INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Analysis of AMCAT Data

About me

- **Name :** KAMMARI SADGURU SAI
- **Specialization :** B.Tech - Computer Science & Engineering (Data Science)
- **GitHub :** kammarisadgurusai
- **Linkedin :** Kammari Sadguru Sai

Objective of the Project

The objective of this project is to perform Exploratory Data Analysis (EDA) on the AMEO dataset to gain insights into the employment outcomes of engineering graduates. Through data visualization and analysis, we aim to understand salary distributions, job titles, job locations, skill assessments, and other relevant factors impacting the employment landscape

Data Description

The dataset is derived from the Aspiring Mind Employment Outcome 2015 (AMEO) study, focusing on employment outcomes among engineering graduates. It includes information on various aspects such as salary, job titles, job locations, standardized scores in cognitive, technical, and personality skills, and demographic features. The dataset contains approximately 4000 data points with around 40 independent variables, comprising both continuous and categorical data

Data Cleaning Steps

Data Loading:

- Imported dataset using `pd.read_csv()`.

Initial Inspection:

- Used `.head()`, `.shape()`, `.describe()`, and `.info()` for structure and data type checks.

Handling Missing/Infinite Values:

- Replaced `inf` values with `NaN` using `replace()` to avoid issues.

Data Type Identification:

- Differentiated numerical and categorical columns using `select_dtypes()`.

Outlier Detection:

- Applied IQR method to find outliers:
    - Calculated Q1, Q3, and IQR.
    - Identified outliers beyond 1.5 * IQR bounds.

Column Renaming:

- Ensured consistent column names for analysis.

Visualization:

- Created histograms and boxplots for data distribution and outlier insights.

Data Manupulation Steps

Data Import & Inspection:

- Loaded the dataset using `pd.read_csv()`.
- Inspected data with `.head()`, `.shape()`, `.describe()`, and `.info()`.

Handling Missing/Infinite Values:

- Replaced `inf` values with `NaN` using `replace()`.

Column Type Identification:

- Used `select_dtypes()` to identify numerical and categorical columns.

Outlier Detection & Handling:

- Applied Interquartile Range (IQR) method for outlier identification.

Data Visualization:

- Utilized histograms, boxplots, and scatterplots for distribution and outlier analysis.
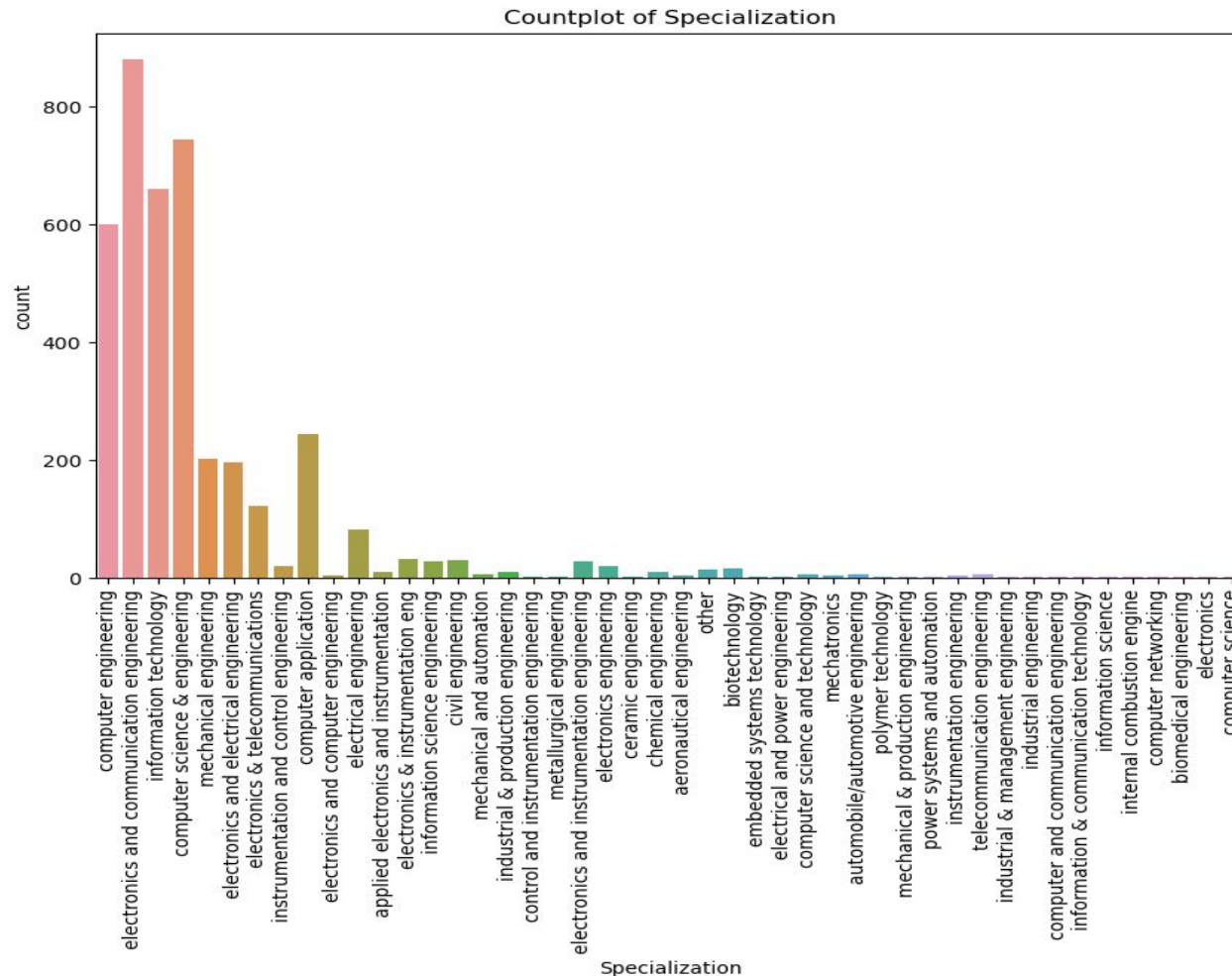
Categorical Data Exploration:

- Created count plots for categorical columns to visualize frequency distributions.

Research-Based Analysis:

- Performed average salary calculations for specific job roles.
- Created contingency tables and applied Chi-square tests to check relationships.

Univariate Analysis Steps

Countplot of Specialization

➔ **Popular Specializations**: "Electronics and Communication Engineering" and "Computer Science and Engineering" have the highest numbers, making them the most popular fields.

➔ **Wide Variety of Fields**: There are many different specializations, though only a few have high counts.

➔ **Engineering Focus**: Most students are in engineering-related fields, showing a strong preference for technical subjects.

➔ **Less Common Fields**: Many specializations have lower counts, meaning fewer students are in these areas.

➔ **Useful for Targeting**: This information can help identify which specializations are most common, which is useful for companies looking for specific skills.
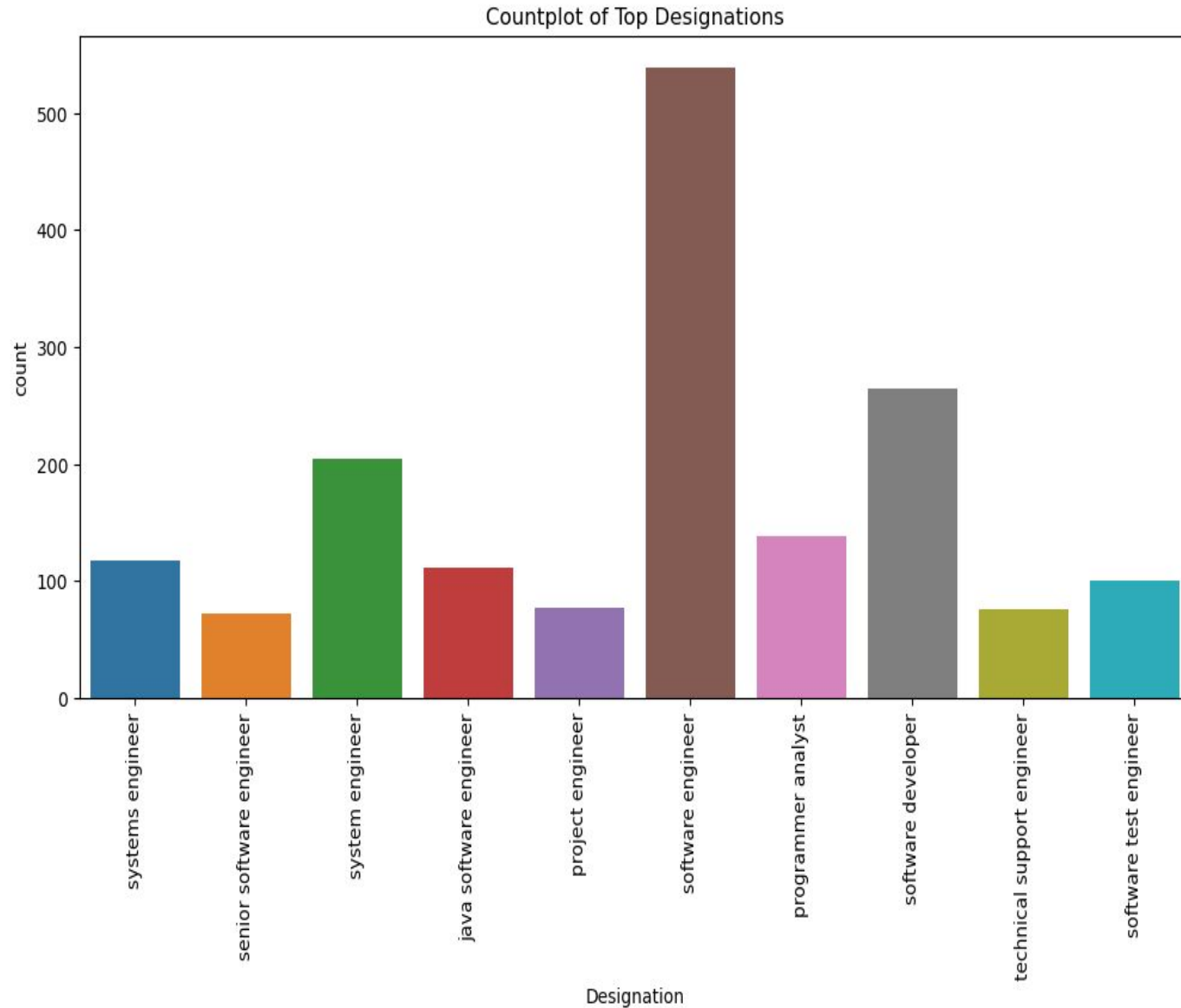
Histogram of Salary
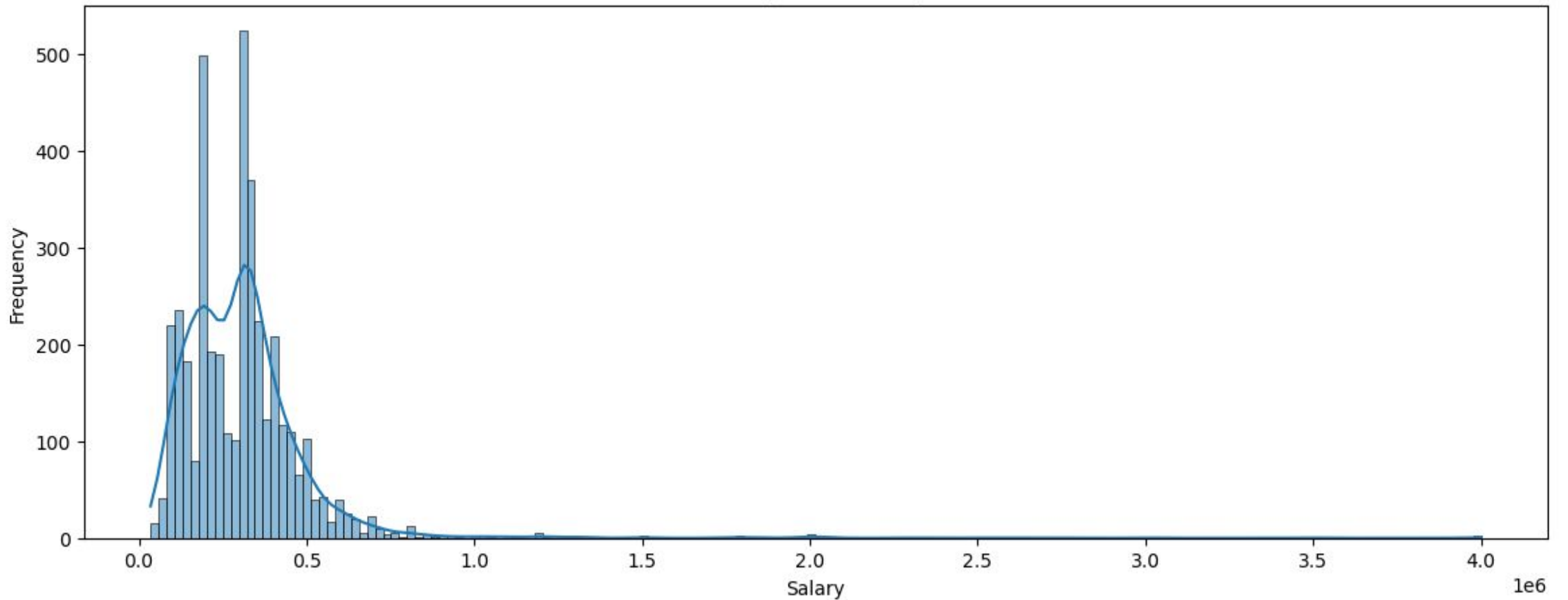
Most salaries are concentrated at the lower end, *between 0 and 0.5 million, showing a high frequency in this range.*

Very few people have high salaries *(over 1 million), resulting in a right-skewed distribution.*

*This chart highlights **salary spread, with most earnings clustered low** and a small tail for higher salaries.*

Countplot of Top Designations

**Software Engineer is the most common designation**, with the highest count among all roles.

System Engineer and Software Developer are also popular, with relatively high counts.

Roles like **Senior Software Engineer, Programmer Analyst, and Technical Support Engineer** have lower counts compared to the top designations.

Java Software Engineer and Software Test Engineer are present, but with moderate frequency.

This plot shows **which job titles are most common**, helpful for understanding workforce distribution across roles.

INNOMATICS
RESEARCH LABS

Bivariate Analysis

Top Designations by Average Salary
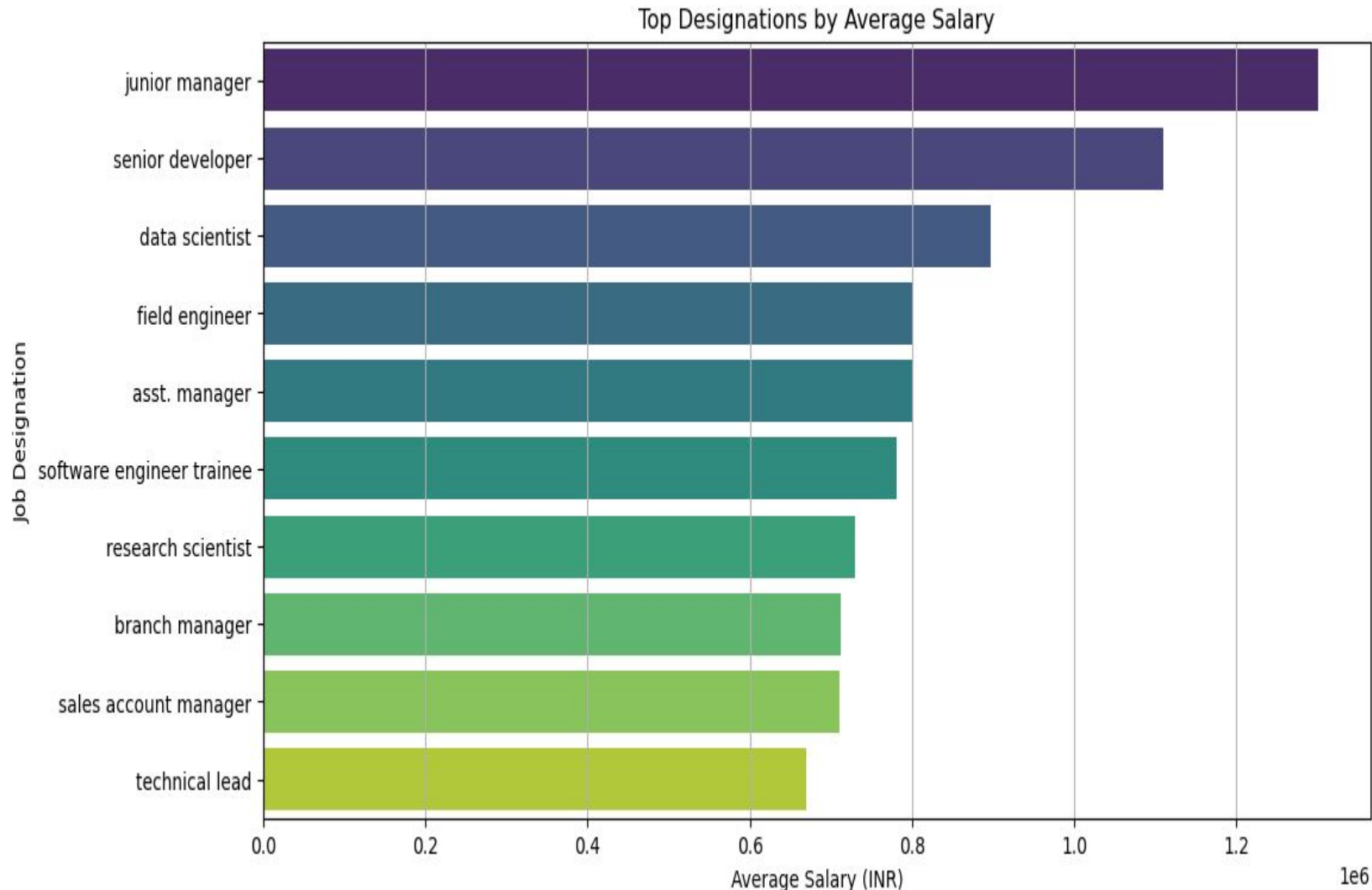
Junior Managers lead with about 1.2 million INR per year, followed closely by Senior Developers at 1.0 million INR. Data Scientists come next at around 0.9 million INR, with Field Engineers and Assistant Managers earning roughly 0.8 million INR each. The other roles, like Software Engineer Trainees, Research Scientists, Branch Managers, Sales Account Managers, and Technical Leads, earn between 0.5 and 0.7 million INR

INNOMATICS
RESEARCH LABS

Conclusion

Key Findings from Exploratory Data Analysis (EDA)

Univariate Analysis:

- **Salary Distribution:** Right skewed; most earn lower salaries, few earn significantly higher.
- **Outliers Detected:** Some individuals earn much higher than average.
- **Common Roles:** Software Engineer and Senior Software Engineer are most prevalent.

Bivariate Analysis:

- **Salary Variations by Role:** Significant differences in salaries across job titles.

Research Questions

Claim from Times of India:

- **Average Salary for Fresh Graduates (Programming Analyst, Software Engineer, etc.):**
  - If average salary is between 2.5 and 3 lakhs, claim is supported; if outside, it's not.

Gender and Specialization Relationship:

- **Chi-square Test Results:**
  - p-value < 0.05: Significant relationship; gender influences specialization.
  - p-value > 0.05: No significant relationship.

Key Takeaways

- **Salary Influencers:** College GPA has a minor effect; designation, location, and experience are more significant.
- **Job Role Variability:** Senior roles offer higher salaries than entry-level positions.
- **Exploring Gender Preferences:** Understanding gender and specialization can provide valuable insights.

INNOMATICS
RESEARCH LABS