Exploratory Data Analysis (EDA) Report

1. Introduction

Exploratory Data Analysis (EDA) is a crucial step in data preprocessing. It allows us to understand the dataset's structure, detect patterns, identify anomalies, and gain insights that drive better decision-making. This report walks through the EDA process applied to the dataset analyzed in the EDA.ipynb notebook.

2. Dataset Overview

2.1 Data Import

- **File Name:** Raw_data/data.xlsx - Sheet1.csv
- **Number of rows:** 3998
- **Number of columns:** 39

Key Columns: 'ID', 'Salary', '10percentage', '12graduation', '12percentage', 'CollegeID', 'CollegeTier', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier', 'GraduationYear','Gender', Designation', 'JobCity','10board', '12board'

Understanding the dataset is the first step in EDA. We begin by loading the data and inspecting its structure to determine the number of observations, data types, and missing values.

2.2 Data Structure

- **Shape:** (3998, 39)
- **Data Types:** Numeric, Categorical, DateTime
- **Missing Values:** Checked & not present
- **Descriptive Statistics:**
    - **Mean, Median, Mode**: Help us understand central tendencies
    - **Standard Deviation**: Measures data spread
    - **Min & Max Values**: Help identify potential outliers

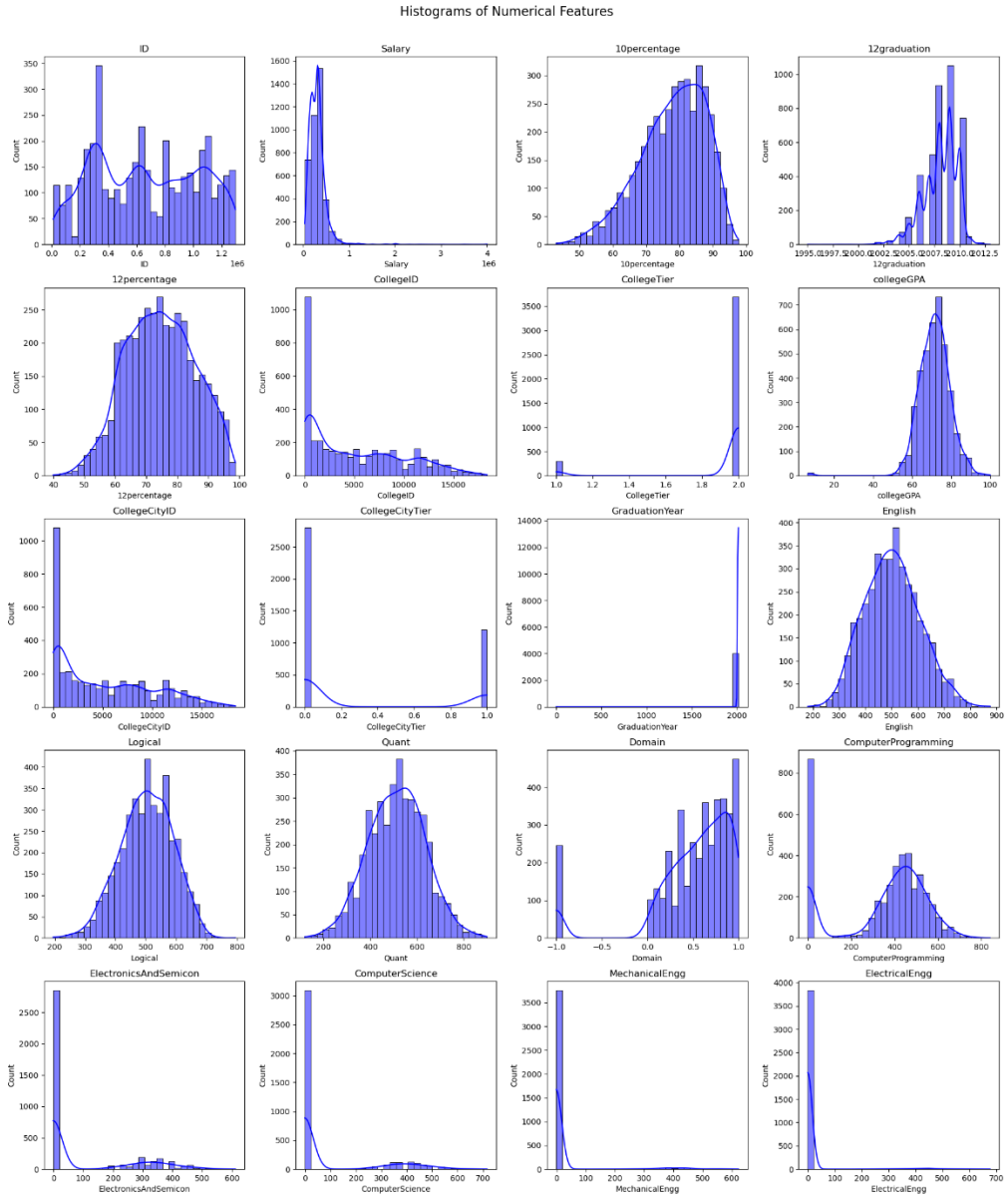
3. Feature Engineering

- Grouped similar categorical values for better analysis (e.g., 'computer engineering' → 'CS').
- Converted Object Columns to DateTime (DOJ, DOB, DOL).
- Replaced 'Present' in DOL with Current Timestamp to standardize date formats.

4. Data Visualization

4.1 Univariate Analysis

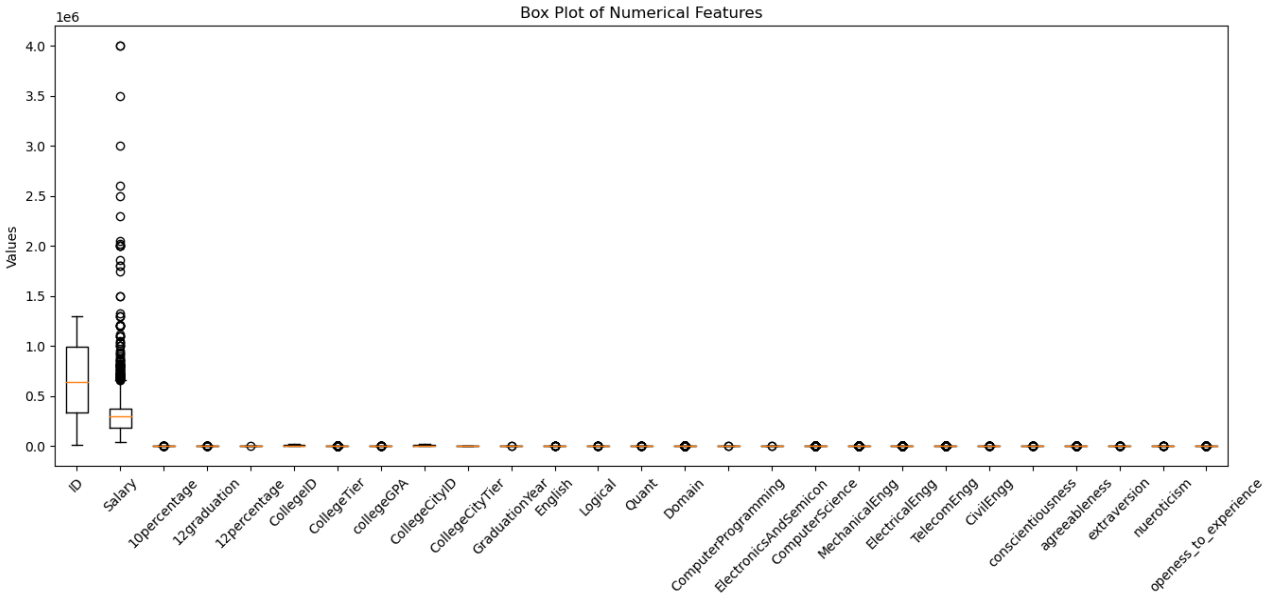
1) **Histograms** for numerical features: Understanding the probability and frequency distribution of numerical



Observations:

- Most numerical features have a normal distribution, but some show skewness.
- Salary and College GPA may have right-skewed distributions.
- Some categorical features have distinct peaks.

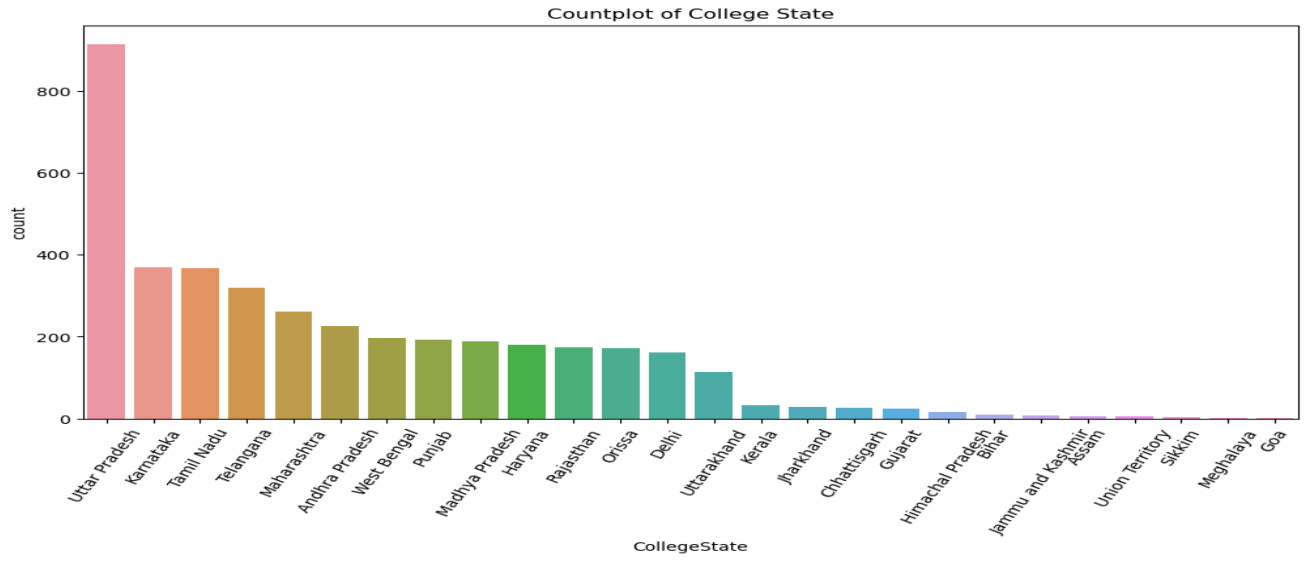
2) Boxplots for outlier detection:



Observation:

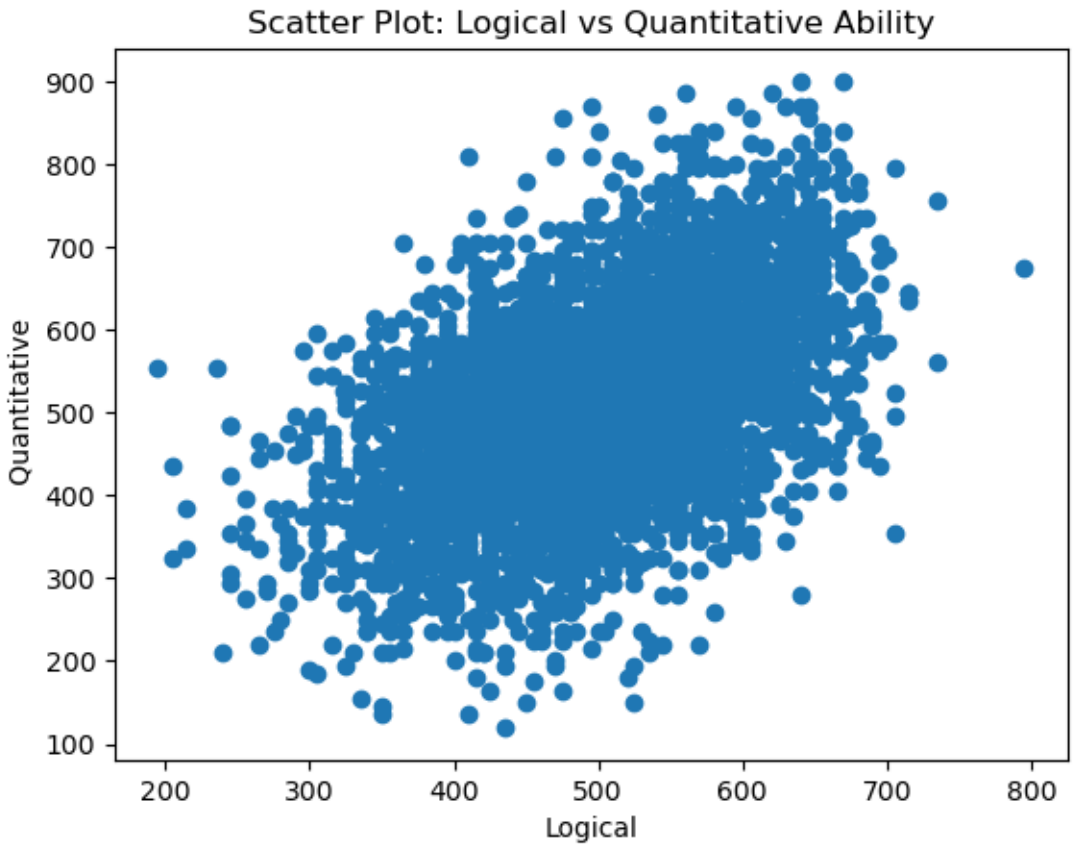
- Outliers are detected in **salary** Column

3) **Count for categorical features:** Display the frequency of different categorical data.

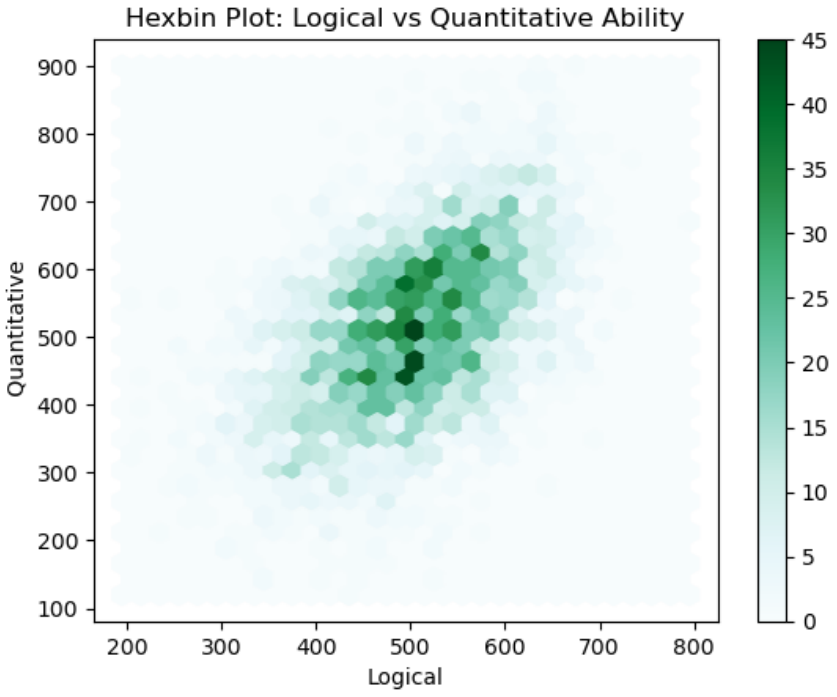


4.2 Bivariate Analysis

- **Scatter plots** to identify relationships between numerical features.



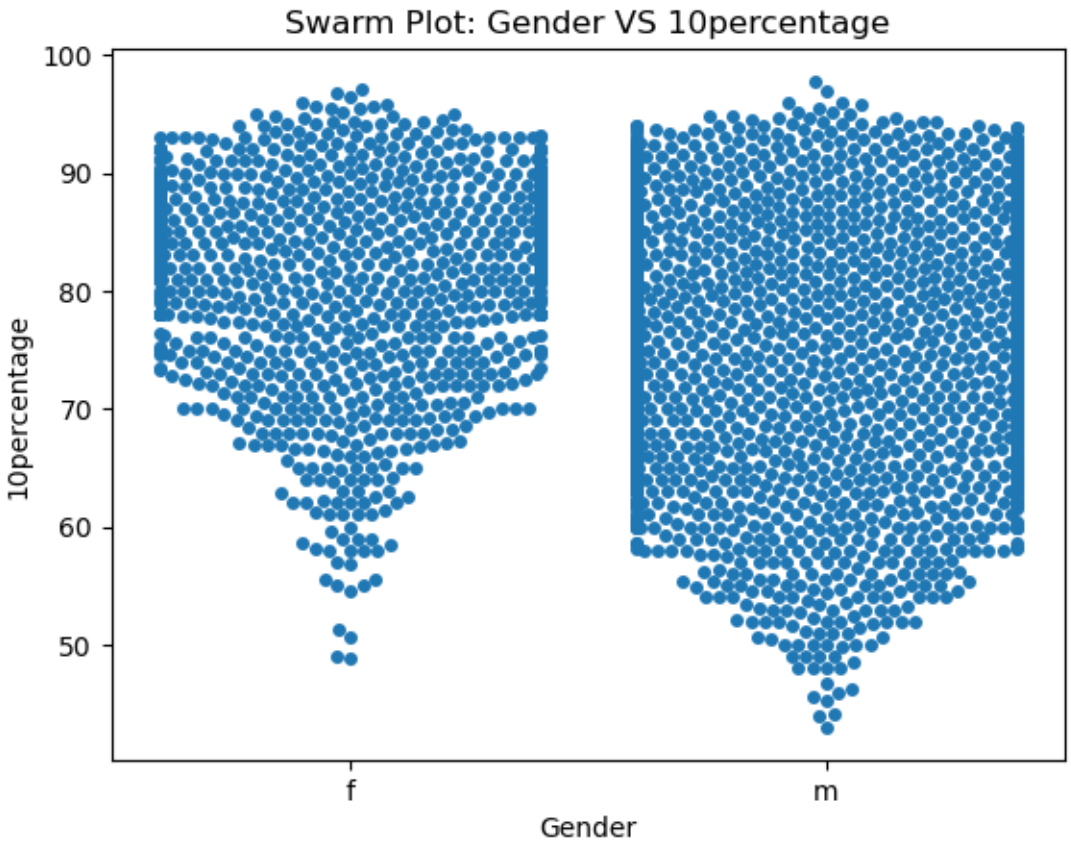
- **Hexbin plot:**



Observations

- People who score **high in Logical Ability** also tend to score **high in Quantitative Ability**.
- The **darkest area** (most common scores) is around **500 Logical & 500 Quantitative**.
- The data forms an **upward pattern**, showing a **positive relationship** between Logical and Quantitative scores.
- Few people have **very low or very high scores** in both categories.

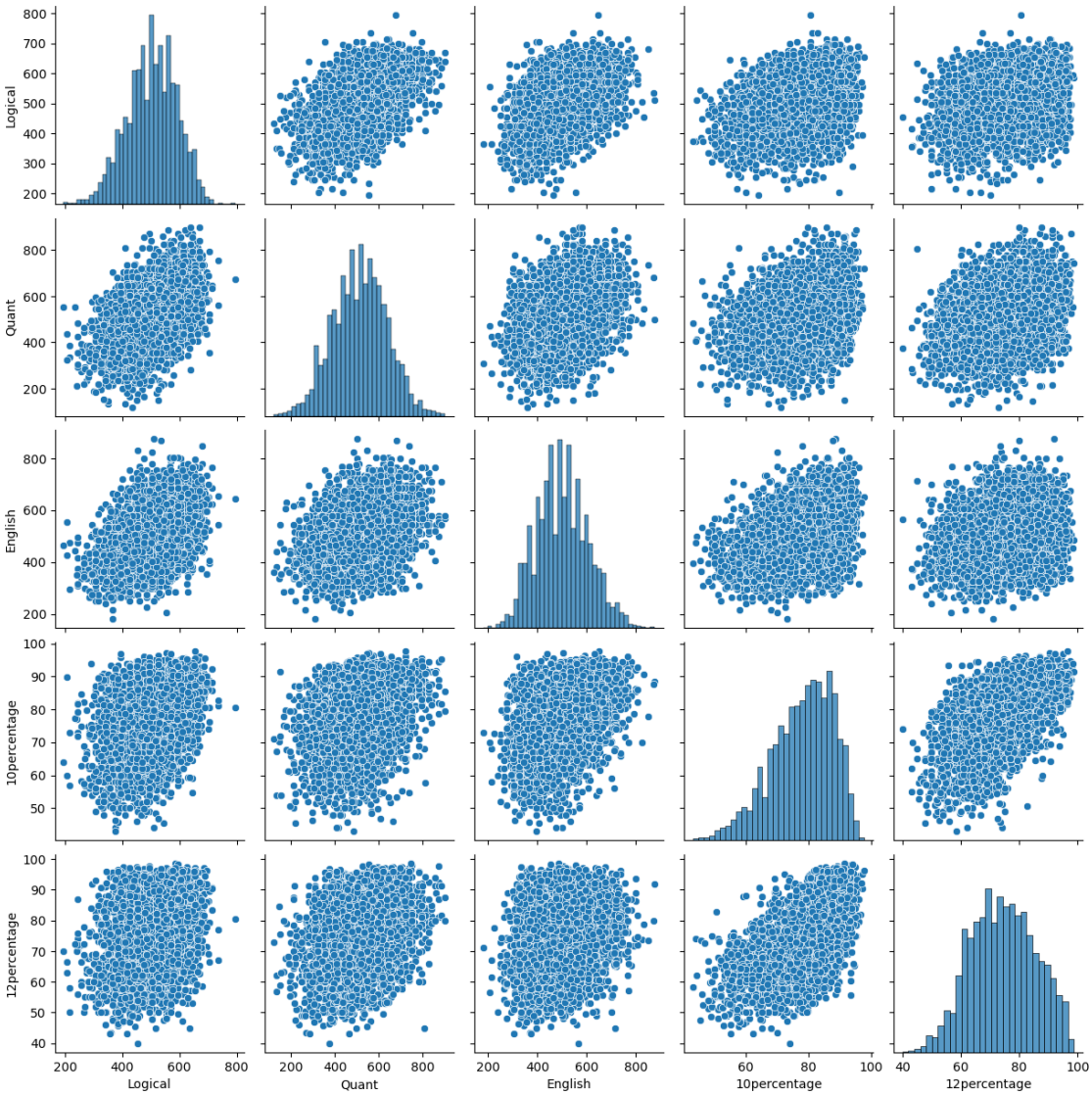
- **Swarm plot :**



Observations:

- Both **Male and Female** have scores between 40% and 95% in their 10th-grade exams.
- Many candidates, in both groups, scored above 80%, showing good performance.
- **Female's** scores are mostly above 60%, while **Male's** scores are more spread out, with some scoring lower.

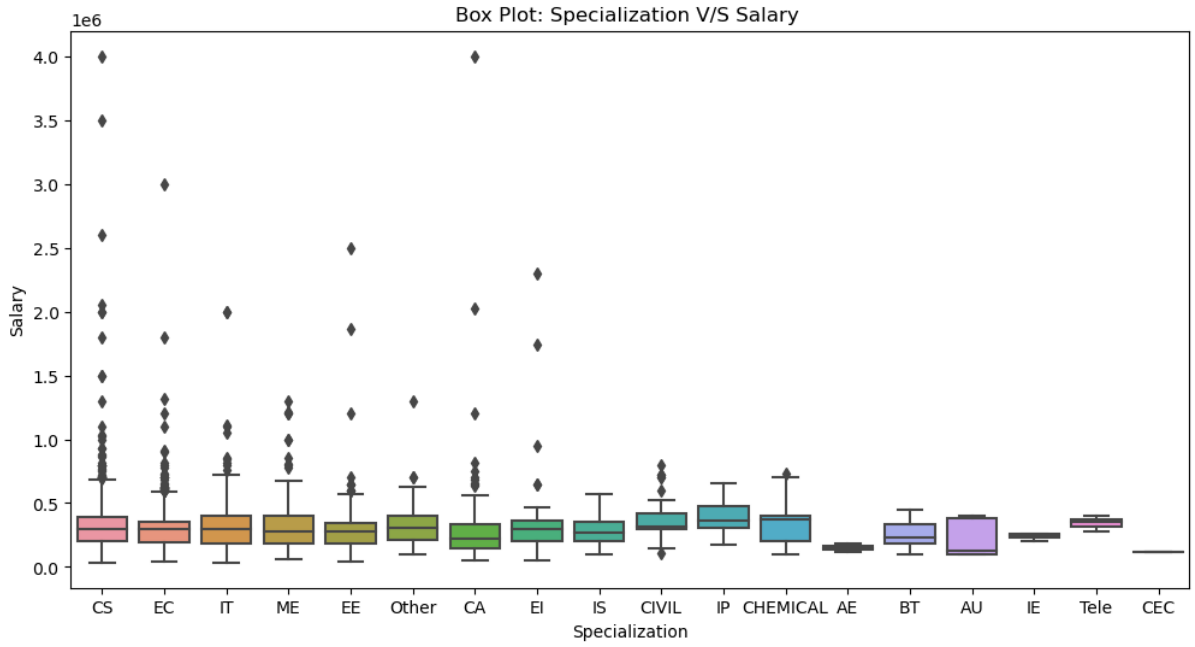
- **Pairplot** for feature interaction: Provide a compact way to explore multiple variable relationships.



Observations from the Pair Plot

- **Diagonal Histograms**: Each subject (Logical, Quant, English, etc.) follows a normal-like distribution.
- **Scatter Plots**:
  - There is a **positive correlation** between Logical, Quant, and English scores.
  - **10th and 12th percentages** also show a mild correlation with the subject scores.
  - The data points are **well-distributed** without extreme clustering or strong outliers.

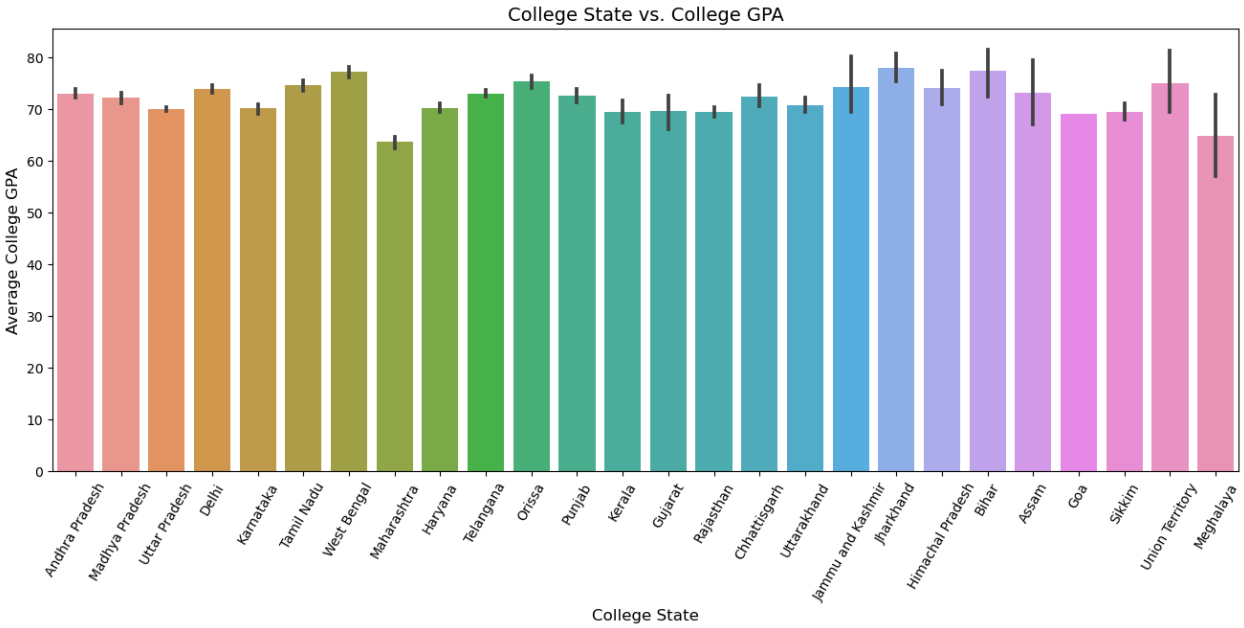
- **Box plot:**



Observation:

- Candidates with **CS** specialization have **highest** Outliers

Bar Plot:



Observation:

- **Jharkhand , Bihar and West Bengal** states colleges have **highest** College GPA

5. Key Findings & Conclusion

- Candidates strong in logical reasoning also perform well in quantitative and English sections.
- Higher scores in 10th grade often predict better performance in 12th grade.
- CS and IT students dominate tech careers, while fields like Civil and Chemical have fewer placements.
- Most students have minimal experience, leading to lower starting salaries.