INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis on AMEO Dataset

About me

A world where chemical reactions whisper their secrets through data, where algorithms predict the perfect film for a cozy night in, and where nature's intricate patterns hold the key to optimizing processes. This is the world I see, the one **I'm eager to build with the chisel of code and the mortar of machine learning**.

I'm not just a Chemical Engineer in the making, I'm a data alchemist. I see molecules not just as building blocks, but as stories waiting to be told. Stories etched in numbers, patterns whispering with potential. And my tools? **AI, ML, and data science – the incantations with which I translate these whispers into real-world solutions.**

My journey began like many IITian sagas – with a hunger for knowledge. Analytics and machine learning became my language, each completed lesson a stepping stone deeper into their intricate, awe-inspiring worlds. But theory alone wouldn't satiate my curiosity. **I craved the alchemy of practice, the thrill of seeing code morph into tangible outcomes.**
My GitHub is a canvas of collaborative ML-DL projects from movie recommendations to predicting wine quality, each project on my GitHub is a testament to my ability to bridge the gap between theory and tangible outcomes. We weren't just building models; we were building bridges between theory and reality.

A data alchemist ready to turn problems into possibilities, challenges into catalysts for change, fuelled by the power of code and data to unlock the universe's secrets.

GitHub Repo → https://github.com/Harsh1629/AMEO_Dataset

INNOMATICS RESEARCH LABS

1. Objective of the Project

This analysis dives deep into the provided dataset to uncover valuable insights and understand how different factors (features) influence the target variable, **Salary**. We aim to achieve the following:

1. Unveiling the Data: We'll start by thoroughly describing the dataset, including its features and their types (numerical, categorical, etc.). This will give us a clear picture of the information we're working with.

2. Spotting Patterns and Trends: We'll become data detectives, searching for hidden patterns and trends in the features. This could involve looking for correlations, distributions, or any other interesting insights the data holds.

3. Connecting the Dots: We'll focus on the relationships between the independent features (like experience or education) and the target variable, **Salary**. This will help us understand what factors are most likely to influence someone's earnings.

4. Detecting Outliers: Sometimes data throws curveballs in the form of outliers. We'll be on the lookout for any data points that seem significantly different from the rest, as they might require further investigation.

By achieving these goals, we'll gain a deeper understanding of the dataset and the factors that contribute to **Salary**. This knowledge can be valuable for various purposes, such as predicting salaries, making informed hiring decisions, or understanding salary trends within specific industries.

2. Summary Of The Data

The AMEO 2015 dataset, courtesy of Aspiring Minds, delves into the careers of engineering graduates. It features **Salary, Job Titles, and Job Locations** as key outcomes, alongside standardized scores for **cognitive, technical, and personality skills**. Packed with around **40 diverse variables** (continuous and categorical) and **4,000 data points**, it also includes **demographic details** and unique identifiers for each graduate. This rich dataset offers valuable insights into the factors shaping engineering careers.

3. Data Cleaning and Preprocessing

Datatype Conversion

To ensure consistent analysis, we transformed the 'Date of Joining' (DOJ) and 'Date of Leaving' (DOL) fields into Date Time objects. Since the survey occurred in 2015, we assumed respondents marked as "present" in DOL had left by the latest recorded date (2015-12-31). Consequently, "present" values in DOL were replaced with this end date.

Validating 0 or -1

Sr.No.	Column Name	Percentage
1	Domain	6.11
2	ComputerProgramming	21.80
3	ElectronicsAndSemicon	71.45
4	ComputerScience	77.49
5	MechanicalEngg	94.06
6	ElectricalEngg	96.06
7	TelecomEngg	90.06
8	CivilEngg	98.94

Null values represented by 0 or -1 were successfully handled in the columns '10board', '12board', **'Graduation Year'**, **'Job City'**, and **'Domain'**.

Columns with over 70 percent null values were removed from the data set. Columns **Domain** and **Computer Programming** were filled the appropriate values

String Data

Performed operations on columns containing string data in order to remove trailing and leading spaces and also to convert all of them into lower case

Collapsing Categories:

To streamline analysis, only the **top 15 most frequent categories** were retained in specific columns. All remaining categories were grouped as "other." This focuses our analysis on the **dominant characteristics** within the dataset.

4. Feature Engineering

Tenure Calculation:

We've enriched the dataset with a new feature, **'tenure'**, calculated by subtracting 'Date of Leaving' (DOL) from 'Date of Joining' (DOJ). This reveals the **duration of an individual's employment** at the company.

Graduation Year Filtering

To ensure data integrity, rows with **graduation years equal to or after** the 'Date of Joining' were removed. This eliminates inconsistencies where graduation appears to occur after employment begins.

Cumulative Distributive Function(CDF)

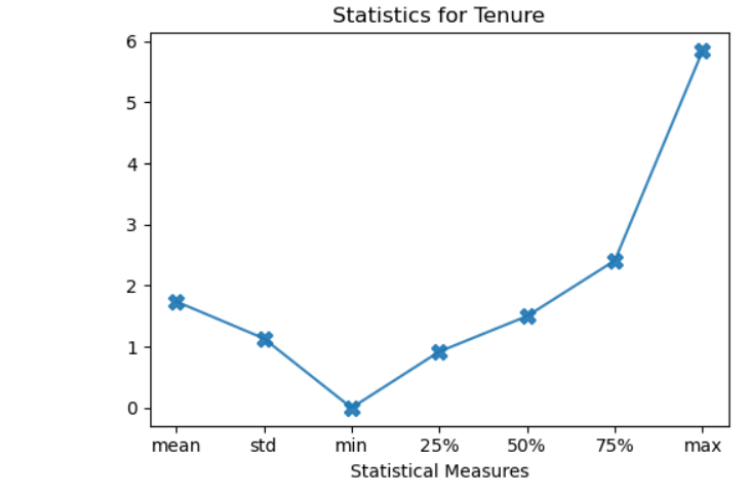
To unlock **deeper insights** into data distribution, we crafted a custom **Cumulative Distribution Function (CDF) calculator**. This tool reveals the **cumulative probability** of values occurring within a variable, empowering us to make **informed decisions** based on the data's statistical landscape.

5. Exploratory Data Analysis:

Univariate Analysis

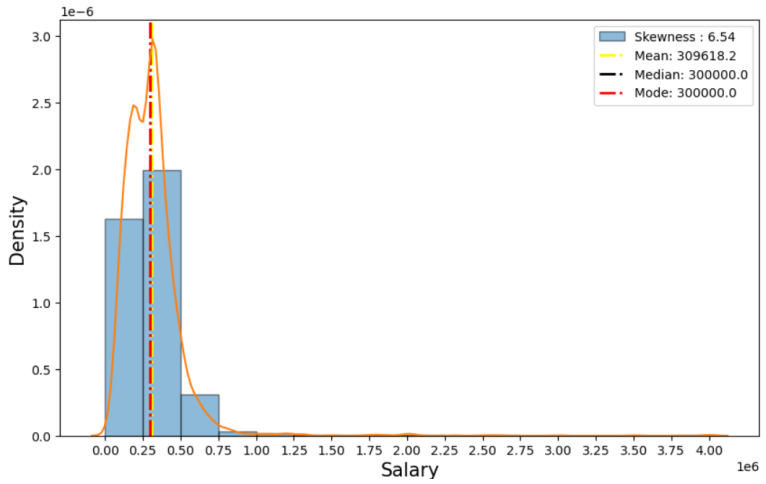
1.1 Tenure:

Dive into tenure patterns with our summary plots! They reveal a **4-year experience range**, with histograms painting a picture of a **positively skewed distribution**. The median tenure of **1.5 years** indicates a dynamic workforce, while outliers hint at longer tenures. Box plots further accentuate these outliers, showcasing the spread of experiences. Notably, the **CDF confirms a non-normal distribution**, adding depth to our understanding of workforce dynamics.



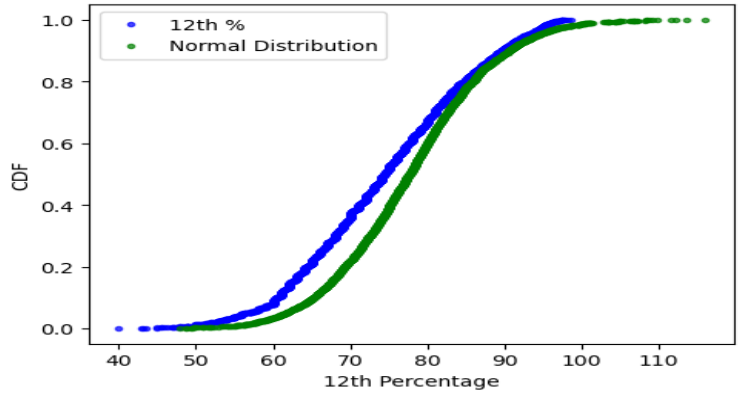
1.2 Salary:

Our data delves into the world of salaries! Summary plots unveil **substantial variation**, while histograms paint a picture of a **strong positive skew**, suggesting a departure from the familiar bell curve. Box plots further emphasize this trend, highlighting a **concentration of high salaries**. Finally, the CDF reinforces this notion, confirming the data's **significant skew** and its deviation from a normal distribution. This paints a fascinating picture of salary trends within the dataset.



1.3 12th Percentage:

Analysis reveals a **bell curve with a twist** for exam scores. Around half scored **78% or below**, showcasing a **sparsity of very low scores**. The majority (69% - 84%) clustered around 70%, peaking at that point. However, an **outlier with an extraordinarily low score** stands out. Notably, the CDF confirms a **departure from normality**, indicating an interesting distribution pattern.

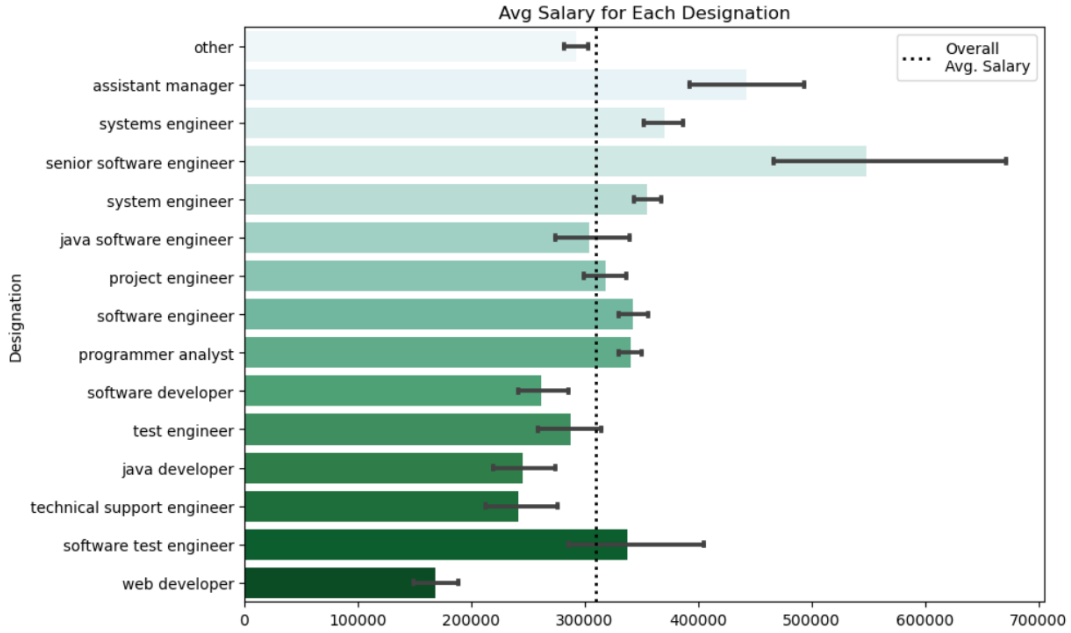


5. Exploratory Data Analysis:

Bivariate Analysis

1.1 Designations and Salary

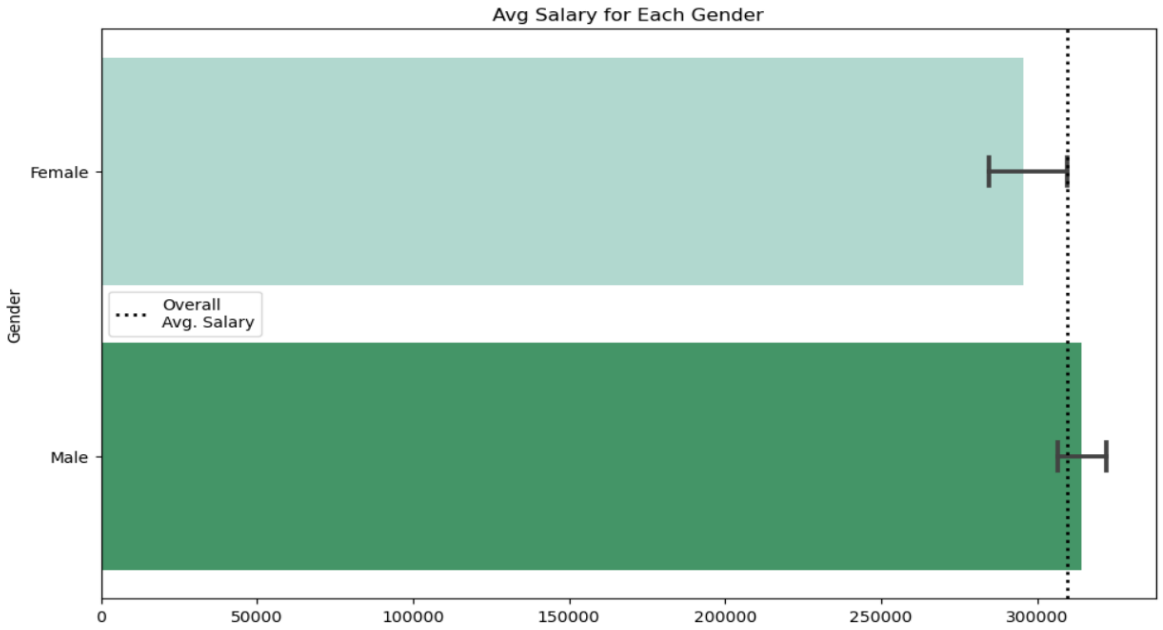
Senior Software Engineers have the highest salary among all designations. Software developer, test engineers, and Java developers have salaries less than average.



Conclusion : Senior Software Engineer has the highest salary.

1.2 Gender and Salary:

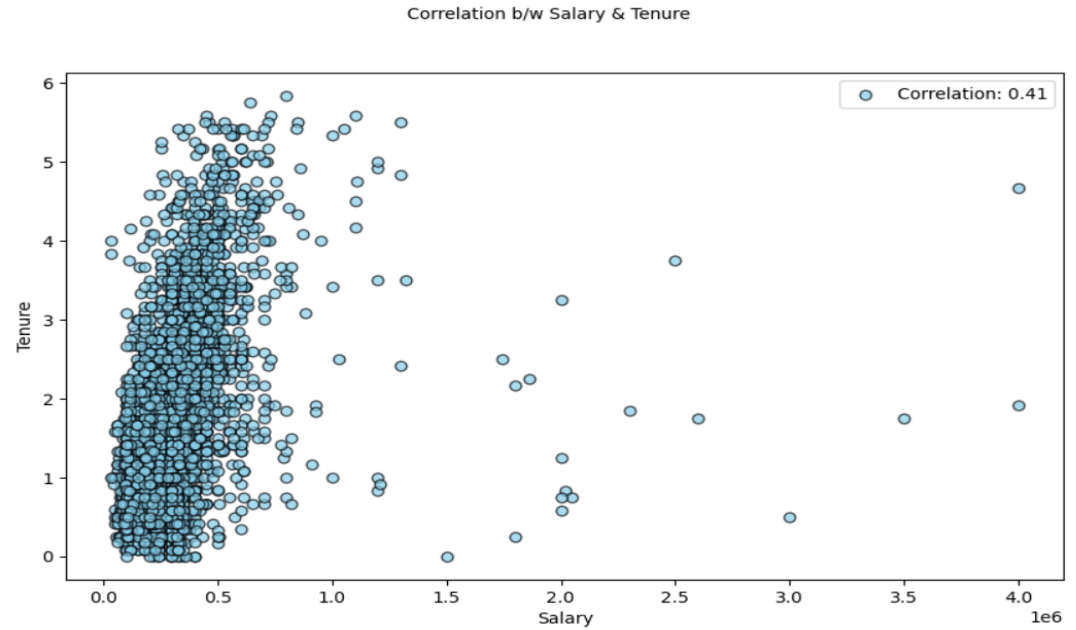
Both male and female salaries are approximately equal on average, suggesting no gender bias overall, though females tend to receive salaries below the overall average



Conclusion : Overall we can say that there was no biasness based on gender in terms of salary

1.3 Salary and Tenure:

There's a positive correlation between tenure and salary, i.e. salary increases with more years of work experience.



Conclusion : We can say that there exists some sort of relationship between salary and tenure(years of experience)

6. Research Outcome:

"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."

Our analysis dives deep into salary distribution across different job roles. We start by grouping the data and calculating the **mean and standard deviation** of salaries for each designation. This reveals fascinating insights!
Software Engineers reign supreme with the **highest mean salary and standard deviation**, suggesting both **higher pay and more variability** compared to Programmer Analysts and Associate Engineers.
Next, we delve deeper with **one-sample t-tests**:

•**Programmer Analysts and Software Engineers:** The test results reject the null hypothesis, meaning their salaries **significantly differ** from the expected range. Wow!
•**Hardware Engineers and Associate Engineers:** The test fails to reject the null hypothesis, implying their salaries **might not significantly deviate** from expectations.
Overall, this analysis paints a clear picture of **salary distribution across job roles** and helps us understand the **significance of salary differences** within the dataset. It's a valuable tool for identifying trends and potential pay gaps!

Designation	t_critical	p_value	Result
Programmer Analyst	12.30	3.51934e-17	There is sufficent evidence to reject the Null Hypothesis
Software Engineer	10.17	8.14734e-21	There is sufficent evidence to reject the Null Hypothesis
Hardware Engineer	NaN	NaN	There is not enough evidence to reject the Null Hypothesis
Associate Engineer	0.61	3.01696e-01	There is not enough evidence to reject the Null Hypothesis

Observations

Test	Value
t_value	3.3686709047035115
t_critical	1.9842169515086827
p_value	0.0007553155406618828

- As the result of the hypothesis testing we see that the claim is false.
- For this claim Null Hupothesis fails.
- The **t_critical** and probability value i.e. **p_value** claiming it as wrong.

THANK YOU

INNOMATICS RESEARCH LABS