# Amazon User Segmentation

**Machine Learning & AI**

**A PROJECT ON**

SUBMITTED IN PARTIAL FULFILMENT OF THE

REQUIREMENTS FOR MINI PROJECT UNDER

BACHELOR OF TECHNOLOGY
**SOFTWARE ENGINEERING**

Submitted by :-

**Davin Braven**
**2K20/SE/09**

Under the supervision of :-
**Prof. Shweta Meena**
**Department of SE, DTU**

Department of Software Engineering
DELHI TECHNOLOGICAL UNIVERSITY
( FORMERLY DELHI COLLEGE OF ENGINEERING )

# Contents

---

## Delhi Technological University

(Delhi College of Engineering)
Bawana, Delhi - 110042

## Candidate Declaration

I Davin Braven (2K20/SE/09), student of B.Tech Software Engineering Dept. declare that the Mini Project (SE391) Report Titled **“Amazon User Segmentation”** which is submitted by me to the Department of Software Engineering, Delhi Technological University, Delhi, is original and not copied from any source without proper citation. This work has not previously formed the basis for the award of any Degree, Diploma, Fellowship or other similar title or recognition.

**Prof. Shweta Meena**
**(Teacher)**

Place : DTU, Delhi, India
Date : 20 December 2022

**Department of Software Engineering**
Delhi Technological University
(Delhi College of Engineering)
Bawana, Delhi - 110042

# Certificate

I hereby certify that the Project Titled “Amazon User Segmentation” submitted by Davin Braven (2K20/SE/09), to Department of Software Engineering, Delhi Technological University, Delhi as part of Mini Project (SE391) is a record of project work carried out by the student under my supervision. To the best of my knowledge, this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

**Prof. Shweta Meena**
**(Teacher)**

Place : DTU, Delhi, India
Date : 20 December 2022

# Acknowledgement

We are thankful to Ms. Shweta Meena and all faculty members of the Software Engineering Dept. of DTU. They all provided immense support and guidance for the completion of the project undertaken by us. It is with their supervision that this work came into existence.

We would also like to express my gratitude to the university for providing the laboratories, infrastructure, test facilities and environment which allowed us to work without any obstructions.

We would also like to appreciate the support provided by our lab assistants, seniors and peer group who aided us with all the knowledge they had regarding various topics.

Name - DAVIN BRAVEN
Roll No. - 2K20/SE/09

# Introduction

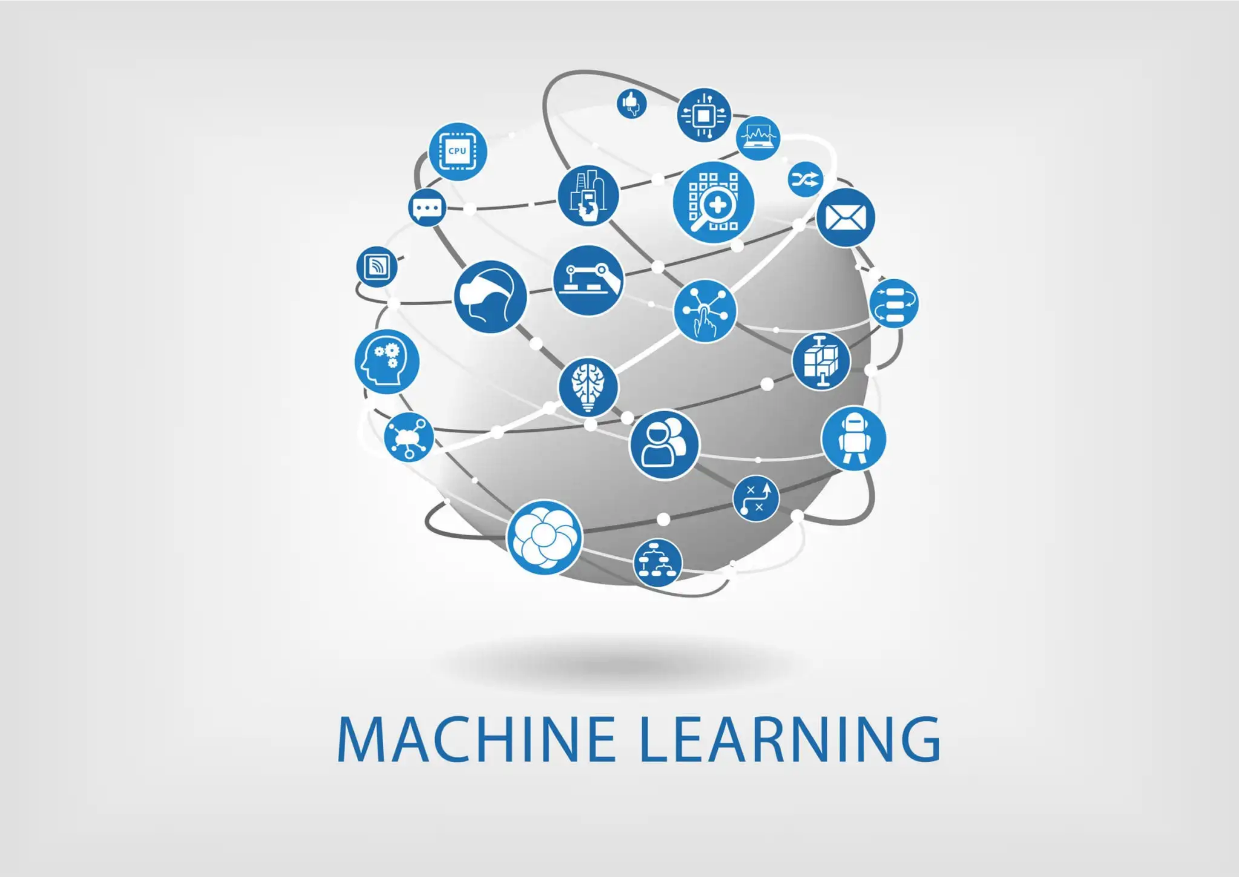
Amazon Personalize now offers intelligent user segmentation which allows you to run more effective prospecting campaigns through your marketing channels. Traditionally, user segmentation has relied on demographic information and manually curated business rules to make assumptions about users' intentions and assign them to pre-defined audience segments. Amazon Personalize uses machine learning techniques to learn about your items, users, and how your users interact with your items. Amazon Personalize segments users based on their preferences for different products, categories, brands, and more. This can help you drive higher engagement with marketing campaigns, increase retention through targeted messaging, and improve the return on investment for your marketing spend.

Our new recipes are simple to use. Provide Amazon Personalize with data about your items and your users' interactions and Amazon Personalize will learn your users' preferences. When given an item or item-attribute Amazon Personalize recommends a list of users sorted by their propensity to interact with the item or items that share the attribute.

Amazon Personalize enables you to personalise your website, app, ads, emails, and more, using the same machine learning technology as used by Amazon, without requiring any prior machine learning experience. To get started with Amazon Personalize, visit our documentation.

# Machine Learning & AI

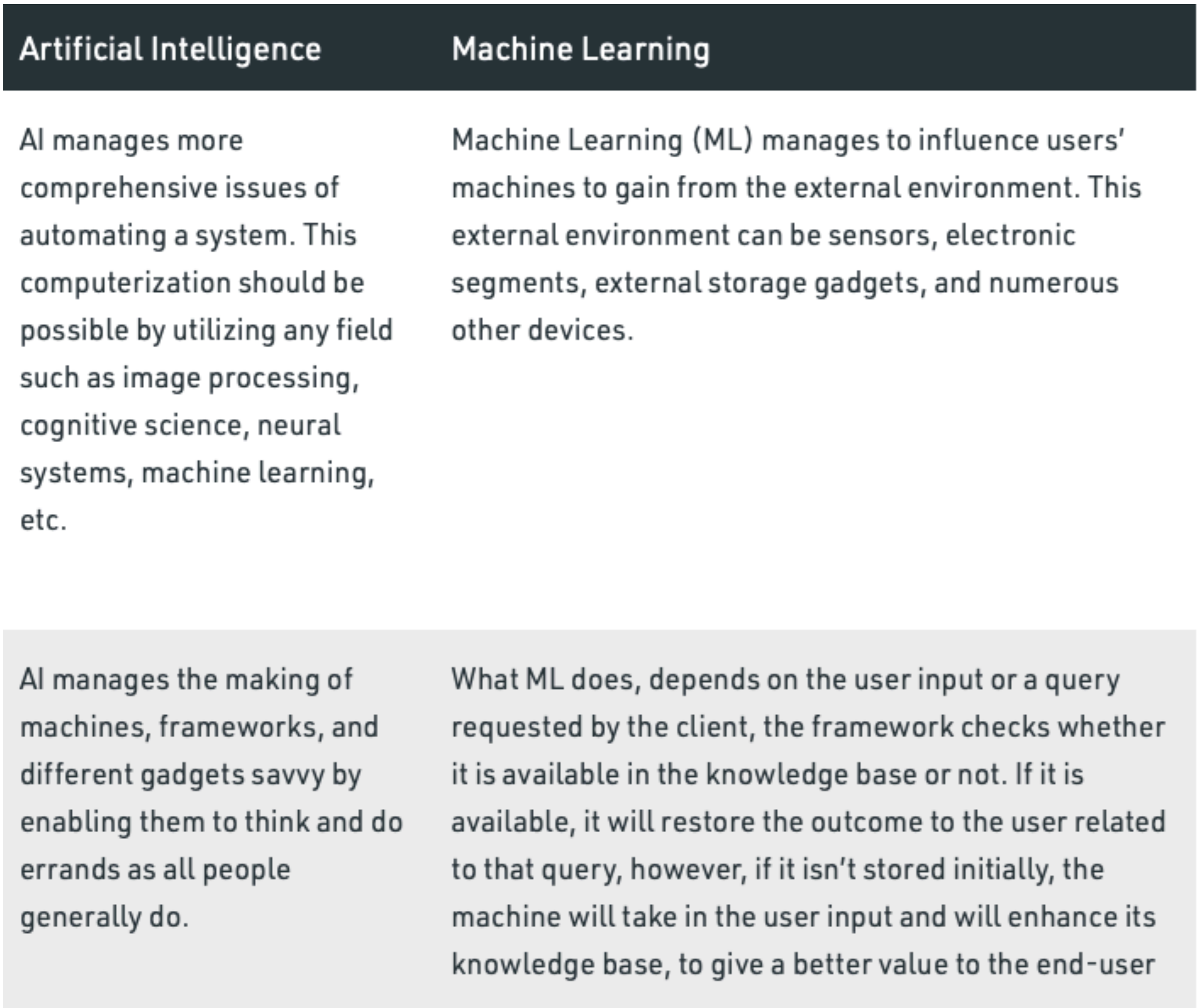
**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that makes it more similar to humans: **The ability to learn**. Machine learning is actively being used today, perhaps in many more places than one would expect.



Arthur Samuel, an early American leader in the field of computer gaming and artificial intelligence, coined the term “Machine Learning ” in 1959 while at IBM. He defined machine learning as “the field of study that gives computers the ability to learn without being explicitly programmed “. However, there is no universally accepted definition for machine learning. Different authors define the term differently.

**Working of Virtual Personal Assistants –**
**Siri**(part of Apple Inc.'s iOS, watchOS, macOS, and tvOS operating systems), **Google Now** (a feature of Google Search offering predictive cards with information and daily updates in the Google app for Android and iOS.), **Cortana** (Cortana is a virtual assistant created by Microsoft for Windows 10) are intelligent digital personal assistants on the platforms like iOS, Android and Windows respectively. To put it plainly, they help to find relevant information when requested using voice. For instance, for answering queries like 'What's the temperature today?' or 'What is the way to the nearest supermarket' etc. and the assistant will react by searching for information, transferring that information from the phone, or sending commands to various other applications.

AI is critical in these applications, as they gather data on the user's request and utilize that data to perceive speech in a better manner and serve the user with answers that are customized to his inclination. **Microsoft says that Cortana "consistently finds out about its user" and that it will in the end build up the capacity to anticipate users' needs and cater to them.** Virtual assistants process a tremendous measure of information from an assortment of sources to find out about users and be more compelling in helping them arrange and track their data. Machine learning is a vital part of these personal assistants as they gather and refine the data based on users' past participation with them. Thereon, this arrangement of information is used to render results that are custom-made to users' inclinations.
Roughly speaking, Artificial Intelligence (AI) is when a computer algorithm does intelligent work. On the other hand, Machine Learning is a part of AI that learns from the data that also involves the information gathered from previous experiences and allows the computer program to change its behavior accordingly. **Artificial Intelligence is the superset of Machine Learning** i.e. all Machine Learning is Artificial Intelligence but not all AI.



| Artificial Intelligence | Machine Learning |
|---|---|
| AI manages more comprehensive issues of automating a system. This computerization should be possible by utilizing any field such as image processing, cognitive science, neural systems, machine learning, etc. | Machine Learning (ML) manages to influence users' machines to gain from the external environment. This external environment can be sensors, electronic segments, external storage gadgets, and numerous other devices. |
| AI manages the making of machines, frameworks, and different gadgets savvy by enabling them to think and do errands as all people generally do. | What ML does, depends on the user input or a query requested by the client, the framework checks whether it is available in the knowledge base or not. If it is available, it will restore the outcome to the user related to that query, however, if it isn't stored initially, the machine will take in the user input and will enhance its knowledge base, to give a better value to the end-user |

**Data in Machine Learning :-**

It can be any unprocessed fact, value, text, sound, or picture that is not being interpreted and analyzed. Data is the most important part of all Data Analytics, Machine Learning, Artificial Intelligence. Without data, we can't train any model and all modern research and automation will go in vain. Big Enterprises are spending lots of money just to gather as much certain data as possible.

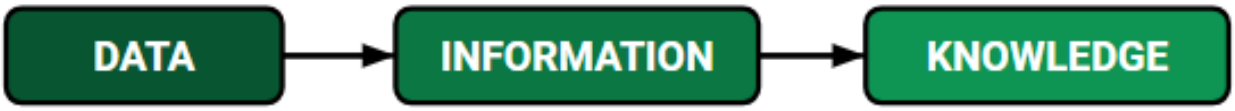
**Example:** Why did Facebook acquire WhatsApp by paying a huge price of $19 billion?

The answer is very simple and logical – it is to have access to the users' information that Facebook may not have but

WhatsApp will have. This information of their users is of paramount importance to Facebook as it will facilitate the task of improvement in their services.

**INFORMATION:** Data that has been interpreted and manipulated and has now some meaningful inference for the users.

**KNOWLEDGE:** Combination of inferred information, experiences, learning, and insights. Results in awareness or concept building for an individual or organization.

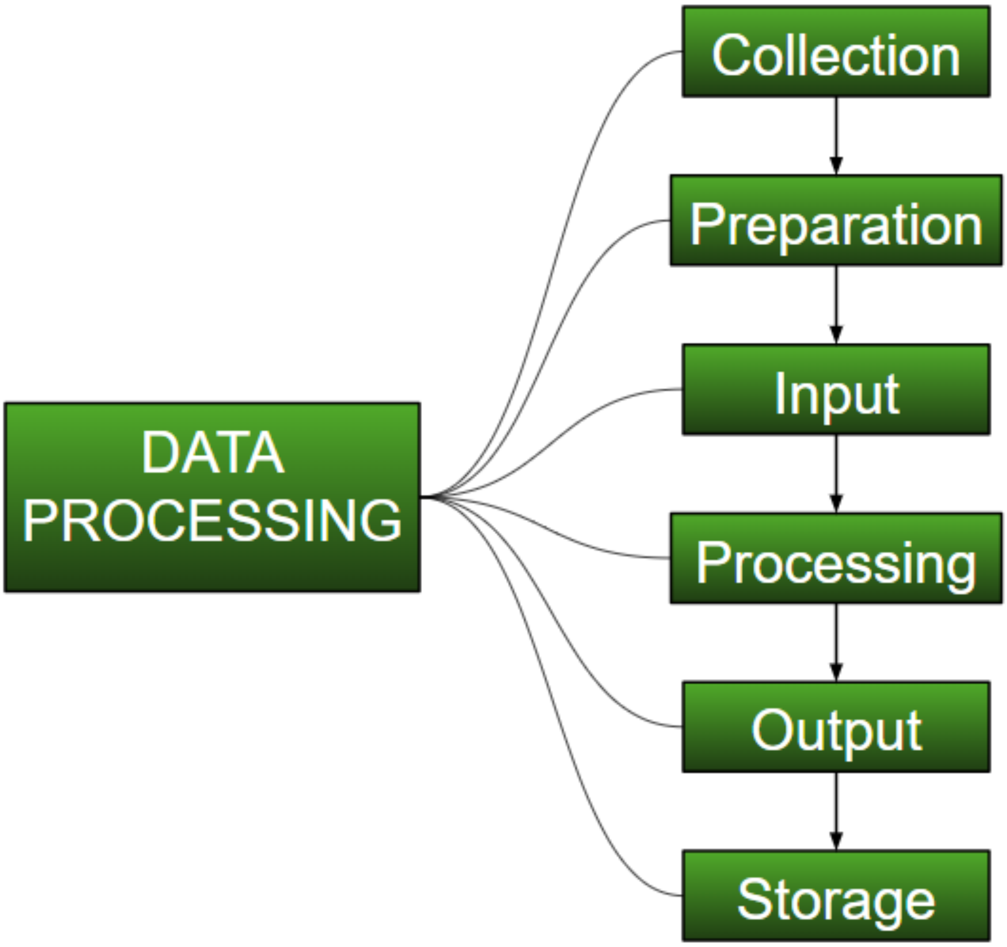


**How we split data in Machine Learning?**

- **Training Data:** The part of data we use to train our model. This is the data that your model actually sees(both input and output) and learns from.
- **Validation Data:** The part of data that is used to do a frequent evaluation of the model, fit on the training dataset along with improving involved hyperparameters (initially set parameters before the model begins learning). This data plays its part when the model is actually training.
- **Testing Data:** Once our model is completely trained, testing data provides an unbiased evaluation. When we feed in the inputs of Testing data, our model will predict some values(without seeing actual output). After prediction, we evaluate our model by comparing it with the actual output present in the testing data. This is how we evaluate and see how much our model has learned from the experiences feed in as training data, set at the time of training.

# Data Pre Processing

Data Processing is the task of converting data from a given form to a much more usable and desired form i.e. making it more meaningful and informative. Using Machine Learning algorithms, mathematical modeling, and statistical knowledge, this entire process can be automated. The output of this complete process can be in any desired form like graphs, videos, charts, tables, images, and many more, depending on the task we are performing and the requirements of the machine. This might seem to be simple but when it comes to massive organizations like Twitter, Facebook, Administrative bodies like Parliament, UNESCO, and health sector organizations, this entire process needs to be performed in a very structured manner. So, the steps to perform are as follows:

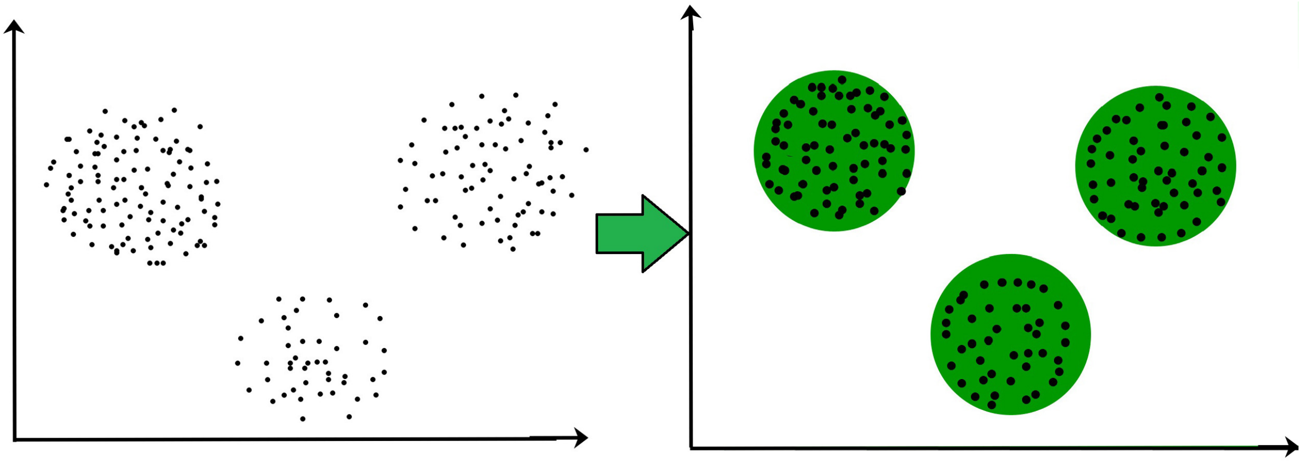


# Clustering in Machine Learning

It is basically a type of unsupervised learning method. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

**Clustering** 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.



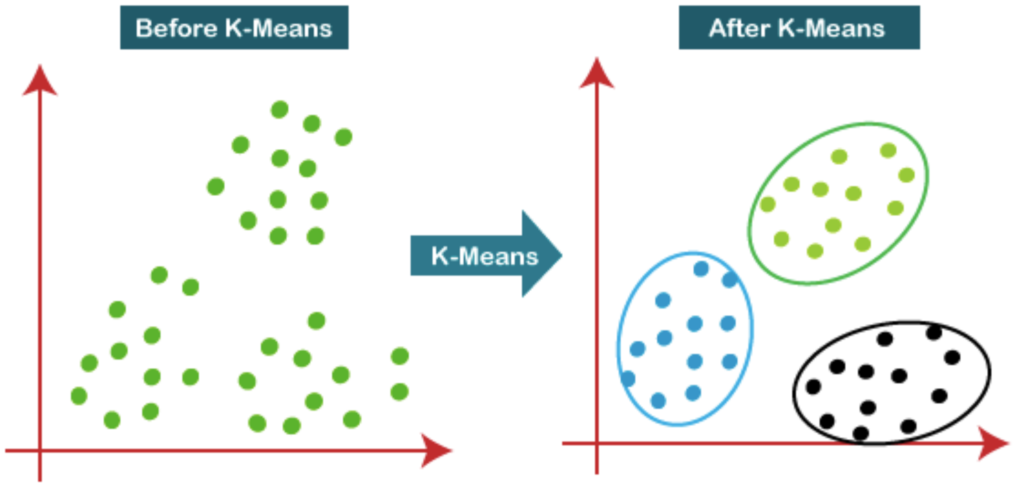
Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, what is the criteria they may use which satisfy their need.

# K - Means Clustering

K-Means Clustering is an Unsupervised Learning algorithm, which groups the unlabeled dataset into different clusters. Here K defines the number of pre-defined clusters that need to be created in the process, as if K=2, there will be two clusters, and for K=3, there will be three clusters, and so on.

> It is an iterative algorithm that divides the unlabeled dataset into k different clusters in such a way that each dataset belongs only one group that has similar properties.

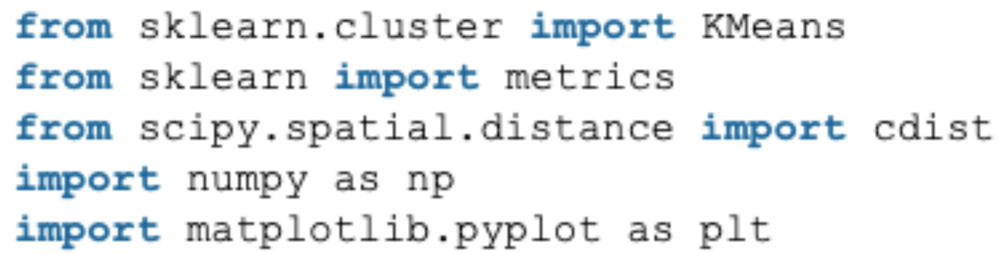
It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training.
It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.
The algorithm takes the unlabeled dataset as input, divides the dataset into k-number of clusters, and repeats the process until it does not find the best clusters. The value of k should be predetermined in this algorithm.
Hence each cluster has datapoints with some commonalities, and it is away from other clusters.
The below diagram explains the working of the K-means Clustering Algorithm:



# Elbow Method for Optimal of K in K-Means

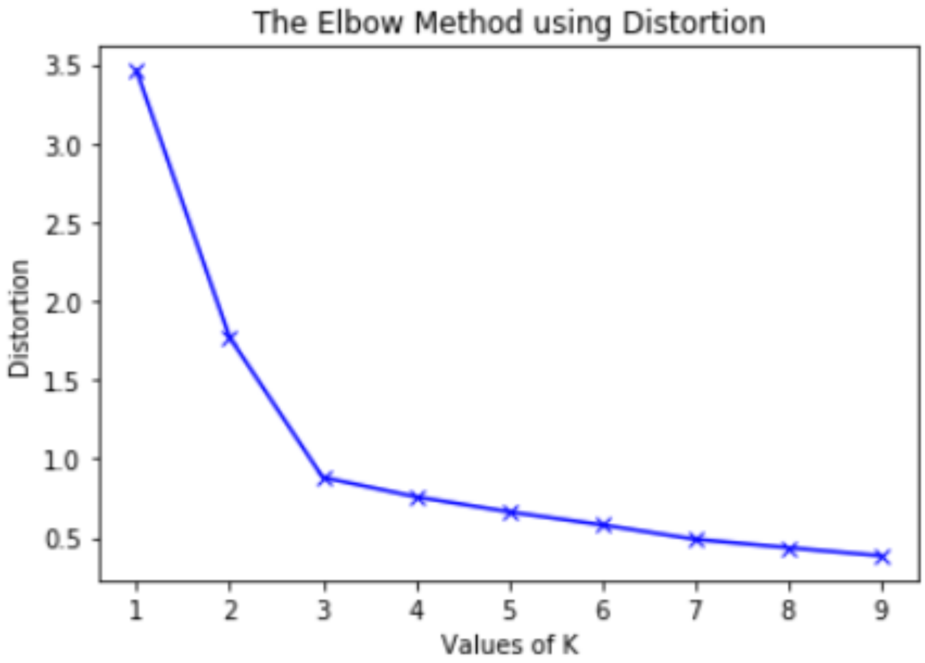
A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The **Elbow Method** is one of the most popular methods to determine this optimal value of k.
We now demonstrate the given method using the K-Means clustering technique using the **Sklearn** library of python.

## Step 1: Importing the required libraries.



```
from sklearn.cluster import KMeans
from sklearn import metrics
from scipy.spatial.distance import cdist
import numpy as np
import matplotlib.pyplot as plt
```

**Step 2: Creating and Visualizing the data.**
**Step 3: Building the clustering model and calculating the values of the Distortion and Inertia.**
**Step 4: Tabulating and Visualizing the results.**

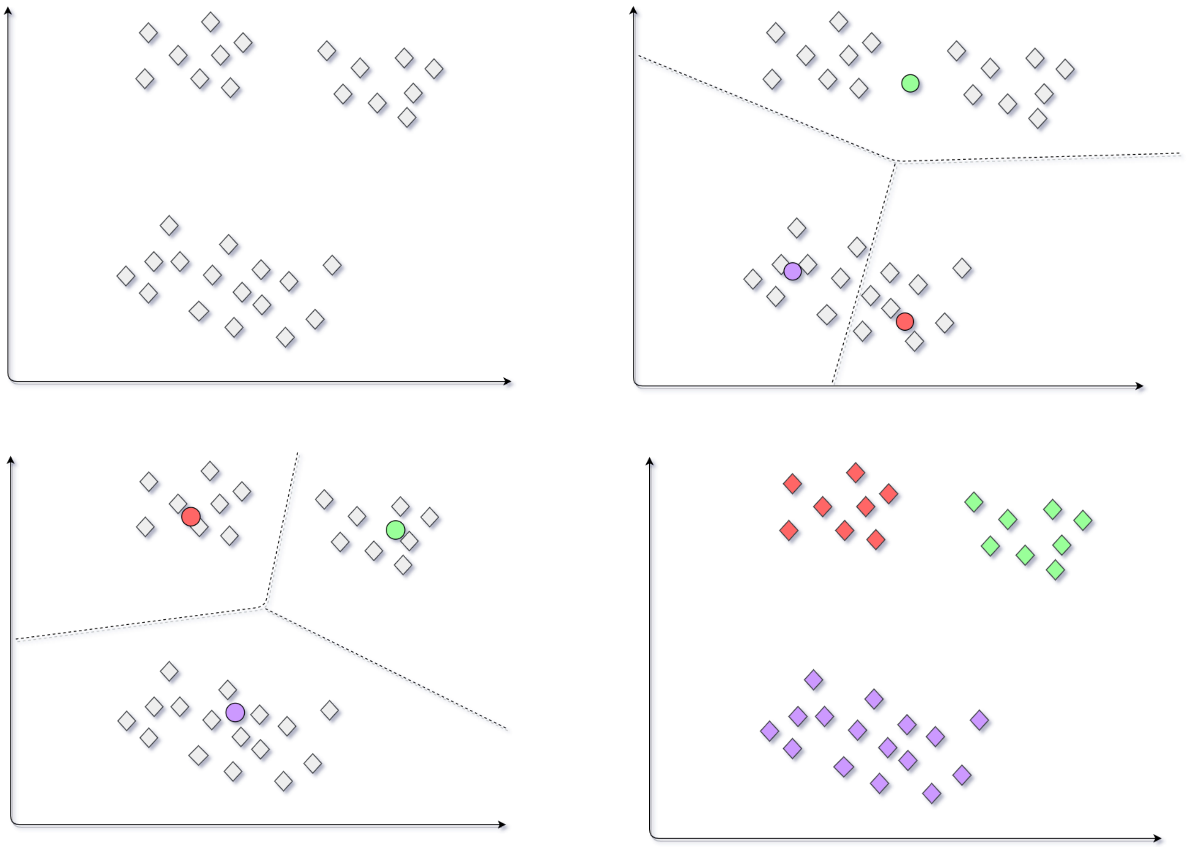


# Random Initialization Trap in k-Means

Random initialization trap is a problem that occurs in the K-means algorithm. In random initialization trap when the centroids of the clusters to be generated are explicitly defined by the User then inconsistency may be created and this may sometimes lead to generating wrong clusters in the dataset. So random initialization trap may sometimes prevent us from developing the correct clusters.

**Example :**

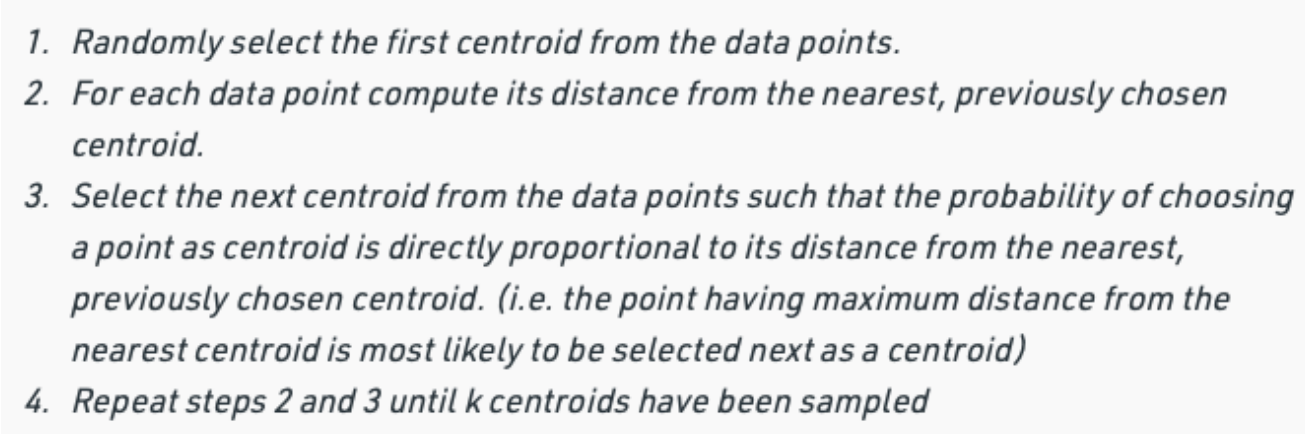
Suppose you have a dataset with the following points shown in the picture and you want to generate three clusters in this dataset according to their attributes by performing K-means clustering. From the figure, we can get the intuition what are the clusters that are required to be generated. K-means will perform clustering on the basis of the centroids fed into the algorithm and generate the required clusters according to these centroids.



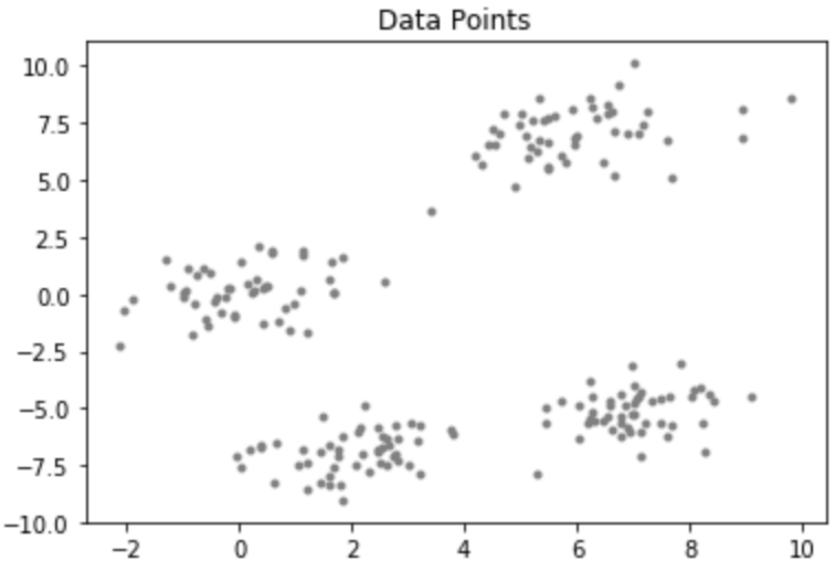
# KMeans++ Algorithm

One disadvantage of the K-means algorithm is that it is sensitive to the initialization of the centroids or the mean points. So, if a centroid is initialized to be a “far-off” point, it might just end up with no points associated with it, and at the same time, more than one cluster might end up linked with a single centroid. Similarly, more than one centroids might be initialized into the same cluster resulting in poor clustering.

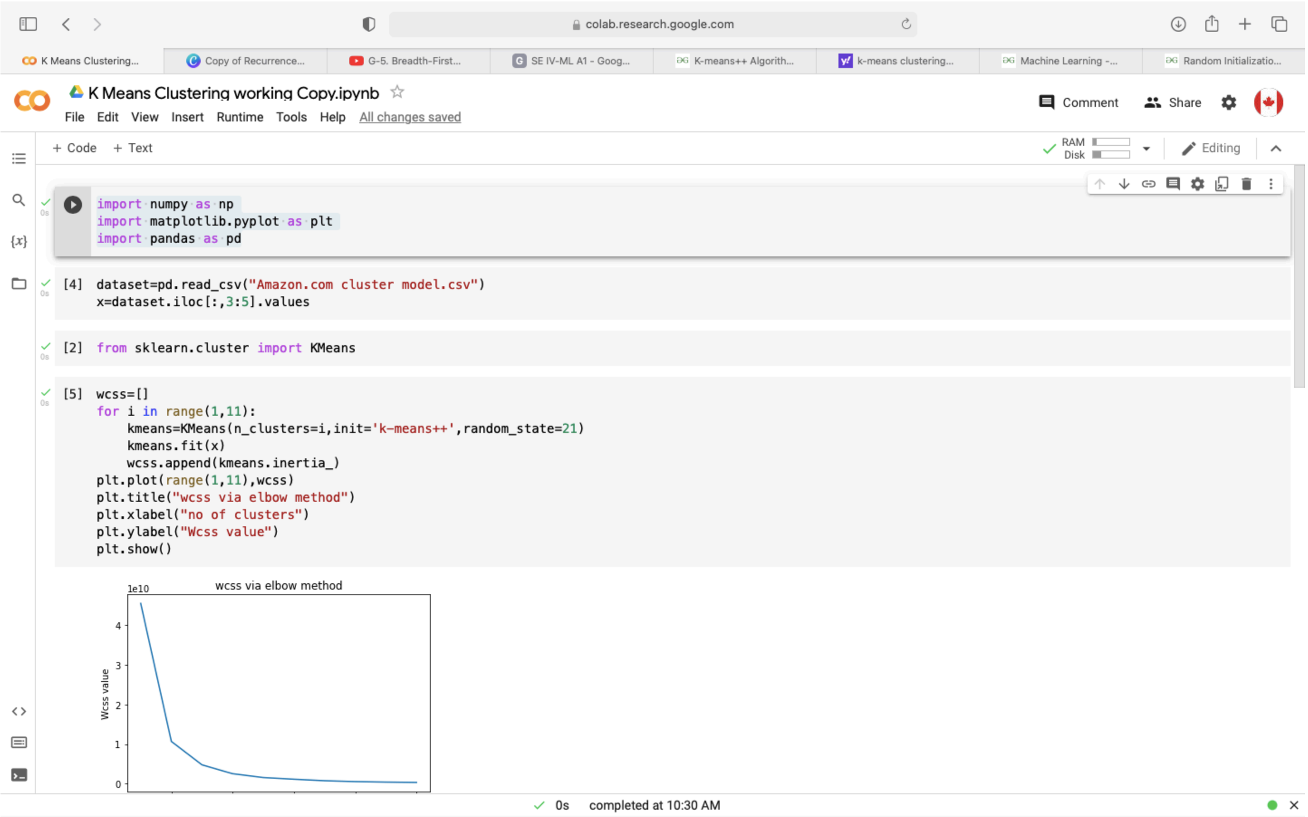
To overcome the above drawback we use K-means++. This algorithm ensures a smarter initialization of the centroids and improves the quality of the clustering. Apart from initialization, the rest of the algorithm is the same as the standard K-means algorithm. That is K-means++ is the standard K-means algorithm coupled with a smarter initialization of the centroids.



1. *Randomly select the first centroid from the data points.*
2. *For each data point compute its distance from the nearest, previously chosen centroid.*
3. *Select the next centroid from the data points such that the probability of choosing a point as centroid is directly proportional to its distance from the nearest, previously chosen centroid. (i.e. the point having maximum distance from the nearest centroid is most likely to be selected next as a centroid)*
4. *Repeat steps 2 and 3 until k centroids have been sampled*



# Project Code with Outputs



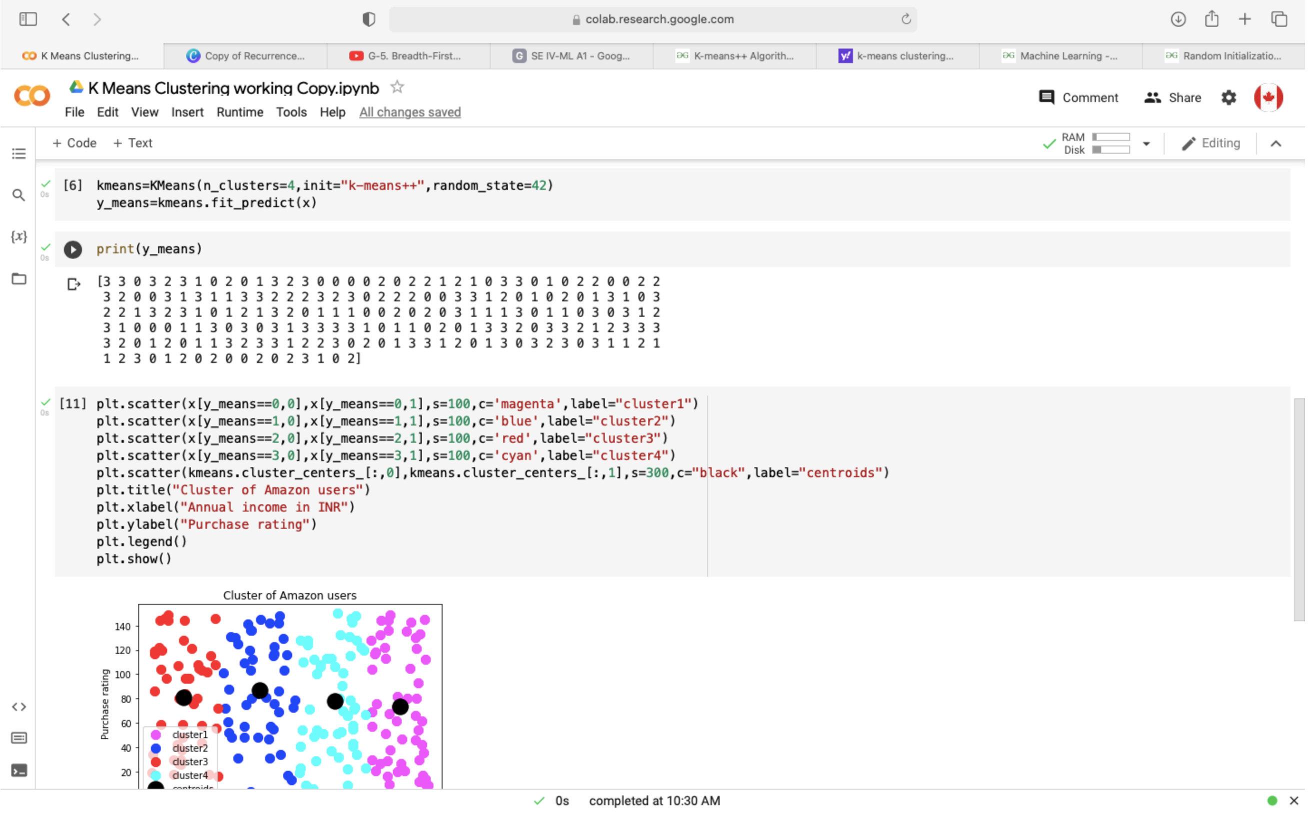
```
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
```

```
[4]  dataset=pd.read_csv("Amazon.com cluster model.csv")
     x=dataset.iloc[:,3:5].values
```

```
[2]  from sklearn.cluster import KMeans
```

```
[5]  wcss=[]
     for i in range(1,11):
         kmeans=KMeans(n_clusters=i,init='k-means++',random_state=21)
         kmeans.fit(x)
         wcss.append(kmeans.inertia_)
     plt.plot(range(1,11),wcss)
     plt.title("wcss via elbow method")
     plt.xlabel("no of clusters")
     plt.ylabel("Wcss value")
     plt.show()
```

wcss via elbow method
1e10
Wcss value
4
3
2
1
0
0s completed at 10:30 AM



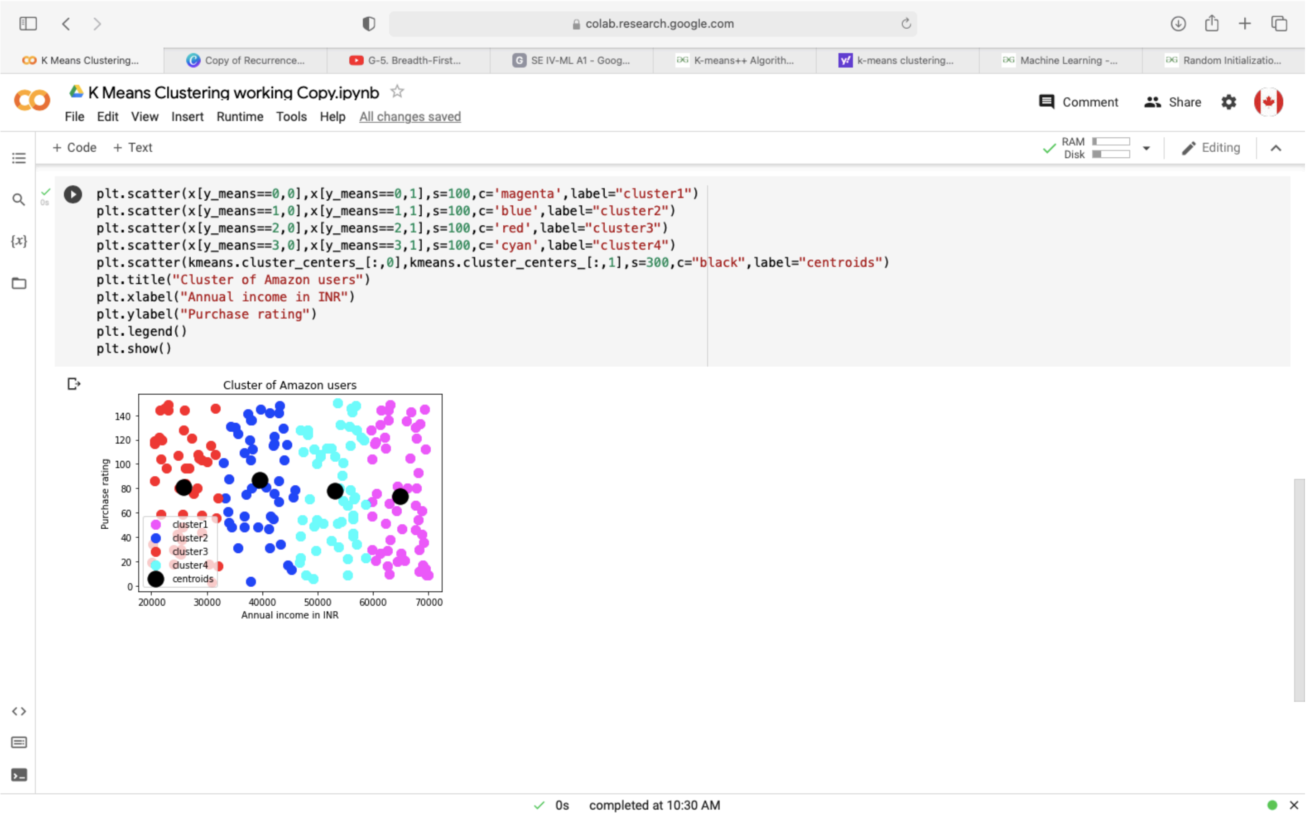
colab.research.google.com
K Means Clustering working Copy.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment
Share
+ Code + Text
Editing

```
[6]  kmeans=KMeans(n_clusters=4,init="k-means++",random_state=42)
     y_means=kmeans.fit_predict(x)
```

```
print(y_means)
```

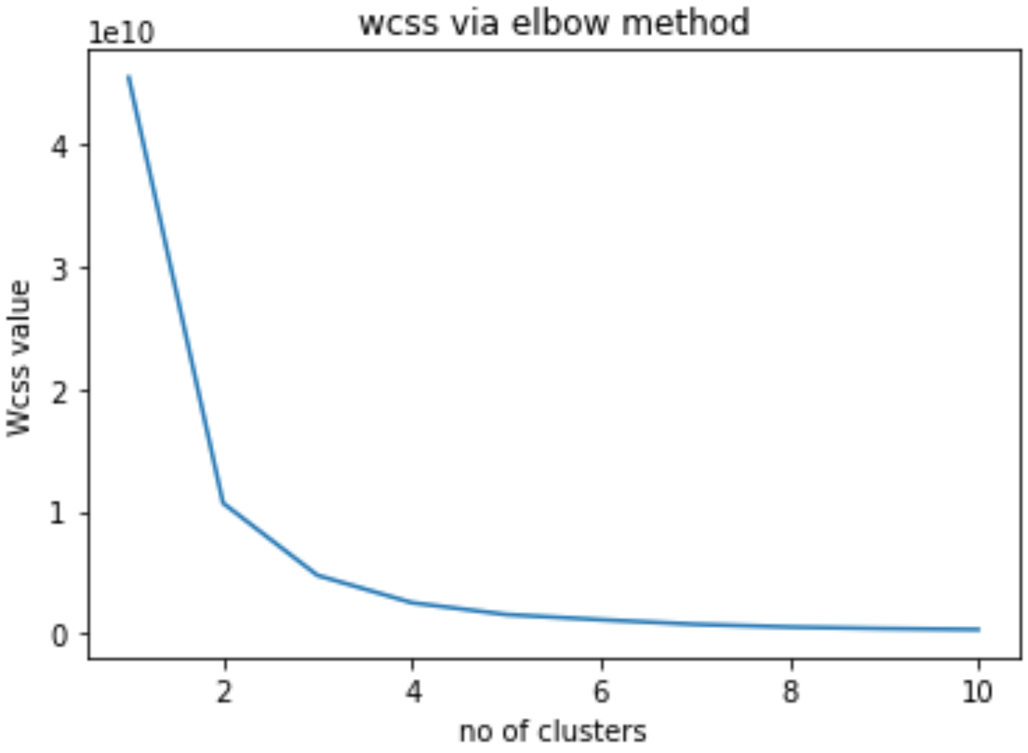
```
[3 3 0 3 2 3 1 0 2 0 1 3 2 3 0 0 0 0 2 0 2 2 1 2 1 0 3 3 0 1 0 2 2 0 0 2 2
 3 2 0 0 3 1 3 1 1 3 3 2 2 2 3 2 3 0 2 2 2 0 0 3 3 1 2 0 1 0 2 0 1 3 1 0 3
 2 2 1 3 2 3 1 0 1 2 1 3 2 0 1 1 1 0 0 2 0 2 0 3 1 1 1 3 0 1 1 0 3 0 3 1 2
 3 1 0 0 0 1 1 3 0 3 0 3 1 3 3 3 3 1 0 1 1 0 2 0 1 3 3 2 0 3 3 2 1 2 3 3 3
 3 2 0 1 2 0 1 1 3 2 3 3 1 2 2 3 0 2 0 1 3 3 1 2 0 1 3 0 3 2 3 0 3 1 1 2 1
 1 2 3 0 1 2 0 2 0 0 2 0 2 3 1 0 2]
```

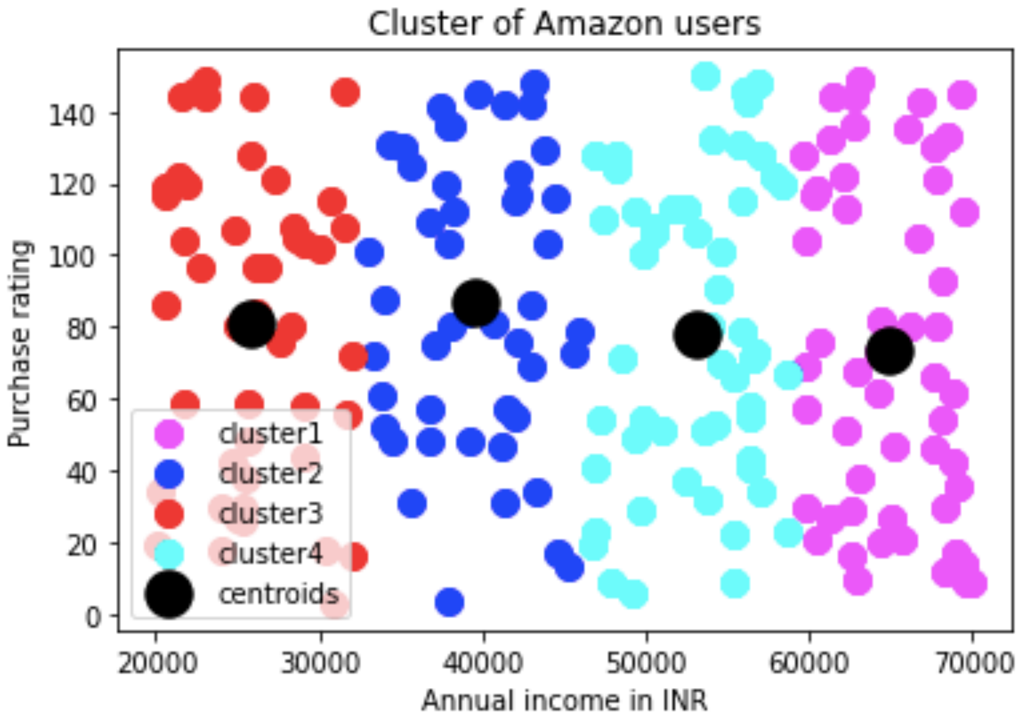
```
[11] plt.scatter(x[y_means==0,0],x[y_means==0,1],s=100,c='magenta',label="cluster1")
     plt.scatter(x[y_means==1,0],x[y_means==1,1],s=100,c='blue',label="cluster2")
     plt.scatter(x[y_means==2,0],x[y_means==2,1],s=100,c='red',label="cluster3")
     plt.scatter(x[y_means==3,0],x[y_means==3,1],s=100,c='cyan',label="cluster4")
     plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=300,c="black",label="centroids")
     plt.title("Cluster of Amazon users")
     plt.xlabel("Annual income in INR")
     plt.ylabel("Purchase rating")
     plt.legend()
     plt.show()
```



```
plt.scatter(x[y_means==0,0],x[y_means==0,1],s=100,c='magenta',label="cluster1")
plt.scatter(x[y_means==1,0],x[y_means==1,1],s=100,c='blue',label="cluster2")
plt.scatter(x[y_means==2,0],x[y_means==2,1],s=100,c='red',label="cluster3")
plt.scatter(x[y_means==3,0],x[y_means==3,1],s=100,c='cyan',label="cluster4")
plt.scatter(kmeans.cluster_centers_[:,0],kmeans.cluster_centers_[:,1],s=300,c="black",label="centroids")
plt.title("Cluster of Amazon users")
plt.xlabel("Annual income in INR")
plt.ylabel("Purchase rating")
plt.legend()
plt.show()
```

✓ 0s completed at 10:30 AM


wcss via elbow method
1e10
Wcss value
4
3
2
1
0
2
4
6
8
10
no of clusters


Cluster of Amazon users
Purchase rating
140
120
100
80
60
40
20
0
20000
30000
40000
50000
60000
70000
Annual income in INR
cluster1
cluster2
cluster3
cluster4
centroids

# Conclusion

Customer segmentation simply means grouping your customers according to various characteristics (for example grouping customers by age).

It’s a way for organizations to understand their customers. Knowing the differences between customer groups, it’s easier to make strategic decisions regarding product growth and marketing.

**The opportunities to segment are endless and depend mainly on how much customer data you have at your use.** Starting from the basic criteria, like gender, hobby, or age, it goes all the way to things like “time spent of website X” or “time since user opened our app”.

There are different methodologies for customer segmentation, and they depend on four types of parameters:

- geographic,
- demographic,
- behavioral,
- psychological.

**Geographic** customer segmentation is very simple, it’s all about the user’s location. This can be implemented in various ways. You can group by country, state, city, or zip code.

**Demographic** segmentation is related to the structure, size, and movements of customers over space and time. Many companies use gender differences to create and market products. Parental status is another important feature. You can obtain data like this from customer surveys.

**Behavioral** customer segmentation is based on past observed behaviors of customers that can be used to predict future actions. For example, brands that customers purchase, or moments when they buy the most. The behavioral aspect of customer segmentation not only tries to understand reasons for purchase but also how those reasons change throughout the year.

**Psychological** segmentation of customers generally deals with things like personality traits, attitudes, or beliefs. This data is obtained using customer surveys, and it can be used to gauge customer sentiment.

# Reference

– Apple Computer Inc. (1987) *Apple Human Interface Guidelines: The Apple Desktop Interface*. Reading, MA: Adison-Wesley.

Baecker, R. M., and Buxton, W. A. S. (1987) 'An Historical and Intellectual Perspective.' In R. M. Baecker & W. A. S. Buxton (Eds.), *Readings in Human-Computer Interaction* San Mateo, California: Morgan Kaufman Publishers.

Bush, V. (1945) 'As we may think.' *Atlantic Monthly*, 76, 1, 101–108.
Card, S., Moran, T. P., and Newell, A. (1983) *The Psychology of Human Computer Interaction.*

Hillsdale, N.J.: Lawrence Erlbaum Associates.
Card, S. K., Moran, T. P., and Newell, A. (1980) 'The keystroke-level model for user

performance time with interactive systems.' *Communication of the ACM*, 23, 396–410.

Diaper, D. (1989) 'Task Analysis for Knowledge Descriptions (TAKD); the method and an example.' In D. Diaper (Eds.), *Task analysis for Human-Computer Interaction,* (pp. 108– 159). Chichester: Ellis-Horwood.

Dix, A., Finlay, J., Abowd, G., and Beale, R. (1993) *Human-Computer Interaction*. New York: Prentice Hall.

Engelbart, D.C., and English, W.K. (1988) A research center for augmenting human intellect. In Greif, I. (Ed.) *Computer-Supported Cooperative Work: A Book of Readings*. (pp. 81–105). Palo Alto: Morgan Kaufmann.

Fitts, P. M. (1954) 'The information capacity of the human motor system in controlling amplitude of movement.' *Journal of Experimental Psychology*, 47, 381–391.

Flower, L. S., and Hayes, J. R. (1980) 'The dynamics of composing: making plans and juggling constraints.' In L. Gregg & E. Steinberg (Eds.), *Cognitive Processes in Writing: an Interdisciplinary Approach* (pp. 31-49). Hillsdale NJ: Lawrence Erlabum.

Harel, D. (1988) 'On visual formalisms.' *Communications of the ACM*, 31, 5, 514–530.

Hewitt, B., Gilbert, N., Jirotka, M., and Wilbur, S. (1990) *Theories of multi-party interaction*. Technical Report, Social and Computer Sciences Research Group, University of Surrey and Queen Mary and Westfield Colleges, University of London.

Hutchins, E. (1990) 'The technology of team navigation'. In Galegher, J., Kraut, R.E. and Edigo, C. (Ed.) *Intellectual Teamwork* (pp. 191–3220). Hillsdale N.J.: Lawrence Erlbaum Associates.

Jeffries, R., Miller, J. R., Wharton, C., and Uyeda, K. M. (1991) 'User interface evaluation in the real world: a comparison of four techniques.' In *ACM CHI '91* (pp. 119–124). New Orleans, LA: ACM, New York.

Jeffries, R., Turner, A. A., Polson, P. G., and Atwood, M. E. (1981) 'The processes involved in designing software.' In J. R. Anderson (Eds.), *Cognitive Skills and their Acquisition* Hillsdale, N.J.: Lawrence Erlbaum.

Kieras, D. E., and Polson, P. G. (1985) 'An approach to the formal analysis of user complexity.' *International Journal of Man-Machine Studies*, 22, 365–394.

Landauer, T. K. (1987) 'Relations between cognitive psychology and computer system design.' In J. M. Carroll (Eds.), *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction* (pp. 1–25). Cambridge, Ma.: MIT Press.

Maguire, M. C. (1990) 'A review of human factors guidelines and techniques for the design of graphical human-computer interfaces.' In J. Preece & L. Keller (Eds.), *Human-Computer Interaction* (pp. 161–184). Hemel Hempstead: Prentice Hall.

Malone, T. W. (1981, December 1981) 'What makes computer games fun?' *BYTE*, p. 258–277. Michie, D. and Johnston, R. (1984) *The Creative Computer: Machine Intelligence and Human*

*Knowledge*. Harmondsworth: Penguin.
Miller, G. A. (1956) 'The magical number seven, plus or minus two: some limits on our
capacity for processing information.' *Psychological Review*, 63, 81–97. Newman, W. M., and Wellner, P. (1992) 'A desk supporting

computer based interaction with

paper documents.' *Technical Report EPC-91-131*, Rank Xerox EuroPARC. Newell, A., and Simon, H. (1972) *Human Problem Solving*. Englewood Cliffs: N.J.: Prentice

Hall.
Norman, D. A. (1986) 'Cognitive Engineering.' In D. A. Norman & S. W. Draper (Eds.), *User*

*Centred System Design* Hillsdale, New Jersey: Lawrence Erlbaum.
Papert, S. (1980) *Mindstorms: Children, Computers and Powerful Ideas*. New York: Basic

Books.

Reece, J. (1993) *Cognitive Processes in the Development of Written Composition Skills: The Role of Planning, Dictation and Computer Tools*. Doctoral Thesis, La Trobe University, Melbourne, Australia.

Sharples, M. (1994) 'A study of breakdowns and repairs in a computer mediated communication system'. *Interacting With Computers,* 5(1), 61–77.

Shneiderman, B. (1992) *Designing the User Interface: Strategies for Effective Human- Computer Interaction*. Reading, Massachusetts: Addison-Wesley.

Spivey, J. M. (1988) *The Z Notation: A Reference Manual*. Hemel Hempstead: Prentice-Hall International.

Sproull, L., and Kiesler, S. (1991, September 1991) 'Computers, networks and work.' *Scientific American*, p. 84–91.

Tesler, L. G. (1991, September 1991) 'Netorked computing in the 1990s.' *Scientific American*, p. 54-61.

# Delhi Technological University

**Delhi Technological University** (**DTU**), formerly known as the **Delhi College of Engineering** (**DCE**) is a state university in Rohini, Delhi, India. It was established in 1941 as Delhi Polytechnic. In 1952, it started giving degrees after being affiliated with the University of Delhi. The institute has been under the Government of Delhi since 1963 and was affiliated with the University of Delhi from 1952 to 2009. In 2009, the college was given university status, thus changing its name to Delhi Technological University.

**Davin Braven**
**2K20/SE/09**

**20 DECEMBER 2022**