Exploratory Data Analysis (EDA) in Python

Presented by: Group 9
Somnath Sharma , Shilpa TS, Gousia Banu, Raghav srivatsav, Sumanth G M and Rakshitha A

Mtech in Data science and Artificial Intelligence
Sec A

Objective

The objective of this project is to select a dataset, perform data cleaning and pre-processing, conduct exploratory data analysis (EDA), and present our findings. This project will help us understand the dataset, uncover underlying patterns, and generate insights that could guide further analysis or decision-making.

Dataset : Placement Prediction
Source : Kaggle

Dataset Details

This dataset contains details of students and their placement status

- Total Number of columns = 12
- Total record count = 10000

Keys

- Student ID (Primary Key)

Numerical columns

- CGPA
- Internships
- Projects
- Workshops/Certifications
- AptitudeTestScore
- SoftSkillsRating
- SSC_Marks
- HSC_Marks

Categorical columns

- ExtraCurricular Activities
- Placement Training
- Placement Status

Data Import and Cleaning

- Read the CSV file into a pandas DataFrame
- The dataset has 10,000 rows and 12 columns.
- It includes numerical (int and float) and categorical (object) columns.
- No missing values are present in any column.
- No duplicate rows were found.
- Categorical columns (ExtracurricularActivities, PlacementTraining, PlacementStatus) were converted to the category data type for efficiency.
- StudentID was converted to string (object type) since it is an identifier and not meant for numerical operations.

Exploratory Data Analysis (EDA)

Descriptive Statistics

- Measures include mean, median, standard deviation, skewness, etc.

Numerical Data : Total 10000 students are there

- Mean CGPA is 7.69 with minimum CGPA as 6.5 amd maximum CGPA as 9.1
- SSC_Marks scored by students: Mean 69.15 , Minimum score is 55 and Max score is 90. 50% of students have got 70
- AptitudeTestScore median(central tendency) is 80 with 66.58 dispersion

Categorical Data :

- ExtracurricularActivities : total 5854 students have participated
- Placement status: 5803 of 7318 who had taken the Placement training have got the successful placement.

Exploratory Data Analysis (EDA)

Group by Analysis

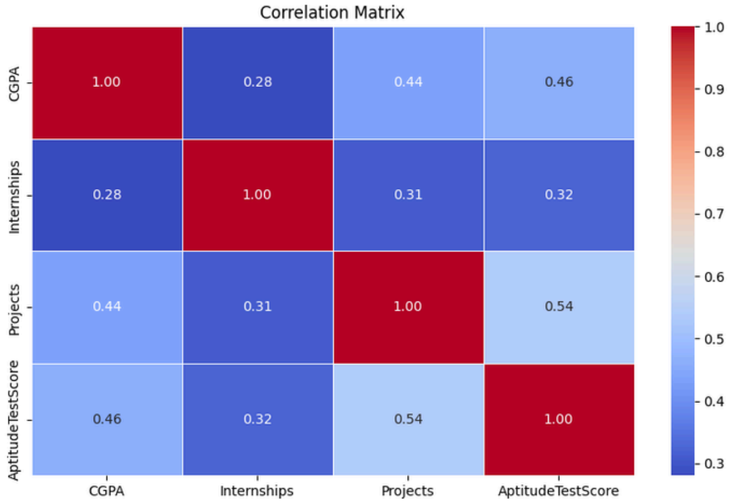
The groupby() function groups the dataset by catogorical features such as PlacementStatus, then calculate the median for the numerical features such as CGPA and Aptitude test score in the data set. Also, to display the median aggregrated data using groupby () function for projects, workshops/certifications in reference to placementstatus.

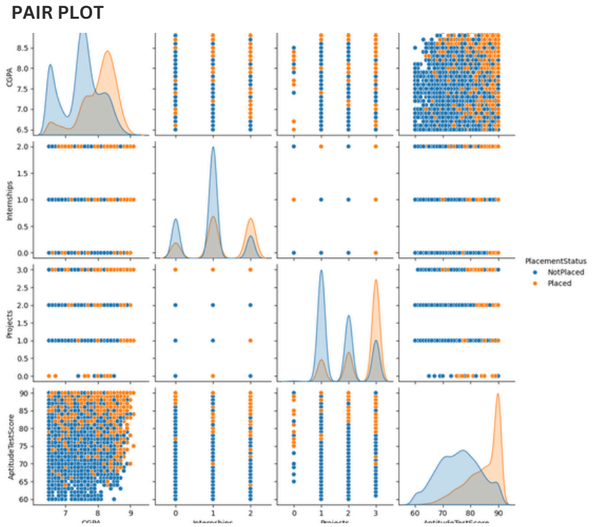
Students placements are dependent on many aspects, however based on categorical features we can determine the students placements as mentioned below :

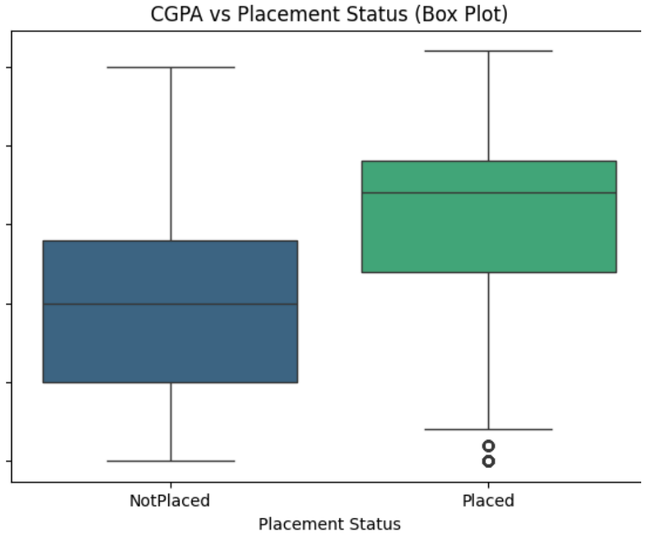
1. Students having a CPGA of 8.2 and above are placed and students having a CGPA of 7.5 and below are not placed.
2. Students having scored 86 marks and above are placed and students having scored 76 marks and below are not placed.
3. Students having worked on 3 or more projects and having 2 or more workshops/certifications have a fair chance of getting placed.

Exploratory Data Analysis (EDA)

Feature Analysis







1.**Pairwise Relationships**: Each subplot in the pairplot represents scatter plots of key numerical features based on the placement status.
2.**Correlation Relationships** : The Heat map represents correlation between the key numerical features in the identify_data set.
3.**Pivot Table** : Pivot table represents mean of key numerical features in the identify_data set based on placement status.
4.**Additional analysis** : Based on Student CGPA analysis placement status using Box plot.

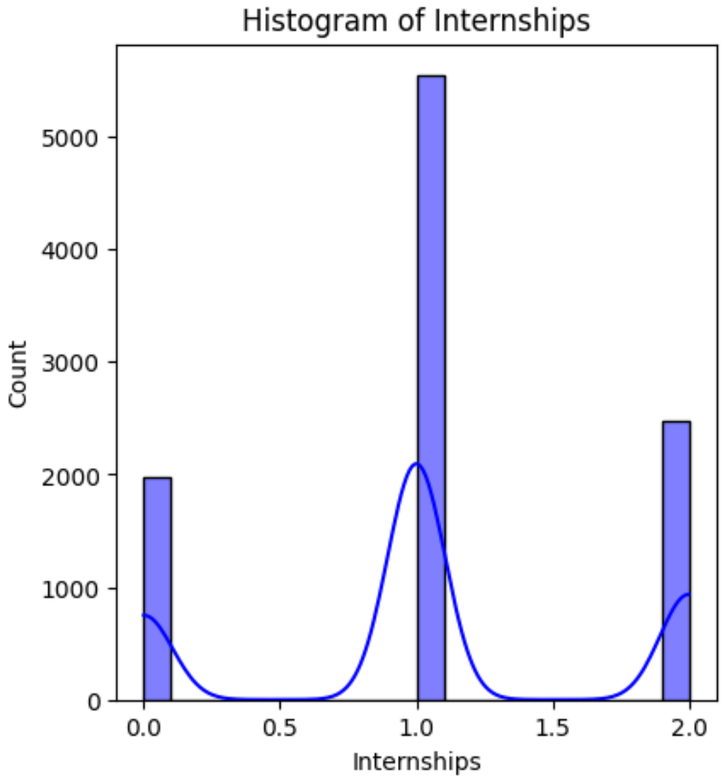
Advanced Python Techniques

- **Lambda Functions** : Create a new column “degree_awarded” based on CGPA.

- **User Defined Functions** : Creates a DataFrame showing how a student's CGPA compares to their peers based on their ID.

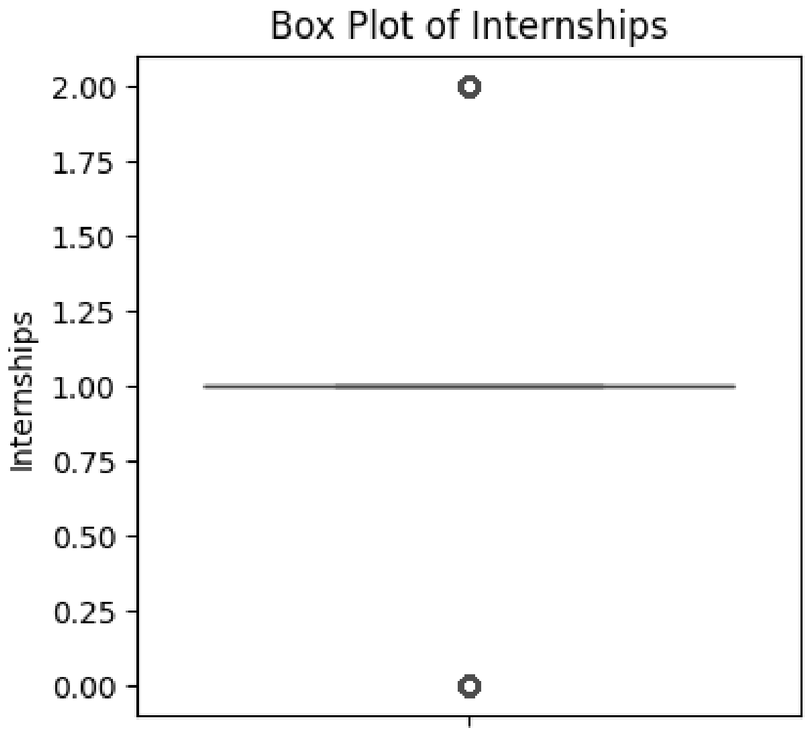
- **List Comprehensions** : Generate List of student who are FCD Holders.

Insights and Conclusions



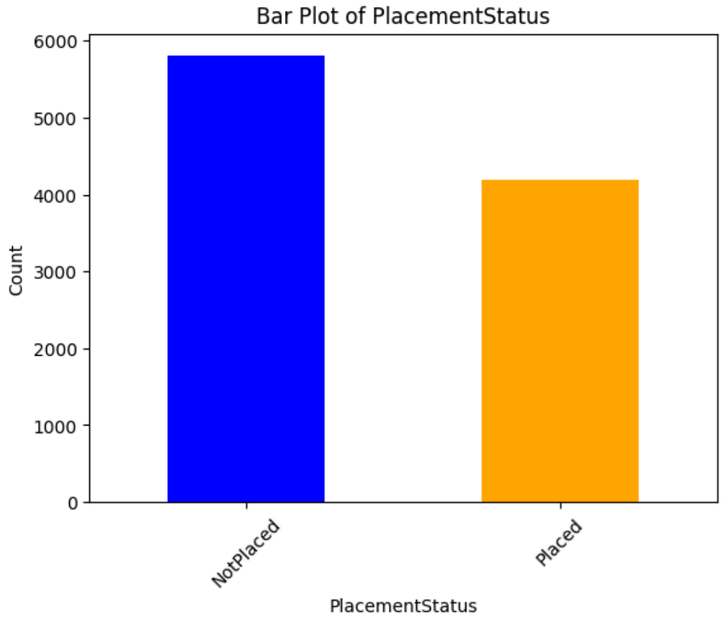
Most students have completed one internship, this suggest that placement is dependent on internships

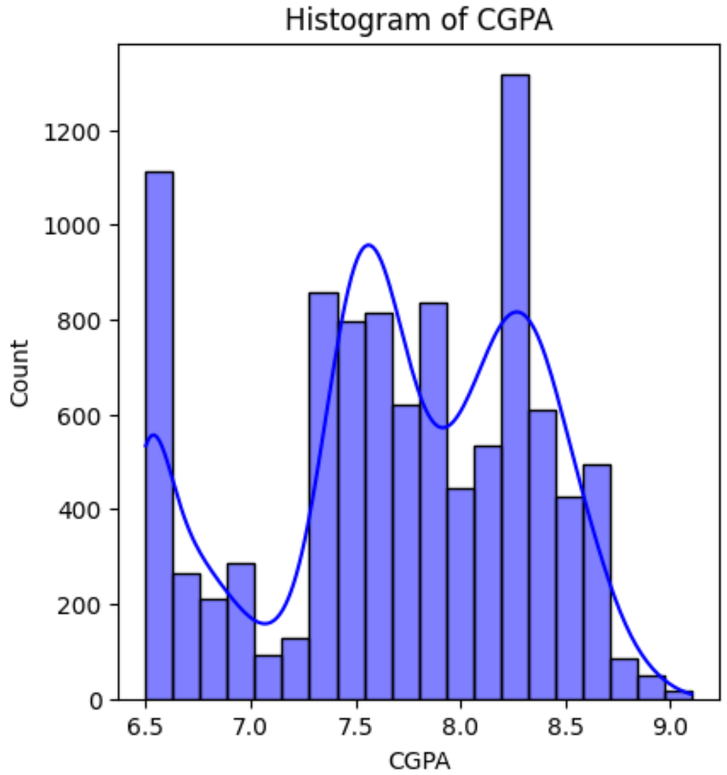
Insights and Conclusions



0 internships (bottom outlier): A few students have no internships,
2 internships (top outlier): A few students have completed more than 1 internship

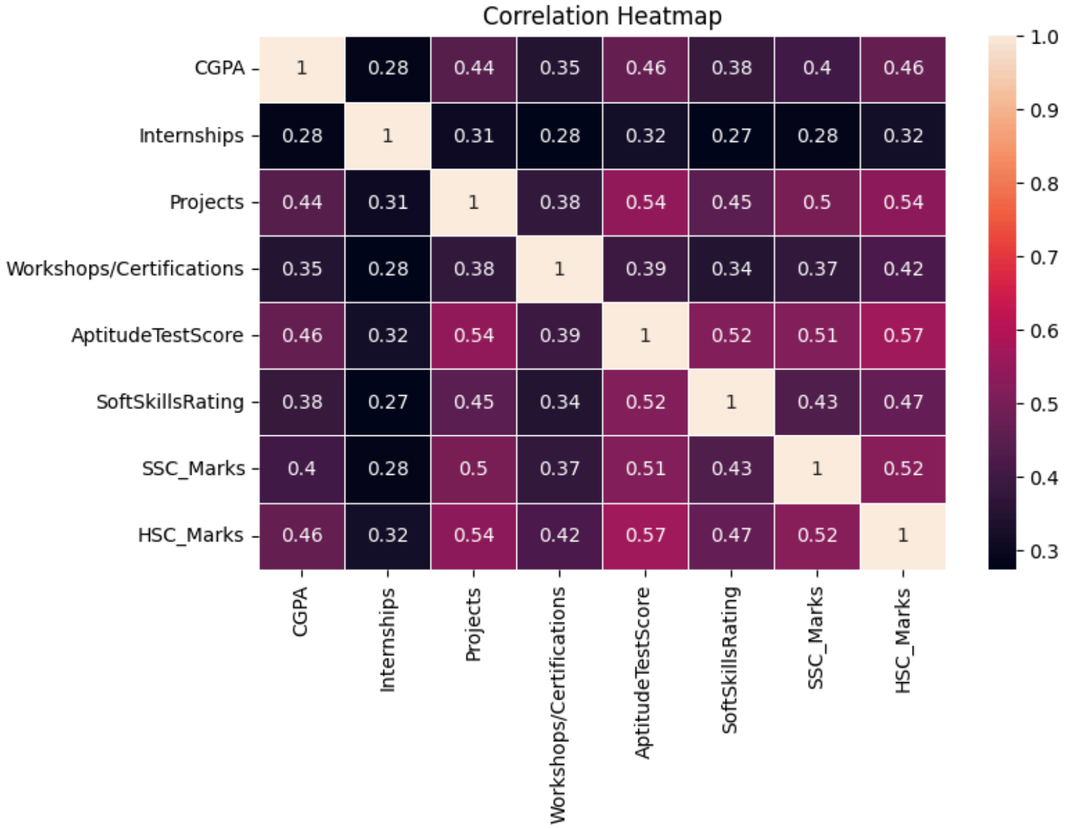
Insights and Conclusions





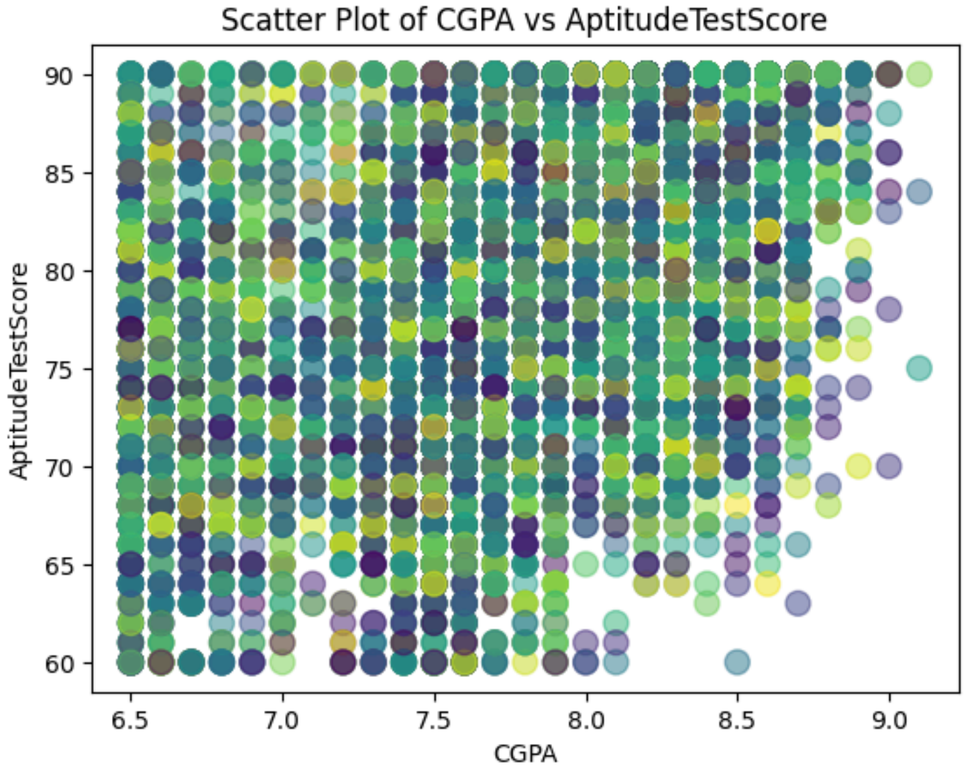
First Class is the most common degree awarded. Since the placement rate is not very high, having a degree alone may not be enough to secure a placement other skills could be necessary.

Insights and Conclusion



A correlation of 0.57 between HSC Marks and Aptitude Test and 0.51 for SSC Marks this indicates a strong positive relationship. This means that students who perform well in high school exams tend to score higher in aptitude tests.

Insights and Conclusion



- The data points are evenly spread, suggesting that there is no strong correlation between CGPA and Aptitude Test Score.
- A few students with high CGPA have lower aptitude scores (<70).
- Some students with lower CGPAhave relatively higher aptitude scores (>80).
- However, a slight upward trend can be observed—higher CGPA values tend to have higher Aptitude Test Scores.

Thank you!