**ACTIVIDAD #1**

**Tipo actividad: Reading comprehension: "Beginner’s Guide To Exploratory Data Analysis" and Matching heading definition.**

**2) Before the reading activity, explain what reading in context is.**

Reading in context refers to the practice of understanding a written passage or text by considering the surrounding information, including the sentences and paragraphs that come before and after a particular segment. It involves looking at the broader context to comprehend the meaning of specific words, phrases, or ideas within the text.

When reading in context, readers pay attention to the overall flow of the information, the relationships between different parts of the text, and the background details that provide a deeper understanding of the subject matter. This approach helps readers extract meaning from the text more accurately and interpret information within the framework of the entire passage.

**3) Socialize key words about: "Beginner’s Guide To Exploratory Data Analysis"**

1. **Exploratory Data Analysis (EDA):** The critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions using summary statistics and graphical representations.

2. **Dichotomous Variable:** A variable that takes only one out of two possible values when measured, such as gender (male/female).

3. **Continuous Variable:** A variable that has an infinite number of possible values, such as temperature.

4. **Univariate Analysis:** The simplest form of analyzing data, focusing on one feature from a set of different features.

5. **Descriptive Statistics:** Statistical measures like mean, median, mode, range, standard deviation, and interquartile range used to describe the central tendency, variability, and distribution of data.

6. **Box Plot:** A graphical representation illustrating the central tendency, variability, and skewness of a distribution, especially useful for detecting outliers.

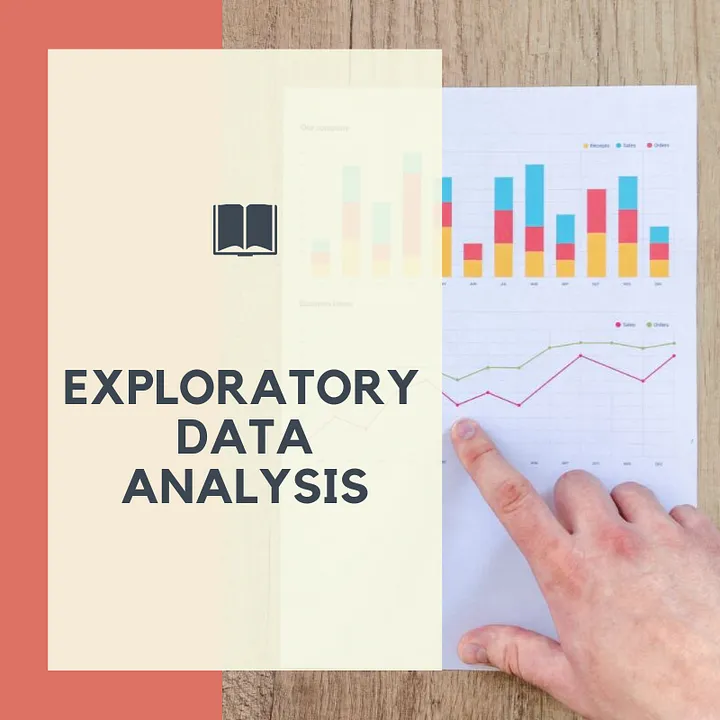
7. **Correlation:** A measure of the relationship between multiple values or features in a dataset, indicating positive, negative, or no correlation.

8. **Scatter Plot:** A graphical representation of individual points to show the relationship between two variables, useful for detecting outliers and extreme values.

**3) Reading comprehension: "Beginner’s Guide To Exploratory Data Analysis"**

**Beginner’s Guide To Exploratory Data Analysis**

Includes a simple case study for better understanding



**TABLE OF CONTENTS:**

INTRODUCTION

TYPES OF DATA

UNIVARIATE ANALYSIS

BIVARIATE ANALYSIS

CONCLUSION

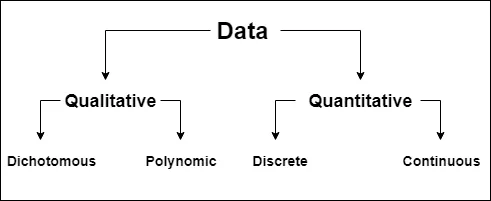
**1. INTRODUCTION:**

Suppose you are looking to book a flight ticket for a trip of yours. Now, you will not go directly to a specific site and book the first ticket that you see. You’ll first search for the tickets on multiple websites on multiple airline service providers. You will then compare the cost of the tickets with the services they are providing. Is there free WiFi available? Are breakfast and lunch complimentary? Is the overall rating of the airlines better than the others?

Whatever measures you will take from thinking about buying a ticket and finding the best ticket option for you and booking it is called “Data Analysis”. The formal definition of Exploratory Data Analysis can be given as:

Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypotheses and to check assumptions with the help of summary statistics and graphical representations.

**2. TYPES OF DATA:**



Types of Data (Image by author)

Dichotomous Variable: A dichotomous variable is a variable that takes only one out of two possible values when measured. For eg. Gender: male/female.

Polynomic Variable: A polynomic variable is a variable that has multiple values to choose from. For eg. Educational Qualifications: Uneducated/ Undergraduate/ Postgraduate/ Doctoral, etc.

Discrete Variable: Discrete variables are countable variables. For eg. your bank balance, no. of employees in an organization, etc.

Continuous Variable: A continuous variable is a variable that has an infinite no. of possible values. Any kind of measure is a continuous variable. For eg. Temperature is a continuous variable. The temperature of a particular area can be described as 30 °C, 30.2 °C, 30.22 °C, 30.221 °C, and so on.

So you must be thinking why do we need to know these different types of data in order to do its analysis? The answer is, statistical methods are designed to work with certain types of data and not others. Many of the methods that one uses to analyze continuous data may not be applicable for categorical data. If you do not know the type of data, you could possibly be providing a wrong analysis of that data! In statistics, there are tons of possible analyses that can be done. Knowing the type of data available to us limits the possibilities and helps us choose the most suitable analysis for that data.

Now that we understand the different types of data, let's dig into Univariate Analysis. Before applying any sort of Machine Learning algorithms, it is very important to understand the type of data you’re dealing with. What kind of features (columns) does the data have? Is it numeric or categorical data? To understand the data, analyzing the data is very important.

**3. UNIVARIATE ANALYSIS:**

Univariate analysis is the simplest form of analyzing the data. It is the analysis of only one feature from a set of different features. It normally gives us a description of a particular feature. It doesn’t deal with cause or relationship, it just takes the data, summarizes it, and represents it in the form of Histograms, Box and Whisker Plots, etc. We’ll be learning about this in the further sections.

**3.a DESCRIPTIVE STATISTICS:**

**Central Tendency:** Central tendency refers to the location of a distribution. It represents a typical value that we normally expect from the participants. We have three choices to describe the central tendency of a distribution.

**Mean:** Mean is also known as the arithmetic mean and calculates the average of all the values in the distribution. The value of the mean is prone to outliers.

**Median (Q2):** The median is the Fiftieth percentile of a distribution. If the distribution has an odd no of values, the middle element is considered and if it has an even no. of values, the average of two middle elements is considered.

**Mode:** Mode is the most frequent value that occurs in the distribution. There can be one or more modes in a data distribution.

Even though measures of central tendency tell us about the center of the distribution, it does not tell us about every value in the distribution. Participants can have values much lower than the mean or substantially higher than the mean. So we should also know about a distribution’s variability.

**Variance:**

The variance of a distribution is specified using the range, standard deviation, and interquartile range.

**Range:** The range is nothing but the difference between the maximum value and minimum value in a distribution.

**Standard Deviation:** Standard Deviation is a quantity that describes how much the members of a group differ from a mean value of the group.

**Interquartile Range:** Interquartile Range is the difference between the third quartile (Q3) and the first quartile.

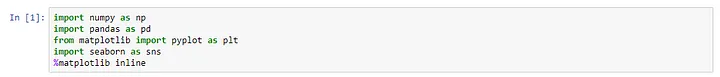
**a. First Quartile (Q1):** The first quartile is defined as the middle value between the smallest number and the median of the dataset.

**b. Second Quartile (Q2):** The second quartile is nothing but the median of the dataset.

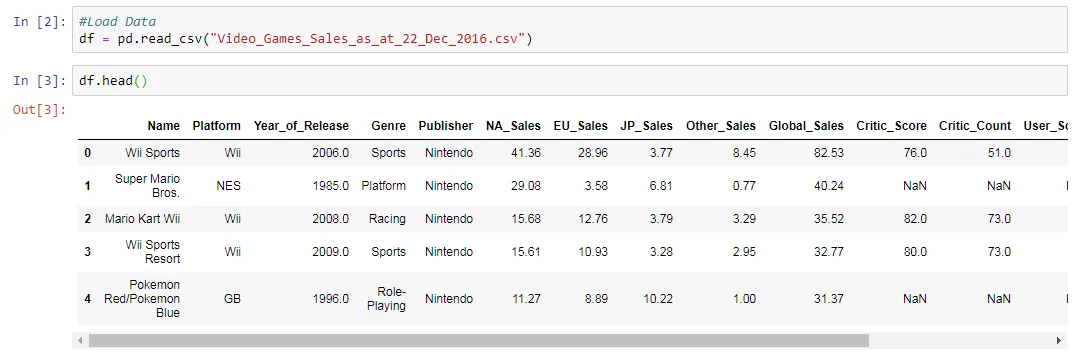
**c. Third Quartile (Q3):** The third quartile is the middle value between the median and the highest value of the dataset.

Lets now understand the different charts and graphs that can be used to describe univariate data with the help of an example. The dataset used is “Video Games Sales with Ratings” and can be found here.

First, let us import the required libraries.



Loading the data:



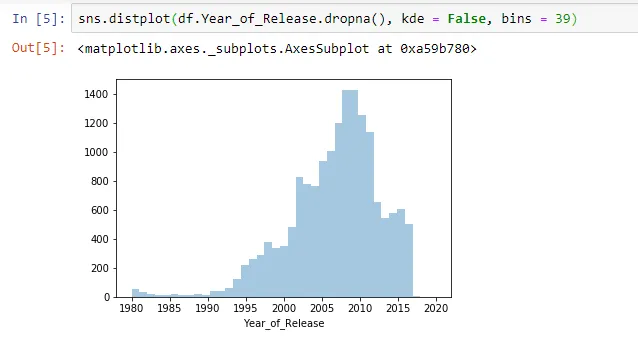
**Five-point summary of all the numeric data:**

Here we can see some of the descriptive statistics such as mean, standard deviation, minimum value, Q1, Q2, Q3, and maximum value of all the numeric data.



**3.b HISTOGRAM:**

We can learn more about the numeric data using a histogram. Mathematical scores have a unimodal and symmetric distribution. A histogram is a good way to illustrate the central tendency, variability, and shape of a distribution. It is also a good way to identify multiple modes if they exist. However, a histogram is not a good way to identify outliers.

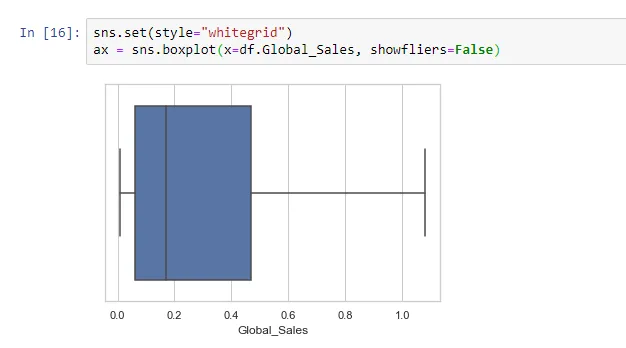


The above histogram shows when a particular game was released. We can see here that the maximum games were launched between the years 2005 and 2012 and then gradually went down.

**3.c BOX PLOT:**

Box plot is an alternative and more robust way to illustrate a continuous variable. The vertical lines in the box plot have a specific meaning. The centerline in the box is the 50th percentile of the data (median). Variability is represented by a box that is formed by marking the first and third quartile. This box represents the interquartile range. Whiskers extend from the box to the left and right. When there are no outliers, the left and right fences represent the minimum and maximum value. When there are any outliers present, it is represented using small circles or stars outside fences.

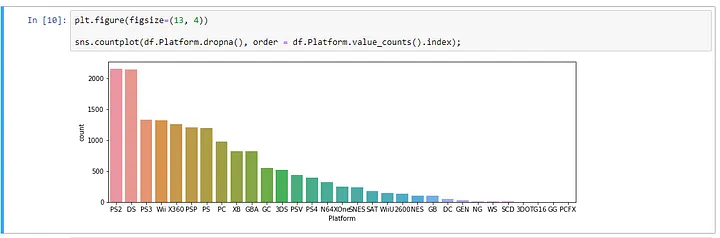
A box plot is a useful way to illustrate the central tendency, variability, and skewness of a distribution. It is also an excellent way to detect outliers and extreme values. Histogram and Box Plots both have certain limitations and restrictions. Hence you should plot both the histogram and box plot when exploring your data but choose only one for the report which gives us better insights about the data. If you have a unimodal distribution with outliers use a box plot, if you have a bimodal distribution use a histogram.



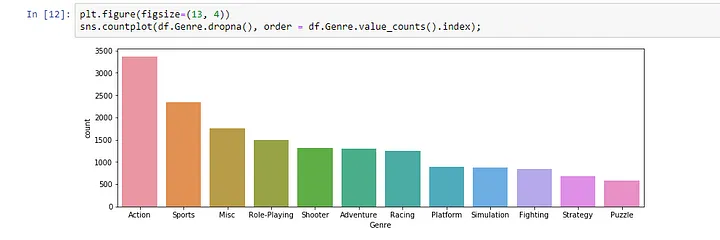
**3.d CATEGORICAL DATA — BAR GRAPH:**

This is a bar graph which shows the different platforms used by the customers to play games. We can see that most customers use PS2 as their console and fewer customers use consoles such as GG or PCFX.

The bar graph is a graphical representation of the information on the table. A key feature of the bar chart is that the bar charts do not touch each other. This characteristic indicates that the values are categorical and not continuous. It is an appropriate way to plot categorical data.



Here’s another example of a bar plot. This plot shows the different genres played by customers. The most favored genre is ‘Action’ whereas the least favored is ‘Puzzle’.



4. BIVARIATE ANALYSIS:

In EDA, we often learn about the relationship between two variables. Some questions can be asked such as:

How correlated is one feature with another feature?

Does a lower value on one variable corresponds to a lower value on another variable?

Does a higher value on one variable corresponds to a higher value on another variable?

What kind of relationship do the two features follow?

Describing the relationship between two variables is more complicated than the methods we have already discussed because we must consider the type of data in both the variables. There might be different combinations of data such as one might be continuous and the other might be categorical or both might be continuous or both might be categorical. These combinations result in different statistical and graphical summaries. The necessary part of the analysis is that both the variables must have some kind of values present in them. If anyone of the variables is missing any value, they cannot be included in the analysis.

Now let’s look at some of the graphical summaries that can be made in bivariate analysis.

4.a CORRELATION:

Data correlation is a way to understand the relationship between multiple values or features in your dataset. Correlation can help in predicting one attribute with the help of another (This can be used to impute missing values in a dataset).

There are three different types of correlations:

Positive Correlation: A correlation is considered to be positive if one feature is directly proportional to the change in the value of another feature.

For eg: If the value of feature A increases then the value of feature B also increases or if the value of feature A decreases then the value of feature B also decreases.

2. Negative Correlation: A correlation is considered to be negative if one feature is inversely proportional to the change in the value of another feature.

For eg: If the value of feature A increases then the value of feature B decreases and vice versa.

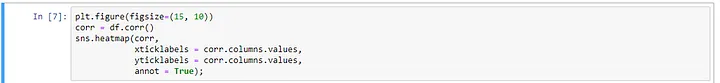
3. No Correlation: There exists no relationship between the two attributes.

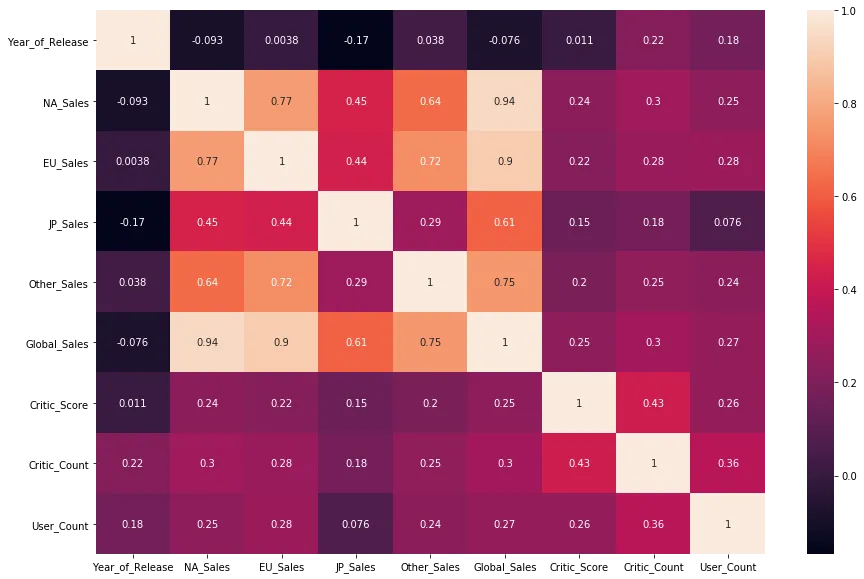
Each of these correlation types exists in a spectrum represented by values from -1 to +1 where slight or high positive correlation features can be like 0.5 or 0.7. A very strong and perfect positive correlation is represented by a correlation score of 0.9 or 1. If there is a strong negative correlation, it will be represented by a value of -0.9 or -1. Values close to zero indicates no correlation.

The reason why we should understand the correlation between features is because of a concept called “Multicollinearity”. If a dataset has a perfectly positive or negative correlated attribute then the model can be impacted by multicollinearity. Multicollinearity happens when one predictor variable can be linearly predicted from the others with a high degree of accuracy. While you might think this is a nice thing to happen, it is not as it can lead to misleading or skewed results.

Some of the ways to deal with this problem are to delete one of the perfectly correlated features or use a dimension reduction algorithm such as Principal Component Analysis (PCA).

The below code shows how we can visualize the correlation between features in a matrix form.



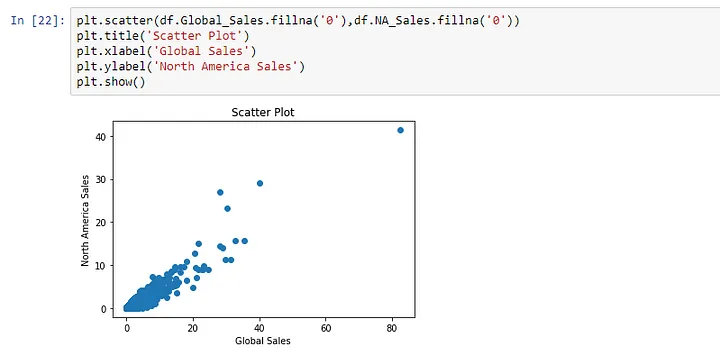


Here we can see that there are some features with slight positive or negative correlation but no perfectly positive or negative correlation. The value of diagonal elements will always be 1 as they are compared against the same feature.

**4.b SCATTER PLOT:**

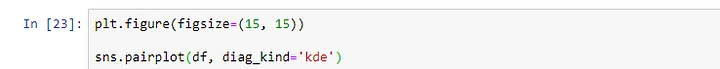
In the scatter plot, the points are represented individually with a dot, circle, or other shapes. Outliers and extreme values can be easily detected using scatter plots.

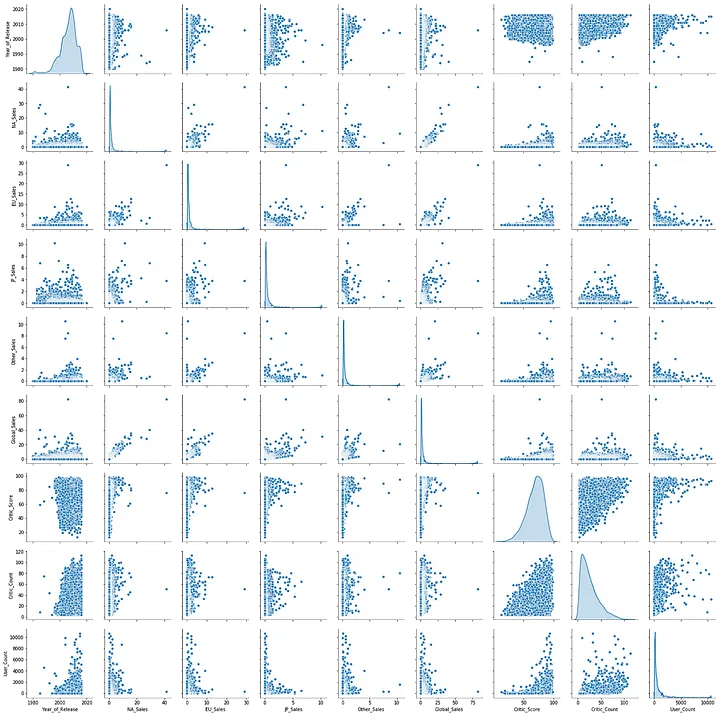
In the figure below, a scatter plot is plotted between Global Sales and North American Sales. We can spot one outlier on the extreme top right.



**4.c PAIRPLOT:**

Instead of plotting to scatter plots individually, we can also plot a pair plot. It’ll plot a scatter plot for every feature with every other feature. Just take note that the default seaborn library only plots pair plots of numerical figures. The histogram on the diagonal allows us to see the distribution of a single variable while the scatter plots on the upper and lower triangles show the relationship (or lack thereof) between two variables.





5. CONCLUSION:

This was just a small introduction to Exploratory Data Analysis. Instead of introducing you to various other types of plots and graphical summaries, my objective was to make you understand what kind of analysis should be done on different types of data. After reading this I’d recommend you pick up any dataset of your choice and perform EDA all by yourself.

Taken from: <https://towardsdatascience.com/exploratory-data-analysis-dcb5e7189c4e>

**4) Matching heading definition.**

Match the following headings with their corresponding definitions from the text.

**Headings:**

1. Introduction to Exploratory Data Analysis (EDA)

2. Types of Data

3. Univariate Analysis

4. Descriptive Statistics

5. Box Plot and Histogram

6. Categorical Data – Bar Graph

7. Bivariate Analysis

8. Conclusion

**Definitions:**

a. The critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypotheses, and check assumptions using summary statistics and graphical representations.

b. A variable that takes only one out of two possible values when measured, such as gender (male/female).

c. The simplest form of analyzing data, focusing on one feature from a set of different features.

d. Statistical measures like mean, median, mode, range, standard deviation, and interquartile range used to describe the central tendency, variability, and distribution of data.

e. A graphical representation illustrating the central tendency, variability, and skewness of a distribution, especially useful for detecting outliers.

f. A measure of the relationship between multiple values or features in a dataset, indicating positive, negative, or no correlation.

g. A bar graph illustrating the different platforms or categories used by customers, providing a graphical representation of categorical data.

h. The concluding remarks summarizing the key concepts and emphasizing the importance of Exploratory Data Analysis.

**5) Socialize key words about: "Real-World Applications of Exploratory Analytics"**

1. **Exploratory Data Analytics (EDA):** Exploratory Data Analytics is an approach to analyzing and exploring datasets to summarize their dominant characteristics, uncover patterns, and gain insights. Developed by John Tukey in 1977, EDA is widely used across various industries.

2. **Data Mining:** Data Mining refers to the process of discovering patterns, anomalies, and useful information from large datasets. It involves the use of statistical techniques, machine learning, and other analytical methods to extract valuable knowledge.

3. **Machine Learning:** Machine Learning is a subset of artificial intelligence that focuses on developing algorithms and models that enable computers to learn and make predictions or decisions without explicit programming. EDA often integrates machine learning tools for data analysis.

4. **Geospatial Analysis:** Geospatial Analysis involves examining and interpreting data that has a geographic or spatial component. It helps in revealing spatial relationships, patterns, and trends, making it valuable in various fields, including urban planning, logistics, and environmental science.

5. **Data Visualization:** Data Visualization is the representation of data in graphical or visual formats, such as charts, graphs, and maps. It aids in the exploration and communication of complex datasets, making patterns and insights more accessible to analysts and stakeholders.