INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMCAT Data

About me

Background:
I am Surya Atrish pursuing B.Tech in Computer Science and Engineering with a strong interest in Python and Data Analytics.

Motivation for Data Science:
Since the start of my undergraduate life, I was deeply interested in Python, its various libraries and frameworks and all the various kinds of problems it can solve. So, I started learning about it, started doing projects. At the same time, I also completed an internship on Data Analytics in Python, SQL Tableau and Power BI. Through research and introspection, I discovered that Data Science resonates deeply with me, offering a perfect fit for my aspirations and skills.

Work Experience:
I am Currently interning at Innomatics Research Labs, transitioning from a computer science and engineering background to data science.

LinkedIn: https://www.linkedin.com/in/surya-atrish/

GitHub: https://github.com/suryaatrish

INNOMATICS RESEARCH LABS

Agenda

Step - 1 - Introduction -> Give a detailed data description and objective

- Dataset Description
- Objective

Step - 2 - Import the data and display the head, shape and description of the data.

- Describing the data
- Data Cleaning
- Conversion of Data Types
- Collapsing Categories
- Feature Engineering

Step - 3 - Univariate Analysis

- Continuous Features
- Categorical Features
- Removing Outliers

Step - 4 - Bivariate Analysis

- Barplots
- Scatter Plots
- Crosstabs
- Pivot Tables
- Step - 5 - Research Questions
- 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.” Test this claim with the data given to you.
- Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)

Step - 6 - Conclusion

Step - 7 - (Bonus) Come up with some interesting conclusions or research questions (such as step-5).

INNOMATICS RESEARCH LABS

Objective:

The primary objective of the project is to conduct Exploratory Data Analysis (EDA) on the provided dataset, with a particular emphasis on understanding the link between various variables and the target variable, Salary. The EDA process involves:

1. Identifying and handling missing values, dealing with outliers, and ensuring data consistency and integrity.
2. Calculating summary statistics (mean, median, standard deviation, etc.) for continuous variables and frequency distributions for categorical variables.
3. Creating visualizations such as histograms, box plots, scatter plots, and heatmaps to explore relationships between variables, identify patterns, and detect outliers.
4. Creating new features or transforming existing ones to better represent the underlying data and improve model performance.
5. Examining the correlations between different variables to understand their relationships and identify potential predictors of employment outcomes.
6. Investigating the relationships between the independent factors and the target variable (salary).

Description:

The dataset contains information on 3,998 individuals, spanning across 39 columns. Each row represents a unique individual, while each column provides specific details about their employment and educational background.

Key columns include:

- **ID**: A unique identifier for each candidate.
- **Salary**: Annual CTC offered to the candidate (in INR).
- **DOJ**: Date of joining the company.
- **DOL**: Date of leaving the company.
- **Designation**: Designation offered in the job.
- **JobCity**: Location of the job (city).
- **Gender**: Candidate's gender.
- **DOB**: Date of birth of the candidate.
- **10percentage**: Overall marks obtained in grade 10 examinations.
- **10board**: The school board whose curriculum the candidate followed in grade 10.
- **12graduation**: Year of graduation from senior year high school.
- **12percentage**: Overall marks obtained in grade 12 examinations.

Description:

- **12board**: The school board whose curriculum the candidate followed in grade 12.
- **CollegeID**: Unique ID identifying the college which the candidate attended.
- **CollegeTier**: Tier of college.
- **Degree**: Degree obtained/pursued by the candidate.
- **Specialization**: Specialization pursued by the candidate.
- **CollegeGPA**: Aggregate GPA at graduation.
- **CollegeCityID**: A unique ID to identify the city in which the college is located.
- **CollegeCityTier**: The tier of the city in which the college is located.
- **CollegeState**: Name of States.
- **GraduationYear**: Year of graduation (Bachelor's degree).
- **English, Logical, Quant**: Scores in AMCAT English, Logical, and Quantitative sections.
- **Domain**: Standardized scores in AMCAT's domain module.
- **ComputerProgramming, ElectronicsAndSemicon, ComputerScience, MechanicalEngg, ElectricalEngg, TelecomEngg, CivilEngg**: Scores in respective sections of AMCAT.
- **Conscientiousness, Agreeableness, Extraversion, Neuroticism, Openness_to_experience**: Standardized scores in different sections of AMCAT's personality test.

Data Cleaning:

After conducting preliminary assessments on the provided data, it has come to my attention that there are some irregularities present within the dataset.

Column Name	Observation
DOL	It have the Value 'present', that means the employee is working on the company now. Replace present with Today's Date
JobCity	Contains '-1' in the column Considered to be a Missing Value and City Names are not proper
10board	Contains '0' in the column considered as Missing Value and having high Cardinality, convert to State. CBSE, and ICSE Board
12board	Contains '0' in the column considered as Missing Value and having high Cardinality, convert to State. CBSE, and ICSE Board
CollegeState	Contains 'Union Territory' in the column considered as Missing Value, we don't know College state belongs to which Union Territory
Domain	Contains '-1' in the column Considered to be a Missing Value
Designation	Contains 'get' in the column Considered to be a Missing Value

Describing the data:

- Shape: (3998. 39)
- Duplicated rows: 0
- Columns: 'Unnamed: 0', 'ID', 'Salary', 'DOJ', 'DOL', 'Designation', 'JobCity', 'Gender', 'DOB', '10percentage', '10board', '12graduation', '12percentage', '12board', 'CollegeID', 'CollegeTier', 'Degree', 'Specialization', 'collegeGPA', 'CollegeCityID', 'CollegeCityTier', 'CollegeState', 'GraduationYear', 'English', 'Logical', 'Quant', 'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion', 'nueroticism', 'openess_to_experience'

Feature Extraction:

1. Since the dataset was release in 2015, we add a age column by subtracting DOB year from 2015. This will add the age as of 2015.
2. Adding a tenure column by subtracting the `DOL` from `DOJ`
3. Dropping the rows where the graduationyear is greater than or equal to date of joining
4. Function to calculate CDF

Univariate Analysis:

1. Continuous Features:

1.1 Tenure

Sr.No.	Conclusion	Inferences
1.	Summary Plot	- The range for experience is 4 years.
2.	Histogram	- The data is positively skewed i.e there exists larger number of respondents with low tenure, 50% data points are below 1.5 years, Average tenure is 1.5 years, The mean, median, and mode lie very close to each other and skewness (0.6) is close to that of a normal (0).
3.	Box Plot	- There are few values with large tenure i.e outliers
4.	CDF	- The data is not normally distributed, We can say that tenure is not normally distributed.

1.2 Salary

Conclusion	Inferences
1. Summary Plot	There is substantial variation in salary across the dataset.
2. Histogram	The data exhibits significant positive skewness, with a skewness value around 6 (approximately), indicating a departure from a normal distribution. The measures of central tendency (mean, median, and mode) are approximately equal.
3. Box Plot	There is a notable concentration of data points with high salaries, as depicted by the box plot.
4. CDF	The cumulative distribution function (CDF) reveals a high degree of skewness in the data, with considerable deviation from a normal distribution pattern.

Univariate Analysis:

1. Continuous Features:

1.3 10th Percentage

Conclusions	Inferences
1. Summary Plot	Around 50% of students achieved scores of approximately 80% or less.
2. Histogram	The histogram depicts a scarcity of students with low percentages, with the majority falling within the 75% to 90% range. The peak frequency occurs at 78%, and the average score hovers around 77%.
3. Box Plot	The presence of a few extreme outliers is evident from the box plot.
4. CDF	The data exhibits some skewness and does not conform to a normal distribution pattern.

1.4 12th Percentage

Conclusions	Inferences
1. Summary Plot	Roughly half of the students achieved scores of approximately 78% or lower.
2. Histogram	The histogram illustrates a scarcity of students with low percentages, with the majority scoring between 69% and 84%. The peak frequency occurs at 70%, and the average score is around 74%.
3. Box Plot	The box plot indicates only one data point with an extremely low score.
4. CDF	The data does not follow a normal distribution pattern.

Univariate Analysis:

1. Continuous Features:

1.5 CollegeGPA

Conclusions	Inferences
1. Summary Plot	75% of students had a GPA of approximately 80% or lower.
2. Histogram	The majority of students had GPAs ranging between 63% and 78%. The highest frequency of students scored 70%, and the average GPA was 74%.
3. Box Plot	The box plot reveals the presence of both low and high extreme values within the dataset.
4. CDF	The data is deemed to be sufficiently normally distributed.

1.6 English

Conclusions	Paraphrased Version
Summary Plot	Half of the students scored below 500 in their English exams.
Histogram	The bulk of the scores fell within the range of 389 to 545. The peak occurred at 475, with an average score of 502.
Box Plot	Both lower and higher extreme values are evident from the distribution representation.
CDF	The data follows a reasonably normal distribution pattern.

Univariate Analysis:

1. Continuous Features:

1.7 Logical

Conclusions	Inferences
Summary Plot	Half of the students scored below 500 in the logical exams.
Histogram	Most scores fell within the range of 454 to 584, peaking at 495, with an average of 502.
Box Plot	Presence of lower extreme values, with only one high extreme value being notable.
CDF	Data closely approximates a normal distribution pattern.

1.8 Quant

Conclusions	Inferences
Summary Plot	75% of students' logical score was less than 600.
Histogram	Majority of the scores were in between 425-608. The maximum number of students scored 605 with an average of 513.
Box Plot	The box plot shows the presence of both low and high extreme values.
CDF	The data is sufficiently close to normally distributed.

Univariate Analysis:

1. Continuous Features:

1.9 Computer Programming

Conclusions	Inferences Version
Summary Plot	50% of students' scores were below 500.
Histogram	The majority of scores ranged between 416 and 459. The peak occurred at 455, with an average score of 452.
Box Plot	The box plot illustrates the presence of numerous low extreme values as well as high extreme values.
CDF	The data does not follow a normal distribution pattern.

1.10 Electronics & Semiconductors

Conclusions	Inferences
Summary Plot	About 75% of students scored less than 250.
Histogram	Most scores fell between 0 and 79. The highest number of students scored 0, with an average score of 96.
Box Plot	The lowest score is equal to the median of the dataset.
CDF	The data does not conform to a normal distribution pattern.

Univariate Analysis:

1. Continuous Features:

1.11 Age

Conclusions	Inferences
Summary Plot	Approximately 75% of students are under 26 years old.
Histogram	The majority of students' ages ranged between 22 and 25. The mean, median, and mode ages are approximately 25.
Box Plot	The box plot indicates the presence of 4 students with very high ages and one with a very low age compared to other data points.
CDF	The age data does not follow a normal distribution pattern.

2. Categorical Features:

2.1 Designation

Software engineer is the most common desgination of all, followed by system engineer and software developer.

2.2 JobCity

The most favourable city for job placements is Bangalore, followed by Noida, Hyderabad and Pune. Mumbai and Kolkata being least favourable.

2.3 Gender

The dataset is not balanced in terms of gender as the population of Male is really larger as compared to the female one.

Univariate Analysis:

2. Categorical Features:

2.4 10board & 12board

CBSE is the most common school board for both 12th and 10th.

2.5 CollegeTier

Almost all the college belongs to Tier 1 only with a percentage of 92.5

2.6 Degree

Most of the students have done their graduation in B.Tech and there are very less students from M.Sc(Tech)

2.7 CollegeCityTier

Majority of the colleges are from Tier 0 city.

2.8 GraduationYear

Maximum number of students were graduated in 2013, followed by the year 2014 and 2012.

3. Removing Outliers:

Number of observation with outliers: 3864

Number of observations without outliers: 2490

Bivariate Analysis:

1. Bar Plots:

1.1 Average Salary for each Designation

Bar plot shows the maximum salary for each Designation. Senior Software Engineer has the highest salary but they also has the maximum standard deviation in their salary. There are only two designations namely, software developer and technical support engineer who has salary lower than average salary.

1.2 Average Salary for each Gender

The average salary for both male and female is approximately equal and it implies that there was no gender bias in terms of salary. It is also plausible to say that Female's get salary below the overall average salary.

2. Scatter Plots:

2.1 Salary & 10^{th} score

There does not exist any correlation between Salary and 10th scores.

2.2 Salary & 12^{th} score

There does not exist any correlation between Salary and 10th scores.

2.3 Salary & CollegeGPAscore

There does not exist any correlation between Salary and CollegeGPA scores.

Bivariate Analysis:

2.3 Salary & Age

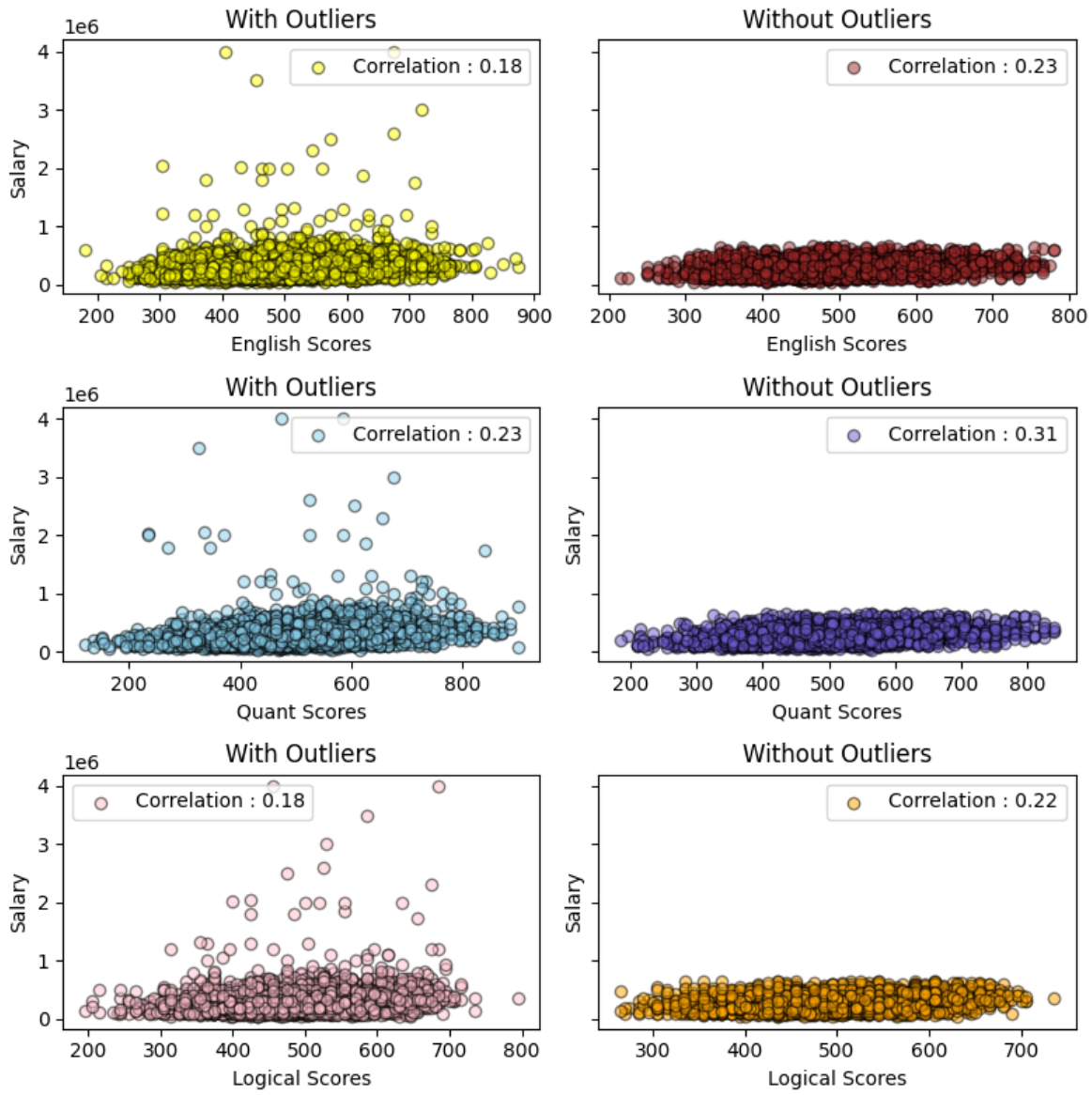
After removing the outliers, it is evident that the salary and age are not related to each other.

2.4 Salary & Tenure

After removing the outliers, it is evident that salary gets about 50% of increment as tenure increase as there is a positive correlation of 0.60.

2.5 Salary with English, Quants & Logical

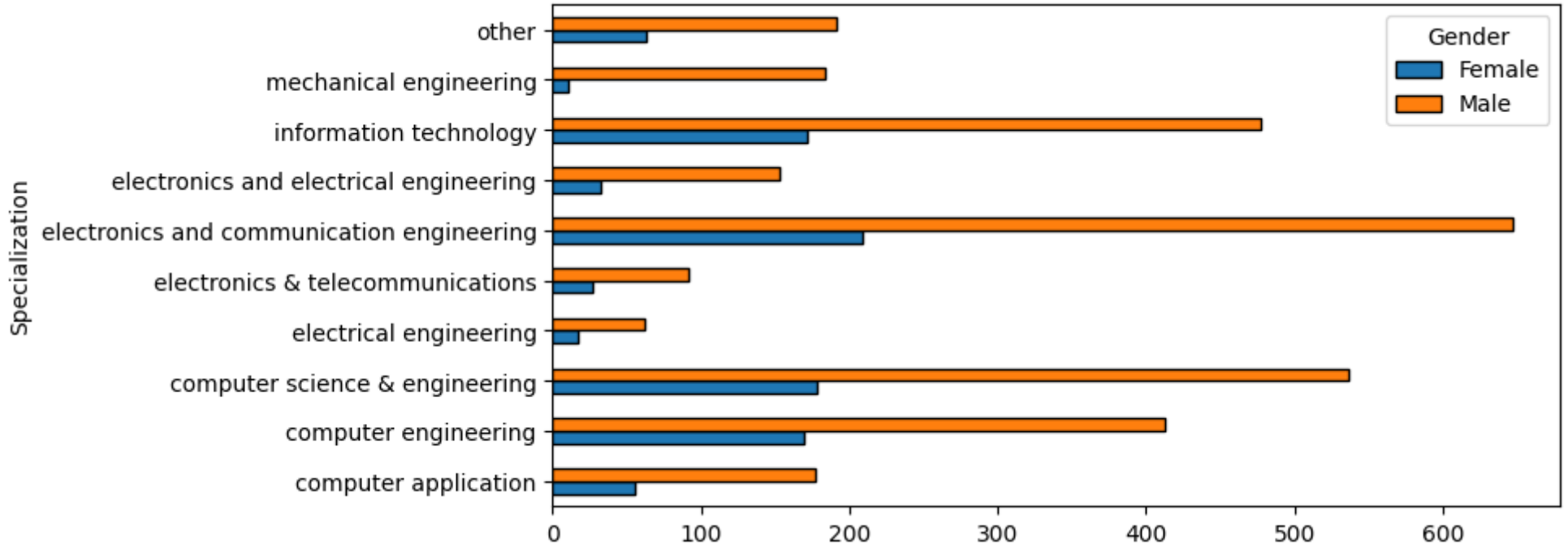
On the right.



INNOMATICS RESEARCH LABS

Bivariate Analysis:

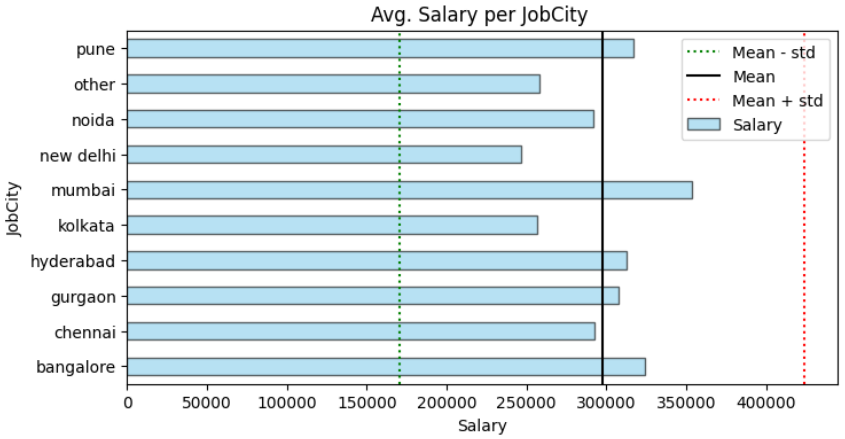
3. Crosstab between gender and specialization:

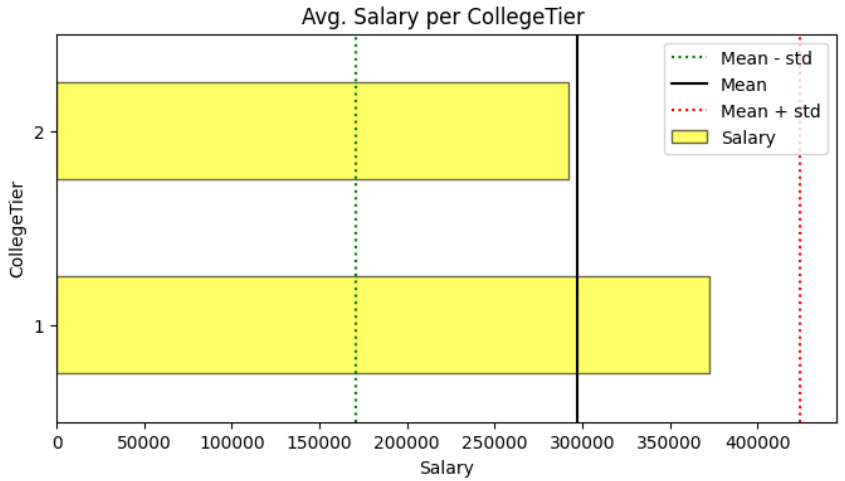


There are almost males 2 times as of females in every specialization. Also, there are very less number of females who opted for mechanical and electronics.

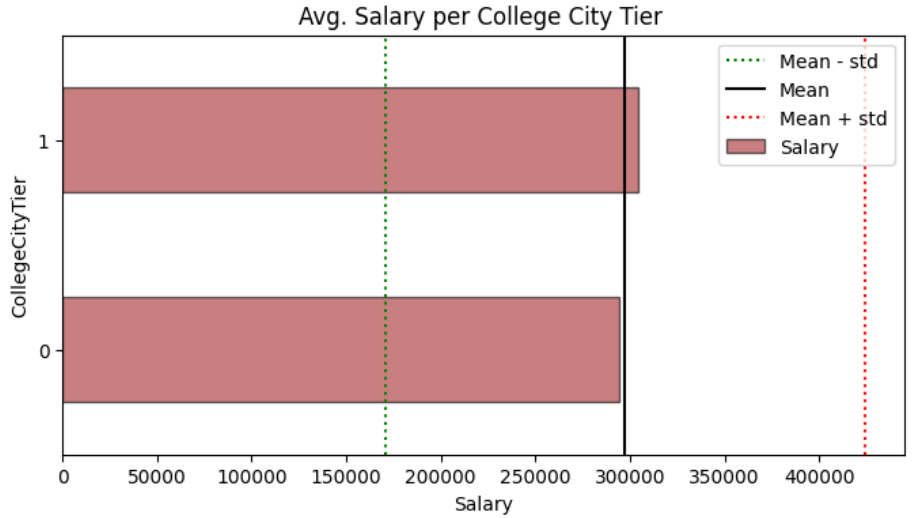
Bivariate Analysis:

4. Pivot Tables:





College within Tier 1 offers high salary as compared to the colleges in TIer 2. Colleges in Tier 2 offers below overall average salary.



Cities under Tier 1 and 2 offers almost same salaries to students.

INNOMATICS RESEARCH LABS

Research Questions:

1. Data Overview:

Observations

Designation	t_critical	p_value	Result
Programmer Analyst	12.77	2.20314e-18	There is sufficent evidence to reject the Null Hypothesis
Software Engineer	10.21	5.81591e-21	There is sufficent evidence to reject the Null Hypothesis
Hardware Engineer	NaN	NaN	There is not enough evidence to reject the Null Hypothesis
Associate Engineer	0.61	3.01696e-01	There is not enough evidence to reject the Null Hypothesis

2. Data Preparation:

Observations

Test	Value
chi2_critical	16.918977604620448
chi2_statistic	48.62141720904882
chi2_p_value	1.9542895953348e-07

- As the result of the second research question we see that there is a relationship between Gender and specialization.
- We test this claim through Chi-Square test and find the result that both the categorical variables are dependent on each other.
- Some specialization or working field does not allow some candidates to work in that field due to some risks.

INNOMATICS RESEARCH LABS

Conclusion:

1. **Data Overview:**

- Within the dataset, the focus lies on exploring the employment outcomes of engineering graduates, with particular attention to the target variable "Salary". Additionally, it encompasses standardized scores representing cognitive, technical, and personality skills.

2. **Data Preparation:**

- Initial scrutiny reveals a dataset comprising 3998 rows and 39 columns.
- To refine the dataset, redundant rows and columns are eliminated to ensure data efficiency.
- The mitigation of missing values (NaN) is prioritized to uphold data integrity.
- Post-cleaning, the next step involves visualization.

3. **Data Visualization:**

- **Univariate Analysis:**
  - Univariate analysis encompasses various plots such as Cumulative Distribution Functions (CDF), Histograms, Box Plots, and Summary Plots.
  - These visualizations effectively illustrate probability and frequency distributions.

- **Bivariate Analysis:**
  - Bivariate analysis incorporates Scatterplots, Barplots, Crosstabs, Pivot tables, and pie charts.
  - It facilitates the comparison of percentages across variables and the identification of outliers, prominently showcased through Boxplots.
  - Countplots are particularly useful in pinpointing outliers within categorical variables like Job City, offering insights into cities with significant employee counts.

THANK YOU

INNOMATICS RESEARCH LABS