MA415 Final Paper

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

Abstract

In this project, we analyzed the dataset provided by Grad cafe website. Our main goal is to investigate the question of how different variables, such as GRE score and undergraduate GPA, relate to the admission decision. In this deliverable, we focus our analysis on graduate applications to US Top 10 computer science programs. We aim to better understand the factors affecting admission decisions by fitting a logistic model to the data. The covariates we use for the model include undergrad GPA, GRE Scores, Student Status, and interaction terms. We used the model to predict the probability of a student getting accepted. The model shows that, for an American student with average grade, the probability of getting is 49%. The probability for an average international student is 39%. For an average international student with a US degree, the probability is 46%. While our model was able to predict to some degree, which students were more likely to be accepted, the predicted probabilities were too variable to be useful for prediction. This is likely due to many variables being missing from the data such as research experience, reccomendations, reputation of undergraduate institution and so on.

Introduction

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.

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.

```
## # A tibble: 416 x 2
## # Groups:   uni_name [416]
##    uni_name                                      n
##    <chr>                                     <int>
##  1 Carnegie Mellon University (CMU)           1414
##  2 Georgia Institute Of Technology (GTech)     985
##  3 University Of California, San Diego (UCSD)  954
##  4 University Of Illinois, Urbana-Champaign (UIUC)  954
##  5 Stanford University                         914
##  6 University Of California, Berkeley (UCB)    844
##  7 Purdue University                           745
##  8 University Of Washington, Seattle (UW)      713
##  9 University Of Texas, Austin (UT Austin)     706
## 10 Cornell University                          674
## # ... with 406 more rows
```

```
## # A tibble: 2,749 x 3
## # Groups:   uni_name, major [2,749]
##    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 (UIUC) Computer Science   582
##  6 University Of Washington, Seattle (UW)    Computer Science   581
##  7 University Of California, San Diego (UCSD)  Computer Science   530
##  8 University Of California, Berkeley (UCB)  Computer Science   521
##  9 University Of California, Los Angeles (UCLA)  Computer Science   498
## 10 Cornell University                        Computer Science   478
## # ... with 2,739 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.

```
##                                                  uni_name accepted   n
## 1                California Institute Of Technology (Caltech)        5 124
## 2  King Abdullah University Of Science And Technology (KAUST)       15 116
## 3                    University Of California, Berkeley (UCB)      155 844
## 4                  Pennsylvania State University (PennState)        8  37
## 5                Massachusetts Institute Of Technology (MIT)      136 600
## 6                  Washington University, St. Louis (WUSTL)       31 135
## 7           Toyota Technological Institute At Chicago (TTIC)        7  30
## 8                    University Of Washington, Seattle (UW)      167 713
## 9                                      Princeton University       91 382
## 10                                          Yale University       43 167
##          rate
## 1  0.04032258
## 2  0.12931034
## 3  0.18364929
## 4  0.21621622
## 5  0.22666667
## 6  0.22962963
## 7  0.23333333
## 8  0.23422160
## 9  0.23821990
## 10 0.25748503
```

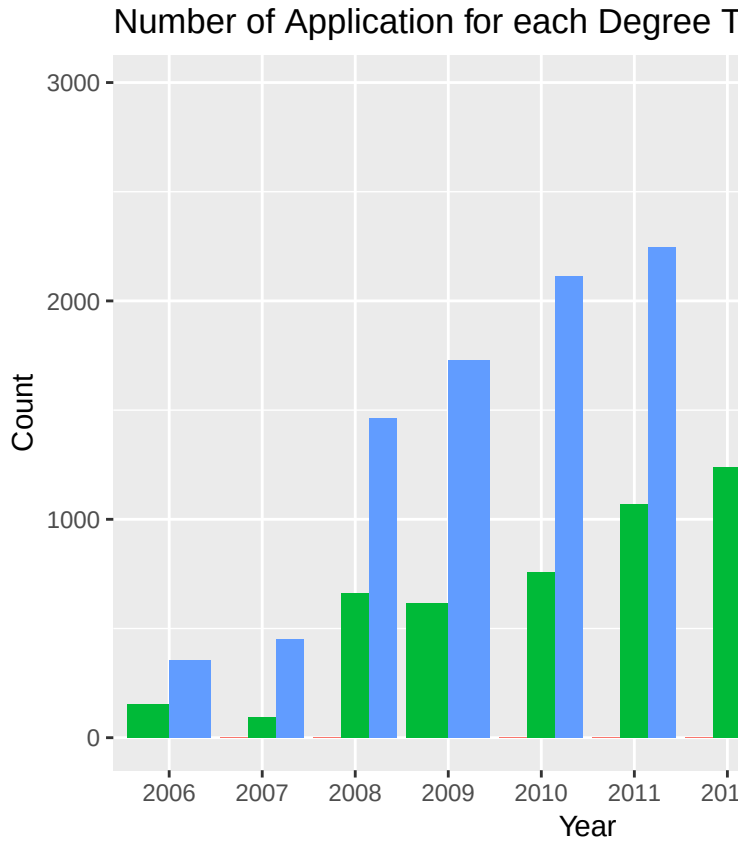
```
##                                       uni_name
## 1    California Institute Of Technology (Caltech)
## 2     Massachusetts Institute Of Technology (MIT)
## 3                 Carnegie Mellon University (CMU)
## 4     Massachusetts Institute Of Technology (MIT)
## 5     Massachusetts Institute Of Technology (MIT)
## 6          University Of Washington, Seattle (UW)
## 7                                Brown University
## 8        University Of California, Berkeley (UCB)
## 9  University Of Illinois, Urbana-Champaign (UIUC)
## 10                                Duke University
##                                           major accepted   n
## 1                                Computer Science        4 100
## 2                           EECS (Computer Science)        1  20
## 3                             Computer Science (CS)        1  19
## 4  Electrical Engineering And Computer Science EECS        1  14
## 5         (ECE) Electrical And Computer Engineering        1  12
## 6         (ECE) Electrical And Computer Engineering        1  12
## 7                                Computer Scicence        1  10
## 8                             Computer Science (CS)        1  10
## 9                                Computer Sceince        1  10
## 10          ECE (Electrical & Computer Engineering)        1   9
##          rate
## 1  0.04000000
## 2  0.05000000
## 3  0.05263158
## 4  0.07142857
## 5  0.08333333
## 6  0.08333333
## 7  0.10000000
```

```
## 8  0.10000000
## 9  0.10000000
## 10 0.11111111
```

```
##                 uni_name accepted   n      rate
## 1 Boston University (BU)       79 184 0.4293478
```

```
##                  uni_name                                     major
## 1  Boston University (BU) (ECE) Electrical And Computer Engineering
## 2  Boston University (BU)                           Computer Science
## 3  Boston University (BU)       Electrical And Computer Engineering
## 4  Boston University (BU)   ECE (Electrical & Computer Engineering)
## 5  Boston University (BU) ECE (Electrical And Computer Engineering)
## 6  Boston University (BU)  ECE(Electrical And Computer Engineering)
## 7  Boston University (BU) Electrical And Computer Engineering (ECE)
## 8  Boston University (BU)                       Computer Engineering
## 9  Boston University (BU) Electrical and Computer Engineering (ECE)
## 10 Boston University (BU)         Electrical & Computer Engineering
##    accepted   n      rate
## 1         2   9 0.2222222
## 2        38 101 0.3762376
## 3         6  13 0.4615385
## 4         5  10 0.5000000
## 5         2   4 0.5000000
## 6         1   2 0.5000000
## 7         1   2 0.5000000
## 8        10  18 0.5555556
## 9         5   9 0.5555556
## 10        4   6 0.6666667
```

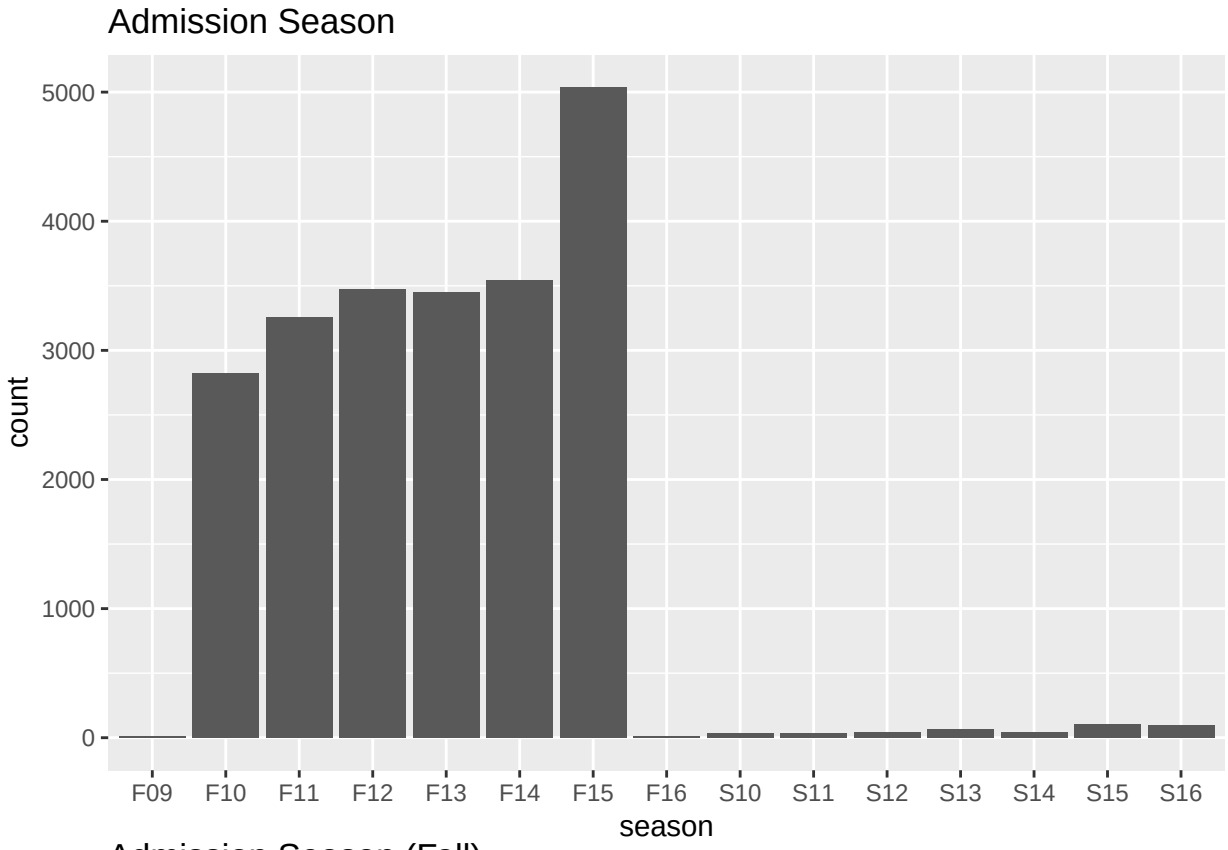
Surprisingly, the most selective grad school and grad program are Virginia Consortium with an acceptance rate of 6.25% and University of Colorado, Boulder, Clinical Psychology with an acceptance rate of 1.89% respectively. We also observe that the BU as a whole is not too selective with an acceptance 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%).

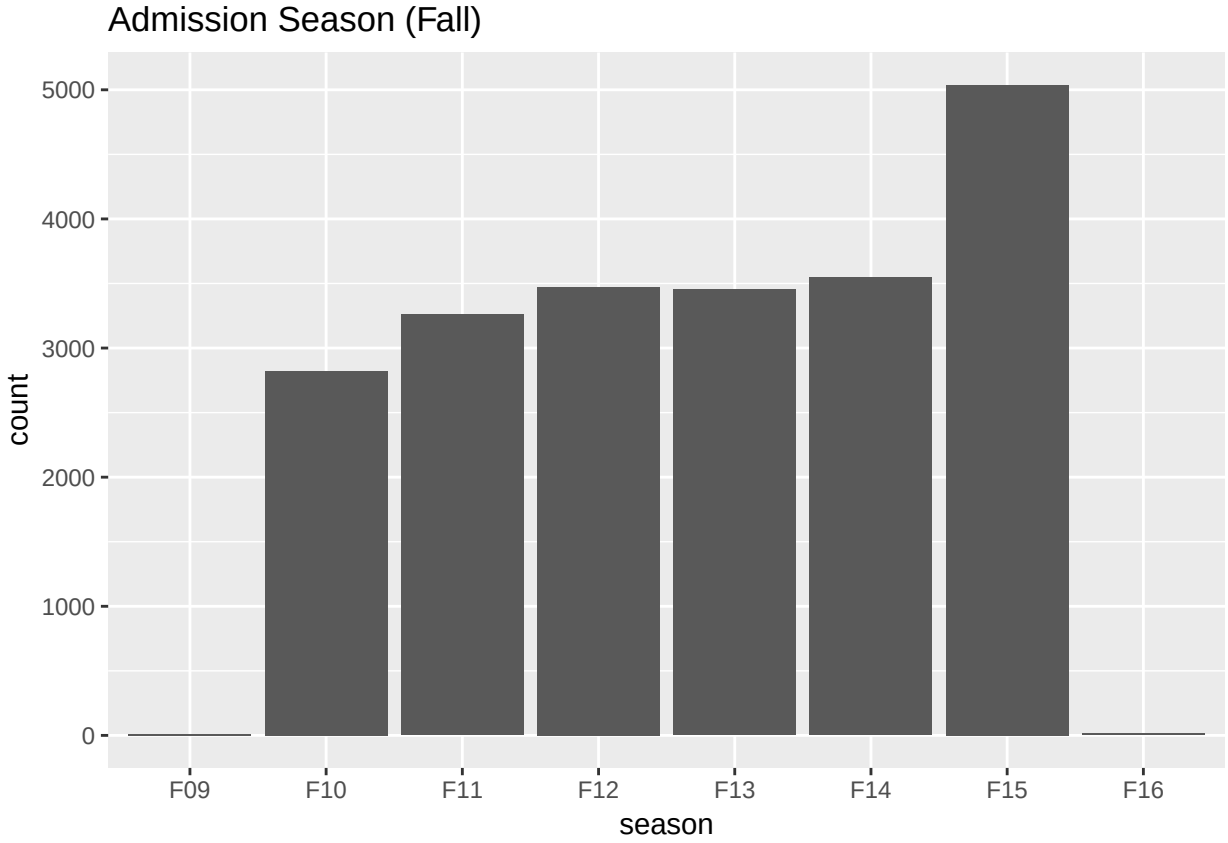


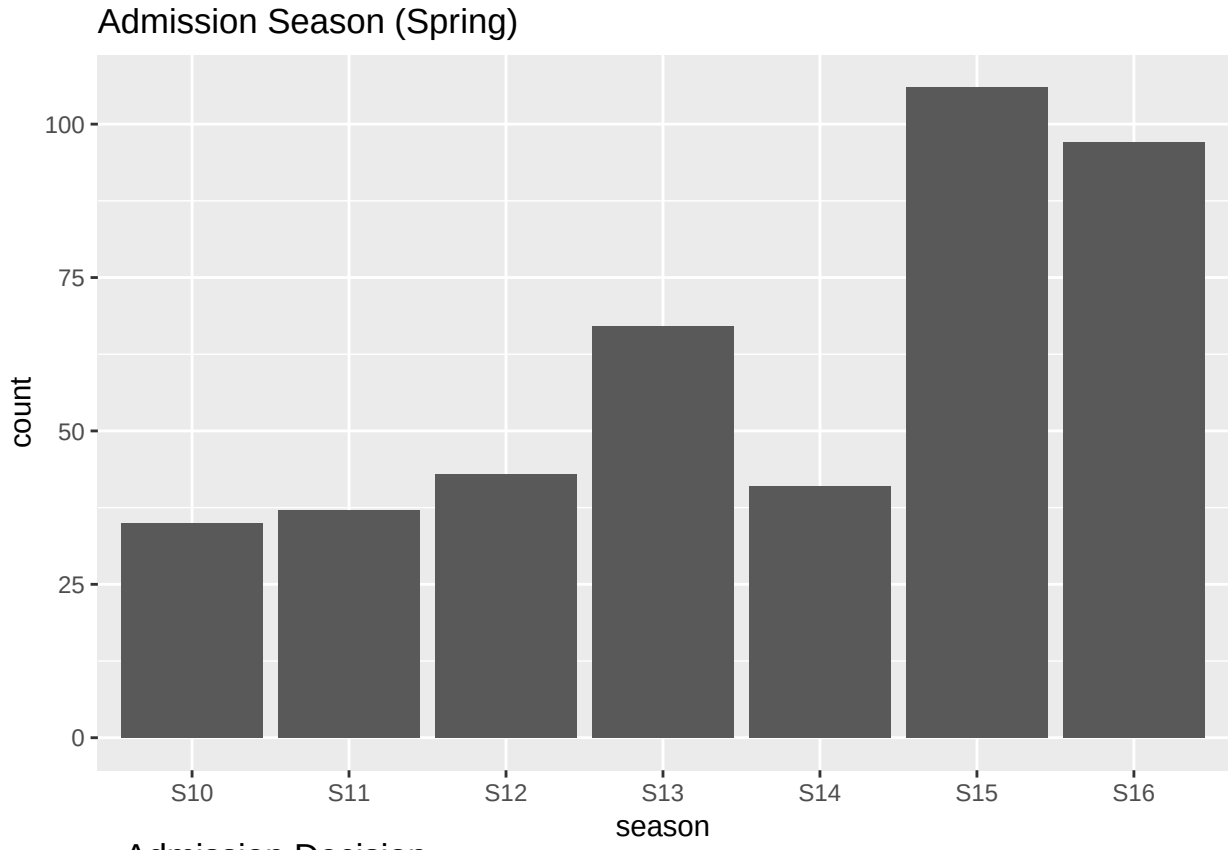
Next, we look at the change in number of application over time.

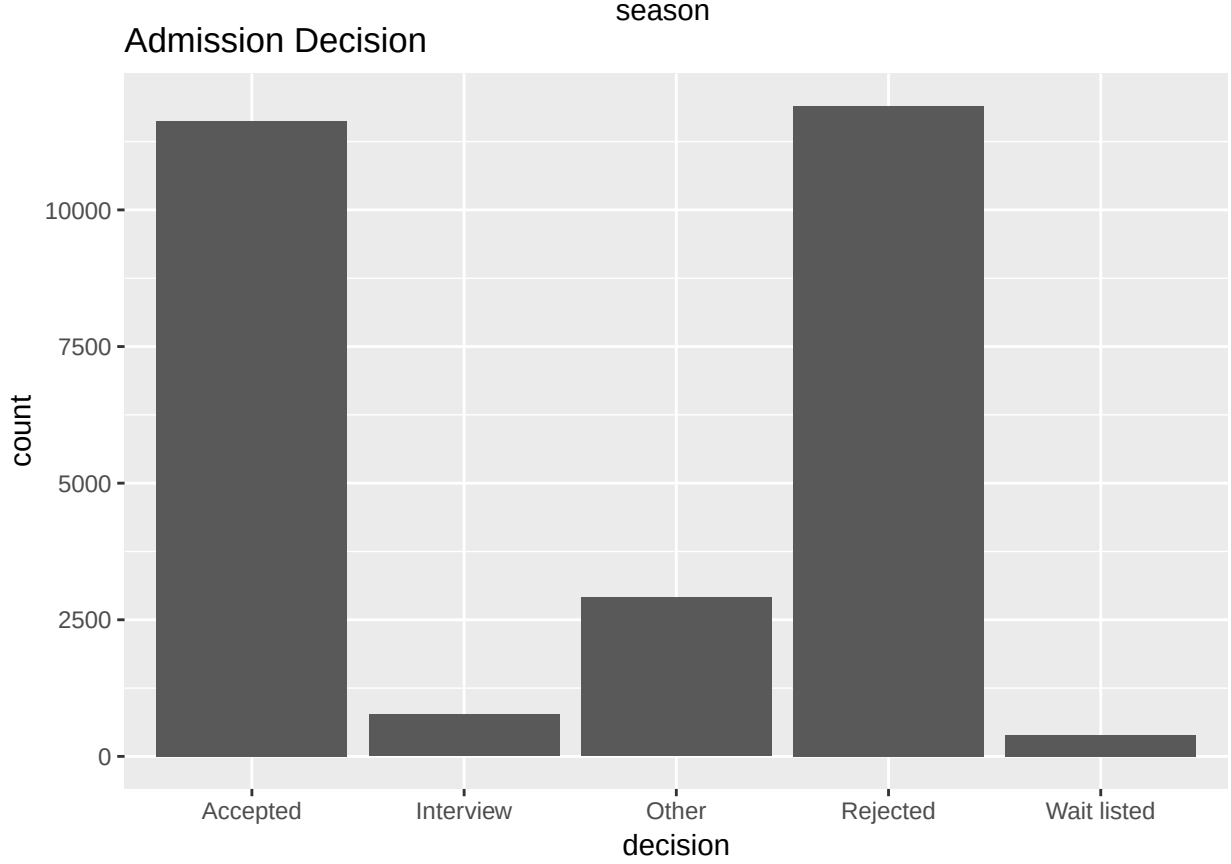
The dataset has official data of report from 2006 to 2015. The application report of three degrees, MFA, MS, PhD increase each year until 2015. The overall shape has a plateauing trend.

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


Admission Season
count
5000
4000
3000
2000
1000
0
F09
F10
F11
F12
F13
F14
F15
F16
S10
S11
S12
S13
S14
S15
S16
season


Admission Season (Fall)
count
5000
4000
3000
2000
1000
0
F09
F10
F11
F12
F13
F14
F15
F16
season



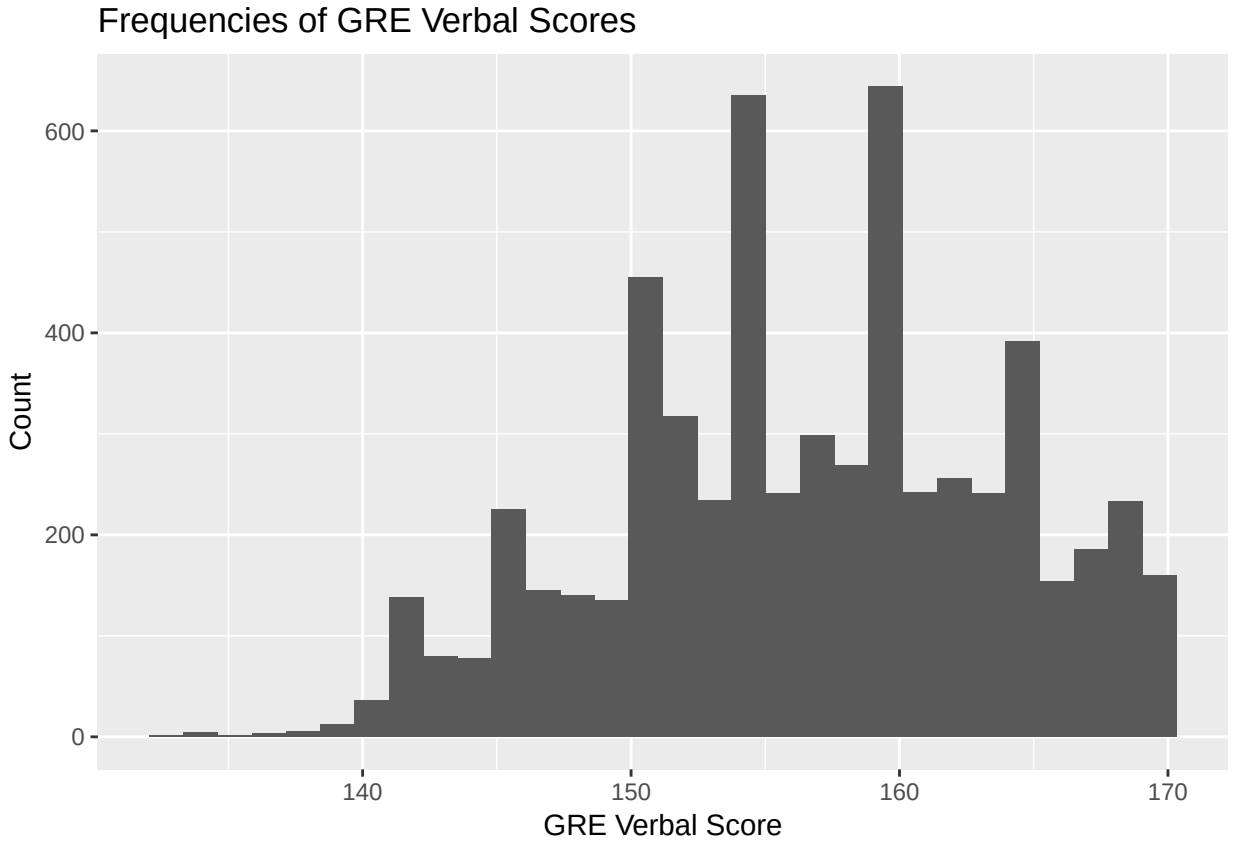


We see that the total number of application are signicantly higher for the fall semester than the Spring

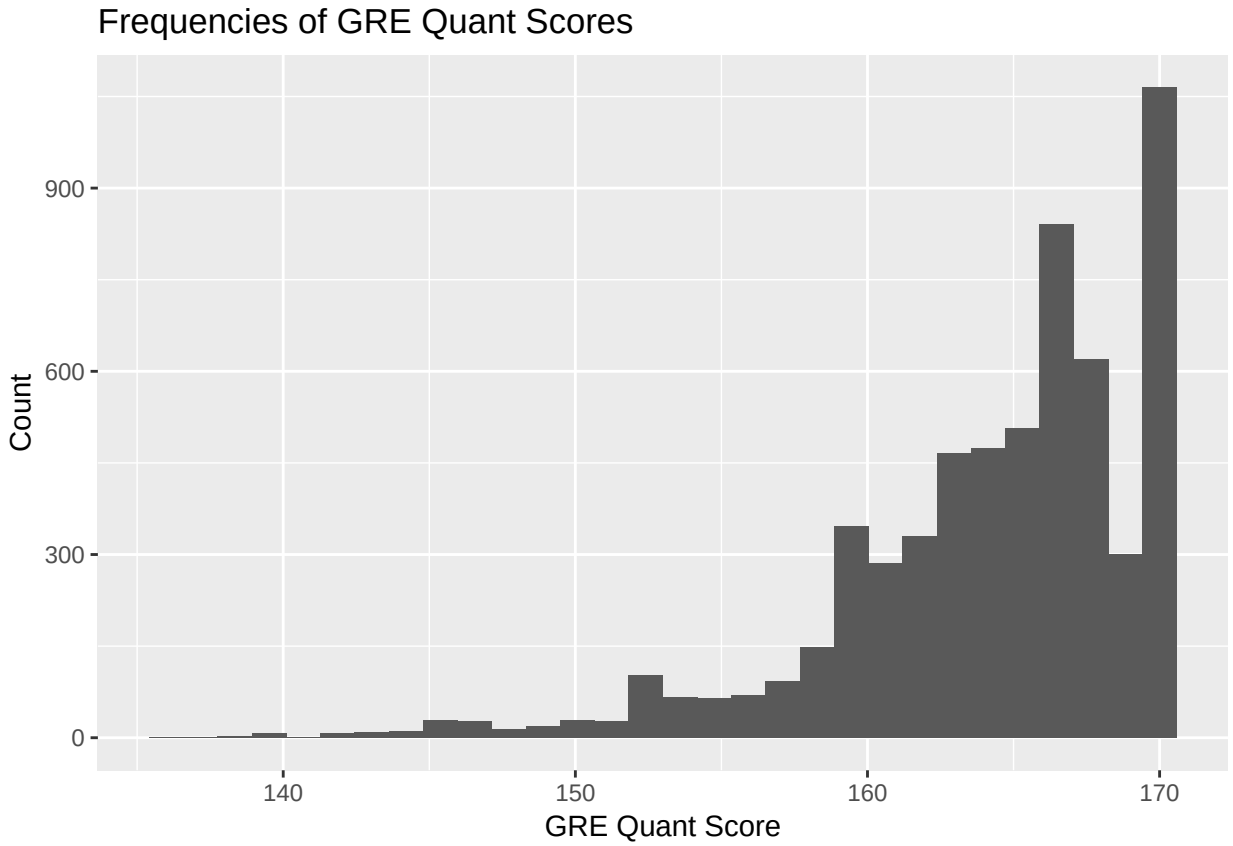
semester. Focusing only on fall semesters, graduate school enrollment clearly has shown a positive trend over the years. Spring enrollment does not show an explicit pattern. When looking at the distribution of decisions, the majority of candidates either recieve of report mainly acceptances and rejections, 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.

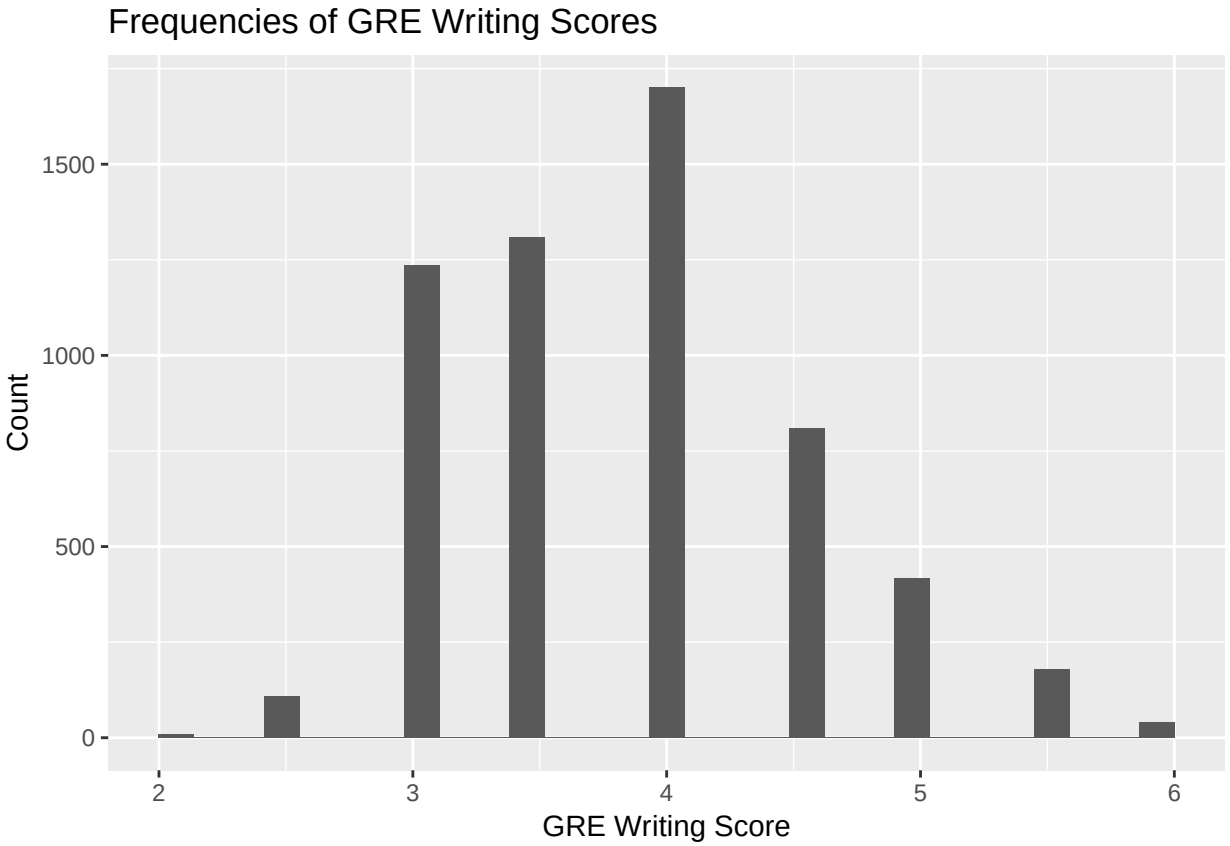
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



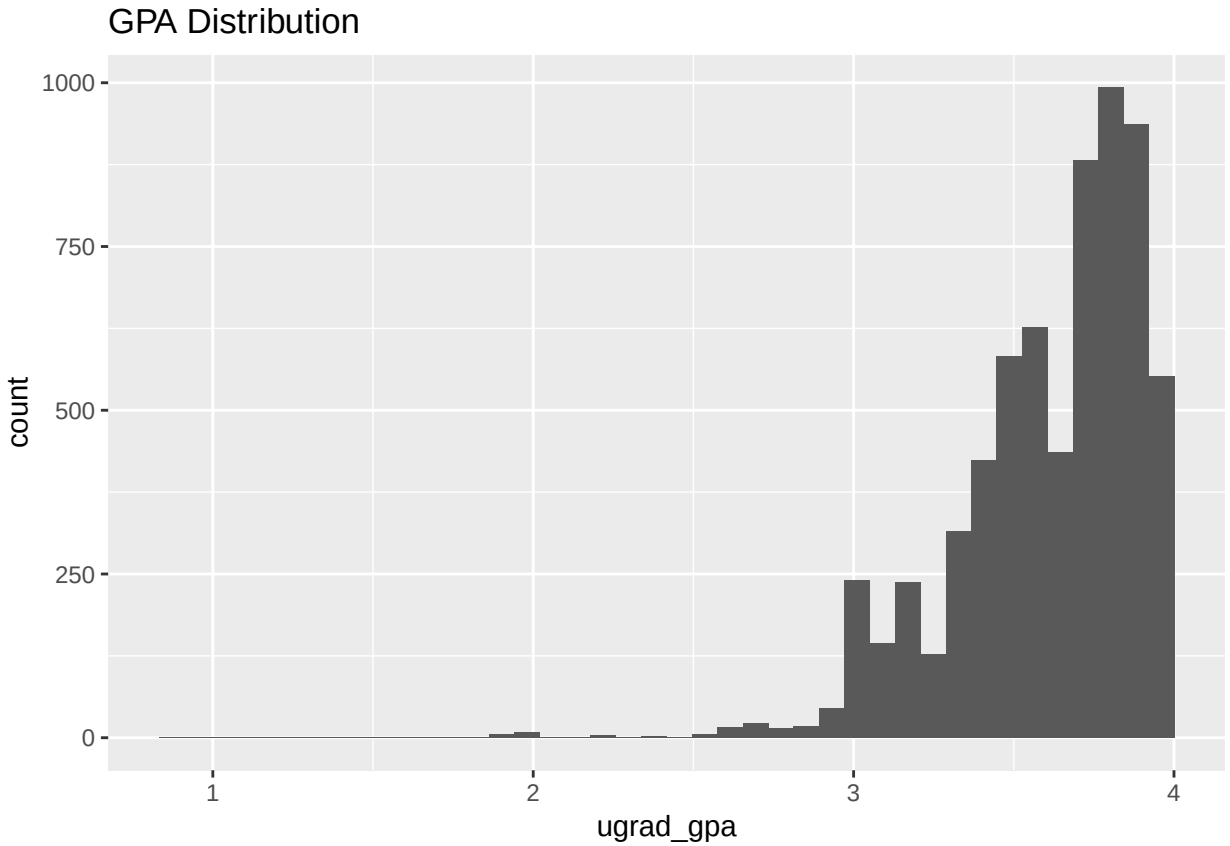

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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

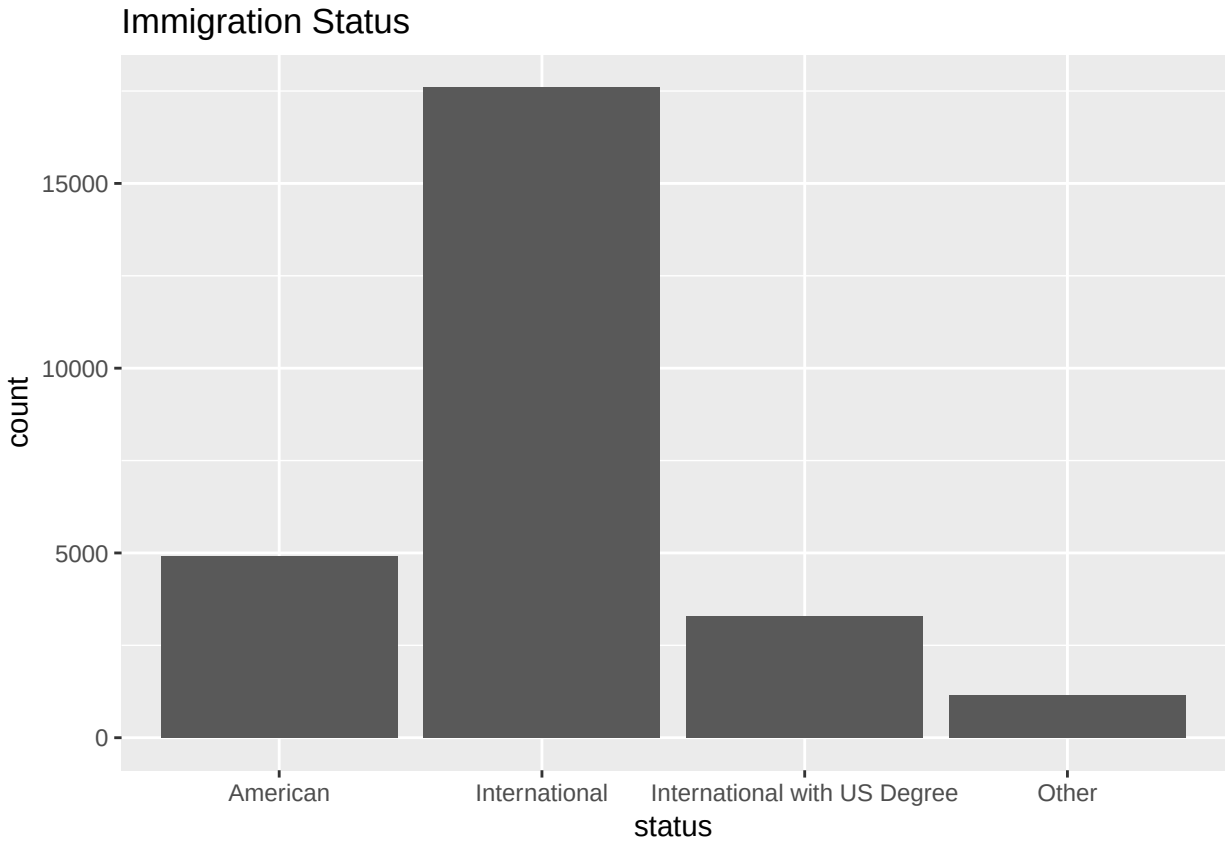


We see from the above histograms that GRE verbal scores range from 130 to 170 with a bell shape. Most of them concentrate 155 - 165. GRE quant score range from 130 to 170 with step like shape. Scores tend to concentrate 160 - 170. GRE writing scores range from 2 to 6 with a bell like shape. Most people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



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)

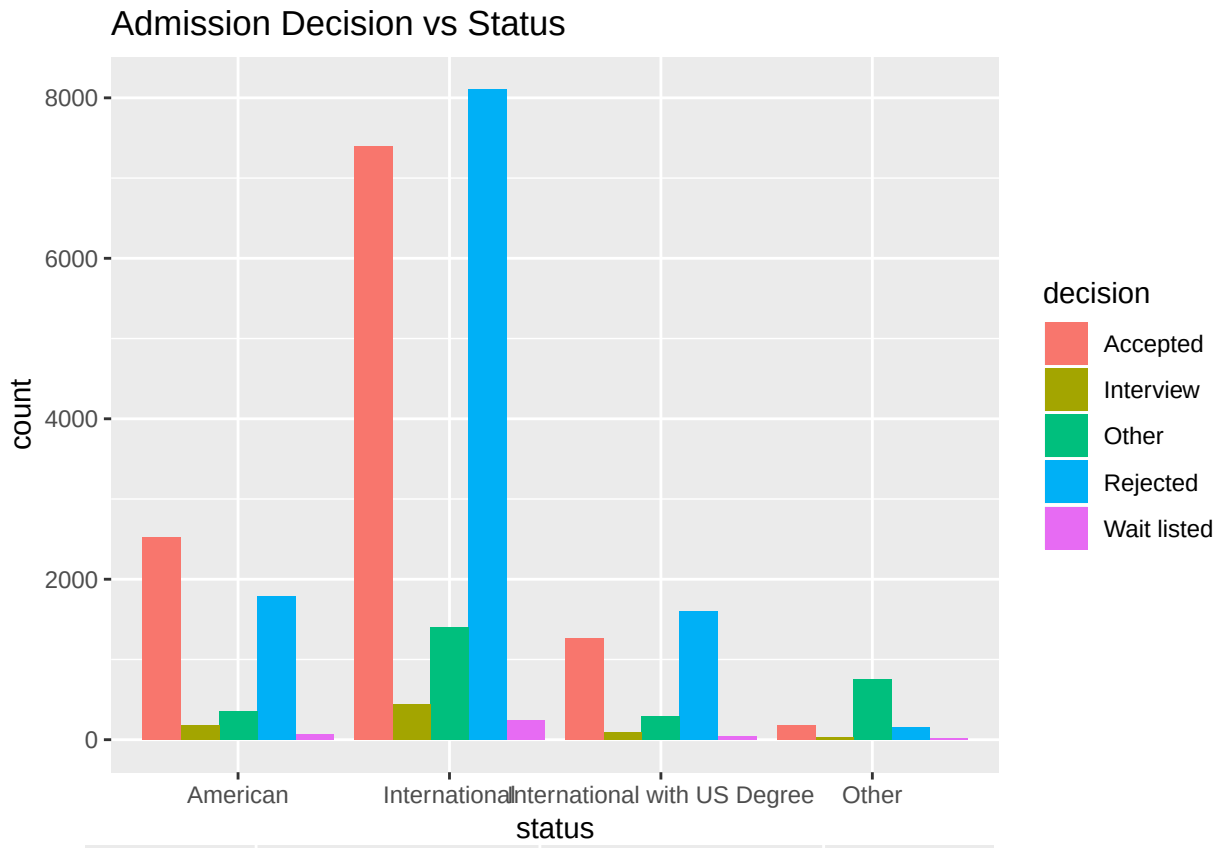


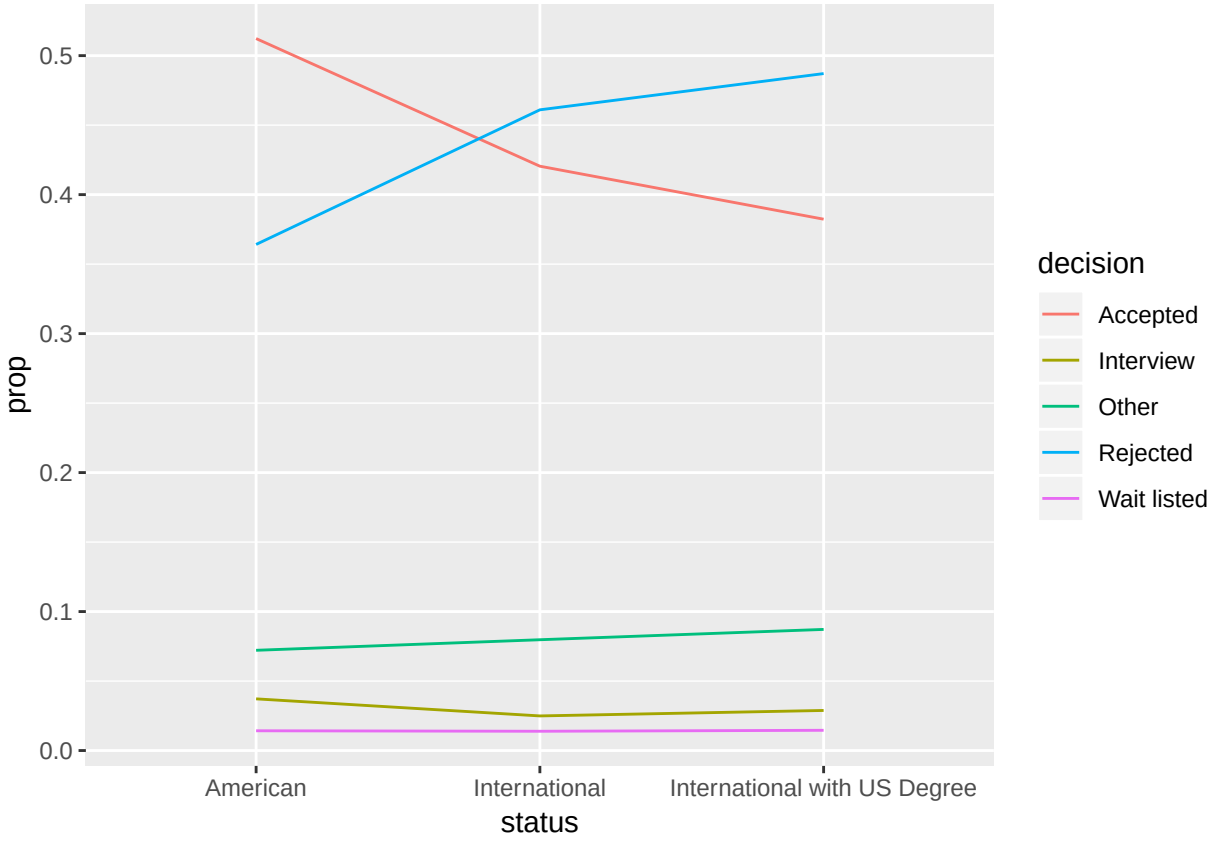
From the chart above, we see that the majority of students applying are American. In Immigration Status, around 60% of applicants are American and the rest of them are international students. We can tell that a big amount of graduate or Ph.D. students are coming from an international background.

Covariation Between Multiple Variables

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

```
## # A tibble: 4 x 6
## # Groups:   status [4]
##   status                     Accepted Interview Other Rejected `Wait listed`
##   <chr>                         <int>     <int> <int>    <int>         <int>
## 1 American                       2521       183   355     1792            70
## 2 International                  7393       438  1402     8106           244
## 3 International with US De~      1259        95   287     1604            48
## 4 Other                           181        26   754      151            23
```

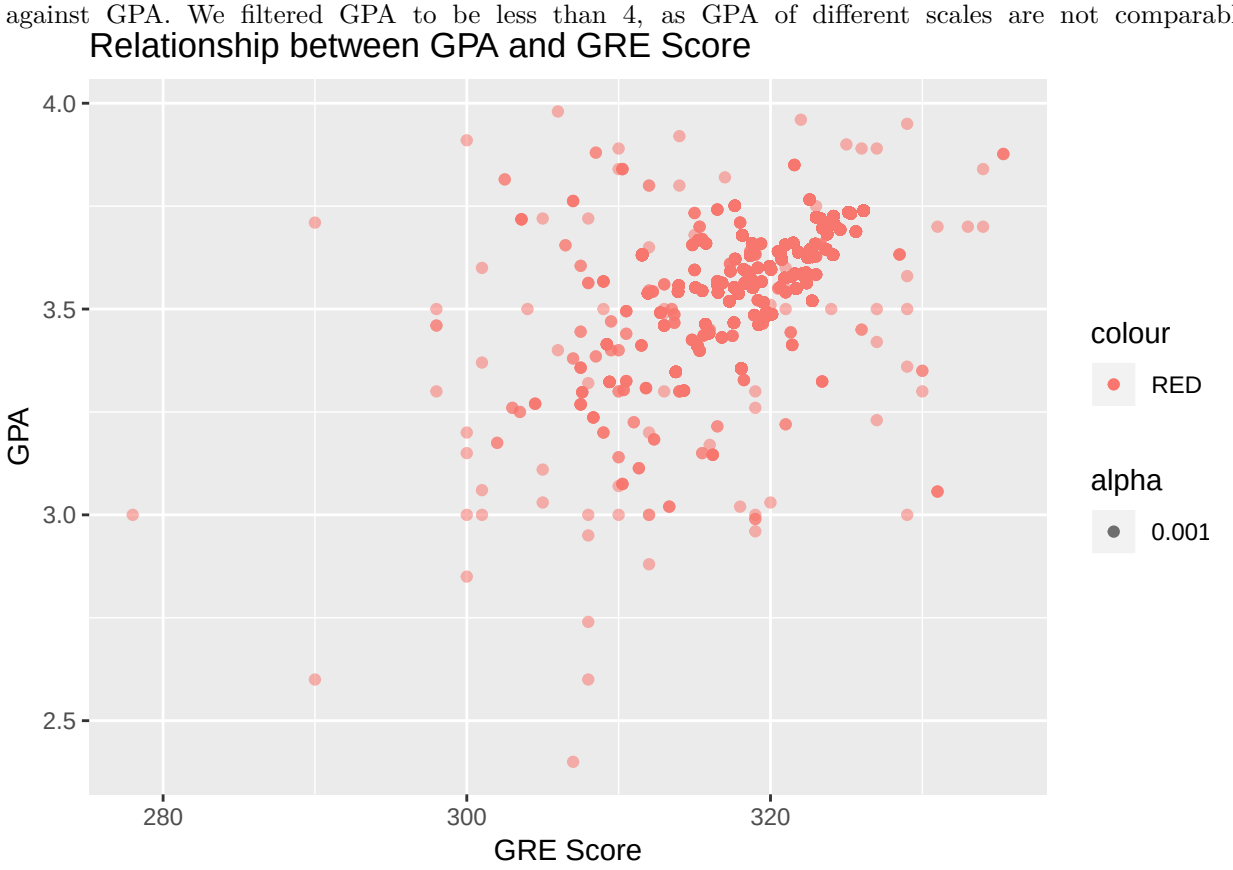




From the charts above, it seems that US based students tend to have higher acceptance rates than international

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.



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.

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.

Research Questions and Modeling Methods

Our general research question remains the same as in deliverable one: "How do different variables relate to admission decision?" However, through our exploratory data analysis, we realised that the full dataset

may be too large to get a clear picture of how different factors affect admissions decision. Because the full dataset contains a wide variety of schools and graduate programs (which would all have different standards for admission and a variety of interactions), we decided to narrow down the dataset to just the top 10 most popular Computer Science programs. The reason for this is that factors related to admission are likely not comparable accross different schools (e.g. highly selective schools vs high acceptance rate schools) or different programs (e.g. factors that may be important to Computer Science programs would likely differ from a Fine arts program).

```
##                                          uni_name accepted    n      rate
## 1                 Carnegie Mellon University (CMU)      523 1414 0.3698727
## 2          Georgia Institute Of Technology (GTech)      413  985 0.4192893
## 3       University Of California, San Diego (UCSD)      349  954 0.3658281
## 4  University Of Illinois, Urbana-Champaign (UIUC)      367  954 0.3846960
## 5                              Stanford University      245  914 0.2680525
## 6         University Of California, Berkeley (UCB)      155  844 0.1836493
## 7                               Purdue University      335  745 0.4496644
## 8           University Of Washington, Seattle (UW)      167  713 0.2342216
## 9          University Of Texas, Austin (UT Austin)      282  706 0.3994334
## 10                             Cornell University      253  674 0.3753709
```

In order to model the probability of acceptance, we decide to use a logisitic regression, as the result we want to model is binary (Accepted vs Not Accepted). For our covariates, we hypothesize that GPA, GRE Scores, and student status (American vs International Student) play a significant role in determining the admission decision. We also suspect that there may be interactions between student status and scores. In order to select variables, we start with a full model containing all of these variables and use the backwards elimination method of stepwise regression to fit a "best" model.

Fitting the Model

We decided to fit a model both with and without interaction to better understand how significant the interaction terms are for prediction. Predictor variables included in the models are GRE total score (GRE verbal + GRE quant), undergraduate GPA, GRE writing score, and student status.

Model With Interaction

```
##
## Call:
## glm(formula = decision1 ~ ugrad_gpa + GRE_Total + gre_writing +
##     status + gre_writing:status - 1, family = binomial, data = grad)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5368  -1.0575  -0.8203   1.2008   1.9261
##
## Coefficients:
##                                              Estimate Std. Error
## ugrad_gpa                                    1.146643   0.283057
## GRE_Total                                    0.031278   0.009388
## gre_writing                                 -0.268174   0.189752
## statusAmerican                             -13.403241   2.973278
## statusInternational                        -12.782405   2.852552
## statusInternational with US Degree         -15.544697   3.081698
## gre_writing:statusInternational             -0.252731   0.225536
```

```
## gre_writing:statusInternational with US Degree   0.540781   0.358197
##                                                 z value Pr(>|z|)
## ugrad_gpa                                         4.051 5.10e-05 ***
## GRE_Total                                         3.332 0.000863 ***
## gre_writing                                      -1.413 0.157571
## statusAmerican                                   -4.508 6.55e-06 ***
## statusInternational                              -4.481 7.43e-06 ***
## statusInternational with US Degree               -5.044 4.55e-07 ***
## gre_writing:statusInternational                  -1.121 0.262468
## gre_writing:statusInternational with US Degree    1.510 0.131112
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1512.4  on 1091  degrees of freedom
## Residual deviance: 1438.4  on 1083  degrees of freedom
## AIC: 1454.4
##
## Number of Fisher Scoring iterations: 4
```

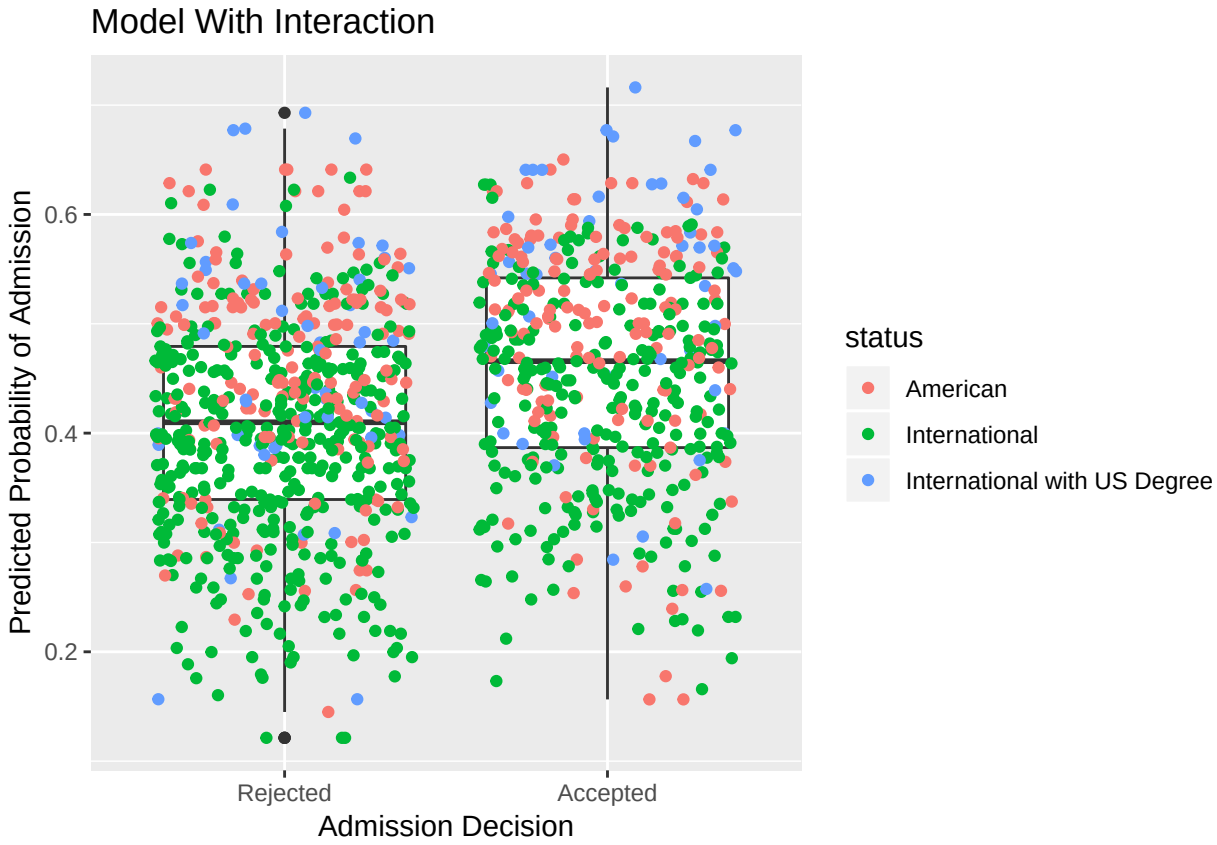
Model Without Interaction

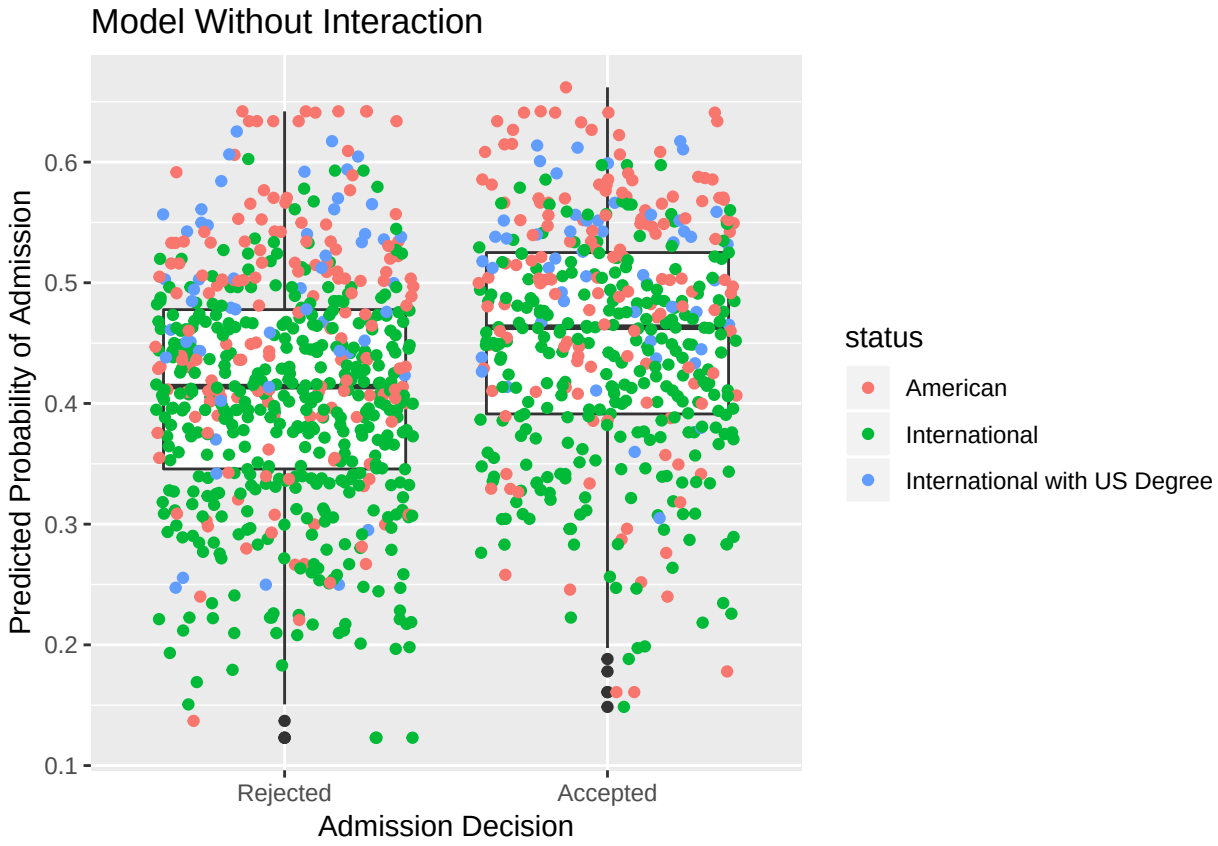
```
##
## Call:
## glm(formula = decision1 ~ ugrad_gpa + GRE_Total + gre_writing +
##     status - 1, family = binomial, data = grad)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4334  -1.0585  -0.8327   1.2071   1.9526
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## ugrad_gpa                          1.168482   0.283037   4.128 3.65e-05
## GRE_Total                          0.030744   0.009342   3.291 0.000998
## gre_writing                       -0.359779   0.110128  -3.267 0.001087
## statusAmerican                   -12.892745   2.846590  -4.529 5.92e-06
## statusInternational              -13.302409   2.834117  -4.694 2.68e-06
## statusInternational with US Degree -12.981663   2.846232  -4.561 5.09e-06
##
## ugrad_gpa                         ***
## GRE_Total                         ***
## gre_writing                       **
## statusAmerican                    ***
## statusInternational               ***
## statusInternational with US Degree ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1512.4  on 1091  degrees of freedom
```

```
## Residual deviance: 1444.7  on 1085  degrees of freedom
## AIC: 1456.7
##
## Number of Fisher Scoring iterations: 4
```

One interesting observation we see is that there is an interaction between GRE Writing and Student Status. This could be due to varying standards for writing ability based on Student Status. American students may be held to a higher standard for writing quality than international students, which is to be expected.

Next, we plot box plots of admission decisions vs predicted probabilities to assess the predictive power of our models.





In model without interaction, the mean of predicted probabilities of rejected students is around 0.41, while the mean of predicted probabilities of accepted students is around 0.43, which is slightly higher than the mean of predicted probabilities of rejected students. Since their interval are overlapped, it means that the prediction may not be significant enough to explain the success of a student being accepted. In addition, the plots are fairly scattered, meaning that there does not exist a certain pattern to explain the trend.

In model with interaction, the mean of predicted probabilities of rejected students is around 0.4, while the mean of predicted probabilities of accepted students is around 0.45, which is slightly higher than the mean of predicted probabilities of rejected students. Since their interval are overlapped, it means that the prediction may not be significant enough to explain the success of a student being accepted. However, the plots are more densely concentrated than the one without interation.

Coefficient Interpretation

For the model that includes interaction:

The regression coefficient for ugrad_gpa is $\hat{\beta_(ugradgpa)} = 1.146643$ meaning that for a one-unit increase in undergraduate GPA the logit-transformed probability of getting accepted to the program will increase by 1.15. Predictor GRE_Total has a coefficient $\hat{\beta_(GREtotal)} = 0.031106$, showing that for a one-unit increase in GRE total scores the log odds will increase by 0.03. We also include categorical variable status representing the applicant's status. The corresponding coefficient $\hat{\beta_(American)} = -13.403241$ shows that if the applicant is an American student, the log odds will decrease by 13.4, holding all other independent variables constant, $\hat{\beta_(International)} = -12.782405$ shows the change in log odds given the student is an international student, and $\hat{\beta_(USdegree)} = -15.544697$ shows the change in log odds given the student is an international student with a US degree.

$\hat{\beta_(GREwriting)} = -0.267686$ is the regression coefficients for GRE writing score, and $\hat{\beta_(GREwriting : International)} = -0.252731$ and for the interaction term $\hat{\beta_(GREwriting : USdegree)} = 0.540781$ are the coefficients of GRE writing scores with respect to students status. However, the hypothesis tests for coefficient indicates that those terms would not significantly impact the prediction of our model.

```
## ugrad_gpa
## 0.4785392
```

```
## ugrad_gpa
## 0.5700877
```

```
## ugrad_gpa
## 0.1562274
```

We next check the prediction for the probability of a student getting accepted at mean level GPA, GRE total score, and writing score. According to our model that includes interaction, there's a 47.9% chance that the student will be admitted to the program if the student is an American student, and 57% and 15.6% respectively if the student is an international student or an international student with a US degree.

For the model that does not include interaction terms:

The regression coefficient for ugrad_gpa is $\hat{\beta_(ugradgpa)} = 1.168482$, which indicates that for a one-unit increase in undergraduate GPA the logit-transformed probability of getting accepted to the program will increase by 1.15. $\hat{\beta_(GREtotal)} = 0.030744$ is the coefficient for predictor GRE_Total showing that for a one-unit increase in GRE total scores the log odds will increase by 0.03. $\hat{\beta_(GREwriting)} = -0.359779$ shows that GRE writing score is negatively related with the probability of acceptance, and for every one unit increase in writing score leads to a 0.36 drop in log odds. If the applicant is an American students, our model predicts a drop equals to $\hat{\beta_(American)} = -12.892745$ in the log odds, holding all other independent variables constant. If the aaplicant is a international student, log odds decreases by $\hat{\beta_(International)} = -13.302409$, and if the student has earned a US degree, log odds drops by $\hat{\beta_(USdegree)} = -12.981663$.

```
## ugrad_gpa
## 0.4909546
```

```
## ugrad_gpa
##  0.390348
```

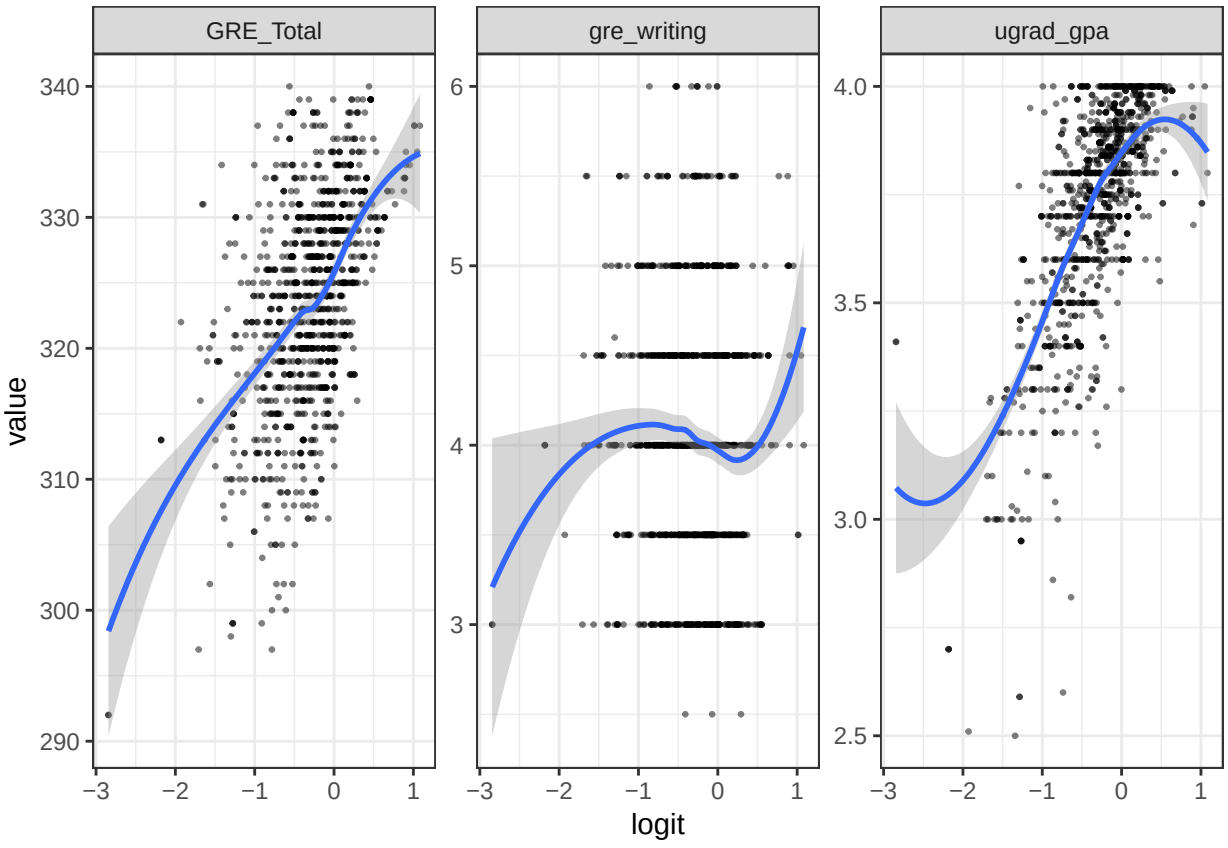
```
## ugrad_gpa
##  0.468765
```

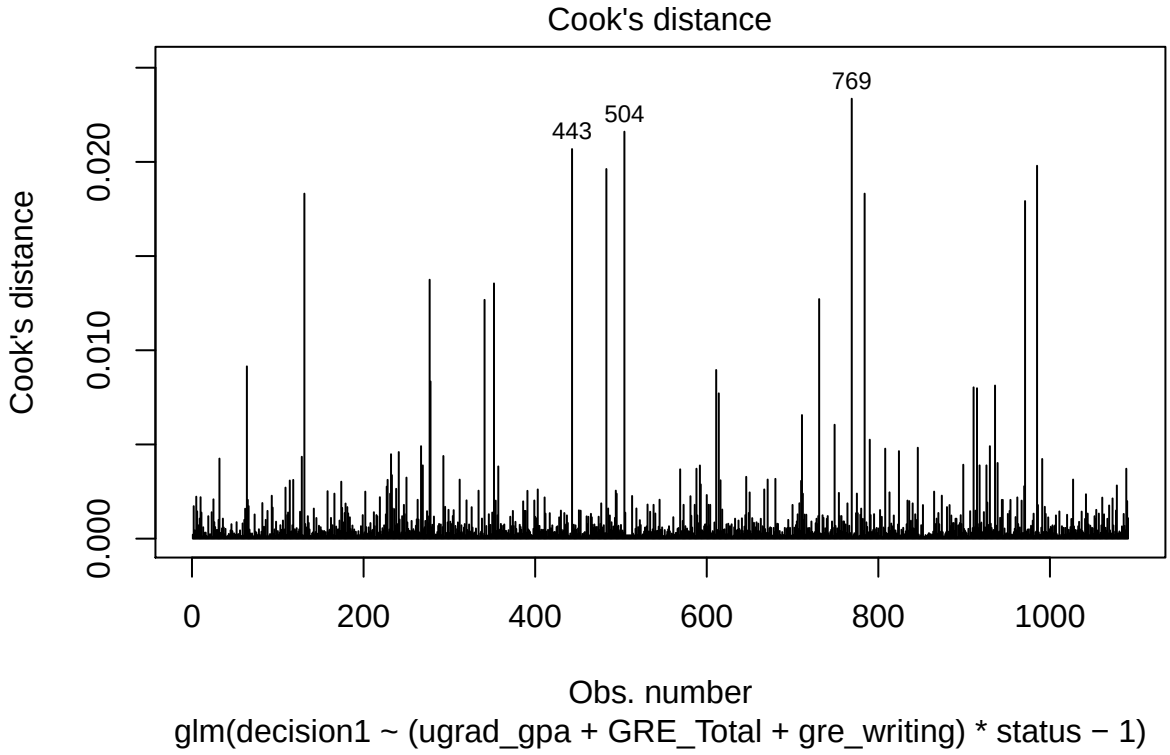
Using same mean level GPA, GRE total score and writing score, our simple logistic model predicts that the probability of an American student getting accepted to the program is 49.1% and the probability for an international student without a US degree and one with a US degree is 39% and 46.9% respectively.

Assumption

Next, to ensure that our models are valid, we check the assumptions of logistic regression:

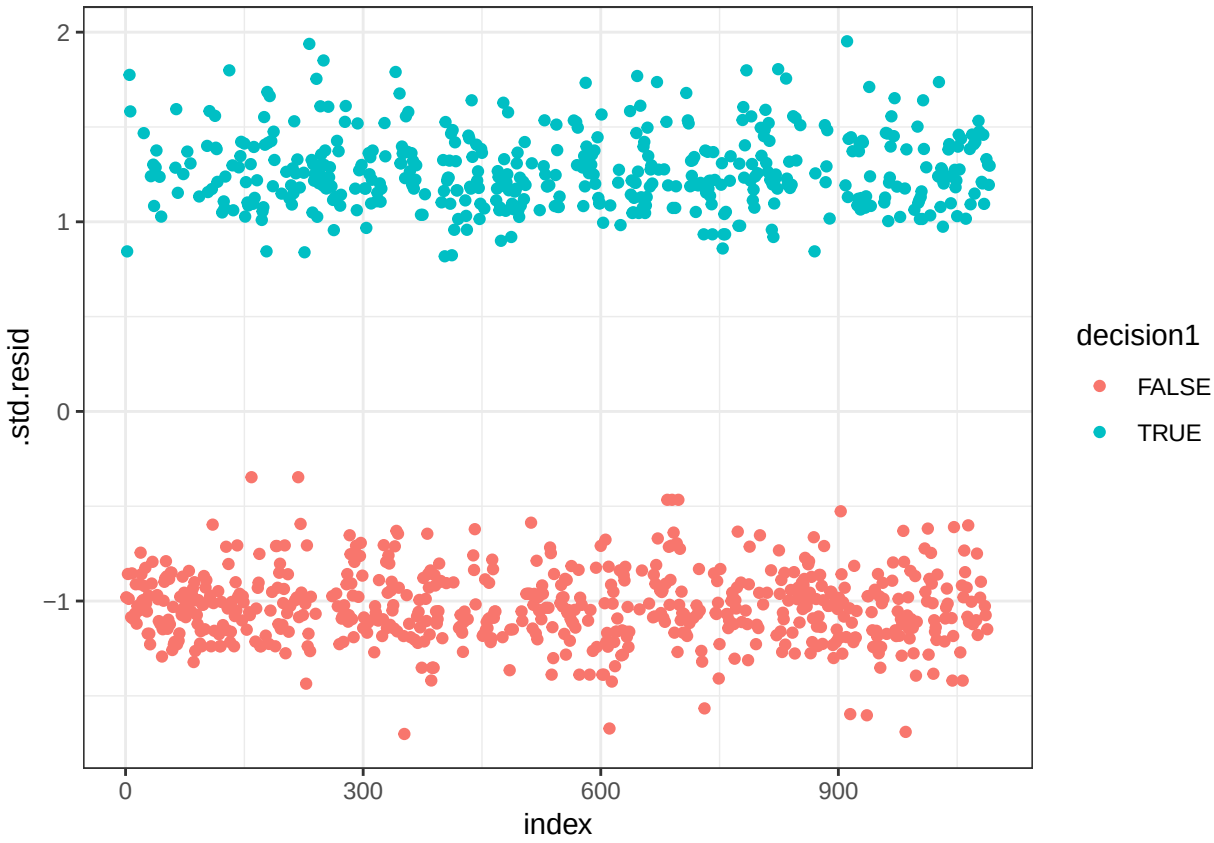
1. Outcome is binary
2. Linear relationship between the logit of the outcome and each predictor variables
3. No influential values
4. No high intercorrelations






```
## # A tibble: 3 x 13
##   decision1 ugrad_gpa GRE_Total gre_writing status  .fitted .se.fit .resid
##   <lgl>         <dbl>     <dbl>       <dbl> <chr>     <dbl>   <dbl>  <dbl>
```

```
## 1 TRUE           3.17       324         3.5 Inter~ -0.299     0.749   1.31
## 2 TRUE           3.2        324         3   Inter~ -0.325     0.755   1.32
## 3 FALSE          3.3        325         5.5 Inter~ -0.00218   0.863  -1.18
## # ... with 5 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, index <int>
```




```
## # A tibble: 0 x 13
## # ... with 13 variables: decision1 <lgl>, ugrad_gpa <dbl>,
## #   GRE_Total <dbl>, gre_writing <dbl>, status <chr>, .fitted <dbl>,
## #   .se.fit <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, index <int>
```

```
## # A tibble: 1,091 x 17
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Purdue ~ (Com~ MS     S16    Rejected 2, 11, 2015         1446440400
##  2 Univers~ Comp~ MS     S16    Accepted 28, 9, 2015         1443412800
##  3 Univers~ (Com~ MS     F15    Rejected 24, 5, 2015         1432440000
##  4 Carnegi~ ( EC~ MS     F15    Other    27, 6, 2015         1435377600
##  5 Carnegi~ Elec~ MS     F15    Accepted 2, 6, 2015          1433217600
##  6 Univers~ Elec~ MS     F15    Accepted 14, 4, 2015         1428984000
##  7 Univers~ Comp~ PhD    F15    Other    20, 4, 2015         1429502400
##  8 Cornell~ Comp~ MS     F15    Rejected 7, 4, 2015          1428379200
##  9 Univers~ Comp~ PhD    F15    Other    16, 4, 2015         1429156800
## 10 Univers~ Comp~ PhD    F15    Other    16, 4, 2015         1429156800
## # ... with 1,081 more rows, and 10 more variables: ugrad_gpa <dbl>,
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <dbl>,
## #   is_new_gre <lgl>, status <chr>, comments <chr>, decision1 <lgl>,
```

```
## #   GRE_Total <dbl>, gre_total <dbl>
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa        1.00       0.19      0.18        0.13
## gre_verbal       0.19       1.00     -0.08        0.54
## gre_quant        0.18      -0.08      1.00        0.03
## gre_writing      0.13       0.54      0.03        1.00
##
## n= 1091
##
##
## P
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa               0.0000     0.0000    0.0000
## gre_verbal  0.0000                0.0069    0.0000
## gre_quant   0.0000     0.0069               0.3114
## gre_writing 0.0000     0.0000     0.3114
```

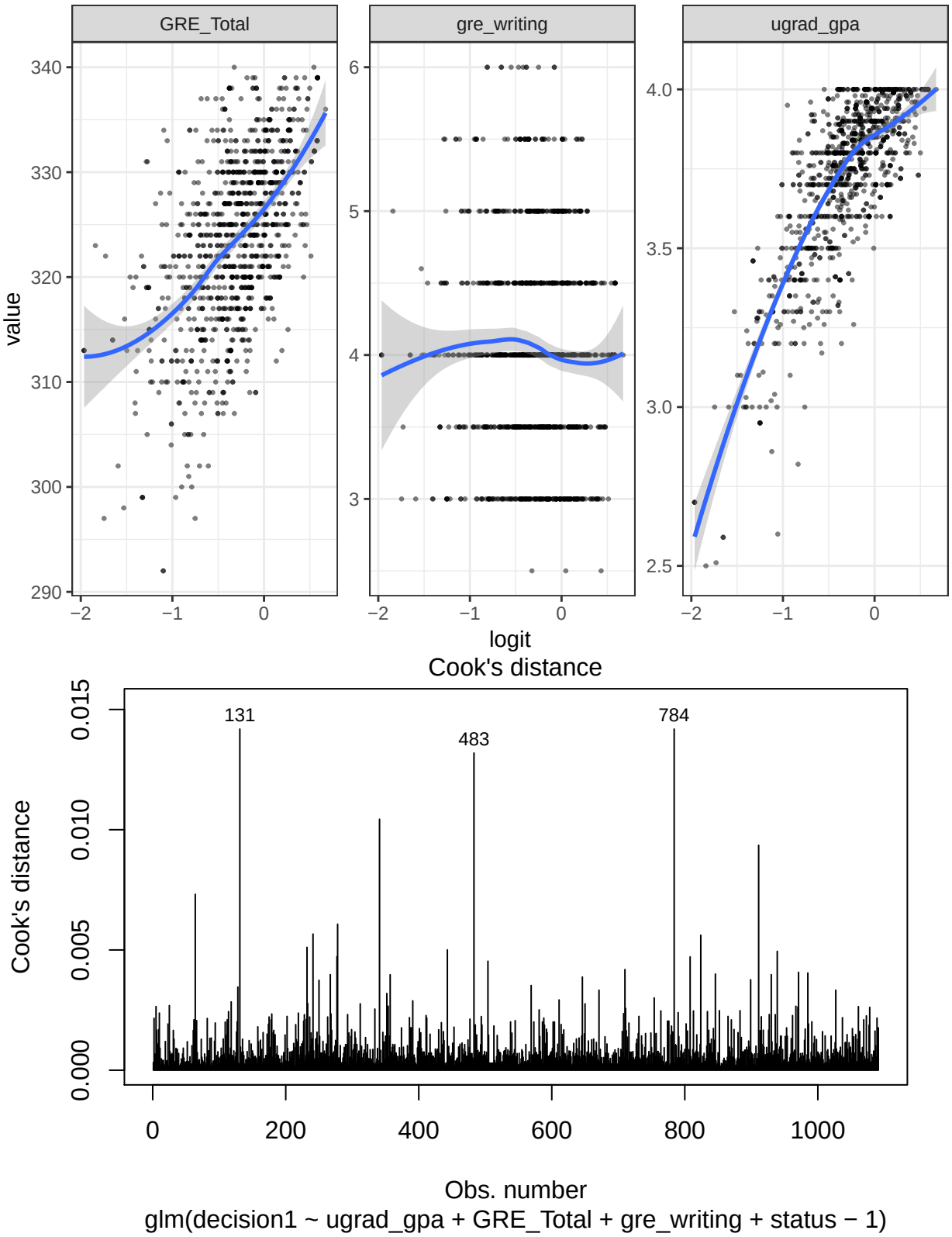
First, since we set the accepted decision as dependent variables and the decision is binary, either 1, accepted or 0, rejected. Therefore, the predicted probability is bind within the interval between 0 and 1. It meets the first assumption of dependent variable to be binary.

Second, logistic regression also assumes the linearity of independent variables.As shown in "The linearity of independent variables", the logit of GRE is quite linear to the accepted probability in logit scale. Even though there exists an U-shaped trend at the end of the parabala, the majority of gpa points associated linearly to the logit outcome of undergraduate gpa. However, the scatter plots of gre_writing shows non_linearity, similar to a cubic term.

Third, some outliers may be influential enough to alter the quality of the logistic regression model. Therefore, we calculated the Cook's distance for each points; the higher the leverage and residuals of that point, the higher its Cook's distance. As demonstrated in Cook's distance graph, there exist couple of spikes in the graph. To further investigate this issue, the deviance residuals plots has ben constructed. Since it does not have any observations whose cook's value is large than 3, we conclude that the dataset does not have any influential outliers.

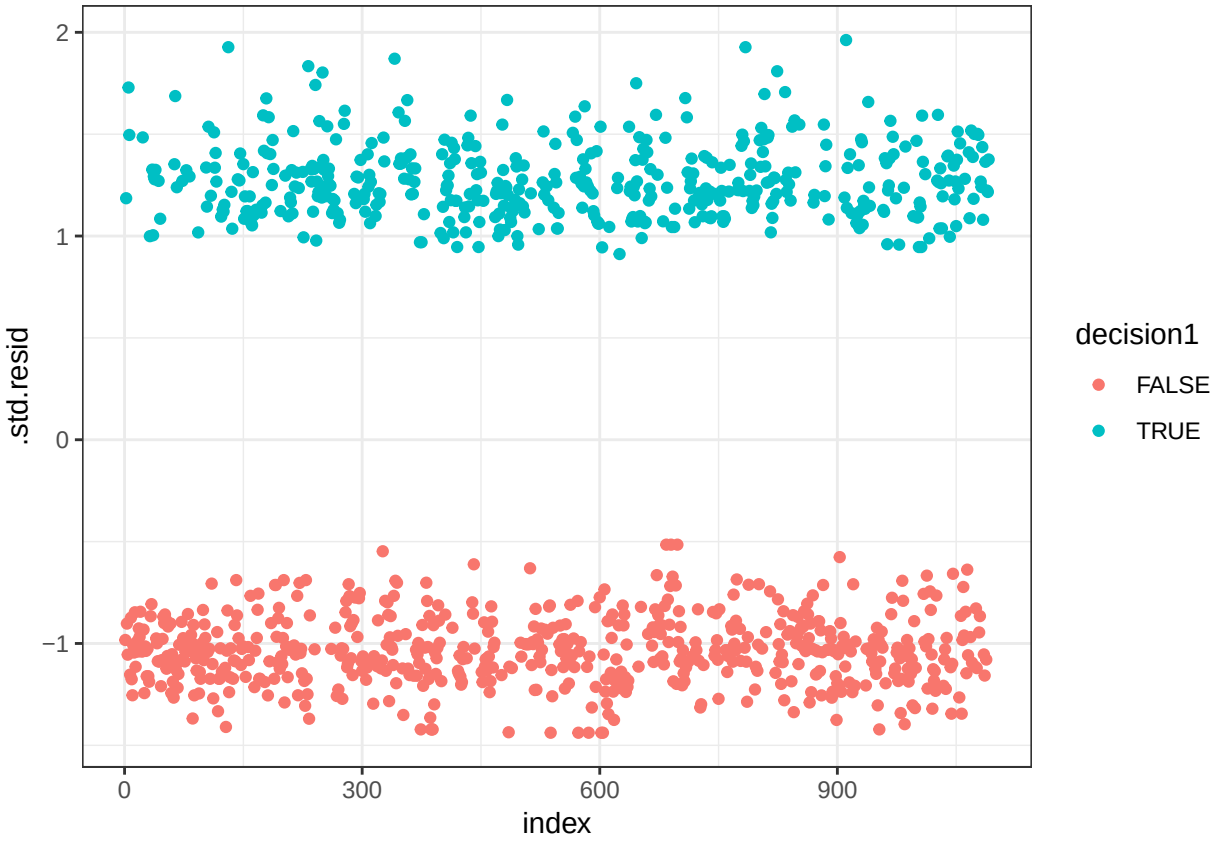
Last but not least, since the variables are intercorrlated, we take this into consideration and use interaction terms to overcome this issue.

Assumption_w/o interation




```
## # A tibble: 3 x 13
```

```
##   decision1 ugrad_gpa GRE_Total gre_writing status .fitted .se.fit .resid
##   <lgl>         <dbl>     <dbl>       <dbl> <chr>    <dbl>   <dbl>  <dbl>
## 1 TRUE           2.59       314           4 Ameri~   -1.65   0.342   1.91
## 2 TRUE           2.6        333           4 Ameri~   -1.06   0.369   1.65
## 3 TRUE           2.59       314           4 Ameri~   -1.65   0.342   1.91
## # ... with 5 more variables: .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, index <int>
```




```
## # A tibble: 0 x 13
## # ... with 13 variables: decision1 <lgl>, ugrad_gpa <dbl>,
## #   GRE_Total <dbl>, gre_writing <dbl>, status <chr>, .fitted <dbl>,
## #   .se.fit <dbl>, .resid <dbl>, .hat <dbl>, .sigma <dbl>, .cooksd <dbl>,
## #   .std.resid <dbl>, index <int>
```

```
## # A tibble: 1,091 x 17
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Purdue ~ (Com~ MS     S16    Rejected 2, 11, 2015         1446440400
##  2 Univers~ Comp~ MS     S16    Accepted 28, 9, 2015         1443412800
##  3 Univers~ (Com~ MS     F15    Rejected 24, 5, 2015         1432440000
##  4 Carnegi~ ( EC~ MS     F15    Other    27, 6, 2015         1435377600
##  5 Carnegi~ Elec~ MS     F15    Accepted 2, 6, 2015          1433217600
##  6 Univers~ Elec~ MS     F15    Accepted 14, 4, 2015         1428984000
##  7 Univers~ Comp~ PhD    F15    Other    20, 4, 2015         1429502400
##  8 Cornell~ Comp~ MS     F15    Rejected 7, 4, 2015          1428379200
##  9 Univers~ Comp~ PhD    F15    Other    16, 4, 2015         1429156800
## 10 Univers~ Comp~ PhD    F15    Other    16, 4, 2015         1429156800
## # ... with 1,081 more rows, and 10 more variables: ugrad_gpa <dbl>,
```

```
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <dbl>,
## #   is_new_gre <lgl>, status <chr>, comments <chr>, decision1 <lgl>,
## #   GRE_Total <dbl>, gre_total <dbl>
```

```
## # A tibble: 1,091 x 3
##    ugrad_gpa gre_writing gre_total
##        <dbl>       <dbl>     <dbl>
##  1      3.5           3.5       325
##  2      3.68          4.5       335
##  3      3.96          5         318
##  4      3.93          5         332
##  5      3.3           4         314
##  6      3.76          5         325
##  7      4             5         337
##  8      3.25          3.5       325
##  9      3.7           3.5       322
## 10      3.7           3         322
## # ... with 1,081 more rows
```

```
##             ugrad_gpa gre_writing gre_total
## ugrad_gpa        1.00        0.13      0.27
## gre_writing      0.13        1.00      0.48
## gre_total        0.27        0.48      1.00
##
## n= 1091
##
##
## P
##             ugrad_gpa gre_writing gre_total
## ugrad_gpa               0           0
## gre_writing  0                      0
## gre_total    0         0
```

```
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa        1.00       0.19      0.18        0.13
## gre_verbal       0.19       1.00     -0.08        0.54
## gre_quant        0.18      -0.08      1.00        0.03
## gre_writing      0.13       0.54      0.03        1.00
##
## n= 1091
##
##
## P
##             ugrad_gpa gre_verbal gre_quant gre_writing
## ugrad_gpa               0.0000     0.0000    0.0000
## gre_verbal  0.0000                 0.0069    0.0000
## gre_quant   0.0000    0.0069                 0.3114
## gre_writing 0.0000    0.0000     0.3114
```

First, since we set the accepted decision as dependent variables and the decision is binary, either 1, accepted or 0, rejected. Therefore, the predicted probability is bind within the interval between 0 and 1. It meets the first assumption of dependent variable to be binary.

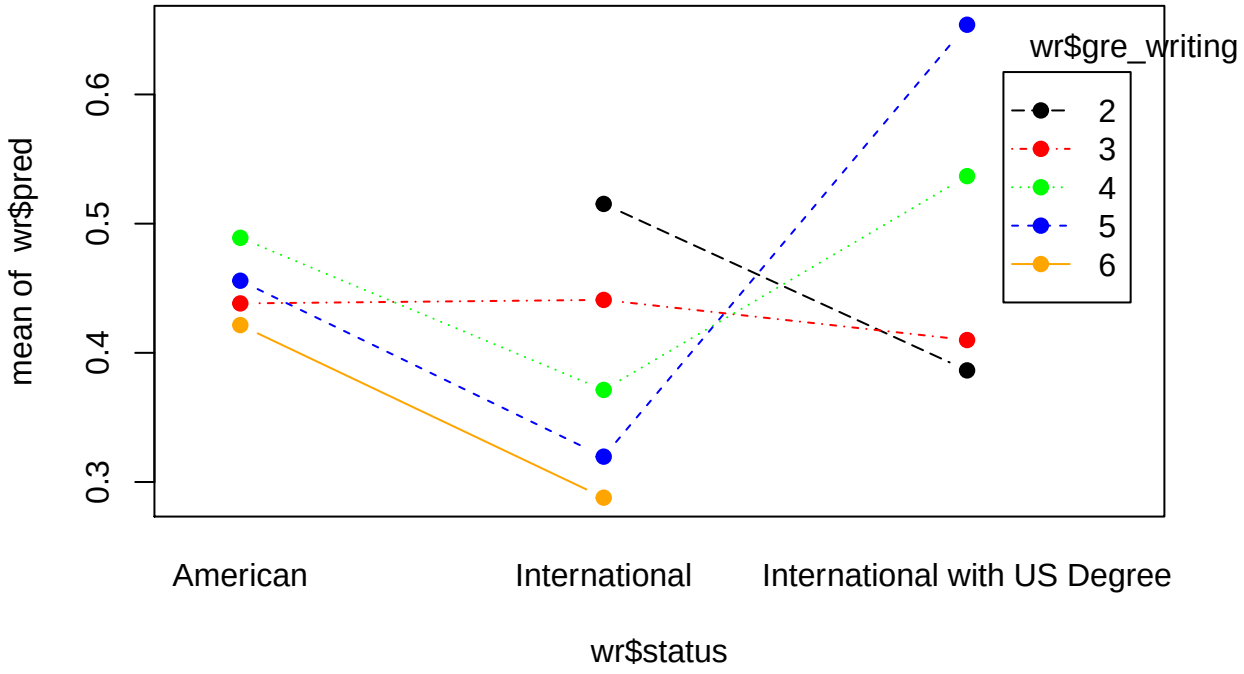
Second, logistic regression also assumes the linearity of independent variables.As shown in "The linearity of independent variables", the logit of GRE and undergraduate gpa are fairly linear to the accepted probability in logit scale. However, the scatter plots of gre_writing fits a parabola, instead of a linear line.

Third, some outliers may be influential enough to alter the quality of the logistic regression model. Therefore, we calculated the Cook's distance for each points; the higher the leverage and residuals of that point, the higher its Cook's distance. As demonstrated in Cook's distance graph, there exist couple of spikes in the graph. To further investigate this issue, the deviance residuals plots has ben constructed. Since it does not have any observations whose cook's value is large than 3, we conclude that the dataset does not have any influential outliers.

Last but not least, from the covariance matrix, we can tell that each term are corrlated with each other since its p value is near 0. Therefore, we incorporate interaction terms in our further model to overcome this disadvantage.

Tests for Significant Interaction

```
## # A tibble: 1,091 x 17
##    uni_name major degree season decision decision_date decision_timest~
##    <chr>    <chr> <chr>  <chr>  <chr>    <chr>                    <dbl>
##  1 Purdue ~ (Com~ MS     S16    Rejected 2, 11, 2015         1446440400
##  2 Univers~ Comp~ MS     S16    Accepted 28, 9, 2015         1443412800
##  3 Univers~ (Com~ MS     F15    Rejected 24, 5, 2015         1432440000
##  4 Carnegi~ ( EC~ MS     F15    Other    27, 6, 2015         1435377600
##  5 Carnegi~ Elec~ MS     F15    Accepted 2, 6, 2015          1433217600
##  6 Univers~ Elec~ MS     F15    Accepted 14, 4, 2015         1428984000
##  7 Univers~ Comp~ PhD    F15    Other    20, 4, 2015         1429502400
##  8 Cornell~ Comp~ MS     F15    Rejected 7, 4, 2015          1428379200
##  9 Univers~ Comp~ PhD    F15    Other    16, 4, 2015         1429156800
## 10 Univers~ Comp~ PhD    F15    Other    16, 4, 2015         1429156800
## # ... with 1,081 more rows, and 10 more variables: ugrad_gpa <dbl>,
## #   gre_verbal <dbl>, gre_quant <dbl>, gre_writing <int>,
## #   is_new_gre <lgl>, status <chr>, comments <chr>, decision1 <lgl>,
## #   GRE_Total <dbl>, pred <dbl>
```




```
## Analysis of Deviance Table
```

```
##
## Model 1: decision1 ~ ugrad_gpa + GRE_Total + gre_writing + status - 1
## Model 2: decision1 ~ (ugrad_gpa + GRE_Total + gre_writing) * status -
##     1
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1085     1444.7
## 2      1079     1435.4  6   9.2779   0.1585
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: decision1
##                     LR Chisq Df Pr(>Chisq)
## ugrad_gpa             17.480  1  2.904e-05 ***
## GRE_Total             11.910  1  0.0005584 ***
## gre_writing           11.308  1  0.0007716 ***
## status                32.334  3  4.449e-07 ***
## ugrad_gpa:status       1.154  2  0.5616588
## GRE_Total:status       2.097  2  0.3504976
## gre_writing:status     3.014  2  0.2215837
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The plot suggests that the effect of GRE writing is not consistent across all three groups of students. For example, a writing score of 5 showed the greater mean probability of acceptance for American and international students with US Degree. for international student, a writing score of 2 gives the highest chance of acceptance. This suggests there may be a meaningful or significant interaction effect, but we will need to do a statistical test to confirm this hypothesis.

Test for the inclusion of a Categorical Variable

H_0: full_mod = full_mod

H_a: full_mod = full_mod_int

Significance Level: 0.05 Pr(>Chi) for two models is 0.1581, which is bigger than siginificant level 0.05. Therefore, two models are not significantly different. Pr(>Chi) for ugrad_gpa, GRE_Total, gre_writing and status are all smaller than siginificant level 0.05, while all the interaction effect is not signficant. Therefore, the anova table indicates that the main effect are significant, and interaction effect is not significant.

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. From the analysis above, we see that while GPA, GRE Scores, and Student Status have a significant affect on admissions decisions, they alone are not great predictors for admission results. We see from the box plots that while the model had a higher average predicted probability for students that were actually accepted, there is too much variance in the resulted predictions. This result is likely due to the fact that the dataset is missing many variables that may also be important for admission decisions, such as research experience, reccomendations, reputation of undergraduate institution and so on. While it may be possible to extract this information from the 'comments', many observations did not include any comments and many more did not mention these factors in the comments. This leads us to believe that the admissions process is more than just a "numbers game," and likely includes many "intangibles" in order to determine the ultimate admission result of each student.