Introduction to data analysis, data analysis process, and clustering in machine learning

Prepared By

Mohamed Eldosouky Ahmed Elamrossy

Fourth grade statistics department

Supervised by

Dr. Ahmed Abo zaid Elbanna

Submitted to Mathematics department, Faculty of science, Tanta university

Table of Contents

Chapter 1

Introduction to data analysis

1.1 Statistics in data analysis

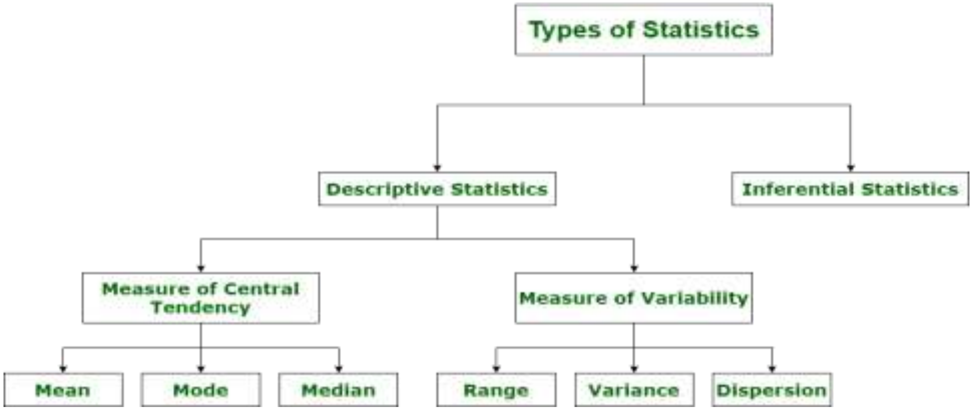
Statistics simply means numerical data, and is field of math that generally deals with collection of data, tabulation, and interpretation of numerical data. It is actually a form of mathematical analysis that uses different quantitative models to produce a set of experimental data or studies of real life. It is an area of applied mathematics concern with data collection analysis, interpretation, and presentation. Statistics deals with how data can be used to solve complex problems. Some people consider statistics to be a distinct mathematical science rather than a branch of mathematics.

Statistics makes work easy and simple and provides a clear and clean picture of work you do on a regular basis.

1.2 Basic terminology of Statistics:

- Population
  It is actually a collection of set of individuals or objects or events whose properties are to be analyzed.
- Sample
  It is the subset of a population.

1.3 Types of Statistics:



1.3.1 Descriptive Statistics :

Descriptive statistics uses data that provides a description of the population either through numerical calculation or graph or table. It provides a graphical summary of data. It is simply used for summarizing objects, etc. There are two categories in this as following below.

1.3.2 inferential statistics

Descriptive Statistics describes data (for example, a chart or graph) and inferential statistics allows you to make predictions ("inferences") from that data. With inferential statistics, you take data from samples and make generalizations about a population.

For example, you might stand in a mall and ask a sample of 100 people if they like shopping at Sears. You could make a bar chart of yes or no answers (that would be descriptive statistics) or you could use your research (and inferential statistics) to reason that around 75-80% of the population (all shoppers in all malls) like shopping at Sears.

1.4 what is data?

- Data are individual facts, statistics, or items of information which may be a text, videos, spreadsheets, databases, pictures, voices and there are many types of data.

1.5 How can we make data useful ?

- Using data is the new way which the world do now .
- Data are used in scientific research, businesses management (e.g., sales data, revenue, profits, stock price), finance, governance (e.g., crime rates, unemployment rates, literacy rates), and in virtually every other form of human organizational activity (e.g., censuses of the number of homeless people by non-profit organizations).
- data are atoms of decision making: they are the smallest units of factual information that can be used as a basis for reasoning, discussion, or calculation. Data can range from abstract ideas to concrete measurements, even statistics. Data are measured, collected, reported, and analyzed, and used to create data visualizations such as graphs, tables or images. Data as a

general concept refers to the fact that some existing information or knowledge is *represented* or *coded* in some form suitable for better usage or processing.

1.6 What are the types of data ?

There are two types of data :

- Qualitative data
- Quantitative data

1.6.1 Qualitative data

Qualitative data is a bunch of information that cannot be measured in the form of numbers. It is also known as categorical data. It normally comprises words, narratives, and we labelled them with names.

It delivers information about the qualities of things in data. The outcome of qualitative data analysis can come in the type of featuring key words, extracting data, and ideas elaboration.

For examples:

- Hair colour - black, brown, red
- Opinion- agree, disagree, neutral
- Martial status ,single or married
- Rating of survey

Qualitative data divided to two types

- Nominal data : which has no order like hair colour
- Ordinal data : which be ordered like rating of survey

1.6.2 Quantitative data

Quantitative data is a bunch of information gathered from a group of individuals and includes statistical data analysis. Numerical data is another name for quantitative data. Simply, it gives information about quantities of items in the data and the items that can be estimated. And, we can formulate them in terms of numbers.

For examples:

- We can measure the height (1.70 meters), distance (1.35 miles) with the help of a ruler or tape.
- We can measure water (1.5 litres) with a jug.

There are two types of Quantitative data :

- Discrete data
- Continuous data

1.6.2.1 Discrete data:

Discrete data is a count that involves integers — only a limited number of values is possible. This type of data cannot be subdivided into different parts. Discrete data includes discrete variables that are finite, numeric, countable, and non-negative integers. In many cases, discrete data can be prefixed with “the number of”. For example:

The number of students who have attended the class;

The number of customers who have bought different products;

The number of groceries people are purchasing every day;

This data type is mainly used for simple statistical analysis because it’s easy to summarize and compute. In most of the practices, discrete data is displayed by bar graphs, stem-and-leaf-plot and pie charts.

1.6.2.2 Continuous data:

Continuous data is considered the complete opposite of discrete data. It’s the type of numerical data that refers to the unspecified number of possible measurements between two presumed points.

The numbers of continuous data are not always clean and integers, as they are usually collected from very precise measurements. Measuring a particular subject is allowing for creating a defined range to collect more data.

Variables in continuous data sets often carry decimal points, with the number stretching out as far as possible. Typically, it changes over time. It can have completely different values at different time intervals, which might not always be whole numbers. Here are some examples:

The weather temperature;

The wind speed;

The weight of the kids;

Continuous data can be measured by using specific tools and displayed in line graphs, skews, histograms.

1.6.2.3 Discrete data .vs. continuous data:

Both data types are important for statistical analysis. However, some major differences need to be noted before drawing any conclusions or making decisions. The key differences are:

Discrete data is the type of data that has clear spaces between values. Continuous data is data that falls in a constant sequence.

Discrete data is countable while continuous — measurable.

To accurately represent discrete data, the bar graph is used. Histogram or line graphs are used to represent continuous data graphically. A diagram of the discrete function shows a distinct point that remains unconnected. While in a continuous function graph, the points are connected with an unbroken line.

Discrete data contains distinct or separate values. Continuous data includes any value within the preferred range.

1.6.2.4 The importance of discrete and continuous data:

Both discrete and continuous data are valuable for all sorts of data-driven decisions. Valuable research and insights are made by combining both sets of data. Here are some examples where discrete and continuous data can be used:

- Marketing and advertising
- Research
- Population analysis
- Product development

1.7 When we analyze the continuous and discrete data we focus on four sides:

1. Center
2. Spread
3. Shape
4. outliers

1.7.1 The Center

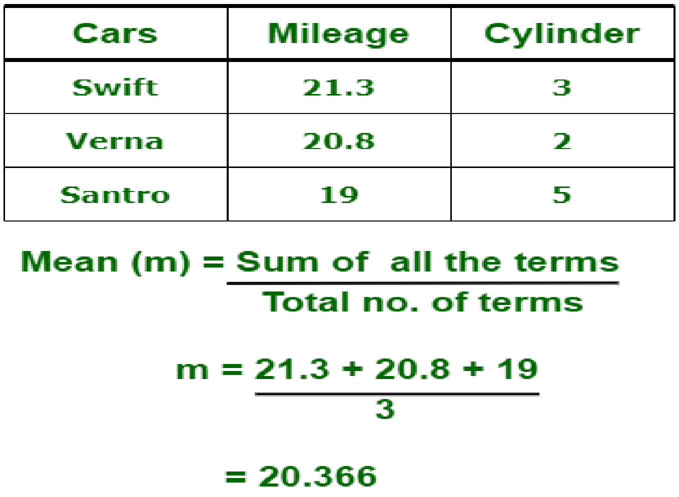
There are three main measurements called **(**Measure of central tendency)

- mean
- median
- mode

1.7.1.1 The mean:

It is measure of average of all value in a sample set and counted by dividing the sum of all terms by the total number of terms

For example



Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5

$$\text{Mean (m)} = \frac{\text{Sum of all the terms}}{\text{Total no. of terms}}$$

$$m = \frac{21.3 + 20.8 + 19}{3}$$

$$= 20.366$$

Note that:

The mean is not the best measure of tendency all time, because it can be neglected in some times

For example:

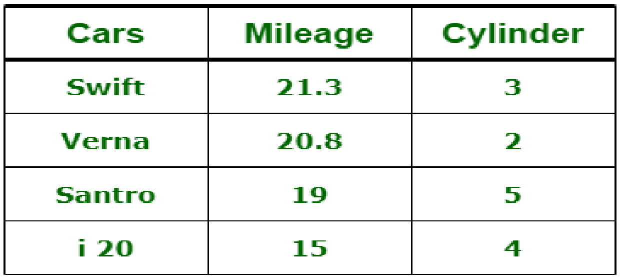
We want to get information about people who attend the match, if we use the mean we may get float number like 1500,353 and this result is not logical for humans.

In this case we can use the median which we will explain.

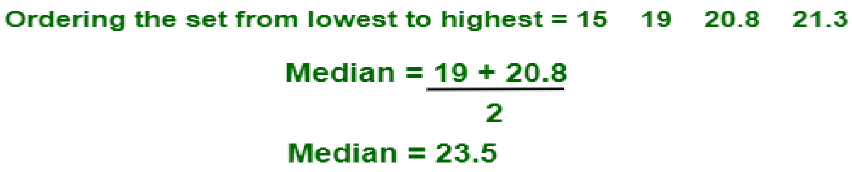
1.7.1.2 The median

- It is measure of central value of a sample set. In these, data set is ordered from lowest to highest value and then finds exact middle.
- Median divides the data to 50% and 50%

For example



Cars	Mileage	Cylinder
Swift	21.3	3
Verna	20.8	2
Santro	19	5
i 20	15	4



Ordering the set from lowest to highest = 15 19 20.8 21.3

$$\text{Median} = \frac{19 + 20.8}{2}$$

Median = 23.5

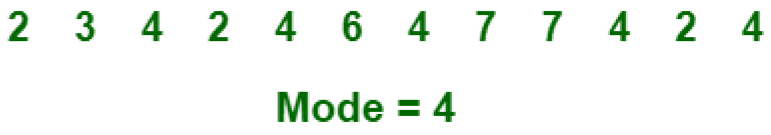
Note that

1. If the size of data is odd number then the median equaled by $\frac{x_{n+1}}{2}$
2. If the size of data is even number then the median equaled by $\frac{x_{\frac{n}{2}}+x_{\frac{n}{2}+1}}{2}$

1.7.1.3 The mode

It is value most frequently arrived in sample set. The value repeated most of time in central set is actually mode.

For example



2 3 4 2 4 6 4 7 7 4 2 4

Mode = 4

1.7.2 spread

There are four measures of spread which measure how far are points from one to onther

- Range
- Five number summary
- Variance
- Standard deviation

1.7.2.1 Range:

It is given measure of how to spread apart values in sample set or data set

Range = Maximum value - Minimum value

1.7.2.2 Five number summary:

In statistics, the five-number summary is mostly used as it gives a rough idea about the dataset. It is basically a summary of the dataset describing some key features in statistics. The five key features are:

1. Minimum value: It is the minimum value in the data set
2. First Quartile, Q1: It is also known as the lower quartile where 25% of the scores fall below it.
3. Median (middle value) or second quartile: It is basically the mid-value in the dataset.
4. Third Quartile, Q3: It is also known as the Upper quartile in which 25% of the data is above it and the rest 75% falls below it.
5. Maximum value: It is the maximum value in the dataset.

1.7.2.3 Variance:

Is the sum of squares of differences between all numbers and means. Deviation for above example. First, calculate the deviations of each data point from the mean, and square the result then divide the result by number of all values by the equation:

$$\sigma^2 = \frac{\Sigma\,(x - \mu)^2}{N}$$ Population Variance

$$s^2 = \frac{\Sigma\,(x - \bar{x})^2}{n - 1}$$ Sample Variance

1.7.2.4 Standard Deviation:

Is square root of variance. It is a measure of the extent to which data varies from the mean.

Std equaled by equation:

Sample

$$S = \sqrt{\frac{\Sigma(x_i - \bar{x})^2}{n - 1}}$$

Population

$$\sigma = \sqrt{\frac{\Sigma(x_i - \mu)^2}{N}}$$

Note that

- If the standard deviation equal zero that mean that all values of data are equal
- Researchers prefer to use std because it measure by the main unit of measure
- Variance square the unit of measure

1.7.3 shape

Measures of shape describe the distribution (or pattern) of the data within a dataset.

- The distribution shape of quantitative data can be described as there is a logical order to the values, and the 'low' and 'high' end values on the x-axis of the histogram are able to be identified.

- The distribution shape of a qualitative data cannot be described as the data are not numeric.

1.7.3.1 Why are measures of shape useful?

- The shape of the distribution can assist with identifying other descriptive statistics, such as which measure of central tendency is appropriate to use.
- If the data are normally distributed, the mean, median and mode are all equal, and therefore are all appropriate measure of central tendency.
- If data are skewed, the median may be a more appropriate measure of central tendency.

1.7.3.2 What are types of shapes?

There are many types of shapes which we can use it to visualize our data like:

- Histogram
- Bar chart
- Pie chart
- Box plot
- Scatter plot
- …

The job of every shape depend on the data type where:

1. If the data is quantitative we can use histogram or box plot
2. If the data is qualitative we can use bar chart or pie chart
3. If we compare between two quantitative variables we can use scatter blot

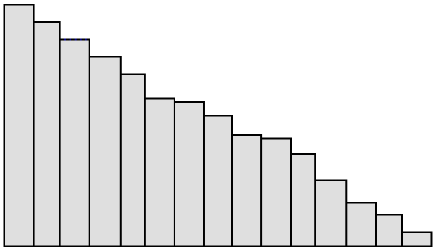
Now we display some figures to explain shapes:

1.7.3.2.1 Histogram:

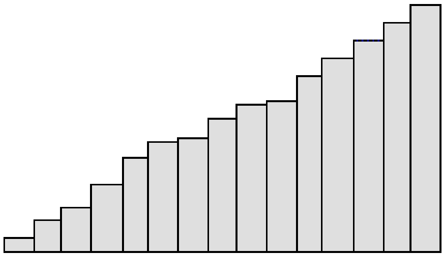
A histogram is the most commonly used graph to show frequency distributions. It looks very much like a bar chart, but there are important differences between them.

There are three types of histogram

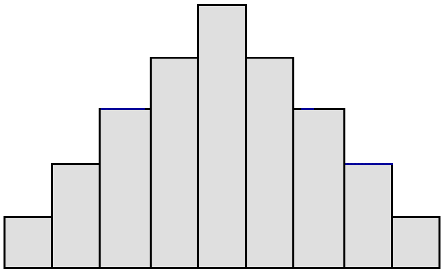
- Right skewed



- Left skewed



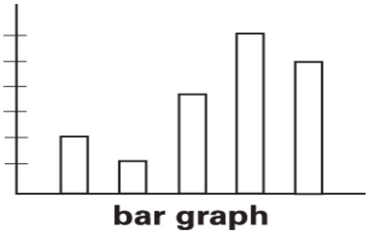
- Symmetric (the bell curve)



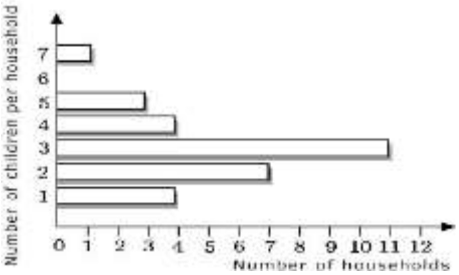
1.7.3.2.2 Bar chart:

A bar chart or a graph is which present the data in the categorical form with rectangular bars with heights or lengths proportional to the values which represent in a graph. The bars can be plotted vertically or horizontally.

- If it plotted vertically

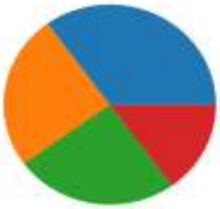


- If it plotted horizontally



1.7.3.2.3 Pie chart:

Pie Chart is a circular chart that shows the categorical data in circular slices

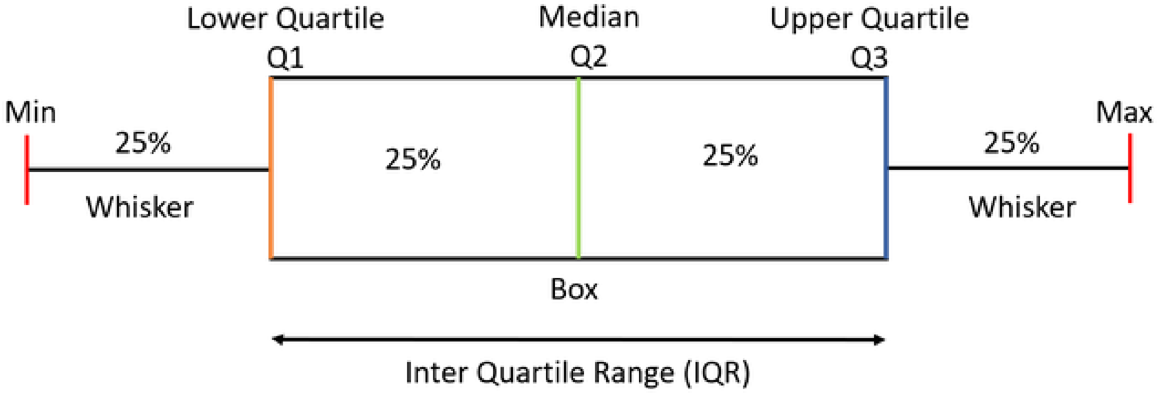


1.7.3.2.4 Box plot:

Box Plot It is a type of chart that depicts a group of numerical data through their quartiles. It is a simple way to visualize the shape of our data. It makes comparing characteristics of data between categories very easy.

A box plot gives a five-number summary of a set of data which is-

- Minimum – It is the minimum value in the dataset excluding the outliers
- First Quartile (Q1) – 25% of the data lies below the First (lower) Quartile.
- Median (Q2) – It is the mid-point of the dataset. Half of the values lie below it and half above.
- Third Quartile (Q3) – 75% of the data lies below the Third (Upper) Quartile.
- Maximum – It is the maximum value in the dataset excluding the outliers.



Note that:

- whisker here refer to variance
- The box plot shown in the above diagram is a perfect plot with no skewness. The plots can have skewness and the median might not be at the center of the box.
- The area inside the box (50% of the data) is known as the Inter Quartile Range. The IQR is calculated as

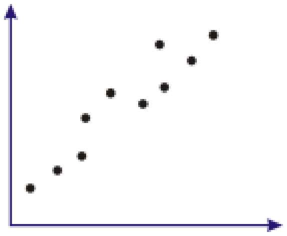
  IQR = Q3-Q1

1.7.3.2.5 Scatter plot:

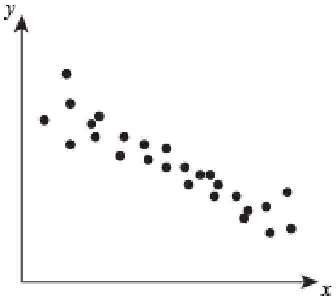
Scatter plot are those charts in which data points are represented horizontally and on vertical axis to show that how one variable affect on another variable used to observe relationships between variables.

There are two types of relationships

1- Positive correlation



2- Negative correlation



There are many details in shapes but we have explained a good summary about it

Now we will explain the fourth side

1.7.4 outlires

To explain the outlires in the data we will give an example, if we want to get the average salary of all employers in facebook company , and after we got it we noticed that the average is a huge number which is not logical for every employee.

What is the problem here?

The problem is the salary of the owner Mark Zuckerberg which make the average of salary is huge and not logical.

Well:

Here the salary of Mark represent an **outlier**

Then what should we do with this outliers?

1- Plot your data
2- Ask some questions about it like
   - shall me remove it?
   - Shall me fix it?
   - Shall me keep it?
3- If the data is normally distributed then we can use mean and std to explore it
4- If the data is skewed then the five number summary is effective than mean and std
5- Note that outliers affect on mean more than median so try to use median

Until here we finish the chapter one about data analysis, in the next chapter we will take more about data analysis process.

Chapter 2

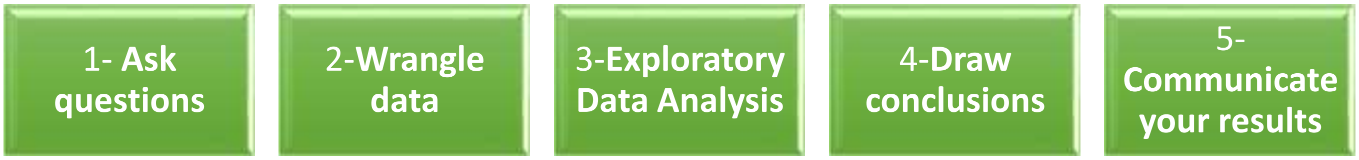
Data analysis process

2.1 What is data analysis?

Data analysis is the process of cleaning, changing, and processing raw data, and extracting actionable, relevant information that helps businesses make informed decisions. The procedure helps reduce the risks inherent in decision-making by providing useful insights and statistics, often presented in charts, images, tables, and graphs.

2.2 Data analysis process steps:

We organized the data analysis process into five steps: Question, Wrangle, Explore, Draw Conclusions, and Communicate and we will practice each step as we go through the next sections:



2.2.1 Step 1: ask questions

- Either you're given data and ask questions based on it, or you ask questions first and gather data based on that later. In both cases, great questions help you focus on relevant parts of your data and direct your analysis towards meaningful insights.

2.2.2 Step 2: wrangle data

- You get the data you need in a form you can work with in three steps: gather, assess, clean. You gather the data you need to answer your questions, assess your data to identify any problems in your data’s quality or structure, and clean your data by modifying, replacing, or removing data to ensure that your dataset is of the highest quality and as well-structured as possible.

2.2.3 Step 3: EDA (exploratory data analysis)

- You explore and then augment your data to maximize the potential of your analyses, visualizations, and models. Exploring involves finding patterns in your data, visualizing relationships in your data, and building intuition about what you're working with. After exploring, you can do things like remove outliers and create better features from your data, also known as feature engineering.

2.2.4 Step 4: draw conclusions

- This step is typically approached with machine learning or inferential statistics that are beyond the scope of this course, which will focus on drawing conclusions with descriptive statistics

2.2.5 Step 5: communicate your results

- You often need to justify and convey meaning in the insights you've found. Or, if your end goal is to build a system, you usually need to share what you've built, explain how you reached design decisions, and report how well it performs. There are many ways to communicate your results: reports, slide decks, blog posts, emails, presentations, or even conversations. Data visualization will always be very valuable.

2.3 Why is data analysis important?

Here is a list of reasons why data analysis is such a crucial part of doing business today.

- Better Customer Targeting: You don't want to waste your business's precious time, resources, and money putting together advertising campaigns targeted at demographic groups that have little to no interest in the goods and services you offer. Data analysis helps you see where you should be focusing your advertising efforts.
- Through data analysis, your business can get a better idea of your target audience's spending habits, disposable income, and most likely areas of interest. This data helps

businesses set prices, determine the length of ad campaigns, and even help project the quantity of goods needed.

- Reduce Operational Costs: Data analysis shows you which areas in your business need more resources and money, and which areas are not producing and thus should be scaled back or eliminated outright.
- Better Problem-Solving Methods: Informed decisions are more likely to be successful decisions. Data provides businesses with information. You can see where this progression is leading. Data analysis helps businesses make the right choices and avoid costly pitfalls.

2.4 Types of data analysis:

- Diagnostic Analysis: Diagnostic analysis answers the question, "Why did this happen?" Using insights gained from statistical analysis (more on that later!), analysts use diagnostic analysis to identify patterns in data. Ideally, the analysts find similar patterns that existed in the past, and consequently, use those solutions to resolve the present challenges hopefully.

- Predictive analysis: Predictive analysis answers the question, "What is most likely to happen?" By using patterns found in older data as well as current events, analysts predict future events. While there's no such thing as 100 percent accurate forecasting, the odds improve if the analysts have plenty of detailed information and the discipline to research it thoroughly.

- Statistical analysis: Statistical analysis answers the question, "What happened?" This analysis covers data collecting, analysis, modeling, interpretation, and presentation using dashboards. The statistical analysis breaks down into two sub-categories:
    - Descriptive: Descriptive analysis works with either complete or selections of summarized numerical data. It illustrates means and deviations in continuous data and percentages and frequencies in categorical data.

- Inferential: Inferential analysis works with samples derived from complete data. An analyst can arrive at different conclusions from the same comprehensive data set just by choosing different samplings.

2.5 Data analysis methods:

Although there are many data analysis methods available, they all fall into one of two primary types:

1. Qualitative Data Analysis
2. Quantitative Data Analysis

2.5.1 Qualitative Data Analysis:

The qualitative data analysis method derives data via words, symbols, pictures, and observations. This method doesn’t use statistics. The most common qualitative methods include:

- Content Analysis, for analyzing behavioral and verbal data.
- Narrative Analysis, for working with data culled from interviews, diaries, surveys.
- Grounded Theory, for developing causal explanations of a given event by studying and extrapolating from one or more past cases.

2.5.2 Quantitative Data Analysis:

Statistical data analysis methods collect raw data and process it into numerical data. Quantitative analysis methods include:

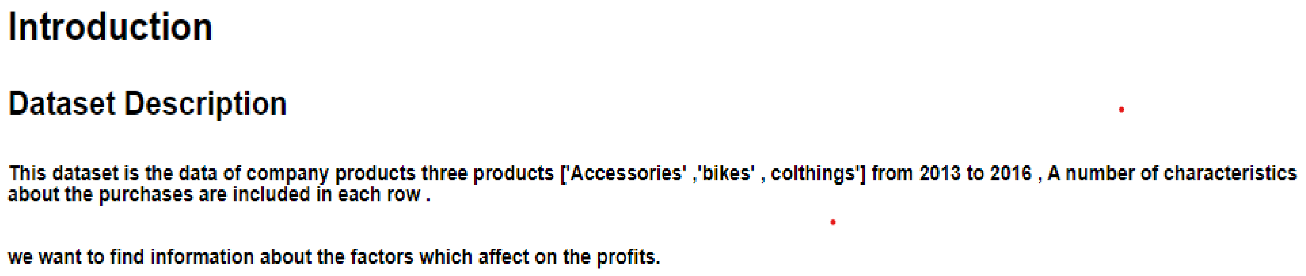
1. Hypothesis Testing, for assessing the truth of a given hypothesis or theory for a data set or demographic.
2. Mean, or average determines a subject’s overall trend by dividing the sum of a list of numbers by the number of items on the list.
3. Sample Size Determination uses a small sample taken from a larger group of people and analyzed. The results gained are considered representative of the entire body.

After this good explain of data analysis we will take a project of real data to apply the data analysis process

2.6 Applying data analysis process in real data:

company products analysis

- The project : https://drive.google.com/file/d/1uqrsvgvY7tRYWcFsg9mLu951-Llz_G7m/view?usp=sharing



Introduction

Dataset Description

This dataset is the data of company products three products ['Accessories' ,'bikes' , colthings'] from 2013 to 2016 , A number of characteristics about the purchases are included in each row .

we want to find information about the factors which affect on the profits.

Question(s) for Analysis :

1- what is the categories of accessories that company products and how many company product for each one

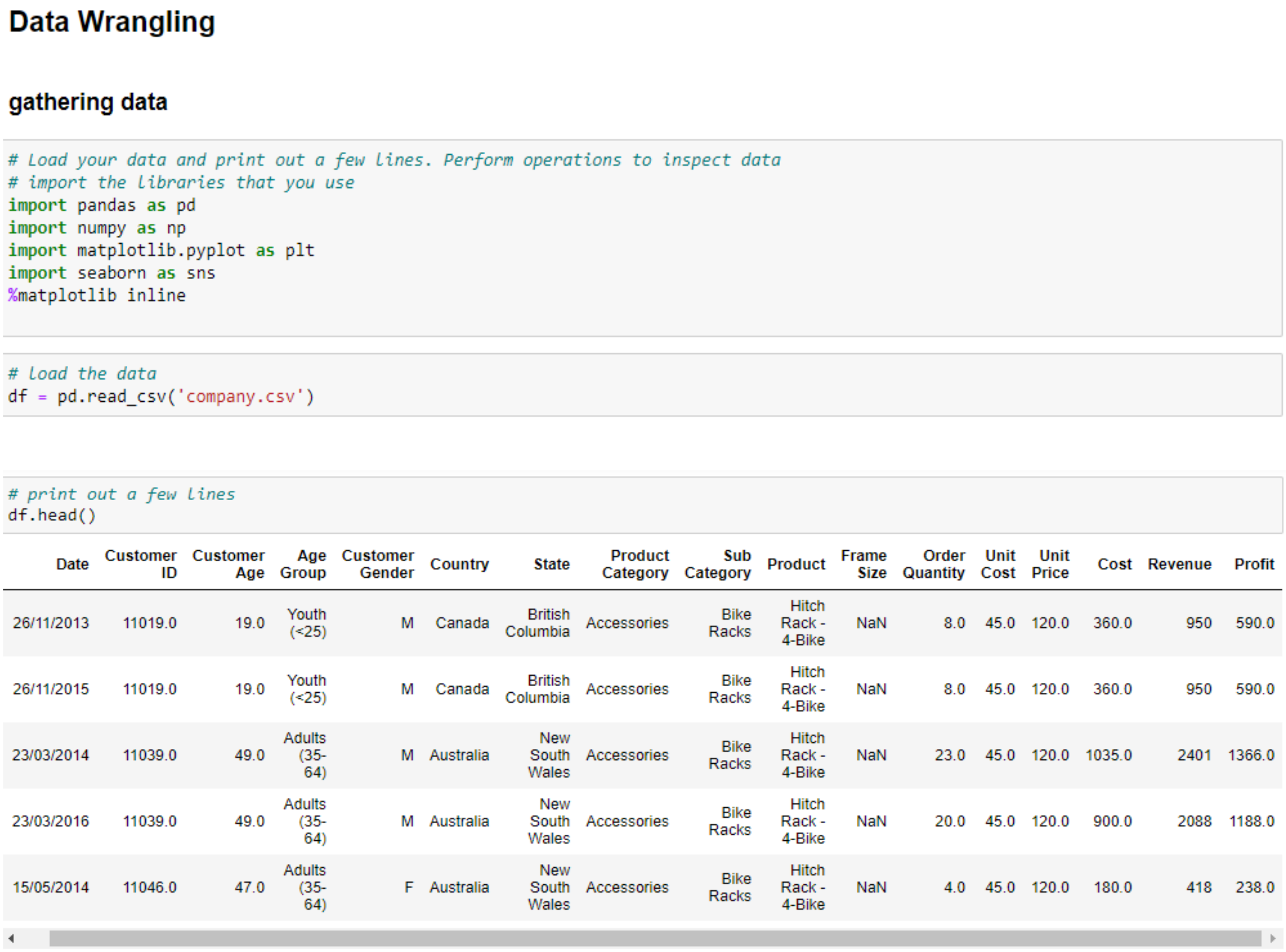
2-what is the categories of Bikes that company products and how many company product for each one

3-what is the categories of Clothing that company products and how many company product for each one

4- what is the best areas which give us a max profit and what is there products

5-what is the best 10 product has the big profit

6-what is the best age group which give us a max profit and what is there products



Data Wrangling

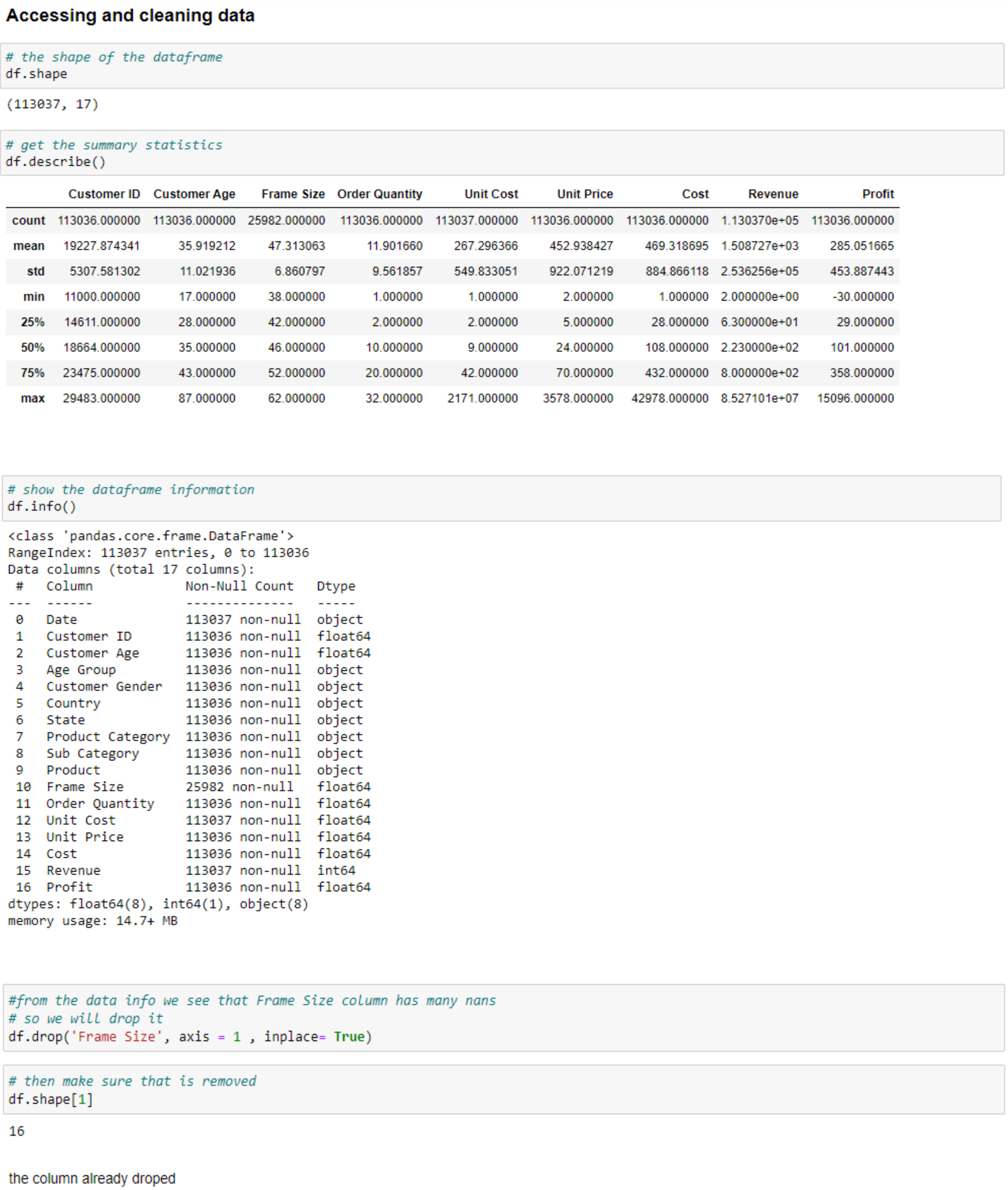
gathering data

```python
# Load your data and print out a few lines. Perform operations to inspect data
# import the libraries that you use
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```python
# load the data
df = pd.read_csv('company.csv')
```

```python
# print out a few lines
df.head()
```

Date	Customer ID	Customer Age	Age Group	Customer Gender	Country	State	Product Category	Sub Category	Product	Frame Size	Order Quantity	Unit Cost	Unit Price	Cost	Revenue	Profit
26/11/2013	11019.0	19.0	Youth (<25)	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack - 4-Bike	NaN	8.0	45.0	120.0	360.0	950	590.0
26/11/2015	11019.0	19.0	Youth (<25)	M	Canada	British Columbia	Accessories	Bike Racks	Hitch Rack - 4-Bike	NaN	8.0	45.0	120.0	360.0	950	590.0
23/03/2014	11039.0	49.0	Adults (35-64)	M	Australia	New South Wales	Accessories	Bike Racks	Hitch Rack - 4-Bike	NaN	23.0	45.0	120.0	1035.0	2401	1366.0
23/03/2016	11039.0	49.0	Adults (35-64)	M	Australia	New South Wales	Accessories	Bike Racks	Hitch Rack - 4-Bike	NaN	20.0	45.0	120.0	900.0	2088	1188.0
15/05/2014	11046.0	47.0	Adults (35-64)	F	Australia	New South Wales	Accessories	Bike Racks	Hitch Rack - 4-Bike	NaN	4.0	45.0	120.0	180.0	418	238.0



Accessing and cleaning data

```
# the shape of the dataframe
df.shape
```

```
(113037, 17)
```

```
# get the summary statistics
df.describe()
```

	Customer ID	Customer Age	Frame Size	Order Quantity	Unit Cost	Unit Price	Cost	Revenue	Profit
count	113036.000000	113036.000000	25982.000000	113036.000000	113037.000000	113036.000000	113036.000000	1.130370e+05	113036.000000
mean	19227.874341	35.919212	47.313063	11.901660	267.296366	452.938427	469.318695	1.508727e+03	285.051665
std	5307.581302	11.021936	6.860797	9.561857	549.833051	922.071219	884.866118	2.536256e+05	453.887443
min	11000.000000	17.000000	38.000000	1.000000	1.000000	2.000000	1.000000	2.000000e+00	-30.000000
25%	14611.000000	28.000000	42.000000	2.000000	2.000000	5.000000	28.000000	6.300000e+01	29.000000
50%	18664.000000	35.000000	46.000000	10.000000	9.000000	24.000000	108.000000	2.230000e+02	101.000000
75%	23475.000000	43.000000	52.000000	20.000000	42.000000	70.000000	432.000000	8.000000e+02	358.000000
max	29483.000000	87.000000	62.000000	32.000000	2171.000000	3578.000000	42978.000000	8.527101e+07	15096.000000

```
# show the dataframe information
df.info()
```

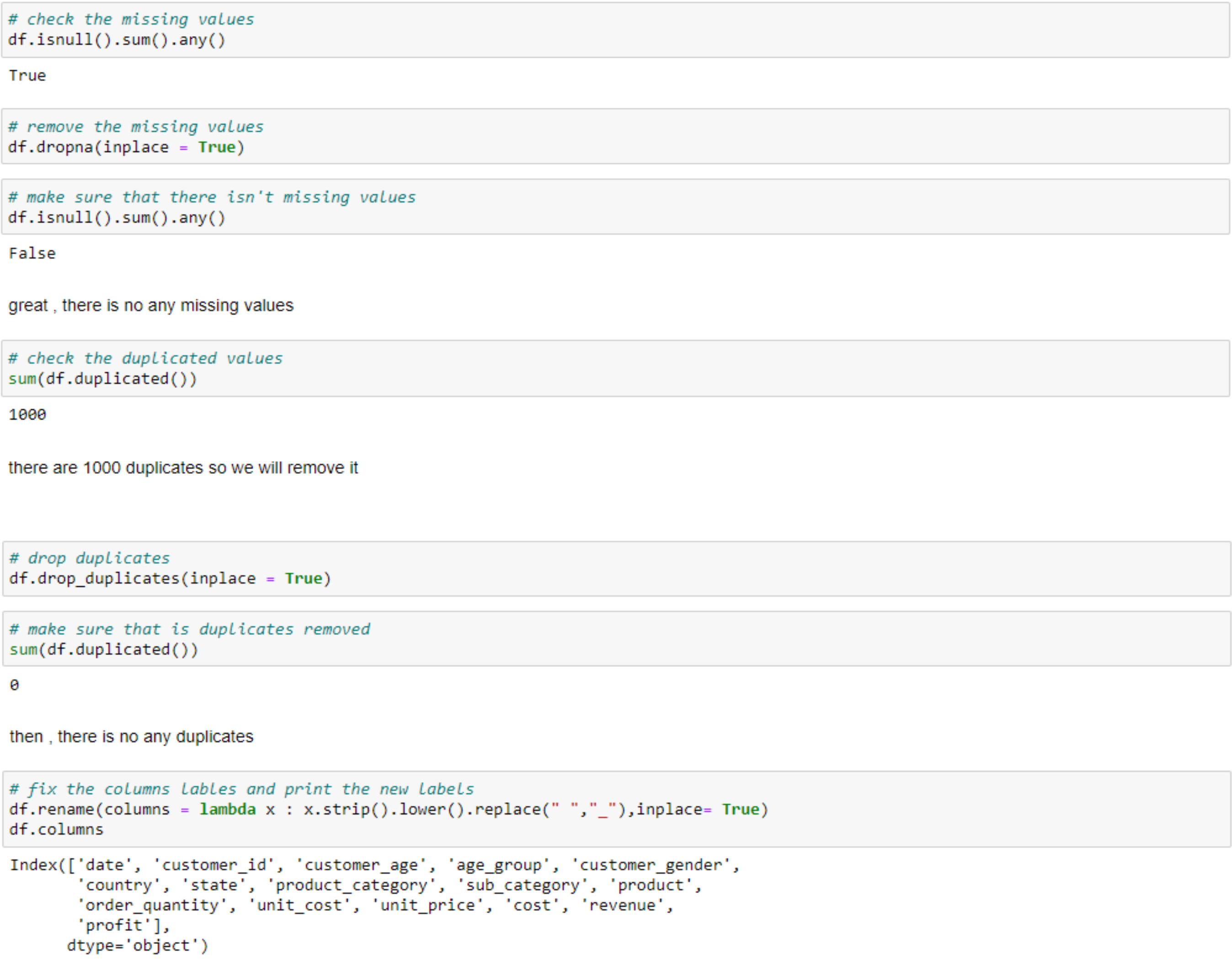
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 113037 entries, 0 to 113036
Data columns (total 17 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   Date              113037 non-null  object 
 1   Customer ID       113036 non-null  float64
 2   Customer Age      113036 non-null  float64
 3   Age Group         113036 non-null  object 
 4   Customer Gender   113036 non-null  object 
 5   Country           113036 non-null  object 
 6   State             113036 non-null  object 
 7   Product Category  113036 non-null  object 
 8   Sub Category      113036 non-null  object 
 9   Product           113036 non-null  object 
 10  Frame Size        25982 non-null   float64
 11  Order Quantity    113036 non-null  float64
 12  Unit Cost         113037 non-null  float64
 13  Unit Price        113036 non-null  float64
 14  Cost              113036 non-null  float64
 15  Revenue           113037 non-null  int64  
 16  Profit            113036 non-null  float64
dtypes: float64(8), int64(1), object(8)
memory usage: 14.7+ MB
```

```
#from the data info we see that Frame Size column has many nans
# so we will drop it
df.drop('Frame Size', axis = 1 , inplace= True)
```

```
# then make sure that is removed
df.shape[1]
```

```
16
```

the column already droped



```
# check the missing values
df.isnull().sum().any()
```

```
True
```

```
# remove the missing values
df.dropna(inplace = True)
```

```
# make sure that there isn't missing values
df.isnull().sum().any()
```

```
False
```

great , there is no any missing values

```
# check the duplicated values
sum(df.duplicated())
```

```
1000
```

there are 1000 duplicates so we will remove it

```
# drop duplicates
df.drop_duplicates(inplace = True)
```

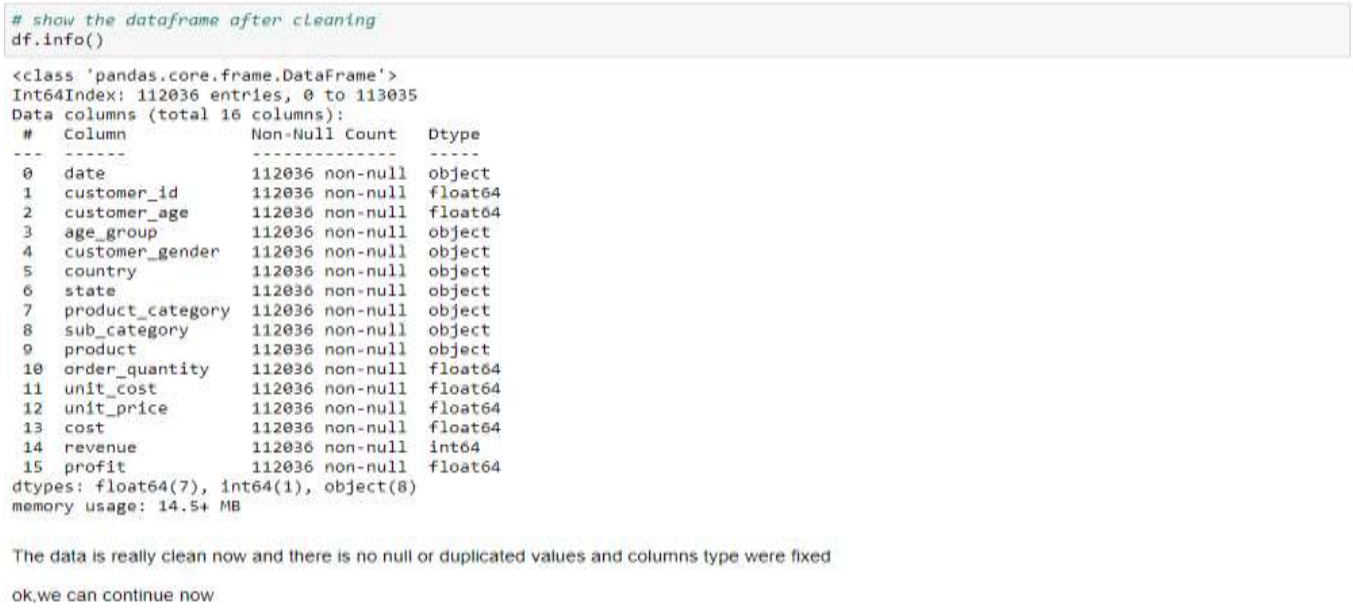
```
# make sure that is duplicates removed
sum(df.duplicated())
```

```
0
```

then , there is no any duplicates

```
# fix the columns lables and print the new labels
df.rename(columns = lambda x : x.strip().lower().replace(" ","_"),inplace= True)
df.columns
```

```
Index(['date', 'customer_id', 'customer_age', 'age_group', 'customer_gender',
       'country', 'state', 'product_category', 'sub_category', 'product',
       'order_quantity', 'unit_cost', 'unit_price', 'cost', 'revenue',
       'profit'],
      dtype='object')
```



```
# show the dataframe after cleaning
df.info()
```

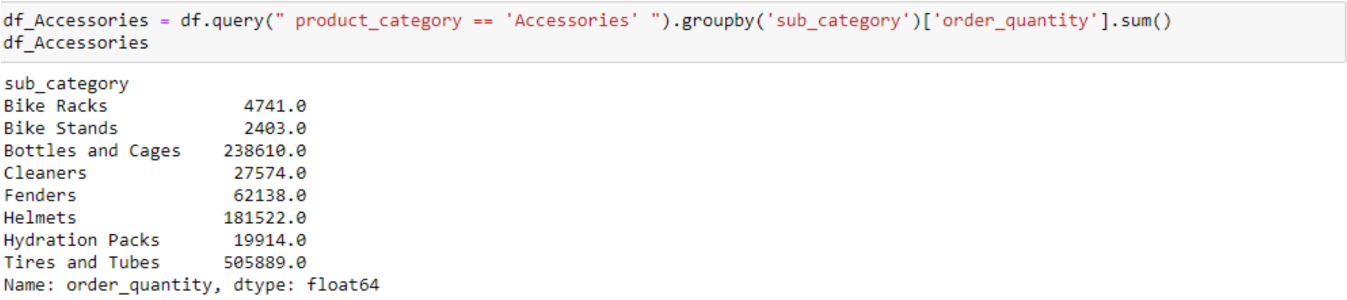
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 112036 entries, 0 to 113035
Data columns (total 16 columns):
 #   Column            Non-Null Count   Dtype  
---  ------            --------------   -----  
 0   date              112036 non-null  object 
 1   customer_id       112036 non-null  float64
 2   customer_age      112036 non-null  float64
 3   age_group         112036 non-null  object 
 4   customer_gender   112036 non-null  object 
 5   country           112036 non-null  object 
 6   state             112036 non-null  object 
 7   product_category  112036 non-null  object 
 8   sub_category      112036 non-null  object 
 9   product           112036 non-null  object 
 10  order_quantity    112036 non-null  float64
 11  unit_cost         112036 non-null  float64
 12  unit_price        112036 non-null  float64
 13  cost              112036 non-null  float64
 14  revenue           112036 non-null  int64  
 15  profit            112036 non-null  float64
dtypes: float64(7), int64(1), object(8)
memory usage: 14.5+ MB
```

The data is really clean now and there is no null or duplicated values and columns type were fixed

ok,we can continue now

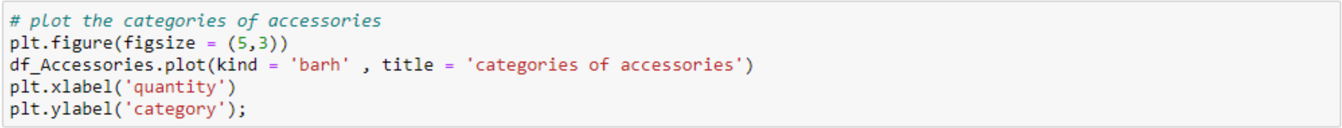
Exploratory Data Analysis

what is the categories of accessories that company products and how many company product for each one

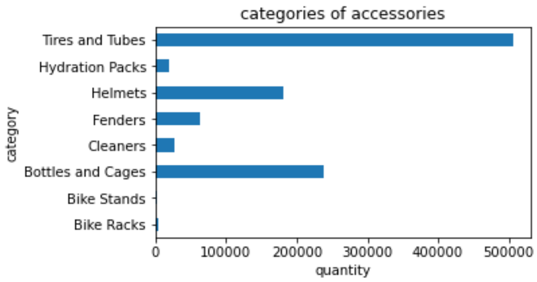


```
df_Accessories = df.query(" product_category == 'Accessories' ").groupby('sub_category')['order_quantity'].sum()
df_Accessories
```

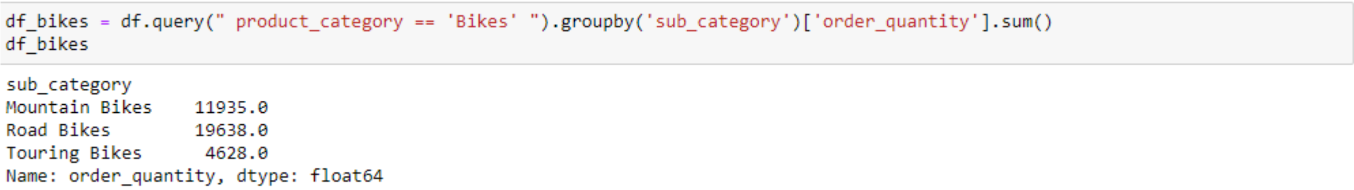
```
sub_category
Bike Racks              4741.0
Bike Stands             2403.0
Bottles and Cages     238610.0
Cleaners               27574.0
Fenders                62138.0
Helmets               181522.0
Hydration Packs        19914.0
Tires and Tubes       505889.0
Name: order_quantity, dtype: float64
```



```
# plot the categories of accessories
plt.figure(figsize = (5,3))
df_Accessories.plot(kind = 'barh' , title = 'categories of accessories')
plt.xlabel('quantity')
plt.ylabel('category');
```

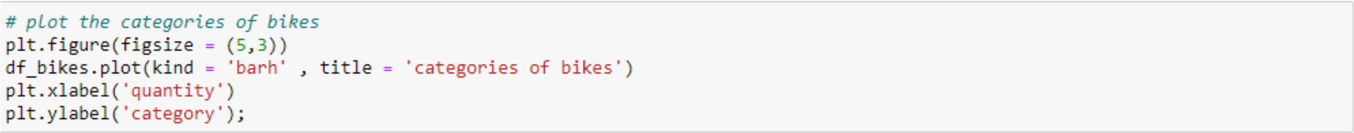


what is the categories of Bikes that company products and how many company product for each one

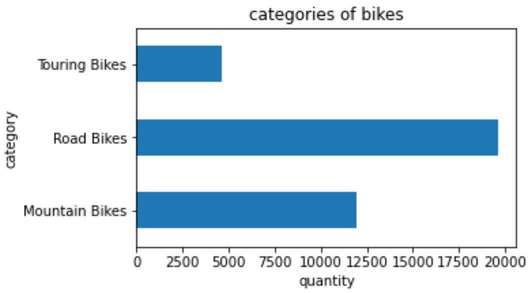


```
df_bikes = df.query(" product_category == 'Bikes' ").groupby('sub_category')['order_quantity'].sum()
df_bikes
```

```
sub_category
Mountain Bikes    11935.0
Road Bikes        19638.0
Touring Bikes      4628.0
Name: order_quantity, dtype: float64
```



```
# plot the categories of bikes
plt.figure(figsize = (5,3))
df_bikes.plot(kind = 'barh' , title = 'categories of bikes')
plt.xlabel('quantity')
plt.ylabel('category');
```

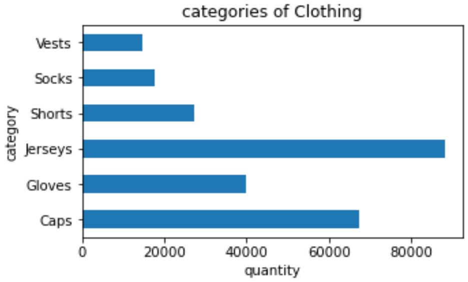


what is the categories of Clothing that company products and how many company product for each one

```
df_Clothing = df.query(" product_category == 'Clothing' ").groupby('sub_category')['order_quantity'].sum()
df_Clothing
```

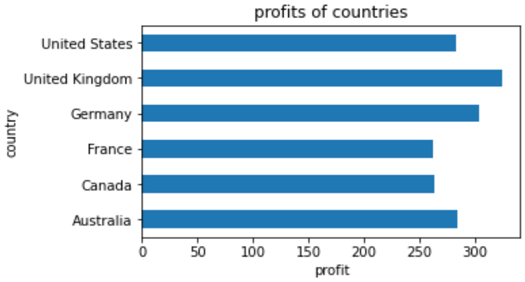
```
sub_category
Caps       67268.0
Gloves     39978.0
Jerseys    88095.0
Shorts     27168.0
Socks      17678.0
Vests      14526.0
Name: order_quantity, dtype: float64
```

```
# plot the categories of clothes
plt.figure(figsize = (5,3))
df_Clothing.plot(kind = 'barh' , title = 'categories of Clothing')
plt.xlabel('quantity')
plt.ylabel('category');
```



what is the best areas which give us a max profit and what is there products

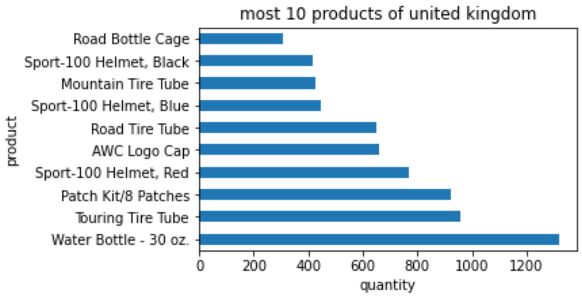
```
areas = df.groupby('country')['profit'].mean()
plt.figure(figsize = (5,3))
areas.plot(kind = 'barh' , title = 'profits of countries')
plt.xlabel('profit')
plt.ylabel('country');
```



then united kingdom has the maximum profit and Australia is the next

so we will get all information about the most 10 products of united kingdom

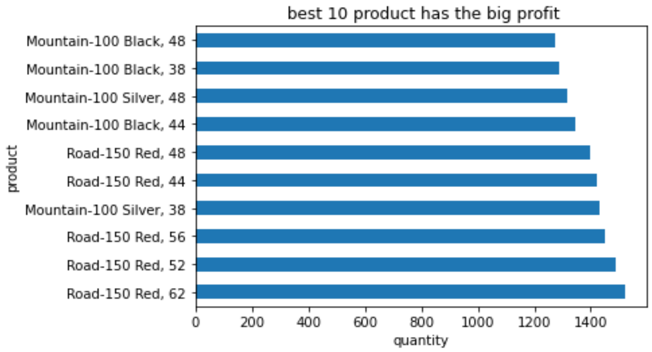
```
UKP= df[df['country']=='United Kingdom']
plt.figure(figsize = (5,3))
UKP['product'].value_counts()[0:10].plot(kind= 'barh', title='most 10 products of united kingdom')
plt.xlabel('quantity')
plt.ylabel('product');
```



what is the best 10 product has the big profit

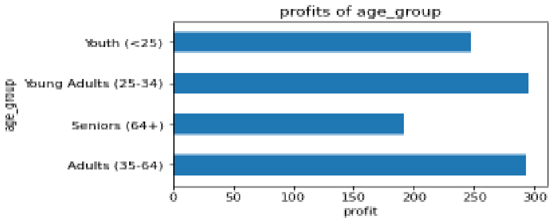
```
products_profits = pd.DataFrame(df.groupby('product')['profit'].mean())
best_10 = products_profits.sort_values(['profit'],ascending=False)
plt.figure(figsize = (5,3))
best_10.head(10).plot(kind = 'barh', title= ' best 10 product has the big profit', legend = False)
plt.xlabel('quantity')
plt.ylabel('product');
```

```
<Figure size 360x216 with 0 Axes>
```



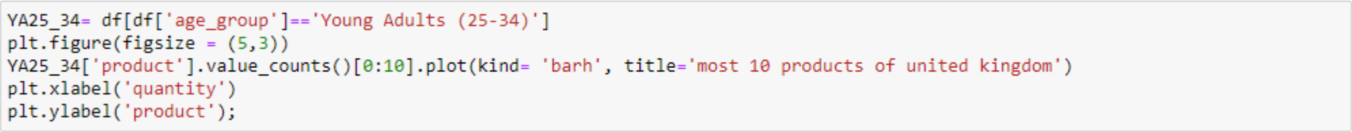
what is the best age_group which give us a max profit and what is there products

```
ages = df.groupby('age_group')['profit'].mean()
plt.figure(figsize = (5,3))
ages.plot(kind = 'barh' , title = 'profits of age_group')
plt.xlabel('profit')
plt.ylabel('age_group');
```

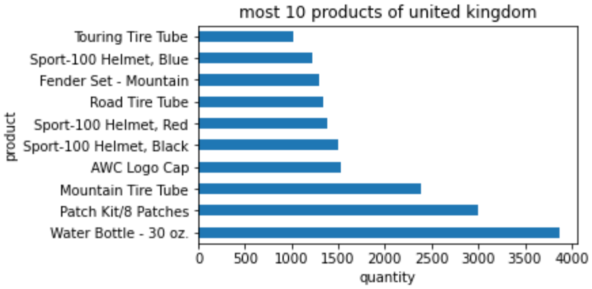


then the most age group has the biggest profit is young adults(25_34)

so we will get all information about the most 10 products of young adults(25_34)



```
YA25_34= df[df['age_group']=='Young Adults (25-34)']
plt.figure(figsize = (5,3))
YA25_34['product'].value_counts()[0:10].plot(kind= 'barh', title='most 10 products of united kingdom')
plt.xlabel('quantity')
plt.ylabel('product');
```



Conclusions

Ok, here we finished our analysis process

let's look over the project and summrize the steps :

step_1 : data wrangling

we gathered the data then moved to access and clean it and we did that :

1- we cleaned the data from the null values

2- we drpoed the duplicated values

3- we removed the outlires and fixed the columns dtype

4- we selected the columns that we need to analysis process and dropped the others

step_2 : EDA

first we had broken the data and get all details about the products , then we started to ask our questions to get ansewrs:

1- the best areas which give us a max profit and we get it the united kingdom and we got all information about the most 10 products of it

2- the best 10 product has the big profit and the best was Road-150 Red ,62

3- the best age_group which give us a max profit and we get it the young adults(25-34) and we got all information about the most 10 products of it .

Chapter 3

Clustering in machine learning

3.1 What is machine learning?

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, uncovering key insights within data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics.

3.2 How machine learning works?

The learning system of a machine learning algorithm has a three main parts.

1. A Decision Process: In general, machine learning algorithms are used to make a prediction or classification. Based on some input data, which can be labelled or unlabeled, your algorithm will produce an estimate about a pattern in the data.
2. An Error Function: An error function serves to evaluate the prediction of the model. If there are known examples, an error function can make a comparison to assess the accuracy of the model
3. An Model Optimization Process: If the model can fit better to the data points in the training set, then weights are adjusted to reduce the discrepancy between the known example and the model estimate. The algorithm will repeat this evaluate and optimize process, updating weights autonomously until a threshold of accuracy has been met.

3.3 Machine learning methods:

Machine learning classifiers fall into two primary categories.

- Supervised Machine learning
- unsupervised Machine learning

3.3.1Supervised Machine learning:

In supervised machine learning, you are interested in predicting a label for your data. Commonly, you might want to predict fraud, customers that will buy a product, or home values in an area.

Supervised learning helps organizations solve for a variety of real-world problems at scale, such as customers that will buy a product, or home values in an area, Some methods used in supervised learning include neural networks, naïve bayes, linear regression, logistic regression, random forest, support vector machine (SVM), and more.

3.3.2 Unsupervised Machine learning:

In unsupervised machine learning, you are interested in clustering data together that isn't already labeled, such as customer segmentation and topical analysis.

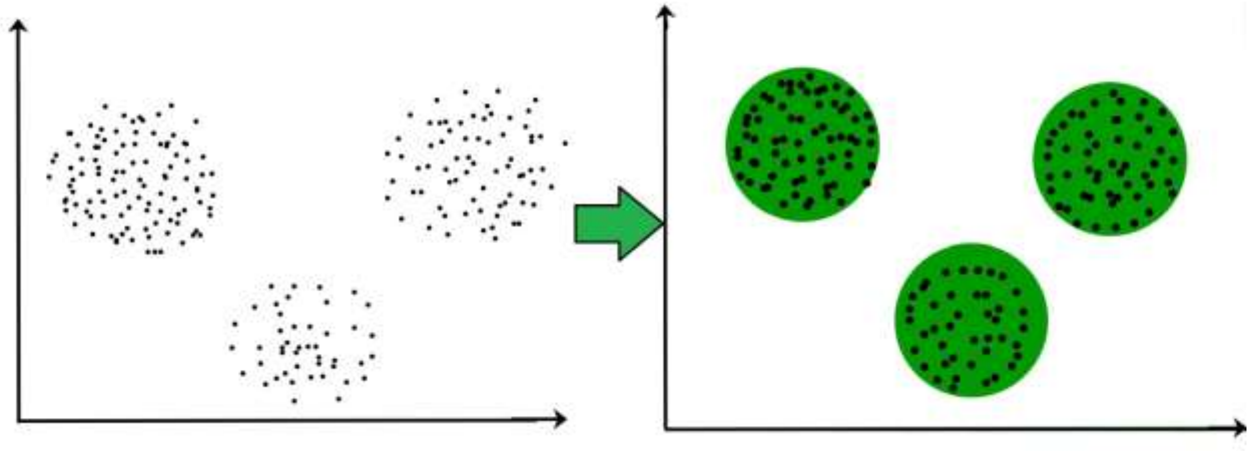
Unsupervised machine learning uses machine learning algorithms to analyze and cluster unlabeled datasets. These algorithms discover hidden patterns or data groupings without the need for human intervention. Its ability to discover similarities and differences in information make it the ideal solution for exploratory data analysis.

Other algorithms used in unsupervised learning include neural networks, k-means clustering, probabilistic clustering methods, and more.

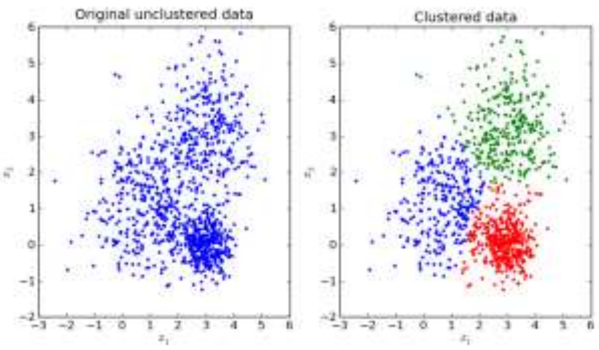
3.4 Clustering:

Clustering is basically a type of unsupervised machine learning and it is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

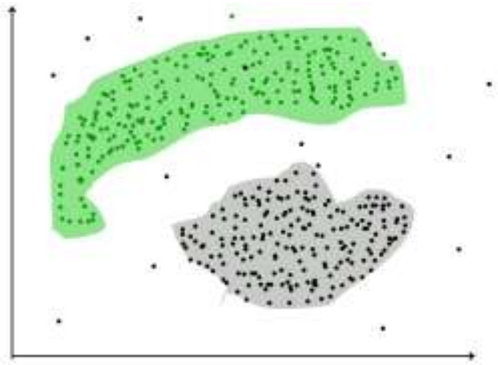
For ex– The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



Where the data in the left side is the original unclustered data (raw data) and the data in the right side is the clustered data like this:



Note that it is not necessary for clusters to be spherical. Such as:

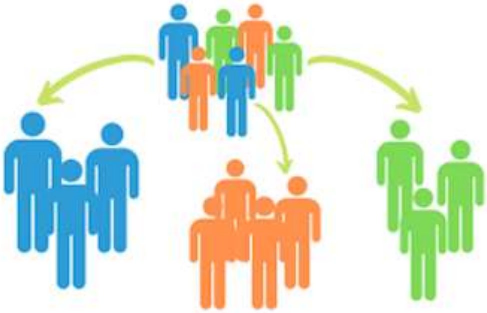


3.5 Why clustering is important?

Clustering is very much important as it determines the intrinsic grouping among the unlabeled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need.

To declare this point we will apply in clients (customers) segmentation where if I in a company And I have 200000 client, then I can`t treat with all clients as the same because I can`t know who is a good customer and who is bad, who will deserve the importance and who is not

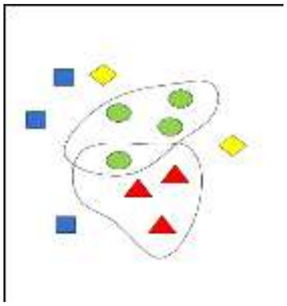
Here if I use the clustering then I can split the 200000 client to several groups such as A, B, and C where the group A represents the clients who interest in high quality and group B represents the clients who interest in price more than quality and group C represents the clients who interest in time of service



Then after using the clustering I have became able to focus in the needs of each group by their interest where I will provide a high quality products to group A and will provide a cheap products to group B and I will work in short service time for them

3.6 Overlapping clustering:

There is a type of clustering has a overlapping form, and this type is executed when the clusters of different groups share some features together as in the next image



In this image we see that there is a shared member between the green group and the red group, in this case the clustering is overlapping.

3.7 Applications of Clustering in different fields:

- Marketing: It can be used to characterize & discover customer segments for marketing purposes.
- Biology: It can be used for classification among different species of plants and animals.

- Libraries: It is used in clustering different books on the basis of topics and information.
- Insurance: It is used to acknowledge the customers, their policies and identifying the frauds.
- City Planning: It is used to make groups of houses and to study their values based on their geographical locations and other factors present.
- Earthquake studies: By learning the earthquake-affected areas we can determine the dangerous zones.

3.8 Clustering algorithms :

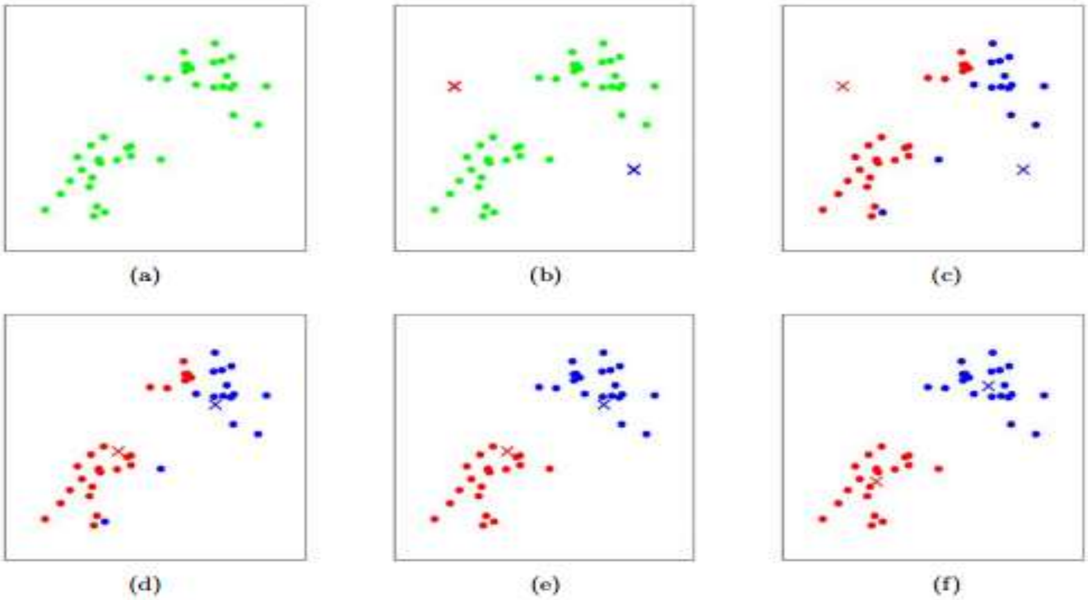
- K-means algorithm
- Hierarchical clustering algorithm
- DBSCAN algorithm

3.8.1 K-means algorithm:

K-means algorithm is the simplest unsupervised learning algorithm that solves clustering problem.

How it works?

We will look to the next image and explain each step:



1- First we select the number of clusters equal (K)

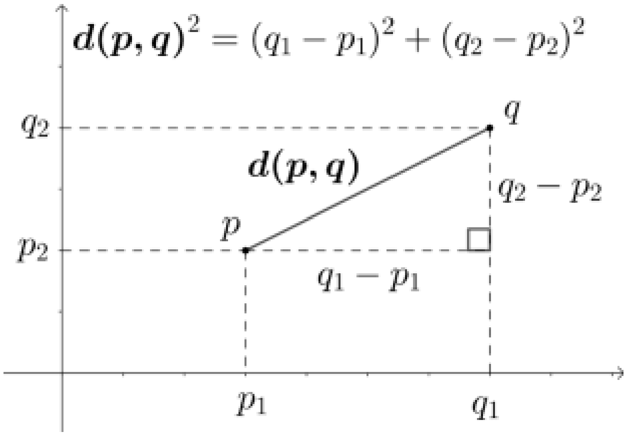
2- In the step (b) the algorithm will select a number (k) of random points in the each cluster and this points called (cluster centroid)

Note that number of clusters (K) = number of cluster centroid (k)

Where two clusters = two center points

3- In the step (c) the algorithm will compute the distance between each point and the two cluster centroids, then every point will belong to the nearest centroid by the Euclidian distance where:

The Euclidian distance is a method in mathematics used to compute the distance between two points by the equation:

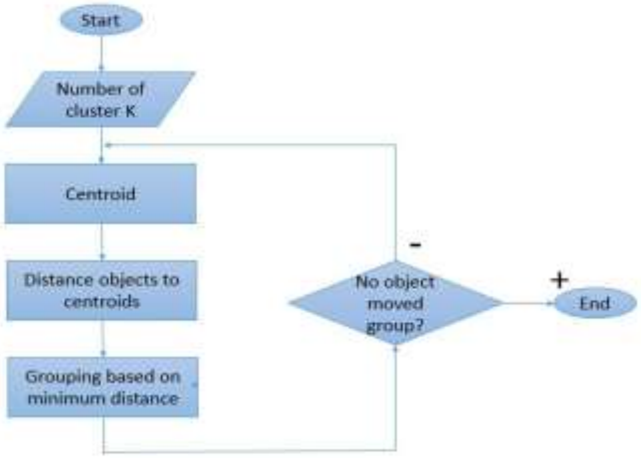


4- In the step (d) the algorithm will do a displacement to the centroids to move every centroid to the center of his cluster which be nearest to the mean point and this way called (arithmetic average or mean)

5- In the step (f) I will repeat the step (c) and step (d) another once and we will see that some points had turned from red to blue and from blue to red and this because the centroids had changed

6- In step (e) I will repeat the step (d) to do a displacement to the new centroids to a nearest point to means of each cluster

Then to summaries the steps in good shape let us show what we actually did:


Start
Number of cluster K
Centroid
Distance objects to centroids
Grouping based on minimum distance
No object moved group?
-
+
End

Then the idea here is that I still repeat selection (step (b)) and displacement (step(c)) until I reach to the perfect shape of clustered data.

Note that:

- We can apply this process over any number (n) of clusters not over two only
- I can make the algorithm select the number of clusters by himself instead of I do this

How the algorithm works**?**

To the algorithm reach to the perfect shape of clustered data it depends on three factors:

1- Optimization objective function
2- The choice of the centroid
3- Local minimum and global minimum values

Then how it works:

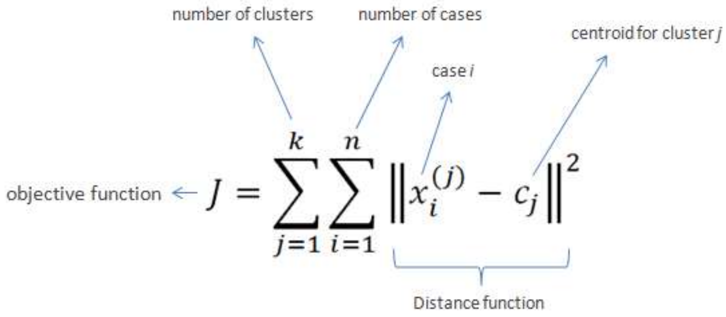
Optimization objective function:

- We refer it by (J)
- The optimization objective function is the function that we should minimize it to reach to perfect shape

- Then smaller J the more each point closer to it`s centroid
- Very important to me to keep my results in the high quality

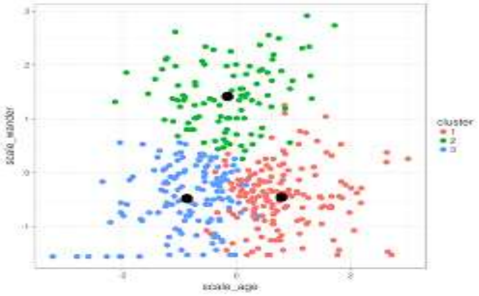
How we measure J mathematically?

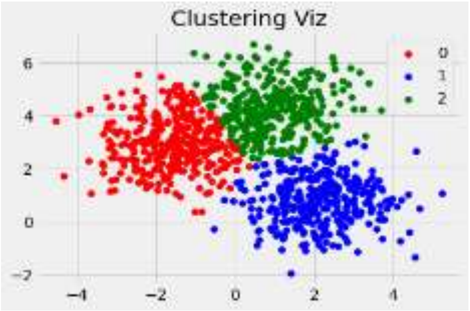
We measure J value by this equation:



$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Then whenever the points are farther of the centroid this get smaller J and whenever the points are closer to the centroid this get greater J like this image:



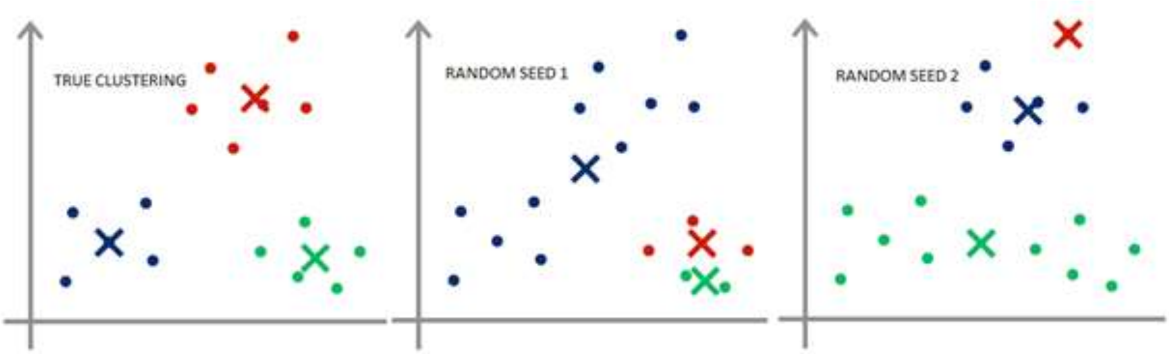


Great J value small J value

The choice of the centroid:

The idea here is the choice is closer to the cluster to help the algorithm to do the perfect clustering, and that because whenever the choice is closer to the cluster then the value of J will be smaller and that help us to reach to the perfect shape.

Look to the next image:



In the first graph the choice of the centroid is fitting

In the second and third graph the choice of the centroid is not fitting

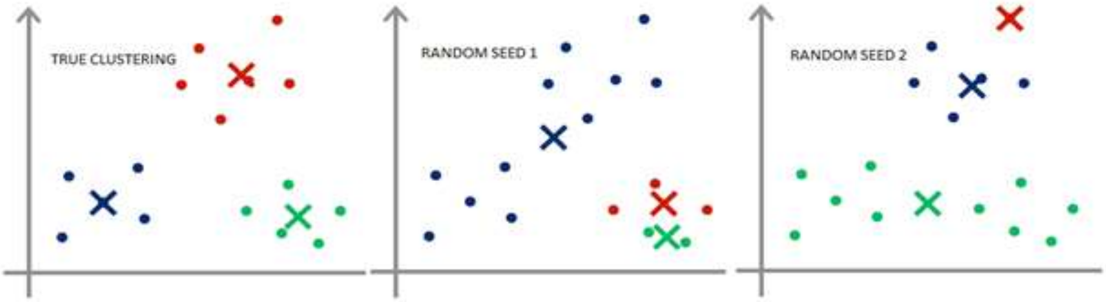
Local minimum and global minimum values:

The third factor is Local minimum and global minimum values, so what is the meaning of Local minimum and global minimum values:

- Local minimum is a small value to J but it is not perfect
- Global minimum is the smallest value of J

Then when the choice of the centroid is perfect the value of J is a global minimum and when the choice of the centroid is not perfect the value of J is a local minimum

Where :



In the first graph the choice is fitting so the value of J is the global minimum

In the second and third graph the choice is not fitting so the value of J is a local minimum

So we get that :

- If the number of clusters is small then we should do the iteration process (selecting and displacement) several times until we reach to the global value of J
- I can`t depend on my eye because sometimes there are a complication of the data

How we can select the number of clusters**?**

In this point we have to ways :

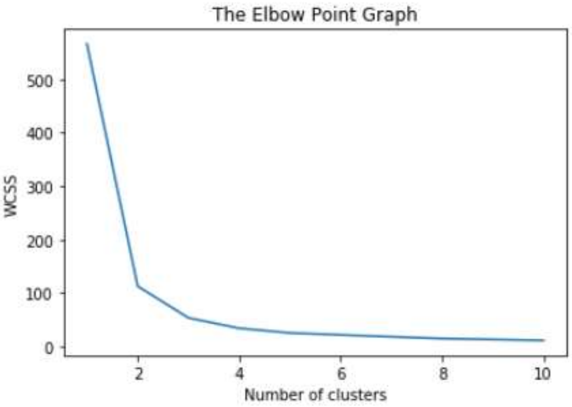
1- By sight
2- By Elbow method

First way : by sight

- Here I depend on my eye in selecting the number of clusters depending on the blocs of the data together
- This way is fitting to small numbers of data
- Sometimes it can`t give me a high quality
- Provide my time and cost

Second way : Elbow method

Here I will do a graph which K number of clusters in the X axis and the value of J on the Y axis like the images:



In the image when I tell the algorithm to do one cluster in X axis it give me a high value of J in Y axis > 500 , and when I tell the algorithm to do two clusters in X axis it give me a lower value of J in Y near to 100 , so I will repeat the process until I get a fit number of clusters and fit value of J to avoid the misleading of quality.

Note that I can`t depend on J value only, but there are a life factors I should take it and to clear it we will take an example:

If I want to divide the clients of the company to number of clusters and the elbow method tell me to make them 6 clusters, but I have only 4 customer service employers , then here I should chose 4 clusters however J value may be higher than 6 clusters

Then now I know how the algorithm works and how I select the number of clusters

We will move to the second algorithm in clustering hierarchical clustering

3.8.2 Hierarchical clustering algorithm:

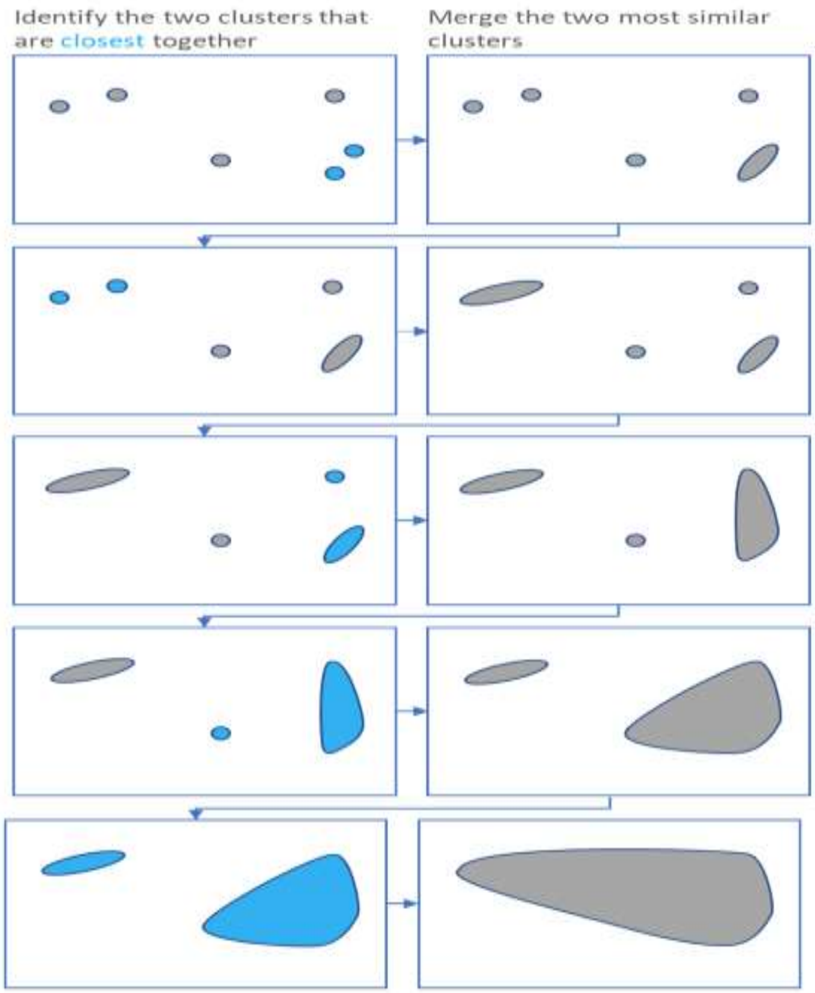
- Hierarchical clustering, also known as hierarchical cluster analysis, is an algorithm that groups similar objects into groups called clusters.
- The endpoint is a set of clusters, where each cluster is distinct from each other cluster, and the objects within each cluster are broadly similar to each other.

How it works**?**

Hierarchical clustering starts by treating each observation as a separate cluster. Then, it repeatedly executes the following two steps:

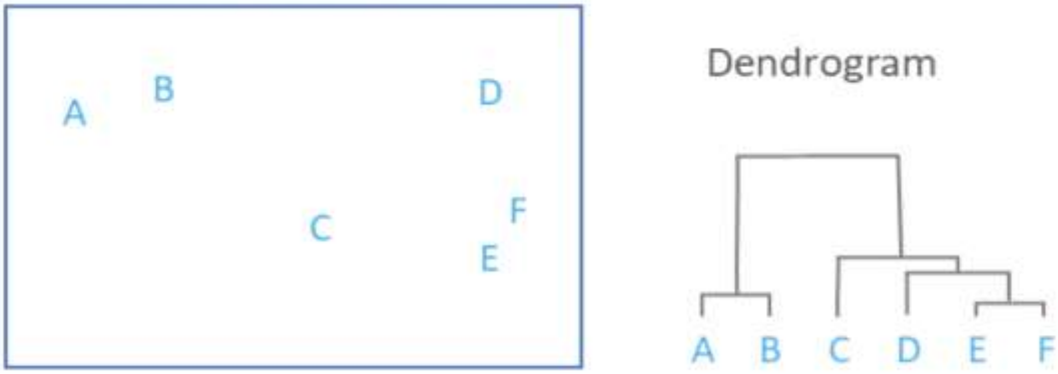
1- Identify the two clusters that are closest together by Euclidian distance

2- Merge the two most similar clusters. This iterative process continues until all the clusters are merged together. This is illustrated in the diagrams below.


Identify the two clusters that are closest together
Merge the two most similar clusters

The output **:**

The main output of Hierarchical Clustering is a Dendrogram, which shows the hierarchical relationship between the clusters:



A dendrogram is a diagram that shows the hierarchical relationship between objects. It is most commonly created as an output from hierarchical clustering The main use of a dendrogram is to work out the best way to allocate objects to clusters.

In the dendrogram above, the height of the dendrogram indicates the order in which the clusters were joined where the heights reflect the distance between the clusters, In the example above, we can see that E and F are most similar, as the height of the link that joins them together is the smallest. The next two most similar objects are A and B. where if the distance between A and B equal 2 then the height between A and B in dendrogram equal 2.

So , we can summarize what happened in the example above:

1- the algorithm computed the distance between each two points by Euclidian method
2- it found that points E and F are most similar so it merged them together in a one cluster (C1)
3- it repeated the step 1 again and found that A and B are most similar so it merged them together in a one cluster (C2)
4- it repeated the step 1 again and found that cluster (C1)and D are most similar so it merged them together in a one cluster (C3)
5- it repeated the step 1 again and found that cluster (C3)and C are most similar so it merged them together in a one cluster (C4)
6- it repeated the step 1 again and found that cluster (C3)and C are most similar so it merged them together in a one cluster (C4)
7- it repeated the step 1 again and found that cluster (C4)and cluster (C2) are most similar so it merged them together in a one cluster (C5)
8- then the last output is the cluster (C5)

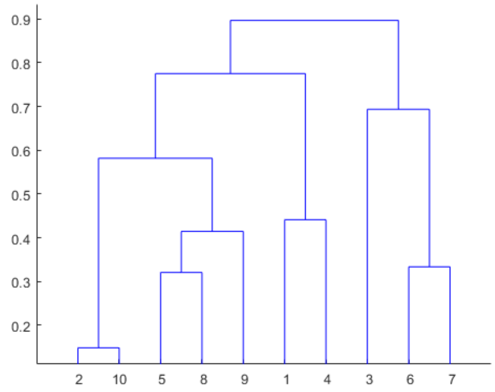
so that reach by us to an important point that **how I select the number of clusters from this data ?**

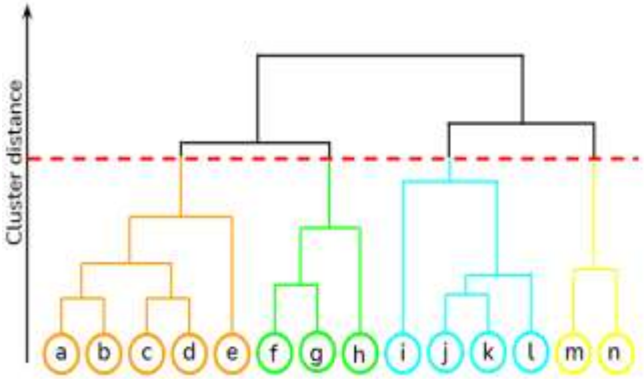
in step 8 we get the cluster (C5) which merges between cluster (C2) combining (A , B) and cluster (C4) combining (C , D , E ,F) so the data will grouped in two clusters.

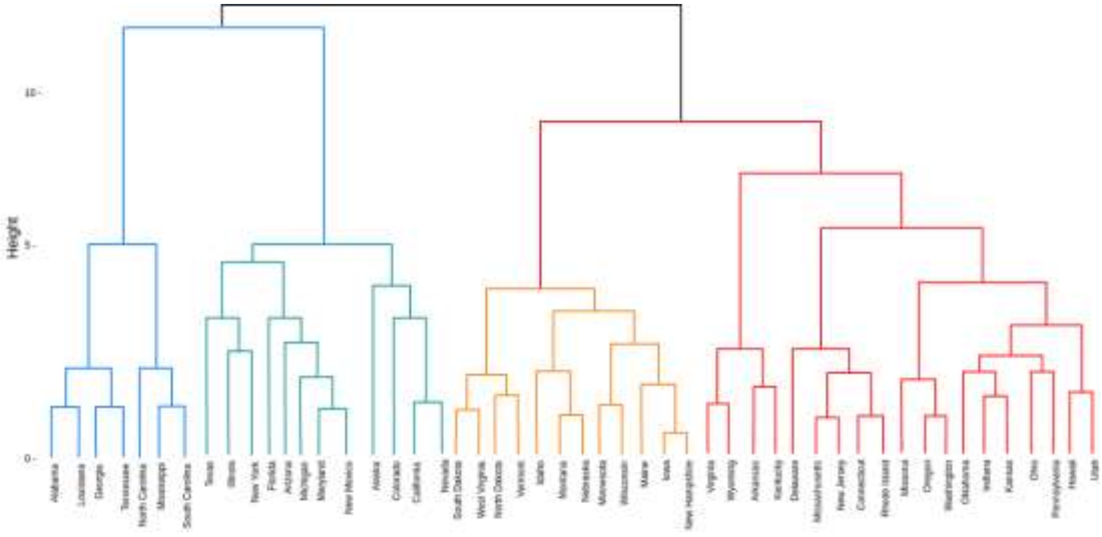
Note that I can consider that c is a one cluster so I will get three clusters which be cluster (C2) combining (A , B) and cluster(C3) combining (D ,E ,F) and point C .

The idea here that I can selected any number of clusters by my sight as we saw in the example above.

Examples of dendrogram:





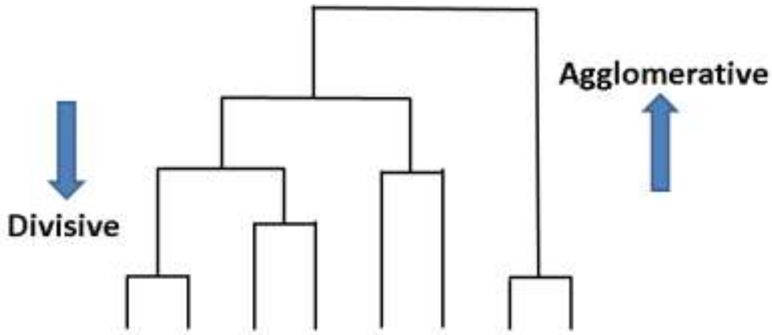


Agglomerative versus divisive algorithms**:**

There are two types of hierarchical clustering:

- Agglomerative hierarchical clustering
- divisive hierarchical clustering

to explain this point see this image:



- Hierarchical clustering typically works by sequentially merging similar clusters, as shown above. This is known as agglomerative hierarchical clustering.
- In theory, it can also be done by initially grouping all the observations into one cluster, and then successively splitting these clusters. This is known as divisive hierarchical clustering.
- Divisive clustering is rarely done in practice.

Here we finished the second type of clustering hierarchical clustering

We will move to the third algorithm in clustering DBSCAN clustering

3.8.3 Density-based spatial clustering of applications with noise (DBSCAN) :

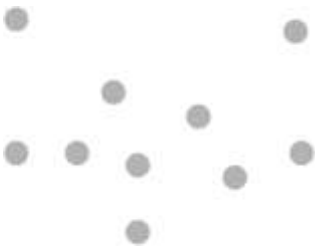
- Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density.
- It groups 'densely grouped' data points into a single cluster. It can identify clusters in large spatial datasets by looking at the local density of the data points.
- The most exciting feature of DBSCAN clustering is that it is robust to outliers**.**
- It also does not require the number of clusters to be told beforehand, unlike K-Means, where we have to specify the number of centroids.

How it works **?**

DBSCAN requires only two parameters epsilon and minPoints.

- Epsilon (eps) is the radius of the circle to be created around each data point to check the density
- minPoints (MinPts) is the minimum number of data points required inside that circle for that data point to be classified as a Core point.

Let's understand it with the help of an example.

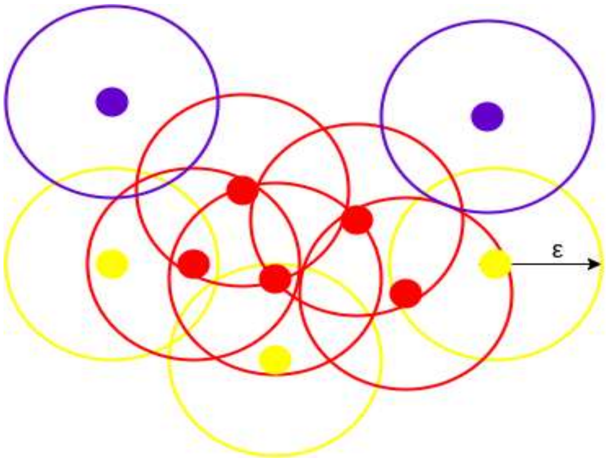


Here, we have some data points represented by grey color. Let's see how DBSCAN clusters these data points.

DBSCAN creates a circle of epsilon radius around every data point and classifies them into **Core** point, **Border** point, and **Noise**.

- A data point is a **Core** point if the circle around it contains at least 'minPoints' number of points.
- If the number of points is less than minPoints, then it is classified as **Border** Point,
- And if there are no other data points around any data point within epsilon radius, then it treated as **Noise**.

See the image:



The above figure shows us a cluster created by DBCAN with **[minPoints = 3].** Here, we draw a circle of equal radius *epsilon* around every data point. These two parameters help in creating spatial clusters.

All the data points with at least 3 points in the circle including itself are considered as **Core** points represented by **red color**

. All the data points with less than 3 but greater than 1 point in the circle including itself are considered as **Border** points. They are represented by **yellow color.**

Finally, data points with no point other than itself present inside the circle are considered as **Noise** represented by the **purple color.**

Then how the algorithm make the clusters**?**

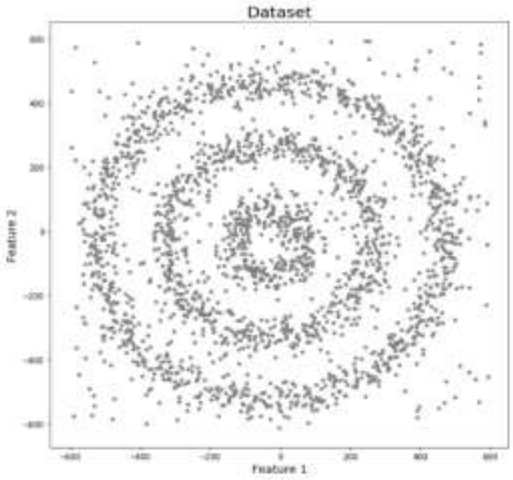
1. Find all the neighbor points within eps and identify the core points or visited with more than MinPts neighbors.
2. For each core point if it is not already assigned to a cluster, create a new cluster.
3. Find recursively all its density connected points and assign them to the same cluster as the core point.

   A point *a* and *b* are said to be density connected if there exist a point *c* which has a sufficient number of points in its neighbors and both the points *a* and *b* are within the *eps distance*. This is a chaining process. So, if *b* is neighbor of *c*, *c* is neighbor of *d*, *d* is neighbor of *e*, which in turn is neighbor of *a* implies that *b* is neighbor of *a*.
4. Iterate through the remaining unvisited points in the dataset. Those points that do not belong to any cluster are noise.

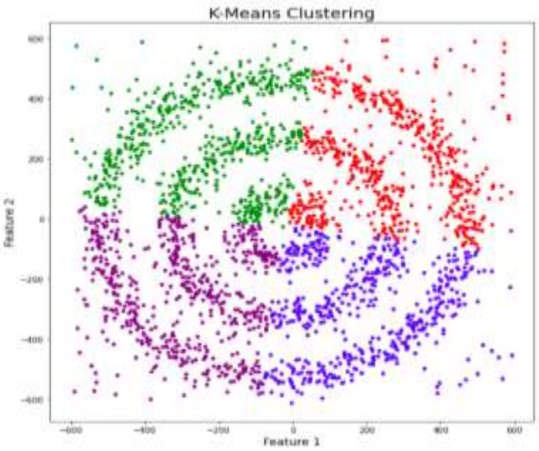
Why we need to use DBSCAN **?**

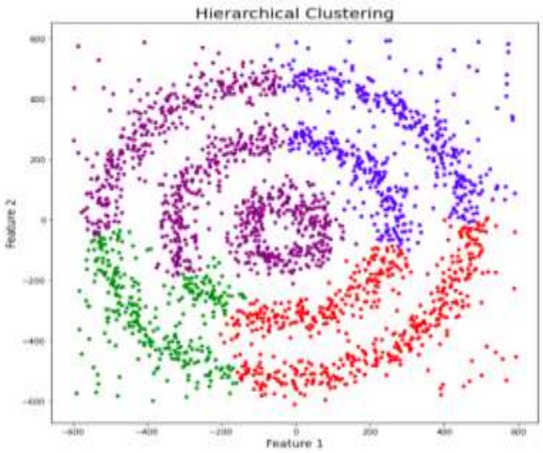
K-Means and Hierarchical Clustering both fail in creating clusters of arbitrary shapes. They are not able to form clusters based on varying densities. That’s why we need DBSCAN clustering.

Let’s try to understand it with an example. Here we have data points densely present in the form of concentric circles:

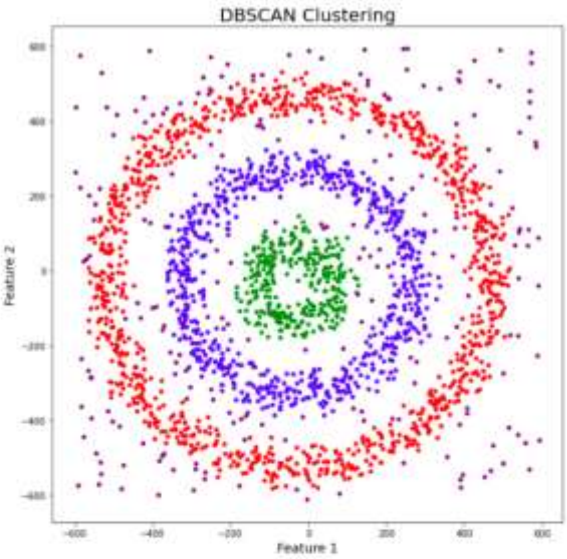


We can see three different dense clusters in the form of concentric circles with some noise here. Now, let’s run K-Means and Hierarchical clustering algorithms and see how they cluster these data points.





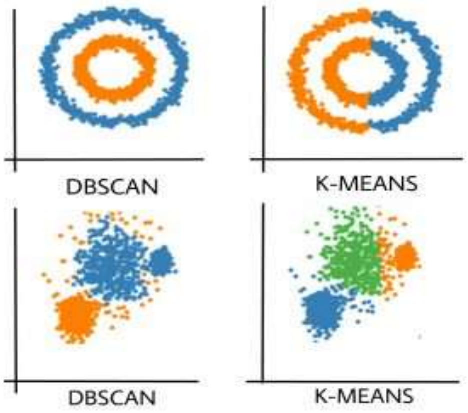
You might be wondering why there are four colors in the graph? That because this data contains noise too, therefore, I have taken noise as a different cluster which is represented by the purple color. Sadly, both of them failed to cluster the data points. Also, they were not able to properly detect the noise present in the dataset. Now, let’s take a look at the results from DBSCAN clustering.



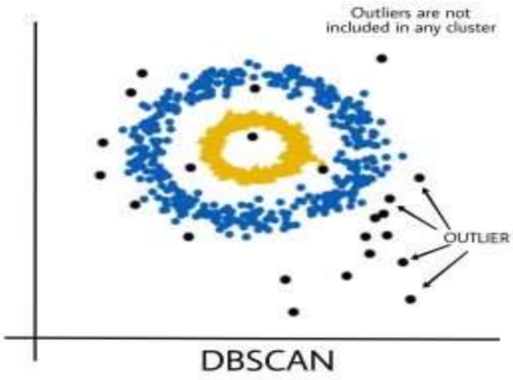
Awesome! DBSCAN is not just able to cluster the data points correctly, but it also perfectly detects noise in the dataset.

3.8.4 DBSCAN vs K-means:

- DBSCAN produces more reasonable results than *k*-means across a variety of different distributions.
- K-Means forms spherical clusters only. This algorithm fails when data is not spherical but DBSCAN is not so see the image:



- K-Means algorithm is sensitive towards outlier. Outliers can skew the clusters in K-Means in very large extent but DBSCAN treat with outliers very well so see the image:



- K-Means algorithm requires one to specify the number of clusters a priory etc but DBSCAN algorithm identifies the dense region by grouping together data points that are closed to each other based on distance measurement.

Until here we explained the three clustering algorithms , now we will explain an application project about it.

This application will be about the customer segmentation , so let`s get started.

3.9 Customer segmentation problem in machine learning:

What we will study?

- Customer Segmentation & Its Types
- Effect Of Customer Segmentation In Marketing Domain
- Keypoints To Remember For Customer Segmentation In Marketing Domain
- Steps to Perform Customer segmentation with Machine Learning Algorithms.
- Conclusion

3.9.1 Customer Segmentation & Its Types:

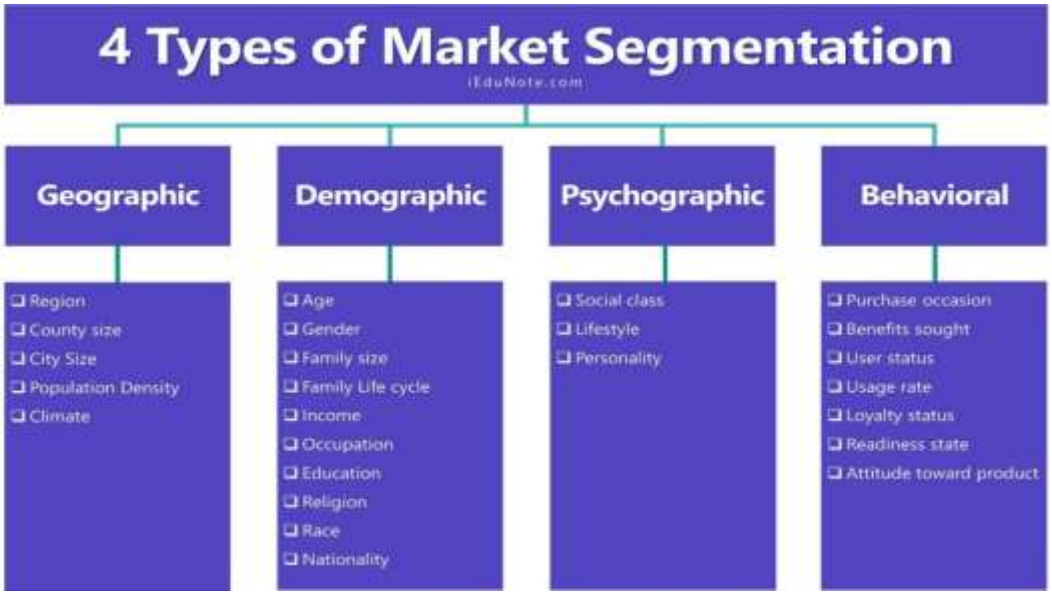
Customer segmentation is the method of distributing a customer base into collections of people based on mutual characteristics so organizations can market to group efficiently and competently individually.

The purpose of segmenting customers is to determine how to correlate to customers in multiple segments to maximize customer benefits. Perfectly done customer segmentation empowers marketers to interact with every customer in the best efficient approach.

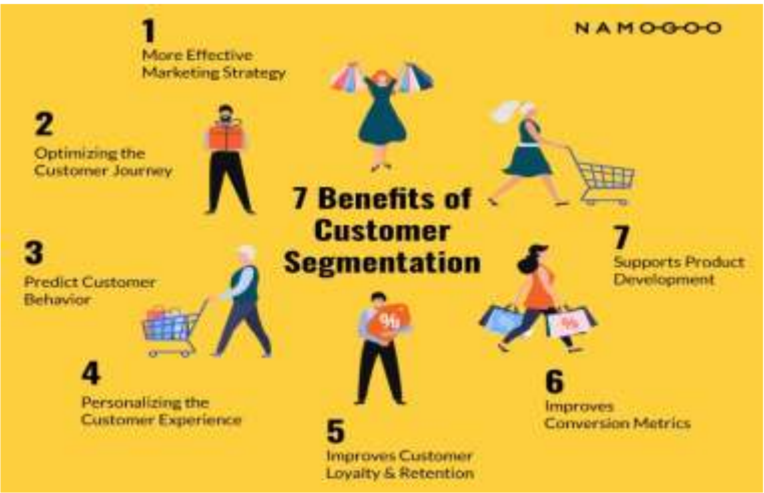
3.9.2 **Types of Customer segmentation :**

- Demographic Segmentation
- Geographic Segmentation
- Psychographic Segmentation
- Behavioral Segmentation

See the image to explain more about the types :



Effect Of Customer Segmentation In Marketing Domain:


NAMOGOO
7 Benefits of Customer Segmentation
1 More Effective Marketing Strategy
2 Optimizing the Customer Journey
3 Predict Customer Behavior
4 Personalizing the Customer Experience
5 Improves Customer Loyalty & Retention
6 Improves Conversion Metrics
7 Supports Product Development

By segmenting users, **marketers** can obtain the most maximum of their operations budgets by targeting the appropriate audiences. You can converse straight to customers who are most assuring to transform without spending money on impressions or users who aren't inclined to purchase the following product.

And you can **decorate marketing messages** & make them *appealing* to sustain prospects down the duct more productively. That work can associate with both intelligence and merchandise development.

Definitely, Segmentation promotes a corporation in the following ways:

- Design and deliver targeted marketing advice that will resonate with particular customer associations but not with others (who will accept notifications according to their requirements and importance, preferably).
- Decide the most reliable communication course for the segment, anything from email, social media posts, radio advertising, or a different procedure, depending on the feature.
- Distinguish methods to promote products or new merchandise or assistance opportunities.
- Build more trustworthy consumer relations to enhance customer assistance.
- Analysis of pricing selections to concentrate on the most influential customers.

3.9.3 Keypoints To Remember For Customer Segmentation In Marketing Domain:

Most companies, when they commence with customer segmentation, they lack a clear vision and a goal. You can try the subsequent measures to obtain segments in customer support on a universal level.

1. Examine the present customers: Knowing the geographic distribution, shopper preferences/beliefs, analyzing website search page analytics, etc.
2. Acquire knowledge of every consumer: Plotting an interactive chart for each customer to an assortment of decisions to explain and foretell their response like the commodities, assistance, and content they would be engaged in.
3. Explain segment possibilities: Once the sections have been established, they should implement a fit business understanding of every segment and its difficulties and possibilities. You can map the entire company's marketing policy to provide to diverse niches of consumers
4. Research the segment: After comparing the description and marketing significance of distinct customer segments, a corporation must realize how to transform its products or

assistance to more valid aid. For illustration, it may determine implementing higher cuts to some buyers than others to develop its existing consumer base.

3.9.4 Steps to Perform Customer segmentation with Machine Learning Algorithms:

1. Design a proper business care before you start
2. Collect & prepare the data
3. performing segmentation using K-Means clustering
4. Tuning the optimal hyperparameters for the model
5. Visualization of the results

Machine learning, a class of artificial intelligence, can investigate data sets of similar customers and interpret the most beneficial and most inadequate performing customer segments.

The subsequent actions are one of many strategies to tackle customer segmentation over machine learning. You can utilize your favorite tools, partners, and skills to handle these methods conveniently.

3.9.4.1 Step 1 :Design a proper business care before you start

In the case research, we need to visualize consumer habits and styles from different perspectives. You don't need to go into this method recklessly. Otherwise, the result will be dirty and disordered.

Alternatively, you require a good business case to start with. The prospect of applying machine learning and artificial intelligence can be thought of with:

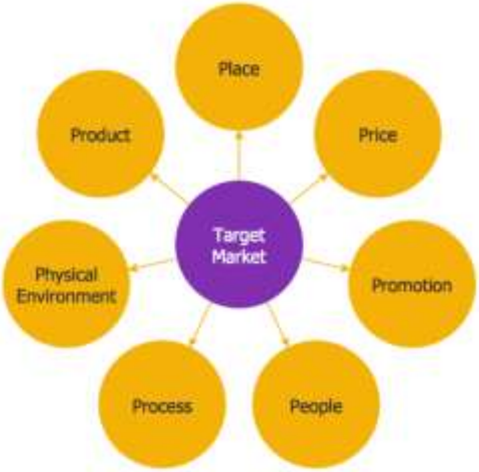
- "Can the consumer support be organized into groups to generate customized connections within them?"
- "Is determining the most vital customer gatherings within the entire pool of consumers worthy? "

To fully appreciate customers' spending and regulation, you can practice with the latter points in mind:

- Amount of commodities ordered
- ordinary return rate
- cumulative spending

Once you've prepared the business case, proceed to the next step.

3.9.4.2 Step 2 : Collect & prepare the data



The next step is to assemble the data to discover more different patterns and biases inside the datasets.

You will also necessitate setting complex characteristics depending on the most relevant metrics for your organization. It may involve:

- Medium lifetime value
- Consumer purchase cost
- Consumer pleasure
- Maintenance rate
- Net earnings

You will need to scale, preprocess and fill the missing values using the open-source tools available in python, such as pandas, NumPy, etc. This step needs to be fixed because they add to the visualization step later.

The more extra customer data you have, the more precise decision you will perform in customer segmentation with machine learning.

That leads us to the next step.

3.9.4.3 Step 3 : performing segmentation using K-Means clustering

K-means clustering is a famous method of unsupervised machine learning. This method obtains all of the diverse “clusters” and clubs them collectively while maintaining them as tiny as attainable.

Algorithms works in this manner:

- First, we randomly initialize the value of k as the number of clusters or n- centroids.
- Next, we allot each data points to the nearest centroid forming separate groups while relocating the center to the middle of all cluster employing euclidian distance.
- While working through the preceding steps, the algorithm checks and tries to reduce the sum of squared distances among clustered-point and middle for all clusters.
- When all data points unite, repetition ends.

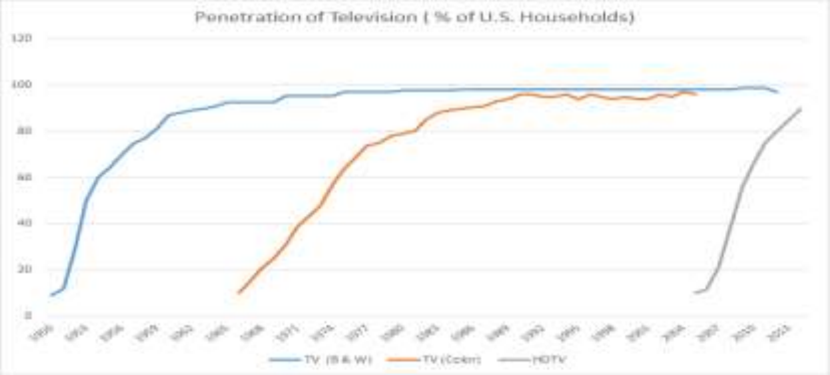
3.9.4.4 Step 4 : Tuning the optimal hyperparameters for the model

Determining the most beneficial kit of **hyperparameters** for an algorithm is the subsequent measure in customer segments with Ml because it assists us in attaining the most genuine and satisfying customer crowds.

While choosing *the **k value***, we will select upon the optimization principles of the K-means, inertia, practicing the **elbow method.**

With the elbow method, we will decide the *k* value wherever the drop in the inertia sustains.

3.9.4.5 Step 5 : Visualization of the results



At last, we visualize the decisions applying the open-source Plotly-Python**,** a plotting library in python for making i*nteractive graphs, plots, and charts.* Then we understand the charts and various graphs to develop our enterprise.

Possessing genuine consumer profiles at your fingertips will help enhance marketing operations targeting, innovation launches, and the merchandise roadmap.

It will provide your organization exceptionally more evident thoughts about which customers have the most effective retention rate, contracts, and additional metrics you initially planned.

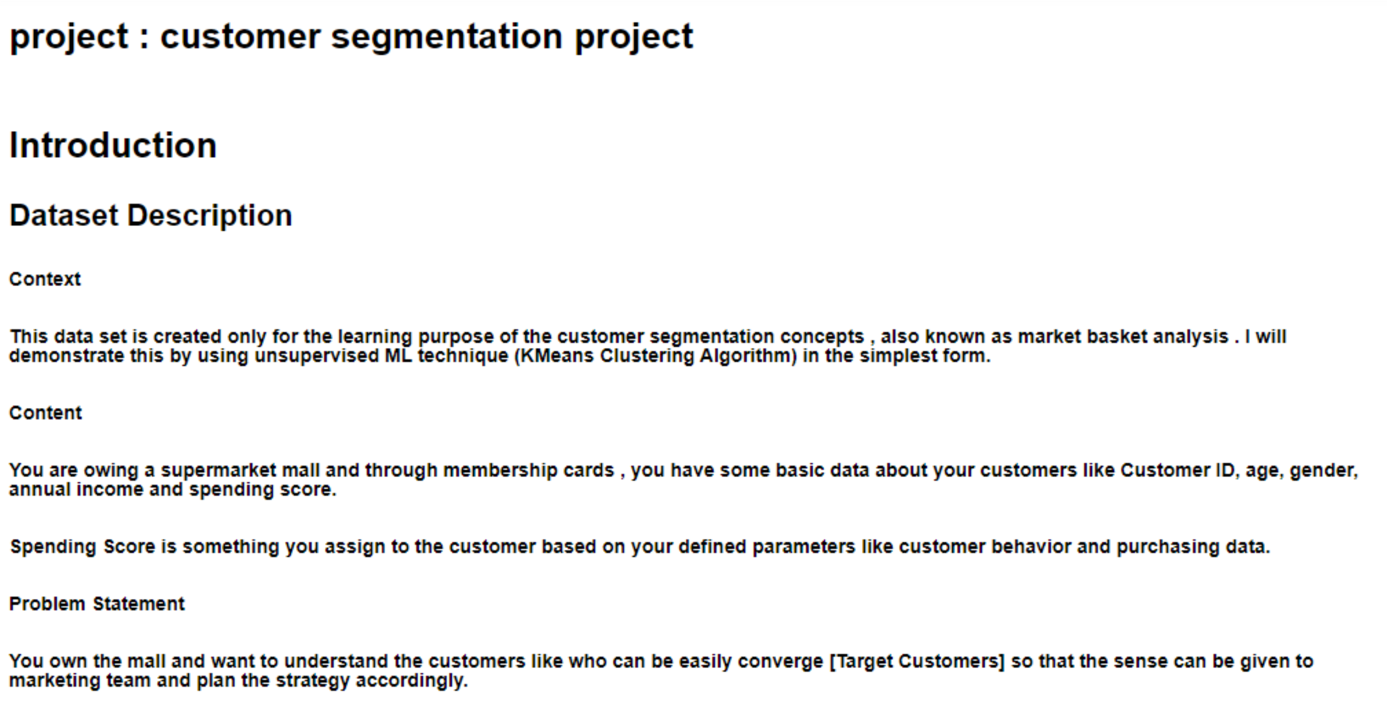
3.9.5 Conclusion

Customer segmentation is essential. Machine learning can get control over the complete process. Discovering all of the different groups that build up a more meaningful customer base permits you to get into customers' brains and give them precisely what they crave, enhancing their participation and expanding profits.

3.10 Applying clustering analysis on real data:

Project : mall customer segmentation

https://drive.google.com/file/d/1xS-qrfCqRlsRebqvs4drbUwXuQB-cDh1/view?usp=sharing



project : customer segmentation project

Introduction

Dataset Description

Context

This data set is created only for the learning purpose of the customer segmentation concepts , also known as market basket analysis . I will demonstrate this by using unsupervised ML technique (KMeans Clustering Algorithm) in the simplest form.

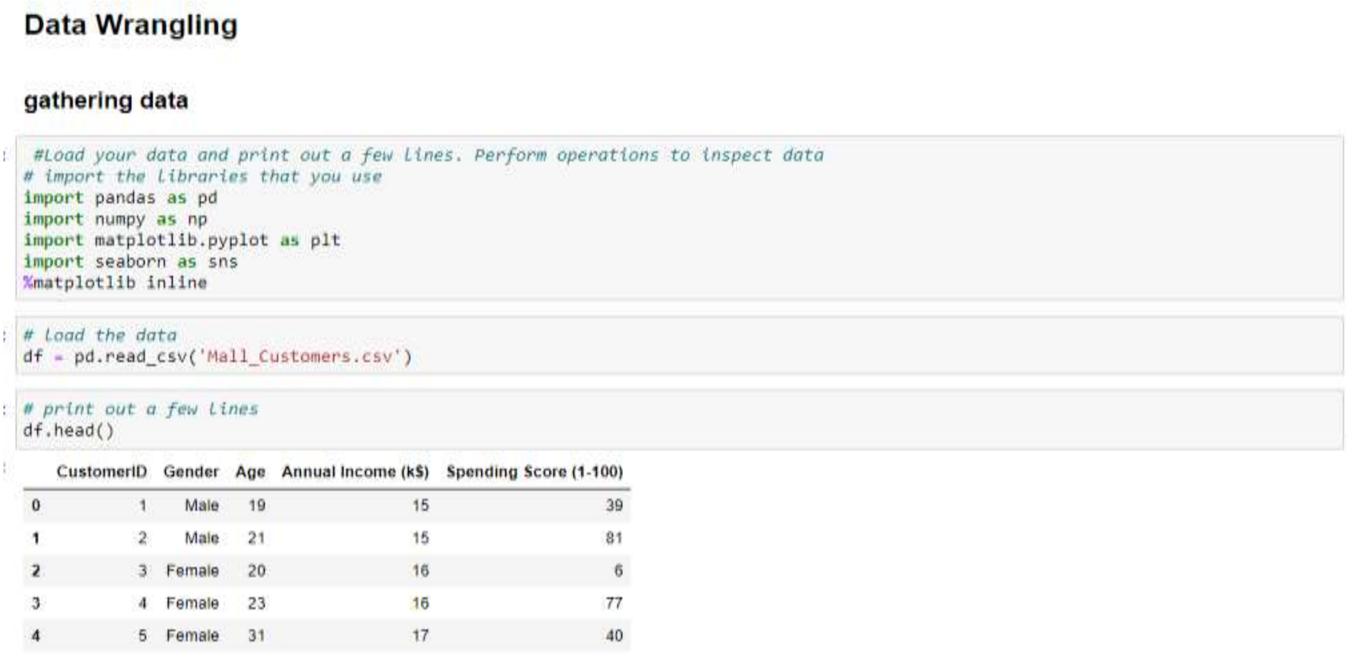
Content

You are owing a supermarket mall and through membership cards , you have some basic data about your customers like Customer ID, age, gender, annual income and spending score.

Spending Score is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.

Problem Statement

You own the mall and want to understand the customers like who can be easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly.



Data Wrangling

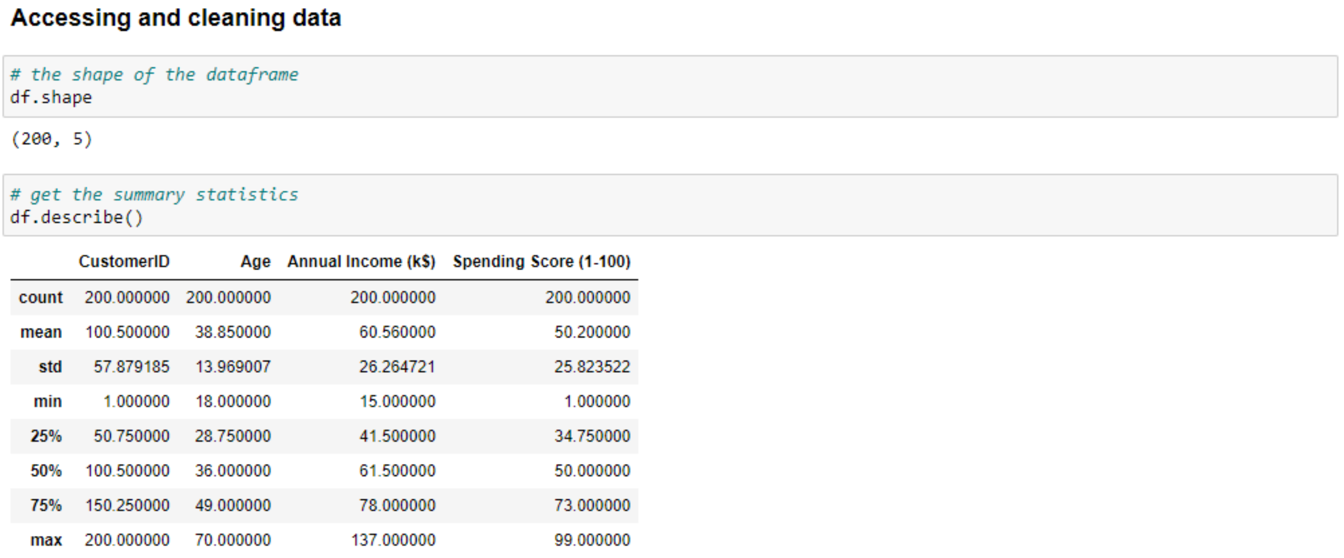
gathering data

```
 #Load your data and print out a few lines. Perform operations to inspect data
# import the libraries that you use
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

```
# Load the data
df = pd.read_csv('Mall_Customers.csv')
```

```
# print out a few lines
df.head()
```

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40



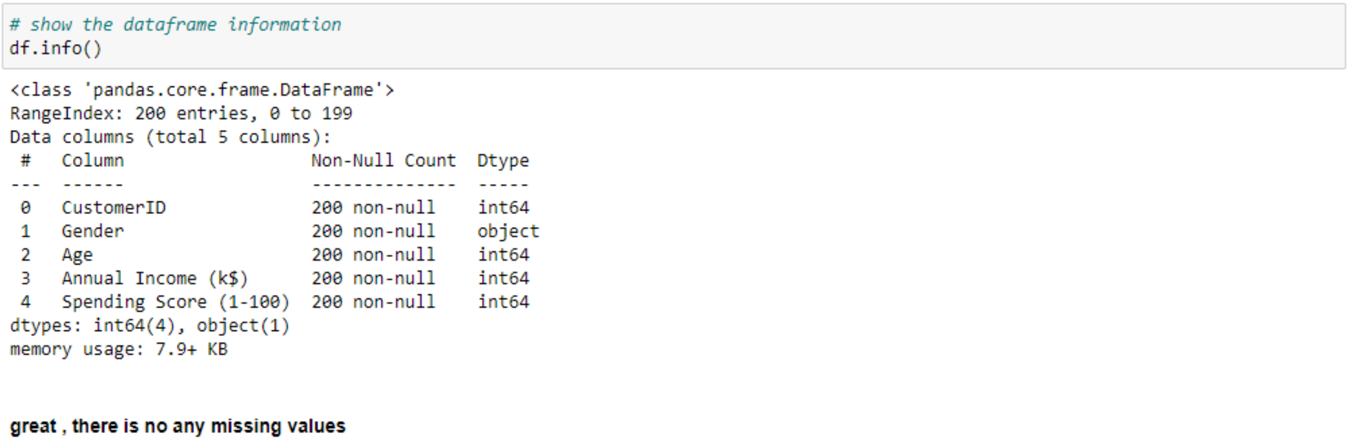
Accessing and cleaning data

```
# the shape of the dataframe
df.shape
```

```
(200, 5)
```

```
# get the summary statistics
df.describe()
```

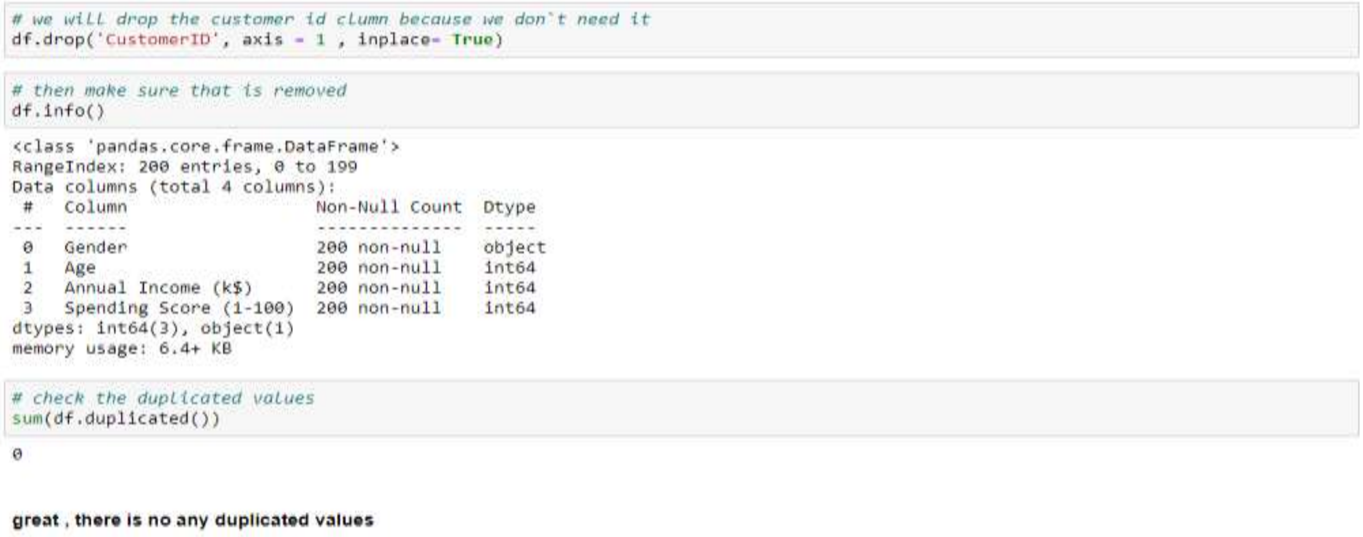
	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	200.000000	200.000000	200.000000
mean	100.500000	38.850000	60.560000	50.200000
std	57.879185	13.969007	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.750000	41.500000	34.750000
50%	100.500000	36.000000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000



```
# show the dataframe information
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   CustomerID              200 non-null    int64 
 1   Gender                  200 non-null    object
 2   Age                     200 non-null    int64 
 3   Annual Income (k$)      200 non-null    int64 
 4   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
```

great , there is no any missing values



```
# we will drop the customer id clumn because we don`t need it
df.drop('CustomerID', axis = 1 , inplace= True)
```

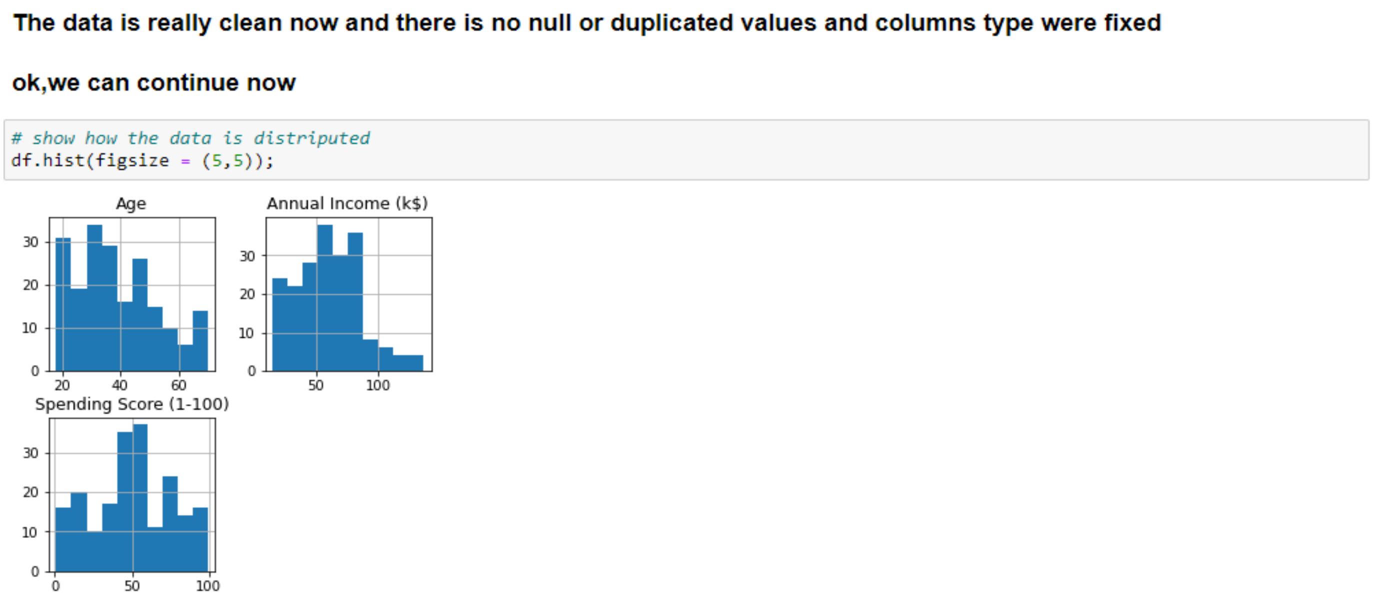
```
# then make sure that is removed
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 4 columns):
 #   Column                  Non-Null Count  Dtype 
---  ------                  --------------  ----- 
 0   Gender                  200 non-null    object
 1   Age                     200 non-null    int64 
 2   Annual Income (k$)      200 non-null    int64 
 3   Spending Score (1-100)  200 non-null    int64 
dtypes: int64(3), object(1)
memory usage: 6.4+ KB
```

```
# check the duplicated values
sum(df.duplicated())
```

```
0
```

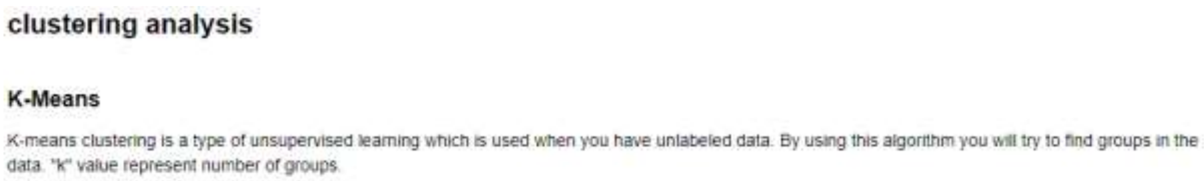
great , there is no any duplicated values



The data is really clean now and there is no null or duplicated values and columns type were fixed

ok,we can continue now

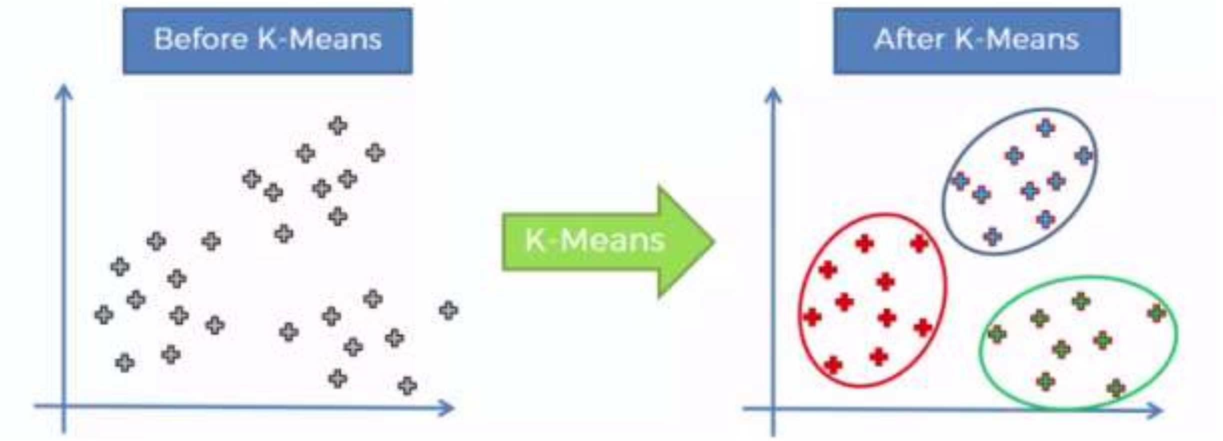
```
# show how the data is distriputed
df.hist(figsize = (5,5));
```



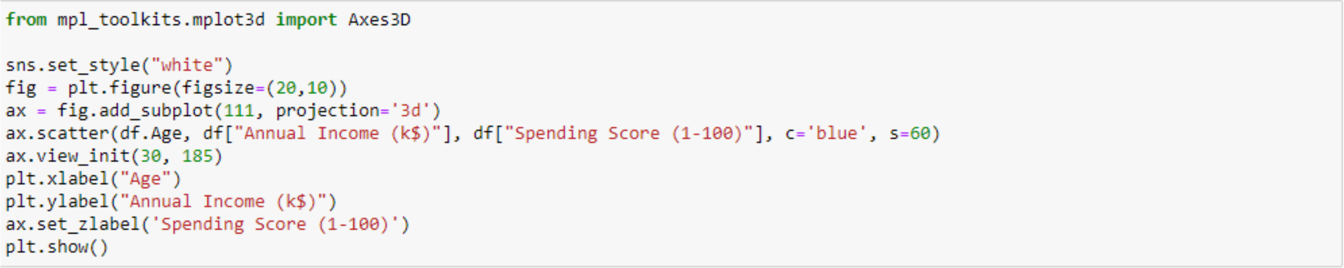
clustering analysis

K-Means

K-means clustering is a type of unsupervised learning which is used when you have unlabeled data. By using this algorithm you will try to find groups in the data. "k" value represent number of groups.

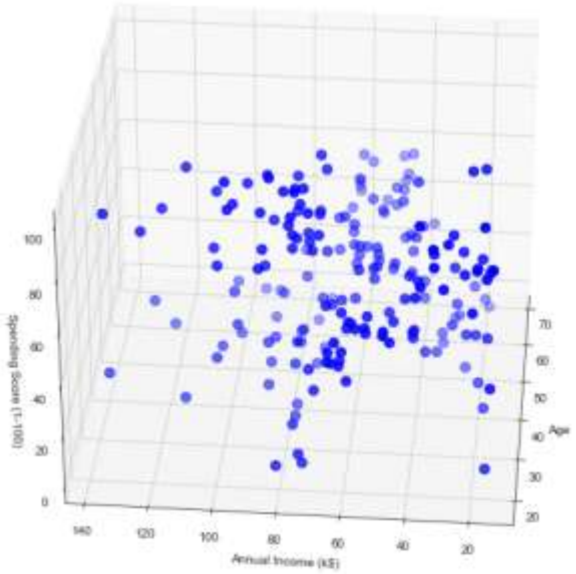


We will use Age, Annual Income and Spending Score for clustering customers. Let's look how our plot is seen without clustering.

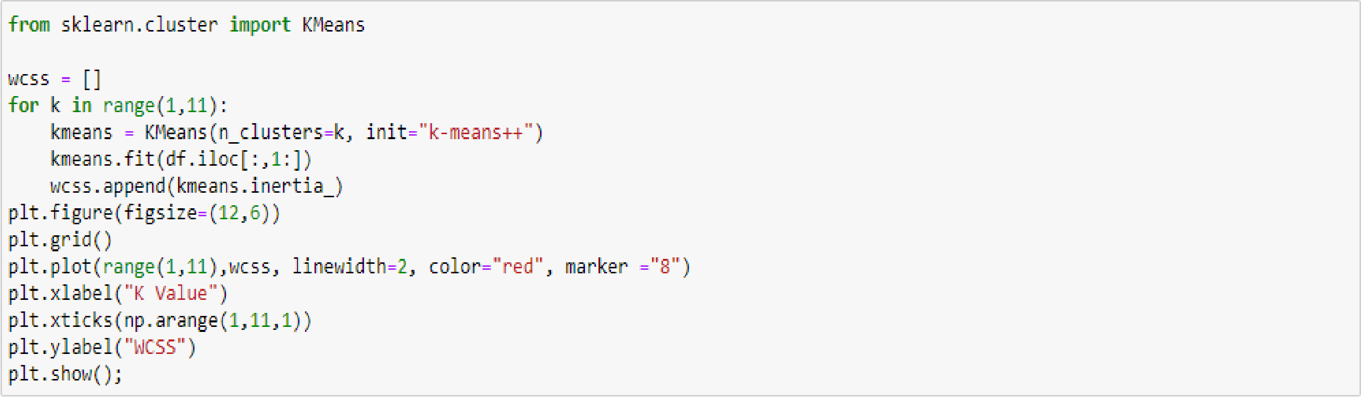


```
from mpl_toolkits.mplot3d import Axes3D

sns.set_style("white")
fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.Age, df["Annual Income (k$)"], df["Spending Score (1-100)"], c='blue', s=60)
ax.view_init(30, 185)
plt.xlabel("Age")
plt.ylabel("Annual Income (k$)")
ax.set_zlabel('Spending Score (1-100)')
plt.show()
```

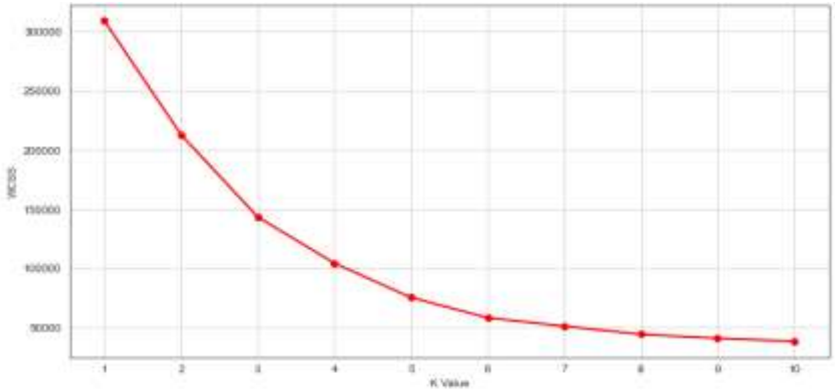


Now we will try to find what "k" value we should use. We will find out it with "elbow method". ¶

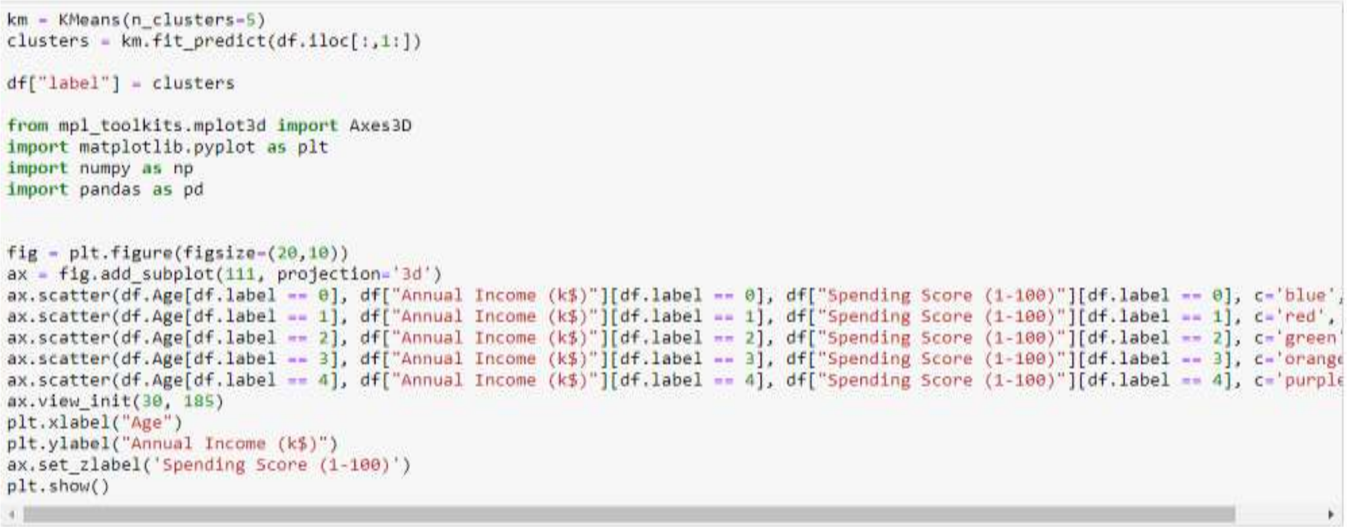


```
from sklearn.cluster import KMeans

wcss = []
for k in range(1,11):
    kmeans = KMeans(n_clusters=k, init="k-means++")
    kmeans.fit(df.iloc[:,1:])
    wcss.append(kmeans.inertia_)
plt.figure(figsize=(12,6))
plt.grid()
plt.plot(range(1,11),wcss, linewidth=2, color="red", marker ="8")
plt.xlabel("K Value")
plt.xticks(np.arange(1,11,1))
plt.ylabel("WCSS")
plt.show();
```



When we use elbow method in the above we may say 5 will be our number of cluster. Let's use K-Means and see how our plot will look like.

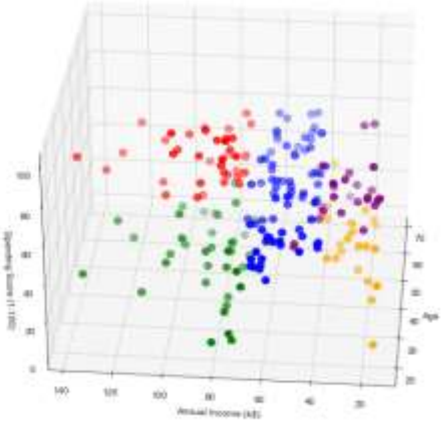


```
km = KMeans(n_clusters=5)
clusters = km.fit_predict(df.iloc[:,1:])

df["label"] = clusters

from mpl_toolkits.mplot3d import Axes3D
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd


fig = plt.figure(figsize=(20,10))
ax = fig.add_subplot(111, projection='3d')
ax.scatter(df.Age[df.label == 0], df["Annual Income (k$)"][df.label == 0], df["Spending Score (1-100)"][df.label == 0], c='blue',
ax.scatter(df.Age[df.label == 1], df["Annual Income (k$)"][df.label == 1], df["Spending Score (1-100)"][df.label == 1], c='red',
ax.scatter(df.Age[df.label == 2], df["Annual Income (k$)"][df.label == 2], df["Spending Score (1-100)"][df.label == 2], c='green'
ax.scatter(df.Age[df.label == 3], df["Annual Income (k$)"][df.label == 3], df["Spending Score (1-100)"][df.label == 3], c='orange
ax.scatter(df.Age[df.label == 4], df["Annual Income (k$)"][df.label == 4], df["Spending Score (1-100)"][df.label == 4], c='purple
ax.view_init(30, 185)
plt.xlabel("Age")
plt.ylabel("Annual Income (k$)")
ax.set_zlabel('Spending Score (1-100)')
plt.show()
```



We have 5 type of customer and we grouped them as you can see above..

conclusion:

we have already finished our research

we talked about introduction to data analysis in the first chapter , then we talked about data analysis process in the second chapter and applied on it by project " company products analysis "and in the last chapter we talked about clustering analysis and it`s algorithms and applied on by a project customer segmentation.

I hope you liked the subject and I great you and thank you for your time.

References:

https://www.ibm.com/cloud/learn/machine-learning

https://www.geeksforgeeks.org/clustering-in-machine-learning/

https://www.geeksforgeeks.org/k-means-clustering-introduction/

https://blog.quantinsti.com/hierarchical-clustering-python/

https://www.saedsayad.com/clustering_kmeans.htm

https://en.wikipedia.org/wiki/Euclidean_distance

https://www.displayr.com/what-is-hierarchical-clustering/

https://www.displayr.com/what-is-dendrogram/

https://www.geeksforgeeks.org/dbscan-clustering-in-ml-density-based-clustering/

https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/

https://www.analyticsvidhya.com/blog/2021/06/how-to-solve-customer-segmentation-problem-with-machine-learning/