LogReg Assignment

Samuel Ivanecky

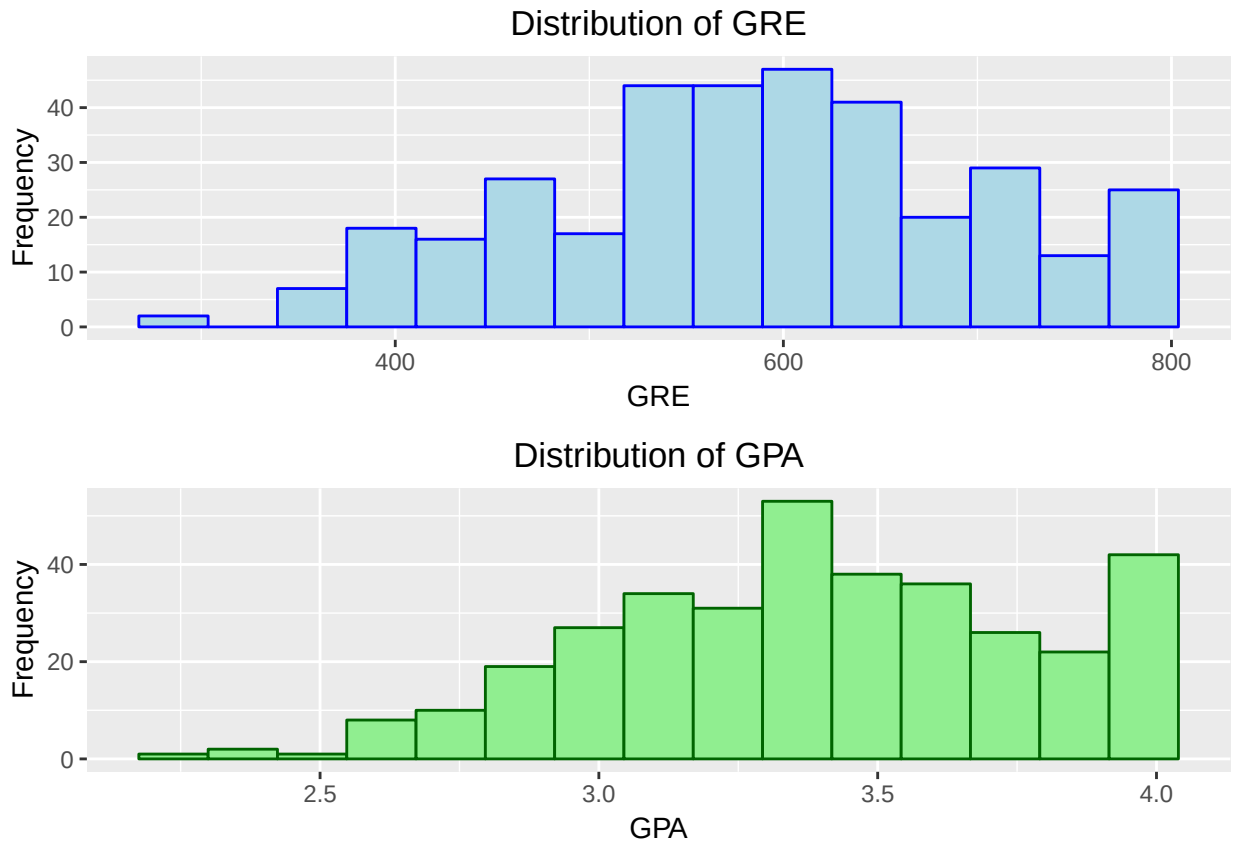
February 8, 2019

Problem 1

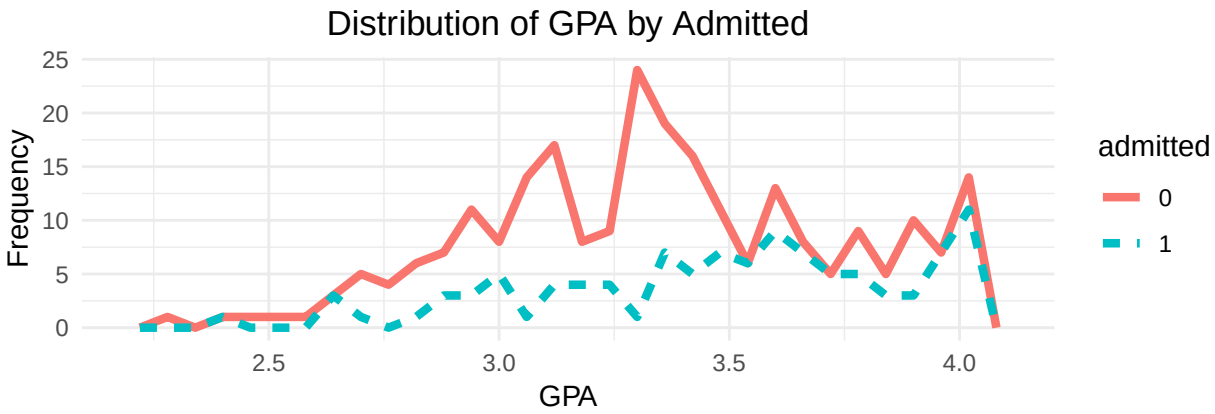
Modeling the probability of winning as a linear combination of predictor variables directly poses a problem because a linear model has potential to produce results with no numerical bounds. This means the model can produce values that are greater than one or less than zero which do not make logical sense in the terms of the problem. A team cannot have over a 100% chance to win a game and similarly they cannot have a negative chance of winning. Therefore, a proper model would constrain the predicted output values on a scale of 0 to 1.

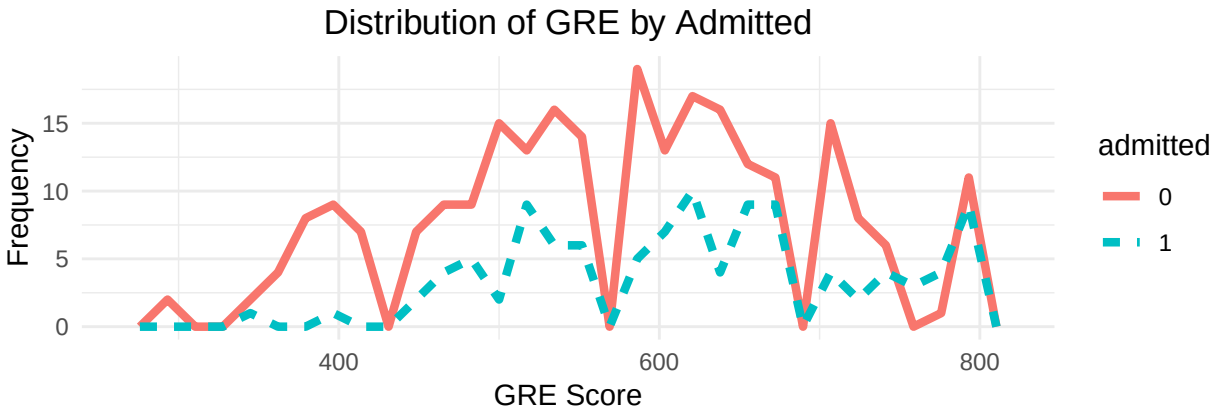
This problem can be handled by using a logistic regression model which equates the linear combination of predictor variables to the term $log\left(\frac{p}{1-p}\right)$, which will only produce results from 0 to 1.

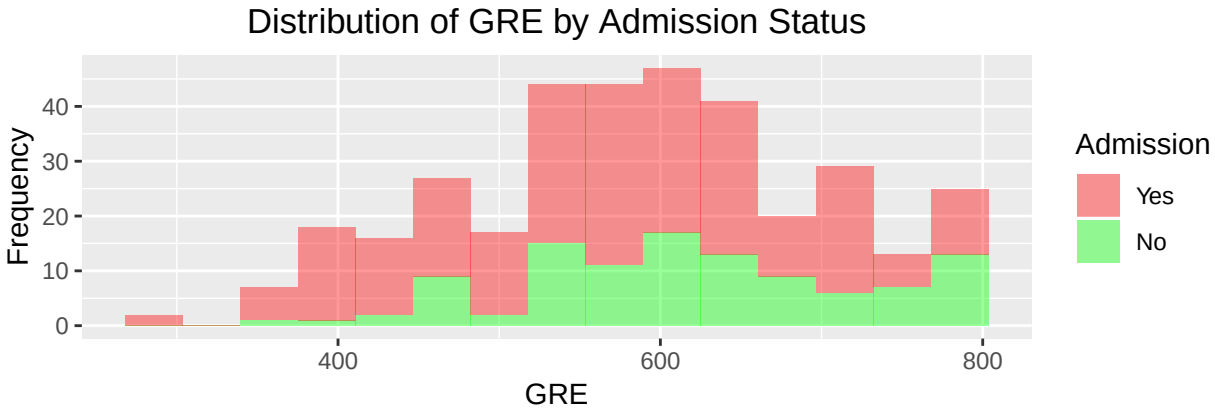
Problem 2

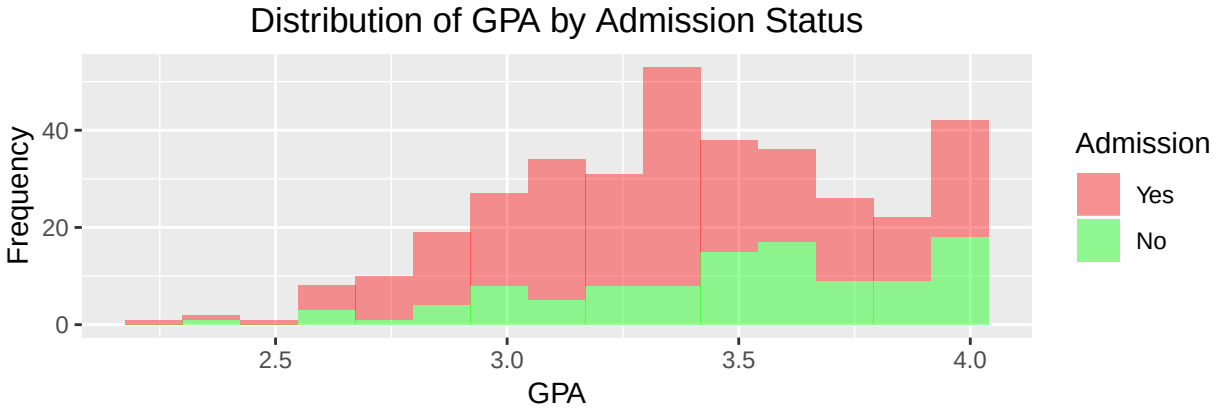



```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


Distribution of GPA by Admitted
Frequency
25
20
15
10
5
0
2.5
3.0
3.5
4.0
GPA
admitted
0
1

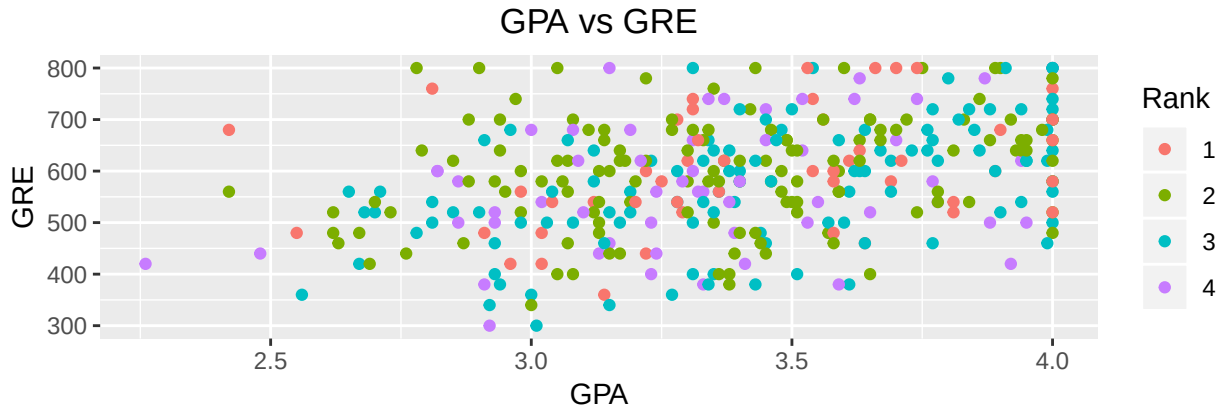

Distribution of GRE by Admitted
Frequency
15
10
5
0
400
600
800
GRE Score
admitted
0
1

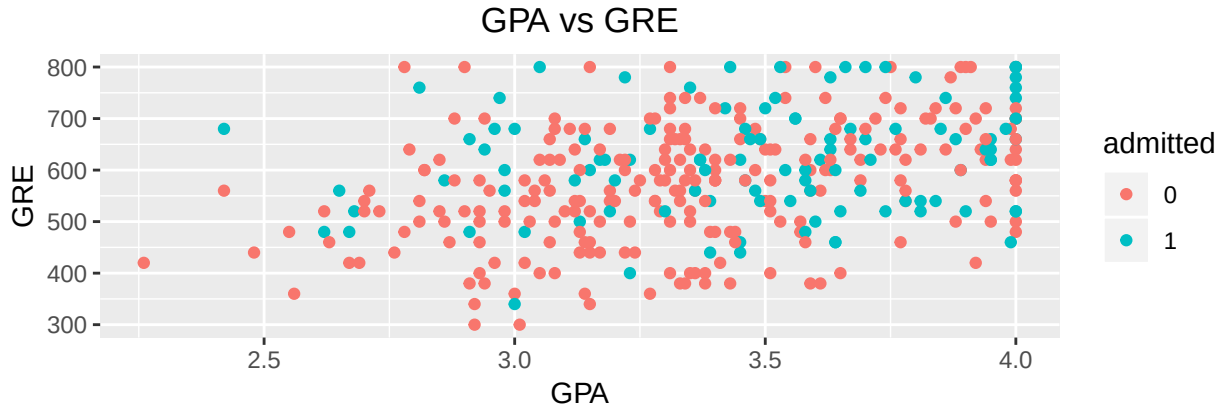




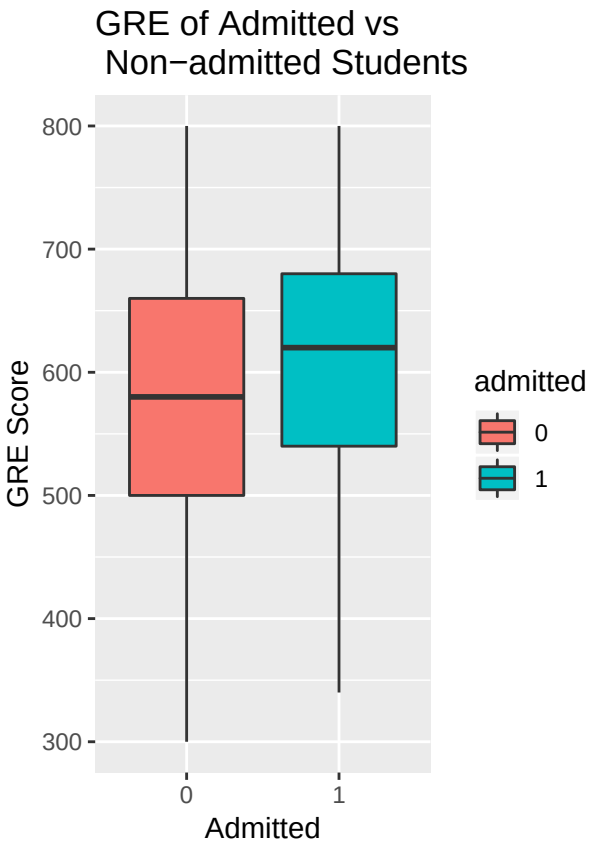
The distributions of GPA and GRE among all students are slightly skewed-left with more students acheiving scores towards the maximum than minimum of either statistic. From the second plots we see the distribution of GPA for those admitted has a larger proportion of data from ~3.5-4.0 whereas the distribution of non-admitted students has the largest proportion between 3.0-3.5. From this it initially appears that the 3.5 GPA mark may be an indicator of admission. This can also be seen in the stacked histograms.

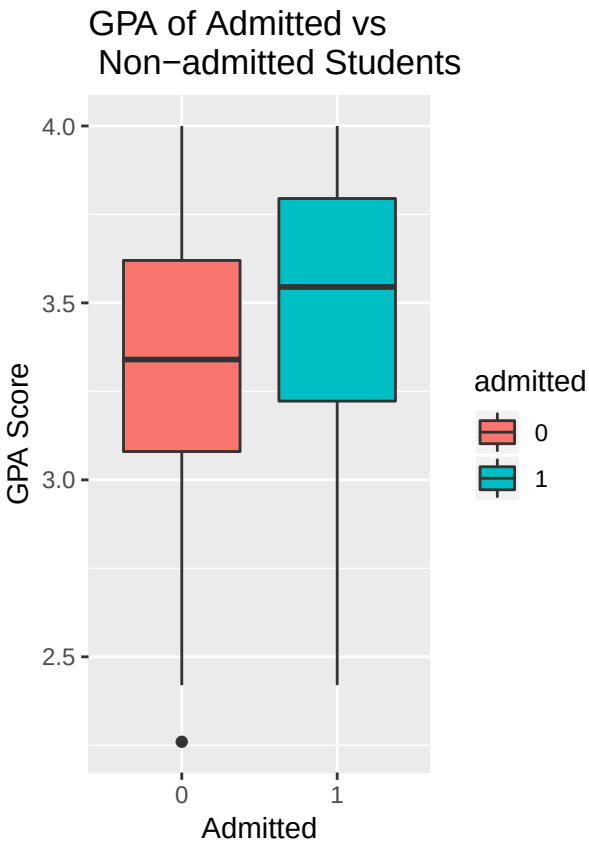
This trend does not appear with GRE. The distribution of GRE between admitted and non-admitted students is roughly identical with one small difference. The proportion of students scoring less than ~450 is much higher in non-admitted students whereas essentially no admitted students scored below this makr.

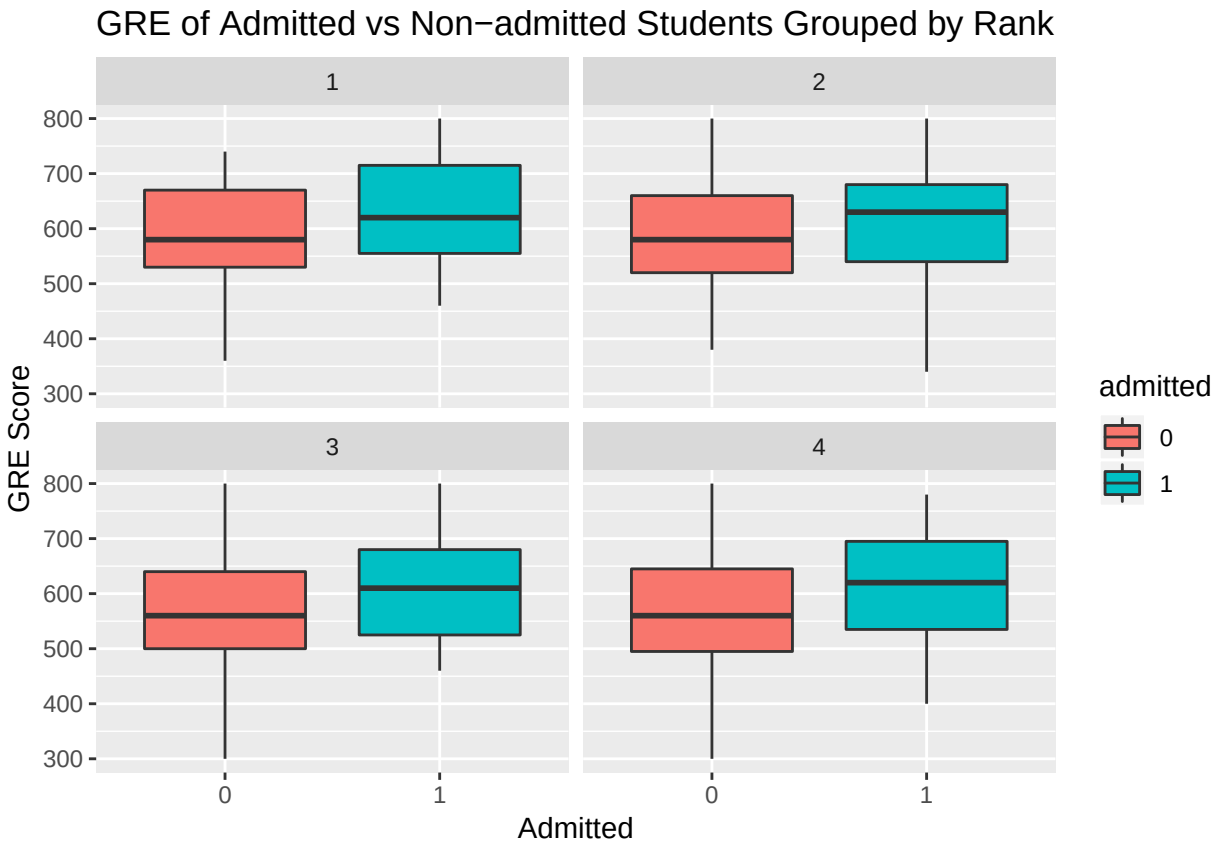


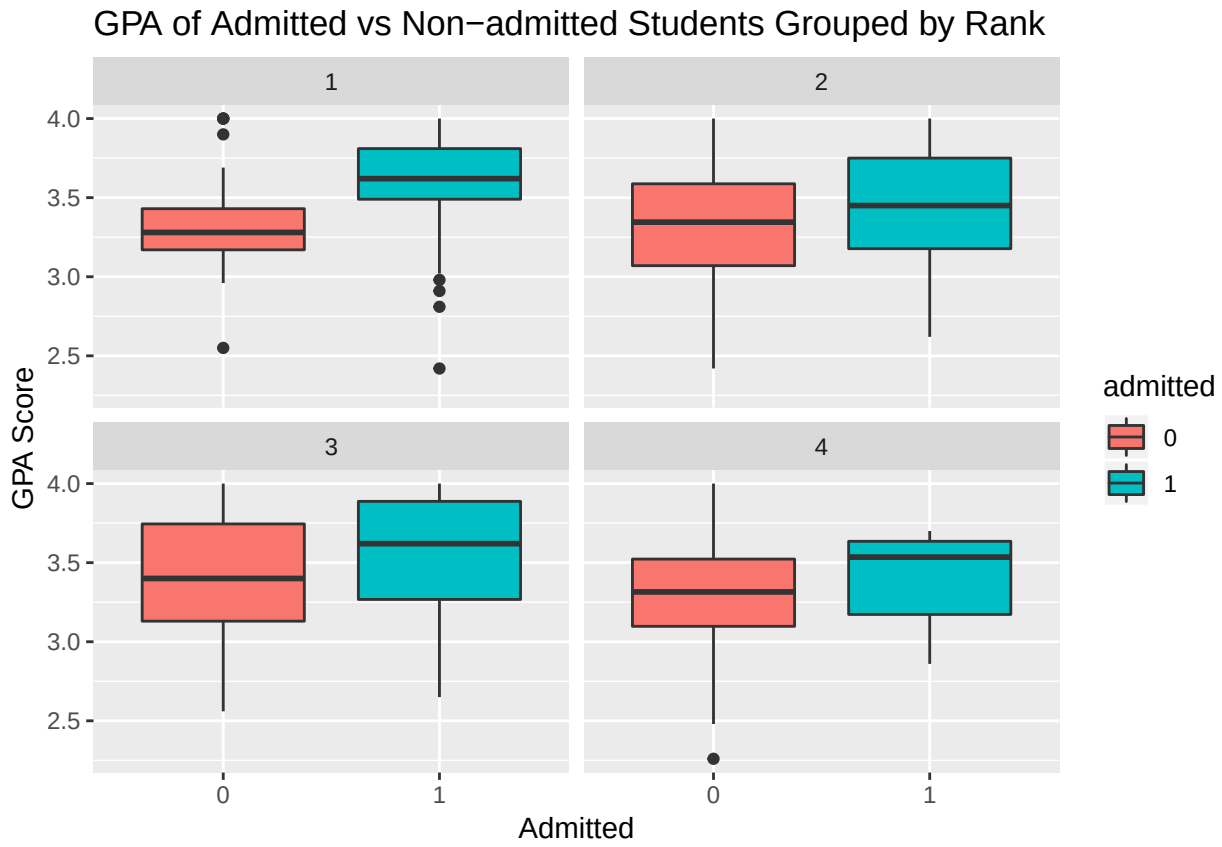


No clear relationships can be drawn between probability of admission and any variables from the scatterplots. It appears there is a higher proportion of students who score well on both the GRE and GPA that are admitted but within this group there are still individuals who are also rejected.

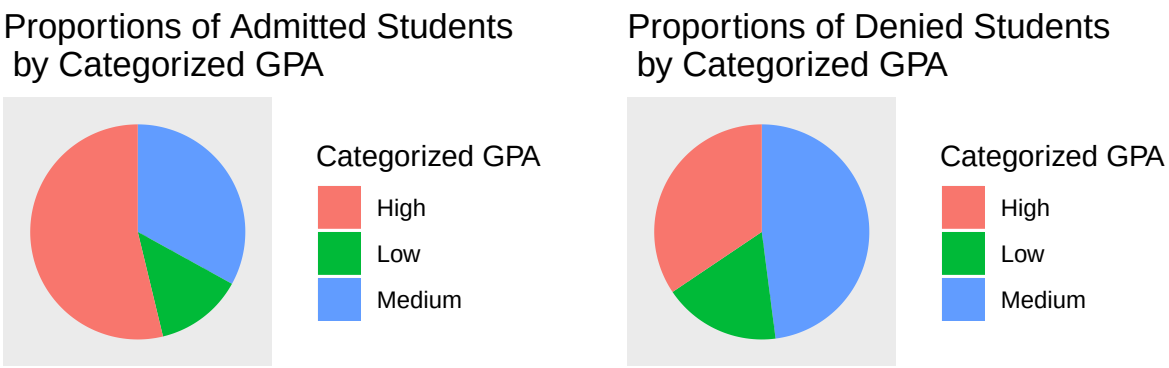

GRE of Admitted vs
Non−admitted Students
800
700
600
500
400
300
GRE Score
0
1
Admitted
admitted
0
1


GPA of Admitted vs
Non−admitted Students
4.0
3.5
3.0
2.5
GPA Score
0
1
Admitted
admitted
0
1


GRE of Admitted vs Non−admitted Students Grouped by Rank
1
2
3
4
800
700
600
500
400
300
GRE Score
0
1
Admitted
admitted
0
1



The boxplots provide little information with regard to rank vs admittance. While the median scores of GRE and GPA tend to be higher in students who are admitted, there is no clear association between rank and admission. One trend highlighted by the boxplots is it appears that GPA has more effect on likelihood of admission than GRE score. The difference in median of admitted vs non-admitted students is much smaller in GRE than in GPA, indicating GPA may be a better predictor of admission. Note that the median GPA is roughly 3.5 which was the mark previously identified. Only slightly more than 25% of non-admitted students scored above this mark, compared to approzximately 50% of admitted students.



The three categories of GPA are classified as: Low: < 3.0 Medium 3.0-3.5 High: > 3.5

The cutoff for the High group matches the cutoff trend from earlier analysis and from the pie charts it is apparaent that this cutoff of 3.5 is significant for admission. Of those admitted, over half had a GPA above 3.5 and almost no individuals were admitted if their GPA was less than 3.0. Compared to the non-admitted students, roughly a third of the students had a GPA in the High range whereas almost half of the group fell between 3.0 and 3.5.

Based on the plots above, the categorized GPA appears to be the most important predictor variable for determining admission. The GRE score also appears to have some weight as admitted students tended to score higher than non-admitted students. Rank appeared to have minimal or no association with admission.

Problem 3

Problem 3.1

```
##
## Call:
## glm(formula = admission ~ GRE + GPAcategorized, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2368  -0.8554  -0.6959   1.2547   2.0463
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)          -2.409566   0.752161  -3.204  0.00136 **
## GRE                   0.003185   0.001147   2.778  0.00547 **
## GPAcategorizedLow    -0.447423   0.369577  -1.211  0.22604
## GPAcategorizedMedium -0.635577   0.267484  -2.376  0.01750 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 429.29  on 349  degrees of freedom
## Residual deviance: 409.88  on 346  degrees of freedom
## AIC: 417.88
##
## Number of Fisher Scoring iterations: 4
```

Based on the model summary, three of the four variables in the model are significant to the admission status based on p-value statistics. The coefficient of GRE being positive indicates that increasing the GRE score of an applicant increases the log odds that the individual is admitted. The value of 0.003185 indicates that for each point increase on the GRE, the likelihood of admission increases by ~1%. Conversely, for GPAcategorizedLow and GPAcategorizedMedium, both variables have negative coefficients which indicated that if an applicant is a member of either group they will have decreased log odds of being admitted. Interstingly, the coefficient for the Medium level is larger than that for the Low level which implies that being in the Medium group decreases odds more than being in the Low group.

Two parameters were estimated for GPAcategorized - Low and Medium. The High level was taken as reference. The estimated parameters for Low and Medium are both negative, indicating that being a member of either group decreases the log odds of an applicant being admitted. As stated above, being a member of the Medium group surprisingly has a larger negative impact on addmission probability than being in the Low group. Based on the coefficient for Low, the lowest group has ~64% of the odds of getting admitted as members of the High group. Those in the Mediu group have only ~53% of the chance of those in the High group.

The GRE coefficient is positive which indicates that increasing the GRE score increases the log odds that an applicant is admitted. Logically this follows as typically students who perform better on the GRE will have a stronger chance of being admitted to graduate programs.

Problem 3.2

```
##
## Call:
## glm(formula = admission ~ GRE + GPAcategorized, family = binomial(link = "logit"),
##     data = df)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.2368  -0.8554  -0.6959   1.2547   2.0463
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -2.856988   0.707198  -4.040 5.35e-05 ***
## GRE                   0.003185   0.001147   2.778  0.00547 **
## GPAcategorizedHigh    0.447423   0.369577   1.211  0.22604
## GPAcategorizedMedium -0.188154   0.369228  -0.510  0.61034
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 429.29  on 349  degrees of freedom
## Residual deviance: 409.88  on 346  degrees of freedom
## AIC: 417.88
##
## Number of Fisher Scoring iterations: 4
```

The estimates related to GPAcategorized changed while the estimate related to GRE did not. The GRE coefficient did not change because the GRE score has no relation to the reference level of the GPA. The GPAcategorizedMedium coefficient changed to indicate that those in the Medium group have ~83% odds of getting accepted as those in the Low group.

Problem 3.3

```
##      351
## 58.39985
```

The probability of an individual getting rejected based on a High categorized GPA and a 650 score on the GRE is 58.39%. This is calculated by adding an individial to the dataframe with the given credentials, then running the *predict* function over the dataframe. This function call generates predictions for the entire dataset and the individual of interest is the 351st observation. We then look up the 351st entry in our probability vector which returns 0.4161, the probability that the individual is accepted. By subtracting this from 1, the value of 0.5839 is calculated which represents the probability that the individual is rejected.

```
##      351
## 58.39985
```

The same steps are performed to calculate the probability of rejection using the second model. This model also yields a result of 58.39% likelihood that the individual is rejected.