QTM2000 Name_____Angelina Cho_____

Case Studies in Business Analytics
Professor Mathaisel
Fall 2021

Mid-Term Exam 1

Individual Exercise

Due: 11:59pm Thursday, October 7.

I pledge my honor that I have neither received nor provided unauthorized assistance during the completion of this work. Please Initial:____AC_____

Note: This is an Individual Exercise. No group submission. In addition to the instructor, feel free to reach out to others for help, but the work must be your original contribution.

1. 10 points
2. 20 points
3. 35 points
4. 35 points

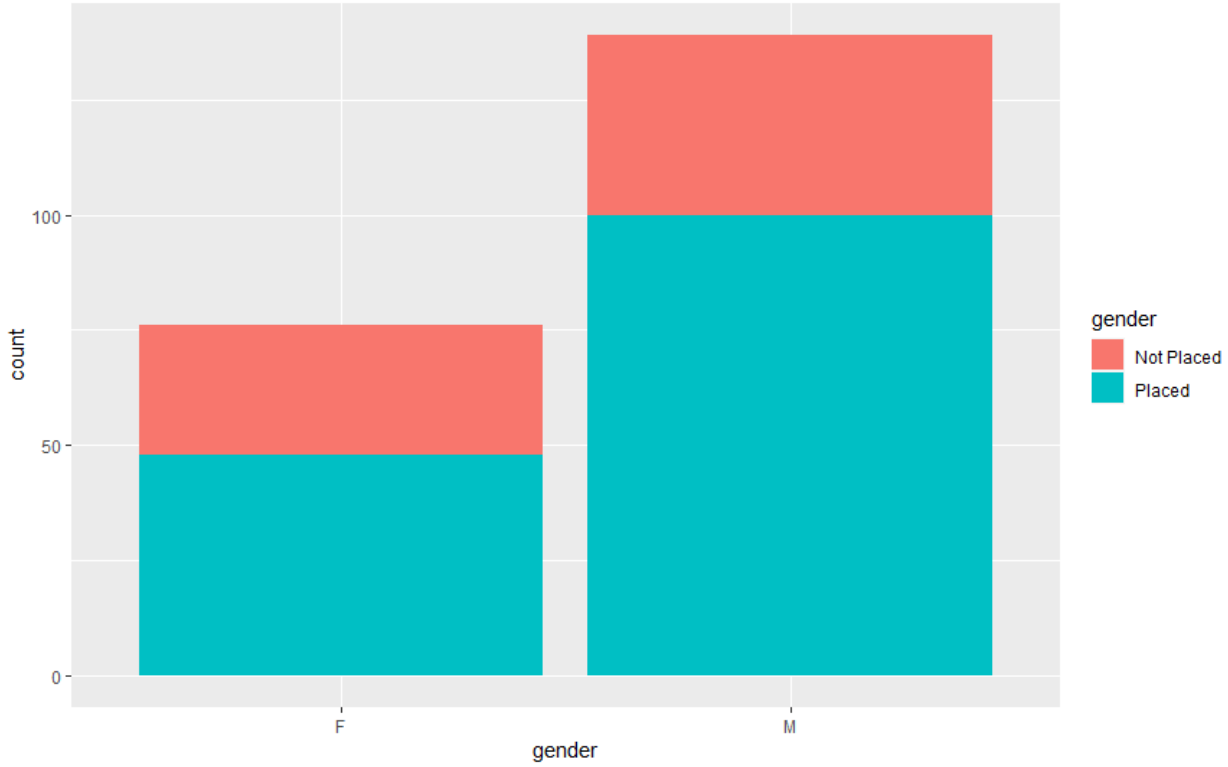
_____

100 points

Introduction

The purpose of this executive summary is to explore the data that showcases the placement data of students in XYZ campus. Attributes observed includes specialization, type, work experience, and salary offers to the placed students from secondary and higher secondary school. The aim of this model is to find which factor or factors influenced a candidate being placed, and how much does each factor influenced the likelihood of a candidate being selected. The dataset was provided by Ben Roshan in Business Analytics at Jain University in Bangalore (dataset available at https://www.kaggle.com/benroshan/factors-affecting-campus-placement).

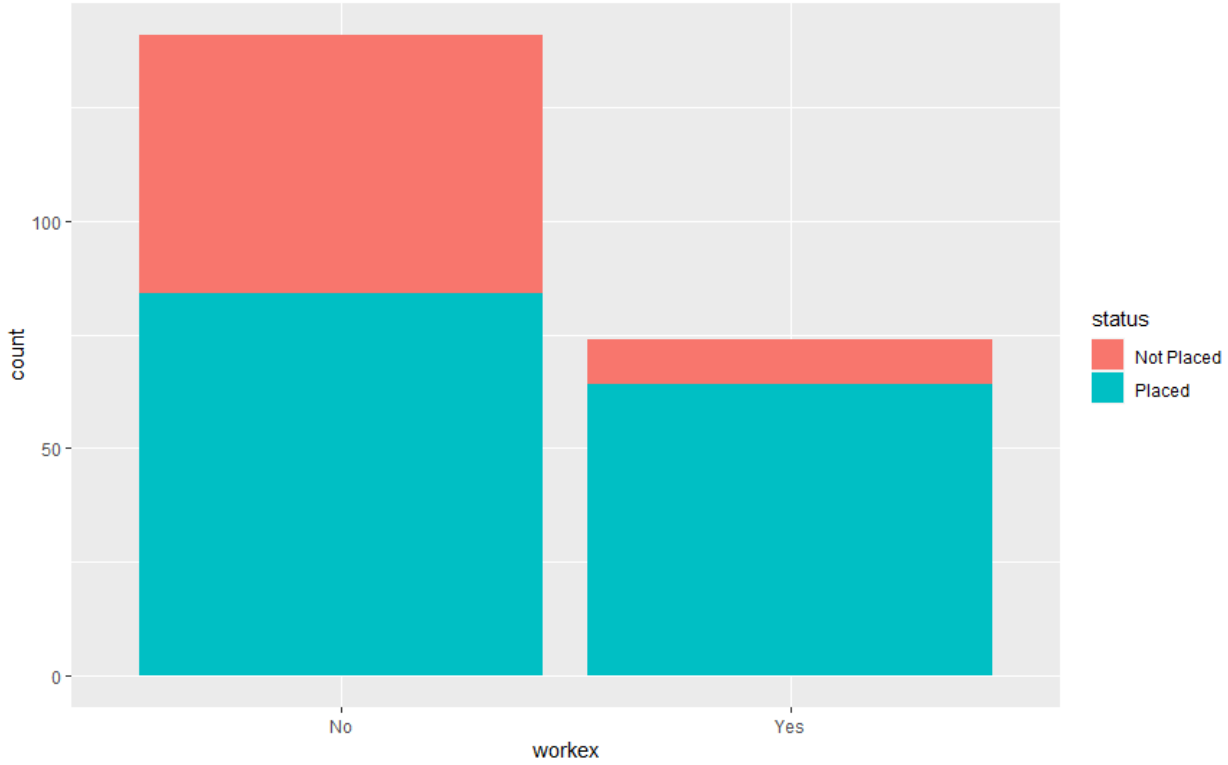
Figure 1.1 Stacked bar plot of female placement and male placement



This stacked bar plot depictes the number of people getting placed and not getting placed. The number of female (approximately 50) getting placed is smaller by double the number of male (approximately 100) getting placed, as we can tell from the number and the spread of the bars. The number of female (approximately 25) not getting placed is also smaller than the number of male (approximately 40) not getting placed, but only by approximately 15 people. However, his does not mean that males are more likely to get placed for jobs on campus because we have to take into account that the total number of female and total number of male in the dataset differs. The number of male in this visualization is almost double the number of female. If we look at the ratio of female getting placed vs. not placed, it's 2:1, so 60% chance of getting placed, according to the plot. While the ratio of male getting

placed v.s not placed is around 2: 0.7, so there's 74% chance of getting placed according to the plot. If we look at the data this way then we can see that the ratio of female not getting placed is a little higher than male. We can not come to a conclusion of whether or not the gender variable influenced the result of getting placed or not just by looking at this graph solely.

Figure 1.2 Stacked bar plot showing work experience vs. placement



This stacked bar plot illustrates the relationship between work experience and placement. Looking at the graph, we see that out of the people getting placed, the differences between having work experience and not having work experience is very small. Out of all the people that has no work experience (approximately 120 people), more than half of the people got placed. As for the group of people that has work experience, almost 90% of the people got placed and only 10% (approximately 15 people) of the people did not get placed. The difference for not getting placed between having work experience and not having work experience is very large. However, the total number of people having work experience is much less than the total number of people that has no work experience. The ratio of getting placed vs. not getting placed for people that have no work experience is around 1.7:1 (around 63% getting placed), while the ratio of getting placed vs. not getting placed for people that have work experience is around 6.5:1. It shows that for the people that have work experience, the chances for them to get placed is very high (around 87% chance), while for the people that don't have work experience, the chances for them to get placed is only around 63%, according to the plot.

Figure 1.3 Stacked bar plot showing undergraduate degree type vs. placement

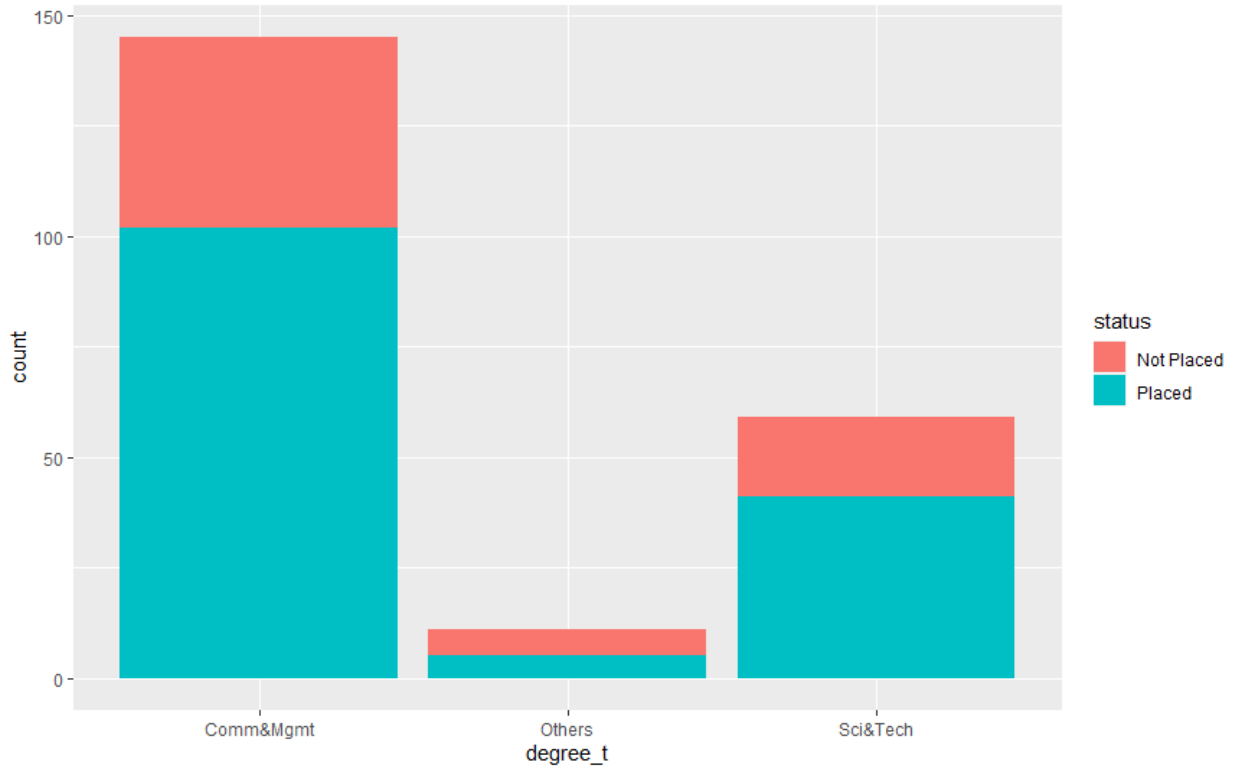


Figure 1.3 shows whether or not people are getting placed for different degrees, like Communication Management, Science & Technology, and other degrees. The total number of people getting placed for different degrees differentiate largely between Communication Management, Others, and Science & Technology, with approximately 140, 20, and 70 students approximately. Communication Management degree has the highest number of students getting placed, while also having the highest number of students in total. The same trend can be found with Science & Technology and Others degree too. The ratio and the percentage of students getting placed for Communication Management degree is around 2:1, and a 60%. The ratio and the percentage of students getting placed for Others degree is approximately 1:1, and a 50%. While the ratio and the percentage of students getting placed for Science & technology degree is around 1.8:0.4, and a 82%. While Communication Management degree has the highest number of students getting placed, Science & technology degree has the highest percentage of students getting placed.

Figure 2.1 Box & whiskers plot showing placement vs. degree percentage

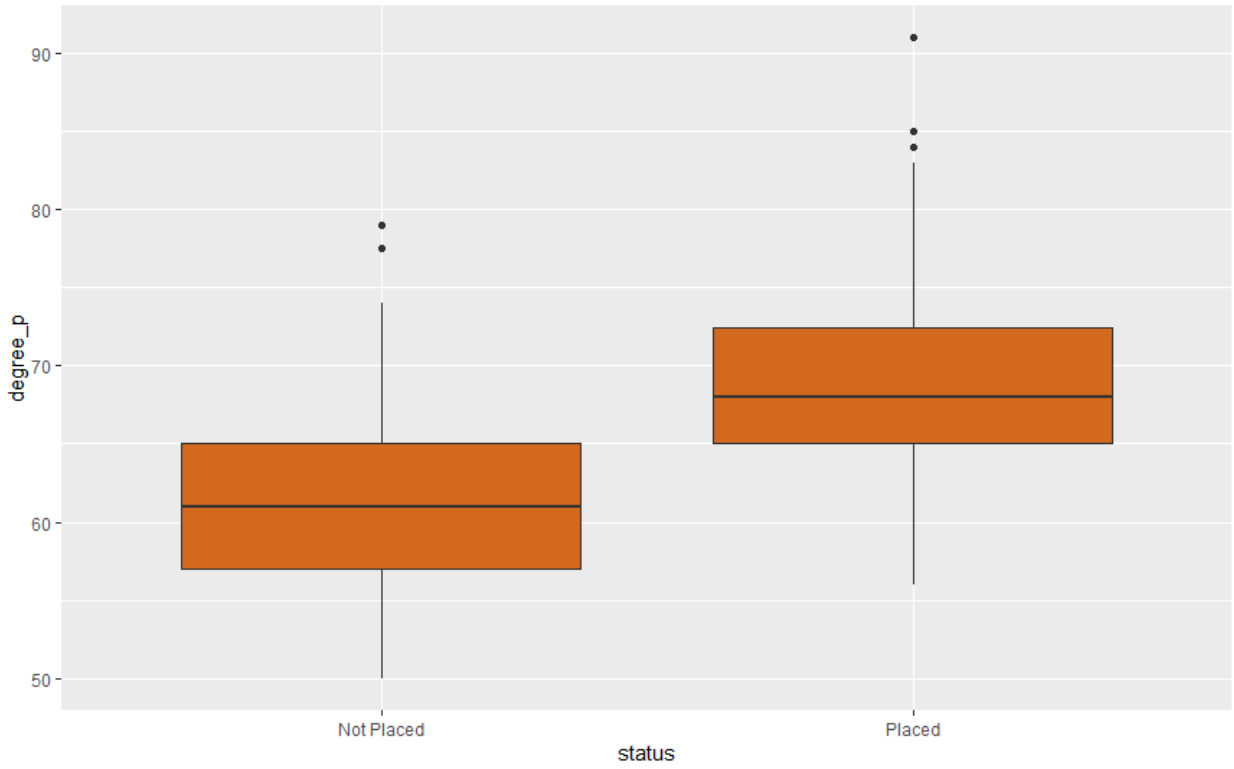
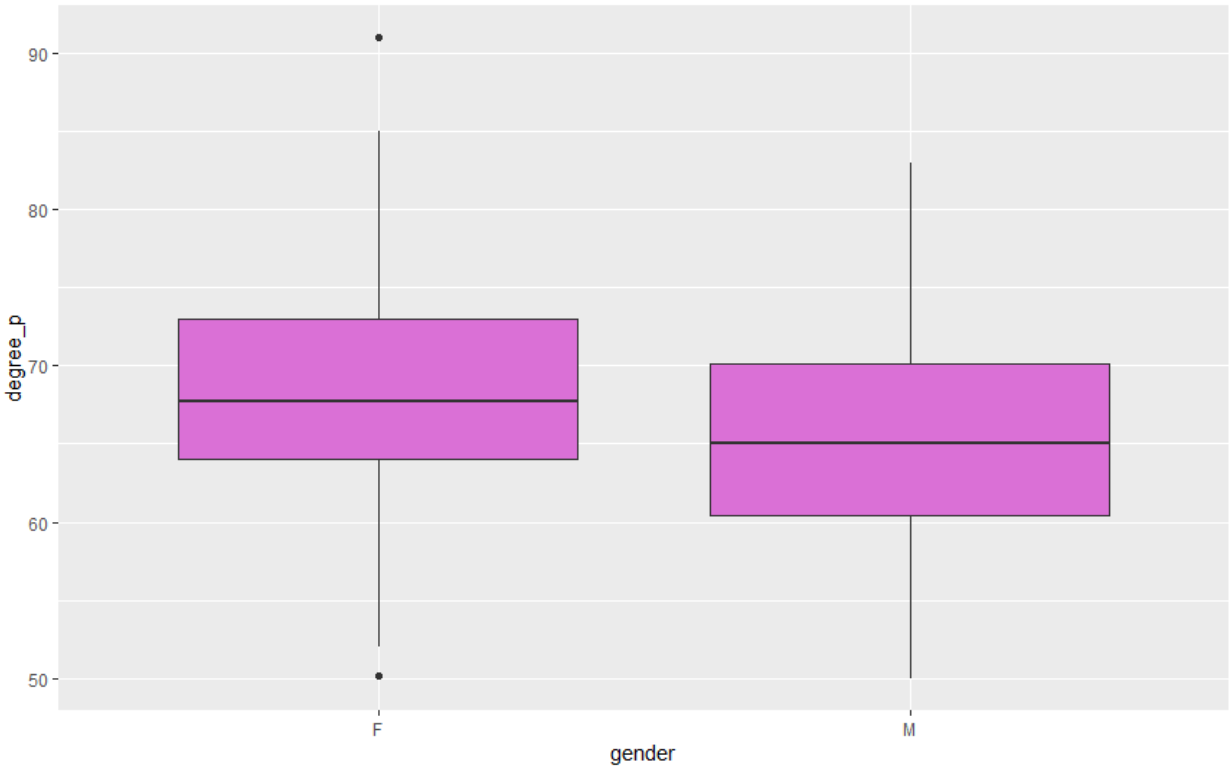


Figure 2.1 is a box & whiskers plot depicting the degree percentage of students that are getting placed and students that are not getting placed. It's split into 4 quartile with each quartile representing 25% of the students' degree percentage. Generally, the students that are not getting placed have a lower degree percentage than the students that are getting placed. If we look closely, the minimum degree percentage of students not getting placed is 50%, lower than the minimum degree percentage of students getting placed, which is around 56%. If we move on to first quartile, we can see that 25% of the students that are not getting placed have a degree percentage lower than 57%. While for the students that are getting placed the lowest of the first quartile is equivalent to the highest of the fourth quartile for the students not getting placed. The interquartile range for students not getting placed is 8% (= 65% (Q3) - 57% (Q1)), and the interquartile range for students getting placed is 7% (= 72% (Q3) - 65% (Q1)). The median for students getting placed is much higher than students not getting placed with 68% in comparison with 61%. The maximum degree percentage for students not getting placed is 79%, and 91% for the students getting placed both being the outliers of each. There are, however, two outliers for students not getting placed at 77% and 79%, and three outliers for students getting placed at 84%, 85%, and 91%.

Figure 2.2 Box & whiskers plot showing gender vs. degree percentage



This box & whiskers plot shows the degree percentage of female and male. We can see that in the female box & whiskers, there is two outliers, one below the minimum and one above the maximum. The lower outlier for female is at 50%, and the higher outlier for female is at 91%. As I've mentioned above, box & whiskers plot split students into four quartile, each representing 25% of the students in the range. The lowest 25% of the female student have a degree percentage between 52% and 64%, and lowest 25% of the male students have a degree percentage between 50% and 60%. The highest 25% of the female students have a degree percentage between 73% and 85%, while the highest 25% male students have a degree percentage between 70% and 83%. On the same note, female students have a minimum and maximum of 50% and 91%, both being the outliers. The male students have a minimum and maximum of 50% and 83%, respectively. The interquartile range for female students is 9% (= 73% (Q3) - 64% (Q1)), and the interquartile range for male students is 10% (= 70% (Q3) - 60% (Q1)). The median degree percentage for females students is 68%, higher than the median degree percentage for male students of 65%. Comparing box & whiskers plots for female and male students can not lead us to a conclusion that there is a significant difference of degree percentage between male and female.

Figure 3 Density plot showing degree percentage frequency

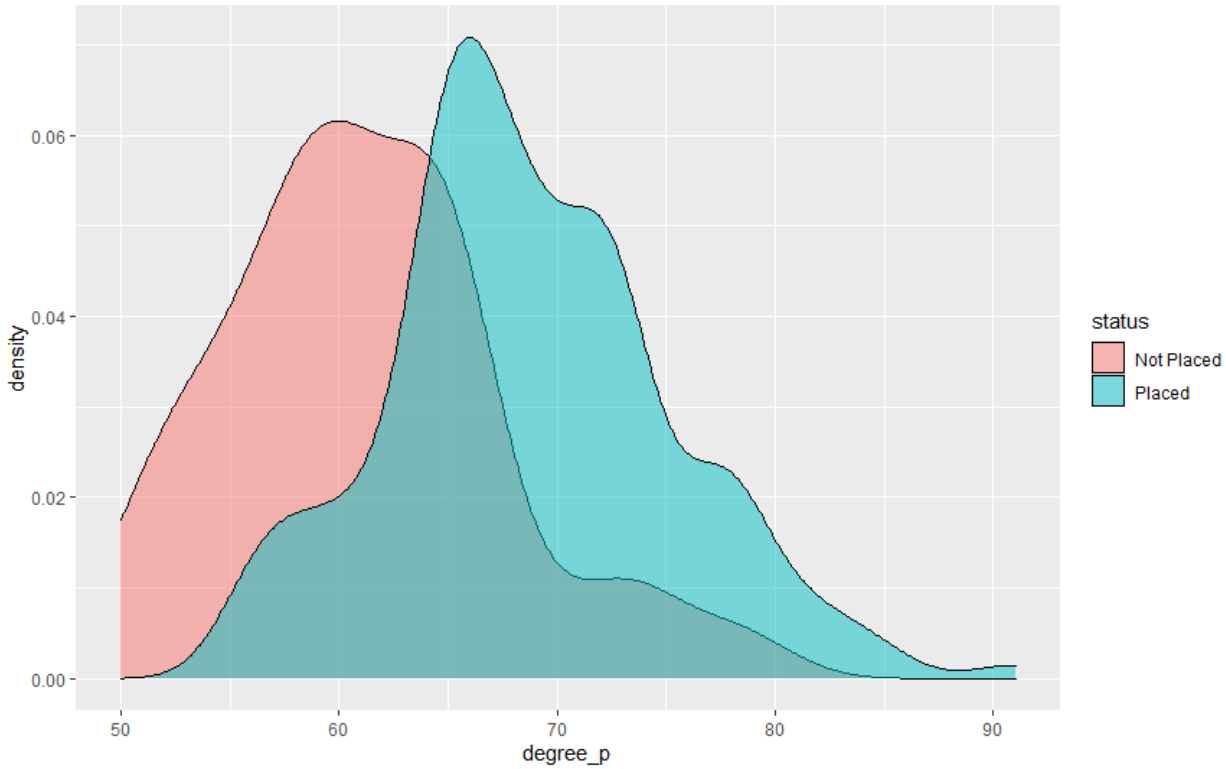


Figure 3 shows the frequency distribution of degree percentage for both students getting placed and not getting placed. We can also see the distribution of data over a continuous period of time. If we look at the density plot for students not getting placed and students getting placed, we can see that they're both skewed to the right. The median degree percentage for students not getting placed is roughly 68%, while the median degree percentage for students not getting placed is roughly around 61%. We can also find out the mean by eyeballing the density plot. Since both graphs are screwed to the right, the mean will be a little larger than the median for both students getting placed and students not getting placed. The mean degree percentage for students not getting placed is roughly around 64%, while the mean degree percentage for students getting placed is roughly around 71%. The highest value of the density plot is around 60% of degree percentage and 66% of degree percentage for students not getting placed and students getting placed, respectively. If we look at the area where the two graphs intersect, it's roughly 0.5 (=0.01*5*10), which is 50% of the degree percentage data falls inside the intersected area between students getting placed and students not getting placed.

Figure 4.1 Violin plot showing higher secondary education percentage vs. placement

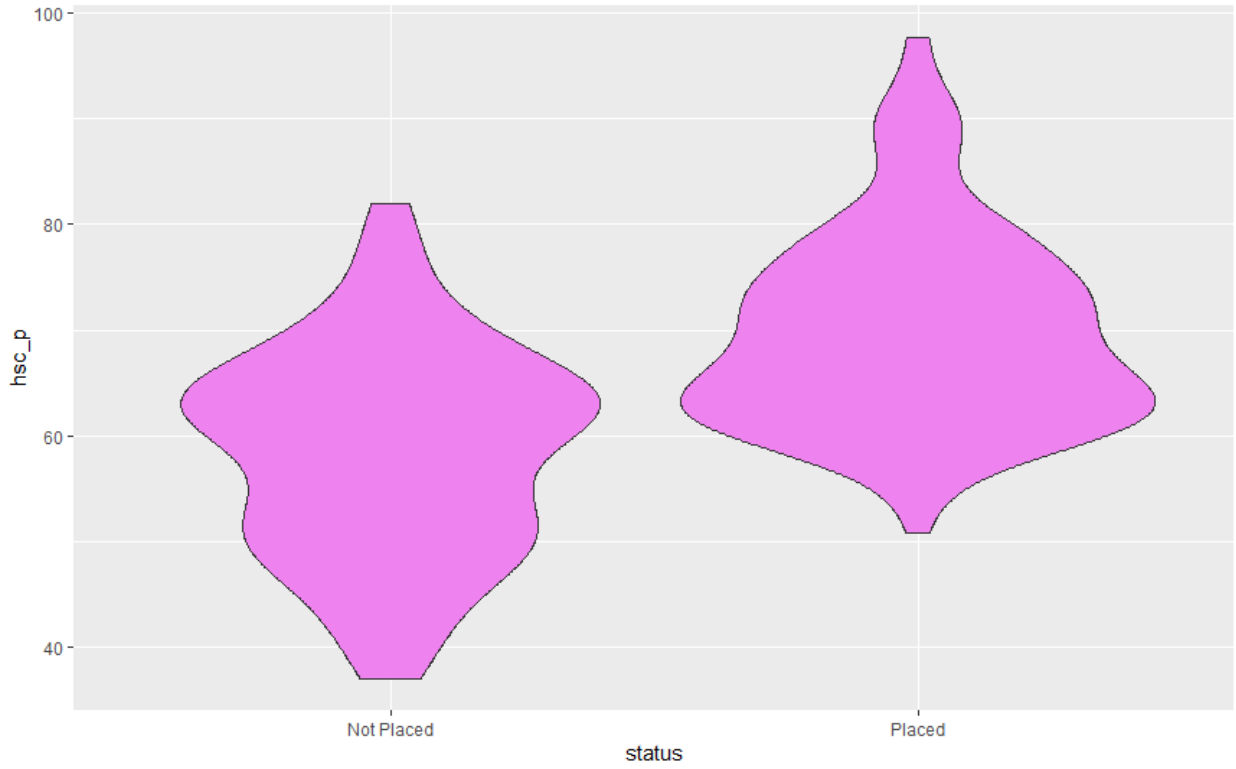


Figure 4.1 is a violin plot, meaning it shows the peaks in the data, as it is a hybrid of a box plot and a kernel density plot. It also has an interquartile range, median, a maximum, and a minimum with outliers. This graph shows the highest frequency of higher secondary education percentageout of the two different placement (getting placed and not getting placed). The median can be found by looking at the widest part of the plot. The median for students not getting placed is roughly 64% as the highest frequency, and the median for students getting placed is also roughly around 64% as the highest frequency. This means that for both students not getting placed and students getting placed, 64% of higher secondary education percentage is the most common. This shows that in terms of higher secondary education percentage, students getting placed and students not getting placed is around the same educational level.

Figure 4.2 Violin plot showing higher secondary education percentage vs. gender

As I've said above, a violin plot is a hybrid box plot. Figure 4.2 shows the highest frequency of higher secondary education percentage out of female students and male students. The median/highest frequency for female students is around 62% and the median/highest frequency for male students is roughly 63%. The male students have a slightly higher percentage for higher secondary education, overall. Yet, we cannot conclude that there is a significant difference between the two as they're so close to each other.

Figure 5 Normality Plot for Degree Percentage

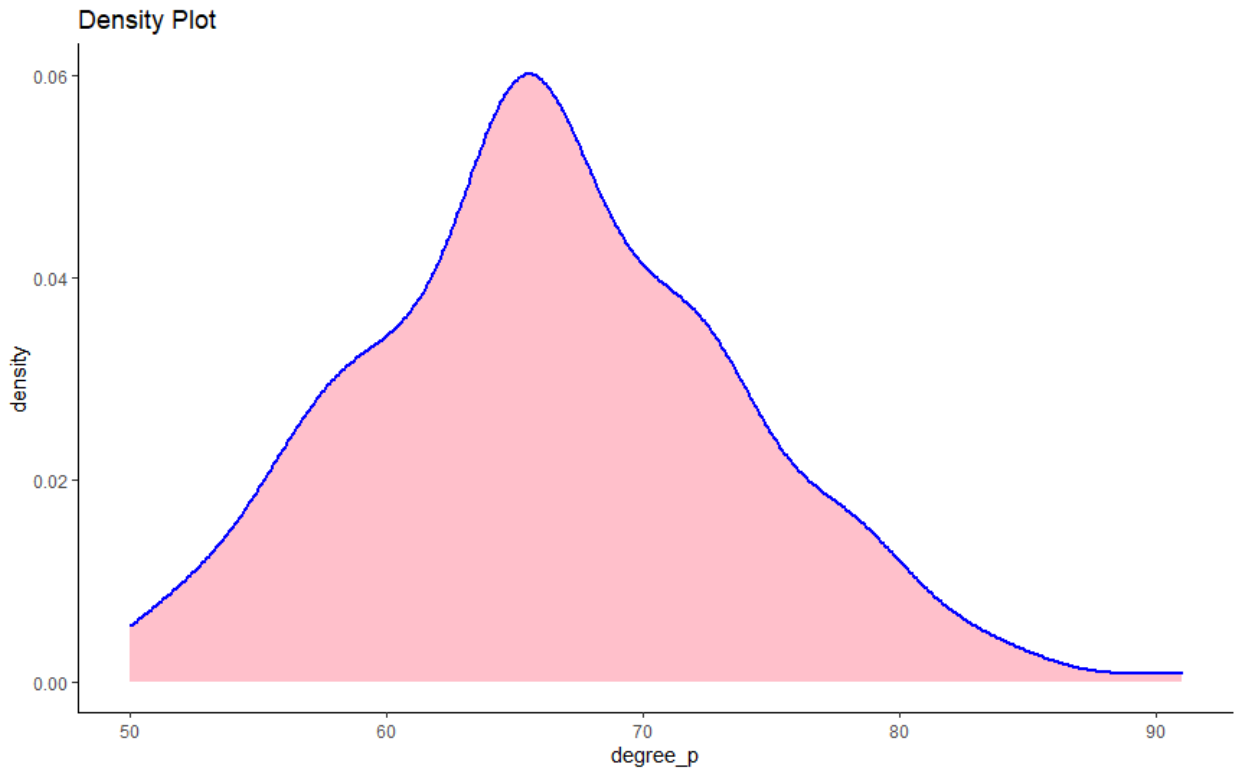


Figure 5 is a density plot for Degree Percentage. This is a roughly normal distribution, with a median of roughly 65% degree percentage. We know that this is a normal distribution because of the bell-shaed curve. The mean should be the same as median if we assume that this is a normal distribution, which is also roughly 65% degree percentage. This means that the average degree percentage out of all the students is 65%.

Figure 6 Correlation Color Chart between 6 variables

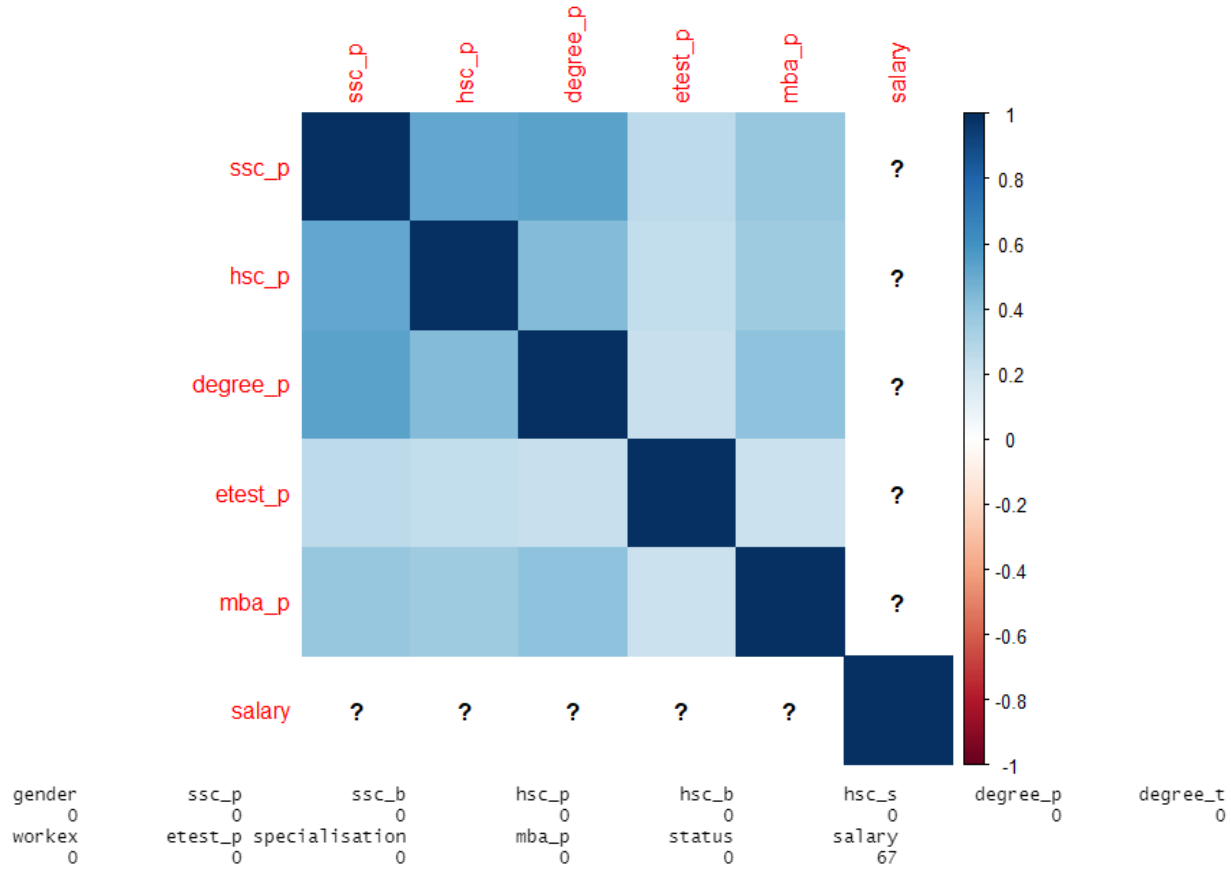


Figure 6 shows the correlation and relationship between the 6 variables, including Secondary Education percentage (10th grade), Higher Secondary Education percentage (12th grade), Degree Percentage, Employability test percentage, MBA percentage, and Salary offered by corporate to candidates. The color key on the right side of the table shows the correlation coefficient (from -1 to 1), and it illustrates whether the variables have strong or weak, negative or positive correlation between 2 variables from the colum and the row of the box. There are question marks shown on the Salary variable column and row because there are 67 missing values in Salary variable. If the data is missing, it is not possible to assess the correlation in the usual way. Due to a lot of values are missing, it is not pssible to conclude the direction of the correlation. The diagonal dark blue boxes across the plot are correlation boxes for their own variables. Therefore, it always has a correlation coefficient of positive 1. There is a moderate positive correlation between Secondary Education percentage and Higher Secondary Education percentage, with a correlation coefficient of roughly 0.6. Additionally, there is also a moderately strong positive correlation between Degree Percentage and Secondary Education percentage, with a correlation coefficient of roughly 0.7. The variable Employability test percentage has a weak positive correlation (correlation coefficient of 0.2) to all other four variables, including Secondary Education percentage, Higher Secondary Education percentage, Degree Percentage, and MBA percentage. This indicates a weaker influence and relationship for the Employability test percentage to the other variables. The MBA

percentage also has a weak positive correlation coefficient of 0.4 with other variables, indicating a stronger influence compared to Employability test percentage, but weaker influence compared to ther other 3 variables.

Figure 7 Pairs Plot for Campus Recruitment

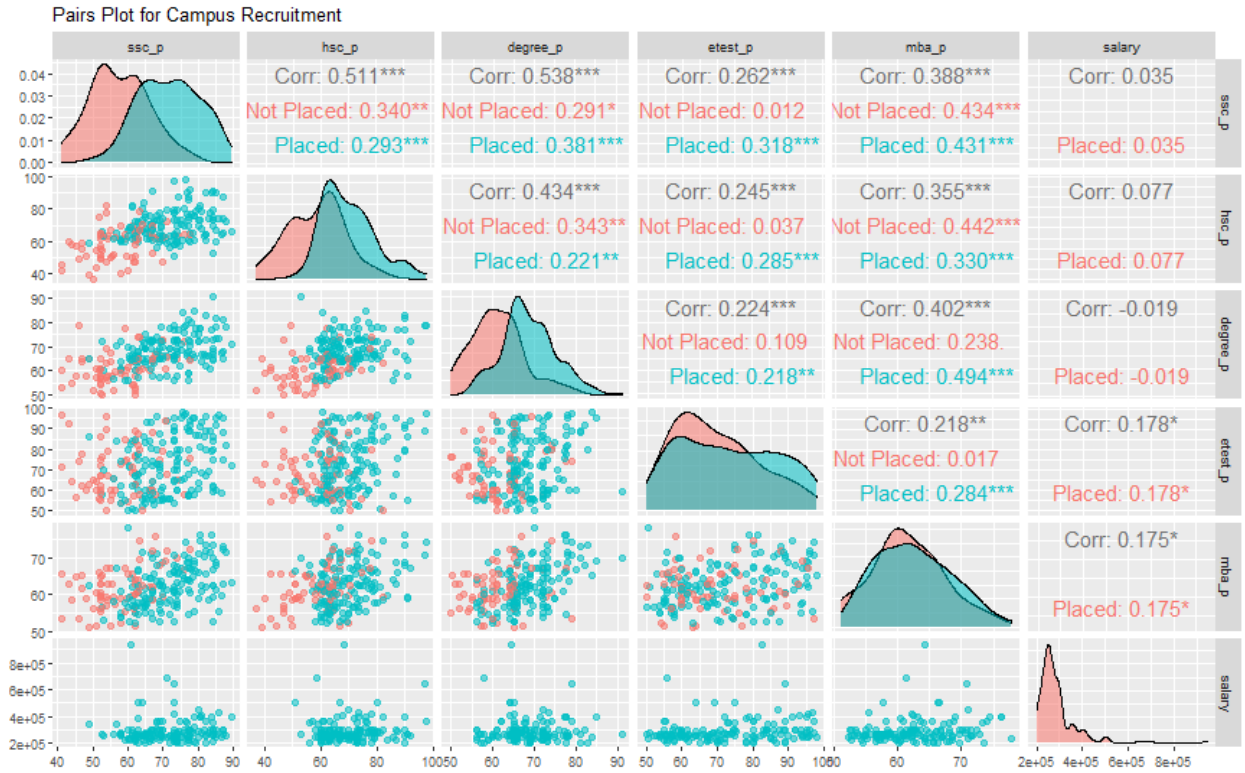


Figure 7 is a multiple scatter plot that displays the relationship between two variables for all six variables. Due to 67 missing values mentioned earlier, the salary column does not have enough data to find the correlation for students not getting placed. The overall correlation is the strongest between Degree Percentage and Secondary Education percentage with a correlation coefficient of 0.538, followed by 0.511 between Higher Secondary Education percentage and Secondary Education percentage. The strongest correlation for students not placed is between MBA percentage and Higher Secondary Education percentage, with a correlation coefficient of 0.442. The strongest correlation for students getting placed is between Degree Percentage and MBA percentage, with a correlation coefficient of 0.494.

Figure 8 Scatter Plot with Regression Line showing Degree Percentage vs. Higher Secondary Education Percentage

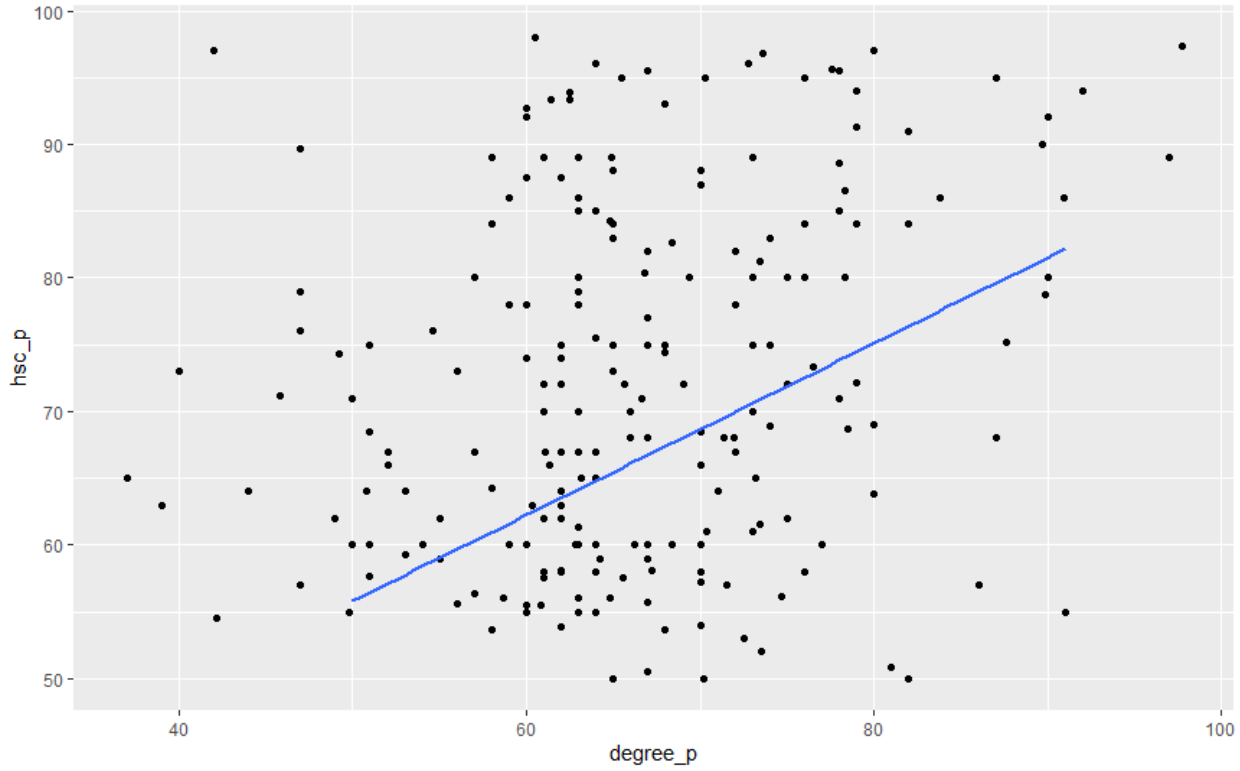


Figure 8 shows the relationship between Degree Percentage and the Higher Secondary Education Percentage (12th grade). The two variables seem to have a weak positive correlation of 0.4, which mirrors what is shown above in Figure 6 (Correlation Color Chart). This means that a student with a higher Degree Percentage is also more likely to have a larger Higher Secondary Education percentage. For example, if we look at the graph, 80% degree percentage is correlated with 75% higher secondary education percentage (according to the regression line). This does not necessarily mean that there is a causal relationship between the two but rather a correlation between the two variables.

Figure 9 Simple Regression

```
Call:
lm(formula = etest_p ~ hsc_p, data = cr)

Residuals:
    Min      1Q  Median      3Q     Max
-26.779 -10.728  -1.105   9.214  32.166

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.29279    5.43978   9.613  < 2e-16 ***
hsc_p        0.29861    0.08093   3.690 0.000285 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.9 on 213 degrees of freedom
Multiple R-squared:  0.06008,   Adjusted R-squared:  0.05567
F-statistic: 13.62 on 1 and 213 DF,  p-value: 0.0002849
```

This is a simple linear regression performed between Employability test percentage and Higher Secondary Education percentage. The residual standard error is 12.9. This means that the average

deviation of the data points around the regression line is 12.9% away from the regression line. If we look at the adjusted r-squared value, 0.05567, meaning 6% of the variation in Employability test percentage can be explained by the linear regression. The F-statistics is 13.62, which means we reject the null hypothesis and states that there is no relationship between the Employability test percentage and Higher Secondary Education percentage. The variable Higher Secondary Education percentage also has the highest t-value, and the only t-value = 3.690, in the simple linear regression. So this means that it has the largest influence on this linear regression model.

Figure 10 Multiple Regression

```
Call:
lm(formula = salary ~ ssc_p + hsc_p + degree_p + etest_p + mba_p,
    data = cr)

Residuals:
    Min      1Q  Median      3Q     Max
-104005  -50454   -9766   14722  619162

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 168690.94   98916.08   1.705   0.0903 .
ssc_p         -641.63    1015.97  -0.632   0.5287
hsc_p           59.53     887.49   0.067   0.9466
degree_p     -2061.50    1368.74  -1.506   0.1343
etest_p       1112.66     600.13   1.854   0.0658 .
mba_p         3548.13    1593.41   2.227   0.0275 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 91730 on 142 degrees of freedom
  (67 observations deleted due to missingness)
Multiple R-squared:  0.06936,   Adjusted R-squared:  0.03659
F-statistic: 2.117 on 5 and 142 DF,  p-value: 0.06682
```

This is a multiple linear regression model performed between Salary, Secondary Education percentage, Higher Secondary Education percentage, Degree percentage, Employability Test percentage, and MBA Percentage. The residual standard error is 91730. This means that the average deviation of the data points around the regression line is $91730 (salary) away from the regression line. If we look at the adjusted r-squared value, 0.03659, meaning roughly 4% of the variation in the Salary variable can be explained by the linear regression. The F-statistics is 2.117, which means we fail to reject the null hypothesis and states that there is likely a relationship between the Salary and one of the other x-variables, such as Secondary Education percentage, Higher Secondary Education percentage, Degree percentage, Employability Test percentage, and MBA Percentage. The variable MBA percentage also has the highest t-value, 2.227, in this multiple linear regression model. So this means that it has the largest influence on this linear regression model.

Figure 11 Correlation matrix as visualization

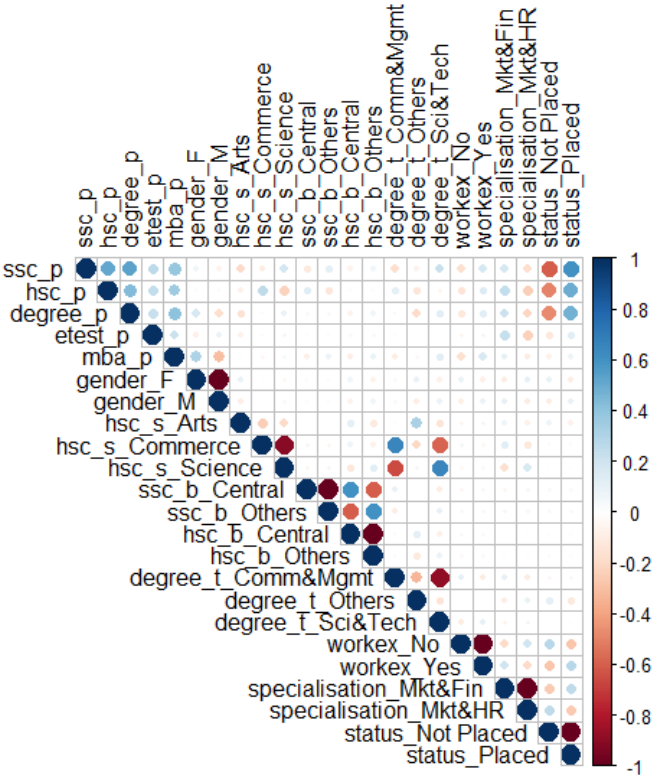


Figure 11 is a correlation matrix that illustrates an observable pattern between all the variables. The line going diagonally from the top left to the bottom right has a correlation coefficient of 1 because it always perfectly correlates to itself. We can also see that gender female and gender male is negatively correlated with a correlation coefficient of -1 because it's a binary variable of either female or male. Higher Secondary Education specializes in Commerce also has a strong negative correlation towards Higher Secondary Education specializes in Science. It is also due to the fact that it is a boolean variable. The same trend follows for the Board of Education (Central vs. Others), Work Experience (Yes vs. No), Post Graduation Specialization (Marketing & Finance vs. Marketing & Human Resources), and Status of Placement (Placed vs. Not Placed). Interestingly, getting placed is moderately and positively correlated with Secondary Education Percentage, Higher Secondary Education Percentage, and Degree Percentage, with a correlation of 0.7, 0.5, and 0.5, respectively. This suggests that the higher percentage you have for the 3 variables mentioned before, the more likely you are to get placed. However, this does not mean that there is a causal relationship but rather a correlational one. There is also a moderate correlation between getting a higher Degree Percentage on specialization and getting a higher Secondary Education Percentage on the same specialization.

Figure 12 Accuracy of k-NN

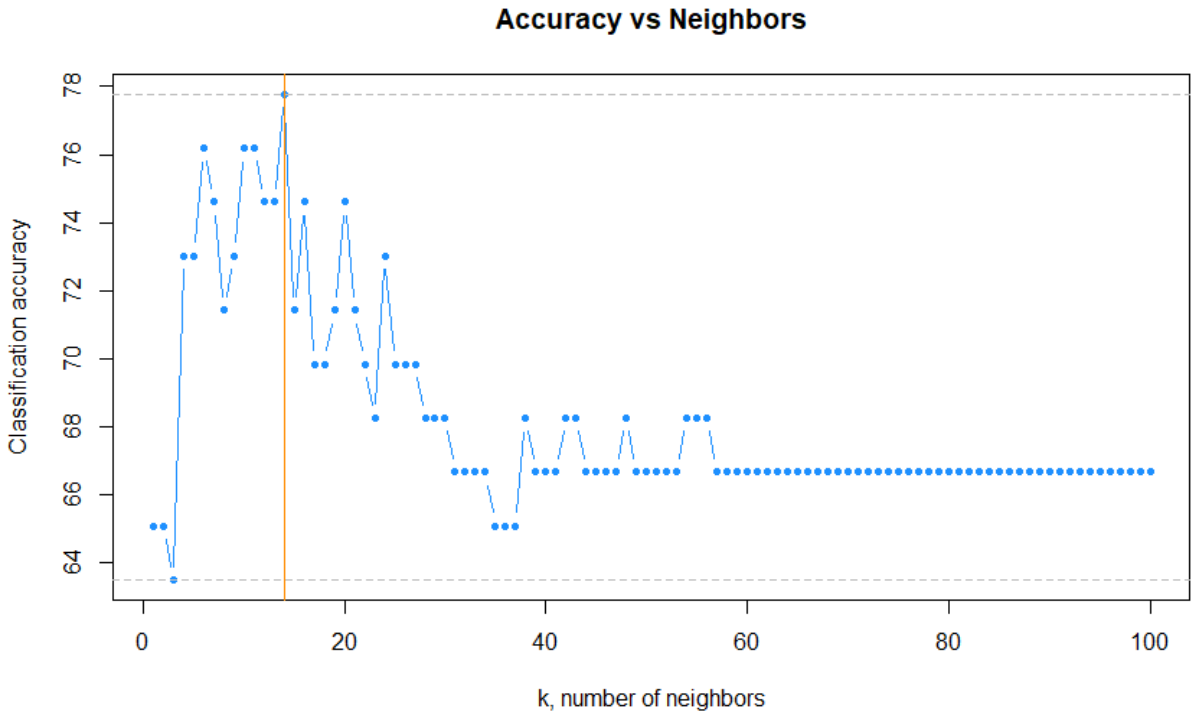


Figure 12 shows the best k value with the highest accuracy is k = 17 (approximately), which is marked by the orange vertical line. The top grey horizontal line marked the maximum accuracy for the k-NN model, which is 78% (approximately). This means the percentage of correct classifications for k is 78% for k = 17. The lowest accuracy is 64% for k = 3 (approximately).

Figure 13 Precision of k-NN

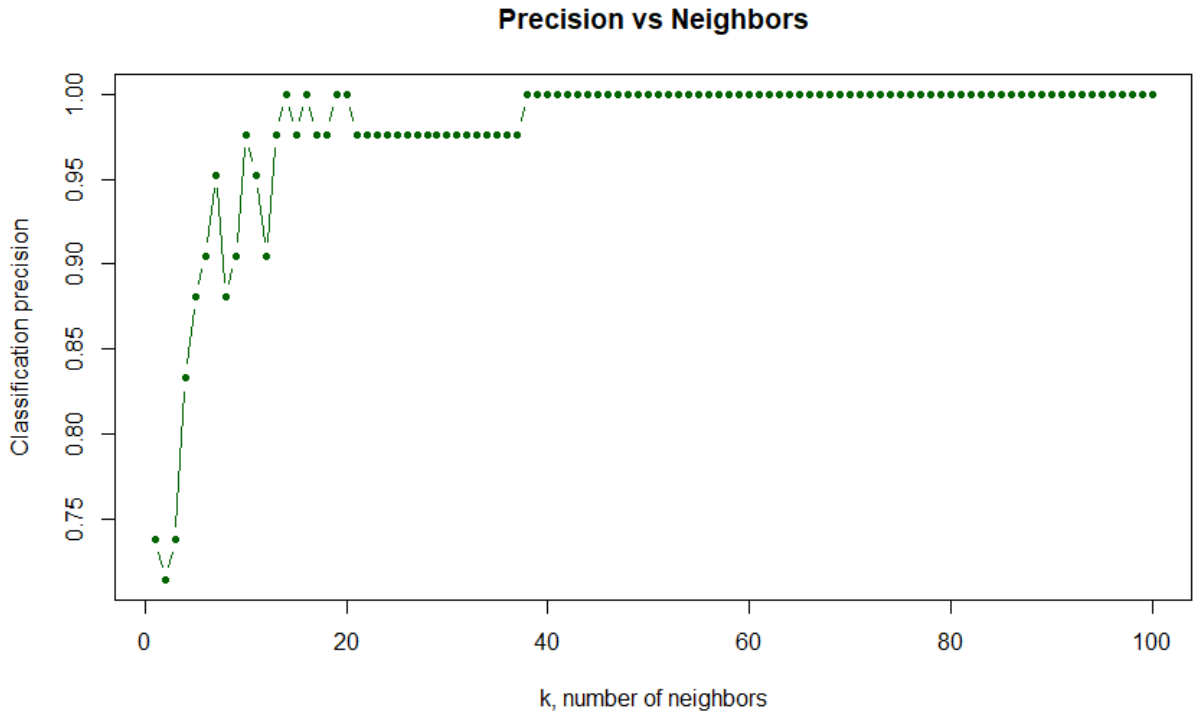


Figure 13 shows the relationship between precision with each k-value. It's also known as the positive predictive value and it showcases how precise your model is at classifying the true cases. The first k value that hits 1.00 precision is k = 17 (approximately). This shows that when k = 17, the accuracy and precision are all relatively high. K values of 19, 21, 22, and 40 to 100 also have the highest classification precision of 1.00. The precision is lowest when k = 3 (approximately), which also corresponds to the previous accuracy graph of having the lowest accuracy at k = 3.

Figure 14 Classification Matrix

```
> knn <- knn(plac.training, plac.test, plac.trainLabels, k=6, prob = TRUE)
> # Accuracy of knn for k = 6
> round(sum(plac.testLabels==knn)/length(plac.testLabels)*100,2)
[1] 79.37
> # Cross Classification / Confusion matrix of knn for k = 6
> table(plac.testLabels, knn)
                knn
plac.testLabels Not Placed Placed
     Not Placed         10     11
     Placed              2     40
```

The classification matrix here describes the misclassification rate of getting placed or not getting placed. The training set and test set are broken into ⅔ and ⅓, respectively. The best value of k is 6, as shown above, for this k-NN matrix. The predicted number of students not getting placed is 12 and the predicted number of students getting placed is 51. However, in reality, there are 21 students not getting placed and only 42 students getting placed. There are a total of 13 misclassified data (2 from Not Placed

and 11 from Placed). The misclassification rate is, therefore, 20.63% (=13/63*100%). This means that, on average, around 79.37% of the observations are classified correctly, which mirrors the data shown above. It also corresponds to Figure 12 Accuracy of the k-NN model where the accuracy is around 78%.

Conclusion

Figure 1.1 shows that although male students have a slightly higher chance of getting placed, there is no clear relationship between gender and placement. In figure 1.2, it shows that students with work experience have a higher chance of getting placed. Although it's not significant, there is likely a correlational relationship between the two variables. In figure 1.3, it is hard to come to a conclusion because of the huge difference between the number of students in different undergraduate degree types. While Science & technology degree does have a higher chance (82%) of getting placed, the number of data for other degrees are not enough for us to be completely sure and to draw a conclusion.

Figure 2.1 shows that there is possibly a significant difference in the Degree Percentage of students getting placed and not getting placed. As the interquartile range does not overlap each other between the two placement variables. The second box & whiskers plot shows the Degree Percentage in female students and male students. Although female students tend to have a higher Degree Percentage, there is no significant difference between the two. Therefore, we can not draw a conclusion of whether or not there is a correlation between Gender and Degree Percentage.

Figure 3 shows that 50% of the Degree Percentage data falls inside the intersected area between students getting placed and not getting placed. Therefore, there is no significant difference between students getting placed and not getting placed in terms of Degree Percentage.

Figure 4.1 shows that there for both getting placed and not getting placed, the Higher Secondary Education percentage is 64%. Therefore, it's likely that the Higher Secondary Education percentage does not have a huge influence on the decision of whether or not a student is getting placed. The second violin plot shows that both female and male students have the highest frequency of 63% for Higher Secondary Education percentage. This means that the Gender variable does not have a relationship with the Higher Secondary Education percentage.

Figure 5 shows that the median for Degree Percentage is 65% for all students. On the other hand, figure 6 is a correlation color chart that shows the relationship between the 6 variables, Secondary Education percentage (10th grade), Higher Secondary Education percentage (12th grade), Degree Percentage, Employability test percentage, MBA percentage, and Salary offered by corporate to candidates. There are 67 missing values in the Salary variable, so it is not possible to assess or conclude the direction of the correlation. Degree Peercentage has a moderately strong correlation with Secondary Education percentage, with a correlation coefficient of roughly 0.7. There is also a moderate positive

correlation between Secondary Education percentage and Higher Secondary Education percentage, with a correlation coefficient of roughly 0.6. Figure 7 is a pairs plot with all six variables. The overall correlation for all students is the strongest between Degree Percentage and Secondary Education percentage with a correlation coefficient of 0.538, followed by 0.511 between Higher Secondary Education percentage and Secondary Education percentage. The strongest correlation for students not placed is between MBA percentage and Higher Secondary Education percentage, with a correlation coefficient of 0.442. The strongest correlation for students getting placed is between Degree Percentage and MBA percentage, with a correlation coefficient of 0.494. Figure 8 shows that there is a weak correlation relationship between Degree Percentage and Higher Secondary Education Percentage.

Figure 11 illustrates that getting placed is moderately and positively correlated with Secondary Education Percentage, Higher Secondary Education Percentage, and Degree Percentage, with a correlation of 0.7, 0.5, and 0.5, respectively. This proves that the moderate correlation mentioned above all makes sense.

Figure 12, 13, and 14 describes the results of k-NN model. The highest classification accuracy is 78% when k = 17, approximately. The lowest classification accuracy is 64% when k = 3. While the highest classification precision is k = 17 with a precision of 1.00. The lowest classification precision is k = 3 and a precision of around 0.7. The two graph information correlates with each other. The classification matrix in figure 14 shows that the misclassification rate is 20.63%, meaning around 79.37% of the observations are classified correctly when k = 6. This correlates with the information from figure 12, where it states the maximum accuracy is approximately 78%.