MATH2319 - Machine Learning

Campus Placement Prediction based on Students' Academic Performances - Phase 1

Project Group 74

Subbiah Soundarapandian(S3825012), Sudershan Ravi(S3829895)

Table of Contents:

Introduction

Campus placement plays a major role in deciding the future of the Under-graduate and Post-graduate students in the countries like India. This analysis aims at predicting if a student would get recruited in the campus placement or not, based on their previous academic achievements and grades. In the phase 1, we will focus at the basic and important step in the process of machine learning such as data cleansing and data preparation as well as have a look at some of the visualizations to get a grip on the analysis to be followed in phase 2.

Dataset Source:

The dataset used in this project is taken from the website kaggle[1]

Dataset Description:

The dataset, **'Placement_Data_Full_Class.csv'** is referred for this project. It contains 13 features and 215 observations which describe the various attributes of each student appearing for the placement interviews. It includes secondary and higher secondary school percentages and specialization along with the under graduation and post graduation percentages and specializations and work experience of the students.

Dataset Features:

Name	Data Type	Units	Brief Description

Name	Data Type	Units	Brief Description
gender	Object	N/A	Gender of the student
ssc_p	Float64	%	Secondary Education Percentage
ssc_b	Object	N/A	Board of Secondary Education
hsc_p	Float64	%	Higher Secondary Education Percentage
hsc_b	Object	N/A	Board of Higher Secondary Education
hsc_s	Object	N/A	Specialization in Higher Secondary Education
degree_p	Float64	%	Under Graduation Degree Percentage
degree_t	Object	N/A	Under Graduation Degree Type (Field of Study)
workex	Object	N/A	Work Experience
etest_p	Float64	%	Employability Test Percentage
specialisation	Object	N/A	Post Graduation (MBA) Specialization
mba_p	Float64	%	MBA Percentage
status	Object	N/A	Status of Placement

Target Feature:

The objective of this project is to predict whether a student would get placed or not. Hence, we choose the feature, **'status'** as the response (target) feature.

Goals and Objectives

The primary objective for modelling this particular data is to try and predict if a student would get placed or not after the completion of his/her post graduation(MBA) depending on his/her field of study, academic performances, work experience and employability score.

Data Cleaning & Preprocessing

All the required python libraries are imported using the 'import' command. Hence, libraries such as 'warnings','pandas','numpy','matplotlib' and 'seaborn' are imported.

In [1]:

```
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

The **'Placement_Data_Full_Class.csv'** dataset is imported for modelling.

In [2]:

```
assignds = pd.read_csv("Placement_Data_Full_Class.csv")
```

Setting the serial number column as index and checking the dataframe.

In [3]:

```
assignds.set_index('sl_no', inplace = True)
assignds.head(10)
```

Out[3]:

	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p

sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p
sl_no										
1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.00
2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.50
3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.00
4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.00
5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.80
6	M	55.00	Others	49.80	Others	Science	67.25	Sci&Tech	Yes	55.00
7	F	46.00	Others	49.20	Others	Commerce	79.00	Comm&Mgmt	No	74.28
8	M	82.00	Central	64.00	Central	Science	66.00	Sci&Tech	Yes	67.00
9	M	73.00	Central	79.00	Central	Commerce	72.00	Comm&Mgmt	No	91.34
10	M	58.00	Central	70.00	Central	Commerce	61.00	Comm&Mgmt	No	54.00

Dimensions of the dataframe is verified.

In [4]:
```
assignds.shape
```

Out[4]:
```
(215, 14)
```

General information of the dataframe

In [5]:
```
assignds.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 215 entries, 1 to 215
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: 25.2+ KB
```

Removing the **'salary'** feature since it is not a predictor, as it is only for placed people.

In [6]:
```
assignds.drop(['salary'],axis = 1,inplace = True)
```

Checking for NA values in all the features.

In [7]:
```
assignds.isna().sum()
```

Out[7]:
```
gender            0
ssc_p             0
ssc_b             0
hsc_p             0
hsc_b             0
hsc_s             0
degree_p          0
degree_t          0
workex            0
etest_p           0
specialisation    0
mba_p             0
status            0
dtype: int64
```

Fortunately, we do not have any missing values in the dataset.

Summary of all the numerical variables.

In [8]:
```
numerical_var = assignds.select_dtypes(include = 'float64')
numerical_var.describe()
```

Out[8]:

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

Detecting outliers using the z-scores.[2]

In [9]:
```
from scipy import stats
z = np.abs(stats.zscore(numerical_var))
print(z)
z.shape
```

```
[[0.02808697 2.2688123  1.14010225 1.29109087 0.59764672]
 [1.11336869 1.10344799 1.51326671 1.08715679 0.6876202 ]
 [0.21323793 0.15331275 0.32284282 0.21890765 0.76947385]
 ...
 [0.02808697 0.06133451 0.90304633 0.98909117 1.27870553]
 [0.61994138 0.03064373 1.14010225 0.15859198 0.35193393]
 [0.49096436 0.76646966 1.82115177 1.27590661 0.3536522 ]]
```

Out[9]:
```
(215, 5)
```

Printing the outliers and then removing them.

In [10]:
```
print(np.where(abs(z)>3))

assignds.drop(197,inplace = True)
```

```
(array([197], dtype=int64), array([2], dtype=int64))
```

Checking the dataframe after removing the outlier.

```
In [11]: assignds.shape
```

```
Out[11]: (214, 13)
```

Checking the number of positive and negative target variables.

```
In [12]: # Counts of Target variable

         assignds['status'].value_counts()
```

```
Out[12]: Placed        147
         Not Placed     67
         Name: status, dtype: int64
```

Data Exploration & Visualisation

Now that the dataset is cleaned and ready for modeling, we can first start exploring the data by visualizing different attributes

Filtering the dataset to analyse the students who got placed.

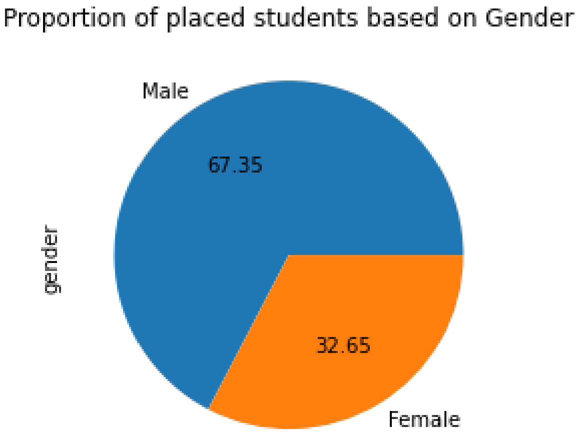
```
In [13]: assignds_filter = assignds.loc[assignds['status'] == 'Placed']
         assignds_filter.shape
```

```
Out[13]: (147, 13)
```

Univariate Visualisation

The proportion of the two major genders in the students who were placed is explored by plotting a pie chart as below. it is discovered that 67.35% of the students placed were male while 32.65% of female students were placed.

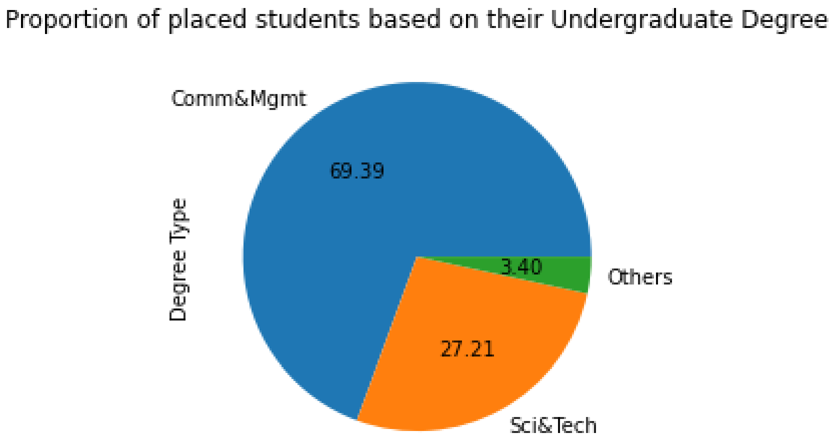
```
In [14]: assignds_filter['gender'].value_counts().plot(kind='pie',autopct='%.2f',labels = ["M
         plt.title('Proportion of placed students based on Gender')
         plt.show()
```



Plotting a pie chart to explore the proportion of students placed based on their under graduation specialization. Upon plotting the proportional distribution of 'Under Graduation Type' in the placed students, it is evident that students with a bachelor's degree in Commerce and Management acquired more placement than any other disciplines with a clear majority of 69.39% which is followed by Science and Technology with 27.21% whereas the other fields of study contributed to 3.40%.

In [15]:

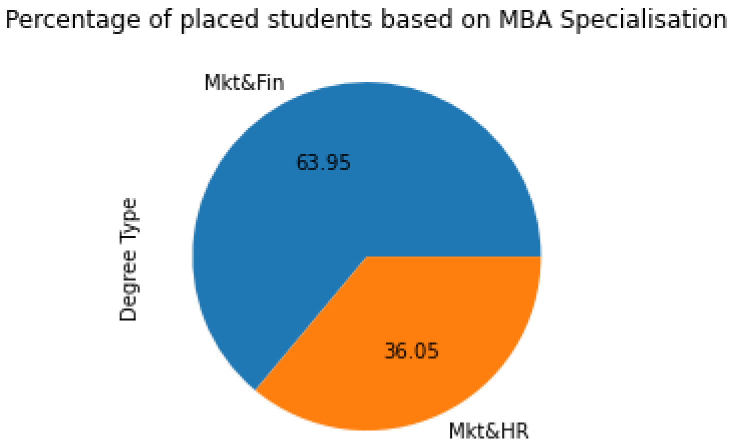
```
assignds_filter['degree_t'].value_counts().plot(kind='pie',autopct='%.2f')
plt.title('Proportion of placed students based on their Undergraduate Degree')
plt.ylabel("Degree Type")
plt.show()
```



The percentage of students placed based on the specialization of MBA is plotted as a pie chart. The below plot provides us a clear perspective about the distribution of proportion of post graduation specialization among the placed students and the fact that students with a master's degree in Marketing and Finance were placed more than their counterparts at Marketing and HR.

In [16]:

```
assignds_filter['specialisation'].value_counts().plot(kind='pie',autopct='%.2f')
plt.title('Percentage of placed students based on MBA Specialisation')
plt.ylabel("Degree Type")
plt.show()
```

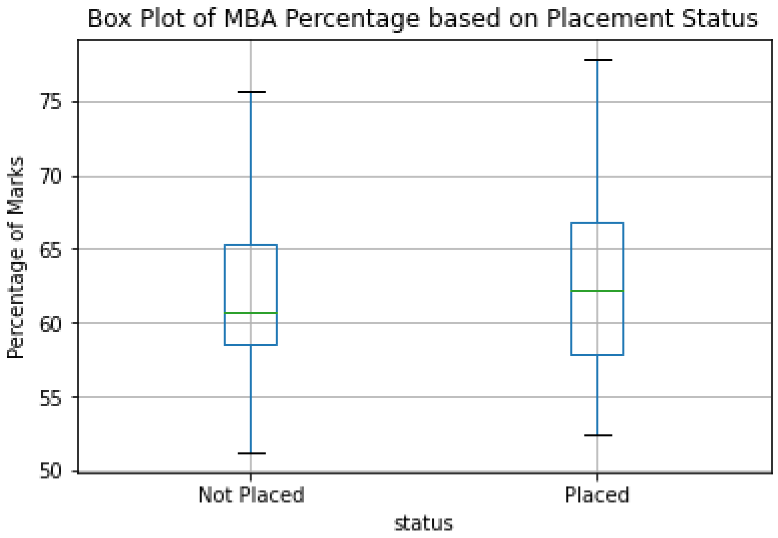


Bivariate Visualisation

The percentages scored by the students who were placed and those who were not placed is compared by plotting a box plot. On reviewing the box plots of students based on placement status, it is surprising to observe that the difference in the average percentage scored in MBA by placed and unplaced is quite lower than expected. This could be of very little significant while analyzing in the phase 2.

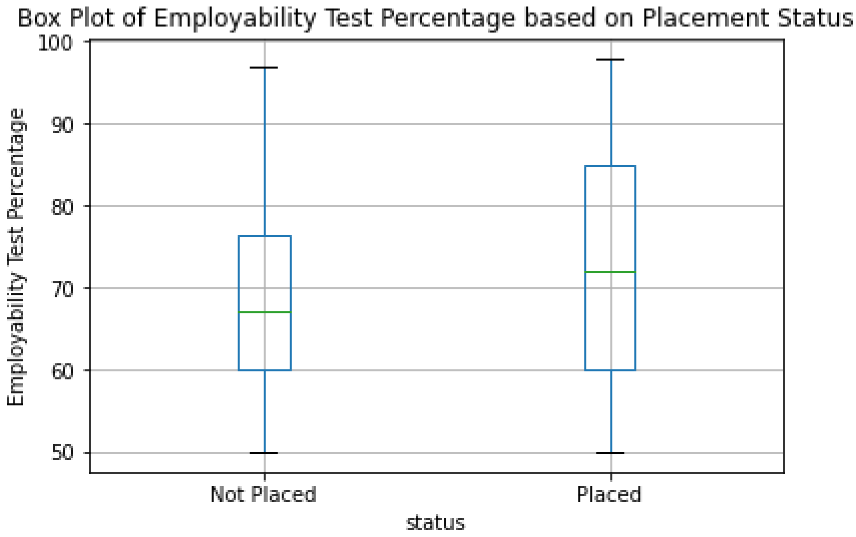
In [17]:

```
assignds.boxplot(column='mba_p',by='status')
plt.title("Box Plot of MBA Percentage based on Placement Status")
plt.ylabel('Percentage of Marks')
plt.suptitle(' ')
plt.show()
```



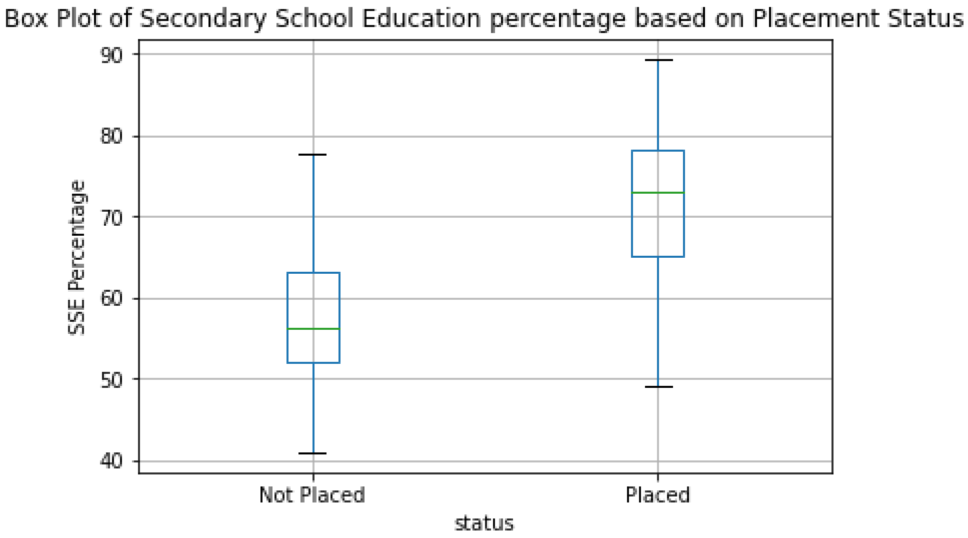
The distribution of students who were placed and not placed were examined based on the percentage scored in the employability test. The box plot makes it evident that the average Employability Test percentage scored by placed students were higher than that of unplaced students. From this, we could sense that the Employability test percentage could be one of the major predictive features deciding the campus placement of a student.

In [18]:
```
assignds.boxplot(column='etest_p',by='status')
plt.title("Box Plot of Employability Test Percentage based on Placement Status")
plt.suptitle(' ')
plt.ylabel('Employability Test Percentage')
plt.show()
```



Another box plot is created to explore the distribution of secondary school percentages of students who were placed and not placed respectively. It is quite evident by reviewing the plot below that the difference in the average percentage scored is quite high. The Students who were placed had scored more marks in their SS Exams than that of the not-placed students.

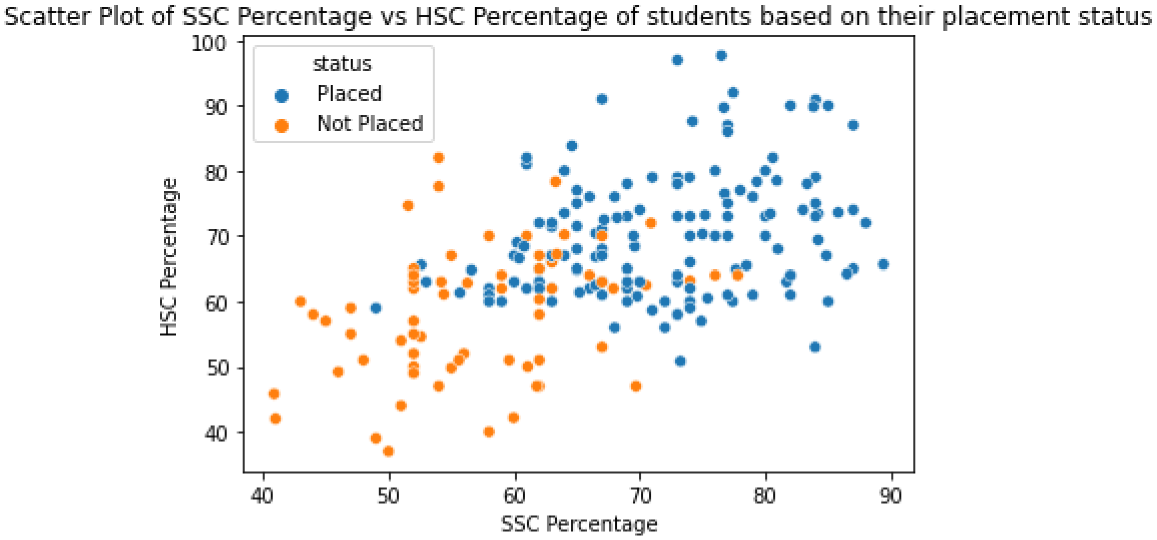
In [19]:
```
assignds.boxplot(column='ssc_p',by='status')
plt.title("Box Plot of Secondary School Education percentage based on Placement Stat
plt.suptitle(' ')
plt.ylabel('SSE Percentage')
plt.show()
```



Trivariate Visualisation

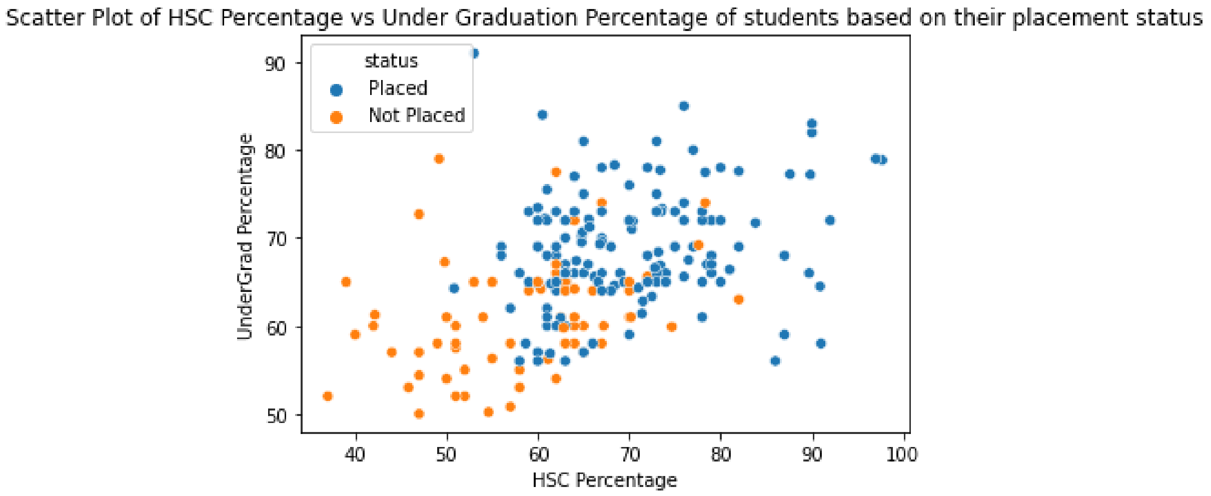
A scatter plot is created to explore the correlation between the percentages scored in secondary school and higher secondary school by students who were placed and not placed respectively. It is quite apparent that there exists, a positive correlation between Higher Secondary School percentage and Secondary School percentage among both the placed and unplaced students. And we could easily see that there are two clusters formed between the placed and non-placed students.

In [20]:
```
sns.scatterplot('ssc_p', 'hsc_p', data=assignds, hue='status')
plt.title("Scatter Plot of SSC Percentage vs HSC Percentage of students based on the
plt.xlabel("SSC Percentage")
plt.ylabel("HSC Percentage")
plt.show()
```



The correlation between the under graduation percentage and higher secondary school percentages of students based on their placement status is explored with the help of a scatter plot. The scatter plot reveals the high positive correlation between the Under Graduation percentage and Higher Secondary School percentage among the placed and unplaced students. We could see that students with more Higher Secondary marks and Undergraduate percentage are more like to get placed in the campus placement that that of students with less marks.

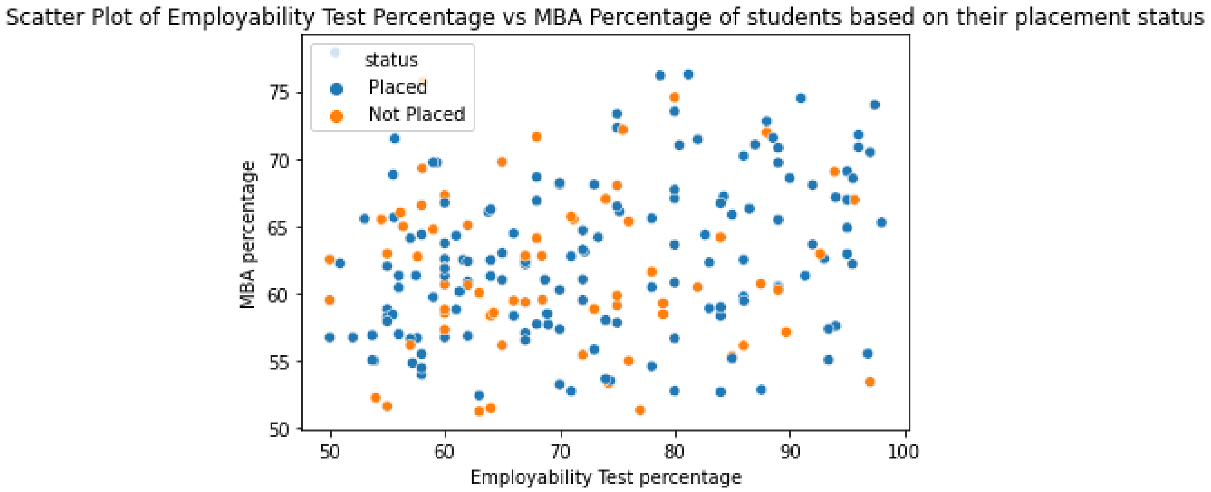
In [21]:
```
sns.scatterplot('hsc_p', 'degree_p', data=assignds, hue='status')
plt.title("Scatter Plot of HSC Percentage vs Under Graduation Percentage of students
plt.xlabel("HSC Percentage")
plt.ylabel("UnderGrad Percentage")
plt.show()
```



It is investigated that if the Employability Test percentage is correlated to the percentage scored in post graduation(MBA) by the students based on the placement status. It is discovered that there is no statistically significant correlation between Post Graduation(MBA) percentage and the Employability Test percentage scored by the students appearing for placements.

In [22]:

```
sns.scatterplot('etest_p', 'mba_p', data=assignds, hue='status')
plt.title("Scatter Plot of Employability Test Percentage vs MBA Percentage of studen
plt.xlabel("Employability Test percentage")
plt.ylabel("MBA percentage")
plt.show()
```



Summary & Conclusions

The initial analysis of the dataset in question revealed that a majority of the students were placed as 68.84% of students were placed. Upon further examination of the features and their interrelationships we acquire a better perspective about the factors affecting the placement statuses of the students. Hence after the completion of phase 1, it is safe to assume that male students with a bachelor's degree with Commerce and Management and a master's degree in Marketing and Finance had the highest probability of being placed. With a good lead, we can proceed with phase 2 of the analysis of campus placement prediction

References

[1] Roshan, B. (2021). Campus Recruitment. Retrieved 5 April 2021, from https://www.kaggle.com/benroshan/factors-affecting-campus-placement

[2] Singh, D., & Outliers, C. (2021). Cleaning up Data Outliers with Python | Pluralsight. Retrieved 7 April 2021, from https://www.pluralsight.com/guides/cleaning-up-data-from-outliers