

# **STEM DATA SET**

## **THE EFFECT OF OUTSIDE VARIABLES ON STEM STUDENT GPA**

STAT 112 - GROUP 5  
DAPHNE HIDLEY (SENIOR 105502680)  
KAILYN NGUYEN (SENIOR 606036735)  
JUNSEONG HWANG (SENIOR 505550240),  
AIDAN OHLSON (JUNIOR 405815712)  
KEVIN LI (JUNIOR 805733229)

# ABSTRACT



For our project, we explored data from the STEM survey dataset. This dataset contains the responses of 1365 UCLA STEM students between 2019-2023 to around 50 questions regarding their academic, social, and family life. The purpose of our research was to find which variables in the survey have the biggest influence on whether a UCLA student's GPA is **low** or **high**.

In order to answer this, we implemented **random forest**, **logistic regression**, and **text mining** to analyze the data.



# DATASET AND VARIABLES

**We narrowed our variables of interest to the eight which we thought would be most influential:**

1. Gender
2. Ethnic Background
3. High School GPA
4. Number of Friends (other + own ethnicity)
5. Amount of Sleep
6. Frequency of Missed Class
7. Family Financial Struggle
8. Academic Anxiety

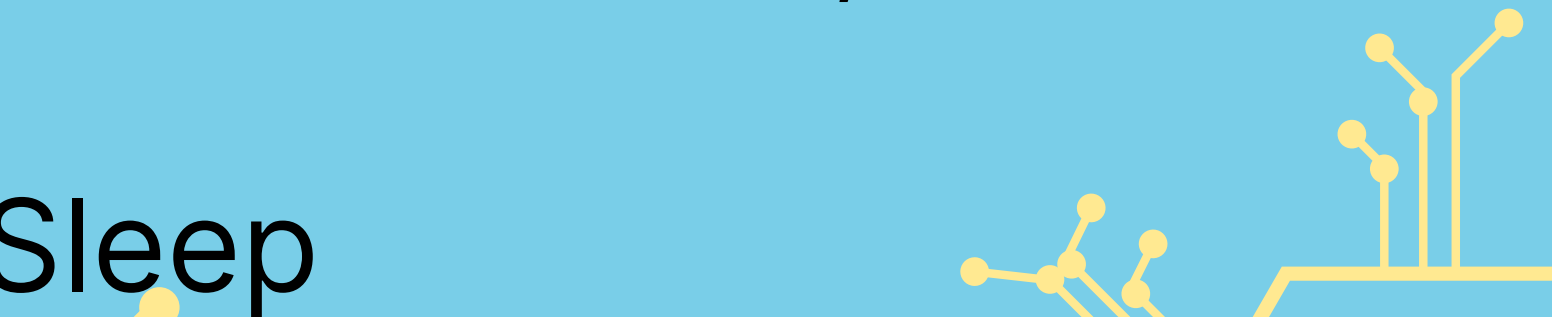
This data was gathered via survey from 2019-2023. After cleaning the data, we were left with around 1,000 responses that were usable in our analysis.

Variables	Variable Types	Measurement	Description
GPA(gpa)	Outcome	Categorical (“>=3.7”, “< 3.7”)	RESPONSE VARIABLE
Gender	Predictor	Categorical ("female", "male", "others")	What is your gender?
Ethnic Background	Predictor	Categorical ("Latino", "Caucasian", "Asian", "Mixed/Other")	What is your ethnic background?
High school GPA	Predictor	Categorical (“0-3.5”, “3.5-4”, “4-5”)	What was your high school GPA?
Friends	Predictor	Categorical (“0 ~ 11”, “12 ~ 21”, “22 ~ 39”, “40 ~”)	How many friends do you have?
Sleep	Predictor	Categorical (“Never”, “Rarely”, “Sometimes”, “Often”, “Always”)	How often do you get enough sleep?
Miss Class	Predictor	Categorical (“Never”, “Rarely”, “Sometimes”, “Often”, “Always”)	How often do you miss classes?
Family Financial Struggle	Predictor	Categorical (“Strongly Agree”, “Agree”, “Neither Agree or Disagree”, “Disagree”, “Strongly Disagree”)	My family struggled financially when I was growing up...
Academic Anxiety	Predictor	Categorical (“Strongly Agree”, “Agree”, “Neither Agree or Disagree”, “Disagree”, “Strongly Disagree”)	I experience anxiety about academics.

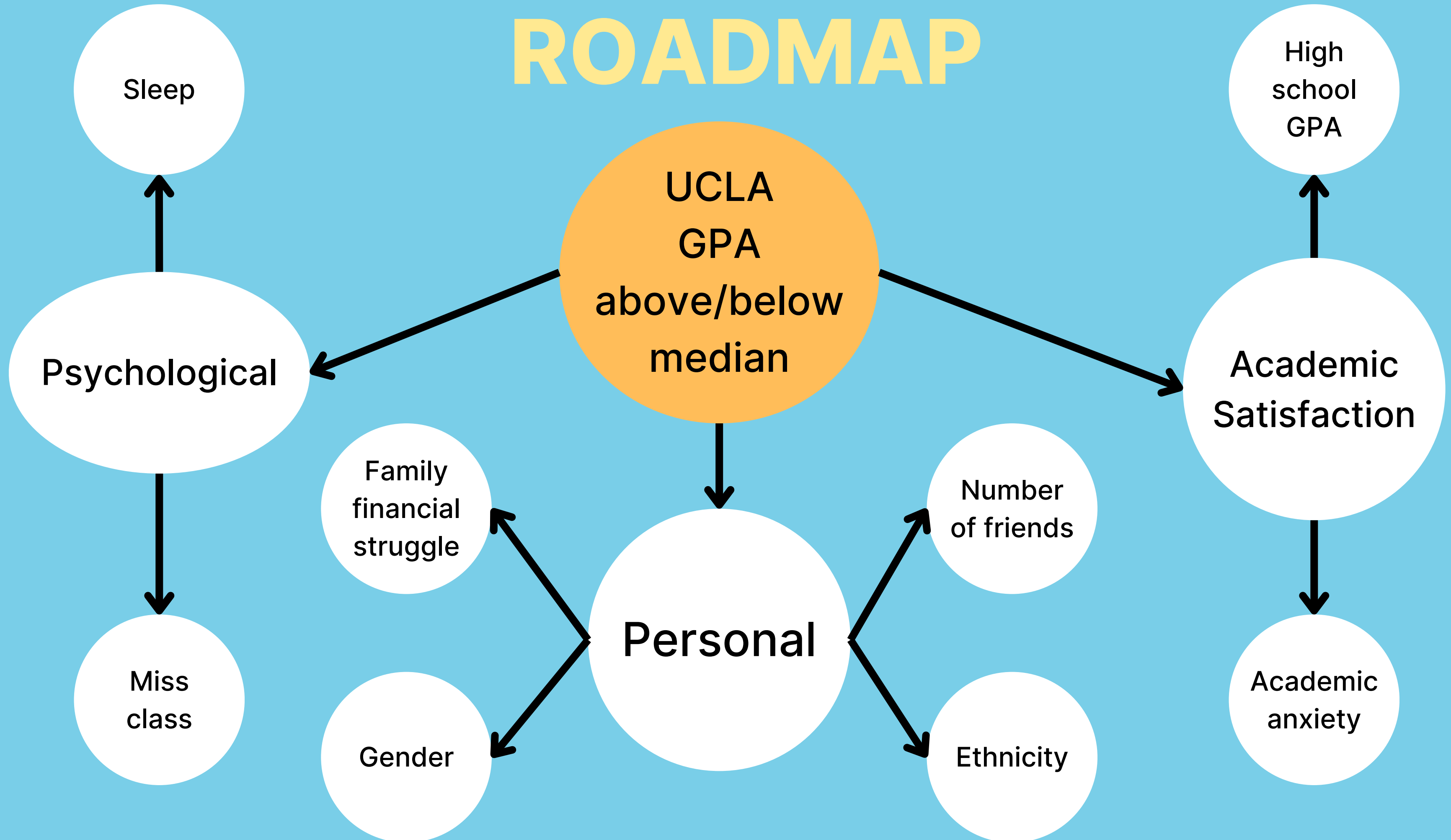


# RESEARCH QUESTION

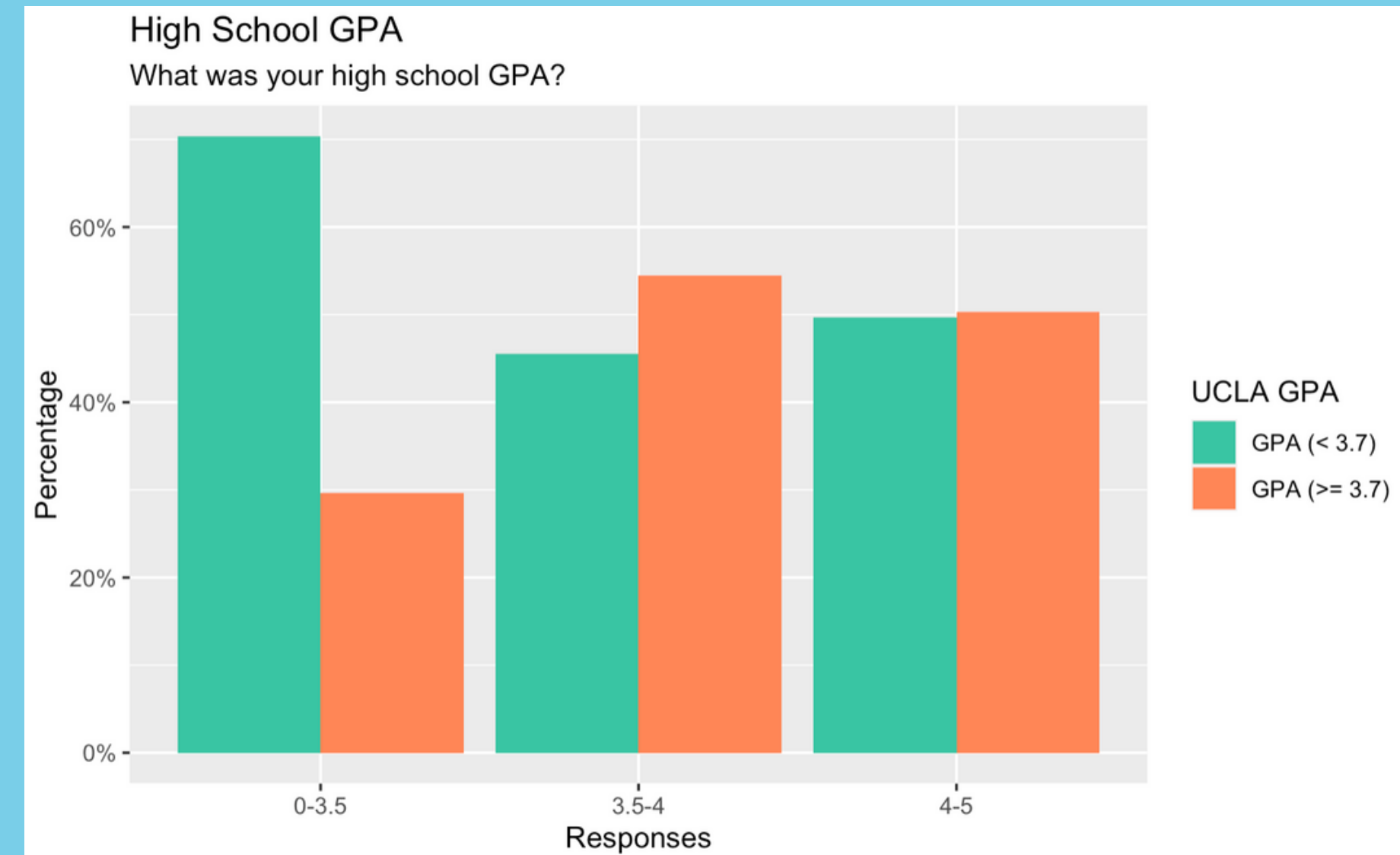
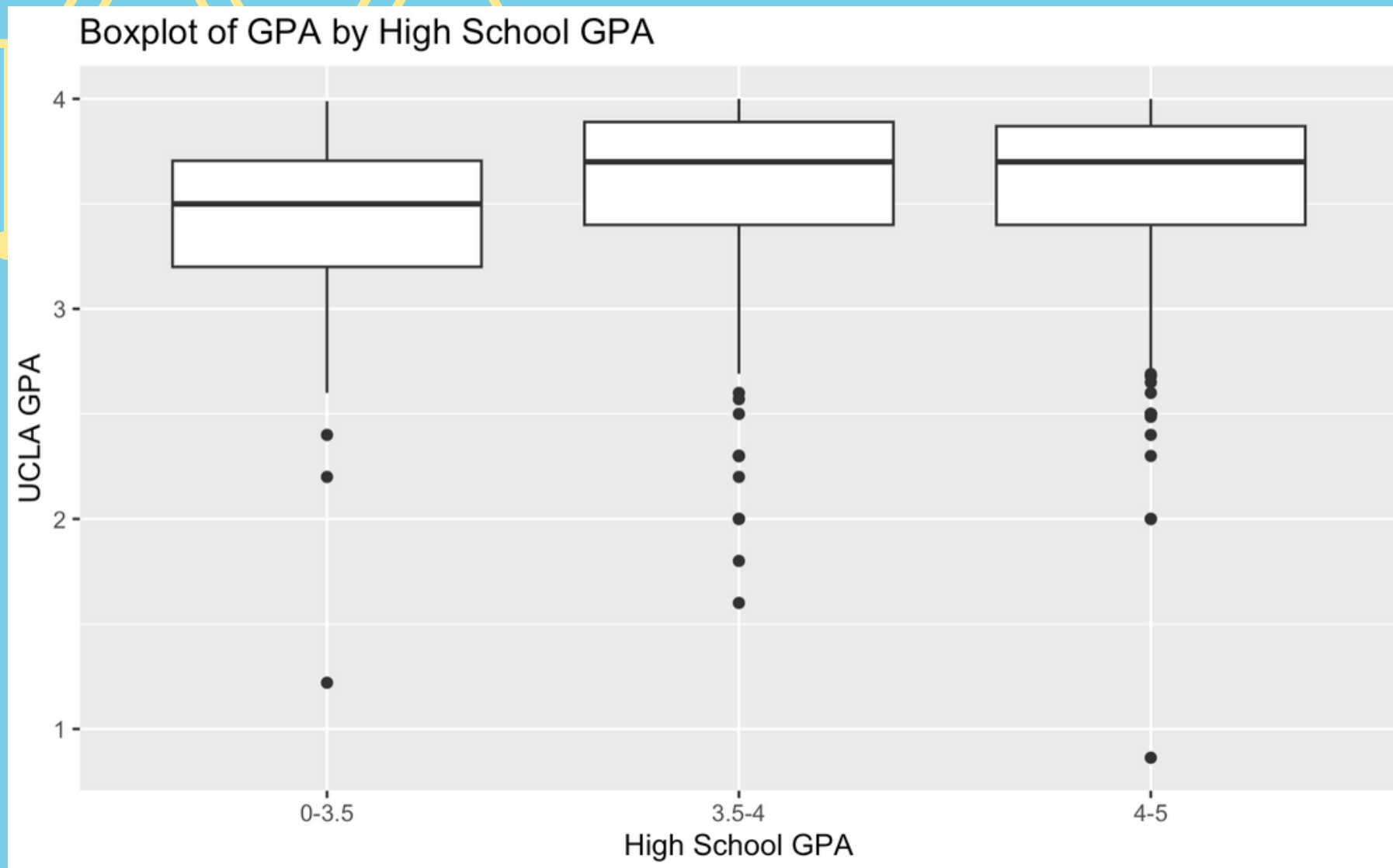
**To what extent can we predict a UCLA student's GPA from the following factors:**

1. Personal - Gender, Ethnic background, Number of friends (other ethnicity + own ethnicity), Family financial background
  2. Academic satisfaction - High school GPA, Academic anxiety
  3. Psychological - Miss class, Sleep
- 

# ROADMAP



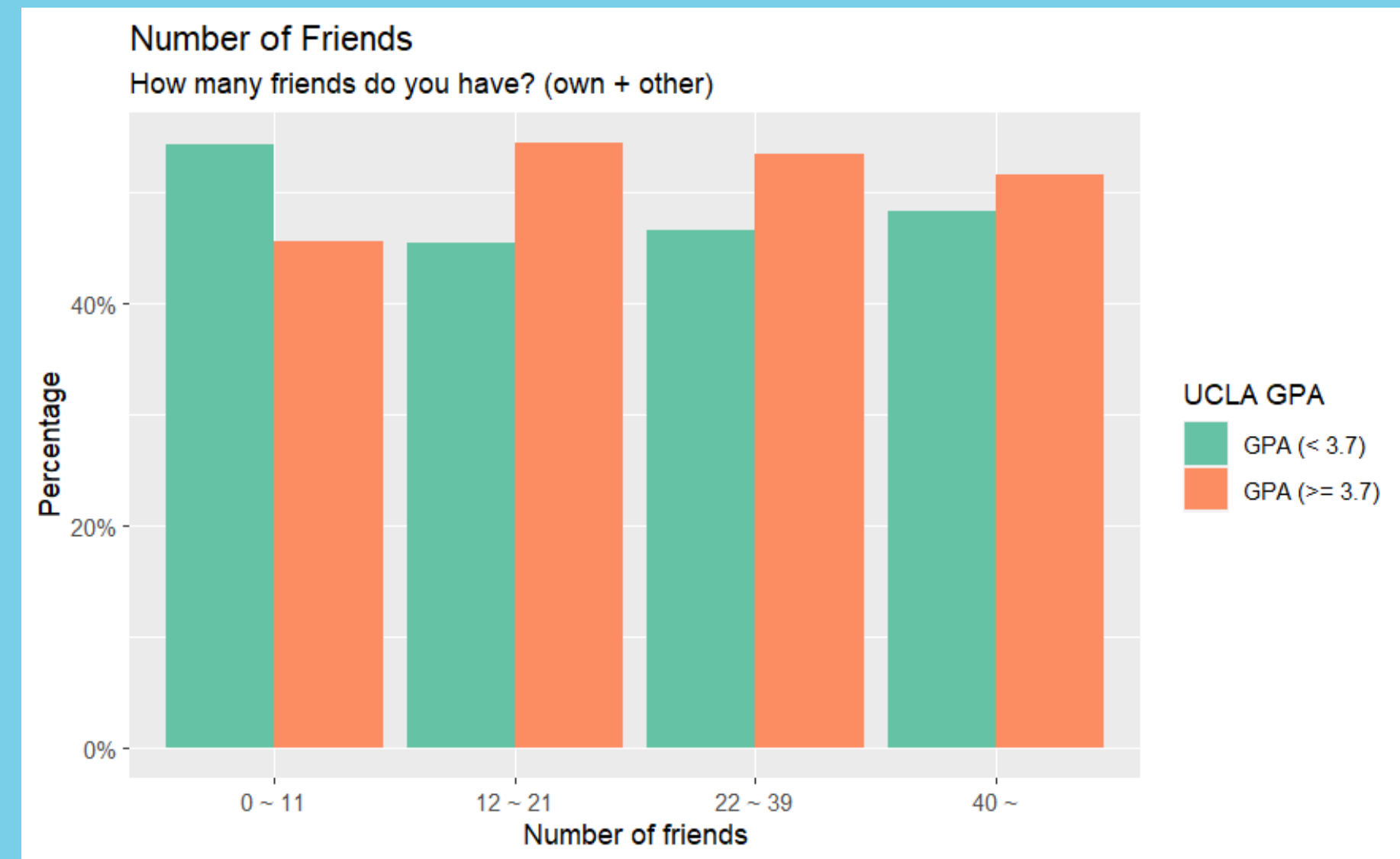
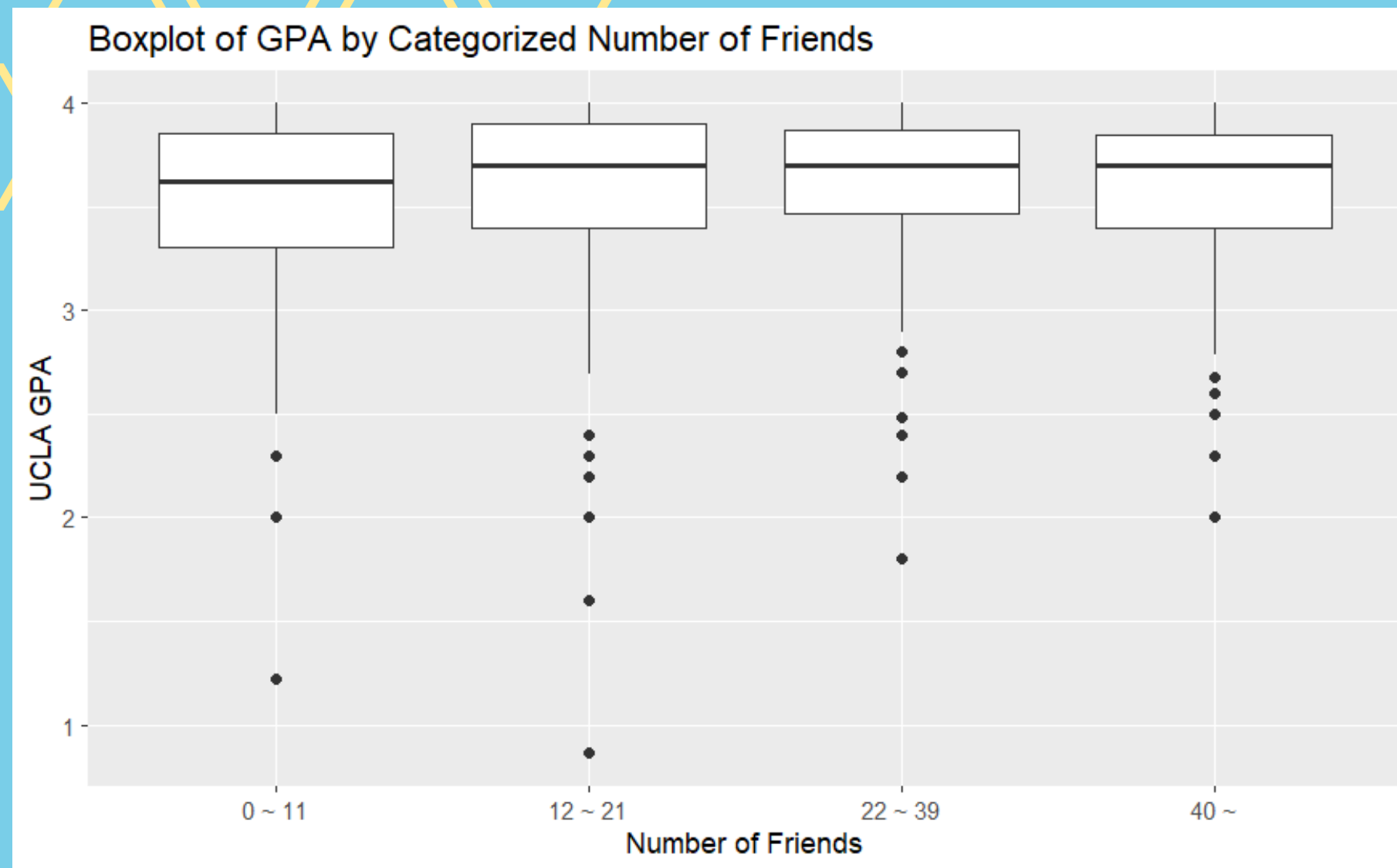
# HIGH SCHOOL GPA



- The median GPA shown in the box plots remains the same for students with a high school GPA of 3.5-5, unlike students with a high school GPA of 0-3.5
- Students with a high school GPA of 0-3.5 significantly have a higher percentage of GPAs below the median
- This shows that a student's high school GPA has a probable influence on their UCLA GPA

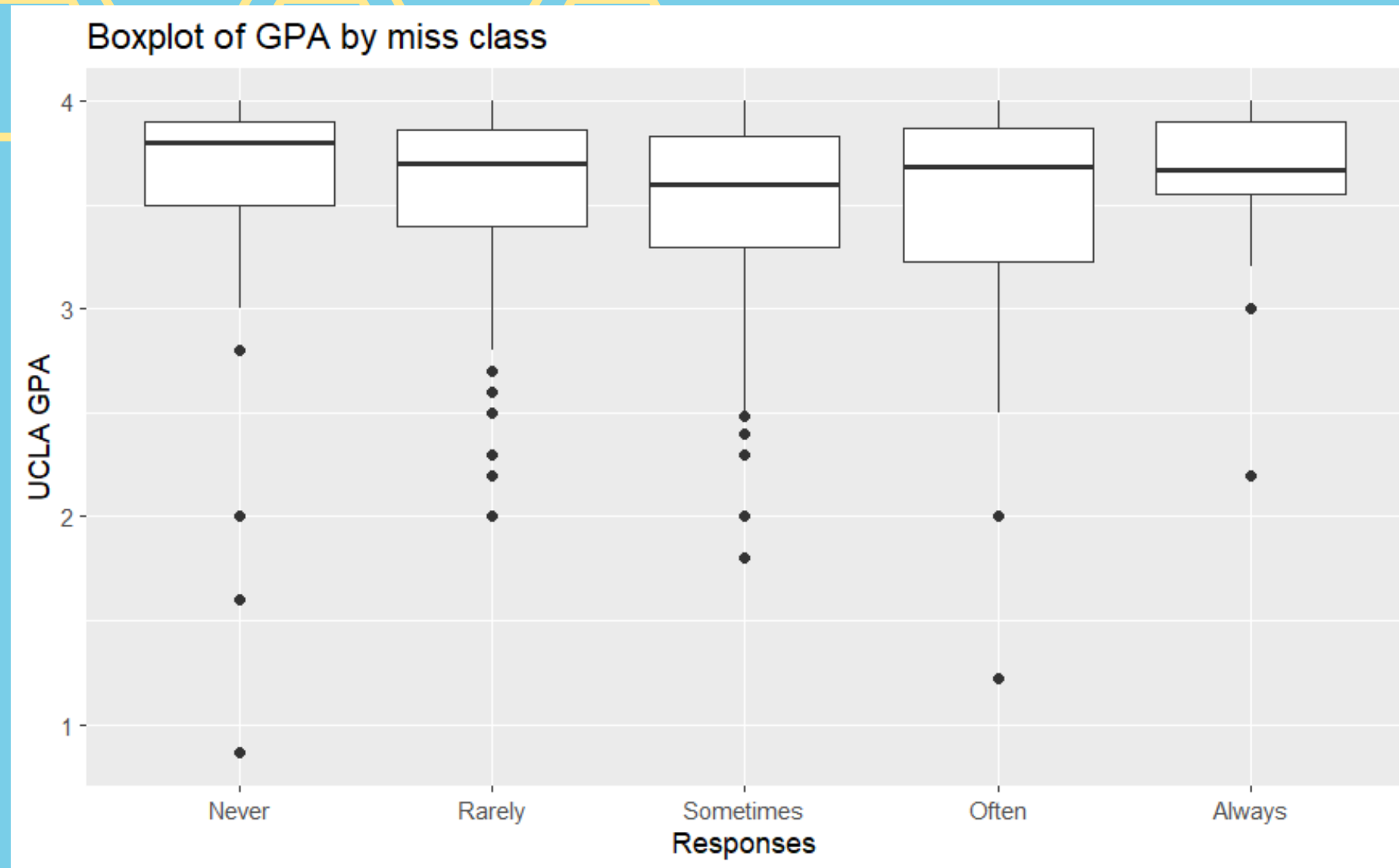


# NUMBER OF FRIENDS



- This variable represents the total count of friends, combining those from one's own ethnic group with those from other ethnicities.
- Students with a small circle of friends (0-11) show a higher percentage with GPAs below the median (3.7).
- A marked balance is observed among students with a large network of friends (40+), with a nearly equal distribution of GPAs above and below the median.
- The median GPA depicted in the boxplot remains relatively stable across all categories of the number of friends.
- A larger number of friends does not necessarily correlate with a higher GPA, as the spread of GPA scores is broad across all friend groups.

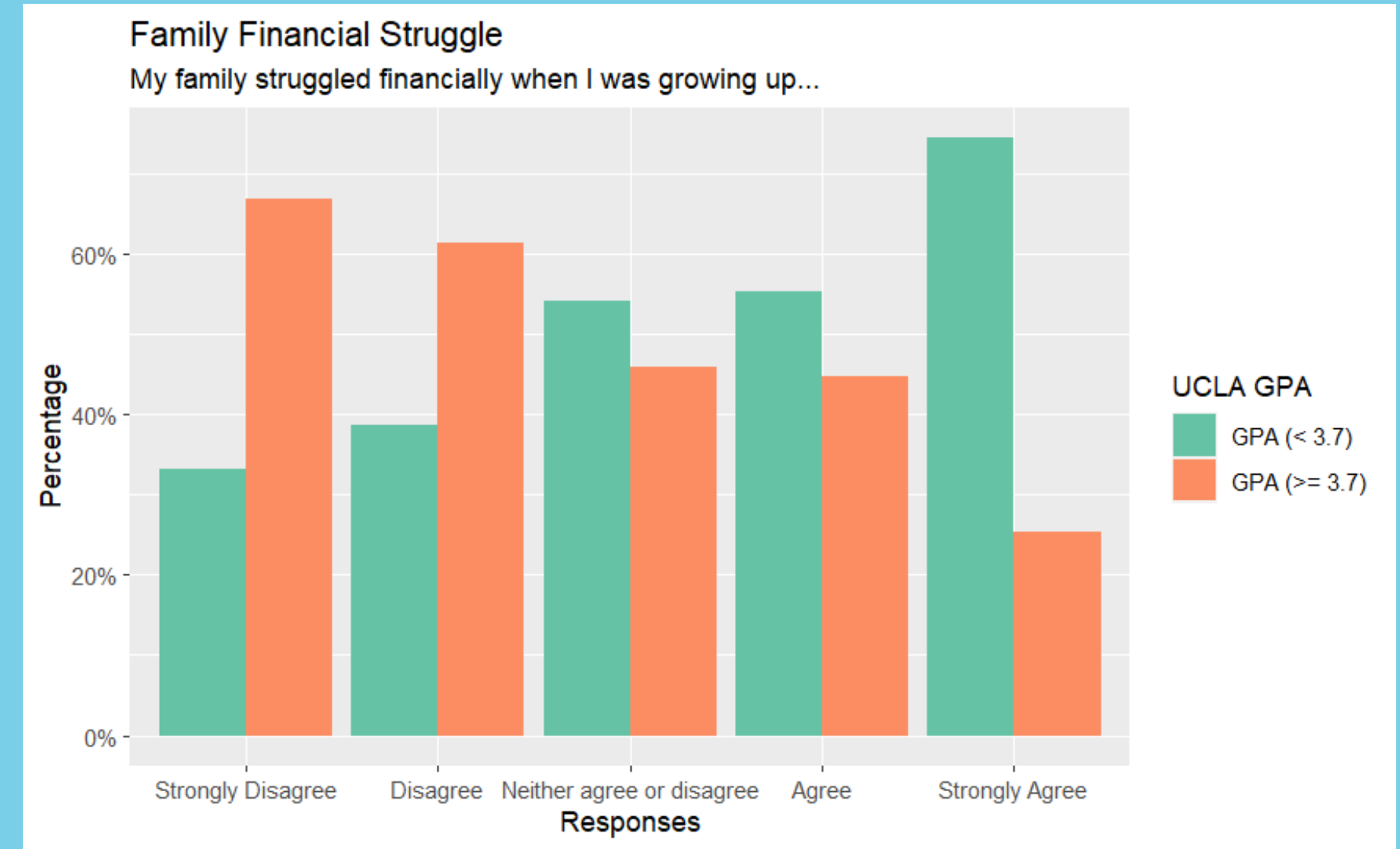
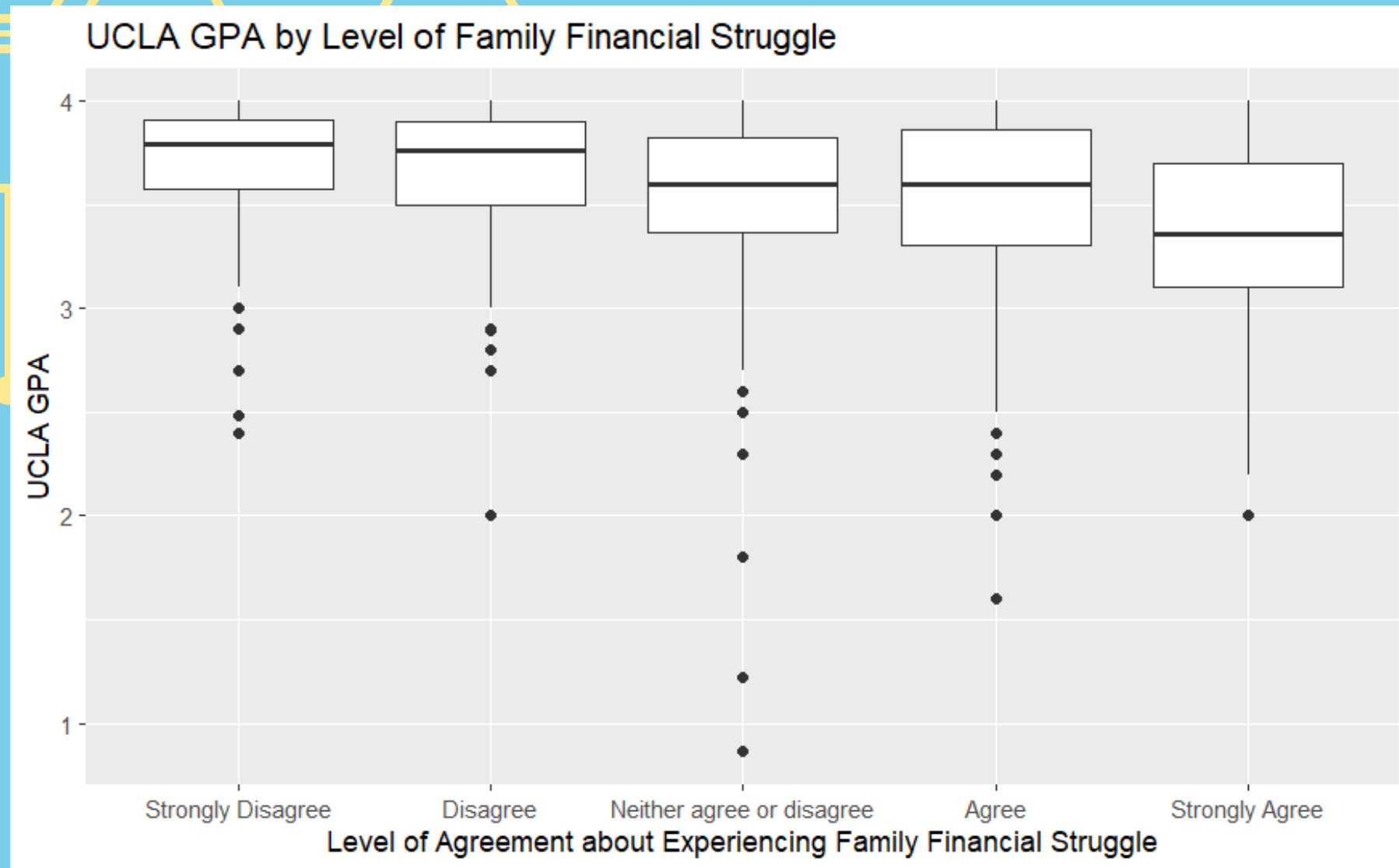
# MISS CLASS



- Boxplot reveals a trend where students with less frequent class absences generally have higher median GPAs.
- The bar graph displays a consistent pattern where students with lower class attendance have a smaller proportion of GPAs above 3.7.
- The 'Never' category shows a markedly higher proportion of students with GPAs above 3.7 compared to other categories.
- The frequency of class absences is not strongly correlated with the GPA except for the 'Never' response.

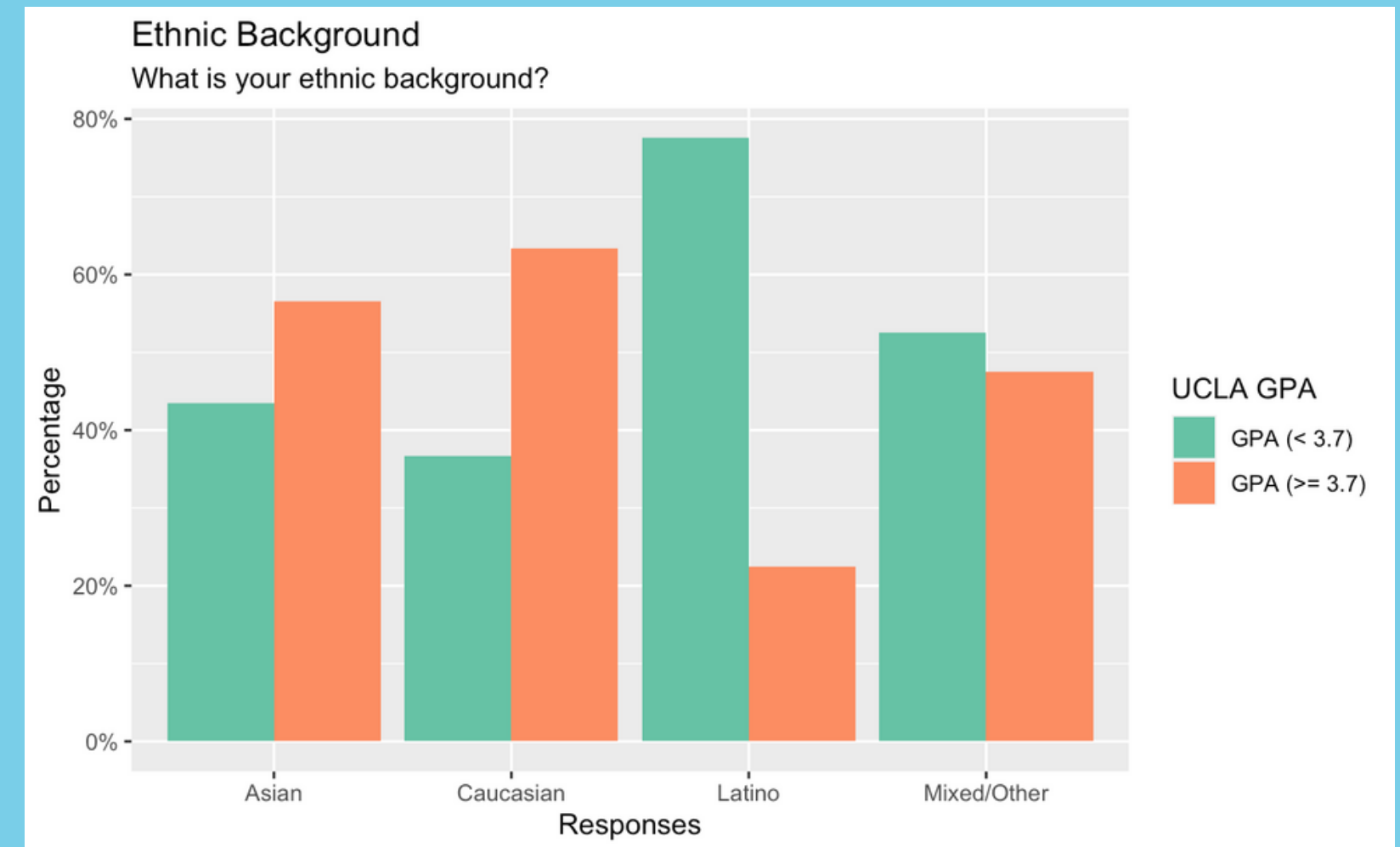
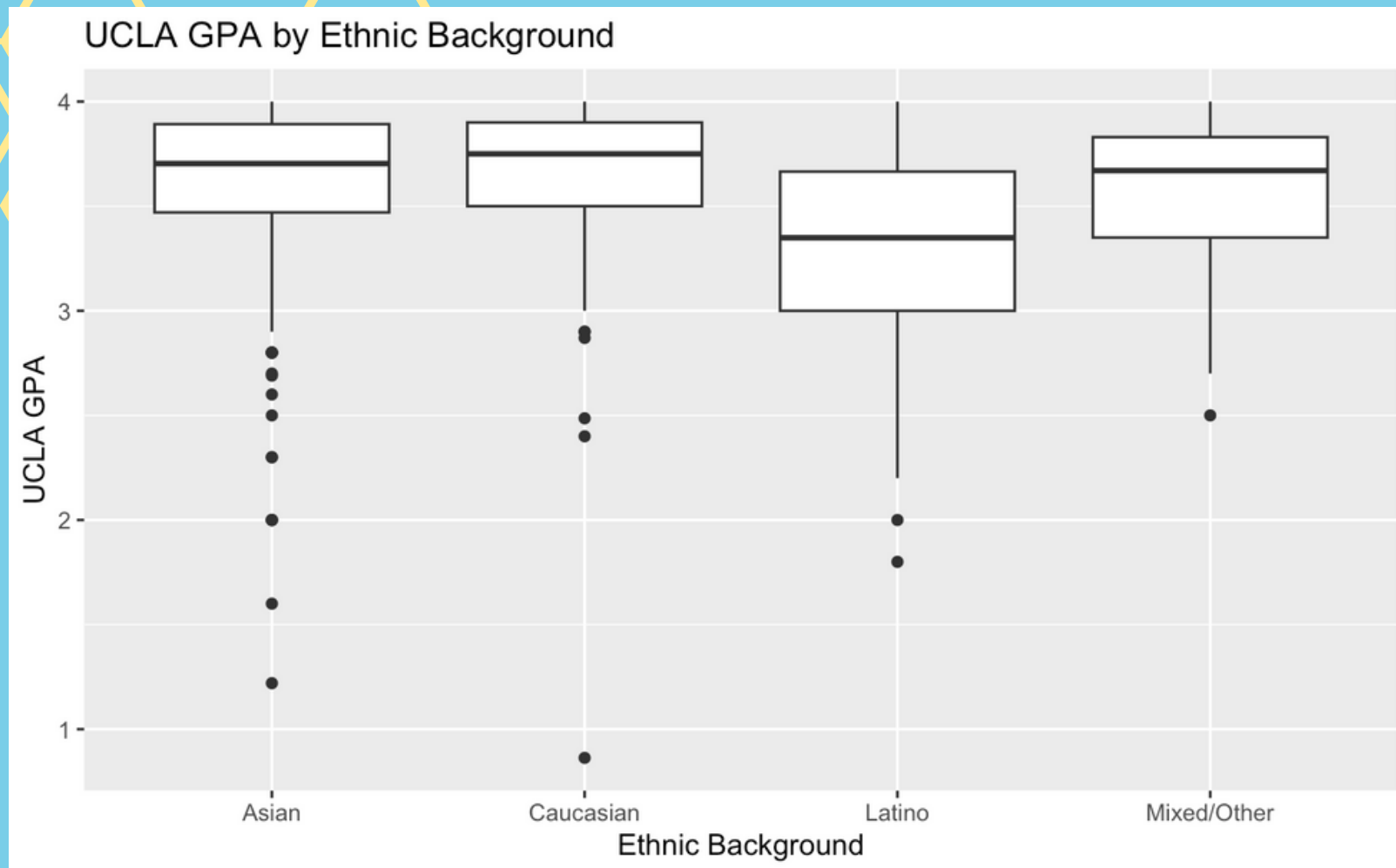


# FAMILY FINANCIAL STRUGGLE



- Students who report no family financial struggle (strongly disagree) tend to have higher median GPAs.
- There is a noticeable disparity in median GPA between students who strongly agree to experiencing family financial struggle and those who don't.
- For those who strongly agree with having experienced financial struggle, there's a greater proportion with below median GPAs.
- In general, having family financial struggle is moderately correlated with a lower GPA.

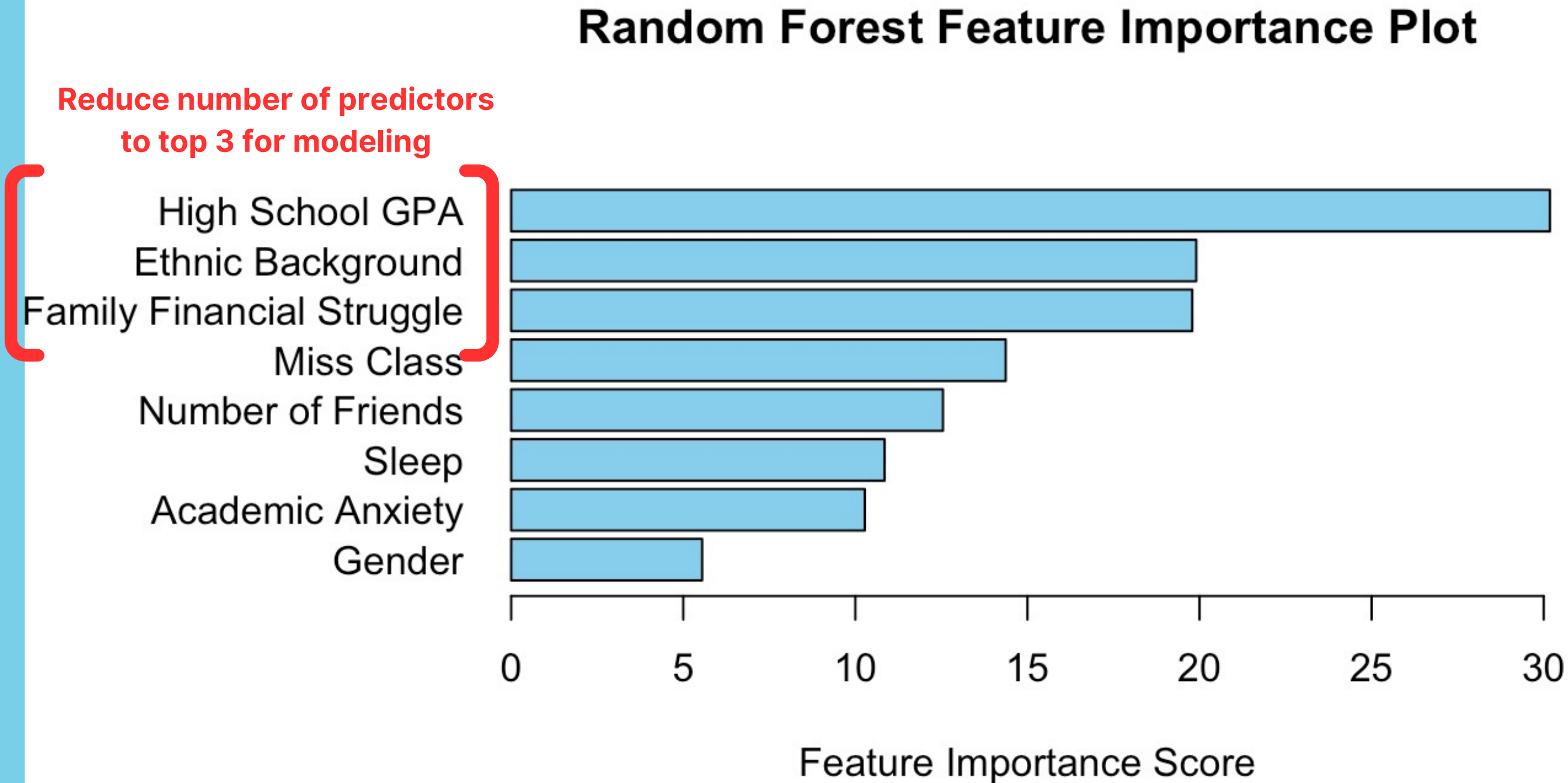
# ETHNIC BACKGROUND



- Caucasian students have the highest median UCLA GPA.
- There exists a relatively large difference between the median UCLA GPA of Latino students and the median UCLA GPA of students from other ethnic backgrounds.
- Among Asian and Caucasian students, there are more students with above median (3.7) UCLA GPA.
- Among Latino students, there are way more students with below median UCLA GPA.
- Among students identifying as Mixed/Other ethnicity, there is a nearly equal distribution between those with above or below median UCLA GPA. However, the number of students with below median UCLA GPA is slightly higher.
- Ethnic background is likely to have an influence on UCLA GPA.

# FEATURE IMPORTANCE

## RANDOM FOREST



# REFLECTION ON IMPORTANT FEATURES

## High School GPA

- most significant predictor
- strong influence of prior academic performance
- one's preparatory performance is a strong factor in how well they will perform in higher education
  - some high schools are more funded than others, providing much more academic resources for students

## Ethnic Background

- second most significant predictor
  - cultural differences in educational approaches
  - socio-economic disparities
  - varying levels of support systems
- systemic challenges and biases in educational systems may disproportionately affect certain ethnic groups

## Family Financial Struggle

- third most significant predictor
- correlation between economic conditions and academic success
- example: working a job could take focus away from academics

# LOGISTIC REGRESSION MODEL

1138 samples

3 predictors: High School GPA, Ethnic Background, Family Financial Struggle

2 classes: 'GPA (< 3.7)', 'GPA (>= 3.7)'

Resampling method: 10-fold Cross-Validation

- The dataset is divided into 10 equal folds.
- The model is trained and evaluated 10 times.
- In each iteration, one of the 10 folds is used as the test set, and the remaining nine folds are used for training.
- The model's performance is recorded for each iteration. The final performance metric is often the average of the metrics obtained in each iteration.

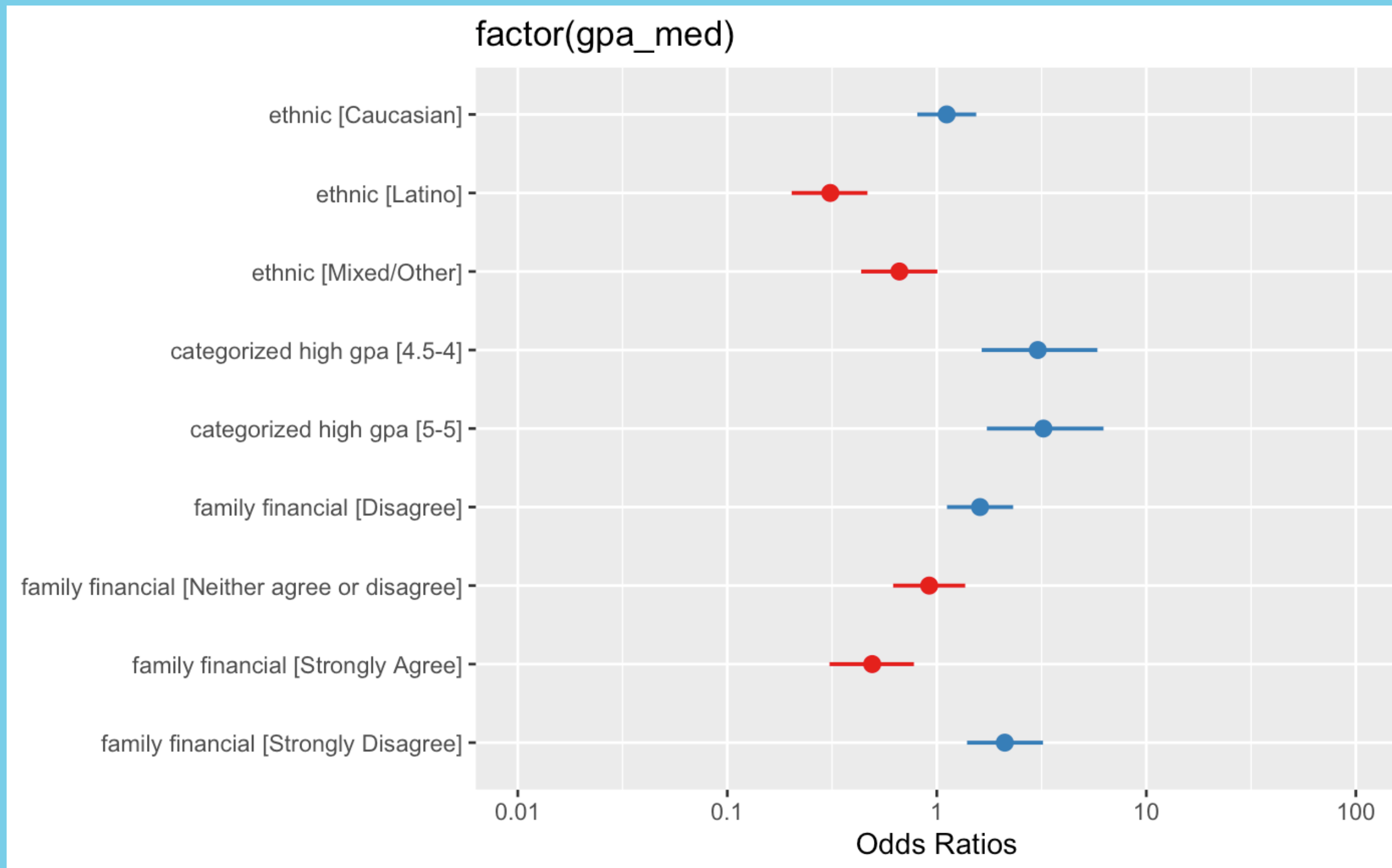
Testing performance of the model:

**Accuracy: 0.6265433**



# ODDS RATIO PLOT

## BASED ON LOGISTIC REGRESSION MODEL



**Reference: Asian, low high school gpa [0-3.5], family financial [Agree]**

**The odds of having UCLA gpa above the median is lower than reference**

- Latino
- Mixed/Other
- strongly agree

**The odds of having UCLA gpa above the median is higher than reference**

- Medium/High high school gpa
- Disagree
- Strongly disagree



# TEXT MINING

- **Students answered:**
  - Please describe the factors that have helped you succeed academically, socially, emotionally, or otherwise at UCLA.
  - Please describe the factors that have inhibited your success academically, socially, emotionally, or otherwise at UCLA.
- **Separated the answers by:**
  - students with GPA below the median (low gpa)
  - students with GPA above the median (high gpa)
- **Analyze the differences in these students responses**
  - word counts
  - word clouds
  - text networks

← **HELP FACTOR**

← **HURT FACTOR**



# TEXT MINING

## TOP TEN WORDS FOR HELP FACTOR

Top Ten Words for Low GPA Help Factor

Word	Frequency
1. Friends	317
2. People	151
3. Support	124
4. Time	94
5. Life	86
6. <u>Classes</u>	85
7. School	84
8. Family	77
9. Study	76
10. <u>Academic</u>	73

Academic theme  
↓  
value friends support  
in their classes

LOW GPA

Top Ten Words for High GPA Help Factor

Word	Frequency
1. Friends	258
2. People	108
3. Support	98
4. School	94
5. Family	65
6. <u>Clubs</u>	64
7. <u>Social</u>	63
8. Time	63
9. Life	57
10. Study	57

Social theme  
↓  
value friends support  
outside of class

HIGH GPA

# TEXT MINING

# WORD CLOUDS FOR HELP FACTOR



# LOW GPA



# HIGH GPA

# TEXT MINING

## TOP TEN WORDS FOR HURT FACTOR

Top Ten Words for Low GPA Hurt Factor

Word	Frequency	Struggle with academic rigor
1. Classes	161	
2. Time	120	
3. School	112	
4. Hard	89	
5. People	87	
6. Students	81	
7. Social	71	
8. Class	67	
9. Anxiety	63	
10. Quarter	61	

**LOW GPA**

Top Ten Words for High GPA Hurt Factor

Word	Frequency	Struggle with school/life balance
1. Time	91	
2. People	85	
3. Classes	78	
4. Students	65	
5. School	62	
6. Social	60	
7. Friends	51	
8. Difficult	50	
9. Stress	50	
10. Class	49	

**HIGH GPA**

# TEXT MINING

# WORD CLOUDS FOR HURT FACTOR



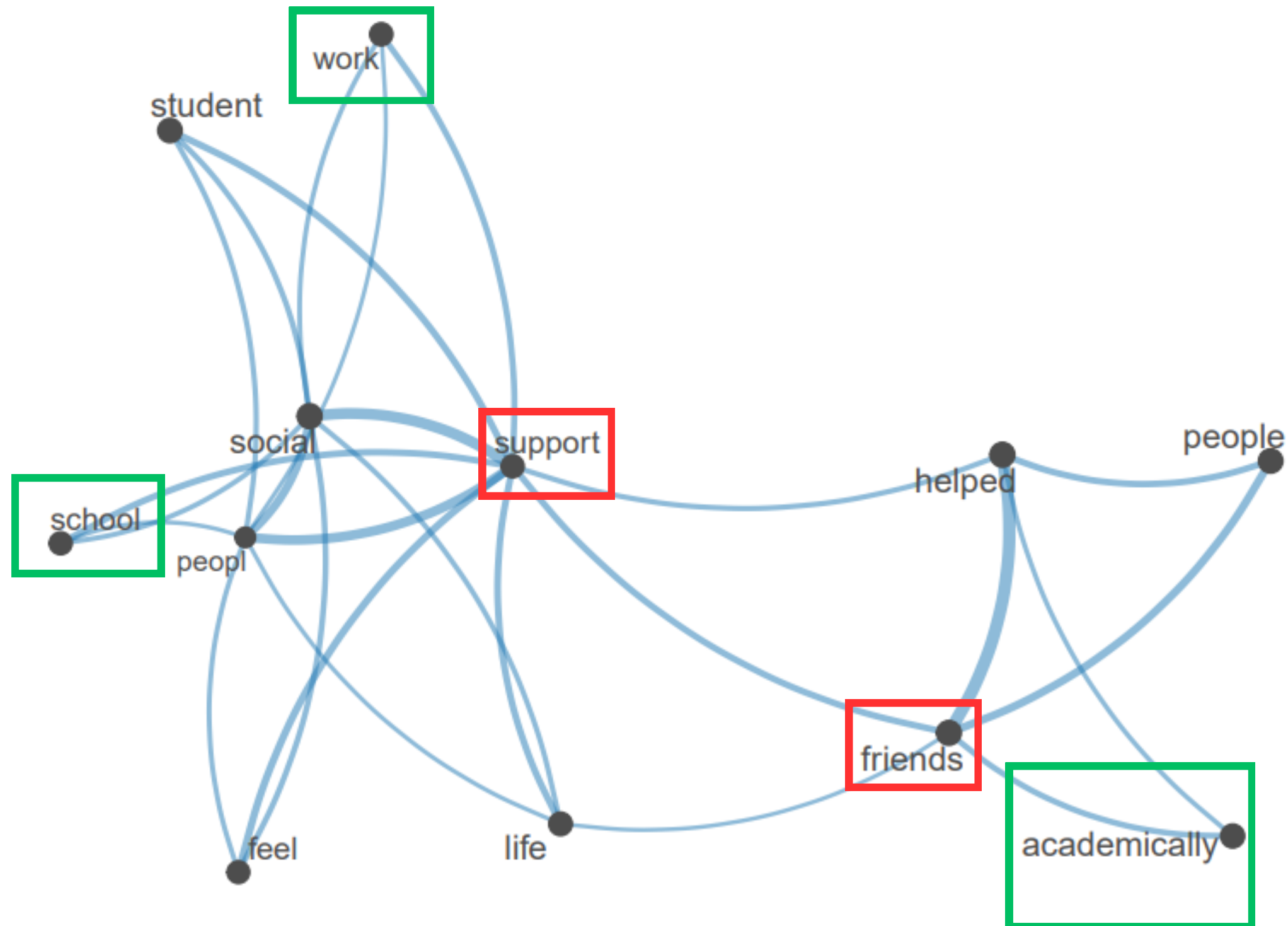
# LOW GPA



# HIGH GPA

# TEXT MINING

## TEXT NETWORKS FOR HELP FACTOR

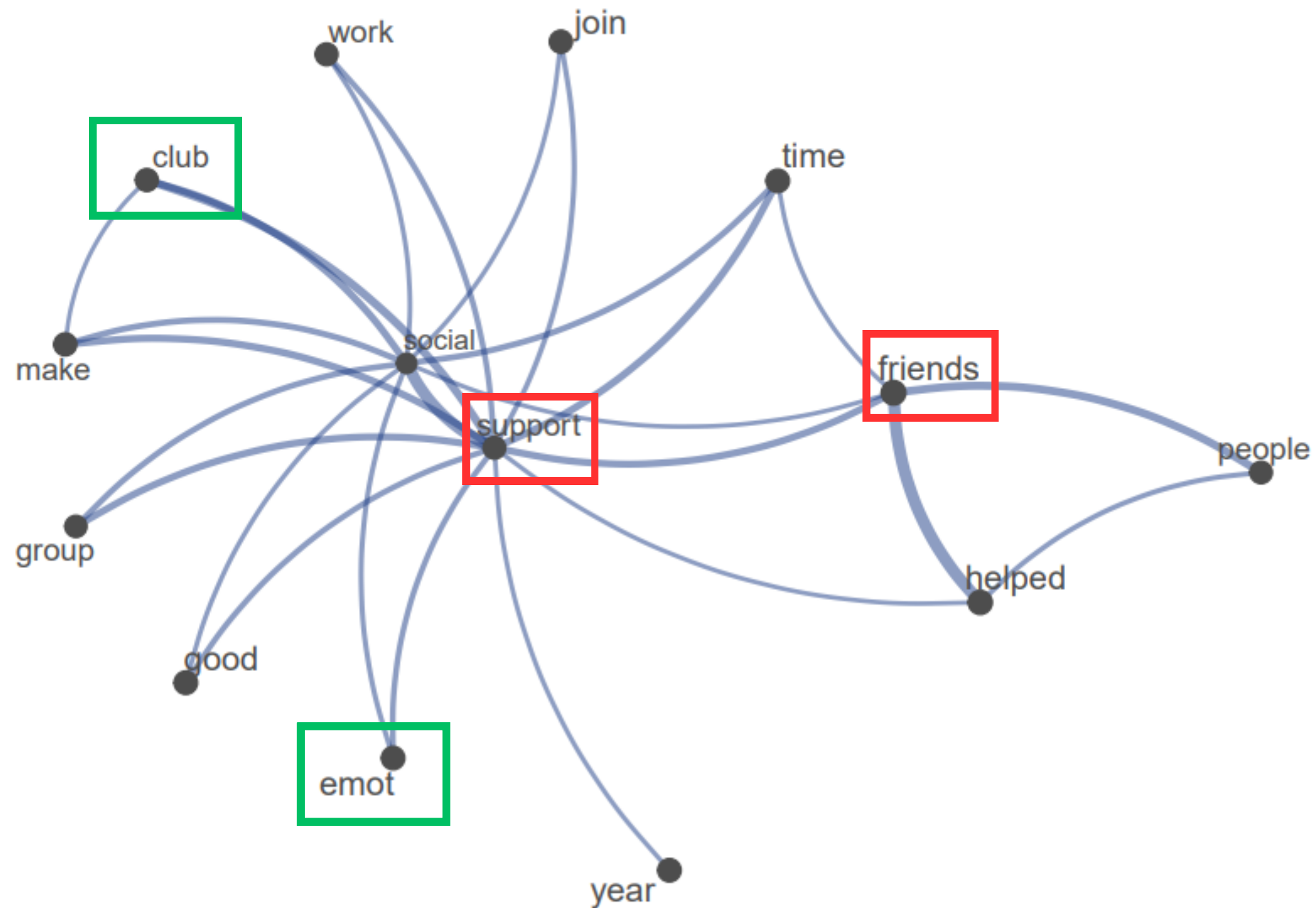


LOW GPA



# TEXT MINING

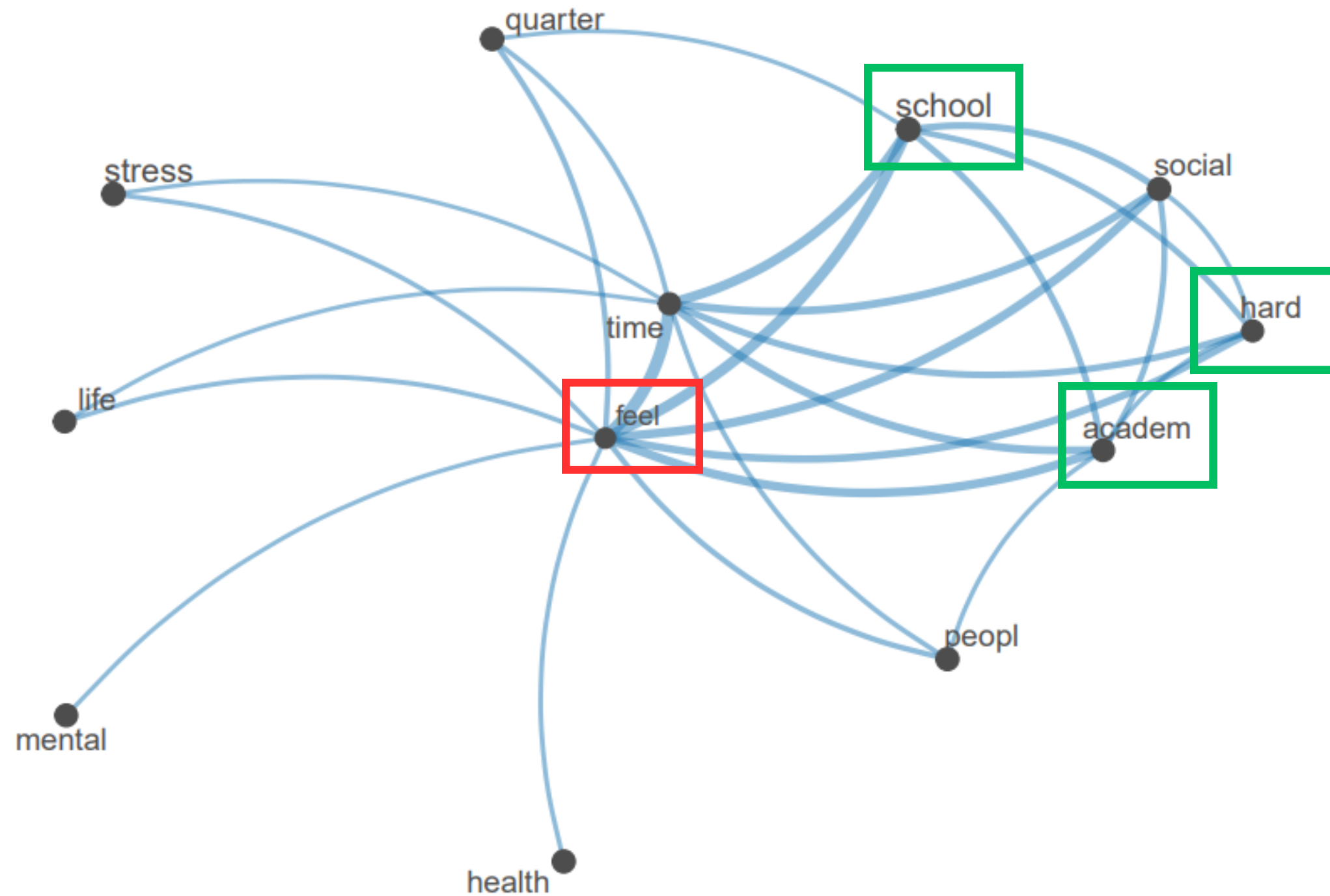
## TEXT NETWORKS FOR HELP FACTOR



HIGH GPA

# TEXT MINING

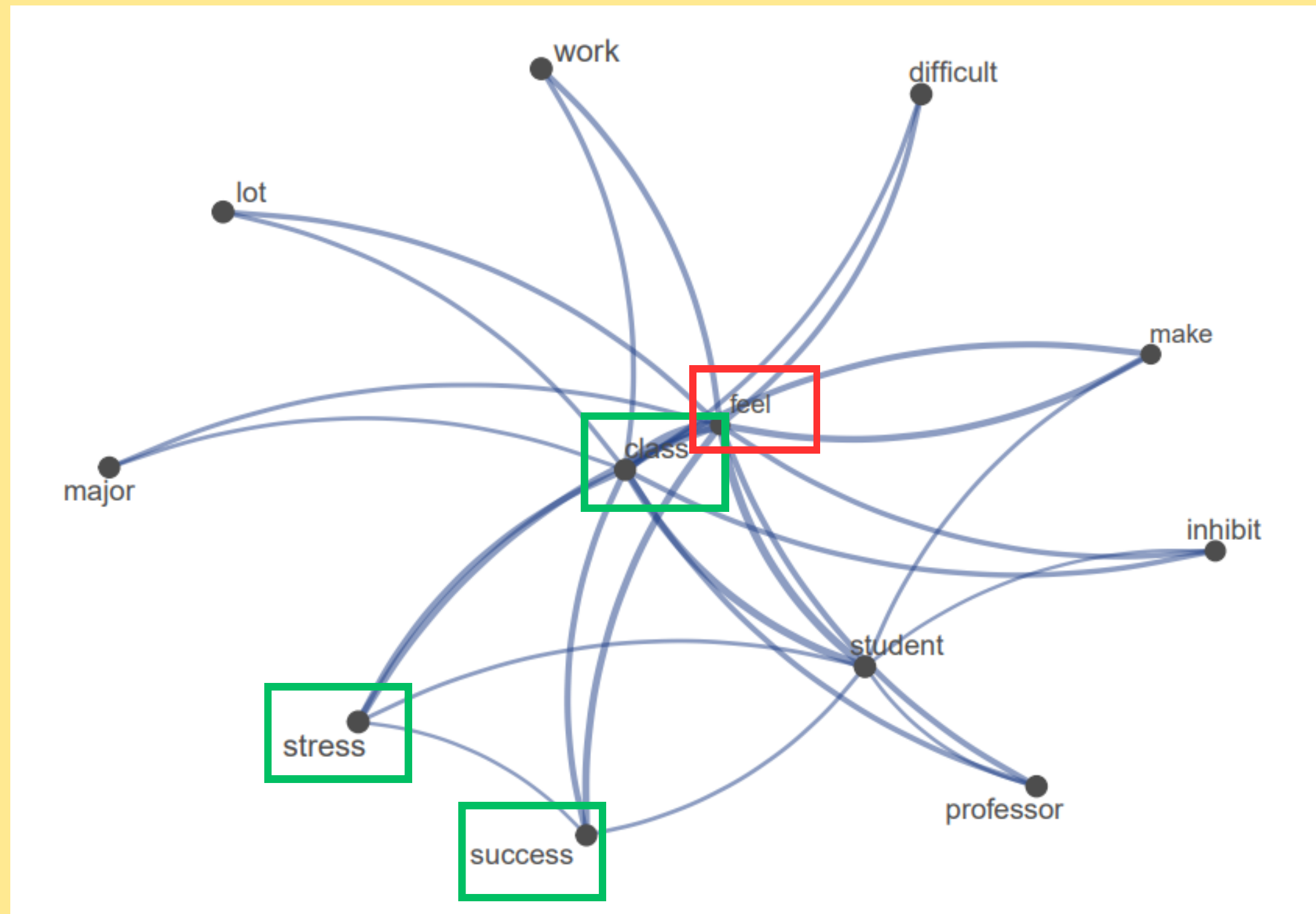
## TEXT NETWORKS FOR HURT FACTOR



LOW GPA

# TEXT MINING

## TEXT NETWORKS FOR HURT FACTOR



HIGH GPA

# TEXT MINING

## SUMMARY OF RESULTS

### LOW GPA

- **support from friends is most important in academics**
- **likely struggle with rigor of UCLA academics**

### HIGH GPA

- **support from friends is most important socially**
- **possibly struggles with school/social life balance**

**Both groups of students place large importance on academics and deeply value the support of friends**

# CONCLUSIONS AND RECOMMENDATIONS

Based on our analyses of the dataset we found that A UCLA student's high school GPA, ethnic background, and family financial struggle had the most influence on their current college GPA. We also found that all UCLA students tend to perceive academics as highly important, while those with low GPA valued academic support from friends and high GPA students valued friendships focused outside of school.

Future studies could explore alternative statistical methods or consider more sophisticated techniques to handle numeric variables without sacrificing their continuous nature. This would provide a more comprehensive understanding of the quantitative aspects of the relationships under investigation.

# **LIMITATIONS**

- **Oversimplification of Numeric Data:**

- Discretizing continuous variables may lead to nuanced information.
- This may potentially impact the accuracy and depth of conclusions.

- **Small Sample Sizes in Certain Categories:**

- Some categories had notably small sample sizes relative to others.
- Findings for these categories may not accurately represent the diversity within the entire population.

- **Loss of Observations:**

- Some observations were lost due to non-responses to questions or the inclusion of values that did not make sense.
- The missing data may introduce bias and affect the completeness of the analysis.



The background is a light blue-grey color, densely populated with various science-related icons in a darker blue-grey color. These icons include laboratory equipment like beakers, flasks, and test tubes; biological elements like a cell, a virus, and a DNA helix; and environmental symbols like a cloud with rain, a sun, and water waves. The icons are scattered across the entire background, creating a thematic pattern.

**THANK YOU  
VERY MUCH!**