CUSTOMER SEGMENTATION USING K – MEANS CLUSTERING

1. INTRODUCTION

Customer Segmentation is one the most important applications of unsupervised learning. Using clustering techniques, companies can identify the several segments of customers allowing them to target the potential user base. In this machine learning project, we will make use of K-means clustering which is the essential algorithm for clustering unlabelled dataset for clustering customers after the exploratory analysis is done.

Customer Segmentation:

Customer Segmentation is the process of division of customer base into several groups of individuals that share a similarity in different ways that are relevant to marketing such as gender, age, interests, and miscellaneous spending habits.

Companies that deploy customer segmentation are under the notion that every customer has different requirements and require a specific marketing effort to address them appropriately. Companies aim to gain a deeper approach of the customer they are targeting. Therefore, their aim has to be specific and should be tailored to address the requirements of each and every individual customer. Furthermore, through the data collected, companies can gain a deeper understanding of customer preferences as well as the requirements for discovering valuable segments that would reap them maximum profit. This way, they can strategize their marketing techniques more efficiently and minimize the possibility of risk to their investment.

The technique of customer segmentation is dependent on several key differentiators that divide customers into groups to be targeted. Data related to demographics, geography, economic status as well as behavioural patterns play a crucial role in determining the company direction towards addressing the various segments.

2. DATASET

This dataset is obtained from Kaggle website (https://www.kaggle.com/shwetabh123/mall-customers) which contains 200 rows of customer data with their age, annual income, gender and spending score is available. The sample dataset is given in the diagram below.

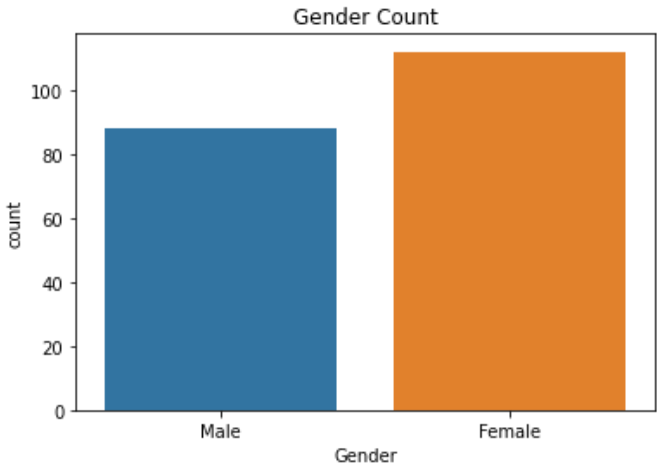
	CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

3. DATA EXPLORATION

The dataset has 200 data entries with zero missing values. 112 entries are female and 88 entries are male with age group from 18 to 70 years with mean age of 38.9

Descriptive

	Age	Annual_Income_(k$)	Spending_Score	Gender
N	200	200	200	200
Missing	0	0	0	0
Mean	38.9	60.6	50.2	1.56
Median	36.0	61.5	50.0	2.00
Minimum	18	15	1	1
Maximum	70	137	99	2



Checking for influence of categorical variable gender on spending score. From the below table, Mean of all the variables are almost same across gender.

Descriptives

	Genre	Annual_Income_(k$)	Spending_Score	Age
N	Female	112	112	112
	Male	88	88	88
Missing	Female	0	0	0
	Male	0	0	0
Mean	Female	59.3	51.5	38.1
	Male	62.2	48.5	39.8
Median	Female	60.0	50.0	35.0
	Male	62.5	50.0	37.0
Minimum	Female	16	5	18
	Male	15	1	18
Maximum	Female	126	99	68
	Male	137	97	70

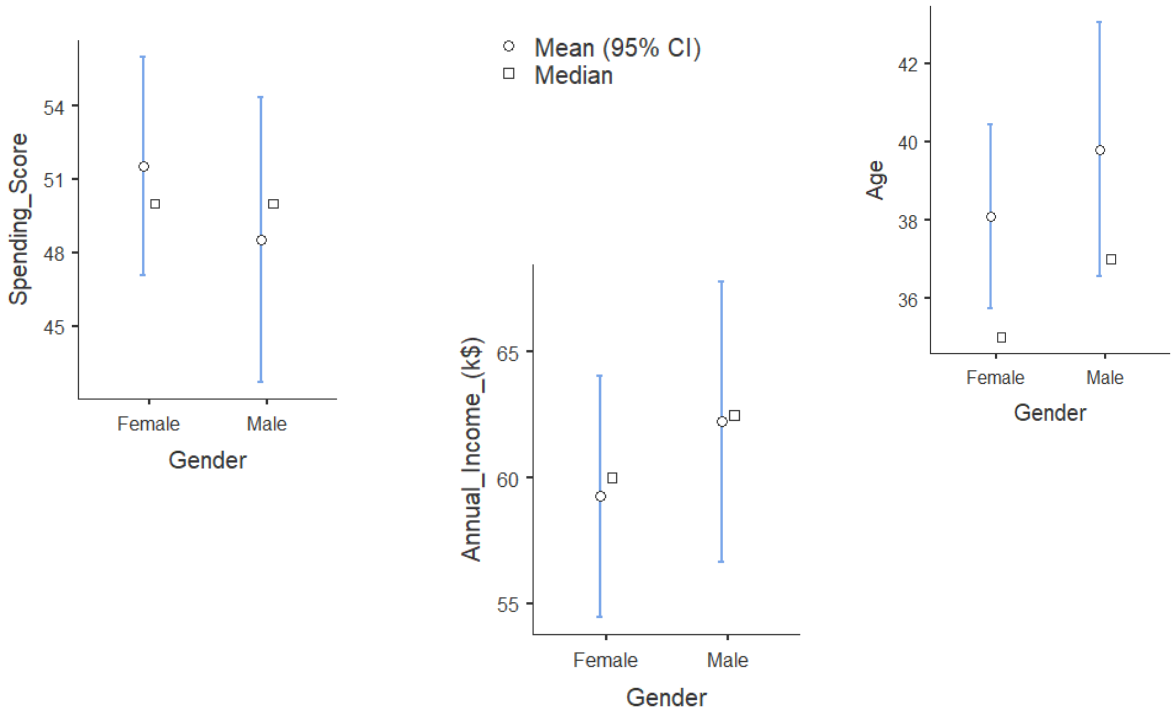
Independent T – Test is done to find the effect of gender in the rest of variables in the table and the result is given below.

Independent Samples T-Test

		Statistic	df	p
Spending_Score	Student's t	0.819	198	0.414
Age	Student's t	-0.858 [a]	198	0.392
Annual_Income_(k$)	Student's t	-0.795	198	0.428

From the above Figure,

P statistic value for all the variables are greater than 0.05 (for 95% significance). Thus, the gender doesn't affect the values of the other variable significantly.



The dependency of Spending score variable with other variables are to be determined by using linear regression. Below shows the results of linear regression.

Model Coefficients - Spending_Score

Predictor	Estimate	SE	t	p
Intercept	73.34785	6.5530	11.1931	< .001
Annual_Income_(k$)	0.00575	0.0662	0.0868	0.931
Age	-0.60479	0.1245	-4.8591	< .001

Model Fit Measures

					Overall Model Test			
Model	R	R^2	Adjusted R^2	RMSE	F	df1	df2	p
1	0.327	0.107	0.0980	24.3	11.8	2	197	< .001

Assumption Checks

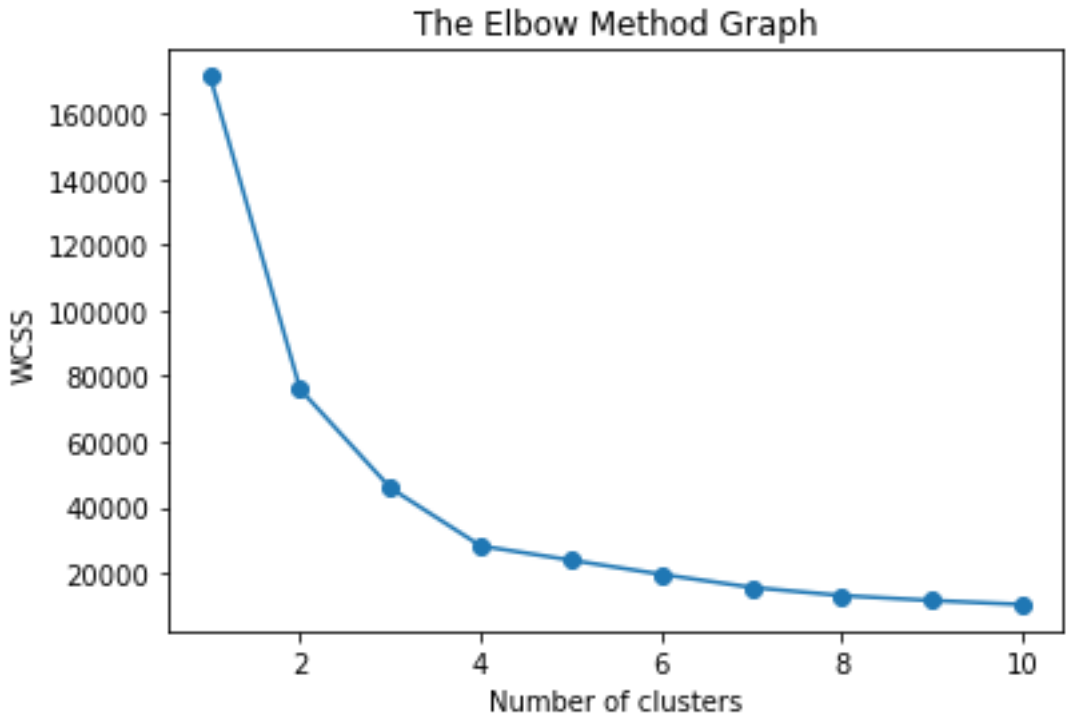
Collinearity Statistics

	VIF	Tolerance
Annual_Income_(k$)	1.00	1.000
Age	1.00	1.000

From this analysis, it is evident that only age is significantly affecting the dependent variable i.e, Spending score.

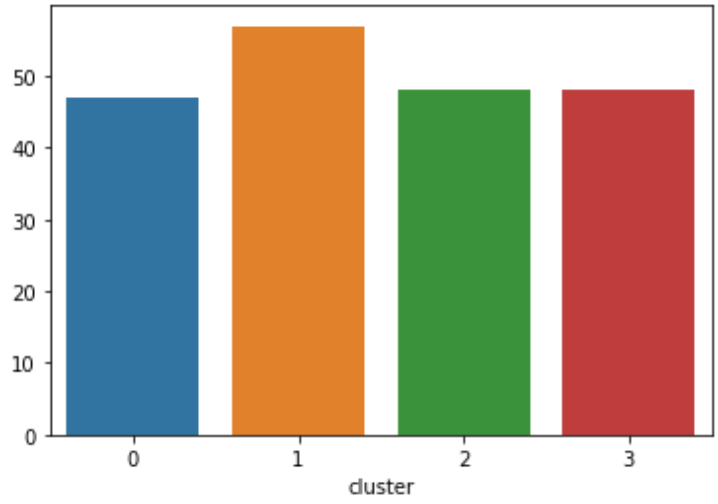
4. K MEANS CLUSTERING

As from the regression analysis, it is found that age is the only factor affecting the spending score. We try to cluster using age and spending scores. Below is the figure of Elbow method graph for 10 clusters.



From the above chart, Number of cluster is chosen as 4 as there is relatively less fall in WCSS error value from 4 to 5.

K- Means clustering is done and the customers are clustered. Below is the results after K-Means Clustering



```
cl0.describe()
```

	CustomerID	index	Age	Annual_Income_(k$)	Spending_Score	cluster
count	47.000000	47.000000	47.000000	47.000000	47.000000	47.0
mean	83.255319	82.255319	27.617021	53.276596	49.148936	0.0
std	40.604919	40.604919	7.167418	17.356495	9.136593	0.0
min	1.000000	0.000000	18.000000	15.000000	29.000000	0.0
25%	52.500000	51.500000	21.000000	42.500000	41.500000	0.0
50%	88.000000	87.000000	27.000000	57.000000	50.000000	0.0
75%	112.500000	111.500000	32.500000	63.500000	55.500000	0.0
max	185.000000	184.000000	41.000000	99.000000	66.000000	0.0

```
cl1.describe()
```

	CustomerID	index	Age	Annual_Income_(k$)	Spending_Score	cluster
count	57.000000	57.000000	57.000000	57.000000	57.000000	57.0
mean	115.157895	114.157895	30.175439	66.070175	82.350877	1.0
std	69.515617	69.515617	5.535995	32.405830	8.913255	0.0
min	2.000000	1.000000	18.000000	15.000000	68.000000	1.0
25%	34.000000	33.000000	27.000000	33.000000	75.000000	1.0
50%	142.000000	141.000000	30.000000	75.000000	81.000000	1.0
75%	172.000000	171.000000	35.000000	87.000000	90.000000	1.0
max	200.000000	199.000000	40.000000	137.000000	99.000000	1.0

```
cl2.describe()
```

	CustomerID	index	Age	Annual_Income_(k$)	Spending_Score	cluster
count	48.000000	48.000000	48.000000	48.000000	48.000000	48.0
mean	115.333333	114.333333	43.291667	66.937500	15.020833	2.0
std	71.062853	71.062853	11.761745	33.346923	8.753090	0.0
• min	3.000000	2.000000	19.000000	16.000000	1.000000	2.0
25%	32.500000	31.500000	36.000000	32.250000	7.750000	2.0
50%	147.000000	146.000000	44.000000	77.500000	14.000000	2.0
75%	173.500000	172.500000	52.000000	87.250000	20.500000	2.0
max	199.000000	198.000000	67.000000	137.000000	32.000000	2.0

```
cl3.describe()
```

	CustomerID	index	Age	Annual_Income_(k$)	Spending_Score	cluster
count	48.000000	48.000000	48.000000	48.000000	48.000000	48.0
mean	85.145833	84.145833	55.708333	54.770833	48.229167	3.0
std	27.188851	27.188851	8.557585	9.852615	6.922795	0.0
min	41.000000	40.000000	43.000000	38.000000	35.000000	3.0
25%	63.750000	62.750000	48.750000	47.000000	43.000000	3.0
50%	82.000000	81.000000	53.500000	54.000000	48.000000	3.0
75%	105.500000	104.500000	65.000000	62.250000	53.500000	3.0
max	161.000000	160.000000	70.000000	79.000000	60.000000	3.0