Analysis of Campus Recruitment Data

By Manthan Bagade

This is my first project in the field of Data Science. I want to learn and apply my learnings on real life datasets. I am starting my data analysis with the help of Python. Python has extensive libraries for data analysis, and which will make my analysis presentable. I have used the dataset available on **Kaggle** with the name **“Campus Recruitment”** by “**Ben Roshan”**. I acknowledge the efforts of the creator of this dataset who collected this data. It is of great help for students like me.

# Objective

The objective of this project is to determine and analyze the **Academic and Employability factors** influencing college placements. The project will present relevant statistics and graphs which will be sufficient to draw out comprehensible conclusions for this dataset. I also aim to answer these questions based on my analysis of the data.

1. Which factor influenced a candidate in getting placed?
2. Do percentage matters for one to get placed?
3. Which degree specialization is much demanded by corporate?
4. Play with the data conducting all statistical tests.

Not to forget, the final aim of this project is to learn and apply the skills which I have developed while starting out my journey to master Data Science.

“The best way to learn Data Science is to do Data Science”

# Dataset Source and Details

The datasets are downloaded from Kaggle. The name of the dataset is **“Campus Recruitment”** by “**Ben Roshan”**. The link to the dataset is attached for the reference.

<https://www.kaggle.com/datasets/benroshan/factors-affecting-campus-placement/>

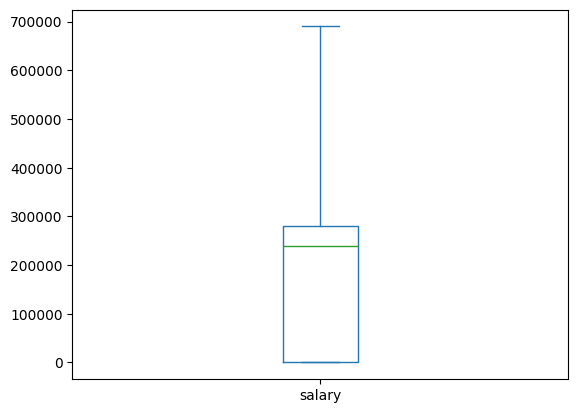
The Dataset is a csv file. The file contains a total of 215 entries and 15 columns.

| Column Name | Datatype | Description |
| --- | --- | --- |
| sl\_no | int | Serial Number |
| gender | categorical(char) | Gender - Male=‘M’, Female=‘F’ |
| ssc\_p | float | Secondary Education percentage- 10th Grade |
| ssc\_b | categorical(string) | Board of Education- Central/ Others |
| hsc\_p | float | Higher Secondary Education percentage- 12th Grade |
| hsc\_b | categorical(string) | Board of Education- Central/ Others |
| hsc\_s | categorical(string) | Specialization in Higher Secondary Education |
| degree\_p | float | Degree Percentage |
| degree\_t | categorical(string) | Under Graduation(Degree type)- Field of degree education |
| workex | categorical(string) | Work Experience |
| etest\_p | float | Employability test percentage ( conducted by college) |
| specialisation | categorical(string) | Post Graduation(MBA)- Specialization |
| mba\_p | float | MBA percentage |
| status | categorical(string) | Status of placement- Placed/Not placed |
| salary | int | Salary offered by corporate to candidates |

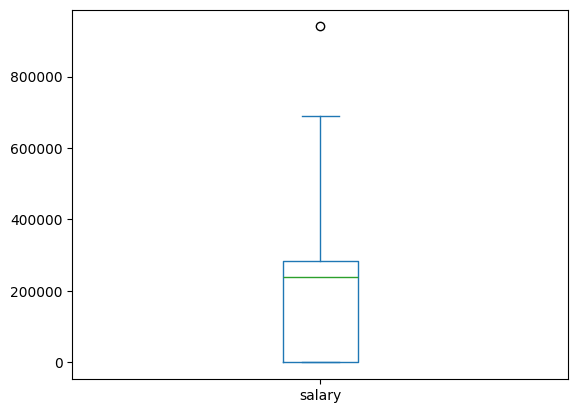
# Dataset Cleaning and Filling Missing Values

All the columns do not have any missing values except the column ‘salary’. It has 67 missing values. I filled all the missing values with 0 since the students who were unplaced did not have any salary.

I also used box plot to identify any outliers using the IQR (Interquartile range) method on the ‘salary’ column. The data was so well collected that I only found one outlier. Dropping the row was the best choice because such high salary as compared to other students cannot be accounted to academic factors. It is a result of personal hard work and can impact the statistics of the data.



Data After Dropping Outlier

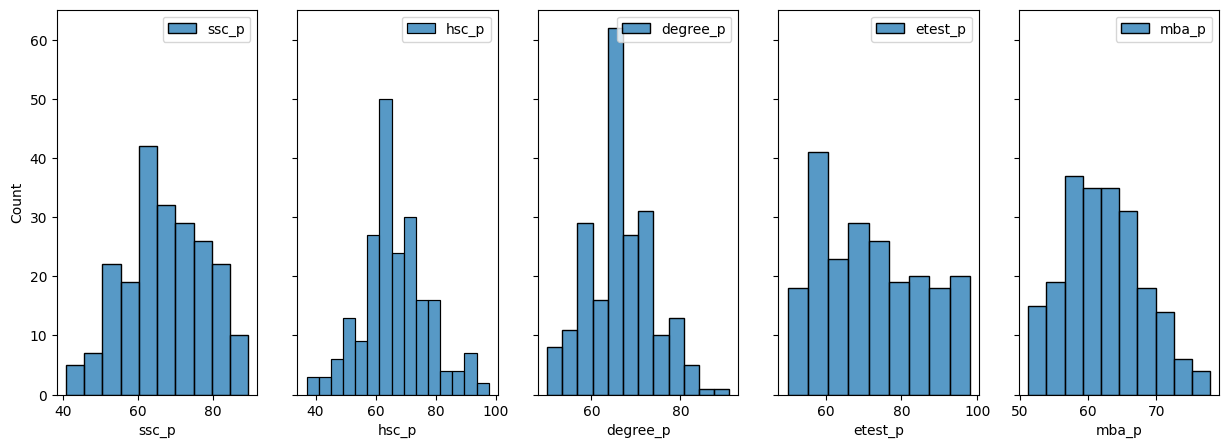


Single Outlier Point

# Data Description

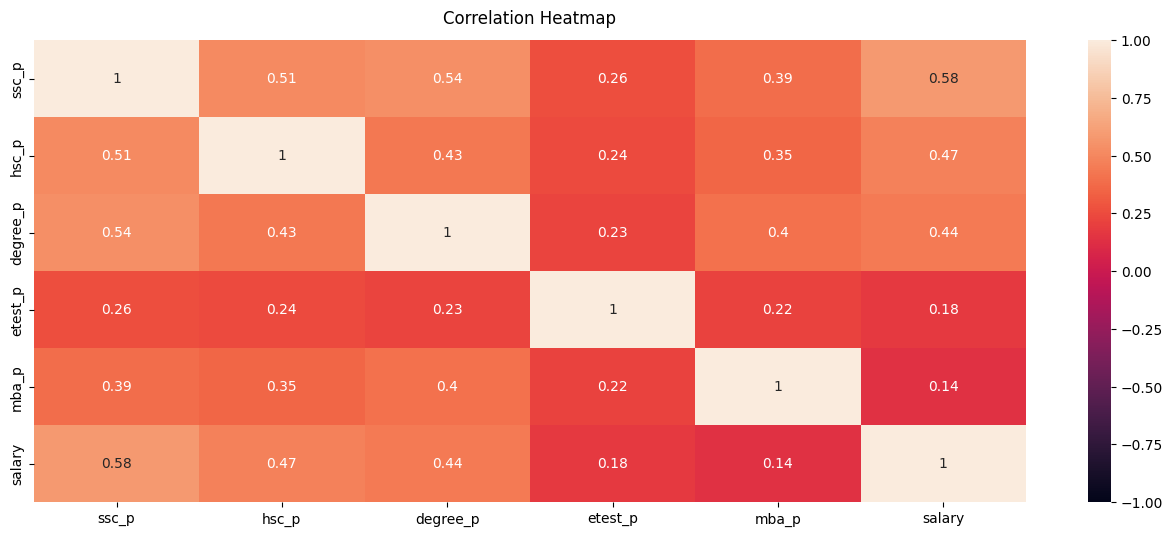
|  | ssc\_p | hsc\_p | degree\_p | etest\_p | mba\_p | salary |
| --- | --- | --- | --- | --- | --- | --- |
| mean | 67.33 | 66.32 | 66.38 | 72.05 | 62.27 | 1,95,238 |
| std | 10.84 | 10.92 | 7.37 | 13.29 | 5.85 | 1,46,553 |
| min | 40.89 | 37.00 | 50.00 | 50.00 | 51.21 | 0 |
| 25% | 60.55 | 60.85 | 61.00 | 60.00 | 57.92 | 0 |
| 50% | 67.00 | 65.00 | 66.00 | 70.50 | 61.95 | 2,40,000 |
| 75% | 75.85 | 73.00 | 72.00 | 83.75 | 66.27 | 2,80,000 |
| Max | 89.40 | 97.70 | 91.00 | 98.00 | 77.89 | 6,90,000 |

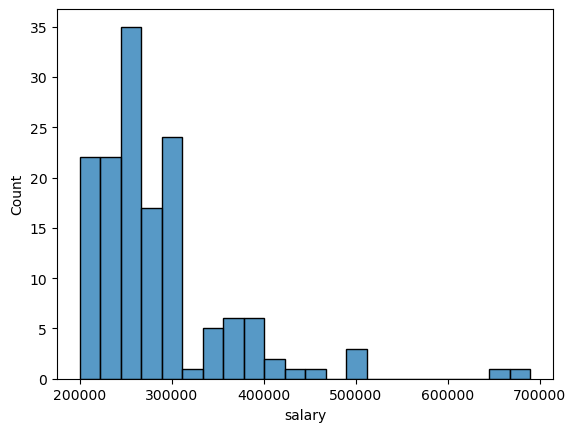
It is clear from the following histogram that maximum number of students had a salary between 2,00,000 to 3,00,000. The other peak at the start is due to the unplaced students for which we have assigned the salary as 0. Thus, we can say that the companies which are coming for the placements are able to pay only that much of the amount. The median salary is 2,40,000 which is verified by histogram.



Histogram of Academic attributes

The other histogram plots show the frequency of the percentages of different examinations.





The result of the Correlation Heatmap is quite surprising. Salary least relates to employability test and MBA percentage. Further, employability test and MBA scores are not related so strongly to any of the other academic factors. The reason for these unexpected results can be the irrelevance of the questions and the syllabus which are asked in tests. The questions asked in employability test could only assess the whether the person is fit for any job rather than evaluating him quantitatively.

Higher Secondary (12th grade) is percentage is quite related to Senior Secondary (10th grade) percentage which is self-explanatory and similarly the degree percentage is related to 10th and 12th grade scores.

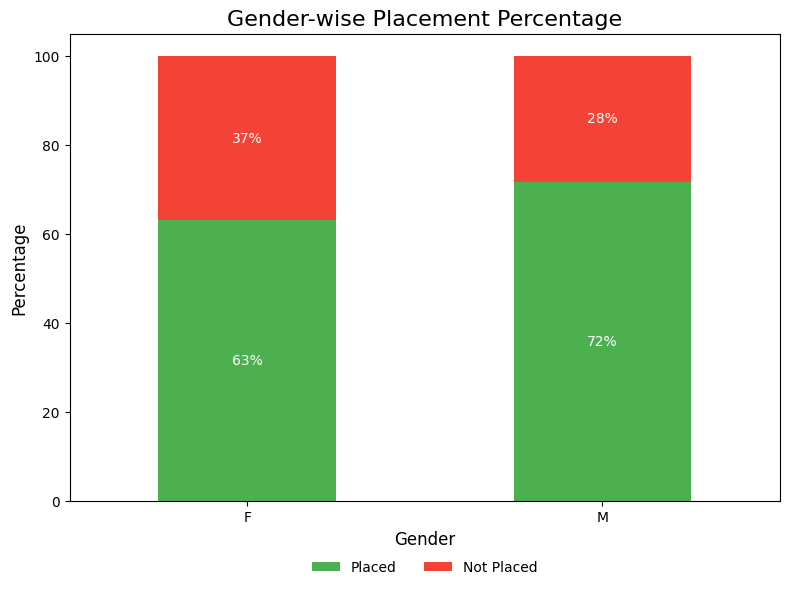
The salary is related to hsc\_p, ssc\_p and degree\_p which points that early development plays a great role in placement season.

# Data Visualization and Interpretation

Let us visualize the answers to the questions in the objectives through charts.

**Gender wise Analysis**

| gender | status | count | mean\_salary | percentage\_placed |
| --- | --- | --- | --- | --- |
| F | Not Placed | 28 | 0 | 36.84% |
| F | Placed | 48 | 2,67,292 | 63.16% |
| M | Not Placed | 39 | 0 | 28.26% |
| M | Placed | 99 | 2,92,434 | 71.74% |

****

From the graphs male students have greater percentage of placements than Female students. They also have slightly higher salary than their counterparts. Generally, male students are more flexible, and this can be the reason they are paid higher average salary.

**Relation between ssc\_p, hsc\_p and salary**