PROJECT BY DEEPIKA.B

Exploratory Analysis Using Univariate, Bivariate, and Multivariate Analysis Techniques

Introduction

Data analysis involves various techniques such as univariate analysis, which is the analysis of a single variable, as well as multivariate analysis, which is the analysis of multiple variables simultaneously. Data is everywhere around us, in spreadsheets, on various social media platforms, in survey forms, and more. The process of cleaning, transforming, interpreting, analyzing, and visualizing this data to extract useful information and gain valuable insights to make more effective business decisions is called Data Analysis.

Data Analysis can be organized into 6 types

1. *Exploratory Analysis*
2. *Descriptive Analysis*
3. *Inferential Analysis*
4. *Predictive Analysis*
5. *Causal Analysis*
6. *Mechanistic Analysis*

Here, we will dive deep into *Exploratory Analysis,*

This article was published as a part of the Data Science Blogathon.

Exploratory Analysis

The preliminary analysis of data to discover relationships between measures in the data and to gain an insight on the trends, patterns, and relationships among various entities present in the data set with the help of statistics and visualization tools is called Exploratory Data Analysis (EDA).

Exploratory data analysis is cross-classified in two different ways where each method is either graphical or non-graphical. And then, each method is either univariate, bivariate or multivariate.

Univariate Analysis

Uni means one and variate means variable, so in univariate analysis, there is only one dependable variable. The objective of univariate analysis is to derive the data, define and summarize it, and analyze the pattern present in it. In a dataset, it explores each variable separately. It is possible for two kinds of variables- Categorical and Numerical.

Some patterns that can be easily identified with univariate analysis are Central Tendency (mean, mode and median), Dispersion (range, variance), Quartiles (interquartile range), and Standard deviation.

Univariate data can be described through:

Ø Frequency Distribution Tables

The frequency distribution table reflects how often an occurrence has taken place in the data. It gives a brief idea of the data and makes it easier to find patterns.

Example:

The list of IQ scores is: 118, 139, 124, 125, 127, 128, 129, 130, 130, 133, 136, 138, 141, 142, 149, 130, 154.

IQ Range	Number
118-125	3
126-133	7
134-141	4
142-149	2
150-157	1

Ø Bar Charts

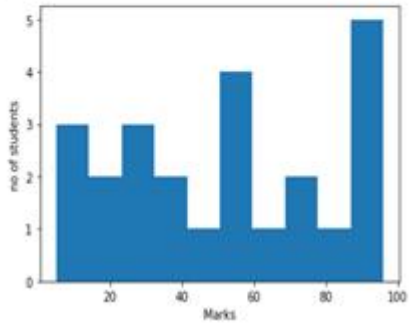
The bar graph is very convenient while comparing categories of data or different groups of data. It helps to track changes over time. It is best for visualizing discrete data.

Ø Histograms

Histograms are similar to bar charts and display the same categorical variables against the category of data. Histograms display these categories as bins which indicate the number of data points in a range. It is best for visualizing continuous data.

SCREEN SHOTS

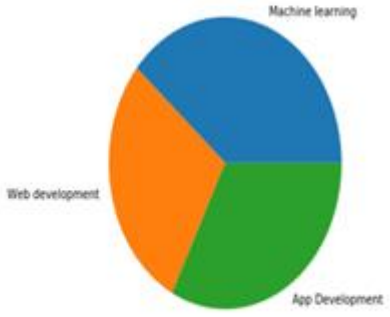
```
In [17]: import matplotlib.pyplot as plt
         import numpy as np
         fig,ax = plt.subplots(1,1)
         a = np.array([22,87,5,43,56,73,55,54,11,20,51,5,79,31,27,63,71,90,92,95,96,32,37,40])
         plt.hist(a)
         ax.set_xlabel('Marks')
         ax.set_ylabel('no of students')
         plt.show()
```



Ø Pie Charts

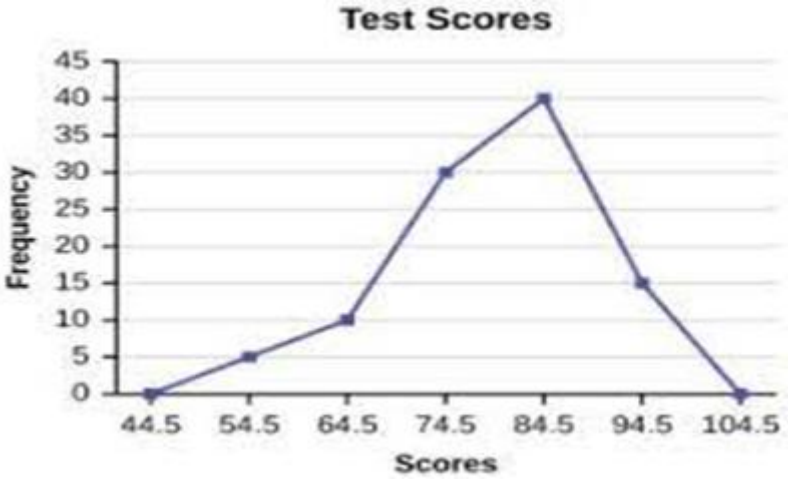
Pie charts are mainly used to comprehend how a group is broken down into smaller pieces. The whole pie represents 100 percent, and the slices denote the relative size of that particular category.

```
In [18]: import matplotlib.pyplot as plt
         fig=plt.figure()
         ax=fig.add_axes([0,0,1,1])
         courses=['Machine learning','Web development','App Development']
         students_enrolled=[50,37,42]
         ax.pie(students_enrolled, labels=courses)
         plt.show()
```



Ø Frequency Polygons

Similar to histograms, a frequency polygon is used for comparing datasets or displaying the cumulative frequency distribution.



Bivariate Analysis

Bi means two and variate means variable, so here there are two variables. The analysis is related to cause and the relationship between the two variables. There are three types of bivariate analysis.

Bivariate Analysis of two Numerical Variables (Numerical-Numerical)

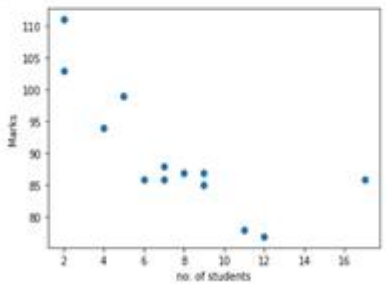
Ø Scatter Plot

A scatter plot represents individual pieces of data using dots. These plots make it easier to see if two variables are related to each other. The resulting pattern indicates the type (linear or non-linear) and strength of the relationship between two variables.

```
In [26]: import matplotlib.pyplot as plt

no_of_students = [5,7,8,7,2,16,2,9,4,11,12,9,6]
marks = [88,86,87,88,79,86,84,87,94,78,77,85,86]

plt.scatter(x, y)
plt.xlabel('no. of students')
plt.ylabel('Marks')
plt.show()
```



Ø *Linear Correlation*

Linear **C**orrelation represents the strength of a linear relationship between two numerical variables. If there is no correlation between the two variables, there is no tendency to change along with the values of the second quantity.

$$r = \frac{Covar(x,y)}{\sqrt{Var(x)Var(y)}}$$

$$Covar(x,y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}$$

$$Var(x) = \frac{\sum(x-\bar{x})^2}{n}$$

$$Var(y) = \frac{\sum(y-\bar{y})^2}{n}$$

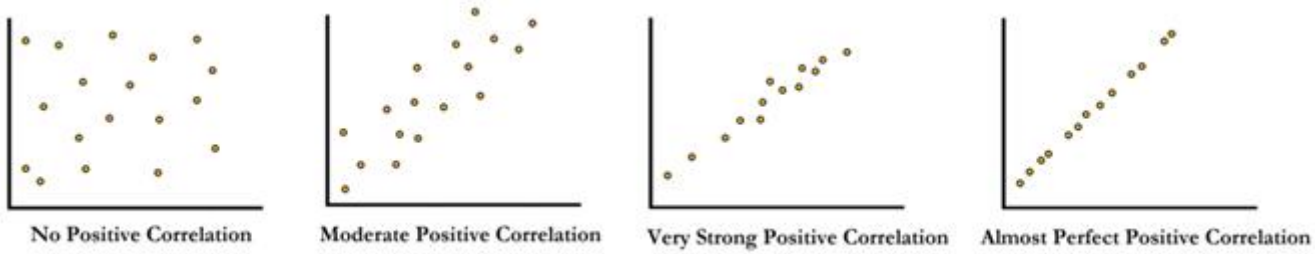
r : Linear Correlation

Covar : Covariance

Var : Variance

Here, r measures the strength of a linear relationship and is always between -1 and 1 where -1 denotes perfect negative linear correlation and +1 denotes perfect positive linear correlation and zero denotes no linear

correlation.



Bivariate Analysis of two categorical Variables (Categorical-Categorical)

Ø Chi-square Test

The chi-square test is used for determining the association between categorical variables. It is calculated based on the difference between expected frequencies and the observed frequencies in one or more categories of the frequency table. A probability of zero indicates a complete dependency between two categorical variables and a probability of one indicates that two categorical variables are completely independent.

Here, subscript c indicates the degrees of freedom, O indicates observed value, and E indicates expected value.

$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Bivariate Analysis of one numerical and one categorical variable (Numerical-Categorical)

Ø Z-test and t-test

Z and T-tests are important to calculate if the difference between a sample and population is substantial.

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$\bar{x}$ = sample mean

μ = population mean

σ = population standard deviation

n = sample size

If the probability of Z is small, the difference between the two averages is more significant.

T-Test

$$t = \frac{\bar{x} - \mu}{s_{\bar{x}}} \qquad \text{where} \qquad s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

where

μ = Proposed constant for the population mean

$\bar{x}$ = Sample mean

n = Sample size (i.e., number of observations)

s = Sample standard deviation

$s_{\bar{x}}$ = Estimated standard error of the mean (s/sqrt(n))

If the sample size is large enough, then we use a Z-test, and for a small sample size, we use a T-test.

Ø ANALYSIS OF VARIANCE (ANOVA)

The ANOVA test is used to determine whether there is a significant difference among the averages of more than two groups that are statistically different from each other. This analysis is appropriate for comparing the averages of a numerical variable for more than two categories of a categorical variable.

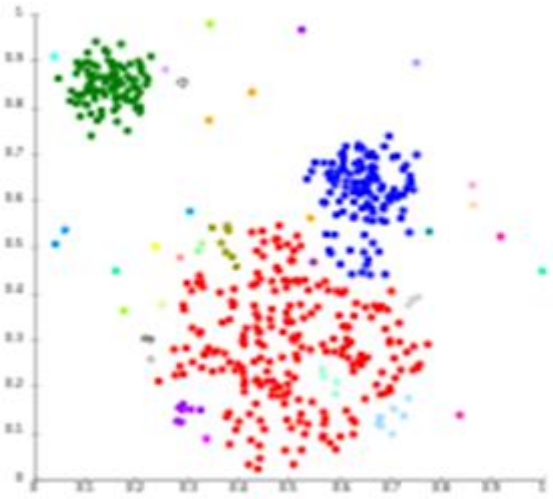
Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SSW = \sum_{j=1}^{k}\sum_{j=1}^{l}(X-\bar{X}_j)^2$	$df_w = k-1$	$MSW = \frac{SSW}{df_w}$	$F = \frac{MSB}{MSW}$
Between	$SSB = \sum_{j=1}^{k}(\bar{X}_j - \bar{X})^2$	$df_b = n-k$	$MSB = \frac{SSB}{df_b}$	
Total	$SST = \sum_{j=1}^{n}(\bar{X}_j - \bar{X})^2$	$df_t = n-1$		

Multivariate Analysis

Multivariate analysis is required when more than two variables have to be analyzed simultaneously. It is a tremendously hard task for the human brain to visualize a relationship among 4 variables in a graph and thus multivariate analysis is used to study more complex sets of data. Types of Multivariate Analysis include Cluster Analysis, Factor Analysis, Multiple Regression Analysis, Principal Component Analysis, etc. More than 20 different ways to perform multivariate analysis exist and which one to choose depends upon the type of data and the end goal to achieve. The most common ways are:

Ø Cluster Analysis

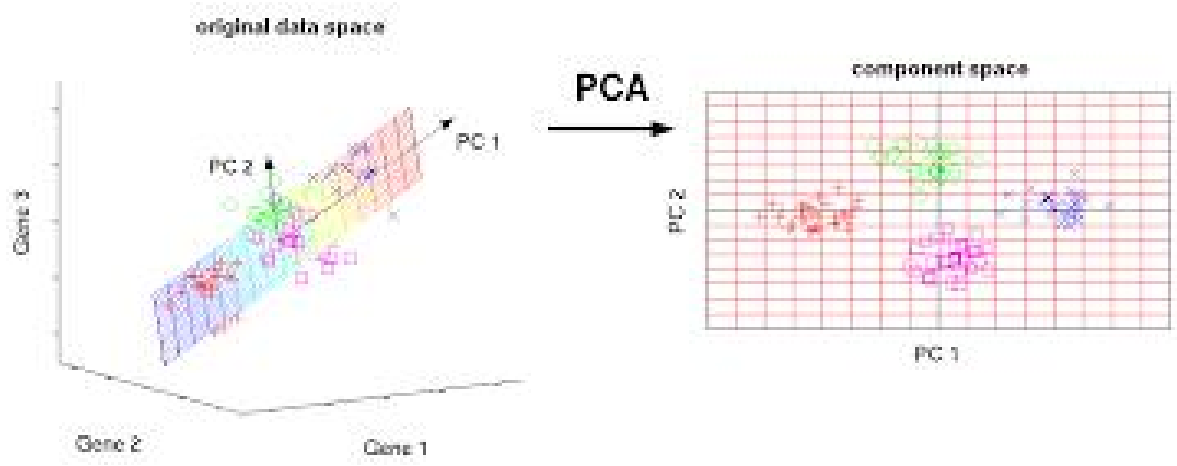
Cluster Analysis classifies different objects into clusters in a way that the similarity between two objects from the same group is maximum and minimal otherwise. It is used when rows and columns of the data table represent the same units and the measure represents distance or similarity.



Ø Principal Component Analysis (PCA)

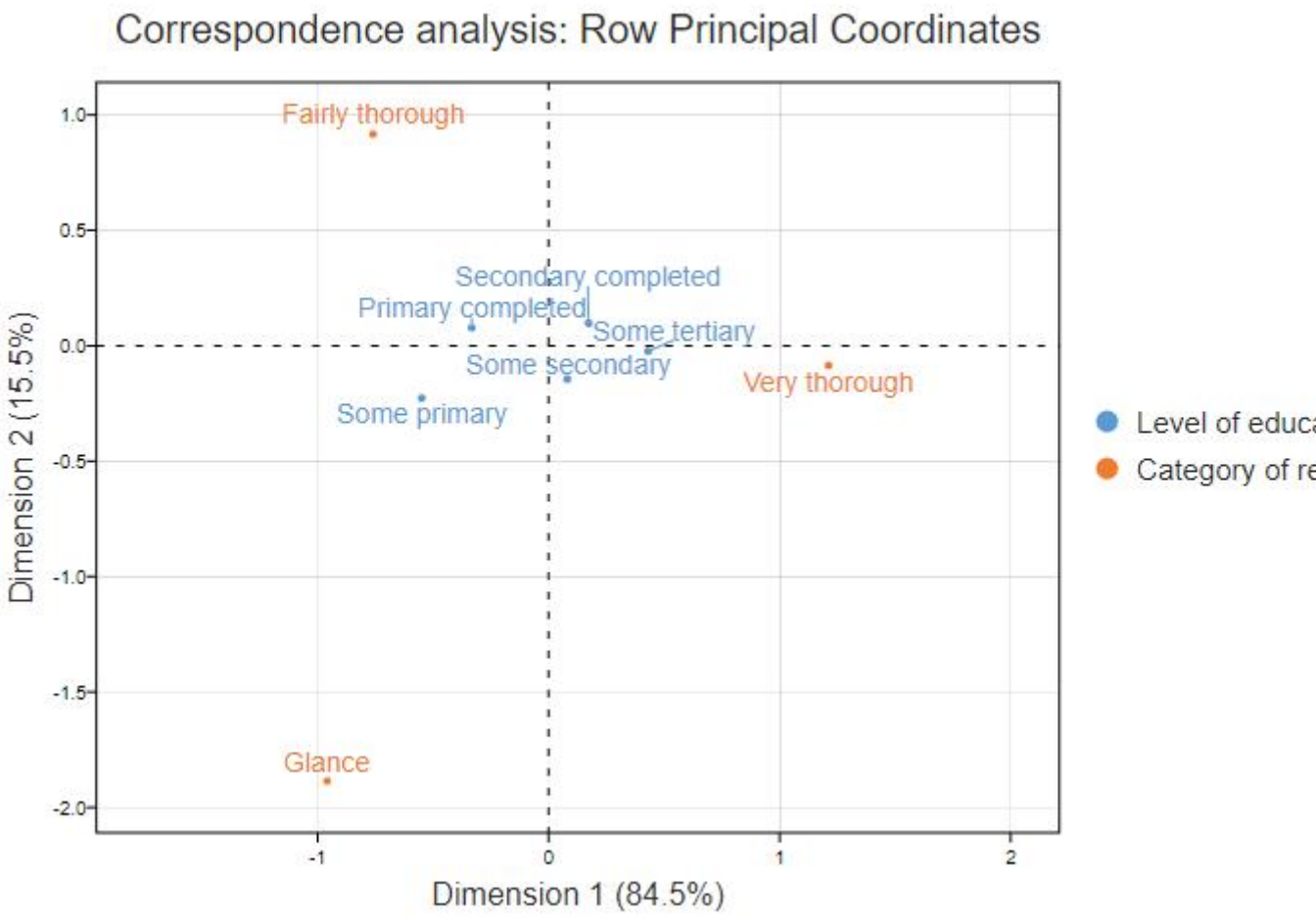
Principal Components Analysis (or PCA) is used for reducing the dimensionality of a data table with a large number of interrelated measures. Here, the original variables are converted into a new set of variables, which are known as the “Principal Components” of Principal Component Analysis.

PCA is used for the dataset that shows multicollinearity. Although least squares estimates are biased, the distance between variances and their actual value can be really large. So, PCA adds some bias and reduces standard error for the regression model.



Ø Correspondence Analysis

Correspondence Analysis using the data from a contingency table shows relative relationships between and among two different groups of variables. A contingency table is a 2D table with rows and columns as groups of variables.



Conclusion

I hope you now have a better understanding of various techniques used in Univariate, Bivariate, and Multivariate Analysis.

The media shown in this article are not owned by Analytics Vidhya and is used at the Author's discretion.