



**EDA Project - AMCAT Data Analysis**

**ABOUT ME:**

My name is Karthik Salikanti, and I'm an undergraduate student majoring in Computer Science with a minor in Data Science. I was captivated by the incredible possibilities that data offers and was drawn to this field due to my keen interest in leveraging data for transformative purposes.

Github: github.com/karthiksalikanti

LinkedIn: linkedin.com/in/karthik-salikanti

Business Problem and Description of the data set:

Objective:

To conduct a comprehensive analysis of the provided dataset, which focuses on the employment outcomes of engineering graduates. The analysis will delve into various aspects of the dataset, including dependent variables such as Salary, Job Titles, and Job Locations, as well as standardized scores from three different areas: cognitive skills, technical skills, and personality traits. Additionally, the analysis will explore demographic features present in the dataset.

Key Goals:

1. Explore and understand the structure of the dataset, including the types and distribution of variables.

2. Investigate the relationships between dependent variables (Salary, Job Titles, and Job Locations) and independent variables (cognitive skills, technical skills, personality traits, and demographic features).

3. Perform descriptive statistics to gain insights into the central tendencies and variability of the data & Inferential Statistics to test the validity of Given

4. Utilize visualization techniques to illustrate patterns, trends, and correlations within the dataset.

5. Conduct subgroup analyses based on demographic characteristics to identify any disparities or trends specific to certain groups.

Overall, the AMCAT Data Analysis project aims to extract valuable insights from the dataset to better understand the dynamics of employment outcomes for engineering graduates and to facilitate evidence-based decision-making in talent management and career development strategies.

EXPLORATORY DATA ANALYSIS:

* DATA CLEANING STEPS:
* After loading the dataset into the Jupyter working environment, thorough cleaning procedures were implemented using the Pandas library. Outliers within each numerical column were identified and subsequently removed, ensuring the integrity of the data for further analysis.
* UNIVARIATE ANALYSIS:
* Focusing on Salary as the target variable, an array of visualization techniques including Probability Density Functions (PDF), Histograms, Boxplots, and Countplots were employed. Key findings from this univariate analysis include:
* Summary statistics of numerical columns reveal significant insights:
* - The Salary column exhibits a wide range spanning from 35k to 4 million, with a considerable standard deviation of 212k, indicating notable variability in salaries.
* - Mean values of 10percentage and 12percentage columns hover around 77-78, with relatively lower standard deviations, indicating less variability.
* - The Domain column displays a mean value of 513 and a standard deviation of 0.51, reflecting a narrow range of values and low variability.
* Initial data inspection reveals noteworthy observations:
* - Salaries vary widely, ranging from 200k to over 1 million.
* - Predominantly, employees possess engineering degrees and occupy tech-related roles.
* - Psychological trait scores such as conscientiousness and openness hold potential insights into job performance.
* - Exploring college tiers and locations may unveil relationships between education quality and career outcomes.
* - Analysis of gender breakdown could illuminate any salary or role disparities.
* Correlation matrix analysis highlights significant relationships between different variables within the dataset. Notably, a weak negative correlation exists between conscientiousness and salary (-0.247), while a weak positive correlation is observed between agreeableness and salary (0.177). Furthermore, a positive correlation is noted between conscientiousness and extraversion (0.175), along with a negative correlation between conscientiousness and neuroticism (-0.146). These correlations offer valuable insights into the interplay between psychological traits and job performance indicators.
* These preliminary observations underscore the dataset's rich diversity, encompassing salary data, educational backgrounds, job roles, and psychological traits. Further analysis is warranted to unearth trends and patterns within the data.
* BIVARIATE ANALYSIS:
* Building upon the univariate analysis, bivariate analysis extends to categorical columns, addressing key inquiries:
* - Relationship exploration between numerical columns employs visualization techniques such as Scatter plots, Hexbin plots, and Pair plots.
* - Patterns between categorical and numerical columns are identified using Swarmplot, Boxplot, and Barplot.
* - Relationships between categorical and categorical columns are unveiled through Stacked Bar plots.
* In essence, the bivariate analysis aims to deepen our understanding of the dataset's interrelations, facilitating nuanced insights into various data dimensions.

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?)

Based on the analysis conducted on the provided data:

1. The average salary for fresh graduates in roles such as Programming Analyst, Software Engineer, Hardware Engineer, and Associate Engineer is approximately ₹339,792. This average salary does not meet the claim mentioned in the Times of India article, which stated that fresh graduates in these roles can earn up to 2.5-3 lakhs.

2. There is a statistically significant relationship between gender and specialization preferences. The Chi-square test of independence resulted in a very low p-value of approximately 0.0000012454, indicating that gender does influence the choice of specialization.

3. The visualization of gender vs. specialization distribution shows distinct patterns in specialization preferences based on gender, further supporting the statistical findings.

In conclusion, the data analysis suggests that the average salary for the specified roles exceeds the claim made in the article, and there is a clear relationship between gender and specialization preferences. Gender plays a role in determining the choice of specialization among individuals.

Note: All the plots related to this Analysis are an integral part of the Jupyter Notebook Attached along with this document

The key highlights and summary of the data analysis made is provided here.