MA415 Lab: Deliverable 1

Superficial Intelligence (Nicholas Tanabe, Haodong Liu, Huiwen He, Xiaoyi Zhang)

Abstract

In this project, we explore graduate admissions data from Grad Cafe. As all group members of Superficial Intelligence have experience with College and graduate school application, we became interested in quantitatively analyzing the graduate admission result. In this project, our main goal is to investigate the question of how do different variables in the dataset, such as GRE score and undergraduate GPA, relate to the admission decision. In the first part, we mainly investigate the distribution of single variables contained in this dataset. We will also extract useful information from the dataset by looking at the applications to different schools, application demographic, and overall application number. We first look at every single variable in the dataset and describe its pattern or trend if it has one. Then we look at the covariation of different variables. For this project deliverable 1, we look at four problems: The decision reported over time, the relationship between acceptance rate and season, the relationship between acceptance rate and student status, the selectivity of different graduate school programs. We use functions in R to plot the variables of interest. We found out as more students tend to report their decision to Grad Cafe these years, 60% of applicants are American students and 40% are international students. According to the figure, American students are more likely to get accepted compared to international students. In the future, we would like to use the logistic model to predict the probability of a student getting accepted by school programs.

Dataset Description

For this project, we will use data on graduate admissions from Grad Cafe provided by Debarghya Das on GitHub. The graduate school admission result database includes admission results and detailed student test scores self-reported by prospective graduate students on https://www.thegradcafe.com/. The dataset contains 345,303 observations and 19 variables with a mix of continuous and categorical data. The dataset contains the following variables:

1. **rowid (integer)** - An integer id of the row.
2. **uni_name (character)** - The name of the university.
3. **major(character)** - The subject of the program self-reported by students.
4. **degree (character)** - The type of degree program. The variable takes one of the following values: MS, MA, PhD, MFA, MBA, MEng, and Other.
5. **season (character)** - The season of application. The first letter indicates whether the program starts from the Fall semester or Spring semester, and then the letter is followed by the last 2 digits of the year the program starts.
6. **decision (character)** - The admission decision. It contains five categories - Accepted, Rejected, Wait-listed, Interview and Other.
7. **decision_method (character)** - The method through which decision was communicated.
8. **decision_date (character)** - The date that the decision was communicated.
9. **decision_timestamp (integer)** - Timestamp of the decision.
10. **ugrad_gpa (double)** - The respondent's undergraduate GPA. The scale of the GPA varies because some students use a 10-point scale while others use a 4-point scale.
11. **gre_verbal (double)** - GRE verbal score, which varies from 130 to 170 for the new GRE and from 200 to 800 for the old GRE.

12. **gre_quant (double)** - GRE quantitative score, which varies from 130 to 170 for the new GRE and from 200 to 800 for the old GRE.
13. **gre_writing (double)** - GRE writing score that ranges from 0 to 6.
14. **is_new_gre (logical)** - Whether or not the applicant took the new GRE.
15. **gre_subject (double)** - GRE subject test score on a 200 to 990 score scale.
16. **status (character)** - Status of the candidate. Can be "International", "International with US Degree", "American" or "Other".
17. **post_data (character)** - The date in which the observation was posted on grad cafe.
18. **post_timestamp (integer)** - Timestamp of the post.
19. **comments (character)** - Applicants' comments.

We decided to drop variables which either contain little information such as 'gre_subject', which few candidates reported, and 'rowid' which is redundant, and variables which are not of interest to us, such as 'comments', 'decision_method', 'post_data', and 'post_timestamp'.

Some problems that we may expect to encounter in the data are missing values, biased data due to self-reporting (it may be possible that positive results are more likely to be reported), and possibly fake data. In addition to this, the data will likely require some cleaning as user fill out forms and may be inconsistent (for example school name might be "Boston University"" or "BU"). Lastly, 'ugrad_gpa' could be based on different scales, such as a 10 point scale that is sometimes used internationally.

Research Questions

Our main question of interest is "How do the different variables relate to admission decision?" We are interesting in understanding how the different factors such as GPA, GRE scores, the degree program you are applying to, or status affect whether you are ultimately chosen for admission. We also aim to answer several "sub-questions" such as:

- How do admissions statistics differ across schools?
- How do admissions differ between Boston University and "top tier" schools?
- How have admissions statistics changed over time? (2015, 2016, 2017)?
- Is there a relationship between acceptance rate and season? (Is it easier to get enrolled in Spring semester or Fall semester?)
- Relationship between acceptance rate and student status (American vs International students; International students with a US degree vs those without)
- Does applying earlier make a difference in getting into a school?

We would like to explore these questions to help all of us who are interested in graduate school to better understand the admission process.

Data Import & Cleaning

We start by importanting the neccesary packages and the dataset.

```
library(tidyverse)
```

```
## -- Attaching packages ----------------------------------------------------------- tidyverse 1.2.1
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  2.0.1     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0
```

```
## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts()
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
#library(plotly)
(grad <- read_csv("data/grad.csv",
    col_types = cols_only(
      uni_name=col_character(),
      major=col_character(),
      degree=col_character(),
      season=col_character(),
      decision=col_character(),
      decision_date=col_character(),
      decision_timestamp=col_double(),
      ugrad_gpa=col_double(),
      gre_verbal=col_double(),
      gre_quant=col_double(),
      gre_writing=col_double(),
      is_new_gre=col_logical(),
      status=col_character())))
```

```
## # A tibble: 345,303 x 13
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Univers~ Ms. ~ MS     S16    Accepted (5, 11, 2015)       1446699600
##  2 Vanderb~ Educ~ MS     F16    Other    <NA>                        NA
##  3 Univers~ Publ~ MS     F16    Accepted (16, 11, 201~       1447650000
##  4 Tufts U~ Comp~ PhD    S16    Accepted (16, 11, 201~       1447650000
##  5 Univers~ Theo~ MS     F16    Accepted (16, 11, 201~       1447650000
##  6 Univers~ Mast~ MS     S16    Rejected (14, 11, 201~       1447477200
##  7 Univers~ Publ~ MS     F16    Accepted (12, 11, 201~       1447304400
##  8 Tufts U~ MALD  MS     S16    Accepted (7, 11, 2015)       1446872400
##  9 New Yor~ Fina~ MS     S16    Accepted (15, 11, 201~       1447563600
## 10 Appalac~ Comm~ MS     S16    Accepted (13, 11, 201~       1447390800
## # ... with 345,293 more rows, and 6 more variables: ugrad_gpa <dbl>,
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <dbl>,
## #   is_new_gre <lgl>, status <chr>
```

```
problems(grad)
```

```
## [1] row      col      expected actual
## <0 rows> (or 0-length row.names)
```

As mentioned previously, we drop variables 'gre_subject','rowid','comments', 'decision_method', 'post_data', and 'post_timestamp'. Aside from this we have no problems regarding data import. While the dataset has some missing data, we keep all data for analyzing variation of single variables.

Variation of Single Variables:

First we plot counts for the most popular grad schools and programs are.

```
grad %>% group_by(uni_name) %>% count(uni_name) %>% arrange(desc(n))
```

```
## # A tibble: 4,535 x 2
## # Groups:   uni_name [4,535]
##    uni_name                                    n
##    <chr>                                   <int>
##  1 Columbia University                     10901
##  2 Stanford University                      9071
##  3 University Of California, Berkeley (UCB)  7796
##  4 University Of Michigan, Ann Arbor (UMich)  7585
##  5 Cornell University                       6983
##  6 University Of California, Los Angeles (UCLA)  6788
##  7 Harvard University                       6587
##  8 University Of Washington, Seattle (UW)   6507
##  9 New York University (NYU)                6291
## 10 University Of Pennsylvania (UPenn)       6043
## # ... with 4,525 more rows
```

```
grad %>% group_by(uni_name,major) %>% count(uni_name, major) %>% arrange(desc(n))
```

```
## # A tibble: 73,296 x 3
## # Groups:   uni_name, major [73,296]
##    uni_name                                    major                  n
##    <chr>                                       <chr>              <int>
##  1 Carnegie Mellon University (CMU)            Computer Science     776
##  2 Stanford University                         Computer Science     765
##  3 Georgia Institute Of Technology (GTech)     Computer Science     610
##  4 Columbia University                         Computer Science     585
##  5 University Of Illinois, Urbana-Champaign (UI~ Computer Science     582
##  6 University Of Washington, Seattle (UW)      Computer Science     581
##  7 Stanford University                         Electrical Engineer~ 559
##  8 Boston University (BU)                      Economics            553
##  9 University Of California, San Diego (UCSD)  Computer Science     530
## 10 Columbia University                         Economics            523
## # ... with 73,286 more rows
```

From the tables above we see that the most popular college for grad applications is Columbia University with 10,901 applications. For the most popular specific grad programs, we see that Carnegie Mellon University, Computer Science is the most popular with 776 applications.

Next we assess the selectivity of different grad schools and programs.

```
grad1 <- grad %>% group_by(uni_name) %>%filter(decision == "Accepted") %>% count(uni_name) %>% arrange(
grad2 <- grad %>% group_by(uni_name) %>% count(uni_name) %>% arrange(desc(n))
colnames(grad1)[2] = "accepted"
merge(grad1,grad2,by =("uni_name")) %>% mutate(rate = accepted/n) %>% arrange(rate) %>% head(10)
```

```
##                                uni_name accepted  n       rate
## 1                    Virginia Consortium        2 32 0.06250000
## 2                              Mannheim        1 15 0.06666667
## 3          University Of Central Oklahoma        2 27 0.07407407
## 4                             Montclair        4 48 0.08333333
## 5                 The Petroleum Institute        1 12 0.08333333
## 6                           James Madison        2 22 0.09090909
## 7                         Sloan Kettering        1 11 0.09090909
## 8                          Calvin College        1 10 0.10000000
```

```
## 9                                Rockhurst        1 10 0.10000000
## 10 Nova South Eastern University (online)        1  9 0.11111111
```

```
grad3 <- grad %>% group_by(uni_name,major) %>% filter(decision == "Accepted") %>% count(uni_name,major)
grad4 <- grad %>% group_by(uni_name,major) %>% count(uni_name,major) %>% arrange(desc(n))
colnames(grad3)[3] = "accepted"
merge(grad3,grad4,by=c("uni_name","major")) %>% mutate(rate = accepted/n) %>% arrange(rate) %>% head(10)
```

```
##                                        uni_name                major
## 1                 University Of Colorado, Boulder   Clinical Psychology
## 2                               Yale University   Clinical Psychology
## 3      University Of Maryland, College Park (UMD) Counseling Psychology
## 4                            Stanford University     Social Psychology
## 5        University Of California, Berkeley (UCB)   Clinical Psychology
## 6                         New York University (NYU) Counseling Psychology
## 7  University Of North Carolina, Chapel Hill (UNC)            Accounting
## 8         University Of Texas, Austin (UT Austin) Counseling Psychology
## 9                            Cornell University     Social Psychology
## 10                             Tufts University     Social Psychology
##    accepted  n       rate
## 1         1 53 0.01886792
## 2         1 52 0.01923077
## 3         1 41 0.02439024
## 4         1 38 0.02631579
## 5         1 37 0.02702703
## 6         1 32 0.03125000
## 7         1 32 0.03125000
## 8         1 30 0.03333333
## 9         1 28 0.03571429
## 10        1 28 0.03571429
```

```
merge(grad1,grad2,by =("uni_name")) %>% mutate(rate = accepted/n) %>% filter(uni_name == "Boston Univers
```

```
##                uni_name accepted    n      rate
## 1 Boston University (BU)     2262 5280 0.4284091
```

```
merge(grad3,grad4,by=c("uni_name","major")) %>% mutate(rate = accepted/n) %>% filter(uni_name == "Bosto
```

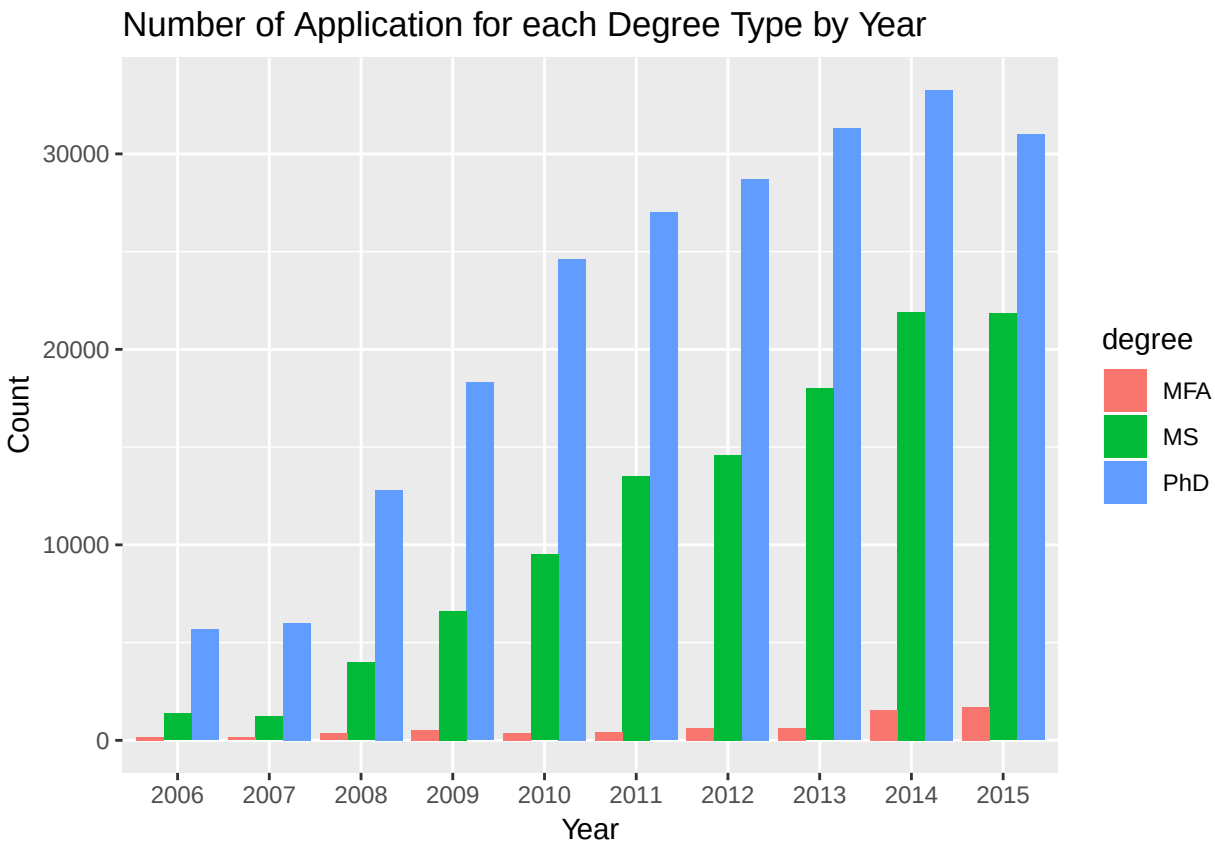
```
##                 uni_name                                          major
## 1  Boston University (BU)                        Genetic Counseling
## 2  Boston University (BU)                                   Finance
## 3  Boston University (BU)                       Counseling Psychology
## 4  Boston University (BU)                                 Marketing
## 5  Boston University (BU)                               Neuroscience
## 6  Boston University (BU)                 Program In Biomedical Sciences
## 7  Boston University (BU)                       Anthropology (cultural)
## 8  Boston University (BU) Communcation Disorders Speech Language Pathology
## 9  Boston University (BU)                                 Counseling
## 10 Boston University (BU)                          European History
##    accepted  n       rate
## 1         1 20 0.05000000
## 2         8 85 0.09411765
## 3         2 21 0.09523810
## 4         1  8 0.12500000
## 5         3 24 0.12500000
## 6         1  8 0.12500000
```

```
## 7          2 14 0.14285714
## 8          1  7 0.14285714
## 9          1  7 0.14285714
## 10         2 14 0.14285714
```

Surprisingly, the most selective grad school and grad program are Virginia Consortium with an acceptance rate of 6.25% and University of Colorado, Boulder, Clinical Psychology with an acceptance rate of 1.89% respectively. We also observe that the BU as a whole is not too selective with an acceptance rate of 42.8%. We also obeserve that the most selective major at BU is genetic counseling with only 1 student accepted out of 20 (5%).

Next, we look at the change in number of application over time.

```
# Decision reported over time (2015, 2016, 2017)?
grad$decision_date <- grad$decision_date %>% str_replace("\\(", "") %>% str_replace("\\)", "")
# Create a dataset for plotting number of application verses year
grad_year = grad %>% select(degree, decision_date)  %>%
  filter(!is.na(decision_date)) %>%
  mutate(yr = str_match(decision_date, "...\\d$")) %>%
  filter(degree == "MFA" | degree == "MS" | degree == "PhD") %>%
  filter(as.integer(yr) < 2016 ) %>% filter(as.integer(yr) > 2005)
grad_year$decision_date <- NULL
# plot
grad_year %>% group_by(yr, degree) %>% ggplot(aes(x = as.factor(yr), fill = degree)) + geom_bar(positio
labs(x ="Year",
     y ="Count",
     title="Number of Application for each Degree Type by Year")
```

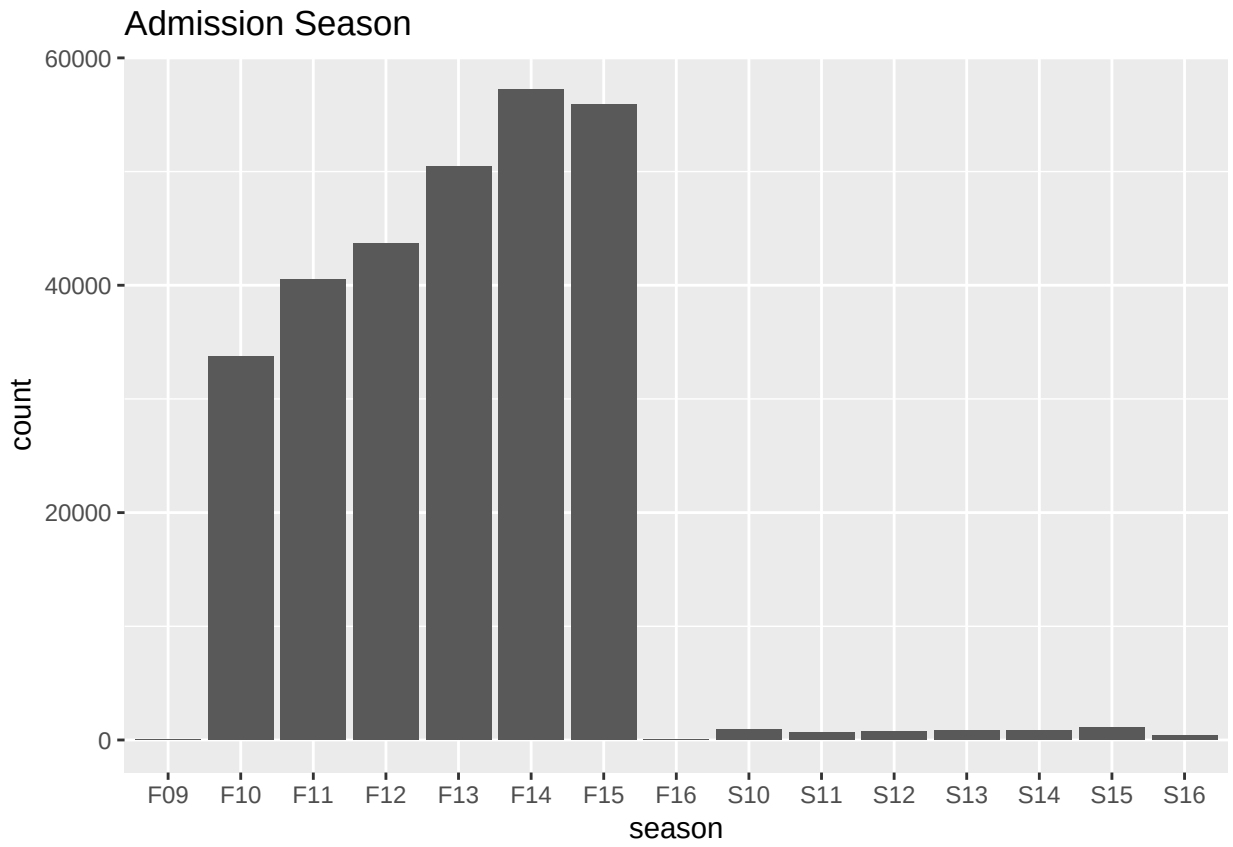


The dataset has official data of report from 2006 to 2015. The application report of three degrees, MFA, MS,

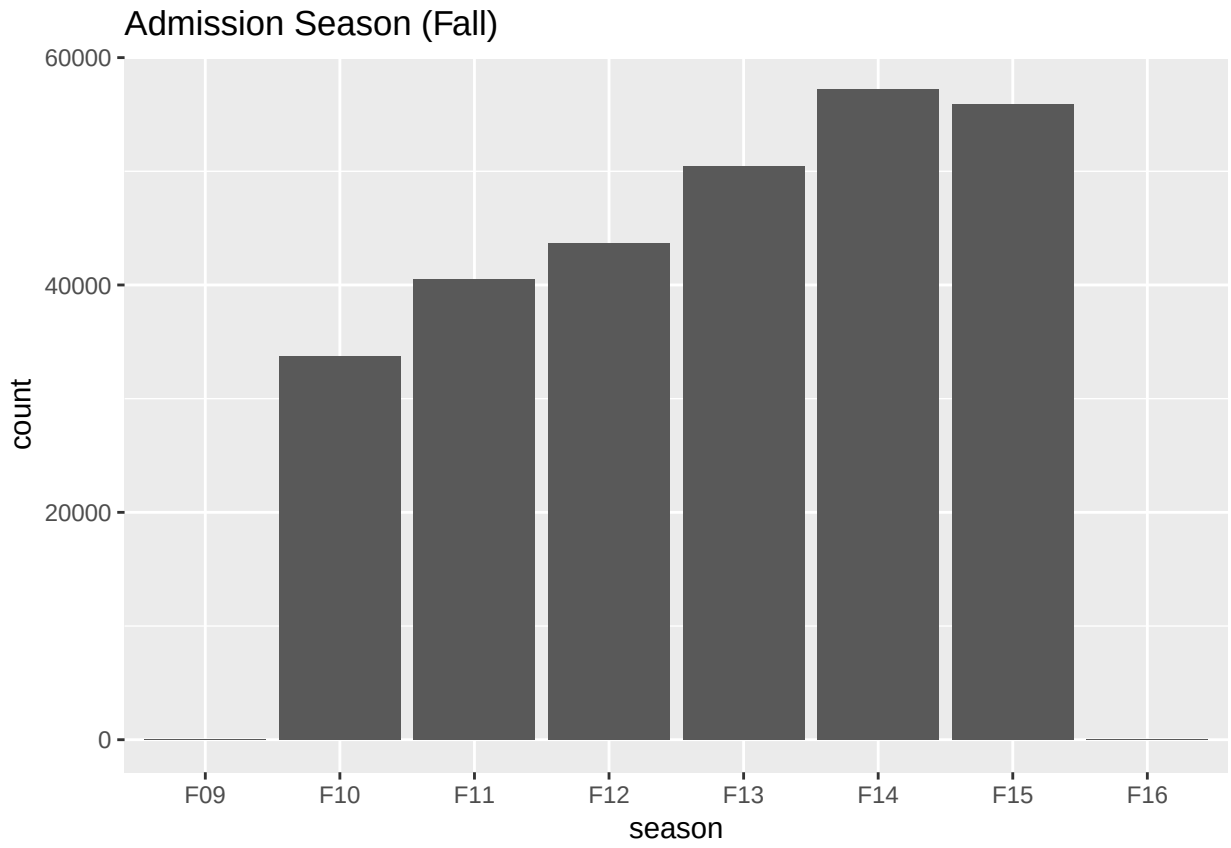
PhD increase each year until 2015. The overall shape has a plateauing trend.

Next, we look at the distribution of applications by season and the counts of each admission decision.

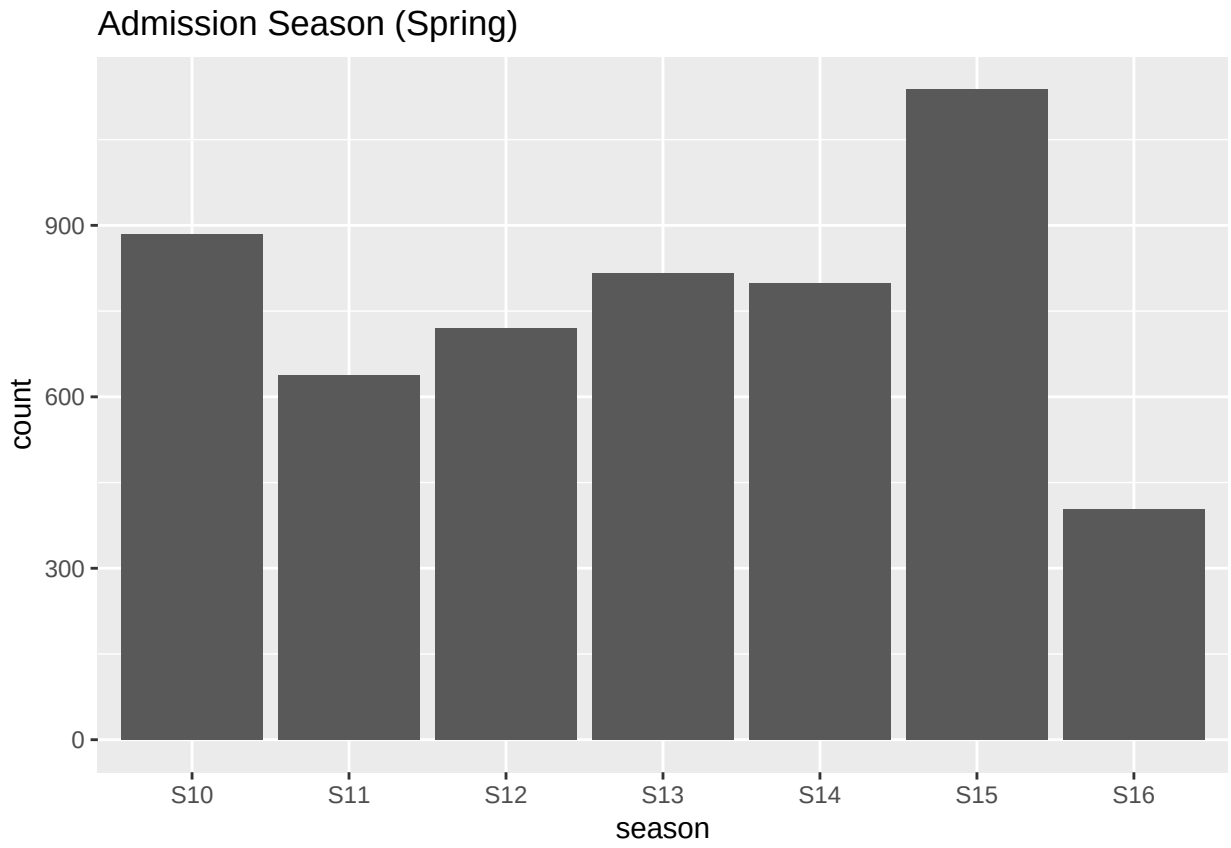
```
# bar chart for season
# bar chart includes both Spring semester and Fall semester
grad %>%
  filter(!is.na(season)) %>%
  ggplot() +
  geom_bar(aes(x = season)) +
  labs(title = "Admission Season")
```



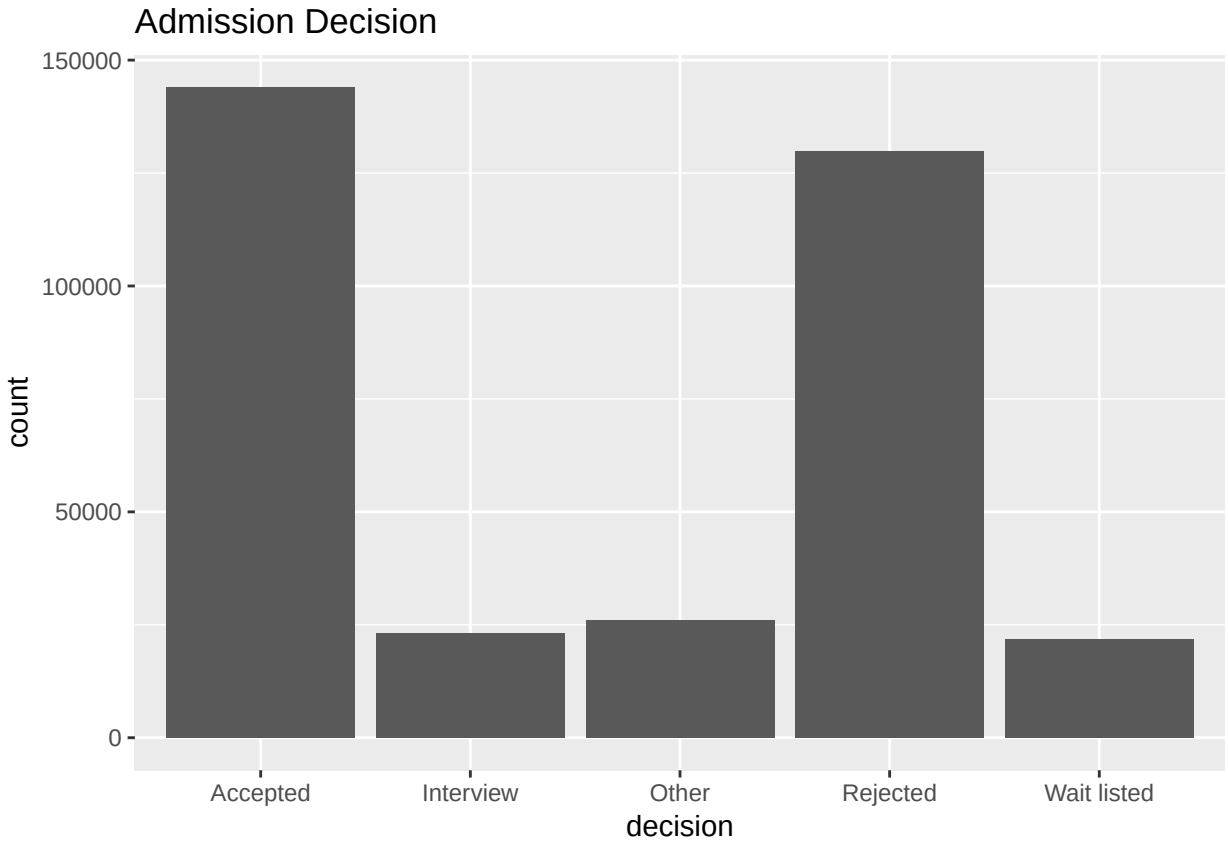

```
# bar chart for Fall semester only
grad %>%
  filter(!is.na(season)) %>%
  group_by(fall = str_match(season, "^\\F..")) %>%
  filter(!is.na(fall)) %>%
  ggplot() +
  geom_bar(aes(x = season)) +
  labs(title = "Admission Season (Fall)")
```




```
# bar chart for Spring semester only
grad %>%
  filter(!is.na(season)) %>%
  group_by(fall = str_match(season, "^\\F..")) %>%
  filter(is.na(fall)) %>%
  ggplot() +
  geom_bar(aes(x = season)) +
  labs(title = "Admission Season (Spring)")
```




```
# bar chart for decision
grad %>%
  filter(!is.na(decision)) %>%
  ggplot() +
  geom_bar(aes(x = decision)) +
  labs(title = "Admission Decision")
```

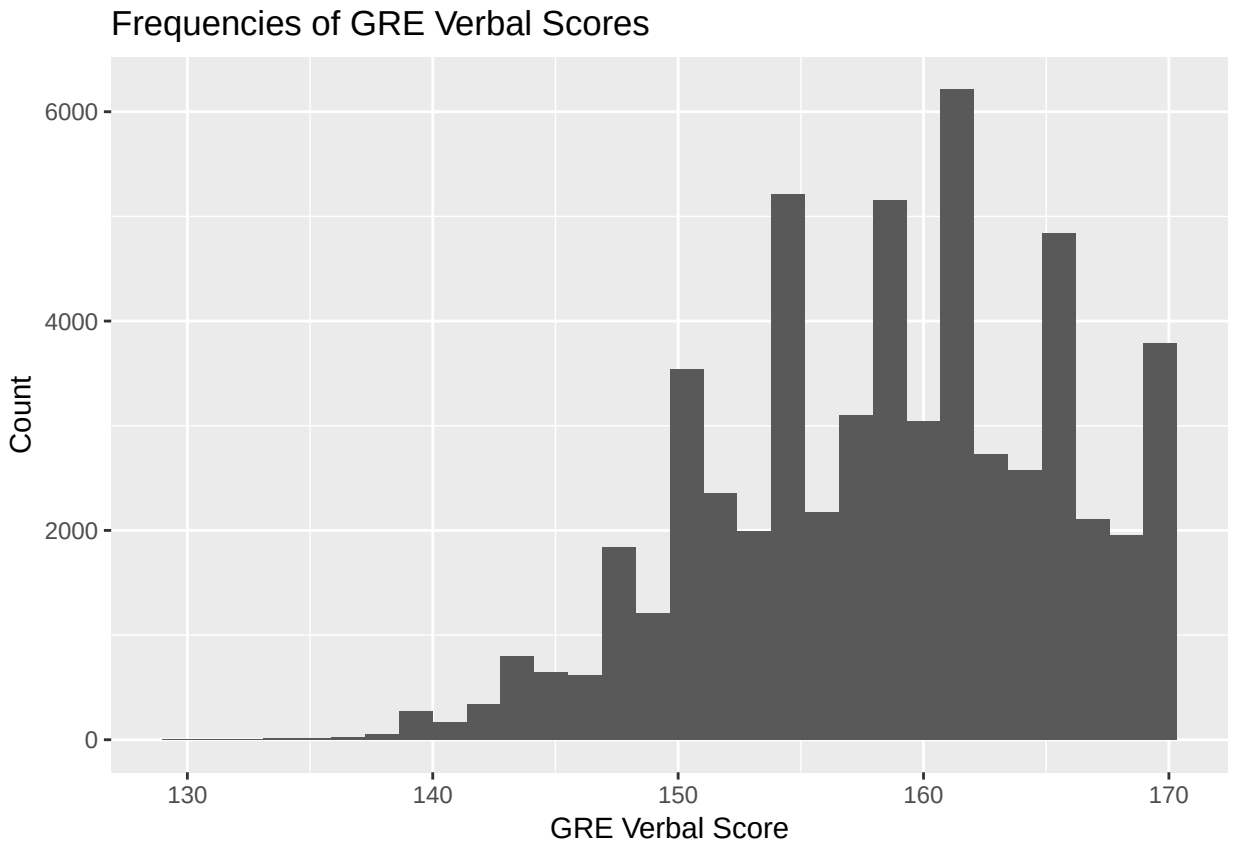


We see that the total number of application are signicantly higher for the fall semester than the Spring semester. Focusing only on fall semesters, graduate school enrollment clearly has shown a positive trend over the years. Spring enrollment does not show an explicit pattern. When looking at the distribution of decisions, the majority of candidates either recieve of report mainly acceptances and rejections, while a few candiates recieve other forms of responses such as waitlist, interview, or "other."

Next, we plot the distribution of GRE test scores, and GPA. Because these is a variable "is_new_gre", which distinguished between old and new GRE, we filter for only new GRE scores, as the majority of observations report new GRE scores.

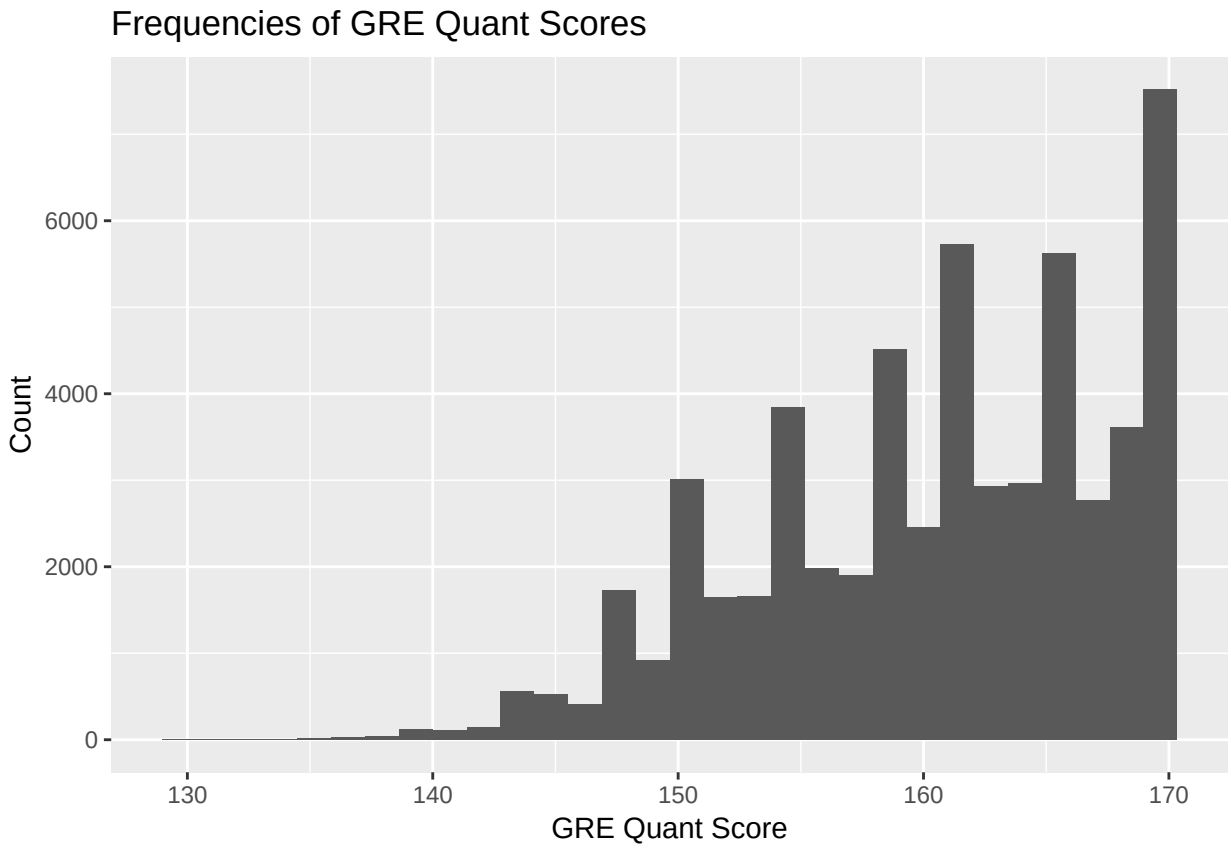
```
# GRE Verbal
grad %>% select(gre_verbal ,is_new_gre) %>%
filter(is_new_gre == TRUE & is.na(gre_verbal)!= TRUE ) %>% ggplot + geom_histogram(aes(gre_verbal)) +
labs(x ="GRE Verbal Score",
     y ="Count",
     title="Frequencies of GRE Verbal Scores")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



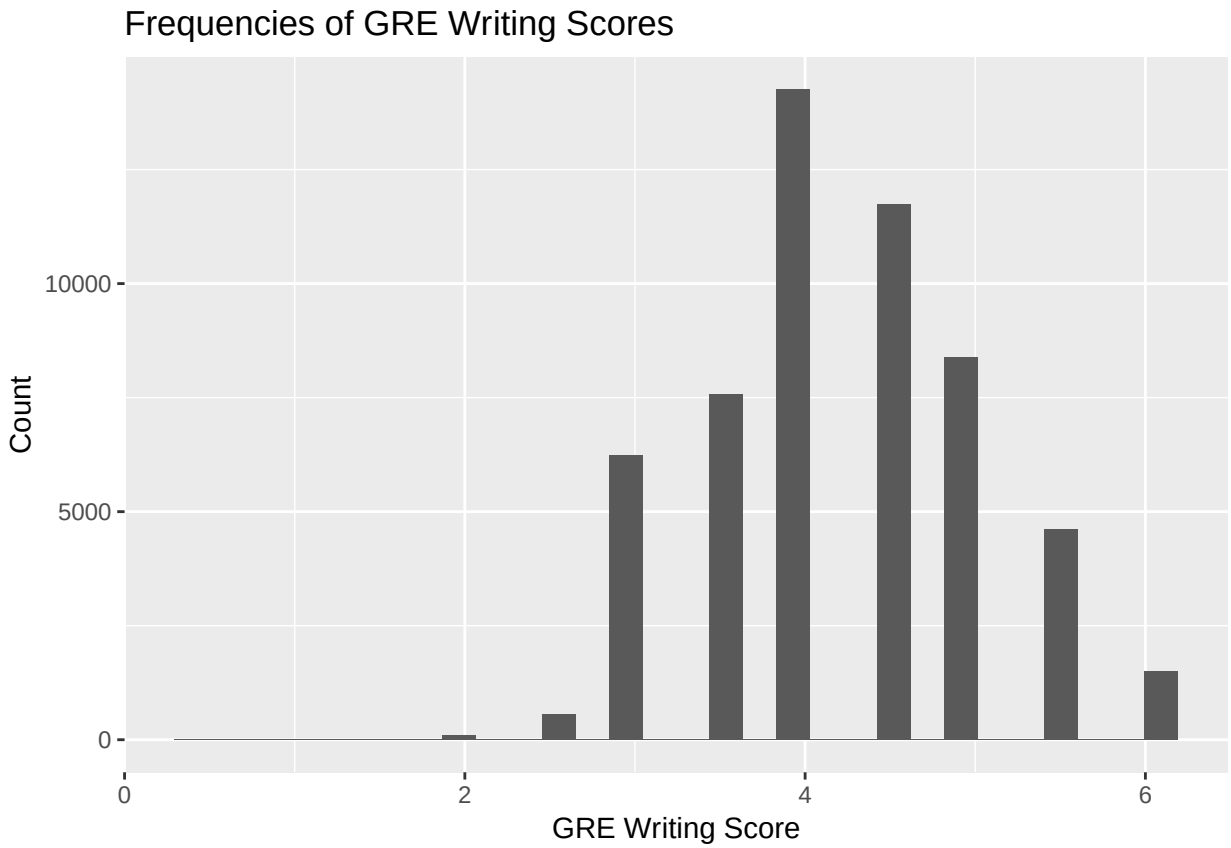

```
# GRE quant
grad %>% select(gre_quant ,is_new_gre) %>%
  filter(is_new_gre == TRUE & is.na(gre_quant)!= TRUE ) %>% ggplot + geom_histogram(aes(gre_quant)) +
labs(x ="GRE Quant Score",
     y ="Count",
     title="Frequencies of GRE Quant Scores")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



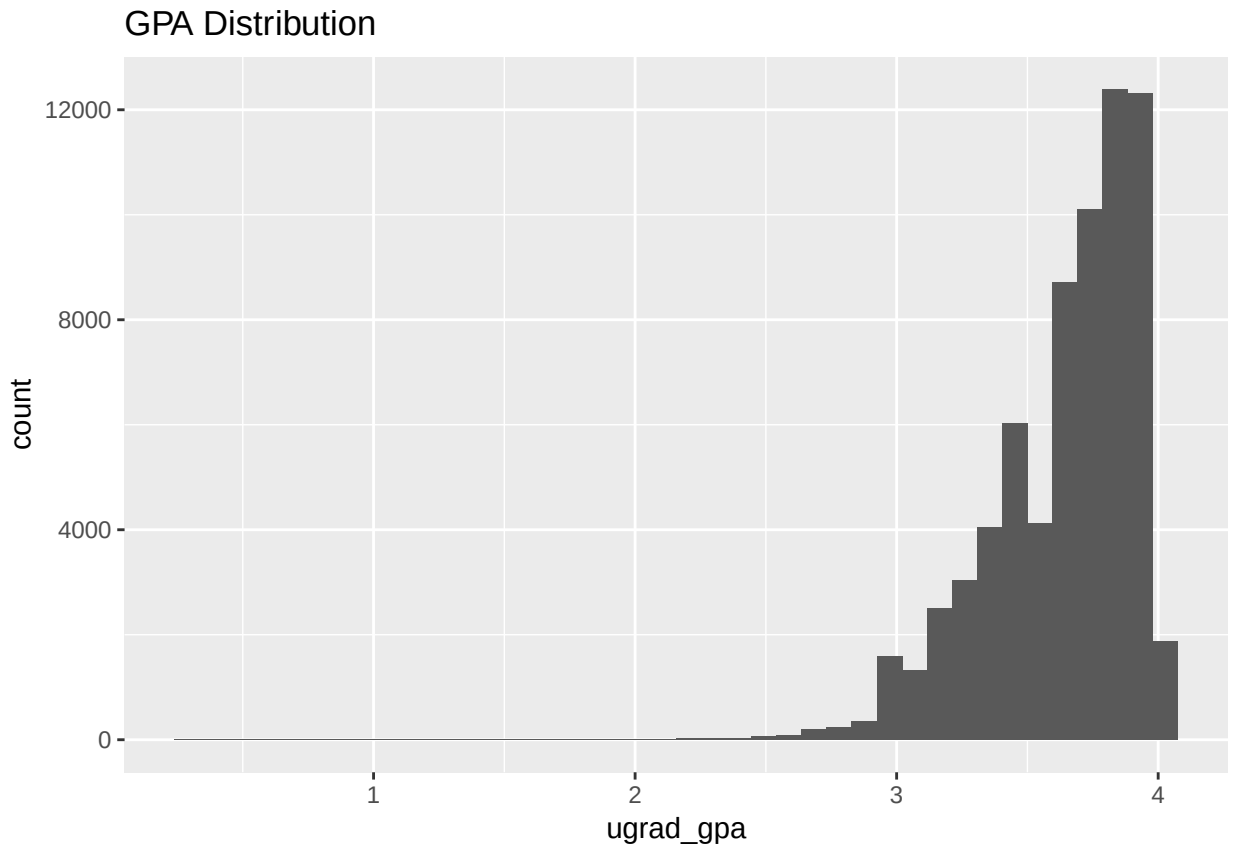

```
# GRE writing
grad %>% select(gre_writing ,is_new_gre) %>%
  filter(is_new_gre == TRUE & is.na(gre_writing)!= TRUE) %>% ggplot + geom_histogram(aes(gre_writing)) +
labs(x ="GRE Writing Score",
     y ="Count",
     title="Frequencies of GRE Writing Scores")
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



We see from the above histograms that GRE verbal scores range from 130 to 170 with a bell shape. Most of them concentrate 155 - 165. GRE quant score range from 130 to 170 with step like shape. Scores tend to concentrate 160 - 170. GRE writing scores range from 2 to 6 with a bell like shape. Most people get a score of 4.

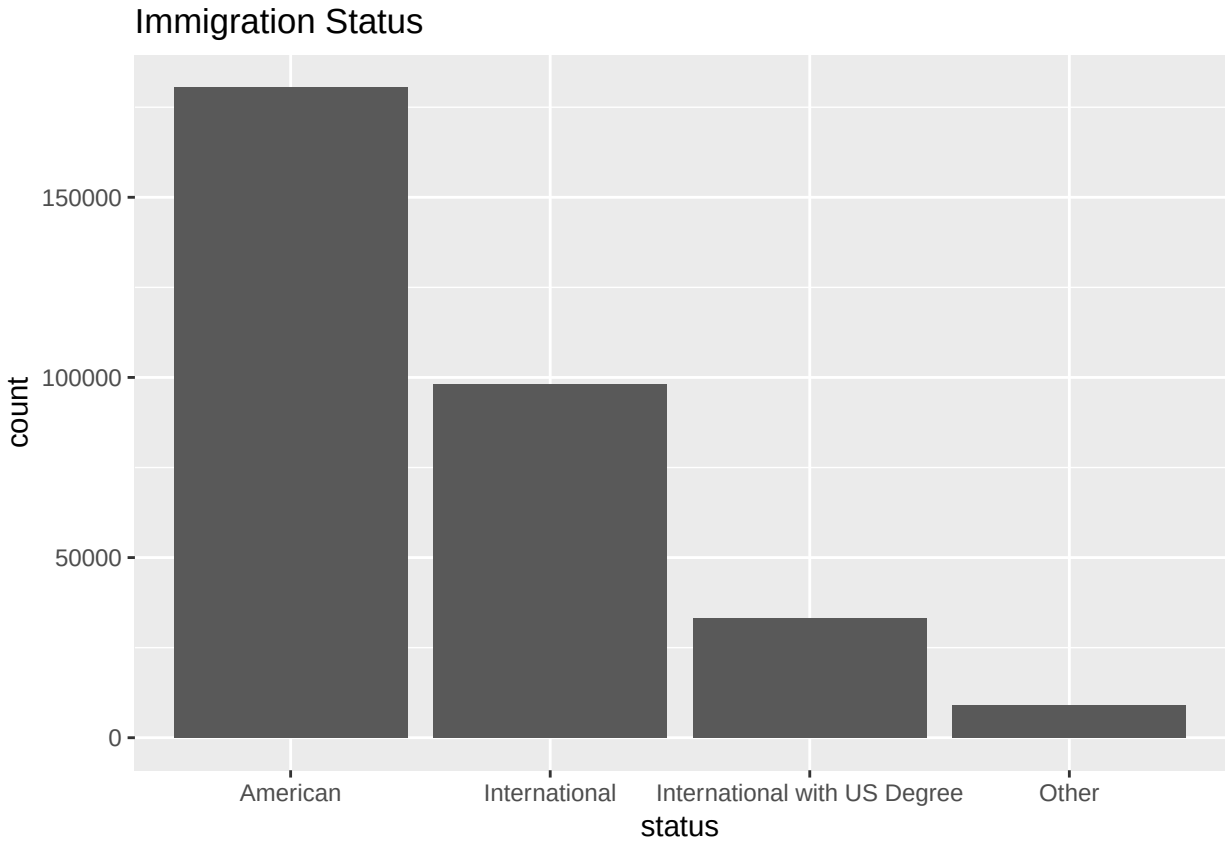
```
grad %>% filter(!is.na(ugrad_gpa) & ugrad_gpa < 4.0) %>%
  ggplot(aes(ugrad_gpa)) + geom_histogram(bins = 40) + labs(titles = "GPA Distribution")
```



We see that the distribution of GPAs for the observations tend to be left skewed, with the majority of candidates having more than 3.6 GPA. This is accepted as grad programs tend to look at GPA as a major factor, and students who aim to attend a grad school would likely have higher GPAs.

Lastly, we look at the distribution of student status (internation, US, international with US degree, etc)

```
grad %>% filter(!is.na(status)) %>%
  mutate(count = n()) %>%
  ggplot(aes(x = status)) + geom_bar() +
  labs(titles = "Immigration Status")
```



From the chart above, we see that the majority of students applying are American. In Immigration Status, around 60% of applicants are American and the rest of them are international students. We can tell that a big amount of graduate or Ph.D. students are coming from an international background.

Covariation Between Multiple Variables

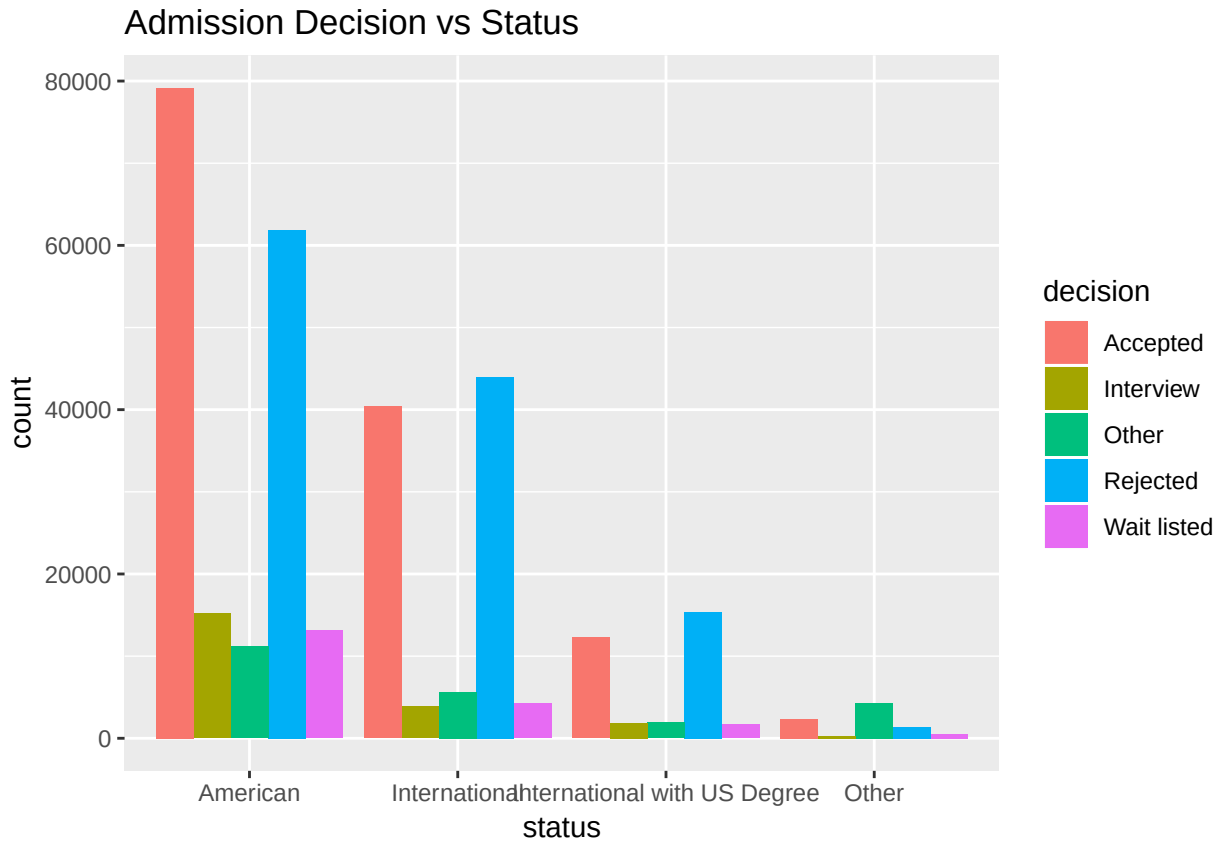
One covariation of interest is the influence of student status (internation, US, etc.) vs admission decision.

```
# student identity vs acceptance rate
# table for student status vs decision
(grid <- grad %>%
  filter(!is.na(status), !is.na(decision)) %>%
  group_by(status, decision) %>%
  summarise(count = n()) %>%
  spread(key = decision, value = count))
```

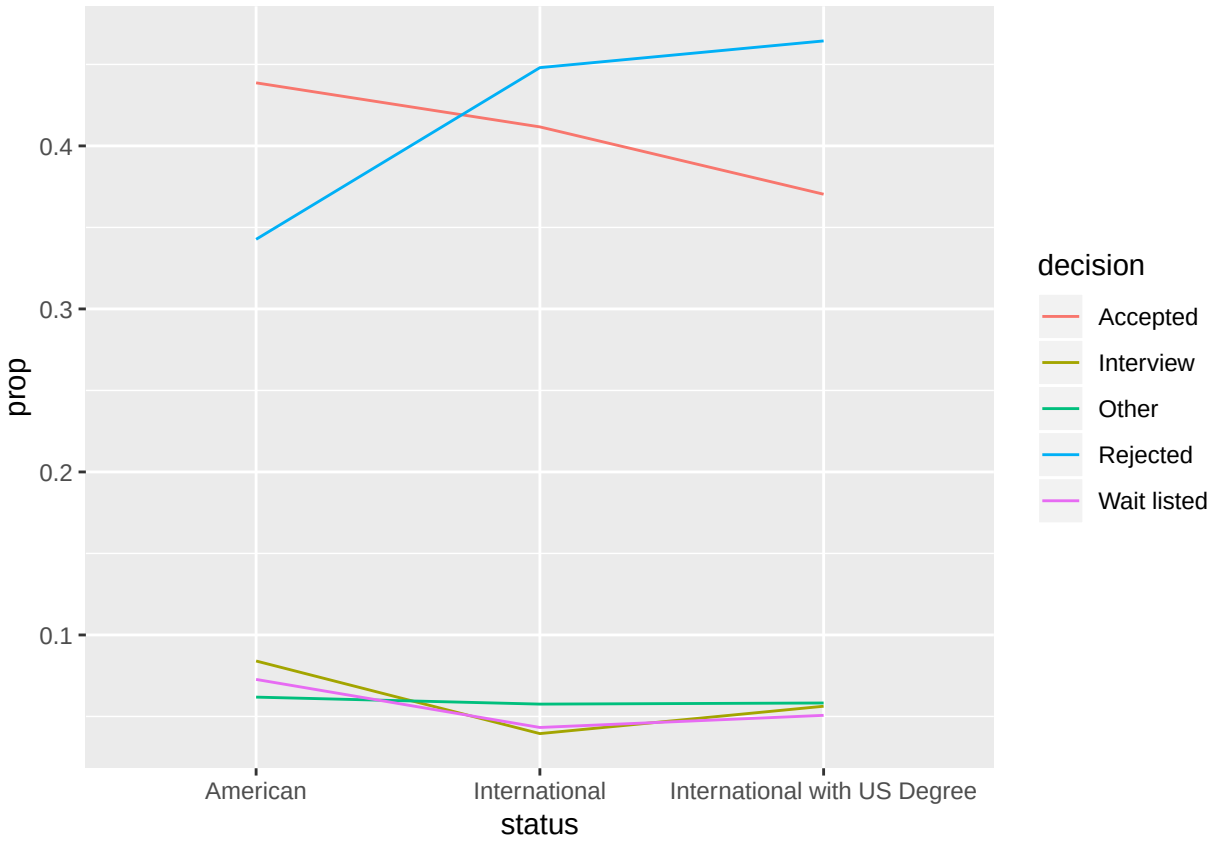
```
## # A tibble: 4 x 6
## # Groups:   status [4]
##   status                    Accepted Interview Other Rejected `Wait listed`
##   <chr>                        <int>     <int> <int>    <int>         <int>
## 1 American                     79164     15164 11163    61841         13123
## 2 International                40404      3875  5651    43973          4245
## 3 International with US De~    12282      1866  1932    15401          1681
## 4 Other                         2351       276  4327     1375           536
```

```
# bar chart
grad %>%
  filter(!is.na(decision), !is.na(status)) %>%
```

```
  ggplot() +
  geom_bar(aes(status, fill = decision), position = "dodge") +
  labs(title = "Admission Decision vs Status")
```



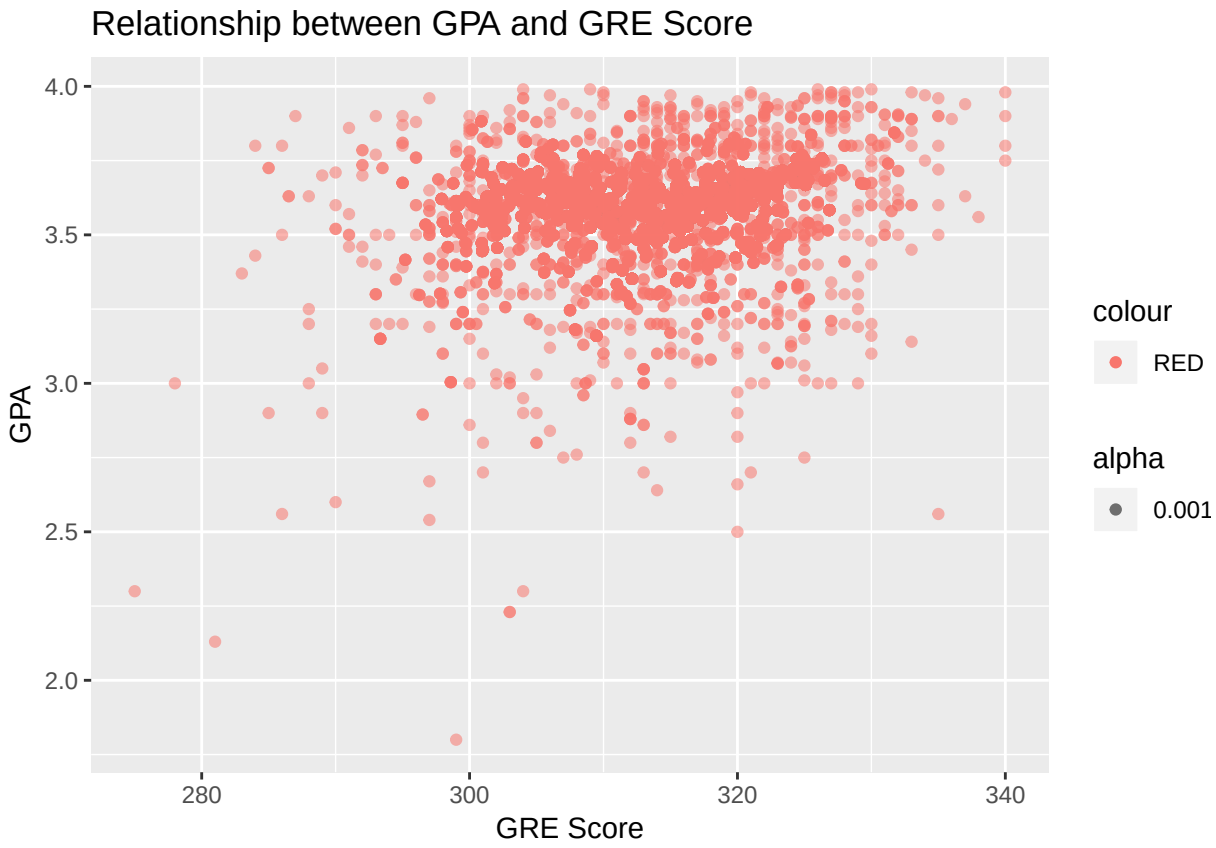

```
# acceptance rate based on student status
grid$sum = rowSums(grid[, c("Accepted", "Interview", "Other", "Rejected", "Wait listed")])
grid %>%
  gather('Accepted', 'Interview', 'Other', 'Rejected', 'Wait listed',
         key = "decision", value = "cases") %>%
  mutate(prop = cases/sum) %>%
  filter(status != "Other") %>%
  ggplot(aes(group = decision, status, prop, color = decision)) +
  geom_line()
```



From the charts above, it seems that US based students tend to have higher acceptance rates than international students, and internation with US degree students. The bar chart shows that, for American students, the number of getting accepted is higher than the number of getting rejected. However, for international students and international students with US degree, the number of acceptance is lower than the number of rejection. To further investigate if international students are treated differently, we calculate the decision rate. For each status category, we divide the total number of each admission decision by total number of students to get the decision rate. From the plot we can tell that the proportion of getting accepted is higher for American students than international students, and the proportion of getting rejected is higher for international students with US degree.

Another covariation of interest is the relationship between GPA and GREE scores. For this we summed GRE verbal and GRE quant to get the full GRE score, and created a scatter plot against GPA. We filtered GPA to be less than 4, as GPA of different scales are not comparable.

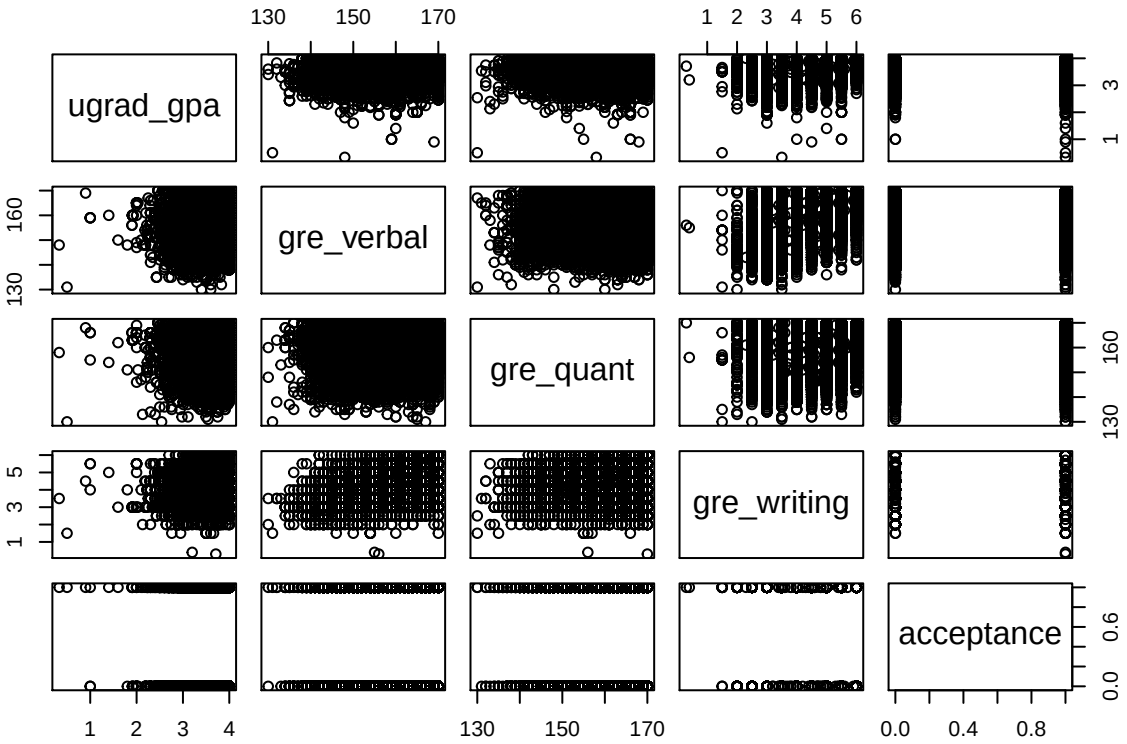
```
grad %>% filter(!is.na(ugrad_gpa|gre_verbal|gre_quant)& ugrad_gpa < 4 & ugrad_gpa >1, is_new_gre == TRU
ggplot(aes(x = mean_GRE, y = mean_gpa)) + geom_point(aes(color = "RED", alpha = 0.001)) +
  labs(titles = "Relationship between GPA and GRE Score",
       y = "GPA",
       x = "GRE Score")
```



From the plot above the relationship between GPA and GRE seems to be positively correlated but is not as strong of a relationship as we expected. Most GPAs tend to be on the higher range: people densely fall into the range between 3.5 and 3.75; GRE scores seem to be more variable across application: scores for all applicants concentrate in the range between 300 and 325 with more outliers.

Lastly, to measure the correlation across all continuous variables, we create a scatterplot matrix, and correlation matrix.

```
g <- grad[complete.cases(grad),] %>% mutate(acceptance = decision == "Accepted") %>% filter(ugrad_gpa<=
```



This plot is somewhat unclear due to the verse dense concentration of the datapoints. In the next step, we will likely use regression to model the probability of acceptance based on the different covariates.

Discussion:

From this exploratory data analysis, we confirm many of the hypotheses that we had going into this project. For example, we confirmed our hypotheses that more students apply in the fall semester than the spring semester and that the majority of students applying (or reported) are American. We confirmed relationships between variables such as GPA and GRE scores. We learned several things as well. For example, we learned the distribution of GPA is left skewed, and GRE scores have an abnormal distribution, with several "spikes" among certain scores. We were surprised to see that American students tended to have higher rates of acceptance than international students.

While this is a very interesting and robust dataset to analyze, there are also several problems we encountered. First, the dataset is not very clean, as it is self-reported. For example, the names of Universities and Majors are not always consistent. For example, some students may write "Boston University (BU)" while others write the name of the specific college at BU such as "Boston University - Metropolitan College." We also noticed that scales of scores and GPA are not always consistent. For example, GPA is most often reported on a 4.0 scale, however, some responses included other scales such as 10 point scale. These will all be problems that we have to work around when going into modeling.