INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

Exploratory Data Analysis Report

On

Aspiring Mind Employment Outcome (AMEO) Dataset

Prepared By – ANKIT RAJ

INNOMATICS
RESEARCH LABS

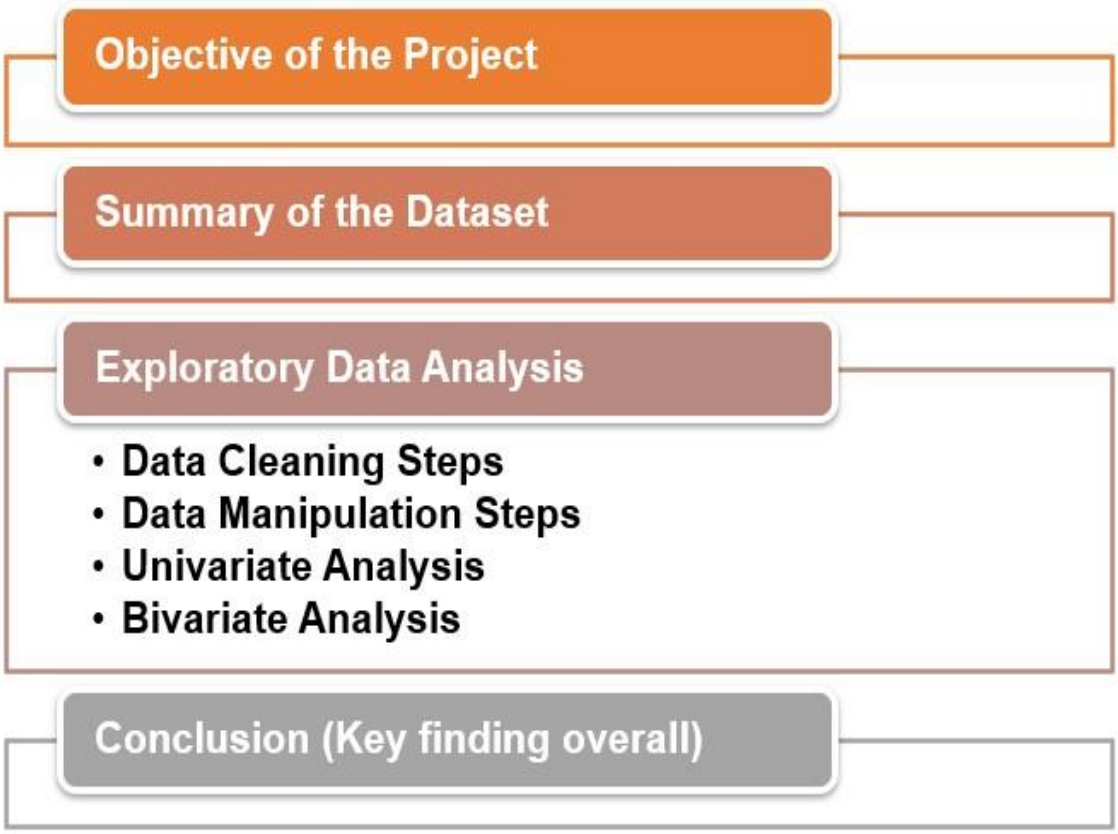
About Me

I've always been a proactive learner, a dedicated learner completed my B.Tech from GITAM in Computer Science Engineering, and prepared to contribute to organizational success while developing new skills and gaining real-world experience. I have outstanding writing, communication, and critical thinking skills. I am also very organized and responsible.

Since data science is the perfect method to merge my love of technology with my passion for problem-solving, I am eager to learn more about it. In today's data-driven culture, a ton of information is created every second, and I'm excited by the prospect of using this data to extract meaningful knowledge that can guide decisions and solve real-world issues.

I've contributed to several solo and group projects across a variety of industries, with a complete stack emphasis being the core focus.

Agenda of Report



Requirements

Programming Language: Python

Libraries: numpy, pandas, matplotlib, seaborn

Objective of the Project

The purpose of this analysis is to learn more about the dataset that has been supplied. Specifically, it will concentrate on the connections between different attributes and the target variable, salary.
The objectives of this analysis are as follows:

- Providing a detailed description of the dataset's attributes.
- Determining whether the data exhibits any patterns or trends.
- Analysing the connections between the target variable (salary) and independent factors.
- Finding any abnormalities or outliers in the data.

Summary of Dataset

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.

Source of Dataset: Innomatics Research Labs

Dataset Detailed Description:
https://docs.google.com/document/d/1xvo6kN5Q4QG-ezZKrNwg2YMD6S_fwlmz/edit?usp=sharing&ouid=101183190533718048323&rtpof=true&sd=true

Exploratory Data Analysis

Step 1: Data Cleaning

- Removing Unwanted Columns
- Data Type Conversion
- Collapsing Categories

Step 2: Data Manipulation

- Adding a Tenure Column
- Imputing Categorical column with mode values
- Validating 10, 12 percentage, and College GPA
- Checking the condition of DOL > DOJ
- Imputing Numerical Columns with median values

Glimpse of Cleaned Data

```
df.head()
```

✓ 0.0s

	Unnamed: 0	ID	Salary	DOJ	DOL	Designation	JobCity	Gender	DOB	10percentage	10board	12graduation	12percentage
0	train	203097	420000.0	6/1/12 0:00	present	senior quality engineer	Bangalore	f	2/19/90 0:00	84.3	board ofsecondary education,ap	2007	95.8
1	train	579905	500000.0	9/1/13 0:00	present	assistant manager	Indore	m	10/4/89 0:00	85.4	cbse	2007	85.0
2	train	810601	325000.0	6/1/14 0:00	present	systems engineer	Chennai	f	8/3/92 0:00	85.0	cbse	2010	68.2
3	train	267447	1100000.0	7/1/11 0:00	present	senior software engineer	Gurgaon	m	12/5/89 0:00	85.6	cbse	2007	83.6
4	train	343523	200000.0	3/1/14 0:00	3/1/15 0:00	get	Manesar	m	2/27/91 0:00	78.0	cbse	2008	76.8

Univariate Analysis

- A univariate analysis was conducted using box plots, histograms, summary plots, and CDFs (Cumulative Distribution Function) to analyze continuous features such as tenure, salary, and college GPA.
- It was noted that the data was not regularly distributed for the majority of the characteristics.
- The Boxplots provided insight into the high number of outliers present, which prompted the removal of outliers.
- Additionally, bar graph plots were created for the categorical features.

Outliers Removal

```
Click here to ask Blackbox to help you code faster
def outlier_treatment(datacolumn):
    sorted(datacolumn)
    Q1,Q3 = np.percentile(datacolumn , [25,75])
    IQR = Q3 - Q1
    lower_range = Q1 - (1.5 * IQR)
    upper_range = Q3 + (1.5 * IQR)
    return lower_range,upper_range
```

```
Click here to ask Blackbox to help you code faster
columns = ['Salary','10percentage','12percentage','English',
           'Logical','Quant','Domain', 'ComputerProgramming',
       'ElectronicsAndSemicon', 'ComputerScience', 'conscientiousness',
       'agreeableness', 'extraversion', 'nueroticism', 'openess_to_experience',
       'Age(2015)', 'Tenure', 'YearGap']
df_copy = df.copy()
```

Observations Summary:

Tenure and Salary Skewness: There is a notable positive skewness in both the tenure and salary data, suggesting that there is a concentration of values towards lower tenures and a bigger proportion of respondents with lower wages. This raises the possibility of problems with the dataset's remuneration and retention policies.

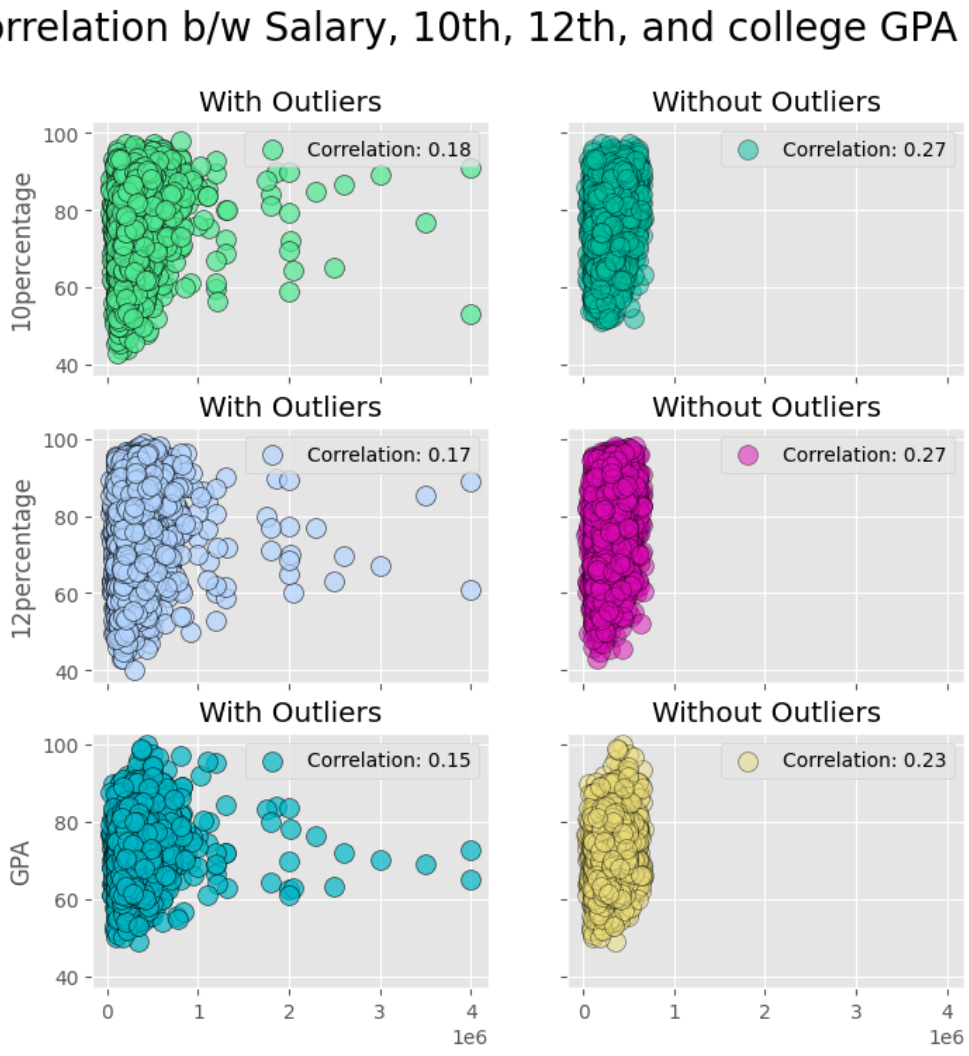
Student Performance Patterns: Although with minor variances, student performance in terms of percentage scores and GPA demonstrates a focus towards higher scores. The distribution of GPA is consistent, while the percentage scores show more variation, suggesting possible variations in grading standards or assessment techniques**.**

Distribution Deviation and Outliers: The appearance of outliers in the tenure and student GPA data points indicates the possibility of exceptional situations that call for more research. Furthermore, there is non-uniformity in the income and percentage score data, which may represent underlying inequities or biases, as indicated by the divergence from normal distribution.

Consequences for Making Decisions: It is essential to comprehend these patterns in order to make wise decisions. In order to improve teaching and assessment techniques, educators should investigate the elements that contribute to different student performance measures. Meanwhile, employers may need to address concerns linked to tenure distribution and wage equity.

Bivariate Analysis

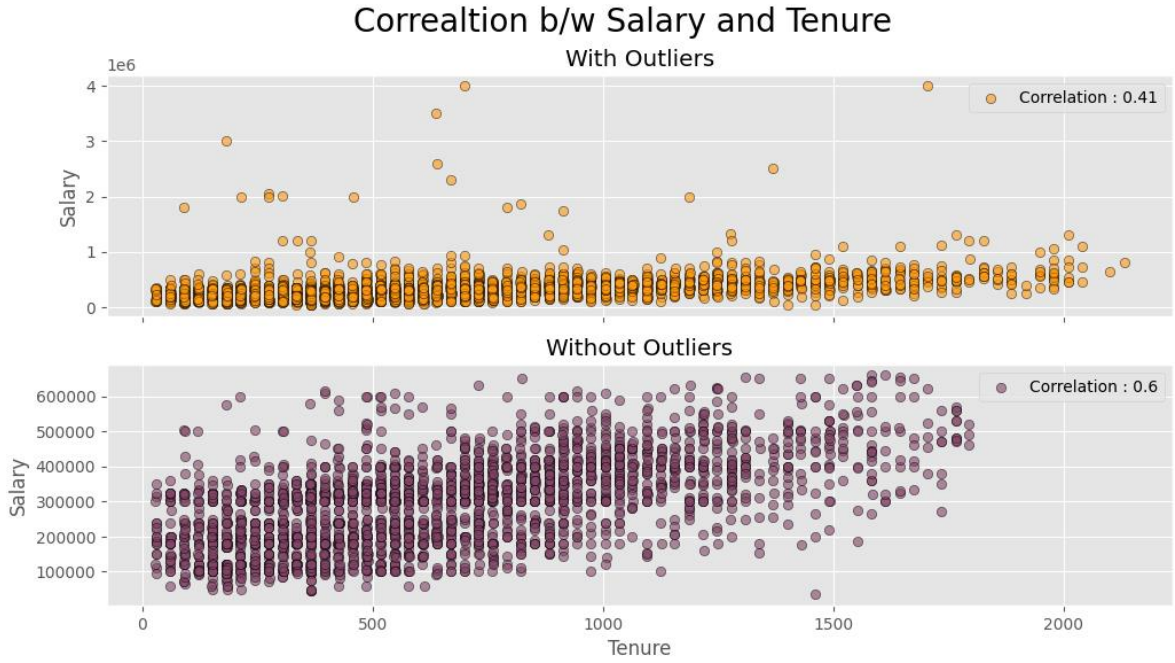
- With salary being one factor, each of the other characteristics was examined separately, and an analysis was conducted.
- In several instances, such as Gender, it was noted that the characteristic did not correlate with salary.
- However, in other instances, such as Designation, it was noticed that the highest salaries were earned by Software Engineers.



Absence of Correlation: Salary and educational results (10th, 12th, and GPA) do not correlate, indicating that academic achievement has little bearing on pay.

Salary Distribution by Designation: Although they make the most money, senior software engineers also have the most income volatility. The only professions with lower average pay are software developers and technical support engineers.

Gender Salary Equality: Average salaries for both genders are approximately equal, indicating no gender bias in salary distribution.



Salary-Tenure Relationship: Once the outliers are eliminated, there is a discernible 50% pay rise as tenure increases, suggesting a positive correlation (0.60) between tenure and compensation.

Salary Disparity between College TiersTier one universities pay more than tier two universities, with tier two universities paying less than the national average.

Research Outcome

The main objective of the project is to compare different percentage indicators by evaluating employee data. Boxplots are one tool we use in our study to find outliers in the dataset. Aside from that, we do several different studies using salary as the primary factor. Specifically, we use countplots to focus on work locations in order to identify the cities with the highest staff concentration.

Overall, this research highlights the necessity for comprehensive pay policies that take into consideration a variety of elements beyond job title or academic credentials and highlights the complexity of salary determinants in the computer industry. Deeper understanding of the complex interactions between these variables may be possible through more investigation into how to effectively structure salaries and manage personnel in the IT industry.