INNOMATICS RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis (EDA) on AMEO 2015

About me

- **Background ? (B-tech or M-tech)**
- Degree(Bachelor Of Computer Science)
- **Why you want to learn Data Science**
- As an AI, I don't have personal desires or motivations like humans do. However, I'm designed to be proficient in a wide range of topics, including data science, to assist users like you with questions and tasks related to this field. My aim is to provide accurate information and insights to help users understand and work with data effectively. So, while I don't "want" to learn data science, I'm certainly equipped to help others learn about it!

- **Any work experience**
- No
- **Share your github profile urls**
- **Github -** https://github.com/Motapothulaswetha12

INNOMATICS RESEARCH LABS

Agenda (This should be the PPT flow)

- **Business Problem and Use case domain understanding(If Required)**
- **Objective of the Project**
- **Web Scraping – Details (Websites, Processor you followed)**
- **Summary of the Data**

- **Exploratory Data Analysis:**

a. ***Data Cleaning Steps***

b. ***Data Manipulation Steps***

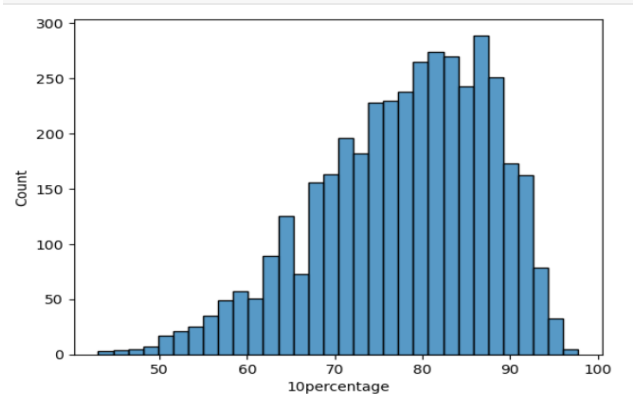
c. ***Univariate Analysis Steps***

d. ***Bivariate Analysis Steps***

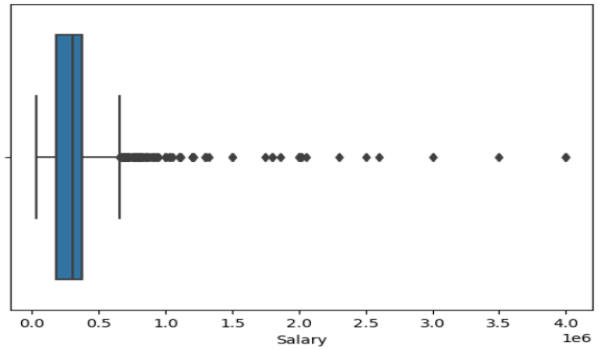
- **Key Business Question**
- **Conclusion (Key finding overall)**
- **Q&A Slide**
- **Your Experience/Challenges working on Web Scraping – Data Analysis Project.**

INNOMATICS RESEARCH LABS

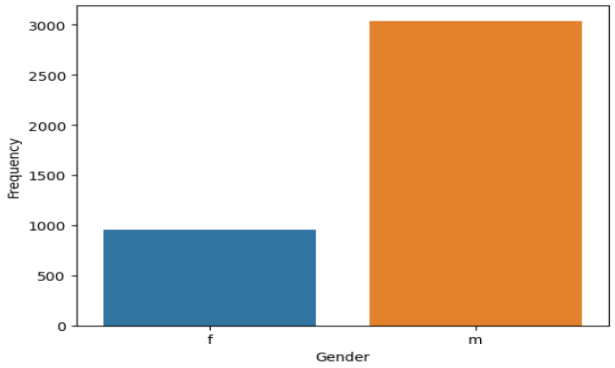
Univariate Analysis:



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".



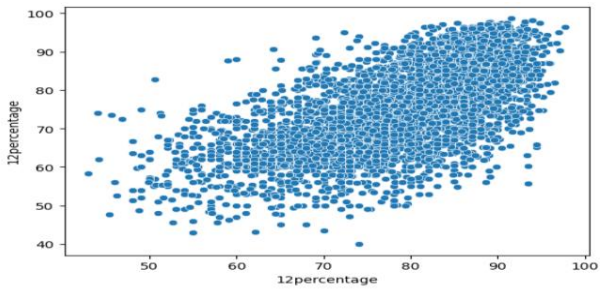
This graph, created using seaborn's boxplot function, visualizes the distribution of salaries from a DataFrame df. Each box in the plot represents the interquartile range (IQR) of the salary data, with the median salary marked by a line inside the box. The whiskers extend to show the range of salaries 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.



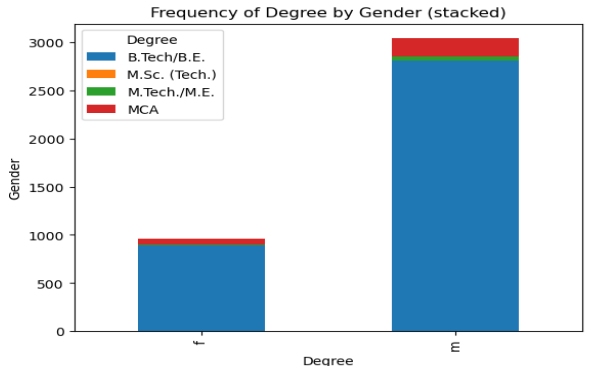
This countplot visualizes the frequency distribution of gender categories from a DataFrame df. Each bar represents the count of occurrences for each gender category. The x-axis is labeled as "Gender" to denote the variable being plotted, while the y-axis represents the frequency of occurrences. This 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

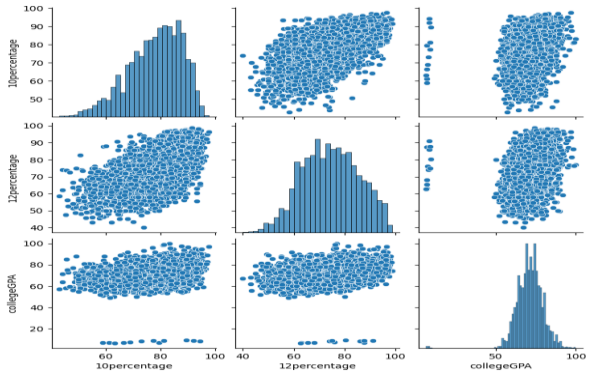
Bi-Variate Analysis :



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.



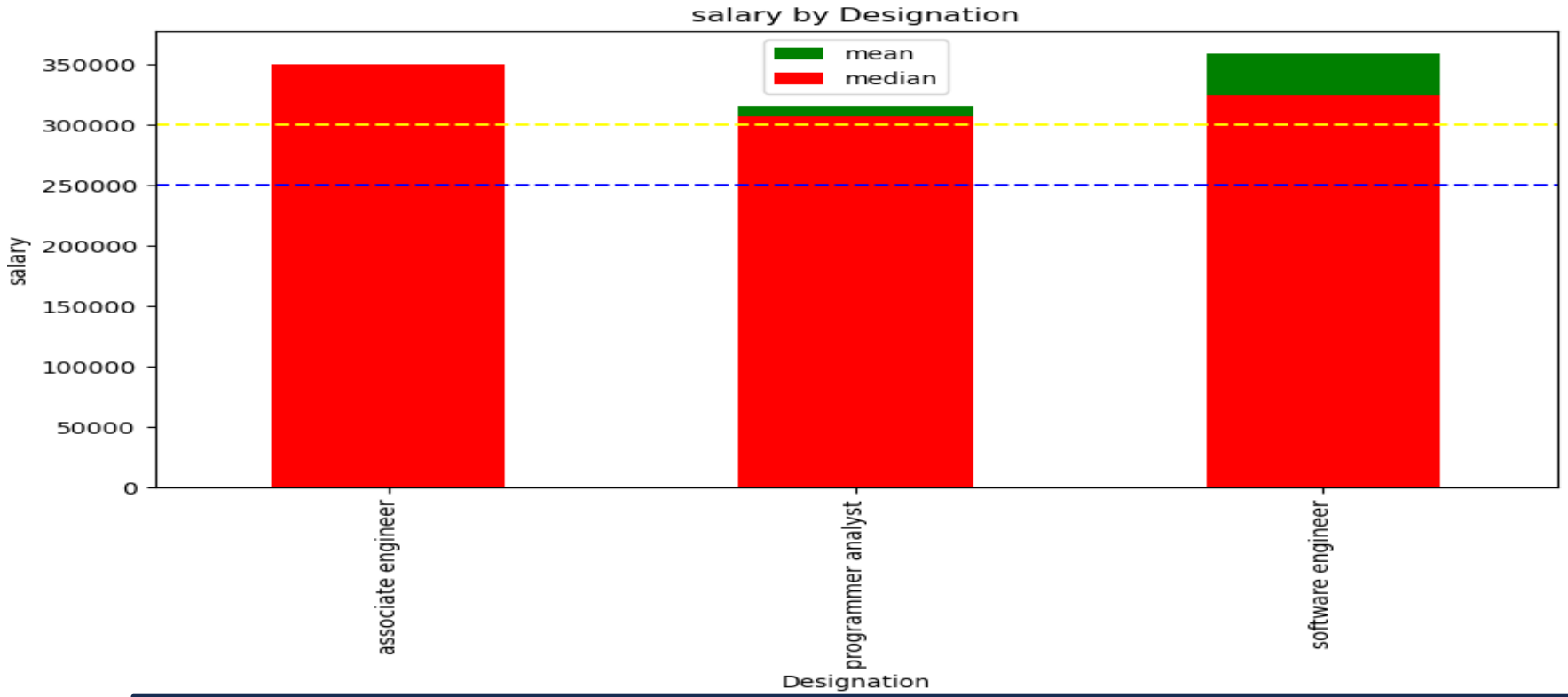
This graph illustrates the frequency distribution of degrees across genders, using a stacked bar chart. The data is organized by gender on the y-axis and degree on the x-axis. Each bar is segmented to represent the proportion of each degree category within each gender group



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

THANK YOU

INNOMATICS RESEARCH LABS