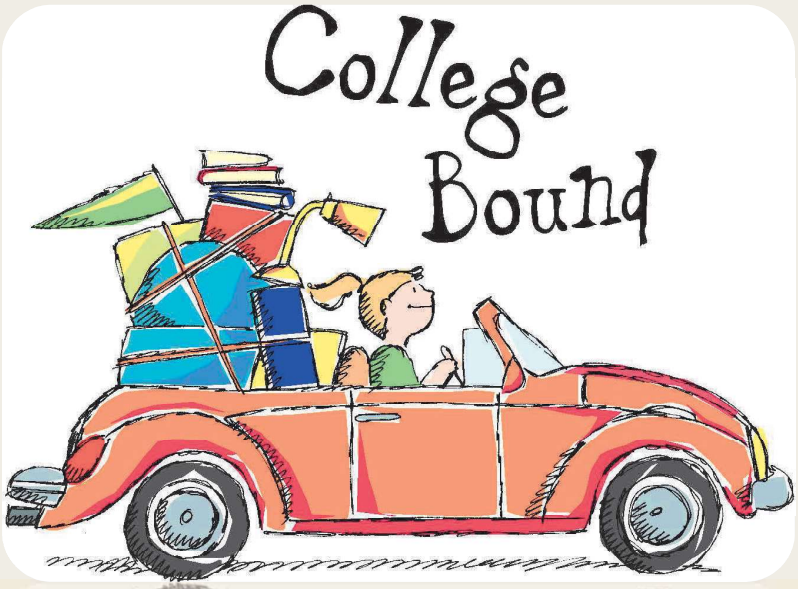
Survival Analysis and Predictive Modeling on college student's course withdrawal rate.



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

DALLAS COLLEGE

After this session, attendees will be able to:

- Help academic advisors identify crucial characteristics of students likely to drop.
- Use predictive analytics as a powerful tool to help identify at-risk students.
- Understand students' course withdrawal behaviors to inform educators in updating withdrawal policies.

PART I:

INTRODUCTION

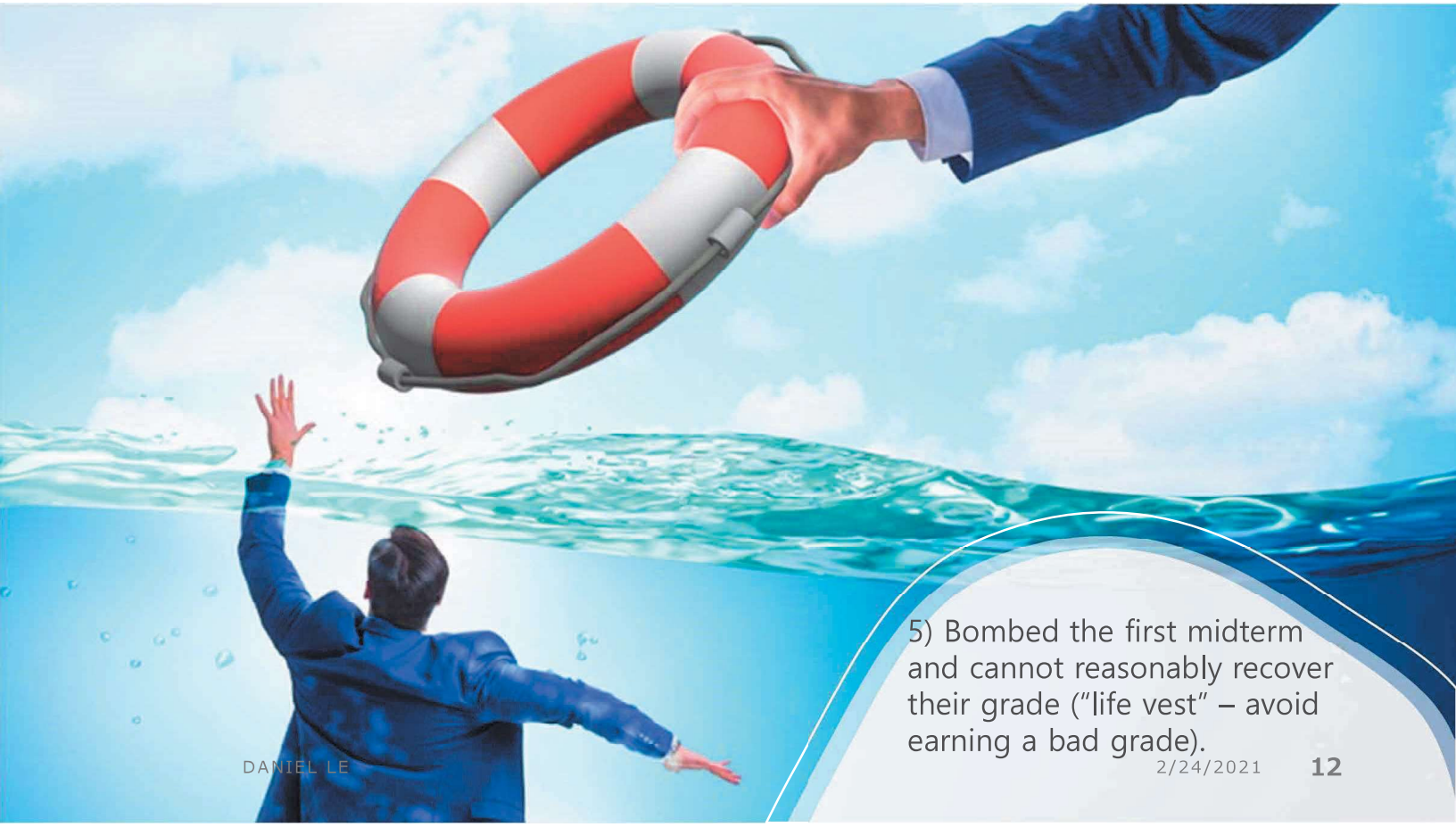
Common reasons why student drops a course

1) Have too many courses in one semester and cannot manage the workload.

2) The timing and overall schedule is not well organized (i.e. too many back to back classes, too spread out, too early, or too late).

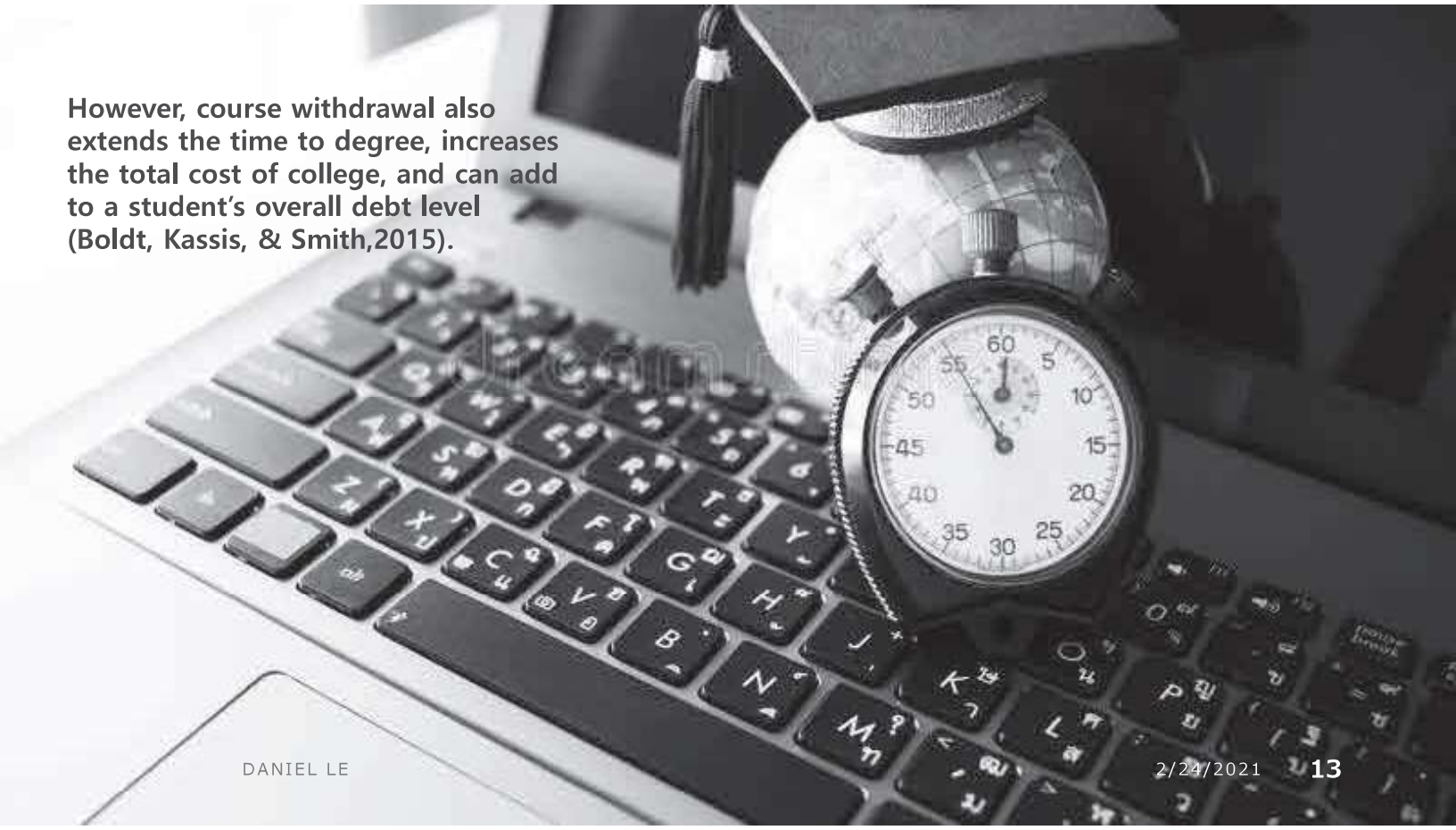
3) The course is not required for their degree, is not relevant to their degree, or is not an acceptable elective.

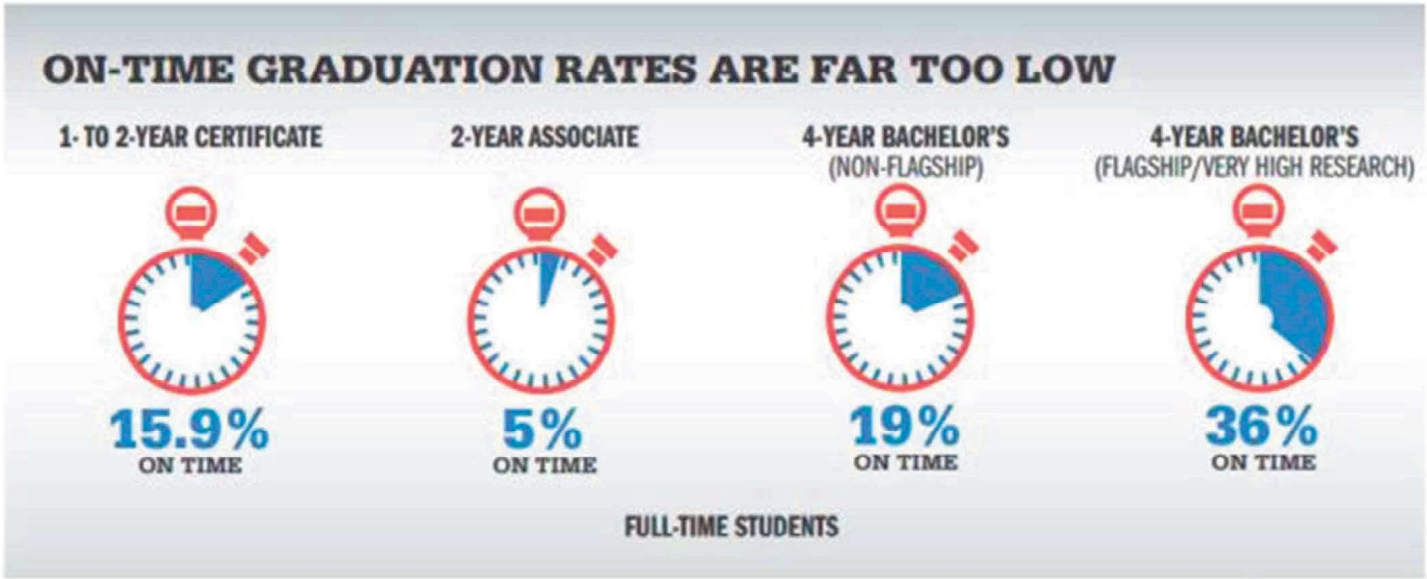
4) They are too far behind in the syllabus and you cannot fathom catching up.



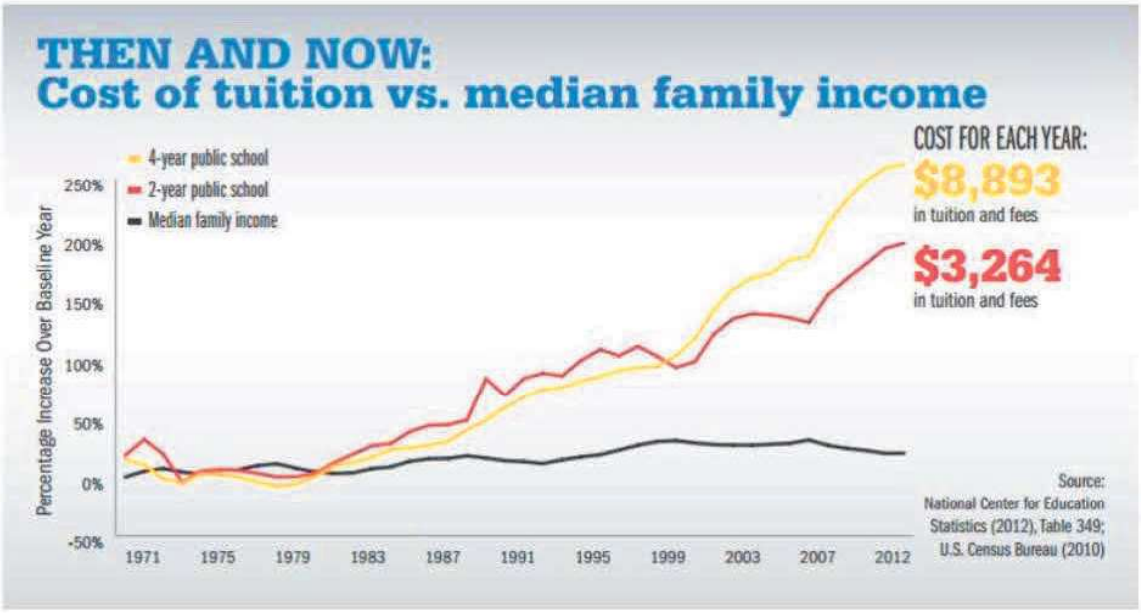
5) Bombed the first midterm and cannot reasonably recover their grade ("life vest" – avoid earning a bad grade).

However, course withdrawal also extends the time to degree, increases the total cost of college, and can add to a student's overall debt level (Boldt, Kassis, & Smith,2015).




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

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.



DATASET

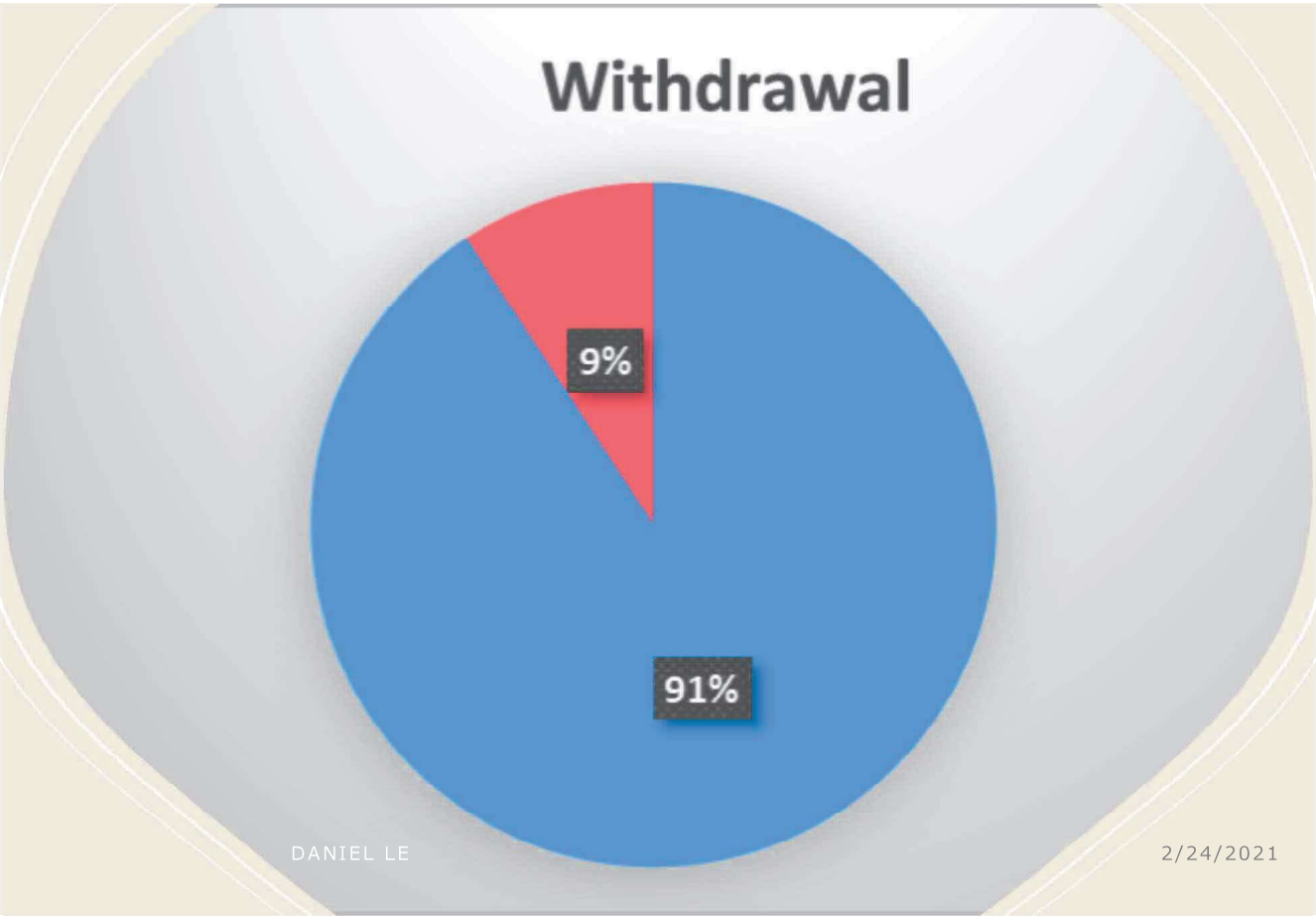
- Following IRB (Institutional review board) approval, a prospectively maintained database of sample of Fall 2020 cohort students is reviewed.

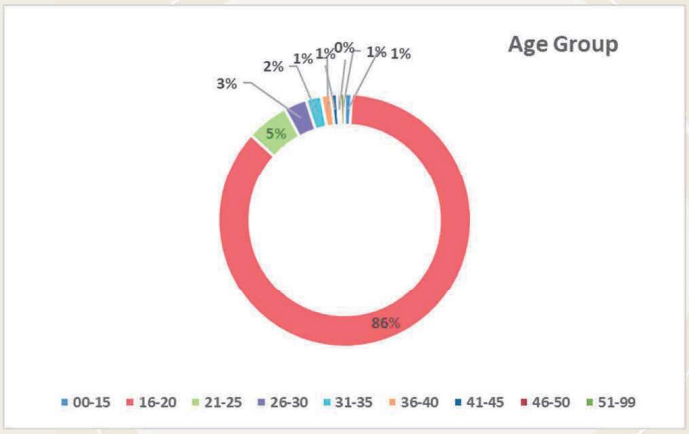
- **Inclusion criteria:** Fall 2020 cohort students who enrolled in Fall 2020 in Dallas College.

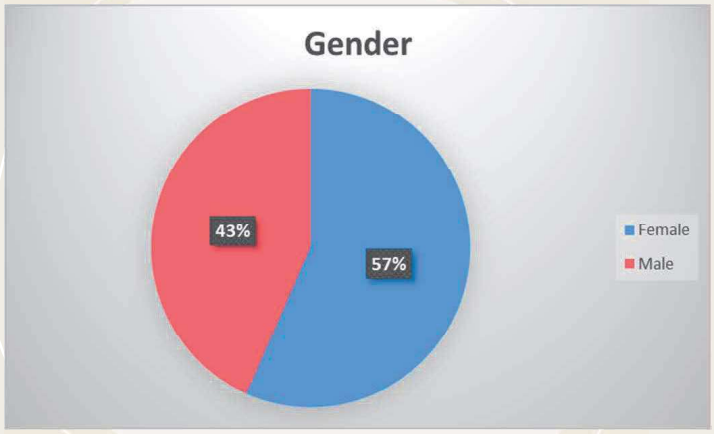
- **Exclusion criteria:** any students who do not meet the above inclusion criteria will be removed from the dataset.

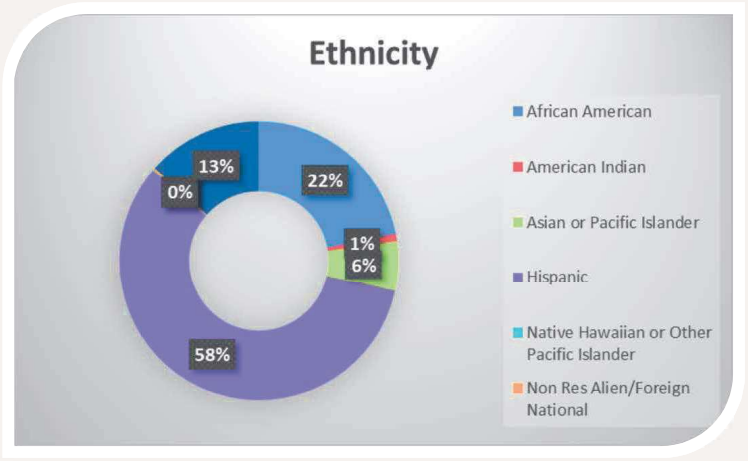
PART II: STATISTICAL ANALYSIS

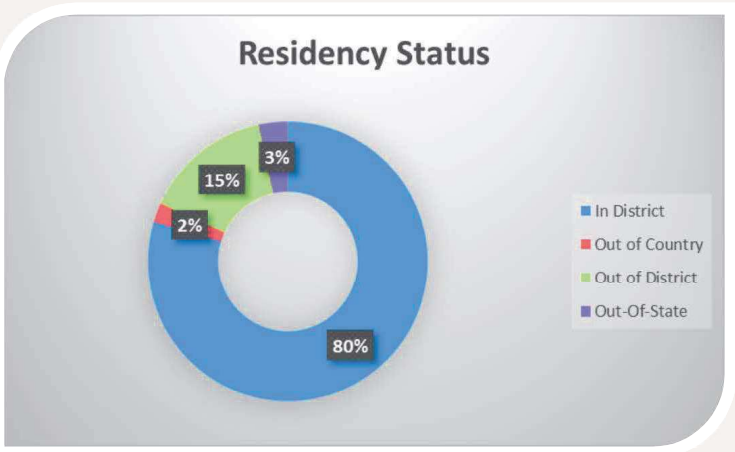
DESCRIPTIVE STATISTICS

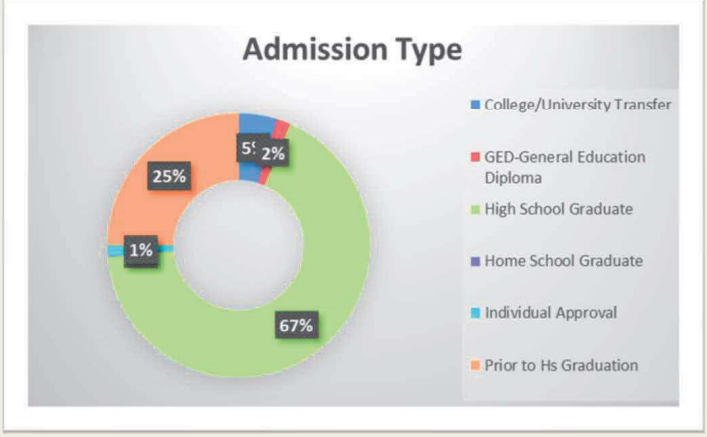

Withdrawal
9%
91%

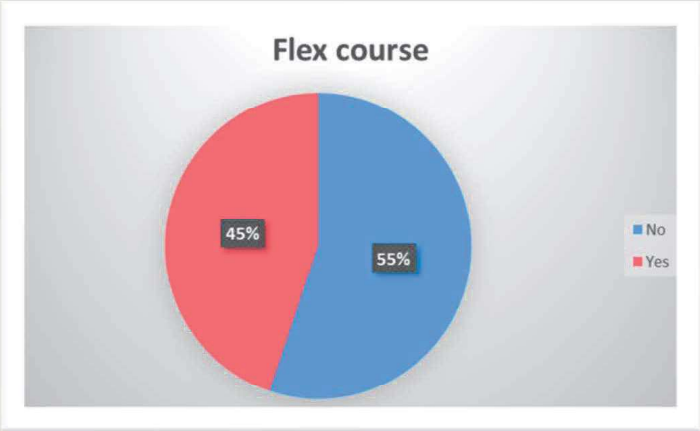

Age Group
3%
2%
1%
1%
0%
1%
1%
5%
86%
00-15
16-20
21-25
26-30
31-35
36-40
41-45
46-50
51-99


Gender
43%
57%
Female
Male


Ethnicity
13%
22%
0%
1%
6%
58%
African American
American Indian
Asian or Pacific Islander
Hispanic
Native Hawaiian or Other Pacific Islander
Non Res Alien/Foreign National


Residency Status
3%
15%
2%
80%
In District
Out of Country
Out of District
Out-Of-State


Admission Type
College/University Transfer
GED-General Education Diploma
High School Graduate
Home School Graduate
Individual Approval
Prior to Hs Graduation
2%
25%
1%
67%

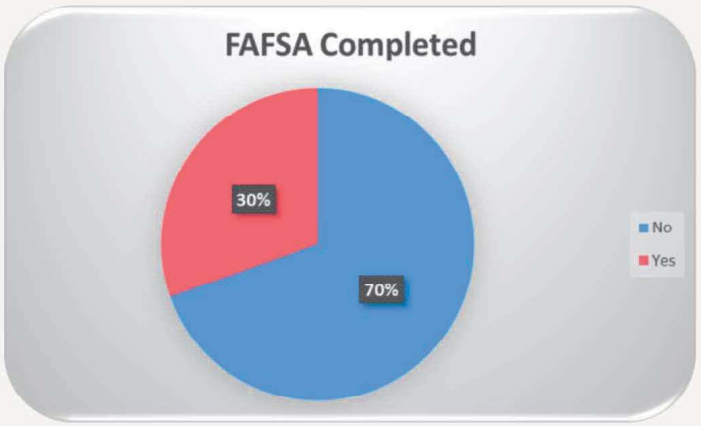

Flex course
No
Yes
45%
55%

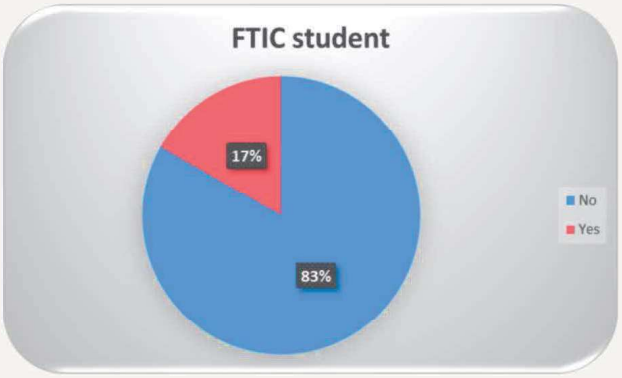
Flex Term Credit Class Schedules

Flex Term classes:

- Let you register throughout the year.
- Last for various lengths of time (not always a whole regular semester).
- Meet once a week, three times a week or every day.
- May help you complete your degree faster through more inter

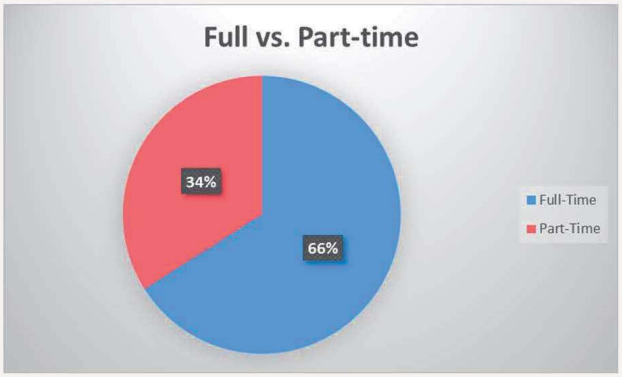

Credit Level
9%
91%
College
Developmental

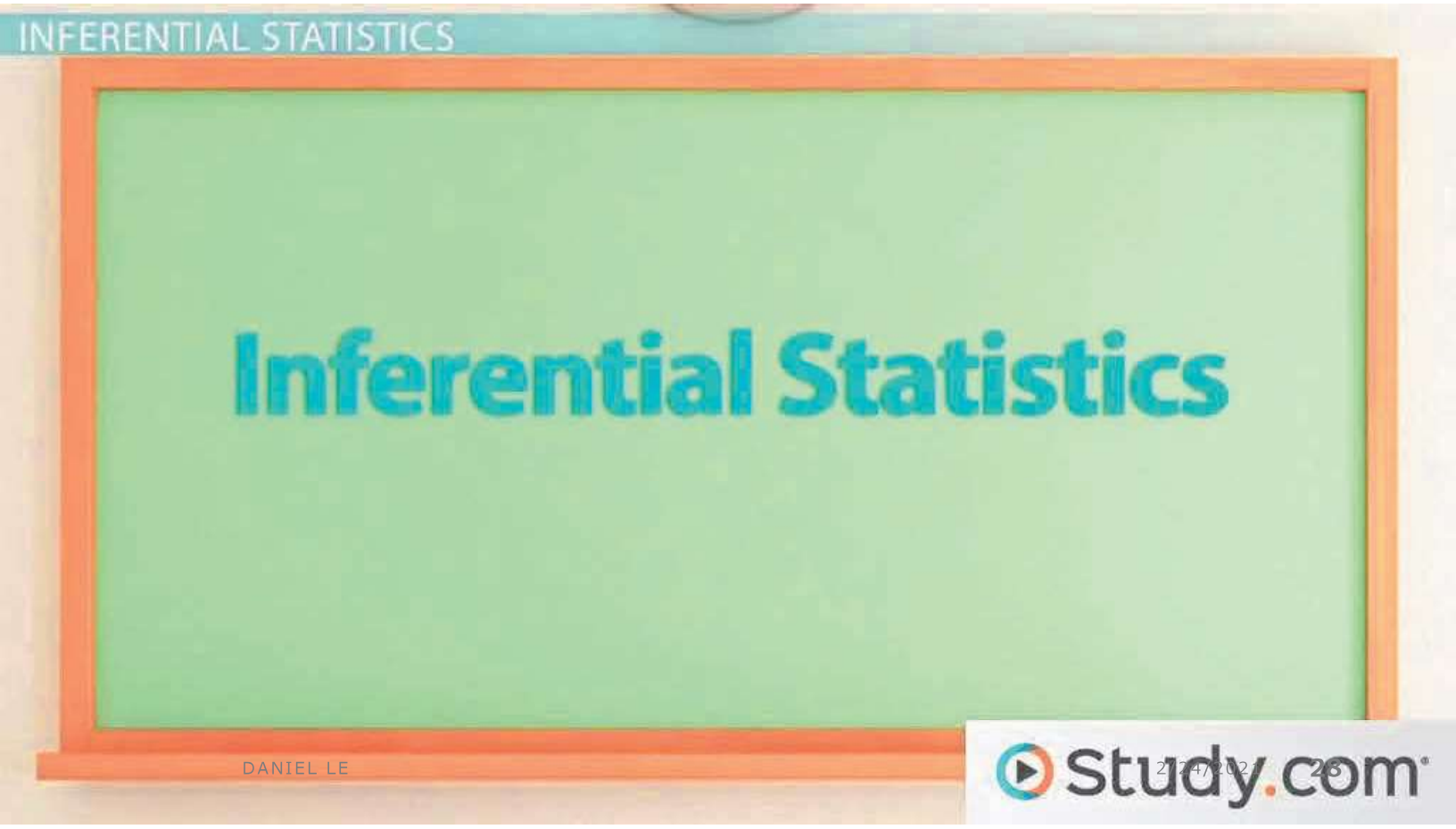

FAFSA Completed
30%
70%
No
Yes


FTIC student
17%
83%
No
Yes

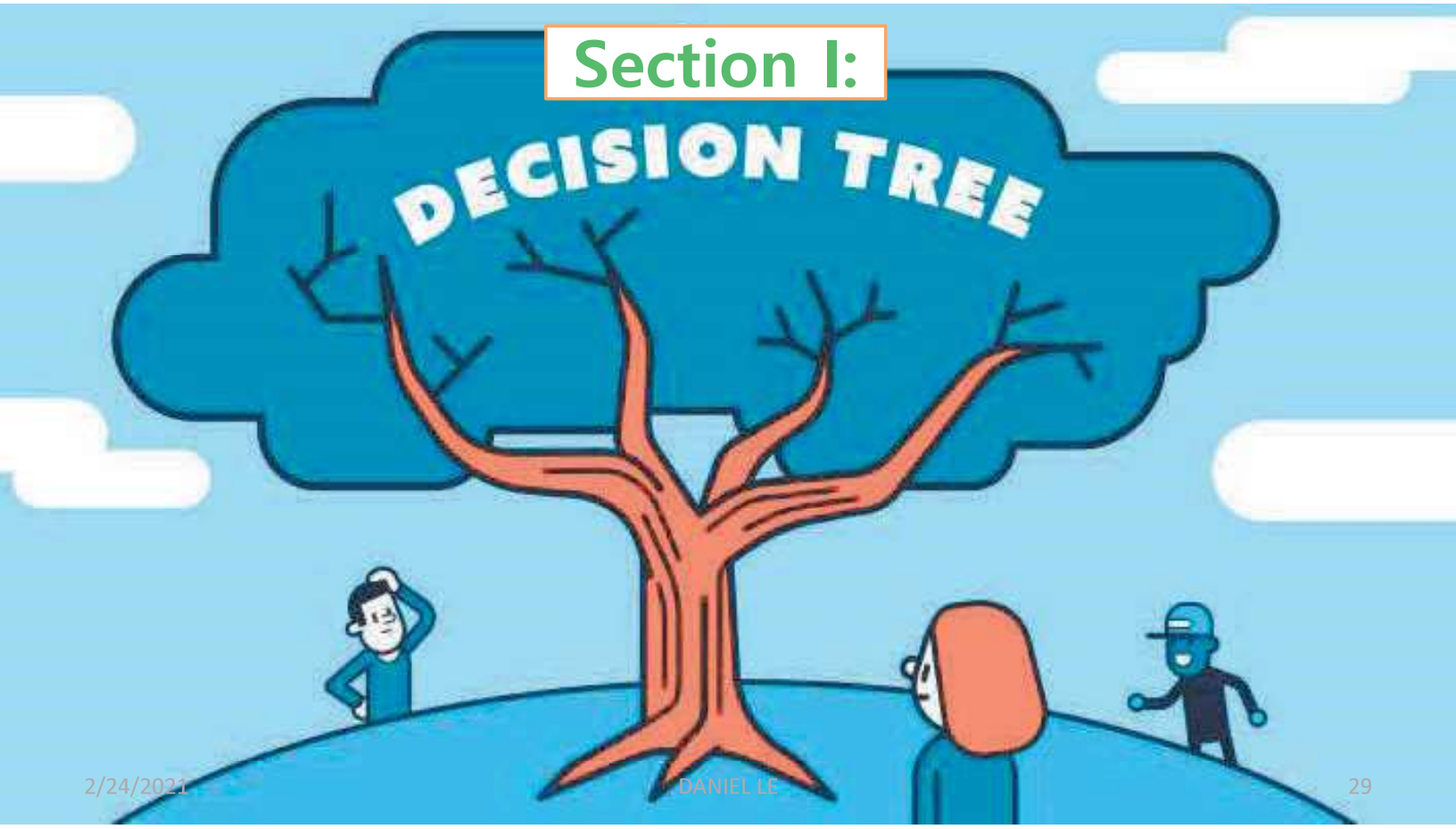

Schedule Type
0%
0%
9%
0%
0%
90%
Blended - 50% Or Less Online
Day
Distance Learning
Flex - Day
Flex Distance Learning
Flex Hybrid
Hybrid - 51% to 84% Online
Night
Off Campus
Self-Paced


Course Level
8%
16%
76%
0
1
2


Full vs. Part-time
34%
66%
Full-Time
Part-Time


Inferential Statistics

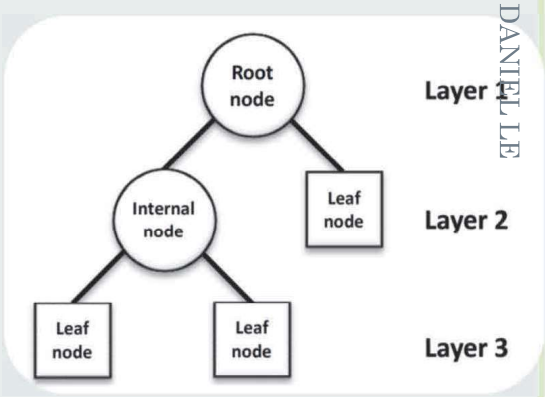
Section I:



DANIEL LE

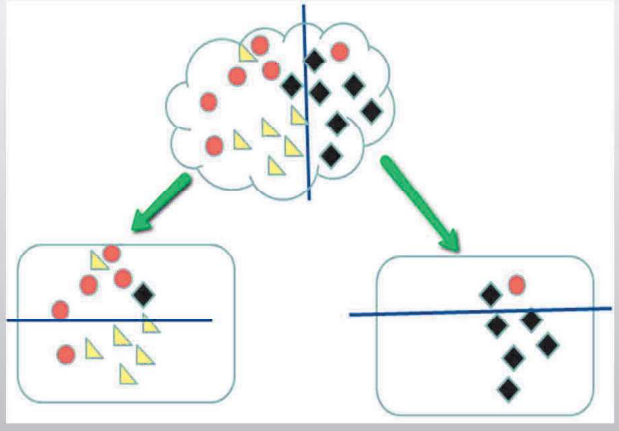
WHAT IS DECISION TREE?

- **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**.



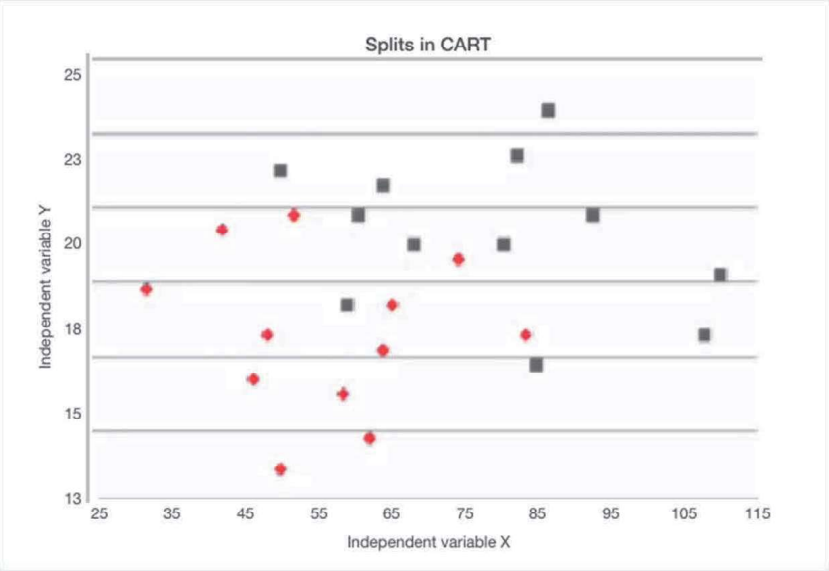
How Does A Decision Tree Work?

- 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.**

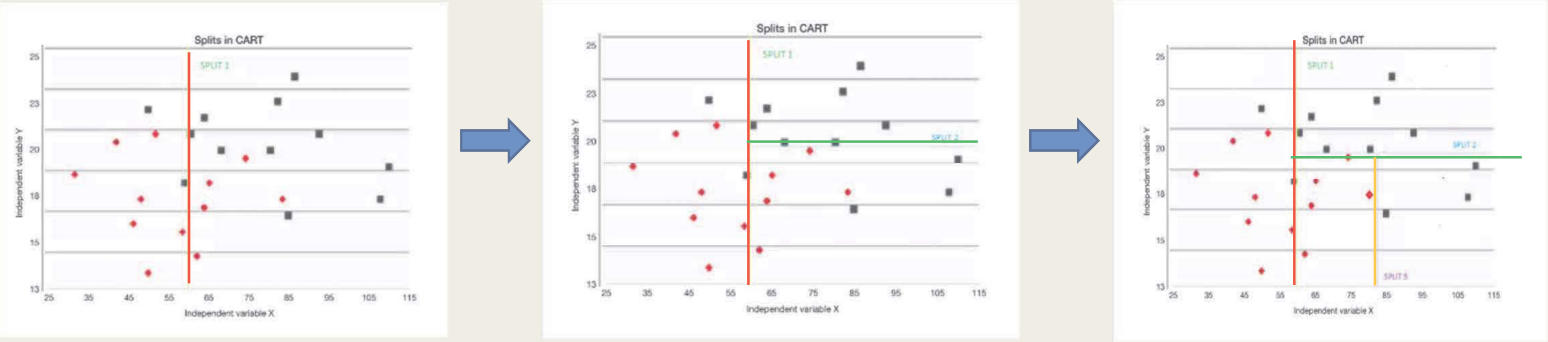


A quick example

- 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



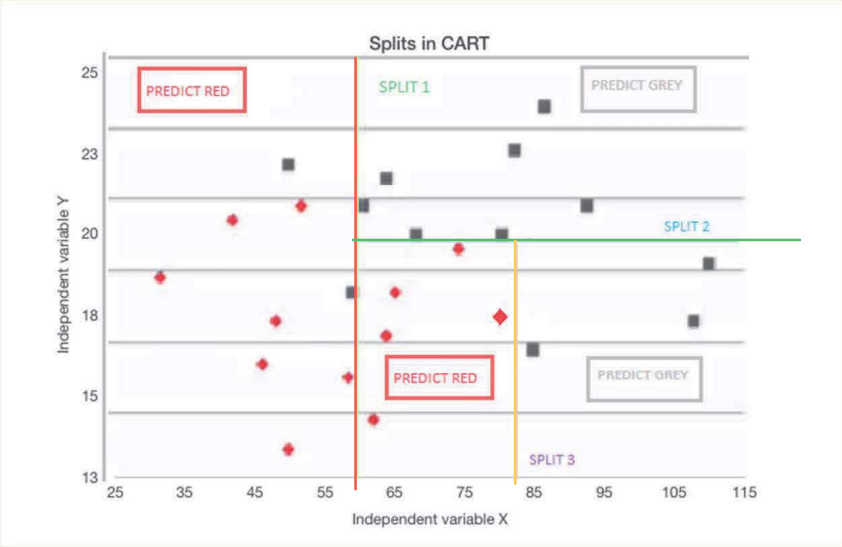
A quick example



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

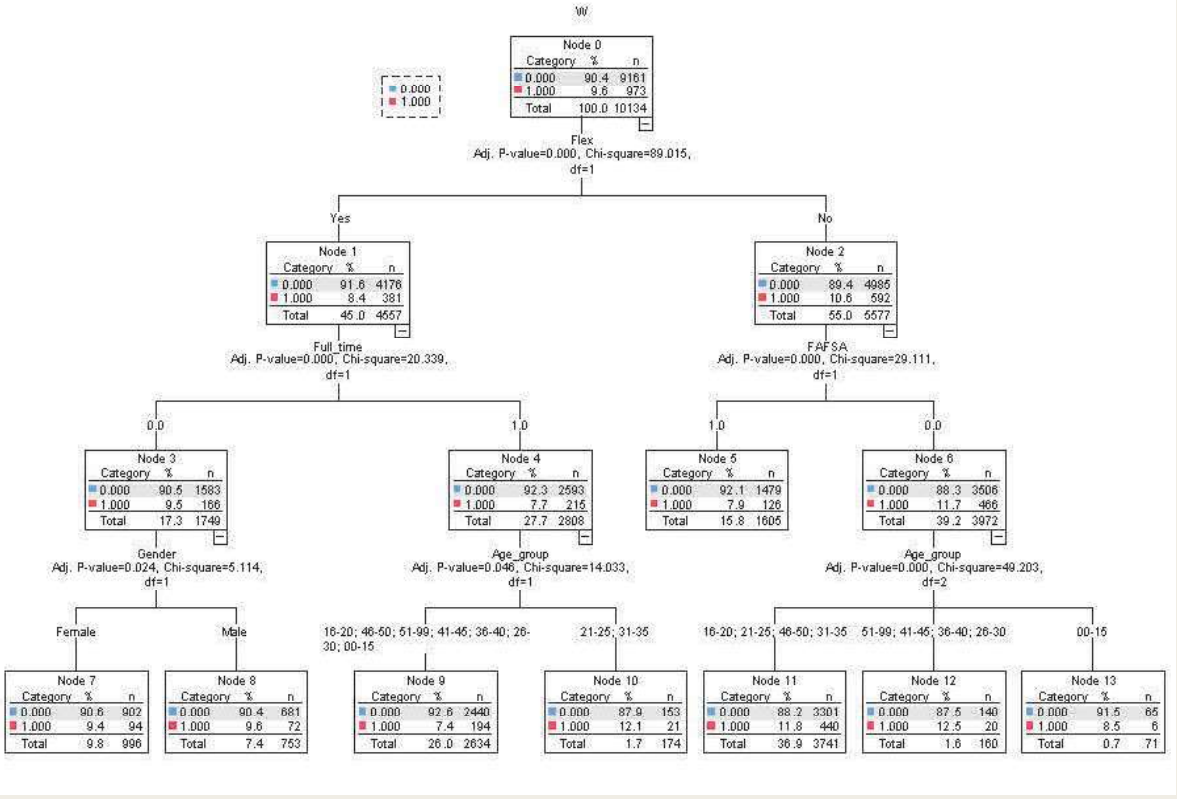
A quick example

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



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.


W
0.000
1.000
Node 0
Category % n
0.000 90.4 9161
1.000 9.6 973
Total 100.0 10134
Flex
Adj. P-value=0.000, Chi-square=89.015, df=1
Yes
No
Node 1
Category % n
0.000 91.6 4176
1.000 8.4 381
Total 45.0 4557
Node 2
Category % n
0.000 89.4 4985
1.000 10.6 592
Total 55.0 5577
Full_time
Adj. P-value=0.000, Chi-square=20.339, df=1
FAFSA
Adj. P-value=0.000, Chi-square=29.111, df=1
0.0
1.0
1.0
0.0
Node 3
Category % n
0.000 90.5 1583
1.000 9.5 166
Total 17.3 1749
Node 4
Category % n
0.000 92.3 2593
1.000 7.7 215
Total 27.7 2808
Node 5
Category % n
0.000 92.1 1479
1.000 7.9 126
Total 15.8 1605
Node 6
Category % n
0.000 88.3 3506
1.000 11.7 466
Total 39.2 3972
Gender
Adj. P-value=0.024, Chi-square=5.114, df=1
Age_group
Adj. P-value=0.046, Chi-square=14.033, df=1
Age_group
Adj. P-value=0.000, Chi-square=49.203, df=2
Female
Male
16-20; 46-50; 51-99; 41-45; 36-40; 26-30; 00-15
21-25; 31-35
16-20; 21-25; 46-50; 31-35
51-99; 41-45; 36-40; 26-30
00-15
Node 7
Category % n
0.000 90.6 902
1.000 9.4 94
Total 9.8 996
Node 8
Category % n
0.000 90.4 681
1.000 9.6 72
Total 7.4 753
Node 9
Category % n
0.000 92.6 2440
1.000 7.4 194
Total 26.0 2634
Node 10
Category % n
0.000 87.9 153
1.000 12.1 21
Total 1.7 174
Node 11
Category % n
0.000 88.2 3301
1.000 11.8 440
Total 36.9 3741
Node 12
Category % n
0.000 87.5 140
1.000 12.5 20
Total 1.6 160
Node 13
Category % n
0.000 91.5 65
1.000 8.5 6
Total 0.7 71

Confusion matrix

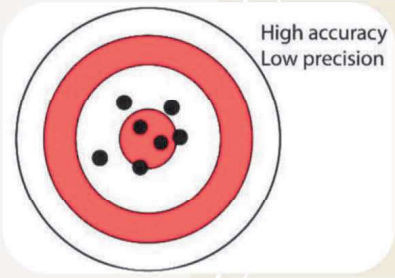
Classification

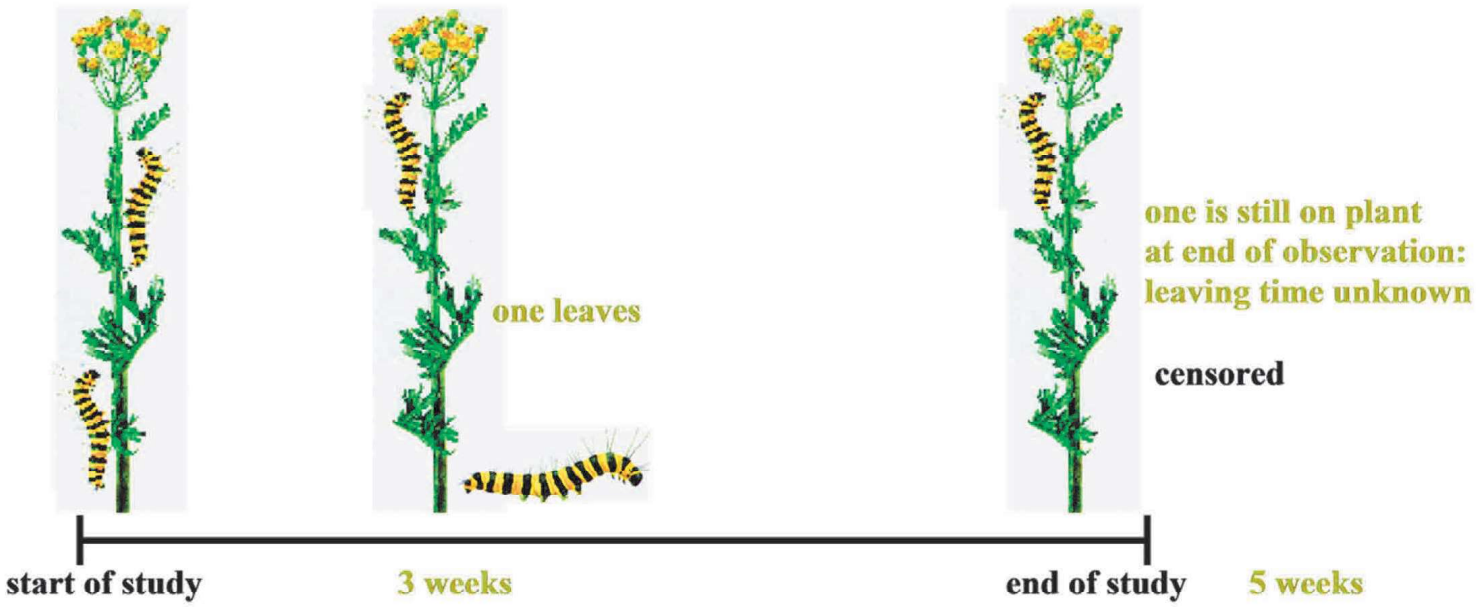
Sample	Observed	Predicted 0	Predicted 1	Percent Correct
Training	0	21534	0	100.0%
	1	2136	0	0.0%
	Overall Percentage	100.0%	0.0%	91.0%
Test	0	9161	0	100.0%
	1	973	0	0.0%
	Overall Percentage	100.0%	0.0%	90.4%

Growing Method: CHAID
Dependent Variable: W

Accuracy

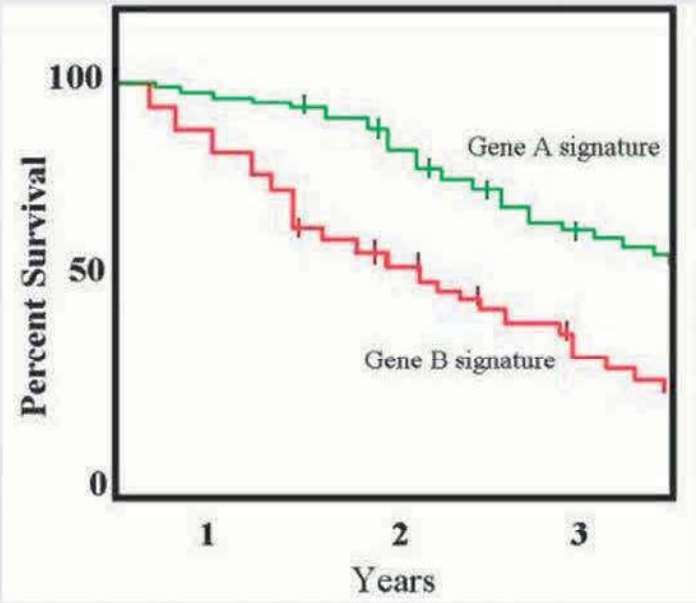
Accuracy





Section II: SURVIVAL ANALYSIS

Kaplan–Meier estimator



https://en.wikipedia.org/wiki/File:Km_plot.jpg

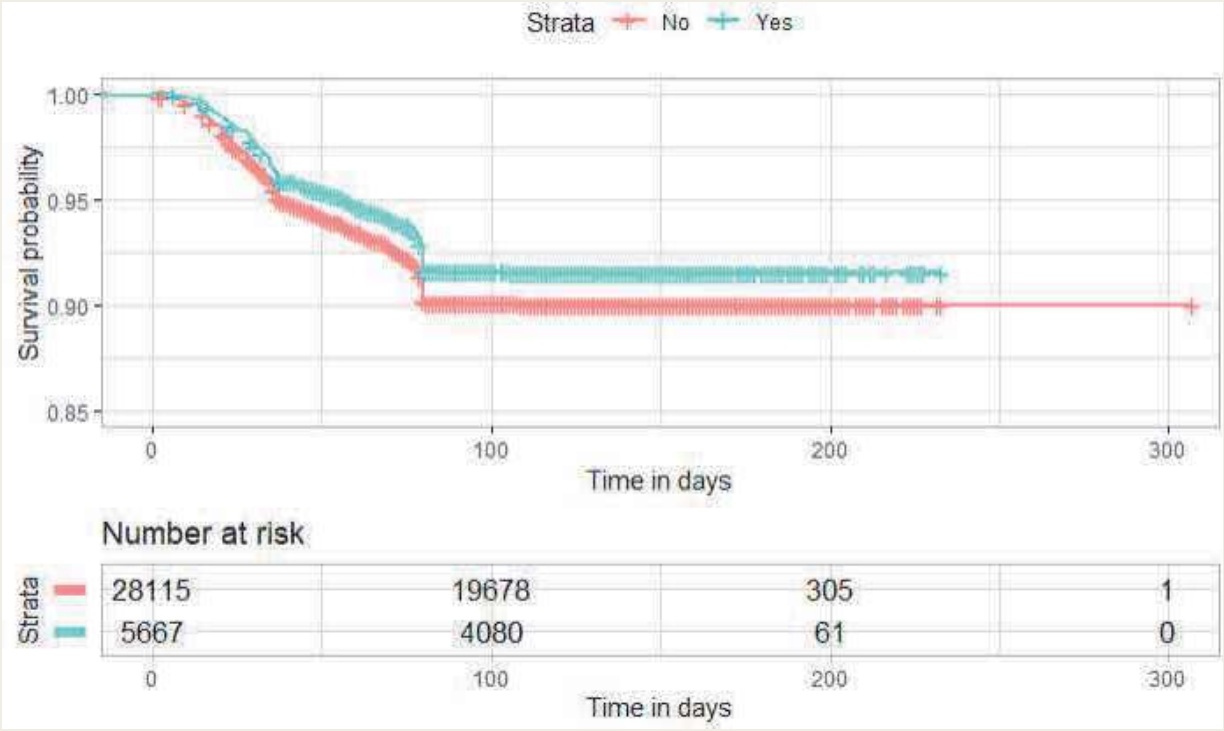
Kaplan – Meier curves

- **Kaplan – Meier estimator** is widely used in clinical and fundamental research to estimate the survival function.

- The visual representation of this function is usually called the **Kaplan-Meier curve**, and it shows what the probability of an event (for example, **survival – probability of students NOT withdraw from their course in this study**) is at a certain time interval.

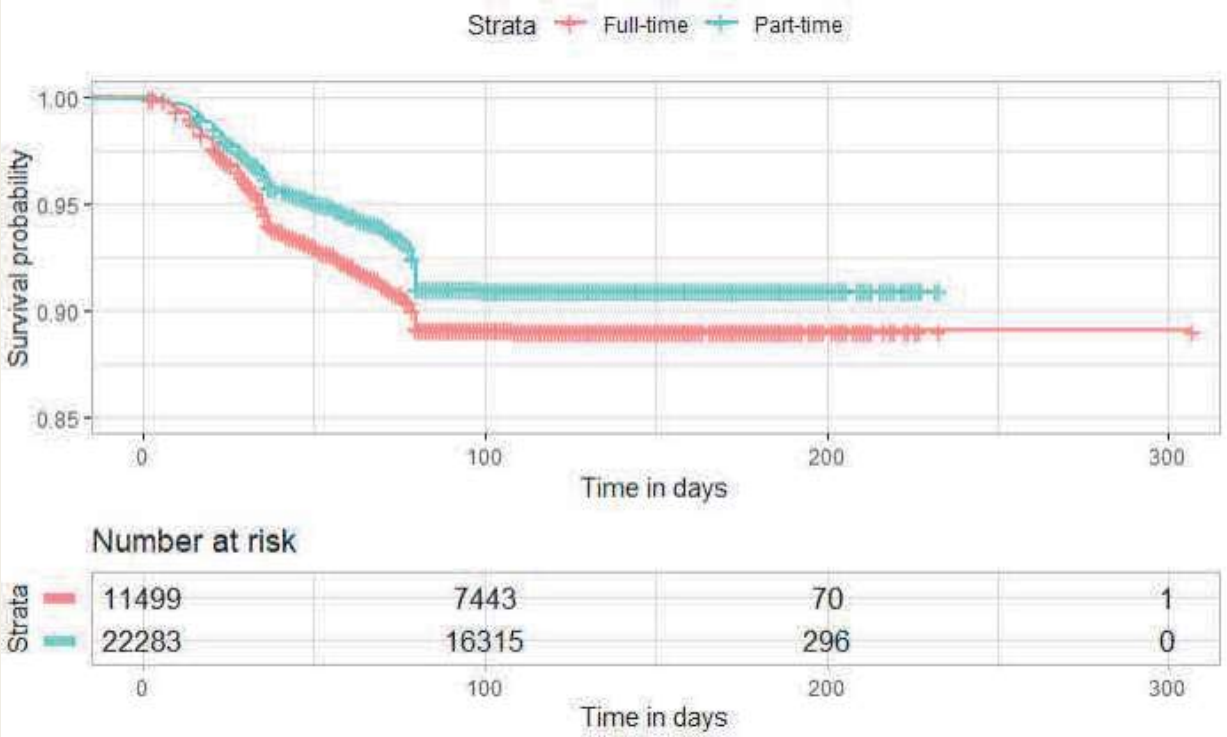
- If the sample size is large enough, the curve should approach the true survival function for the population under investigation.

- It is usually compared two or more groups in a study.



Log-rank
P-value = $4e^{-4}$

FTIC



Log-rank
P-value = $9e^{-10}$

Full-time vs. Part-time

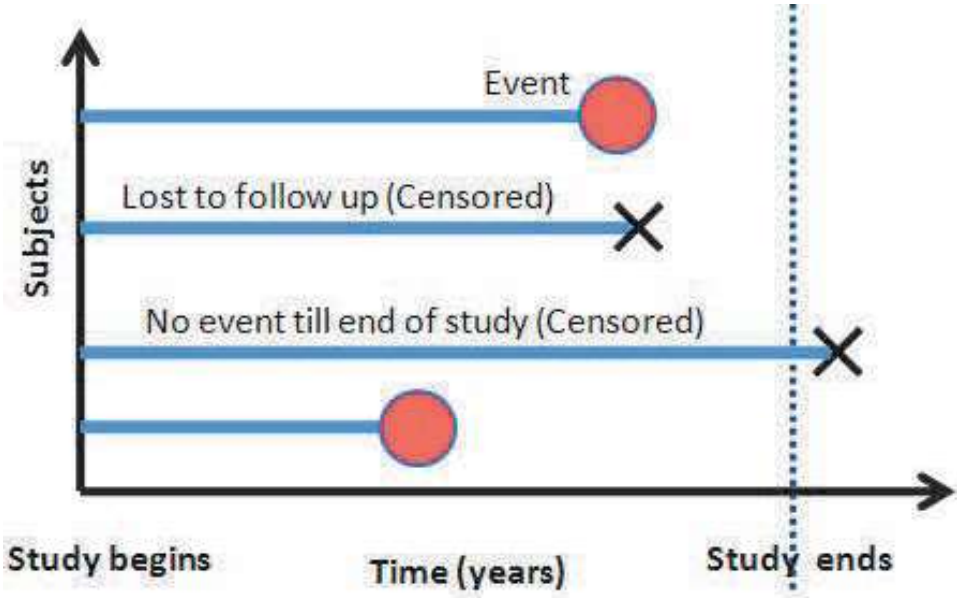
QUESTION:

Which factor(s) have a significant impact on the student course withdrawal rate?

THE COX PROPORTIONAL HAZARDS MODEL

The **Cox proportional-hazards model** (Cox, 1972) is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables.

- **Subject:** student
- **Censored:** when student completes the course
- **Event of interest:** withdraw
- **Time to event:** how many days from the course start date until the date student withdraw from their course.
- **Hazard rate** = course withdrawal rate



A Cox Proportional Hazards full model (with all 14 variables).

n= 33804, number of events= 3109

	coef	exp(coef)	se(coef)	z	Pr(>|z|)	
factor(Admit)Home	2.965e-01	1.345e+00	5.237e-01	0.566	0.571195	
factor(Admit)HSG	1.422e-01	1.153e+00	1.364e-01	1.043	0.297127	
factor(Admit)IA	3.048e-01	1.356e+00	1.860e-01	1.639	0.101222	
factor(Admit)PHSG	1.201e-01	1.128e+00	1.430e-01	0.840	0.401126	
factor(Age_group)16-20	8.380e-01	2.312e+00	2.620e-01	3.198	0.001385	**
factor(Age_group)21-25	9.716e-01	2.642e+00	2.717e-01	3.576	0.000349	***
factor(Age_group)26-30	1.097e+00	2.994e+00	2.780e-01	3.944	8.01e-05	***
factor(Age_group)31-35	7.609e-01	2.140e+00	2.947e-01	2.582	0.009810	**
factor(Age_group)36-40	1.057e+00	2.878e+00	2.985e-01	3.541	0.000398	***
factor(Age_group)41-45	1.508e+00	4.519e+00	3.057e-01	4.934	8.06e-07	***
factor(Age_group)46-50	4.598e-01	1.584e+00	4.248e-01	1.083	0.279025	
factor(Age_group)51-99	9.676e-01	2.632e+00	3.456e-01	2.800	0.005112	**
factor(`Credit Level`)Dev	-1.564e-01	8.552e-01	8.757e-02	-1.786	0.074171	.
factor(`Credit Level`)Other	-1.173e+01	8.035e-06	8.928e+02	-0.013	0.989516	
factor(Ethnicity)API	-5.720e-01	5.644e-01	1.014e-01	-5.640	1.70e-08	***
factor(Ethnicity)FOR	-3.572e-01	6.996e-01	5.079e-01	-0.703	0.481860	
factor(Ethnicity)HI	-1.391e-01	8.701e-01	4.443e-02	-3.132	0.001736	**
factor(Ethnicity)IA	-5.525e-02	9.463e-01	1.928e-01	-0.287	0.774475	
factor(Ethnicity)NH	-9.462e-03	9.906e-01	5.052e-01	-0.019	0.985057	
factor(Ethnicity)WH	-9.201e-02	9.121e-01	6.047e-02	-1.522	0.128103	
factor(FAFSA)1	-2.128e-01	8.083e-01	4.126e-02	-5.157	2.51e-07	***
factor(FTIC)1	-1.222e-01	8.850e-01	5.635e-02	-2.168	0.030151	*

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).

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.

Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
Reference: No (Flex)				
Yes	3.75e-14	0.7313	(0.6744, 0.7930)	~ 26.9% lower

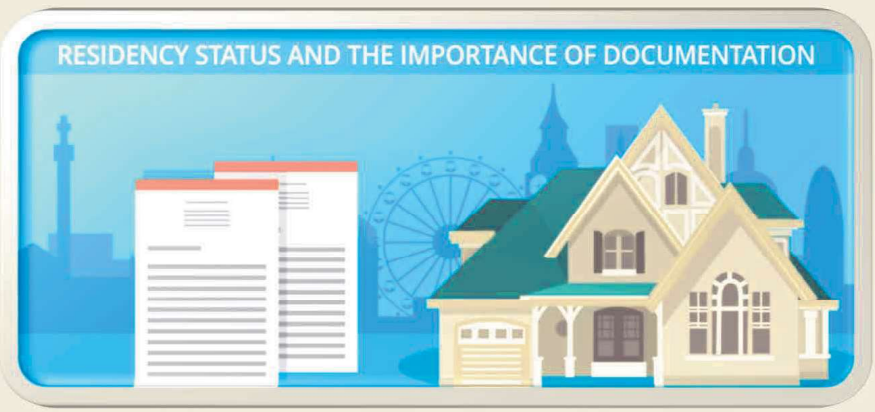
***Note:** to interpret the hazard ratio, compare it to 1. We can also subtract 1 from it to compute the percentage difference. If the result is positive, it is a higher hazard rate. Otherwise, it is a lower hazard rate. For example, the first number is 0.7313 – 1 = - 0.2687.

Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
Reference: College (Credit level)				
Developmental	0.0299	0.8271	(0.6969, 0.9817)	~ 17.3% lower

Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
Reference: African American (Ethnicity)				
Asian or Pacific Islander	1.05e-8	0.5600	(0.4592, 0.6831)	~ 44% lower
Hispanic	0.00081	0.8618	(0.7899, 0.9402)	~ 13.8% lower

Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
Reference: No (FAFSA Completed)				
Yes	3.28e-7	0.8105	(0.7477, 0.8786)	~ 19% lower

Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
	Reference: No (FTIC)			
Yes	0.01895	0.8854	(0.7998, 0.9801)	~ 11.5% lower



Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
Reference: In District (Residency Status)				
Out of country	0.00743	0.6609	(0.4880, 0.8950)	~ 33.9% lower

Significant Variable	P-value	Hazard Ratio	95% CI	Interpretation
Reference: Part – Time				
Full – Time	3.39e-9	0.7988	(0.7415, 0.8606)	~ 20.1% lower

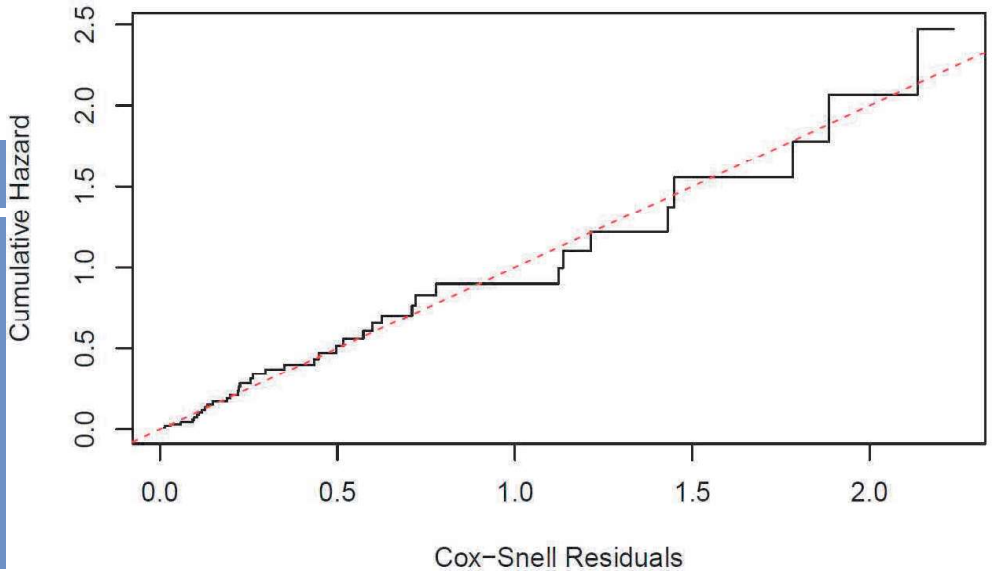
GOODNESS OF FIT OF THE FINAL MODEL

We can evaluate the fit of the model by using the **Cox-Snell residuals**.

We graph the **Nelson-Aalen cumulative hazard function** so that we can compare the hazard function to the diagonal line.

If the hazard function follows the diagonal (45^{o}) line, we know that it approximately has an exponential distribution with a hazard rate of one and that the model fits the data well.

The Nelson-Aalen cumulative hazard function.



- **Note:** We see that the hazard function follows the 45 degrees line very closely except for some very large values of time. It is very common for models with censored data to have some wiggling at large values of time and it is not something which should cause much concern. Overall, we would conclude that the final model fits the data very well.

SUMMARY

Seven Significant factors: Flex, Credit Level, Ethnicity, FAFSA, FTIC, Residency Status, Full - Time.

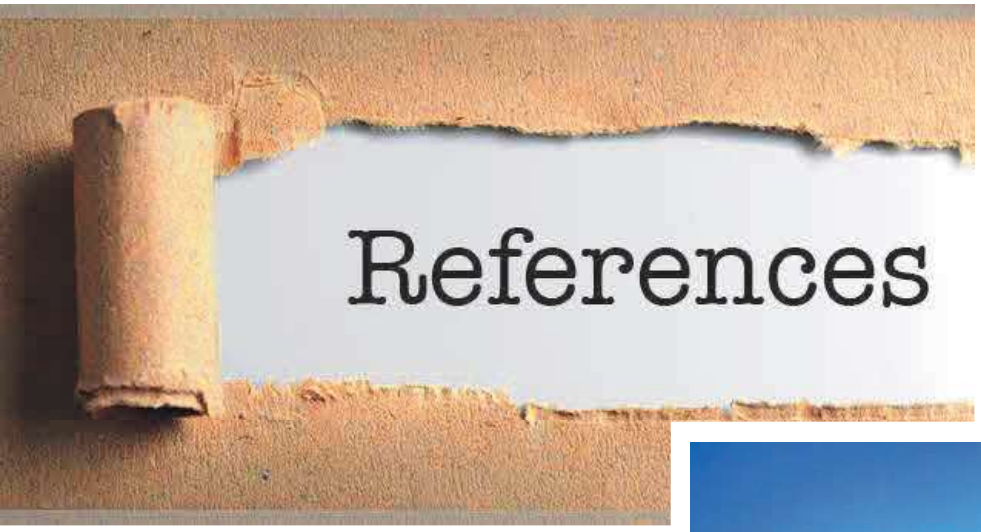
African American students, Not completed FAFSA students, and part – time students are three at – risk groups that need more supports.


Conclusion

- **Cox Hazards Proportional Model** points out the hazard ratio of some significant factors on course withdrawal rate. However, we need to make assumption that all students will eventually withdraw their course, given the follow up period is long enough (main assumption of this regression model).

- **Plan for future research study:** extend the inclusive criteria, adding more students in many different types of courses, and include more of their characteristics to feed into the regression model.

- Hope the result can benefit both students and college advisors/ administrators to improve student success rate.


References

- Adelman, C. (2006). The toolbox revisited: Paths to degree completion from high school through college. Washington, D.C.: U.S. Department of Education.
- Adelman, C. (2005). Moving into town—and moving on: The community college in the lives of traditional–age students. Washington, DC: U.S. Department of Education.
- Boldt, D.J., Kassis, M.M., & Smith, W.J. (2015). Factors impacting the likelihood of withdrawal in core business classes. Journal of College Student Retention, 1-16.
- Intuitive Machine Learning, "Decision Tree: Important things to know", https://www.youtube.com/watch?v=JcI5E2Ng6r4&ab_channel=IntuitiveMachineLearning
- McKinney, L., Novak, H., Hagedorn, L.S. et al. Giving Up on a Course: An Analysis of Course Dropping Behaviors Among Community College Students. Res High Educ 60, 184–202 (2019). https://doi.org/10.1007/s11162-018-9509-z
- Nguyen, Ethan, "Risk factors for infant mortality using the CDC's National Center for Health Statistics 2012 data", UTA Master Project, 2020.
- https://towardsdatascience.com/kaplan-meier-curves-c5768e349479
- https://www.usnews.com/news/blogs/data-mine/2014/12/01/report-too-much-freedom-hurts-college-graduation-rates
- https://onlineacademiccommunity.uvic.ca/myuviclife/2016/02/25/5-reasons-you-can-drop-a-class-and-5-reasons-you-probably-shouldnt/#:~:text=%205%20Reasons%20You%20Can%20Drop%20a%20Course%3A,degree%2C%20or%20isn%E2%80%99t%20an%20acceptable%20elective.%20More


THANK YOU
FOR LISTENING!