INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMCAT Dataset

Omkar Arun Shinde
ID : IN9240411

About me

I am currently pursuing a **B.Tech degree** in **Computer Engineering** at **Pimpri Chinchwad College of Engineering, Pune.**

Why you want to learn Data Science ?
With a strong foundation in mathematics, having been selected for the *Indian National Mathematical Olympiad (INMO) in 2020*, I have always been deeply passionate about problem-solving and analytical thinking. This love for mathematics naturally led me to explore data science, where I can combine my analytical skills with programming to uncover patterns, insights, and solutions in complex datasets.
I am particularly drawn to data science because it offers a perfect blend of my interests in statistics, machine learning, and data-driven decision-making. The field's ability to transform raw data into actionable knowledge fascinates me, and I love the challenge of making sense of numbers, finding trends, and predicting outcomes.

About me

Beyond academics, I am also passionate about poetry, novel reading, and acting, all of which help me cultivate a creative and critical mindset. These interests have shaped my approach to data science, allowing me to think outside the box and approach problems with a fresh perspective.

This is my first internship, and I am currently enjoying every moment of it. I am gaining a wealth of knowledge, particularly in exploratory data analysis, Python, and machine learning. It has been an enriching experience, providing me with hands-on learning opportunities and a deeper understanding of the data science field.

https://www.linkedin.com/in/omkararunshinde/

https://github.com/Omkar-Shinde12/

Agenda

- Business Problem and Use case domain understanding(If Required)
- Objective of the Project
- Web Scraping – Details (Websites, Processor you followed)
- Summary of the Data

- Exploratory Data Analysis:

a. Data Cleaning Steps

b. Data Manipulation Steps

c. Univariate Analysis Steps

d. Bivariate Analysis Steps

- Key Business Question
- Conclusion (Key finding overall)
- Q&A Slide
- Your Experience/Challenges working on Web Scraping – Data Analysis Project.

Business Problem Statement

Engineering graduates face a competitive job market, with varying outcomes in salary, job titles, and locations. The challenge is to identify which factors, such as cognitive, technical, personality skills, and demographics, most influence employment success.

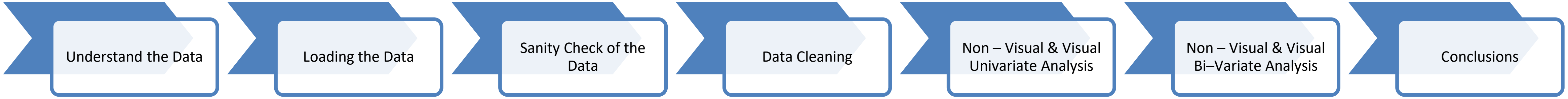
Using the Aspiring Minds Employment Outcome 2015 dataset, we will analyze how these factors impact job outcomes. By identifying key predictors, we will provide insights for universities to improve curricula and for employers to refine hiring practices. Predictive models will help guide career development and recruitment strategies.

Our mission is to bridge the gap between education and employment by empowering institutions, employers, and students with data-driven insights to improve career success for engineering graduates.

Objective

- The project aims to investigate the impact of educational background on salary by analyzing how different qualifications and institutions affect earnings among engineering graduates.
- It will also examine the influence of demographic factors, such as age, gender, and location, on salary variations.
- To gain deeper insights, univariate analysis will be conducted to explore the distribution of salary concerning individual factors, while bivariate analysis will investigate relationships between salary and multiple influencing factors, including education and personality traits.
- Overall, the goal is to identify key trends and correlations that contribute to salary outcomes.

Workflow:-


Understand the Data
Loading the Data
Sanity Check of the Data
Data Cleaning
Non – Visual & Visual Univariate Analysis
Non – Visual & Visual Bi–Variate Analysis
Conclusions

Summary of data

The AMCAT data initiative seeks to examine the relationships between various determinants and salary outcomes for individuals.

1. Essential factors for analysis include educational background, which encompasses the level of education, the field of study, and the reputation of the college attended.
2. Demographic characteristics, such as age, gender, and geographic location.
3. The Big Five personality traits are: Openness to Experience, Conscientiousness, Extraversion, Agreeableness, Neuroticism
4. Areas of specialization, indicating the specific disciplines in which individuals have expertise.
5. The project focuses on salary as the primary variable of interest, aiming to uncover which of these factors exert the strongest influence on salary levels.
6. The insights generated from this analysis are intended to aid individuals in making well-informed career choices and to enhance the understanding of what drives salary variations in the labor market.

Univariate Numerical Analysis

Outliers Detection and Removal:

- Descriptive Statistics: Calculated measures such as *mean, median, mode, standard deviation, and range* to summarize the main features of the numerical data (*Shape – (3998, 39)*).

Outlier Detection:

- Boxplots: Utilized boxplots to visually detect outliers in the dataset.

Outlier Removal:

- Outliers identified as being *less than 5%* of the data were *directly removed* (*Shape – (3711, 39)*).
- For outliers *between 5% and 10%,* a *combination of the Interquartile Range (IQR) and Z-score methods* was applied to assess their impact

  (*Shape – (3685, 39)*).
- After this initial filtering, the remaining data was further analyzed using the *IQR method* to identify and remove any additional outliers

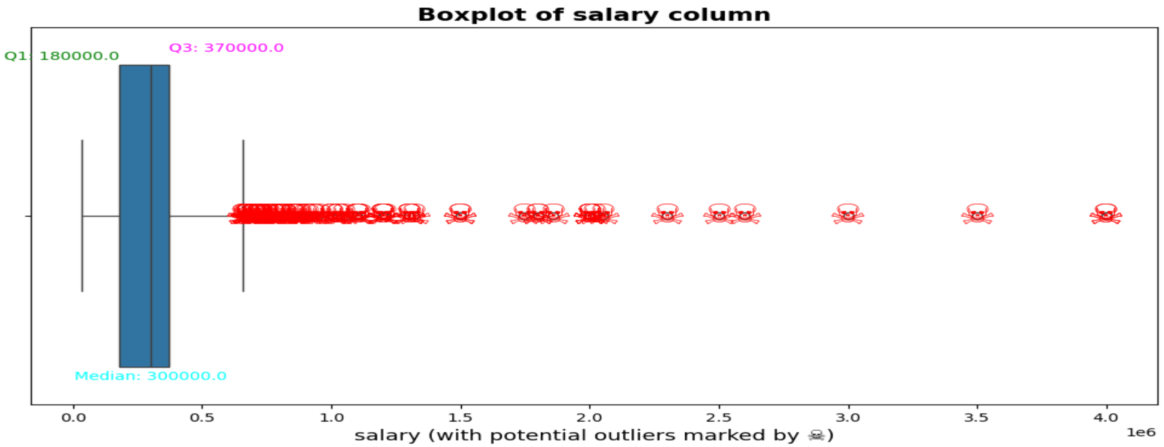
  (*Shape – (3685, 39)*).

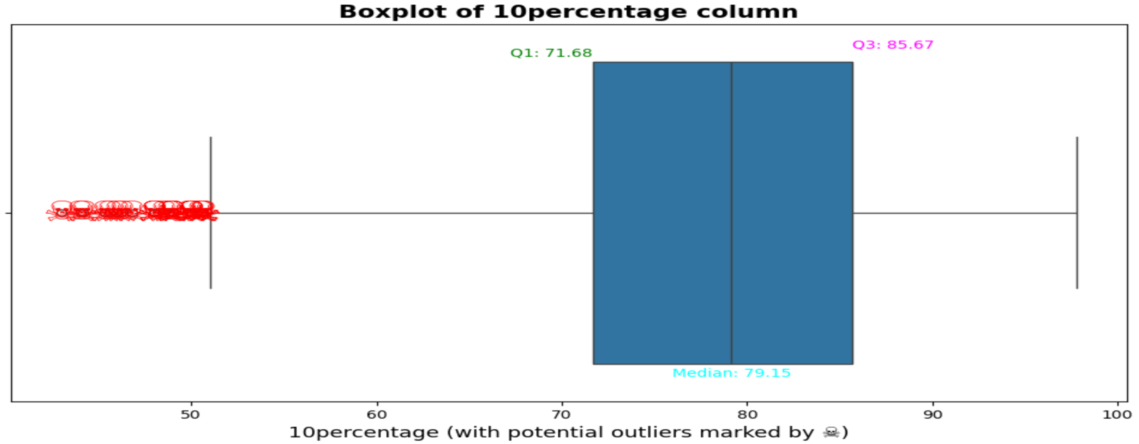
Data Visualization:

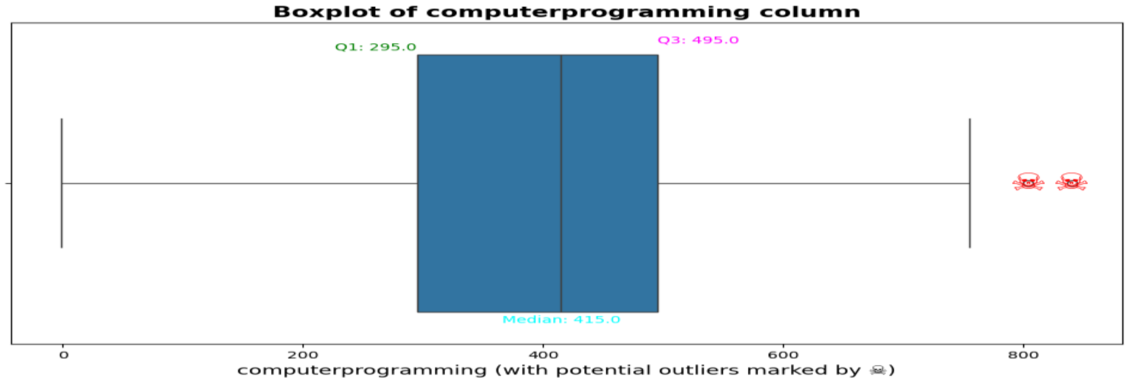
- Generated boxplots to provide visual insights into the data distribution and the effect of outlier removal on the dataset.

Final Dataset:

- Prepared a cleaned dataset for further analysis, ensuring that the influence of outliers was minimized while retaining relevant data points.







INNOMATICS RESEARCH LABS

Univariate Numerical Analysis

Plotting Frequency Distribution Graphs like Histogram, Density Graph, KDE:

Here are the inferences I derived after removing outliers and analyzing the data with bar plots and frequency distribution density plots:

1. Salary Distribution: The salary graph is *right-skewed*, with *most salaries falling below 1,000,000 (10 lakhs)*. Those with salaries exceeding this amount are primarily outliers, indicating a *leptokurtic distribution*.

2. 10th Percentile Scores: The graph for the 10th percentile is *left-skewed*, showing that *nearly 50% of individuals scored above 80% (median = 79.40)*. This distribution is *platykurtic.*

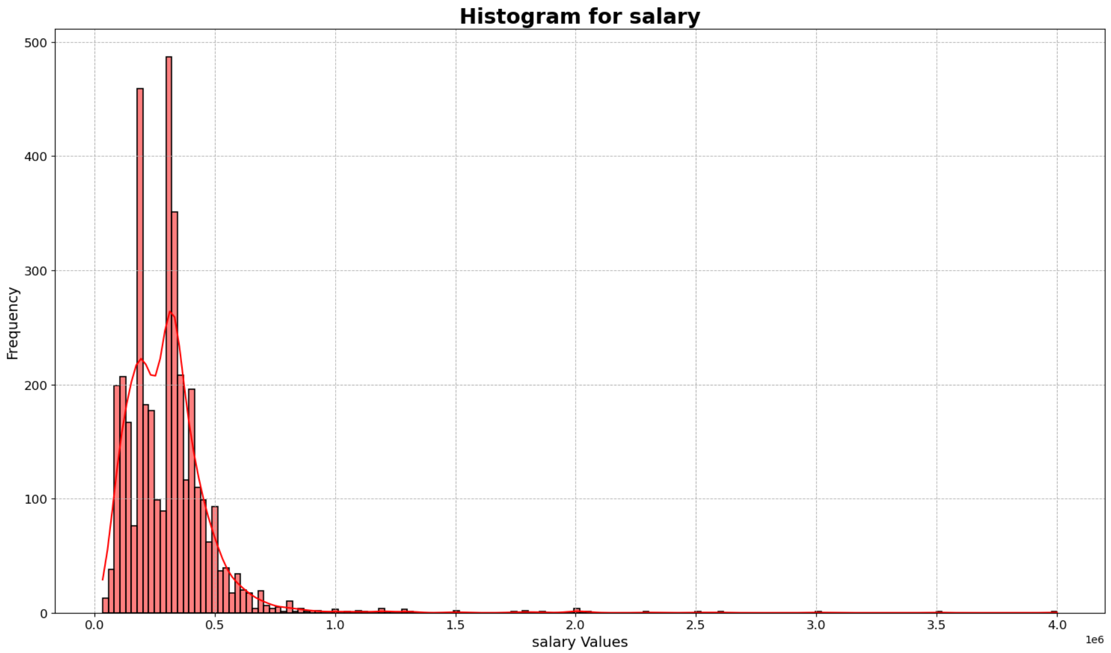
3. 12th Percentile Scores: The graph for the 12th percentile is *slightly left-skewed*, with *nearly 50% of individuals scoring above 75% (median = 74.40)*, also indicating a *platykurtic* distribution.

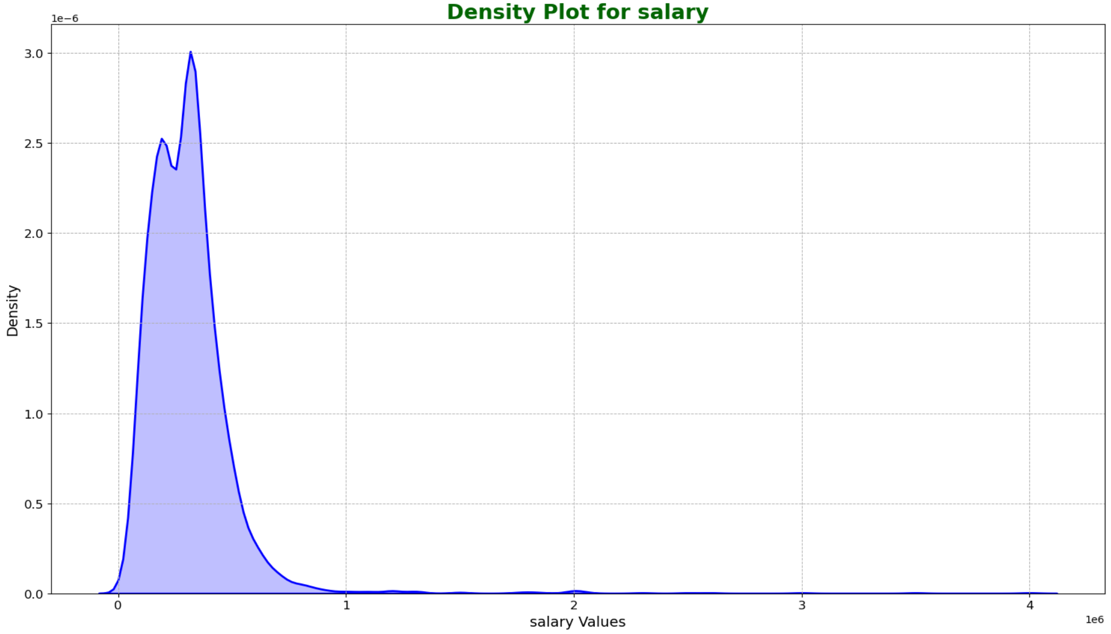
4. College GPA: The college GPA graph is *slightly right-skewed*, with *nearly 50% of individuals scoring above 70% (median = 71.90)*, suggesting a *platykurtic* distribution.

5. Graduation Year: A significant number of individuals graduated from the 12th grade between *2006 and 2011*, accounting for *nearly 50%* of the dataset.

6. College Tier: Most students (*approximately 80%)* attended *tier-2 colleges*, indicating a predominance of this college tier among the graduates.

These inferences provide valuable insights into the dataset and the educational background of the individuals analyzed.





INNOMATICS
RESEARCH LABS

Univariate Categorical Analysis

Plotting Frequency Distribution Graphs like Count Plot:

Here are inferences from the count plot for univariate categorical analysis:

1. Designation Distribution: The highest job count is observed for *software engineers*, followed by *software developers* as the second highest and *system engineers* in third place.

2. Job City Distribution: *Bangalore* has the highest number of jobs, with *Noida* ranking second and *Hyderabad* third.

3. Gender Distribution: The analysis shows that there are more *working men* compared to *working women* in the dataset.

4. 10th Board Distribution: Most individuals reported a local board name for their 10th grade, with *CBSE* being the most common, followed by *ICSE*.

5. 12th Board Distribution: Similar to the 10th board, most individuals reported a local board name for their 12th grade, with *CBSE* again being the most prevalent, followed by *ICSE*.

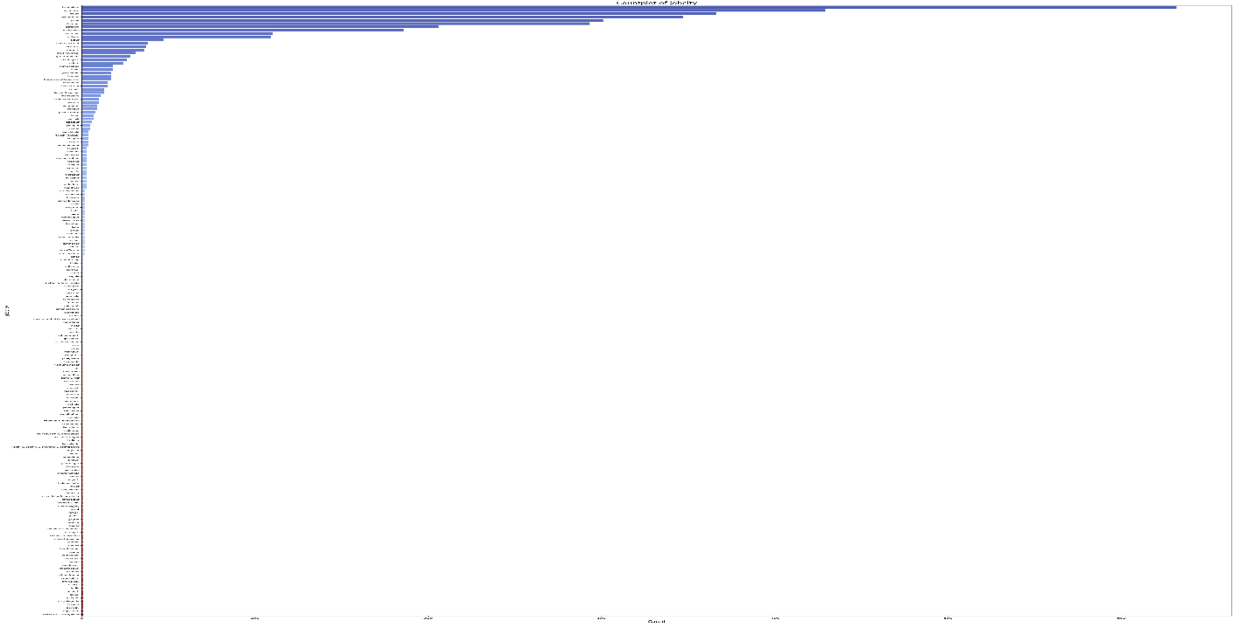
6. Educational Background: The majority of individuals in the dataset are *B.Tech or B.E. graduates*.

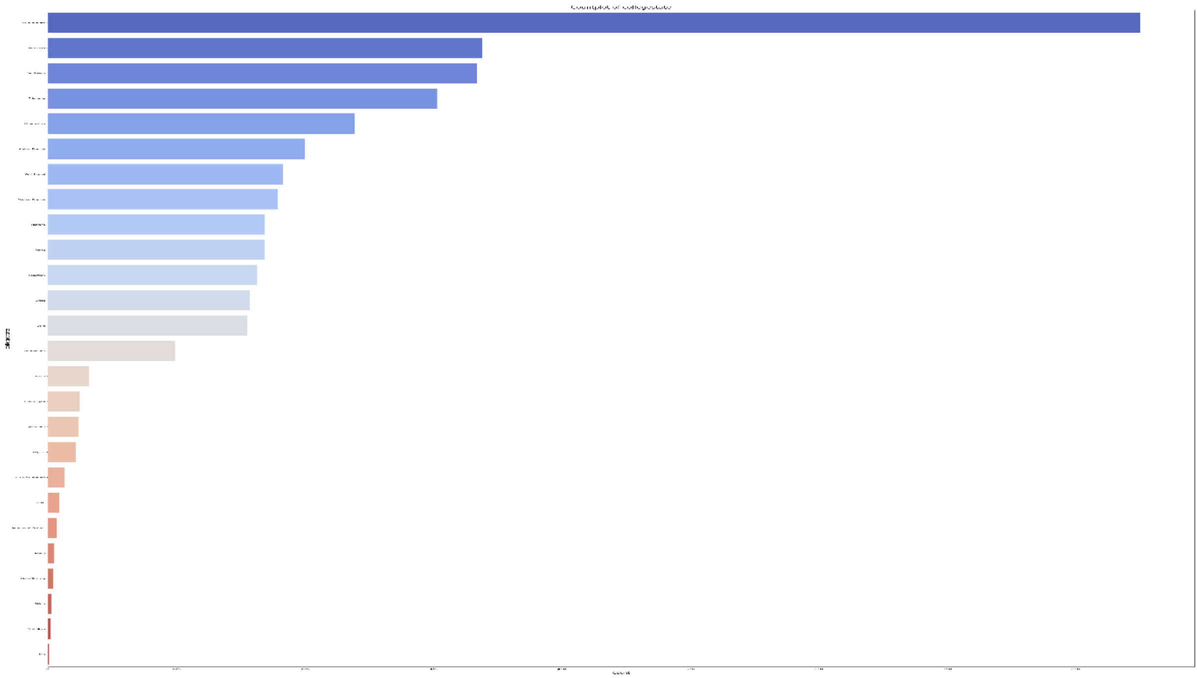
7. Branch Distribution: Most individuals belong to the *CSE (Computer Science Engineering)* branch, followed by *ENTC (Electronics and Telecommunication).*

8. College State Distribution: *Uttar Pradesh* has the highest number of graduates, followed by *Karnataka* and *Tamil Nadu*.

9. Common Roles Distribution: The analysis indicates that most individuals occupy *engineering roles*, with *developer roles* and *analyst roles* following.

These insights provide a comprehensive overview of the categorical data in your analysis.





Bi-Variate Analysis : Numerical V/s Numerical

Plotting Graphs like Bar Plot and Box Plot:

Here are inferences from the bivariate analysis using numerical vs. numerical bar plots and box plots:

1. Salary vs. College Tier:

Individuals who graduated from *college tier 1* have a mean salary of *433,603.0*, significantly higher than the mean salary of *296,680.0* for those from *college tier 2*. This indicates that the tier of the college attended plays a crucial role in salary outcomes.

2. Salary vs. 10th Percentile Scores:

Individuals who scored between *95 and 100* 9in the 10th standard have a mean salary of *381,389.1*. This is followed by those who scored between *90 and 95*, with a mean salary of *364,318.5*, and individuals who scored between *85 and 90*, who have a mean salary of *349,674.2*. The results suggest that higher 10th percentile scores are associated with higher salaries.

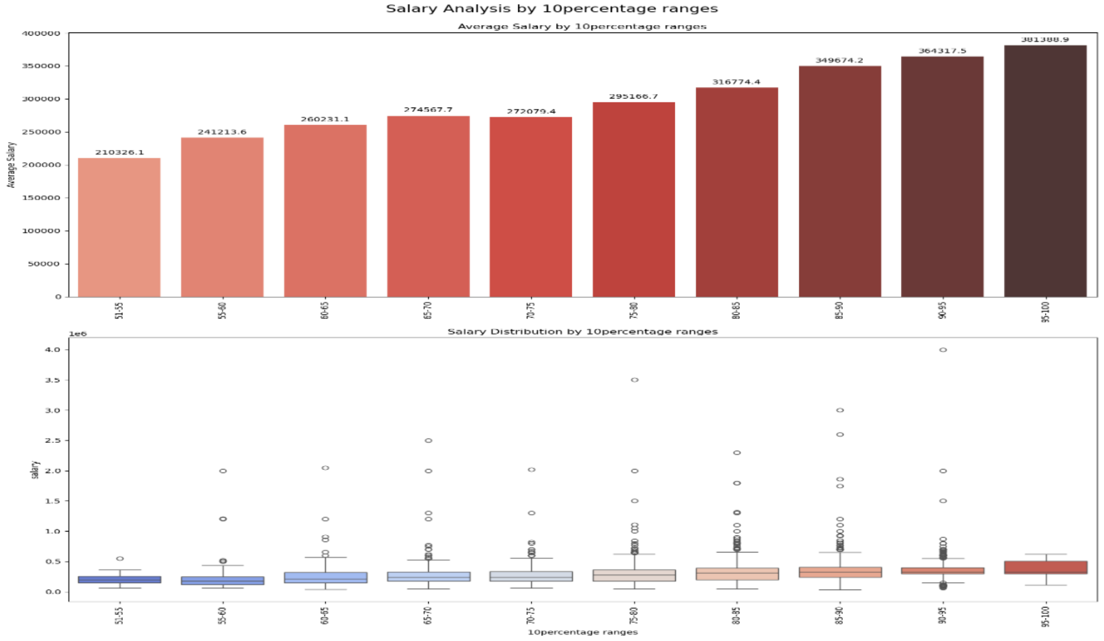
3. Salary vs. 12th Percentile Scores:

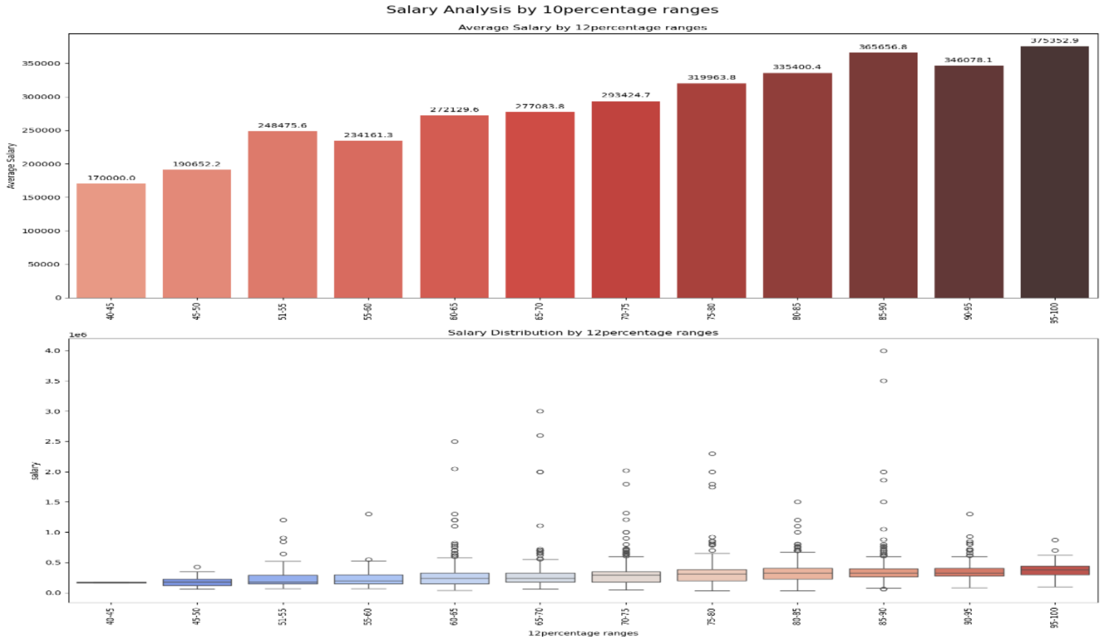
Individuals who scored between *95 and 100* in the 12th standard have a mean salary of *375,353.0*. Following this, those who scored between *90 and 95* have a mean salary of *365,657.0*, while individuals who scored between *85 and 90* have a mean salary of *346,078.0*. This indicates a similar trend as with the 10th percentile scores, where higher performance correlates with increased salary levels.

4. Salary vs. College GPA:

Individuals who scored between *85 and 90* in college GPA have a mean salary of *370,042.0*. This is followed by those who scored between *80 and 85*, with a mean salary of *348,780.0*, and individuals who scored between *90 and 95*, who have a mean salary of *347,750.0*. This suggests that better GPA performance in college is also linked to higher salary outcomes.

These analyses provide insight into how educational performance and college quality are associated with salary outcomes in your dataset.





INNOMATICS RESEARCH LABS

Bi-Variate Analysis : Numerical V/s Categorical

Plotting Graphs like Bar Plot and Box Plot:

Here are inferences from the bivariate analysis of numerical vs. categorical variables using bar plots and box plots:

1. Salary vs. Common Role:

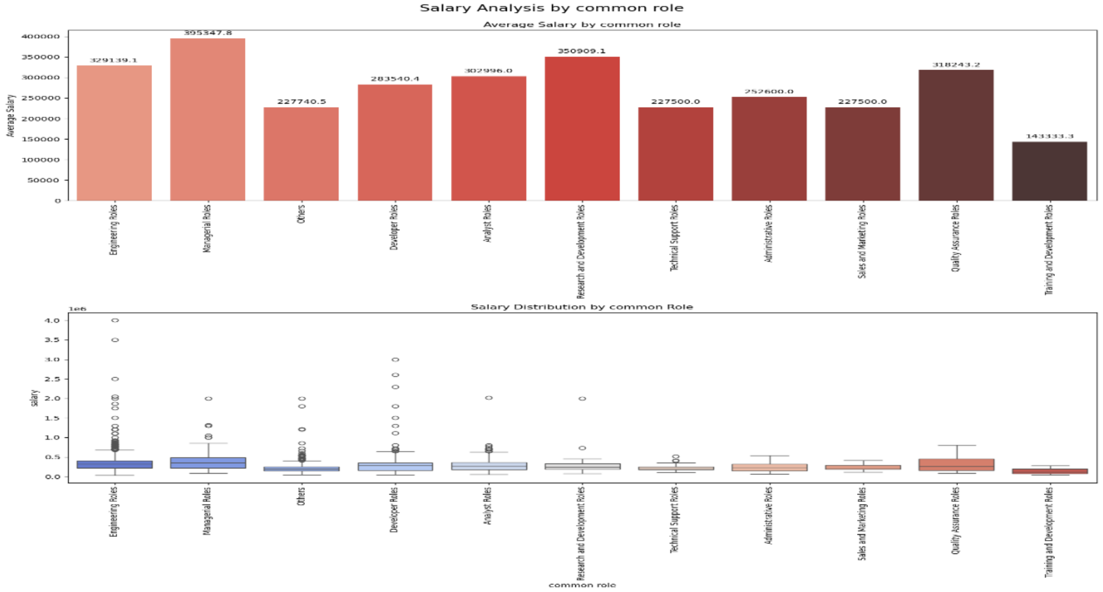
Managerial roles have the highest mean salary of *395,348.0*. This is followed by *research and development roles*, with a mean salary of *350,909.0*. *Engineering roles* come in third, with a mean salary of *329,139.1*, indicating that individuals in managerial and research positions tend to earn more than those in engineering roles.

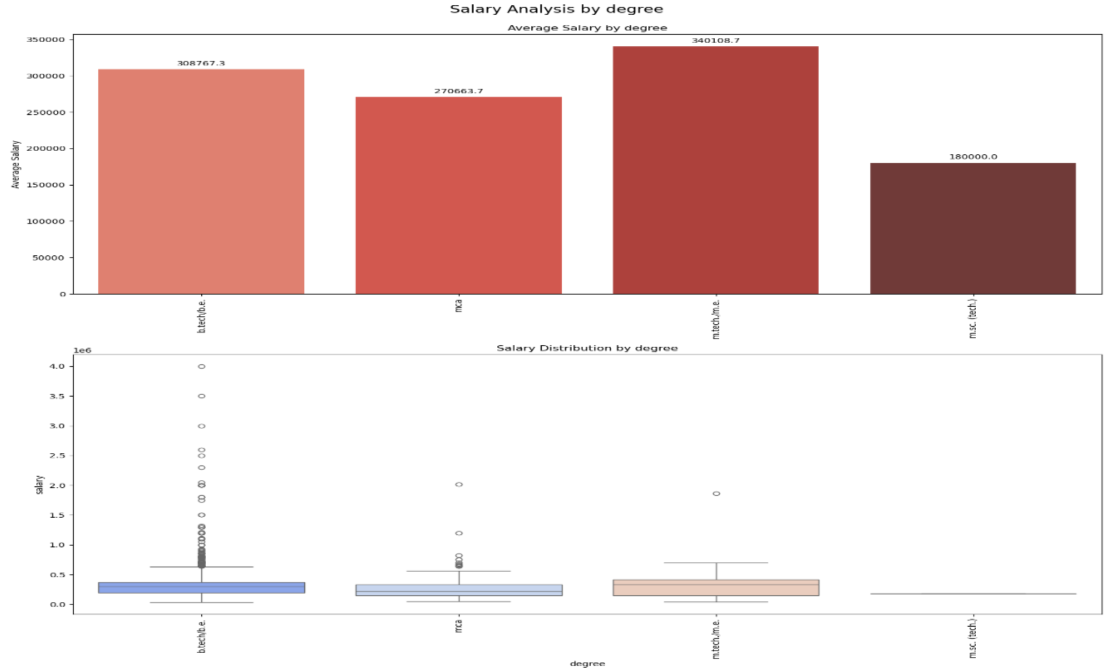
2. Salary vs. Specialization: *Civil engineers* have the highest mean salary, at *370,000.0*, followed by *computer science engineers*, with a mean salary of *311,437.6*. *Mechanical engineers* rank third, with a mean salary of *310,271.4*, suggesting that specialization in civil engineering leads to higher salary outcomes compared to other fields.

3. Salary vs. Degree: Individuals with an *M.Tech/M.E degree* have a mean salary of *340,109.7*, making them the highest earners. This is followed by those with a *B.Tech/B.E degree*, with a mean salary of *308,767.3*. *MCA degree* holders rank third, with a mean salary of *270,663.7*, indicating that advanced degrees generally lead to higher salaries.

4. Salary vs. Gender: *Male employees* have a higher mean salary of *311,416.9,* while *female employees* have a mean salary of *292,564.0*. This suggests a gender disparity in earnings, with male employees earning more on average than their female counterparts.

These inferences provide insights into how roles, specialization, education, and gender impact salary levels in your dataset.





Bi-Variate Analysis : Categorical V/s Categorical

Plotting Graphs like Stacked Bar Plot:

Here are inferences from the bivariate analysis of categorical vs. categorical variables:

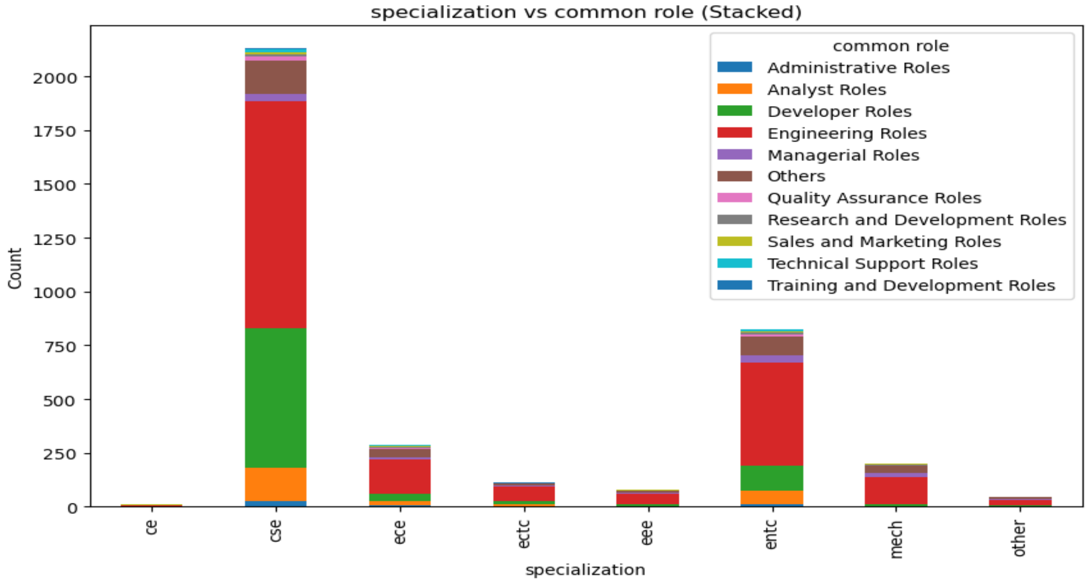
1. Specialization vs. Common Role:

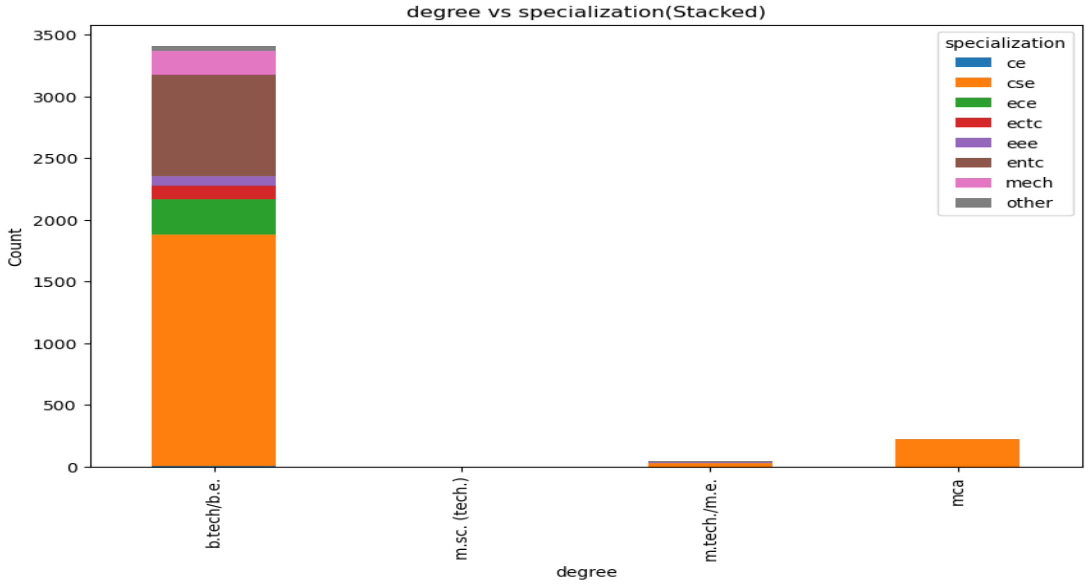
Most of the *engineering roles* are held by students specializing in *Computer Science Engineering (CSE),* followed by those specializing in *Electronics and Telecommunication Engineering (ENTC).* The same pattern is observed for *developer roles*, where *CSE students* dominate, followed by *ENTC students*.

2. Degree vs. Specialization:

A majority of individuals who pursued a *B.Tech/B.E degree* have specialized in *CSE*, followed by those who specialized in *ENTC*. This indicates that CSE and ENTC are the most popular specializations among engineering graduates in this dataset.

These insights highlight trends in how specific specializations are distributed across job roles and degrees.

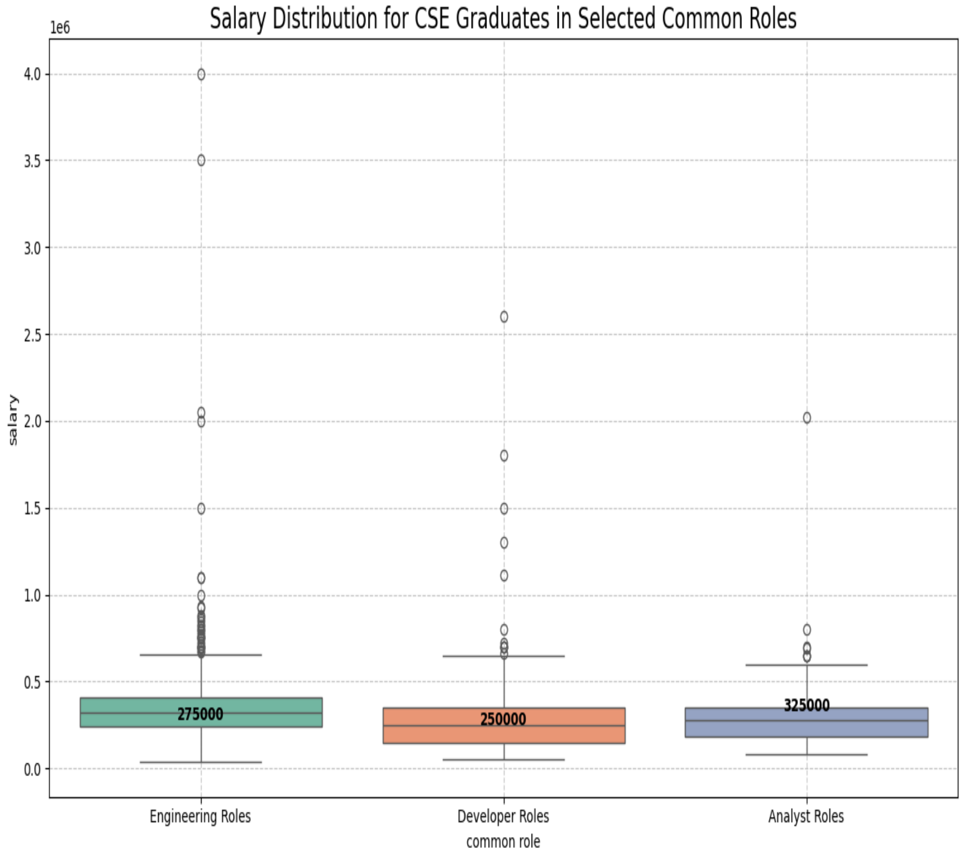


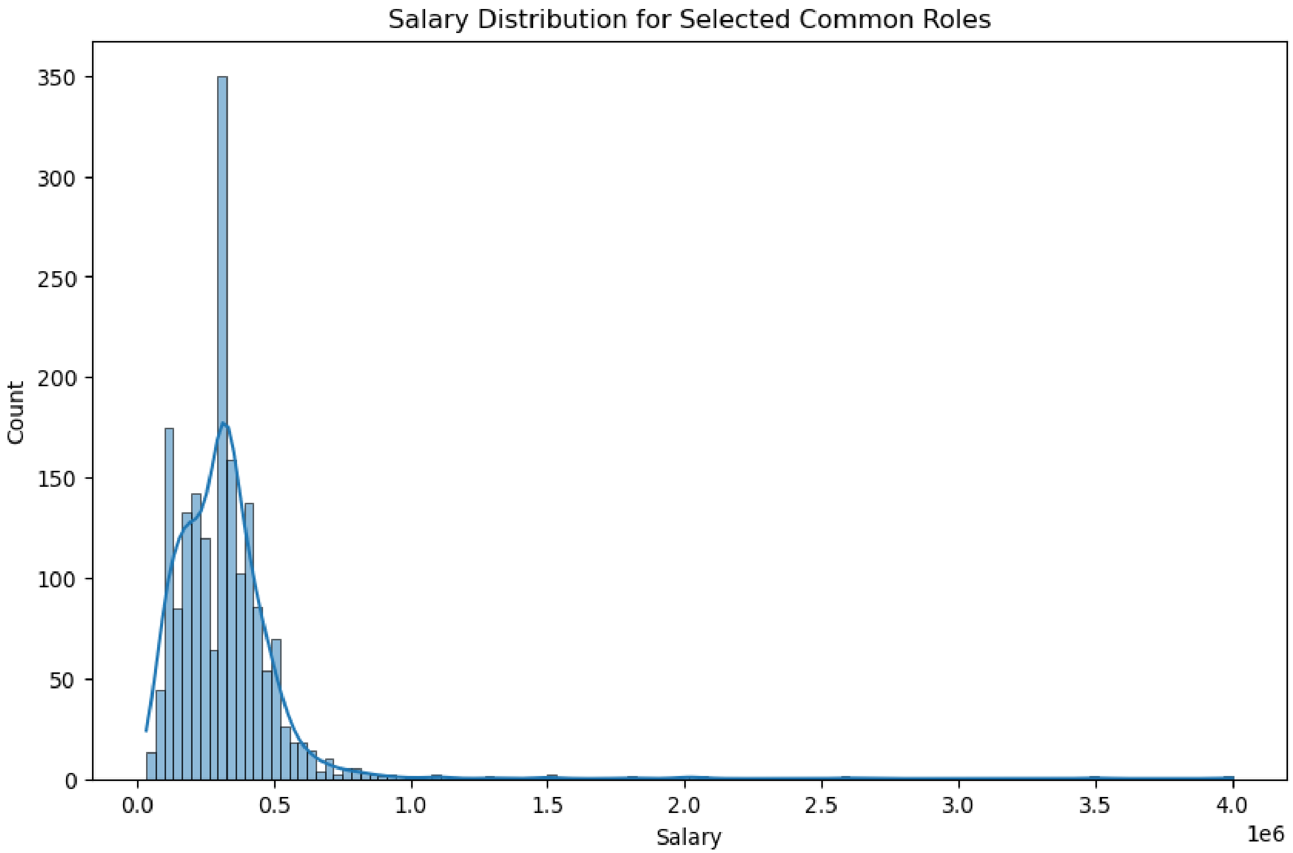


INNOMATICS RESEARCH LABS

Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate... Test this claim with the data given to you.

Fail to Reject Null Hypothesis (H1): The claim that Recent graduates can earn up to 2.5 - 3 lakhs is supported by the data.

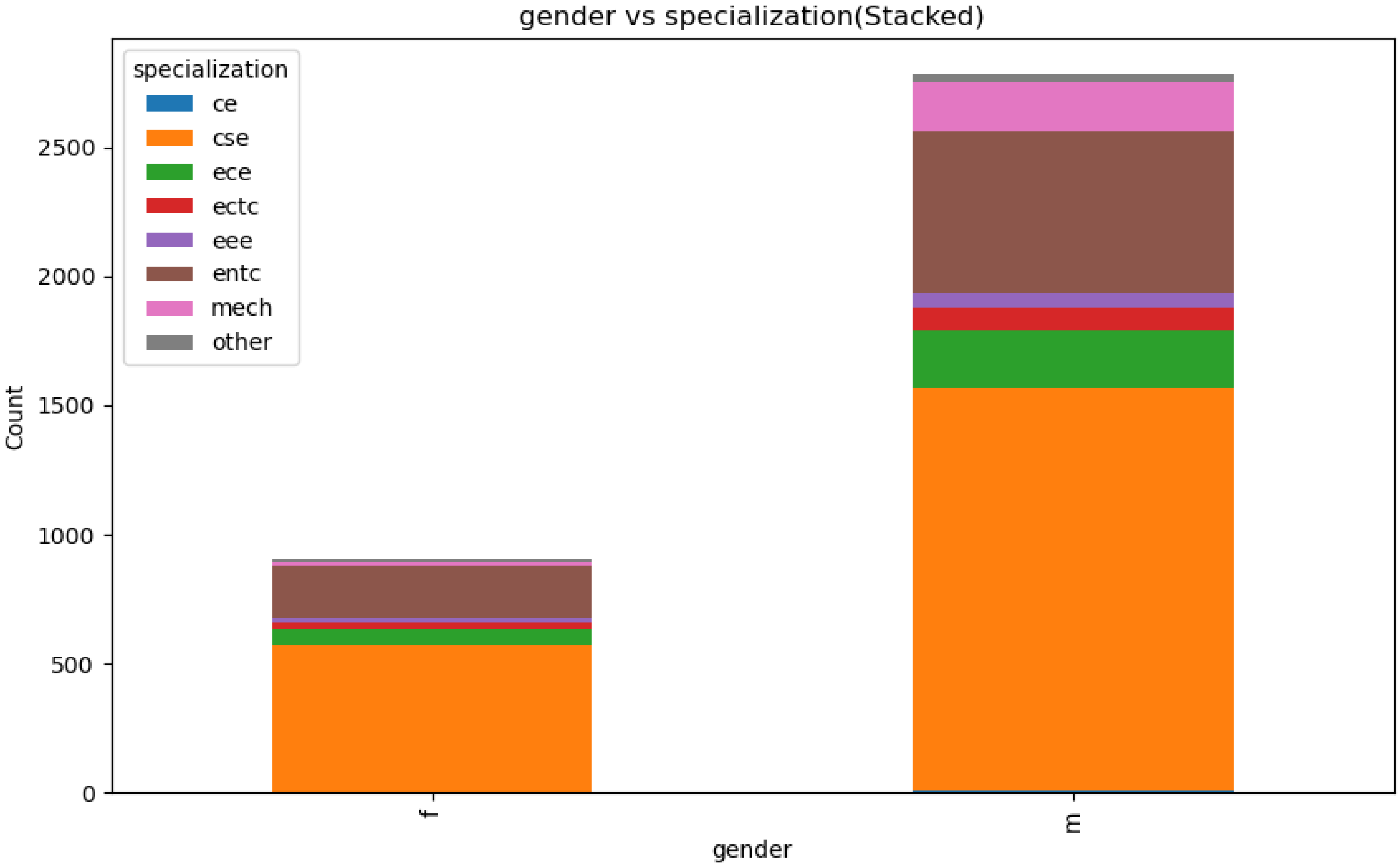




Gender V/s Specialization :

Here’s the inference from the Gender vs. Specialization analysis:

The graph shows that while both *male and female students* prefer specializing in *Computer Science Engineering (CSE), male students* significantly *outnumber female students* in this field. Other specializations, such as *Electronics and Telecommunication Engineering (ENTC)* and *Electronics and Communication Engineering (ECE)*, are also pursued by *both genders*, but *CSE* remains the most popular overall, especially among *male students*. This suggests a gender disparity in specialization choices, with *more male representation in technical fields like CSE.*



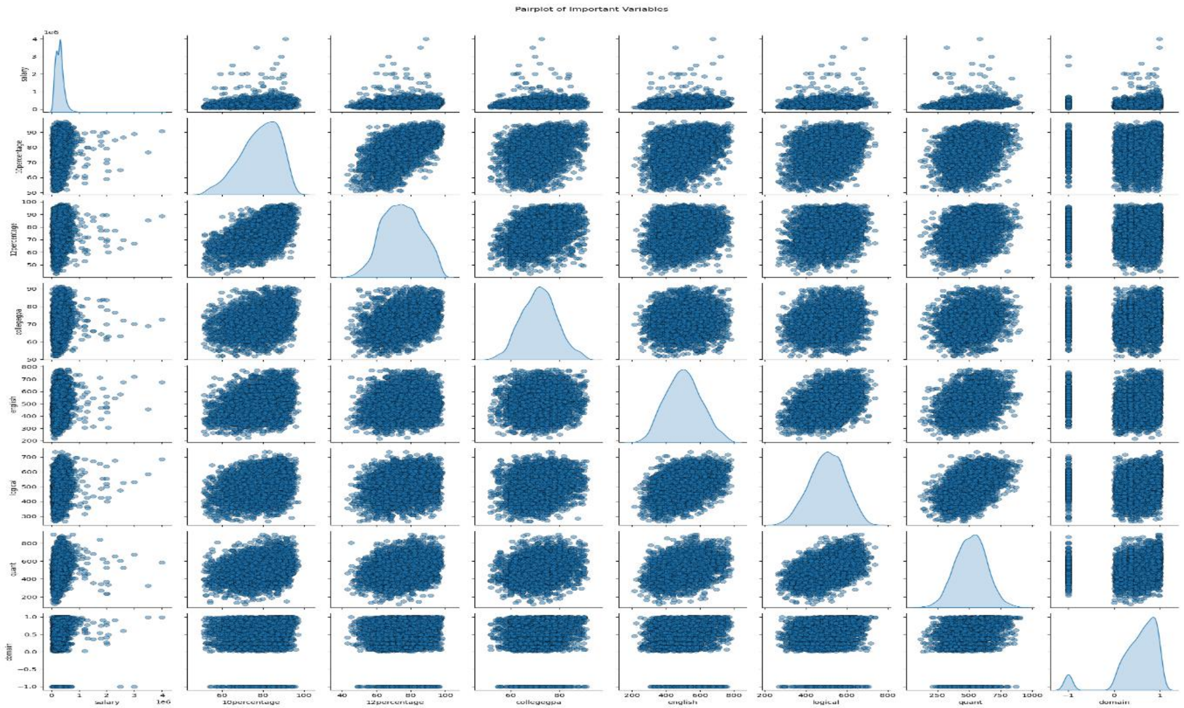
Bonus Insights :

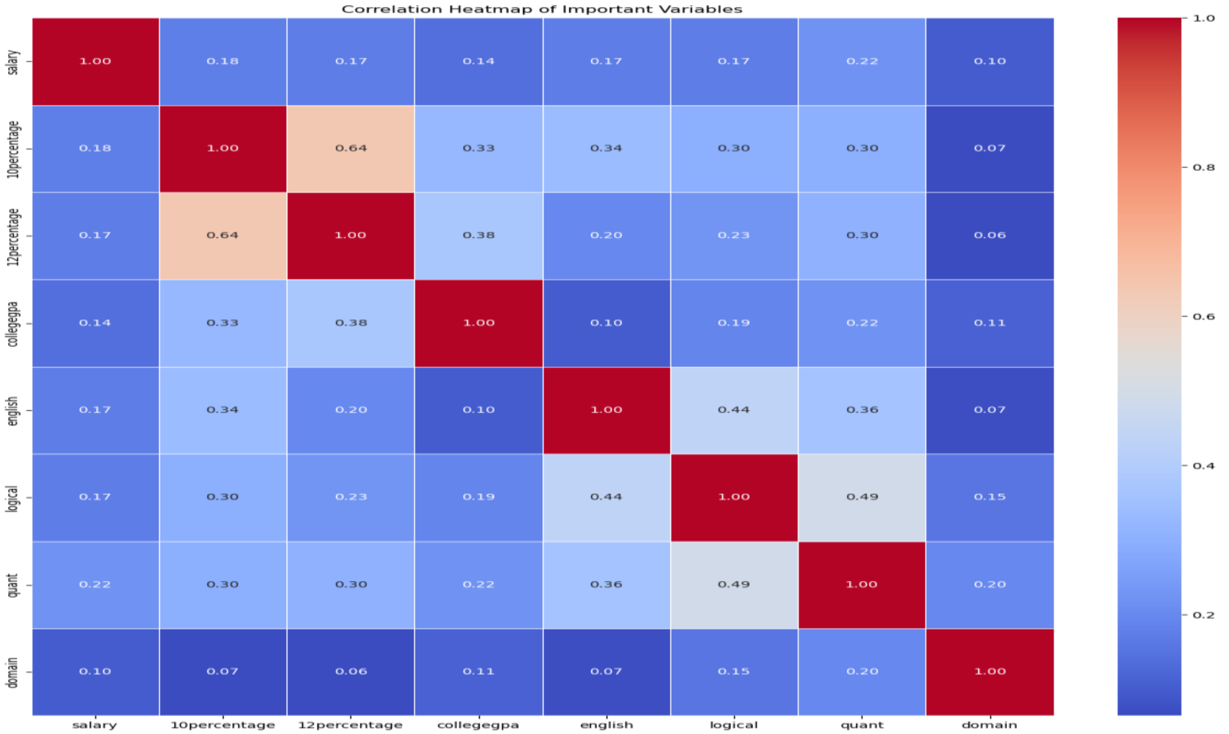
Here are your inferences from the bivariate analysis using pair plots and heatmaps:

1. Salary Relationships: The salary attribute *does not exhibit* a clear positive or negative relationship with the other numerical columns. However, it demonstrates a *positive correlation* with the *10th percentile scores, 12th percentile scores, and college GPA.*

2. Skill Correlations: The attributes for *English, logical reasoning, and quantitative skills* show *a positive relationship with each other*, indicating that higher proficiency in one area is associated with higher proficiency in the others.

These insights provide valuable information about the interrelationships among numerical variables in your dataset.





Final Conclusion

1. High Demand for Technical Expertise:
Bachelor of Technology/Engineering graduates dominate the job market due to the increasing need for strong technical skills.
2. Computer Science & Engineering (CSE) Leads in Salary:
CE specialization boasts the highest median salary at *₹311437.0* annually.
3. Managerical roles in High Demand:
The *Managerical role* employs the largest number of graduates, indicating strong market demand for this position.
4. Realities of Graduate Salaries:
Contrary to popular belief, the data does support the assumption of ₹2.5-3 lakh starting salaries for fresh Computer Science graduates.
Graduates with a *B.Tech/B.E. degree* generally expect an average salary of *₹250,000* annually.
5. Impact of College Tier on Earnings:
Graduates from *Tier-1 colleges* consistently earn more than those from lower-tier institutions, demonstrating the significant advantage of attending top-ranked colleges.
6. Gender-Based Salary Differences:
On average, *females earn ₹292564.0*, slightly lower than the *₹311417.0 earned by males*.
The gap between genders is minor, though the reasons behind this disparity require further investigation.
7. No Significant Relationship Between Gender and Specialization:
There is no notable correlation between *gender and specialization* preferences, contradicting common assumptions.
Females tend to specialize in *Information Technology (IT),* while males lean toward *Computer Science*.
8. Managerial and R & D Roles Are Highest Earners:
Managerial and R & D positions offer the highest salaries, reflecting the value placed on leadership and specialized expertise.

Experience and Challenges :

Experiences:
Throughout this project, I gained valuable experience in several key areas:
Python Programming: My knowledge of Python helped me effectively manage the dataset. I made use of libraries like Pandas for data handling, Matplotlib and Seaborn for visualization, and NumPy for numerical analysis, which streamlined the entire data analysis process.
Data Analysis & Exploration (EDA): I focused on cleaning the dataset, exploring patterns and relationships, and conducting detailed Exploratory Data Analysis (EDA) to uncover insights from the data.
Graph Interpretation: Using bar plots, box plots, heatmaps, and pair plots, I examined the relationships between salary and factors like education, specialization, and college tier, gaining key insights through visualization.
Complete EDA Workflow: I followed the full workflow from data preprocessing, univariate and bivariate analysis, to generating meaningful conclusions. This helped in understanding the data's structure and patterns effectively.

Challenges:
Time Management: Balancing my academic work at PCCOE, Pune, with internship responsibilities proved challenging. Managing deadlines and finding time for both required significant effort, but it helped me improve my time management skills.
LinkedIn Profile Merging Issue: A technical problem with merging my LinkedIn profiles delayed my ability to post project updates. I had to wait for this issue to be resolved to ensure that my posts weren’t lost during the merge process, contributing to the delay.
Late Submission: Due to the time constraints from balancing college and internship work, along with the LinkedIn issue, I apologize for the delay in submitting this project. I appreciate your understanding and patience.

THANK YOU