INNOMATICS® RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis (EDA) on AMEO 2015

By

Yarramsetti . Mohana Manikanteswari

About me

- **Background**?
  B.Tech(Civil Engineering)

- **Why you want to learn Data Science?**
  Learning data science enables me to uncover insights, predict trends, and make informed decisions in today's data-driven world. It's about turning raw data into valuable knowledge, whether to understand consumer behaviour, optimize business processes, or advance scientific research. It's the key to unlocking the potential of data and shaping the future.

- **Any Work Experience?**
  Fresher (No Experience)

- **Share your LinkedIn, GitHub Profile URL's**

  LinkedIn: https://www.linkedin.com/in/yarramsetti-mohanamanikanteswari-50b032236
  GitHub : https://github.com/MohanaYarramsetti12

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***

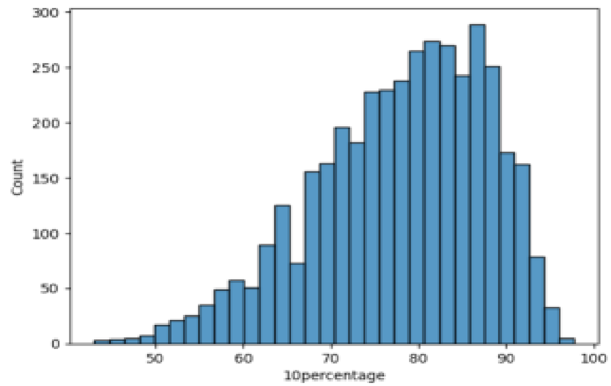
c. ***Univariate Analysis Steps***

d. ***Bivariate Analysis Steps***

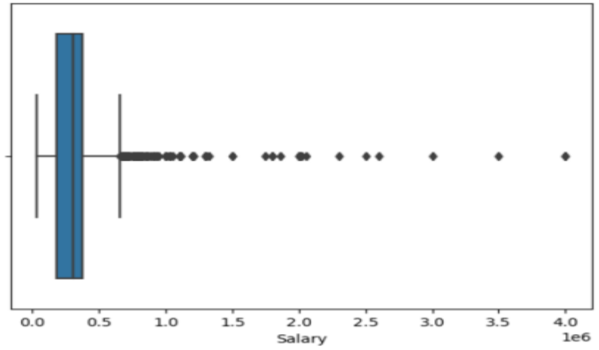
- **Key Business Question**
- **Conclusion (Key finding overall)**
- **Q&A Slide**
- **Your Experience/Challenges working on Web Scraping – Data Analysis Project.**

INNOMATICS RESEARCH LABS

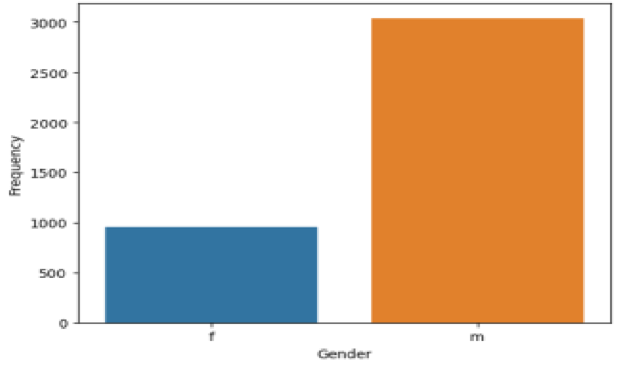
Univariate Analysis:



This histogram visualizes the distribution of values in the "10percentage" column from a Data Frame 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".

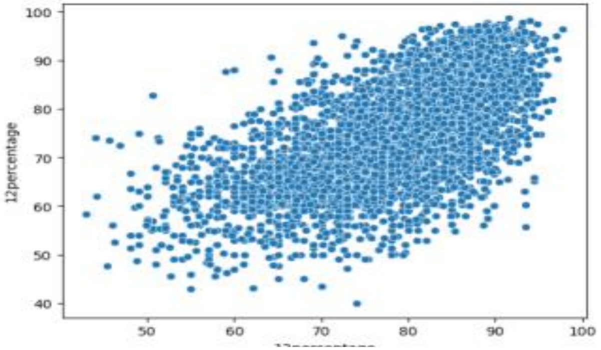


This graph, created using seaborn's boxplot function, visualizes the distribution of salaries from a Data Frame df. Each box in the plot represents the interquartile range (IQR) of the salary data, with the median salary marked by a line inside the box. The whiskers extend to show the range of salaries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

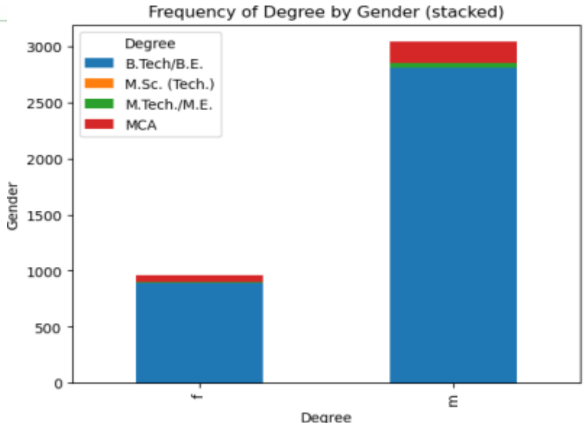


This countplot visualizes the frequency distribution of gender categories from a Data Frame df. Each bar represents the count of occurrences for each gender category. The x-axis is labelled as "Gender" to denote the variable being plotted, while the y-axis represents the frequency of occurrences. This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

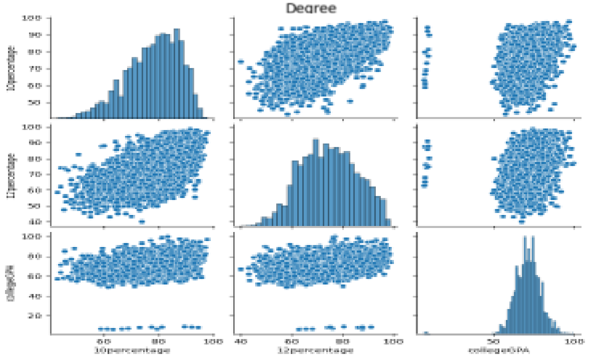
Bi-Variate Analysis:



This scatterplot visualizes the relationship between two variables, "10percentage" and "12percentage", from the Data Frame 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 labelled as "10percentage" and "12percentage" respectively, providing clarity to the plot.

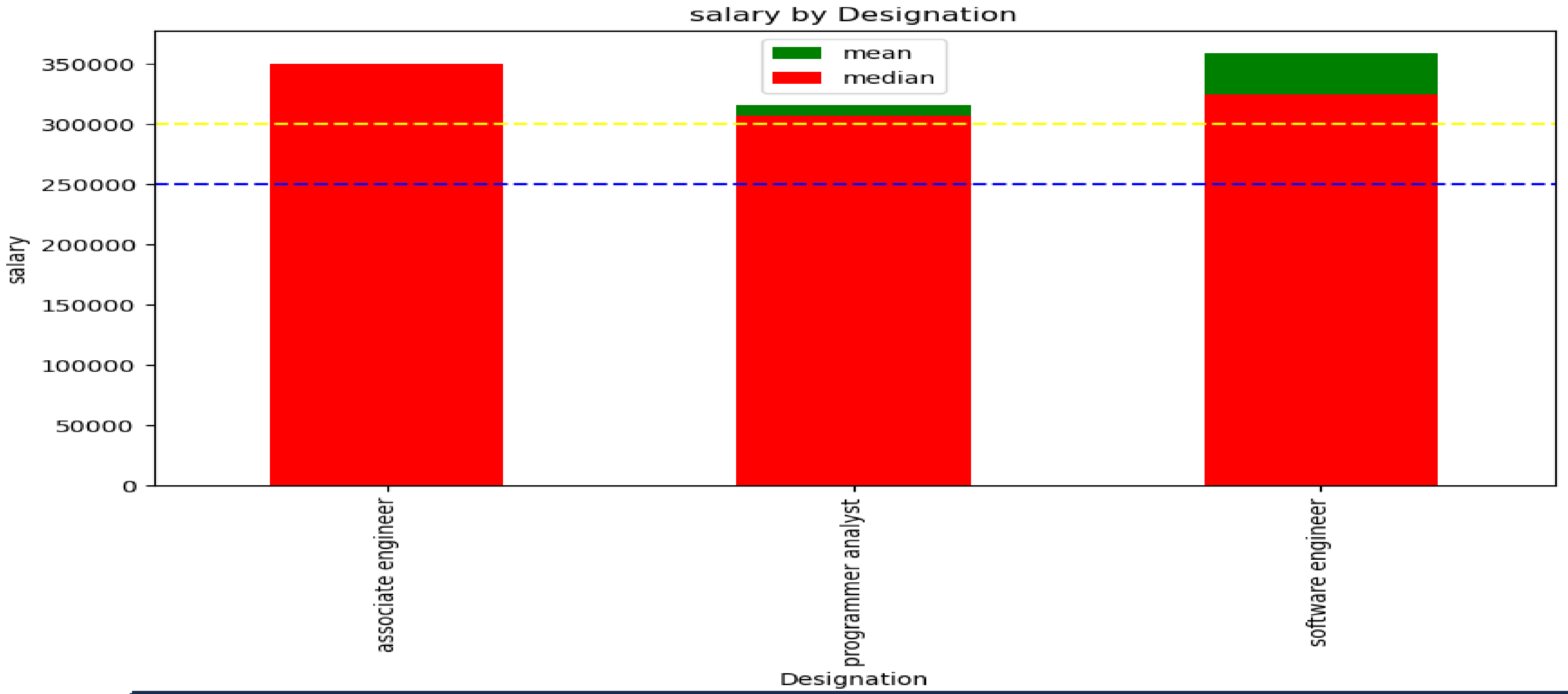


This graph illustrates the frequency distribution of degrees across genders, using a stacked bar chart. The data is organized by gender on the y-axis and degree on the x-axis. Each bar is segmented to represent the proportion of each degree category within each gender group



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

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 Data Frame to include only relevant data points based on specialization, job designation, and year of joining (DOJ) matching graduation year

THANK YOU

INNOMATICS
RESEARCH LABS