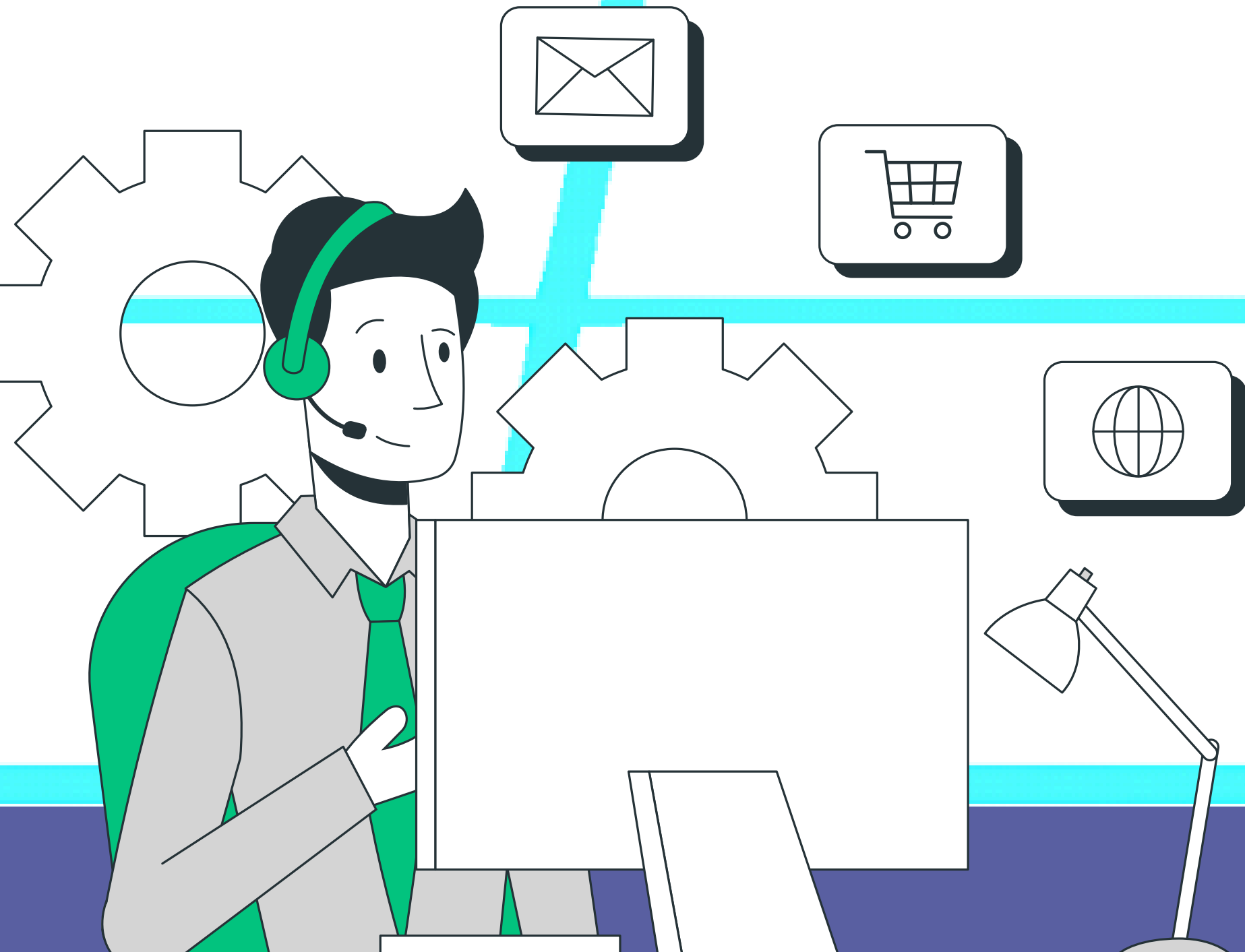


- K-Means Clustering algorithm
- Mean-shift algorithm
- DBSCAN Algorithm
- Principal Component Analysis
- Independent Component Analysis

By Angad's Group

# UNLOCKING CUSTOMER INSIGHTS: BEHAVIORAL SEGMENTATION WITH K- MEANS CLUSTERING






# What is Clustering ?

Clustering is the process of grouping data points into clusters based on their similarity. This technique is useful for identifying patterns and relationships in data without the need for labeled examples.

## What is K-Means Clustering ?

One method of data grouping is K-Means Clustering. The K-Means Clustering method divides the data into groups and can accept inputs of data without class labels. This method partitions the data into clusters/groups so that data that have the same characteristics are grouped into the same cluster and data that have different characteristics are grouped into other groups.



# Why Clustering?

**1. Unsupervised Learning:** No Labeled Data Required: Clustering doesn't require pre-labeled data, making it useful when you don't have labeled examples or want to explore the data's structure without predefined categories.

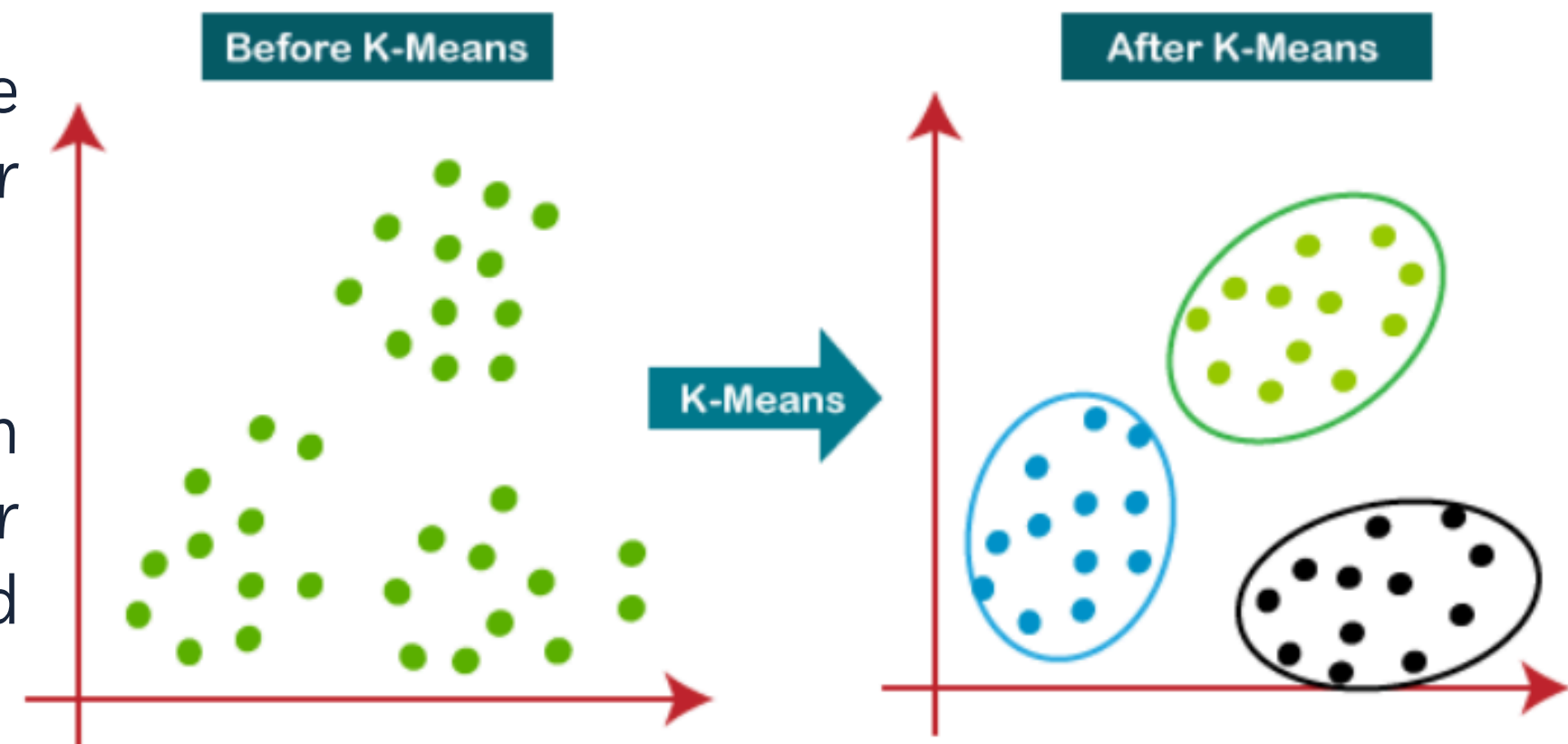
**2. Discovering Patterns:** Uncover Hidden Structures: It helps in discovering hidden patterns or natural groupings within the data, which might not be apparent through other methods.

**3. Data Reduction:** Simplify Complex Data: By grouping similar data points together, clustering can reduce the complexity of the data and make it easier to analyze or visualize.

**4. Preprocessing for Other Algorithms:** Feature Engineering: Clustering results can be used as features for supervised learning algorithms to improve performance.

**5. Segmentation and Personalization:**

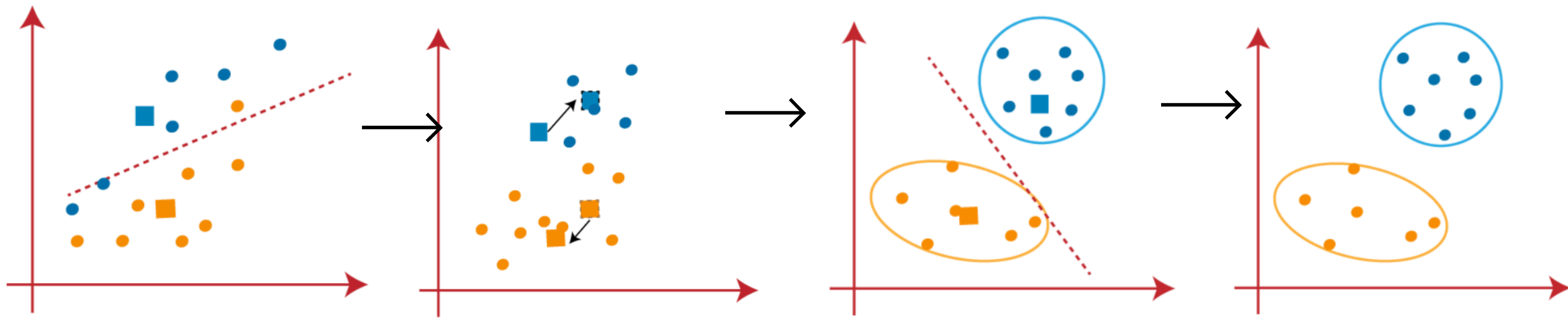
**Customer Segmentation:** In marketing, clustering can segment customers into groups with similar behaviors or preferences, allowing for targeted strategies and personalized recommendations.





# How K-Means Clustering Works ?

- Step 1: Let's choose the number  $k$  of clusters, i.e.,  $K=2$ , to segregate the dataset and put them into different respective clusters. We will choose some random 2 points which will act as centroids to form the cluster.
- Step 2: Now, we will assign each data point to a scatter plot based on its distance from the closest  $K$ -point or centroid.
- Step 3: Points closer to the blue centroid form the blue cluster; those closer to the yellow centroid form the yellow cluster.
- Step 4: Repeat the process by choosing a new centroid.
- Step 5: Reassign each data point to the nearest new centroid. Repeat until clusters stabilize using a median line.

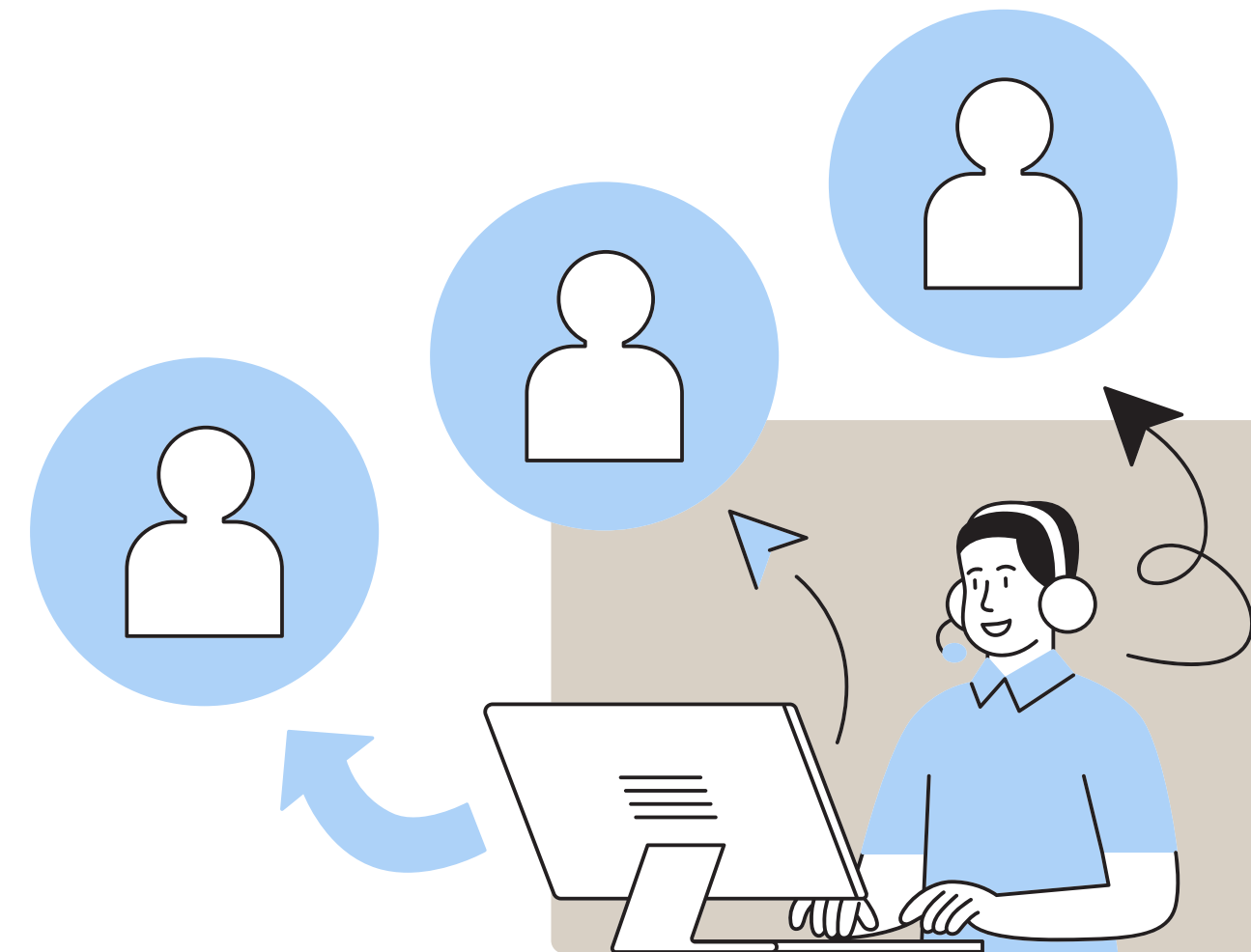


# Behavioral Segmentation Using K-Means Algorithm

Behavioral segmentation divides a market based on consumer behaviors, such as purchase patterns, product usage, and brand loyalty. It helps in understanding how different groups interact with products or services.

Behavioral segmentation is like dividing people into groups based on how they act or behave. Instead of looking at who they are (like age or gender), you look at what they do. For example:

- How often they buy things
- How much they spend
- What products they prefer



# Detailed Explanation

jupyter Behavioral Segmentation By Angad's Group (autosaved)

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Save Add Cut Copy Paste Undo Redo Run Stop Restart Code

```
In [ ]: # Loading the data from csv to a pandas data frame:
customer_data = pd.read_csv('/content/Mall_Customers.csv')
```

```
In [ ]: # Inspect the dataframe
customer_data.head()
```

```
Out[3]:
```

|   | 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                     |

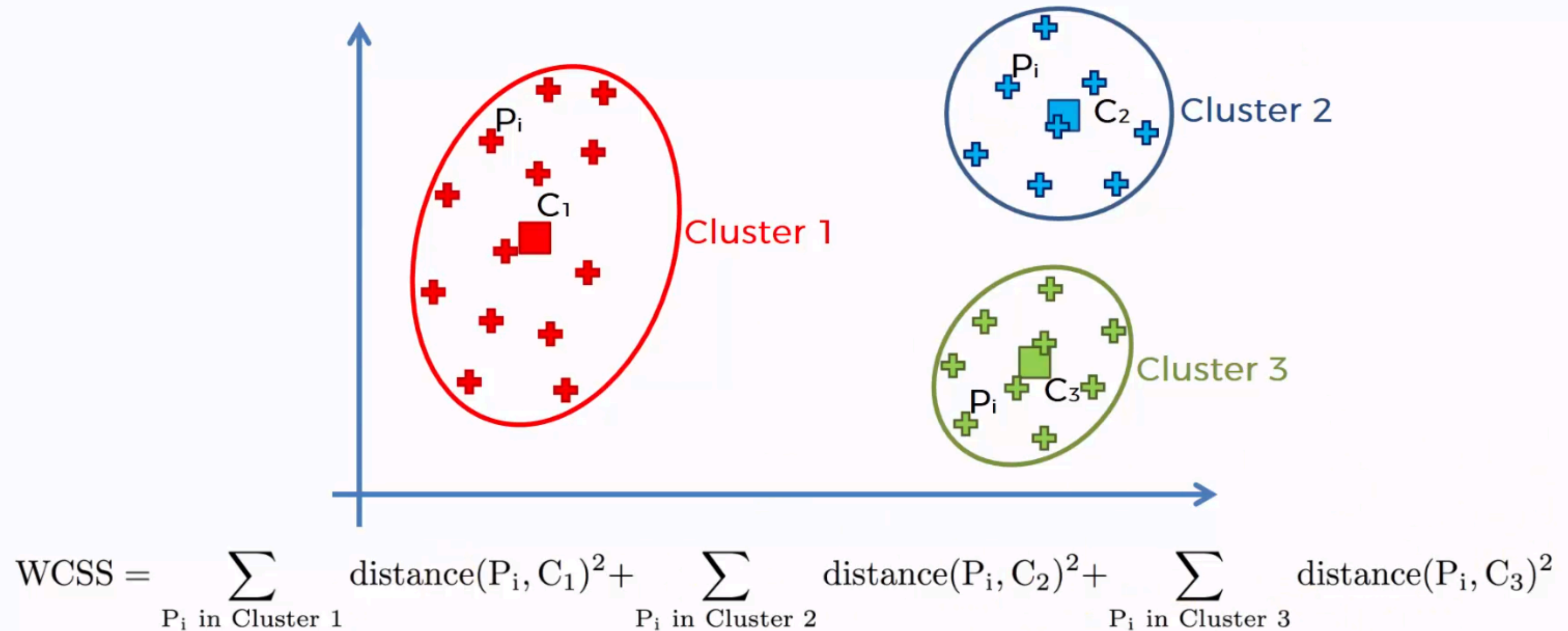
```
In [ ]: # Count Rows & Columns
# Result: (rows,cols)
customer_data.shape
```

```
Out[4]: (200, 5)
```

Active



To choose the best number of clusters, we use WCSS (Within-Cluster Sum of Squares). WCSS measures how spread out the data points are within each cluster. A lower WCSS means the clusters are tighter and more compact, while a higher WCSS means the clusters are more spread out. So, we aim for the number of clusters that gives us the smallest WCSS, indicating well-defined, compact clusters.



# Visualizing all the clusters!



**Cluster 4(Violet):** Don't have High annual income but high spending score (Very less such people)

**Cluster 5(Blue):** Don't have High annual income also low spending score (People with less income buy less -> Understandable)

Give Offers and better Discounts with some membership card.

**Cluster 1(Green):** Have High annual income still low spending score (Are conscious of what they buy -> which has obviously made them richer)

***Get Business insights and place products that this cadre of people tend to buy, which is generally branded and high quality product.***

**Cluster 3(Yellow)** Have High annual income along with high spending score (They are comfortable in spending)

**Cluster 2(Red):** Have average annual income and spending score (Most cutomers fall in this category, which shows a upper-middle class people who can't be manipulated easily)

# K-Means Clustering offers several advantages:

- **Simplicity:** Easy to understand and implement.
  - **Speed:** Efficient for large datasets and scales well with the number of data points.
  - **Scalability:** Handles large datasets effectively.
  - **Flexibility:** Works with different types of data and distances.
  - **Interpretability:** Easy to interpret the results, as each cluster is represented by its centroid.
  - **Partitioning:** Clearly partitions data into distinct groups.
- 

## Challenges and Limitations

- **Choosing the Number of Clusters (k):** It's not always clear how many clusters to use. Too few or too many can give poor results.
- **Initialization Sensitivity:** Different starting points can lead to different clusters.
- **Scalability:** K-Means can struggle with very large datasets or high-dimensional data (data with many features).

# CONCLUSION

K-Means Clustering is a useful method for grouping customers based on their behaviors. It helps businesses understand different customer groups without needing labeled data. By finding patterns and simplifying data, companies can create targeted marketing strategies and personalized recommendations. K-Means Clustering is easy to use, fast, and works well with large datasets. However, it's important to choose the right number of groups and be aware that different starting points can affect the results. Overall, K-Means Clustering helps businesses gain valuable insights into their customers and make better decisions.





# REFERENCES

- [https://github.com/thisisankit27/Insightify/blob/master/Customer\\_Segmentation\\_using\\_K\\_Means\\_Clustering.ipynb](https://github.com/thisisankit27/Insightify/blob/master/Customer_Segmentation_using_K_Means_Clustering.ipynb)
- <https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/>
- <https://www.ibm.com/topics/clustering#:~:text=Clustering%20is%20an%20unsupervised%20machine%20learning%20algorithm%20that,groups%20or%20clusters%20based%20on%20similarities%20or%20patterns.>
- <https://towardsdatascience.com/unsupervised-learning-and-data-clustering-eeecb78b422a?gi=168fc50866c0>
- [https://www.canva.com/design/DAGLw6plVLg/xWBx8tDmmHA57kb8PNiuQw/editutm\\_content=DAGLw6plVLg&utm\\_campaign=designshare&utm\\_medium=link2&utm\\_source=sharebutton](https://www.canva.com/design/DAGLw6plVLg/xWBx8tDmmHA57kb8PNiuQw/editutm_content=DAGLw6plVLg&utm_campaign=designshare&utm_medium=link2&utm_source=sharebutton)





Thank you!

