ANALYSIS OF FACTORS INFLUENCING CAMPUS PLACEMENTS

CSE3020 – DATA VISUALISATION

PROJECT BASED COMPONENT REPORT

By

Shruti Garg
Rahul Gudivada
Glenn Varghese George
Advika Srivastava
Shreya Rastogi

Register Number

19BCE0994
19BCE2469
19BCE2495
19BCE2217
19BCE0756

School of Computer Science and Engineering

May 2021

DECLARATION

I hereby declare that the report entitle "Analysis of Factors Influencing Campus Placement" submitted by me, for the CSE3020 DATA VISUALISATION (EPJ) to VIT is a record of bonafide work carried out by me under the supervision of Dr.S.VENGADESWARAN .

I further declare that the work reported in this report has not been submitted and will not be submitted, either in part or in full, for any other courses in this institute or any other institute or university.

Place : Vellore

Date : 07/06/2021

Signature of the Candidate

Shruti Garg
Rahul Gudivada
Glenn Varghese George
Advika Srivastava
Shreya Rastogi

CONTENTS

1. ABSTRACT:

Campus placements help the students to get a platform for themselves and they don't have to struggle themselves in the search for a job. Hence it is important to properly analyze the whole process of placements. In order to perform thorough data analysis, we chose a data set that consists of Placement data of students on the campus. It includes secondary and higher secondary school percentages and specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students.

2. INTRODUCTION TO THE PROJECT:

2.1 OBJECTIVE:

It is important to have prior research about the placement trends before applying for the same as it will help achieve better results in the process. The process of data visualization holds a lot of significance in doing so. Since there are a lot of students in a college and understanding the data trends manually will cause great difficulty, it is much easier to analyze and understand data if its in a visual form like a bar graph or pie chart rather than in a textual form like spreadsheets. Understanding data quickly also meansthat students can make decisions based on that data much more quickly as well. It is sometimes possible to even estimate future trends using Data Visualization. This gives a huge edge to students as they can move ahead of their competitors by analyzing future placement trends.

2.2 PROBLEM STATEMENT:

To understand the process of placements in a college, how the system works, and what are the major factors that affect the placement statistics and influence the candidates participating in it.

2.3 FUNCTIONAL REQUIREMENTS:

Along with analysis and visualization of data, our project will also provide a prediction model which will help students understand their chances of getting placed or not. The applicability of this feature will help students to understand what are the factors they need to improve in order to get a good placement package.

3. DATA ABSTRACTION:

Dataset Type:

The data set used for this project is a table which consists of Placement data of students in a XYZ campus. It includes secondary and higher secondary school percentages with specialization. It also includes degree specialization, type and Work experience and salary offers to the placed students.

Data Types:

The data types used here are attributes and items.

The dataset table of campus placements contains 15 attributes and 215 items.

Attribute Semantics:

1. **Sl_no:** Serial Number.
2. **gender:** The gender of the student. Male=”M”, Female=”F”
3. **ssc_p:**Secondary Education percentage of students in 10th Grade
4. **ssc_b:** Board of Education for 10th Grade - Central/ Others
5. **hsc_p:** Higher Secondary Education percentage of students in 12th Grade
6. **hsc_b:** Board of Education for 12th grade - Central/ Others
7. **hsc_s:**Specialization in Higher Secondary Education
8. **degree_p:** Degree Percentage scored by student
9. **degree_t:** Under Graduation(Degree type)- Field of degree education
10. **workex:** Work Experience
11. **etest_p:** Employability test percentage (conducted by college)
12. **specialisation:** Post Graduation(MBA)- Specialization
13. **mba_p:** MBA percentage
14. **status:** Status of placement- Placed/Not placed
15. **salary:** Salary offered by corporate to candidates

The attributes can be classified as categorical and ordered as follows:

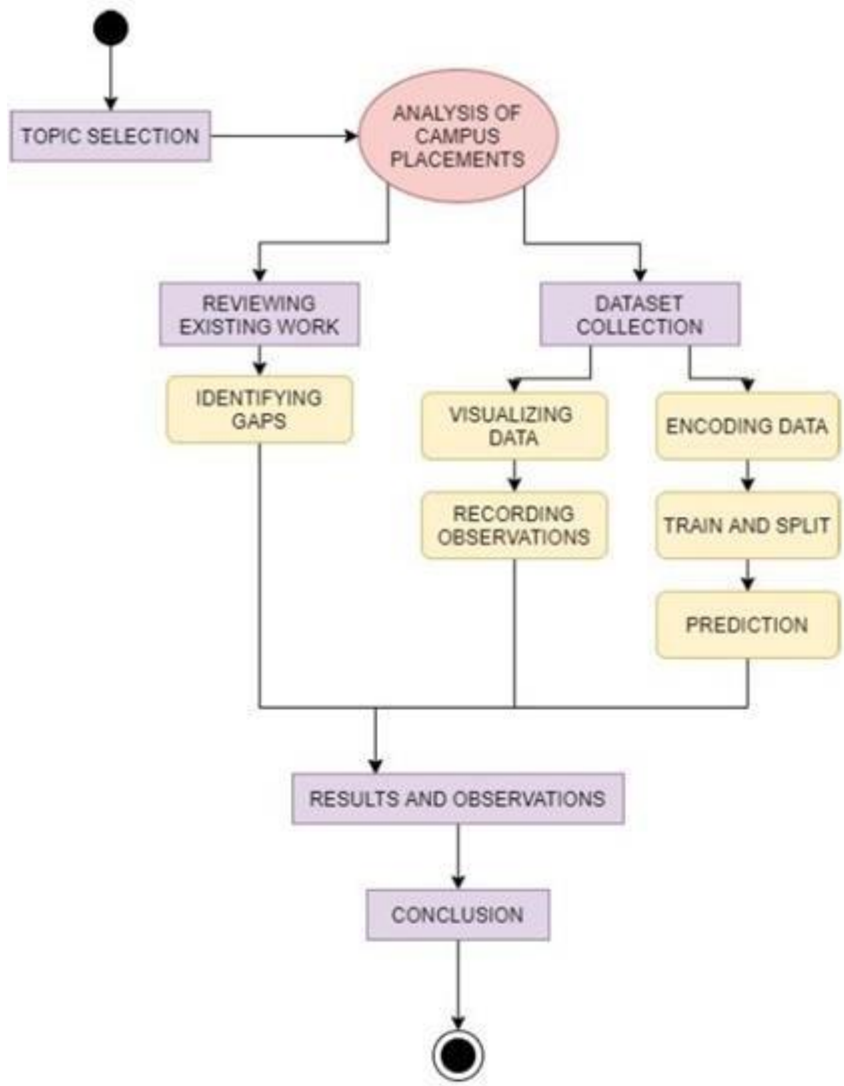
Categorical	Ordered
Gender : Male- M , Female- F	Sl_no
ssc_b: Central / Other	ssc_p
hsc_b: Central / Other	hsc_p
hsc_s: Science/ Commerce/ Arts	degree_p
degree_t: Sci&Tech / Comm&Mgmt	etest_p
workex: Yes / No	mba_p
specialisation: Mkt&HR/ Mkt&Fin	
status: Placed / Not placed	

Target Identification:

We have focused on the following objectives or targets for our project:

- Does the candidate's **gender** (male or female) have any role in placement?
- Which **factors** influenced a candidate in getting placed?
- Does **10th and 12th percentage** matter for one to get placed?
- Which **degree specializatio**n is much demanded by corporate?
- Determine the **average salary** offered during placements and its factors.
- Play with the data conducting all **statistical tests**.

4. DESIGN OF THE PROPOSED SYSTEM:



Dataset Collection:

We use a dataset from kaggle – "Factors affecting campus placements" for our data.

Visualizing data:

Here we find the best graphical representation of our data. i.e. bar charts, pie charts, box plot etc.

Recording Observations:

We record the observations from the data.

Encoding Data:

Here we translate the data into a visual element on the plots we are making.

Train and split:

We divide the training sessions by body regions into two: one for training and one for testing.

Prediction:

Here we Predict whether the candidate will be placed or not based on some

predictors.

5. ALGORITHMIC DESIGN:

Classification Algorithms Used:

1. Logistic Regression: Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

2. Decision Tree: Decision trees use multiple algorithms to decide to split a node into two or more sub-nodes. The creation of sub-nodes increases the homogeneity of resultant sub-nodes. The decision tree splits the nodes on all available variables and then selects the split which results in most homogeneous sub-nodes.

3. Ensemble Model: After finding out the accuracy scores from logistic regression and decision tree, we found they both have an equal score of 86.15%. We then use ensemble model to combine both the models for better prediction. The accuracy score was raised to 92.3%.

6. <u>TASK ABSTRACTION:</u>

1) ANALYZE -

a) Consume:

We have chosen a dataset which provides us with a table containing factors which influence the campus placements. These factors can be analyzed and compared to find out some trends and important factors.

a) Produce:

Our production goal is to “derive”. From the present dataset we can produce a prediction model which will help students understand their chances of getting placed or not. The applicability of this feature will help students to understand what are the factors they need to improve in order to get a good placement package.

2) SEARCH -

We are doing an analysis of all the factors to find out their role in placements and thus don’t have any fix target. Also the location is not known as any of the factor might play a prime role. So as both target and location are unknown, our search method is “explore”.

3) QUERY-

After the searching mechanism we have provided a comparison of all the factors as well as a provide a comprehensive view i.e. “summary” of these factors through histogram.

Using one or two variables:

- Does the candidate's gender (male or female) have any role in placement?
- Does 10th and 12th percentage matter for one to get placed?
- Which degree specialization is much demanded by corporate?
- Determine the average salary offered during placements and its factors.

Using all the attributes:

- Which factors influenced a candidate in getting placed?
- Play with the data conducting all statistical tests.
- To provide a summary of all the factors and show correlation between them.

7. DASHBOARD IMPLEMENTATION:

Using VOILA:

ANALYSIS OF FACTORS INFLUENCING CAMPUS PLACEMENTS

VIT
Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)

Reading Dataset

```
Number of rows in data : 215
Number of columns in data : 14
```

Generating The Data Types of the columns

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 215 entries, 0 to 214
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   gender          215 non-null    object
 1   ssc_p           215 non-null    float64
 2   ssc_b           215 non-null    object
 3   hsc_p           215 non-null    float64
 4   hsc_b           215 non-null    object
 5   hsc_s           215 non-null    object
 6   degree_p        215 non-null    float64
 7   degree_t        215 non-null    object
 8   workex          215 non-null    object
 9   etest_p         215 non-null    float64
 10  specialisation  215 non-null    object
 11  mba_p           215 non-null    float64
 12  status          215 non-null    object
 13  salary          148 non-null    float64
dtypes: float64(6), object(8)
memory usage: 23.6+ KB
```

What are the percentage of Candidates that are not placed?

```
Salary column has 31.16% null values.
```

This tells us that around 31% candidates were not placed

let's see what were the reasons

What is the average placement package of the college ?

	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
count	215.000000	215.000000	215.000000	215.000000	215.000000	148.000000
mean	67.303395	66.333163	66.370186	72.100558	62.278186	288655.405405

What is the average placement package of the college ?

	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
count	215.000000	215.000000	215.000000	215.000000	215.000000	148.000000
mean	67.303395	66.333163	66.370186	72.100558	62.278186	288655.405405
std	10.827205	10.897509	7.358743	13.275956	5.833385	93457.452420
min	40.890000	37.000000	50.000000	50.000000	51.210000	200000.000000
25%	60.600000	60.900000	61.000000	60.000000	57.945000	240000.000000
50%	67.000000	65.000000	66.000000	71.000000	62.000000	265000.000000
75%	75.700000	73.000000	72.000000	83.500000	66.255000	300000.000000
max	89.400000	97.700000	91.000000	98.000000	77.890000	940000.000000

Average Salary Offered: 288655

Min Salary Offered:200000

Max Salary Offered:940000

Analysis of classes (unique values) of columns in the dataset.

```
-gender--
M     139
F      76
Name: gender, dtype: int64
-ssc_b--
Central    116
Others      99
Name: ssc_b, dtype: int64
-hsc_b--
Others     131
Central     84
Name: hsc_b, dtype: int64
-hsc_s--
Commerce    113
Science      91
Arts         11
Name: hsc_s, dtype: int64
-degree_t--
Comm&Mgmt    145
Sci&Tech      59
Others        11
Name: degree_t, dtype: int64
-workex--
No     141
Yes     74
Name: workex, dtype: int64
-specialisation--
Mkt&Fin    120
Mkt&HR      95
Name: specialisation, dtype: int64
-status--
Placed        148
```

OBSERVATION:

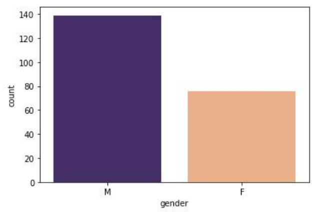
hsc_s and degree_t have 3 classes.

All other columns have 2 classes each

Imbalanced data:148 placed students and 67 not placed students, showing higher placement Rate

EXPLORING COLUMNS THROUGH VISUALIZATION

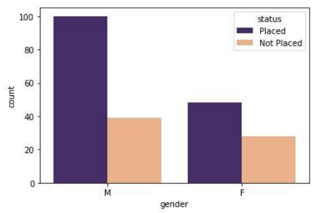
Campare the Male and Female candidates who applied for Placement



OBSERVATION:

More number of male candidates applied for the placement process than female candidates.

What is the placement Status of male and Female?



OBSERVATIONS:

(I) Number of male students are almost double as compared to female.

(II) Fraction of placed vs not placed for female candidates is significantly low as compared to male candidates.

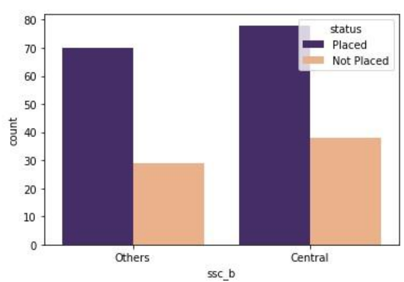
OBSERVATIONS:

(I) Number of male students are almost double as compared to female.

(II) Fraction of placed vs not placed for female candidates is significantly low as compared to male candidates.

Hence we can conclude male candidates are accepted more often than female.

Is there an impact of taking a specific board in 10th grade on placements?

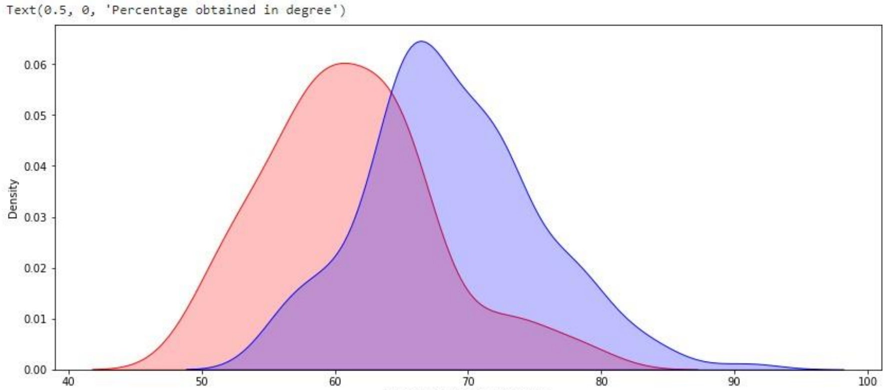


OBSERVATIONS: (I) There is count of central board students is very high as compared to all other boards. (II) The count of placed students from central board is little more than others category which doesn't say

OBSERVATION: (I) Packages with salary: 300000 were offered in highest number. (II)High Salary Packages Have a very low count

Does CGPA and Degree Percentage Matter in Placements?

```
Text(0.5, 0, 'Percentage obtained in degree')
```

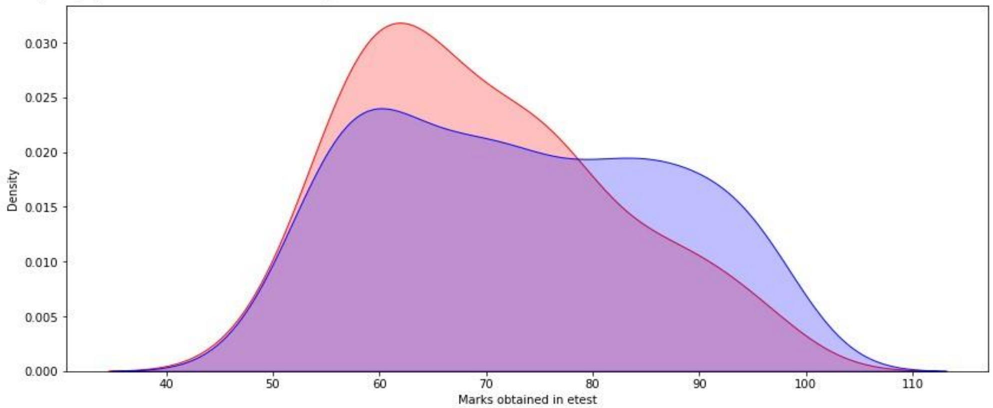


Percentage obtained in degree

OBSERVATIONS: (I) Students with percentages from 90-100 are fully placed. (II) Students with percentages from 40-50 are not at all placed.

Does Etest marks Matter in Placements?

```
Text(0.5, 0, 'Marks obtained in etest')
```

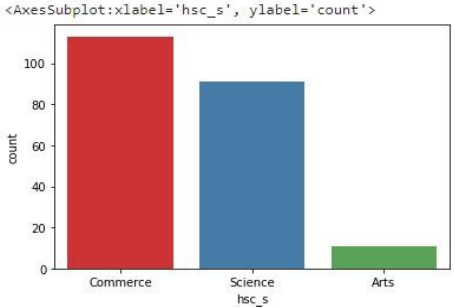


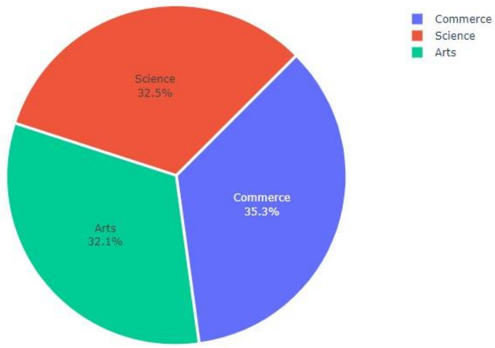
OBSERVATION:

Etest marks cannot be considered as a significant factor as the marks are even distributed along with the placement status

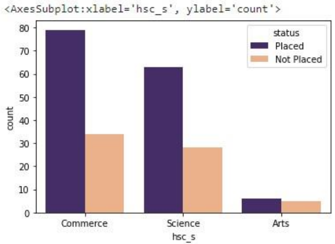
What is the impact of hsc specializations in placement?

```
<AxesSubplot:xlabel='hsc_s', ylabel='count'>
```





<AxesSubplot:xlabel='hsc_s', ylabel='count'>



OBSERVATIONS: (I) The most popular branch turns out to be commerce or maybe as most of students get average marks so they were admitted to got commerce on based of their marks. Science is the second most popular and the least popular is arts.

(II) Almost every branch students performed equally but commerce students have slightly better score than other two.

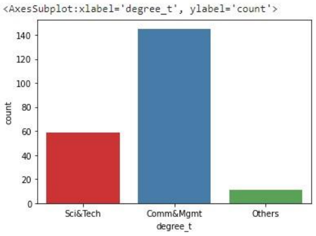
OBSERVATIONS: (I) The most popular branch turns out to be commerce or maybe as most of students get average marks so they were admitted to got commerce on based of their marks. Science is the second most popular and the least popular is arts.

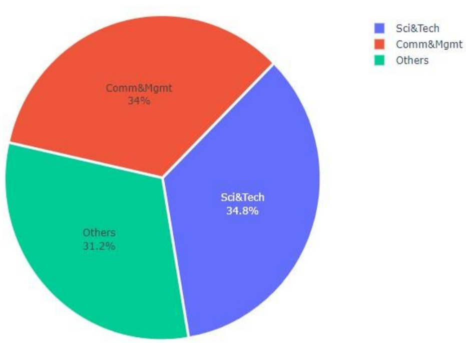
(II) Almost every branch students performed equally but commerce students have slightly better score than other two.

(III) Looking at the fraction of placed and not placed we can say that science branch students have more chance of getting placed than commerce students and most around 45% of the students in arts are not placed

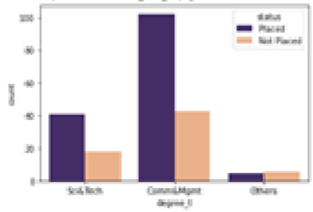
What is the impact of Degree type specializations in placement?

<AxesSubplot:xlabel='degree_t', ylabel='count'>





<AxesSubplot:xlabel='degree_t', ylabel='count'>



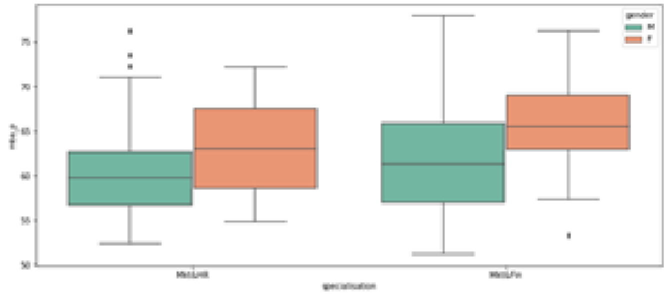
OBSERVATION:

(I) The students opted for following fields:

Science and Technology (must be science students) Commerce and management (might be a mixture of commerce and Arts) Others (II) There is not much difference in performace of students from Science and Commerce but there but students who opted for "Others" have low p

(III) Looks like Commerce and Science degree students are preffered by companies which is obvious. Students who opted for Others have very low placement chance

What is the distribution of students based on their specialization?



OBSERVATION:

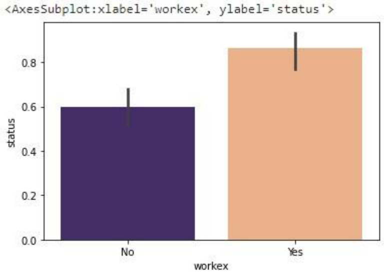
Females of Mkt and Fin are having higher average mba percentages Males of Mkt and HR are having lowest average mba percentages

OBSERVATION:

Females of Mkt and Fin are having higher average mba percentages Males of Mkt and HR are having lowest average mba percentages

Does Work Experience Matter in Campus Placement?

<AxesSubplot:xlabel='workex', ylabel='status'>

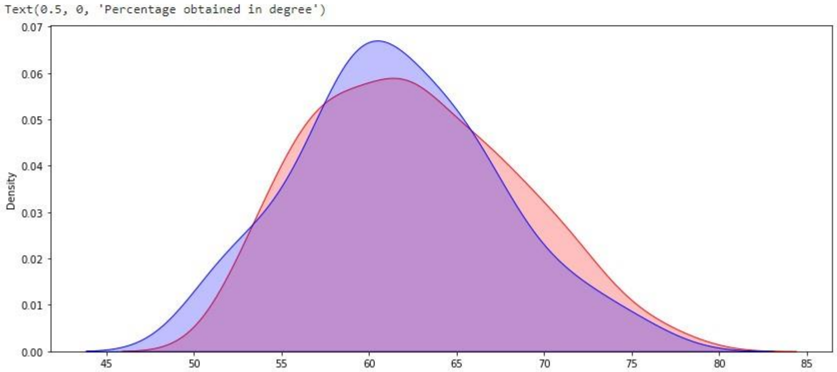


OBSERVATION:

Companies prefer candidates with work experience so the students with internships and past job experience have better chances of being placed.

If i have high MBA percentage, will I get placed?

Text(0.5, 0, 'Percentage obtained in degree')



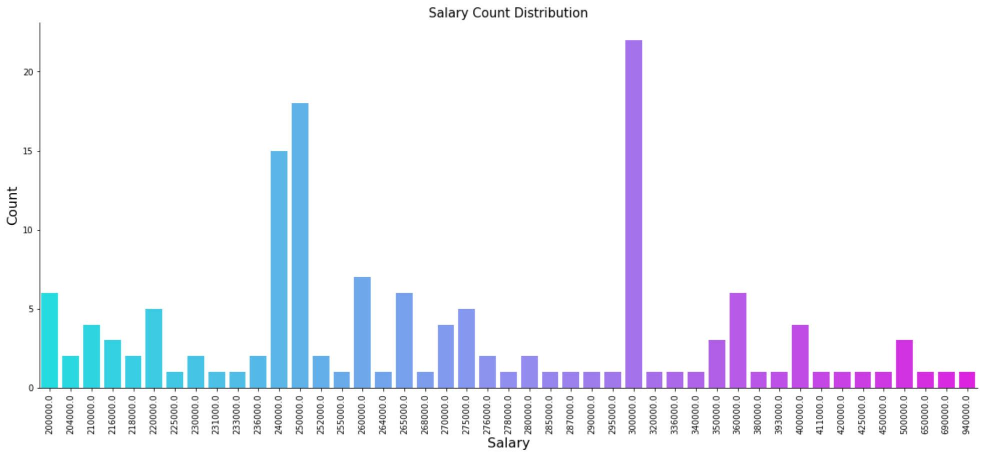
OBSERVATION:

OBSERVATION:

We can see that getting good percentages in MBA does not guarantee placement of the candidate

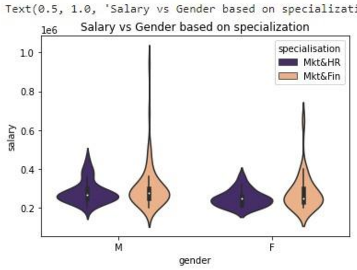
SALARY ANALYSIS

What is the package recieved by maximum number of students?



Salary vs Gender based on specialisation

```
Text(0.5, 1.0, 'Salary vs Gender based on specialization')
```



OBSERVATIONS:

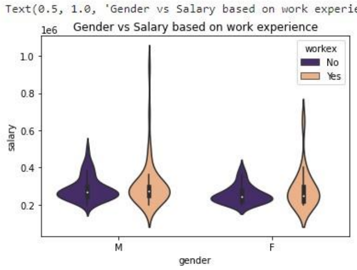
(I) Salary column for male candidates seems to have more outliers than females which means that a lot more male candidates got more than the average CTC.

(II) Mean salary is somewhere around 220k.

(III) Mkt&Fin students are given higher salaries as compared to Mkt&HR.

Gender vs Salary based on work experience

```
Text(0.5, 1.0, 'Gender vs Salary based on work experience')
```



OBSERVATIONS:

(I) Work Experience is a clear indicator as more work experience results in higher CTC jobs.

(II) The maximum salary in male candidates with experience is >1M and for female it is ~700k. The maximum salary in male candidates without experience is ~550k and for female it is ~430k.

Salary vs Gender based on Board in 10th grade

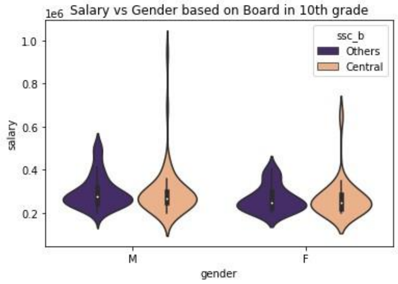
OBSERVATIONS:

(I) Work Experience is a clear indicator as more work experience results in higher CTC jobs.

(II) The maximum salary in male candidates with experience is >1M and for female it is ~700k. The maximum salary in male candidates without experience is ~550k and for female it is ~430k.

Salary vs Gender based on Board in 10th grade

```
Text(0.5, 1.0, 'Salary vs Gender based on Board in 10th grade')
```

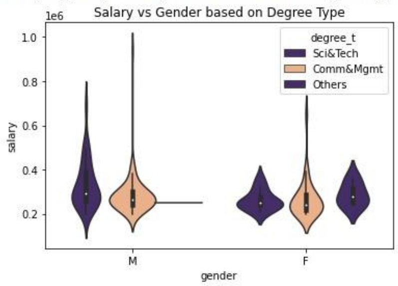


OBSERVATION:

Both Male and Female candidates from Central board got higher CTC as compared to other boards thus we can that central board in 10th grade might fetch you higher CTCs.

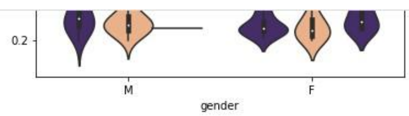
Salary vs Gender based on Degree Type

```
Text(0.5, 1.0, 'Salary vs Gender based on Degree Type')
```



OBSERVATIONS:

(I) Both male and female candidate got high CTCs choosing Comm&Mgmt as their degree.

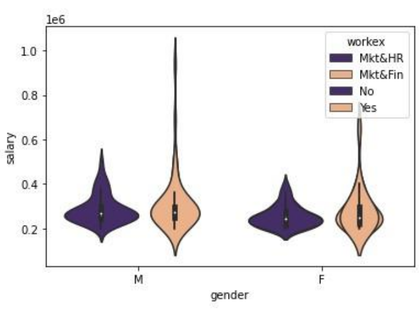


OBSERVATIONS:

(I) Both male and female candidate got high CTCs choosing Comm&Mgmt as their degree.

(II) Male candidates from Sci&Tech got high CTCs as compared to Female candidates.

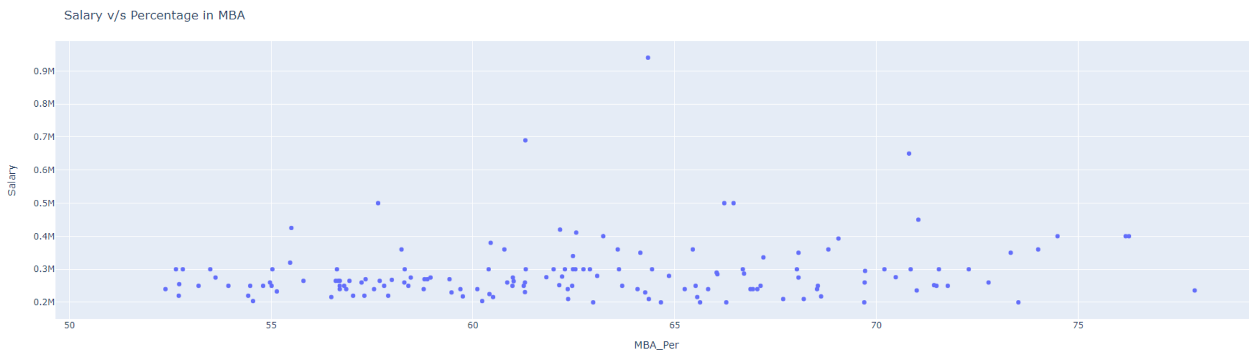
(III) None of the male candidates got placed from "Others" category whereas for female candidates the package is close to what female Sci&Tech candidates got.



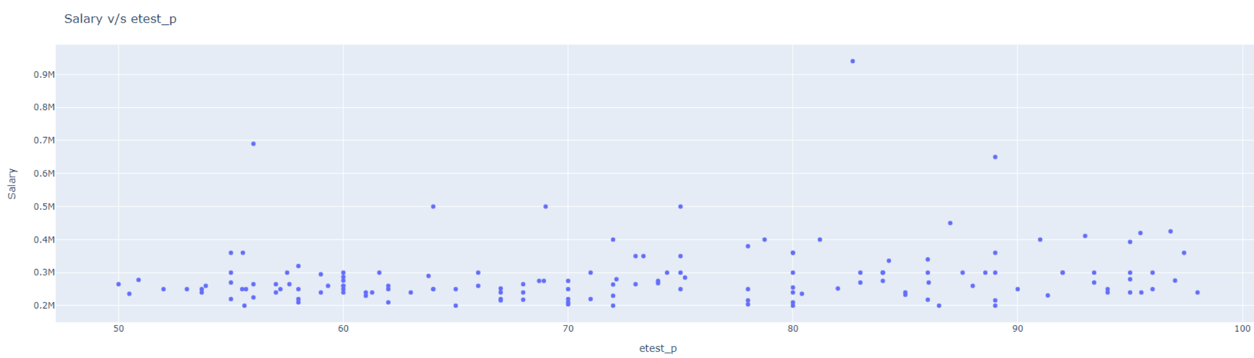
x specialisation

```
<AxesSubplot:xlabel='gender', ylabel='salary'>
```

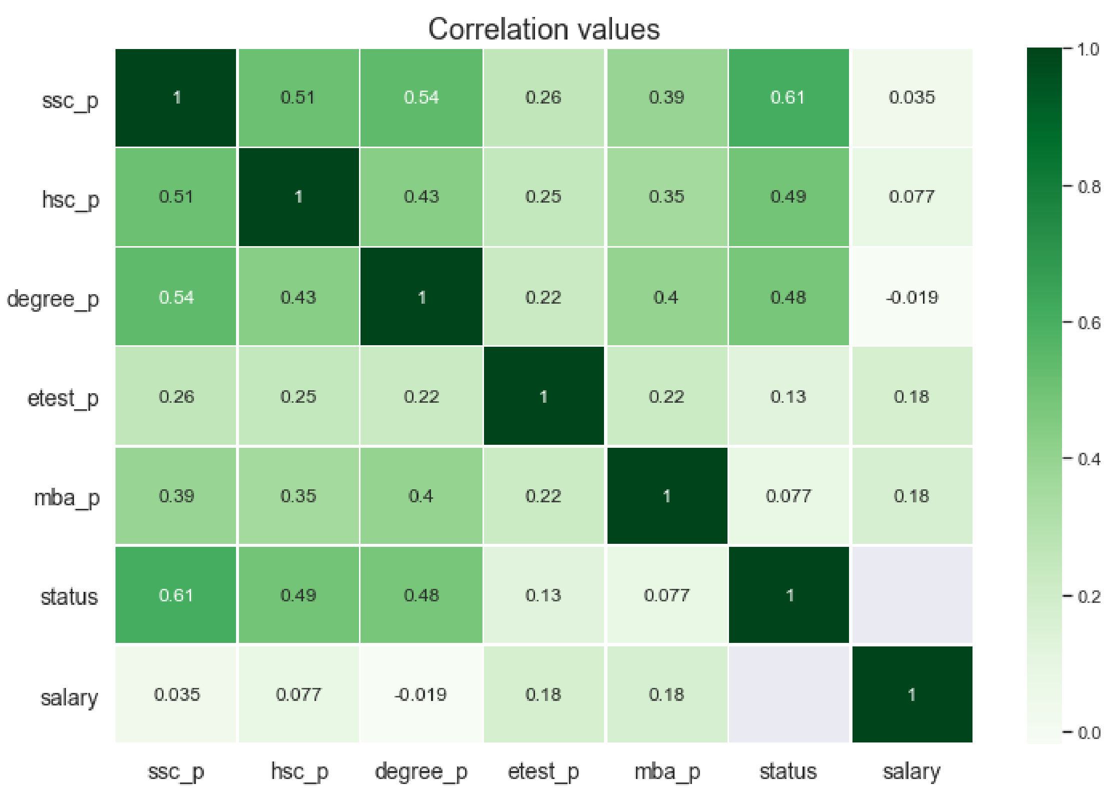
Salary v/s Percentage in MBA



Salary v/s etest_p



Is it possible to find relation between numerical values in data set?



OBSERVATIONS:

The ssc_p,hsc_p,degree_p have higher correlation with status, hence affect the placement procedure more.

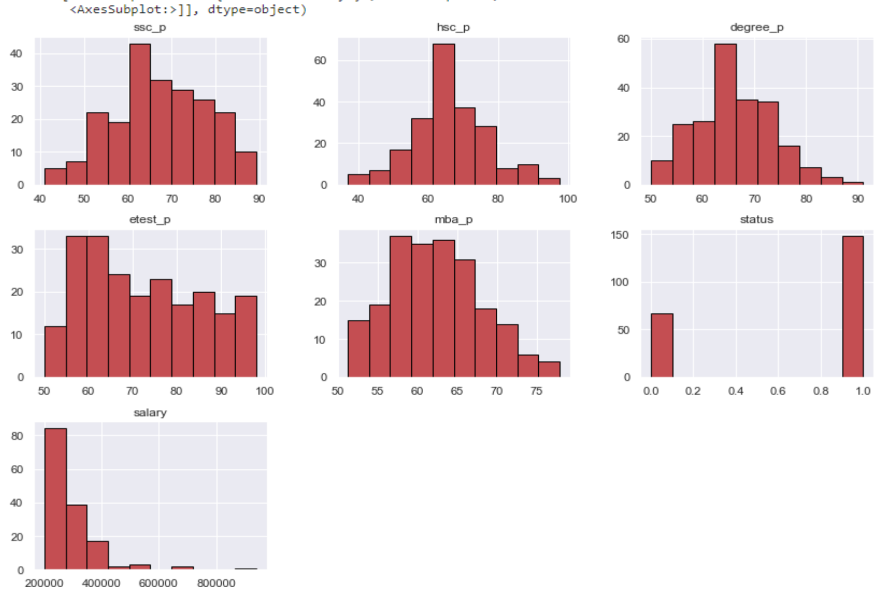
ssc_p hsc_p degree_p etest_p mba_p status salary

OBSERVATIONS:

The ssc_p,hsc_p,degree_p have higher correlation with status, hence affect the placement procedure more.

Summary (Histogram Distribution)

```
array([[<AxesSubplot:title={'center':'ssc_p'}>,
        <AxesSubplot:title={'center':'hsc_p'}>,
        <AxesSubplot:title={'center':'degree_p'}>],
       [<AxesSubplot:title={'center':'etest_p'}>,
        <AxesSubplot:title={'center':'mba_p'}>,
        <AxesSubplot:title={'center':'status'}>],
       [<AxesSubplot:title={'center':'salary'}>, <AxesSubplot:>,
        <AxesSubplot:>]], dtype=object)
```



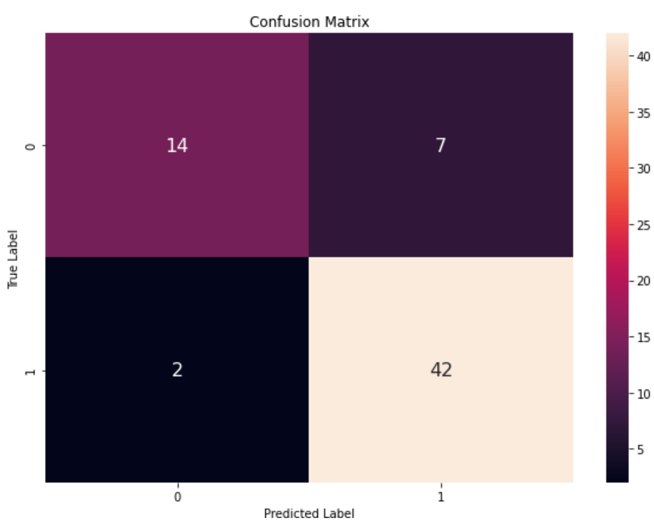
PREDICTING WHETHER A STUDENT WILL GET PLACED OR NOT

Encoding Data

	gender	ssc_p	hsc_p	degree_p	workex	etest_p	specialisation	mba_p	status	Arts	Commerce	Science	Comm&Mgmt	Others	Sci&Tech
0	0	67.00	91.00	58.00	0	55.0	0	58.80	Placed	0	1	0	0	0	1
1	0	79.33	78.33	77.48	1	86.5	1	66.28	Placed	0	0	1	0	0	1
2	0	65.00	68.00	64.00	0	75.0	1	57.80	Placed	1	0	0	1	0	0
3	0	56.00	52.00	52.00	0	66.0	0	59.43	Not Placed	0	0	1	0	0	1
4	0	85.80	73.60	73.30	0	96.8	1	55.50	Placed	0	1	0	1	0	0

```
X-Train: (150, 12)
X-Test: (65, 12)
Y-Train: (150,)
Y-Test: (65,)
```

```
LogisticRegression()
```



OBSERVATIONS: Our confusion Matrix looks decent. We have correctly predicted 42 (placed) + 14 (not-placed) correct predictions and 7 (not placed as placed) + 2(placed as not-placed) incorrect predictions.

We need to decrease these incorrect predictions because a good candidate can be rejected (false positive) and a unfit candidate can be selected (false negatives)

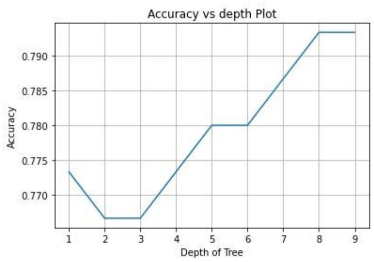
```
              precision    recall  f1-score   support

           0       0.88      0.67      0.76        21
           1       0.86      0.95      0.90        44

    accuracy                           0.86        65
   macro avg       0.87      0.81      0.83        65
weighted avg       0.86      0.86      0.86        65
```

```
The accuracy : 86.15%
```

Decision Tree Classifier

Let's try some decision trees now and see how well they perform but as Decision trees are easy to overfit so I will use K-FOLD CV first to find the best depth.

```
The optimal depth value is:  8
```




```
Accuracy scores for each depth value is :  [0.773 0.767 0.767 0.773 0.78  0.78  0.787 0.793 0.793]

The accuracy on test set using optimal depth = 8 is 86.154%
```

We achieved 86% accuracy which is similiar to what we achieved using logistic regression so they seem to work equally well.

What if we could combine the power of two models to get better results?

Ensemble Modelling

We will train a voting classifier using our previously trained logistic regeression and Decision tree model

```
Training the LogisticRegression()
Training the DecisionTreeClassifier(max_depth=8, random_state=42)

[0.8615384615384616, 0.8615384615384616]

VotingClassifier(estimators=[('log_reg', LogisticRegression()),
                             ('dt_tree',
```


What if we could combine the power of two models to get better results?

Ensemble Modelling

We will train a voting classifier using our previously trained logistic regeression and Decision tree model

```
Training the LogisticRegression()
Training the DecisionTreeClassifier(max_depth=8, random_state=42)

[0.8615384615384616, 0.8615384615384616]

VotingClassifier(estimators=[('log_reg', LogisticRegression()),
                             ('dt_tree',
```


```
                              DecisionTreeClassifier(max_depth=8,
                                                     random_state=42))])

The accuracy on test set using voting classifier is 92.31%
```

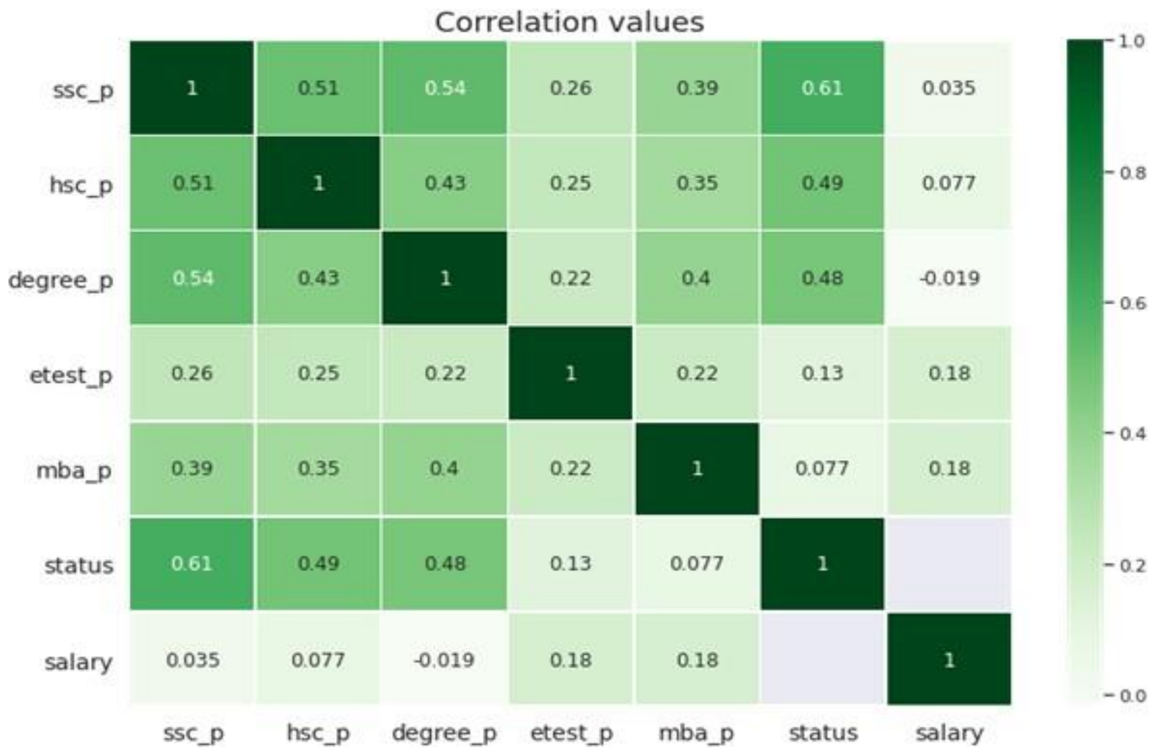
We went from 86.4% to 92.3% accuracy score!

Hence, ensemble modelled voting classifier of Logistic and decision tree helped us increase the accuracy of the prediction model

Conclusions Drawn

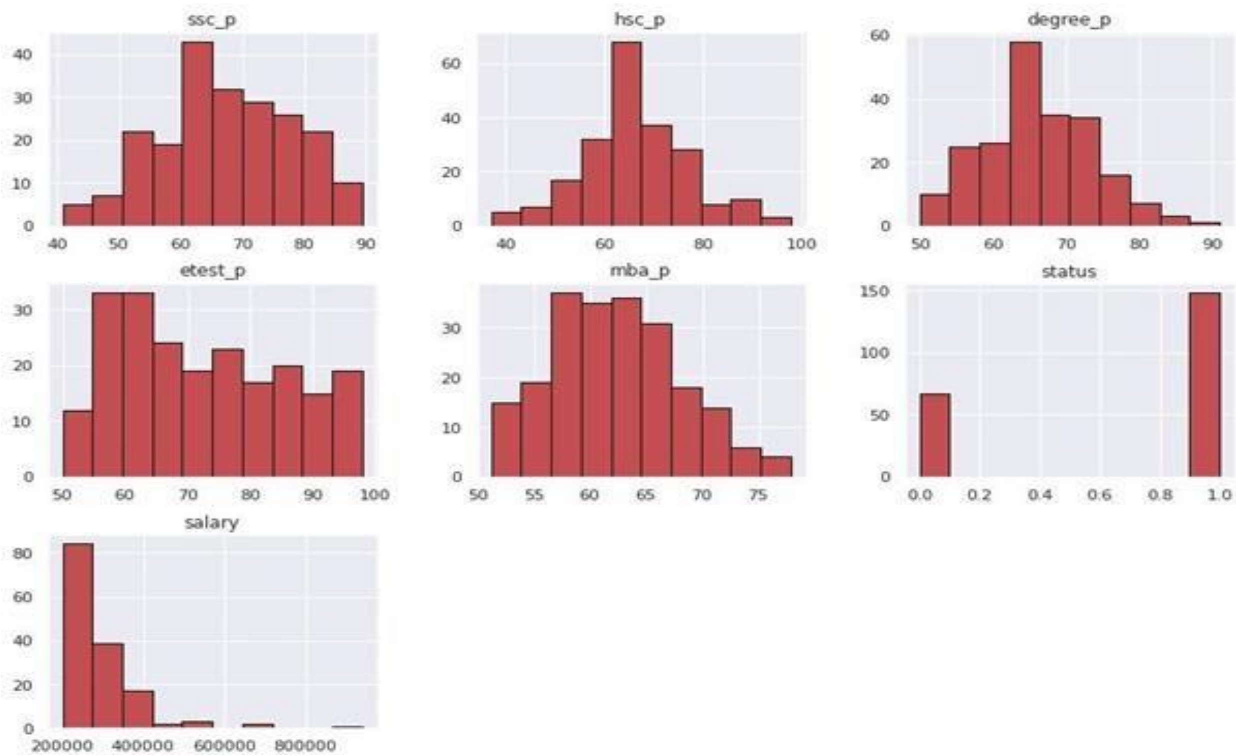
- More male candidates got placed as compared to female candidates.
- Male Candidates got higher CTCs as compared to female candidates.
- Type of Board choosen does not have any effect on placements thus we can drop in preprocessing steps.
- Most of the students preferred Central board in 10th grade whereas other boards in 12th grade.
- Candidates with higher percentages have better chance of placements.
- Choosing Science and Commerce as Specialisation seems to have perk when it comes to placments.
- Maximum package was bagged by male candidate from Mkt&Fin branch which is around 940k.
- Commerce is the most popular branch among candidates.
- Mean CTC is around 220k for male and female candidates individually.
- Choosing Sci&Tech and Comm&Mngmt as degree will fetch you higher CTCs.
- Mkt&Fin major have higher salaries and more placement chance as compared to Mkt&HR.
- Employability test percentage and MBA percentage does not effect the placements

8. RESULT ANALYSIS:



The ssc_p,hsc_p,degree_p have higher correlation with status, hence affect the placement procedure more.

Summary (Histogram Distribution)



9. CONCLUSION:

● Male Candidates got higher CTCs as compared to female candidates.

●More male candidates got placed as compared to female candidates.

● Type of Board chosen does not have any effect on placements thus we can drop in preprocessing steps.

● Most of the students preferred the Central board in 10th grade whereas other boards in 12th grade.

● Candidates with higher percentages have better chances of placements.

● Choosing Science and Commerce as Specialisation seems to have perks when it comes to placements.

● Maximum package was bagged by male candidate from Mkt&Fin branch which is around 940k.

● Commerce is the most popular branch among candidates.

● Mean CTC is around 220k for male and female candidates individually.

● Choosing Sci & Tech and Comm Mgmt as degrees will fetch you higher CTCs.

● Mkt&Fin major have higher salaries and more placement chance as compared to Mkt&HR.

● Employability test percentage and MBA percentage does not affect the placements.

● Ensemble Modelling gives better accuracy when predicting the data

10. APPENDIX:

- ## SAMPLE CODING:

FACTORS AFFECTING CAMPUS PLACEMENTS

importing libraries

In [1]:
```
import warnings
warnings.filterwarnings('ignore')

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix, classification_report, accuracy_score
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import roc_auc_score
from sklearn.metrics import roc_curve,auc
```

Reading Dataset

In [2]:
```
data = pd.read_csv("C:/Users/rahul/Downloads/Placement_Data_Full_Class.csv")

data.drop("sl_no", axis=1, inplace=True)
```

Checking Total rows and columns

In [3]:
```
print("Number of rows in data :",data.shape[0])
print("Number of columns in data :", data.shape[1])
```

```
Number of rows in data : 215
Number of columns in data : 14
```

Generating The Data Types of the columns

In [4]:
```
data.info()
```

In [4]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 215 entries, 0 to 214
Data columns (total 14 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   gender          215 non-null    object
 1   ssc_p           215 non-null    float64
 2   ssc_b           215 non-null    object
 3   hsc_p           215 non-null    float64
 4   hsc_b           215 non-null    object
 5   hsc_s           215 non-null    object
 6   degree_p        215 non-null    float64
 7   degree_t        215 non-null    object
 8   workex          215 non-null    object
 9   etest_p         215 non-null    float64
 10  specialisation  215 non-null    object
 11  mba_p           215 non-null    float64
 12  status          215 non-null    object
 13  salary          148 non-null    float64
dtypes: float64(6), object(8)
memory usage: 23.6+ KB
```

What are the percentage of Candidates that are not placed?

In [5]:
```
p = data['salary'].isnull().sum()/(len(data))*100

print(f"Salary column has {p.round(2)}% null values.")
```

```
Salary column has 31.16% null values.
```

This tells us that around 31% candidates were not placed

let's see what were the reasons

What is the average placement package of the college ?

```
In [6]: data.describe()
```

Out[6]:

	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
count	215.000000	215.000000	215.000000	215.000000	215.000000	148.000000
mean	67.303395	66.333163	66.370186	72.100558	62.278186	288655.405405
std	10.827205	10.897509	7.358743	13.275956	5.833385	93457.452420
min	40.890000	37.000000	50.000000	50.000000	51.210000	200000.000000
25%	60.600000	60.900000	61.000000	60.000000	57.945000	240000.000000
50%	67.000000	65.000000	66.000000	71.000000	62.000000	265000.000000
75%	75.700000	73.000000	72.000000	83.500000	66.255000	300000.000000
max	89.400000	97.700000	91.000000	98.000000	77.890000	940000.000000

```
Average Salary Offered: 288655

Min Salary Offered:200000

Max Salary Offered:940000
```

Analysis of classes (unique values) of columns in the dataset.

```
In [7]:
object_columns = data.select_dtypes(include=['object']).columns


for col in object_columns:
    print( '-'+ col +'-', end='-')
    display(data[col].value_counts())
```

```
-gender--

M    139
F     76
Name: gender, dtype: int64

-ssc_b--

Central    116
Others      99
Name: ssc_b, dtype: int64

-hsc_b--

Others     131
Central     84
```

OBSERVATION:

hsc_s and degree_t have 3 classes,

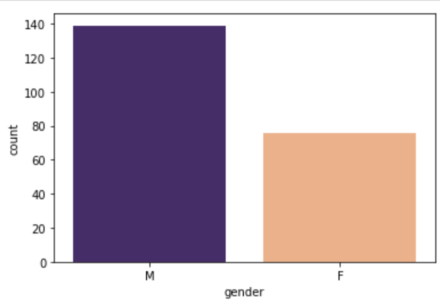
All other columns have 2 classes each

Imbalanced data:148 placed students and 67 not placed students, showing higher placement Rate

EXPLORING COLUMNS THROUGH VISUALIZATION

Campare the Male and Female candidates who applied for Placement

```
In [8]: sns.countplot("gender", data = data,palette=['#432371',"#FAAE7B"])
        plt.show()
```



OBSERVATION:

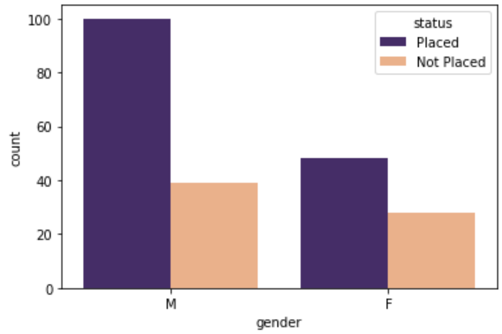
More number of male candidates applied for the placement process than female candidates.

What is the placement Status of male and Female?

```
In [9]: sns.countplot("gender", hue="status", data=data,palette=['#432371',"#FAAE7B"])
        plt.show()
```

What is the placement Status of male and Female?

```
In [9]: sns.countplot("gender", hue="status", data=data,palette=['#432371',"#FAAE7B"])
        plt.show()
```



OBSERVATIONS:

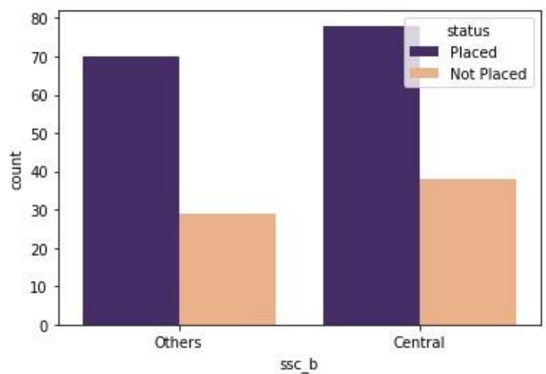
(I) Number of male students are almost double as compared to female.

(II) Fraction of placed vs not placed for female candidates is significantly low as compared to male candidates.

Hence we can conclude male candidates are accepted more often than female.

Is there an impact of taking a specific board in 10th grade on placements?

```
In [10]: sns.countplot("ssc_b", hue="status", data=data,palette=['#432371',"#FAAE7B"])
         plt.show()
```



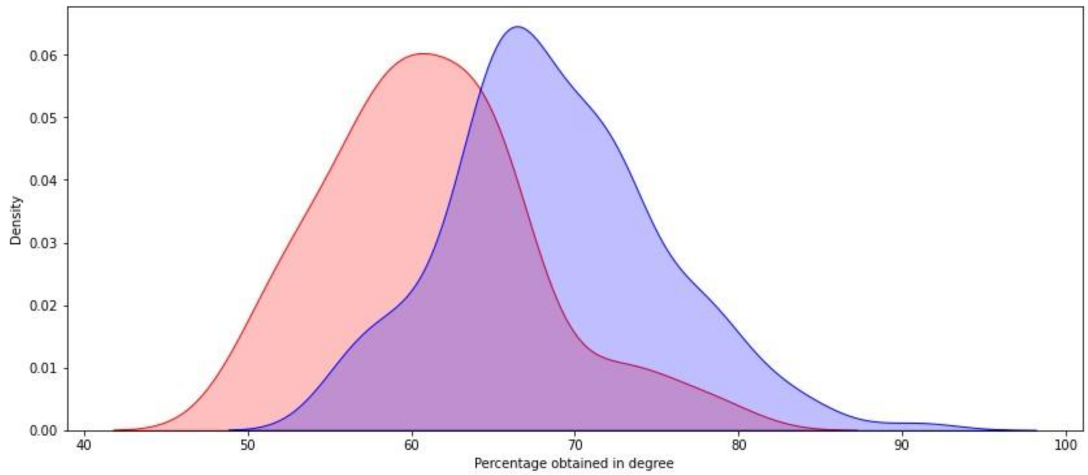
OBSERVATIONS:
(I) There is count of central board students is very high as compared to all other boards.
(II) The count of placed students from central board is little more than others category which doesn't say much.

OBSERVATION:
(I) Packages with salary: 300000 were offered in highest number.
(II)High Salary Packages Have a very low count

Does CGPA and Degree Percentage Matter in Placements?

In [11]:
```
placed_df = data[data['status']=="Placed"]
not_placed = data[data['status']=="Not Placed"]
plt.figure(figsize = (14,6))
sns.kdeplot(not_placed['degree_p'], label = 'Students not placed', color = 'r', shade = True)
sns.kdeplot(placed_df['degree_p'],label='Students who got placed',color = 'b',shade=True)
plt.xlabel('Percentage obtained in degree')
```

Out[11]: `Text(0.5, 0, 'Percentage obtained in degree')`



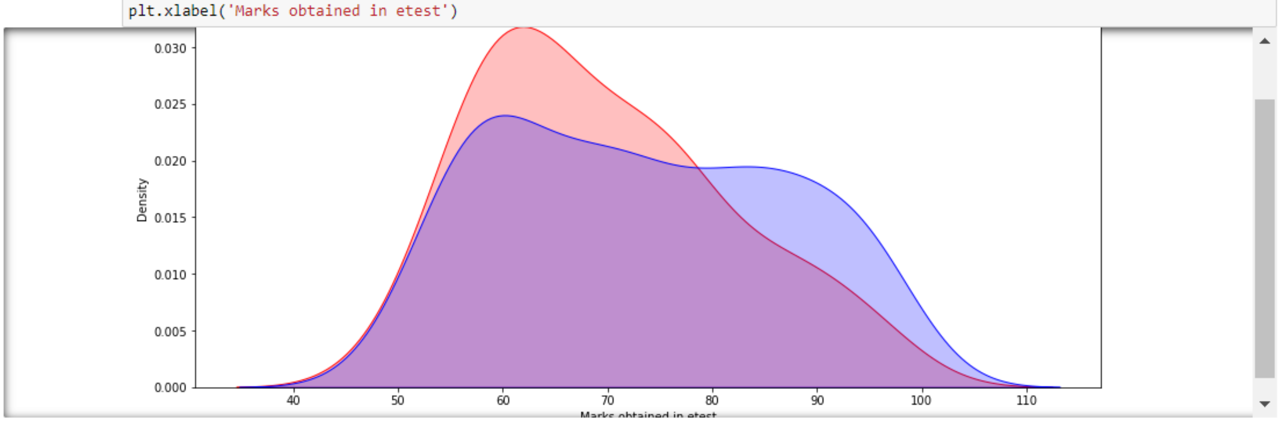
OBSERVATIONS:

(I) Students with percentages from 90-100 are fully placed.

(II) Students with percentages from 40-50 are not at all placed.

Does Etest marks Matter in Placements?

In [12]:
```
plt.figure(figsize = (14,6))
sns.kdeplot(not_placed['etest_p'], label = 'Students not placed', color = 'r', shade = True)
sns.kdeplot(placed_df['etest_p'],label='Students who got placed',color = 'b',shade=True)
plt.xlabel('Marks obtained in etest')
```



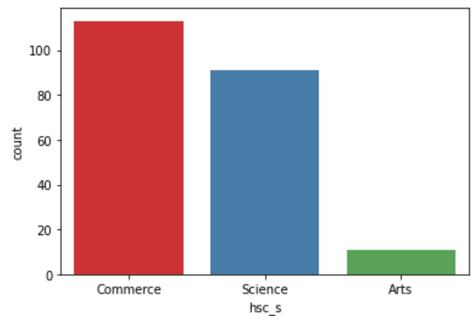
OBSERVATION:

Etest marks cannot be considered as a significant factor as the marks are even distributed along with the placement status

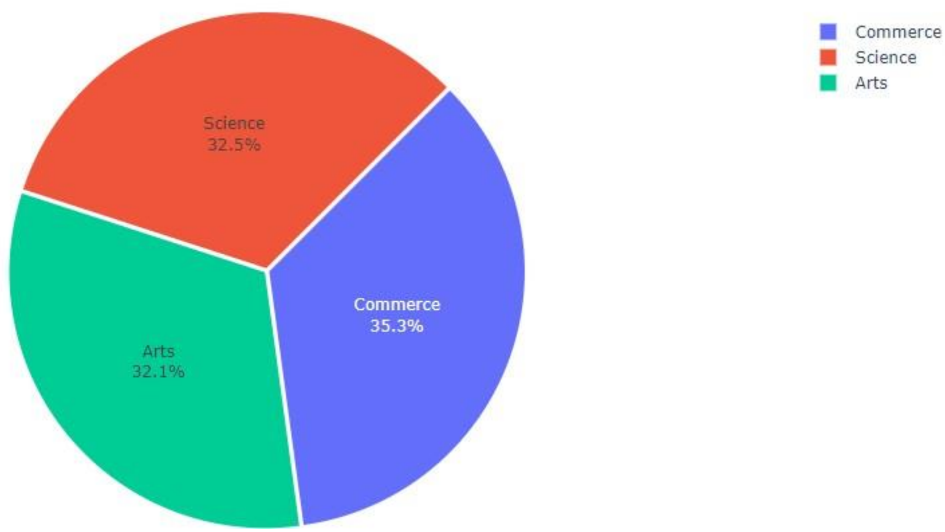
What is the impact of hsc specializations in placement?

In [13]:
```
#count for each specialization
sns.countplot("hsc_s", data=data,palette="Set1")
```

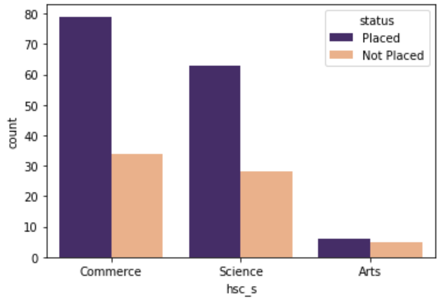
Out[13]: <AxesSubplot:xlabel='hsc_s', ylabel='count'>




```
import plotly.express as px
grdsp = data.groupby(["hsc_s"])[["hsc_p"]].mean().reset_index()

fig = px.pie(grdsp,values="hsc_p",names="hsc_s",)
fig.update_traces(rotation=45, pull=0.01, textinfo="percent+label")
fig.show()
```



In [15]:
```
sns.countplot("hsc_s", hue="status", data=data,palette=['#432371',"#FAAE7B"])
```
Out[15]: <AxesSubplot:xlabel='hsc_s', ylabel='count'>



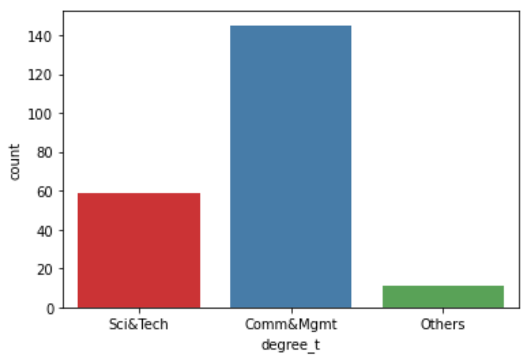
OBSERVATIONS: (I) The most popular branch turns out to be commerce or maybe as most of students get average marks so they were admitted to got commerce on based of their marks. Science is the second most popular and the least popular is arts.

(II) Almost every branch students performed equally but commerce students have slightly better score than other two.

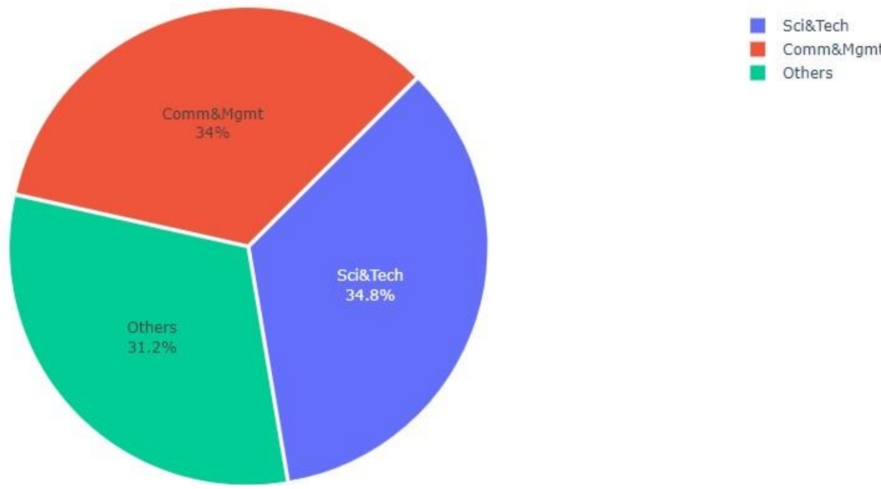
(III) Looking at the fraction of placed and not placed we can say that science branch students have more chance of getting placed than commerce students and most around 45% of the students in arts are not placed

What is the impact of Degree type specializations in placement?

In [17]:
```
sns.countplot("degree_t", data=data,palette="Set1")
```
Out[17]: <AxesSubplot:xlabel='degree_t', ylabel='count'>

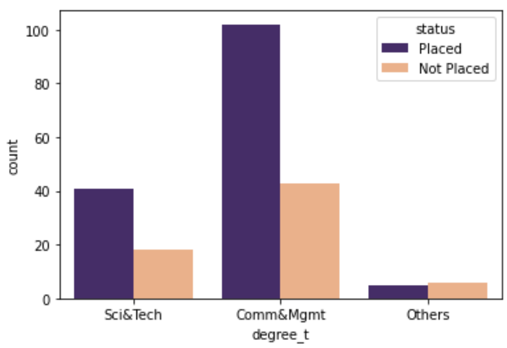


In [16]:
```
grdsp = data.groupby(["degree_t"])[["degree_p"]].mean().reset_index()

fig = px.pie(grdsp,values="degree_p",names="degree_t")
fig.update_traces(rotation=45, pull=0.01, textinfo="percent+label")
fig.show()
```




```
sns.countplot("degree_t", hue="status", data=data,palette=['#432371',"#FAAE7B"])
```

<matplotlib.axes._subplots.AxesSubplot at 0x7f316a015a20>



OBSERVATION:

(I) The students opted for following fields:

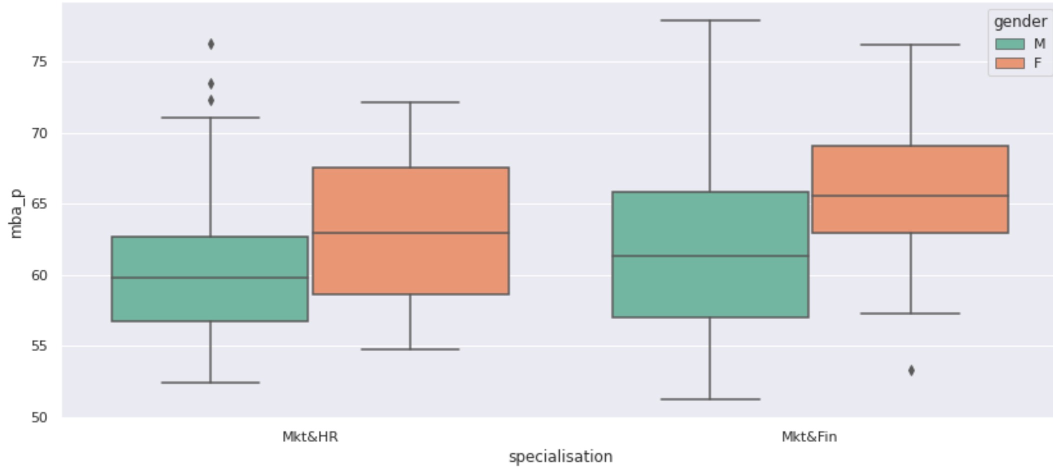
- Science and Technology (must be science students)
- Commerce and management (might be a mixture of commerce and Arts)
- Others

(II) There is not much difference in performace of students from Science and Commerce but there but students who opted for "Others" have low performance

(III) Looks like Commerce and Science degree students are preffered by companies which is obvious. Students who opted for Others have very low placement chance.

What is the distribution of students based on their specialization?

```
[ ]: plt.figure(figsize=(14,6))
     ax = sns.boxplot(x="specialisation", y="mba_p", hue="gender",
                      data=data, palette="Set2")
```



OBSERVATION:

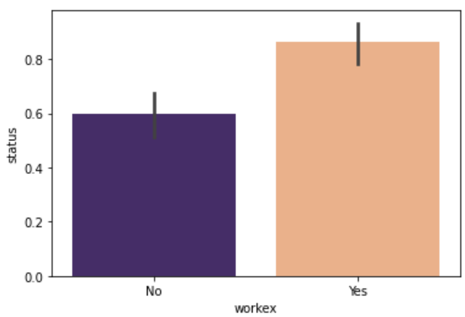
- Females of Mkt and Fin are having higher average mba percentages
- Males of Mkt and HR are having lowest average mba percentages

Does Work Experience Matter in Campus Placement?

Does Work Experience Matter in Campus Placement?

```
In [25]: data['status'] = data['status'].map( {'Placed':1, 'Not Placed':0})
         sns.barplot(x="workex", y="status",data=data,palette=['#432371',"#FAAE7B"])
```

```
Out[25]: <AxesSubplot:xlabel='workex', ylabel='status'>
```



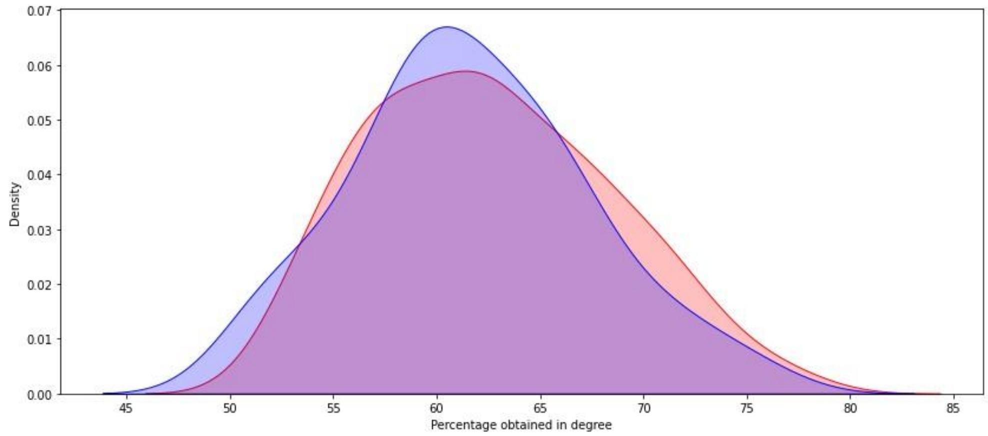
OBSERVATION:

Companies prefer candidates with work experience so the students with internships and past job experience have better chances of being placed.

If i have high MBA percentage, will I get placed?

In [ ]:
```
placed_df = data[data['status']==0]
not_placed = data[data['status']==1]
plt.figure(figsize = (14,6))
sns.kdeplot(not_placed['mba_p'], label = 'Students not placed', color = 'r', shade = True)
sns.kdeplot(placed_df['mba_p'],label='Students who got placed',color = 'b',shade=True)
plt.xlabel('Percentage obtained in degree')
```

Out[30]:
```
Text(0.5, 0, 'Percentage obtained in degree')
```



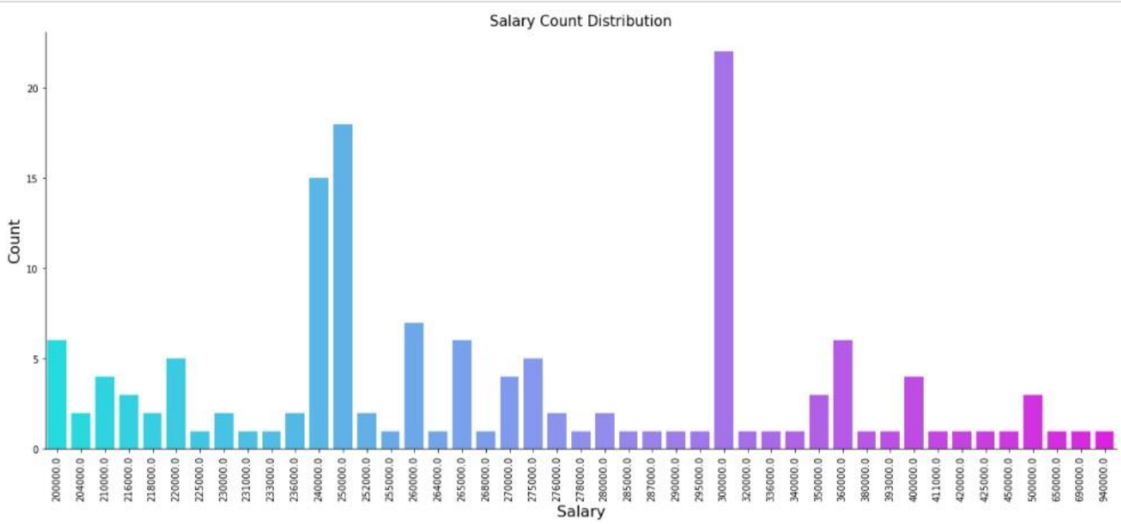
OBSERVATION:

We can see that getting good percentages in MBA does not guarantee placement of the candidate.

SALARY ANALYSIS

What is the package recieved by maximum number of students?

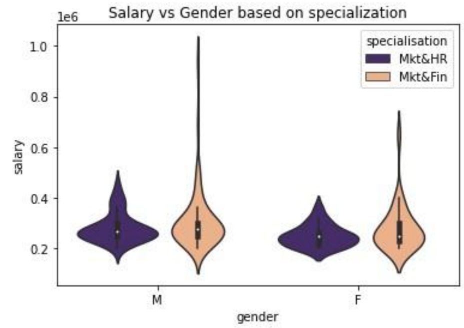
In [ ]:
```
var = 'salary'
fig, ax = plt.subplots()
fig.set_size_inches(20, 8)
plt.xticks(rotation=90);
sns.countplot(x = var,palette="cool", data = data)
ax.set_xlabel('Salary', fontsize=16)
ax.set_ylabel('Count', fontsize=16)
ax.set_title('Salary Count Distribution', fontsize=15)
sns.despine()
```



Salary vs Gender based on specialisation

In [ ]:
```python
sns.violinplot(x=data["gender"], y=data["salary"], hue=data["specialisation"],palette=['#432371',"#FAAE7B"])
plt.title("Salary vs Gender based on specialization")
```

Out[27]: `Text(0.5, 1.0, 'Salary vs Gender based on specialization')`



OBSERVATIONS:

(I) Salary column for male candidates seems to have more outliers than females which means that a lot more male candidates got more than the average CTC.

(II) Mean salary is somewhere around 220k.

(III) Mkt&Fin students are given higher salaries as compared to Mkt&HR.

Gender vs Salary based on work experience

In [ ]:
```python
sns.violinplot(x=data["gender"], y=data["salary"], hue=data["workex"],palette=['#432371',"#FAAE7B"])
plt.title("Gender vs Salary based on work experience")
```

Out[28]: `Text(0.5, 1.0, 'Gender vs Salary based on work experience')`

OBSERVATIONS:

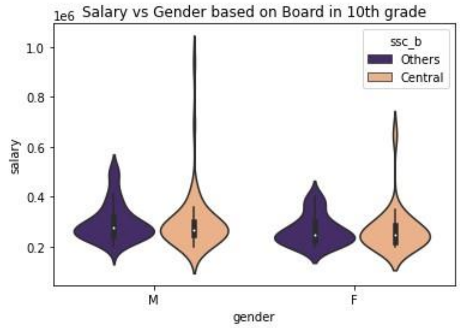
(I) Work Experience is a clear indicator as more work experience results in higher CTC jobs.

(II) The maximum salary in male candidates with experience is >1M and for female it is ~700k. The maximum salary in male candidates without experience is ~550k and for female it is ~430k.

Salary vs Gender based on Board in 10th grade

```
In [ ]: sns.violinplot(x=data["gender"], y=data["salary"], hue=data["ssc_b"],palette=['#432371',"#FAAE7B"])
        plt.title("Salary vs Gender based on Board in 10th grade")
```

```
Out[29]: Text(0.5, 1.0, 'Salary vs Gender based on Board in 10th grade')
```

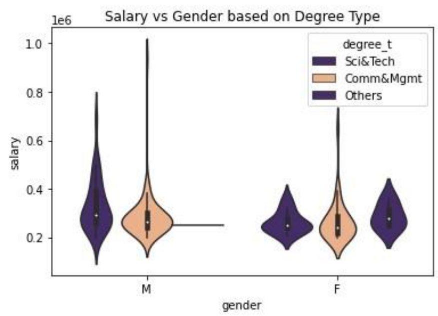


OBSERVATION:

Both Male and Female candidates from Central board got higher CTC as compared to other boards thus we can that central board in 10th grade might fetch you higher CTCs.

Run Code Voila

```
Out[30]: Text(0.5, 1.0, 'Salary vs Gender based on Degree Type')
```



OBSERVATIONS:

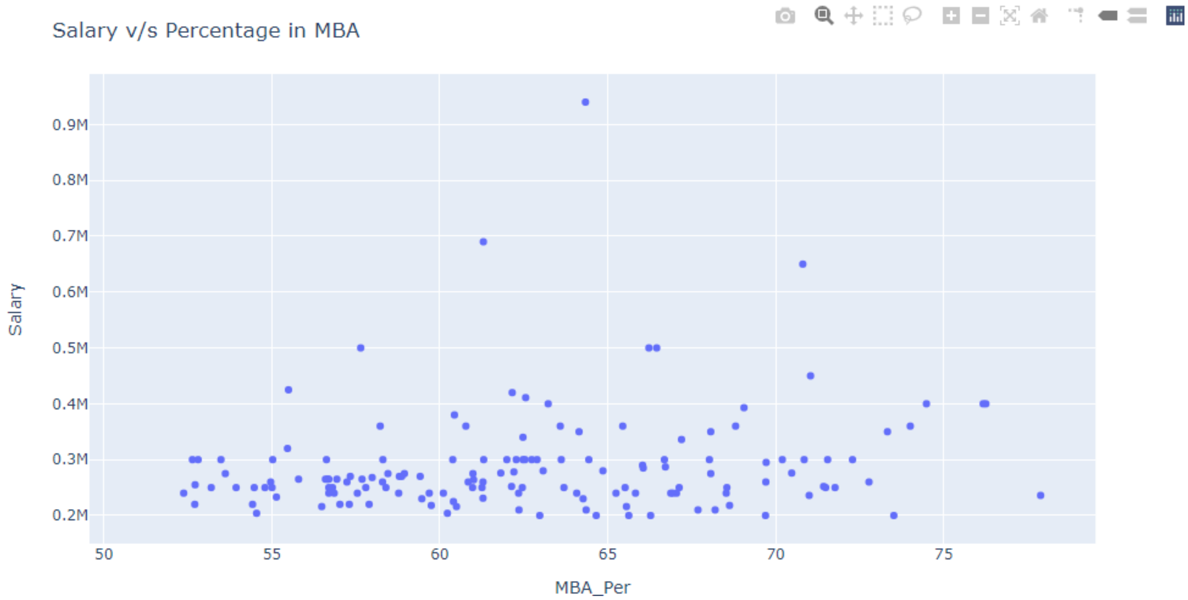
(I) Both male and female candidate got high CTCs choosing Comm&Mgmt as their degree.

(II) Male candidates from Sci&Tech got high CTCs as compared to Female candidates.

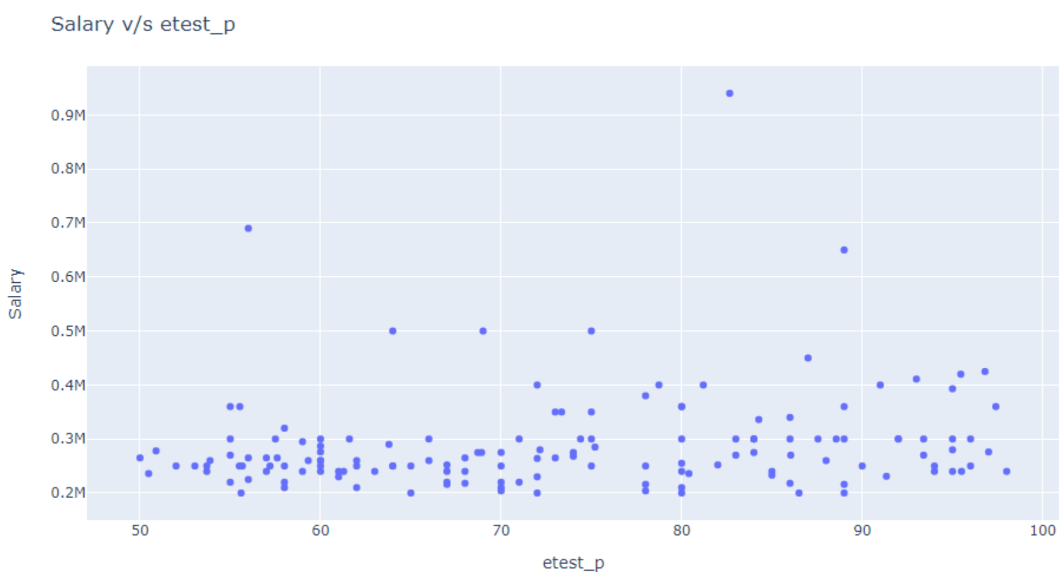
(III) None of the male candidates got placed from "Others" category whereas for female candidates the package is close to what female Sci&Tech candidates got.

got.

In [17]:
```
fig = px.scatter(data,x='mba_p', y='salary')
fig.update_layout(title='Salary v/s Percentage in MBA',xaxis_title="MBA_Per",yaxis_title="Salary")
fig.show()
```

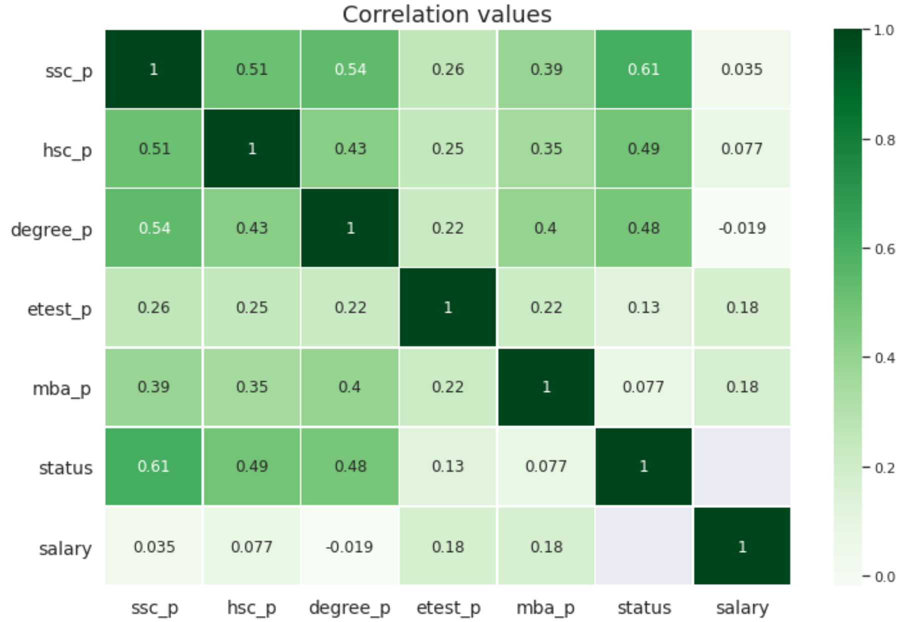


In [18]:
```
fig = px.scatter(data,x='etest_p', y='salary')
fig.update_layout(title='Salary v/s etest_p',xaxis_title="etest_p",yaxis_title="Salary")
fig.show()
```



Is it possible to find relation between numerical values in data set?

In [ ]:
```
plt.figure(figsize=(12, 8))
sns.set(font_scale=1)
correlations = data.corr()
sns.heatmap(correlations,cmap="Greens",linewidths=.5, annot=True)

plt.xticks(fontsize=14, rotation = 0)
plt.yticks(fontsize=14, rotation = 0)
plt.title('Correlation values', fontsize=18)
plt.show()
```



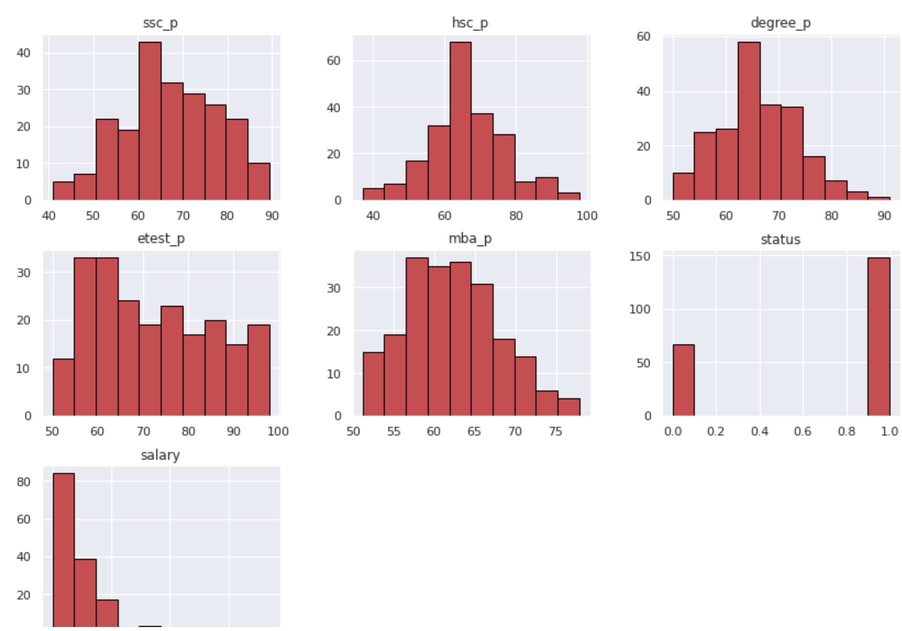
OBSERVATIONS:

The ssc_p,hsc_p,degree_p have higher correlation with status, hence affect the placement procedure more.

Summary (Histogram Distribution)

In [ ]:
```
data.hist(color='r',figsize=(14,10),ec="black")
```

Out[57]:
```
array([[<matplotlib.axes._subplots.AxesSubplot object at 0x7f5589f457b8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f55891dd3c8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f55891e4e80>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f558691c2e8>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f558691da58>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5589f49208>],
       [<matplotlib.axes._subplots.AxesSubplot object at 0x7f55868f8978>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5584df80f0>,
        <matplotlib.axes._subplots.AxesSubplot object at 0x7f5584df8160>]],
      dtype=object)
```



PREDICTING WHETHER A STUDENT WILL GET PLACED OR NOT

Encoding Data

In [4]:
```python
data.drop(['ssc_b','hsc_b', 'salary'], axis=1, inplace=True)
data["gender"] = data.gender.map({"M":0,"F":1})

data["workex"] = data.workex.map({"No":0, "Yes":1})
data["specialisation"] = data.specialisation.map({"Mkt&HR":0, "Mkt&Fin":1})
for column in ['hsc_s', 'degree_t']:
    dummies = pd.get_dummies(data[column])
    data[dummies.columns] = dummies
data.drop(['degree_t','hsc_s'], axis=1, inplace=True)
data.head()
```

Out[4]:

	gender	ssc_p	hsc_p	degree_p	workex	etest_p	specialisation	mba_p	status	Arts	Commerce	Science	Comm&Mgmt	Others	Sci&Tech
0	0	67.00	91.00	58.00	0	55.0	0	58.80	Placed	0	1	0	0	0	1
1	0	79.33	78.33	77.48	1	86.5	1	66.28	Placed	0	0	1	0	0	1
2	0	65.00	68.00	64.00	0	75.0	1	57.80	Placed	1	0	0	1	0	0
3	0	56.00	52.00	52.00	0	66.0	0	59.43	Not Placed	0	0	1	0	0	1
4	0	85.80	73.60	73.30	0	96.8	1	55.50	Placed	0	1	0	1	0	0

In [5]:
```python
data["status"] = data.status.map({"Not Placed":0,"Placed":1})
```

In [6]:
```python
data.drop(['Others', 'Arts'], axis=1, inplace=True)
```

In [8]:
```python
y = data['status']
data.drop('status', axis = 1, inplace = True)
sc = StandardScaler()
X = sc.fit_transform(data)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, shuffle=True)
```

In [9]:
```python
print("X-Train:",X_train.shape)
print("X-Test:",X_test.shape)
print("Y-Train:",y_train.shape)
print("Y-Test:",y_test.shape)
```

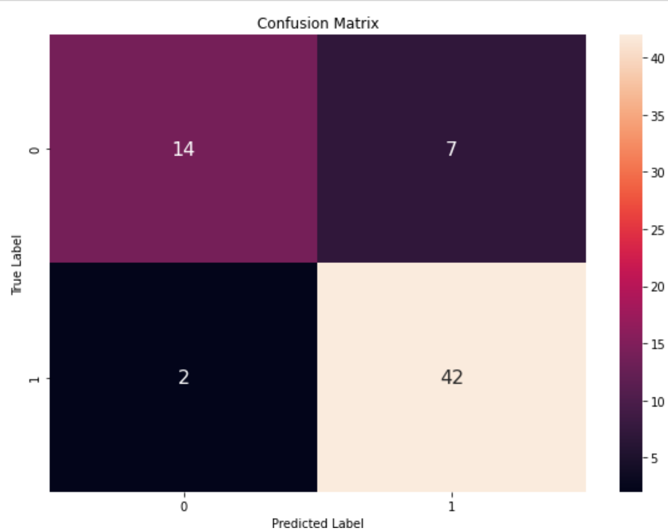
```
X-Train: (150, 12)
X-Test: (65, 12)
Y-Train: (150,)
Y-Test: (65,)
```

In [10]:
```python
log_reg = LogisticRegression()
log_reg.fit(X_train, y_train)
```

Out[10]: `LogisticRegression()`

In [11]:
```python
y_pred=log_reg.predict(X_test)
```

In [12]:
```python
conf_mat = pd.DataFrame(confusion_matrix(y_test, y_pred))
fig = plt.figure(figsize=(10,7))
sns.heatmap(conf_mat, annot=True, annot_kws={"size": 16}, fmt='g')
plt.title("Confusion Matrix")
plt.xlabel("Predicted Label")
plt.ylabel("True Label")
plt.show()
```



OBSERVATIONS: Our confusion Matrix looks decent. We have correctly predicted 42 (placed) + 14 (not-placed) correct predictions and 7 (not placed as placed) + 2(placed as not-placed) incorrect predictions.

We need to decrease these incorrect predictions because a good candidate can be rejected (false positive) and a unfit candidate can be selected (false negatives)

In [13]:
```
print(classification_report(y_test, y_pred))
```

```
              precision    recall  f1-score   support

           0       0.88      0.67      0.76        21
           1       0.86      0.95      0.90        44

    accuracy                           0.86        65
   macro avg       0.87      0.81      0.83        65
weighted avg       0.86      0.86      0.86        65
```

In [14]:
```
accuracy = accuracy_score(y_pred, y_test)

print(f"The accuracy : {np.round(accuracy, 4)*100.0}%")
```

```
The accuracy : 86.15%
```

Decision Tree Classifier

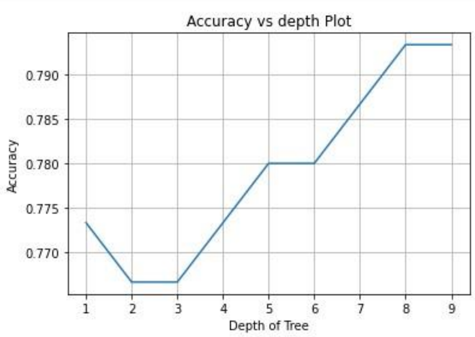
Let's try some decision trees now and see how well they perform but as Decision trees are easy to overfit so I will use K-FOLD CV first to find the best depth.

In [15]:
```
depth = list(range(1,10))
cv_scores = []
for d in depth:
  dt = DecisionTreeClassifier(criterion="gini", max_depth=d, random_state=42)
  scores = cross_val_score(dt, X_train, y_train, cv=10, scoring='accuracy', n_jobs = -1)
  cv_scores.append(scores.mean())
# finding the optimal depth
optimal_depth = depth[cv_scores.index(max(cv_scores))]
print("The optimal depth value is: ", optimal_depth)
```

```
The optimal depth value is:  8
```


```
The optimal depth value is:  8
```


In [16]:
```
# plotting accuracy vs depth
plt.plot(depth, cv_scores)
plt.xlabel("Depth of Tree")
plt.ylabel("Accuracy")
plt.title("Accuracy vs depth Plot")
plt.grid()
plt.show()

print("Accuracy scores for each depth value is : ", np.round(cv_scores, 3))
```




```
Accuracy scores for each depth value is :  [0.773 0.767 0.767 0.773 0.78  0.78  0.787 0.793 0.793]
```

In [17]:
```
dt_optimal = DecisionTreeClassifier(criterion="gini", max_depth=optimal_depth, random_state=42)


dt_optimal.fit(X_train,y_train)


y_pred = dt_optimal.predict(X_test)


accuracy = accuracy_score(y_test, y_pred)*100
print(f"The accuracy on test set using optimal depth = {optimal_depth} is {np.round(accuracy, 3)}%")
```

```
The accuracy on test set using optimal depth = 8 is 86.154%
```

```python
accuracy = accuracy_score(y_test, y_pred)*100
print(f"The accuracy on test set using optimal depth = {optimal_depth} is {np.round(accuracy, 3)}%")
```

```
The accuracy on test set using optimal depth = 8 is 86.154%
```

We achieved 86% accuracy which is similiar to what we achieved using logistic regression so they seem to work equally well.

What if we could combine the power of two models to get better results?

Ensemble Modelling

We will train a voting classifier using our previously trained logistic regeression and Decision tree model

In [18]:
```python
ensembles = [log_reg, dt_optimal]


for estimator in ensembles:
    print("Training the", estimator)
    estimator.fit(X_train,y_train)
```

```
Training the LogisticRegression()
Training the DecisionTreeClassifier(max_depth=8, random_state=42)
```

In [19]:
```python
scores = [estimator.score(X_test, y_test) for estimator in ensembles]

scores
```

Out[19]:
```
[0.8615384615384616, 0.8615384615384616]
```

In [20]:
```python
from sklearn.ensemble import VotingClassifier

named_estimators = [
    ("log_reg",log_reg),
    ("dt_tree", dt_optimal),

]
```

In [21]:
```python
voting_clf = VotingClassifier(named_estimators)
```

In [22]:
```python
voting_clf.fit(X_train,y_train)
```

Out[22]:
```
VotingClassifier(estimators=[('log_reg', LogisticRegression()),
                             ('dt_tree',
                              DecisionTreeClassifier(max_depth=8,
                                                     random_state=42))])
```


```
]
```

In [21]:
```python
voting_clf = VotingClassifier(named_estimators)
```

In [22]:
```python
voting_clf.fit(X_train,y_train)
```

Out[22]:
```
VotingClassifier(estimators=[('log_reg', LogisticRegression()),
                             ('dt_tree',
                              DecisionTreeClassifier(max_depth=8,
                                                     random_state=42))])
```


In [25]:
```python
acc = voting_clf.score(X_test,y_test)
print(f"The accuracy on test set using voting classifier is {np.round(acc, 4)*100}%")
```

```
The accuracy on test set using voting classifier is 92.31%
```

We went from 86.4% to 92.3% accuracy score!

Hence, ensemble modelled voting classifier of Logistic and decision tree helped us increase the accuracy of the prediction model

Conclusions Drawn

- More male candidates got placed as compared to female candidates.
- Male Candidates got higher CTCs as compared to female candidates.
- Type of Board choosen does not have any effect on placements thus we can drop in preprocessing steps.
- Most of the students preferred Central board in 10th grade whereas other boards in 12th grade.
- Candidates with higher percentages have better chance of placements.
- Choosing Science and Commerce as Specialisation seems to have perk when it comes to placments.
- Maximum package was bagged by male candidate from Mkt&Fin branch which is around 940k.
- Commerce is the most popular branch among candidates.
- Mean CTC is around 220k for male and female candidates individually.
- Choosing Sci&Tech and Comm&Mngmt as degree will fetch you higher CTCs.
- Mkt&Fin major have higher salaries and more placement chance as compared to Mkt&HR.
- Employability test percentage and MBA percentage does not effect the placements

In [ ]:
```python

```