R Project on Graduates Placement

Jenita Jeyakumar

SECTION-1 INTRODUCTION:

The data set used for this project is from Kaggle.com. It contains information about the placement status of graduates from an MBA course. The variables include the percentage, board and/or stream of secondary school, higher secondary school, degree, and MBA, of male and female students. The data also contains Employability test scores and work experience status for all students and the salary of the placed students.

RESEARCH QUESTION: As part of the main analysis of this data set, we will look into factors that determine whether or not a student gets placed.

Data source: https://www.kaggle.com/benroshan/factors-affecting-campus-placement

```
## Warning: package 'tidyverse' was built under R version 4.1.2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.2
```

Taking a look at the data using the "**glimpse**" function:

```
glimpse(placement)
```

```
## Rows: 215
## Columns: 15
## $ sl_no          <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, ~
## $ gender         <chr> "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M", ~
## $ ssc_p          <dbl> 67.00, 79.33, 65.00, 56.00, 85.80, 55.00, 46.00, 82.00,~
## $ ssc_b          <chr> "Others", "Central", "Central", "Central", "Central", "~
## $ hsc_p          <dbl> 91.00, 78.33, 68.00, 52.00, 73.60, 49.80, 49.20, 64.00,~
## $ hsc_b          <chr> "Others", "Others", "Central", "Central", "Central", "O~
## $ hsc_s          <chr> "Commerce", "Science", "Arts", "Science", "Commerce", "~
## $ degree_p       <dbl> 58.00, 77.48, 64.00, 52.00, 73.30, 67.25, 79.00, 66.00,~
## $ degree_t       <chr> "Sci&Tech", "Sci&Tech", "Comm&Mgmt", "Sci&Tech", "Comm&~
## $ workex         <chr> "No", "Yes", "No", "No", "No", "Yes", "No", "Yes", "No"~
## $ etest_p        <dbl> 55.00, 86.50, 75.00, 66.00, 96.80, 55.00, 74.28, 67.00,~
## $ specialisation <chr> "Mkt&HR", "Mkt&Fin", "Mkt&Fin", "Mkt&HR", "Mkt&Fin", "M~
## $ mba_p          <dbl> 58.80, 66.28, 57.80, 59.43, 55.50, 51.58, 53.29, 62.14,~
## $ status         <chr> "Placed", "Placed", "Placed", "Not Placed", "Placed", "~
## $ salary         <dbl> 270000, 200000, 250000, NA, 425000, NA, NA, 252000, 231~
```

SECTION-2 DATA ANALYSIS PLAN:

- The outcome or response variable (Y) is the "status" which is a categorical variable that gives info about the status of placement of the students. It has two values, either placed or not placed. It can also be treated as a binary variable since it has two values- placed and not placed.

  The explanatory or predictor variables (X) could be any of the following :

1. SSC or senior secondary school percentage
2. HSC or higher secondary school percentage
3. Degree percentage and specialization
4. Employability test percentage
5. MBA percentage and specialization
6. Work experience

- The entire data set can be grouped using any of the following variables:

1. Gender (Female and Male)
2. Work experience (Yes and No)
3. MBA specialization (Mkt&Fin and Mkt&HR)

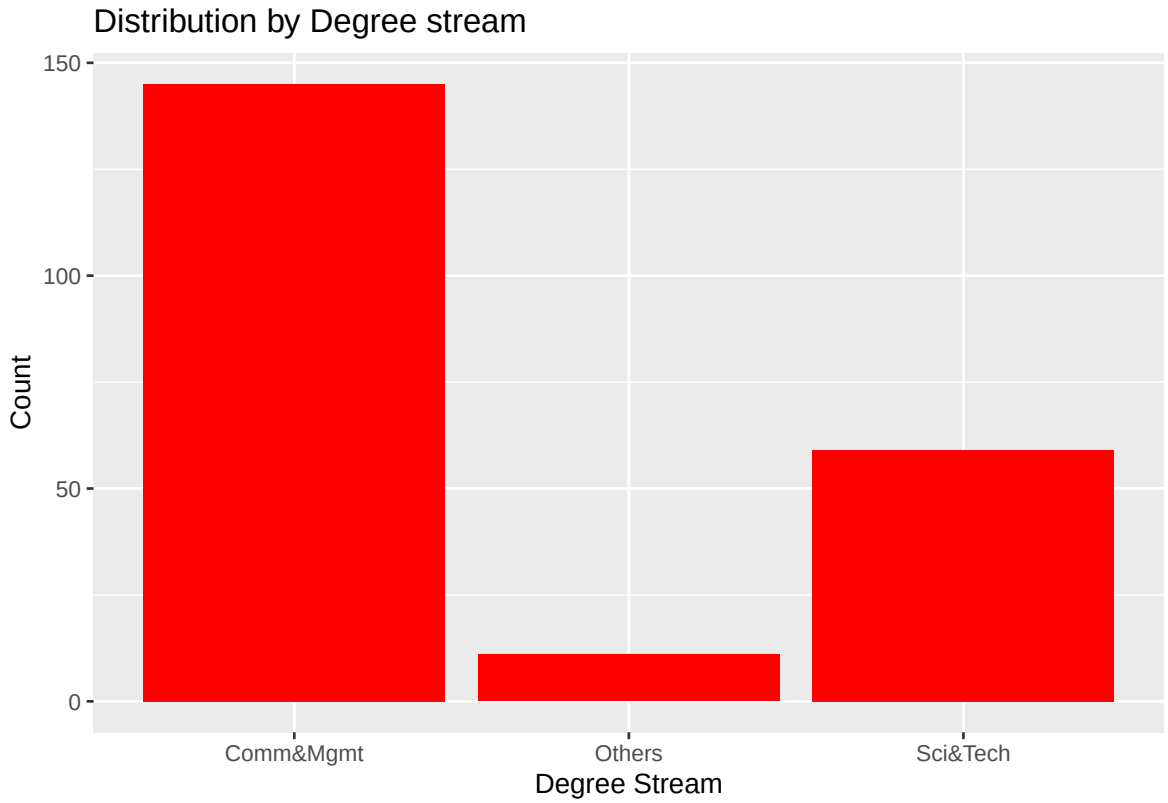
- The preliminary analysis will include some basic univariate analysis.
  Below gives the summary statistics for the dataset:

```
summary(placement)
```

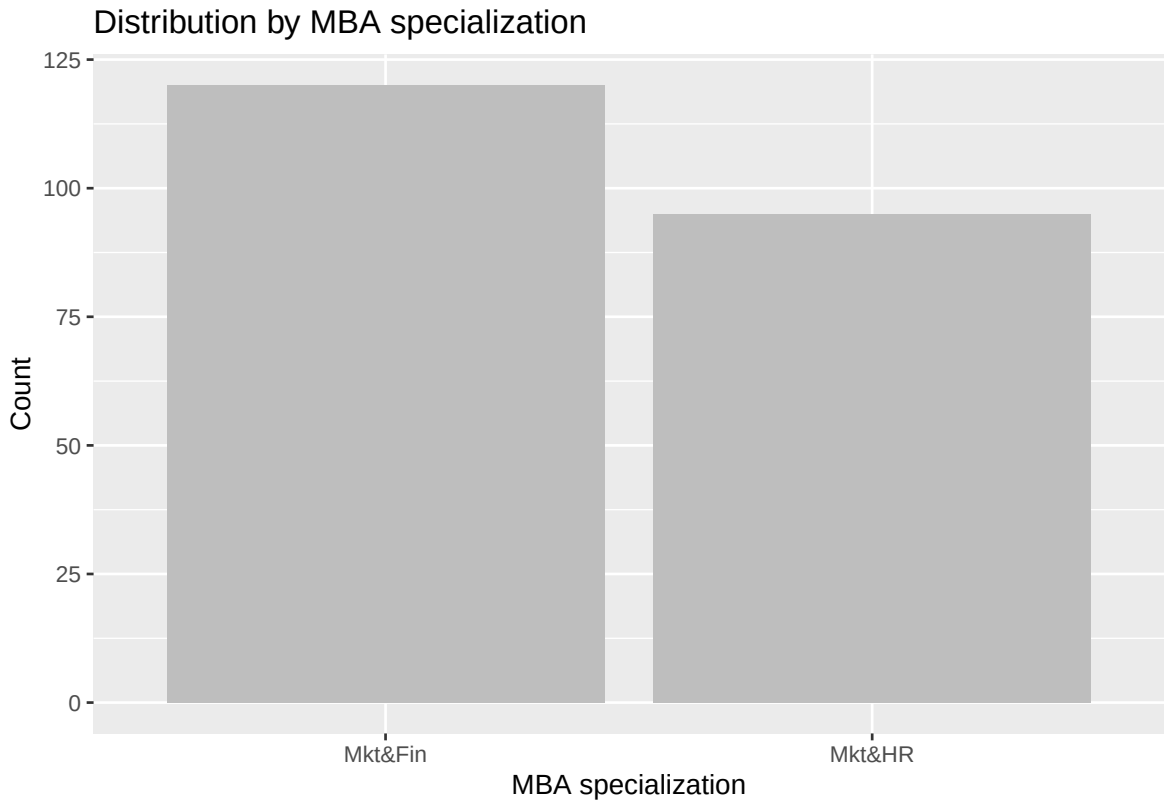
```
##      sl_no          gender              ssc_p          ssc_b
##  Min.   :  1.0   Length:215         Min.   :40.89   Length:215
##  1st Qu.: 54.5   Class :character   1st Qu.:60.60   Class :character
##  Median :108.0   Mode  :character   Median :67.00   Mode  :character
##  Mean   :108.0                      Mean   :67.30
##  3rd Qu.:161.5                      3rd Qu.:75.70
##  Max.   :215.0                      Max.   :89.40
##
##      hsc_p          hsc_b              hsc_s              degree_p
##  Min.   :37.00   Length:215         Length:215         Min.   :50.00
##  1st Qu.:60.90   Class :character   Class :character   1st Qu.:61.00
##  Median :65.00   Mode  :character   Mode  :character   Median :66.00
##  Mean   :66.33                                         Mean   :66.37
##  3rd Qu.:73.00                                         3rd Qu.:72.00
##  Max.   :97.70                                         Max.   :91.00
##
##    degree_t           workex             etest_p     specialisation
##  Length:215         Length:215         Min.   :50.0   Length:215
##  Class :character   Class :character   1st Qu.:60.0   Class :character
##  Mode  :character   Mode  :character   Median :71.0   Mode  :character
##                                        Mean   :72.1
##                                        3rd Qu.:83.5
##                                        Max.   :98.0
##
##      mba_p          status              salary
##  Min.   :51.21   Length:215         Min.   :200000
##  1st Qu.:57.95   Class :character   1st Qu.:240000
##  Median :62.00   Mode  :character   Median :265000
##  Mean   :62.28                      Mean   :288655
##  3rd Qu.:66.25                      3rd Qu.:300000
##  Max.   :77.89                      Max.   :940000
##                                     NA's   :67
```

PRELIMINARY ANALYSIS:

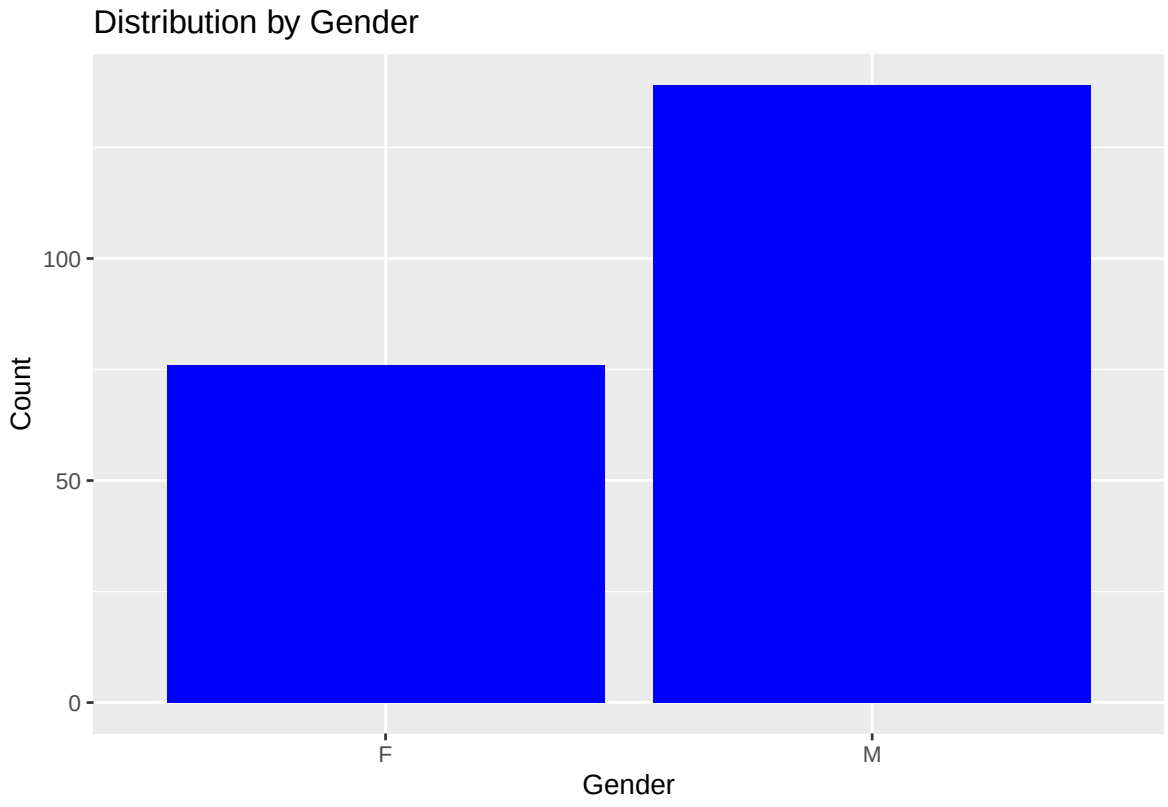
- Below is the distribution of students by Degree stream:



- Below is the distribution of students by MBA specialization:



– Below is the distribution of students by Gender:



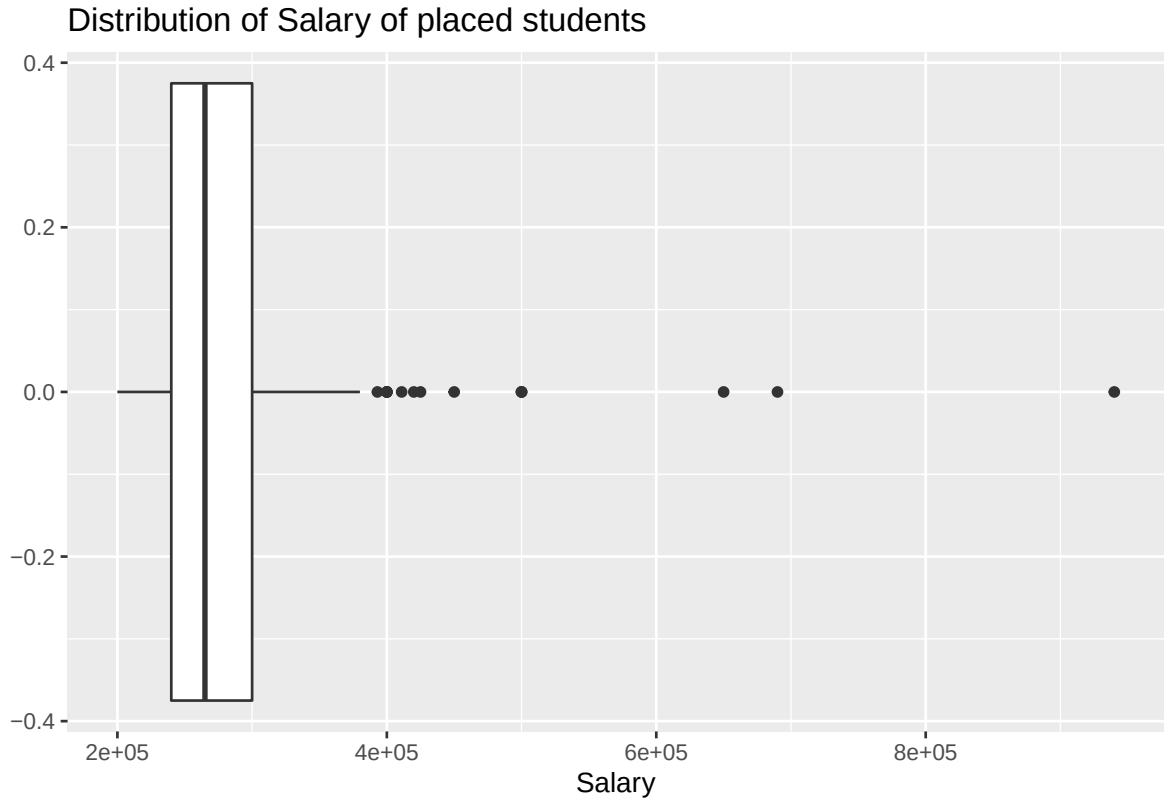
- We will determine if Data cleaning and transformation steps are required and the same will be executed:
  - The summary shows that there are NA values in the 'salary' variable. These NA's correspond to the salary for students that are not placed, hence not making it necessary to clean the NA values because it makes sense to have them.
- We will check if there are outliers for the variable 'salary' in the data using a boxplot, and will see how to handle them.
- The statistical method 'Linear regression' can be used to determine relationship between various factors that lead to placement.

PROJECT- ANALYSIS and VISUALIZATIONS

A. DATA CLEANING AND TRANSFORMATION:

- The summary of the data set shows that there are 67 NA values in the 'salary' variable. These NA's correspond to the salary for students that are not placed, hence not making it necessary to clean the NA values because it makes sense to have them.
- We will check if there are outliers for the variable 'salary' in the data using a boxplot, and will see how to handle them.

```
## Warning: Removed 67 rows containing non-finite values (stat_boxplot).
```



Let's check the observations corresponding to the top 5 salary:

```
## # A tibble: 5 x 15
##   sl_no gender ssc_p ssc_b   hsc_p hsc_b   hsc_s degree_p degree_t workex etest_p
##   <dbl> <chr>  <dbl> <chr>   <dbl> <chr>   <chr>    <dbl> <chr>    <chr>    <dbl>
## 1   120 M       60.8 Central  68.4 Central Comm~     64.6 Comm&Mg~ Yes       82.7
## 2   151 M       71   Central  58.7 Central Scie~     58   Sci&Tech Yes       56
## 3   178 F       73   Central  97   Others  Comm~     79   Comm&Mg~ Yes       89
## 4    78 M       64   Others   80   Others  Scie~     65   Sci&Tech Yes       69
## 5   164 M       63   Others   67   Others  Scie~     64   Sci&Tech No        75
## # ... with 4 more variables: specialisation <chr>, mba_p <dbl>, status <chr>,
## #   salary <dbl>
```

The high values of salary are justified for these observations as the percentages are high for these students, and the top 4 students have work experience too.

- The 'status' variable says whether or not the student is placed. We can use the mutate function to make a new column to obtain a numeric representation of the 'status' variable, where the value **'not placed' is 1 and 'placed' is 2.**

```
## # A tibble: 215 x 16
##    sl_no gender ssc_p ssc_b   hsc_p hsc_b   hsc_s degree_p degree_t workex etest_p
##    <dbl> <chr>  <dbl> <chr>   <dbl> <chr>   <chr>    <dbl> <chr>    <chr>    <dbl>
##  1     1 M       67   Others   91   Others  Comm~     58   Sci&Tech No        55
##  2     2 M       79.3 Central  78.3 Others  Scie~     77.5 Sci&Tech Yes       86.5
##  3     3 M       65   Central  68   Central Arts      64   Comm&Mg~ No        75
##  4     4 M       56   Central  52   Central Scie~     52   Sci&Tech No        66
##  5     5 M       85.8 Central  73.6 Central Comm~     73.3 Comm&Mg~ No        96.8
##  6     6 M       55   Others   49.8 Others  Scie~     67.2 Sci&Tech Yes       55
##  7     7 F       46   Others   49.2 Others  Comm~     79   Comm&Mg~ No        74.3
##  8     8 M       82   Central  64   Central Scie~     66   Sci&Tech Yes       67
##  9     9 M       73   Central  79   Central Comm~     72   Comm&Mg~ No        91.3
## 10    10 M       58   Central  70   Central Comm~     61   Comm&Mg~ No        54
## # ... with 205 more rows, and 5 more variables: specialisation <chr>,
## #   mba_p <dbl>, status <chr>, salary <dbl>, status_int <dbl>
```

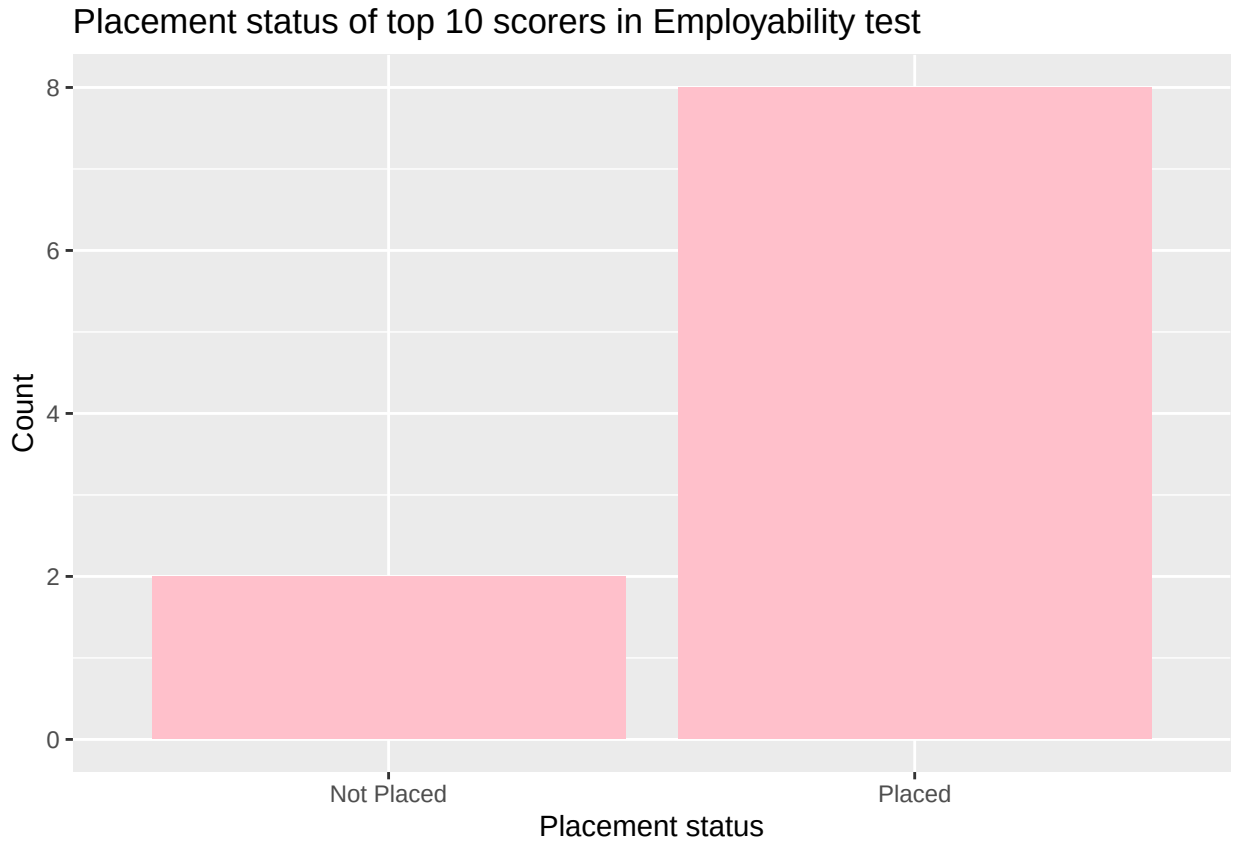
B. ANALYSIS QUESTIONS:

B.1. What is the placement status of the top 10 students in the employability test:

We can see that 20% of the students are not placed in the top 10 performers of the employability test. Indicating that the test is not the only determining factor for the students' placement.

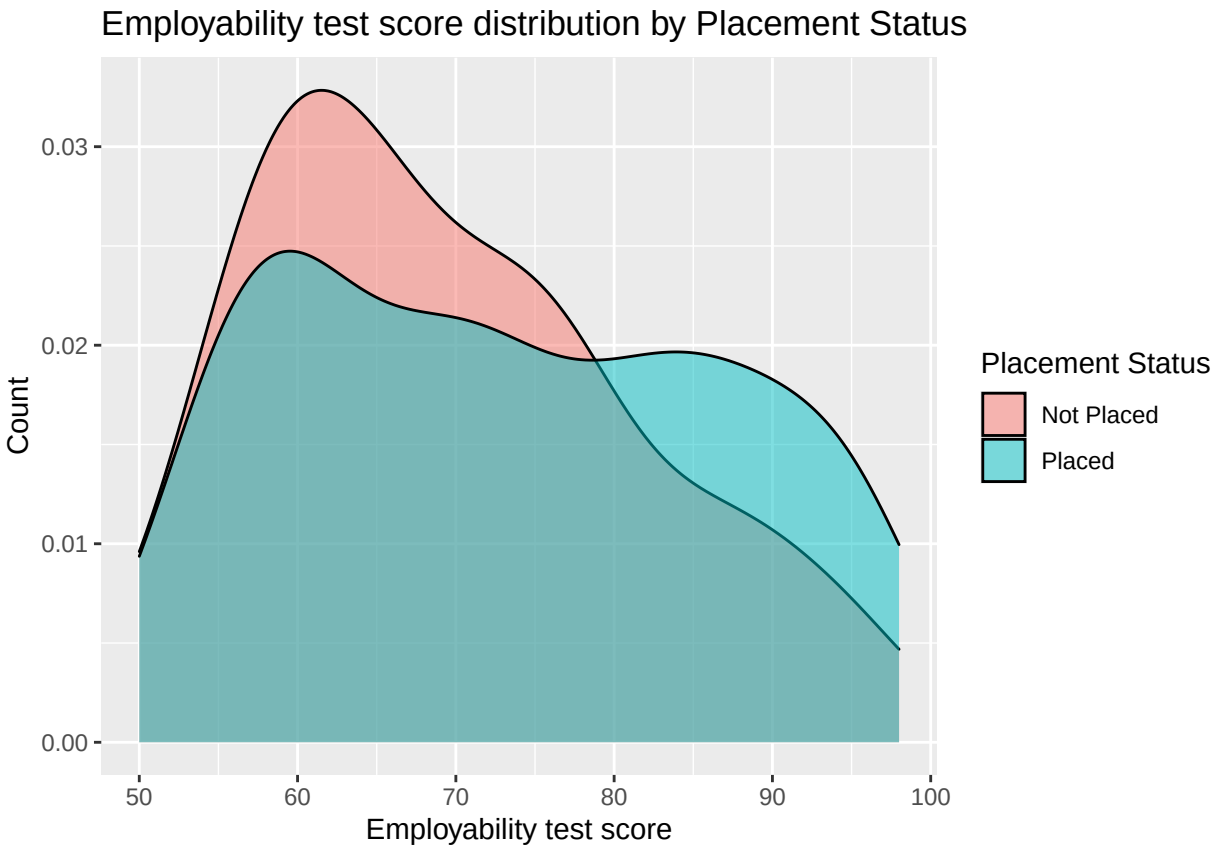
```
## # A tibble: 10 x 8
##    gender degree_p workex etest_p mba_p status     salary status_int
##    <chr>     <dbl> <chr>    <dbl> <dbl> <chr>       <dbl>      <dbl>
##  1 F          84   No        98    65.2 Placed     240000          2
##  2 M          78.9 No        97.4  74.0 Placed     360000          2
##  3 M          78   Yes       97    70.5 Placed     276000          2
##  4 M          60   No        97    53.4 Not Placed     NA          1
```

```
##  5 M          73.3 No        96.8  55.5 Placed     425000          2
##  6 M          66.6 Yes       96    70.8 Placed     300000          2
##  7 F          73   Yes       96    71.8 Placed     250000          2
##  8 F          69.2 No        95.6  66.9 Not Placed     NA          1
##  9 M          78   No        95.5  68.5 Placed     240000          2
## 10 M          65   Yes       95.5  62.2 Placed     420000          2
```



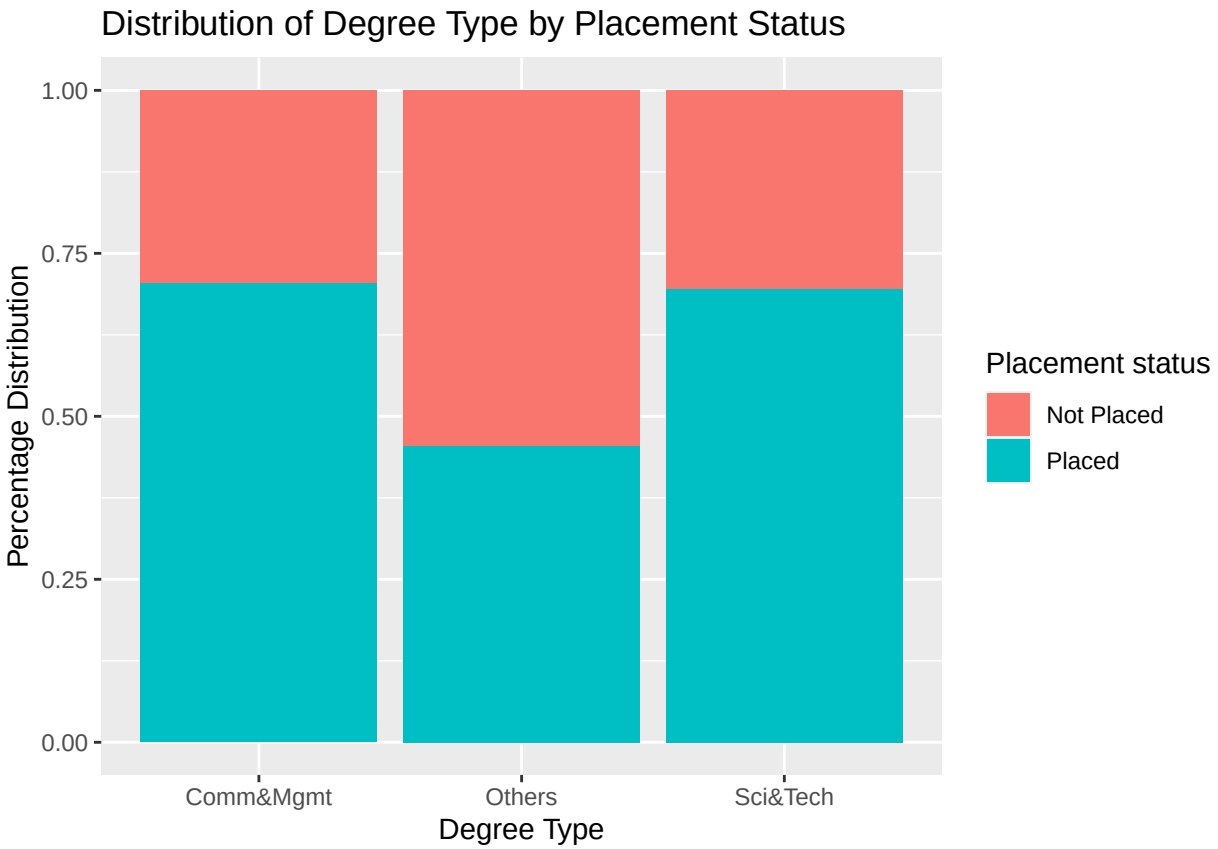
To further prove the above inference let us plot the distribution of the **Employability test scores for all students** as follows:

As we can observe from the below plot, the peak for the students not placed is of lower value but that of students placed has a bimodal distribution, with lower and higher peaks of test scores. This indicates that the Employability test scores is NOT the only determining factor for the placement of students.



B.2. Which stream of Degree/UG has a larger proportion of students not placed?

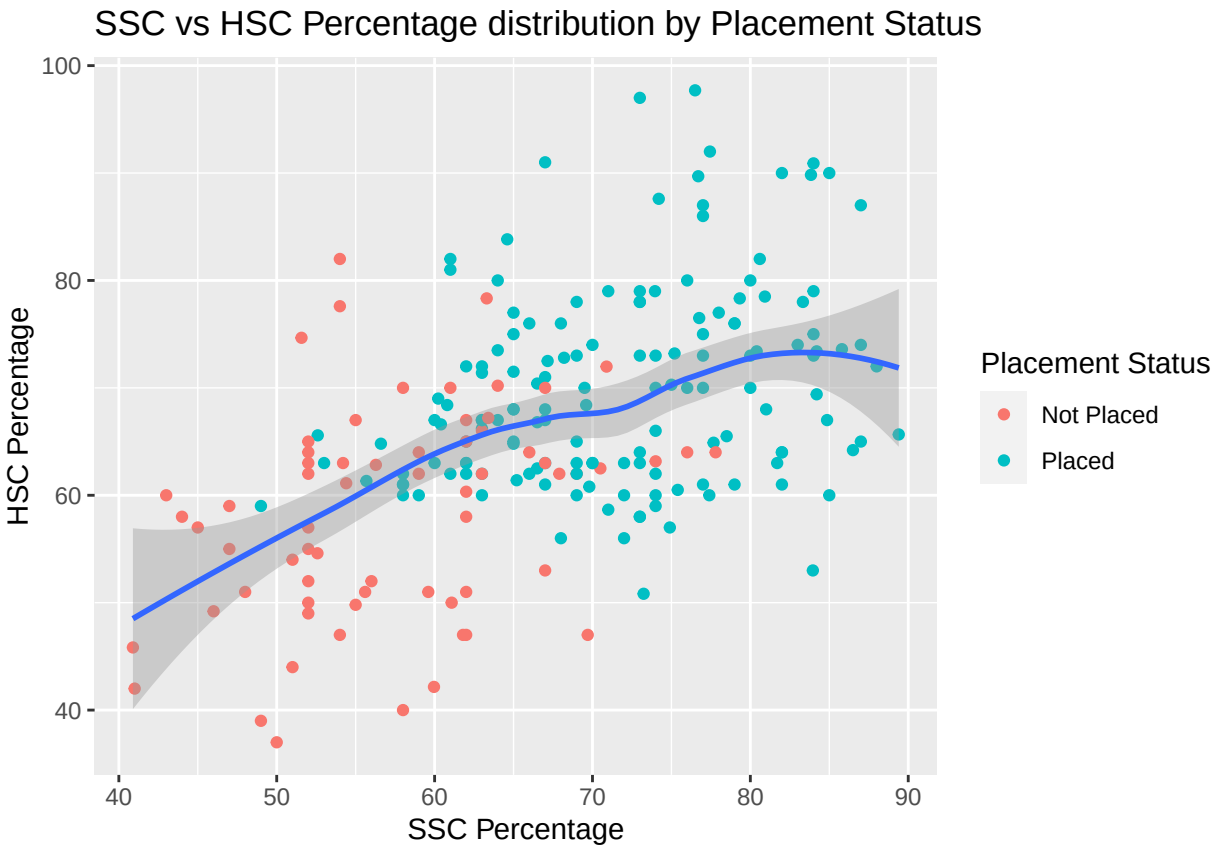
The students in Other streams of Degree/UG apart from Comm&Mgmt and Sci&Tech have faced a larger proportion of unemployment after completing their PG studies. While the proportion of students placed from the streams Comm&Mgmt and Sci&Tech is the same at around 75%.



B.3. Do higher percentages in SSC and HSC help in getting placement offers?

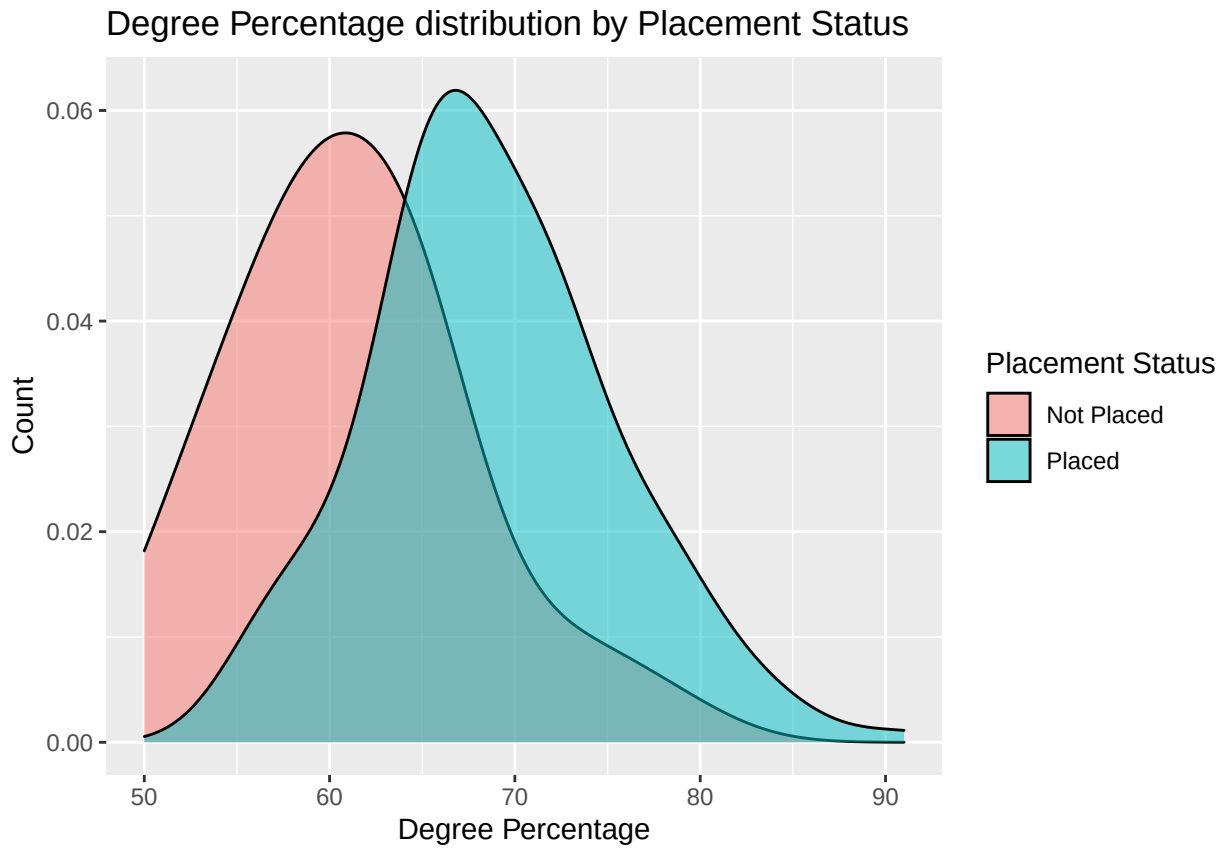
As we can observe from the scatter plot below, the students with lower percentages are not placed whereas, the ones with higher scores in SSC and HSC have gotten placed.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



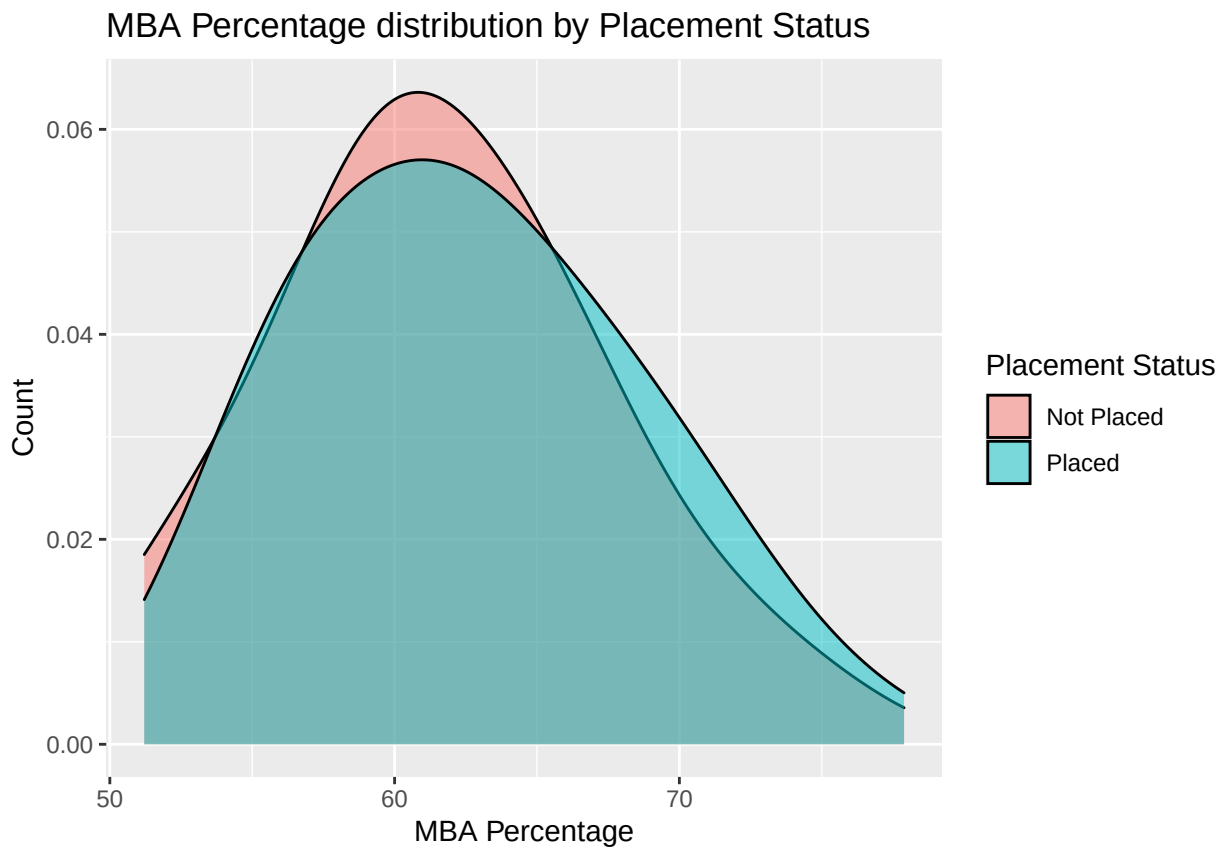
B.4. Does the percentage obtained in Degree impact the placement status?

As we can observe from the below plot, the peak for the students placed is distinctly farther from that of students not placed. This indicates that the Degree percentage was a major deciding factor for the placement of the students.



B.5. Does the percentage obtained in MBA impact the placement status?

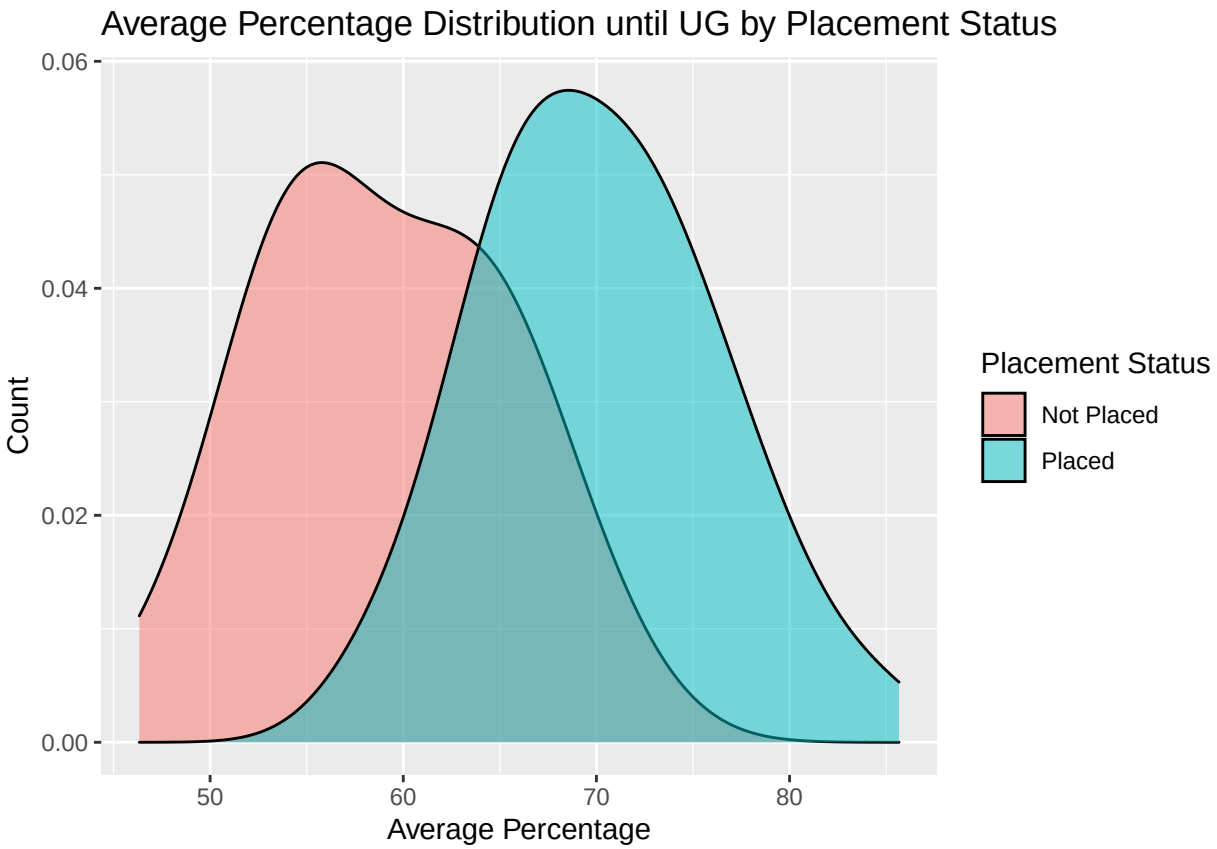
As we can observe from the below plot, the peak for the students placed is the same as that of students not placed. This indicates that the MBA percentage was NOT a major deciding factor for the placement of the students.



B.6. From the above points, we can see that the SSC, HSC and Degree percentages are impacting the placements. Let us see if the average value has the same behavior as each individual entity.

The below graph proves that the behavior of the average is the same as previously mentioned individual entities like SSC, HSC and Degree percentage. Higher the average percentage, better are the placement records.

```
## # A tibble: 215 x 17
##    sl_no gender ssc_p ssc_b   hsc_p hsc_b   hsc_s degree_p degree_t workex etest_p
##    <dbl> <chr>  <dbl> <chr>   <dbl> <chr>   <chr>    <dbl> <chr>    <chr>    <dbl>
##  1     1 M       67   Others   91   Others  Comm~     58   Sci&Tech No        55
##  2     2 M       79.3 Central  78.3 Others  Scie~     77.5 Sci&Tech Yes       86.5
##  3     3 M       65   Central  68   Central Arts      64   Comm&Mg~ No        75
##  4     4 M       56   Central  52   Central Scie~     52   Sci&Tech No        66
##  5     5 M       85.8 Central  73.6 Central Comm~     73.3 Comm&Mg~ No        96.8
##  6     6 M       55   Others   49.8 Others  Scie~     67.2 Sci&Tech Yes       55
##  7     7 F       46   Others   49.2 Others  Comm~     79   Comm&Mg~ No        74.3
##  8     8 M       82   Central  64   Central Scie~     66   Sci&Tech Yes       67
##  9     9 M       73   Central  79   Central Comm~     72   Comm&Mg~ No        91.3
## 10    10 M       58   Central  70   Central Comm~     61   Comm&Mg~ No        54
## # ... with 205 more rows, and 6 more variables: specialisation <chr>,
## #   mba_p <dbl>, status <chr>, salary <dbl>, status_int <dbl>, avg_ug <dbl>
```

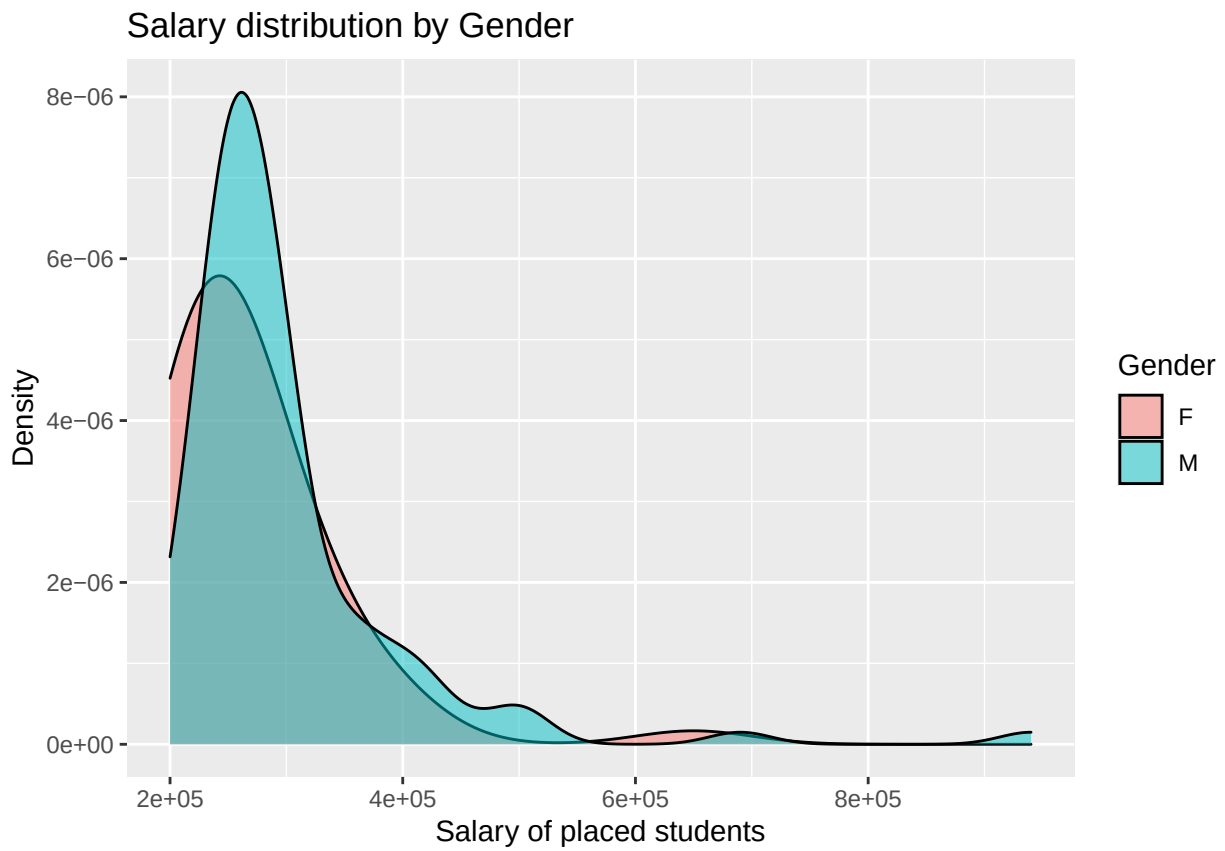


B.7. Does having work experience boost one's chances of getting placed?

Yes, having work experience has bettered the chances of students to get job offers. As we can see from the 100% stacked bar graph, the proportion of students placed is higher for those having work experience.

B.8. Is there gender bias in the pay for students who are placed?

From the graph below, we can observe that the pay for Female students is slightly lesser than that of the Male counterparts, as the peaks are at a lower Salary value for the females.



C. STATISTICAL MODEL: LINEAR REGRESSION

The Equation for the LINEAR REGRESSION between the target or predicted variable (Y) i.e. "status_int" vs the explanatory variable (X) i.e. "avg_ug" is as follows:

(predicted mean value of Y) = intercept + (slope)(mean value of X)

Let's call the lm method to generate a linear regression model that gives the intercept and slope values as coefficients:

```
##
## Call:
## lm(formula = status_int ~ avg_ug, data = placement)
##
## Coefficients:
## (Intercept)       avg_ug
##    -0.84761      0.03804
```

Hence, our equation now becomes:

status_int = -0.84761 + (0.03804)*avg_ug

```
##
```

```
## Call:
## lm(formula = status_int ~ avg_ug, data = placement)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.88647 -0.22219  0.05812  0.28337  0.65406
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.847611   0.204613  -4.143 4.95e-05 ***
## avg_ug       0.038038   0.003048  12.481  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3536 on 213 degrees of freedom
## Multiple R-squared:  0.4224, Adjusted R-squared:  0.4197
## F-statistic: 155.8 on 1 and 213 DF,  p-value: < 2.2e-16
```

The above summary function when called gives the residuals' IQR range, coefficient values for intercept and slope, and the Coefficient of Determination value i.e. R-square is equal to 0.4224.

In the column that was created "status_int", the value **'not placed' is 1 and 'placed' is 2.**

- The slope can be interpreted as follows:

  For every increase in unit of average percentage (i.e. avg_ug), it is expected that the status_int would increase on an average by 0.092209 units. Thus increasing the chances of getting placed.

- The intercept can be interpreted as follows:

  When the average percentage scored would be zero, it is expected on average that value of status_int is -0.84761 which can never be possible since percentage values are nowhere zero.

D. CONCLUSION:

The analysis above can be concluded as follows:

- Many factors were taken into consideration to determine what impacts the placement of students in a class who are graduating with an MBA.
- The Employability test scores is NOT the only determining factor for the placement of students.
- The students in Other streams of Degree/UG apart from Comm&Mgmt and Sci&Tech have faced a larger proportion of unemployment.
- While 75% of students from the streams Comm&Mgmt and Sci&Tech are placed.
- The students with higher percentages in SSC and HSC had better placement numbers.
- The Degree percentage was also a major deciding factor for the placement of the students.
- Whereas, their MBA percentage was NOT a major deciding factor for the placement of the students.

- The behavior of the average of SSC, HSC and Degree percentage is the same as the previously mentioned individual entities SSC, HSC and Degree percentage. Higher the average percentage, better are the placement records.
- Having work experience has bettered the chances of students to get job offers.
- Gender bias in pay is still a controversial topic in today's times. The pay for Female students is slightly lesser than that of the Male counterparts.
- On the whole, students with good average percentage until their UG, who are from either Comm&Mgmt or Sci&Tech Degree background, and have had prior work experience have had better placement records.

–**Thank You** –