CSE CSE3020

Data visualisation

Theory

Digital Assessment 2

TOPIC: Multi-variate data analysis

Name: Makesh Srinivasan
Registration number: 19BCE1717
Slot: F1 + TF1
Date: 08-April-2022-Friday
Faculty: Prof. Parvathi

Directions:

1. Define Multivariate Data
2. Example of Multivariate Data with case study
3. Different Visualization can be used for Multivariate Data (Mention 5 different visualisation)
4. Implement the same using python or R using the same case study given in point no 2
5. Give the dataset link and coding (co-lab link or GitHub link)
5. Conclusion

Table of contents

As part of our Computer Science Engineering curriculum, we have worked with many datasets where there was one label or prediction attribute but multiple independent variables. We have even see univariate data where there is only one attribute. For example, age distribution of students in my class as a vector. But less rarely, we have also seen something called as multivariate data.

I) Definition

Multivariate data is a dataset that involves more than 2 dependent or prediction variables, resulting in a single outcome. Multiple independent variables can be used to determine two or more dependent variables using a function or a model. Multivariate data analytics involve simultaneous observation and analysis of more than one outcome variables. This allows us to find patterns and correlations, to get a deeper and more complex understanding of the scenario described in the dataset.

The same is depicted in Figure 1. Independent variables X1 and X2 are used to predict Y1, which is a dependent variable. We then make use of two independent variables X3 and X4, and one dependent variable Y1 to predict Y2, which is also a dependent variable. This is an example of multivariate dataset.

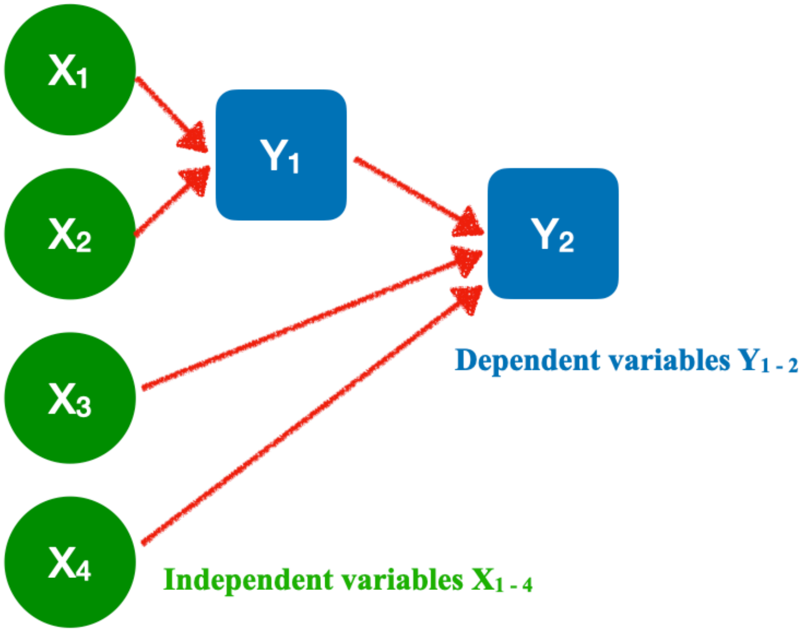
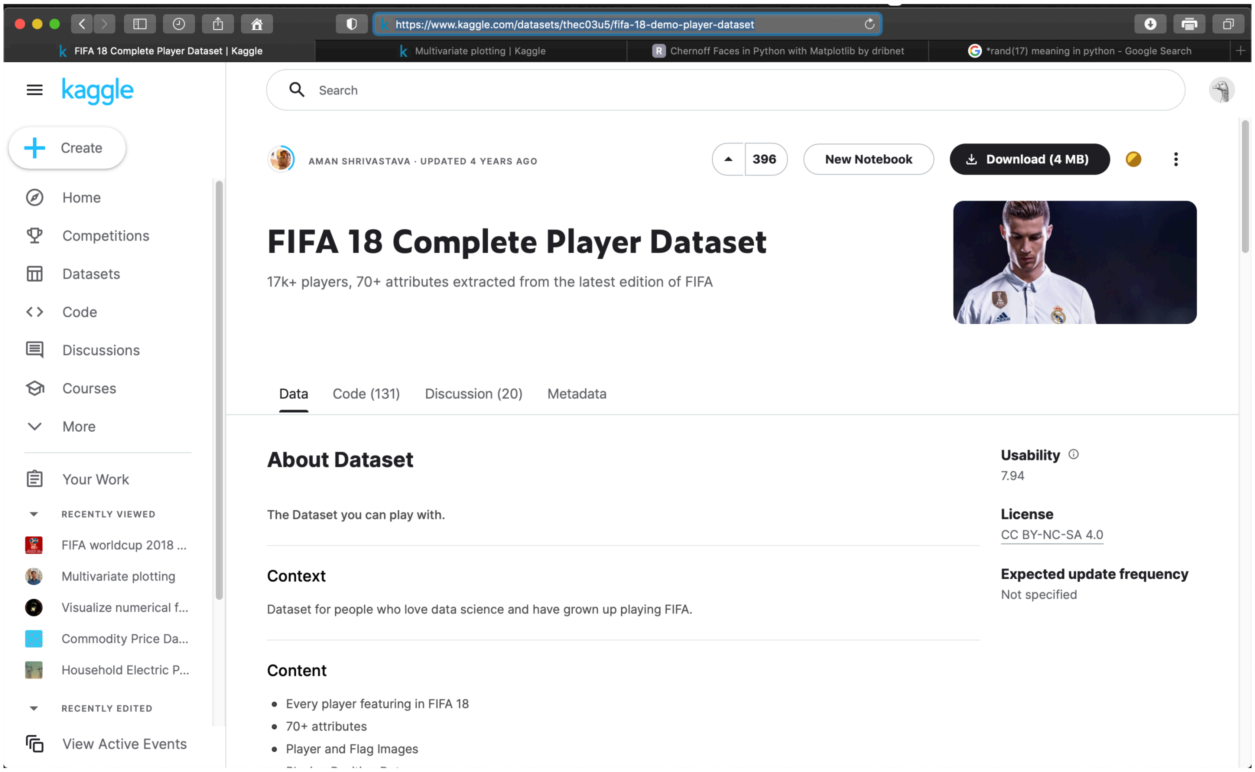

X_1
X_2
X_3
X_4
Y_1
Y_2
Dependent variables Y_{1-2}
Independent variables X_{1-4}

Figure 1: Multivariate dataset structure

II) Example of a multivariate dataset with case study

For the purpose of this digital assessment, I have used the FIFA dataset that consists of 75 attributes about all the football players in the FIFA tournament. It describes features such as player speed, dribble abilities, crossing skills, accuracy, stamina, nationality, age, position, etc. We can perform several analysis using this dataset such as analyse the performance of the team in a given season, estimate the potential of a team in a tournament, or find relations between various factors such as age and nationality with respect to performance.

The dataset [1] was acquired from Kaggle which was in turn scraped from the sofifa portal that provides details of players and teams. The dataset depicts the players data of EA's FIFA-18.



AMAN SHRIVASTAVA · UPDATED 4 YEARS AGO

FIFA 18 Complete Player Dataset

17k+ players, 70+ attributes extracted from the latest edition of FIFA

About Dataset

The Dataset you can play with.

Context

Dataset for people who love data science and have grown up playing FIFA.

Content

- Every player featuring in FIFA 18
- 70+ attributes
- Player and Flag Images

Usability
7.94

License
CC BY-NC-SA 4.0

Expected update frequency
Not specified

Image 1: FIFA 18 players dataset on Kaggle

III) Different visualisations for multivariate dataset

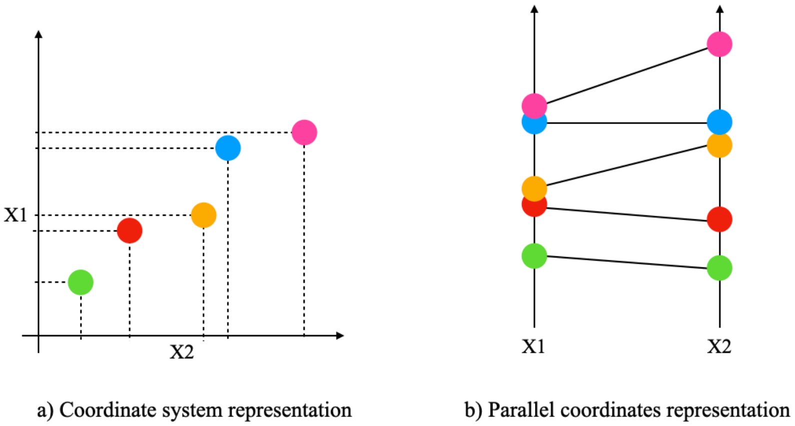
In our three-dimensional universe, we are able to perceive objects in three spatial dimensions. This means, we are able to visualise the width, depth and the height of an entity in our universe. We also define a fourth dimension for time. This allows to determine the properties of the object with respect to time. But our abilities to visualise dimensions beyond these 3 (4 if time is considered) is limited by our cognition. We are unable to directly visualise higher dimensions in our three dimensional world. This is why we use multivariate visualisation techniques.

Multivariate data visualisation techniques allows us to interpret and visualise multiple attributes encoded as a dimension simultaneously. It is generally used to understand the interaction and relation between more than 3 variables. The objective of this technique is to enable effective comprehension of multivariate data using several retinal elements such as colours, shapes, sizes, angles, etc. In this digital assessment, I have implemented and described the following techniques.

1. Parallel coordinates
2. Star plots
3. Scatter plot matrix
4. Chernoff faces
5. Pixel oriented representation

1) Parallel coordinates

They depict the orthogonal axis of the coordinates system and lay them in parallel to one another. This enables us to perceive higher dimensional data effectively. The attributes or axis are normalised so that the top represents the higher values while the bottom represents the lower



a) Coordinate system representation

b) Parallel coordinates representation

Figure 2: Coordinate system representation to parallel coordinates representation

values. The points on the line correspond to the values of the attribute encoded in the axis. By doing this, we can spread the data points of a multi-dimensional coordinate system on a two dimensional parallel coordinates system. Figure 2 shows how a 2D coordinate system (cartesian plane) is mapped to a parallel coordinate system. By increasing the dimensions of the coordinate system, we see how we reach a limit in our spatial and cognitive abilities in perceiving more than 3 dimensions. However, parallel coordinates system can handle this effectively as it is only a matter of adding one more parallel axis.

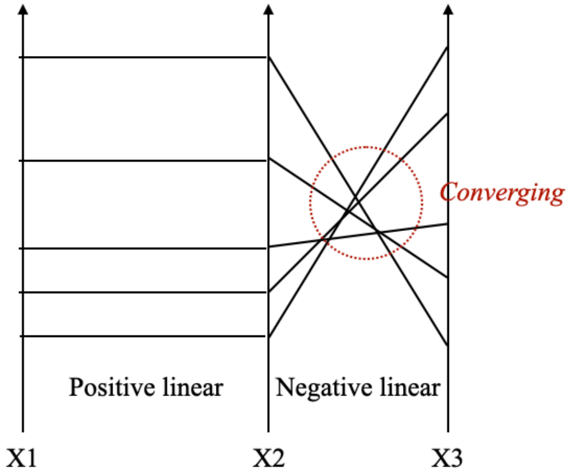


Figure 3: Relationship - positive and negative on parallel coordinates

Once the points are mapped to the parallel coordinates, we draw a line between the corresponding pairs of points to represent a duality, in which the line in parallel coordinates system draws a parallel to the orthogonal coordinate system in a way that the points of the latter are mapped as lines in the parallel coordinates system. One of the features in this type of visualisation is collinearity. If the lines are perfectly parallel between two parallel axis, we say it is perfect positive linear relationship. If the lines cross each other, we say it is perfect negative linear relationship. This is depicted in Figure 3. The attributes X1 and X2 are perfect positive linear while X2 and X3 are perfect negative linear.

2) Star plots

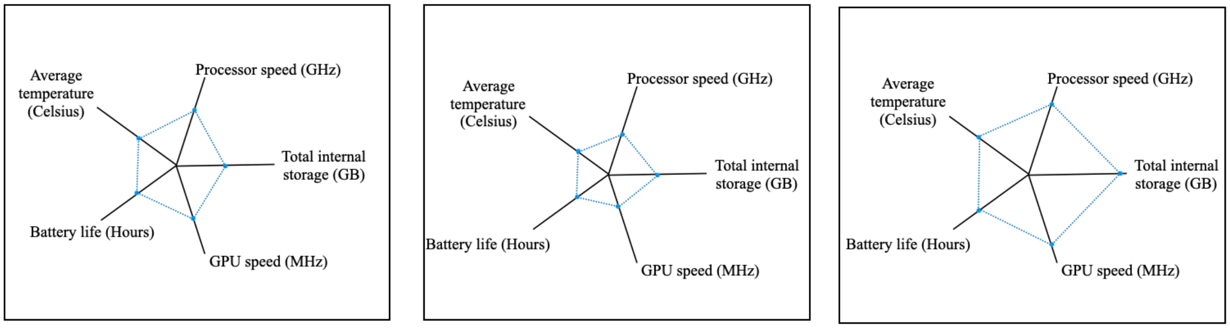
In this plot, each data record is represented as a star-shaped figure with one ray for each variable. The length of each ray is proportional to the value of its corresponding variable. Hence, the variables have to be normalised between a small range such as 0 and 1 for the diagram to be

Table 1

Attributes	Laptop 1	Laptop 2	Laptop 3
Processor speed (GHz)	2.4	1.8	3.6
Total internal storage (GB)	512	512	1024
GPU speed (MHz)	3.2	2.8	3.6
Battery life (Hours)	10	8	12
Average temperature (Celsius)	53	49	60

effectively understood. Once the data points of a particular object is encoded on each ray of the star plot, the adjacent points are connected to create a closed loop. This is depicted in the following example scenario below.

Assume we have three laptops and we want to understand the performance of each of them. We have identified and calculated several important attributes such as processor speed, Total internal storage, GPU speed, battery life, and average temperature of the system. The sample data is shown in Table 1. Each attribute is represented as a ray in the star plot. For each laptop, we create a star plot. The star plots for laptops 1, 2 and 3 are given below in Figure 4. By observing the star plot, we can analyse and infer details about one laptop in comparison to another. However, one drawback of this visualisation technique is that if we add more attributes, it becomes difficult to distinguish the rays.



Laptop 1 **Laptop 2** **Laptop 3**

Figure 4: Star plots for the laptops

3) Scatter plot matrix

This is one of the very first plotting technique a data analyst would implement to understand the dataset presented. This is a very simple yet rich in information and can drastically change the way we proceed with the dataset. It's an excellent tool for comparing similar data sets quickly (as they are each arranged next to each other in vertical and horizontal directions). This method is easier to use for analysis in some circumstances than parallel coordinates, but it has its own problems, including the inability to label the individual axes of smaller scatterplots due to space constraints and legibility requirements. A scatterplot matrix is an attempt to add dimensions to the typical 2D scatterplot. It works by storing all potential scatterplots formed from pairs of variables in a matrix with all possible scatterplots created from pairs of variables in the data set.

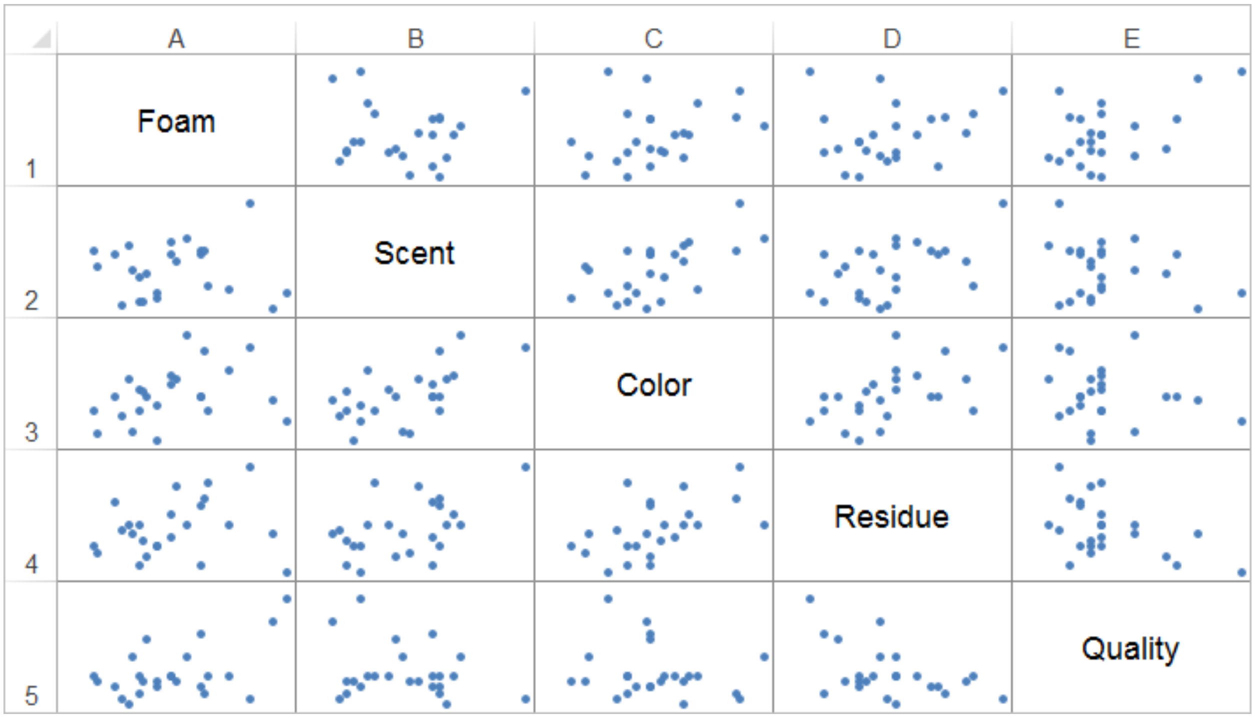


Figure 5: Scatter plot matrix of 5 attributes

In the given scenario in Figure 5, we present the scatter plot between the attributes foam, scent, colour, residue, and quality with each other. If the scatterplot follows an upward trend between a given two distinct attributes, we say that there is a positive linear trend observed. If the scatterplot is downward trending, then the attributes are negatively or inversely proportional. If the scatterplot is seemingly random, there is no trend observed between the attributes and hence no correlation.

4) Chernoff faces

Chernoff faces is to display multiple variables at once by using parts of the human face, such as ears, hair, eyes, and nose, based on numbers in a dataset. The orientation, position, shape, colour, angle and size of different parts of the human can be used to represent different attributes from the dataset. The reason for this technique is that the human face is one of the easiest and most intuitive shape recognisable. The assumption here is that by encoding data as human face, we can infer and understand the data more naturally.

Let us take an example to understand how this is done. I have used a very commonly used example of Chernoff faces - Life in Los Angeles. There are four attributes that are to be visualised on a geographical map of Los Angeles. The four attributes are affluence - the amount of money one has, unemployment rate, urban stress, and proportion white population. The facial attributes that are encoded for the four mentioned features are shape of the face, shape of the smile, size of the eyes, and proportion of white population. Now, let us try to interpret the plot shown in figure 6.

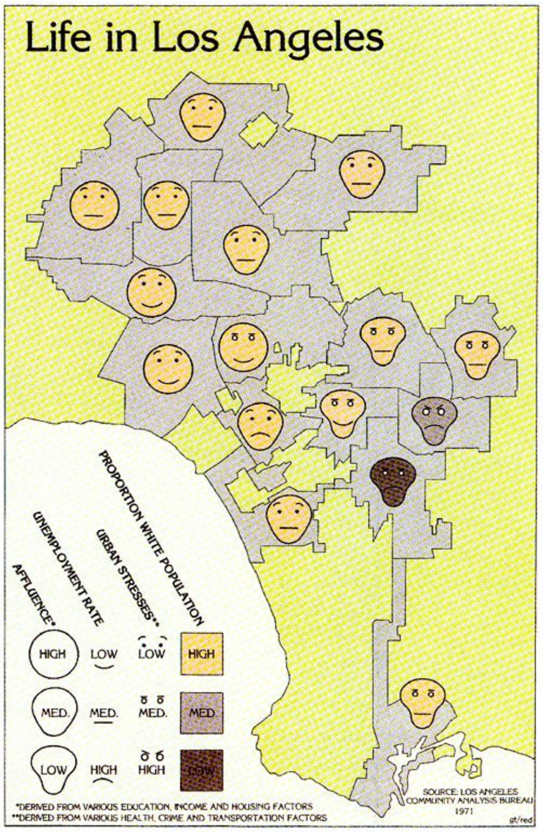
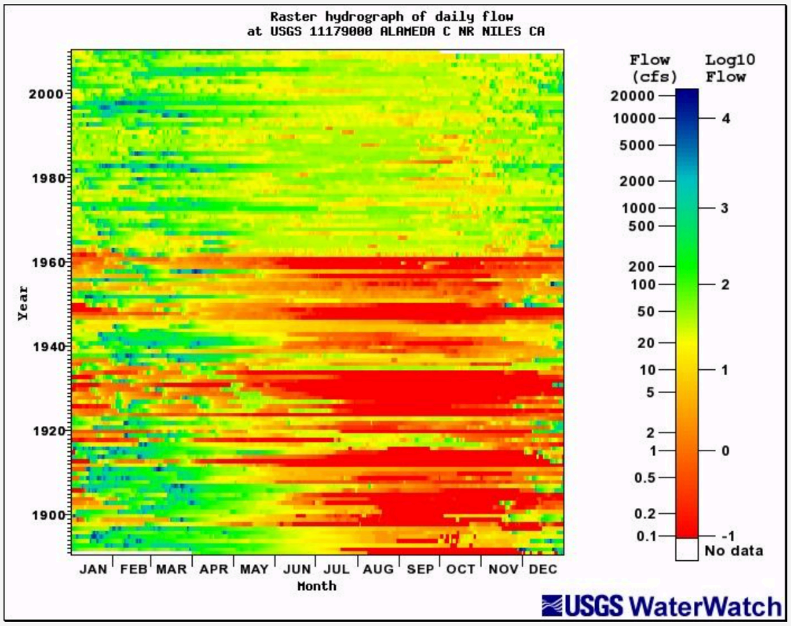

Life in Los Angeles
PROPORTION WHITE POPULATION
URBAN STRESSES**
UNEMPLOYMENT RATE
AFFLUENCE*
HIGH
LOW
LOW
HIGH
MED.
MED.
MED.
MED.
LOW
HIGH
HIGH
LOW
*DERIVED FROM VARIOUS EDUCATION, INCOME AND HOUSING FACTORS
**DERIVED FROM VARIOUS HEALTH, CRIME AND TRANSPORTATION FACTORS
SOURCE: LOS ANGELES COMMUNITY ANALYSIS BUREAU 1971
gt/red

Figure 6: Chernoff faces of Life in Los Angeles

If the face is more round, then the general population in the district has higher affluence (more money). If the face is shrunken, the region has lower affluence. We can draw a parallel to the side that if there is more money, they are able to eat more and thus, the face is round and chubby. Whereas, if there is not much affluence, there is not enough money to eat and thus, the face is shrunken and thin. The second attribute unemployment rate is encoded by the smile. If the unemployment rate is low, the smile is wider, and if it is high, the lips look frowned (inverted smile). Similarly, if the size of the eyes are small, the urban stress is low; and if the size of the eyes are large, the urban stress is high. Lastly, if the proportion white population is skewed towards white, the faces look whiter and if the proportion is skewed towards other races, the colour is on the darker side of the spectrum.

5) Pixel oriented representation

If we have a very large time series dataset that spans multiple years or even decades, and we would like to understand the days in which the sales were profitable, we can use this multivariate visualisation technique. The primary characteristic of this plot is the ability to depict a large set of data in a clear and effective way. The plot is basically a matrix whose rows depict year while the column represents the months. Thus, each cell is essentially a day, known as a pixel. We can colour each cell in a diverging ordering method to represent the magnitude of profits and loss. If the cell is extremely red, the loss is very high; if the cell is extremely green, the profits is very high. The diverging point or the neutral point depicting zero profits or loss will be a colour in-between green and red. This is depicted in Figure 7 using a different scenario. The diagram depicts the daily flow of US65 hydrography. Red represents low while blue represents high flow (sequential ordering technique).



Author/Copyright holder: Richard Koehler. Copyright terms and licence: All rights reserved. Public Domain.

Figure 7: Pixel oriented representation of a hydrography flow

IV) Implementation of the above five visualisation techniques

I have used Python (version 3.7) to perform the visualisation techniques mentioned above and the link for the project is given in the following section. Tools used include Jupiter Notebook for editing and several packages such as pandas, numpy and matplotlib were used as well.

V) GitHub link to the project

Link: https://github.com/Makesh-Srinivasan/Datavisualisation-DA-2

VI) Conclusion

In conclusion, we see that in most real-world scenarios we deal with multivariate datasets, and it is very important to interpret such data effectively. We have explored 5 different strategies to infer data from these methods. They are Scatterplot matrix, Chernoff faces, Pixel oriented representation, star plots and parallel coordinates. We say where and how to apply them and the advantages and disadvantages. We saw several examples of each visualisations as well. These techniques were also implemented using Python in Jupiter Notebook and the same is attached in the repository shared.

VII) References

[1] https://www.kaggle.com/datasets/thec03u5/fifa-18-demo-player-dataset

[2] www.youtube.com/watch?v=WWfbwiO8NNs

[3] www.youtube.com/watch?v=2B6uz0H7BSo&t=370s

APPENDIX

[4] Figure 5: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.qimacros.com%2Fscatter-plot-excel%2Fscatter-plot-matrix%2F&psig=AOvVaw267lQ94u-FALILBGYRs6pI&ust=1649871312345000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCKjrweCHj_cCFQAAAAAdAAAAABAD

[5] Figure 6: https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.pinterest.com%2Fpin%2F447545281695447237%2F&psig=AOvVaw276uTxR6dKgGnrRAWaQi00&ust=1649873797499000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCLiaxf-Qj_cCFQAAAAAdAAAAABAD

NOTE: All the diagrams except Figures 5, 6, and 7 were drawn by me so the link/references are not given. For these, links are mentioned in the appendix section. Image 1 was taken as screenshot. No content was copied or plagiarised in doing this DA 2.