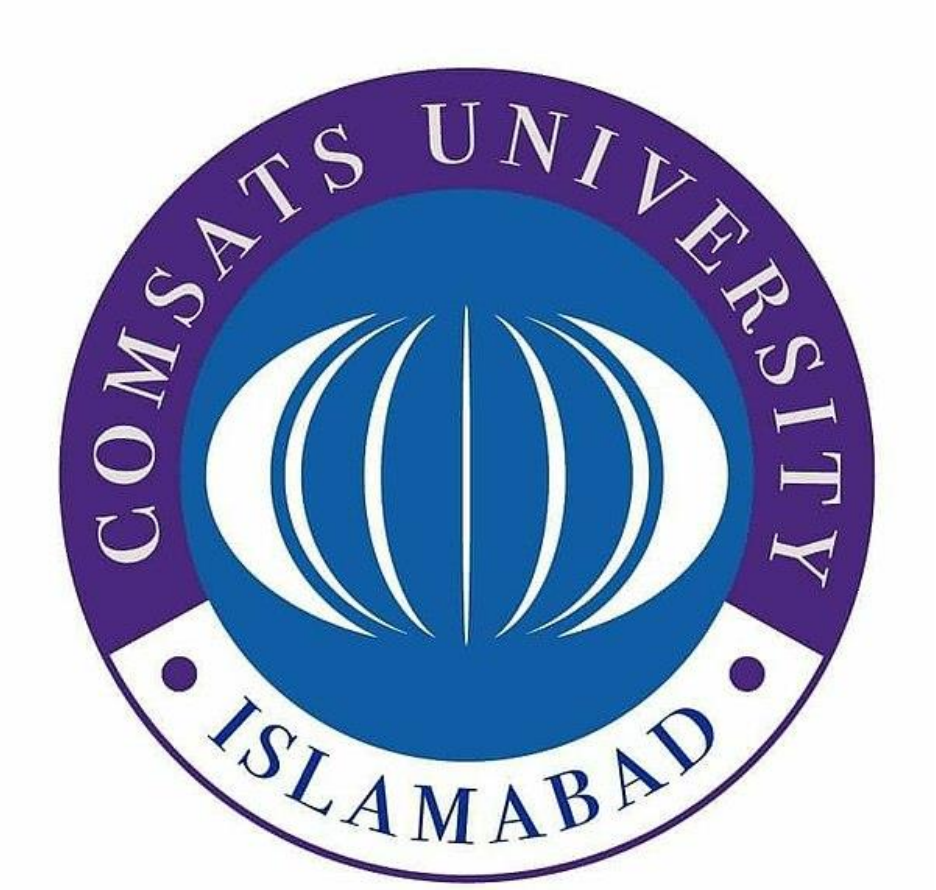

INTRODUCTION TO DATA SCIENCE

-ASSIGNMENT # 1-

Registration Number: SP20-BCS-095

Student Name: Rohaan Farooq

Course: Introduction to Data Science (CSC461)

Group: BCS 6 - G2-A

Assignment #: 1

Submitted to: Prof. Muhammad Sharjeel

Date: 21st October 2022

Manually Viewing the Dataset (Analysis)

The original dataset is not fit to apply data science techniques. A few observations are as follows:

- In the HSSC marks there are various formats of data (marks). e.g. in the HSSC marks there are various formats of data (marks). Some of the marks are out of only the total of that particular year. Some marks in HSSC-2 are out of total of both years. Some marks are in percentages, etc.
- Names are written differently, some names are Capitalized, while some are not. Some names have the name Muhammad abbreviated as M, some names are all in uppercase, while some in all lowercase, etc.
- Favorite colors are also written in various format, like in all uppercase, all lowercase, Capitalized, and-separated (in case of more than one favorite color), space-separated (in case of more than one favorite color), etc.
- Birth months are also in different formats like capitalized, all lowercase, month abbreviations, also numbered months, etc.
- The weights are also in different formats, some are as integers, some as floats, some have the units Kg, some don't, etc.

Therefore, the dataset needs to be cleaned in order to effectively apply data science methods

Cleaning the Dataset

Although this dataset could have more easily be corrected manually, but for scalability, and generality, the dataset was cleared using programmatic approach.

The dataset was hence cleaned/normalized using Regular Expressions, the codes for which can be found at the start of the ipynb file (after importing the libraries) included in the zip file. A basic overview is provided in the points below:

- The HSSC-1 and HSSC-2, marks were mostly corrected using the "split()" function, to split the string at '/' or '%'.
- There was a discrepancy in the CGPA as well which was "2.84." that was also corrected using "split()" function.
- The Weights had a couple of records with the units Kg attached with them, which were also removed using the "split()" function.
- For the birth months, two arrays were created, one containing the names the months in lowercase, and the other containing the names in Capitalized. Then then names or their abbreviations was matched with the corresponding entries to correctly fill the record. Also some month names had whitespaces at the start or end, they were dealt with "strip()" function.
- Favorite Colors had to be cleaned manually as it was getting exceedingly difficult to cleaning the colors due to the wild differences in formats and patterns and whitespaces.

What other things (insights) you can get from the dataset?

- It can be noted from the dataset that a few values in hssc-1 have been entered incorrectly, which makes the analysis of both hssc-1 and hssc-2 of those particular records useless.
- The dataset shows the different ways and formats that people tend to enter data.
- The incorrectly entered data can also signify that sometimes, people are not willing to share their actual data, so they provide false information instead, which can ultimately have a negative impact on the overall analysis of the data as a whole.

Notice

For the source code, kindly refer to the ipynb file, as it has the solution to all the given problems along with the problem statement commented at the top of each cell.

The End