**Course – Data Analytics Open Elective**

|  |  |
| --- | --- |
| **UID** | 2021300126, 2021300125 |
| **Name** | Pranay Singhvi, Yoshita Singh |
| **Class and Batch** | TE Computer Engineering - Batch A |
| **Date** | 23-01-2024 |
| **Lab #** | 1 |
| **Aim** | Perform EDA And Plot Graphs |
| **Objective** | Perform EDA such as number of data samples, number of features, number of classes, number of data samples per class, removing missing values, conversion to numbers, using the seaborn library to plot different graphs LIKE Histogram, Pie Chart, Bar Plot, Parallel coordinates, etc. |
| **Theory** | Exploratory Data Analysis (EDA) is a crucial process that involves examining and summarizing data to reveal patterns and characteristics. This is achieved through various techniques such as data visualization, summary statistics, and the exploration of univariate, bivariate, and multivariate analyses.  The importance of EDA in data science lies in providing analysts with a profound understanding of the dataset. It enables the identification of trends, outliers, and anomalies, guiding informed decisions in data preprocessing, model selection, and hypothesis formulation. EDA ensures that data is well-prepared for further analysis, leading to more accurate and meaningful results.  Different types of EDA include univariate analysis, focusing on one variable at a time; bivariate analysis, studying relationships between two variables; and multivariate analysis, considering multiple variables simultaneously to explore complex interactions. Techniques like data visualization, employing charts and graphs, and summary statistics, calculating measures like mean and standard deviation, contribute to a comprehensive EDA approach.  Specifically, visual tools like parallel sets, depicting relationships between categories through parallelograms, and parallel coordinates, presenting observations across variables, aid in comparisons and distribution analyses. These techniques enhance the exploration of data patterns, contributing to a more thorough understanding of datasets for effective decision-making in data science projects. |
| **Data Set** | <https://www.kaggle.com/datasets/trolukovich/nutritional-values-for-common-foods-and-products> |
| **Colab File Link** | <https://colab.research.google.com/drive/1n3kQIQqfKdOaQ9YUpYi53MVSM3f7iFVM?usp=sharing> |
| **Purpose** | Nutritional Research: Analyzing relationships between different nutrients and exploring trends in food composition for scientific and academic research. |
| **Code (Part 1)** | **Repair and Normalize Data with Pandas** 1)Dataset of Nutritional Value of common food products    2) Except for 'name', all variables appear to have numeric values. However, as we see in the datatypes table, most of the columns are of type character. This happens because in these variables, together with the value, they also show the unit in which it is expressed (for example, it shows 13g of protein for x food. That 'g' (of grams) is a character, so the function that we use to load the data assumes it is of type character. The idea is to pass those variables from 'character' to 'numeric', and the way we're going to do that is by first finding those unit characters (g, mg, mcg, UI) and removing them, and then just converting the type from data to numeric.    3) Checking for Null Values and Duplicate Values   1. Assign each column name its unit of measure. As we eliminate all the units of measurement (UM) in which the values ​​are expressed (point B), we are losing valuable information. Since within the same column, the UM does not vary, we can add that information directly in the column name.      1. Dropping Columns with same values for all rows 2. Normalizing data with numerical values using the Min Max Scaler   A screen shot of a computer screen  Description automatically generated |
| **Code(Part 2)** | **Data Visualization and EDA with Pandas and Seaborn**   1. How saturated fatty acids and fibers are in some selected food?   A graph with blue and orange bars  Description automatically generated  From above graph we can conclude that all selected food has more fiber compared to saturated fatty acid.   1. Histogram of different attributes   Histograms provide a visual representation of the distribution of each feature. Understanding the distribution of nutritional components helps in identifying patterns, potential outliers, and the overall shape of the data.  A group of blue and white graphs  Description automatically generated   1. Correlation Features   Correlation features reveal relationships between variables, aiding insights. High correlation implies association; low correlation suggests independence.  **A screenshot of a computer screen  Description automatically generated**   1. Relation between Calories and Fats(Scatter Plot)   A graph of blue dots  Description automatically generated  Scatter plots visualize data points' distribution, indicating patterns or trends. X and Y axes display variables, revealing relationships graphically. |
| **Code(Part 3)** | **Working on the Titanic dataset for parallel coordinates and parallel sets:**    **A close-up of a grid  Description automatically generated**  **A black and white screen with white text  Description automatically generated**  **A blue net with many strings  Description automatically generated with medium confidence**  **A blue and white wire frame  Description automatically generated with medium confidence**  From above graph we get to know that there were 549 passengers didn’t survive the disaster.  This visualization lets you explore and analyze relationships between categorical variables within the dataset, facilitating pattern recognition and insights generation that inform data-driven decision-making. |
| **Conclusion** | During the Exploratory Data Analysis (EDA), I investigated essential dataset attributes, like sample count, features, and classes. Ensuring data integrity involved handling missing values and converting categories to numerical formats. Seaborn visualizations, such as histograms and pie charts, offered insights into feature distributions and class compositions. Analyzing class balances and feature interactions aids subsequent modeling efforts. This EDA provides a holistic grasp of the dataset, tackling data quality, exposing patterns, and offering visual insights for informed decision-making in subsequent analyses. |