INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis ON AMEO(2015) Data

About me

- **My name is B.Prasanna Lakshmi.**
- **I did my graduation in Bsc(Data Science) and passed out in the year 2023**

- **The reasons I chose to study data science are :**

- **Demand: Data Scientists are in High Demand.**
- **Growth: Data Science Careers Have High Earning Potential.**
- **Opportunity: Data Science Has a Range of Potential Job Opportunities.**
- ***LinkedIn***
- ***Github***

Objective

- **The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates.**
- **The dataset contains the employment outcomes of engineering graduates as dependent variables (Salary, Job Titles, and Job Locations) along with the standardized scores from three different areas – cognitive skills, technical skills and personality skills.**
- **With around 40 independent variables and 4000 data points, these variables encompass both continuous and categorical data. The dataset also includes demographic features and unique identifiers for each candidate.**
- **The objective of this Exploratory Data Analysis (EDA) is to comprehensively explore the Aspiring Mind Employment Outcome 2015 (AMEO) dataset, focusing on understanding the relationship between various features and the target variable, Salary.**

Key Goals

- **Dataset Overview:** Provide a detailed overview of the dataset's features, including continuous and categorical variables, as well as demographic information and unique identifiers.
- **Pattern Identification:** Identify any discernible patterns or trends within the dataset through visualizations and statistical analysis.
- **Relationship Investigation:** Investigate the relationships between independent variables and the target variable (Salary) to uncover insights into factors influencing salary outcomes.
- **Outlier Detection:** Perform outlier detection to identify any anomalies or abnormalities in the dataset that may affect the analysis.
- **Insights and Recommendations:** Offer practical insights and recommendations based on the analysis findings, aiming to provide valuable insights for stakeholders and decision-makers.

Key Features

- **ID:** Unique identifier for each candidate**.**
- **Salary:** Annual CTC (Cost to Company) offered to the candidate (in INR).
- **DOJ:** Date of joining the company.
- **DOL:** Date of leaving the company.
- **Designation:** Job title or position offered to the candidate.
- **JobCity:** Location of the job (city).
- **Gender:** Gender of the candidate.
- **DOB:** Date of birth of the candidate**.**
- **10percentage:** Overall marks obtained in grade 10 examinations.
- **10board:** School board whose curriculum the candidate followed in grade 10.
- **12graduation:** Year of graduation from senior year high school.
- **CollegeID:** Unique identifier for the college attended by the candidate.
- **CollegeTier:** Tier of the college attended.
- **Degree:** Degree obtained or pursued by the candidate.
- **Specialization:** Field of specialization pursued by the candidate.
- **CollegeGPA:** Aggregate GPA (Grade Point Average) at graduation.
- **CollegeCityID:** Unique identifier for the city where the college is located.
- **CollegeCityTier:** Tier of the city where the college is located.
- **CollegeState:** Name of the state where the college is located.
- **GraduationYear:** Year of graduation with a bachelor's degree.
- **12board:** School board whose curriculum the candidate followed in grade 12..
- **12percentage:** Overall marks obtained in grade 12 examinations.

DATA CLEANING AND PRE-PROCESSING

- **Handling Missing Values:** Identified and addressed missing values in various columns using appropriate techniques such as imputation or removal. This included replacing 0 and -1 values through imputation using mode for categorical variables and mean or median for numerical variables, as well as dropping unwanted columns.

- **Addressing Outliers:** Identified outliers through visualization and statistical methods, and decided whether to remove or adjust them based on context.

- **Standardizing Data:** Standardized data formats, such as date formats, to ensure consistency and compatibility across the dataset.

- **Feature Engineering:** Derived new features or transformed existing ones to extract meaningful insights and improve model performance.

- **Data Validation:** Validated the dataset to ensure it aligns with expectations and domain knowledge, correcting any discrepancies or anomalies as necessary.

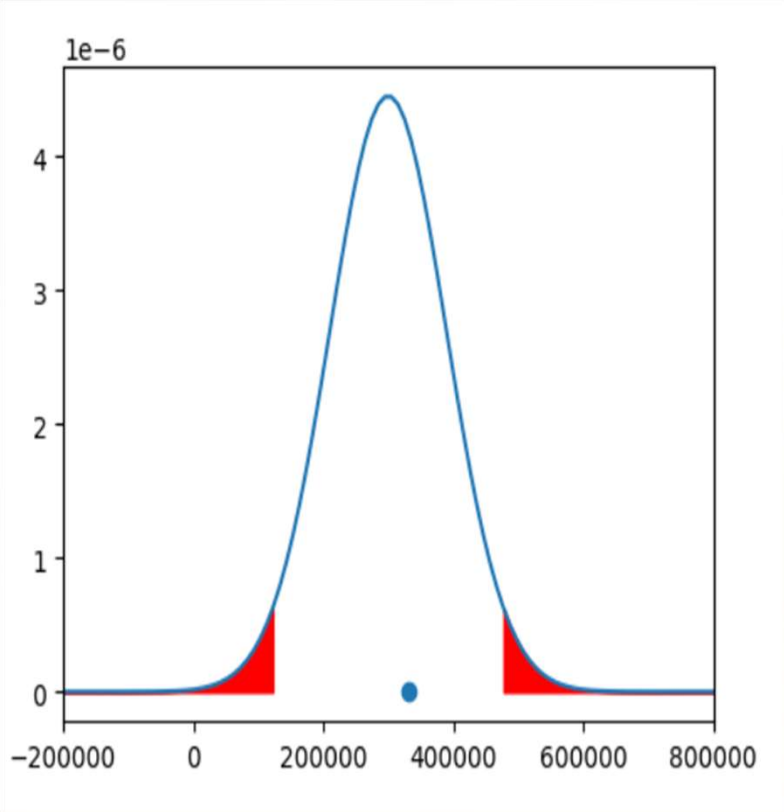
EXPLORING INDIVIDUAL VARIABLES: INSIGHTS FROM ANALYSIS

- **The dataset predominantly consists of candidates from CBSE schools, highlighting a preference for CBSE curriculum.**
- **Candidates from state boards follow CBSE in representation, showcasing widespread state board education adoption.**
- **Employees born in 1991 form the largest group, followed by those born in 1992, indicating diverse age distribution.**
- **Highest employee joining observed in 2014, with a rising trend from 2010 to 2014.**
- **Significant gender imbalance with more male employees than female.**
- **Tier 0 college cities have higher representation than Tier 1, implying varied employment opportunities.**
- **College ID 272 has the most employees, followed by College ID 64, indicating recruitment variations.**
- **2024 saw the highest employee attrition, akin to 2015 trends, showing fluctuating attrition rates.**
- **Programming Analyst and Software Engineer are prevalent job roles.**
- **Electronics and Communication Engineering is the most common specialization.**
- **UP, Karnataka, and Tamil Nadu lead in employee representation.**
- **Bangalore is the primary employment hub, followed by Noida and Hyderabad.**
- **UP employees are disproportionately represented, indicating regional prominence.**
- **Educational background influences employment prospects and career trajectories.**
- **Gender, age, and region shape employment patterns and opportunities.**
- **Demographic and geographic trends inform recruitment and training strategies.**
- **The dataset offers insights into labor market dynamics for informed decision-making.**
- **Continuous analysis is crucial to anticipate future workforce needs.**
- **Data-driven insights optimize recruitment, engagement, and retention.**
- **Understanding employment factors drives sustainable growth and societal value.**

Exploring Salary Trends Across Factors

- **Salary vs. Age:** The plot reveals a concentration of younger individuals with lower salaries, indicating potential entry-level positions. However, there are outliers with higher salaries across different age groups.
- **Salary vs. Tenure:** There's a positive correlation between salary and tenure, suggesting that longer-tenured employees tend to earn higher salaries. However, there's variability in salaries even at similar tenure levels.
- **Salary vs. Gender:** While there are more male employees than female employees, the median salary lines for both genders are similar, indicating comparable salary distributions.
- **Salary vs. College Tier:** Employees from Tier 0 colleges earn higher average salaries compared to Tier 1 colleges, indicating potential differences in educational quality or employer preferences.
- **Salary vs. College ID :** Certain colleges, such as those with IDs 239 and 38, show higher average salaries compared to others, suggesting potential differences in academic reputation or alumni networks.
- **Salary vs. College GPA:** While GPA is important, it may not have a direct correlation with salary, as individuals with varying GPAs exhibit similar salary distributions.
- **Correlation Matrix Insights:** Salary correlates positively with logical and quantitative skills but negatively with English proficiency. Personality traits show positive correlations among themselves but no consistent relationship with salary.
- **Salary vs. Graduation Year:** The analysis shows an increasing trend in average salaries from 2007 to 2010, followed by a decline till 2014 and a slight increase in 2016, indicating fluctuations in job market demand over time.
- **Salary vs. College State:** States like Jharkhand, Assam, and Goa show relatively high average salaries, potentially reflecting regional economic factors or industry presence.
- **Salary vs. Job City:** Certain cities like Kohlapur, Bhopal, and Ghaziabad exhibit high average salaries, indicating strong job markets or higher cost of living.
- **Salary vs. 10th/12th Board:** Employees from CBSE and state boards show higher representation, suggesting a preference for these educational systems among employers.
- **Salary vs. Degree:** B.Tech/BE and M.Tech/ME graduates show higher average salaries, with some outliers, while MCA graduates also have notable salary levels.
- **Salary vs. Specialization:** Certain specializations like Polymer Technology and Computer Networking show higher average salaries, indicating industry demand or specialized skills.
- **Salary vs. State:** Uttar Pradesh has the highest representation, followed by Karnataka and Tamil Nadu, with UP showing significantly higher numbers, potentially reflecting population demographics or economic factors.

Research Questions



Null Hypothesis(h1): 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.¶

Alternative Hypothesis(h0): After doing your Computer Science Engineering if you take up jobs as a Programming Analyst, Software Engineer, Hardware Engineer and Associate Engineer you can't earn up to 2.5-3 lakhs as a fresh graduate.¶

```
confidence_level = 0.95
alpha = 1 - confidence_level

chi2_critical = chi2.ppf(1 - alpha, dof)

print(chi2_critical)
if(chi2_stat > chi2_critical):
    print("Failed to reject the Null Hypothesis")
else:
    print(" reject the Null Hypothesis")
```

```
61.65623337627955
Failed to reject the Null Hypothesis
```

Conclusions :

- **Educational Choices: Consider CBSE and state boards for better career prospects.**
- **Geographic Considerations: Explore job opportunities in states like Jharkhand, Assam, and Goa, and cities like Kohlapur, Bhopal, and Ghaziabad.**
- **Specialization Selection: Opt for specializations like Polymer Technology and Computer Networking for higher-paying roles.**
- **Career Paths: Aim for roles like Programming Analyst and Software Engineer for higher average salaries.**
- **Skill Development: Focus on improving logical and quantitative skills and domain knowledge.**
- **Language Proficiency: Maintain strong communication skills, though English proficiency may not directly impact salary.**
- **Personality Traits: Develop a well-rounded personality and interpersonal skills for career success.**
- **Continuous Learning: Emphasize lifelong learning to adapt to changing job market demands.**
- **Networking: Leverage alumni networks and connections from colleges with higher average salaries.**
- **Job Market Trends: Stay informed about market trends and seek employment during favorable periods.**