National University of Computer and Emerging Sciences

Lab Exercise
CL461-Artificial Intelligence Lab

Department of Computer Science
FAST-NU, Lahore, Pakistan

Table of Contents

1 Objectives 3

2 Task Distribution 4

3 Exploratory Data Analysis 4

3.1 Significance of EDA 4

3.2 Properties of Attributes in a dataset 5

3.2.1 Missing Values 5

3.2.2 Outliers 5

3.3 Basic Steps 5

3.3.1 Variable Identification 6

3.3.2 Missing Values Treatment 7

3.3.3 Outlier Treatment 7

4 Exercise (20 Marks) 7

4.1 Perform Exploratory Data Analysis using python libraries (NumPy and Pandas) on the dataset provided. 7

4.1.1 Instructions 7

5 Submission Instructions 8

1 Objectives

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

- ✓ Anaconda Introduction
- ✓ Application of NumPy
- ✓ Application of Pandas
- ✓ Dataset handling
- ✓ Exploratory Data Analysis

2 Question 1:

Form a queue such that it works in FIFO order

2.1 Question 2:

Create a class for rectangle shape that calculates its area based upon the length and width

3 Question 3

Write a Python class named Circle constructed by a radius and two methods which will compute the area and the perimeter of a circle.

4 Question 4

Create a custom iterator that returns numbers, starting with 1, and each sequence will increase by one (returning 1,2,3,4,5 etc.):

Question 5

Create a class for rectangle shape that calculates its area based upon the length and width. Make a sub class of rectangle called Trapezium, such that it inherits the functionality of rectangle class and implements area method of its own. Length and width should be defined in the constructor of rectangle class.

Area of Rectangle =Length+Width

Area of Trapezium=½*(l+w)*h

After creation of the class, define the relevant attributes. Define a function for area computation and then a function for displaying area. Incorporate your knowledge of class and objects here.

5 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

6 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.

6.1 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.

6.2 Properties of Attributes in a dataset

6.2.1 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.

6.2.2 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.

6.3 Basic Steps

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

1. Variable Identification

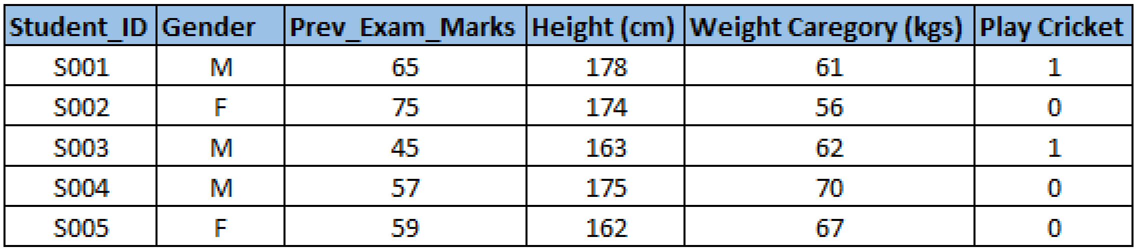
2. Missing values treatment
3. Outlier treatment

6.3.1 Variable Identification

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

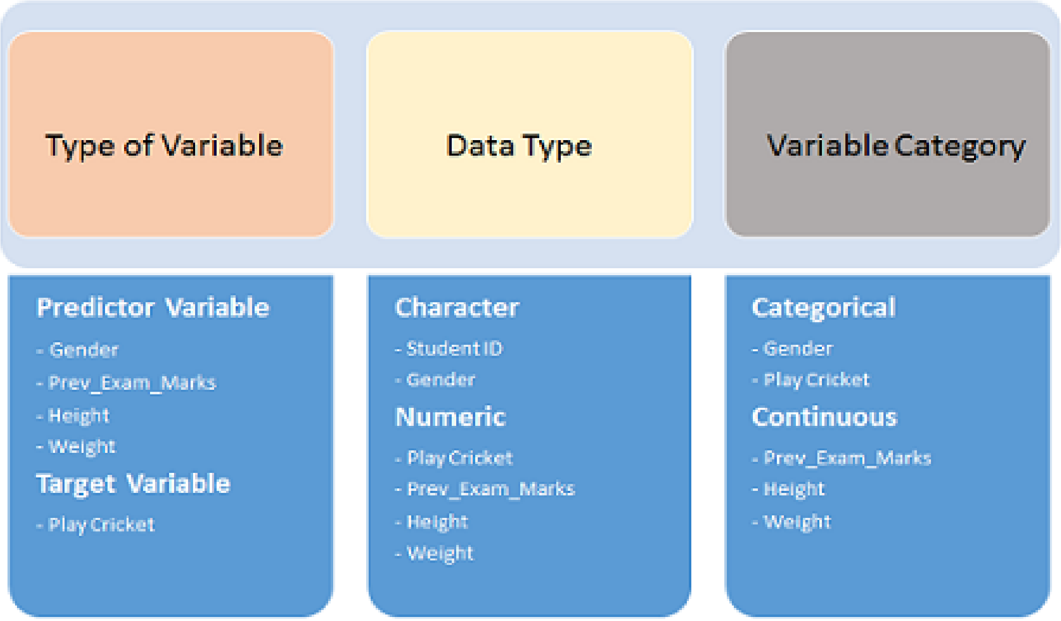
Example:

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.



Student_ID	Gender	Prev_Exam_Marks	Height (cm)	Weight Caregory (kgs)	Play Cricket
S001	M	65	178	61	1
S002	F	75	174	56	0
S003	M	45	163	62	1
S004	M	57	175	70	0
S005	F	59	162	67	0

Below, the variables have been defined in different category:



Type of Variable

Data Type

Variable Category

Predictor Variable
- Gender
- Prev_Exam_Marks
- Height
- Weight

Target Variable
- Play Cricket

Character
- Student ID
- Gender

Numeric
- Play Cricket
- Prev_Exam_Marks
- Height
- Weight

Categorical
- Gender
- Play Cricket

Continuous
- Prev_Exam_Marks
- Height
- Weight

6.3.2 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.

6.3.3 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.

7 Exercise

7.1 NumPy is a library for the Python which adds support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays. Create two arrays A= [1,2,3,2,3,4,3,4,5,6] and B= [7,2,10,2,7,4,9,4,9,8]. You need to get the positions where elements of A and B match. Use a numpy function or implement your own logic.

Desired Outcome:

```
[1, 3, 5, 7]
```

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

7.2.1 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

.

Use the automobile data set provided in the Google class room to explore EDA functions.

1. Check first five entries of data set. *# data.head().*
2. Check last five entries of dataset.*# data.tail().*
3. Check the columns of data set.*# data.columns.*
4. Check unique values for each column *# data.numique().*
5. Check the missing values for each column. *# data.isnull().sum().*

6. Drop unnecessary data columns *# new_data= data.drop([‘column_name’,‘second_column_name’], axis=1).*
7. Drop null value. `#df.dropna(subset=df.columns,inplace=True)`

7.2.2 Data Analysis is the process of exploring and analyzing large datasets to make predictions and decisions. It involves a broad set of activities to clean, process, transform a data collection to learn from it and derive meaningful insights. Its profound application can be seen in analysing consumer behaviour in retail industry to reach out to the right customers and perform targeted marketing to increase sales. One sample dataset has been provided which lists various features of cars. You need to use your data analysis skills in answering the questions given ahead.

(Libraries Involved: Numpy, Pandas)
Dataset: **automobile_data.csv**
Initial Steps:

- Import necessary libraries
- Upload/Read the csv file using pandas
- Review the dataset for identifying any missing values
- Observe the different attributes and entries

Questions

a. Find the most expensive car from the dataset and display its price
b. Calculate total cars per manufacturer and show the result
c. Read the details of vehicles against Toyota manufacturer and print them
d. Arrange the cars according to the prices (highest-lowest) and display relevant information for first **5** rows only

7.2.3 Create a dataframe. Make relevant columns, populate values in the data set, write it as .csv file and download the .csv file to your computer.

Visit this site for step wise Exploratory Data Analysis on a dataset.

8 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