INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

# PROJECT ON

Explorartory Data Analysis on AMCAT Dataset

# About me

I am Narendra Kotapati, a proactive, responsible, and results-oriented professional who recently completed my bachelor's of engineering. Following my graduation, I pursued a Data Science course to deepen my understanding and expertise in the field of data analysis and data science. My interests lie in solving technical issues, conducting research, and innovating new technologies. I thrive in team environments and enjoy connecting with new individuals, leveraging my kind and outgoing nature to foster collaboration. Additionally, I am a quick learner and excel in working under pressure, possessing excellent stress management skills.

My enthusiasm for Data Science stems from its transformative potential across various industries. I am intrigued by the prospect of extracting valuable insights from vast amounts of data, which can drive informed decision-making and innovation. Data Science offers a powerful toolkit to uncover patterns, trends, and correlations that can significantly impact businesses and society. Through my ongoing journey in the field of Data Science, I aim to contribute to meaningful projects and leverage data-driven approaches to drive positive change.

## Objective of the Project :

This analysis aims to gain insights and understanding from the provided dataset, particularly focusing on the relationship between various features and the target variable, which is Salary.

Specifically, the goals of this analysis include: • Describing the dataset and its features comprehensively. • Identifying any patterns or trends present in the data. • Exploring the relationships between independent and target variables (Salary). • Identifying any outliers or anomalies in the data

## Summary Of The Data :

The Aspiring Mind Employment Outcome 2015 (AMEO) dataset, released by Aspiring Minds, focuses on employment outcomes for engineering graduates. It includes dependent variables such as Salary, Job Titles, and Job Locations, along with standardized scores in 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

## Data Preprocessing:

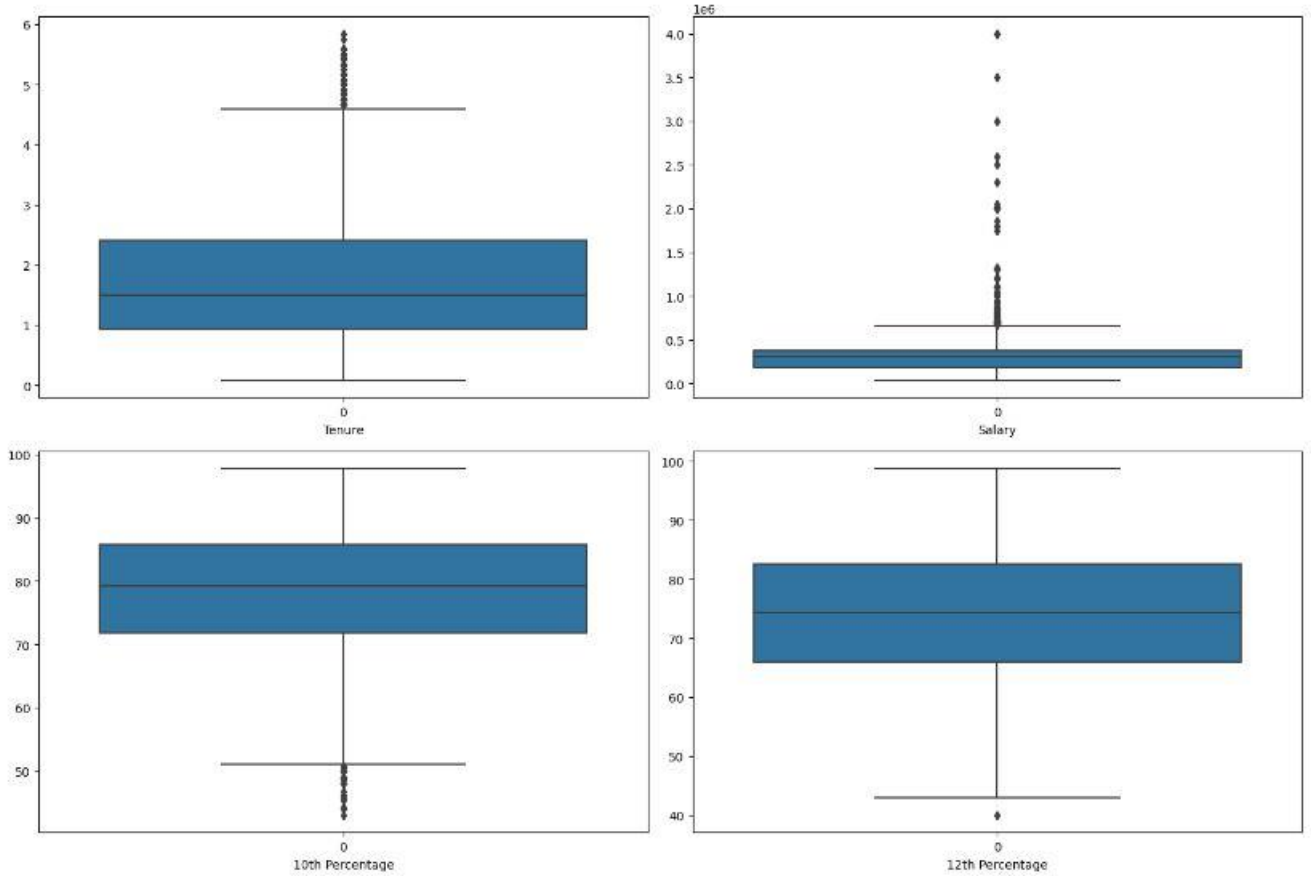
The dataset contains employee information including tenure, salary, educational qualifications, and specialization.
Data preprocessing steps include handling missing values, data type conversions, and outlier detection/removal.
The data handling process has been successfully completed. Firstly, the columns '10board', '12board', 'GraduationYear', 'JobCity', and 'Domain' were processed to replace null values represented by 0 or -1
.

## Feature Engineering:

Tenure Calculation: Another new feature, 'tenure', has been introduced by subtracting the 'Date of Leaving' (DOL) from the 'Date of Joining' (DOJ). This indicates the duration of an individual's employment within the company.
Graduation Year Filtering: Rows where the graduation year is greater than or equal to the date of joining have been removed. This ensures data integrity by excluding instances where the graduation year suggests a date after the individual's employment start date.

## Univariate analysis

**1) Numerical continuous Features:**
**Tenure Box Plot:** Majority of employees have a tenure within a certain range with few outliers.
**Salary Box Plot:** Salary distribution shows significant variability with some high outliers.
**10th & 12th Percentage Box Plots:** Wide range of percentages observed with some outliers at the lower end.
**College GPA Box Plot:** Moderate variability in GPA with a few lower outliers.
**Computer Programming & Quant Score Box Plots:**
Scores exhibit variability with some high outliers.



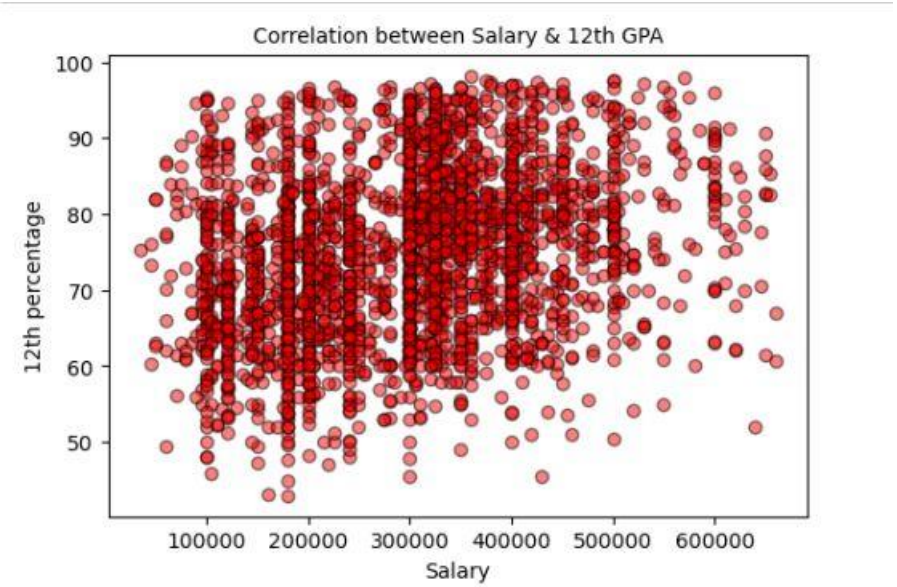

**2) Categorical Features:**
**Gender Distribution Pie Chart:** Shows the distribution of employees based on gender.
**Degree Count Plot:** Most individuals have a degree in Engineering.
**Specialization Count Plot:** Electronic communication Engineering and Computer Science & Engineering are the most common specializations.

## Bivariate analysis :

**Average Salary per Gender Bar Plot:** Males tend to have higher average salaries compared to females.

**Average Salary for Top Designations Bar Plot:** Certain designations have higher average salaries compared to others.

**Correlation Plots:** Weak positive correlations observed between salary and 10th percentage, 12th percentage, and college GPA. No apparent correlation observed between salary and tenure.



## Conclusion :

The analysis provides valuable insights into the employee data, highlighting key factors such as tenure, salary, and educational background.
Recommendations include:
Further investigation into factors influencing salary discrepancies based on gender and designation.
Implementation of targeted skill development programs based on educational background and specialization.
Regular monitoring of employee tenure and satisfaction to improve retention rates.

# THANK YOU

INNOMATICS RESEARCH LABS