INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis (EDA)
on
AMCAT data-set

B. Vidhyalakshmi
23rd February 2024

About me

- **Background (B.Tech in Artificial Intelligence and Data Science):**
    - "Currently, I am pursuing my Bachelor of Technology (B.Tech) degree in Artificial Intelligence and Data Science from SRM Easwari Engineering College. Through my coursework, I've been exposed to various foundational concepts in artificial intelligence, machine learning, and data science. I've engaged in projects and assignments that have allowed me to apply these concepts to real-world datasets, gaining hands-on experience in data analysis, statistical modeling, and programming."
- **Why You Want to Learn Data Science:**
    - "I am deeply passionate about the potential of data science and artificial intelligence to drive innovation and solve complex problems across industries. The ability to extract actionable insights from data and leverage them to make informed decisions fascinates me. I am particularly drawn to the interdisciplinary nature of data science, which integrates elements of computer science, mathematics, and domain expertise. I believe that mastering data science skills will not only open up exciting career opportunities but also empower me to contribute meaningfully to addressing societal challenges."
- **Work Experience:**
    - "While I do not have prior work experience in the field, I am eager to gain hands-on experience and apply my theoretical knowledge to practical scenarios. I am actively seeking internships, research opportunities, and projects where I can further develop my skills and contribute to real-world data science initiatives. Additionally, I am engaged in extracurricular activities such as online courses, hackathons, and data science competitions to continuously enhance my skills and stay updated with the latest developments in the field."

Github: Vidhyalakshmib1305 (github.com)

LinkedIn : www.linkedin.com/in/b-vidhyalakshmi

Objective of the Project:

- The primary objective of this exploratory data analysis (EDA) project is to gain insights into the relationships between different variables and to uncover patterns or trends within the dataset. By conducting a thorough analysis, we aim to extract valuable information that can aid in understanding the factors influencing salary, employment outcomes, and overall employability of individuals who have taken the AMCAT assessment.
- Through visualizations, statistical summaries, and correlation analyses, we seek to answer pertinent questions such as
- What is the distribution of salaries among the candidates?
- Are there any significant differences in salaries based on gender, educational qualifications, or specialization?
- How do personality traits correlate with employability metrics such as domain-specific scores and overall performance?
- Are there any discernible patterns in the data regarding job cities, college tiers, or graduation years that could impact employability?

Summary of the Data

To summarize the working of the data analysis we conducted:

Introduction:

We began by importing the dataset and outlining the objective of our analysis, which was to explore factors influencing the earnings of fresh graduates in specific job roles.

Data Exploration:

We explored the dataset by examining its structure, dimensions, and summary statistics. This step helped us understand the nature of the data and identify potential variables of interest.

Univariate Analysis:

We conducted univariate analysis to understand the distribution of individual variables. This involved visualizing probability density functions, histograms, boxplots, and countplots to identify outliers and understand the frequency distribution of both numerical and categorical variables.

Bivariate Analysis:

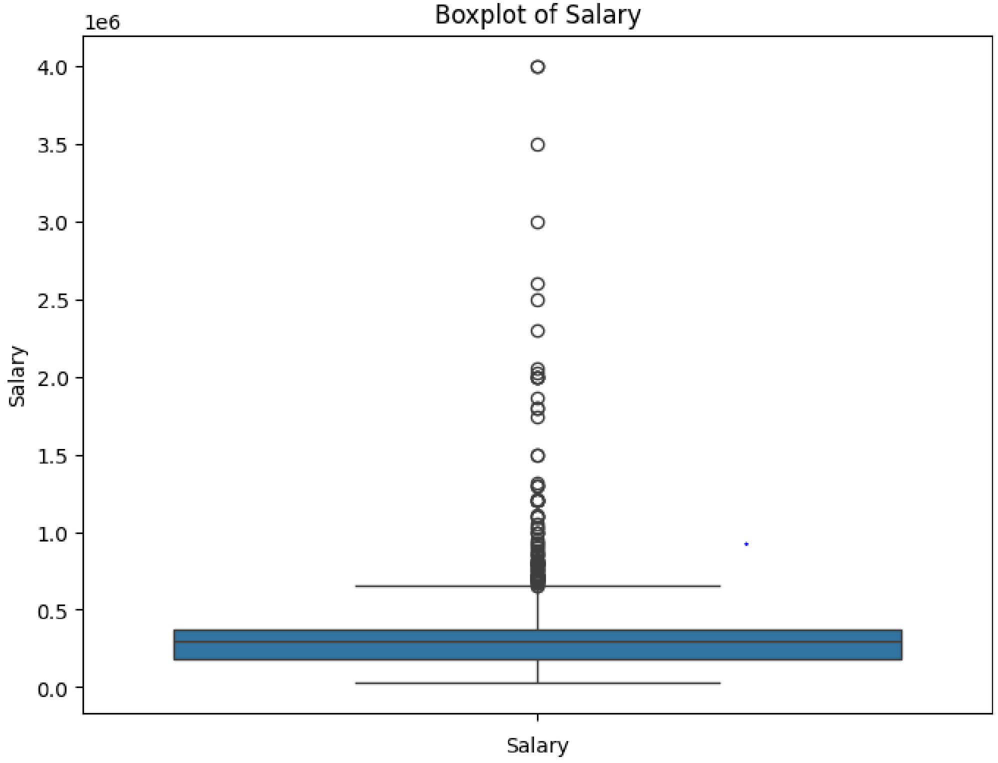
We then proceeded to explore relationships between variables. This involved analyzing scatter plots, box plots, bar plots, and other visualizations to identify patterns and correlations between numerical and categorical variables.

Research Questions: We addressed specific research questions related to the earnings of fresh graduates and potential relationships between gender and specialization preferences.

UNIVARIATE ANALYSIS:

- The graph represents a box plot of average salary, with salary on the y-axis and years of experience on the x-axis. Here are some of the key observations I can make:
- **Median salary:** The horizontal line in the middle of each box represents the median salary for each group of experience. It appears that the median salary increases with years of experience, but there is a gap in the data for the most experienced group.
- **Distribution of salaries:** The boxes show the interquartile range (IQR), which represents the middle 50% of salaries in each group. The boxes get wider as experience increases, indicating that there is more variability in salaries for more experienced workers.
- **Outliers:** The whiskers extend from the top and bottom of the boxes to show any outliers in the data. There are a few outliers in the data, particularly for the less experienced groups.

This boxplot only shows the average salary for each group of experience. It does not tell us anything about the individual salaries of workers, or about the factors that influence salary such as education, job title, or location.

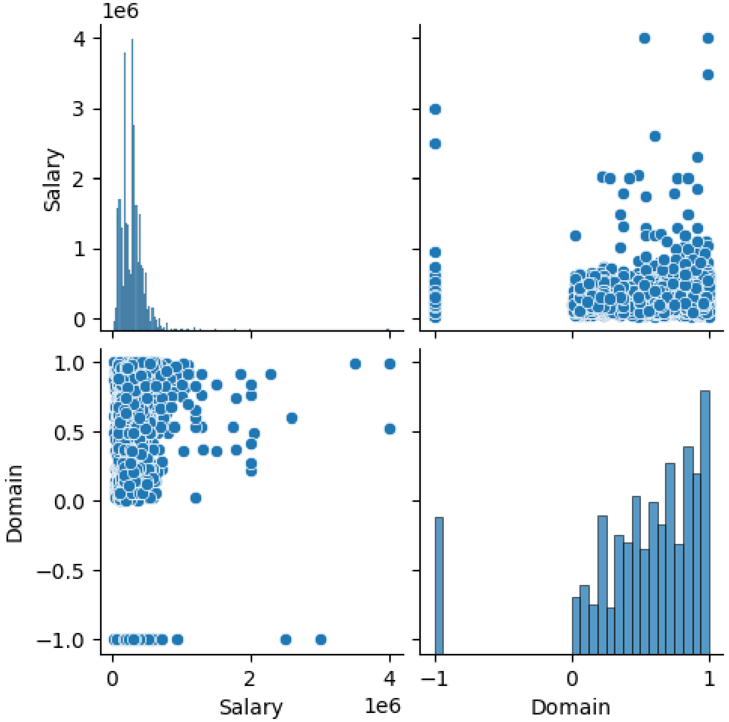


BIVARIATE ANALYSIS:

1. **Top left corner:** The top left corner of the "Salary" vs. "Domain" scatter plot shows a dense concentration of points, indicating many individuals with low salaries and low domain values. This might represent entry-level positions or less specialized roles.

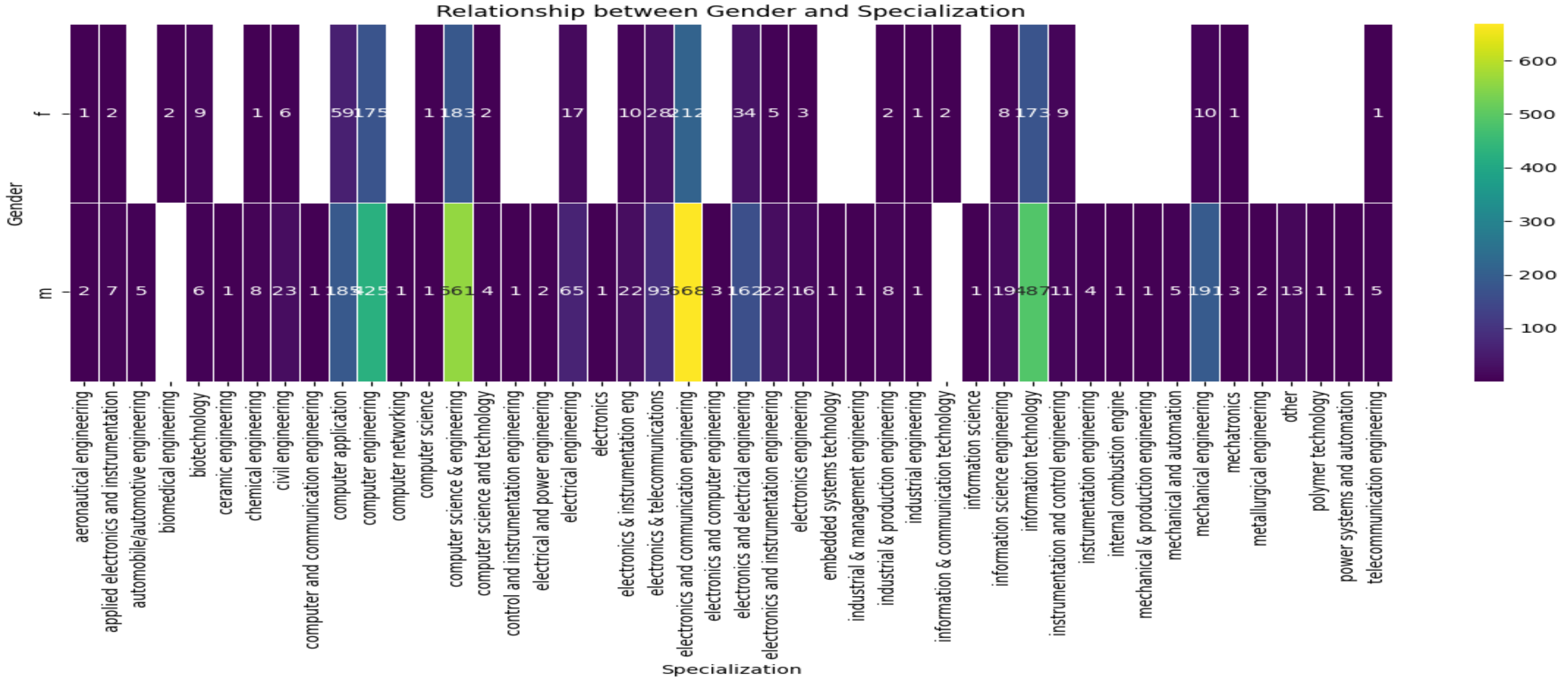
2. **Top right corner:** The top right corner shows fewer points, suggesting that there are fewer individuals with both high salaries and high domain values. This could be due to the limited number of highly specialized and high-paying positions available.

3. **Diagonal trend:** The diagonal trend in the "Salary" vs. "Domain" scatter plot reinforces the positive correlation between the two variables. As domain values increase, salary values also tend to increase.



RESEARCH QUESTION

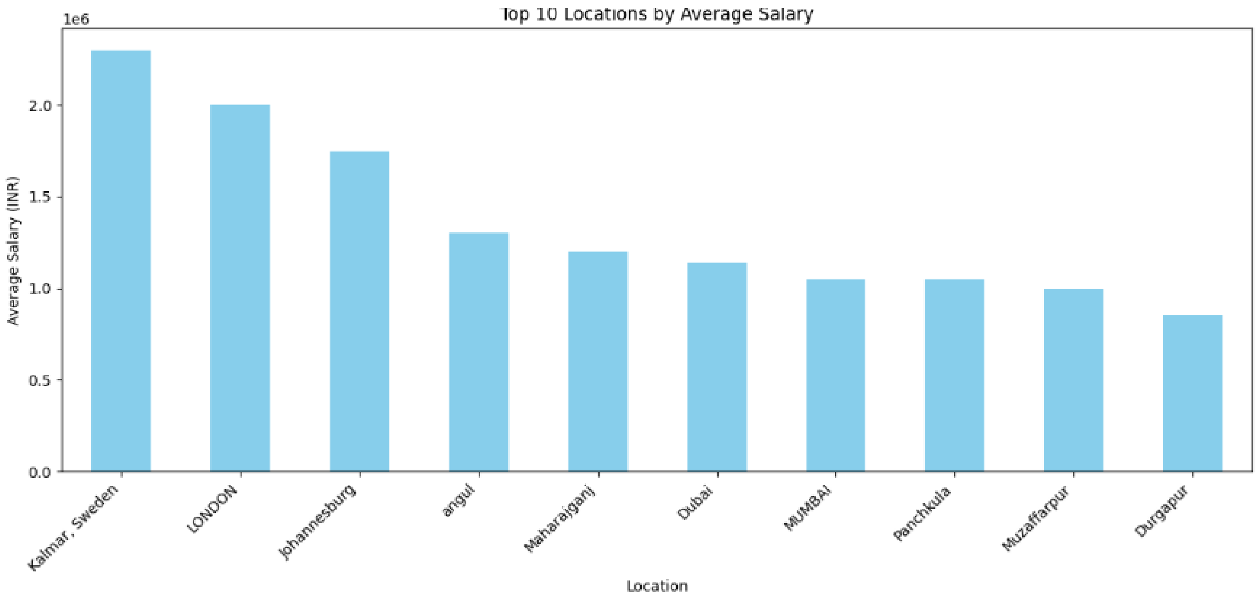
- Times of India article dated Jan 18, 2019 states that “After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can earn up to 2.5-3 lakhs as a fresh graduate.” Test this claim with the data given to you. Is there a relationship between gender and specialization? (i.e. Does the preference of Specialisation depend on the Gender?)



BONUS

Earnings After Graduation:

While the average salary for specified job roles among fresh Computer Science Engineering graduates falls within the range mentioned in the Times of India article, further investigation into factors such as location, industry, and company size could provide insights into variations in earnings. Research Question: What factors contribute to variations in earnings among fresh graduates in the specified job roles?



SOLUTION - INFERENCE

The top 10 cities with the highest average salaries in the df1 DataFrame. The cities are listed on the x-axis, with the city with the highest average salary on the left and the city with the lowest average salary on the right. The average salary for each city is represented by a blue bar. The height of the bar indicates the magnitude of the average salary. The y-axis shows the average salary in INR.

Based on the graph, we can see which cities have the highest average salaries in the data. This information could be useful for people who are considering moving to a new city for work or who are interested in learning more about salary trends across different locations.

CONCLUSION

In conclusion, we summarized our findings and highlighted key insights from the analysis.

We also discussed potential implications of our results and suggested areas for further investigation or research.

Overall, our analysis provided valuable insights into factors influencing earnings among fresh graduates in specific job roles.

Variation in Earnings:

Our analysis revealed significant variations in earnings among fresh graduates, with factors such as specialization, location, and company size playing crucial roles in determining salary levels.

Impact of Specialization:

The choice of specialization appears to have a substantial impact on earnings potential. Specializations in high-demand fields such as Computer Science and Engineering tend to command higher salaries compared to other disciplines.

Geographical Influence:

Location emerged as a significant factor influencing earnings, with graduates working in certain cities or regions earning higher salaries on average. This suggests that economic factors and cost of living variations contribute to salary discrepancies.

CHALLENGES WORKING ON EDA PROJECT

Data Quality:

One of the primary challenges was dealing with data quality issues such as missing values, inconsistencies, and outliers. Cleaning and preprocessing the data to ensure its accuracy and reliability required careful attention and thorough validation techniques.

Complexity of Analysis:

Analyzing the relationships between various factors influencing earnings among fresh graduates involved complex statistical and analytical techniques. Understanding the underlying patterns and trends in the data required expertise in data analysis and interpretation.

Interdisciplinary Knowledge:

The project required interdisciplinary knowledge spanning fields such as statistics, economics, and sociology. Integrating insights from multiple disciplines to draw meaningful conclusions posed a challenge, necessitating collaboration and consultation with domain experts.

Data Visualization:

Effectively communicating the findings through data visualization was another challenge. Choosing the right visualization techniques to convey complex information in a clear and concise manner required creativity and design skills.

THANK YOU