Stats 101A - Final Project

Taro Iyadomi, (fill in)

1. Introduction

Question: What factors affect college students' class performance, and how do they do so?

The goal of this project is to understand what factors affect college students' performance in class as well as understanding the relationships between those factors and the class performance metric, GPA. These insights can help us find any discrepancies in student learning due to socioeconomic and social identity differences, which we can take action on to help improve the quality of education for all students.

2. Exploratory Data Analysis

```
## Reading in data
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.5
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
data <- read.csv("diversity-2.csv", stringsAsFactors=T)

## Shape of our data
data %>% dim()
```

```
## [1] 939  81
```

```
## Types of variables
variable_types <- sapply(data, class)
convert_name <- c(character = "Categorical", numeric = "Quantitative")
variable_types <- convert_name[variable_types]
variable_types %>% table()
```

```
## .
## Quantitative
##            6
```

```
## We can remove variables with >20% missing values. For the rest, we can just remove NA observations.
library(VIM)
```

Understanding Missing Values

```
## Warning: package 'VIM' was built under R version 4.2.2
```

```
## Loading required package: colorspace
```

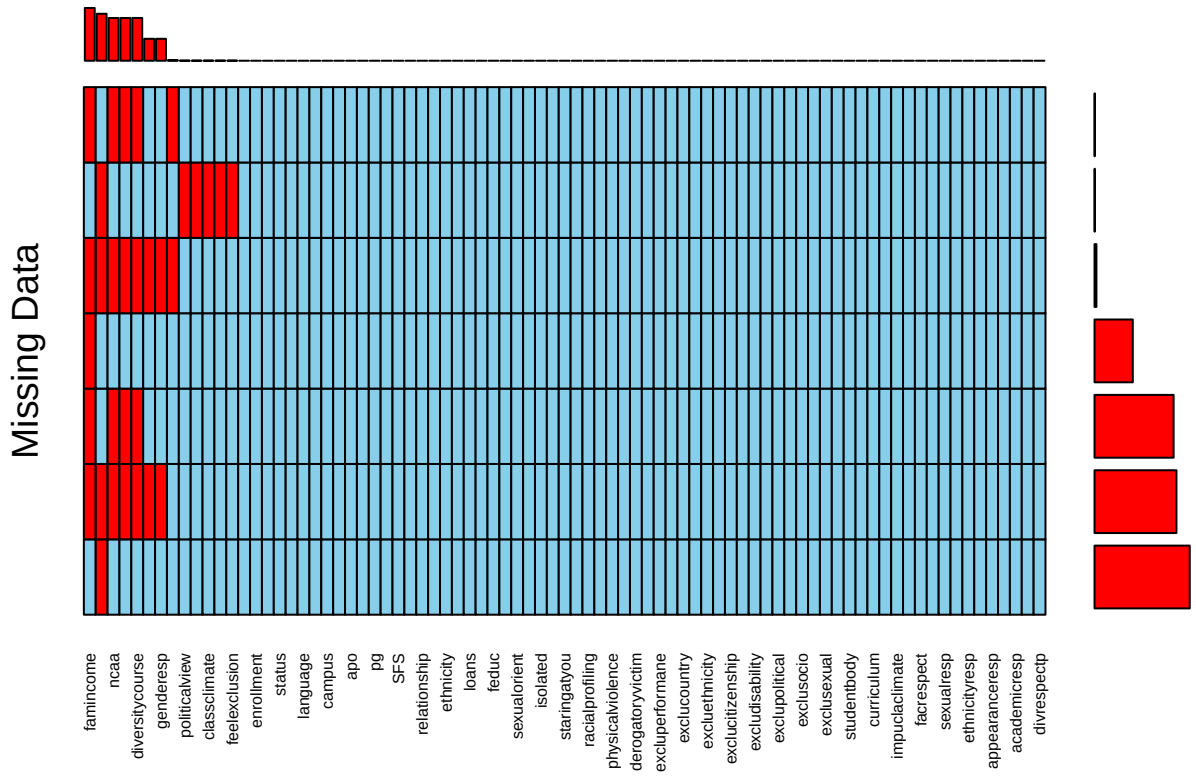
```
## Loading required package: grid
```

```
## VIM is ready to use.
```

```
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
```

```
##
## Attaching package: 'VIM'

## The following object is masked from 'package:datasets':
##
##     sleep
```

```
NA_plot <- aggr(data, labels=names(data), cex.axis=0.5, ylab=c("Missing Data", "Pattern"), combined=T, s
```




```
##
##  Variables sorted by number of missings:
##            Variable Count
##           famincome   637
##      diversityclass   567
##                ncaa   516
##     obsexclusionary   516
##     diversitycourse   516
##      admiunderstand   265
##           genderesp   265
##                year     7
##       politicalview     1
##         uclaclimate     1
##        classclimate     1
##         leavingucla     1
##       feelexclusion     1
```

```
##              Course     0
##          enrollment     0
##            transfer     0
##              status     0
##                 gpa     0
##            language     0
##          discipline     0
##              campus     0
##         participate     0
##                 apo     0
##               imccg     0
##                  pg     0
##                 RSO     0
##                 SFS     0
##      participationp     0
##        relationship     0
##              gender     0
##           ethnicity     0
##        financialaid     0
##               loans     0
##               meduc     0
##               feduc     0
##            socioeco     0
##        sexualorient     0
##            religion     0
##            isolated     0
##         intimidated     0
##        staringatyou     0
##              feared     0
##     racialprofiling     0
##         crimevictim     0
##    physicalviolence     0
##             stalking     0
##    derogatoryvictim     0
##           ucladiscp     0
##     excluperformane     0
##             excluage     0
##        exclucountry     0
##        excluenglish     0
##      excluethnicity     0
##            exclurace     0
##    exclucitizenship     0
##         exclumental     0
##     excludisability     0
##  excluparticipation     0
##      exclupolitical     0
##       exclureligion     0
##          exclusocio     0
##         exclugender     0
##         exclusexual     0
##   uclaexclusionaryp     0
##         studentbody     0
##       crosscultural     0
##          curriculum     0
```

```
##        facultydiver     0
##      impuclaclimate     0
##       facunderstand     0
##          facrespect     0
##            channels     0
##          sexualresp     0
##         countryresp     0
##       ethnicityresp     0
##        religionresp     0
##      appearanceresp     0
##           socioresp     0
##        academicresp     0
##       politicalresp     0
##         divrespectp     0
```

```
NA_df <- NA_plot$missings %>% filter(Count > 0)

predictor_classes <- data.frame("Variable"=names(data), "Type" = sapply(data, class))

NA_df <- NA_df %>%
  inner_join(predictor_classes, by="Variable") %>%
  mutate(Percent_Missing = Count / nrow(data) * 100) %>%
  arrange(desc(Count)) %>%
  print()
```

```
##           Variable Count    Type Percent_Missing
## 1        famincome   637 numeric      67.8381257
## 2   diversityclass   567 integer      60.3833866
## 3             ncaa   516 integer      54.9520767
## 4  obsexclusionary   516 integer      54.9520767
## 5  diversitycourse   516 integer      54.9520767
## 6   admiunderstand   265 integer      28.2215122
## 7        genderesp   265 integer      28.2215122
## 8             year     7  factor       0.7454739
## 9    politicalview     1  factor       0.1064963
## 10     uclaclimate     1  factor       0.1064963
## 11    classclimate     1  factor       0.1064963
## 12     leavingucla     1 integer       0.1064963
## 13   feelexclusion     1  factor       0.1064963
```

```
factors_to_remove <- NA_df %>%
  filter(Percent_Missing > 20) %>%
  select(Variable) %>%
  print()
```

```
##          Variable
## 1       famincome
## 2  diversityclass
## 3            ncaa
## 4 obsexclusionary
## 5 diversitycourse
## 6  admiunderstand
## 7       genderesp
```

```r
data <- data %>%
  select(-factors_to_remove[[1]]) %>%
  na.omit()

sapply(data, function(x) sum(is.na(x)))
```

```
##             Course               year         enrollment           transfer
##                  0                  0                  0                  0
##             status                gpa           language         discipline
##                  0                  0                  0                  0
##             campus        participate                apo              imccg
##                  0                  0                  0                  0
##                 pg                RSO                SFS     participationp
##                  0                  0                  0                  0
##       relationship             gender          ethnicity       financialaid
##                  0                  0                  0                  0
##              loans              meduc              feduc          socioeco
##                  0                  0                  0                  0
##       sexualorient           religion      politicalview        uclaclimate
##                  0                  0                  0                  0
##       classclimate        leavingucla      feelexclusion           isolated
##                  0                  0                  0                  0
##        intimidated       staringatyou             feared    racialprofiling
##                  0                  0                  0                  0
##        crimevictim   physicalviolence           stalking   derogatoryvictim
##                  0                  0                  0                  0
##          ucladiscp    excluperformane           excluage       exclucountry
##                  0                  0                  0                  0
##       excluenglish     excluethnicity           exclurace   exclucitizenship
##                  0                  0                  0                  0
##        exclumental    excludisability excluparticipation     exclupolitical
##                  0                  0                  0                  0
##      exclureligion         exclusocio        exclugender        exclusexual
##                  0                  0                  0                  0
##  uclaexclusionaryp        studentbody      crosscultural         curriculum
##                  0                  0                  0                  0
##       facultydiver     impuclaclimate      facunderstand         facrespect
##                  0                  0                  0                  0
##           channels         sexualresp        countryresp      ethnicityresp
##                  0                  0                  0                  0
##       religionresp     appearanceresp          socioresp       academicresp
##                  0                  0                  0                  0
##      politicalresp        divrespectp
##                  0                  0
```

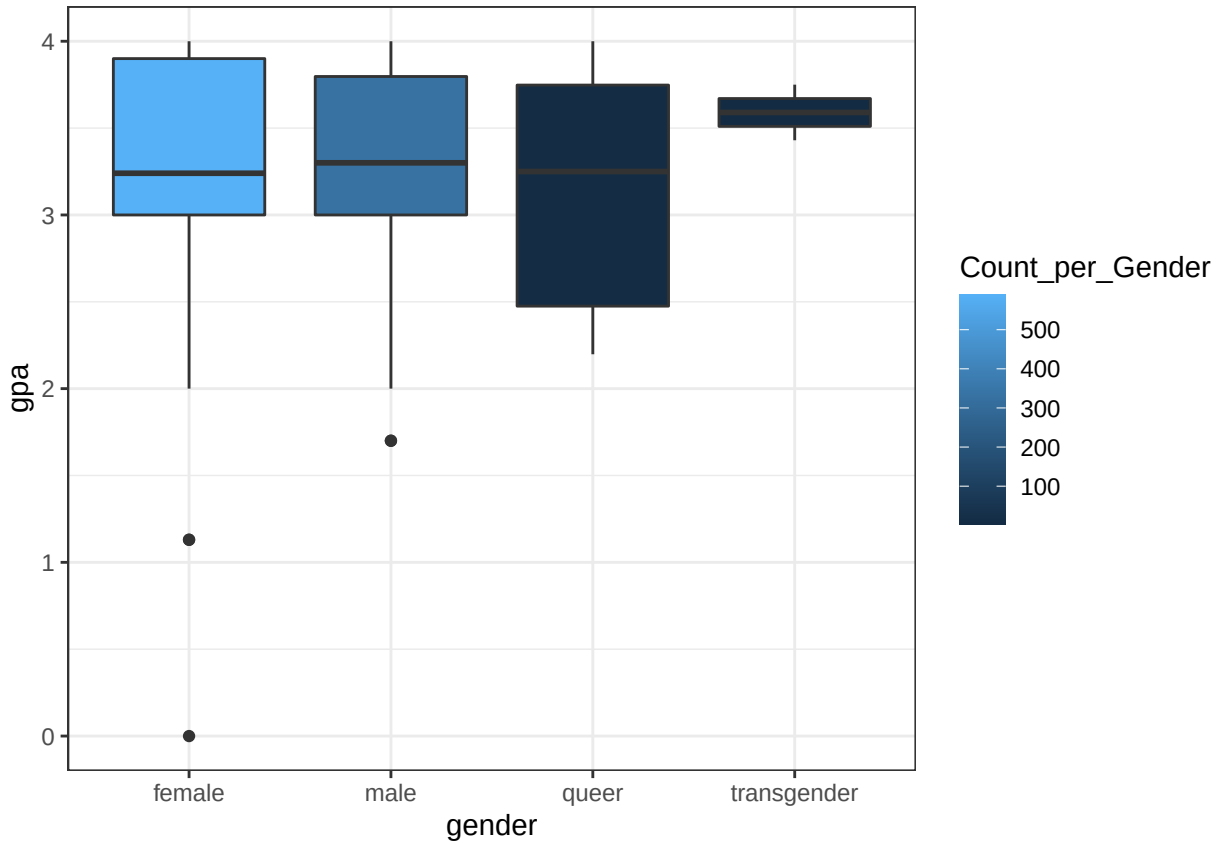
Visualizing Relationships between Suspect Factors and GPA Before any data preprocessing or modeling, we suspected a few variables to be associated with GPA based on our intuition. These variables were gender, discrimination, relationship status, and ethnicity.

```r
## GPA vs Gender
data$gender %>% levels()
```

```
## [1] "female"             "female/transgender" "genderqueer"
```

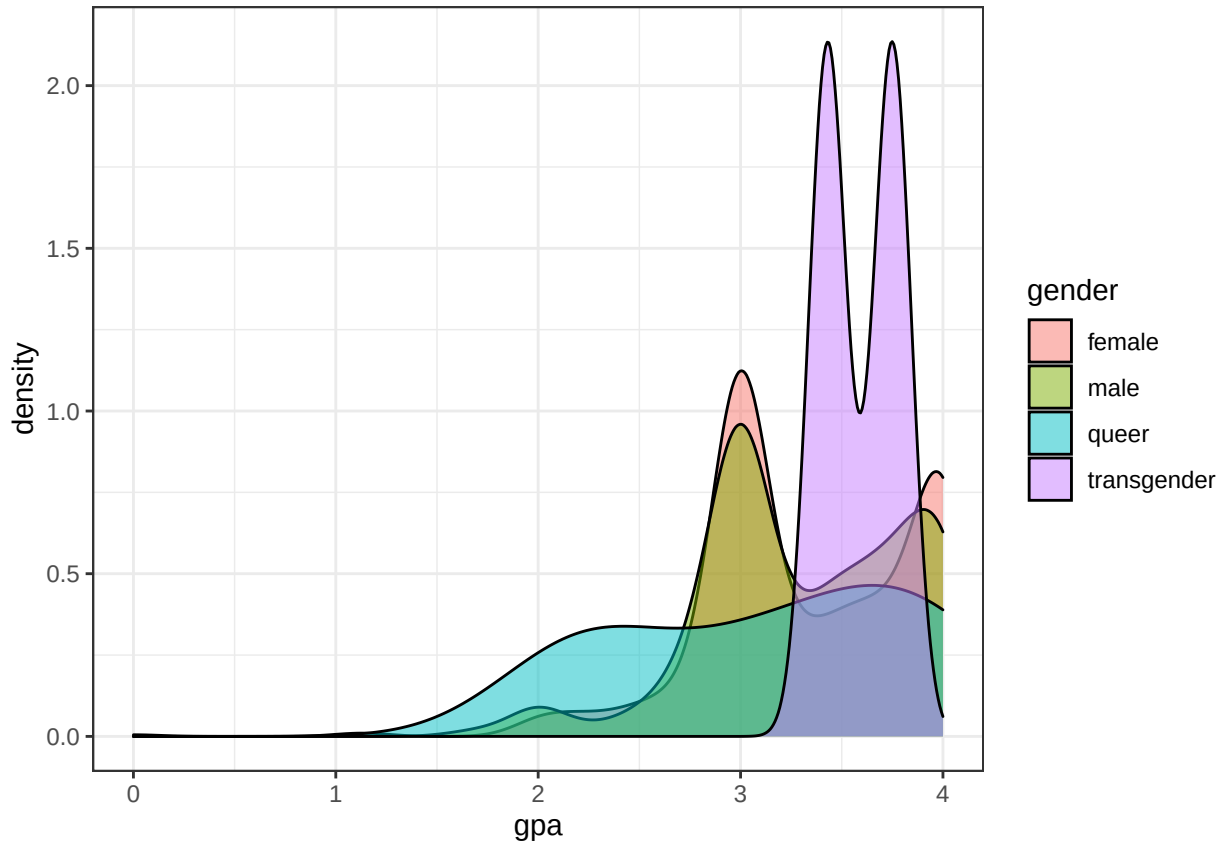
```
## [4] "male"                "other"               "queer"
## [7] "transgender"
```

```r
# Reorganizing Gender
gender_categories <- c("female" = "female",
                       "male" = "male",
                       "queer" = "queer",
                       "genderqueer" = "queer",
                       "female/transgender" = "transgender",
                       "other" = "queer",
                       "transgender" = "transgender")

data$gender <- gender_categories[as.character(data$gender)]

data %>%
  group_by(gender) %>%
  mutate(Count_per_Gender = n()) %>%
  ggplot(aes(x=gender, y=gpa)) + geom_boxplot(aes(fill=Count_per_Gender)) + theme_bw()
```



Here we can see that while the the median GPAs for female, male, and queer identifying students are relatively similar, we find that transgender students have a much higher median GPA than the other students. That being said, there aren't that many transgender students in our data, so it may not be representative for all transgender students.

```
# Density plot
data %>%
  ggplot(aes(x = gpa, fill=gender)) +
  geom_density(alpha = 0.5) +
  theme_bw()
```



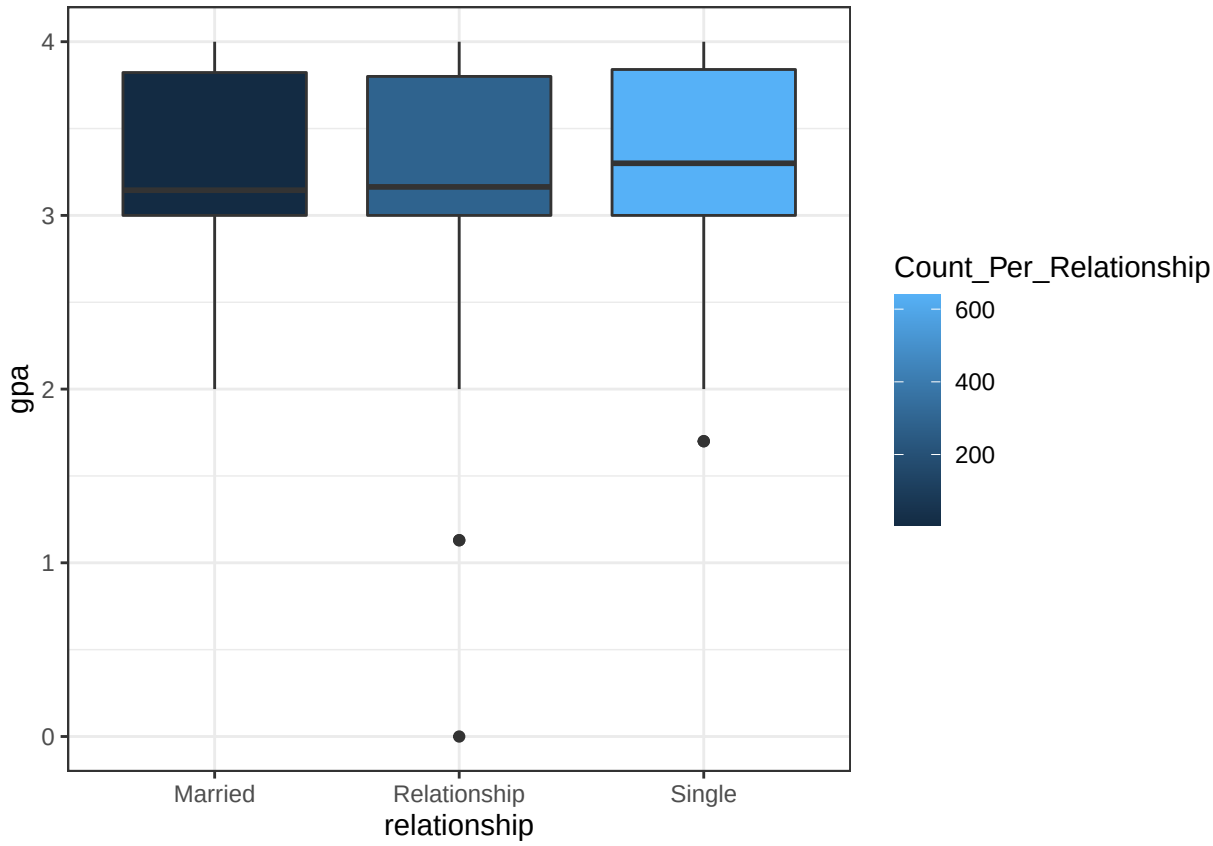
When we compare the GPA densities of each gender, we find that while male and female students have overlapping distributions, queer students show greater proportions of lower gpa students, while transgender students tend to stay in the middle of 3.0 and 4.0 gpas. They all show a bimodal distribution.

```
## Relationship Status

data$relationship %>% unique()
```

```
## [1] Single            In a relationship Other             Partnered
## [5] Married
## Levels: In a relationship Married Other Partnered Single
```

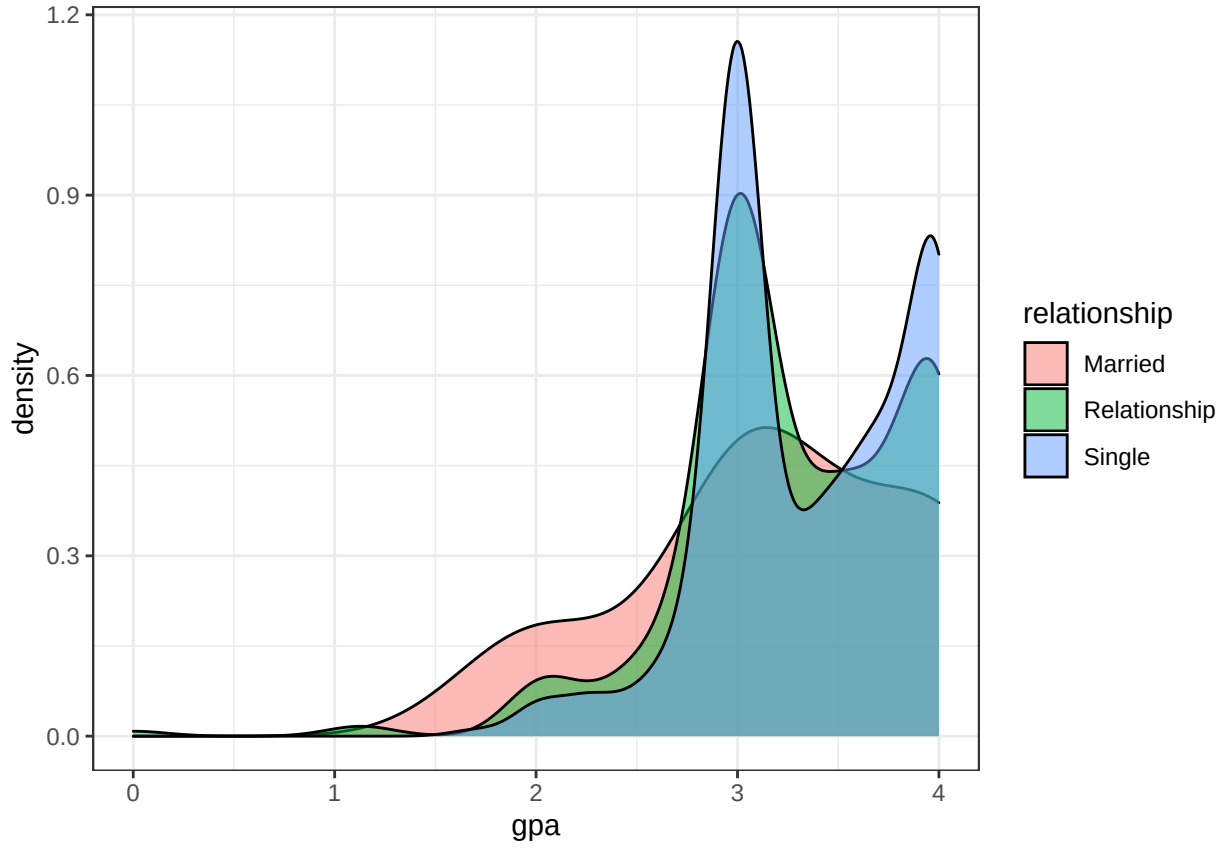
```
# Reorganizing relationship
relationship_categories <- c("Single" = "Single",
                             "In a relationship" = "Relationship",
                             "Partnered" = "Married",
                             "Married" = "Married",
                             "Other" = "Married")
```

```
data$relationship <- relationship_categories[as.character(data$relationship)]

data %>%
  group_by(relationship) %>%
  mutate(Count_Per_Relationship = n()) %>%
  ggplot(aes(relationship, gpa, fill=Count_Per_Relationship)) +
  geom_boxplot() +
  #coord_flip() +
  theme_bw()
```



Here we can see that there is a slight increase in GPA among students that are single, with similar performances by married students and students in relationships.

```
data %>%
  ggplot(aes(x = gpa, fill=relationship)) +
  geom_density(alpha = 0.5) +
  theme_bw()
```



While Single and Relationship students show similar bimodal distributions, married students show a more normal distribution. In this context, it means that while a lot of married students have a GPA around 3.0, there isn't a peak around 4.0, indicating that married students may prioritize having perfect GPAs less than non-married students.

Evaluation of Visualizations While we found some unique differences between the variables here and there, the majority of those variables don't appear to be that useful in predicting GPA since they have a lot of overlap. So, we must select features another way.

3. Early Models

After we preprocessed the data, we can build some preliminary models to serve as a baseline for any future models we might consider.

Selected Variables Initially, we prioritized the variables gender, relationship status, and ethnicity.

```
set.seed(5)
train_i <- sample(nrow(data), 0.8*nrow(data))
selected_data <- data %>% select(
  gpa, gender, relationship, ethnicity
)
selected_train <- selected_data[train_i, , drop=F]
selected_test <- selected_data[-train_i, , drop=F]

selected_lm <- lm(gpa~., selected_train)
summary(selected_lm)
```

```
##
## Call:
## lm(formula = gpa ~ ., data = selected_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.76273 -0.41815  0.04343  0.42876  1.09035
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        3.79603    0.53165   7.140 2.26e-12 ***
## gendermale                        -0.04691    0.03752  -1.250    0.212
## genderqueer                       -0.23156    0.17483  -1.324    0.186
## gendertransgender                  0.18751    0.34502   0.543    0.587
## relationshipRelationship          -0.15987    0.22353  -0.715    0.475
## relationshipSingle                -0.09603    0.22215  -0.432    0.666
## ethnicityasian                    -0.21801    0.48406  -0.450    0.653
## ethnicityblack                    -0.56693    0.49416  -1.147    0.252
## ethnicityhispanic/latino          -0.74343    0.48508  -1.533    0.126
## ethnicitymiddle east/north africa -0.31775    0.49218  -0.646    0.519
## ethnicitymultiple                 -0.37701    0.49103  -0.768    0.443
## ethnicityother                    -0.55747    0.52939  -1.053    0.293
## ethnicitypacific islander          0.24691    0.68411   0.361    0.718
## ethnicitywhite                    -0.20503    0.48489  -0.423    0.673
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.483 on 730 degrees of freedom
## Multiple R-squared:  0.1636, Adjusted R-squared:  0.1487
## F-statistic: 10.99 on 13 and 730 DF,  p-value: < 2.2e-16
```

```
yhat_selected <- predict(selected_lm, selected_test)

selected_MSE <- mean((yhat_selected - selected_test$gpa)^2)
selected_MSE
```

```
## [1] 0.2825577
```

```
set.seed(5)
train_i <- sample(nrow(data), 0.8 * nrow(data))
train_data <- data[train_i, ]
test_data <- data[-train_i, ]

all_lm <- lm(gpa~., train_data)

summary(all_lm)
```

All Variables

```
##
## Call:
## lm(formula = gpa ~ ., data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5149 -0.2432 -0.0021  0.2619  1.0091
##
## Coefficients: (5 not defined because of singularities)
##                                                    Estimate
## (Intercept)                                       3.4980247
## Coursestats101b                                   0.1590387
## Coursestats112                                   -0.1465723
## Coursestats13                                    -0.0235750
## Coursestats13M                                   -0.2767280
## yearJunior                                       -0.2908266
## yearOther                                        -0.9472383
## yearSenior                                       -0.3483830
## yearSophomore                                    -0.2297829
## enrollmentInternational                           0.0689929
## enrollmentOut of state                           -0.0240236
## transfer                                          0.0244899
## statusPart-time                                  -0.0444425
## languageEnglish and other Language(s)            -0.0064486
## languageEnglish only                             -0.0184406
## languageOther than English                        0.0073805
## disciplineBusiness                                0.0571783
## disciplineEngineering and computer science        0.0320361
## disciplineLinguistics                            -0.0368137
## disciplineMathematics                            -0.1254281
## disciplineOthers                                  0.0142237
## disciplinesciene related                          0.0801834
## disciplineSocial science                          0.0025830
## campussouth                                      -0.0180426
## participate                                      -0.0055818
## apo                                               0.0444710
## imccg                                             0.0436085
## pg                                               -0.0734700
```

```
## RSO                                                          -0.0650573
## SFS                                                          -0.1682898
## participationp                                                       NA
## relationshipRelationship                                      0.0084183
## relationshipSingle                                           0.0311792
## gendermale                                                   -0.0428934
## genderqueer                                                  -0.4918616
## gendertransgender                                            0.2916995
## ethnicityasian                                               0.2742265
## ethnicityblack                                               0.2686294
## ethnicityhispanic/latino                                     -0.0025975
## ethnicitymiddle east/north africa                            0.1730478
## ethnicitymultiple                                            0.1663540
## ethnicityother                                               0.3050663
## ethnicitypacific islander                                    0.4005417
## ethnicitywhite                                               0.3162673
## financialaid                                                 -0.0246446
## loans                                                        -0.0257067
## meducGraduate / post Graduate                                0.0361834
## meducHigh school or less                                     -0.0762107
## meducTwo-year college                                        -0.0344124
## feducGraduate / Post Graduate                                0.0303651
## feducHigh school or less                                     -0.0911735
## feducTwo-year college                                        0.0050271
## socioecolower middle                                         0.1508285
## socioecoMiddle class                                         0.0985212
## socioecoUpper middle class/professional                      0.1059956
## socioecoWealthy                                              -0.1056680
## socioecoWorking class                                        0.0494669
## sexualorientBisexual                                         0.0181370
## sexualorientGay                                              0.1619282
## sexualorientHeterosexual                                     0.1268016
## sexualorientHomosexual                                       0.0839080
## sexualorientLesbian                                          -0.0601358
## sexualorientOther                                            0.0342779
## sexualorientQueer                                            0.1628545
## sexualorientQuestioning                                      0.0378419
## religionEastern religion                                     -0.0366746
## religionJewish                                               0.0608357
## religionMuslim                                               -0.0664408
## religionNot particularly spiritual                           0.1115137
## religionOther                                                0.1347469
## religionSpiritual but not associated with a major religion   -0.0005192
## politicalviewFar left                                        -0.0976850
## politicalviewLiberal                                         -0.1051486
## politicalviewModerate                                        -0.0608993
## politicalviewOther                                           -0.0651130
## uclaclimateSomewhat comfortable                              0.0248390
## uclaclimateUncomfortable                                     0.0276370
## uclaclimateVery comfortable                                  -0.1168002
## uclaclimateVery uncomfortable                                -0.1102823
## classclimateNeither comfortable nor uncomfortable            -0.0342478
## classclimateUncomfortable                                    -0.0882874
## classclimateVery comfortable                                 0.1459384
```

```
## classclimateVery uncomfortable                                      -0.3250177
## leavingucla                                                          -0.0928996
## feelexclusionYes, and it interfered with my ability to work or learn  0.1766895
## feelexclusionYes, but it did not interfere with my ability to work or learn -0.0035826
## isolated                                                             0.0112250
## intimidated                                                          -0.0550722
## staringatyou                                                         -0.0451510
## feared                                                               -0.0528159
## racialprofiling                                                      0.0285717
## crimevictim                                                          0.0671453
## physicalviolence                                                     -0.0391264
## stalking                                                             -0.0097236
## derogatoryvictim                                                     -0.0253148
## ucladiscp                                                                   NA
## excluperformane                                                      -0.0105988
## excluage                                                             -0.0445566
## exclucountry                                                         0.0029958
## excluenglish                                                         0.0713175
## excluethnicity                                                       -0.0082451
## exclurace                                                            0.0266471
## exclucitizenship                                                     -0.0459800
## exclumental                                                          -0.0582380
## excludisability                                                      0.1128268
## excluparticipation                                                   0.0067394
## exclupolitical                                                       -0.0190387
## exclureligion                                                        0.0953391
## exclusocio                                                           -0.0516317
## exclugender                                                          0.0622909
## exclusexual                                                          0.0117330
## uclaexclusionaryp                                                           NA
## studentbody                                                          -0.0227274
## crosscultural                                                        0.0036984
## curriculum                                                           -0.0064063
## facultydiver                                                         0.0421978
## impuclaclimate                                                              NA
## facunderstand                                                        0.0481313
## facrespect                                                           -0.0599968
## channels                                                             0.0273684
## sexualresp                                                           -0.0418891
## countryresp                                                          0.0237614
## ethnicityresp                                                        0.0334083
## religionresp                                                         -0.0419970
## appearanceresp                                                       -0.0768678
## socioresp                                                            -0.0414790
## academicresp                                                         0.1016151
## politicalresp                                                        0.0102422
## divrespectp                                                                 NA
##                                                                     Std. Error
## (Intercept)                                                          0.5873627
## Coursestats101b                                                      0.0887253
## Coursestats112                                                       0.0897896
## Coursestats13                                                        0.0698585
## Coursestats13M                                                       0.0866386
## yearJunior                                                           0.0548944
```

```
## yearOther                                                    0.4479929
## yearSenior                                                   0.0668397
## yearSophomore                                                0.0501389
## enrollmentInternational                                      0.0657554
## enrollmentOut of state                                       0.0514798
## transfer                                                     0.0734301
## statusPart-time                                              0.1556883
## languageEnglish and other Language(s)                        0.0612325
## languageEnglish only                                         0.0582996
## languageOther than English                                   0.0598708
## disciplineBusiness                                           0.1790659
## disciplineEngineering and computer science                   0.2179699
## disciplineLinguistics                                        0.3051377
## disciplineMathematics                                        0.1863820
## disciplineOthers                                             0.1776283
## disciplinesciene related                                     0.1707659
## disciplineSocial science                                     0.1667928
## campussouth                                                  0.0678361
## participate                                                  0.0450970
## apo                                                          0.0382708
## imccg                                                        0.0446256
## pg                                                           0.0673754
## RSO                                                          0.0517901
## SFS                                                          0.0541084
## participationp                                                      NA
## relationshipRelationship                                     0.2301581
## relationshipSingle                                           0.2282710
## gendermale                                                   0.0368433
## genderqueer                                                  0.1758501
## gendertransgender                                            0.3218615
## ethnicityasian                                               0.4507586
## ethnicityblack                                               0.4642292
## ethnicityhispanic/latino                                     0.4497216
## ethnicitymiddle east/north africa                            0.4574666
## ethnicitymultiple                                            0.4565464
## ethnicityother                                               0.4994829
## ethnicitypacific islander                                    0.6278034
## ethnicitywhite                                               0.4543934
## financialaid                                                 0.0535699
## loans                                                        0.0422268
## meducGraduate / post Graduate                                0.0478472
## meducHigh school or less                                     0.0574491
## meducTwo-year college                                        0.0659069
## feducGraduate / Post Graduate                                0.0470323
## feducHigh school or less                                     0.0604982
## feducTwo-year college                                        0.0757239
## socioecolower middle                                         0.0914660
## socioecoMiddle class                                         0.0560040
## socioecoUpper middle class/professional                      0.0734149
## socioecoWealthy                                              0.1124243
## socioecoWorking class                                        0.0845325
## sexualorientBisexual                                         0.1145115
## sexualorientGay                                              0.2144456
## sexualorientHeterosexual                                     0.0839291
```

```
## sexualorientHomosexual                                                    0.1363841
## sexualorientLesbian                                                       0.3352740
## sexualorientOther                                                         0.1329559
## sexualorientQueer                                                         0.2506298
## sexualorientQuestioning                                                   0.2403852
## religionEastern religion                                                  0.1331821
## religionJewish                                                            0.1346109
## religionMuslim                                                            0.1663732
## religionNot particularly spiritual                                        0.0460924
## religionOther                                                             0.0550341
## religionSpiritual but not associated with a major religion                0.0595660
## politicalviewFar left                                                     0.1243536
## politicalviewLiberal                                                      0.0609415
## politicalviewModerate                                                     0.0599042
## politicalviewOther                                                        0.0823477
## uclaclimateSomewhat comfortable                                           0.0466799
## uclaclimateUncomfortable                                                  0.0973561
## uclaclimateVery comfortable                                               0.0531003
## uclaclimateVery uncomfortable                                             0.2467039
## classclimateNeither comfortable nor uncomfortable                         0.0473273
## classclimateUncomfortable                                                 0.0840903
## classclimateVery comfortable                                              0.0562967
## classclimateVery uncomfortable                                            0.2362343
## leavingucla                                                               0.0435420
## feelexclusionYes, and it interfered with my ability to work or learn      0.0810078
## feelexclusionYes, but it did not interfere with my ability to work or learn 0.0464310
## isolated                                                                  0.0201671
## intimidated                                                               0.0271369
## staringatyou                                                              0.0196922
## feared                                                                    0.0257331
## racialprofiling                                                           0.0229685
## crimevictim                                                               0.0393998
## physicalviolence                                                          0.0477063
## stalking                                                                  0.0425390
## derogatoryvictim                                                          0.0413660
## ucladiscp                                                                        NA
## excluperformane                                                           0.0435376
## excluage                                                                  0.0548890
## exclucountry                                                              0.0507820
## excluenglish                                                              0.0496577
## excluethnicity                                                            0.0454187
## exclurace                                                                 0.0406813
## exclucitizenship                                                          0.0461639
## exclumental                                                               0.0683942
## excludisability                                                           0.0771294
## excluparticipation                                                        0.0662895
## exclupolitical                                                            0.0516990
## exclureligion                                                             0.0450879
## exclusocio                                                                0.0569213
## exclugender                                                               0.0588630
## exclusexual                                                               0.0524121
## uclaexclusionaryp                                                                NA
## studentbody                                                               0.0227349
## crosscultural                                                             0.0244918
```

```
## curriculum                                                0.0248642
## facultydiver                                              0.0241515
## impuclaclimate                                                   NA
## facunderstand                                             0.0261540
## facrespect                                                0.0281716
## channels                                                  0.0229775
## sexualresp                                                0.0289067
## countryresp                                               0.0244053
## ethnicityresp                                             0.0237627
## religionresp                                              0.0215942
## appearanceresp                                            0.0212364
## socioresp                                                 0.0202006
## academicresp                                              0.0173599
## politicalresp                                             0.0217715
## divrespectp                                                      NA
##                                                           t value
## (Intercept)                                                 5.955
## Coursestats101b                                             1.792
## Coursestats112                                             -1.632
## Coursestats13                                              -0.337
## Coursestats13M                                             -3.194
## yearJunior                                                 -5.298
## yearOther                                                  -2.114
## yearSenior                                                 -5.212
## yearSophomore                                              -4.583
## enrollmentInternational                                     1.049
## enrollmentOut of state                                     -0.467
## transfer                                                    0.334
## statusPart-time                                            -0.285
## languageEnglish and other Language(s)                      -0.105
## languageEnglish only                                       -0.316
## languageOther than English                                  0.123
## disciplineBusiness                                          0.319
## disciplineEngineering and computer science                  0.147
## disciplineLinguistics                                      -0.121
## disciplineMathematics                                      -0.673
## disciplineOthers                                            0.080
## disciplinesciene related                                    0.470
## disciplineSocial science                                    0.015
## campussouth                                                -0.266
## participate                                                -0.124
## apo                                                         1.162
## imccg                                                       0.977
## pg                                                         -1.090
## RSO                                                        -1.256
## SFS                                                        -3.110
## participationp                                                NA
## relationshipRelationship                                    0.037
## relationshipSingle                                          0.137
## gendermale                                                 -1.164
## genderqueer                                                -2.797
## gendertransgender                                           0.906
## ethnicityasian                                              0.608
## ethnicityblack                                              0.579
```

```
## ethnicityhispanic/latino                                                -0.006
## ethnicitymiddle east/north africa                                        0.378
## ethnicitymultiple                                                        0.364
## ethnicityother                                                           0.611
## ethnicitypacific islander                                                0.638
## ethnicitywhite                                                           0.696
## financialaid                                                            -0.460
## loans                                                                   -0.609
## meducGraduate / post Graduate                                            0.756
## meducHigh school or less                                                -1.327
## meducTwo-year college                                                   -0.522
## feducGraduate / Post Graduate                                            0.646
## feducHigh school or less                                                -1.507
## feducTwo-year college                                                    0.066
## socioecolower middle                                                     1.649
## socioecoMiddle class                                                     1.759
## socioecoUpper middle class/professional                                  1.444
## socioecoWealthy                                                         -0.940
## socioecoWorking class                                                    0.585
## sexualorientBisexual                                                     0.158
## sexualorientGay                                                          0.755
## sexualorientHeterosexual                                                 1.511
## sexualorientHomosexual                                                   0.615
## sexualorientLesbian                                                     -0.179
## sexualorientOther                                                        0.258
## sexualorientQueer                                                        0.650
## sexualorientQuestioning                                                  0.157
## religionEastern religion                                                -0.275
## religionJewish                                                           0.452
## religionMuslim                                                          -0.399
## religionNot particularly spiritual                                       2.419
## religionOther                                                            2.448
## religionSpiritual but not associated with a major religion              -0.009
## politicalviewFar left                                                   -0.786
## politicalviewLiberal                                                    -1.725
## politicalviewModerate                                                   -1.017
## politicalviewOther                                                      -0.791
## uclaclimateSomewhat comfortable                                          0.532
## uclaclimateUncomfortable                                                 0.284
## uclaclimateVery comfortable                                             -2.200
## uclaclimateVery uncomfortable                                           -0.447
## classclimateNeither comfortable nor uncomfortable                       -0.724
## classclimateUncomfortable                                               -1.050
## classclimateVery comfortable                                             2.592
## classclimateVery uncomfortable                                          -1.376
## leavingucla                                                             -2.134
## feelexclusionYes, and it interfered with my ability to work or learn     2.181
## feelexclusionYes, but it did not interfere with my ability to work or learn -0.077
## isolated                                                                 0.557
## intimidated                                                             -2.029
## staringatyou                                                            -2.293
## feared                                                                  -2.052
## racialprofiling                                                          1.244
## crimevictim                                                              1.704
```

```
## physicalviolence                                  -0.820
## stalking                                          -0.229
## derogatoryvictim                                  -0.612
## ucladiscp                                            NA
## excluperformane                                   -0.243
## excluage                                          -0.812
## exclucountry                                       0.059
## excluenglish                                       1.436
## excluethnicity                                    -0.182
## exclurace                                          0.655
## exclucitizenship                                  -0.996
## exclumental                                       -0.852
## excludisability                                    1.463
## excluparticipation                                 0.102
## exclupolitical                                    -0.368
## exclureligion                                      2.115
## exclusocio                                        -0.907
## exclugender                                        1.058
## exclusexual                                        0.224
## uclaexclusionaryp                                    NA
## studentbody                                       -1.000
## crosscultural                                      0.151
## curriculum                                        -0.258
## facultydiver                                       1.747
## impuclaclimate                                       NA
## facunderstand                                      1.840
## facrespect                                        -2.130
## channels                                           1.191
## sexualresp                                        -1.449
## countryresp                                        0.974
## ethnicityresp                                      1.406
## religionresp                                      -1.945
## appearanceresp                                    -3.620
## socioresp                                         -2.053
## academicresp                                       5.853
## politicalresp                                      0.470
## divrespectp                                          NA
##                                                 Pr(>|t|)
## (Intercept)                                     4.35e-09
## Coursestats101b                                 0.073543
## Coursestats112                                  0.103103
## Coursestats13                                   0.735878
## Coursestats13M                                  0.001474
## yearJunior                                      1.63e-07
## yearOther                                       0.034879
## yearSenior                                      2.54e-07
## yearSophomore                                   5.55e-06
## enrollmentInternational                         0.294479
## enrollmentOut of state                          0.640907
## transfer                                        0.738859
## statusPart-time                                 0.775389
## languageEnglish and other Language(s)           0.916161
## languageEnglish only                            0.751875
## languageOther than English                      0.901930
```

```
## disciplineBusiness                                  0.749596
## disciplineEngineering and computer science          0.883200
## disciplineLinguistics                               0.904010
## disciplineMathematics                               0.501222
## disciplineOthers                                    0.936203
## disciplinesciene related                            0.638840
## disciplineSocial science                            0.987649
## campussouth                                         0.790348
## participate                                         0.901534
## apo                                                 0.245679
## imccg                                               0.328847
## pg                                                  0.275935
## RSO                                                 0.209527
## SFS                                                 0.001955
## participationp                                           NA
## relationshipRelationship                            0.970835
## relationshipSingle                                  0.891400
## gendermale                                          0.244786
## genderqueer                                         0.005317
## gendertransgender                                   0.365135
## ethnicityasian                                      0.543167
## ethnicityblack                                      0.563031
## ethnicityhispanic/latino                            0.995393
## ethnicitymiddle east/north africa                   0.705356
## ethnicitymultiple                                   0.715702
## ethnicityother                                      0.541579
## ethnicitypacific islander                           0.523706
## ethnicitywhite                                      0.486677
## financialaid                                        0.645645
## loans                                               0.542896
## meducGraduate / post Graduate                       0.449799
## meducHigh school or less                            0.185136
## meducTwo-year college                               0.601762
## feducGraduate / Post Graduate                       0.518763
## feducHigh school or less                            0.132308
## feducTwo-year college                               0.947090
## socioecolower middle                                0.099652
## socioecoMiddle class                                0.079040
## socioecoUpper middle class/professional             0.149304
## socioecoWealthy                                     0.347633
## socioecoWorking class                               0.558639
## sexualorientBisexual                                0.874204
## sexualorientGay                                     0.450475
## sexualorientHeterosexual                            0.131344
## sexualorientHomosexual                              0.538626
## sexualorientLesbian                                 0.857711
## sexualorientOther                                   0.796636
## sexualorientQueer                                   0.516074
## sexualorientQuestioning                             0.874964
## religionEastern religion                            0.783122
## religionJewish                                      0.651472
## religionMuslim                                      0.689774
## religionNot particularly spiritual                  0.015835
## religionOther                                       0.014624
```

```
## religionSpiritual but not associated with a major religion                  0.993048
## politicalviewFar left                                                       0.432435
## politicalviewLiberal                                                        0.084953
## politicalviewModerate                                                       0.309735
## politicalviewOther                                                          0.429417
## uclaclimateSomewhat comfortable                                             0.594838
## uclaclimateUncomfortable                                                    0.776600
## uclaclimateVery comfortable                                                 0.028203
## uclaclimateVery uncomfortable                                               0.655015
## classclimateNeither comfortable nor uncomfortable                           0.469561
## classclimateUncomfortable                                                   0.294168
## classclimateVery comfortable                                                0.009758
## classclimateVery uncomfortable                                              0.169372
## leavingucla                                                                 0.033270
## feelexclusionYes, and it interfered with my ability to work or learn        0.029548
## feelexclusionYes, but it did not interfere with my ability to work or learn 0.938521
## isolated                                                                    0.578001
## intimidated                                                                 0.042842
## staringatyou                                                                0.022191
## feared                                                                      0.040545
## racialprofiling                                                             0.213988
## crimevictim                                                                 0.088844
## physicalviolence                                                            0.412444
## stalking                                                                    0.819270
## derogatoryvictim                                                            0.540782
## ucladiscp                                                                         NA
## excluperformane                                                             0.807746
## excluage                                                                    0.417242
## exclucountry                                                                0.952977
## excluenglish                                                                0.151454
## excluethnicity                                                              0.856007
## exclurace                                                                   0.512697
## exclucitizenship                                                            0.319631
## exclumental                                                                 0.394818
## excludisability                                                             0.144022
## excluparticipation                                                          0.919055
## exclupolitical                                                              0.712804
## exclureligion                                                               0.034870
## exclusocio                                                                  0.364721
## exclugender                                                                 0.290360
## exclusexual                                                                 0.822940
## uclaexclusionaryp                                                                 NA
## studentbody                                                                 0.317860
## crosscultural                                                               0.880020
## curriculum                                                                  0.796761
## facultydiver                                                                0.081096
## impuclaclimate                                                                    NA
## facunderstand                                                               0.066201
## facrespect                                                                  0.033590
## channels                                                                    0.234072
## sexualresp                                                                  0.147812
## countryresp                                                                 0.330626
## ethnicityresp                                                               0.160250
## religionresp                                                                0.052247
```

```
## appearanceresp                                  0.000319
## socioresp                                       0.040457
## academicresp                                    7.81e-09
## politicalresp                                   0.638206
## divrespectp                                           NA
##
## (Intercept)                                     ***
## Coursestats101b                                 .
## Coursestats112
## Coursestats13
## Coursestats13M                                  **
## yearJunior                                      ***
## yearOther                                       *
## yearSenior                                      ***
## yearSophomore                                   ***
## enrollmentInternational
## enrollmentOut of state
## transfer
## statusPart-time
## languageEnglish and other Language(s)
## languageEnglish only
## languageOther than English
## disciplineBusiness
## disciplineEngineering and computer science
## disciplineLinguistics
## disciplineMathematics
## disciplineOthers
## disciplinesciene related
## disciplineSocial science
## campussouth
## participate
## apo
## imccg
## pg
## RSO
## SFS                                             **
## participationp
## relationshipRelationship
## relationshipSingle
## gendermale
## genderqueer                                     **
## gendertransgender
## ethnicityasian
## ethnicityblack
## ethnicityhispanic/latino
## ethnicitymiddle east/north africa
## ethnicitymultiple
## ethnicityother
## ethnicitypacific islander
## ethnicitywhite
## financialaid
## loans
## meducGraduate / post Graduate
## meducHigh school or less
```

```
## meducTwo-year college
## feducGraduate / Post Graduate
## feducHigh school or less
## feducTwo-year college
## socioecolower middle                                                    .
## socioecoMiddle class                                                    .
## socioecoUpper middle class/professional
## socioecoWealthy
## socioecoWorking class
## sexualorientBisexual
## sexualorientGay
## sexualorientHeterosexual
## sexualorientHomosexual
## sexualorientLesbian
## sexualorientOther
## sexualorientQueer
## sexualorientQuestioning
## religionEastern religion
## religionJewish
## religionMuslim
## religionNot particularly spiritual                                      *
## religionOther                                                           *
## religionSpiritual but not associated with a major religion
## politicalviewFar left
## politicalviewLiberal                                                    .
## politicalviewModerate
## politicalviewOther
## uclaclimateSomewhat comfortable
## uclaclimateUncomfortable
## uclaclimateVery comfortable                                             *
## uclaclimateVery uncomfortable
## classclimateNeither comfortable nor uncomfortable
## classclimateUncomfortable
## classclimateVery comfortable                                            **
## classclimateVery uncomfortable
## leavingucla                                                             *
## feelexclusionYes, and it interfered with my ability to work or learn    *
## feelexclusionYes, but it did not interfere with my ability to work or learn
## isolated
## intimidated                                                             *
## staringatyou                                                            *
## feared                                                                  *
## racialprofiling
## crimevictim                                                             .
## physicalviolence
## stalking
## derogatoryvictim
## ucladiscp
## excluperformane
## excluage
## exclucountry
## excluenglish
## excluethnicity
## exclurace
```

```
## exclucitizenship
## exclumental
## excludisability
## excluparticipation
## exclupolitical
## exclureligion                                                        *
## exclusocio
## exclugender
## exclusexual
## uclaexclusionaryp
## studentbody
## crosscultural
## curriculum
## facultydiver                                                         .
## impuclaclimate
## facunderstand                                                        .
## facrespect                                                           *
## channels
## sexualresp
## countryresp
## ethnicityresp
## religionresp                                                         .
## appearanceresp                                                       ***
## socioresp                                                            *
## academicresp                                                         ***
## politicalresp
## divrespectp
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4143 on 620 degrees of freedom
## Multiple R-squared:  0.4775, Adjusted R-squared:  0.3738
## F-statistic: 4.606 on 123 and 620 DF,  p-value: < 2.2e-16
```

```
yhat_all <- predict(all_lm, test_data)
```

```
## Warning in predict.lm(all_lm, test_data): prediction from a rank-deficient fit
## may be misleading
```

```
all_MSE <- mean((yhat_all - test_data$gpa)^2)
all_MSE
```

```
## [1] 0.2697427
```

4. Preprocessing the Data

Before selecting features, we can manually change categorical data to numeric data.

```
## Create dummy variables for Course
data$Course %>% levels()
```

```
## [1] "stats10"   "stats101b" "stats112"  "stats13"   "stats13M"
```

```
data <- data %>%
  mutate(stats10 = ifelse(Course == "stats10", 1, 0),
         stats101b = ifelse(Course == "stats101b", 1, 0),
         stats112 = ifelse(Course == "stats112", 1, 0),
         stats13 = ifelse(Course == "stats13", 1, 0),
         stats13M = ifelse(Course == "stats13M", 1, 0),) %>%
  select(-Course)

## Convert Year to Numeric Values
data$year %>% levels()
```

```
## [1] "Freshman"  "Junior"    "Other"     "Senior"    "Sophomore"
```

```
year_to_num <- c("Freshman" = 1,
                 "Sophomore" = 2,
                 "Junior" = 3,
                 "Senior" = 4,
                 "Other" = 5)

data$year <- year_to_num[as.character(data$year)]

## Convert Enrollment to dummy variables
data$enrollment %>% levels()
```

```
## [1] "In state"      "International" "Out of state"
```

```
data <- data %>%
  mutate(inState = ifelse(enrollment == "In state", 1, 0),
         international = ifelse(enrollment == "International", 1, 0)) %>%
  select(-enrollment)

## Convert Status
data$status %>% levels()
```

```
## [1] "Full-time" "Part-time"
```

```
data <- data %>%
  mutate(fullTime = ifelse(status == "Full-time", 1, 0)) %>%
  select(-status)

## Convert Language
data$language %>% levels()
```

```
## [1] "English and other"             "English and other Language(s)"
## [3] "English only"                  "Other than English"
```

```
data <- data %>%
  mutate(multilingual = ifelse(language == "English only", 0, 1)) %>%
  select(-language)

## Convert Discipline
data$discipline %>% levels()
```

```
## [1] "Art and architecture"             "Business"
## [3] "Engineering and computer science" "Linguistics"
## [5] "Mathematics"                      "Others"
## [7] "sciene related"                   "Social science"
```

```
stem_disciplines <- c("Engineering and computer science", "Mathematics", "sciene related")
humanities_disciplines <- c("Art and architecture", "Social science", "Linguistics")
business_disciplines <- c("Business")
other_disciplines <- c("Art and architecture", "Others")

data <- data %>%
  mutate(stem = ifelse(discipline %in% stem_disciplines, 1, 0),
         humanities = ifelse(discipline %in% humanities_disciplines, 1, 0),
         business = ifelse(discipline %in% business_disciplines, 1, 0),
         otherDiscipline = ifelse(discipline %in% other_disciplines, 1, 0)) %>%
  select(-discipline)

## Convert Campus
data$campus %>% levels()
```

```
## [1] "north" "south"
```

```
data <- data %>%
  mutate(northCampus = ifelse(campus == "north", 1, 0)) %>%
  select(-campus)

## Convert Relationship
data$relationship %>% unique()
```

```
## [1] "Single"       "Relationship" "Married"
```

```
data <- data %>%
  mutate(single = ifelse(relationship == "Single", 1, 0),
         relationship = ifelse(relationship == "Relationship", 1, 0),
         married = ifelse(relationship == "Married", 1, 0)) %>%
  select(-relationship)

## Convert Gender
data$gender %>% unique()
```

```
## [1] "female"      "male"        "transgender" "queer"
```

```
data <- data %>%
  mutate(female = ifelse(gender == "female", 1, 0),
         male = ifelse(gender == "male", 1, 0),
         transgender = ifelse(gender == "transgender", 1, 0),
         queer = ifelse(gender == "queer", 1, 0)) %>%
  select(-gender)

## Convert Ethnicity
data$ethnicity %>% unique()
```

```
## [1] hispanic/latino          asian                     white
## [4] black                    middle east/north africa other
## [7] multiple                 american indian           pacific islander
## 9 Levels: american indian asian black ... white
```

```
data <- data %>%
  mutate(hispanicLatino = ifelse(ethnicity == "hispanic/latino", 1, 0),
         asian = ifelse(ethnicity == "asian", 1, 0),
         white = ifelse(ethnicity == "white", 1, 0),
         black = ifelse(ethnicity == "black", 1, 0),
         ME_NA = ifelse(ethnicity == "middle east/north africa", 1, 0),
         otherEthnicity = ifelse(ethnicity == "other", 1, 0),
         multipleEthnicity = ifelse(ethnicity == "multiple", 1, 0),
         americanIndian = ifelse(ethnicity == "american indian", 1, 0),
         pacific_islander = ifelse(ethnicity == "pacific islander", 1, 0)) %>%
  select(-ethnicity)

## Convert Mother/Father education
data$feduc %>% levels()
```

```
## [1] "Four-year college"        "Graduate / Post Graduate"
## [3] "High school or less"      "Two-year college"
```

```
levels(data$meduc) <- c(3, 4, 1, 2)
levels(data$feduc) <- c(3, 4, 1, 2)

data$meduc <- as.numeric(as.character(data$meduc))
data$feduc <- as.numeric(as.character(data$feduc))

## Convert Socioeco
socioeco_to_num <- c("Low income" = 1,
                     "lower middle" = 1,
                     "Working class" = 1,
                     "Middle class" = 2,
                     "Upper middle class/professional" = 3,
                     "Wealthy" = 3)

data <- data %>%
  mutate(socioeco = socioeco_to_num[socioeco])

## Convert Sexualorient
data$sexualorient %>% levels()
```

```
## [1] "Asexual"      "Bisexual"     "Gay"          "Heterosexual" "Homosexual"
## [6] "Lesbian"      "Other"        "Queer"        "Questioning"
```

```
data <- data %>%
  mutate(asexual = ifelse(sexualorient == "Asexual", 1, 0),
         bisexual = ifelse(sexualorient == "Bisexual", 1, 0),
         homosexual = ifelse(sexualorient == "Homosexual" | sexualorient == "Gay" | sexualorient == "Les
         heterosexual = ifelse(sexualorient == "Heterosexual", 1, 0),
         otherSexualOrient = ifelse(sexualorient == "Other" | sexualorient == "Questioning", 1, 0),
         queer = ifelse(sexualorient == "Queer", 1, 0)) %>%
  select(-sexualorient)

## Convert Religion
data$religion %>% levels()
```

```
## [1] "Christian"
## [2] "Eastern religion"
## [3] "Jewish"
## [4] "Muslim"
## [5] "Not particularly spiritual"
## [6] "Other"
## [7] "Spiritual but not associated with a major religion"
```

```
data <- data %>%
  mutate(christian = ifelse(religion == "Christian", 1, 0),
         jewish = ifelse(religion == "Jewish", 1, 0),
         muslim = ifelse(religion == "Muslim", 1, 0),
         otherReligion = ifelse(religion == "Eastern religion" | religion == "Other" | religion == "Spi
         nonreligious = ifelse(religion == "Not particularly spiritual", 1, 0)) %>%
  select(-religion)

## Convert Political View
data$politicalview %>% levels()
```

```
## [1] "Conservative" "Far left"     "Liberal"      "Moderate"     "Other"
```

```
data <- data %>%
  mutate(conservative = ifelse(politicalview == "Conservative", 1, 0),
         farLeft = ifelse(politicalview == "Far left", 1, 0),
         liberal = ifelse(politicalview == "Liberal", 1, 0),
         moderate = ifelse(politicalview == "Moderate", 1, 0),
         otherPoliticalView = ifelse(politicalview == "Other", 1, 0)) %>%
  select(-politicalview)

## Convert UCLA Climate / Class Climate
data$uclaclimate %>% levels()
```

```
## [1] "Comfortable"          "Somewhat comfortable" "Uncomfortable"
## [4] "Very comfortable"     "Very uncomfortable"
```

```
data$classclimate %>% levels()
```

```
## [1] "Comfortable"
## [2] "Neither comfortable nor uncomfortable"
## [3] "Uncomfortable"
## [4] "Very comfortable"
## [5] "Very uncomfortable"
```

```
climate_to_num <- c("Very uncomfortable" = 1,
                    "Uncomfortable" = 2,
                    "Somewhat comfortable" = 3,
                    "Neither comfortable nor uncomfortable" = 3,
                    "Comfortable" = 4,
                    "Very comfortable" = 5)

data <- data %>%
  mutate(uclaclimate = climate_to_num[uclaclimate],
         classclimate = climate_to_num[classclimate])

## Convert Feelexclusion
data$feelexclusion %>% levels()
```

```
## [1] "No"
## [2] "Yes, and it interfered with my ability to work or learn"
## [3] "Yes, but it did not interfere with my ability to work or learn"
```

```
feelexclusion_to_num <- c("No" = 1,
                          "Yes, and it interfered with my ability to work or learn" = 3,
                          "Yes, but it did not interfere with my ability to work or learn" = 2)

data <- data %>%
  mutate(feelexclusion = feelexclusion_to_num[feelexclusion])

table(sapply(data, class)) #All variables now integer or numeric
```

```
##
## integer numeric
##      50      56
```

5. Feature Selection

Significant Features Our first approach is choosing the significant variables shown in the summary(all_lm) call in section 3.

The significant factors from our all-variables linear model were: - stats101b - stats13M - year - SFS - queer - socioeco - otherReligion - nonreligious - liberal - uclaclimate - classclimate - leavingucla - feelexclusion - intimidated - staringatyou - feared - crimevictim - exclureligion - facultydiver - facunderstand - facrespect - religionresp - appearanceresp - socioresp - academicresp

Lasso Regression In this section, we'll be using Lasso Regression to select important features from our data. This method is similar to weighted least squares, except instead of assigning weights to minimize RSS ($\sum(y_i - \hat{y}_i)^2$), Lasso seeks to minimize $RSS + \lambda \sum_{j=1}^{p} |\beta_j|$ where $\lambda \in (0, 1)$.

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
## Loaded glmnet 4.1-4
```

```
#train_data %>% head()
x_train <- model.matrix(gpa~., train_data)
x_train <- cbind("(Intercept)" = x_train[, 1], scale(x_train[, -1]))
y_train <- train_data$gpa

cv_lasso <- cv.glmnet(x = x_train, y = y_train, alpha=1)
coef(cv_lasso)
```

```
## 130 x 1 sparse Matrix of class "dgCMatrix"
##                                                                          s1
## (Intercept)                                                    3.328497e+00
## (Intercept)                                                    .
## Coursestats101b                                                .
## Coursestats112                                                -2.660791e-02
## Coursestats13                                                  .
## Coursestats13M                                                -2.031459e-02
## yearJunior                                                    -1.284806e-02
## yearOther                                                     -1.173257e-02
## yearSenior                                                    -1.944496e-02
## yearSophomore                                                  .
## enrollmentInternational                                        4.546067e-05
## enrollmentOut of state                                         .
## transfer                                                       .
## statusPart-time                                                .
## languageEnglish and other Language(s)                          .
```

```
## languageEnglish only                                                .
## languageOther than English                                          .
## disciplineBusiness                                                  .
## disciplineEngineering and computer science                          .
## disciplineLinguistics                                               .
## disciplineMathematics                                         -1.541475e-02
## disciplineOthers                                                    .
## disciplinesciene related                                            .
## disciplineSocial science                                            .
## campussouth                                                         .
## participate                                                         .
## apo                                                            4.008048e-03
## imccg                                                          6.435471e-04
## pg                                                                  .
## RSO                                                                 .
## SFS                                                           -3.497779e-02
## participationp                                                      .
## relationshipRelationship                                      -1.954237e-04
## relationshipSingle                                                  .
## gendermale                                                          .
## genderqueer                                                         .
## gendertransgender                                                   .
## ethnicityasian                                                      .
## ethnicityblack                                                      .
## ethnicityhispanic/latino                                      -9.085422e-02
## ethnicitymiddle east/north africa                                   .
## ethnicitymultiple                                                   .
## ethnicityother                                                      .
## ethnicitypacific islander                                           .
## ethnicitywhite                                                      .
## financialaid                                                  -3.263187e-02
## loans                                                         -3.815209e-03
## meducGraduate / post Graduate                                  9.366556e-05
## meducHigh school or less                                      -1.622938e-02
## meducTwo-year college                                               .
## feducGraduate / Post Graduate                                  1.031925e-02
## feducHigh school or less                                      -4.982047e-02
## feducTwo-year college                                               .
## socioecolower middle                                                .
## socioecoMiddle class                                           1.706925e-02
## socioecoUpper middle class/professional                             .
## socioecoWealthy                                                     .
## socioecoWorking class                                               .
## sexualorientBisexual                                                .
## sexualorientGay                                                     .
## sexualorientHeterosexual                                            .
## sexualorientHomosexual                                              .
## sexualorientLesbian                                                 .
## sexualorientOther                                                   .
## sexualorientQueer                                                   .
## sexualorientQuestioning                                             .
## religionEastern religion                                            .
## religionJewish                                                      .
## religionMuslim                                                      .
```

```
## religionNot particularly spiritual                                        2.486278e-02
## religionOther                                                              1.414465e-02
## religionSpiritual but not associated with a major religion                 .
## politicalviewFar left                                                      .
## politicalviewLiberal                                                       .
## politicalviewModerate                                                      .
## politicalviewOther                                                         .
## uclaclimateSomewhat comfortable                                            .
## uclaclimateUncomfortable                                                   .
## uclaclimateVery comfortable                                                .
## uclaclimateVery uncomfortable                                             -1.784404e-04
## classclimateNeither comfortable nor uncomfortable                         -5.040454e-04
## classclimateUncomfortable                                                  .
## classclimateVery comfortable                                               3.390694e-03
## classclimateVery uncomfortable                                            -8.492999e-03
## leavingucla                                                               -2.369777e-02
## feelexclusionYes, and it interfered with my ability to work or learn       2.731321e-03
## feelexclusionYes, but it did not interfere with my ability to work or learn .
## isolated                                                                   .
## intimidated                                                               -2.380366e-02
## staringatyou                                                              -4.441377e-02
## feared                                                                     .
## racialprofiling                                                            .
## crimevictim                                                                .
## physicalviolence                                                           .
## stalking                                                                   .
## derogatoryvictim                                                           .
## ucladiscp                                                                  .
## excluperformane                                                            .
## excluage                                                                   .
## exclucountry                                                               .
## excluenglish                                                               2.644308e-03
## excluethnicity                                                             .
## exclurace                                                                  .
## exclucitizenship                                                           .
## exclumental                                                                .
## excludisability                                                            .
## excluparticipation                                                         .
## exclupolitical                                                             .
## exclureligion                                                              1.439992e-02
## exclusocio                                                                 .
## exclugender                                                                1.299414e-03
## exclusexual                                                                .
## uclaexclusionaryp                                                          .
## studentbody                                                                .
## crosscultural                                                              .
## curriculum                                                                 .
## facultydiver                                                               .
## impuclaclimate                                                             .
## facunderstand                                                              .
## facrespect                                                                -4.999709e-03
## channels                                                                   .
## sexualresp                                                                 .
## countryresp                                                                .
```

```
## ethnicityresp                                                    .
## religionresp                                                -1.655105e-03
## appearanceresp                                              -4.878858e-02
## socioresp                                                        .
## academicresp                                                 9.912954e-02
## politicalresp                                                    .
## divrespectp                                                      .
```

Here, we see that the most influential variables are year, financialaid, meduc, feduc, staringatyou, academicresp, and hispanicLatino.

Stepwise Regression ? (might do later)

6. Improved Models

```
significant_factors <- c("gpa", "stats101b", "stats13M", "year", "SFS", "queer", "socioeco", "otherReli

significant_features_data <- data %>%
  select(any_of(significant_factors))

significant_train <- significant_features_data[train_i, ]
significant_test <- significant_features_data[-train_i, ]

significant_lm <- lm(gpa~., significant_train)

summary(significant_lm)
```

Significant Features Model

```
##
## Call:
## lm(formula = gpa ~ ., data = significant_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00173 -0.33765  0.00554  0.34307  1.24066
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.8497586  0.1777776  21.655  < 2e-16 ***
## stats101b       0.1414380  0.0771781   1.833 0.067274 .
## stats13M       -0.1569057  0.0717903  -2.186 0.029167 *
## year           -0.1134189  0.0183631  -6.176 1.10e-09 ***
## SFS            -0.1197495  0.0494203  -2.423 0.015636 *
## queer          -0.1639677  0.2130815  -0.770 0.441846
## socioeco        0.0323986  0.0272508   1.189 0.234871
## otherReligion   0.0632245  0.0425369   1.486 0.137628
## nonreligious    0.1605447  0.0426504   3.764 0.000181 ***
## liberal        -0.0526674  0.0351672  -1.498 0.134670
## uclaclimate    -0.0136661  0.0236860  -0.577 0.564140
## classclimate   -0.0003451  0.0234959  -0.015 0.988285
## leavingucla    -0.1068120  0.0422390  -2.529 0.011660 *
## feelexclusion   0.0638142  0.0330351   1.932 0.053789 .
## intimidated    -0.0801116  0.0261058  -3.069 0.002231 **
## staringatyou   -0.0685746  0.0182838  -3.751 0.000191 ***
## feared         -0.0031947  0.0248329  -0.129 0.897671
## crimevictim     0.0143883  0.0284752   0.505 0.613510
## exclureligion   0.0935944  0.0417320   2.243 0.025218 *
## facultydiver    0.0173872  0.0160344   1.084 0.278567
## facunderstand   0.0134244  0.0254556   0.527 0.598102
## facrespect     -0.0337532  0.0271431  -1.244 0.214080
## religionresp   -0.0415773  0.0198456  -2.095 0.036518 *
```

```
## sexualresp     -0.0298031  0.0246047  -1.211 0.226190
## appearanceresp -0.0905760  0.0203299  -4.455 9.72e-06 ***
## socioresp       0.0309217  0.0182924   1.690 0.091385 .
## academicresp    0.1151061  0.0169136   6.806 2.13e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4556 on 717 degrees of freedom
## Multiple R-squared:  0.2692, Adjusted R-squared:  0.2428
## F-statistic: 10.16 on 26 and 717 DF,  p-value: < 2.2e-16
```

```
yhat_significant <- predict(significant_lm, significant_test)

significant_MSE <- mean((yhat_significant - significant_test$gpa)^2)
significant_MSE
```

```
## [1] 0.3127873
```

```
filtered_data <- data %>%
  select(gpa, year, financialaid, meduc, feduc, staringatyou, academicresp, hispanicLatino)

filtered_train <- filtered_data[train_i, ]
filtered_test <- filtered_data[-train_i, ]

x_train <- filtered_train %>% select(-gpa) %>% scale()
y_train <- filtered_train$gpa

unscaled_lm <- lm(gpa~., filtered_train)
scaled_lm <- lm(gpa~., data.frame("gpa" = y_train, x_train))

summary(unscaled_lm)
```

Lasso Regression Chosen Model

```
##
## Call:
## lm(formula = gpa ~ ., data = filtered_train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08097 -0.30942  0.04301  0.32892  1.06499
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.33071    0.10496  31.733  < 2e-16 ***
## year           -0.09434    0.01652  -5.711 1.63e-08 ***
## financialaid   -0.08564    0.04205  -2.037 0.042019 *
## meduc           0.01853    0.01991   0.931 0.352115
## feduc           0.04594    0.01980   2.320 0.020602 *
## staringatyou   -0.05692    0.01542  -3.692 0.000239 ***
```

```
## academicresp    0.07791    0.01370   5.685 1.88e-08 ***
## hispanicLatino -0.28591    0.05057  -5.653 2.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4458 on 736 degrees of freedom
## Multiple R-squared:  0.2816, Adjusted R-squared:  0.2748
## F-statistic: 41.22 on 7 and 736 DF,  p-value: < 2.2e-16
```

```
summary(scaled_lm)
```

```
##
## Call:
## lm(formula = gpa ~ ., data = data.frame(gpa = y_train, x_train))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08097 -0.30942  0.04301  0.32892  1.06499
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.32850    0.01634 203.646  < 2e-16 ***
## year           -0.09431    0.01651  -5.711 1.63e-08 ***
## financialaid   -0.04284    0.02103  -2.037 0.042019 *
## meduc           0.02182    0.02344   0.931 0.352115
## feduc           0.05710    0.02461   2.320 0.020602 *
## staringatyou   -0.06305    0.01708  -3.692 0.000239 ***
## academicresp    0.09875    0.01737   5.685 1.88e-08 ***
## hispanicLatino -0.11121    0.01967  -5.653 2.25e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4458 on 736 degrees of freedom
## Multiple R-squared:  0.2816, Adjusted R-squared:  0.2748
## F-statistic: 41.22 on 7 and 736 DF,  p-value: < 2.2e-16
```

```
yhat_unscaled <- predict(unscaled_lm, newdata=filtered_test %>% select(-gpa))

unscaled_MSE <- mean((yhat_unscaled - filtered_test$gpa)^2)
unscaled_MSE
```

```
## [1] 0.2509912
```

```
x_test <- filtered_test %>% select(-gpa) %>% scale()
y_test <- filtered_test$gpa

yhat_scaled <- predict(scaled_lm, newdata=data.frame("gpa" = y_test, x_test))

scaled_MSE <- mean((yhat_scaled - filtered_test$gpa)^2)
scaled_MSE
```

```
## [1] 0.2510066
```

In this case, scaling the data did not improve the performance of the model. However, the resulting MSE for the unscaled model is lower than our previous best MSE from the all-variables linear model! That being said, the Adjusted R-squared is lower than the all-variables model.

7. Applying Transformations

8+ Anything else we might do