INL1 Analytics Portfolio

Paul Gaillard 23001
20030704-T194

In this portfolio, I will show you my knowledge of applied analysis, sharing with you my own definition of analysis and the areas in which it can be useful, then I will take up Dahlin's diamond model view of applied analysis and explain in my own words what it is. In this portfolio, I wanted to show my skills in applying the knowledge I have learnt in my courses, but also in applying the many hours I have spent on the internet to broaden the spectrum of my knowledge of R code. All this by taking a dataset of different academic results of Indian students to get valuable information. I want to show the analytical path I undertake to arrive at interesting and sometimes non-intuitive information from a dataset, in particular I wanted to understand whether some of the student outcomes could have a schematic link to their future work placement status or not through data clustering, and in a second example, I ran a logistic regression model to understand what the determinants of that same work placement were, how those determinants affected the probabilities of being placed in a job, and then I created a predictive model from those determinants to know the hypothetical placement of new observations.

Course name: Business Analytics *Course code:* FOA197
Submission date: January 12th, 2024

Analysis is a detailed and meticulous study whose aim is to identify the elements that make up a whole in order to make it more comprehensible and enlightened. It is often used to explain complex systems by breaking them down and establishing the links between its elements. Analysis often involves reconstituting the elements of the whole through what we might call synthesis.
Analysis is applied in a variety of fields, each adapting it to its own needs. For example, it can be used in politics, where we seek to assess the consequences of decisions on society or the economy, or in project management, finance or security, to identify and mitigate risks, It is also essential in data science, where we seek to interpret data sets using techniques such as clustering or regression. In short, there's no single way to use analysis.
But is there a common structure or model on which all these types of analysis can be based?
Well, that's precisely what Dr. Peter Dahlin proposes with his diamond model of applied analysis. Peter Dahlin with his diamond model of applied analysis, applied analysis enables us to create a pathway to decision-making, decision-making in business being all too often influenced by, or even solely made up of, intuitions and habits that managers carry with them, and if we look at this aspect, we can well admit that intuition is an elementary strength for a person having to deal with multitudes of situations, which are never the same, and which often require a reaction within a limited timeframe, but we can't deny that certain decisions can't be based on intuition alone, and that these intuitions are sometimes biased by information that escapes the intuitive sides that humans have, which can have excessively negative consequences. Good decision-making most certainly lies in a balance between the two, which is why it must be complemented by solid, factual elements on which to rely, and this is where applied analysis comes in.
The diamond model of applied analysis offers us a structure, so that each element of the analysis is linked and the whole creates meaning, starting with a question that will serve as a guiding thread throughout the creation of this analysis, it represents the reason and objective of the analysis.
The question must be as precise as possible, as it is with this that the results are guided and will define the rest of the analysis - it is therefore very important.
A good way to define whether the question is good or not is to ask yourself: "What is the objective of answering this question?" and see if it corresponds to the problems for which we want to provide a result.
Then we want clear ideas, concepts, which are paths of reasoning that help to orientate the analysis, give it a clear and defined perspective, they specify which aspects must be considered, which aspects are relevant and which data must be used for the analysis, to give it a certain direction.
Analytics is the use of tools and techniques guided by the concepts defined earlier, to create meaning from data sets, this use of data will create a result that should answer the question thanks to the concepts. It's the use of data to create value and answers.
The results part of the analysis presents the results, and answers the question defined at the outset, responding to the problem by adding value through recommendations, reports, graphs, or even a simplified visualization of the result. But the results must not differ from the initial objective, i.e. the question. It's important to stay within the framework defined above.
The analysis approach applied based on D Peter. Dahlin's Diamond Model allows us to be guided through an analytical process that will support decision-making in any field where analysis can be applied

Example 1

The following analysis aims to explore the diversity of academic performance of Indian students at ssc and hsc level by grouping similar observations using the k-means clustering algorithm. The aim of this clustering of students by combining the scores of the first two tests (SSC: Secondary School Leaving Certificate Examination, and HSC: Higher Secondary School Leaving Certificate Examination) they took during their course of study is to observe, with the help of visual graphs, whether the groups that stand out were subsequently placed in a job or not.

What are the different groups (clusters) of students that emerge from the clustering analysis? And is there a pattern between the clusters identified and the students' placement status?

#Step 1: Data preparation.

The "Student_data" dataset was loaded from a CSV file, containing information on the grades of Indian students at different levels of their academic career.
Two commands were written to understand the data structure of the dataset:

```
str(Student_data)
summary(Student_data)
```

These commands have been used to identify the variables that can be analyzed as part of a clustering process.

Selection of two variables for analysis:

SSC: secondary school leaving examination
HSC: Higher Secondary School Leaving Examination

```
selected_data <- Student_data[, c("ssc_p", "hsc_p")]
```

#Step 2: Data exploring

An exploratory analysis was conducted to understand general trends in student scores by comparing them to their placement status.

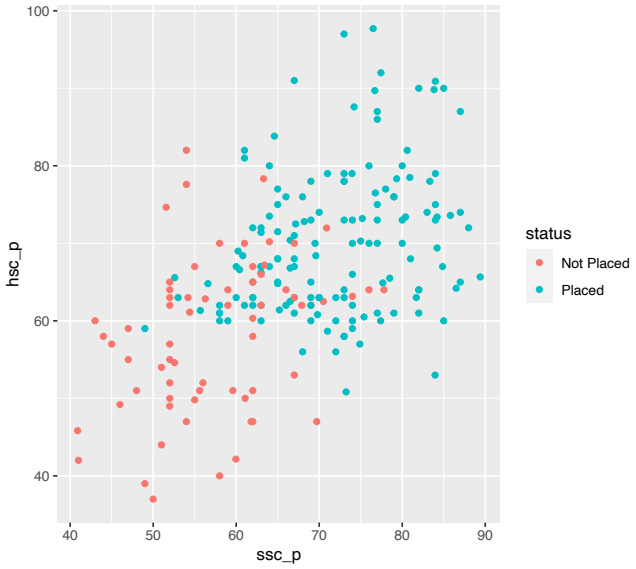
- Descriptive summary of means by status:

```
Student_data %>%
  group_by(status) %>%
  summarise(mean_ssc_p = mean(ssc_p, na.rm = TRUE),
            mean_hsc_p = mean(hsc_p, na.rm = TRUE))
```

	status	mean_ssc_p	mean_hsc_p
	<chr>	<dbl>	<dbl>
1	Not Placed	57.5	58.4
2	Placed	71.7	69.9

- Scatterplot to visualize the distribution of student scores based on their status:

```
ggplot(Student_data, aes(x=ssc_p, y=hsc_p, color=status)) +
  geom_point()
```



#Step 3: K-means Clustering

The k-means clustering algorithm was applied to group students into clusters based on their academic performance.

- Data standardisation:

```
str(selected_data)
Scaled_data <- scale(selected_data)
```

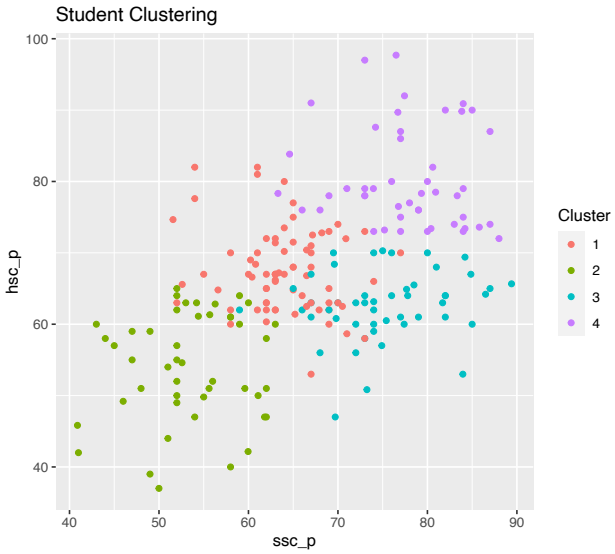
- Application of the K-means algorithm:

```
set.seed(123)
kmeans_result <- kmeans(Scaled_data, centers = 4)
kmeans_result
selected_data$cluster <- kmeans_result$cluster
```

The cluster number associated with each observation has been added to the "selected_data" dataset.

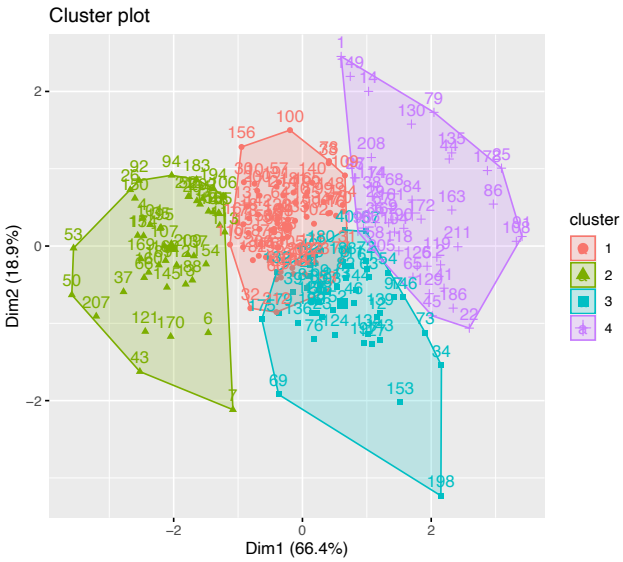
- Scatter plot:

```
ggplot(selected_data, aes(x = ssc_p, y = hsc_p, color = factor(cluster))) +
  geom_point() +
  labs(title = "Student Clustering",
       x = "ssc_p",
       y = "hsc_p",
       color = "Cluster")
```



- Visualization of clustering algorithm results:

```
fviz_cluster(kmeans_result, data = Scaled_data)
```



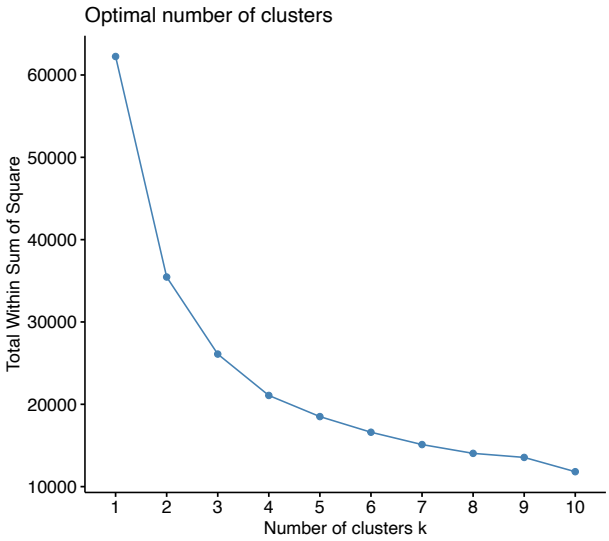
We can see that the clusters overlap.

#Step 4: Optimal selection of k.

Use the Elbow method to determine the optimal number of clusters.

- Elbow Method:

```
fviz_nbclust(selected_data2, kmeans, method = "wss")
```

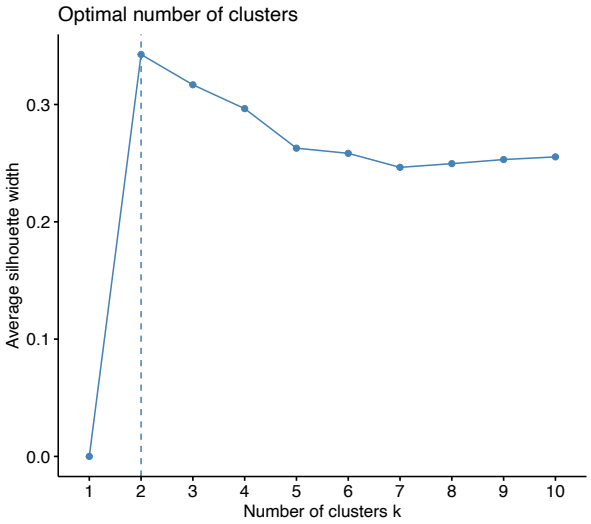


The elbow seems to be in 3, but the graph is still rather confusing.

Using a secondary method to determine the optimal number of clusters

- Silhouette Method :

```
fviz_nbclust(selected_data2, kmeans, method = "silhouette")
```



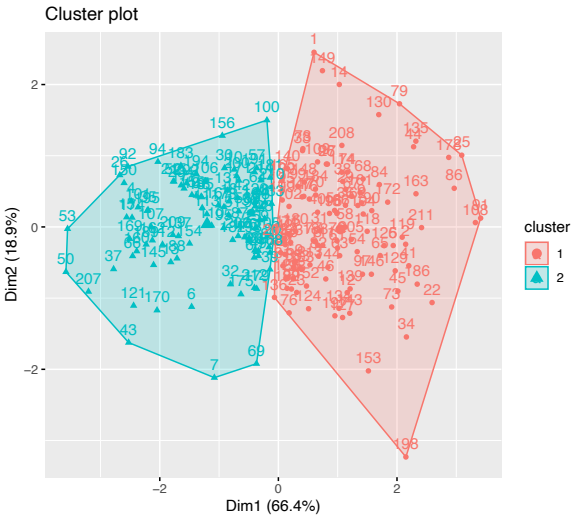
The "silhouette" method indicates that the optimum number of clusters is 2.

As the ideal number of clusters remains quite unclear, the rest of the analysis will use both 2 and 3 clusters in the visual graphs. This will enable us to better understand and integrate the data, particularly through comparison.

#Step 5: Cluster visualization.

- Model visualization using 2 clusters:

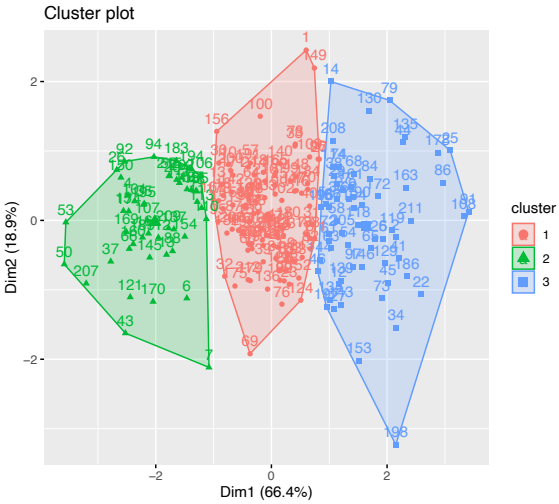
```
set.seed(123)
kmeans_result_2c <- kmeans(Scaled_data, centers = 2)
kmeans_result_2c
selected_data$cluster <- kmeans_result_2c$cluster
fviz_cluster(kmeans_result_2c, data = Scaled_data)
```



- Model visualization using 3 clusters:

```
set.seed(123)
kmeans_result_3c <- kmeans(Scaled_data, centers = 3)
kmeans_result_3c
selected_data2$cluster <- kmeans_result_3c$cluster

fviz_cluster(kmeans_result_3c, data = Scaled_data)
```



- Adding results to the original dataset:

```
Student_data$cluster_model2c <- kmeans_result_2c$cluster
Student_data$cluster_model3c <- kmeans_result_3c$cluster
```

#Step 6: Analysis of potential link with placement status.

To assess the potential relationship between the identified clusters and students' placement status, a comparison was made using cross-tabulations and visual graphs.

Hypothesis: We hypothesized that there might be some semblance of a link between Cluster 3, assumed to contain students with the highest grades, and "Placed" status. The aim was to determine whether students in clusters combining higher grades are more frequently placed than those with lower grades.

- Comparison of status by cluster with (3-cluster model):

Using cross-tabulations to examine the distribution of placement status across the 3 clusters

```
table(Student_data$status, Student_data$cluster_model3c)
```

```
            1  2  3
Not Placed 26 39  2
Placed     80  8 60
```

- Comparison of status by cluster (2-cluster model):

Using cross-tabulation to examine the distribution of placement status across the 2 clusters

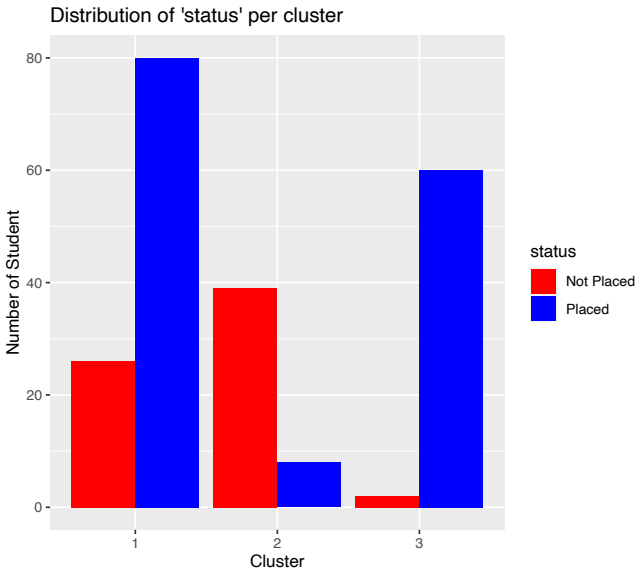
```
table(Student_data$status, Student_data$cluster_model2c)
```

```
              1   2
Not Placed    9  58
Placed      106  42
```

- Visualization of results (3-cluster model):

Creation of bar graphs to visualize the distribution of "Placed" and "Not Placed" status by cluster.

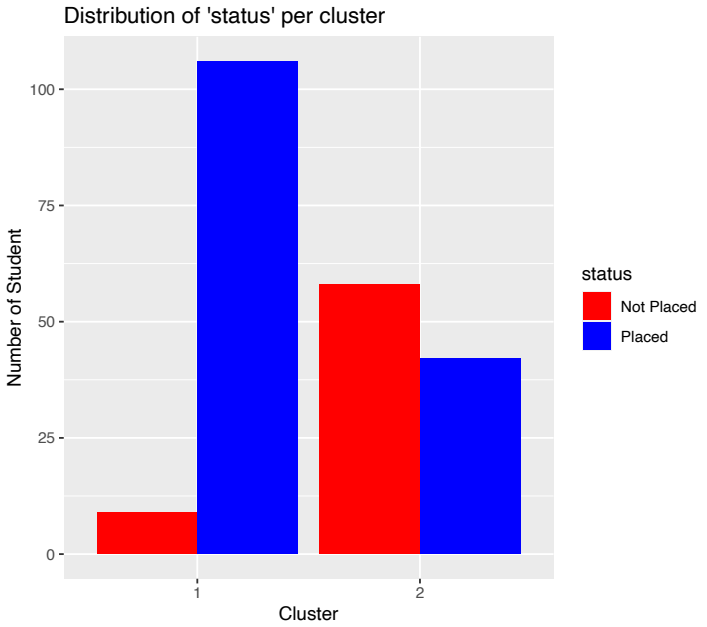
```
ggplot(Student_data, aes(x = factor(cluster_model3c), fill = status)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of 'status' per cluster",
       x = "Cluster",
       y = "Number of Student") +
  scale_fill_manual(values = c("Placed" = "blue", "Not Placed" = "red"))
```



- Visualization of results (2 cluster model):

Creation of bar graphs to visualize the distribution of "Placed" and "Not Placed" status by cluster.

```
ggplot(Student_data, aes(x = factor(cluster_model2c), fill = status)) +
  geom_bar(position = "dodge") +
  labs(title = "Distribution of 'status' per cluster",
       x = "Cluster",
       y = "Number of Student") +
  scale_fill_manual(values = c("Placed" = "blue", "Not Placed" = "red"))
```



The results do not support the hypothesis of a link between clusters assumed to have a combination of the best scores and "Placed" status.
Students in clusters where the combination of “ssc” and “hsc” scores is the lowest also have a high rate of "Placed" status, even higher than those in clusters where the combination of scores is the best.
This suggests that other factors, beyond high school graduation (SSC) and high school graduation (HSC) grades alone, influence students' placement status.

Reflection on the analysis of example 1:

The clustering analysis undertaken aimed to identify distinct groups (clusters) among students based on their academic performance. The process involved several steps, including data preparation, exploration, the application of the k-means clustering algorithm and the visualization between scores and status. The goal was to uncover patterns in the students' academic achievements and assess if these patterns seemed to be linked to their placement status post-graduation.

The utilization of the elbow method and silhouette method to determine the optimal number of clusters demonstrated a thoughtful approach to the analysis. The aim was to be able to determine the optimal number of clusters to implement in the clustering model, but also to be able to solve the cluster overlap problem.

The visual representation of the clusters in scatter diagrams using the fviz_cluster function provided a clear understanding of how students were grouped according to their academic performance. Secondly, the visualization of students by score combination group based on their status (Placed or Not placed) allowed us to observe whether there appeared to be a higher number of students placed in clusters with a better score combination, thus improving the overall interpretability of the results.

Testing the hypothesis concerning the relationship between the identified groups and placement status was an important aspect of the analysis. However, the results were not consistent with the hypothesis that the group with the best academic results would necessarily have a higher placement rate. This discrepancy led to a valuable insight: ssc and hsc grades may not be the only determinant of placement after graduation.
An examination of the placement status of the various groups, presented in tabular and visual form, facilitated a nuanced interpretation of the results.

In conclusion, while the analysis has led to a better understanding of student groupings based on academic performance, it has also highlighted the complexity of factors influencing post-graduation placement. The exploration of clustering, coupled with the assessment of placement outcomes, contributes to a better understanding of the dynamics within student groups. Moving forward, it is essential to recognize the limitations of the analysis, such as dependence on a specific data set, and to consider other variables that may influence placement status. This thorough reflection underlines the importance of a thoughtful approach to data analysis and offers valuable insights for future research into the reasons significantly influencing student status.

Example 2:

The logistic regression analysis presented aims to explore and understand the determinants of Indian students' work placement, based on academic indicators and other attributes, and then to be able to create a model to predict work placement status based on these same determinants.

What are the most significant variables in determining student placement status? Do these determinants allow us to predict the placement status of future observations?

In the following example, code created using code generation tools has been marked in blue

#Step 1 : Data preparation

The "Student_data" dataset was loaded from a CSV file, containing information on the grades of Indian students at different levels of their academic career and their placement outcomes

- Two commands were written to understand the data structure of the dataset:

```
View(Student_data)
summary(Student_data)
str(Student_data)
```

```
'data.frame':	215 obs. of  15 variables:
 $ sl_no         : int  1 2 3 4 5 6 7 8 9 10 ...
 $ gender        : chr  "M" "M" "M" "M" ...
 $ ssc_p         : num  67 79.3 65 56 85.8 ...
 $ ssc_b         : chr  "Others" "Central" "Central" "Central" ...
 $ hsc_p         : num  91 78.3 68 52 73.6 ...
 $ hsc_b         : chr  "Others" "Others" "Central" "Central" ...
 $ hsc_s         : chr  "Commerce" "Science" "Arts" "Science" ...
 $ degree_p      : num  58 77.5 64 52 73.3 ...
 $ degree_t      : chr  "Sci&Tech" "Sci&Tech" "Comm&Mgmt" "Others" ...
 $ workex        : chr  "No" "Yes" "No" "No" ...
 $ etest_p       : num  55 86.5 75 66 96.8 ...
 $ specialisation: chr  "Mkt&HR" "Mkt&Fin" "Mkt&Fin" "Mkt&HR" ...
 $ mba_p         : num  58.8 66.3 57.8 59.4 55.5 ...
 $ status        : chr  "Placed" "Placed" "Placed" "Not Placed" ...
 $ salary        : int  270000 200000 250000 NA 425000 NA NA 252000 231000 NA ...
```

The variables "hsc_s" and "degree_t" will not be included in the creation of the model, as their elements could create collinearity problems in the logistic regression model. In addition, given their context, these variables are very unlikely to be influential in determining student status, and we will also remove the variables "sl_no" and "salary" which, given their context, are not determinant in the outcome of student status either.

- Selection of variables for analysis:

```
selected_variables1 <- Student_data[, c(2, 3, 4, 5, 6, 8, 10, 11, 12, 13, 14)]
```

#Step 2: Data transformation

The selected variables are converted into factual variables for use in the logistic regression model.

- Transformation of qualitative variables into binary variables:

```
Student_data$gender_binary <- ifelse(selected_variables1$gender == "M", 1,0) #Var2
Student_data$status_binary <- ifelse(selected_variables1$status == "Placed", 1, 0) #Var14
Student_data$ssc_b_binary <- ifelse(selected_variables1$ssc_b == "Central", 1, 0) #Var4
Student_data$hsc_b_binary <- ifelse(selected_variables1$hsc_b == "Central", 1, 0) #Var6
Student_data$workex_binary <- ifelse(selected_variables1$workex == "Yes", 1, 0) #Var10
Student_data$specialisation_binary <- ifelse(selected_variables1$specialisation == "Mkt&HR", 1, 0) #Var12
```

- Create a new dataset of usable variables:

```
Linear_regression_Student_data <- Student_data[, c(16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26)]
```

#Step 3: Logistic regression model.

Creation of a first logistic regression model to determine the statistical significance of each variable.

- Establish the response variable (dependent variable):

```
response_variable <- "status_binary"
```

- Creating the logistic regression model:

```
logistic_model <- glm(paste(response_variable, "~."), data = Linear_regression_Student_data, family = "binomial")
```

- Results of the logistic regression model:

```
summary(logistic_model)
```

```
Call:
glm(formula = paste(response_variable, "~."), family = "binomial",
    data = Linear_regression_Student_data)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          -17.18854    4.95997  -3.465 0.000529 ***
gender_binary          0.87000    0.61421   1.416 0.156645
ssc_p_verified         0.20927    0.04287   4.882 1.05e-06 ***
ssc_b_binary          -0.38514    0.64680  -0.595 0.551538
hsc_p_verified         0.11696    0.03657   3.198 0.001383 **
hsc_b_binary          -0.11875    0.63930  -0.186 0.852635
degree_p_verified      0.15852    0.05145   3.081 0.002065 **
workex_binary          2.05269    0.65924   3.114 0.001847 **
etest_p_verified      -0.01169    0.02123  -0.551 0.581766
specialisation_binary -0.40480    0.54187  -0.747 0.455036
mba_p_verified        -0.21179    0.05392  -3.928 8.58e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 266.77  on 214  degrees of freedom
Residual deviance: 106.94  on 204  degrees of freedom
AIC: 128.94

Number of Fisher Scoring iterations: 7
```

The model summary focuses on the individual significance of the coefficients. Variables with stars indicate statistical significance; the more stars, the greater the certainty of the variable's effect on the dependent variable.
In this analysis, we consider that a p-value > 0.05 is not statistically significant.

#Step 4: Significance of variables and interpretation.

Interpretation of the model summary, overall significance of the variables in the model using the Wald test and interpretation of the Odds ratios.

The model summary indicated that the variables "gender_binary," "ssc_b_binary," , "hsc_b_binary" , "etest_binary" and "specialisation_binary" do not appear to be significant, i.e. student gender, type of examination board location at secondary and upper secondary level, employability test and specialisation at upper cycle university study level (Master), do not appear to be significant in determining student status.

To complete the analysis of variable significance in the model summary, a Wald test is performed.
The Wald test evaluates the overall significance of each variable in the model. This is done by comparing the deviance of the full model (including all variables) with the deviance of a reduced model (without the variable in question).

- Wald test:

```
wald_test <- anova(logistic_model, test = "Chisq")
print(wald_test)
```

```
> print(wald_test)
Analysis of Deviance Table

Model: binomial, link: logit

Response: status_binary

Terms added sequentially (first to last)

                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                  214     266.77
gender_binary        1    1.746       213     265.02  0.186370
ssc_p_verified       1  103.144       212     161.88 < 2.2e-16 ***
ssc_b_binary         1    0.275       211     161.60  0.599838
hsc_p_verified       1   18.447       210     143.16 1.747e-05 ***
hsc_b_binary         1    0.052       209     143.11  0.819448
degree_p_verified    1    8.730       208     134.38  0.003129 **
workex_binary        1    8.263       207     126.11  0.004046 **
etest_p_verified     1    0.608       206     125.50  0.435683
specialisation_binary 1   0.631       205     124.88  0.427056
mba_p_verified       1   17.939       204     106.94 2.281e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Differences between the p-values of the model summary and those of the Wald test may arise due to differences in test methodology, but both approaches should provide consistent and similar information regarding the significance of variables.

The results of the Wald test confirm the observations made in the model summary, they confirm the statistical significance of the important variables identified in the model summary, by providing an overall assessment of the contribution of each variable to the model fit, this information complements the robustness of our results and is useful for a thorough interpretation of the logistic regression model.

- Odds ratios of the logistic regression model:

```
odds_ratios <- exp(coef(logistic_model))
print(odds_ratios)
```

```
> print(odds_ratios)
         (Intercept)        gender_binary       ssc_p_verified         ssc_b_binary       hsc_p_verified
        3.428562e-08         2.386912e+00         1.232778e+00         6.803542e-01         1.124074e+00
        hsc_b_binary    degree_p_verified         workex_binary    etest_p_verified specialisation_binary
        8.880257e-01         1.171771e+00         7.788794e+00         9.883757e-01         6.671076e-01
      mba_p_verified
        8.091366e-01
```

- Interpretation of odds ratios for significant variables:

"ssc_p_verified" (Odds ratio ≈1.23)
For every one-unit increase in SSC score, success ratings increase by around 23%. Students with higher SSC scores therefore have around 23% more chance of achieving the "Placed" outcome of the dependent variable compared with those with lower scores.

"hsc_p_verified" (Odds ratio ≈1.12)
For every one-unit increase in the HSC score, success ratings increase by around 12%. This suggests that students with higher HSC grades have around 12% more chance of achieving the "Placed" outcome of the dependent variable compared with those with lower grades.

"degree_p_verified" (Odds ratio ≈1.17)
For every one-unit increase in degree grade, success scores increase by around 17%. This suggests that students with higher diploma grades have around 17% more chance of achieving the "Placed" outcome of the dependent variable compared with those with lower grades.

"worked_binary" (Odds ratio ≈7.79)
Students with work experience are approximately 7.79 times more likely to achieve the "Placed" outcome of the dependent variable compared to those without work experience.

"mba_p_verified" (Odds ratio ≈0.81)
Students who have validated an MBA are around 19% less likely to achieve the "Placed" outcome of the dependent variable compared with those who have not validated an MBA.

The negative relationship between MBA score and employment probability is surprising and may seem counter intuitive. There may be several explanations for this, such as the presence of variables not included in the model, general trends among students with high scores that would make them less inclined to actively seek employment immediately after graduation, perhaps preferring to pursue further studies, start their own business or other, but also possibly labor market conditions specific to the data period, or complex interactions with other variables in the model.

A thorough analysis of the specific context of the data needs to be considered to better understand this surprising relationship.

#Step 5: Model simplification

Creation of a second simplified model with the variables considered significant

The variables "ssc_p_verified", "hsc_p_verified", "degree_p_verified", "workex_binary" and "mba_p_verified" were saved in the simplified model.
The variables "gender_binary", "ssc_b_binary", "hsc_b_binary", "etest_p_verified", "specialization_binary" were removed to simplify the complexity of the model, as their p-value was greater than 0.05.

- Selection of significant variables:

```
Linear_regression_Student_data2 <- Linear_regression_Student_data[, c(2, 3, 5, 7, 8, 11)]
```

- Creating the simplified model:

```
logistic_model2 <- glm(paste(response_variable, "~."), data = Linear_regression_Student_data2, family = "binomial")
```

- Displaying the results of the simplified model:

```
summary(logistic_model2)
```

```
> summary(logistic_model2)

Call:
glm(formula = paste(response_variable, "~."), family = "binomial",
    data = Linear_regression_Student_data2)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -15.48852    3.96106  -3.910 9.22e-05 ***
ssc_p_verified     0.19214    0.03761   5.109 3.25e-07 ***
hsc_p_verified     0.11948    0.03499   3.415 0.000637 ***
degree_p_verified  0.14975    0.04769   3.140 0.001690 **
workex_binary      2.26667    0.64652   3.506 0.000455 ***
mba_p_verified    -0.22805    0.04985  -4.575 4.76e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 266.77  on 214  degrees of freedom
Residual deviance: 112.47  on 209  degrees of freedom
AIC: 124.47

Number of Fisher Scoring iterations: 7
```

The summary of the simplified model shows that all the variables are significant, with a p-value of "Pr(>|z|)" < 0.05.

#Step 6: Interpretation of the simplified model

Interpretation of the overall significance of the variables in the model using the Wald test, interpretation of the Odds ratios.

- Wald test :

```
wald_test2 <- anova(logistic_model2, test = "Chisq")
```

```
print(wald_test2)
```

```
> print(wald_test2)
Analysis of Deviance Table

Model: binomial, link: logit

Response: status_binary

Terms added sequentially (first to last)


                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                                214     266.77
ssc_p_verified     1   96.497       213     170.27 < 2.2e-16 ***
hsc_p_verified     1   15.954       212     154.32 6.491e-05 ***
degree_p_verified  1    5.163       211     149.16  0.023079 *
workex_binary      1   10.677       210     138.48  0.001085 **
mba_p_verified     1   26.010       209     112.47 3.397e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The results of the Wald test on the simplified model provide an overall assessment of the contribution of each variable to the fit of the model.

- Odds ratios:

```
odds_ratios2 <- exp(coef(logistic_model2))
print(odds_ratios2)
```

```
> print(odds_ratios2)
      (Intercept)    ssc_p_verified    hsc_p_verified degree_p_verified     workex_binary    mba_p_verified
     1.876814e-07      1.211846e+00      1.126915e+00      1.161544e+00      9.647202e+00      7.960811e-01
```

When comparing the full model and the simplified model, it is important to note that slight variations can be observed in the Odds ratios. These differences may be linked to correlations between variables, control effects in the full model and interactions not included in the simplified model. However, it is important to point out that despite minor variations, the general direction of the effects remains the same and is completely consistent. These variations are not significant and are therefore considered acceptable for the rest of the analysis.

#Step 7: Confusion Matrix for Simplified Model

Creation of a confusion matrix with the model considered to be the most balanced.

Previously two models were run, the first included all variables, some of which indicated low statistical significance, however the second model was simplified to include only statistically significant variables, and presents in its summary a lower AIC (Akaike Information Criterion) (AIC = 124 .47) compared to the first model (AIC = 128.94). This indicates that the second model has a better fit of the data to the model while minimising complexity. In other words, the second model is preferred for the rest of the analysis because of its better balance between fit and parsimony.

The confusion matrix will allow us to evaluate our model.

- Creating predictions on model 2 (Simplified):

```
predictions2 <- predict(logistic_model2, newdata =
Linear_regression_Student_data2, type = "response")
predicted_classes2 <- ifelse(predictions2 > 0.5, 1, 0)
```

- Creating a confusion matrix :

```
conf_matrix2 <- confusionMatrix(as.factor(predicted_classes2),
as.factor(Linear_regression_Student_data2$status_binary))
```

- Displaying the confusion matrix :

```
print(conf_matrix2)
```

```
> print(conf_matrix2)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  50  10
         1  17 138

               Accuracy : 0.8744
                 95% CI : (0.8226, 0.9156)
    No Information Rate : 0.6884
    P-Value [Acc > NIR] : 1.58e-10

                  Kappa : 0.6987

 Mcnemar's Test P-Value : 0.2482

            Sensitivity : 0.7463
            Specificity : 0.9324
         Pos Pred Value : 0.8333
         Neg Pred Value : 0.8903
             Prevalence : 0.3116
         Detection Rate : 0.2326
   Detection Prevalence : 0.2791
      Balanced Accuracy : 0.8394

       'Positive' Class : 0
```

The confusion matrix gives us these results:
True negative = 50
False negative = 10
False positive = 17
True positive = 138

For example, the confusion matrix tells us that the predictive model correctly predicted that the student would be Placed 138 times, correctly predicted that the student would be Not Placed 50 times, incorrectly predicted that the student would be Placed 17 times, and incorrectly predicted that the student would be Not Placed 10 times.

The performance metric shows a total proportion of correct predictions (Accuracy) of 87.44%, for a sensitivity of 74.63%, i.e. the proportion of students with a correctly identified 'Not Placed' status, and a specificity of 93.24%, i.e. the proportion of students with a correctly identified 'Placed' status.

#Step 8: Prediction model

Use the simplified model to make a prediction about a new observation, i.e. to predict the status of a student created from scratch.

- Creating a new observation:

```
new_observation <- data.frame(
```

```
  ssc_p_verified = 67.30,
  hsc_p_verified = 65.00,
  degree_p_verified = 66.37,
  workex_binary = 0,
  mba_p_verified = 62.28)
```

The new observation simulates a student with SSC 67.3, HSC 65, Degree 66.37, MBA 62.28 and no work experience (workex_binary = 0).

- Prediction with the simplified model:

```
new_prediction <- predict(logistic_model2, newdata = new_observation, type = "response")
```

- Determinate the student's status:

```
NewObs_status <- ifelse(new_prediction > 0.5, "Placed", "Not Placed")
```

Determination of the confidence rate attributed by the model to the prediction of "Placed" status set at 0.5

- Display of the probability of the "Placed" status and display of the status of the new observation:

```
cat("Placement prediction", new_prediction, "\n")
cat("Placement Status :", NewObs_status, "\n")
```

```
> cat("Placement prediction", new_prediction, "\n")
Placement prediction 0.7201021
> cat("Placement Status :", NewObs_status, "\n")
Placement Status : Placed
```

Interpretation: The probability of obtaining "Placed" status for the new observation created from scratch is approximately 72%.

Reflection on the analysis of the example 2:

The logistic regression analysis undertaken aimed to explore and understand in depth the determinants of Indian students' work placement, based on academic indicators and other attributes, and then to be able to create a model that would predict work placement status based on these same determinants. This process was marked by meticulous steps, highlighting the complexity involved in understanding the factors influencing placement outcomes.
The starting point was the preparation of the data, by selecting the relevant variables for the creation of a logistic regression model. The choice of variables was guided by a thorough understanding of their context, as well as the problem.

Setting up the initial model structure required close attention to each variable, to see if their implementation made sense.
The process then turned to identifying the significant variables. Summarizing the model and performing a significance test (Wald test) revealed that certain variables were not relevant for predicting placement status.
The measurement of variable significance continued with the use of odds ratios. This step enabled us to translate the statistical results into more tangible, concrete information. The odds ratios acted like a magnifying glass, offering a detailed view of the impact of each variable on the probability of investment. These steps enabled us to better understand how student placement was defined.

Attention was then focused on the adjustments needed to achieve a more balanced model, leading to the creation of a simplified version. A second model comprising only variables with significance in determining the student's professional status. Model evaluation was also a key element of the analysis, offering valuable insights into the robustness and significance of the simplified model's results.

The creation of a confusion matrix added a practical dimension to model evaluation. This step quantified the accuracy of the model, highlighting true positives, false positives, true negatives and false negatives. This quantitative analysis reinforced confidence in the model's ability to accurately predict student placement status.

The analysis concluded with the application of the model to a new hypothetical observation. This served as a demonstration of the model's ability to generalize its predictions to new cases.

In conclusion, this analysis contributes to understanding the nuanced dynamics of factors influencing post-graduation placement. The iterative process of model refinement and interpretation has unveiled a more meaningful model through comparison of their respective assessments made with AIC, model summary and Wald test interpretations. However, recognition of the limitations must be mentioned, including the model’s accuracy limit, the specificity of the data set, and suggests avenues for future research. Future investigations should delve deeper into unexplored variables and external factors to improve the predictive accuracy of the model and broaden understanding of the determinants of employability among Indian students.