National University of Computer and Emerging Sciences



Lab Exercise 08

AL2002-Artificial Intelligence Lab

|  |  |
| --- | --- |
| Course Instructor | Dr. Hajra Waheed |
| Lab Instructor (s) | Hamza Ayub  Muhammad Ahmed |
| Sections | B1 - B2 |
| Semester | Spring 2023 |

Department of Computer Science

FAST-NU, Lahore, Pakistan

Table of Contents

[1 Objectives 3](#_Toc69063793)

[2 Task Distribution 3](#_Toc69063794)

[3 Exploratory Data Analysis 3](#_Toc69063795)

[3.1 Significance of EDA 3](#_Toc69063796)

[3.2 Properties of Attributes in a dataset 4](#_Toc69063797)

[3.2.1 Missing Values 4](#_Toc69063798)

[3.2.2 Outliers 4](#_Toc69063799)

[3.3 Basic Steps 4](#_Toc69063800)

[3.3.1 Variable Identification 5](#_Toc69063801)

[3.3.2 Missing Values Treatment 6](#_Toc69063802)

[3.3.3 Outlier Treatment 6](#_Toc69063803)

[4 Exercise (20 Marks) 6](#_Toc69063804)

[4.1 Perform Exploratory Data Analysis using python libraries (NumPy and Pandas) on the dataset provided. 6](#_Toc69063805)

[4.1.1 Instructions 6](#_Toc69063806)

[5 Submission Instructions 6](#_Toc69063807)

# Objectives

After performing this lab, students shall be able to understand Python concepts and applications:

* Anaconda Introduction
* Application of NumPy
* Application of Pandas
* Dataset handling
* Exploratory Data Analysis

# Task Distribution

| **Total Time** | **170 Minutes** |
| --- | --- |
| Anaconda Introduction | 10 Minutes |
| Overview of NumPy and Pandas | 20 Minutes |
| Dataset handling | 20 Minutes |
| Review of Dataset | 20 Minutes |
| Exercise | 90 Minutes |
| Online Submission | 10 Minutes |

# Exploratory Data Analysis

Exploratory data analysis (EDA) is used by data scientists to analyze and investigate data sets and summarize their main characteristics, often employing data visualization methods. It helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

## Significance of EDA

The main purpose of EDA is to help look at data before making any assumptions. It can help identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables. Data scientists can use exploratory analysis to ensure the results they produce are valid and applicable to any desired business outcomes and goals.

EDA also helps stakeholders by confirming they are asking the right questions. EDA can help answer questions about standard deviations, categorical variables, and confidence intervals. Once EDA is complete and insights are drawn, its features can then be used for more sophisticated data analysis or modeling, including machine learning.

## Properties of Attributes in a dataset

### Missing Values

Some entries can be missing because of the following reasons:

1. **Data Extraction**: It is possible that there are problems with extraction process. In such cases, we should double-check for correct data with data guardians. Errors at data extraction stage are typically easy to find and can be corrected easily as well.
2. **Data collection**: These errors occur at time of data collection and are harder to correct.

### Outliers

Outlier is an observation that appears far away and diverges from an overall pattern in a sample. Let’s take an example, we do customer profiling and find out that the average annual income of customers is $1 lakh. But, there are two customers having annual income of $4 and $4.2 million. These two customers annual income is much higher than rest of the population. These two observations will be seen as Outliers.

Outliers can be due to the following reasons:

1. **Data Entry Errors:** Human errors such as errors caused during data collection, recording, or entry can cause outliers in data. For example: Annual income of a customer is $100,000. Accidentally, the data entry operator puts an additional zero in the figure. Now the income becomes $1,000,000 which is 10 times higher. Evidently, this will be the outlier value when compared with rest of the population.
2. **Measurement Error:**It is the most common source of outliers. This is caused when the measurement instrument used turns out to be faulty. For example: There are 10 weighing machines. 9 of them are correct, 1 is faulty. Weight measured by people on the faulty machine will be higher / lower than the rest of people in the group. The weights measured on faulty machine can lead to outliers.
3. **Experimental Error:** Another cause of outliers is experimental error. For example: In a 100m sprint of 7 runners, one runner missed out on concentrating on the ‘Go’ call which caused him to start late. Hence, this caused the runner’s run time to be more than other runners. His total run time can be an outlier.
4. **Data Processing Error:**Whenever we perform data mining, we extract data from multiple sources. It is possible that some manipulation or extraction errors may lead to outliers in the dataset.
5. **Sampling error:** For instance, we have to measure the height of athletes. By mistake, we include a few basketball players in the sample. This inclusion is likely to cause outliers in the dataset.

## Basic Steps

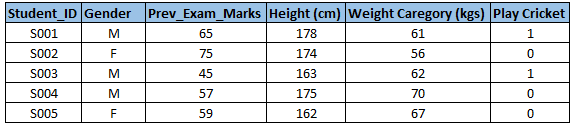
Below are the steps involved to understand, clean and prepare your data for building any predictive model:

1. Variable Identification
2. Missing values treatment
3. Outlier treatment

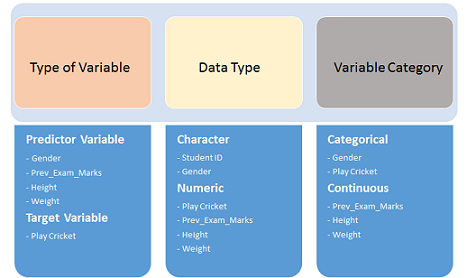
### Variable Identification

Variable identification means to identify the data type and category of the variables in a dataset.

**Example:**

Suppose, we want to predict, whether the students will play cricket or not (refer below data set). Here you need to identify predictor variables, target variable, data type of variables and category of variables.[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_11.png)

Below, the variables have been defined in different category:

[](https://www.analyticsvidhya.com/wp-content/uploads/2015/02/Data_exploration_2.png)

### Missing Values Treatment

1. **Deletion:** Delete entry with missing value.
2. **Mean/ Mode/ Median Imputation**: Imputation is a method to fill in the missing values with estimated ones. The objective is to employ known relationships that can be identified in the valid values of the data set to assist in estimating the missing values. Mean / Mode / Median imputation is one of the most frequently used methods. It consists of replacing the missing data for a given attribute by the mean or median (quantitative attribute) or mode (qualitative attribute) of all known values of that variable.

### Outlier Treatment

1. **Deleting observations:**We delete outlier values if it is due to data entry error, data processing error or outlier observations are very small in numbers. We can also use trimming at both ends to remove outliers.
2. **Transforming and binning values:**Transforming variables can also eliminate outliers.

# Exercise (20 Marks)

## Perform Exploratory Data Analysis using python libraries (NumPy and Pandas) on the dataset provided.

### Instructions

You need to find insights about data using as many different techniques as you can. Don’t use libraries that haven’t been covered yet.

Hint: Explore the dataset, look for the outliers, missing values, etc.

Visit [this](https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15#:~:text=Exploratory%20Data%20Analysis%20refers%20to,summary%20statistics%20and%20graphical%20representations.) site for step wise Exploratory Data Analysis on a dataset.

Watch a video tutorial [here](#_Exercise_(_Marks)) to learn more about practical implementation of EDA.

# Submission Instructions

Always read the submission instructions carefully.

* Rename your Jupyter notebook to your roll number and download the notebook as **.ipynb** extension.
* To download the required file, go to **File->Download .ipynb**
* Only submit the **.ipynb** file. DO NOT **zip** or **rar** your submission file
* Submit this file on Google Classroom under the relevant assignment.
* Late submissions will not be accepted