



INNOVATION. AUTOMATION. ANALYTICS

## PROJECT ON

EDA Project - AMCAT Data Analysis

# About me

- Background ?

Basically, I am pursuing my bachelor's Degree in Chandigarh University.

- Why you want to learn Data Science

Data Science is such a interesting domain, and it deals with the real world data and focuses on getting the meaningful insights. And it provides the opportunities to meet different organisations and can make a good network and grow together.

- Any work experience

I am fresher with no experience but good in Analytical Skills and Machine Learning concepts.

- Share your linkedin and github profile urls

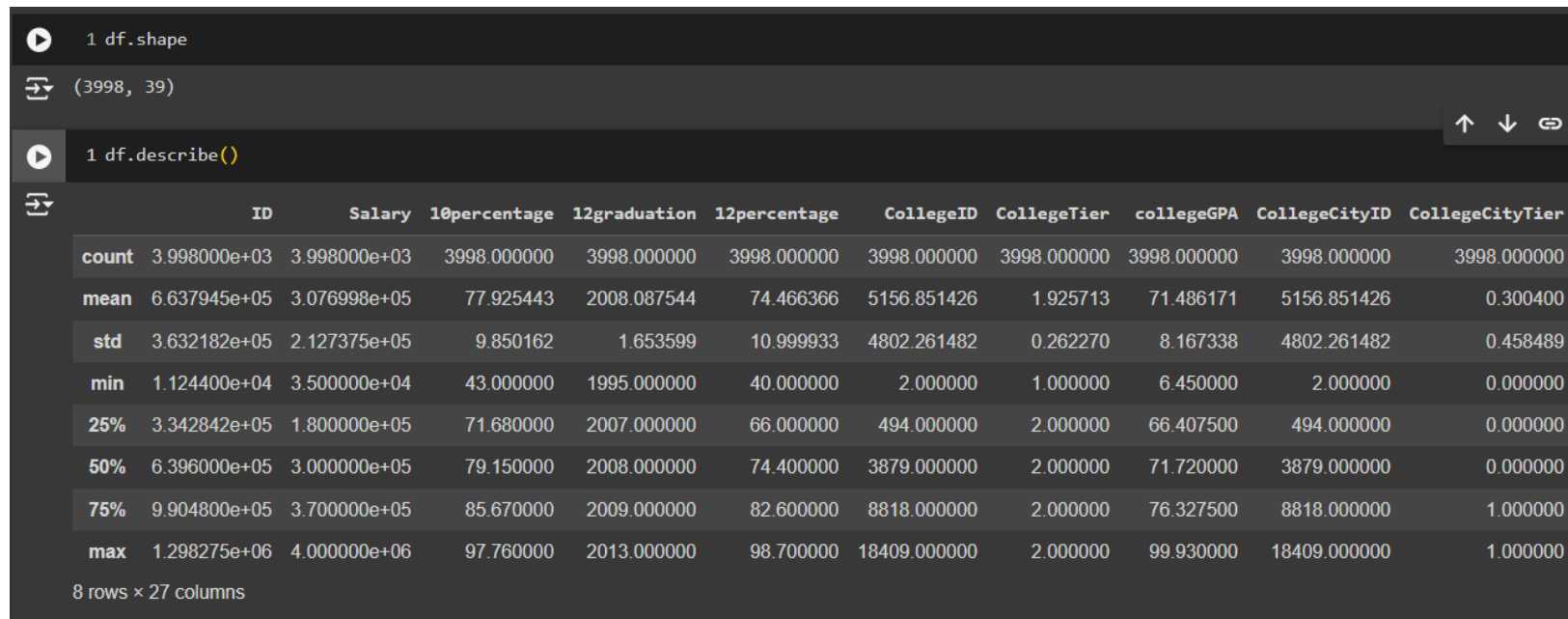
Git-hub: <https://github.com/Maheendra-mj>

Linkedin: <https://www.linkedin.com/in/maheendra-kada/>

# Agenda (This should be the PPT flow)

## About Dataset:

This dataset contains information about 3998 individuals, including their demographics, education, job experience, skills, and personality traits. It covers continuous variables such as salary and academic scores, as well as categorical variables like gender, job city, and educational specialization. The goal is to analyze this data, perform various statistical analyses, and answer specific research questions such as salary expectations and specialization preferences based on gender.



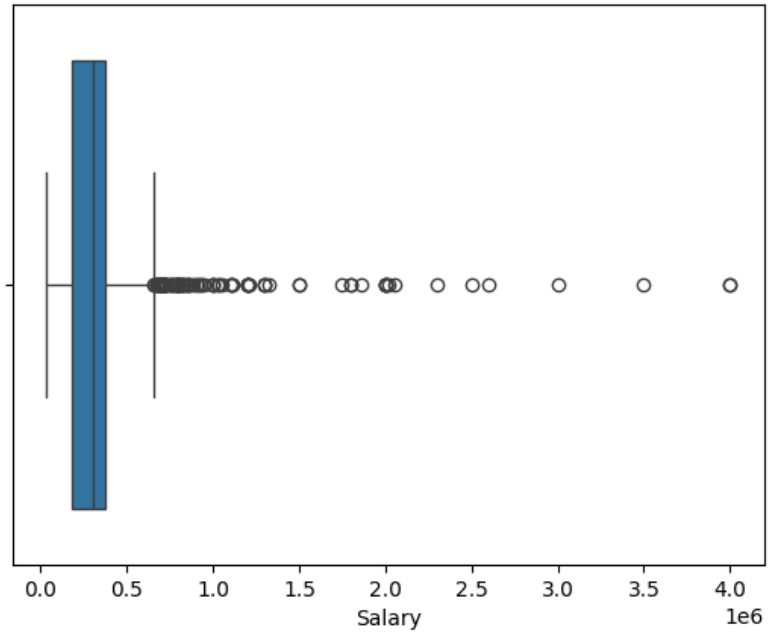
The screenshot shows a Jupyter Notebook interface with two code cells. The first cell contains `df.shape` and the second contains `df.describe()`. Below the code cells, a summary table is displayed, showing statistical data for 11 variables: ID, Salary, 10percentage, 12graduation, 12percentage, CollegeID, CollegeTier, collegeGPA, CollegeCityID, and CollegeCityTier. The table includes rows for count, mean, std, min, 25%, 50%, 75%, and max. At the bottom, it indicates '8 rows x 27 columns'.

	ID	Salary	10percentage	12graduation	12percentage	CollegeID	CollegeTier	collegeGPA	CollegeCityID	CollegeCityTier
count	3.998000e+03	3.998000e+03	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000	3998.000000
mean	6.637945e+05	3.076998e+05	77.925443	2008.087544	74.466366	5156.851426	1.925713	71.486171	5156.851426	0.300400
std	3.632182e+05	2.127375e+05	9.850162	1.653599	10.999933	4802.261482	0.262270	8.167338	4802.261482	0.458489
min	1.124400e+04	3.500000e+04	43.000000	1995.000000	40.000000	2.000000	1.000000	6.450000	2.000000	0.000000
25%	3.342842e+05	1.800000e+05	71.680000	2007.000000	66.000000	494.000000	2.000000	66.407500	494.000000	0.000000
50%	6.396000e+05	3.000000e+05	79.150000	2008.000000	74.400000	3879.000000	2.000000	71.720000	3879.000000	0.000000
75%	9.904800e+05	3.700000e+05	85.670000	2009.000000	82.600000	8818.000000	2.000000	76.327500	8818.000000	1.000000
max	1.298275e+06	4.000000e+06	97.760000	2013.000000	98.700000	18409.000000	2.000000	99.930000	18409.000000	1.000000

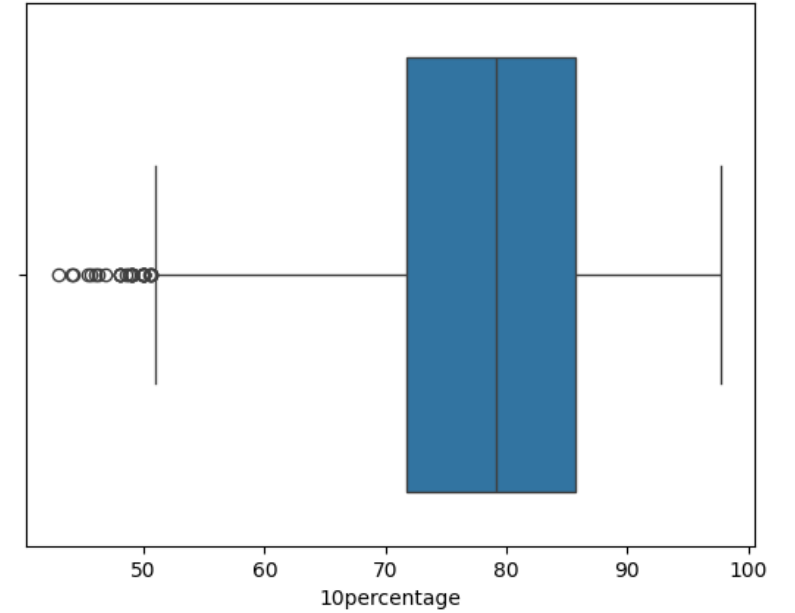
8 rows x 27 columns

# Outlier Identification:

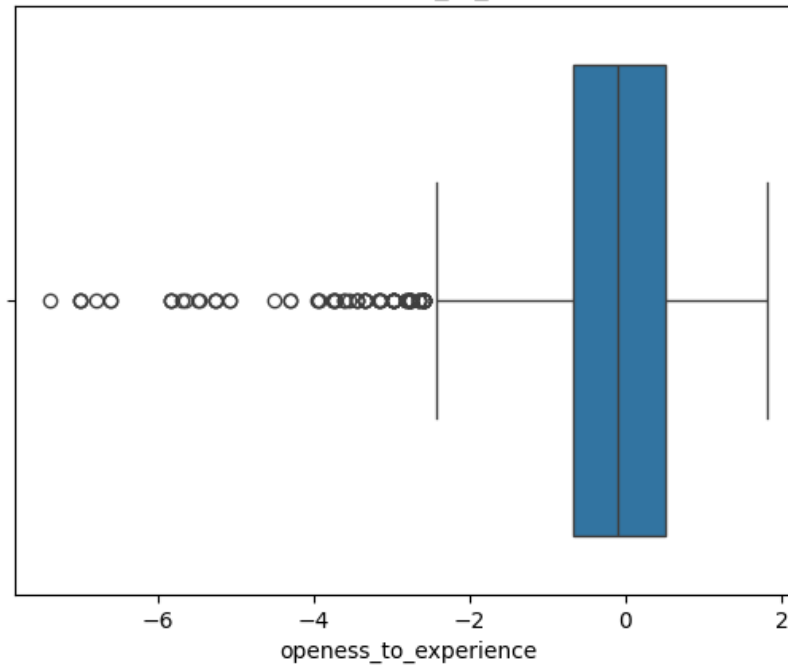
Boxplot of Salary



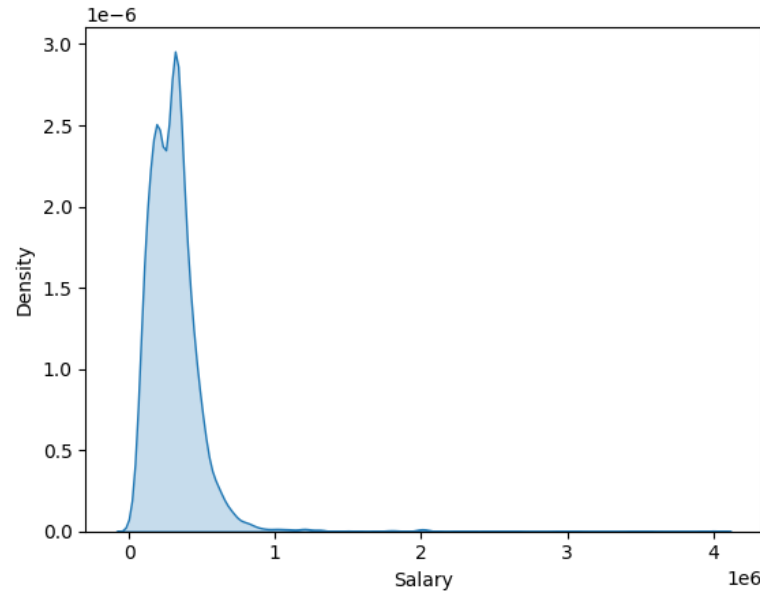
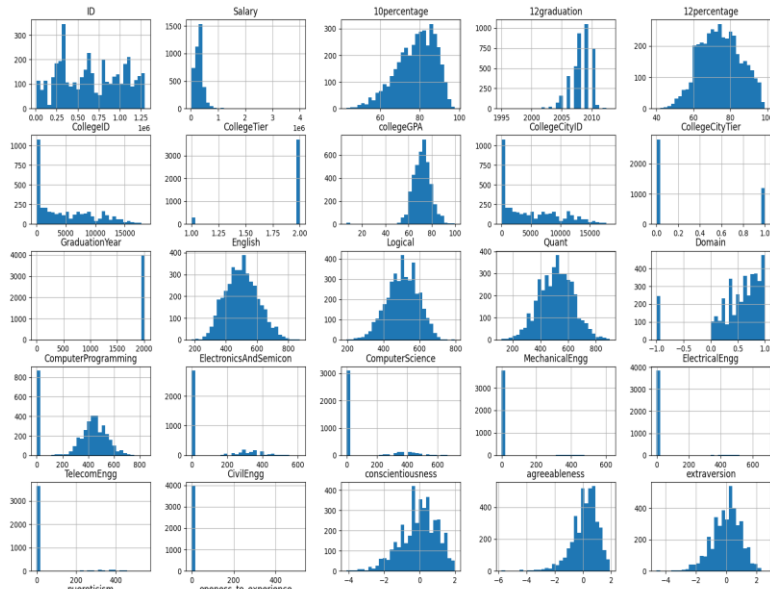
Boxplot of 10percentage



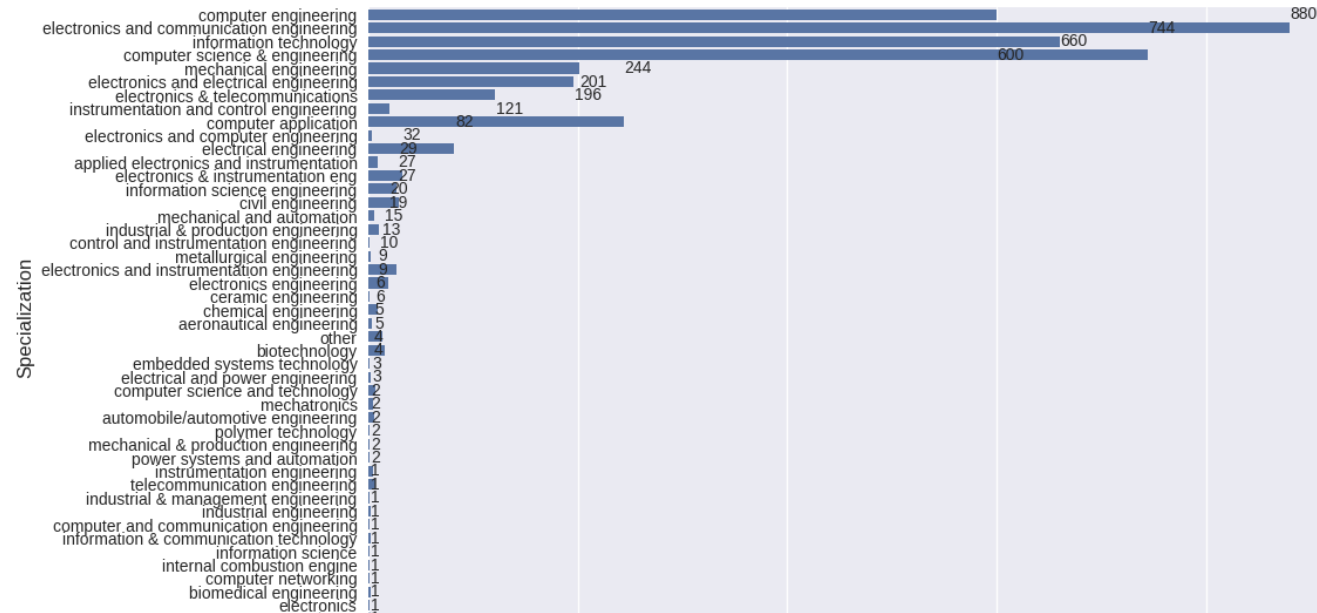
Boxplot of openness\_to\_experience



# Uni-Variate Analysis:



Distribution of Specialization Categories



**Salary, 10percentage, 12graduation, 12percentage:** These variables have a right-skewed distribution, indicating that a majority of individuals have lower values, with a smaller number having significantly higher values.

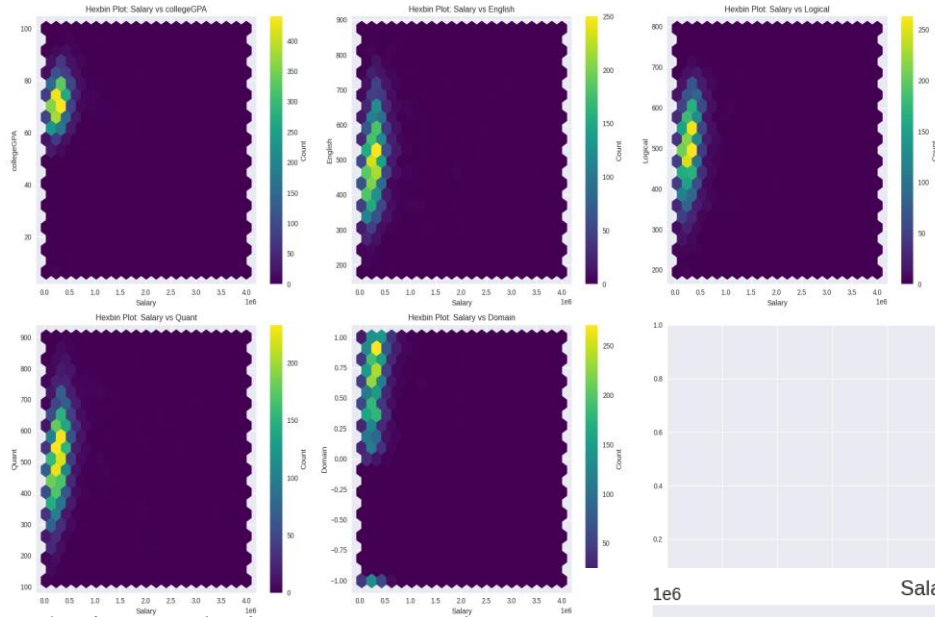
**English, Logical, Quant, Domain:** These variables have a similar distribution, with a peak around 200-300 and a general right skew, suggesting that most individuals score within this range.

**ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg:** These variables also show a similar distribution, with a peak around 200-400 and a general right skew, indicating that most individuals have scores within this range.

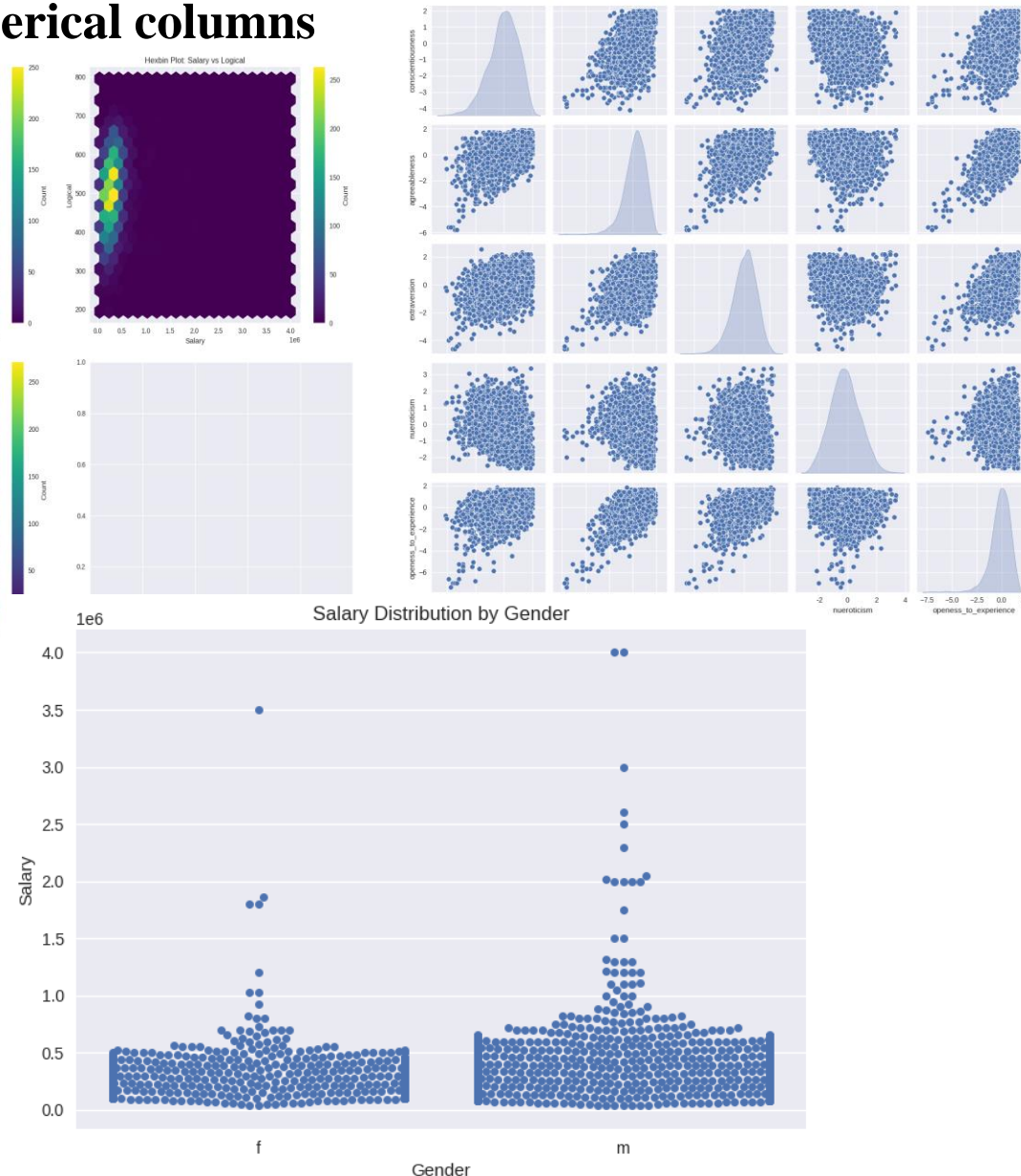
**CollegeTier, CollegeCityTier:** These variables have a right-skewed distribution, suggesting that most colleges or cities belong to lower tiers, with fewer belonging to higher tiers.

# Bi-Variate Analysis:

## Relation between 2 numerical columns



The image depicts a swarm plot illustrating the salary distribution by gender. Plot says individuals generally earn higher salaries than female individuals, with a significant number of male earners exceeding the 2 million salary threshold. While there is some overlap between the salary ranges of both genders, overall trend indicates a gender-based disparity in earnings.

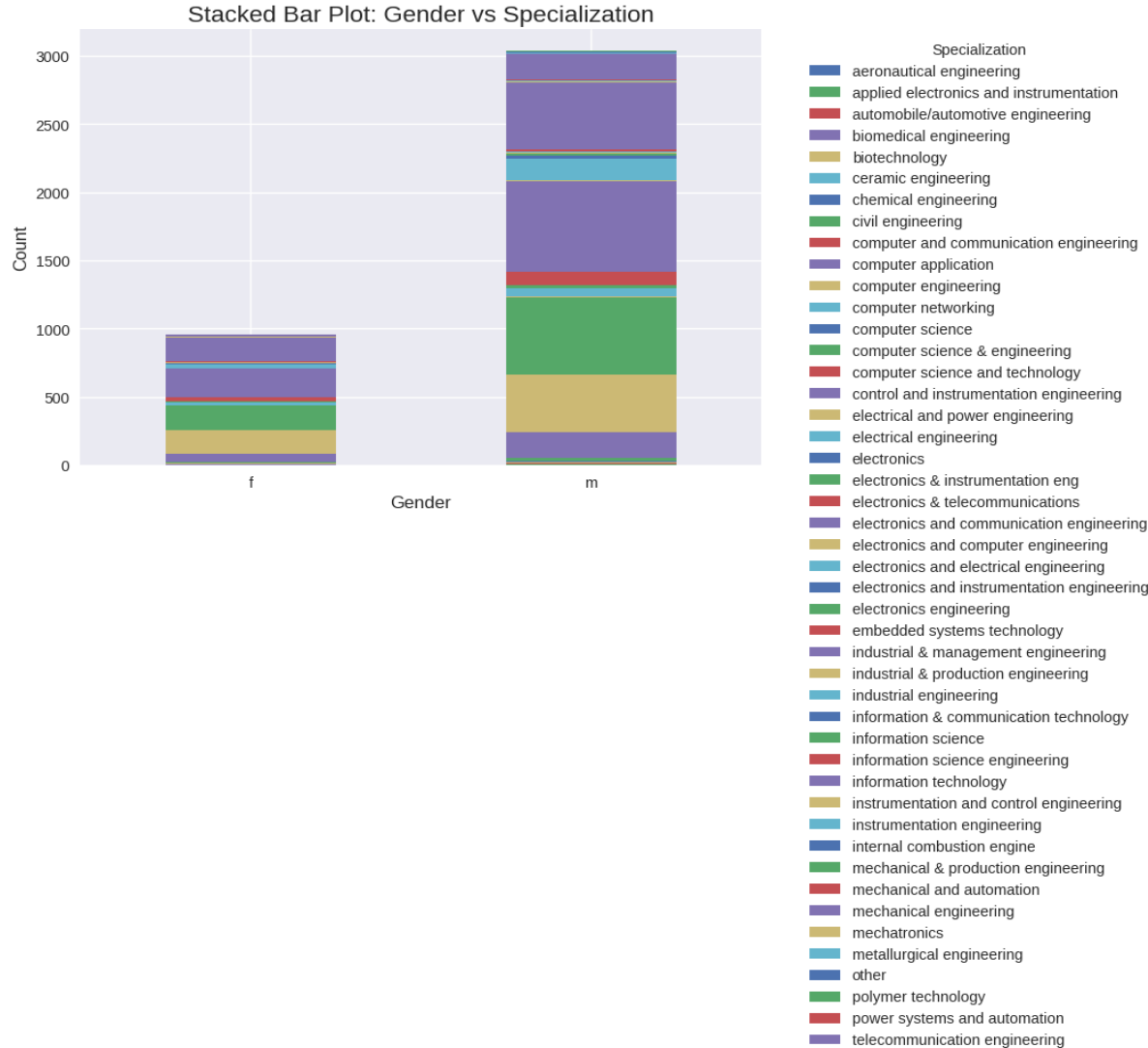


The analysis reveals strong positive correlations between conscientiousness, agreeableness, and extraversion, suggesting that individuals who possess these traits tend to exhibit a cluster of related behaviors. Conversely, neuroticism is negatively correlated with extraversion and agreeableness, indicating that individuals high in neuroticism are less likely to be outgoing or cooperative. Furthermore, the data reveals a multimodal distribution for the ID variable, suggesting the presence of distinct groups or categories within the data.



# Bi-variate Analysis:

## Categorical vs Categorical:



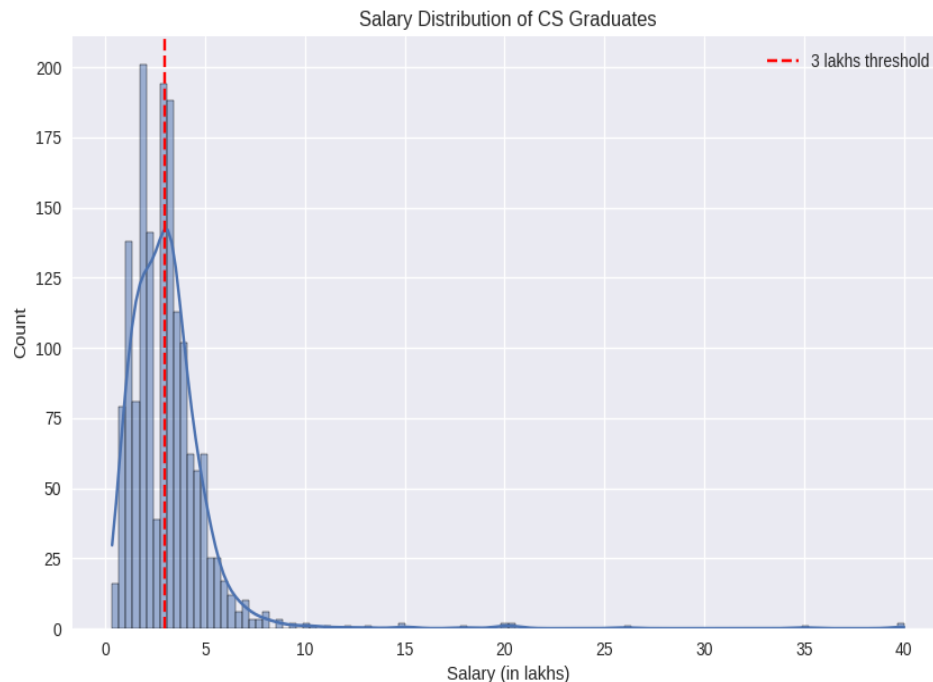
**Gender Disparity:** The plot reveals a significant gender disparity in certain specializations. For instance, there is a higher proportion of males in specializations like Computer Science & Engineering, Computer Science, and Information Technology, while there is a higher proportion of females in specializations like Biotechnology and Biomedical Engineering.

**Dominant Specializations:** The plot also highlights the dominant specializations in terms of the number of individuals pursuing them. Computer Science & Engineering, Computer Science, and Information Technology appear to be the most popular choices among both genders.

**Gender-Specific Preferences:** Certain specializations seem to have a more pronounced preference for one gender over the other. For example, Biotechnology and Biomedical Engineering appear to be more popular among females, while Computer Science & Engineering and Information Technology are more popular among males.

## Research Question:

Times of India article dated Jan 18, 2019 states that “After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate.



- Right-Skewed Distribution:** The salary distribution is significantly right-skewed, indicating that a majority of CS graduates earn salaries within the lower range, while a smaller proportion earns significantly higher salaries.

- Salary Concentration:** A significant portion of CS graduates earn salaries between 3 and 6 lakhs, with a peak around 4 lakhs.

- High-Salaried Graduates:** A smaller number of graduates earn salaries above 10 lakhs, with a few outliers earning significantly higher salaries.

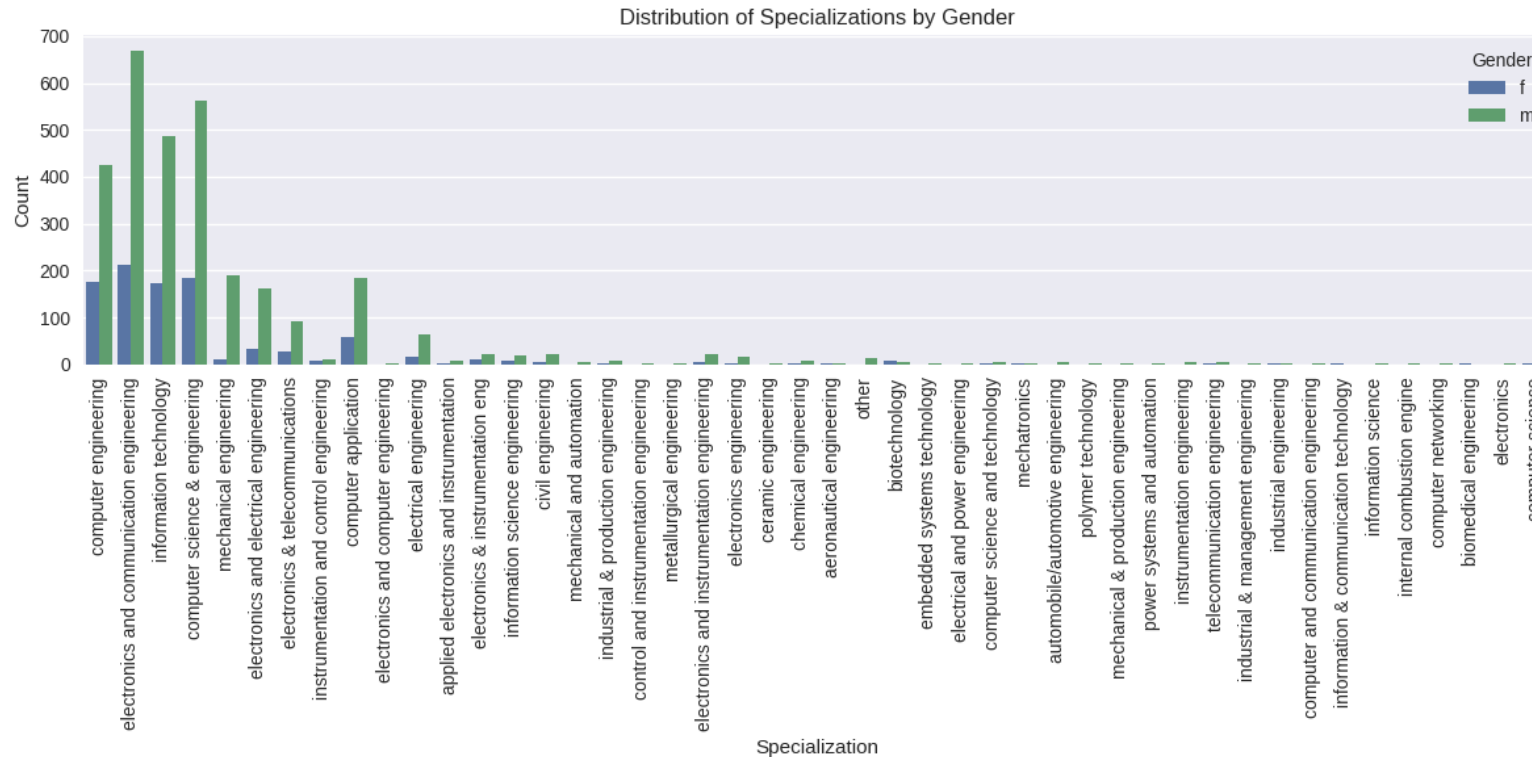
- 3 Lakhs Threshold:** The vertical red line at the 3 lakhs mark highlights a potential threshold for high-earning graduates. A significant number of graduates earn above this threshold, indicating that a considerable portion of CS graduates achieve salaries above the average.

**Overall, the histogram reveals a skewed salary distribution among CS graduates, with a majority earning within the lower to mid-range salaries and a smaller proportion achieving significantly higher salaries.**



# Research Question-B

Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)



Since the p-value is significantly smaller than 0.05, we can reject the null hypothesis (which states that there is no association) and conclude that there is a significant relationship between gender and specialization. In other words, the preference for a particular specialization does depend on the gender.

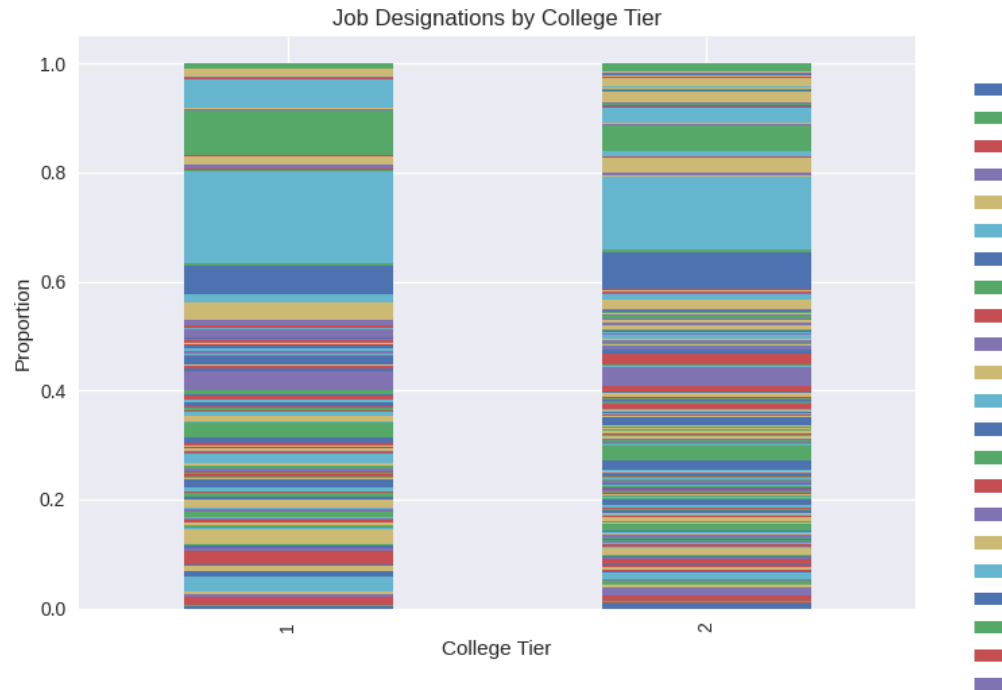
**Gender Disparity:** The plot reveals a significant gender disparity in certain specializations. For instance, there is a higher proportion of males in specializations like Computer Science & Engineering, Computer Science, and Information Technology, while there is a higher proportion of females in specializations like Biotechnology and Biomedical Engineering.

**Dominant Specializations:** The plot also highlights the dominant specializations in terms of the number of individuals pursuing them. Computer Science & Engineering, Computer Science, and Information Technology appear to be the most popular choices among both genders.

**Gender-Specific Preferences:** Certain specializations seem to have a more pronounced preference for one gender over the other. For example, Biotechnology and Biomedical Engineering appear to be more popular among females, while Computer Science & Engineering and Information Technology are more popular among males.

# Additional Research Questions:

## Q1. Impact of College Tier on Career Outcomes

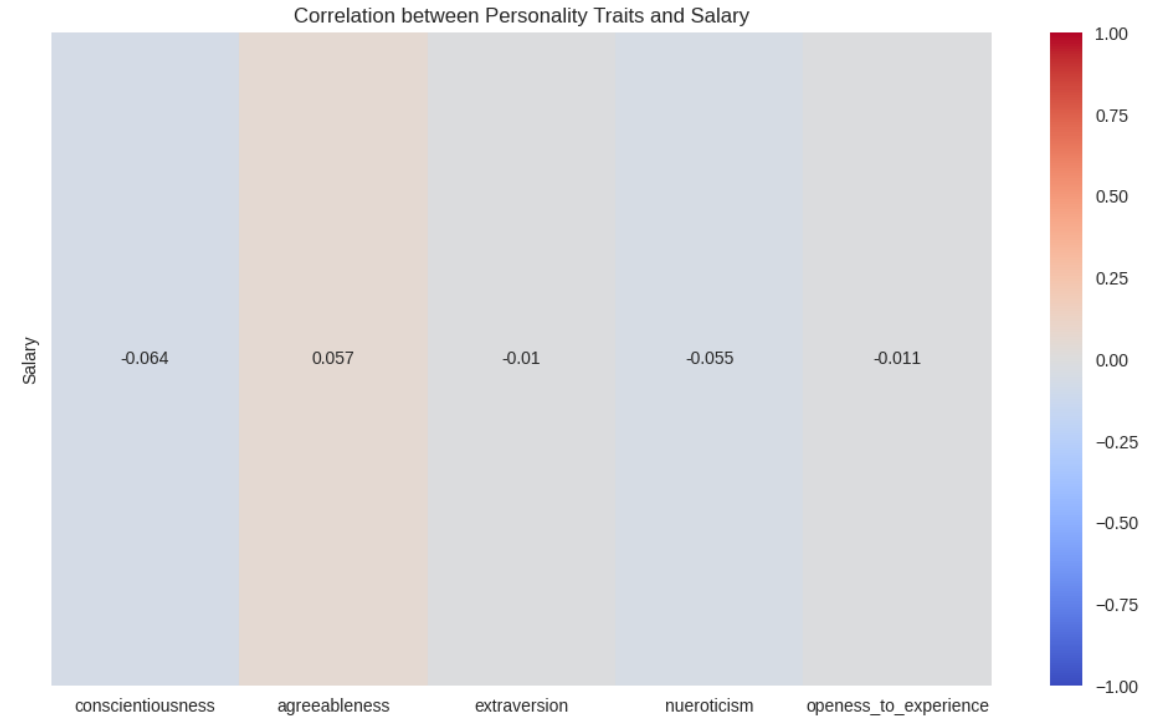


**Job Designation Distribution:** The plot reveals that the distribution of job designations varies across the two college tiers. While there are some common job designations in both tiers, the proportions of individuals in each designation differ.

**Tier-Specific Preferences:** Certain job designations appear to be more prevalent in one college tier compared to the other. For example, some job designations may be more common among graduates from tier 1 colleges, while others may be more common among graduates from tier 2 colleges.

**Job Designation Diversity:** The plot also highlights the diversity of job designations within each college tier. Both tiers have a wide range of job designations, indicating that graduates from both tiers pursue various career paths.

## Q2. Influence of Personality Traits on Salaries.



We can observe that there is less correlation between the salary earning and the personality traits

# Conculsion:

In this project, a detailed data analysis was conducted to explore the various factors that influence salary and job placements for fresh graduates. The analysis uncovered interesting insights, such as the significant impact of college reputation (CollegeTier) on starting salary, where graduates from Tier 1 colleges tended to secure higher-paying jobs compared to their counterparts from Tier 2 and Tier 3 institutions. Furthermore, the analysis revealed that certain specializations, such as Computer Science, were associated with higher salary ranges and faster career progression. This suggests that specialization choice plays a crucial role in determining the future career trajectory of graduates. Additionally, there was a clear trend showing that job location (JobCity) also influenced salary, with metropolitan cities offering higher salaries compared to smaller towns.

The project also tackled broader social issues, such as the existence of a gender pay gap. The findings revealed a disparity in average salary based on gender, with male graduates typically earning higher salaries across several designations. However, further analysis of the relationship between gender and specialization suggested that preferences for certain specializations were not strongly gender-dependent. Moreover, the analysis of extra-curricular scores, such as English and Logical reasoning, indicated a positive correlation with salary, implying that well-rounded skill sets are valued in the job market. Overall, this project highlights the importance of various academic and non-academic factors in shaping the career opportunities and salaries of fresh graduates.

THANK  
YOU

