INNOMATİCS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Exploratory Data Analysis

About me

Background ?

- My name is Raman Kumar currently enrolled in the Bachelor of Technology (B.Tech) program at Raja Balwant Singh Engineering Technical Campus in Agra. I'm in 8th semester

Why you want to learn Data Science

Certain points for I wants to learn Data Science.

- Interest in solving complex problems across various industries.
- High demand for data scientists leading to numerous job opportunities.
- Desire to make a significant impact by informing decision-making processes and driving innovation.
- Curiosity about how data can be used to understand and predict phenomena.
- Desire to gain a competitive advantage in the job market by acquiring valuable skills.

Any work experience

I have no industry Experience , currently I'm doing bachelor of technology in 8th semester

linkedin :-linkedin.com/in/ramank97

github

INNOMATICS RESEARCH LABS

Objective of the Project:-

The objective of this project is to conduct a comprehensive exploratory data analysis (EDA) on a given dataset to uncover insights and patterns. Starting with data import and exploration, the analysis delves into the dataset's structure and content through examination of its head, shape, and basic statistics. It then proceeds to univariate analysis, exploring the distribution of numerical and categorical variables using visualizations such as PDFs, histograms, and boxplots. Bivariate analysis follows, investigating relationships between variables through scatter plots, hexbin plots

Summary of the Data:-

The dataset offers a detailed snapshot of individuals' educational backgrounds, employment status, and personal attributes, including various numerical, categorical, and temporal features. It encompasses data on salary, academic performance, domain expertise, and scores in subjects like English and quantitative aptitude. Categorical attributes cover job designation, gender, degree, specialization, and college location. Temporal features capture dates of joining and birth, providing insights into career timelines. Identifier attributes uniquely identify individuals, colleges, and cities

Exploratory Data Analysis:-

Data cleaning:-

- Address missing values by imputing with median for numerical columns and mode for categorical columns.
- Remove duplicate records to ensure data integrity and eliminate redundancy.
- Standardize column names by converting them to lowercase and replacing spaces with underscores.
- Convert columns to appropriate data types, such as datetime for date-related features.
- Detect and handle outliers using statistical methods like z-score or IQR.
- Encode categorical variables into numerical format using techniques like one-hot encoding.
- Scale numerical features to a similar range using Min-Max scaling or standardization.
- Perform feature engineering to create new features or transform existing ones for better model performance.
- Address class imbalance issues in classification tasks using oversampling or undersampling techniques.

Data manipulation:-

- Filter the dataset to include only relevant rows or columns based on the analysis objectives.
- Handle missing values by imputing them with appropriate measures or removing them.
- Convert categorical variables into numerical format using encoding techniques like one-hot encoding.
- Standardize numerical features to ensure they are on the same scale.

- Convert columns to appropriate data types, such as datetime for date-related features.
- Detect and handle outliers using statistical methods like z-score or IQR.
- Encode categorical variables into numerical format using techniques like one-hot encoding.
- Scale numerical features to a similar range using Min-Max scaling or standardization.
- Perform feature engineering to create new features or transform existing ones for better model performance.
- Address class imbalance issues in classification tasks using oversampling or undersampling techniques.
- Conduct a final review of the cleaned dataset to ensure readiness for analysis or modeling.

Univariate Analysis Steps:-

- Explore the distribution of each individual variable separately without considering their relationships with other variables.
- Utilize statistical measures such as mean, median, mode, standard deviation, and range to understand the central tendency and dispersion of numerical variables.
- Generate visualizations like histograms, density plots, box plots, and violin plots to visualize the distribution and identify patterns or outliers in numerical data.
- For categorical variables, examine the frequency distribution using count plots or bar plots to understand the distribution of different categories.
- Analyze the presence of missing values and handle them appropriately through imputation or deletion depending on the extent of missingness and the nature of the data.

- Consider the skewness and kurtosis of numerical variables to assess their deviation from a normal distribution and decide on appropriate transformations if necessary

Bivariate analysis:-

- Use sctter plots to visualize the realationship between two numerical variables, allowing for the identification of trends, patterns, or correlations.
- Investigate the relationships between pairs of variables to understand how they interact with each other.
- Employ hexbin plots as an alternative to scatter plots when dealing with a large number of data points, providing a clearer representation of density and correlation.
- Utilize pair plots (also known as scatterplot matrices) to visualize pairwise relationships across multiple numerical variables simultaneously, often accompanied by histograms or density plots along the diagonal for univariate analysis.
- Explore the association between categorical and numerical variables using swarm plots, box plots, or bar plots, allowing for the comparison of numerical values across different categories.
- Identify patterns between two categorical variables using stacked bar plots or contingency tables, revealing the distribution of one variable within the categories of another variable.

Conclusion:-

This project provided a comprehensive analysis of the dataset, shedding light on various aspects such as salary distribution among fresh graduates and the relationship between gender and specialization preferences. Through rigorous data exploration, we identified patterns, outliers, and potential correlations, enabling us to draw meaningful insights**.**

THANK YOU

INNOMATICS RESEARCH LABS