Abstract

The zeitgeist of modern era is innovation, where everyone is embroiled into competition to be better than others. Today's business run on the basis of such innovation having ability to enthral the customers with the products, but with such a large raft of products leave the customers confounded, what to buy and what to not and also the companies are nonplussed about what section of customers to target to sell their products. This is where machine learning comes into play, various algorithms are applied for unravelling the hidden patterns in the data for better decision making for the future. This elude concept of which segment to target is made unequivocal by applying segmentation. The process of segmenting the customers with similar behaviours into the same segment and with different patterns into different segments is called customer segmentation. In this paper K-means clustering is been implemented to segment the customers. A python program has been developed and the program is been trained by applying standard scalar onto a dataset having 5 features of two hundred training sample. By applying 5 segments of cluster have been formed.

TABLE OF CONTENTS

INTRODUCTION

In the contemporary day and age, the importance of treating customers as the principal asset of an organization is increasing in value. Organizations are rapidly investing in developing strategies for better customer acquisition, maintenance and development. The concept of business intelligence has a crucial role to play in making it possible for organizations to use technical expertise for acquiring better customer insight for outreach programs. In this scenario, the concept of CRM (Customer Relationship Management) garners much attention since it is a comprehensive process of acquiring and retaining customers, using business intelligence, to maximize the customer value for a business enterprise. One of the two most important objectives of CRM is customer development through customer insight. This objective of CRM entails the usage of analytical approach in order to correctly assess customer information and analysis of the value of customers for better customer insight. Keeping up with the changing times, organizations are modifying their business flow models by employing systems engineering as well as change management and designing information technology (IT) solutions that aid them in acquiring new customers, help retain the present customer base and boost the customer's lifelong value. due to the diverse range of products and services available in the market as well as the intense competition among organizations, customer relationship management has come to play a significant role in the identification and analysis of a company's best customers and the adoption of best marketing strategies to achieve and sustain competitive advantage.

One of the most useful techniques in business analytics for the analysis of consumer behaviour and categorization is customer segmentation. By using clustering techniques, customers with similar means, end and behaviour are grouped together into homogeneous clusters. Customer Segmentation helps organizations in identifying or revealing distinct groups of customers who think and function differently and follow varied approaches in their spending and purchasing habits.

Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behaviour, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. This study aims to explore the avenues of using customer segmentation, as a

business intelligence tool within the CRM framework as well as the use of clustering techniques for helping organizations redeem a clearer picture of the valuable customer base. The concepts of customer relationship management, customer segmentation as a core function of CRM as well as the approach of segmenting customers using clustering techniques are discussed.

The available clustering models for business analysis in the context of customer segmentation, the advantages and disadvantages of the two main models chosen for our study- K-Means, as well as the possibility of developing a hybrid model which can outperform the individual models is surveyed.

EXISTING METHOD

Customer Segmentation is the subdivision of a market into discrete customer groups that share similar characteristics. Customer Segmentation can be a powerful means to identify unsatisfied customer needs. Using the above data companies can then outperform the competition by developing uniquely appealing products and services.

The most common ways in which businesses segment their customer base are:

1. **Demographic information**, such as gender, age, familial and marital status, income, education and occupation.
2. **Geographical information**, which differs depending on the scope of the company. For localized business, this info might pertain to specific towns or countries. For larger companies, it might mean a customer's city, state, or even country of residence.
3. **Psychographics**, such as social class, lifestyle, and personality traits.
4. **Behavioural data**, such as spending and consumption habits, product/service usage, and desired benefits.
5. **Decision-making status**, especially useful for B2B companies, knowing a customer's job title can make the difference between barking up the wrong tree and making a sale quickly. But decision-making status also works for B2C companies. If you're trying to sell children's products, for example, you'll need to target those with the expendable income.
6. **Personal interactions with customers**, Support team can help us designate your power users, advocates, and even your difficult tickets. These groups are perfect test segments to start your experiments.

Advantages of Customer Segmentation

1. Determine appropriate product pricing.
2. Develop customized marketing campaigns.
3. Design an optimal distribution strategy.
4. Choose specific product features for deployment.
5. Prioritize new product development efforts.

The Challenge

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. You want to understand the customers like who are the target customers so that the sense can be given to marketing team and plan the strategy accordingly.

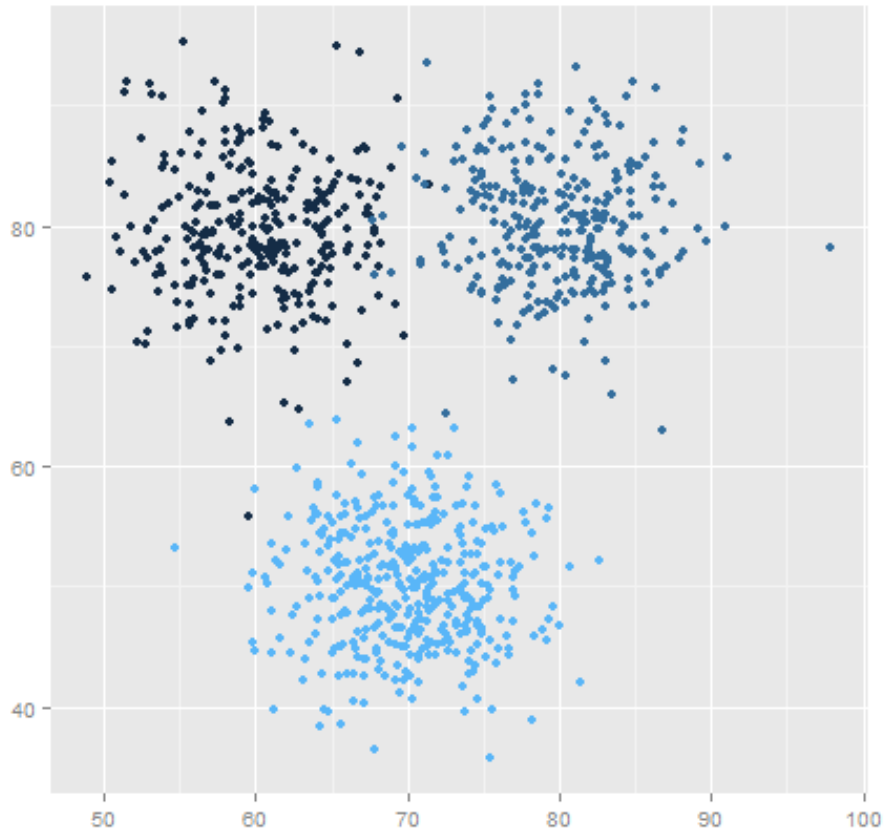
PROPOSED METHOD

Clustering for segmentation purposes: Clustering techniques reveal internally homogeneous and externally heterogeneous groups. Customers vary in terms of behaviour, needs, wants and characteristics and the main goal of clustering techniques is to identify different customer types and segment the customer base into clusters of similar profiles so that the process of target marketing can be executed more efficiently. Here we use k-means clustering for customer segmentation.

K Means Clustering Algorithm:

1. Specify number of clusters K.
2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points for the centroids without replacement.
3. Keep iterating until there is no change to the centroids. i.e. assignment of data points to clusters isn't changing.

(we should use Standard Scalar onto the dataset before apply K Means)



K Means Clustering where K=3

Challenges with K-Means Clustering Algorithm:

The main challenge with this algorithm is to find out the optimal number of clusters(k-value).

Solution:

To overcome the challenge, we need to plot Within Cluster Sum of Squares (WCSS) against the number of clusters (K Value) to figure out the optimal number of clusters value. WCSS measures sum of distances of observations from their cluster centroids which is given by the below formula.

$$WCSS = \sum_{i \in n} (X_i - Y_i)^2$$

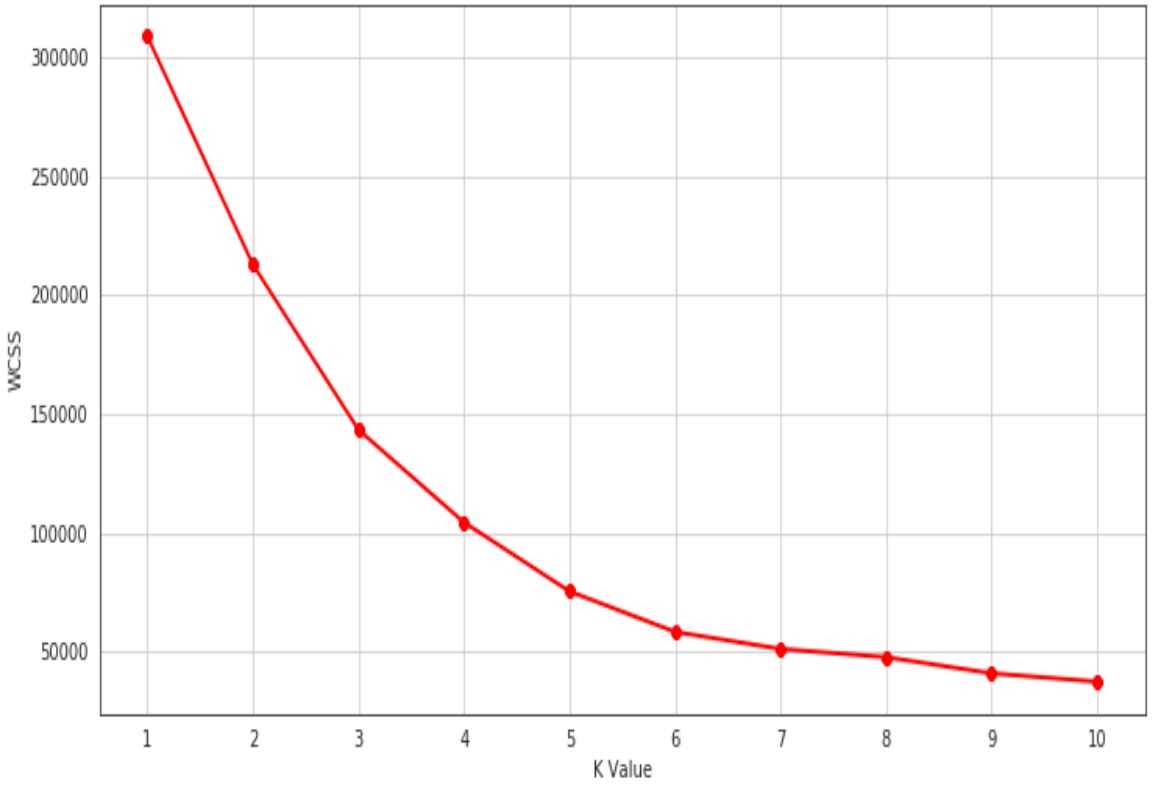
where Yi is centroid for observation Xi. The main goal is to maximize number of clusters and in limiting case each data point becomes its own cluster centroid.

Elbow Method:

Calculate the Within Cluster Sum of Squared Errors (WSS) for different values of k, and choose the k for which WSS first starts to diminish. In the plot of WSS-versus k, this is visible as an elbow.

The steps can be summarized in the below steps:

1. Compute K-Means clustering for different values of K by varying K from 1 to 10 clusters.
2. For each K, calculate the total within-cluster sum of square (WCSS).
3. Plot the curve of WCSS vs the number of clusters K.
4. The location of a bend (knee) in the plot is generally considered as an indicator of the appropriate number of clusters.



The optimal K value is found to be 5 using the elbow method.

Methodology

A common customer segmentation targeting and positioning process developed is given below:

Developing Quantitative Segmentation Criteria tailored to local market needs		Developing Quantitative Segmentation Criteria (look for global guidance)		Complete and Validate Segmentation	
Step 1	**Step 2**	**Step 3**	**Step 4**	**Step 5**	**Step 6**
Define the full potential of the market	**Develop quantitative factors for segmentation**	**Consider potential criteria for segmentation based on customer behaviour, values & attitudes**	**Select a customer segmentation based on 'qualitative' factors**	**Plot key customers from Step 2 onto the segmentation map from Step 4**	**Validate Segmentation**
1. How do you define your market? • Number of customers/end users? 2. How many potential end users or customers are there in the defined market? Consider all sources • Nationally though industry sources official statistics • Regionally through sales rep visits 3. Define and quantify all the key customers who may benefit from our product offering? • Are we missing any customers (i.e. those traditionally served by our competitions) ?	1. For the whole market, group customers into the appropriate number of groups, based on quantitative factors that define the potential value for us. 2. Potential ways to group customers with similar characteristics could include: • Industry • Customer potential • Demographic (e.g. age or occupation) • Geographic (e.g. regional or city sized)	1. Evaluate a number of different ways (1-3) of segmenting the market based on quantitative factors (e.g. customers' behavioural values or attitudes) • Some qualitative factors that could be used to segment your customers include: • Psychographic (e.g. personality, ego driven, or service driven) • Behavioural or attitude (benefits, prescription habits or attitude or product (e.g. scientific driven) 2. Assess the company strategy. 3. Use a cross functional team (e.g. sales and marketing) to 'brainstorm' some alternative approaches.	1. Select the 'bestway' to group customers based on qualitative criteria developed in Step 3. 2. Represent these qualitative factors graphically by developing an 'empty' segmentation map (e.g. you may have two criteria that can be used in a 2×2 matrix)	1. Using the qualitative approach developed in Step 4, plot your quantified customers onto the empty segmentation map.	1. Consider supporting evidence for your segmentation approach, including market research data. • Does the market really work according to the way you have segmented? • Are the most successful competitions taking a different segmentation approach and can we learn from them? • What market data could help you confirm your current segmentation? 2. Are customers effectively grouped into homogeneous segments with the same needs and wants so that marketing programmes can be tailored for each group?

	Targeting		Positioning	Planning	
	Step 7	**Step 8**	**Step 9**	**Step 10**	**Step 11**
	Confirm the market potential of each segment, prioritize target in target segments	**Which product/ brand is being used by each customer the most attractive?**	**Decide on product positioning for each of the targeted segments**	**Develop for each targeted segment a marketing mix action plain to support the brand positioning**	**Develop a detailed Key Account Management Action Plan with Sales Management**
What questions do we need to answer and what do we need to do?	1. Evaluate the attractiveness of each segment • What is the purchasing size and power of each segment? • How easy will it be to convert customers in each segment to Serono product offering? • How responsive is the segment to marketing actions? • What would be the cost of serving the segment? • How are the competitors targeting these customers and what is the hurdle to convert? 2. Decide which segments you will target. Have you clearly explained, justified and tested your logic?	1. What competitors are present in each segment? • Which products and what is their market share? 2. How are our competitors positioning their brand in each segment? • What brand values do they promote? • What is the current perceptual map for our competitors in this segment? • How do we currently positioning itself (if at all)?	1. What are the Value Components for our brands and how will you match or position these to the wants and needs of this customer segment? • How have we defined the Core Value Components (CVCs - the intrinsic elements of value common to all regions and countries) of its brands and how can they be positioned in your segments? • How will you customize the elements of value for our brand (e.g. through different services) by defining Augmented Value Components (AVCs - the customized elements of value that are expected to tip the scales in favour of our brand versus the alternatives)? 2. Decide on a clear product positioning for our brand in each segment to gain a competitive advantage. • Establish objectives that can be monitored in step 11.	1. Develop a marketing mix action plan for each of our segments to support our desired positioning? • What marketing objectives, strategies and actions are required for each segment? • Will these actions change the perceptions of the customers in your segment to enhance sales? • How will you review whether the needs and wants of the customers in this segment are being satisfied and how their perceptions to our products have changed? 2. Measure progress against objectives established in step 9.	1. Marketing should support sales to develop a Key Account Management Action Plan that is consistent with the segmentation, targeting and positioning outline in the marketing plan. This helps the sales team to : • Target customer accounts following segmentation guidelines • Prioritize sales activities to maximize sales (e.g. call rate activity by segment and selling tactics)

Developing customer segmentation using quantitative factors (steps 1-2):

The first two steps of the segmentation methodology involve the gathering of quantitative information (unique to your market) to define the full po-tential of your market and to develop a way of segmenting the market, based on these quantitative factors.

Example: For a Jewellery mall – Families with MHI of Rs 25,000, credit card, car.

Developing customer segmentation using qualitative factors (steps 3-4):

Steps 3 and 4 of the segmentation process consider (and select) additional ways of segmenting customers based on qualitative factors such as cus-tomer behaviour, values or attitudes.

As these qualitative factors may be common across all or many of your global markets, you should first look for guidance from your own company business strategy before you de-velop your segmentation. This part of the process should develop a frame-work or a 'segmentation map' based on qualitative factors that quantified customers can be plotted onto in step 5 of the process.

Example: For a Jewellery mall – Families with modern society values and acceptance of modern retail format for shopping.

Plot key customers onto the segmentation map and validate segmentation (steps 5-6):

The next stage of the segmentation process is to complete and validate the chosen qualitative segmentation.

The targeting (of segments) and positioning (within segments) process:

Once segmentation is complete, the next consideration is which segment(s) to target and how to position your product offering into the target seg-ment. Steps 8 to 1 0 of the process consider the positioning of the brand(s) by segment as well as developing marketing mix by segment.

Example: For a Jewellery mall – Families with MHI of Rs 15000-25000 staying within 10 km radius of the mall with belief in modern retail format. The mall will be positioned as one stop point for all kind of metal, artificial and modern jewellery needs with best branded and unbranded stores.

Implementation

Code:

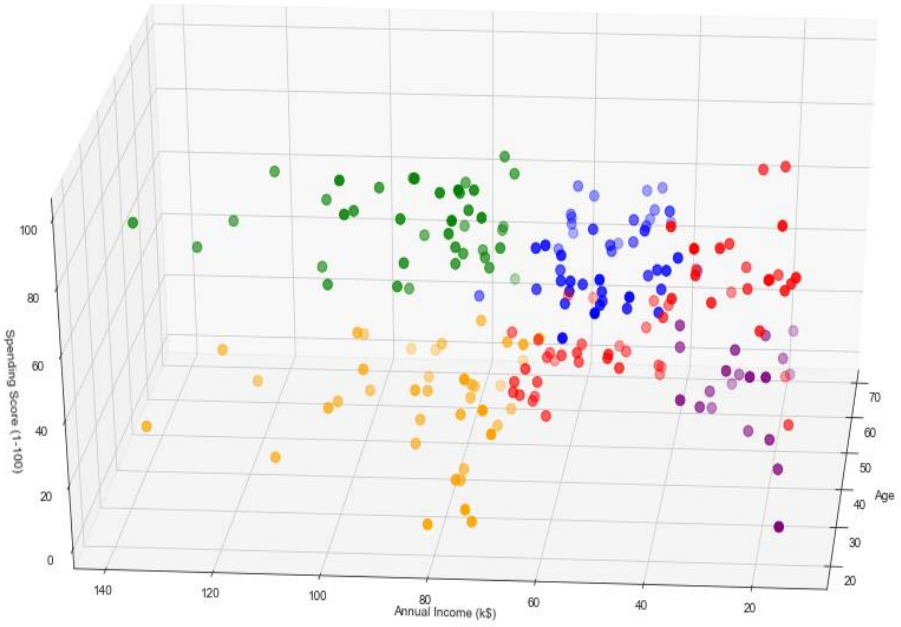
```
import pandas as pd
import numpy as np
df=pd.read_csv("Mall_Customers.csv") //read the dataset
df.head()
df.info()
df.describe()
df.columns
df['Gender'].value_counts()
from sklearn.preprocessing import LabelEncoder
le=LabelEncoder()
le.fit(df['Gender'])
df['Gender']=le.transform(df['Gender'])
df.head()
from sklearn.preprocessing import StandardScaler
st=StandardScaler()
df.columns
X=df[['Age','Annual Income (k$)','Spending Score (1-100)']]
X=st.fit_transform(X)
X[0:5]
X.shape
X.shape[0]
from sklearn.cluster import KMeans
from scipy.spatial.distance import cdist
inner,dist=[],[]
```

```
c=range(1,10)
for i in c:
    k=KMeans(n_clusters=i,n_init=12,random_state=4)
    k.fit(X)
    inner.append(k.inertia_)
    dist.append(sum(np.min(cdist(X,k.cluster_centers_,'euclidean'),
    axis=1))/  X.shape[0])
import matplotlib.pyplot as plt
len(inner)
plt.plot(c,inner)
plt.xlabel('number of cluster')
plt.ylabel('innertia')
plt.show()
plt.plot(c,dist)
plt.xlabel('number of cluster')
plt.ylabel('distance')
plt.show()
K=KMeans(n_clusters=5)
K.fit(X)
labels=K.predict(X)
labels[0:5]
labels
df['label']=labels
df['label'].value_counts()
import seaborn as sns
sns.set_style('darkgrid')
```

```
sns.barplot(['Female','Male'],df['Gender'].value_counts().values)
plt.ylabel('Number of customer')
plt.show()
age18_30=df.Age[(df.Age>=18) & (df.Age<=30)].value_counts().sum()
age31_40=df.Age[(df.Age>=31) & (df.Age<=40)].value_counts().sum()
age41_50=df.Age[(df.Age>=41) & (df.Age<=50)].value_counts().sum()
age51_60=df.Age[(df.Age>=51) & (df.Age<=60)].value_counts().sum()
age60=df.Age[(df.Age>=61)].value_counts().sum()
xaxis=['age18_30','age31_40','age41_50','age51_60','age60+']
yaxis=[age18_30,age31_40,age41_50,age51_60,age60]
sns.barplot(x=xaxis,y=yaxis)
plt.ylabel('Number of customer')
plt.show()
plt.figure(figsize=(15,6))
plt.subplot(1,2,1)
sns.boxplot(df['Annual Income (k$)'],orient='v',color='red')
plt.subplot(1,2,2)
sns.boxplot(df['Spending Score (1-100)'],orient='v')
plt.show()
ai1_30=df['Annual Income (k$)'][(df['Annual Income (k$)']>=1) &
(df['Annual Income (k$)']<=30)].value_counts().sum()
ai31_60=df['Annual Income (k$)'][(df['Annual Income (k$)']>=31) &
(df['Annual Income (k$)']<=60)].value_counts().sum()
ai61_90=df['Annual Income (k$)'][(df['Annual Income (k$)']>=61) &
(df['Annual Income (k$)']<=90)].value_counts().sum()
ai91_120=df['Annual Income (k$)'][(df['Annual Income (k$)']>=91) &
(df['Annual Income (k$)']<=120)].value_counts().sum()
```

```
ai121_150=df['Annual Income (k$)'][(df['Annual Income (k$)']>=121) & (df['Annual Income (k$)']<=150)].value_counts().sum()

xaxis=["$ 0 - 30,000", "$ 30,001 - 60,000", "$ 60,001 - 90,000", "$ 90,001 - 120,000", "$ 120,001 - 150,000"]

yaxis=[ai1_30,ai31_60,ai61_90,ai91_120,ai121_150]

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

sns.barplot(x=xaxis,y=yaxis)

plt.ylabel('Number of customers')

plt.xlabel('Annual Income')

plt.show()

ss1_20=df['Spending Score (1-100)'][(df['Spending Score (1-100)']>=1) & (df['Spending Score (1-100)']<=20)].value_counts().sum()

ss21_40=df['Spending Score (1-100)'][(df['Spending Score (1-100)']>=21) & (df['Spending Score (1-100)']<=40)].value_counts().sum()

ss41_60=df['Spending Score (1-100)'][(df['Spending Score (1-100)']>=41) & (df['Spending Score (1-100)']<=60)].value_counts().sum()

ss61_80=df['Spending Score (1-100)'][(df['Spending Score (1-100)']>=61) & (df['Spending Score (1-100)']<=80)].value_counts().sum()

ss81_100=df['Spending Score (1-100)'][(df['Spending Score (1-100)']>=81) & (df['Spending Score (1-100)']<=100)].value_counts().sum()

xaxis=["1-20", "21-40", "41-60", "61-80", "81-100"]

yaxis=[ss1_20,ss21_40,ss41_60,ss61_80,ss81_100]

plt.figure(figsize=(16,5))

sns.barplot(xaxis,yaxis)

plt.xlabel('Spending Score')

plt.ylabel('Number of customers')

plt.show()

sns.set_style('white')
```

```
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', s=60)

ax.scatter(df.Age[df.label == 1], df["Annual Income (k$)"][df.label == 1],
df["Spending Score (1-100)"][df.label == 1], c='red', s=60)

ax.scatter(df.Age[df.label == 2], df["Annual Income (k$)"][df.label == 2],
df["Spending Score (1-100)"][df.label == 2], c='green', s=60)

ax.scatter(df.Age[df.label == 3], df["Annual Income (k$)"][df.label == 3],
df["Spending Score (1-100)"][df.label == 3], c='orange', s=60)

ax.scatter(df.Age[df.label == 4], df["Annual Income (k$)"][df.label == 4],
df["Spending Score (1-100)"][df.label == 4], c='purple', s=60)

ax.view_init(30, 185)

plt.xlabel("Age")

plt.ylabel("Annual Income (k$)")

ax.set_zlabel('Spending Score (1-100)')

plt.show()
```



Conclusions

K means clustering is one of the most popular clustering algorithms and usually the first thing practitioners apply when solving clustering tasks to get an idea of the structure of the dataset. The goal of K means is to group data points into distinct non-overlapping subgroups. One of the major application of K means clustering is segmentation of customers to get a better understanding of them which in turn could be used to increase the revenue of the company.