3.336666667	2.492361251
TERM_GPA2_CREDS	4	3	11
TERM_GPA2	2.341348023	4.000000000	2.491987891

Number of Cases in each Cluster

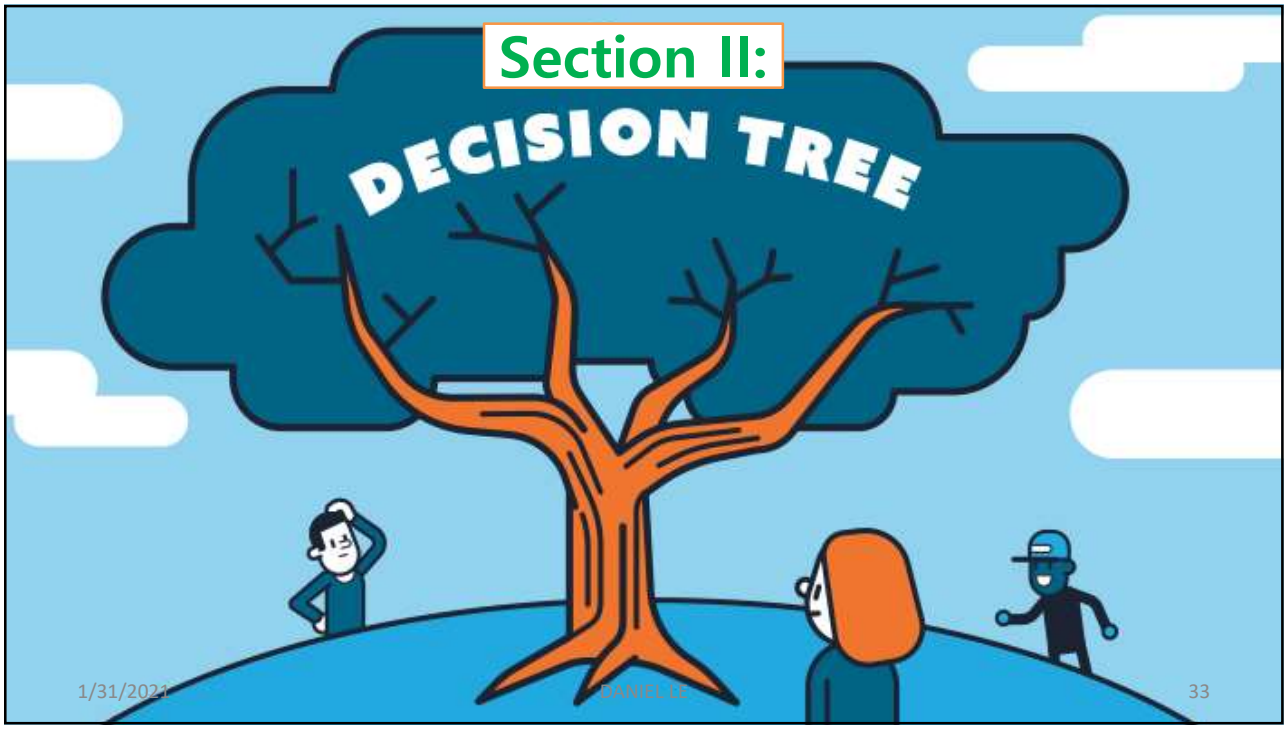
Cluster	1	1543.000
	2	3.000
	3	991.000
Valid		2537.000
Missing		.000

31

Cluster Membership

Case Number	Cluster	Distance
1	1	1.700
2	2	1.432
3	3	13.627
4	1	2.717
5	2	3.353
6	1	4.854
7	1	2.301
8	3	5.051
9	1	2.717
10	1	2.685
11	3	5.339
12	1	2.717
13	1	3.221
14	1	3.622
15	1	2.220

32

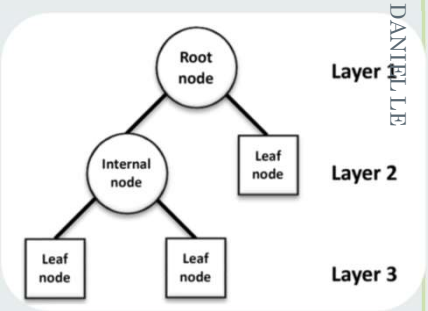


33

34

WHAT IS DECISION TREE?

- **Decision tree learning** is a graphical representation of all possible solutions to a decision based on certain conditions.
- It is used for either **classification (**categorical target variable**)** or **regression (**continuous target variable**)** ****CART****
- Trees are drawn upside down. The final regions are termed **leaves**. The points inside the tree where a split occurs is an **interval node**. Finally, segments that connect nodes are **branches**.

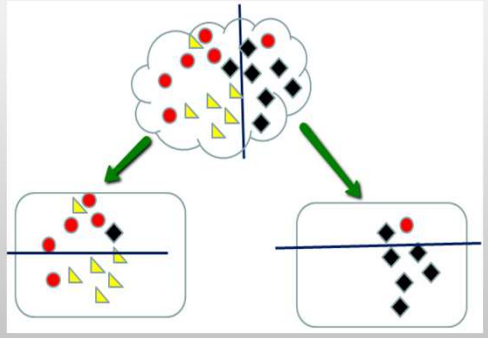


35

35

How Does A Decision Tree Work?

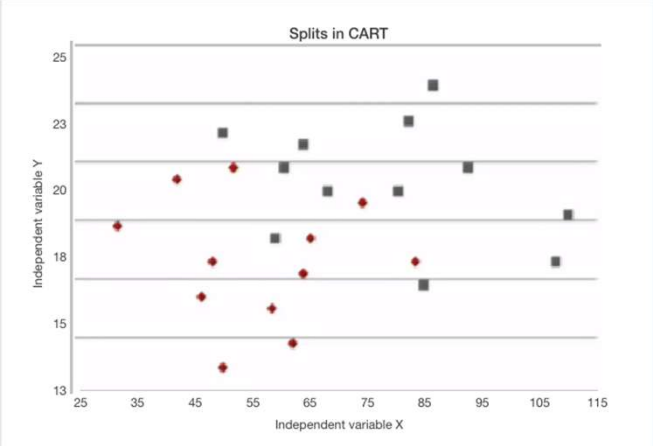
- Repeatedly partitioning the data into multiple sub-spaces so that the outcomes in each final sub-space is as homogeneous as possible.
- This is called **recursive partitioning.**



36

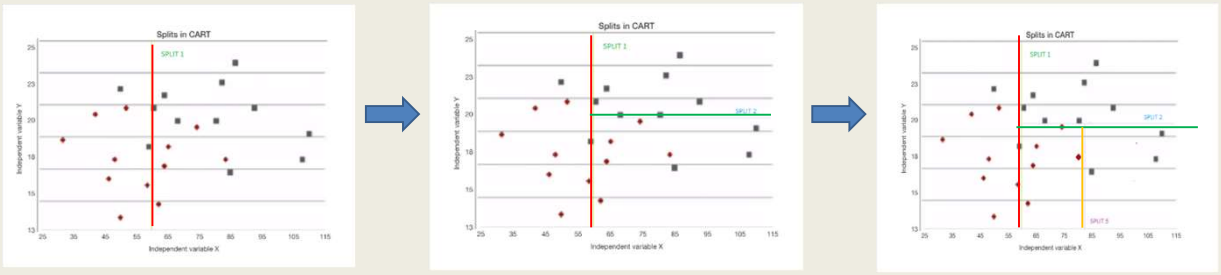
A quick example

- The plot shows a sample data for two independent variables, x, and y, and each data point is colored by the outcome variable, red or grey



37

A quick example

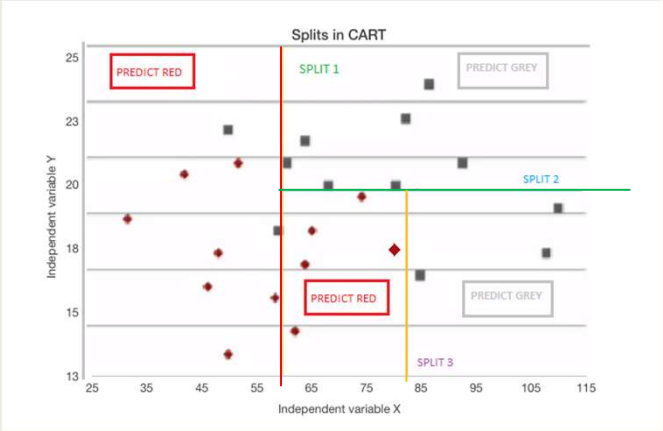


- **CART** tries to split this data into subsets so that each subset is as **homogeneous** as possible.

38

A quick example

- If a new observation fell into any of the subsets, it would now be decided by most of the observations in that subset.

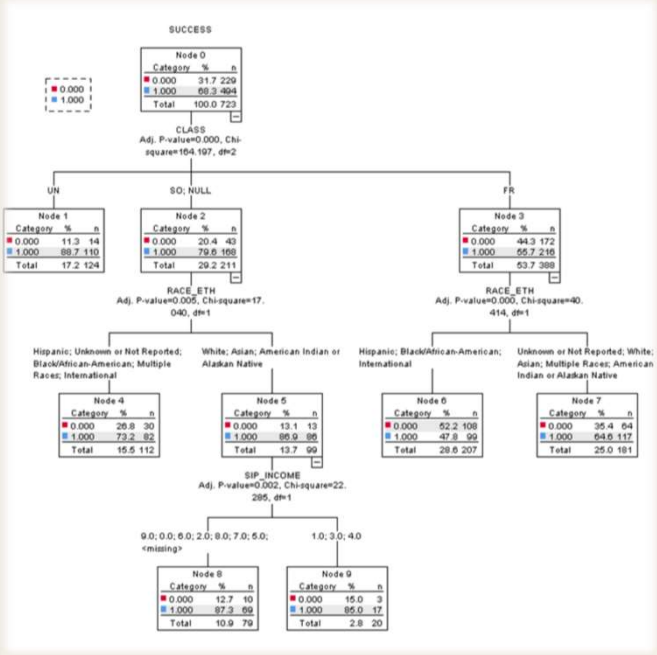


39

Chi-square Automatic Interaction Detector (CHAID)

- **CHAID** decision trees are nonparametric procedures that make no assumptions of the underlying data.
- **CHAID** algorithm operates using a series of merging, splitting, and stopping steps based on user-specified criteria.

40



41

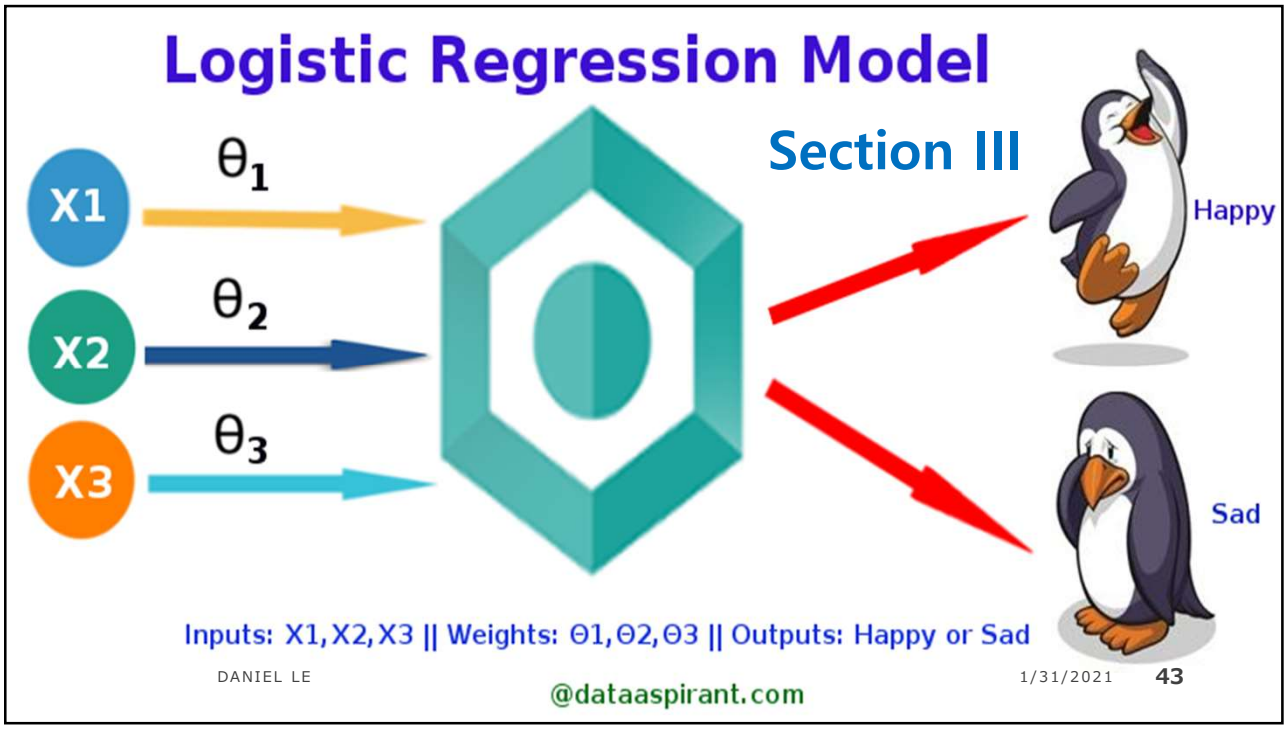
Confusion matrix

Classification

Sample	Observed	Predicted 0	Predicted 1	Percent Correct
Training	0	276	278	49.8%
	1	247	1013	80.4%
	Overall Percentage	28.8%	71.2%	71.1%
Test	0	108	121	47.2%
	1	99	395	80.0%
	Overall Percentage	28.6%	71.4%	69.6%

Growing Method: CHAID
Dependent Variable: SUCCESS

42



43

- The binomial distribution is convenient to model a binary classification problem. Suppose we have a single predictor xi. In the logistic regression, we assume

$$(Y \mid x_i) \sim \text{Bin}(m_i, \theta(x_i)) \quad i = 1, \ldots, n$$

$$\text{E}(y_i/m_i|x_i) = \theta(x_i) \text{ and } \text{Var}(y_i/m_i|x_i) = \theta(x_i)(1-\theta(x_i))/m_i$$

- Here Y_i is the number of "successes". When $m_i = 1$, Y_i is either 1 or 0.
- Our goal is to estimate $\theta(x_i)$. Since y_i/m_i is an unbiased estimate of $\theta(x_i)$, we shall consider it as a response variable.
  - If possible, it is recommend to consider (y_i, m_i) as a vector rather that the proportion.

44

Odds ratio

- When θ is a probability, the quantity θ/(1-θ) is called odds. The concept of odds has two forms.
- Suppose θ is a probability of "success".
  - 1. We define $\text{Odds in favor of success} = \frac{P(\text{success})}{1-P(\text{success})} = \frac{\theta}{1-\theta}.$
  - 2. We define $\text{Odds against success} = \frac{1-P(\text{success})}{P(\text{success})} = \frac{1-\theta}{\theta}.$

Case Processing Summary

Unweighted Cases[a]		N	Percent
Selected Cases	Included in Analysis	2413	95.1
	Missing Cases	124	4.9
	Total	2537	100.0
Unselected Cases		0	.0
Total		2537	100.0

a. If weight is in effect, see classification table for the total number of cases.

(Binary) Logistic regression full model

Variables in the Equation

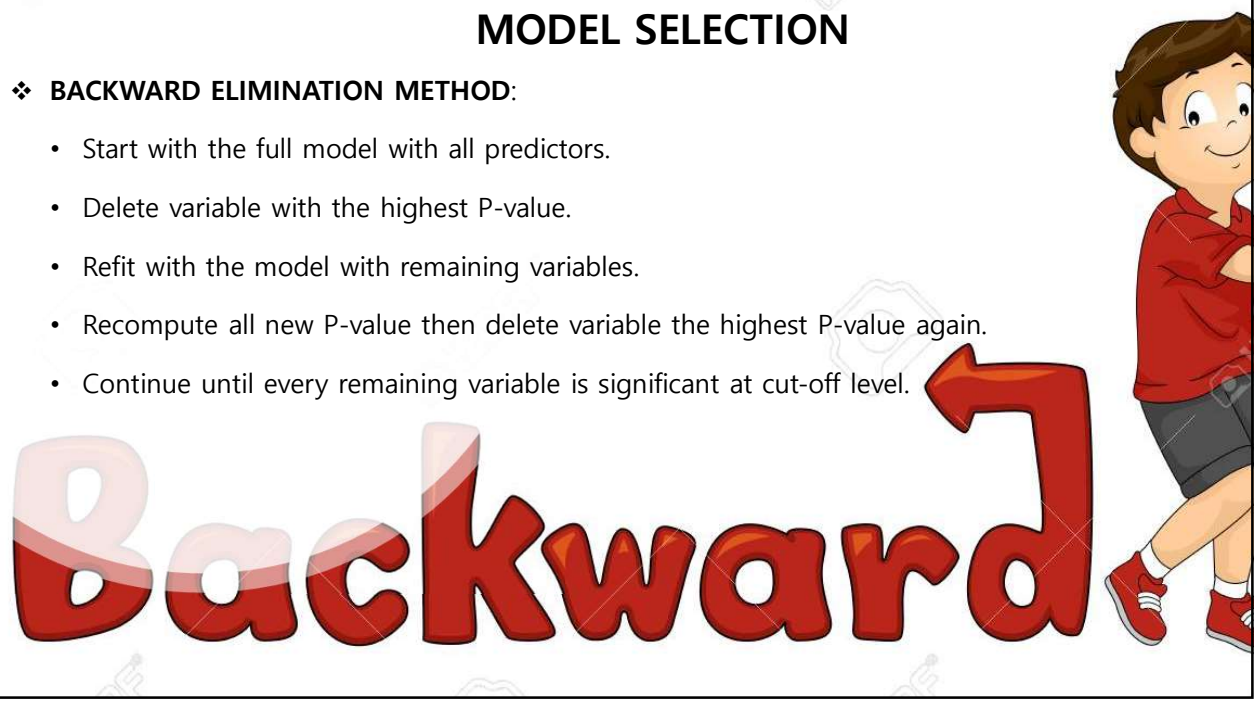
		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SIP_DEPENDANTS	-.053	.028	3.509	1	.061	.948
	SIP_INCOME			26.753	9	.002	
	SIP_INCOME(1)	-.467	.151	9.499	1	.002	.627
	SIP_INCOME(2)	-.177	.172	1.054	1	.305	.838
	SIP_INCOME(3)	-.146	.198	.545	1	.460	.864
	SIP_INCOME(4)	-.121	.210	.330	1	.566	.886
	SIP_INCOME(5)	.070	.217	.105	1	.746	1.073
	SIP_INCOME(6)	.226	.229	.970	1	.325	1.253
	SIP_INCOME(7)	.472	.281	2.828	1	.093	1.603
	SIP_INCOME(8)	.040	.297	.018	1	.893	1.041
	SIP_INCOME(9)	.317	.166	3.654	1	.056	1.373
	SIP_EMP_STATUS			9.072	7	.248	
	SIP_EMP_STATUS(1)	-.250	.161	2.404	1	.121	.779
	SIP_EMP_STATUS(2)	-.230	.199	1.337	1	.248	.795
	SIP_EMP_STATUS(3)	-.525	.183	8.278	1	.004	.591
	SIP_EMP_STATUS(4)	-.222	.163	1.851	1	.174	.801
	SIP_EMP_STATUS(5)	-.415	.323	1.655	1	.198	.660
	SIP_EMP_STATUS(6)	-.431	.877	.241	1	.624	.650
	SIP_EMP_STATUS(7)	-.570	.755	.570	1	.450	.566
	SIP_EDUC_MOTHER			20.484	7	.005	
	SIP_EDUC_MOTHER(1)	-.238	.157	2.298	1	.130	.788
	SIP_EDUC_MOTHER(2)	-.417	.161	6.726	1	.010	.659
	SIP_EDUC_MOTHER(3)	-.874	.224	15.250	1	.000	.417
	SIP_EDUC_MOTHER(4)	.009	.190	.002	1	.962	1.009
	SIP_EDUC_MOTHER(5)	-.236	.182	1.670	1	.196	.790
	SIP_EDUC_MOTHER(6)	-.246	.249	.975	1	.323	.782
	SIP_EDUC_MOTHER(7)	-.448	.372	1.449	1	.229	.639

A brief version of the full model variables. Most of the P-value is NOT significant at 5% level (P-value is greater than 0.05).

47

MODEL SELECTION

- **BACKWARD ELIMINATION METHOD**:
  - Start with the full model with all predictors.
  - Delete variable with the highest P-value.
  - Refit with the model with remaining variables.
  - Recompute all new P-value then delete variable the highest P-value again.
  - Continue until every remaining variable is significant at cut-off level.



48

Step 4[a]	SIP_DEPENDANTS	-.054	.028	3.868	1	.049	.947
	SIP_INCOME			33.592	9	.000	
	SIP_INCOME(1)	-.499	.150	11.078	1	.001	.607
	SIP_INCOME(2)	-.222	.169	1.723	1	.189	.801
	SIP_INCOME(3)	-.174	.193	.816	1	.366	.840
	SIP_INCOME(4)	-.139	.205	.464	1	.496	.870
	SIP_INCOME(5)	.091	.214	.182	1	.670	1.095
	SIP_INCOME(6)	.216	.227	.910	1	.340	1.241
	SIP_INCOME(7)	.481	.276	3.030	1	.082	1.618
	SIP_INCOME(8)	.060	.295	.041	1	.840	1.061
	SIP_INCOME(9)	.359	.163	4.864	1	.027	1.432
	SIP_EDUC_MOTHER			27.040	7	.000	
	SIP_EDUC_MOTHER(1)	-.303	.150	4.077	1	.043	.739
	SIP_EDUC_MOTHER(2)	-.456	.151	9.152	1	.002	.634
	SIP_EDUC_MOTHER(3)	-.861	.192	20.183	1	.000	.423
	SIP_EDUC_MOTHER(4)	-.042	.182	.054	1	.816	.959
	SIP_EDUC_MOTHER(5)	-.201	.178	1.272	1	.259	.818
	SIP_EDUC_MOTHER(6)	-.568	.210	7.303	1	.007	.566
	SIP_EDUC_MOTHER(7)	-.447	.362	1.528	1	.216	.639
	AGE	.016	.006	6.676	1	.010	1.016
	RACE_ETH			82.142	8	.000	
	RACE_ETH(1)	.444	.799	.308	1	.579	1.558
	RACE_ETH(2)	.481	.217	4.908	1	.027	1.618
	RACE_ETH(3)	-.931	.126	54.622	1	.000	.394
	RACE_ETH(4)	-.393	.134	8.585	1	.003	.675
	RACE_ETH(5)	-2.203	.853	6.665	1	.010	.110
	RACE_ETH(6)	-.461	.276	2.792	1	.095	.630
	RACE_ETH(7)	19.691	28151.138	.000	1	.999	356113499.2
	RACE_ETH(8)	-.621	.216	8.294	1	.004	.538
	Constant	1.232	.216	32.520	1	.000	3.427

a. Variable(s) entered on step 1: SIP_DEPENDANTS, SIP_INCOME, SIP_EMP_STATUS, SIP_EDUC_MOTHER, SIP_EDUC_FATHER, AGE, RACE_ETH, ENROLL_STATUS.

49

Significant Variable	P-value	Odds Ratio	Interpretation
Reference: Not reported (Annual Family Income)			
Less than $17,820	0.001	0.607	39.3% lower
$61,335 or higher	0.027	1.432	43.2% higher

***Note:** to interpret the odds ratio, compare it to 1. We can also subtract 1 from it to compute the percentage difference. If the result is positive, it is a higher odd of success. Otherwise, it is a lower odd. For example, the first number is 0.607 – 1 = - 0.393.

50

Significant Variable	P-value	Odds Ratio	Interpretation
Reference: Bachelor's degree or higher (Mother's education level)			
Attend College	0.043	0.739	26.1% lower
Graduated High School	0.002	0.634	36.6% lower
Attended High School	0.000	0.423	57.7% lower

51

Significant Variable	P-value	Odds Ratio	Interpretation
Reference: 16 (Age)			
Age + 1	0.010	1.016	1.6% higher

52

Significant Variable	P-value	Odds Ratio	Interpretation
Reference: White (Race/Ethnicity)			
Asian	0.027	1.618	61.8% higher
African American	0.000	0.394	60.6% lower
Hispanic	0.003	0.675	32.5% lower
International	0.010	0.110	89% lower
Unknown	0.004	0.538	46.2% lower