Revelation 23 BrainDead

BrainDead is the flagship Data Analysis and Machine Learning competition of Revelation '23.

Binary Brains

Problem Statement 1 : Analyze Placement Data

Challenge Description:

In this challenge, you are supposed to analyze the placement records of the students of a MBA college. The dataset includes secondary and higher secondary school percentages and specializations. It also contains degree specialization, work experience, and the salary offered to the students. Your main task is to analyze the factors that affect the placement and salary of students.

Introduction

The purpose of this report is to analyze the placement records of the students of a MBA college. The dataset includes secondary and higher secondary school percentages and specializations, degree specialization, work experience, and the salary offered to the students. The main task is to analyze the factors that affect the placement and salary of students.

Dataset Exploration:

The dataset consists of 215 rows and 8 columns. The columns include Secondary Education Percentage, Higher Secondary Education Percentage, Degree Specialization, Work Experience, MBA Percentage, MBA Specialization, Status, and Salary. The dataset was cleaned and checked for missing values and outliers.

Link: https://drive.google.com/drive/folders/1aB9Z6frlz3-F2_P32pg5TF7YLMiJXpKZ?usp=sharing

The columns and their description:

- sl_no: Serial Number
- gender: Gender- Male='M',Female='F'
- ssc_p: Secondary Education percentage- 10th Grade
- ssc_b: Board of Education- Central/ Others
- hsc_p: Higher Secondary Education percentage- 12th Grade
- hsc_b: Board of Education- Central/ Others
- hsc_s: Specialization in Higher Secondary Education
- degree_p: Degree Percentage
- degree_t: Under-Graduation(Degree type)- Field of degree education
- workex: Work Experience
- etest_p: Employability test percentage (conducted by the college)
- specialisation: Post Graduation(MBA)- Specialization
- mba_p: MBA percentage
- status: Status of placement- Placed/Not placed
- salary: Salary offered by corporate to candidates

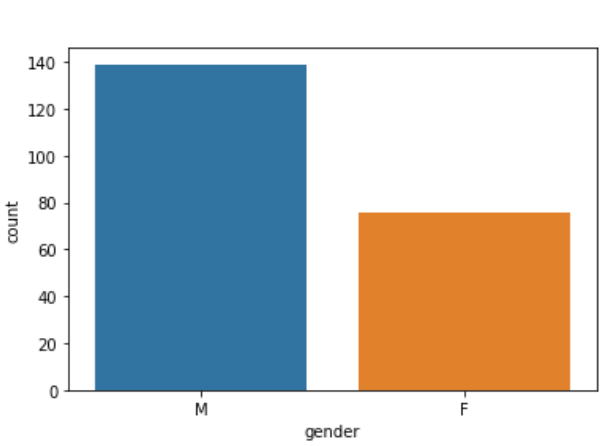
	sl_no	gender	ssc_p	ssc_b	hsc_p	hsc_b	hsc_s	degree_p	degree_t	workex	etest_p	specialisation	mba_p	status	salary
0	1	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed	270000.0
1	2	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed	200000.0
2	3	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed	250000.0
3	4	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed	NaN
4	5	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed	425000.0

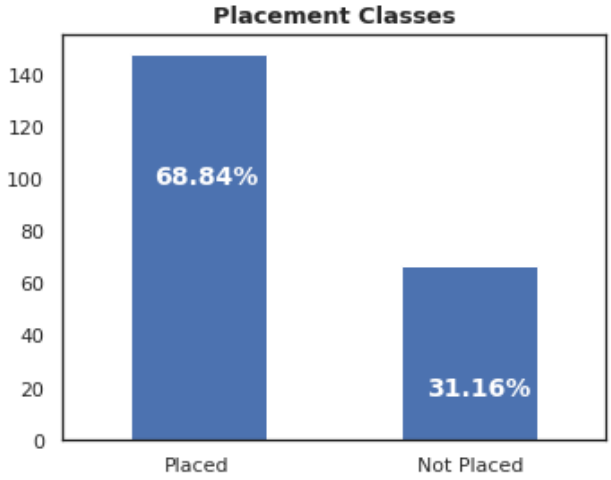
	sl_no	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
count	215.000000	215.000000	215.000000	215.000000	215.000000	215.000000	148.000000
mean	108.000000	67.303395	66.333163	66.370186	72.100558	62.278186	288655.405405
std	62.209324	10.827205	10.897509	7.358743	13.275956	5.833385	93457.452420
min	1.000000	40.890000	37.000000	50.000000	50.000000	51.210000	200000.000000
25%	54.500000	60.600000	60.900000	61.000000	60.000000	57.945000	240000.000000
50%	108.000000	67.000000	65.000000	66.000000	71.000000	62.000000	265000.000000
75%	161.500000	75.700000	73.000000	72.000000	83.500000	66.255000	300000.000000
max	215.000000	89.400000	97.700000	91.000000	98.000000	77.890000	940000.000000

```
df.gender.value_counts()
```

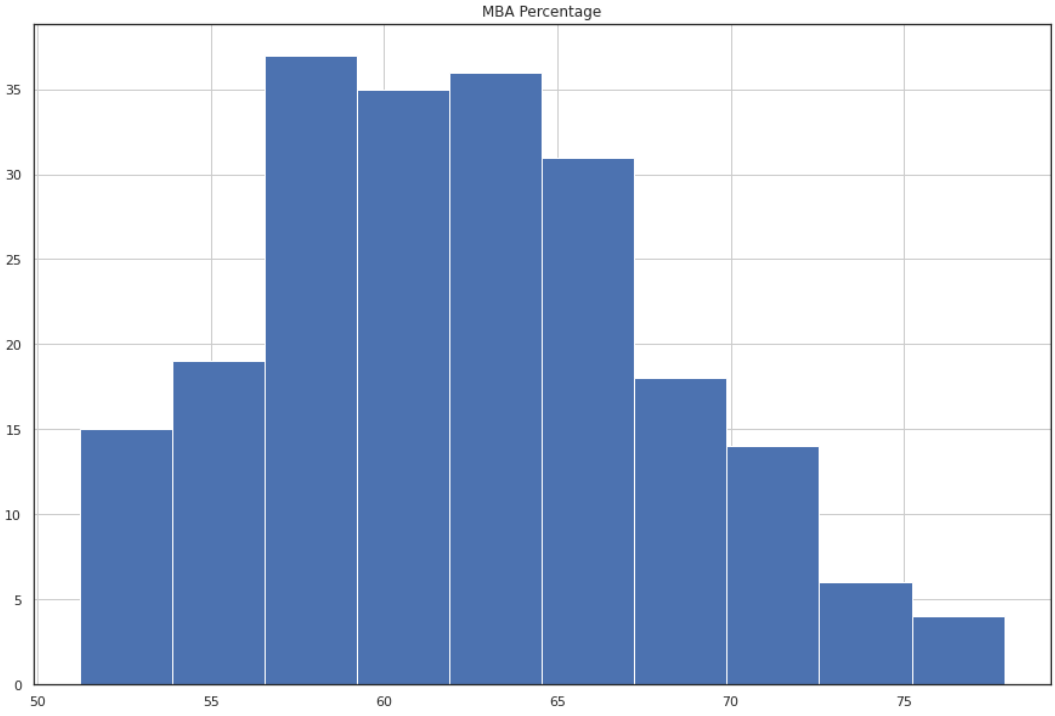
```
M    139
F     76
Name: gender, dtype: int64
```

Describes the dataset and gives the sex ratio of the Placements with 139 total male candidates sitting for placements to 76 female candidates sitting for placement.

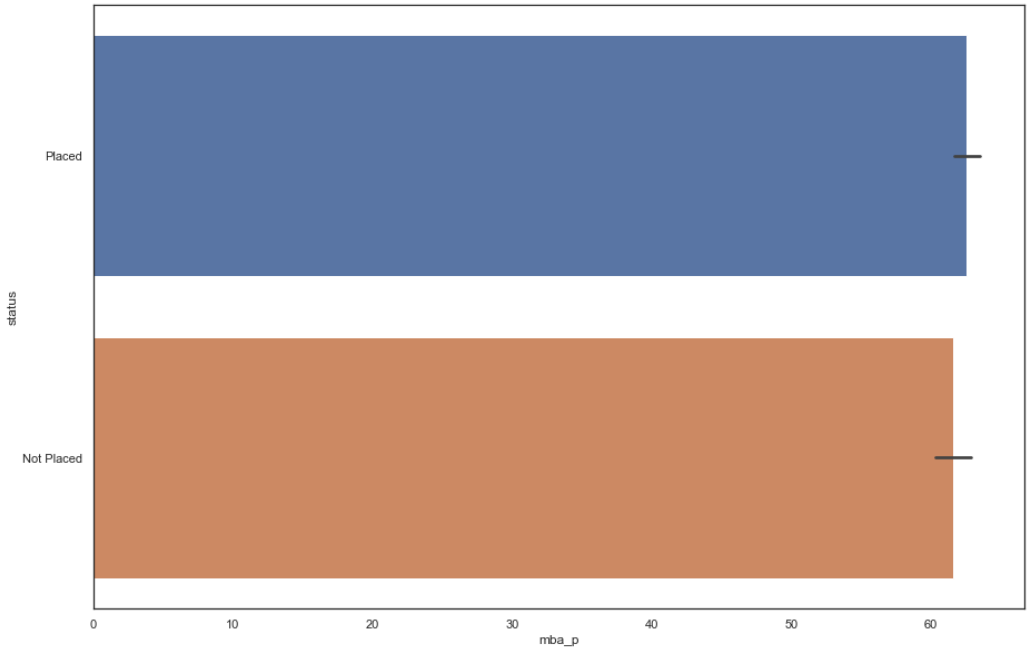


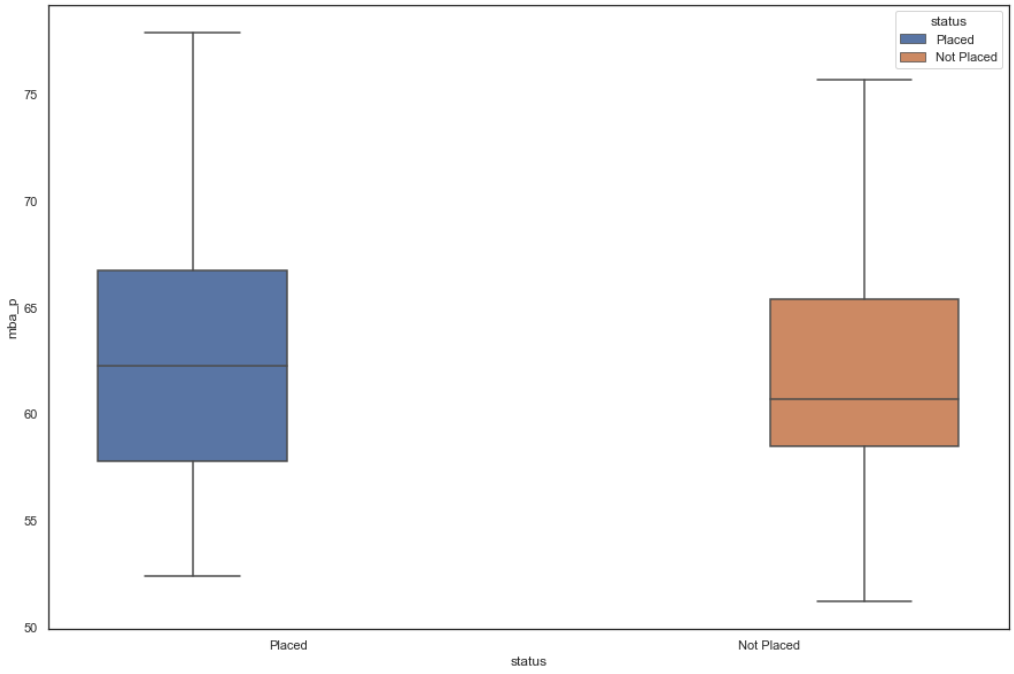


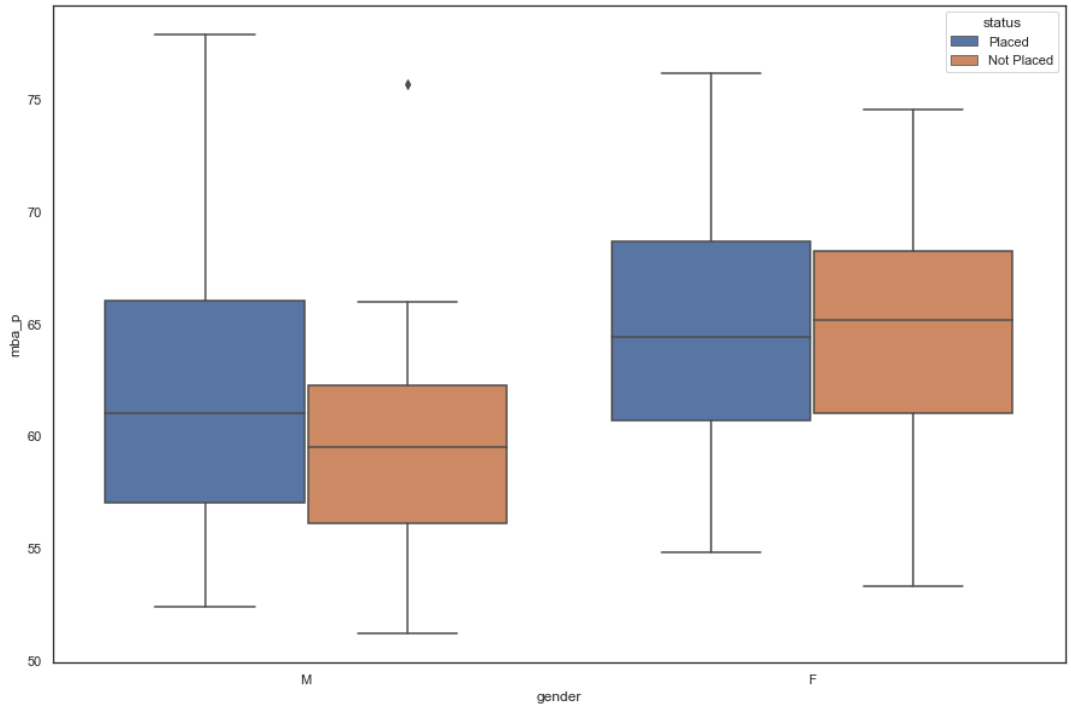
The Above graph shows the gender count for for placement process and the percentage count of the getting placed and not placed.



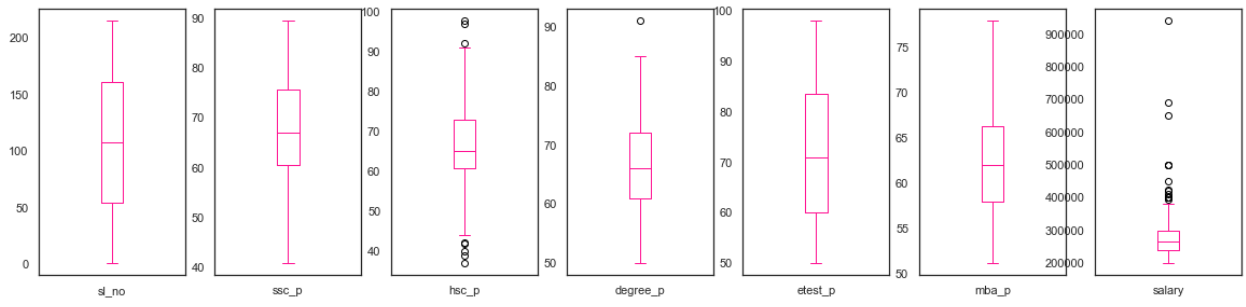
The Above graph shows the MBA percentage histogram distribution for all the applicants and this shows that the average MBA percentage of the applicants lies between 55% to 70% and the distribution varies with majority saturation in 60% area.


Placed
Not Placed
status
0
10
20
30
40
50
60
mba_p

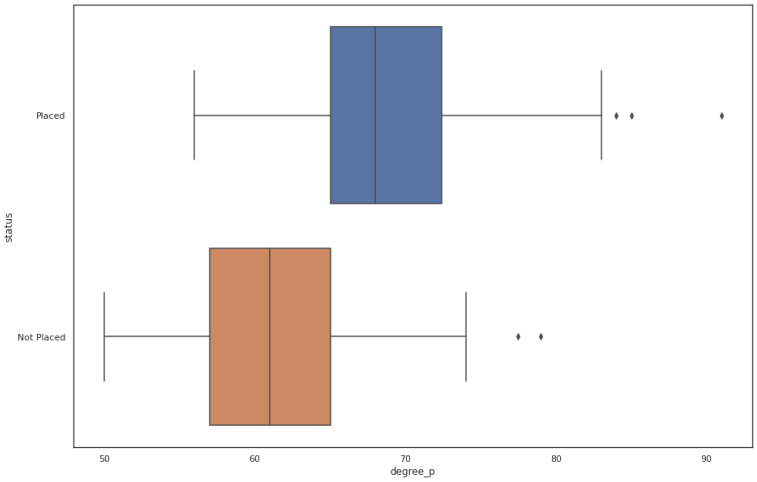

status
Placed
Not Placed
75
70
65
60
55
50
mba_p
Placed
Not Placed
status

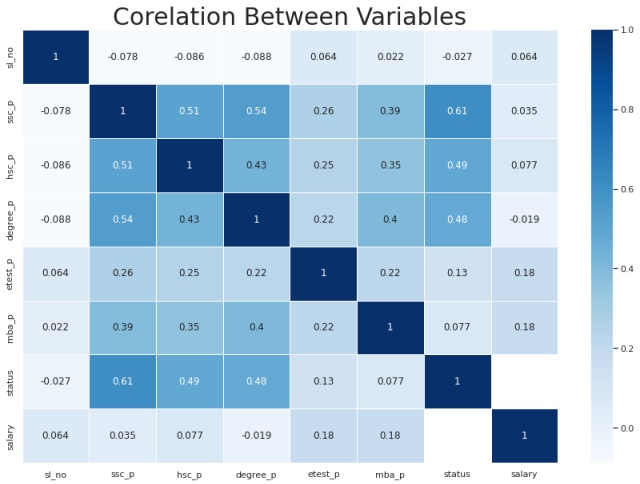


The above graphs shows that the percentage of MBA is slightly less significant for the placement of the students as the students getting placed and not placed have almost similar percentage distribution. Box plot graph for the mba_px vs gender also clarifies that female candidate on an average scored more then the male candidate and still many we not placed.

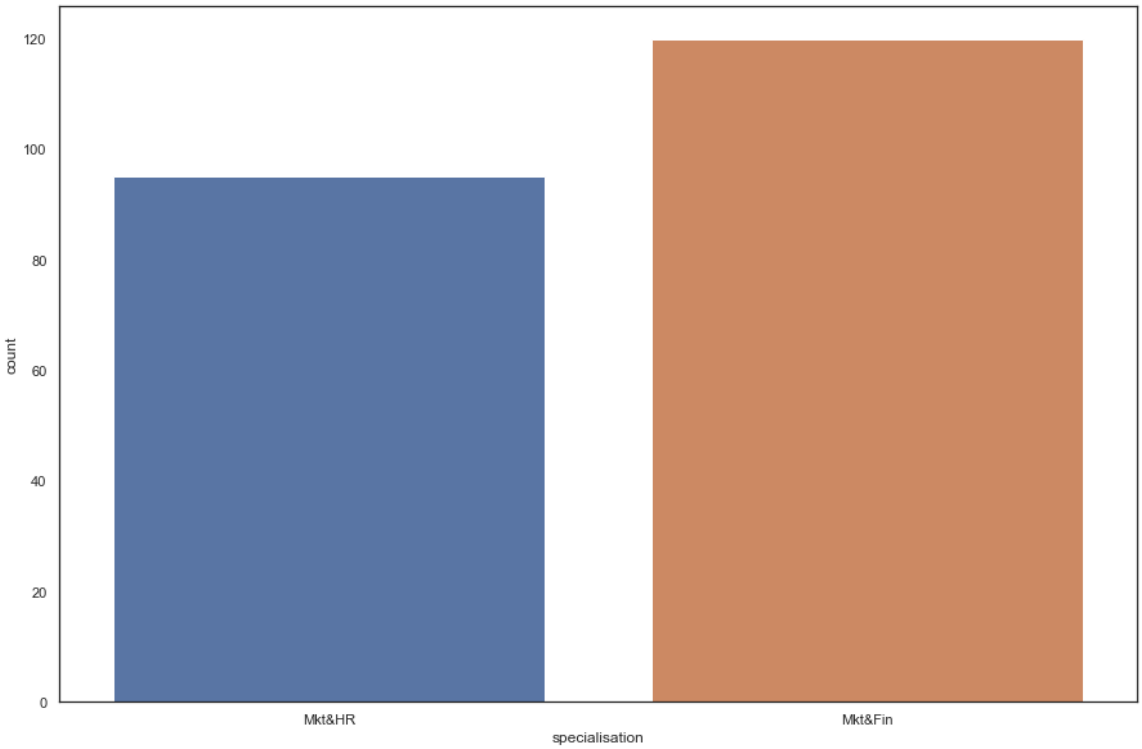


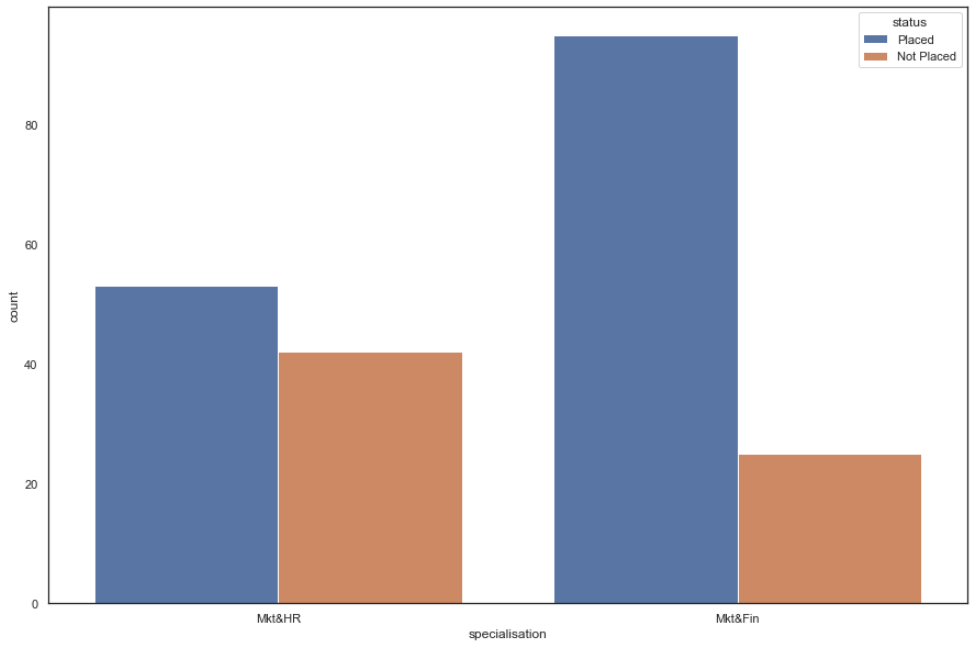
Outliers are identified in the dataset. Most of the outliers in the dataset are in the salary column as the salary column has relatively high numbers then the other columns.



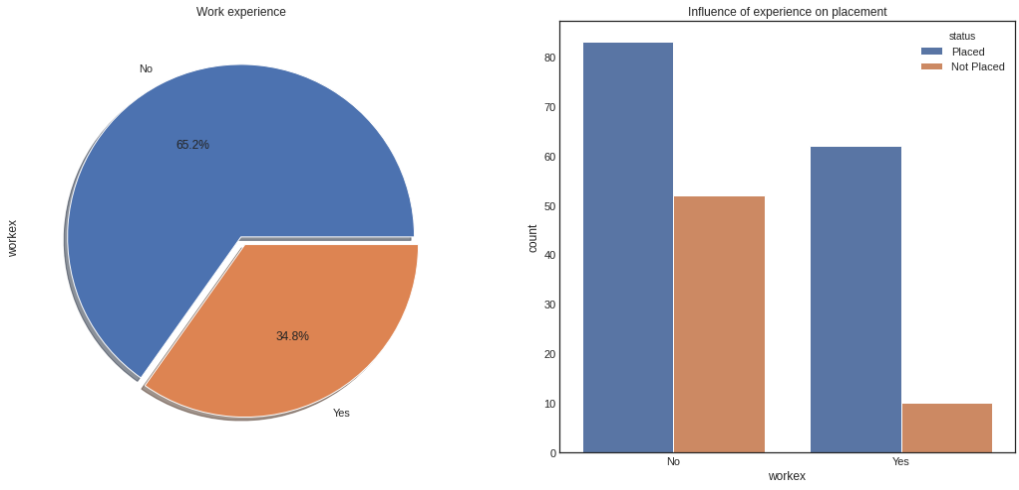


The above box plot shows that degree_p(degree percentage) criteria matters for the placement of the candidates as it is seen that the scores of the placed students starts almost from where the scores of the not placed students end. The confusion matrix also suggests that degree_p is significant.

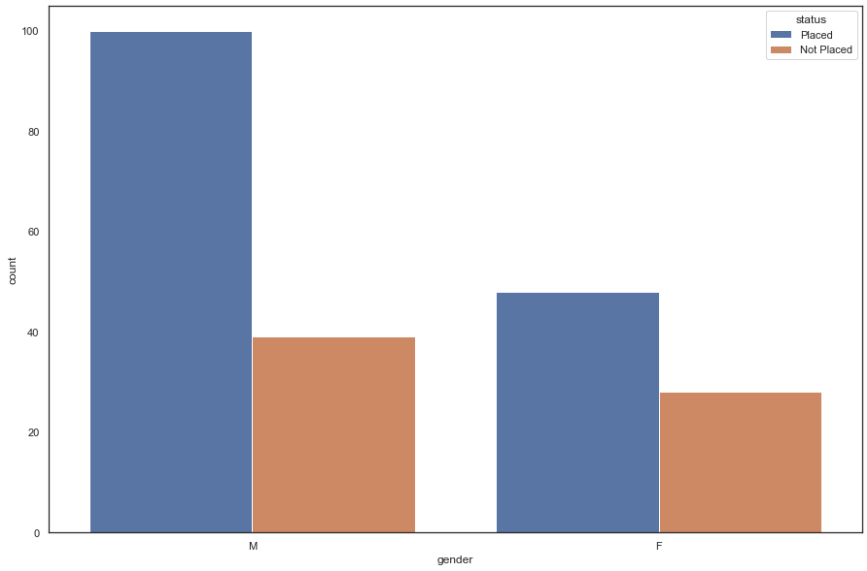




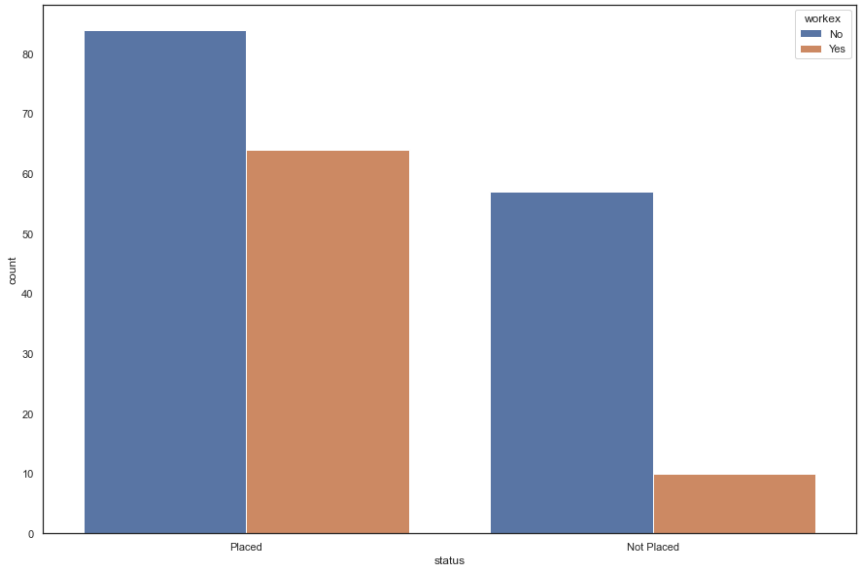
The above graph shows the count distribution of candidates in the two specializations - MKT&HR and MKT%Fin. From the graphs it can be seen that MKTFin has more demand than HR in the market and the maximum number of students getting placed are from MKTFin and the minimum number of students getting placed are also from MKTFin than MKT&HR thus it can be concluded that Mkt&Fin is getting placed more often.



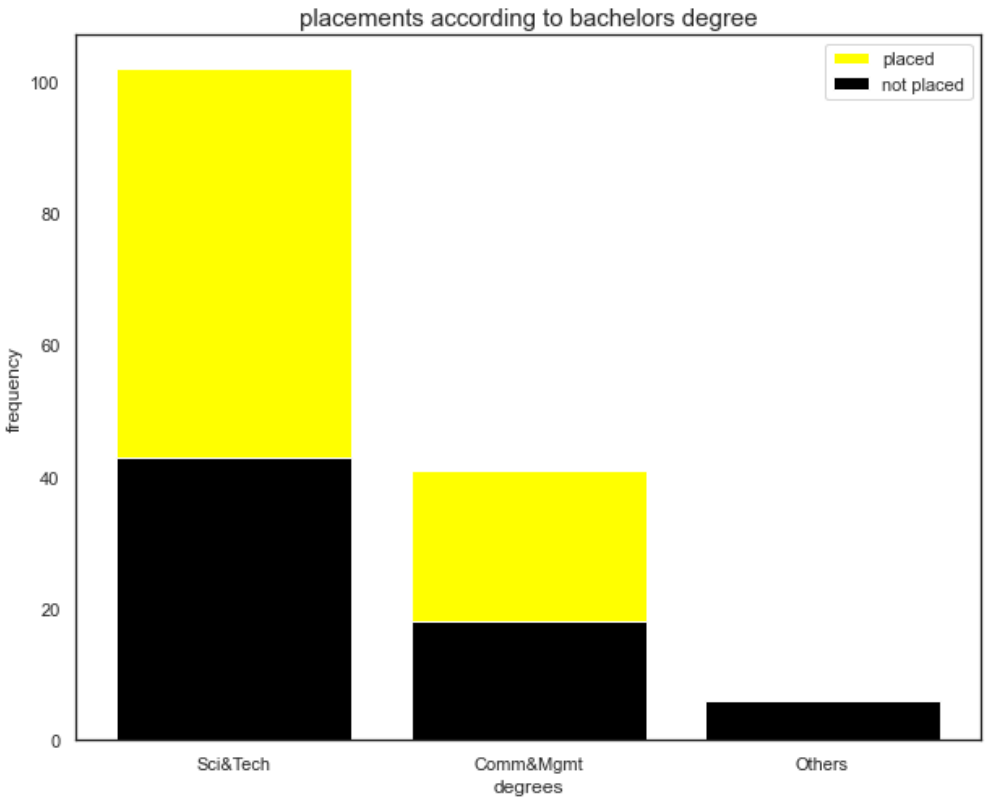
We can conclude that work experience doesn't influence a candidate in the recruitment process



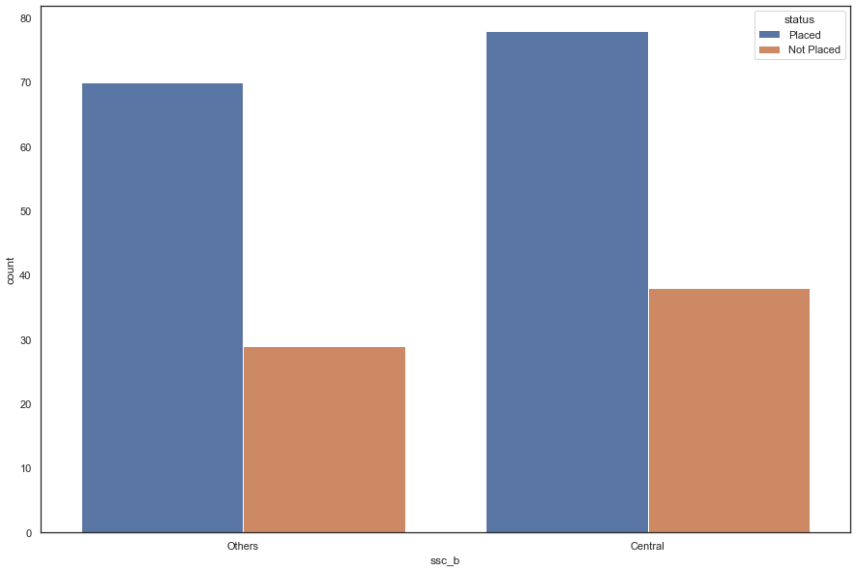
Above graph shows that male candidates are getting placed slightly higher than female candidates in the industry.

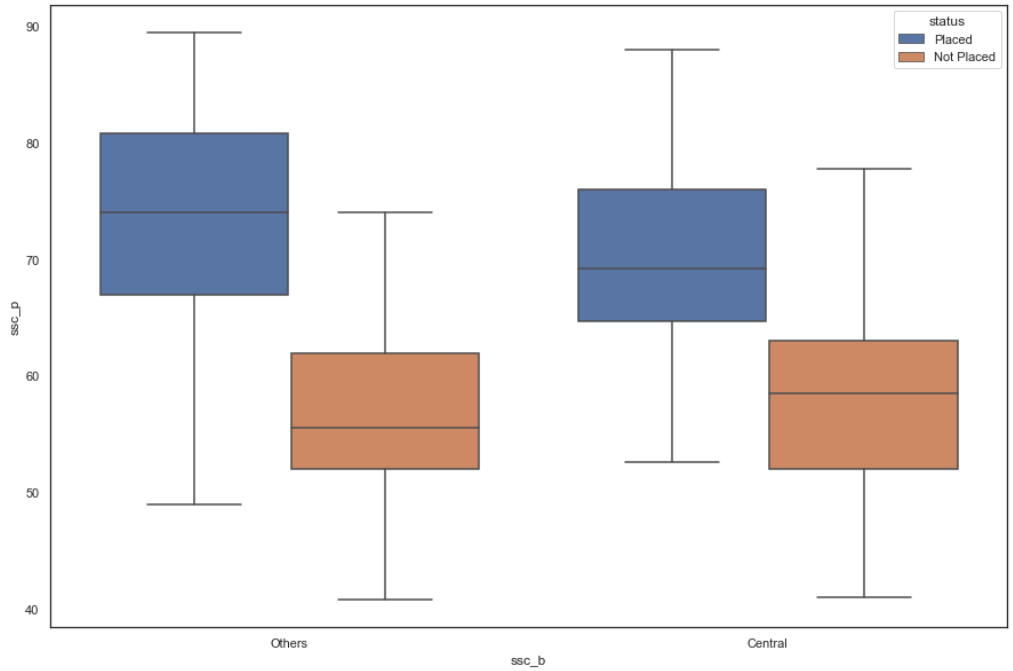


Above graph shows that workex is not a significant parameter for the placement of candidates as it can be seen that significant number of candidates with workex are not getting placed whereas 90+ candidates with no work experience are getting placed.

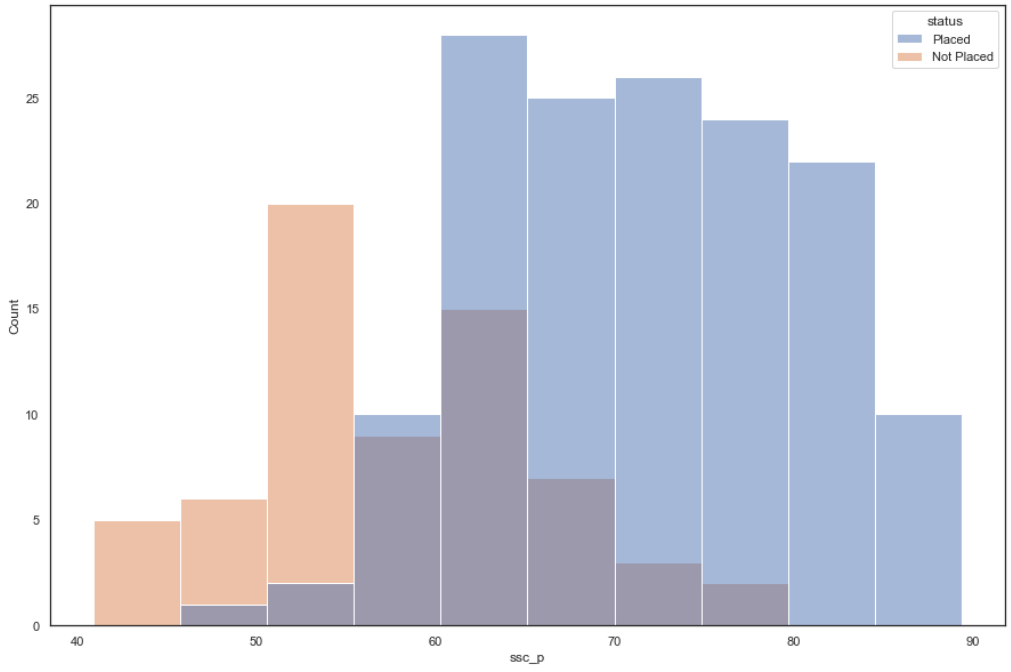


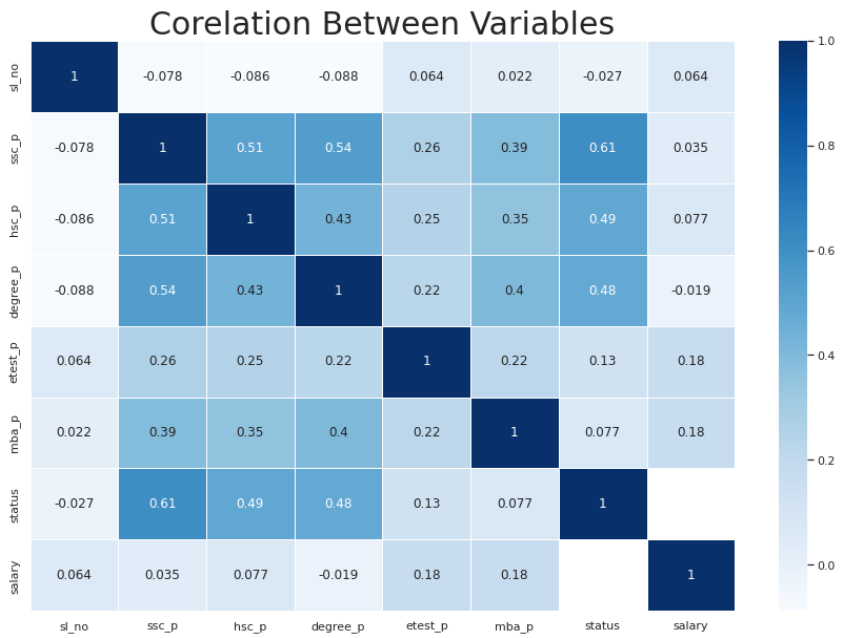
The above graph shows that candidates having bachelor degree in Sci&Tech are getting placed more in the industry than any other candidates followed by Comm&Mgmt and then Other degrees. Thus bachelor degree in Sci & Tech has more demand in the industry



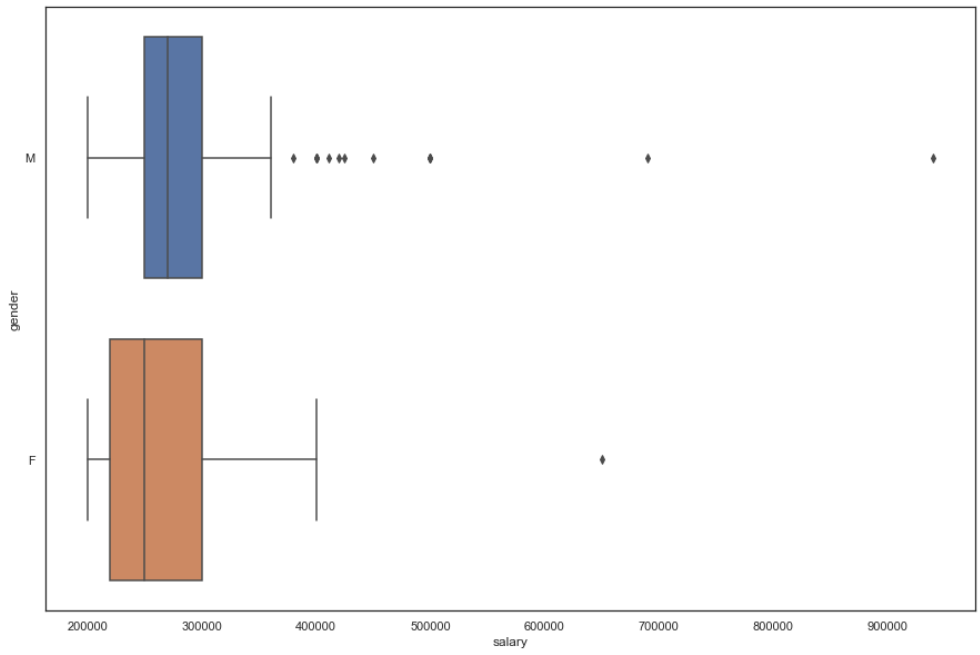


The above graph shows that ssc board of the candidates is not a significant factor for their placements as the number of students getting placed from the central board and other board are almost the same. The box plot confirms further that irrespective of the ssc_b variable , the not placed data is on the similar lines.





The above graph shows that ssc_p is a significant attribute for the placement of the candidates as the students having higher ssc_p are fully getting placed. The confusion matrix also gives the highest degree 0.61.



The above graph shows that the avg salary for Male is more than that of the female and male candidates are getting offered more high packages than female candidates.

Method

Exploratory Data Analysis was performed on the dataset to analyze the factors affecting the placement of a student, which degree specializations are in high demand in the industry, and whether MBA percentage matters in placement. Various data visualization libraries like Matplotlib , seaborn and various graphs like bar graph , box plot , count plot etc were used for the Exploratory Data Analysis for Problem Statement -1.

DATA

Link : https://www.kaggle.com/datasets/revelation2k23/brain-dead-placement-data

RESULTS

- There are more male applicants than female ones.
- Applicants with a science background and applicants who studied business for their high school diplomas come in second in both categories.
- Candidates with dual specializations in marketing and finance are in high demand.
- The majority of our candidates in our dataset lack any professional experience.
- The majority of our dataset's candidates were hired by a corporation.
- Many candidates who were hired received packages ranging from 2L to 4L PA.
- Only one candidate received close to 10L PA.
- The typical compensation is little higher than 2LPA.
- The majority of candidates with scores around 60% received respectable offers of about 3 lakhs PA.
- Few applicants obtained salaries of more than 4 lakhs PA.
- Candidates with no prior work experience have received more job offers than those with experience.
- We can conclude that a candidate's work experience has no bearing on the hiring process because there is a relationship between work experience and status.
- Many candidates with similar e_test scores but lack job experience did not receive placement.
- The majority of candidates with work experience were hired.

- The top salaries were given to male
- The average salary offered were also higher for male
- More male candidates were placed compared to female candidates
- Candidates who has high score in higher secondary and undergrad got placed
- Whomever got high scores in their schools got placed

CONCLUSION

1. Educational percentages are highly influential for a candidate to get placed
2. Past work experience doesn't influence much on your masters final placements
3. There are no gender discrimination while hiring, but higher packages were given to male
4. Academic percentages have no relation towards salary package.

Thus analysis on Campus Recruitment dataset was done and the detailed report is made which given insights regarding the placement patterns for the candidates.

Problem Statement 2 : Analyze Placement Data

Challenge Description

Social media platforms are widely used by individuals and organizations to express emotions, opinions, and ideas. These platforms generate vast amounts of data, which can be analyzed to gain insights into user behavior, preferences, and sentiment. Accurately classifying the sentiment of social media posts can provide valuable insights for businesses, individuals, and organizations to make informed decisions.

To accomplish this task, a customized private cartoon dataset (original images) of social media posts has been provided, which contains labels for each post's emotion category, such as happy, angry, sad, or neutral.

The task is to build and fine-tune a machine-learning model that accurately classifies social media posts into their corresponding emotion categories, using synthetic images.

Introduction

The goal of this task is to build a machine learning model that accurately classifies social media posts into their corresponding emotion categories such as happy, angry, sad, or neutral. We have been provided with a customized private cartoon dataset of social media posts that contain labels for each post's emotion category. We will use synthetic images to augment the dataset and increase its size. The model's performance will be evaluated using standard evaluation metrics, and the best performing model will be selected based on the evaluation scores.

Dataset Exploration

The dataset consists of cropped cartoon face images, and it has been pre-processed and cleaned. The dataset is divided into four categories of emotions, including happy, angry, sad, and neutral. The given data set is used for synthetic generation of 10,000 images for training and testing purposes.

Link : https://www.kaggle.com/datasets/revelation2k23/brain-dead-emotion-detection

Methods

To augment the dataset and increase its size, synthetic images were generated using GAN (Generative Adversarial Network). Images in each emotion category were used to synthetically generate similar images. The synthetic images were then combined with the original images to build a machine learning model.

The machine learning model was built using a convolutional neural network (CNN) algorithm. The CNN algorithm was trained using the Adam optimizer and a categorical cross-entropy loss function. The model was fine-tuned using the original and synthetic images, and the model's hyperparameters were optimized using grid search.

Results

1) The model trained on CNN gave an accuracy of 0.90 on the training dataset and accuracy of 0.67 on testing dataset
2) The classification of the images was done with confidence level ranging from 70% to 90%.
3) The confusion matrix showed that the model correctly classified most of the images in each emotion category.
4) The ROC curve and AUC ROC score showed that the model had good performance in distinguishing between the different emotion categories.
5) Synthetically generated images generated are not up to the mark.
6) GAN model is not able to make clear images to help the model train and have a better accuracy
7) The synthetic images provided additional examples of each emotion category, which improved the model's ability to generalize and classify new images accurately.

Link :
https://drive.google.com/drive/folders/1A9kf-QhPwmy8e-TBGGBah0DShHsXbz6O?usp=sharing

Conclusion

In conclusion, the use of synthetic images can be an effective technique for improving the classification accuracy of machine learning models. However, it is important to note that the quality of the synthetic images generated and the size of the augmented dataset can significantly impact the model's performance. Further research is needed to investigate the optimal techniques for generating synthetic images and their impact on model performance.

The model made by us for emotion classification needs more training with the help of synthetic images of good quality. The GAN model used to generate the images could not give the best possible images thus the model could not be trained to a good accuracy level for the testing data.