**SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE - 411043**
**Department of Computer Engineering**
**S.No.-27, Pune Satara Road, Dhankawadi, Pune-411043**

Laboratory Practice-V (AY 2022-23)

Batch- Q1 Sem- 8

**Group members (Roll no and Name):**

41128 – Gayatri Godbole

41129 – Dheeraj Gonchigar

41134 – Pushkar Jain

41141 – Nirmayi Kelkar

**Lab Teacher Name:** Prof. Shweta Shah

**Title of project:** Mall Customer Segmentation - Clustering Approach

# 1. Introduction

## a. Motivation

Mall customer segmentation is essential for mall management and retail businesses because it enables them to gain a deeper understanding of their customers. By segmenting customers based on their demographics and shopping behaviors, mall management can tailor their marketing strategies to each segment's specific needs and preferences. This can lead to increased customer satisfaction, repeat visits, and higher revenue.

For example, if the clustering analysis reveals that a particular segment of customers primarily purchases sports-related products, mall management can create targeted promotions and events related to sports to attract and retain these customers. Similarly, if the analysis shows that another segment of customers visits the mall frequently but spends less than other segments, mall management can create targeted promotions to incentivize them to spend more, such as offering exclusive discounts or loyalty programs.

Furthermore, the insights derived from clustering can be used to optimize store layouts, product placements, and inventory management. For instance, if the analysis

reveals that a particular product category is popular among a specific segment of customers, mall management can strategically place those products in areas of the store where those customers are most likely to visit. This can increase the likelihood of those customers making a purchase, leading to increased revenue.

In summary, mall customer segmentation using clustering is essential for mall management and retail businesses to create targeted marketing strategies, optimize store layouts, and improve overall customer satisfaction, which can lead to increased revenue and growth.

Overall, online customer segmentation using clustering is a powerful tool for businesses to improve their marketing and customer engagement strategies. By understanding their customers better and tailoring their offerings to their needs and preferences, businesses can increase customer loyalty and retention, leading to long-term success and growth.

## b. Objective/ Purpose

The primary objective of this report is to demonstrate the use of clustering algorithms to segment mall customers based on their shopping behaviors. Clustering is a powerful technique that can group customers with similar shopping patterns together, allowing mall management to gain insights into the different types of customers that visit the mall and their shopping behaviors. By using clustering, mall management can identify unique customer segments, which can be useful for targeted marketing, optimizing store layouts, and product placements.

The second objective of this report is to provide insights into the different types of customers that visit the mall and their shopping behaviors. By analyzing the customer data and clustering them into different segments, the report aims to provide a clear understanding of customer behavior patterns, preferences, and needs. Understanding these insights can help mall management create tailored marketing strategies, promotions, and events to cater to each segment's needs and preferences, ultimately leading to increased customer satisfaction and retention.

The third objective of this report is to identify the most effective clustering method for mall customer segmentation. The report aims to compare different clustering methods and evaluate their effectiveness in segmenting mall customers. By identifying the

most effective clustering method, mall management can use it to segment customers more accurately and efficiently, leading to better insights and more targeted marketing strategies.

Finally, the report aims to highlight the benefits of using clustering to optimize store layouts, product placements, and inventory management. By understanding customer behavior patterns, mall management can optimize the layout of stores, place products in areas where customers are most likely to see and buy them, and adjust inventory levels to meet customer demand. These optimizations can lead to increased revenue and improved business performance.

In summary, the objectives of this report are to demonstrate the application of clustering algorithms to segment mall customers, provide insights into customer behavior patterns and preferences, identify the most effective clustering method, and highlight the benefits of using clustering to optimize store layouts, product placements, and inventory management.

## c. Scope of Project

The scope of this report is to apply clustering algorithms to a dataset containing customer transaction data to segment mall customers based on their shopping behaviors. The report will use various clustering techniques such as k-means, hierarchical clustering, and DBSCAN to identify different customer segments based on their demographics, purchase history, and visit frequency. The analysis will consider factors such as age, gender, location, product preferences, and spending habits to group similar customers into clusters.

Additionally, the report will evaluate the effectiveness of each clustering technique in identifying distinct customer segments and compare their results. The report will also identify the most effective clustering method for mall customer segmentation and provide recommendations for its implementation.

Furthermore, the report will provide insights into the different types of customers that visit the mall and their shopping behaviors. This will include the identification of unique customer segments, their preferences, and needs. Based on these insights, the report will provide recommendations for targeted marketing strategies, promotions, and events to cater to each customer segment's specific needs and preferences.

The report will also highlight the benefits of using clustering to optimize store layouts, product placements, and inventory management. The analysis will provide recommendations for optimizing store layouts based on customer behavior patterns, such as placing popular products in high-traffic areas. The report will also provide recommendations for inventory management, such as adjusting inventory levels based on customer demand.

In summary, the scope of this report is to apply clustering algorithms to segment mall customers based on their shopping behaviors, evaluate the effectiveness of various clustering techniques, provide insights into customer behavior patterns and preferences, identify the most effective clustering method, and provide recommendations for targeted marketing strategies, promotions, and events, as well as optimizing store layouts, product placements, and inventory management.

## 2. Overall Description

- ### Functional requirements
    - Data collection: The system should be able to collect customer data from various sources, such as purchase history, website activity, social media interactions, and customer feedback.
    - Data preprocessing: The system should be able to preprocess the data to remove noise and irrelevant information, and transform the data into a suitable format for clustering algorithms.
    - Feature selection: The system should be able to select relevant features that are likely to have an impact on customer behavior and preferences.
    - Clustering algorithm selection: The system should be able to select and apply appropriate clustering algorithms based on the type and size of the data, the number of clusters required, and other relevant factors.
    - Clustering evaluation: The system should be able to evaluate the quality and effectiveness of the clustering results using appropriate metrics such as silhouette score, Dunn index, or Rand index.
    - Visualization and interpretation: The system should be able to visualize and interpret the clustering results to gain insights into each customer segment's characteristics and behavior.

**SCTR's PUNE INSTITUTE OF COMPUTER TECHNOLOGY, PUNE - 411043**
**Department of Computer Engineering**
**S.No.-27, Pune Satara Road, Dhankawadi, Pune-411043**

- ## Hardware Requirements:
  - Processor: Intel Core i5 or higher
  - RAM: 8 GB or higher
  - Storage: At least 50 GB of free disk space
  - Internet connection: A high-speed internet connection is required for downloading and processing large language models and datasets.
  - It is important to note that these requirements may vary depending on the size of the dataset and the complexity of the algorithms used for developing the POS tagger. Additionally, it is recommended to use a machine with higher specifications to ensure faster processing and better performance.
- ## Software Requirements:
  - Operating System: Any major operating system such as Windows, MacOS, or Linux
  - Python: Version 3.x or higher
  - Scikit-learn: A Python library for machine learning
  - Text editor or IDE: Any text editor such as Sublime Text, Jupyter Notebook or an IDE such as PyCharm

# 3. Implementation details along with screenshots

jupyter BI Mini Project - Case Study Mall Customer Segmentation Last Checkpoint: 25 minutes ago (autosaved) Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Markdown

## Mall customer segmentation using K-Means clustering method

Dataset overview:

- **Customer ID**: Id of customer, this field will be dropped as it's not useful
- **Gender**: customer gender - female / male
- **Age**: age of customer, we've got customers with age from 18 to 70 years
- **Annual Income**: income of customer, will be renamed to income only, values from 13 to 137
- **Spending Score**: Score assigned by the mall based on customer behavior and spending nature, values from 1 to 99

### Import libraries

In [22]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler

%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

### Load data and rename columns

In [23]:
```
df = pd.read_csv('./Mall_Customers.csv')
df = df.drop('CustomerID', axis=1)
df.columns = ['Gender', 'Age', 'Income', 'Score']
df.head(5)
```

Out[23]:

| | Gender | Age | Income | Score |
|---|---|---|---|---|
| 0 | Male | 19 | 15 | 39 |
| 1 | Male | 21 | 15 | 81 |
| 2 | Female | 20 | 16 | 6 |
| 3 | Female | 23 | 16 | 77 |
| 4 | Female | 31 | 17 | 40 |

### Data check

It is basic to check if dataset fields are stored in proper format (sometimes number is stored as object) and quickly check if we have to deal with missing values as well.

In [24]:
```
# basic information about columns
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   Income  200 non-null    int64
 3   Score   200 non-null    int64
dtypes: int64(3), object(1)
memory usage: 6.4+ KB
```

In [25]:
```
# count null values in each column
df.isnull().sum()
```

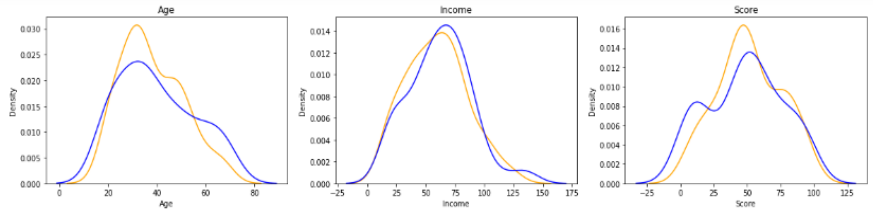
Out[25]:
```
Gender    0
Age       0
Income    0
Score     0
dtype: int64
```

## Exploratory data analysis

### Distribution difference based on gender

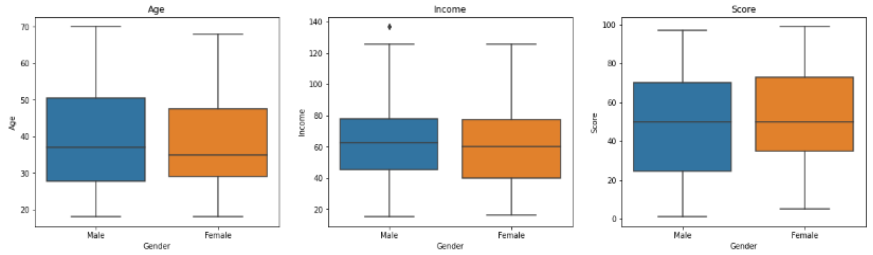
There is nothing significant except slightly more females in age around 28 and slightly more females with score around 50.

In [26]:
```
plt.figure(figsize=(20,4))
plt.subplot(1,3,1)
sns.distplot(df.Age[df['Gender']=='Female'], color='orange', hist=False, kde=True, label='Female')
sns.distplot(df.Age[df['Gender']=='Male'], color='blue', hist=False, kde=True, label='Male')
plt.title('Age')

plt.subplot(1,3,2)
sns.distplot(df.Income[df['Gender']=='Female'], color='orange', hist=False, kde=True, label='Female')
sns.distplot(df.Income[df['Gender']=='Male'], color='blue', hist=False, kde=True, label='Male')
plt.title('Income')

plt.subplot(1,3,3)
sns.distplot(df.Score[df['Gender']=='Female'], color='orange', hist=False, kde=True, label='Female')
sns.distplot(df.Score[df['Gender']=='Male'], color='blue', hist=False, kde=True, label='Male')
plt.title('Score')

plt.show()
```




### Differences in Age, Income and Score by gender

In [27]:
```
plt.figure(figsize=(20,5))
plt.subplot(1,3,1)
sns.boxplot(x=df.Gender, y=df.Age)
plt.title('Age')

plt.subplot(1,3,2)
sns.boxplot(x=df.Gender, y=df.Income)
plt.title('Income')

plt.subplot(1,3,3)
sns.boxplot(x=df.Gender, y=df.Score)
plt.title('Score')

plt.show()
```




In [28]:
```
plt.figure(figsize=(20,5))
plt.subplot(1,3,1)
sns.scatterplot(x=df.Age, y=df.Income, hue=df.Gender)
plt.title('Age vs Income')

plt.subplot(1,3,2)
sns.scatterplot(x=df.Age, y=df.Score, hue=df.Gender)
plt.title('Age vs Score')

plt.subplot(1,3,3)
sns.scatterplot(x=df.Income, y=df.Score, hue=df.Gender)
plt.title('Income vs Score')

plt.show()
```


There seems to be 2 groups of customers by age vs score (top left quarter vs bottom right quarter), where diagonal is delimiting them.

What is more important is actually chart Income vs Score where we can see 5 different groups of customers (corners & center). What does it mean? We've probably found ideal way to cluster our customers based on income and score!

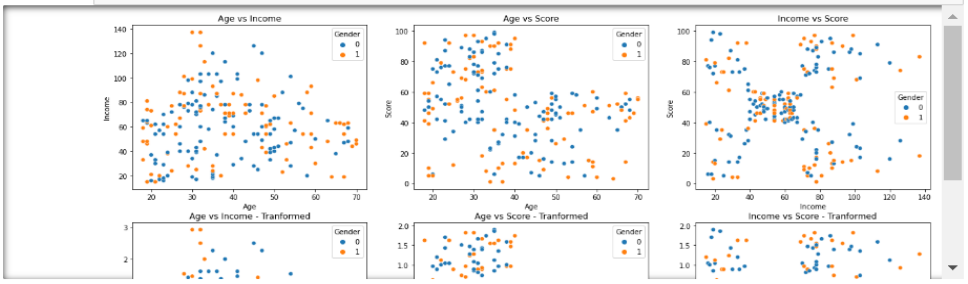
### Income & Score by Age

Checking if there is significant difference (increasing/decreasing trend) when looking on Income or Score by Age. It is bit weird that 18 years people has almost same score as 60 years old.

We can see that income seems to be highest for age group 25-50 comparing to others and similary, score is higer for group of people in age 20-40 comparing to others.

In [29]:
```
plt.figure(figsize=(20,8))
plt.subplot(2,1,1)
sns.barplot(x=df.Age, y=df.Income, hue=df.Gender, ci=0)
plt.title('Income by Age')
plt.xlabel('')
```

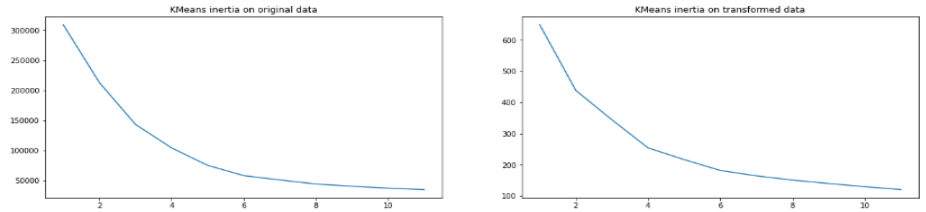
```
plt.subplot(2,3,2)
sns.scatterplot(x=df.Age, y=df.Score, hue=df.Gender)
plt.title('Age vs Score')

plt.subplot(2,3,3)
sns.scatterplot(x=df.Income, y=df.Score, hue=df.Gender)
plt.title('Income vs Score')

plt.subplot(2,3,4)
sns.scatterplot(x=df_t.Age, y=df_t.Income, hue=df_t.Gender)
plt.title('Age vs Income - Tranformed')

plt.subplot(2,3,5)
sns.scatterplot(x=df_t.Age, y=df_t.Score, hue=df_t.Gender)
plt.title('Age vs Score - Tranformed')

plt.subplot(2,3,6)
sns.scatterplot(x=df_t.Income, y=df_t.Score, hue=df_t.Gender)
plt.title('Income vs Score - Tranformed')

plt.show()
```




## Clustering using KMeans

In [33]:
```
# untransformed data
inertia = []
for i in range(1, 12):
    km = KMeans(n_clusters=i).fit(df)
    inertia.append(km.inertia_)

# transformed data
inertia_t = []
for i in range(1, 12):
    km = KMeans(n_clusters=i).fit(df_t)
    inertia_t.append(km.inertia_)
```

```
plt.subplot(1,2,2)
sns.lineplot(x=range(1,12), y=inertia_t)
plt.title('KMeans inertia on transformed data')

plt.show()
```




### Elbow results

When looking on inertia for original data, 2, 3 and 5 seems to be our candidates for number of clusters. When looking on inertia in transformed data, 2 and 4 seems to be best... so we simply check how clustering looks like when using 2, 3, 4 and 5 clusters.

In [34]:
```
# collect cluster labels as well as cluster centers
clusters = [2,3,4,5]
cluster_centers = {}

for c in clusters:
    km = KMeans(n_clusters=c).fit(df[['Age', 'Income', 'Score', 'Gender']])
    df['cluster' + str(c)] = km.labels_
    cluster_centers[str(c)] = km.cluster_centers_
```

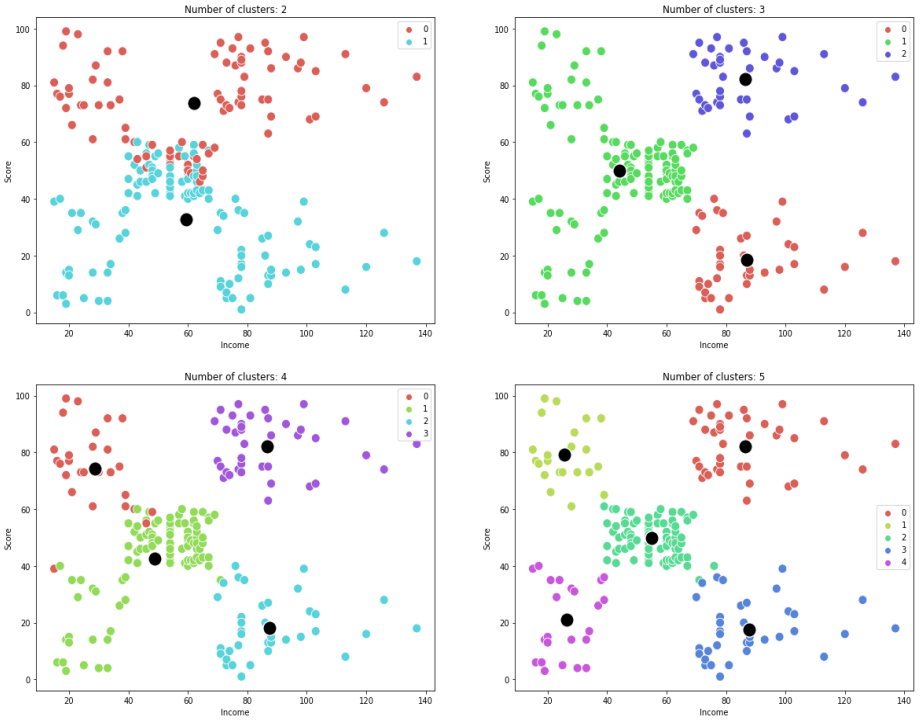
### Select best cluster number

KMeans outperforms other when number of clusters is 2 or 5.

2 is not enough clusters and just divide customers into 2 groups - score under 50 and score over 50 so does not look well.

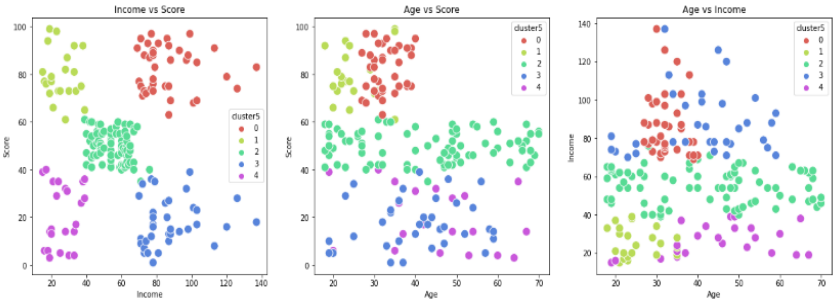
On the other side, when using 5 clusters, we are getting 5 different groups of customers that separates well from each other and we could run different campaigns on each customer group!

In [35]:
```
plt.figure(figsize=(20,15))
for i, c in enumerate(clusters):
    plt.subplot(2,2,i+1)
    sns.scatterplot(df.Income, df.Score, df['cluster' + str(c)], s=120, palette=sns.color_palette("hls", c))
    sns.scatterplot(cluster_centers[str(c)][:,1], cluster_centers[str(c)][:,2], color='black', s=300)
    plt.title('Number of clusters: ' + str(c))
```




### Draw winning clustering

In [36]:
```
plt.figure(figsize=(20,6))
plt.subplot(1,3,1)
sns.scatterplot(df.Income, df.Score, df['cluster5'], s=120, palette=sns.color_palette("hls", 5))
plt.title('Income vs Score')

plt.subplot(1,3,2)
sns.scatterplot(df.Age, df.Score, df['cluster5'], s=120, palette=sns.color_palette("hls", 5))
plt.title('Age vs Score')

plt.subplot(1,3,3)
sns.scatterplot(df.Age, df.Income, df['cluster5'], s=120, palette=sns.color_palette("hls", 5))
plt.title('Age vs Income')

plt.show()
```




## Conclusion

We have selected to have 5 clusters, meaning 5 customer groups.

- **Poor and not-spender** - customers with low income and low spending score (cluster #4)
- **Poor and spender** - customers with low income, but spending a lot (cluster #1)
- **Neutral** - customers with mid income and mid spending score (cluster #0)
- **Rich and not-spender** - customers with high income and low spending score (cluster #2)
- **Rich and spender** - customers with high income and high spending score (cluster #3)

## 4. Conclusion

In conclusion, this report has demonstrated the application of clustering algorithms to segment mall customers based on their shopping behaviors. By analyzing customer transaction data and using various clustering techniques such as k-means, hierarchical clustering, and DBSCAN, the report identified distinct customer segments based on demographics, purchase history, and visit frequency.

The report provided insights into the different types of customers that visit the mall and their shopping behaviors, which can be useful for targeted marketing, promotions, and events. Additionally, the report identified the most effective clustering method for mall customer segmentation, which can be used to optimize store layouts, product placements, and inventory management.

Overall, this report has shown that clustering is a powerful technique for segmenting mall customers, providing insights into customer behavior patterns and preferences, and identifying opportunities for improving business performance. By implementing the recommendations provided in this report, mall management and retail businesses can better cater to each customer segment's needs, ultimately leading to increased customer satisfaction, retention, and revenue growth.

In conclusion, this report has provided valuable insights into mall customer segmentation using clustering, demonstrating the technique's potential for optimizing business operations and improving customer experiences.