STATISTICAL ANALYSIS OF FACTORS AFFECTING ON PLACEMENT

A Project report submitted in partial fulfillment of the requirements for the degree of M.Sc.(Statistics) with specialization in Industrial Statistics

Submitted by

Ms. Badgujar Shivani Vijay (382752)

Ms. Patil Gayatri Pravin (382782)

Mr. Patil Prajwal Rajendra (382788)

Project Guide

Mr. Manoj C. Patil

in the

Department of Statistics, School of Matehmatical Sciences

Kavayitri Bahinabai Chaudhari North Maharashtra University,

Jalgaon-425001

(Academic Year : 2021-2022)

CERTIFICATE

This is to certify that **Ms. Badgujar Shivani Vijay, Ms. Patil Gayatri Pravin and Mr. Patil Prajwal Rajendra** are the student of **M.Sc. Statistics** (with specialization in Industrial Statistics) at Department of Statistics, School of Mathematical Sciences Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon have successfully completed their project entitled **“Statistical Analysis of factors affecting on placement”** under my guidance and supervision during the academic year 2021-2022.

Mr. M. C. Patil

(Project Guide)

ACKNOWLEDGEMENT

On the completion of this project we must acknowledge from the core of our heart to Dr. R. L. Shinde, Head of the Department of Statistics, School of Mathematical Sciences, Kavayitri Bahinabai Chaudhari North Maharashtra University, Jalgaon for seeking us the desire permission for this project.

We take this opportunity to express our sense of gratitude to our project guide Mr. M. C. Patil for his valuable guidance, immense support, motivation and encouragement to which we could complete our project work successfully.

We have thanks to our parents, friends and classmates to give us moral support. We have also thankful to all for directly or indirectly help for project work.

Ms. Badgujar Shivani Vijay (382752)

Ms. Patil Gayatri Pravin (382782)

Mr. Patil Prajwal Rajendra (382788)

Contents

Chapter 1

Introduction

1.1 Introduction

Nowadays placement plays an important role in this world full of unemployment. Even the ranking and rating of institutes depend upon the amount of average package and amount of placement they are providing. The main aim of every academia enthusiast is placement in a reputed organization or company. So, any system that will predict the placement of the students will be positive and decreases the workload of any institute's training and placement office (TPO). With the help of machine learning techniques, the knowledge can be extracted from past placed students and can be predicted. As we live in a world where tremendous competition is present. In this world, everyone is busy building a secure future by getting a good job in a reputed organization or a company. Every student is taking education and after that, they go for the interview. This project assesses student's perception concerning the level of campus placement activities and determines the order of importance of various factors, as provided by students, relating to the employer's selection criteria. For the student who is studying, they should know on which factors they should focus for getting placed, and which social media they should use for placement. All this we know from the previous data of placed students. So this will help them to identify that factors, and according to that, they will prepare for the interview.

1.2 Motivation

Being a student of M.Sc.II(Statistics) with a specialization in industrial statistics during our M.Sc. we are interested in knowing the factors affecting the placement. There are so many factors affecting the placement that we have to find out the factors which affect the most the placement of the student. For that, we take the primary data as we want to learn how to prepare questionnaire, how to collect primary data and finally to how to handle the primary data, How to face the challenges in primary data this all things we want to learn. So for this purpose we take this topic.

1.3 What is placement?

Placement means getting position in a organization/company it may be private or government. Also, campus placement or campus recruiting is a program conducted within universities or other educational institutions to provide jobs to students nearing completion of their studies. In this type of program, the educational institutions partner with corporations who wish to recruit from the student population.

https://tinyurl.com/y357bmrm

Types of placement:

1. On-Campus: On-Campus placements are those where educational institutes set up programs for placement drives. They partner and invite leading business enterprises to their campus for recruitment.

2. Off-Campus: Off-campus placement is when you get placed at a company without your college being involved in the process. In case of off-campus placements, you need to directly send job applications to the companies that you wish to join.

1.4 Objectives

1. To check if there is any significant impact of SSC, HSC, Degree and Master's percentage on placement.
2. To check if there is any significant impact of gender or age on placement.
3. To check if there is any significant impact of school type and medium of education on placement.
4. To determine characteristics affecting placement.
5. To predict whether the student getting placed or not.
6. To learn how to handle primary data.
7. Play with data conducting statistical tests.
8. To find which social media app is mostly used by students for placement.
9. To check if there is any significant change between the first salary offered and the current salary.
10. Play with primary data and apply statistical tools.

1.5 Data Collection and Description

Data Collection Method

We have used the online method for data collection and sampling design used is snowball sampling. We have collected primary data in which we get a data of 523 respondents was interviewed through a google form. For this, first we created a google form (questionnaire).Then we collect the mobile numberd and email ID's of our college alumni who are placed from the placement cell department only for sending the google form to them. For sending google forms we use excel macros. In this way, we send this google form to 8,484 mobile numbers and 5,460 email addresses. Also we forward this google form to other universities students, other college students, etc . And we send this google form to other friends in our contact, our seniors, juniors etc. through E-mail, telegram, what's app, linkedIn and tell them to forward this google form to their friends in order to get more data we do this. In this way, we get the data of 523 students by using the snowball sampling method.

Data Description

1. The data contains 523 rows and 72 columns. In the data, there are some numerical variables and some are categorical variables.
2. In our data there is academic information on the placed, non-placed, and the students who are studying now.
3. In academics information we collect a percentage of SSC, HSC, Degree, and Masters if done. And also collect some information related to their status placement or not.
4. We collect information about their view regarding to factors affecting to placement.

The primary data which we collect is:
https://github.com/shivani108/Statistical-Analysis-Of-Factors-Affecting-On-\
Placement-Primary-Dataset./blob/main/Primary%20Data.xlsx

1.6 Questionnaire

Study of Factors Affecting Student Placements

- What is your Gender? ○ Male ○ Female
- What is your Age? .
- Your Hometown? ○ Village ○ City ○ Small town
- In which primary/secondary school did you study? ○ Government ○ Private
- What is your SSC percentage? .
- What is your HSC percentage? .
- What is your Educational board? .

	CBSE	ICSE	State Board
10th	○	○	○
12th	○	○	○

- Your Highest education?:
- Degree Percentage: .
- Name of the college you graduated from: .
- Masters Percentage?(if Applicable) .
- University Name?
- Year of Higher Education completion?
- Your Stream?
- Your specialization?(like,Maths,Civics,Biology,etc) .
- How good are you in communication in following languages?

	Poor	Good	Fluent
Marathi	○	○	○
Hindi	○	○	○
English	○	○	○

- Rate yourself in the following:

	1	2	3	4	5
Critical Thinking	○	○	○	○	○
Teamwork	○	○	○	○	○
Problem Solving	○	○	○	○	○
Communication Skills	○	○	○	○	○
Technical Skills	○	○	○	○	○
Creativity	○	○	○	○	○
Leadership	○	○	○	○	○

- According to you,which of the following Skills are required for Placement?

	Not Required	Required But Not Compulsory	Required	Essential	Very Essential Add on
Communication Skills	○	○	○	○	○
Confidence	○	○	○	○	○
Critical Thinking	○	○	○	○	○
Technical Skills	○	○	○	○	○
Soft Skills	○	○	○	○	○
Right Attitude	○	○	○	○	○
Literacy Skills	○	○	○	○	○
Team Work	○	○	○	○	○
Sincerity	○	○	○	○	○
Truthfullness	○	○	○	○	○
Medium of Education	○	○	○	○	○
Experience	○	○	○	○	○
Coding/Programming	○	○	○	○	○

- According to you,do the following Parameters matters for getting Placed?

	Does't Affect	May Have Impact	Yes It Matters
Geographical Location of the Candidate	○	○	○
Teaching Quality of Teachers	○	○	○
Learning Environment	○	○	○
Course Design	○	○	○
Requirement Campaign	○	○	○
Oppurtunities Getting for Candidate	○	○	○
Mock Interviews	○	○	○
Financial Background	○	○	○
Group Discussion	○	○	○

- Does your Department/School/University Oraganization Seminars/Workshop Which may be helpfull for the Placement? ○ Yes ○ No.
- Are You Placed? ○ Yes ○ No.
- Sector you are Placed in? ○ Government ○ Private
- Name of the Organization/Company in which you are placed? .
- Post Offered at the time of your Placement(First Time): .
- Your Current Post/Designation: .
- First salary Offered?(per month)(Conditional) .
- Now your salary is-(conditional) .
- How many Switches you have done till now? .
- How many Interviews did you appear before getting placed? .
- Mode of your Placement drive: ○ Off-campus ○ On-campus
- Are you working in the field you studied? ○ Yes ○ No
- Did you get the job you desired for? ○ Yes ○ No

- Did you pursue any extra courses for Placement? ○ Yes ○ No
- If YES,which are those courses? .
- Did you go through any internship? ○ Yes ○ No
- If YES,which internship you went through? .
- Which website do you prefferd for Placement?
- What are you suggestions to boost the placement?

....
....
....
....
....
....

Chapter 2

Missing Data Imputation

2.1 What is missing data?

Some of the values in the data set are either lost or not observed or not available due to natural or non natural reasons is called as missing data.

2.1.1 Types of Missing Data

There are three types of missing data,

1. Missing Completely At Random (MCAR).
2. Missing At Random (MAR).
3. Missing Not At Random (MNAR).

1. Missing Completely At Random implies that the reason for the missingness of a field is completely random, and that we probably not predict that value from any other value in the dataset.

for example,

a. Questionnaire might be lost.

b. Blood sample might be damaged in the lab.

If values for observations are missing completely at random, then disregarding those cases would not bias the inferences made.

2. Missing Not At Random implies that there was a reason why the respondent didn't fill up that field. That is there is a relationship between the value to be missing and its value.

for example,

a. Suppose the study is not effective for reducing the blood pressure, their may be a chance of subject drop out.

b. In a study of income,respondent with low or high income might be less inclined to report their income.

In this type of data we can find more data.

3. Missing At Random implies that the missingness of the field can be explained by the values in other columns, but not from that column. MAR occures when there is a relationship between the missing values and the observed data. In other words we can say that the probability of missing depends on the available information.

for example,

a. If a child does not attend an educational assessment because the child is (genuinely) ill, this might be predictable from other data we have about the child's health, but it would not be related to what we would have measured had the child not been ill.

b. Missing blood pressure measurement may be lower than measured blood pressure younger people because younger people may be more likel to have missing blood pressur measurement.

In these situation, if we decide to proceed with the variable with missing values we might benefit from including the other variable to control bias in the missing observation.

Understanding the mechanism by which data is missing is important to decide which method to use to impute the missing values.

There are some thumb rule to decide whether imputation(act of replacing missing data with statistical estimates of the missing values) is necessary or not.

The goal of any imputation technique is to produce a complete dataset for analysis of the data.

Principle for dealing with missing data are :

1. Analysing only the available data.
2. Replacing the missing data that is doing imputation.

As mentioned in above, analysing only the available data we can also do that but by seeing the whole data we come to know that student had done degree, masters still not filling the value so we impute this missing values.

General Thumb Rule : **Imputation**

1. If MAR assumption is fulfilled: The missing data mechanism is said to ignorable, which basically means that there is no need to model the missing data mechanism as part of the estimation process. These are the method this report will cover.
2. If MAR assumption is not fulfilled: The missing data mechanism is said to be non-ignorable and, thus, it must be modeled to get good estimates of the parameters of interest. This requires a very good understanding of the missing data process.

Simply delete all cases that have any missing values at all, so you are left only with observations with all values observed.

These can be done by two methods as :

1. List wise deletion: In these the entire record is excluded in which the missing value is present.
2. Pair wise deletion: In these the missing values are excluded.

Advantages

1. Essay to understand and analysis is simple.

Disadvantages

1. Loss of information.
2. Produces bias in results.

CCA is generally used when missing data is 10percent of the total data.

Single Imputation : Single imputation procedure are those where one value for a missing data element is filled without defining an explicit model for the partially missing data.

Single Imputation are as : 1. Mean imputation : Used when missing data does not have any outlier. As mean is affected by extreme observation. 2. Median imputation : Used when there is outlier in the data. 3. Mode imputation : Used when data is categirical type.

Last observation carried forward (LOCF) takes the last available response and substitutes the values into all subsequent missing values.

Advantages

1. It generates a complete data set.
2. Easy to implement.

Disadvantages

1. Produce biased estimates.
2. Not sensible when the data are MCAR

Regression Imputation

In **Regression Imputation** the missing value is fill with the predicted value obtained by regressing the missing variable on other variables.

Advantages

1. It generates a complete data set.
2. Preserves relationship among variables involves in the imputation model.

Disadvantages

1. Not preserves variability around predicted values.

2.2 Missing Data Imputation

In our dataset there are no missing values in the SSC Percentage, HSC Percentage and HD (Highest Degree) and HDCY (Highest degree qualification year).There are missing values in degree percentage and masters percentage and year of placement. In degree percentage 52 values are missing and in masters percentage 263 values are missing. In this 52 degree percentage missing value some are missing because they did not done degree or they are pursuing degree and some are students are not fill their degree percentage marks still they did degree. After seeing the data we come to know that this 52 students had done degree but still not fill the percentage.In Masters also the condition is same, out of 263 missing values in masters degree some students highest education is B.degree only, some are doing masters so the values are missing and some students are not fill their marks.

Here we use multiple linear regression to impute the missing values as follows:

2.2.1 Steps in multiple linear regresssion missing data imputation method

1. Separate the Null values from the data frame and consider the test data.
2. Drop the Null values from the data frame and consider the train data.
3. Create ‘X train’ and ‘y train’ from the train data.
4. Build the linear regression model.
5. Create the ‘X train’ from the test data.
6. Apply the model on X test data and make predictions.
7. Replace the missing values with predicted values.

Let us see what is train data and test data.

Train Data: Training data is large dataset than test data that is used to teach a machine learning model. Training data is used to teach prediction models that use machine learning algorithms. It teaches a machine that how to extract features that are relevant to specific business goals.

Test Data: Once your machine learning model is built (with your training data), you need unseen data to test your model. This data is called testing data, and you can use it to evaluate the performance and progress of your algorithms’ training and adjust or optimize it for improved results.

Now, let us impute the missing value by multiple linear regression method.

Here,

response variable is y=Degree Percentage

regressors are

x_1= SSC Percentage

x_2= HSC Percentage

2.2.2 Implementation of missing data imputation using Python

Importing libraries

[1]:

```
import pandas as pd
import numpy as np
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
from sklearn.linear_model import LinearRegression
```

Importing dataset

```
[2]: df2=pd.read_csv("new.csv")
     df2.head()
```

```
[2]:    Sr.No  SSC Percentage  HSC Percentage  Degree Percentage  \
     0      1           88.80            47.0              69.00
     1      2           60.00            65.0              78.00
     2      3           74.80            65.0                NaN
     3      4           76.88            72.0                NaN
     4      5           80.00            74.0              87.84

        Masters Percentage                     HD  HDCY Year of placement
     0                 NaN                 B.Tech  2015               NaN
     1                55.0      Master Of Science  2019               NaN
     2                 NaN      Master Of Science  2022              2022
     3                 NaN  Bachelor Of Commerce  2018               NaN
     4                 NaN   Bachelor Of Science  2021               NaN
```

Here,

HD= Highest Degree

HDCY= Highest Degree Completion Year

```
[3]: # Droping the column sr.no
     df2.drop("Sr.No",axis=1,inplace=True)
```

```
[4]: # Checking null values in the dataframe
     df2.isnull().sum()
```

```
[4]: SSC Percentage          0
     HSC Percentage          0
     Degree Percentage      52
     Masters Percentage    263
     HD                      0
     HDCY                    0
     Year of placement     360
     dtype: int64
```

```
[5]: df=df2.drop(['Masters Percentage','HD','HDCY','Year of␣
       ↪placement'],axis=1)
     df.head()
```

```
[5]:    SSC Percentage  HSC Percentage  Degree Percentage
     0           88.80            47.0              69.00
     1           60.00            65.0              78.00
     2           74.80            65.0                NaN
     3           76.88            72.0                NaN
     4           80.00            74.0              87.84
```

1. Separate the NULL values

```
[6]: test_data=df[df["Degree Percentage"].isnull()]
```

```
[7]: test_data.isnull().sum()
```

```
[7]: SSC Percentage        0
     HSC Percentage        0
     Degree Percentage    52
     dtype: int64
```

```
[8]: test_data.shape
```

```
[8]: (52, 3)
```

The test data contains 52 missing values.

2. Drop the null values from the dataframe and consider as train data.

```
[9]: df.dropna(inplace=True)
```

Train data is a data from dataframe in which there is no missing values in the degree percentage.

```
[10]: df.shape
```

```
[10]: (470, 3)
```

check the null values from the train dataset.

```
[11]: df.isnull().sum()
```

```
[11]: SSC Percentage       0
      HSC Percentage       0
      Degree Percentage    0
      dtype: int64
```

So there is no missing value in the train dataset.

Create “X_train” and “y_train” from dataframe.

```
[12]: #Create the y_train
      # y_train means Rows from df['Degree Percentage'] with Non-Null values.
      y_train=df['Degree Percentage']
```

```
[13]: y_train.shape
```

```
[13]: (470,)
```

```
[14]: # X_train means datasets except features with non-null values.
      x_train=df.drop("Degree Percentage",axis=1)
      x_train.head()
```

```
[14]:    SSC Percentage  HSC Percentage
      0            88.8           47.00
      1            60.0           65.00
      4            80.0           74.00
      5            89.0           68.00
      6            91.4           62.15
```

```
[15]: x_train.shape
```

```
[15]: (470, 2)
```

Build the model Preparing the machine learning model(Linear Regression) on training the data set and predicting the missing in column Degree percentage.

```
[17]: lr=LinearRegression()
```

```
[18]: x_train.shape,y_train.shape
```

```
[18]: ((470, 2), (470,))
```

```
[19]: lr.fit(x_train,y_train)
```

```
[19]: LinearRegression()
```

5. Create the X_test from the test_data

```
[20]: x_test=test_data.drop("Degree Percentage",axis=1)
```

```
[21]: # x_test means dataset except df["Degree Percentage"] feature with Null␣
        ↪value
      x_test.head()
```

```
[21]:     SSC Percentage  HSC Percentage
      2            74.80         65.0000
      3            76.88         72.0000
      10           82.60         69.0000
      11           62.80         67.6666
      32           89.80         75.5000
```

```
[22]: x_test.shape
```

```
[22]: (52, 2)
```

```
[23]: test_data.shape
```

```
[23]: (52, 3)
```

6. Apply the model on x_test and predicting the missing values.

[24]:
```
#Apply the trained model on x_test
y_pred=lr.predict(x_test)
```

[25]:
```
y_pred
```

[25]:
```
array([75.52494563, 77.81064325, 77.02495355, 76.0205011 , 79.30189974,
       75.41720835, 70.91023956, 78.04878049, 69.68806695, 73.54127189,
       73.4725059 , 71.30167759, 77.19624191, 76.71408474, 75.4773843 ,
       77.13315717, 77.53925994, 77.44248003, 74.82949035, 76.57223437,
       73.36476862, 78.5916406 , 76.54112801, 75.04181961, 80.27107227,
       77.34536845, 74.30061382, 78.13676222, 75.99047531, 71.05208994,
       75.61879282, 79.92708482, 76.29863231, 73.13493651, 76.08324289,
       78.98116541, 80.43184456, 78.14191535, 76.36586405, 77.77427747,
       80.40478725, 78.52087512, 76.674119  , 74.01416399, 80.07297257,
       70.32369465, 75.87498272, 76.71887059, 83.39774729, 71.13224291,
       77.90214426, 80.62061853])
```

[26]:
```
y_pred.shape
```

[26]:
```
(52,)
```

[27]:
```
test_data.head()
```

[27]:
```
    SSC Percentage  HSC Percentage  Degree Percentage
2            74.80         65.0000                NaN
3            76.88         72.0000                NaN
10           82.60         69.0000                NaN
11           62.80         67.6666                NaN
32           89.80         75.5000                NaN
```

7. Replacing the missing values with predicted values.

[28]:
```
#Replacing the missing values with predicated values.
test_data['Degree Percentage']=y_pred
```

[29]:
```
test_data
```

[29]:
```
    SSC Percentage  HSC Percentage  Degree Percentage
2            74.80         65.0000          75.524946
3            76.88         72.0000          77.810643
10           82.60         69.0000          77.024954
11           62.80         67.6666          76.020501
32           89.80         75.5000          79.301900
50           92.80         63.0000          75.417208
54           80.00         50.0000          70.910240
70           85.00         72.0000          78.048780
72           60.00         48.0000          69.688067
89           77.60         58.5000          73.541272
90           59.00         60.0000          73.472506
```

92	50.00	54.0000	71.301678
124	91.80	68.6900	77.196242
125	72.00	69.0000	76.714085
138	89.00	63.5400	75.477384
140	67.00	70.7800	77.133157
141	89.30	70.0000	77.539260
145	86.00	70.0000	77.442480
151	72.76	63.0000	74.829490
171	78.00	68.0000	76.572234
174	77.00	58.0000	73.364769
176	71.00	75.0000	78.591641
190	84.20	67.3300	76.541128
209	80.00	63.0000	75.041820
216	80.80	79.3800	80.271072
220	66.00	71.5400	77.345368
251	76.40	61.0000	74.300614
252	88.00	72.0000	78.136762
279	69.00	67.0000	75.990475
286	74.00	51.0000	71.052090
306	78.00	65.0000	75.618793
310	82.40	78.1500	79.927085
312	82.00	66.7700	76.298632
313	80.00	57.0000	73.134937
322	83.00	66.0000	76.083243
324	88.40	74.6200	78.981165
363	90.40	79.0000	80.431845
371	60.00	74.6000	78.141915
376	81.80	67.0000	76.365864
383	75.64	72.0000	77.774277
394	82.00	79.6900	80.404787
397	83.00	73.6700	78.520875
407	84.40	67.7300	76.674119
410	68.80	60.8000	74.014164
425	89.00	78.0000	80.072973
430	60.00	50.0000	70.323695
442	79.80	65.6400	75.874983
500	83.00	68.0000	76.718871
503	94.00	88.0000	83.397747
510	77.60	50.9200	71.132243
517	80.00	72.0000	77.902144
520	86.00	80.0000	80.620619

Conclusion: In this way we replace the missing values in degree percentage by using the multiple linear regression imputation method. Now in our data there is no missing value in degree percentage.

Chapter 3

Visualization of Data

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand and pull insights from. The main goal of data visualization is to make it easier to identify patterns, trends and outliers in large data sets.
Here, in this chapter we see the nature of our dataset.How our data is? what kind of correlation is held by the attributes of data. Visualization allows to recognize relationships between the data, providing greater meaning to it.
Tools used for data visualization

- Simple bar diagram
- Multiple bar diagram
- Pie diagram
- Box plot
- Distribution plot
- Pivots

Let us visualize our data by drawing pivots, graphical representation in order to know our data well.

3.1 Data visualization using Python and conclusions

Importing libraries

[30]:

```
import warnings
warnings.filterwarnings('ignore')
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
```

Importing data

[31]:
```
df=pd.read_csv("Final Data11.csv")
```

[32]:
```
print("Number of rows in data :",df.shape[0])
print("Number of columns in data :", df.shape[1])
```

```
Number of rows in data : 523
Number of columns in data : 72
```

[33]:
```
df.shape
```

[33]:
```
(523, 72)
```

In our dataset there are 523 rows and 72 columns.

[34]:
```
df.isnull().sum()
```

[34]:
```
Gender
0
Age
0
 Hometown
0
School Type
0
SSC percentage
0
HSC percentage
0
SSC Board
0
HSC Board
0
Degree Percentage
0
Degree College
0
Highest Degree
0
Masters Percentage
248
University Name
51
Higher Education Completion Year
0
 Stream Of Education
0
Specialization
```

```
2
Marathi
0
English
0
Hindi
0
Critical Thinking
0
Team Work
0
Problem Solving
0
Communication Skills
0
Technical SKills
0
Creativity
0
Leadership
0
Communication Skills Rating
0
Confidence Rating
0
Critical Thinking Rating
0
Technical Skills Rating
0
Soft skills Rating
0
Right Attitude Rating
0
Literacy Skills Rating
0
Team Work  Rating
0
Sincerity  Rating
0
Truthfullness Rating
0
Medium of Education Rating
0
Experience Rating
0
Coding/Programming Rating
0
```

```
Geographical Location Of The Candidate Rating
0
Teaching Quality Of Teachers Rating
0
Learning Environment Rating
0
Course Design Rating
0
Recruitment Campaign Rating
0
Opportunities Getting For Candidate Rating
0
Mock Interviews Rating
0
Financial Background Rating
0
Group Discussion Rating
0
Does your Department/ School /University organize seminars/workshop␣
  ↪which may be helpful for the placement? 6
If Yes, was it truly helpful?
66
Placed?
0
Placed Sector
360
Year of placement
360
Placed Company/Organization
365
First Post offered
361
Current Post
362
First Salary Offered
416
Current Salary
438
How many switches you have done till now?
419
How many interviews did you appear before getting placed?
360
Placement Mode
360
Are you working in the field you studied ?
360
Did you get the job you desired for?
```

```
360
Did you pursue any extra courses for placement?
360
If YES, Which are those courses ?
473
Did you go through any internship?
360
If YES, which internship you went through?
469
Which website do you prefer for placements?
360
What are your suggestions to boost the placement?
361
```

Conclusions

1. In first section of our questionarrie which is for all placed and non-placed students in that there are missing value in some of the columns such as Master's percentage, University name, specialization.
2. In second section which is only for placed students in that there are missing values in some of the columns such as company name, first salary offered, current salary etc.

```
[35]: dfc=df.groupby('Gender').size().reset_index().rename(columns={0:
        ↪'count'})
      sns.countplot('Gender', data = df)
      plt.show()
      dfc
```

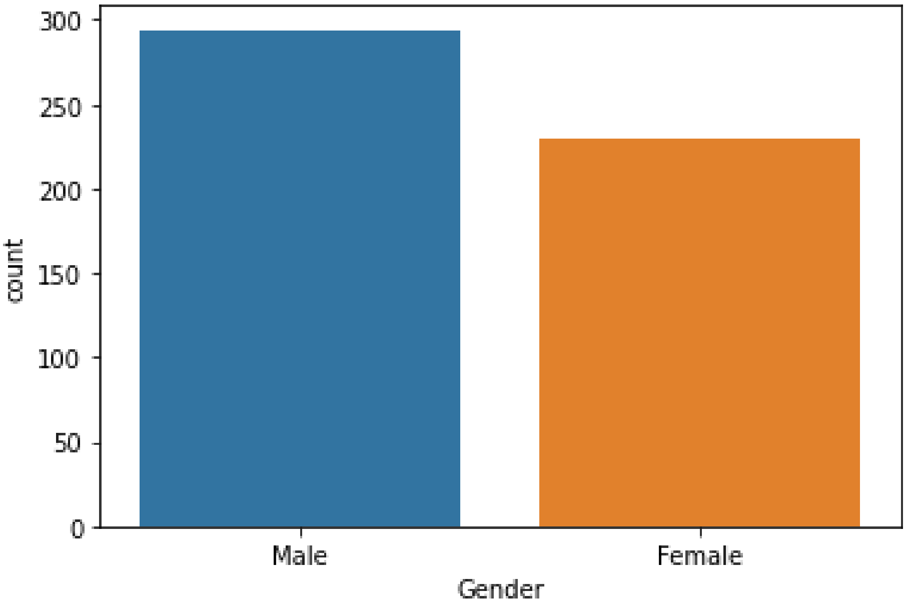


Figure 3.1: Genderwise Barplot

```
[35]:    Gender  count
      0  Female    229
      1    Male    294
```

Conclusions from figure (3.1)

In our data, there are 294 male candidates that is (56.21%) and 229 female candidates that is (43.78%).

```
[36]: plt.figure(figsize=(7,7))
      sns.countplot('Gender',hue='Placed?',data=df)
      df1=df.groupby(['Gender','Placed?']).size().reset_index().
        ↪rename(columns={0:'Count'})
      df1
```

```
[36]:    Gender Placed?  Count
      0  Female      No    172
      1  Female     Yes     57
      2    Male      No    188
      3    Male     Yes    106
```

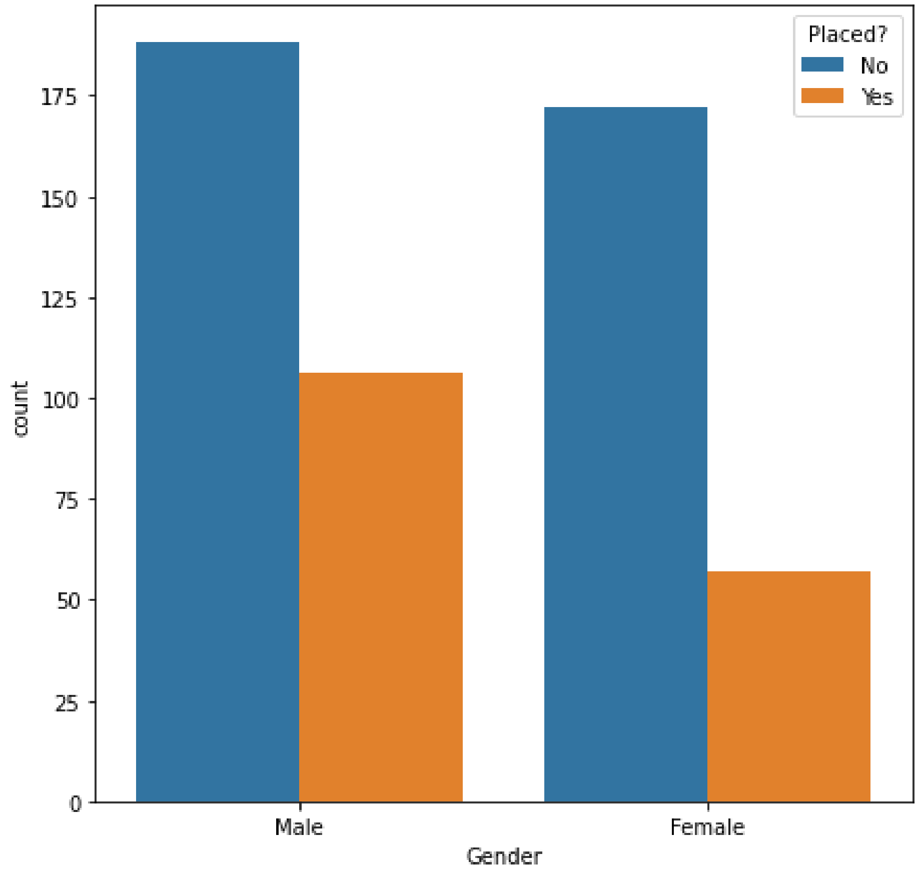
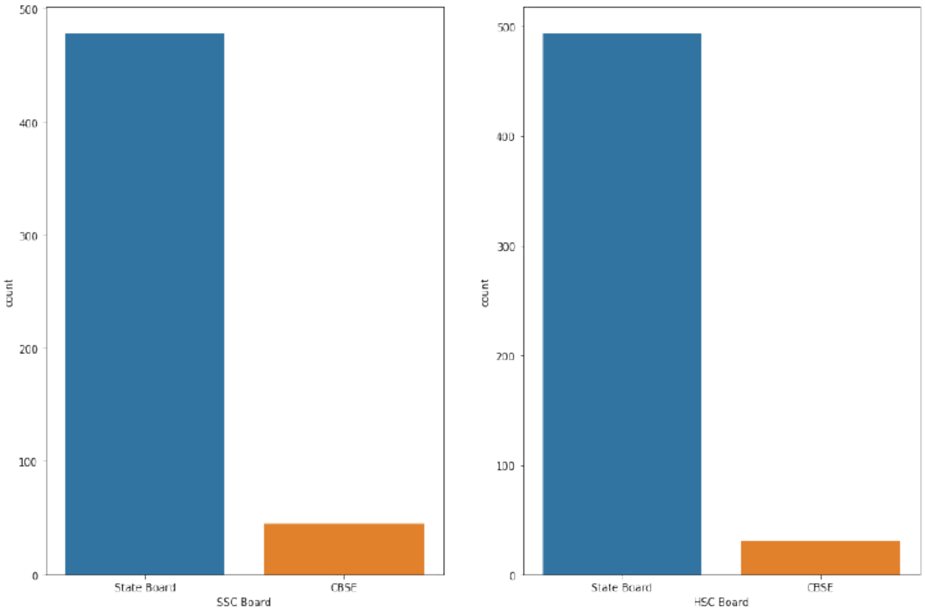


Figure 3.2: Genderwise placed non-placed multiple bar diagram

Conclusions from figure (3.2)
Out of 229 (43.78%) female candidates 57 (24.89%) females are placed and 172 (75.10%) female candidates are not placed. Out of 294 (56.21%) male candidates 106 (36.05%) males are placed and 188 (63.94%) male candidates are not placed.

[37]:

```
fig, axes = plt.subplots(1,2,figsize=(15,10))
sns.countplot('SSC Board',data=df,ax=axes[0])
sns.countplot('HSC Board',data=df,ax=axes[1])
plt.show()
```



2.

Figure 3.3: Boardwise bar diagram

[38]:

```
c=df['SSC Board'].value_counts()
d=df['HSC Board'].value_counts()
print(c)
print(d)
```

```
State Board    478
CBSE            45
Name: SSC Board, dtype: int64
State Board    493
CBSE            30
Name: HSC Board, dtype: int64
```

1. **Conclusions from figure (3.3)**
   In state board there are 478 (91.39%) students in SSC.

2. In CBSE board there are 45 (8.60%) students in SSC.
3. In state board there are 493 (94.26%) students in HSC.
4. In CBSE board there are 30 (5.73%) students in HSC.

[39]:
```
fig, axes = plt.subplots(1,2,figsize=(15,10))
sns.countplot('SSC Board',hue='Placed?',data=df,ax=axes[0])
sns.countplot('HSC Board',hue='Placed?',data=df,ax=axes[1])
plt.show()
```

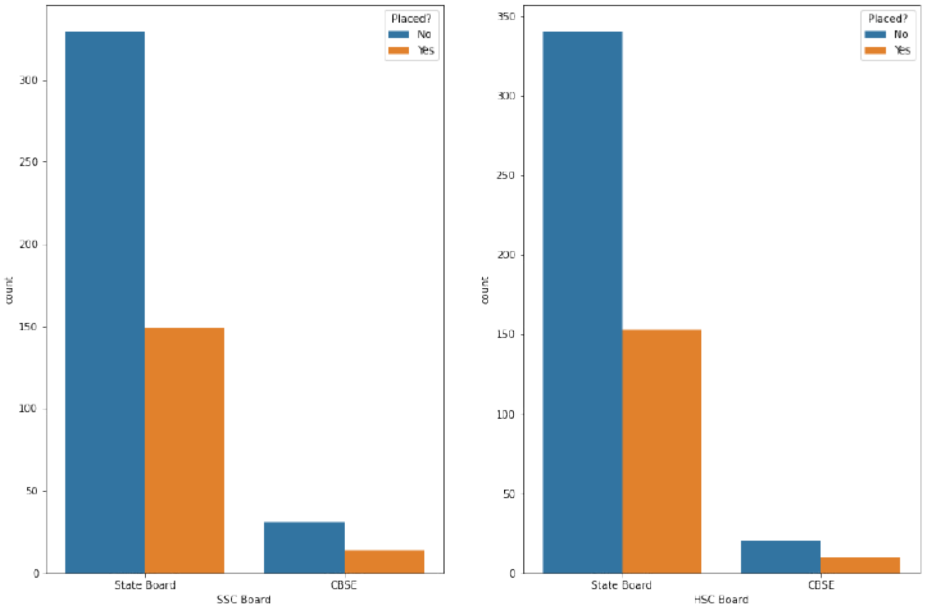


Figure 3.4: Boardwise placed non-placed multiple bar diagram

[40]:
```
df1=df.groupby(['SSC Board','Placed?']).size().reset_index().
  ↪rename(columns={0:'Count'})
df1
```

[40]:
```
      SSC Board Placed?  Count
0         CBSE      No     31
1         CBSE     Yes     14
2  State Board      No    329
3  State Board     Yes    149
```

[41]:
```
df2=df.groupby(['HSC Board','Placed?']).size().reset_index().
  ↪rename(columns={0:'Count'})
df2
```

[41]:
```
      HSC Board Placed?  Count
0         CBSE      No     20
1         CBSE     Yes     10
```

```
2  State Board      No    340
3  State Board     Yes    153
```

Conclusions from figure (3.4)

1. The students in state board in SSC are 478 out of that 149 are placed.
2. The student in CBSE board in SSC are 45 out of that 14 are placed.
3. The placement from State board is 31.17% and from CBSE board is 31.11%.
4. The students in state board in HSC are 493 out of that 153 are placed.
5. The student in CBSE board in HSC are 30 out of that 10 are placed.
6. The placement from State board is 31.03% and from CBSE board is 33.33%.

[42]:
```
fig, axes = plt.subplots(1,3,figsize=(15,10))
sns.boxplot(y = 'Placed?',x = 'SSC percentage',data=df, whis=np.
 ↪inf,ax=axes[0])
sns.boxplot(y = 'Placed?',x = 'HSC percentage',data=df, whis=np.
 ↪inf,ax=axes[1])
sns.boxplot(y = 'Placed?',x = 'Degree Percentage ',data=df, whis=np.inf)
plt.show()
```

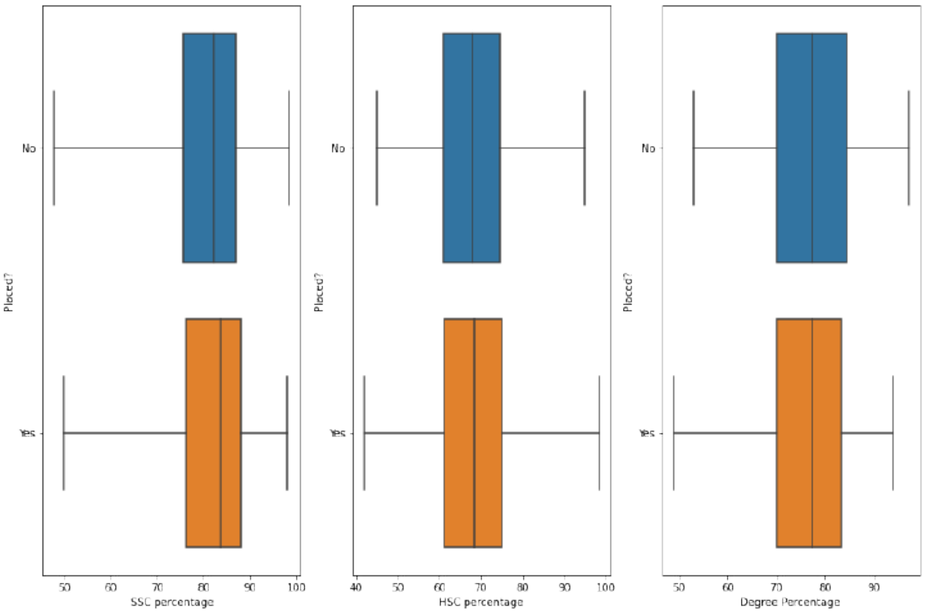


Figure 3.5: Percentagewise box plot of placed non-placed students

Conclusions from figure (3.5)

1. In SSC, 50% placed students score more than 84%.
2. In HSC, 50% placed students score more than 68%.
3. In degree, 50% placed students score more than 78%.

[43]:
```
data1=df.groupby(['University Name','Placed?']).size().reset_index().
 ↪rename(columns={0:'Count'})
```

```
data1.head(25)
```

```
[43]:                               University Name Placed?  Count
      0                       Annamalai University      No      1
      1            Banaras Hindu University,Banaras      No      2
      2            Banaras Hindu University,Banaras     Yes      1
      3                    CCS University Meerut,UP     Yes      1
      4                    DDU University,Gorakhpur      No      1
      5   Delhi Pharmaceutical Sciences and Research Uni...      No      1
      6   Delhi Pharmaceutical Sciences and Research Uni...     Yes      1
      7        Dr. B. A Technological University,Lonere      No      1
      8     Dr. Babasaheb Ambedkar Marathawada University      No      1
      9                     Dr.PDK Vidyapeeth,Akola      No      1
      10                  Govt Girls College Dhar,MP      No      1
      11                  Gujrat University,Ahmedabad     Yes      1
      12                                  IIT Bombay     Yes      1
      13  Indian Institute of Plantation Management, Ben...      No      1
      14                            KBCNMU, Jalgaon      No    273
      15                            KBCNMU, Jalgaon     Yes    107
      16                  KES Pratap College, Amalner      No      1
      17                         KTHM College,Nashik      No      1
      18                              MIT WPU,Pune      No      2
      19                              MIT WPU,Pune     Yes      1
      20                                     MSBTE      No      1
      21                                     MSBTE     Yes      1
      22                     Mumbai University,Mumbai      No      3
      23                     Mumbai University,Mumbai     Yes      3
      24                    R.C.Patel College,Shirpur      No      2
```

```
[44]: data2=data1[data1['Placed?']=='Yes']
      lab=list(data2['University Name'].unique())
```

```
[45]: plt.pie(data2['Count'],radius=3,labels=lab,autopct="%0.2f%%")
      plt.show()
```

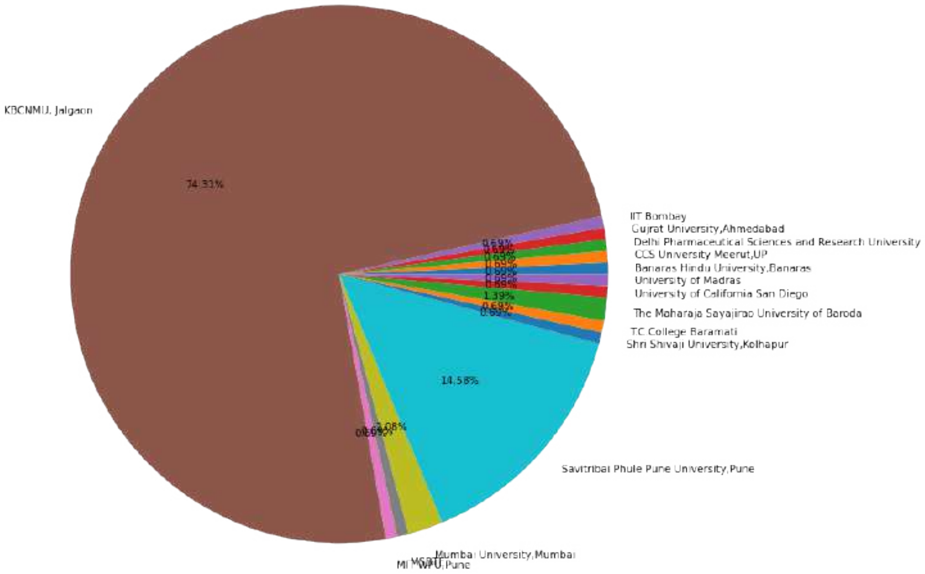


Figure 3.6: Pie chart on university names

Conclusions from figure (3.6)
In our data,(380) 74.31% candidates are from KBCNMU and (43) 14.58% students are from Savitribai Phule University,Pune,(3) 2.08% from the Mumbai university and others are from University of Madras, Gujarat University, IIT Bombay etc.

```
[46]: fig, axes = plt.subplots(3,3,figsize=(15,10))
sns.distplot(df['SSC percentage'],ax=axes[0,0])
sns.distplot(df['HSC percentage'],ax=axes[0,1])
sns.distplot(df['Degree Percentage '],ax=axes[0,2])
sns.distplot(df['Critical Thinking'],ax=axes[1,0])
sns.distplot(df['Team Work'],ax=axes[1,1])
sns.distplot(df['Problem Solving'],ax=axes[1,2])
sns.distplot(df['Technical SKills'],ax=axes[2,0])
sns.distplot(df['Creativity'],ax=axes[2,1])
sns.distplot(df['Leadership'],ax=axes[2,2])
plt.show()
```

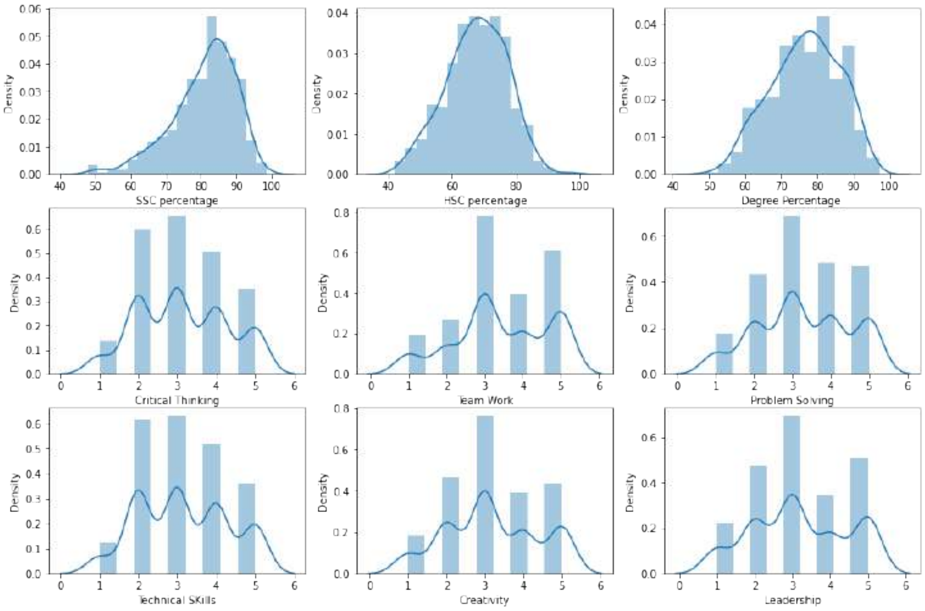


Figure 3.7: Some distribution plots

Conclusions from figure (3.7)

1. The distribution of SSC percentage in our data seems to be negatively skewed.
2. The distribution of HSC percentage in our data seem to be normally distributed.
3. The distribution of Degree percentage in our data seem to be normally distributed.
4. In distribution plot of critical thinking, team work, problem solving, technical skills, creativity and leadership in all that the highest peak is at 3(Required), that is this most of the students thinks that the above mentioned characteristics are required for placement.
5. In critical thinking and technical skills there is second highest peak at 2(required but not compulsory) it means that some of the students think that critical thinking and technical skill are required but if you don't have it still there is a chance of being placed.
6. In team work and leadership there is second highest peak at 5(very essential) it means that some of the students think that team work and leadership are very essential for placement.

[47]:

```
fig, axes = plt.subplots(3,3,figsize=(15,15))
sns.boxplot(df['SSC percentage'],ax=axes[0,0])
sns.boxplot(df['HSC percentage'],ax=axes[0,1])
sns.boxplot(df['Degree Percentage '],ax=axes[0,2])
sns.boxplot(df['Critical Thinking'],ax=axes[1,0])
sns.boxplot(df['Team Work'],ax=axes[1,1])
sns.boxplot(df['Problem Solving'],ax=axes[1,2])
sns.boxplot(df['Technical SKills'],ax=axes[2,0])
sns.boxplot(df['Creativity'],ax=axes[2,1])
sns.boxplot(df['Leadership'],ax=axes[2,2])
plt.show()
```

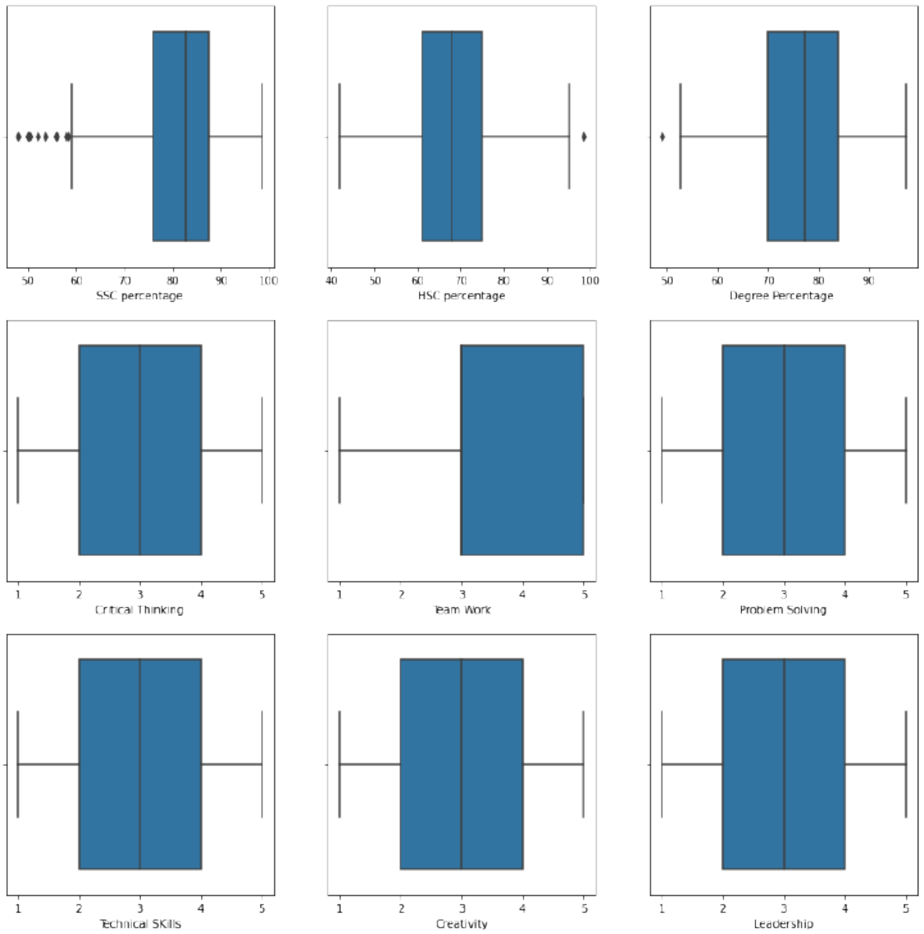


Figure 3.8: Box-plot on students skills and percentage

Conclusions from figure (3.8)

1. The SSC Percentage of the students are lies between 60 to 98 and some student have 52 to 60 percentage in SSC.
2. The HSC Percentage of the students are lies between 40 to 95 and and one have greater than 95 percentage.
3. The degree Percentage of the students are lies between 55 to 95 and one student have less than 55 percentage.

[48]:
```
d2=df.groupby(['SSC percentage','HSC percentage', 'Degree Percentage␣
 ↪','Placed?']).size().reset_index().rename(columns={0:'count'})
d2=d2.drop(['count'],axis=1)
d2.head()
```

[48]:
```
   SSC percentage  HSC percentage  Degree Percentage  Placed?
0            48.0           58.60               83.2       No
1            50.0           54.00               65.0      Yes
2            50.0           75.00               90.0      Yes
3            50.6           60.61               80.0       No
4            50.6           82.00               75.0       No
```

Conclusions

1. The student which has 50% in SSC, 54% in HSC and 65% in degree is placed.
2. The student which has 50.6% in SSC, 82% in HSC and 75% in degree is not placed.

```
[49]: d2.describe()
```

```
[49]:        SSC percentage  HSC percentage  Degree Percentage
      count      523.000000      523.000000         523.000000
      mean        80.826616       67.684729          76.614264
      std          9.111252        9.562524           9.524415
      min         48.000000       42.000000          49.000000
      25%         76.000000       61.085000          70.000000
      50%         82.730000       68.000000          77.290000
      75%         87.445000       75.000000          83.965000
      max         98.400000       98.400000          97.200000
```

```
[50]: h=df['Highest Degree'].value_counts()
      h.head(12)
```

```
[50]: Master Of Science       305
      Bachelor Of Science      85
      B.Tech                   43
      Other                    22
      MBA                      17
      MCA                      14
      M.Tech                   14
      Bachelor Of Commerce      5
      Master Of Art             5
      Diploma                   4
      Bachelor Of Art           3
      BCA                       2
      Name: Highest Degree, dtype: int64
```

```
[51]: l=df.groupby(['Highest Degree','Placed?']).size().reset_index().
       ↪rename(columns={0:'Count'})
      l.head(23)
```

```
[51]:           Highest Degree Placed?  Count
      0                 B.Tech      No     26
      1                 B.Tech     Yes     17
      2                    BCA      No      1
      3                    BCA     Yes      1
      4        Bachelor Of Art      No      3
      5   Bachelor Of Commerce      No      4
      6   Bachelor Of Commerce     Yes      1
      7    Bachelor Of Science      No     79
      8    Bachelor Of Science     Yes      6
      9                Diploma      No      2
```

```
10             Diploma     Yes      2
11              M.Pharm      No      1
12               M.Tech      No      7
13               M.Tech     Yes      7
14                  MBA      No     10
15                  MBA     Yes      7
16                  MCA      No      5
17                  MCA     Yes      9
18        Master Of Art      No      3
19        Master Of Art     Yes      2
20   Master Of Commerce      No      2
21    Master Of Science      No    202
22    Master Of Science     Yes    103
```

Conclusion
Out of 163 placed students,

1. 103 (63.19%) students had done master's degree.
2. 79 (48.46%) students had done degree.
3. 26 (15.95%) students had done b.Tech as a highest education.

[52]:
```
f=df[' Stream Of Education '].value_counts()
f
```

[52]:
```
Science & Technology     476
Commerce & Management     20
Arts & Humanities         27
Name:  Stream Of Education , dtype: int64
```

1. There are 476 (91.01%) students are from science branch.
2. There are 20 (3.82%) students are from commerce branch.
3. There are 27 (5.16%) students are from arts branch.

[53]:
```
sns.countplot(' Stream Of Education ', hue='Placed?', data=df)
```

[53]:

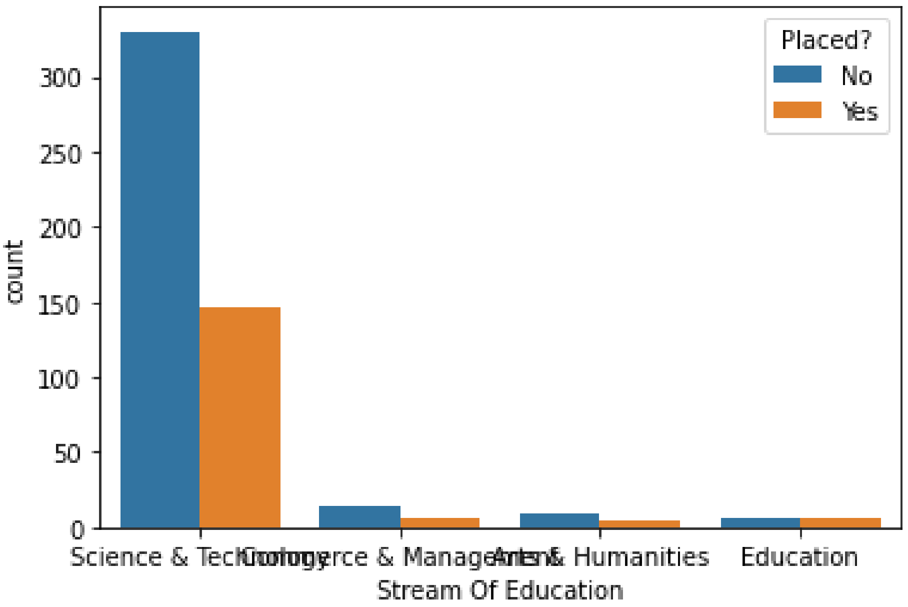


Figure 3.9: Streamwise placed non-placed multiple bar diagram

[54]:
```
g=df.groupby([' Stream Of Education ','Placed?']).size().reset_index().
  ↪rename(columns={0:'count'})
g
```

[54]:
```
     Stream Of Education  Placed?  count
0      Arts & Humanities       No     10
1      Arts & Humanities      Yes      5
2   Commerce & Management      No     14
3   Commerce & Management     Yes      6
4               Education      No      6
5               Education     Yes      6
6    Science & Technology      No    330
7    Science & Technology     Yes    146
```

Conclusion from figure (3.9)

1. There are 476 science branch students from that 146 (30.67%) are placed.
2. There are 20 commerce branch students from that 6 (30%) are placed.
3. There are 27 art's branch students from that 11 (40.74%) are placed.

[55]:
```
e=df['Specialization '].value_counts()
e
```

[55]:
```
Statistics                                            224
Mathematics                                            78
Chemical Engineering                                   17
Chemistry                                              17
```

```
Computer Science                              14
Actuarial Science                             13
Food Technology                               13
Paint Technology                              10
Physics                                        9
Environmental Science                          8
Information Technology                         8
Finance                                        6
Computer                                       6
Computational Mathematics                      5
Marketing                                      5
Computer Applications                          4
Mechanical                                     3
Economics                                      3
Mass Communication And Journalism              3
Environmental Sciences                         2
Microbiology                                   2
Environmental Science and Technology           2
Math                                           2
Biology                                        2
Computer Application                           2
Geography                                      2
Electrical                                     2
HRM                                            1
Food Science                                   1
Chemistry ( Pesticides And Agrochemicals)      1
Paints Technology                              1
Physics ( Material Science)                    1
Acturial Science                               1
Environment science and Technology             1
Oil Technology                                 1
Computers                                      1
B.tech Paint Technology                        1
Oil, Fats and Waxes                            1
Applied Statistics and Informatics             1
Electrical and Electronics                     1
Computer Engineering                           1
Machine learning and Data Science              1
Agriculture                                    1
Physics (Material Science)                     1
Instrumentation Science                        1
Accounting and Finance                         1
Mechanical engineering                         1
Big data analytics                             1
Pharmacy                                       1
Civics                                         1
MBA HR                                         1
```

```
PCMB                                              1
Statistics(Applied Statistics)                    1
Plastic Technology                                1
Paint and Coating Technology                      1
Analytical Chemistry                              1
Polymer Chemistry                                 1
Chemistry ( Polymer Chemistry)                    1
Organic Chemistry                                 1
Food Processing Business Management               1
Civil Engineering                                 1
Biotechnology                                     1
Geology                                           1
Public Administration                             1
Applied Geography                                 1
Enviornmental Science                             1
Social Work                                       1
Education                                         1
Master of Computer Application                    1
Chemical                                          1
Accounts                                          1
Polymer                                           1
Mass Communication and Journalism                 1
Applied Geology                                   1
Electronics                                       1
PHYSICS                                           1
Political Science                                 1
URCD                                              1
Commerce                                          1
MBA Marketing                                     1
Environment Science                               1
Computer application                              1
Community Development                             1
Polymers                                          1
Computer applications                             1
Computer science                                  1
Name: Specialization , dtype: int64
```

Conclusions

1. There are 224 (42.82%) students from the statistics subject.
2. There are 78 (14.91%) students from the mathematics subject.

[56]:

```
b=df.groupby([ 'Specialization ','Placed?']).size().reset_index().
↪rename(columns={0:'Count'})
b
```

[56]:

```
                          Specialization  Placed?  Count
0                  Accounting and Finance       No      1
1                                Accounts       No      1
```

```
2                             Actuarial Science      No     10
3                             Actuarial Science     Yes      3
4                              Acturial Science      No      1
5                                   Agriculture      No      1
6                          Analytical Chemistry      No      1
7                             Applied Geography     Yes      1
8                               Applied Geology      No      1
9            Applied Statistics and Informatics     Yes      1
10                      B.tech Paint Technology      No      1
11                           Big data analytics     Yes      1
12                                      Biology     Yes      2
13                                Biotechnology      No      1
14                                     Chemical      No      1
15                         Chemical Engineering      No     14
16                         Chemical Engineering     Yes      3
17                                    Chemistry      No     14
18                                    Chemistry     Yes      3
19    Chemistry ( Pesticides And Agrochemicals)     Yes      1
20               Chemistry ( Polymer Chemistry)      No      1
21                                       Civics     Yes      1
22                            Civil Engineering     Yes      1
23                                     Commerce      No      1
24                        Community Development     Yes      1
25                    Computational Mathematics      No      4
26                    Computational Mathematics     Yes      1
27                                     Computer      No      1
28                                     Computer     Yes      5
29                         Computer Application      No      2
30                        Computer Applications      No      2
31                        Computer Applications     Yes      2
32                         Computer Engineering     Yes      1
33                             Computer Science      No      7
34                             Computer Science     Yes      7
35                         Computer application      No      1
36                        Computer applications     Yes      1
37                             Computer science      No      1
38                                    Computers     Yes      1
39                                    Economics      No      3
40                                    Education      No      1
41                                   Electrical     Yes      2
42                   Electrical and Electronics     Yes      1
43                                  Electronics      No      1
44                        Enviornmental Science      No      1
45                          Environment Science     Yes      1
46           Environment science and Technology      No      1
47                        Environmental Science      No      8
48         Environmental Science and Technology     Yes      2
```

```
49                      Environmental Sciences      No      2
50                                     Finance      No      3
51                                     Finance     Yes      3
52         Food Processing Business Management      No      1
53                                Food Science      No      1
54                             Food Technology      No      8
55                             Food Technology     Yes      5
56                                   Geography      No      2
57                                     Geology     Yes      1
58                                         HRM      No      1
59                      Information Technology      No      6
60                      Information Technology     Yes      2
61                     Instrumentation Science     Yes      1
62                                      MBA HR      No      1
63                               MBA Marketing      No      1
64           Machine learning and Data Science     Yes      1
65                                   Marketing      No      3
66                                   Marketing     Yes      2
67           Mass Communication And Journalism      No      2
68           Mass Communication And Journalism     Yes      1
69           Mass Communication and Journalism     Yes      1
70              Master of Computer Application     Yes      1
71                                        Math      No      1
72                                        Math     Yes      1
73                                 Mathematics      No     63
74                                 Mathematics     Yes     15
75                                  Mechanical     Yes      3
76                      Mechanical engineering     Yes      1
77                                Microbiology      No      2
78                              Oil Technology      No      1
79                         Oil, Fats and Waxes     Yes      1
80                           Organic Chemistry      No      1
81                                        PCMB      No      1
82                                     PHYSICS      No      1
83                            Paint Technology      No      5
84                            Paint Technology     Yes      5
85                Paint and Coating Technology      No      1
86                           Paints Technology      No      1
87                                    Pharmacy      No      1
88                                     Physics      No      8
89                                     Physics     Yes      1
90                 Physics ( Material Science)      No      1
91                  Physics (Material Science)      No      1
92                          Plastic Technology      No      1
93                           Political Science     Yes      1
94                                     Polymer     Yes      1
95                           Polymer Chemistry     Yes      1
```

```
96                             Polymers     Yes      1
97                Public Administration      No      1
98                          Social Work      No      1
99                           Statistics      No    154
100                          Statistics     Yes     70
101      Statistics(Applied Statistics)     Yes      1
102                                URCD     Yes      1
```

Conclusions

1. 70 (42.94%) students from statistics subject are placed.
2. 63 (17.5%) students from mathematics subject are not placed.

```
[57]: data1=df.groupby(['Placed Company/Organization ','Placed?']).size().
      ↪reset_index().rename(columns={0:'Count'})
      data1
```

```
[57]:                            Placed Company/Organization  Placed?  Count
      0                                  ADANI Wilmar Ltd.      Yes      2
      1                             Aarco Engineering India     Yes      1
      2                               Acadecraft Pvt. Ltd.      Yes      1
      3                                            Academic     Yes      1
      4                  ActionEdge Research Services LLP.      Yes      1
      5                          Alkem Laboratories Ltd.        Yes      1
      6                                              Amdocs     Yes      1
      7                      Amoli Organics Vapi Pvt. Ltd.      Yes      1
      8                        Anchanto Services Pvt. Ltd.      Yes      1
      9                                          Ankleshwar     Yes      1
      10                                  Atos Syntel ,Pune     Yes      1
      11             Australian Broadcasting Corporation        Yes      1
      12                                          Axis Bank     Yes      1
      13                       BAJAJ HOUSING FINANCE Ltd.       Yes      1
      14                         Baidyanath Ayurveda Nagpur     Yes      1
      15                       Bajaj Allianz Life Insurance     Yes      1
      16            Bharati Vidyapeeth Medical College,Pune     Yes      1
      17                                            Bitwise     Yes      1
      18          Brose India Automotive Systems Pvt. Ltd.      Yes      1
      19                                 CBSE SCHOOL JALGAON    Yes      1
      20                                           CLINEXEL     Yes      1
      21                                 Capgemini Pvt. Ltd.    Yes      2
      22                             Capital Foods Pvt. Ltd.    Yes      1
      23                                Carraro India Ltd.      Yes      1
      24                      Cartesian Consulting Pvt. Ltd.    Yes      1
      25       Centre For Youth Development and Activities      Yes      1
      26                               Chiprn IT Solutions      Yes      1
      27                                         Chubb Ltd.     Yes      1
      28                            Cliantha Research Ltd.      Yes      2
      29                                       Cognizant IT     Yes      1
      30          Coromandal International Fertilizer Ltd.      Yes      1
```

```
31                                Cytel Pune Pvt. Ltd.     Yes      1
32                     Datamatics Global Services Ltd.     Yes      1
33                              Dell Technologies LLC.     Yes      1
34                           Department of Post, India     Yes      2
36
Directorate of Economics and Statistics     Yes      1
37                      Dolat Capital Market Pvt. Ltd.     Yes      1
38   Dr. D. Y. Patil Medical College, Hospital and ...     Yes      1
39                               EFTEC India Pvt. Ltd.     Yes      1
40                                      Evident Health     Yes      1
41                                   FARE Labs Gurgaon     Yes      1
42                                      Fenoplast Ltd.     Yes      1
43                 Food Fortification Initiative,India     Yes      1
44                                   Force Motors Ltd.     Yes      2
45                           Future Marketing Analysis     Yes      1
46                                             Genpact     Yes      1
47              Government Engineering College Jalgaon     Yes      1
48                                 Government Official     Yes      1
49                                   Government Sector     Yes      1
50                          Grauer and Weil India Ltd.     Yes      2
51                         Gujarat Ambuja Exports Ltd.     Yes      1
52                                 Hans Solutions Ltd.     Yes      1
53                                 Hindustan Petroleum     Yes      1
54                                                 IBM     Yes      1
55                                    IDBI Intech Ltd.     Yes      1
56                                                IHMP     Yes      1
57                         ISRI Technologies Pvt. Ltd.     Yes      1
58                                           IndiaMART     Yes      2
59                       Inetllify Solutions Pvt. Ltd.     Yes      1
60                                             Infosys     Yes     10
61                           Intelii Systems Pvt. Ltd.     Yes      1
62                         Jain Irrigation System Ltd.     Yes      1
63                                             Jalgaon     Yes      1
64                   Jalgaon Janata Sahakari Bank Ltd.     Yes      1
65                                    Kaldin Solutions     Yes      1
66                                              Kantar     Yes      3
67                          Kirloskar Oil Engines Ltd.     Yes      1
68                                 Kotak Mahindra Bank     Yes      1
69                      Krios Info Solutions Pvt. Ltd.     Yes      2
70                                             Labcorp     Yes      1
71                             Local Fund Audit Office     Yes      1
72                              M. J. College, Jalgaon     Yes      1
73                                 MAGFinserv Co. Ltd.     Yes      1
74                              Mahyco Seeds Pvt. Ltd.     Yes      1
75                     Masterline Lubricants Pvt. Ltd.     Yes      1
76     Max Super Speciality Hospital, Saket, New Delhi     Yes      1
77   Ministry of Statistics and Programme Implement...     Yes      1
```

```
78               Multi-Act Equity Consultancy Pvt. Ltd.     Yes      1
79                              Nielsen India Pvt. Ltd.     Yes      3
80                        Nova Agriscience Pvt.Ltd. Hyderabad     Yes     ␣
 ↪1
83                             Nupeak IT Solutions LLP.     Yes      1
84                                               Oracle     Yes      1
85             Orion English Medium State Board Jalgaon     Yes      1
86                                            Pinterest     Yes      1
87                  Precision Seals Manufacturing Ltd.     Yes      1
88                      Principal Global Services Pune     Yes      1
89                      Protogene Consulting Pvt. Ltd.     Yes      1
90                                   Qualys Pvt. Ltd.     Yes      1
91                                  RWS India Pvt. Ltd.     Yes      1
92                            RYAN International School     Yes      1
93                     Real Estate Company Legal Advisor     Yes      1
94                             Reliance Industries Ltd.     Yes      2
95                                   Reliance Pvt. Ltd.     Yes      1
96                        S.P.D.M. ACS College, Shirpur     Yes      1
97                      SY Minerals Pvt. Ltd. Adilabad     Yes      1
98                                  Shadowfax Pvt. Ltd.     Yes      1
99               Siddhivinayak Technical Campus,Shegaon     Yes      1
100                                        SmartAnalyst     Yes      1
101                           St. Teresas School Jalgaon     Yes      1
102  State Bureau of Health Intelligence and Vital ...     Yes      1
103                     State Health Society Maharashtra     Yes      1
104                           Sycamore Software Pvt. Ltd     Yes      1
105                                                 TCS     Yes     15
106                                                 TFS     Yes      1
107                             The Aakanksha Foundation     Yes      1
108               The Outlook Group (Summer Internship)     Yes      1
109                          Tntra Innovation Ecosystem     Yes      1
110                                           UPL Ltd.     Yes      1
111                   Upthink Edutech Services Pvt. Ltd.     Yes      2
112                         Veeda Clinical Research Ltd.     Yes      1
113                   Vermittler IT Consulting Pvt. Ltd.     Yes      1
114                                Vodafone Idea Ltd.     Yes      1
115                                             WhizAI     Yes      1
116                                         Wipro Ltd.     Yes      3
117                               Worley India Pvt. Ltd.     Yes      1
118     Yashwantrao Chavan Maharashtra Open University     Yes      1
119                                       Z. P. School     Yes      1
120                                 Zensar Technologies     Yes      1
121                                eClerx Services Ltd.     Yes      1
```

Conclusions

1. In our data there are 163 students candidates who are placed. There are 121 different organization in which they are working.
2. 10 (6.13%) are working in Infosys,15 (9.20%) are working in TCS,3 (1.84%) and

other students are working in other companies/ organizations as Genpact,Fenoplast Ltd,Department of Post office, India etc.

[58]:
```
data4=df.groupby(['Placed Sector ','Placed?']).size().reset_index().
  ↪rename(columns={0:'Count'})
data4.head()
```

[58]:
```
  Placed Sector  Placed?  Count
0     Government      Yes     13
1        Private      Yes    150
```

[60]:
```
plt.pie(data4['Count'],radius=2,labels=lab2,autopct="%0.2f%%")
plt.show()
```

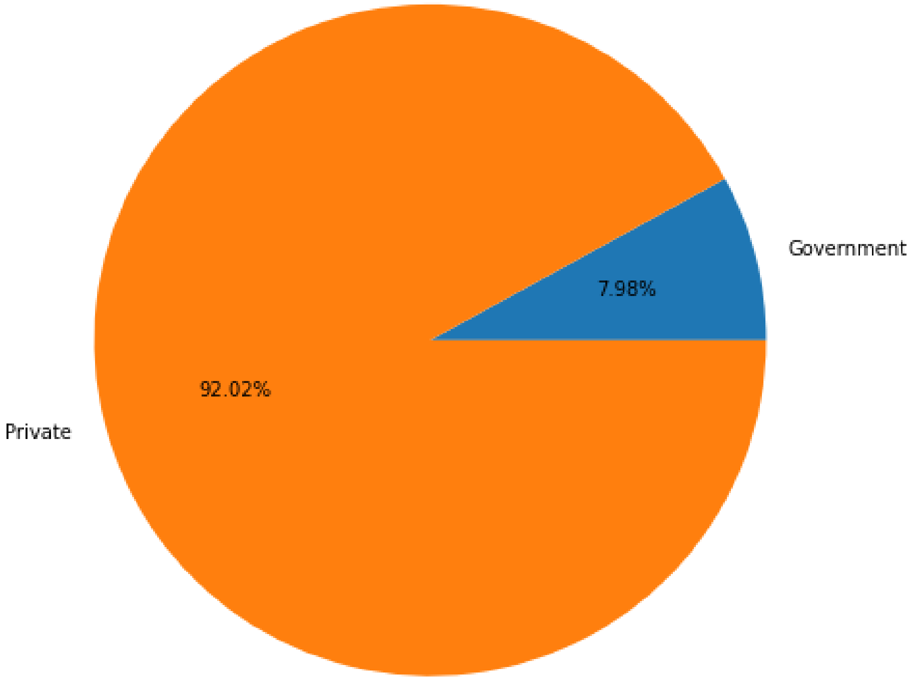


Figure 3.10: Placed sector pie diagram

Conclusions from figure (3.10)

1. Out of 163 placed candidates 13 (7.97%) are placed in Government sector and 150 (92.02%) candidates are placed in private sector.

[61]:
```
data7=df.groupby([' Placement Mode ','Placed?']).size().reset_index().
  ↪rename(columns={0:'Count'})
data7.head()
```

[61]:
```
   Placement Mode  Placed?  Count
0      Off-Campus      Yes    113
1       On-Campus      Yes     50
```

```
[62]: ['Off-Campus', 'On-Campus']
```

```
[63]: plt.pie(data7['Count'],labels=lab3,autopct='%0.2f%%',radius=2)
      plt.show()
```

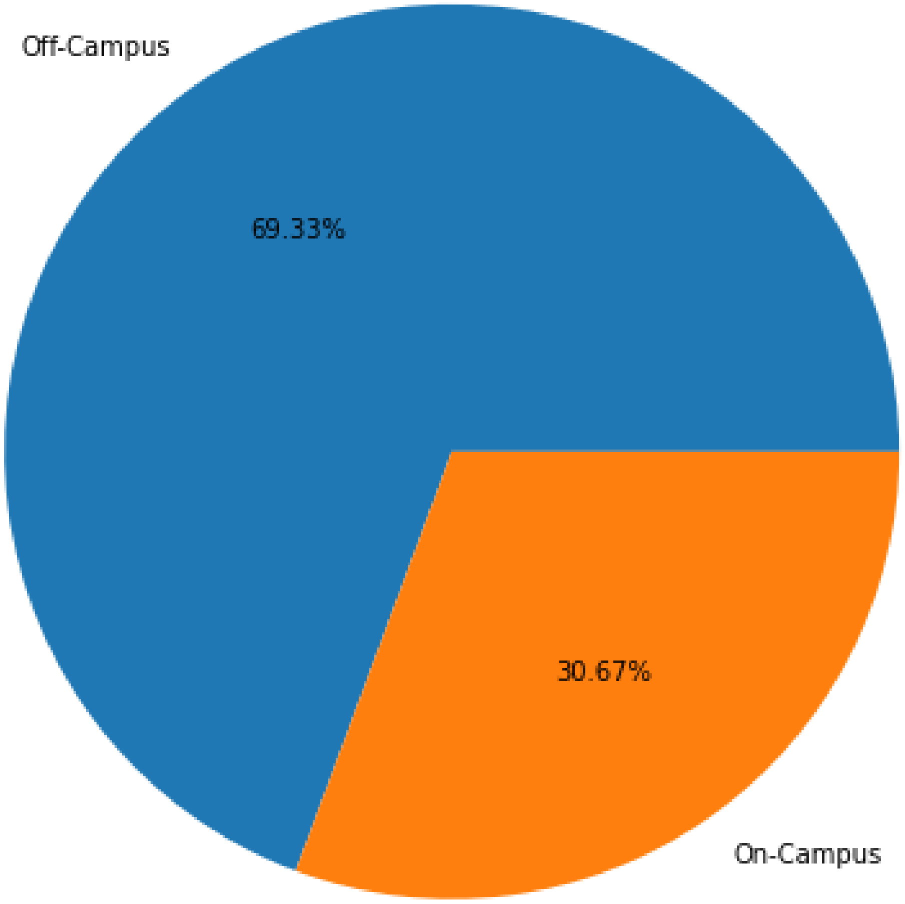


Figure 3.11: Mode of placement pie diagram

Conclusions from figure (3.11)

1. 113 (69.32%) candidates are placed through off-campus mode.
2. 50 (30.67%) candidates are placed through on-campus mode.

```
[64]: data5=df.groupby(['How many switches you have done till now?','Placed?
       ↪']).size().reset_index().rename(columns={0:'Count'})
      data5
```

```
[64]:   How many switches you have done till now? Placed?  Count
      0                                         0     Yes     56
      1                                         1     Yes     23
      2                                         2     Yes     12
      3                                         3     Yes      7
      4                                         4     Yes      2
```

```
5                              6 or more     Yes      4
```

Conclusions
Out of 163 placed candidates

1. 56 (34.35%) candidates are still working in the organization or their first campany.
2. 23 (14.11%) candidates were switched their first job.
3. 12 (7.36%) candidates were switched their second job.
4. 7 (4.29%) candidates were switched theur third job.
5. 2 (1.22%) candidates were switched their fourth job.
6. 4 (2.45%) candidates were switched their fifth or more than fifth job.

[65]:
```
data6=df.groupby(['How many interviews did you appear before getting␣
 ↪placed?','Placed?']).size().reset_index().rename(columns={0:'Count'})
data6
```

[65]:
```
  How many interviews did you appear before getting placed? Placed? ␣
 ↪Count

0                                              0          Yes  15
1                                              1          Yes  42
2                                              2          Yes  41
3                                              3          Yes  19
4                                              4          Yes  14
5                                              5          Yes  7
6                                      6 or more          Yes  25
```

Conclusions
Out of 163 placed candidates,

1. 15 (9.20%) students got job in first interview.
2. 42 (25.76%) students got job in second interview.
3. 41 (25.15%) students got job in third interview.
4. 19 (11.65%) students got job in fourth interview.
5. 14 (8.58%) students got job in fifth interview.
6. 25 (15.33%) students got job in sixth or more than sixth interview.

[66]:
```
data8=df.groupby(['Are you working in the field you studied ?','Placed?
 ↪']).size().reset_index().rename(columns={0:'Count'})
data8.head()
```

[66]:
```
  Are you working in the field you studied ? Placed?  Count
0                                         No     Yes     27
1                                        Yes     Yes    136
```

```
[68]: labels1 = [' Not got job in their field ',' Got job in their field']
      plt.pie(data8['Count'],labels=labels1,autopct='%0.2f%%',radius=2)
      plt.show()
```

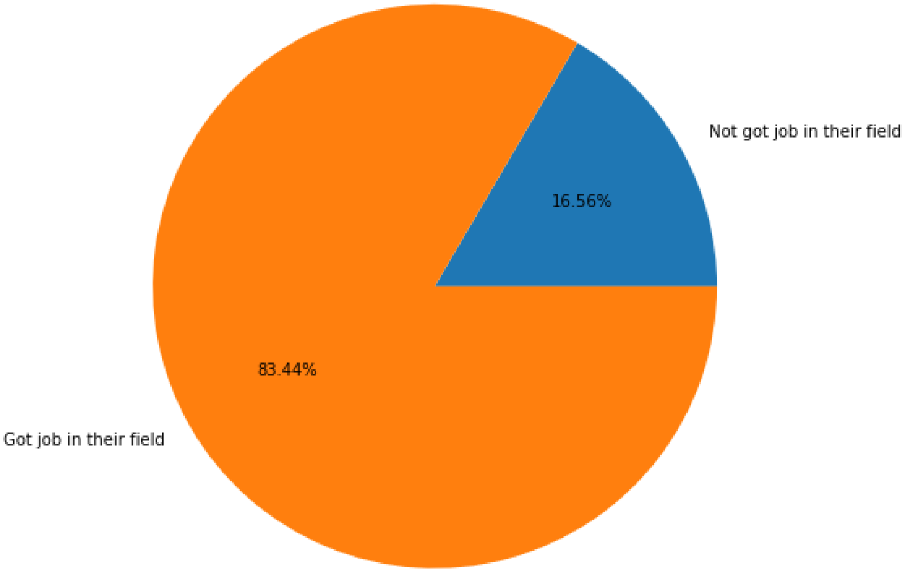


Figure 3.12: Pie diagram on their field job got or not

Conclusions from figure (3.12)

1. 136 (83.43%) candidates got the job in which they had done their education.
2. 27 (16.56%) candidates not got the job in which they had done their education.

```
[69]: data9=df.groupby(['Did you get the job you desired for?','Placed?']).
        ↪size().reset_index().rename(columns={0:'Count'})
      data9.head()
```

```
[69]:   Did you get the job you desired for? Placed?  Count
      0                                   No     Yes     29
      1                                  Yes     Yes    134
```

```
[71]: labels = ['Not got desired job',' Got desired job']
      plt.pie(data9['Count'],labels=labels,autopct='%0.2f%%',radius=2)
      plt.show()
```

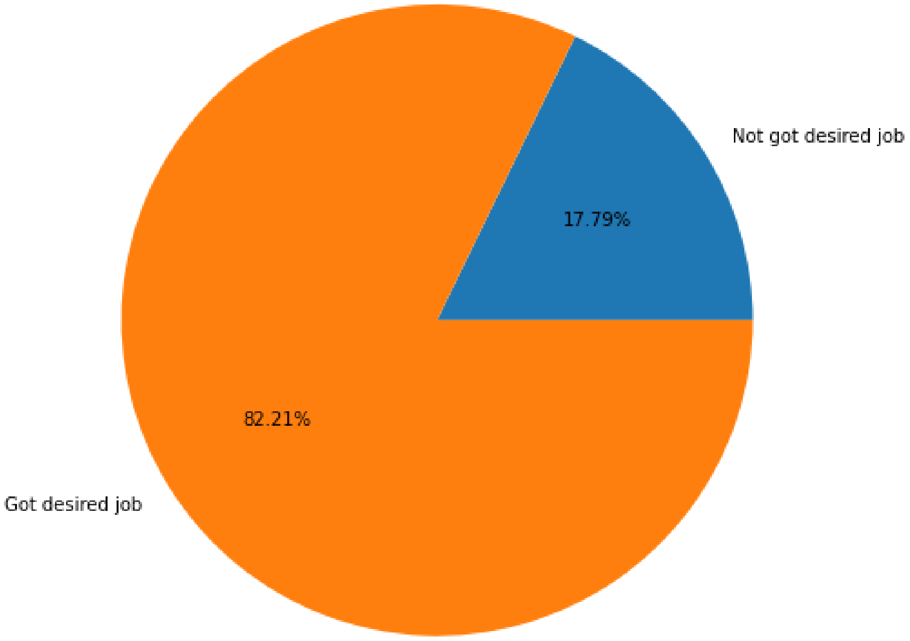


Figure 3.13: Pie diagram on ’Got desired job or not?’

Conclusions from figure (3.13)

1. 134 (82.20%) candidates got the job which they want.
2. 29 (17.79%) candidates not got the job which they want.

```
[72]: data10=df.groupby(['Did you pursue any extra courses for placement?
        ↪','Placed?']).size().reset_index().rename(columns={0:'Count'})
      data10.head()
```

```
[72]:   Did you pursue any extra courses for placement? Placed?  Count
      0                                              No     Yes    112
      1                                             Yes     Yes     51
```

```
[74]: labels2 = [' Done extra courses',' Not done extra courses']
      plt.pie(data10['Count'],labels=labels2,autopct='%0.2f%%',radius=2)
      plt.show()
```

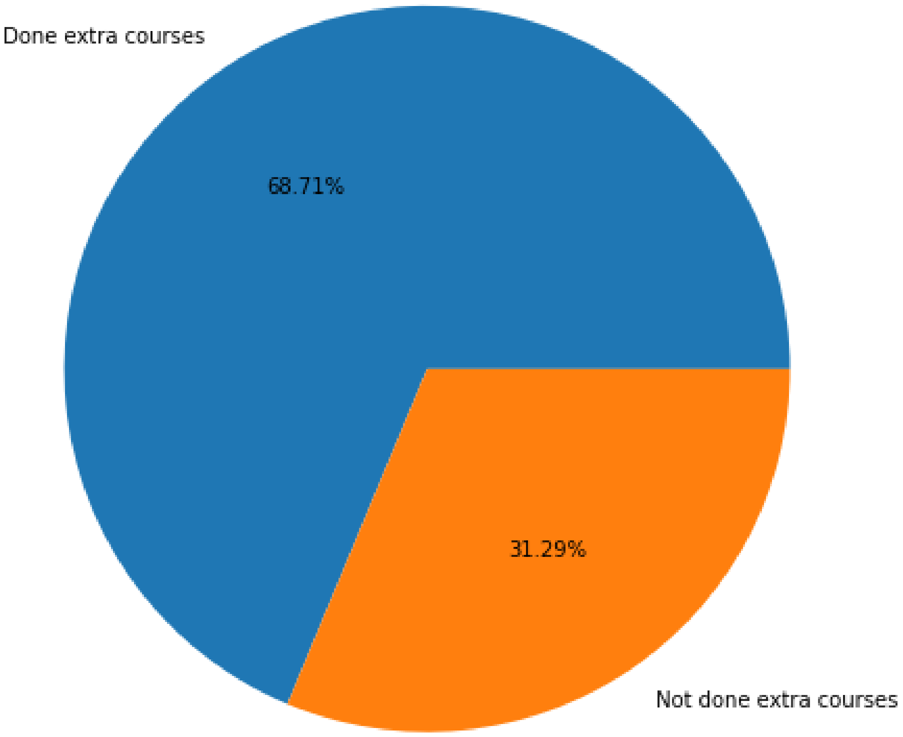


Figure 3.14: Pie diagram on done extra course or not

Conclusions from figure (3.14)

1. 112 (68.71%) candidates had done extra courses for placement.
2. 51 (31.28%) candidates had not done extra courses for placement.

```
[75]: data11=df.groupby([ 'Did you go through any internship?','Placed?']).
        ↪size().reset_index().rename(columns={0:'Count'})
      data11.head()
```

```
[75]:   Did you go through any internship? Placed?  Count
      0                                 No     Yes    103
      1                                Yes     Yes     60
```

```
[77]: labels3 = [' Done internship',' Not done internship']
      plt.pie(data11['Count'],labels=labels3,autopct='%0.2f%%',radius=2)
      plt.show()
```

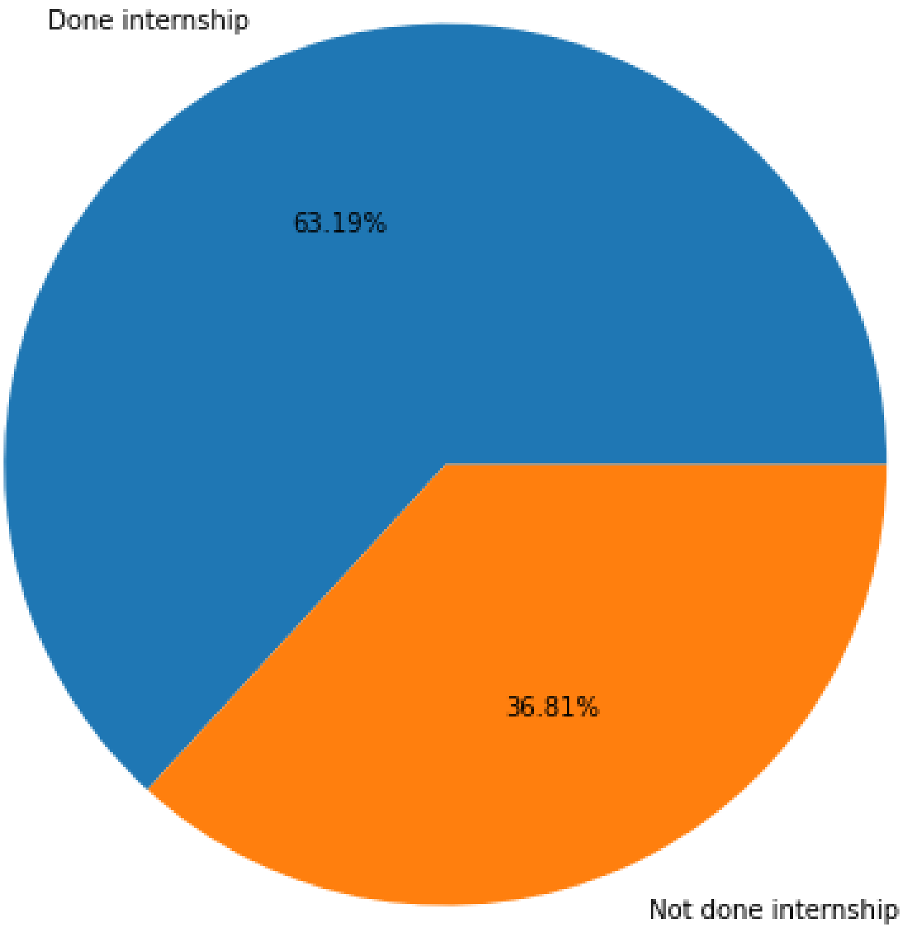


Figure 3.15: Pie diagram on done internship or not

Conclusions on figure (3.15)

1. 103 (63.19%) candidates had done intership for placement.
2. 60 (36.80%) candidates had not done efor placement.

[78]:
```
s1=pd.read_csv("First salary.csv")
```

[79]:
```
s2=pd.read_csv("Second salary.csv")
```

[80]:
```
fig, axes = plt.subplots(1,2,figsize=(15,10))
sns.distplot(s1['First Salary Offered '],ax=axes[0],kde_kws={'clip': (0.
  ↪0, 1.0)})
sns.distplot(s2['Current Salary '],ax=axes[1],kde_kws={'clip': (0.0, 1.
  ↪0)})
plt.show()
```

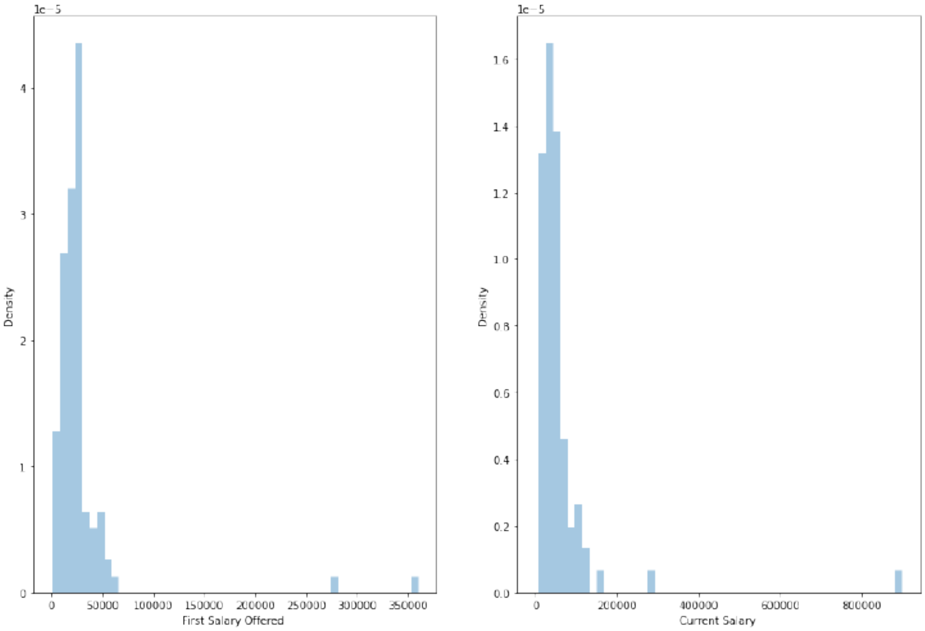


Figure 3.16: Distribution plots of salaries

```
[81]: fig, axes = plt.subplots(1,2,figsize=(15,5))
      sns.boxplot(s1['First Salary Offered '],ax=axes[0])
      sns.boxplot(s2['Current Salary '],ax=axes[1])
```

[81]:

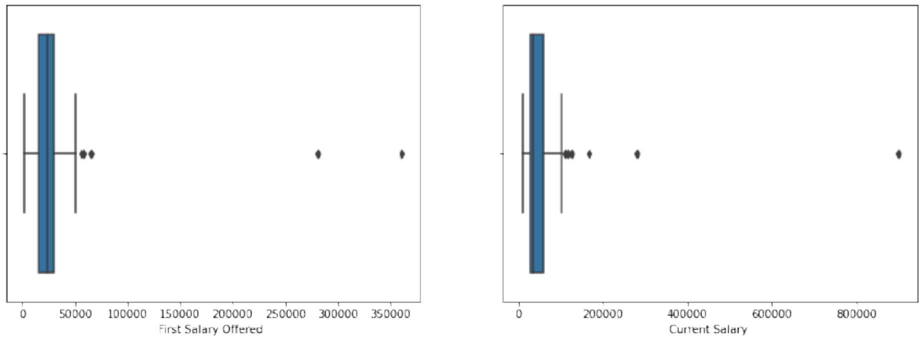


Figure 3.17: Box-plots of salaries

```
[82]: s1.describe()
```

```
[82]:        First Salary Offered
      count            109.000000
      mean           28614.403670
```

```
std          42158.241065
min           1600.000000
25%          15000.000000
50%          23000.000000
75%          30000.000000
max         360000.000000
```

[83]:
```
s2.describe()
```

[83]:
```
       Current Salary
count       85.000000
mean     57980.047059
std     100462.669220
min       7000.000000
25%      25000.000000
50%      35000.000000
75%      56000.000000
max     900000.000000
```

1. The average first salary is 28,614 per month with standard error 42,158.
2. The average current salary is 57,980 per month with standard error 1,00,462.

[84]:
```
data11=df.groupby([ 'Which website do you prefer for placements?
 ↪','Placed?']).size().reset_index().rename(columns={0:'Count'})
data11.head(10)
```

[84]:
```
        Which website do you prefer for placements? Placed?  Count
0                                       Google form     Yes      1
1                                            Indeed     Yes      1
2                                          LinkedIn     Yes     32
3                            LinkedIn, Freshersworld     Yes      1
4                               LinkedIn, Glassdoor     Yes      1
5                LinkedIn, Glassdoor, Freshersworld     Yes      1
6  LinkedIn, Glassdoor, Freshersworld, W3School ,...     Yes      1
7                                  LinkedIn, Indeed     Yes      3
8                LinkedIn, Indeed, Devnetjobsindia     Yes      1
9                                 LinkedIn, Naukari     Yes     47
```

Conclusions:
In 163 placed students,

1. 32 (19.63%) students uses LinkedIn website for placements.
2. 47 (28.83%) students uses LinkedIn and Naukari website for placements.
3. 18 (11.04%) students uses Naukari website for placements.
4. Other websites such as Glassdoor, Freshersworld, Indeed are also use by the students for placements.

Chapter 4

Statistical Analysis of Data

4.1 Statistical Analysis of data

A hypothesis is used in an experiment to define the relationship between two variables. The purpose of a hypothesis is to find the answer to a question. A formalized hypothesis will force us to think about what results we should look for in an experiment.
In this chapter, we perform hypothesis testing by using the following test:

Tests used for hypothesis testing

1. Understanding the relationship between categorical variables.
2. Understanding the relationship between continous variables.

1.Understanding the relationship between categorical variables

A family of hypothesis tests that compare the observed distribution of your categorical data to their expected distribution under the null hypothesis. Different chi-square tests exist to understand the relationship between categorical varibles:

1. Test if a statistical model fits your data.
2. Test the independence between categorical variables. For example, a manager of three customer support call centers wants to know if a successful resolution of a customer's problem (yes or no) depends on which branch receives the call. The manager tallies the successful and unsuccessful resolutions for each branch in a table and performs a chi-square test of independence on the data. In this case, the chi-square statistic quantifies how the observed distribution of counts varies from the distribution you would expect if no relationship exists between call center and a successful resolution.

Chi square test for independence of attribute:

Computing system for independency/dependency:

In this section, we make a tabular form of n levels of attribute A & m levels of attribute B such as,

A\B	B_{1} B_{2} ———B_{3}———B_{m}	Total
A_{1}	O_{11} O_{12} ———-O_{1j}—–O_{1m}	(A_{1})
A_{2}	O_{21} O_{22}————O_{2j}—–O_{2m}	(A_{2})
.	.	.
.	.	.
A_{i}	O_{i1} O_{i2}————O_{i3}—–O_{im}	(A_{i})
.	.	.
.	.	.
A_{n}	O_{n1} O_{n2}————O_{n3}—-O_{nm}	(A_{n})
Total	(B_{1}) (B_{2})————(B_{i})—–(B_{m})	N

O_{ij}=Observed frequency corresponding to (i^{th}) row and (j^{th}) column
i.e. corresponding to $(i, j)^{th}$ cell.
i = 1, 2........., n., j=1, 2........m.
A_i= Total of observed frequency in the i^{th} row.
B_j=Total of observed frequency in the j^{th} column.
Here to test,
Ho: Two attributes A and B are independent.
V/S
H1: Two attributes A and B are not independent.
Fix- Level of significance= α
To carrying out above test we compute test statistic as follows,
Oij=Observed frequency corresponding to i^{th} row and $j^{th} column$
i.e. corresponding to $(i, j)^{th}$ cell.
i= 1, 2........., n., j=1, 2........m.

(A$_i$) =$\sum_{j=1}^{m} Oij$ **= Total of observed frequency in the ith row.**
(B$_j$) =$\sum_{i=1^n} Oij$ **= Total of observed frequency in the jth row.**
Here to test,
Ho: Two attributes A and B are independent.
V/S
H1: Two attributes A and B are not independent.
Fix- Level of significance=α
To carrying out above test we compute test statistic as follows,

$$= \sum_{i=1}^{n} = Oij \sum_{j=1}^{m} = Oij \frac{Oij}{2Eij} - N \chi^2 = \sum_{i=1}^{n} = Oij \sum_{j=1}^{m} = Oij \frac{(Oij-Eij)^2}{Eij}$$

$$= \sum_{i=1}^{n} = Oij \sum_{j=1}^{m} = Oij \frac{Oij}{2Eij} - N$$

Where,
Eij=expected frequency corresponding to $(i, j)^{th}$ **cell.**
$\chi^2_{calculated}$=$\frac{(Ai)(Bj)}{N}$,$i = 1,2......n, j = 1,2.....m$
N=Total Frequency
Also,under H_0
$X2 \sim \chi^2((n-1)(m-1)$ **Decision rule**
We reject if, $\chi^2_{calculated} > \chi^2_{(n-1)(m-1),\alpha}$

let's perform testing of hypothesis by using chi-square test using python

[85]:

```
import warnings
warnings.filterwarnings('ignore')
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
import pandas as pd
import numpy as np
import scipy.stats as st
```

[86]:

```
df=pd.read_csv("Placement csv.csv")
```

To test the following hypothesis

Ho:There is no relationship between placement and gender.

v/s

H1:There is relationship between placement and gender.

[87]:

```
df_table=pd.crosstab(df[ 'Gender'],df['Placed?'])
print(df_table)
```

```
Placed?   No  Yes
Gender
Female   172   57
Male     188  106
```

[88]:

```
st.chi2_contingency(df_table)
```

[88]:

```
(6.966966800781471,
 0.008302811927709253,
 1,
 array([[157.6290631,  71.3709369],
        [202.3709369,  91.6290631]]))
```

χ^2_{cal}: 6.96
p-value :0.0083
DF: 1
Here,p-value 0.0083 is less than 0.05, So reject null hypothesis.
There is relationship between placement and gender.

Ho1:There is no relationship between placement and hometown.

v/s

H1:There is relationship between placement and homwtown.

[89]:

```
df1_table=pd.crosstab(df[ ' Hometown'],df['Placed?'])
print(df1_table)
```

```
Placed?       No  Yes
 Hometown
City         153   66
Small Town    41   21
Village      166   76
```

```
[90]: st.chi2_contingency(df1_table)
```

```
[90]: (0.3259952566931311,
       0.8495932033528395,
       2,
       array([[150.7456979 ,  68.2543021 ],
              [ 42.67686424,  19.32313576],
              [166.57743786,  75.42256214]]))
```

χ^2_{cal}: 0.3259
p-value :0.8495
DF: 2
Here, pvalue 0.8495 is greater than 0.05, So fail to reject null hypothesis. **There is no relationship between placement and hometown.**

Ho2:There is no relationship between placement and School Type.

v/s

H1:There is relationship between placement and School Type.

```
[91]: df2_table=pd.crosstab(df['School Type '],df['Placed?'])
      print(df2_table)
```

```
Placed?         No  Yes
School Type
Government     227  100
Private        133   63
```

```
[92]: st.chi2_contingency(df2_table)
```

```
[92]: (0.07604735550295974,
       0.782727146705599,
       1,
       array([[225.08604207, 101.91395793],
              [134.91395793,  61.08604207]]))
```

χ^2_{cal}: 0.0760
p-value :0.7827
DF: 1
Here, pvalue is 0.7827 greater than 0.05. So, fail to reject null hypothesis. **There is no relationship between placement and School Type.**

Ho3:There is no relationship between placement and SSC board.

v/s

H1:There is relationship between placement and SSC board.

[93]:
```
df3_table=pd.crosstab(df['SSC Board'],df['Placed?'])
print(df3_table)
```

```
Placed?       No  Yes
SSC Board
CBSE          28   13
ICSE           3    1
State Board  329  149
```

[94]:
```
st.chi2_contingency(df3_table)
```

[94]:
```
(0.07649622833054835,
 0.962474107803633,
 2,
 array([[ 28.22179732,  12.77820268],
        [  2.75334608,   1.24665392],
        [329.0248566 , 148.9751434 ]]))
```

χ^2_{cal}: 0.0764
p-value :0.9624
DF: 2
Here, pvalue is 0.9624 greater than 0.05. So we fail to reject null hypothesis. **There is no relationship between placement and SSC Board.**

Ho4:There is no relationship between placement and HSC board.

v/s

H1:There is relationship between placement and HSC board.

[95]:
```
df4_table=pd.crosstab(df['HSC Board'],df['Placed?'])
print(df4_table)
```

```
Placed?       No  Yes
HSC Board
CBSE          19   10
ICSE           1    0
State Board  340  153
```

[96]:
```
st.chi2_contingency(df4_table)
```

[96]:
```
(0.6054521307853232,
 0.7388014519387585,
 2,
 array([[1.99617591e+01, 9.03824092e+00],
        [6.88336520e-01, 3.11663480e-01],
        [3.39349904e+02, 1.53650096e+02]]))
```

χ^2_{cal}: 0.6054
p-value :0.7388

DF: 2
Here,pvalue is 0.7388 greater than 0.05.So we fail to reject null hypothesis.

There is no relationship between placement and HSC Board.

Ho5:There is no relationship between placement and Stream of education.

v/s

H1:There is relationship between placement and Stream of education.

```
[97]: df5_table=pd.crosstab(df[ ' Stream Of Education '],df['Placed?'])
      print(df5_table)
```

```
Placed?                 No  Yes
 Stream Of Education
Arts & Humanities       10    5
Commerce & Management   14    6
Education                6    6
Science & Technology   330  146
```

```
[98]: st.chi2_contingency(df5_table)
```

```
[98]: (2.083780300122776,
       0.555201082197833,
       3,
       array([[ 10.3250478 ,   4.6749522 ],
              [ 13.7667304 ,   6.2332696 ],
              [  8.26003824,   3.73996176],
              [327.64818356, 148.35181644]]))
```

χ^2_{cal}: 2.0837
p-value :0.5552
DF: 3

Here,pvalue is 0.5552 greater than 0.05.So we fail to reject null hypothesis.

There is no relationship between placement and Stream of education.
Conclusion

Group 1	Group 2	P-value	Decision	Association
Placed?	Gender	0.008	Reject H0	Yes
Placed?	Hometown	0.8495	Fail to Reject H0	No
Placed?	School Type	0.7827	Fail to Reject H0	No
Placed?	SSC Board	0.9624	Fail to Reject H0	No
Placed?	HSC Board	0.7388	Fail to Reject H0	No
Placed?	Stream of Education	0.5552	Fail to Reject H0	No

In this way we check the association between categorical variables.

2. Understanding the relationship between continous variables.

A t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two groups, which may be related in certain features. The t-test is one of many tests used for the purpose of hypothesis testing in statistics.

The basic idea for calculating a t-test is to find the difference between the means.That is to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another.

Assumptions

1. Independence: The observations in one sample are independent of the observations in the other sample.
2. Normality: Both samples are approximately normally distributed.
3. Homogeneity of Variances: Both samples have approximately the same variance.

- H0:The average SSC Percentage of placed students are same as average SSC Percentage of Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average SSC Percentage of placed students are Greater than average SSC Percentage of Non-placed students.
  i.e $\mu_1 > \mu_2$
- H0:The average HSC Percentage of placed students are same as average HSC Percentage of Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average HSC Percentage of placed students are Greater than average HSC Percentage of Non-placed students.
  i.e $\mu_1 > \mu_2$
- H0:The average Degree Percentage of placed students are same as average Degree Percentage of Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average Degree Percentage of placed students are Greater than average Degree Percentage of Non-placed students.
  i.e $\mu_1 > \mu_2$
- H0:The average Master's Percentage of placed students are same as average Master's Percentage of Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average Master's Percentage of placed students are Greater than average Master's Percentage of Non-placed students.
  i.e $\mu_1 > \mu_2$
  Here,
  No. of placed students= $n_1 = 163$
  No.of Non Placed students= $n_2 = 360$

	Placed		Non-Placed				
Percentage	Mean	Varience	Mean	Varience	tcal	|tcal|	P-value
SSC	82.91	71.67	80.50	83.45	2.93	2.93	0.0017
HSC	70.36	88.02	67.71	84.27	3.01	3.01	0.0013
Degree	70.80	87.41	76.41	114.61	1.503	1.503	0.0667
Master's	69.61	88.89	70.50	125.08	-0.948	0.948	0.1718

from p-value in the above table, we conclude that

(a) The average SSC Percentage of placed students are Greater than average SSC Percentage of Non-placed students.

(b) The average HSC Percentage of placed students are Greater than average HSC Percentage of Non-placed students.

(c) The average Degree Percentage of placed students are same as average Degree Percentage of Non-placed students.

(d) The average Master's Percentage of placed students are Greater than average Master's Percentage of Non-placed students,

- H0:The average SSC Percentage of placed male students are same as average SSC Percentage of placed female students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average SSC Percentage of placed female students are Greater than average SSC Percentage of placed male students.
  i.e $\mu_1 > \mu_2$

- H0:The average HSC Percentage of Male placed students are same as average HSC Percentage of Female placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average HSC Percentage of female placed students are Greater than average HSC Percentage of male placed students.
  i.e $\mu_1 > \mu_2$

- H0:The average Degree Percentage of Male placed students are same as average Degree Percentage of Female placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average Degree Percentage of female placed students are Greater than average Degree Percentage of male placed students.
  i.e $\mu_1 > \mu_2$

Placed							
	Male		Female				
Percentage	Mean	Varience	Mean	Varience	tcal	|tcal|	P-value
SSC	80.76	88.26	82.93	63.89	-1.553	1.553	0.0061
HSC	66.18	110.68	70.27	88.43	-2.539	2.539	0.0060
Degree	74.64	97.62	79.30	72.29	-3.150	3.150	0.00097

from p-value in the above table, we conclude that

(a) The average SSC Percentage of male placed students are same as the average SSC Percentage of male placed students.

(b) The average HSC Percentage of Female placed students are Greater than mean of HSC Percentage of Non-placed.

(c) The average Degree Percentage of placed students are Greater than mean of Degree Percentage of Non-placed.

- H0:The average SSC Percentage of Male Non-placed students same as average SSC Percentage of Female Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average SSC Percentage female Non-placed students are Greater than average SSC Percentage of male Non-placed students.
  i.e $\mu_1 > \mu_2$

- H0:The average HSC Percentage of Male Non-placed students are same as average HSC Percentage of Female Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average HSC Percentage of female Non-placed students are Greater than average HSC Percentage of male Non-placed students.
  i.e $\mu_1 > \mu_2$

- H0:The average Degree Percentage of Male Non-placed students are same as average Degree Percentage of Female Non-placed students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average Degree Percentage of female Non-placed students are Greater than average Degree Percentage of male Non-placed students.
  i.e $\mu_1 > \mu_2$

Non-Placed							
	Male		Female				
Percentage	Mean	Varience	Mean	Varience	tcal	|tcal|	P-value
SSC	80.19	84.35	80.84	82.85	-0.492	0.492	0.311
HSC	66.965	94.63	68.53	71.64	-1.80	1.80	0.036
Degree	73.768	124.8	79.29	87.45	-6.27	6.27	5.02E

from p-value in the above table, we conclude that

(a) The average SSC Percentage of female non-placed students are same as the average SSC percentage of non-placed students.

(b) The average HSC Percentage of placed students are Greater than mean of HSC Percentage of Non-placed.

(c) The average Degree Percentage of placed students are Greater than mean of Degree Percentage of Non-placed.

- H0:The average SSC Percentage of Male students are same as average SSC Percentage of Female students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average SSC Percentage of female students are Greater than average SSC Percentage of male students.
  i.e $\mu_1 > \mu_2$

- H0:The average HSC Percentage of Male students are same as average HSC Percentage of Female students.
  i.e $\mu_1 = \mu_2$
  vs

  H1:The average HSC Percentage of female students are Greater than average HSC Percentage of male students.
  i.e $\mu_1 > \mu_2$

- H0:The average Degree Percentage of Male students are same as average Degree Percentage of Female students.
  i.e $\mu_1 = \mu_2$
  vs
  H1:The average Degree Percentage of female students are Greater than average Degree Percentage of male students. students.
  i.e $\mu_1 > \mu_2$

	Male		Female				
Percentage	Mean	Varience	Mean	Varience	tcal	|tcal|	P-value
SSC	80.403	84.84	81.369	78.49	-1.211	1.211	0.1131
HSC	66.682	100.56	68.97	76.38	-2.78	2.78	0.027
Degree	74.085	115.21	79.30	83.68	-5.992	5.992	1.9E-0.9

from p-value in the above table, we conclude that

(a) The average SSC Percentage of Male students are same as average SSC Percentage of Female students.

(b) The average HSC Percentage of placed females students Greater than mean of HSC Percentage of males.

(c) The average Degree Percentage of placed female students are Greater than the average of Degree Percentage of male.

- H0:The average of first salary offered is same as the average of current salary offered.
  vs
  H1:The average of first salary offered is less than the average of current salary offered.

First Salary offered		Current Salary				
Mean	Varience	Mean	Varience	tcal	|tcal|	P-value
28614.40	1761011626	57980.05	9974009696	-2.54158	2.54158	0.005914

From the p value in the above table, we can conclude that,

(a) The average of first salary offered is less than the average of current salary offered.

Chapter 5

Analysis of Data Using Factor Analysis in SPSS

5.1 What is factor analysis?

Factor Analysis (FA) is an exploratory data analysis method used to search influential underlying factors from a set of observed variables. It helps in data interpretations by reducing the number of variables. It extracts maximum common variance from all variables and puts them into a common score.

Use factor analysis, like principal components analysis, to summarize the data covariance structure in a few dimensions of the data. However, the emphasis in factor analysis is the identification of underlying "factors" that might explain the dimensions associated with large data variability. Primarily used to examined the structure of data by explainig the correlations among variables. Factor analysis summarizes data into a few dimensions by condensing many variables into a smaller set of latent variables or factors. Factor analysis is commonly used in social science, market research, advertising, psychology, finance,operation research and other industries that used large datasets. **Market researchers** use factor analysis to identify price-sensitive customers, identify brand features that influence consumer choice, and helps in understanding channel selection criteria for the distribution channel.

Consider a example of credit card company that creates a survey to evaluate customer satisfaction. The survey is designed to answer questions in three categories: timeliness of service, accuracy of the survice and courteousness of phone operators. The company can used factor analysis to ensure that the survey items address these three areas before sending the survey to many customers. If the survey does not adequately measure the three factors, then the company should reevaluate the questions and retest the survey before sending it to customers.

Mathematically, factor analysis is somewhat similar to multiple regression analysis in that each variable is expressed as linear combination of underlying factors. The amount of variance a variable share with all other variables included in the analysis is referred to as communality.

The covariation among the variables is described in terms of a small number of common factors plus a unique factor for each variable. These factors are not overtly observed. If the variables are standardized, the factor model may be represented as:

$X_i = A_{i1}F1 + A_{i2}F_2 + A_{i3}F_3 + ... + A_{im}F_m + V_1U_1$
Where,

X_i=standardized variable.

A_{ij}=Standardized multiple regression coefficient of variable i on common factor j.

F = common factor.

V_i=standardized regression coefficient of variable i on unique factor.

U_i=the unique factor for variable i.

m= number of common factors.
The unique factors are correlated with each other and with the common factors. The common factors themselves can be expressed as linear combinations of the observed variables.

$F_1 = W_{i1}X_2 + W_{i2}X_1 + W_{i3}X_3 + ... + W_{ik}X_k$
where,

F_i= estimate of ith factor.

W_i= weight or factor score coefficient.

k= number of variables.

The key Statistics associated with factor analysis are as follows:

Kaiser-Meyer -Olkin(KMO) measure of sampling adequacy:

Kaiser-Meyer -Olkin (KMO) measure of sampling adequacy is an index used to examine the appropriate Kaiser-Meyer-Olkin of factor analysis. Values between 0.5 and 0.7 shows Mediocre Sample size, value greater than 0.7 shows good sample size and value greater than 0.8 shows best sample size. And factor analysis is appropriate for given data. Values below 0.5 imply that factor analysis may not be appropriate.

Bartlett's test of sphericity:

Bartlett's test of sphericity is a test statistic used to examine the hypothesis that the variables are uncorrelated in the population.

Communality:

Communality is the amount of variance a variable share with all the other variables being considered. This is also the proportion of variance explained by the common factors.

Correlation matrix:

A correlation matrix is a lower triangle matrix showing the simple correlations, r, between all possible pairs of variables included in the analysis. The diagonal elements, which are all 1, are usually omitted.

Eigen value:

The eigenvalue represents the total variance explained by each factor.

Factor loadings:

Factor loadings are simple correlations between the variables and the factors. Factor loadings indicate how much a factor explains a variable. Loadings can range from -1 to 1.

Factor loading plot:

A factor loading plot is a plot of the original variables using the factor loadings as coordinates.

Factor matrix:

A factor matrix contains the factor loadings of all the variables on all the factors extracted.

Factor scores:

Factor scores are composite scores estimated for each respondent on the derived factors.

Scree plot:

A scree plot is a plot of the Eigen-values against the number of factors in order of extraction.

Percentage of variance:

The percentage of the total variance attributed to each factor. In this approach the number of factors extracted is determined so that the cumulative percentage of variance extracted by the factors reaches a satisfactory level

However, it is recommended that the factors extracted should account for at least 60% of the variance.

Let's perform Factor Analysis

5.2 Factor Analysis on students rating about themselves

In this chapter, we use factor analysis on the data of rating of students to themselves in given 10 factors. That is how good they are in the following factors is used in this factor analysis.
Factors are:
F1: Marathi communication.
F2: English communication.
F3: Hindi communication.
F4: Critical thinking.
F5: Team work
F6: Problem solving.
F7: Communication skills.
F8: Technical skills.
F9: Creativity.
F10: Leadership.

Before you perform factor analysis, you need to evaluate the **"factorability"** of our data set. Factorability means "we an find the factors in the data set which influence more.That is in the study a sample of size 523 respondents were taken but it is needed to be tested whether

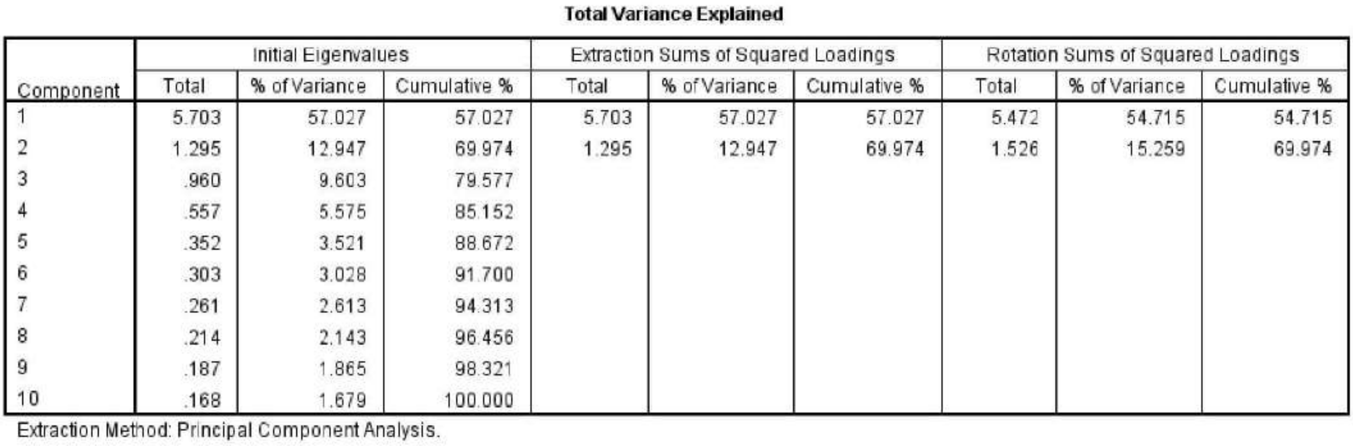
it is adequate or not for the factor analysis. To check this, there are two methods to check the factor ability or sampling adequacy.
1. Bartlett's Test
2. Kaiser-Meyer-Olkin Test.

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		*0.932*
3**Bartlett's Test Sphericity*	*Approx. Chi-Square*	*3733.914*
	df	*45*
	Sig.	*0.000*

KMO estimates the proportion of variance among all the observed variable.Lower proportion id more suitable for factor analysis. KMO values range between 0 and 1. Value of KMO less than 0.6 is considered inadequate.

Conclusions: In the above table, as the value of KMO is 0.932, it shows that the sample size is adequate and the factor analysis is appropriate for the given data.
Under the Bartlett's test we test the hypothesis that:
H_0 : The variables are uncorrelated in the population.
H_1: The variables are correlated in the population.
Here value of test statistics is large which is favour the rejection of null hypothesis.

So, the variables are correlated in the population.
It means that our data is adequate for factor analysis.
The below table shows that the Eigen values for a factor which indicates the total variance attributed to the factor. Here first factor accounts 69.974% of total variance accounted by the 10 variables.



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.703	57.027	57.027	5.703	57.027	57.027	5.472	54.715	54.715
2	1.295	12.947	69.974	1.295	12.947	69.974	1.526	15.259	69.974
3	.960	9.603	79.577						
4	.557	5.575	85.152						
5	.352	3.521	88.672						
6	.303	3.028	91.700						
7	.261	2.613	94.313						
8	.214	2.143	96.456						
9	.187	1.865	98.321						
10	.168	1.679	100.000						

Extraction Method: Principal Component Analysis.

Figure 5.1: Total variance explained table

Conclusions: The above table lists the eigenvalues associated with each factors before extraction, after extraction and after rotation. Before extraction, SPSS has identified 10 factors within the data set. The eigenvalues associated with each factor represents the variance explained by that particular factor. Also, the eigenvalues are displays in terms of the percentage of variance explained. So factor 1 explains 57.02% of total variance. It should be clear that the first two factors explains relatively large amount of variance whereas subsequent factors explains only small amounts of variance. SPSS then extracts all factors with eigenvalues greater than 1, which leaves us with 2 factors. The eigenvalues associated with these factors are again displayed in extracted sums of squared loadings table.The values in this part of the table are same as the values before extraction. In the final part of the table(rotated sum of square loadings), the eigenvalues of the factors after rotation displayed. Rotation has the effect of optimizing the factor structure and one consequence for these data is that the relative importance of the four factors is equalized. Before rotation factor 1 accounted for considerably more variance than the remaining one factor(57.027%), howeve after extraction it accounts for only 54.715%. Also, we can conclude that about 69.974% of variation in the data is explained by the 4 factors only.

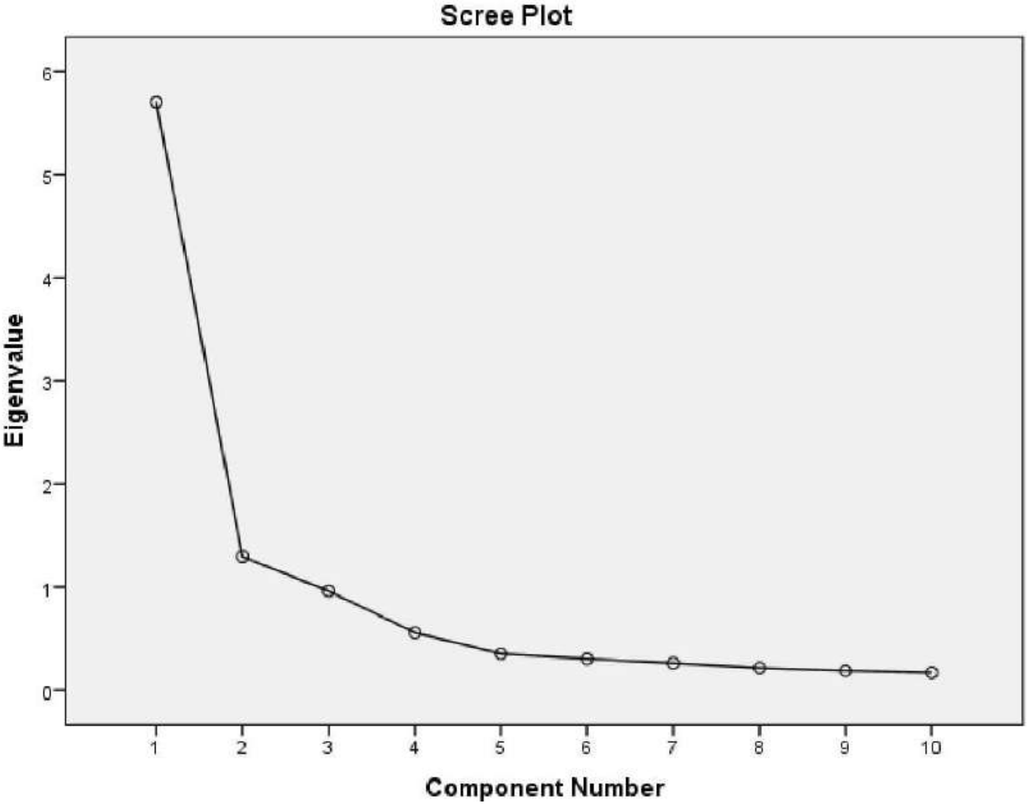
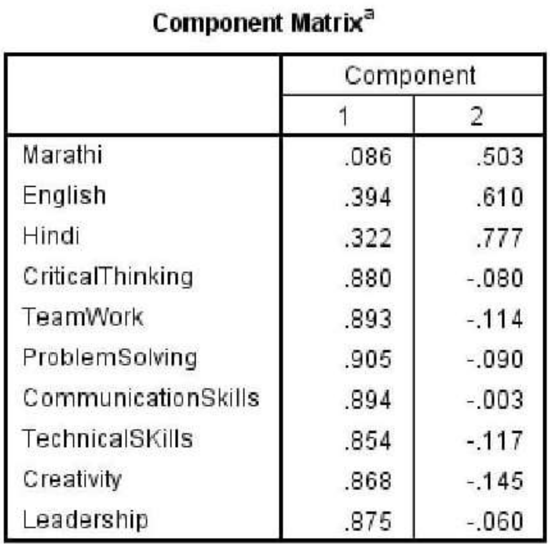


Figure 5.2: Scree Plot

Conclusion: This scree plot shows that the first two factors account for most of the total variability in data (given by the eigenvalues). The eigenvalues for the first two factors are all greater than 1.The remaining factors account for a very small proportion of the variability and are likely unimportant.

Component Matrix

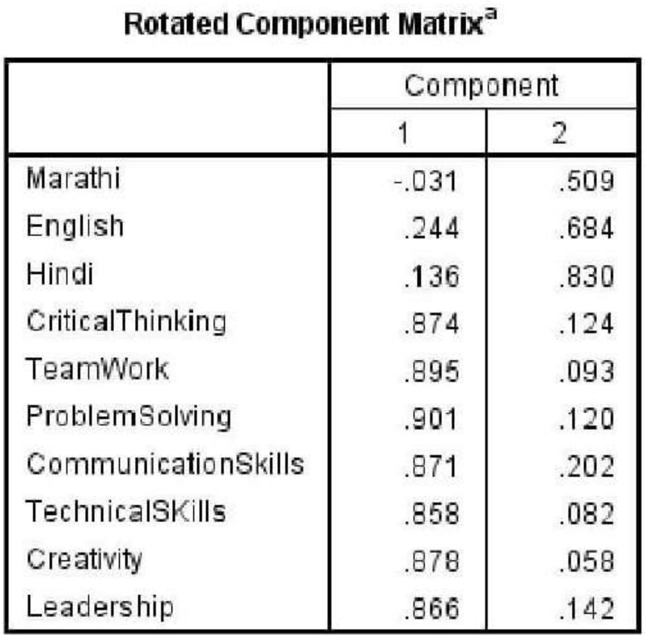


Component Matrix[a]

	Component	
	1	2
Marathi	.086	.503
English	.394	.610
Hindi	.322	.777
CriticalThinking	.880	-.080
TeamWork	.893	-.114
ProblemSolving	.905	-.090
CommunicationSkills	.894	-.003
TechnicalSKills	.854	-.117
Creativity	.868	-.145
Leadership	.875	-.060

Figure 5.3: Component matrix table

Extraction Method: Principal Component Analysis.
a. 2 components extracted.

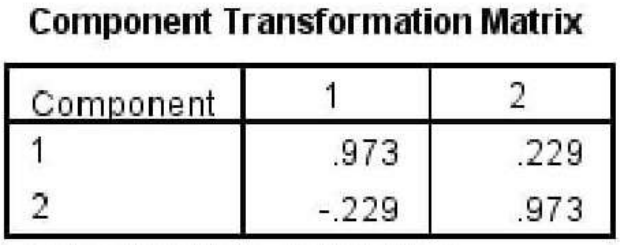


Rotated Component Matrix[a]

	Component	
	1	2
Marathi	-.031	.509
English	.244	.684
Hindi	.136	.830
CriticalThinking	.874	.124
TeamWork	.895	.093
ProblemSolving	.901	.120
CommunicationSkills	.871	.202
TechnicalSKills	.858	.082
Creativity	.878	.058
Leadership	.866	.142

Figure 5.4: Rotated component matrix

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 3 iterations



Component Transformation Matrix

Component	1	2
1	.973	.229
2	-.229	.973

Figure 5.5: Component transformation matrix

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

Conclusion: From the results of factor analysis, we can conclude that the factors affecting for placement which are given in data description (F1-F10) can be classified into two components. Component 1 Includes factors such as Hindi(0.830) and English (0.684) (language rating about themselves which can be named as "language related" which has accounted for the 54.715% of the data. Component 2 Includes factors such as problem solving (0.901) and team work (0.895) which can be named as the "Skills of the students" which has accounted the 12.947% of information.

Components	Factors	Explained variation
Communication Language	Hindi+English	54.715%
Skills of the Students	Problem solving+Team-work	12.947%

5.3 Factor analysis on students view about factors required for placement

In this section, we use factor analysis on the data of students view about factors required for placement in given 20 factors. That is how good they should be the following factors for placement.

Factors are:

F1: Communication skills.

F2: Confidence.

F3: Critical thinking.

F4: Technical skills.

F5: Soft skills.

F6: Literacy skills.

F7: Team work.

F8: Sincerity.

F9: Truth fullness

F10: Medium of education.

F11: Experience.

F12: Coding programming.

F13: Geographical location of the candidate.

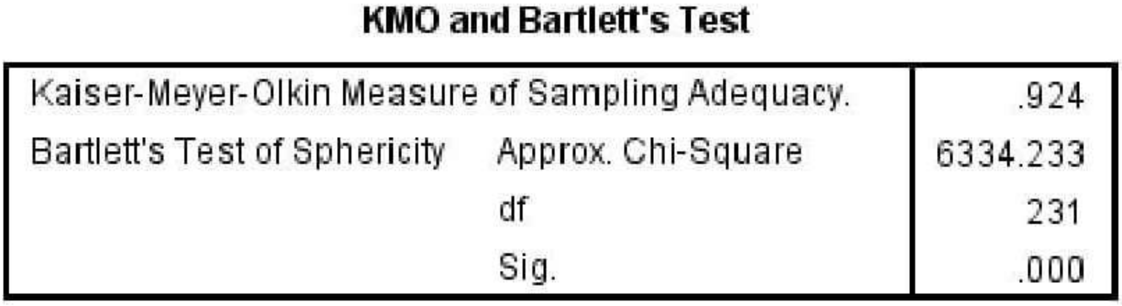
F14: Teaching quality of teachers.

F15: Course design.

F16: Recruitment campaign.

F17: Opportunities getting to the candidate.

F18: Mock interviews.
F19: Financial Background.
F20: Group discussion.



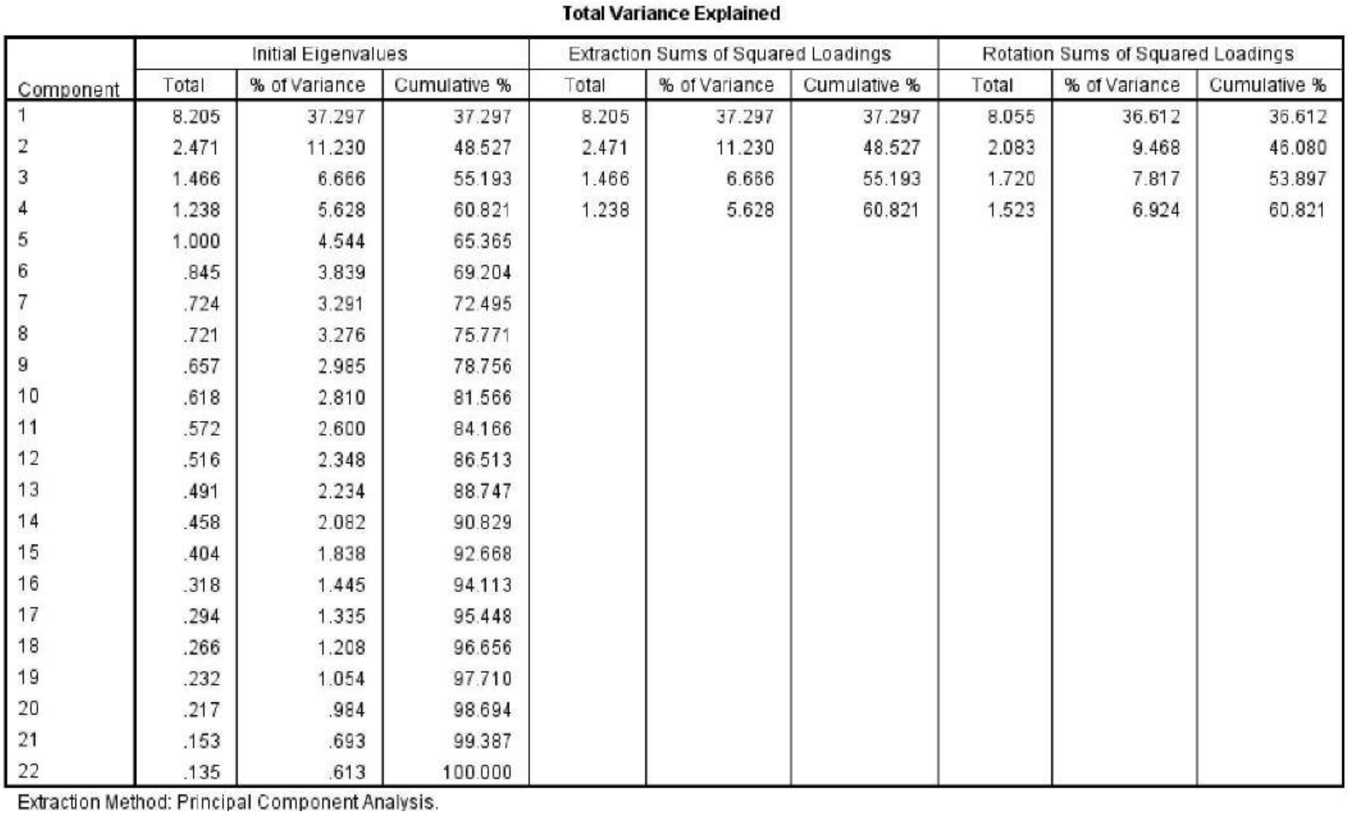
KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.924
Bartlett's Test of Sphericity	Approx. Chi-Square	6334.233
	df	231
	Sig.	.000

Figure 5.6: KMO and Bartlett's test table

From above table we observe that, the data is adequate for factor analysis as KMO value is 0.924 which is greater than 0.6. Also, the value of Bartlett's test of Sphericity is 6334.233 which indicates that the variables in population are correlated.

From this we say that, our data is adequate for Factor analysis.



Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	8.205	37.297	37.297	8.205	37.297	37.297	8.055	36.612	36.612
2	2.471	11.230	48.527	2.471	11.230	48.527	2.083	9.468	46.080
3	1.466	6.666	55.193	1.466	6.666	55.193	1.720	7.817	53.897
4	1.238	5.628	60.821	1.238	5.628	60.821	1.523	6.924	60.821
5	1.000	4.544	65.365						
6	.845	3.839	69.204						
7	.724	3.291	72.495						
8	.721	3.276	75.771						
9	.657	2.985	78.756						
10	.618	2.810	81.566						
11	.572	2.600	84.166						
12	.516	2.348	86.513						
13	.491	2.234	88.747						
14	.458	2.082	90.829						
15	.404	1.838	92.668						
16	.318	1.445	94.113						
17	.294	1.335	95.448						
18	.266	1.208	96.656						
19	.232	1.054	97.710						
20	.217	.984	98.694						
21	.153	.693	99.387						
22	.135	.613	100.000						

Extraction Method: Principal Component Analysis.

Figure 5.7: Total variance explained table

Conclusions: The above table lists the eigenvalues associated with each factors before extraction, after extraction and after rotation. Before extraction, SPSS has identified 22 factors within the data set. The eigenvalues associated with each factor represents the variance explained by that particular factor. Also, the eigenvalues are displays in terms of the percentage of variance explained. So factor 1 explains 37.297% of total variance. It should be clear that the first four factors explains relatively large amount of variance whereas subsequent factors explains only small amounts of variance. SPSS then extracts all factors with eigenvalues greater than 1, which leaves us with 4 factors. The eigenvalues associated with these factors are again displayed in extracted sums of squared loadings table.The values in this part of the table are same as the values before extraction. In the final part of the table(rotated sum of square loadings), the eigenvalues of the factors after rotation displayed. Rotation has the effect of optimizing the factor structure and one consequence for these data is that the relative importance of the four factors is equalized. Before rotation factor 1 accounted for considerably more variance than the remaining one factor(37.297%), howeve after extraction it accounts for only 36.612%. Also, we can conclude that about 60.82% of variation in the data is explained by the 4 factors only.

From the below scree plot also we can see that the first 4 components explains more than half of the variation in the data

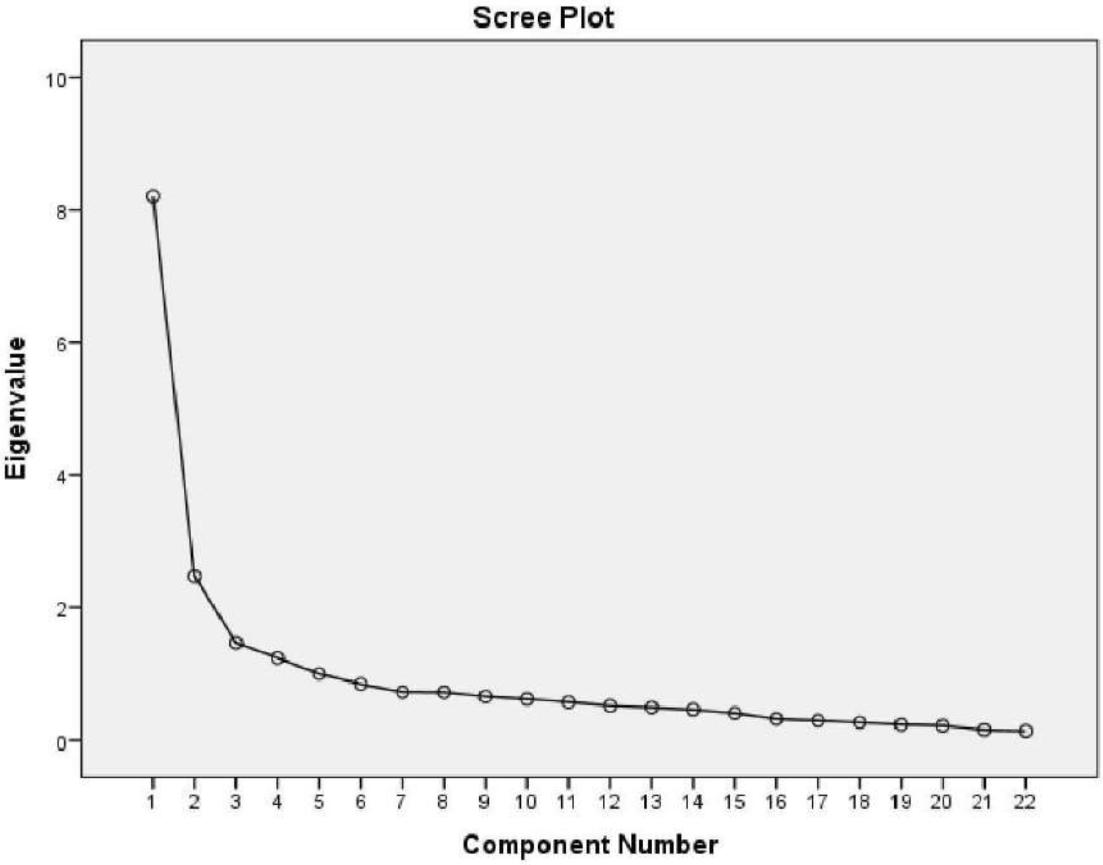
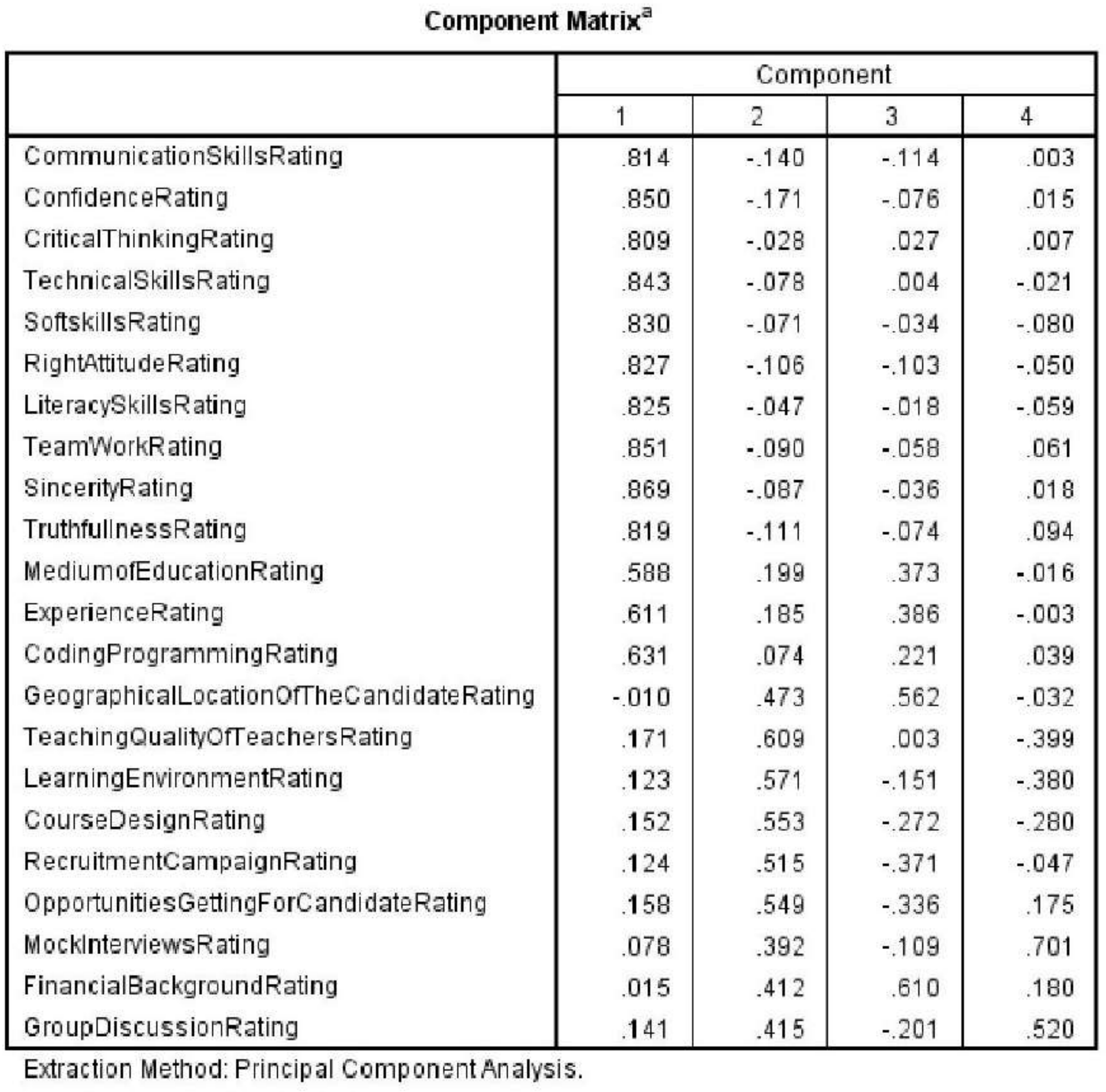


Figure 5.8: Scree Plot

Conclusion: This scree plot shows that the first four factors account for most of the total variability in data (given by the eigenvalues). The eigenvalues for the first four factors are all greater than 1.The remaining factors account for a very small proportion of the variability and are likely unimportant .



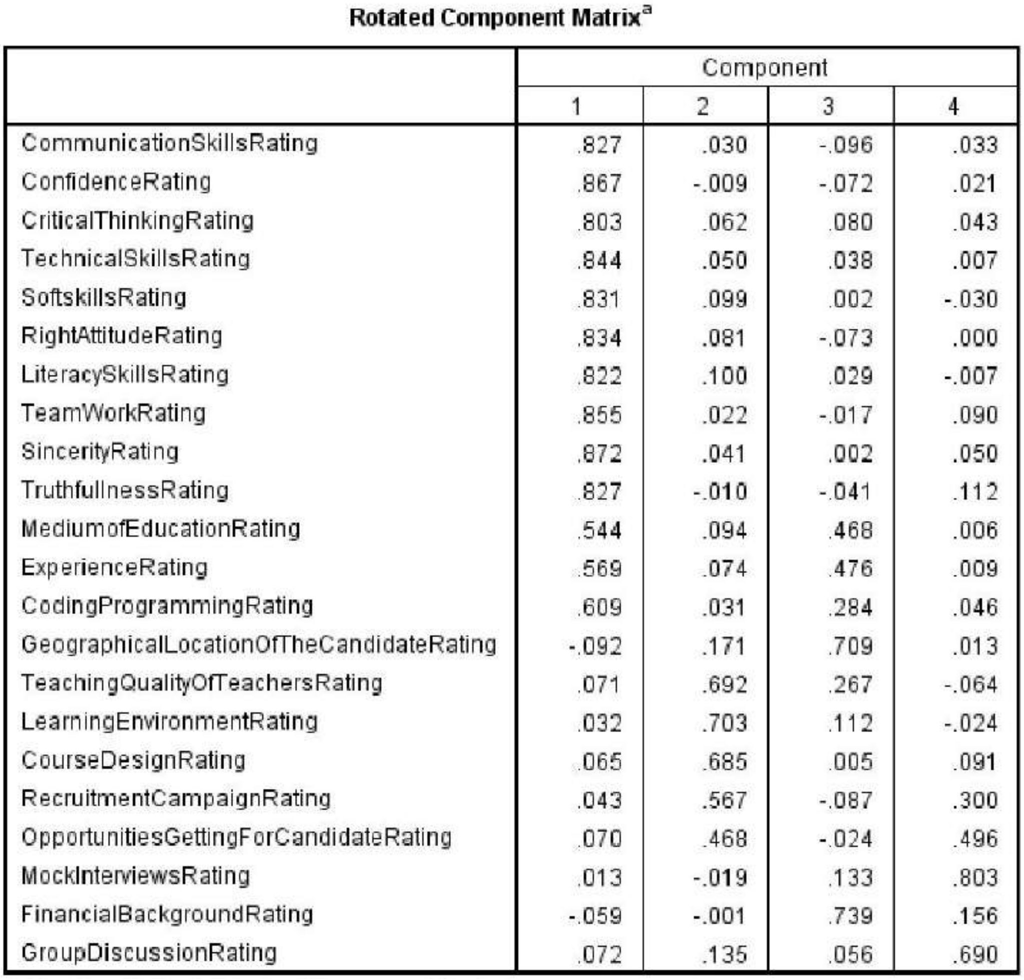
Component Matrix[a]

	Component			
	1	2	3	4
CommunicationSkillsRating	.814	-.140	-.114	.003
ConfidenceRating	.850	-.171	-.076	.015
CriticalThinkingRating	.809	-.028	.027	.007
TechnicalSkillsRating	.843	-.078	.004	-.021
SoftskillsRating	.830	-.071	-.034	-.080
RightAttitudeRating	.827	-.106	-.103	-.050
LiteracySkillsRating	.825	-.047	-.018	-.059
TeamWorkRating	.851	-.090	-.058	.061
SincerityRating	.869	-.087	-.036	.018
TruthfullnessRating	.819	-.111	-.074	.094
MediumofEducationRating	.588	.199	.373	-.016
ExperienceRating	.611	.185	.386	-.003
CodingProgrammingRating	.631	.074	.221	.039
GeographicalLocationOfTheCandidateRating	-.010	.473	.562	-.032
TeachingQualityOfTeachersRating	.171	.609	.003	-.399
LearningEnvironmentRating	.123	.571	-.151	-.380
CourseDesignRating	.152	.553	-.272	-.280
RecruitmentCampaignRating	.124	.515	-.371	-.047
OpportunitiesGettingForCandidateRating	.158	.549	-.336	.175
MockInterviewsRating	.078	.392	-.109	.701
FinancialBackgroundRating	.015	.412	.610	.180
GroupDiscussionRating	.141	.415	-.201	.520

Extraction Method: Principal Component Analysis.

Figure 5.9: Component matrix

a. 4 components extracted.

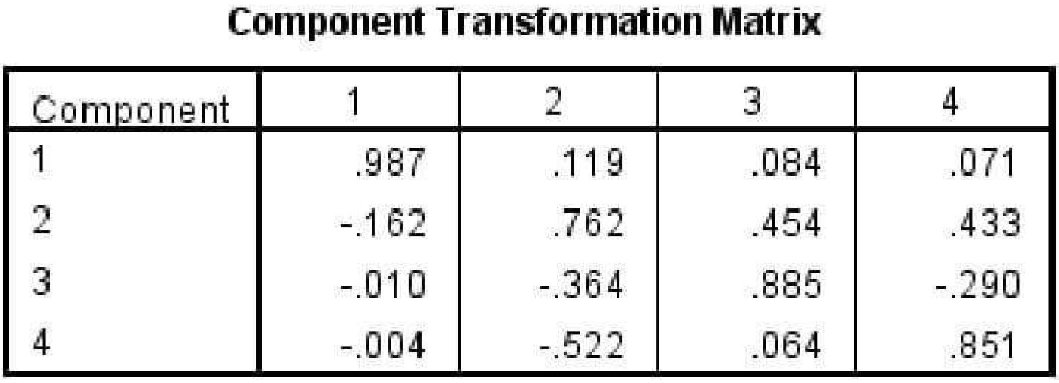


Rotated Component Matrix[a]

	Component			
	1	2	3	4
CommunicationSkillsRating	.827	.030	-.096	.033
ConfidenceRating	.867	-.009	-.072	.021
CriticalThinkingRating	.803	.062	.080	.043
TechnicalSkillsRating	.844	.050	.038	.007
SoftskillsRating	.831	.099	.002	-.030
RightAttitudeRating	.834	.081	-.073	.000
LiteracySkillsRating	.822	.100	.029	-.007
TeamWorkRating	.855	.022	-.017	.090
SincerityRating	.872	.041	.002	.050
TruthfullnessRating	.827	-.010	-.041	.112
MediumofEducationRating	.544	.094	.468	.006
ExperienceRating	.569	.074	.476	.009
CodingProgrammingRating	.609	.031	.284	.046
GeographicalLocationOfTheCandidateRating	-.092	.171	.709	.013
TeachingQualityOfTeachersRating	.071	.692	.267	-.064
LearningEnvironmentRating	.032	.703	.112	-.024
CourseDesignRating	.065	.685	.005	.091
RecruitmentCampaignRating	.043	.567	-.087	.300
OpportunitiesGettingForCandidateRating	.070	.468	-.024	.496
MockInterviewsRating	.013	-.019	.133	.803
FinancialBackgroundRating	-.059	-.001	.739	.156
GroupDiscussionRating	.072	.135	.056	.690

Figure 5.10: Rotated component matrix

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 5 iterations.



Component Transformation Matrix

Component	1	2	3	4
1	.987	.119	.084	.071
2	-.162	.762	.454	.433
3	-.010	-.364	.885	-.290
4	-.004	-.522	.064	.851

Figure 5.11: Component transformation matrix

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
Conclusions: From the results of factor analysis, we can conclude that the factors affecting on placement which are given in data description (F1-F20) can be classified into four components. Component 1 Includes factors such as confidence (0.867), sincerity (0.72) which can be named as "soft skills". Which has accounted for the 36.612% in the data component 2 Includes factor such as teaching quality of teachers(0.692) and leaning environment (0.703) which has accounted for 9.468% in the data. Component 3 Includes factor such as geographical location of the students (0.562), experience (0.86) which has accounted 7.87% in the data. And component 4 includes group disccusion which has accounted 6.924% in the data. All they total explaines 60.82% variation in the data.

Components	Factors	Explained variation
1	Confidence+Sincerity	36.612%
2	Teaching Quality+Learning Environment	9.468%
3	Geographical location+Experience	7.87%
4	Group Discussion	6.924%

Chapter 6

Classification of Data

6.1 What is classification of data?

Classification is a process of categorizing a given set of data into classes. It can be performed on both structured or unstructured data. The process starts with predicting the class of given data points. The classes are often referred to as target, label or categories.
Where is classification used?
One of the most common uses of classification is filtering emails into "spam" or "non-spam." In short, classification is a form of "pattern recognition," with classification algorithms applied to the training data to find the same pattern (similar words or sentiments, number sequences, etc.) in future sets of data.
In this chapter we use different classification techniques for classification of placed and non-placed students. That is we used it for the purpose of prediction. Whether the student get placed or not.

6.2 Classification of students whether they will be placed or not?

Let's first see the different classification algorithms in python.

1. Decision tree classifier
2. Random forest classifier
3. K-Nearest Neighbors(KNN) classifier
4. Support vector classifier
5. Logistic classifier

Here we apply different types of classification algorithms to our data set and we see the accuracy score of the classification model among them the model whose accuracy score is maximum we choose this model for the classification purpose.

First of all let us see the different classification techniques as follows:

1. Decision Tree classifier

In general, Decision tree analysis is a predictive modelling tool that can be applied to across many areas. Decision tree can be contrusted by an algorithmic approch that can split the dataset in different ways based on different conditions. Decision trees are the most powerful algorithm that falls under the category of supervised algorithms.

They can be used for both classification and regression tasks. The two main entities of a tree are decision nodes, where the data is split and leaves, where we got outcome. In classification kind of decision tree, the decision variable is categorical. Decision tree classification prefers the features values of be categorical.

Decision tree is a flowchart like a structure in which each internal node represents a “test” on an attribute and each branch represents the outcome of the test, the each leaf node represents a class label.

2. Random Forest Classifier

Random forest is an supervised learning algorithm. Random forest classifier creates a set of decision tress from randomly selected subset of training set. It aggregates the votes from different decision trees to decides the final class of the test object.

The basic parameter of the Random Forest Classifier are total number of trees to be generated and decision tree related parameters like minimum split, split criteria. It can be used for both classification and regression problem.

3. K-Nearest Neighbors (KNN) Classifier

K-Nearest is an supervised learning algorithm. This algorithms are use classification as well as regression problem. In K-nearest neighbors method, the classifier identifies K observations in the training dataset that are similar to a new record that we wish to classify. KNN algorithm is non-parametric method. In KNN algorithm nearest distance by using Euclidean distance formula.

4. A Support Vector Classifier

A support vector machine is an algorithm for the classification of bothe linear and non-linear data. It transforms the original data in a higher dimension, from where it can be find a hyperplane for the separation of the data by classi using essential training tuples which are called as support vectors. It uses a nonlinear mapping to transform the original training data into higher dimenstions. We use the linear kernel to fitting the model.

The SVM searches the hyperplane with the largest margin that means the one with maximum distance between the nearest training tuples. The associated margin gives the largest separation between classes. An SVM with small number of support vectors can have good generalization.

5. Logistic Regression Classifier

When in the data there are quantitative regressors and qualitative response in that case we are using the logistic regression to build a model of prediction. In our dataset the reponse variable placed or non-placed is a categorical type so in this situation one can use logistic regression.

It is also possible that the dependent variable be a dichotomous variable. That is the output is as pass or fail, yes or no, success or failure etc.In this situation we commonly use a statistical technique like linear regression called logistic regression. Logistic regression shares the same objective as linear regression. Logistic and linear regression are statistical methods used to predict the dependent variable based on one or more independent variables. They both produce a regression model to summarize the relationship between these variables, as well as other Statistics to describe how well the model fits the data.

Logistic Regression Assumptions

1. Binary logistic regression requires the dependent variable to be binary.
2. For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.
3. Only the meaningful variables should be included.
4. The independent variables should be independent of each other. That is, the model should have little or no multicollinearity.
5. The independent variables are linearly related to the log odds.
6. Logistic regression requires quite large sample sizes. Keeping the above assumptions in mind, let's look at our dataset.

The process of modelling the data

1. Importing the model.
2. Fitting the model.
3. Predicting the output observations.
4. Classification metrics.

Score matrics for classification

Accuracy

Accuracy is the matric we plot after doing classification to see that how good our model is. It is pretty eassy to understand from accuracy matrix about the model fitting.

Accuracy matrix is the ratio of the ratio of correct predictions to total predictions made.It is often presented as a percentage by multiplying the result by 100.

```
Accuracy=(TP+TN)/(TP+TN+FP+FN)
```

Where,

TP: True Positive

TN: True Negative

FP: False Positive

TN: True Negative

Precision

Precision shows us that what proportion of **predicted positive** is truly positive.

```
Precision=TP/(TP+TF)
```

Recall

Recall shows us that what proportion of **actual positive** is correctly clssified.

```
Recall=TP/(TP+FN)
```

f1 Score

f1 score is the number between 0 and 1, the harmonic mean of precision and recall.

```
f1=2*(Precision*recall)/(Presicion+recall)
```

ROC Curve

A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. The method was originally developed for operators of military radar receivers starting in 1941, which led to its name.

What ROC curve explains?

An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate.

In general, an AUC of 0.5 suggests no discrimination (i.e., ability to diagnose patients with and without the disease or condition based on the test), 0.7 to 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and more than 0.9 is considered outstanding.

How is a ROC curve generated?

The ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds. For example, in logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class.

Classifiers that give curves closer to the top-left corner indicate a better performance. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

6.3 Implementation of classification algorithms in Python

Feature Selection

Using the following features
1. Gender
2. Age
3. Hometown
4. School Type
5. SSC Percentage

6. HSC Percentage
5. SSC Board
7. HSC Board
8. Degree percentage
9. Highest Degree
10. Stream of education
11. Specialization
12. Communication skill in Marathi
13. Communication skill in English
14. Communication skill in Hindi
15. Technical skills 16. Creativity
17. Leadership

Importing required libraries

[99]:
```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
pd.set_option('display.max_columns', None)
pd.set_option('display.max_rows', None)
plt.rc("font", size=14)
sns.set(style="white")
sns.set(style="whitegrid", color_codes=True)
from sklearn import preprocessing
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from xgboost import XGBClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve,auc
from sklearn.metrics import confusion_matrix,accuracy_score
```

Importing dataset

[100]:
```
df2=pd.read_csv('Data.csv')
```

[101]:
```
print("Number of rows in data :",df2.shape[0])
print("Number of columns in data :", df2.shape[1])
```

```
Number of rows in data : 523
Number of columns in data : 23
```

Defining X (regressors) and y (response) variables.

[105]:
```
X = df2.loc[:, df2.columns != 'Placed?']
y = df2.loc[:, df2.columns == 'Placed?']
```

```
[108]: label_encoder = preprocessing.LabelEncoder()
       df2['Placed?']= label_encoder.fit_transform(df2['Placed?'])
```

```
[109]: y = df2.loc[:, df2.columns == 'Placed?']
```

```
[111]: sum(df2.duplicated())
```

```
[112]: X.dtypes
```

```
[112]: Gender                    object
       Age                        int64
        Hometown                  object
       School Type               object
       SSC percentage           float64
       HSC percentage           float64
       SSC Board                 object
       HSC Board                 object
       Degree Percentage        float64
       Highest Degree            object
        Stream Of Education      object
       Specialization            object
       Marathi                   object
       English                   object
       Hindi                     object
       Critical Thinking          int64
       Team Work                  int64
       Problem Solving            int64
       Communication Skills       int64
       Technical SKills           int64
       Creativity                 int64
       Leadership                 int64
       dtype: object
```

Here , the data type of Some features is object so first we want to convert it into float or integer by doing label encoding.

So, let's do label encoding first.

Label Encoding

```
[113]: l=[  'Gender',' Hometown', 'School Type ','SSC Board','HSC␣
         ↪Board','Highest Degree',' Stream Of Education ',
           'Specialization ','Marathi','English','Hindi'  ]
```

```
[114]: # label_encoder object knows how to understand word labels.
       label_encoder = preprocessing.LabelEncoder()
       for i in l:
           X[i]= label_encoder.fit_transform(X[i])
           X[i].unique()
```

[115]:

```
X.head(3)
```

[115]:

```
   Gender  Age   Hometown  School Type   SSC percentage  HSC percentage␣
 ↪ \
0       1   30          0            1            88.8            47.0
1       0   25          2            0            60.0            65.0
2       0   22          0            0            74.8            65.0

   SSC Board  HSC Board  Degree Percentage   Highest Degree  \
0          2          2              69.00                0
1          2          2              78.00               12
2          2          2              72.05               12

    Stream Of Education   Specialization   Marathi  English  Hindi  \
0                     3               45         0        0      0
1                     3               83         0        1      2
2                     3               60         0        2      1

   Critical Thinking  Team Work  Problem Solving  Communication Skills ␣
 ↪\
0                  3          5                4                     5
1                  2          3                2                     1
2                  1          1                1                     1

   Technical SKills  Creativity  Leadership
0                 4           4           4
1                 2           3           2
2                 4           1           1
```

[116]:

```
X.dtypes
```

[116]:

```
Gender                     int32
Age                        int64
 Hometown                  int32
School Type                int32
SSC percentage           float64
HSC percentage           float64
SSC Board                  int32
HSC Board                  int32
Degree Percentage        float64
Highest Degree             int32
 Stream Of Education       int32
Specialization             int32
Marathi                    int32
English                    int32
Hindi                      int32
Critical Thinking          int64
Team Work                  int64
```

```
Problem Solving            int64
Communication Skills       int64
Technical SKills           int64
Creativity                 int64
Leadership                 int64
dtype: object
```

Now our data is ready for classification.

1.Splitting the dataset into train and test

```
[117]: #Train Test Split
       X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3)
```

```
[118]: from sklearn.metrics import confusion_matrix,accuracy_score
```

1.Decision Tree clssification

```
[119]: dtree = DecisionTreeClassifier(criterion='entropy')
       dtree.fit(X_train, y_train)
       y_pred = dtree.predict(X_test)
```

```
[120]: a=accuracy_score(y_test, y_pred)
       a
```

```
[120]: 0.7006369426751592
```

```
[121]: confusion_matrix = confusion_matrix(y_test, y_pred)
```

```
[122]: print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.76      0.82      0.79       107
           1       0.54      0.44      0.48        50

    accuracy                           0.70       157
   macro avg       0.65      0.63      0.64       157
weighted avg       0.69      0.70      0.69       157
```

```
[123]: conf_mat = pd.DataFrame(confusion_matrix)
       fig = plt.figure(figsize=(10,7))
       sns.heatmap(conf_mat, annot=True, annot_kws={"size": 16}, fmt='g')
       plt.title("Confusion Matrix for Decision Tree Algorithm")
       plt.xlabel("Predicted Label")
       plt.ylabel("True Label")
       plt.show()
```

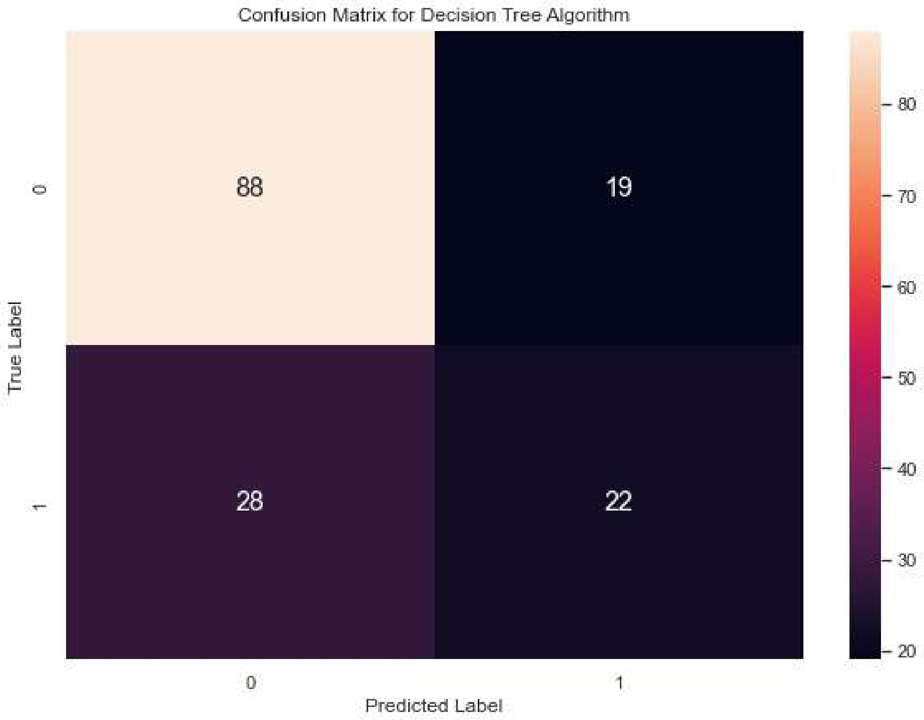


Figure 6.1: Heatmap for confusion matrix in Decision Tree classifier

Decision Tress algorithm gives us 70.06% **accuracy**

2.Randoam Forest (RF) Classification

[124]:
```
#Using Random Forest Algorithm
random_forest = RandomForestClassifier(n_estimators=100)
random_forest.fit(X_train, y_train)
y_pred = random_forest.predict(X_test)
```

[125]:
```
from sklearn.metrics import confusion_matrix,accuracy_score
confusion_matrix = confusion_matrix(y_test, y_pred)
accuracy_score= accuracy_score(y_test,y_pred)
```

[126]:
```
b=accuracy_score
b
```

[126]:
```
0.6942675159235668
```

[127]:
```
confusion_matrix
```

[127]:
```
array([[100,   7],
       [ 41,   9]], dtype=int64)
```

[128]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support
```

```
           0       0.71      0.93      0.81       107
           1       0.56      0.18      0.27        50

    accuracy                           0.69       157
   macro avg       0.64      0.56      0.54       157
weighted avg       0.66      0.69      0.64       157
```

[129]:

```
conf_mat = pd.DataFrame(confusion_matrix )
fig = plt.figure(figsize=(10,7))
sns.heatmap(conf_mat, annot=True, annot_kws={"size": 16}, fmt='g')
plt.title("Confusion Matrix  for Random Forest Algorithm")
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.show()
```

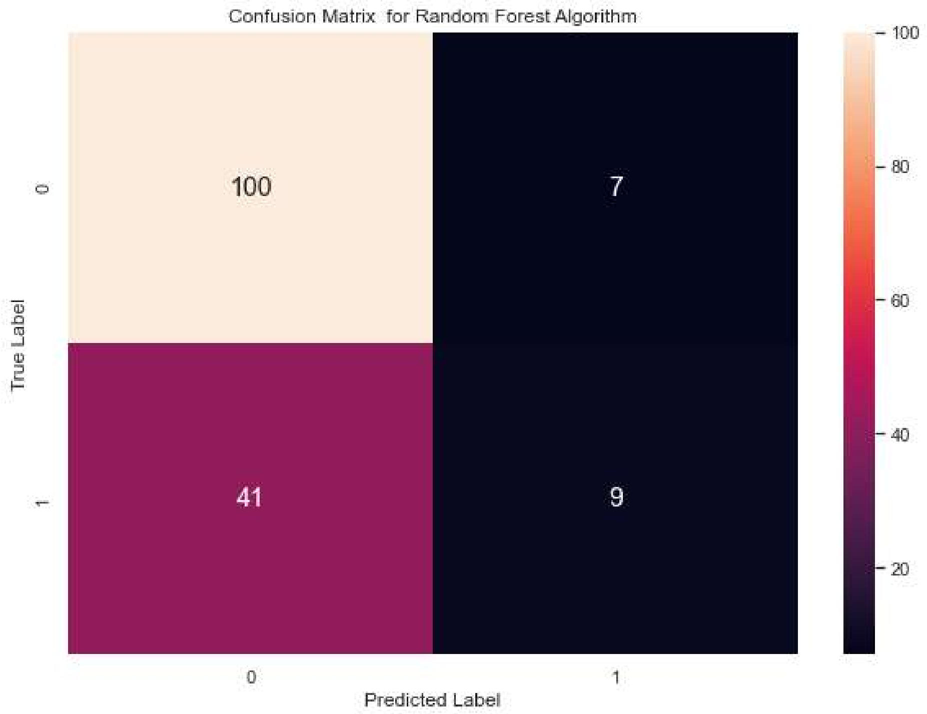


Figure 6.2: Heatmap for confusion matrix in RF Classifier.

Random Forest algorithm gives us 69.42% **accuracy**

3.K-Nearest Neighbors(KNN) Classification

[130]:

```
from sklearn.neighbors import KNeighborsClassifier
classifier = KNeighborsClassifier(n_neighbors = 5, metric =␣
  ↪'minkowski', p = 2)
classifier.fit(X_train, y_train)
```

```
[130]: KNeighborsClassifier()
```

```
[131]: from sklearn.metrics import confusion_matrix,accuracy_score
        confusion_matrix = confusion_matrix(y_test, y_pred)
        accuracy_score= accuracy_score(y_test,y_pred)
```

```
[132]: c=accuracy_score
        c
```

```
[132]: 0.6942675159235668
```

```
[133]: confusion_matrix
```

```
[133]: array([[100,   7],
               [ 41,   9]], dtype=int64)
```

```
[134]: print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.71      0.93      0.81       107
           1       0.56      0.18      0.27        50

    accuracy                           0.69       157
   macro avg       0.64      0.56      0.54       157
weighted avg       0.66      0.69      0.64       157
```

```
[135]: conf_mat = pd.DataFrame(confusion_matrix)
        fig = plt.figure(figsize=(10,7))
        sns.heatmap(conf_mat, annot=True, annot_kws={"size": 16}, fmt='g')
        plt.title("Confusion Matrix for KNN Algorithm ")
        plt.xlabel("Predicted Label")
        plt.ylabel("True Label")
        plt.show()
```

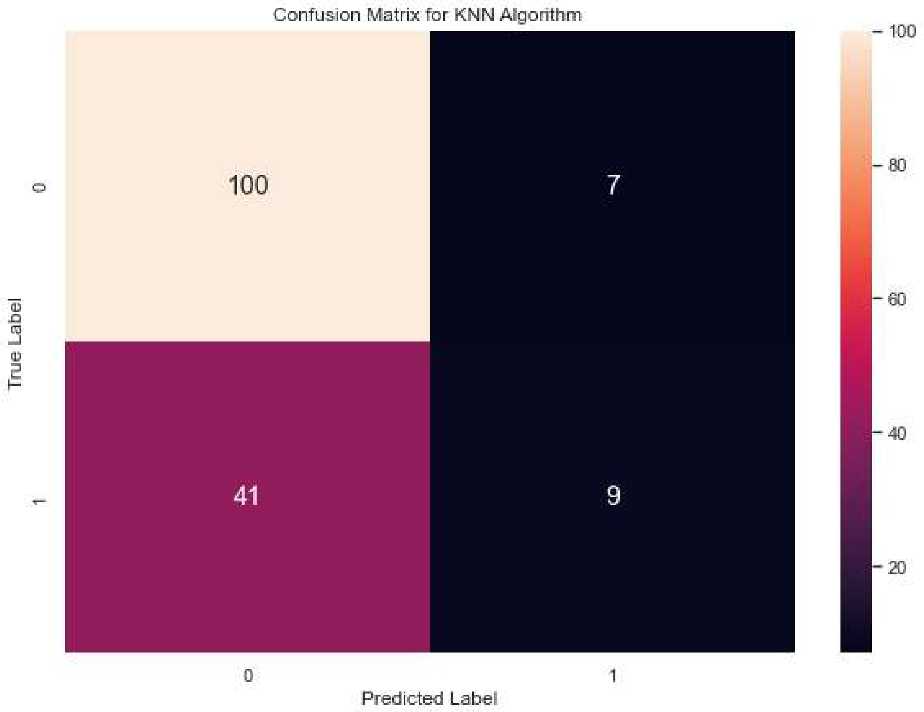


Figure 6.3: Heatmap for confusion matrix in KNN algorithm

KNN algorithm gives us 69.42% **accuracy**

4.A Support Vector Classification

```
[136]: #Import svm model
       from sklearn import svm
       #Create a svm Classifier
       clf = svm.SVC(kernel='linear') # Linear Kernel
       #Train the model using the training sets
       clf.fit(X_train, y_train)
       #Predict the response for test dataset
       y_pred = clf.predict(X_test)
```

```
[137]: from sklearn.metrics import confusion_matrix,accuracy_score
       confusion_matrix = confusion_matrix(y_test, y_pred)
       accuracy_score= accuracy_score(y_test,y_pred)
```

```
[138]: confusion_matrix
```

```
[138]: array([[99,  8],
              [41,  9]], dtype=int64)
```

```
[139]: d=accuracy_score
       d
```

```
[139]: 0.6878980891719745
```

[140]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.71      0.93      0.80       107
           1       0.53      0.18      0.27        50

    accuracy                           0.69       157
   macro avg       0.62      0.55      0.54       157
weighted avg       0.65      0.69      0.63       157
```

[141]:
```
conf_mat = pd.DataFrame(confusion_matrix)
fig = plt.figure(figsize=(10,7))
sns.heatmap(conf_mat, annot=True, annot_kws={"size": 16}, fmt='g')
plt.title("Confusion Matrix for SVM algorithm")
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.show()
```

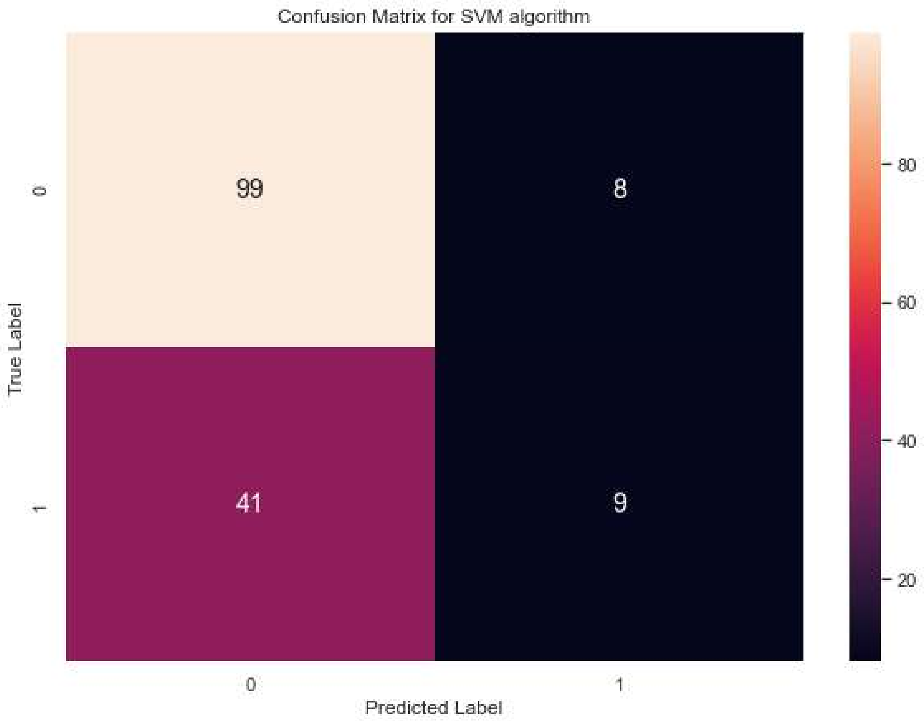


Figure 6.4: Heatmap for confusion matrix in SVM algorithm

SVM algorithm gives us 69.42% **accuracy**

5.Logistic Regression Classification

[142]:
```
# creating our model instance
log_reg = LogisticRegression()
```

```
# fitting the model
log_reg.fit(X_train, y_train)
```

```
[142]: LogisticRegression()
```

[143]:
```
# predicting the target vectors
y_pred=log_reg.predict(X_test)
y_pred
# creating confusion matrix heatmap
```

```
[143]: array([0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
              0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
              0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
              1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
              0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
              0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
              0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
              0, 1, 0])
```

[144]:
```
from sklearn.metrics import confusion_matrix,accuracy_score
cm = confusion_matrix(y_test, y_pred)
e= accuracy_score(y_test,y_pred)
e
```

```
[144]: 0.6624203821656051
```

[145]:
```
cm
```

```
[145]: array([[95, 12],
              [41,  9]], dtype=int64)
```

[146]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.70      0.89      0.78       107
           1       0.43      0.18      0.25        50

    accuracy                           0.66       157
   macro avg       0.56      0.53      0.52       157
weighted avg       0.61      0.66      0.61       157
```

[147]:
```
conf_mat = pd.DataFrame(cm)
fig = plt.figure(figsize=(10,7))
sns.heatmap(conf_mat, annot=True, annot_kws={"size": 16}, fmt='g')
plt.title("Confusion Matrix for Logistic Regression algorithm ")
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.show()
```

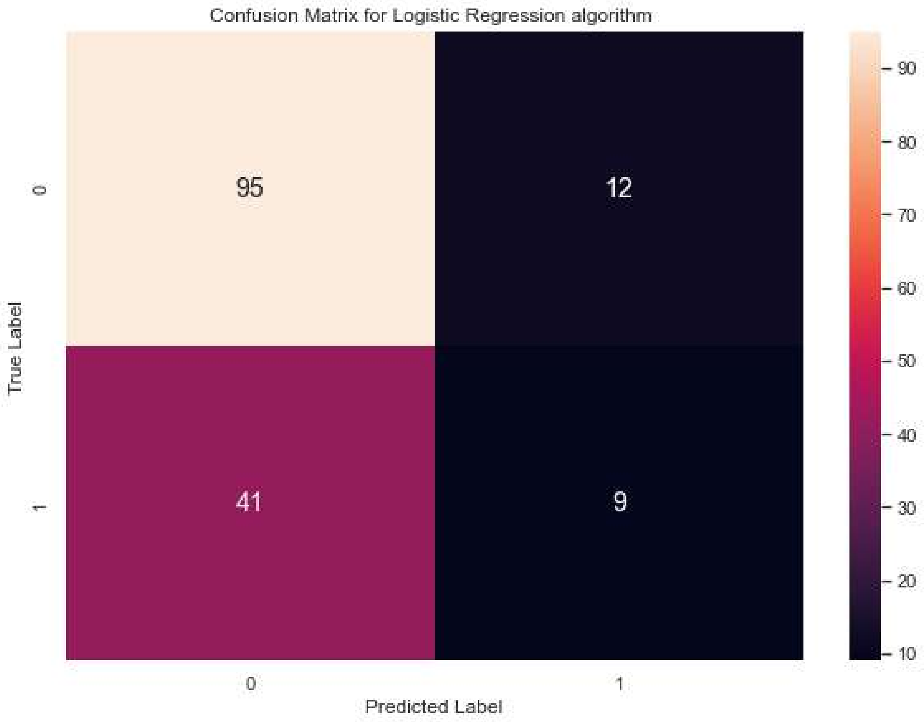


Figure 6.5: Heatmap for confusion matrix for logistic regression

Conclusion

Random Forest algorithm gives us 68.78% **accuracy** Our confusion Matrix looks decent. We have correctly predicted 95 (placed) + 9 (not-placed) correct predictions and 12 (not placed as placed) + 41(placed as not-placed) incorrect predictions.

We need to decrease these incorrect predictions because a good candidate can be rejected (false positive) [Type I error] and a unfit candidate can be selected (false negatives) [Type II Error].

```
[148]: print("Accuracy score of Decision tree algorithm is",a)
       print("Accuracy score of Random Forest algorithm is",b)
       print("Accuracy score of KNN algorithm is",c)
       print("Accuracy score of SVM algorithm is",d)
       print("Accuracy score of logistic algorithm is",e)
```

```
Accuracy score of Decision tree algorithm is 0.7006369426751592
Accuracy score of Random Forest algorithm is 0.6942675159235668
Accuracy score of KNN algorithm is 0.6942675159235668
Accuracy score of SVM algorithm is 0.6878980891719745
Accuracy score of logistic algorithm is 0.6624203821656051
```

```
[149]: label_encoder = preprocessing.LabelEncoder()
       y_test = label_encoder.fit_transform(y_test)
```

```
[150]: y_test=y_test.reshape(len(y_test),1)
```

```
[151]: label_encoder = preprocessing.LabelEncoder()
       y_pred = label_encoder.fit_transform(y_pred)
       y_pred
```

```
[151]: array([0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
              0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1,
              0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
              1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
              0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
              0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
              0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
              0, 1, 0], dtype=int64)
```

```
[152]: log_reg.predict_proba(X_test)[:,1]
```

```
[152]: array([0.09655022, 0.22436638, 0.36155369, 0.11380713, 0.43188201,
              0.3941052 , 0.23906786, 0.41878121, 0.83545712, 0.37664972,
              0.09398466, 0.09446159, 0.20000055, 0.41020177, 0.06148275,
              0.28616722, 0.2500308 , 0.14331831, 0.20289615, 0.2065301 ,
              0.34801143, 0.09348149, 0.32113982, 0.33340479, 0.28927221,
              0.10978924, 0.06106735, 0.18868867, 0.09173241, 0.14503679,
              0.35406714, 0.1236219 , 0.36285779, 0.26969192, 0.07040934,
              0.31333663, 0.25871657, 0.63373456, 0.28122385, 0.30243481,
              0.34496008, 0.4642903 , 0.15851687, 0.75934777, 0.44063586,
              0.25055878, 0.81853098, 0.47057462, 0.52021737, 0.50564909,
              0.09418966, 0.54057525, 0.09603089, 0.08990964, 0.23742844,
              0.09717017, 0.13327443, 0.30756257, 0.20346116, 0.30053486,
              0.44175729, 0.16879218, 0.24681302, 0.22121602, 0.08952118,
              0.26785664, 0.55857584, 0.20884104, 0.19674485, 0.71716436,
              0.1180427 , 0.35672827, 0.35275094, 0.86660976, 0.15458902,
              0.28958553, 0.21835329, 0.23387634, 0.13502529, 0.23182931,
              0.05020644, 0.34924433, 0.20805589, 0.26451144, 0.30796101,
              0.65129996, 0.07758494, 0.48188656, 0.47567004, 0.13096984,
              0.72766764, 0.07917856, 0.3073834 , 0.11122995, 0.57931627,
              0.30116294, 0.24636574, 0.60541779, 0.49159191, 0.40246228,
              0.39726098, 0.15066401, 0.17676648, 0.13852732, 0.28646138,
              0.30750146, 0.67292996, 0.1700256 , 0.39894917, 0.17917955,
              0.24333474, 0.6585855 , 0.38458859, 0.49235293, 0.22537323,
              0.25331683, 0.46920773, 0.31973832, 0.2562058 , 0.61922867,
              0.42922978, 0.73889282, 0.41773948, 0.36654385, 0.24244085,
              0.30609422, 0.28003556, 0.40422775, 0.12157507, 0.50851988,
              0.17942085, 0.25936704, 0.27070504, 0.43509392, 0.36619913,
              0.16318997, 0.25462431, 0.28001001, 0.16404297, 0.09583073,
              0.10761904, 0.1378797 , 0.37854315, 0.16322561, 0.29971851,
              0.08516014, 0.16212342, 0.14217155, 0.40517364, 0.36729043,
              0.85551444, 0.37569693, 0.44703018, 0.25658321, 0.05735779,
        0.69332327, 0.21640365])
```

[153]:
```
print("Actual Values:",list(y_pred))
print("Prediction:",list(map(round,y_pred)))
```

```
Actual Values: [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,␣
 ↪0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0,␣
 ↪0, 1, 0,
1, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0,␣
 ↪0, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0,␣
 ↪0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0,␣
 ↪0, 0, 0,
1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,␣
 ↪0, 0, 1,
0]
Prediction: [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,␣
 ↪0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0,␣
 ↪1, 0, 1,
1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0,␣
 ↪1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 0, 0,␣
 ↪0, 0, 0,
0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0,␣
 ↪0, 0, 1,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,␣
 ↪0, 1, 0]
```

[154]:
```
accuracy_score(y_test,y_pred)
```

[154]:
```
0.6624203821656051
```

[155]:
```
# plotting the ROC curve
auc_roc = roc_auc_score(y_test, y_pred)
fpr, tpr, thresholds = roc_curve(y_test, log_reg.predict_proba(X_test)[:
 ↪,1])
plt.plot(fpr, tpr, color='darkorange', lw=2,
         label='Average ROC curve (area = {0:0.3f})'.format(auc_roc))
plt.plot([0, 1], [0, 1], color='black', lw=2, linestyle='--',
         label= 'Average ROC curve (area = 0.500)')
plt.xlim([0.0, 1.0])
plt.ylim([0.0, 1.05])
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('Receiver operating characteristic')
plt.legend(loc="lower right")
plt.show()
```

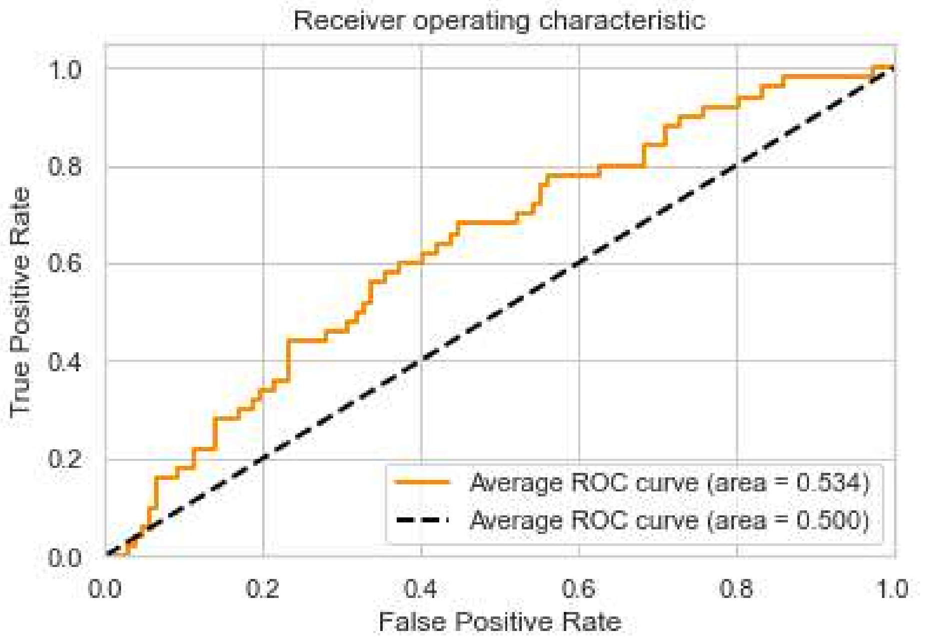


Figure 6.6: ROC curve for logistic classifier

From the above ROC curve we can say that, the area under the curve is 0.535 it means that the model has the ability to classify True positive and false negative.

Classification Algorithm	Accuracy Score
Decision Tree	70.063
Random Forest	69.426
KNN	69.426
SVM	68.789
Logistic	66.242

Conclusion
The accuracy score of **decision tree** algorithm is large as compaire to random forest, KNN, SVM and logistic algorithms.

6.4 Decision Tree

let see the python code for decision tree.

[156]:
```
from sklearn.tree import DecisionTreeClassifier
dt = DecisionTreeClassifier(criterion="gini", max_depth=3)
dt = dt.fit(X_train,y_train)
```

```
y1_pred = dt.predict(X_test)
```

[157]:
```
from sklearn.metrics import confusion_matrix,accuracy_score
```

[158]:
```
accuracy_score(y_test, y_pred)
```

[158]:
```
0.7006369426751592
```

[159]:
```
!pip install pydotplus
```

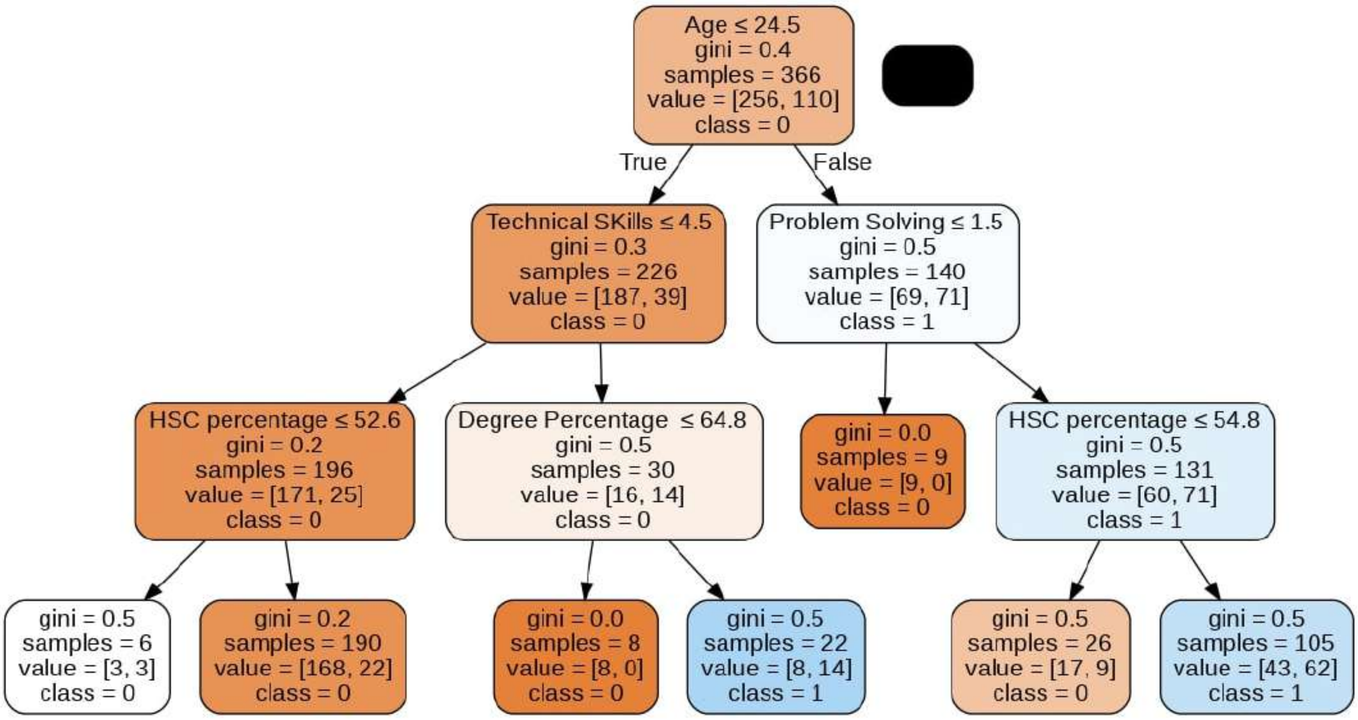
[160]:
```
feature_cols=['Gender', 'Age', ' Hometown', 'School Type ', 'SSC␣
  ↪percentage',
       'HSC percentage', 'SSC Board', 'HSC Board', 'Degree Percentage ',
       'Highest Degree', ' Stream Of Education ', 'Specialization ',␣
  ↪'Marathi',
       'English', 'Hindi', 'Critical Thinking', 'Team Work', 'Problem␣
  ↪Solving',
       'Communication Skills', 'Technical SKills', 'Creativity',␣
  ↪'Leadership']
```

[161]:
```
!pip install graphviz
```

[163]:
```
import six
import sys
sys.modules['sklearn.externals.six'] = six
```

[165]:
```
from sklearn.tree import export_graphviz
from sklearn.externals.six import StringIO
from IPython.display import Image
import pydotplus

dot_data = StringIO()
export_graphviz(dt, out_file=dot_data,
                filled=True, rounded=True,
                special_characters=True,feature_names =␣
  ↪feature_cols,class_names=['0','1'], precision=1)
graph = pydotplus.graph_from_dot_data(dot_data.getvalue())
Image(graph.create_png())
```



6.5 Conclusions:

From this decision tree, we classify **’Whether the candidate is going to be placed or not?’** Here 0 means student is not placed, 1 means student is placed.

- We have three sets of placed and non-placed students.
- It has been splitted based on Technical skills, problem-solving skills, HSC percentage, Degree percentage.

Chapter 7

Conclusions, Suggestions and References

7.1 Conclusions

1. There is significant impact of SSC, HSC, Degree and Master's percentage on placement.
2. There is significant impact of gender on placement.
3. There is no significant impact of School type and stream of education on placement.
4. The most used social media app for placement is LinkedIn and Majhi Naukari.
5. There is significant change in first salary offered and current salary.
6. The most affecting factors on placement are communication skills, soft skills, teaching quality and learning environment.

7.2 Some suggestions given by the students for the placement of students.

In this chapter, we see some suggestions given by the placed students in the google form for the placement of the students.
In our google form as there are 163 students who are placed out of this 153 students give some suggestions for the placement of the student. Some of them are as follows.

1. "If you are interested in taking a placement then first of all target the posts you are interested in like if you are interested in data science then there are posts like Data Scientists, Data Analysts, and Business Analysts. Then Study according to that like developing some basic coding skills, and learning about Data Science, and Machine Learning. Take help of the sensors which are placed in different companies. Start studying for Placement on your own as soon as possible, give mock interviews it will help a lot."-user 53
2. "As a student, we need communication skills nothing else. Hardworking, helping nature, and problem-solving are by default skills that we have already pursued in our blood. I would suggest every student should at least get an internship somewhere so

that he/she will get to know the working culture, market trends, exposure get some soft skills."-user 58

3. "I think students should study consistently and also they should do some extra-curricular activities for increasing their chances of placement."-user 60
4. "Be honest with yourself and study honestly. Immerse yourself in study till you succeed."-user 64
5. "Be confident in your field. Learn more practical applications from your field and try to apply them to real-life examples. This helps to stand out during interviews. Also, another suggestion would be to focus on problem-solving and try to think of all different scenarios while approaching any problem during interviews."-user 72
6. "Build your resumes strong and create a strong profile on LinkedIn as it's impacting the most. Organize the guest lectures and mock interviews as much as possible."-user 76
7. "Give too much time on LinkedIn and concentrate on technical skills."-user 121
8. "Develop skills when you are learning. Make connections on Linkedln with Seniors. Do at least one internship in your degree or master."-user 121
9. "Be smart and study well."-user 145
10. "Understand all the concepts you studied with real-life examples."-user 151
    In this way, the students who are placed are says that the student should do study consistently and focus on some social media such as LinkedIn and Majhi Naukri etc. for placement. Learn new skills, never stop learning and make contacts with seniors. Work hard you will get success.

7.3 Suggestions

Some suggestions from this survey by observing responses and analysis are as follows:

- From factor analysis, we say that students should focus on communication languages such as Hindi and English for placement.
- Student should develop their problem-solving attitude, confidence and have ability to work in a team, to lead a team.
- As teaching quality and learning environment also affects placement. So one should focus on these factors.
- Give some time to Linkedln and Majhi Naukari websites for being placed.

7.4 References

- https://journals.sagepub.com/doi/abs/10.1177/0950422220950191
- https://journals.sagepub.com/doi/abs/10.1177/097226290601000104?journalCode=visa
- https://www.projectguru.in/interpretation-of-factor-analysis-using-spss/

- https://www.projectpro.io/article/7-types-of-classification-algorithms-in-machine-learning/435

7.5 Software's used for Analysis:

- Python
- SPSS
- Minitab
- Excel
- Latex

https://tinyurl.com/2p8pbkv5