INNOMATİCS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

EDA Project - Analysis of AMCAT Data

About me

- **Background ? (B-tech or M-tech)**
  My Academic Background is BSc (Bachelor of Science)

- **Why you want to learn Data Science**
- *Data science is like being a detective of information, equipped with powerful analytical tools and techniques to solve complex puzzles. Whether it's predicting customer behavior, optimizing business operations, or advancing scientific research, data science empowers us to extract meaningful knowledge from the sea of information that surrounds us.*

- **Any work experience**
- *I have gained valuable practical insights from my internship experience at Shanxi Zhijie Software Engineering Co., Ltd. working as a full time Front-end Engineer in Health Insurance Department. The most valuable lesson that I have learned from industrial professionals is the fact that "Data cannot make an impact without proper manipulation", and I have dedicated three months in designing a more efficient healthcare information system to maximize the value derived from medical data. During this process, I have experienced the need for further research and development in Information Systems, especially when different departments in a hospital have varying analytical needs for the same medical data.*

- **Share your linkedin and github profile urls**
- *https://www.linkedin.com/in/codereric1955/*
- *https://github.com/00AKun*

INNOMATICS RESEARCH LABS

Agenda

- **The Introduction of Project**
  - *Objective of the Project*
  - *Summary of the Data*

- **Exploratory Data Analysis**
  - *Data Cleaning*
  - *Data Manipulation*
  - *Univariate Analysis*
  - *Bivariate Analysis*

- **Research Questions**
  - *Claim Key Questions*
  - *Conclusion (Findings)*
  - *Experience/Challenges working on Analysis of AMCAT Data*

INNOMATICS RESEARCH LABS

The Introduction of Project

Objective of the Project

We used Exploratory Data Analysis to develop the Data Analysis of the Dataset released by Aspiring Minds under the Aspiring Mind Employment Outcome ***2015*** (AMEO), including Data Cleansing, Data Manipulation, Univariate Analysis and Bivariate Analysis.

Summary of the Data

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. The dataset also contains demographic features. The dataset contains around ***40*** independent variables and ***4000*** data points. The independent variables are both continuous and categorical in nature. The dataset contains a unique identifier for each candidate. Below mentioned table contains the details for the original dataset.

Exploratory Data Analysis

Data Cleaning

We analysed all Continuous Variables in all data.csv for Missing Values, Outliers and Anomalies, eliminated those values that behaved abnormally, and generated a cleaned new dataset cleaned_data.csv.

Data Manipulation

Import the Data and display the Head, Shape, and Description.

```python
import pandas as pd

# Read the CSV file
data = pd.read_csv('data.csv')

# Check for missing values
missing_values = data.isnull().sum()
print("Missing Values:")
print(missing_values)

# Check for outliers and anomalies in continuous variables
continuous_variables = ['Salary', '10percentage', '12percentage']

for variable in continuous_variables:
    # Calculate the z-score for each value in the variable
    z_scores = (data[variable] - data[variable].mean()) / data[variable].std()

    # Identify outliers using a threshold (e.g., z-score > 3 or < -3)
    outliers = data[variable][(z_scores > 3) | (z_scores < -3)]

    print(f"Outliers in {variable}:")
    print(outliers)

    # Remove outliers from the dataset
    data = data.drop(outliers.index)

# Save the cleaned data to a new CSV file
data.to_csv('cleaned_data.csv', index=False)
```

```python
import pandas as pd

# Import the dataset
data = pd.read_csv("cleaned_data.csv")

# Display the head of the data
print(data.head())

# Display the shape of the data
print("Shape of the data:", data.shape)

# Display the description of the data
print(data.describe())
```

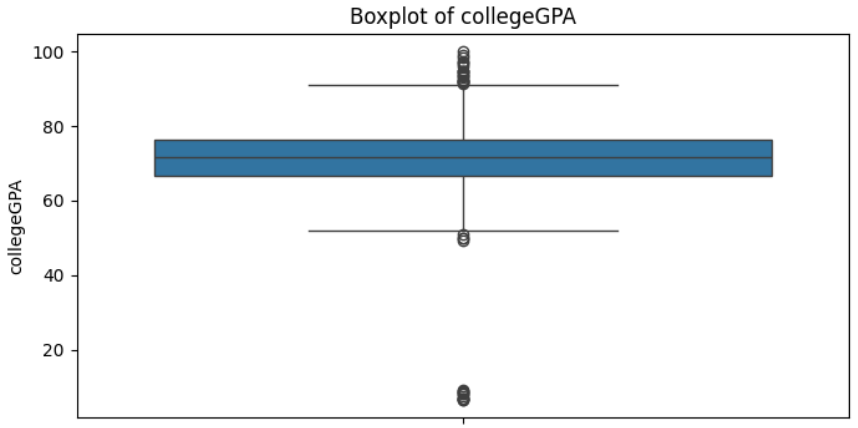
```
  Unnamed: 0       ID    Salary          DOJ          DOL  \
0      train   203097  420000.0  6/1/12 0:00      present
1      train   579905  500000.0  9/1/13 0:00      present
2      train   810601  325000.0  6/1/14 0:00      present
3      train   343523  200000.0  3/1/14 0:00  3/1/15 0:00
4      train  1027655  300000.0  6/1/14 0:00      present
```

INNOMATICS RESEARCH LABS

Exploratory Data Analysis

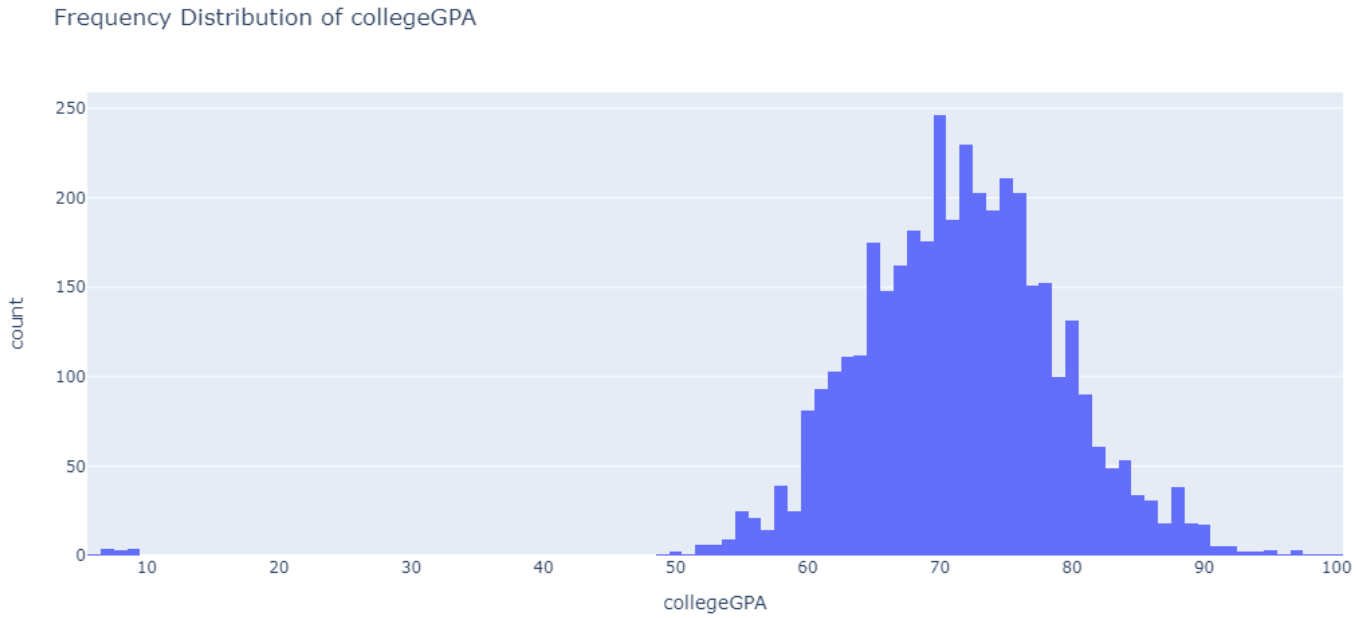
Univariate Analysis

- Find the outliers in each numerical column
- Understand the probability and frequency distribution of each numerical column
- Understand the frequency distribution of each categorical Variable/Column



Univariate Analysis

```
[63] # Importing the required libraries
     import pandas as pd
     import matplotlib.pyplot as plt
     import seaborn as sns

     import plotly.express as px
     import pandas as pd

     # Loading the dataset
     df = pd.read_csv('cleaned_data.csv')

     # Numerical Columns
     numerical_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA', 'English', 'Logical',
                          'Quant', 'Domain', 'ComputerProgramming', 'ElectronicsAndSemicon',
                          'ComputerScience', 'MechanicalEngg', 'ElectricalEngg', 'TelecomEngg',
                          'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion',
                          'nueroticism', 'openess_to_experience']

     # Categorical Columns
     categorical_columns = ['Designation', 'JobCity', 'Gender', 'CollegeTier', 'Degree',
                            'Specialization', 'CollegeCityTier', 'CollegeState']

     # Univariate Analysis - Numerical Columns
     for column in numerical_columns:
         # Outliers
         plt.figure(figsize=(8, 4))
         sns.boxplot(data=df, y=column)
         plt.title(f'Boxplot of {column}')
         plt.show()

         # Probability Distribution
         plt.figure(figsize=(8, 4))
         sns.histplot(data=df, x=column, kde=True)
         plt.title(f'Distribution of {column}')
         plt.xlabel(column)
         plt.show()
```



INNOMATICS RESEARCH LABS

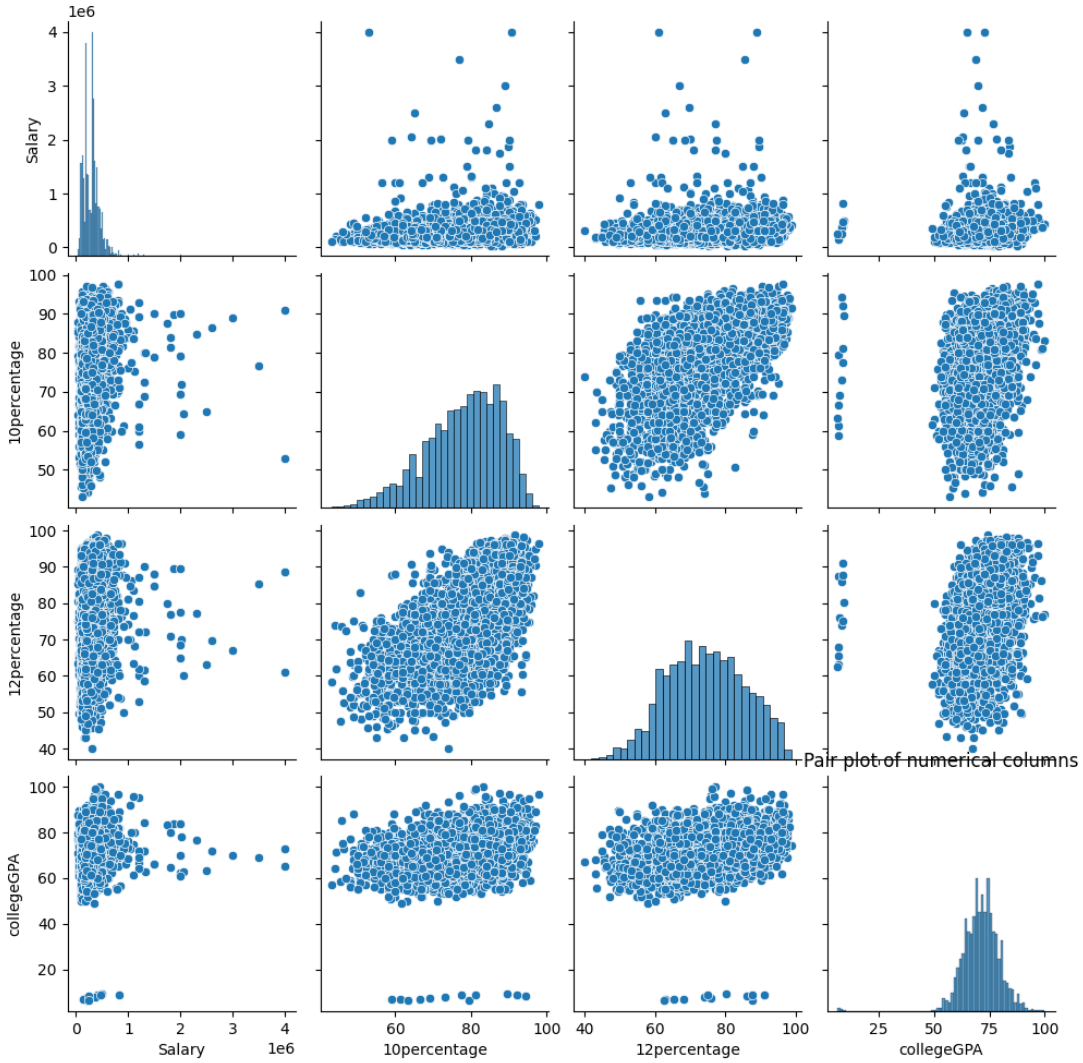
Exploratory Data Analysis

Bivariate Analysis

- Discover the relationships between numerical columns using Scatter plots, hexbin plots, pair plots, etc..
- Identify the patterns between categorical and numerical columns using swarmplot, boxplot, barplot, etc..
- Identify relationships between categorical and categorical columns using stacked bar plots.

Bivariate Analysis

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Load the dataset
data = pd.read_csv('cleaned_data.csv')

# Numerical Columns
numerical_columns = ['Salary', '10percentage', '12percentage', 'collegeGPA', 'English', 'Logical', 'Quant', 'Domain',
                     'ComputerProgramming', 'ElectronicsAndSemicon', 'ComputerScience', 'MechanicalEngg',
                     'ElectricalEngg', 'TelecomEngg', 'CivilEngg', 'conscientiousness', 'agreeableness', 'extraversion',
                     'nueroticism', 'openess_to_experience']

# Scatter plot of Salary vs 10percentage
plt.scatter(data['10percentage'], data['Salary'])
plt.xlabel('10percentage')
plt.ylabel('Salary')
plt.title('Scatter plot of Salary vs 10percentage')
plt.show()

# Hexbin plot of Salary vs CollegeGPA
sns.jointplot(data=data, x='collegeGPA', y='Salary', kind='hex')
plt.xlabel('collegeGPA')
plt.ylabel('Salary')
plt.title('Hexbin plot of Salary vs CollegeGPA')
plt.show()
```



INNOMATICS RESEARCH LABS

Research Questions

Claim Key Questions

- 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?)
- Is there a relationship between gender and designation (i.e. Does the preference of designation depend on the Gender?)

Conclusion (Findings)

- The claim is not supported by the data. The average salary for fresh graduates in these job roles does not fall within the mentioned range.
- There is a significant relationship between gender and specialization.
- There is a significant relationship between gender and designation.

Experience/Challenges working on Analysis of AMCAT Data

- Choosing the right Plots to analyse the degree of correlation between different types of data.
- Marking may not be clear due to improper data setting of x-axis or y-axis.
- Data intervals for the x-axis and y-axis are not correctly labelled resulting in the image not displaying enough data points correctly.

THANK YOU

INNOMATICS RESEARCH LABS