INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

Exploratory Data Analysis Report

On

# Aspiring Mind Employment Outcome (AMEO) Dataset

Prepared By – Mahima Churi



INNOMATICS
RESEARCH LABS

## About Me

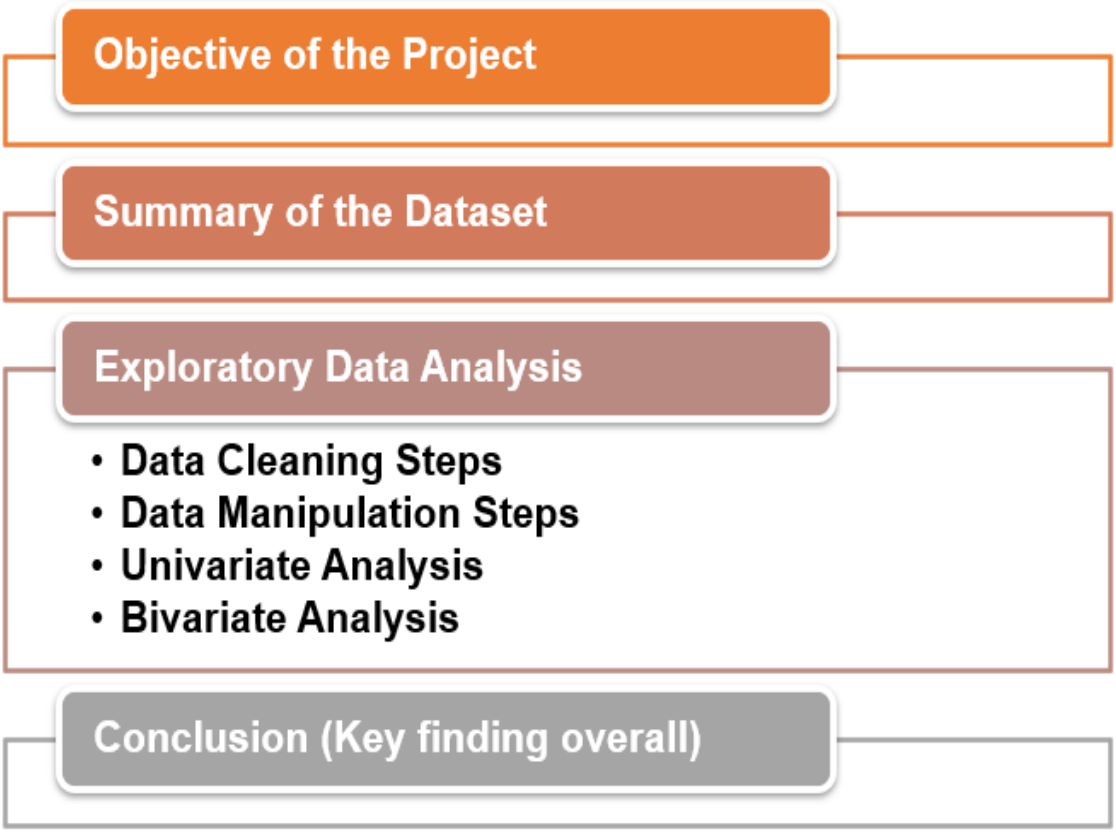
I've always been a proactive learner, a dedicated student currently finalizing my B.E. in Computer Engineering, and prepared to contribute to organizational success while developing new skills and gaining real-world experience. I am highly responsible and organized with excellent writing, communication, and critical thinking abilities.

I am excited to learn data science because it is an ideal way to combine my love of technology and my passion for solving problems. There is a tonne of information generated every second in today's data-driven society, and I'm enthusiastic about the possibility of gaining insightful knowledge from this data to inform choices and address practical problems.

I have worked on many individual and team projects in different domains, full stack being the primary focus and have developed my coding skills. I am also familiar with a few popular programming languages like C++, Python, PHP, Javascript, etc. I'm looking forward to connecting with outstanding people in the industry to work with them and explore more!!

## Connect with me

## Agenda of Report




## Requirements

**Programming Language:** Python

**Libraries:** numpy, pandas, matplotlib, seaborn, statsmodels

## 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 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

Click here to ask Blackbox to help you code faster

```
data.head()
```

Python

| | Salary | DOJ | DOL | Designation | JobCity | Gender | DOB | 10percentage | 10board | 12graduation | 12percentage | 12board | CollegeTier | Degree | Specialization | collegeGPA | Colle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 420000 | 2012-06-01 | 2015-12-31 | other | bangalore | f | 2/19/1990 0:00 | 84.3 | other | 2007 | 95.8 | other | 2 | B.Tech/B.E. | computer science & engineering | 78.00 | |
| 1 | 500000 | 2013-09-01 | 2015-12-31 | other | other | m | 10/4/1989 0:00 | 85.4 | cbse | 2007 | 85.0 | cbse | 2 | B.Tech/B.E. | electronics & telecommunications | 70.06 | |
| 2 | 325000 | 2014-06-01 | 2015-12-31 | system engineer | chennai | f | 8/3/1992 0:00 | 85.0 | cbse | 2010 | 68.2 | cbse | 2 | B.Tech/B.E. | information technology | 70.00 | |
| 3 | 1100000 | 2011-07-01 | 2015-12-31 | senior software engineer | gurgaon | m | 12/5/1989 0:00 | 85.6 | cbse | 2007 | 83.6 | cbse | 1 | B.Tech/B.E. | computer science & engineering | 74.64 | |
| 4 | 200000 | 2014-03-01 | 2015-03-01 | other | other | m | 2/27/1991 0:00 | 78.0 | cbse | 2008 | 76.8 | cbse | 2 | B.Tech/B.E. | electronics & telecommunications | 73.90 | |

## Univariate Analysis

- Univariate Analysis for Continuous Features like Tenure, Salary, College GPA, etc was done by plotting Summary Plots, Histogram, Box Plot and CDF (Cumulative Distribution Function).
- For most of the features it was observed that the data was not normally distributed.
- The Boxplots gave an idea of the presence of a high amount of outliers which led to the further process of Removal of outliers.
- Bar graph plots for Categorical Features were also done.

### Outliers Removal

[106]
```
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
```

[107]
```
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 = data.copy()
```

```
Click here to ask Blackbox to help you code faster
for cols in columns:
    #exctracting upper fence and lower fence
    lowerbound,upperbound = outlier_treatment(df_copy[cols])
```

## Observations Summary:

**Tenure and Salary Skewness**: Both tenure and salary data exhibit significant positive skewness, indicating a concentration of values towards lower tenures and a higher number of respondents with lower salaries. This suggests potential issues with retention and compensation distribution within the dataset.

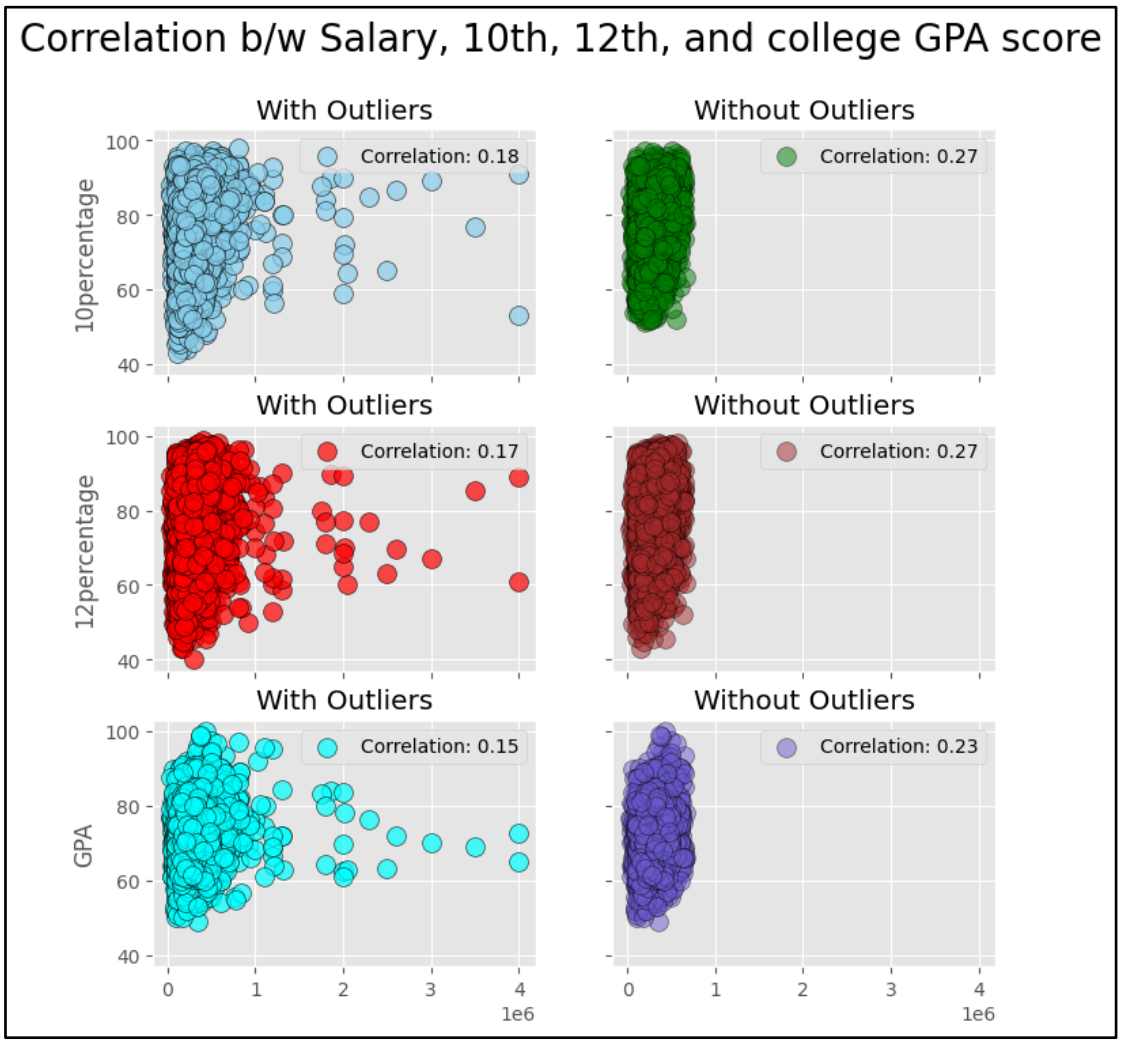
**Student Performance Patterns:** Student performance in both percentage scores and GPA shows a concentration towards higher scores, albeit with slight variations. While there's consistency in GPA distribution, the percentage scores exhibit a broader spread, highlighting potential differences in assessment methods or grading criteria.

**Outliers and Distribution Deviation:** The presence of outliers in both tenure and student GPA data points suggests the existence of extreme cases that may warrant further investigation. Additionally, the deviation from normal distribution in salary and percentage score data indicates non-uniformity in these metrics, potentially reflecting underlying disparities or biases.

**Implications for Decision-Making:** Understanding these patterns is crucial for informed decision-making. Employers may need to address issues related to tenure distribution and compensation fairness, while educators could explore factors contributing to varying student performance metrics to enhance teaching and assessment strategies.

## Bivariate Analysis

- Considering Salary as one element the other attributes were taken into consideration one by one and the analysis was done.
- For some cases like Gender it was observed that the gender attribute was independent of the Salary
- But for some cases like Designation it was seen that Software Engineers had the highest Salary

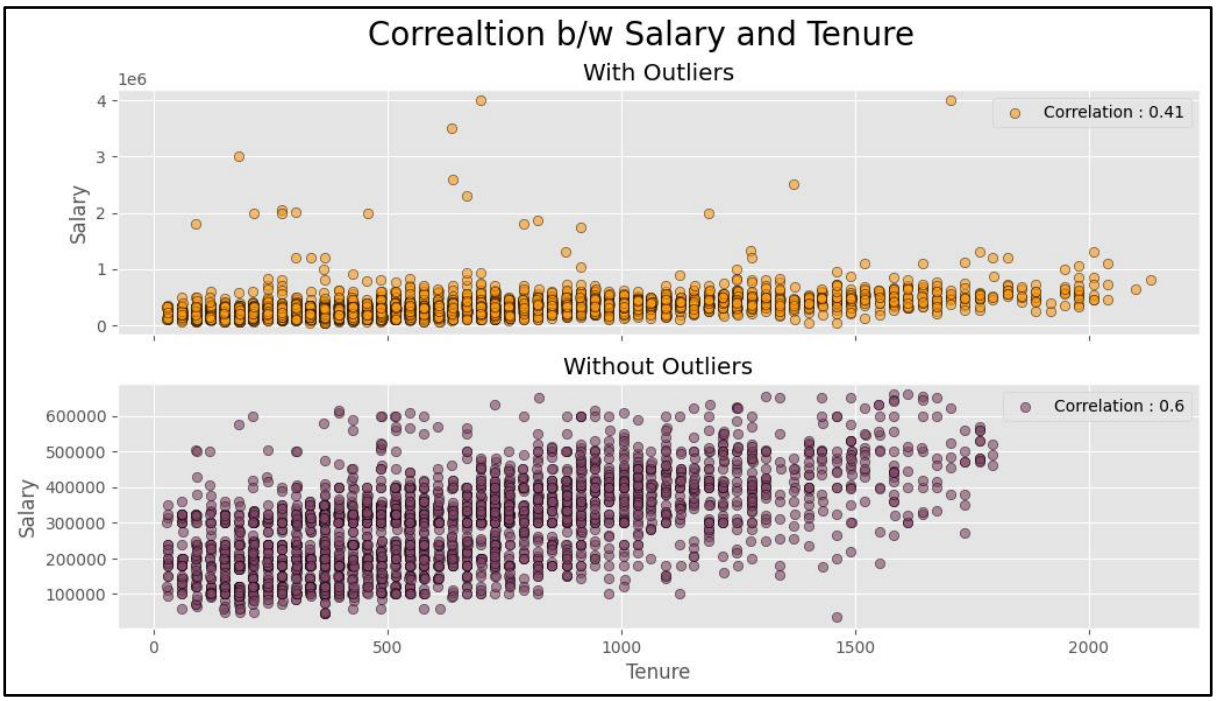



### Observations Summary:

**Absence of Correlation:** No correlation exists between salary and educational scores (10th, 12th, and GPA), suggesting that academic performance does not significantly impact salary.

**Salary Distribution by Designation:** Senior Software Engineers earn the highest salaries but also exhibit the highest salary variability. Only Software Developers and Technical Support Engineers earn below the average salary.

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




**Salary-Tenure Relationship:** After removing outliers, there's a noticeable 50% salary increment with increased tenure, indicating a positive correlation (0.60) between tenure and salary.

**Salary Disparity between College Tiers:** Tier one colleges offer higher salaries compared to tier two colleges, with tier two colleges offering salaries below the overall average.

## Research Outcome

The overarching goal of our project involves analyzing employee data to compare various percentage metrics. One aspect of our analysis includes identifying outliers within the dataset using boxplots. Also, we do many other analyses by keeping Salary as the main element apart from that, we aim to determine which cities have a higher concentration of employees by utilizing countplots, specifically focusing on job locations.

Overall, this analysis emphasizes the complexity of salary determinants in the tech industry and underscore the need for comprehensive compensation policies that account for various factors beyond job title or academic credentials. Further research into the nuanced interplay of these factors could provide deeper insights into effective salary structuring and talent management practices in the tech sector.