International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 09 | Sep 2021 www.irjet.net p-ISSN: 2395-0072

THE CUSTOMER DATA ANALYSIS USING SEGMENTATION WITH SPECIAL REFERENCE MALL

Dr. C K GOMATHY[1], Ms. D. ABIRAMI[2], Mr. J. BALAMURUGAN[3], Ms. DONDAPATI TEJASWI[4]

ABSTRACT:-

To develop the business and marketing in an efficient and successful manner we need to analyze the customer data. For that process the grouping of customers into small segments of individuals who share the common interest &characteristics' are known as customer segmentation. To analyze more efficiently we need to segment the customers based upon various types of segmentation.1.Demographic segmentation: Segmenting the market based upon Age, Gender, Income, Financial status, Education, Family status and so on. 2. Geographic segmentation: as the name itself suggested that this kind of segmentation is done based upon the physical location of person.3Behavioral segmentation, this kind of segmentation is based on the behavioral data of the customers like Purchasing habits, spending habits, Brand interaction are used in this type.

Keywords: Target customers, Spending Score, Unsupervised learning, clusters, visual data representation.

I. INTRODUCTION

This project is based on real-world data. Data is one of the important features of every organization because it helps business leaders to make decisions based on facts, statistical numbers and trends. So in this project we are taking mall customers data which means collecting the data from customers who are visiting the malls. Based on this data applying the segmentation process on it. This segmentation enables marketers to create targeted marketing messages for a specific group of customers which increases the chances of the person buying a product. It allows them to create and use specific communication channels to communicate with different segments to attract them. A simple example would be that the companies try to attract the younger generation through social media posts and older generation with maybe radio advertising. This helps the companies in establishing better customer relationships and their overall performance as an organization.

II. EXISTING SYSTEM

In the existing system, for business all the data accumulation and handling are done manually. For knowing details about a customer, there is no proper means of collecting the data in the business. Even though if a business wants to cluster the data about their customers, then they need to manually go through all the customer purchase bills. By doing this the business can't derive any useful or meaningful data that can help the business. This type of data contains zero insights about the customer and makes difficult to extract useful data.

III. PROPOSED SYSTEM

Learning methodologies are a great tool for analyzing customer dataand finding insights and patterns. Artificially intelligent models are powerful tools for decision-makers. They can precisely identify customer segments, which is much harder to do manually or with conventional analytical methods. There are many machine learning algorithms, each suitable for a specific typeof problem. One very common machine learning algorithm that's suitable for customer segmentation problems is the k-means clustering algorithm. The Advantages are: Whole process will be automated, so human error will be avoided

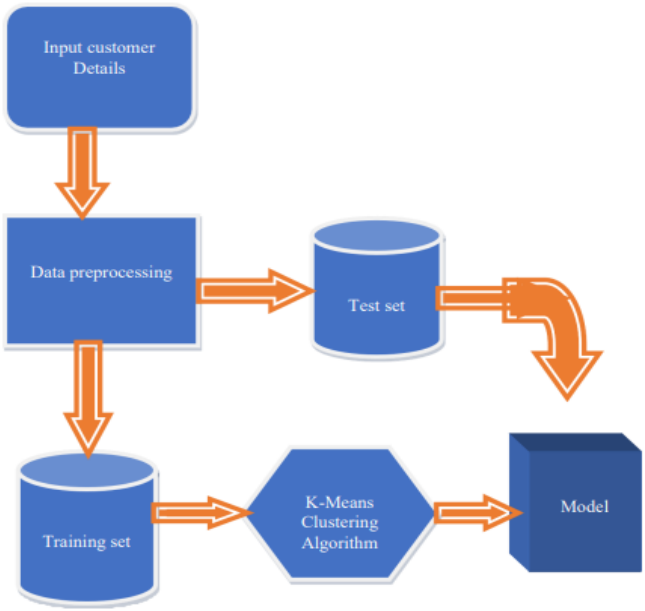


Fig.1 System Architecture

K - Means Clustering Algorithm

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with thePython implementation of k-means clustering. It allows us to cluster the data into different groups and a convenient way to discover the categories of groups in the unlabeled dataset on its own without the need for any training. It is a centroid-based algorithm, where each cluster is associated with a centroid. The main aim of this algorithm is to minimize the sum of distances between the data point and their corresponding clusters.

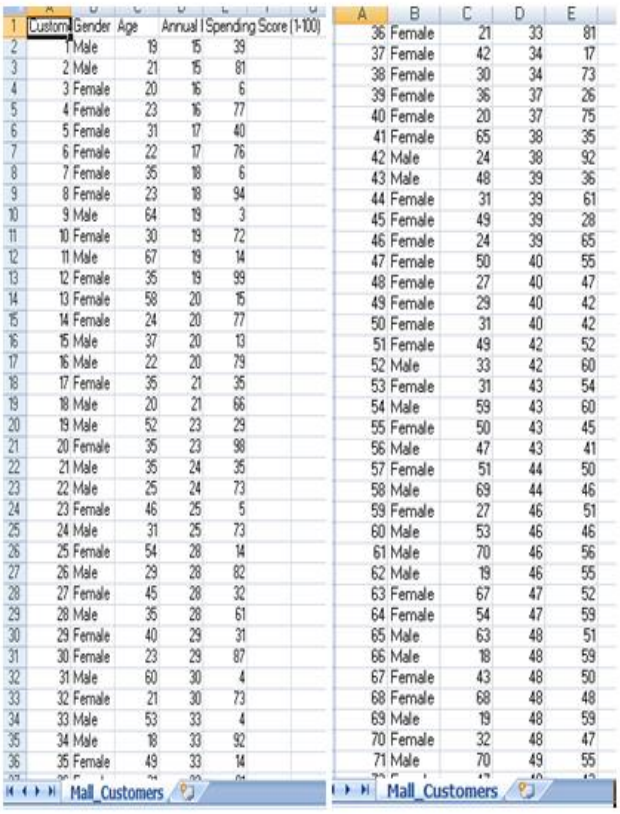
IV. IMPLEMENTATION TECHNIQUES

Pre processing:

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. In order to overcoming these issues we use map function.

Train model on training data set:

Now we will train the model on the training dataset and make predictions for the test dataset. But can we validate these predictions? One way of doing this is we can divide our train dataset into two parts: train and validation. We can train the model on this training part and using that make predictions for the validation part. In this way, we can validate our predictions as we have the true predictions for the validation part (which we do not have for the test dataset)



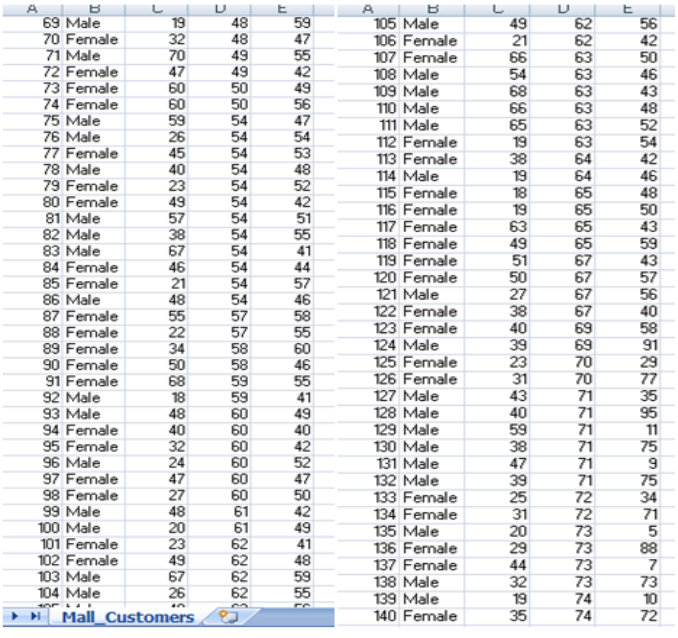
	Custom	Gender	Age	Annual I	Spending Score (1-100)
2	1	Male	19	15	39
3	2	Male	21	15	81
4	3	Female	20	16	6
5	4	Female	23	16	77
6	5	Female	31	17	40
7	6	Female	22	17	76
8	7	Female	35	18	6
9	8	Female	23	18	94
10	9	Male	64	19	3
11	10	Female	30	19	72
12	11	Male	67	19	14
13	12	Female	35	19	99
14	13	Female	58	20	15
15	14	Female	24	20	77
16	15	Male	37	20	13
17	16	Male	22	20	79
18	17	Female	35	21	35
19	18	Male	20	21	66
20	19	Male	52	23	29
21	20	Female	35	23	98
22	21	Male	35	24	35
23	22	Male	25	24	73
24	23	Female	46	25	5
25	24	Male	31	25	73
26	25	Female	54	28	14
27	26	Male	29	28	82
28	27	Female	45	28	32
29	28	Male	35	28	61
30	29	Female	40	29	31
31	30	Female	23	29	87
32	31	Male	60	30	4
33	32	Female	21	30	73
34	33	Male	53	33	4
35	34	Male	18	33	92
36	35	Female	49	33	14

Mall_Customers

	A	B	C	D	E
36	Female	21	33	81	
37	Female	42	34	17	
38	Female	30	34	73	
39	Female	36	37	26	
40	Female	20	37	75	
41	Female	65	38	35	
42	Male	24	38	92	
43	Male	48	39	36	
44	Female	31	39	61	
45	Female	49	39	28	
46	Female	24	39	65	
47	Female	50	40	55	
48	Female	27	40	47	
49	Female	29	40	42	
50	Female	31	40	42	
51	Female	49	42	52	
52	Male	33	42	60	
53	Female	31	43	54	
54	Male	59	43	60	
55	Female	50	43	45	
56	Male	47	43	41	
57	Female	51	44	50	
58	Male	69	44	46	
59	Female	27	46	51	
60	Male	53	46	46	
61	Male	70	46	56	
62	Male	19	46	55	
63	Female	67	47	52	
64	Female	54	47	59	
65	Male	63	48	51	
66	Male	18	48	59	
67	Female	43	48	50	
68	Female	68	48	48	
69	Male	19	48	59	
70	Female	32	48	47	
71	Male	70	49	47	

Mall_Customers

Fig.2 Dataset - 1

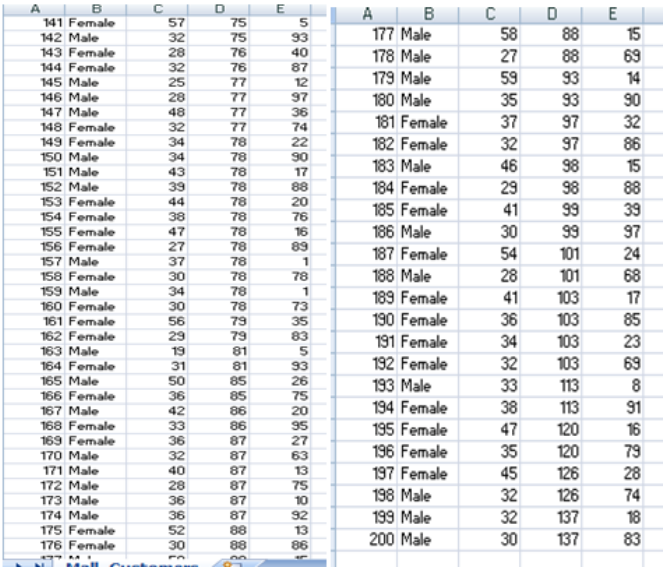


	A	B	C	D	E
69	Male	19	48	59	
70	Female	32	48	47	
71	Male	70	49	55	
72	Female	47	49	42	
73	Female	60	50	49	
74	Female	60	50	56	
75	Male	59	54	47	
76	Male	26	54	54	
77	Female	45	54	53	
78	Male	40	54	48	
79	Female	23	54	52	
80	Female	49	54	42	
81	Male	57	54	51	
82	Male	38	54	55	
83	Male	67	54	41	
84	Female	46	54	44	
85	Female	21	54	57	
86	Male	48	54	46	
87	Female	55	57	58	
88	Female	22	57	55	
89	Female	34	58	60	
90	Female	50	58	46	
91	Female	68	59	55	
92	Male	18	59	41	
93	Male	48	60	49	
94	Female	40	60	40	
95	Female	32	60	42	
96	Male	24	60	52	
97	Female	47	60	47	
98	Female	27	60	50	
99	Male	48	61	42	
100	Male	20	61	49	
101	Female	23	62	41	
102	Female	49	62	48	
103	Male	67	62	59	
104	Male	26	62	55	

Mall_Customers

	A	B	C	D	E
105	Male	49	62	56	
106	Female	21	62	42	
107	Female	66	63	50	
108	Male	54	63	46	
109	Male	68	63	43	
110	Male	66	63	48	
111	Male	65	63	52	
112	Female	19	63	54	
113	Female	38	64	42	
114	Male	19	64	46	
115	Female	18	65	48	
116	Female	19	65	50	
117	Female	63	65	43	
118	Female	49	65	59	
119	Female	51	67	43	
120	Female	50	67	57	
121	Male	27	67	56	
122	Female	38	67	40	
123	Female	40	69	58	
124	Male	39	69	91	
125	Female	23	70	29	
126	Female	31	70	77	
127	Male	43	71	35	
128	Male	40	71	95	
129	Male	59	71	11	
130	Male	38	71	75	
131	Male	47	71	9	
132	Male	39	71	75	
133	Female	25	72	34	
134	Female	31	72	71	
135	Male	20	73	5	
136	Female	29	73	88	
137	Female	44	73	7	
138	Male	32	73	73	
139	Male	19	74	10	
140	Female	35	74	72	

Fig.3 Dataset - 2



	A	B	C	D	E
141	Female	57	75	5	
142	Male	32	75	93	
143	Female	28	76	40	
144	Female	32	76	87	
145	Male	25	77	12	
146	Male	28	77	97	
147	Male	48	77	36	
148	Female	32	77	74	
149	Female	34	78	22	
150	Male	34	78	90	
151	Male	43	78	17	
152	Male	39	78	88	
153	Female	44	78	20	
154	Female	38	78	76	
155	Female	47	78	16	
156	Female	27	78	89	
157	Male	37	78	1	
158	Female	30	78	78	
159	Male	34	78	1	
160	Female	30	78	73	
161	Female	56	79	35	
162	Female	29	79	83	
163	Male	19	81	5	
164	Female	31	81	93	
165	Male	50	85	26	
166	Female	36	85	75	
167	Male	42	86	20	
168	Female	33	86	95	
169	Female	36	87	27	
170	Male	32	87	63	
171	Male	40	87	13	
172	Male	28	87	75	
173	Male	36	87	10	
174	Male	36	87	92	
175	Female	52	88	13	
176	Female	30	88	86	

	A	B	C	D	E
177	Male	58	88	15	
178	Male	27	88	69	
179	Male	59	93	14	
180	Male	35	93	90	
181	Female	37	97	32	
182	Female	32	97	86	
183	Male	46	98	15	
184	Female	29	98	88	
185	Female	41	99	39	
186	Male	30	99	97	
187	Female	54	101	24	
188	Male	28	101	68	
189	Female	41	103	17	
190	Female	36	103	85	
191	Female	34	103	23	
192	Female	32	103	69	
193	Male	33	113	8	
194	Female	38	113	91	
195	Female	47	120	16	
196	Female	35	120	79	
197	Female	45	126	28	
198	Male	32	126	74	
199	Male	32	137	18	
200	Male	30	137	83	

Fig.4 Dataset - 3

Correlating attributes; - Based on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant caninclude Property area, education, loan amount, and lastly credit History, which is since by intuition it is considered as important. The correlation among attributes can be identified using corplot and boxplot in Python platform

Algorithms:

K-means algorithm is used in this project to analyze and form clusters of customers based on their income and spending score features.

Model:

K-means model is used and is hyper tuned parameters liken_clusters=5 using elbow method to find the optimal number of clusters also init='k-means++' to avoid random initialization traps.

Programming and Environment:

Programming Language: Python 3.6 Environment (Libraries and Technologies):

Numpy, Pandas, Matplotlib, Seaborn, JupyterNotebook, Google Colab

Predicting the outcomes:

Using K-MEANS clustering algorithm, the outcomes of all applicant can get theresult.

V. RESULTS

Here we are clustering the customer data based on different criteria's and representing them using various visual data analysis techniques.

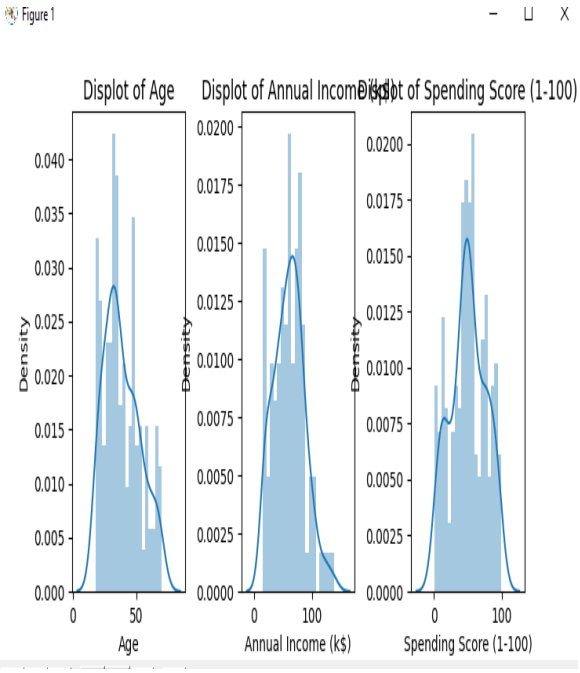


Fig.5 Displot of age, income and spending score

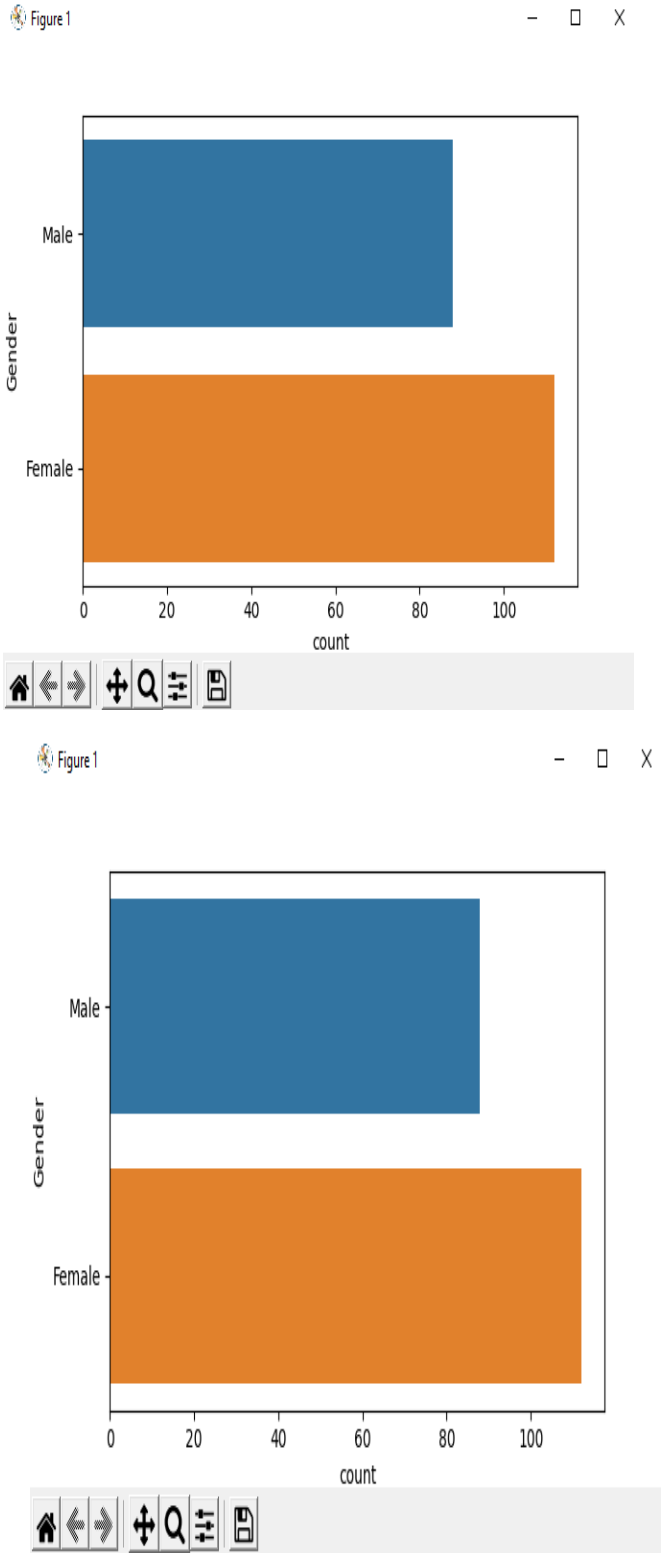


Fig.6 Gender to count graph

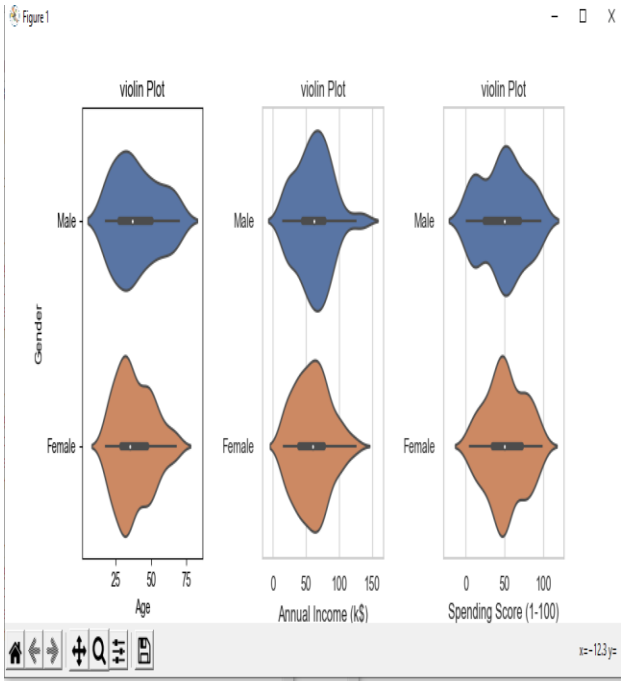


Fig.7 Violin plot for age, income and spending score

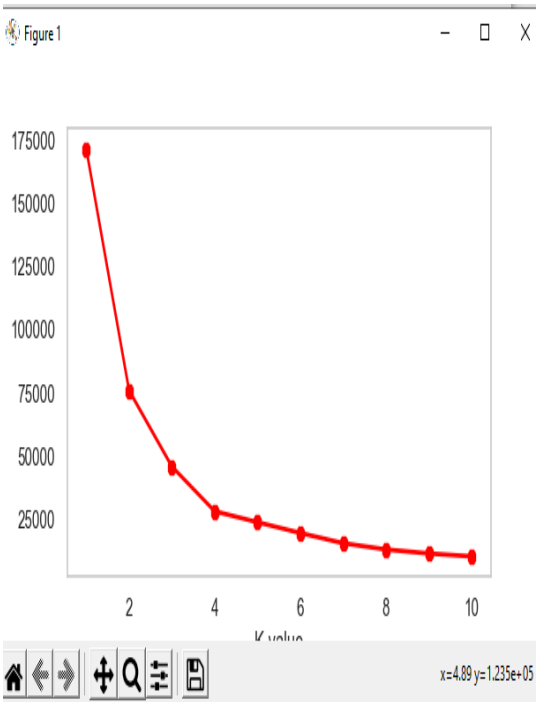


Fig. 9 K means elbow method

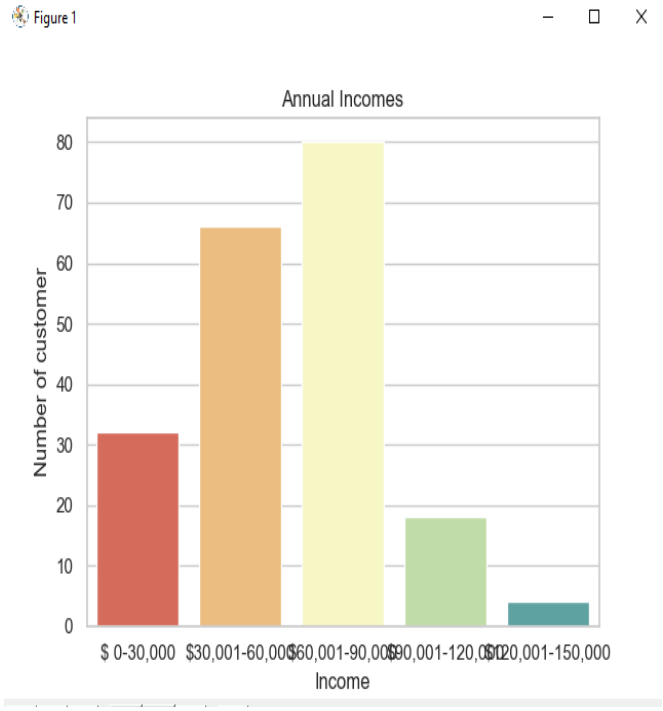


Fig. 8 Number of customers income

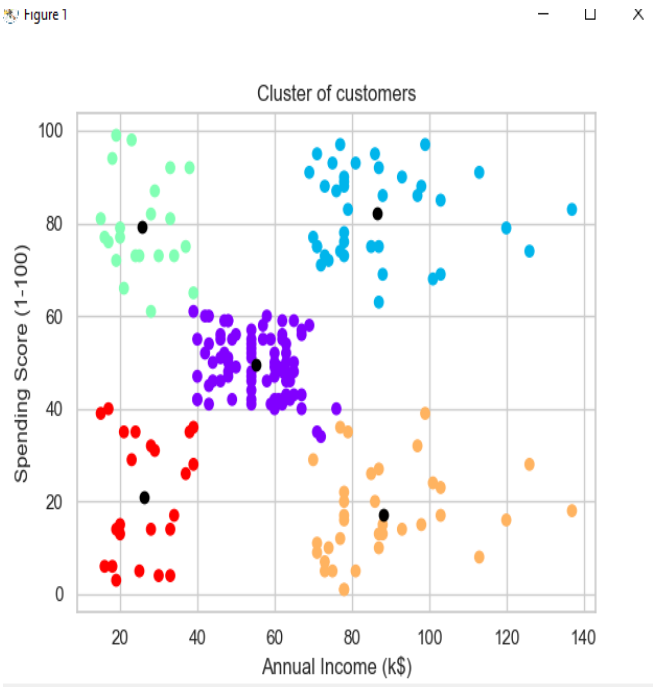


Fig .10 Scatter plot for customers income to spending score

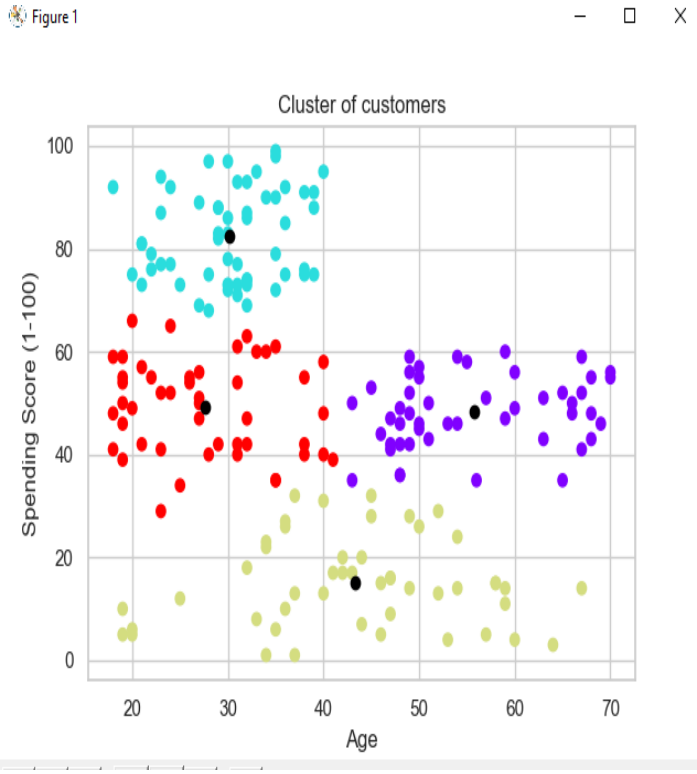


Fig .11 Scatter plot for customers age to spending score

VI. CONCLUSION

In this project we have successfully segregated and represented the data for customers who are visiting mall. This project clusters the customer data based on different criteria and represents them using visual data analysis techniques. At present the Customer segmentation plays a vital role in business and marketing, we are using this customer segmentation mainly in higher levels of business & marketing, using this process we can derive meaningful and useful business insights of a company.

VII. REFERENCES

1. Ilung Pranata and Geo↵ Skinner. “Segmenting and targeting customers through clusters selection & analysis”. In: Advanced Computer Science and Information Systems (ICACSIS), 2015 International Conference on. IEEE. 2015, pp. 303–308

2. J. Kleinberg, "An impossibility theorem for clustering", Proc. 2002Conf. Advances in Neural Information Processing Systems, vol. 15, pp. 463-470, 2002.

3. Minghua Han. “Customer segmentation model based on retail consumer behavior analysis”. In: Intelligent Information Technology Application Workshops, 2008. IITAW’08. International Symposium on. IEEE. 2008, pp. 914–917.

4. Rui Xu and Donald Wunch, "survey of clustering algorithms", IEEE transactions on neural networks, vol. 16, no. 3, may 2005

5. Dr.C K Gomathy, Article: Supply chain-Impact of importance and Technology in Software Release Management, International Journal of Scientific Research in Computer Science Engineering and Information Technology (IJSRCSEIT) Volume 3 | Issue 6 | ISSN : 2456-3307, P.No:1-4, July-2018

6. Dr.C K Gomathy, Article: A Study on the recent Advancements in Online Surveying ,International Journal of Emerging technologies and Innovative Research (JETIR) Volume5 | Issue 11 | ISSN : 2349-5162, P.No:327-331, Nov-2018

Author's Profile:-

1.  Ms. D.Abirami, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her Area of Big data analytics

2.  Mr .Balamurugan, Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. His Area of Big data analytics

3.  Ms. Dondapati Tejaswi Student, B.E. Computer Science and Engineering, Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her Area of Big data analytics

4.  Dr.C.K.Gomathy is Assistant Professor in Computer Science and Engineering at Sri Chandrasekharendra SaraswathiViswa Mahavidyalaya deemed to be university, Enathur, Kanchipuram, India. Her area of interest is Software Engineering, Web Services, Knowledge Management and IOT.