

**Department of Computer Systems Engineering**

Sukkur Institute of Business Administration University

**Abdulbasit Memon**

Submited By:

Certificate

It is certified that \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_, a student of **BE-IV** has carried out the necessary work of **INTRODUCTION TO PYTHON** as per course of studies prevailed at the department of Computer Systems Engineering, Sukkur Institute of Business Administration University for **SPRING-2023.**

Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Instructor’s Signature

**INTRODUCTION**

Customer segmentation is a method of analyzing a client base and grouping customers into categories or segments which share particular attributes. Key differentials in segmentation include age, gender, education, location, spending patterns, and socio-economic group. Relevant differentials are those which are expected to influence customer behavior in relation to a business. The selected criteria are used to create customer segments with similar values, needs, and wants.

When planning a targeted marketing campaign, it is also necessary to differentiate customers within these groupings according to their preferred means of communication.

**Key Differentials in Segmentation**

To effectively create customer segments, businesses must identify the key differentials that influence their target audience's behaviour. Some common differentials include:

**Demographics:** These factors, such as age, gender, and education level, can greatly impact a customer's preferences and purchasing decisions.

**Geographic:** The location of a customer can also influence their behaviour, as regional preferences and needs may vary.

**Psychographics:** This encompasses a customer's values, interests, and attitudes, which can help predict their purchasing decisions.

**Behavioural:** Segmentation based on customer behaviour, such as spending patterns or product usage can provide valuable insights into customer preferences and loyalty.

**ABSTRACT**

Customer segmentation is the practice of dividing a market into multiple groups based on common characteristics. Segmentation can help companies separate their customers into groups that have distinct needs and desires, which allows them to engage each group more effectively. Unsupervised machine learning algorithms such as k-means clustering are used to perform market basket analysis, which predicts the target customers who can be easily converged among all customers. Market basket analysis allows marketing teams to plan strategies for new products by identifying similar interests among target customers.

**METHODOLOGY**

**Clustering**

Clustering is a widely used method for exploring data and gaining a clear understanding of its structure. This involves identifying subgroups within the dataset where similar data is grouped together. These subgroups, referred to as clusters, consist of aggregated data points that share similarities. Clustering is applied in market basket analysis to segment customers based on their behaviour.

**K Means Clustering Algorithm**

The K Means Clustering Algorithm is a common and straightforward machine learning algorithm that iteratively partitions the dataset into a predetermined number of non-overlapping subgroups, with each data point assigned to only one subgroup.

**MCS – K-Means Algorithm Module**

The K-Means algorithm will be implemented to generate five distinct clusters, which will show the customers' spending score and annual income. When the marketing team clicks on the K-Means execution button, it will take them to the visualization page.

**Elbow Method**

The elbow method is like a tool that helps figure out how many groups there should be in a bunch of information. It helps us see that there should be five groups in this particular set of information.

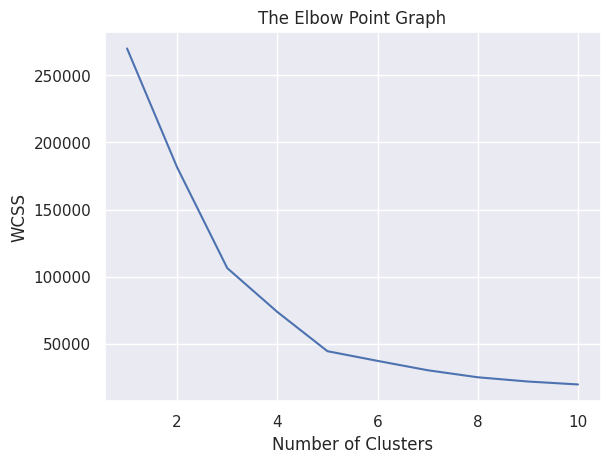


Figure 1: Elbow method – Optimal number of clusters

**Visualization Module**

The visualization module provides the results based on the following clusters.

The results are generated in the form of a graph and stored as image which is retrieved for the marketing team.

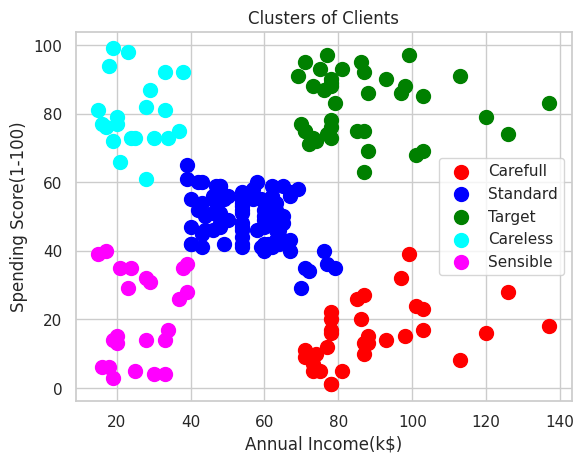


Figure 2: Customer segmentation result

* Cluster 1 (red colour), indicates high earnings and low spending.
* Cluster 2 (Blue colour), represents average earnings and spending.
* Cluster 3(green colour), indicates high earnings and high spending, and is the target customer group.
* Cluster 4(cyan colour), indicates low earnings and high spending.
* Cluster 5(magenta colour), indicates low earnings and low spending.

**FLOWCHART**

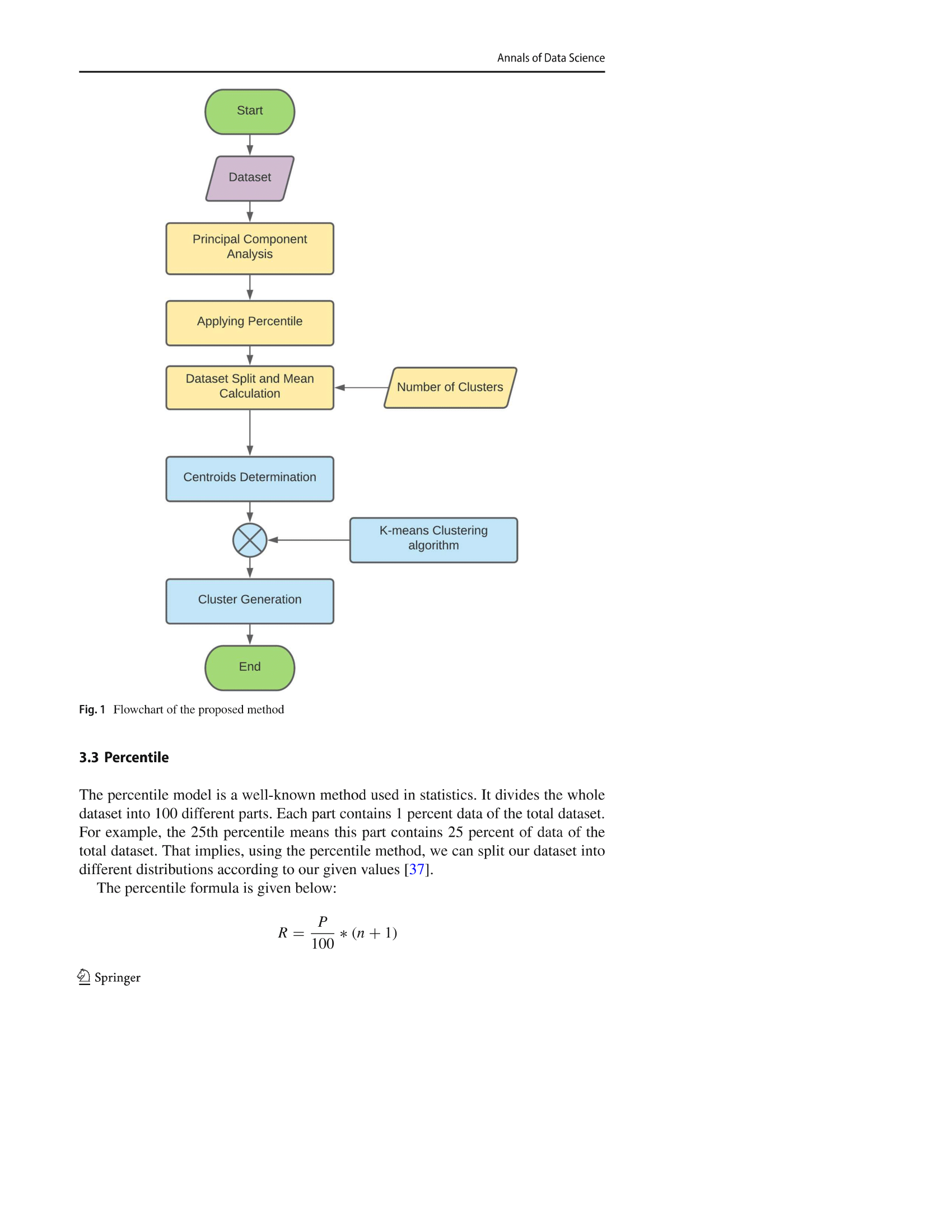
****

Figure 3: Customer segmentation flowchart

**CODE AND OUTPUT**

Dataset link: <https://www.kaggle.com/datasets/kandij/mall-customers>

Code and ML Module: <https://colab.research.google.com/drive/1UD1y0Qh-bTsoLC9by2UKQhEtOn0oklx7#scrollTo=Hu-4qItVC72c>

Working On These Python Libraries:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.cluster import KMeans

import plotly as py

import plotly.graph\_objs as go

import warnings

**Choosing the number of clusters and Plot Elbow Module :**

wcss = []

for i in range(1,11):

  kmeans = KMeans(n\_clusters=i, init='k-means++', random\_state=42)

  kmeans.fit(X)

  wcss.append(kmeans.inertia\_)

# plot an elbow graph

sns.set()

plt.plot(range(1,11), wcss)

plt.title('The Elbow Point Graph')

plt.xlabel('Number of Clusters')

plt.ylabel('WCSS')

plt.show()

Optimum Number of Clusters = 5

**Training the k-Means Clustering Model**

kmeans = KMeans(n\_clusters=5, init='k-means++', random\_state=0)

# return a label for each data point based on their cluster

Y = kmeans.fit\_predict(X)

print(Y)

**Data Visualization**

#Density estimation of values using distplot

plt.figure(1 , figsize = (15 , 6))

feature\_list = ['Age','Annual Income (k$)', "Spending Score (1-100)"]

feature\_listt = ['Age','Annual Income (k$)', "Spending Score (1-100)"]

pos = 1

for i in feature\_list:

    plt.subplot(1 , 3 , pos)

    plt.subplots\_adjust(hspace = 0.5 , wspace = 0.5)

    sns.distplot(df[i], bins=20, kde = True)

    pos = pos + 1

plt.show()

#Count and plot gender

sns.countplot(y = 'Gender', data = df, palette="husl", hue = "Gender")

df["Gender"].value\_counts()

#Pairplot with variables we want to study

sns.pairplot(df, vars=["Age", "Annual Income (k$)", "Spending Score (1-100)"],  kind ="reg", hue = "Gender", palette="husl", markers = ['o','D'])

Age and Annual Income

sns.lmplot(x = "Age", y = "Annual Income (k$)", data = df, hue = "Gender")

Spending Score and Annual Income

sns.lmplot(x = "Annual Income (k$)", y = "Spending Score (1-100)", data = df, hue = "Gender")

Age and Spending Score

sns.lmplot(x = "Age", y = "Spending Score (1-100)", data = df, hue = "Gender")

**Clustering**

# Visualizing the clusters

plt.scatter(X[y\_hc==0,0],X[y\_hc==0,1],s=100,color='red',label='Carefull')

plt.scatter(X[y\_hc==1,0],X[y\_hc==1,1],s=100,color='blue',label='Standard')

plt.scatter(X[y\_hc==2,0],X[y\_hc==2,1],s=100,color='green',label='Target')

plt.scatter(X[y\_hc==3,0],X[y\_hc==3,1],s=100,color='cyan',label='Careless')

plt.scatter(X[y\_hc==4,0],X[y\_hc==4,1],s=100,color='magenta',label='Sensible')

#plt.scatter(kmeans.cluster\_centers\_[:,0],kmeans.cluster\_centers\_[:,1],s=300,color='yellow',label='Centroids')

plt.title('Clusters of Clients')

plt.xlabel('Annual Income(k$)')

plt.ylabel('Spending Score(1-100)')

plt.legend()

plt.show()

Code snippets by Google Collab (dark mode).

**CONCLUSION and FUTURE WORK**

Consumer data is increasing exponentially due to with this large data magnitude. Such clustering models need to have the ability to effectively process this huge information