INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis (EDA) on AMCET 2015

About me

- **P.Bhargavi(M.com)**
- **My enthusiasm for data science lies in its ability to drive informed decision-making and foster innovation across industries, ultimately making a tangible impact on business success and societal progress.**
- **linkedin : - www.linkedin.com/in/panchamurthy-bhargavi**
- **github : - https://github.com/bhargavipanchamurthy**

INNOMATICS RESEARCH LABS

Problem of Statement:-

AMEO dataset provides anonymized bio data information along with their respective skill scores and employment outcome information. As part of the challenge, the aim is to find the relationship between salary vs remaining features to know which are more affecting the salary column. Based on the analysis finding the insights.

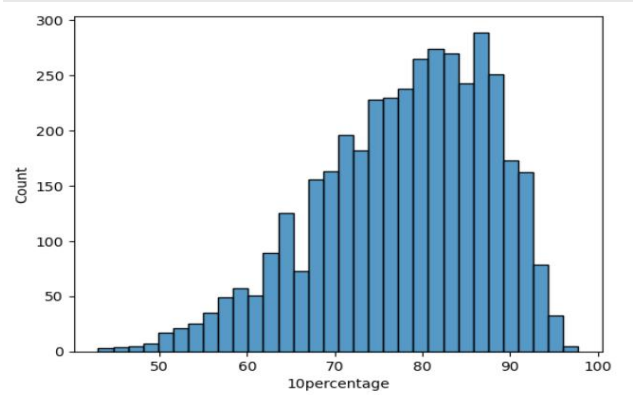
Brief About Data

- Most of the engineering students in the data are males.
- Majority of the students in 10th and 12th standard are from state board followed by CBSE board.
- Tier 2 colleges dominate the data just like they dominate in the real world meaning that only a few colleges are tier 1 colleges
- B Tech/ B.E clearly dominates other degrees in terms of frequency of occurrence
- The specialization has been manually categorized into 5 buckets with majority of students specializing in computer science, followed by Electronics/Electrical, IT, Mechanical and others.
- Majority of the data belongs to students who went to a college in Uttar Pradesh Data was checked for all the assumptions of linear regression and all of them were duly satisfied!

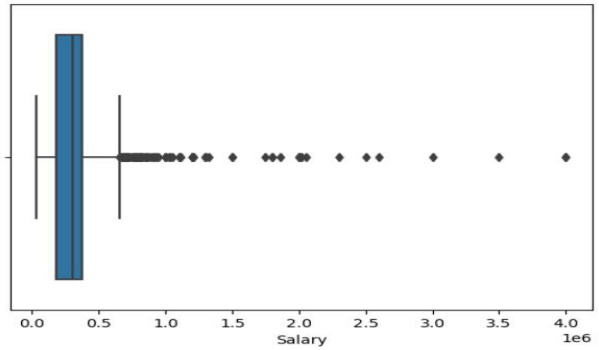
Data Cleaning and Manipulation

- Dropping College and college city ID' -- they are same (around 1350 unique id’s)
- Dropping columns which are unnecessary or may not be known prior to receiving a job offer
- Changing the data type as required
- On observing the data, some graduation years were mentioned as 0, replacing graduation year 0 with the modal graduation year
- we have many boards so i have converted these multi class to binary as cbse or state board
- We have many specializations . I have to converted these many classes to 6 specializations.

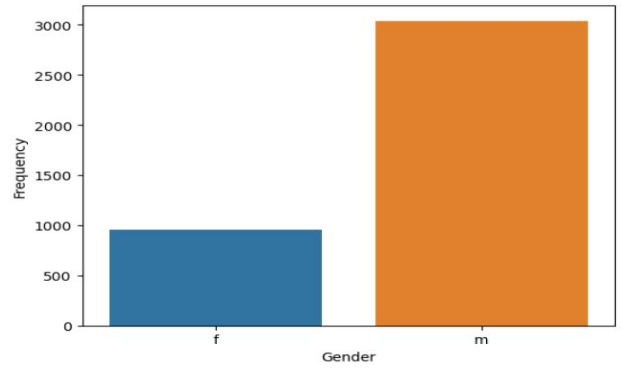
Univariate Analysis



This histogram visualizes the distribution of values in the "10percentage" column from a DataFrame df. The x-axis represents the "10percentage" values, while the height of each bar indicates the frequency of occurrence of those values in the dataset. The plot provides insights into the spread and concentration of data points for the variable "10percentage".



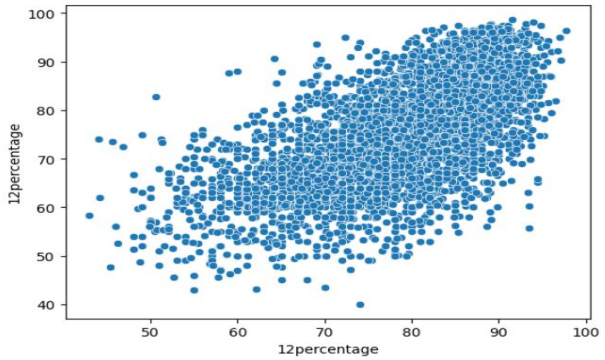
This graph, created using seaborn's boxplot function, visualizes the distribution of salaries from a DataFrame df. Each box in the plot represents the interquartile range (IQR) of the salary data, with the median salary marked by a line inside the box. The whiskers extend to show the range of salaries within 1.5 times the IQR. Any outliers beyond this range are plotted individually. By labeling the x-axis as "Salary", the plot is appropriately annotated for clarity.



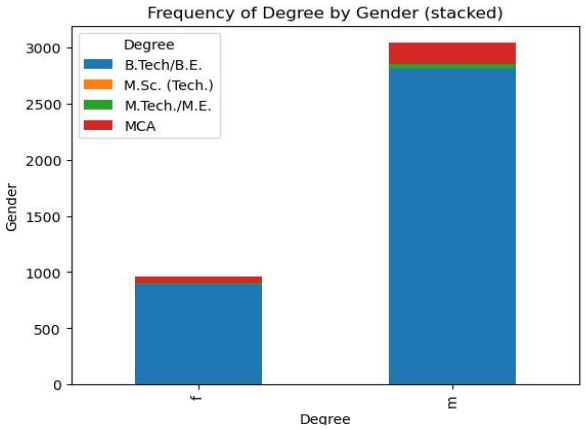
This countplot visualizes the frequency distribution of gender categories from a DataFrame df. Each bar represents the count of occurrences for each gender category. The x-axis is labeled as "Gender" to denote the variable being plotted, while the y-axis represents the frequency of occurrences. This graph provides a clear comparison of the number of data points for each gender category, facilitating quick insights into the distribution of gender within the dataset.

INNOMATICS RESEARCH LABS

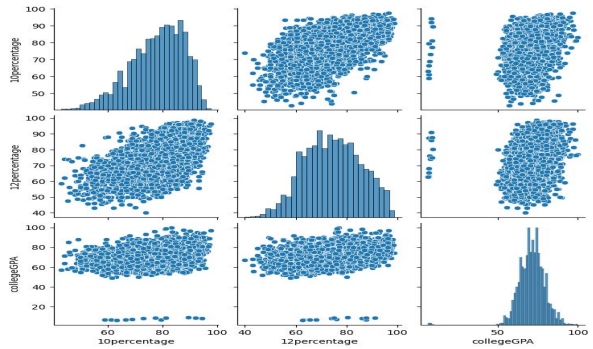
Bi-Variate Analysis :



This scatterplot visualizes the relationship between two variables, "10percentage" and "12percentage", from the DataFrame df. Each point on the plot represents an individual data entry, with the x-axis corresponding to the "10percentage" values and the y-axis corresponding to the "12percentage" values. By examining the distribution of points, one can assess any patterns or trends between these two variables. The x-axis and y-axis are appropriately labeled as "10percentage" and "12percentage" respectively, providing clarity to the plot.



This graph illustrates the frequency distribution of degrees across genders, using a stacked bar chart. The data is organized by gender on the y-axis and degree on the x-axis. Each bar is segmented to represent the proportion of each degree category within each gender group



The pairplot visualizes the pairwise relationships between the variables "10percentage", "12percentage", and "collegeGPA" from the DataFrame df. Each scatter plot in the grid represents the relationship between two variables, while the diagonal shows the distribution of each individual variable

INNOMATICS RESEARCH LABS

Bonus Question :

This code generates a plot that focuses on individuals with a specialization in "Computer Science & Engineering" and certain job designations ("Programmer Analyst", "Software Engineer", "Associate Engineer") who started working right after graduation. It filters the DataFrame to include only relevant data points based on specialization, job designation, and year of joining (DOJ) matching graduation year

Insights

•Out of all the given variables, it appears that quant skills are the most important in terms of Pearson's correlation coefficient. This means higher the score of candidate in quant, higher are the chances of having better salary prospects!

•10th percentage is positively correlated with 12th percentage

•Scores in AMCAT's Logical aptitude test is having a positive correlation (moderate) with English and Quant tests.

•In the AMCAT's personality test, it appears that openness to experience is having a 44% positive correlation with agreeableness in the candidates.

•Mumbai, Bangalore, Pune and Hyderabad have the highest entry level salary while it is surprising to note that cities like Delhi and Kolkata have the lowest entry level salaries for engineering graduates. This was a shocker since these are the metropolitans!

•Mean salary of M.tech / M.E and B.tech / B.E students appear to be slightly higher than other degrees.

•Another interesting insight is that lower salary quartile (first quartile) of civil engineering students is higher than any other specialization. So do advice your juniors to choose civil engineering !

THANK YOU

INNOMATICS RESEARCH LABS