**Week 2**

Sandish Khoju Shrestha

Department of IT, Westcliff University

Presidential Graduate School

TECH 405: Artificial Neural Network & Deep Learning

Professor Acharya

November 3, 2024

**Introduction**

The code gives us t full processing and visualization of dataset. This is accomplished by starting with some imports for necessary packages as follows: pandas, NumPy and matplotlib, seaborn.It then imports ,cleaned dataset californi\_a\_housing\_test.csv and gives basic summary information of it. By importing essential libraries like pandas, NumPy, matplotlib, and seaborn. Then, it loads the dataset california\_housing\_test.csv' and explores its basic information. This code paves all the initial rows of the data, gives summarisation, and descriptive statistics. It also verifies if the data is missing, or data type and column names. In data visualization, histograms are used to determine the distribution of numerical features where the code acts as follows. For categorical data it produces count plots that show frequency of each categorical variable. A correlation matrix is used to determine whether or not there is a relationship between numerical columns. Laststly,aping derives pair plot and boxplot in order to gain relational check and detect outlier in analytical numeral.

**Findings**

1. Basic Structure of the DatasetThe dataset is composed of several features which include both numerical and categorical features. The first five rows give an insight of each column’s value orientation.
2. Data Types : There are numerical and categorical features in the data frame, where numerical ones are such as ‘median\_income’ and ‘total\_rooms’, and categorical such as ‘ocean\_proximity’. We need to make sure we know which data types work with which kind of visualizations, so come up with suitable technique.
3. Missing Values: Checking of missing values gets established to check the level of completeness of the dataset. Preprocessing activities before analysis requires determination of columns containing missing values.
4. Statistical Summary: Meaningful numerical data contains information about the values of the data and its dispersion by looking at the mean, median, minimum, maximum, and quartiles of the numerical columns. For example, ‘median\_house\_value’ may fluctuate and show a wide range which may depict the variation of price in housing.
5. Distribution of Numerical Features: The histogram plots of the numerical columns bring out the number of data points with regards to each of the numerical characteristic.
6. Categorical Data Analysis: To show the relative frequency distribution of related categories, count plots for categorical columns are used. It can show which of the categories is more present and how they might influence other values.
7. Correlation Matrix: It turns out that using the heatmap of the correlation matrix we can define the relation between the numbers. Key observations might include:

Signs, for example, ‘median\_income’ with ‘median\_house\_value-(positional relationships).

1. Pair Plot: The pair plot enable us to look for relationships between two numerical variables at once. It gives a clue on linear relationship and grouping of the data.
2. Outlier Detection: The general box plots of the numerical features of a dataset allow the identification of outliers within the data set. Observations might include:

**Approach**

1. Data Loading and Initial Exploration:The data set was then read into a Pandas Data frame as it allowed easier manipulation of the data. Data was explored using general properties, the initial few rows of the dataset were looked at using ‘df.head ()’ which gave an idea of the nature and arrangement of the dataset. Function like, ‘df.info()’ was used to give an overview about the structure of the database, how many records it contains, the whole data type and measurement of memory space it occupies.
2. Statistical Summary: The ‘describe()’ method was used on the numerical columns to output summary statistics which included central tendencies and dispersion. This step drawn attentions to the distributions of the numerical features.
3. Missing Values Assessment: Missing values are one of the biggest problems that one encounters while analyzing the data. To check the number of missing values in the each column the ‘isnull().sum()’ was applied. The above information is essential in determining how to proceed with regard to imputation procedures or whether records or features need to be deleted completely.
4. Data Type Verification: The type of each column was also determined using the command: ‘dtypes from df’ It was particularly important to make sure that all calculation fields were properly recognized as numerical in order to benefit from further statistical processing and data visualization.
5. Visualization of Distributions: For numerical features, histogram were used to present the distribution of these features. This enabled identification of skewness, modality (unimodal, bimodal) and outliers.
6. Correlation Matrix: In fact, it appears that the heatmap of the correlation matrix allows defining the relation between the numbers.
7. Pair Plot: Both of the pair plot allow to search for some relationship of two numerical variables simultaneously. It provides an indication on the linear association and the grouping of the data set.
8. Outlier Detection: The general box plots of the numerical features of a datasets make it possible to determine data points that are either in or out of the data set.

**Conclusion**

Upon analyzing the sample data for California housing dataset, the following insights were obtained from exploratory data analysis (EDA). There are no missing values in the dataset which enables the analysis of the data set without any interruption. Numerical variables, displayed skewness positive values showing that a large number of values are located on the smaller side, meaning that some transformation could be necessary during modeling. Performing both correlation analysis, histogram and box plot it was found that some features were highly correlated that might depict multicollinearity while some of the features had outliers that could skew the results of statistical test. Frequency distributions of categorical variables offered a mechanism for identifying how categories of different variables are distributed so as to reveal any issue that might influence classification. In conclusion, the EDA proved to be very insightful in giving an overview of the dataset that will allow for better feature selection as well as better strategies in modeling the relationships to exist in the data set as obtained in the subsequent analysis.

**Reference**

Kaggle. (2019). *Kaggle: Your Home for Data Science*. Kaggle.com. <https://www.kaggle.com>