INTERNSHIP REPORT

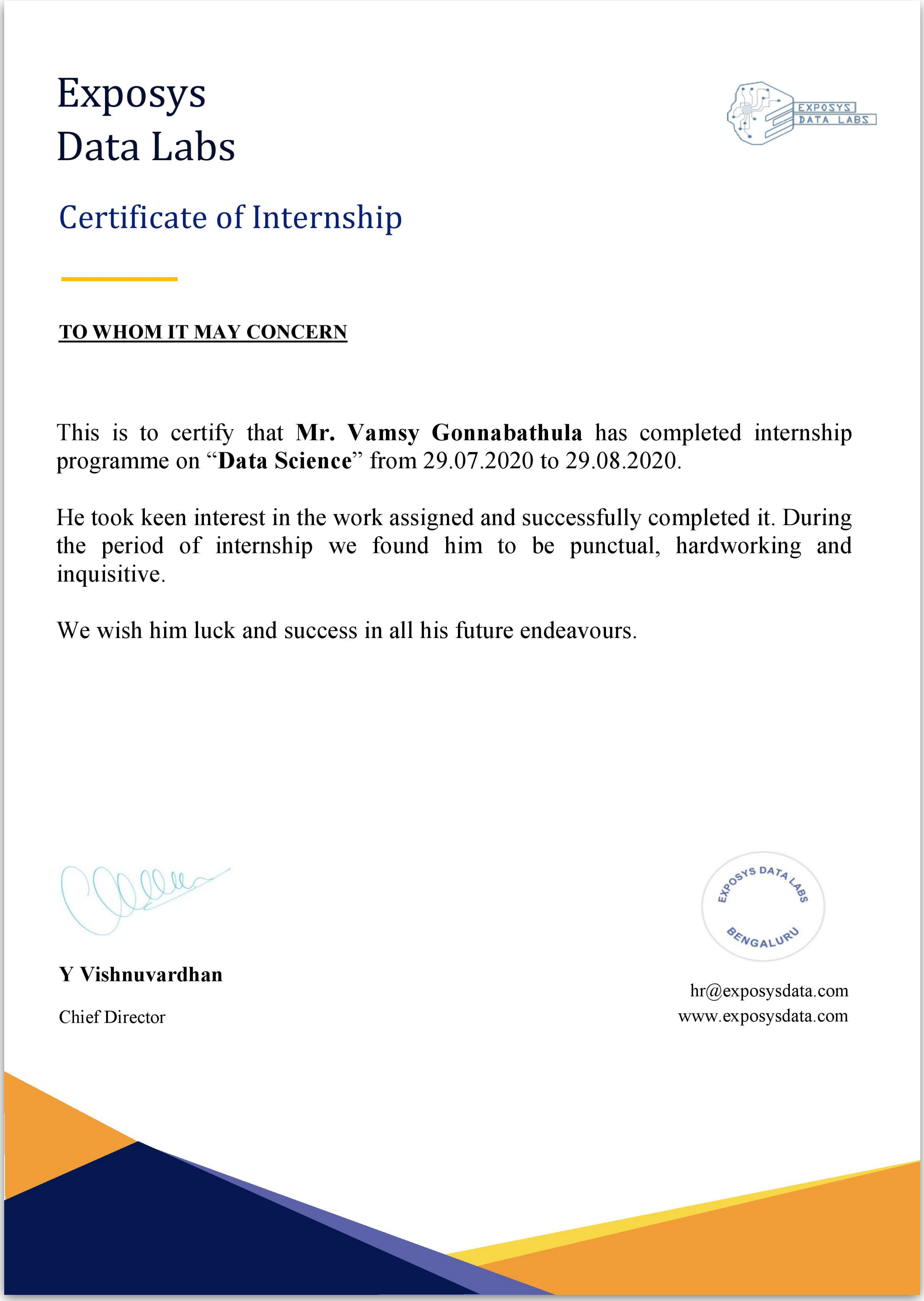
Company: ***Exposys Data Labs, Bengaluru.***
Role: ***Data Science Intern***
Duration: ***1 month [29.07.2020 to 29.08.2020]***

BY,

GONNABATHULA V. S VAMSY

(121710306015) B-6

CERTIFICATE



Exposys
Data Labs

EXPOSYS DATA LABS

Certificate of Internship

TO WHOM IT MAY CONCERN

This is to certify that **Mr. Vamsy Gonnabathula** has completed internship programme on “**Data Science**” from 29.07.2020 to 29.08.2020.

He took keen interest in the work assigned and successfully completed it. During the period of internship we found him to be punctual, hardworking and inquisitive.

We wish him luck and success in all his future endeavours.

EXPOSYS DATA LABS
BENGALURU

Y Vishnuvardhan
Chief Director

hr@exposysdata.com
www.exposysdata.com

ACKNOWLEGEMENT

The satisfaction that accompanies the successful completion of any task would be incomplete without the mentioning of people who made it possible and for the constant guidance and encouragement crown all the effort with success.

I want to thank Mr. Y Vishnuvardhan, (Chief Director) of Exposys Data Labs for welcoming me into this internship. I would also like to thank Dr. Arvind Kumar (Research Scientist) for providing me with technical clarifications and support during the entire timeline of this internship.

ABSTRACT

In the past years, research in the fields of big data analysis, machine learning and data mining techniques is getting more frequent. This report describes a customer segmentation approach in case of Mall Customers. These customer groups are based on user interactions with items in the marketplace such as views and “likes” or a membership card. A major goal of this thesis was to construct a feed for many customers where the items are derived from the user groups which can be achieved by Unsupervised Learning.

The customer segmentation method discussed in this report is based on the K-Means clustering algorithms and data visualisation of different distributions. Then input the K-means clustering and Hierarchal clustering we analyse customers groups annual incomes and spending scores. A visualisation tool was also constructed in order to get a better picture of the data and the resulting clusters. In order to visualise the given dataset many packages were imported and installed to achieve this process. This then serves the marketing goal of targeting different customer groups.

TABLE OF CONTENTS

1. Introduction

1.1 WHAT IS CUSTOMER SEGMENTATION?

- Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways relevant to marketing, such as age, gender, interests and spending habits.

- Companies employing customer segmentation operate under the fact that every customer is different and that their marketing efforts would be better served if they target specific, smaller groups with messages that those consumers would find relevant and lead them to buy something. Companies also hope to gain a deeper understanding of their customers' preferences and needs with the idea of discovering what each segment finds most valuable to more accurately tailor marketing materials toward that segment.

- Customer segmentation relies on identifying key differentiators that divide customers into groups that can be targeted. Information such as a customers' demographics (age, race, religion, gender, family size, ethnicity, income, education level), geography (where they live and work), psychographic (social class, lifestyle and personality characteristics) and behavioural (spending, consumption, usage and desired benefits) tendencies are taken into account when determining customer segmentation practices.

1.2 ADVANTAGES OF CUSTOMER SEGMENTATION

1. *Personalisation*
    - *Personalisation ensures that you provide exceptional customer experience.*
2. *Customer Retention*
    - *It is 16 times as costly to build a long-term business relationship with a new customer than simply to cultivate the loyalty of an existing customer.*
3. *Better ROI for marketing*
    - *Affirmations that right marketing messages are sent to the right people based on their life cycle stage.*
4. *Reveal new opportunities*
    - *Customer segmentation may reveal new trends about products and it may even give the first mover's advantage in a product segment.*

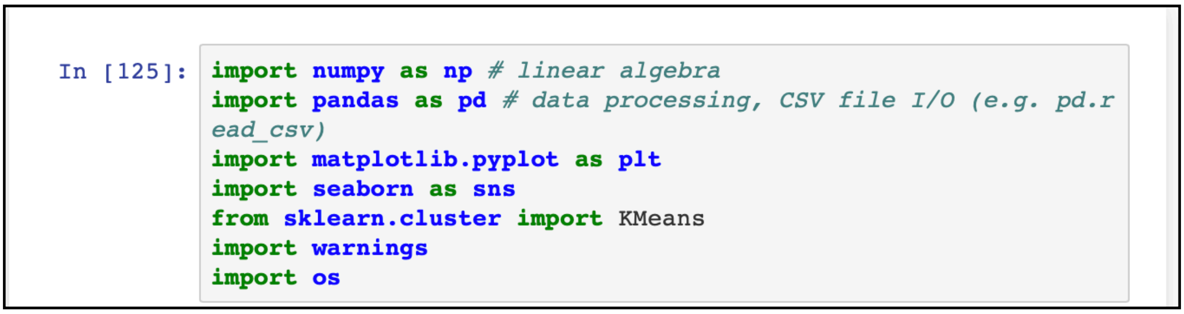
1.3 APPROACH USING MACHINE LEARNING

- Unsupervised Learning is a class of Machine Learning techniques to find the patterns in data. The data given to unsupervised algorithm are not labelled, which means only the input variables(X) are given with no corresponding output variables. In unsupervised learning, the algorithms are left to themselves to discover interesting structures in the data.

- There are some analytics techniques that can help you with segmenting your customers. These are useful especially when you have a large number of customers and it's hard to discover patterns in your customer data just by looking at transactions. The two most common ones are:

1. Clustering
    - Clustering is an exploration technique for datasets where relationships between different observations may be too hard to spot with the eye.
2. Principal Component Analysis (PCA)
    - PCA is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.

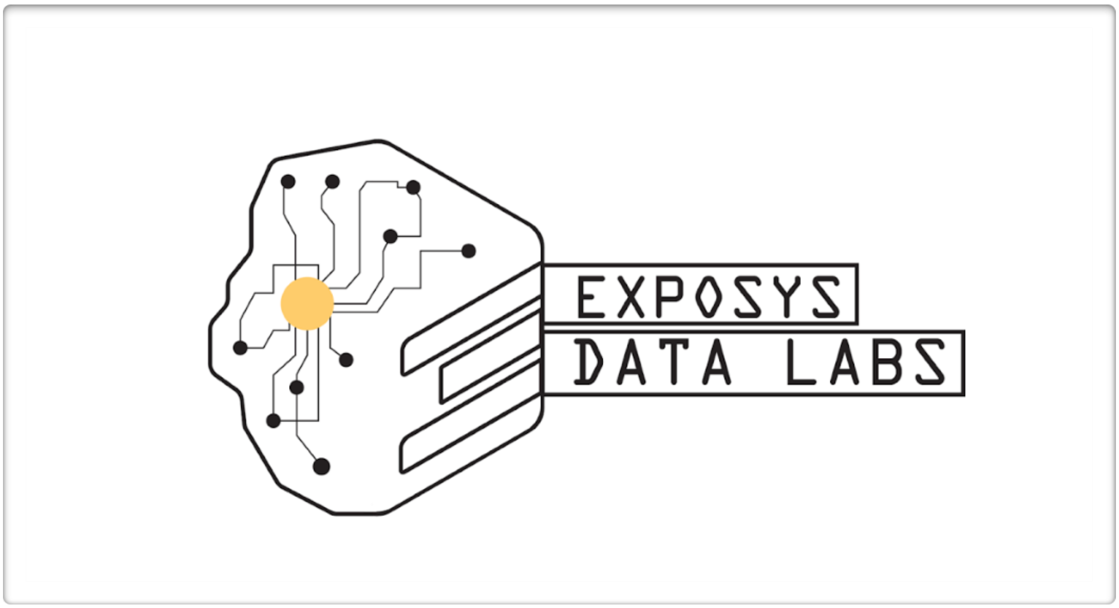
The following code takes advantage of the Mall Customer Segmentation Data to demonstrate the ability of K-Means clustering algorithm to identify customer's segments.



```
In [125]: import numpy as np # linear algebra
          import pandas as pd # data processing, CSV file I/O (e.g. pd.r
          ead_csv)
          import matplotlib.pyplot as plt
          import seaborn as sns
          from sklearn.cluster import KMeans
          import warnings
          import os
```

- matplotlib.pyplot is a comprehensive library for creating static, animated and interactive visuals in notebook.
- Seaborn is used to drawing attractive and informative statistical graphics. Used for data visualisation.
- K means is imported through tensor flow which is used for clustering of data to associate the data in more efficient manner. It is the common technique in Unsupervised Learning.
- Warning is base category about feature that will be deprecated in future i.e ignored by default.
- Os interacts with the file system.

2. PROFILE OF THE COMPANY



- Exposys Data Labs, Bangalore aims to solve real world business problems like Research, Automation, Big Data and Data Science. The core team of experts in various technologies help businesses to identify issues, opportunities and prototype solutions for Energy, Humanoid, Blockchain Technology, Marketing, etc. Using trending technologies like AI, ML, Deep Learning and Data Science.

- **Website:** http://www.exposysdata.com
- **Industry:** Computer Software
- **Company size:** 11-50 employees
- **Headquarters:** Bengaluru, Karnataka
- **Type:** Public Company
- **Founded:** 2019

3.Exisiting method

3.1 LIBRARIES AND TOOLS INSTALLED:

-NumPy (1.18.5)

-Pandas (1.0.5)

-matplotlib.pyplot (3.2.2)

-seaborn (0.10.1)

-Keras(2.3.1)

-warnings(1.2.0)

-os (0.1.0)

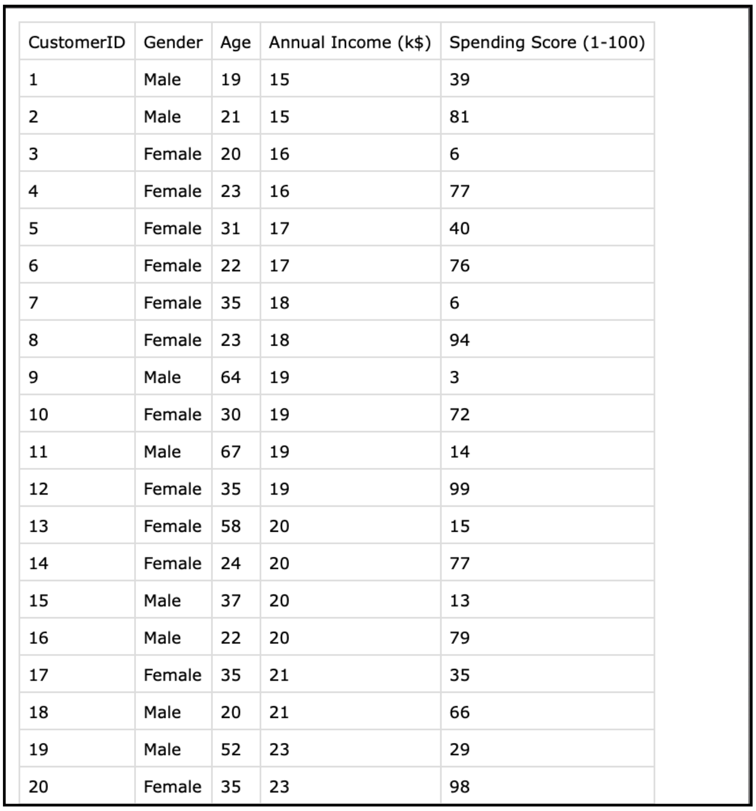
`Programming Language:` *Python v3.6*
`Environment:` *Jupyter Notebook v6.0.3*
`Tools used:` *Anaconda Navigator*

3.2 ANALYZING DATASET PROVIDED:-

Here we have the following features :

1. CustomerID: It is the unique ID given to a customer(200 CUSTOMERS DATA PROVIDED)
2. Gender: Gender of the customer
3. Age: The age of the customer
4. Annual Income(k$): It is the annual income of the customer
5. Spending Score: It is the score(out of 100) given to a customer by the mall authorities, based on the money spent and the behavior of the customer.

Here's how the data set is provided.



CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40
6	Female	22	17	76
7	Female	35	18	6
8	Female	23	18	94
9	Male	64	19	3
10	Female	30	19	72
11	Male	67	19	14
12	Female	35	19	99
13	Female	58	20	15
14	Female	24	20	77
15	Male	37	20	13
16	Male	22	20	79
17	Female	35	21	35
18	Male	20	21	66
19	Male	52	23	29
20	Female	35	23	98

4. PROPOSED METHOD WITH ARCHITECTURE

4.1 PACKAGES USED:

Below is the snapshot of environments used in Jupyter Notebook. While using machine learning models we can use Python or R as suitable language to commence tasks. I have used Python as language as it is more user friendly and easy to implement and understand.

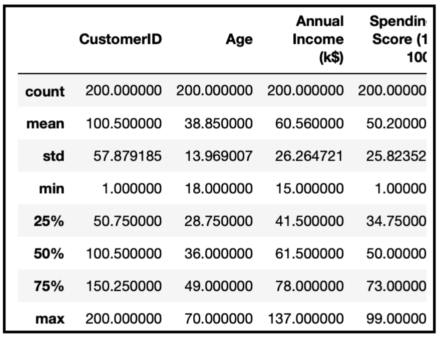
We first import (NumPy) which does the functions of linear algebra I.e mathematical functions in the dataset.

Pandas is imported for one main reason that is for data processing. As we deal with concept of data science. It also deals with data manipulation and analysis

4.2 USING PYTHON TO ACCESS DATASET:

In order to access the dataset I have first made READ DATA into the notebook by using using the command data = pd.read_csv('Mall_Customers.csv') and data head() to read and display the .csv file.

Then we find the shape of data in matrix form which gives (200,5) and followed by data.describe() We get like the following:



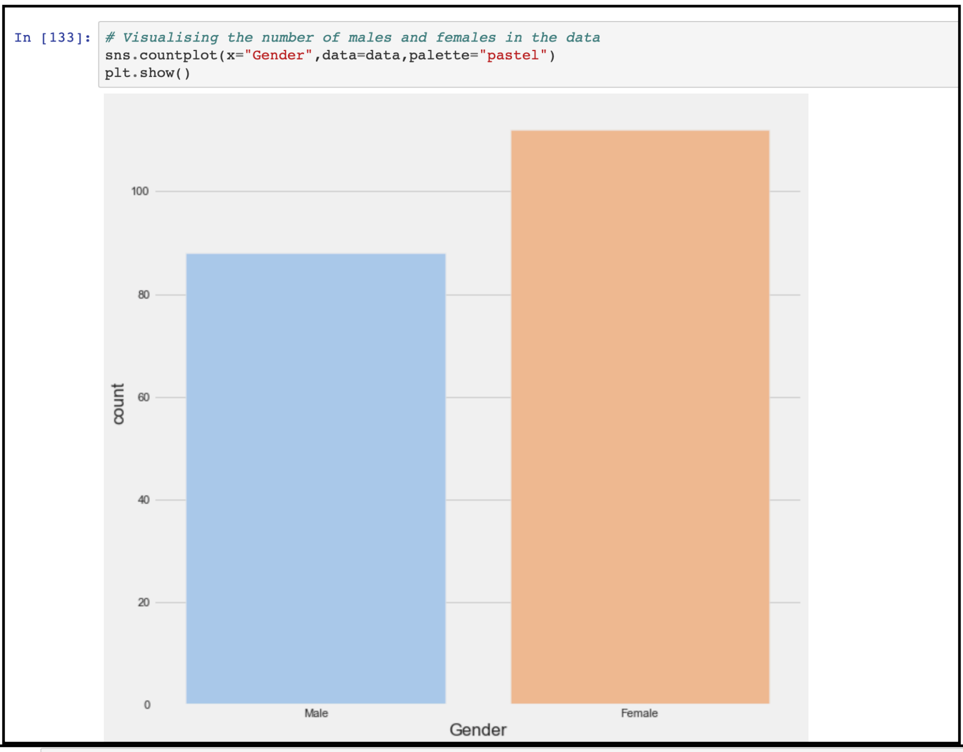
	CustomerID	Age	Annual Income (k$)	Spendin Score (1 10(
count	200.000000	200.000000	200.000000	200.00000
mean	100.500000	38.850000	60.560000	50.20000
std	57.879185	13.969007	26.264721	25.82352
min	1.000000	18.000000	15.000000	1.00000
25%	50.750000	28.750000	41.500000	34.75000
50%	100.500000	36.000000	61.500000	50.00000
75%	150.250000	49.000000	78.000000	73.00000
max	200.000000	70.000000	137.000000	99.00000

- The count is 200 means we have records of 200 customers with us.
- The minimum age of customer in our data is 18 yrs and maximum age is 70.
- The mean here is 38 and median is 36.Here Mean>Median means our data has high outliers ie more of youngsters prefer to go malls.
- The minimum annual income of customer is 15k$ and maximum is 137k$.T
- The mean and median here is 60k$ and 61k$ respectively.
- Spending Score is something you assign to the customer based on your defined parameters like customer behaviour and purchasing data.Here the minimum spending score assigned is 1 and maximum ranges till 99.Both mean and median is 50.

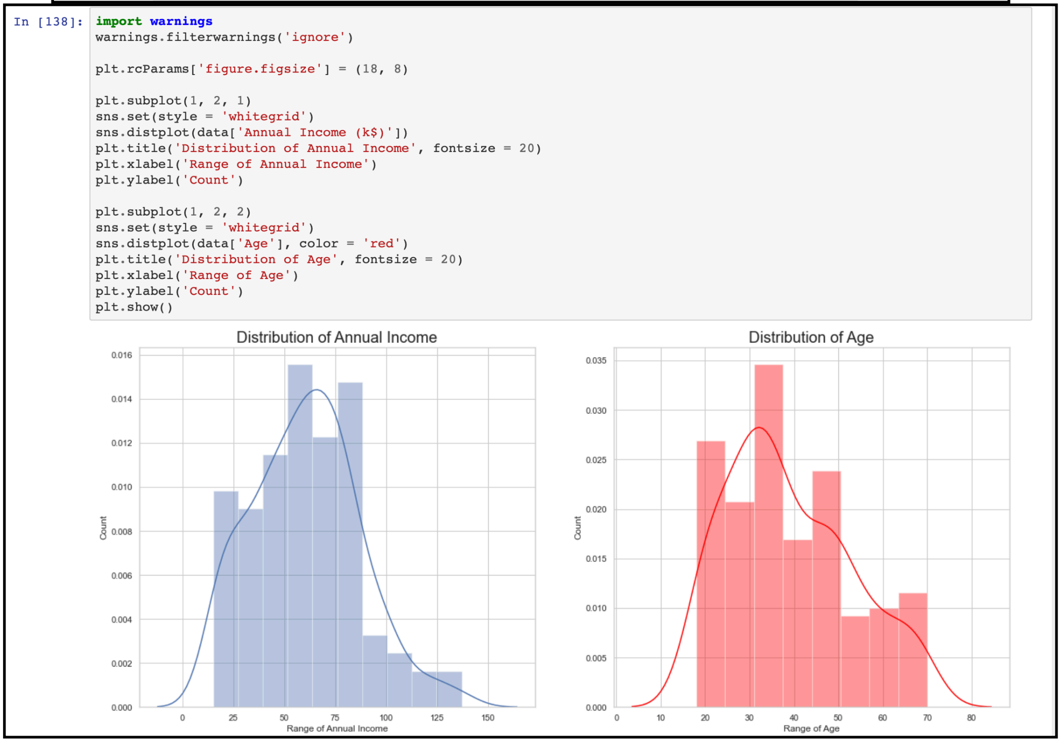
4.3 VISUALIZING DATASET:

Visualising data is the first step to reach goal in machine learning. Using the installed libraries seaborn.sns

- I have made a graph showing Male and Female data which clearly shows that female count ratio is higher than male ratio.



```
In [133]: # Visualising the number of males and females in the data
          sns.countplot(x="Gender",data=data,palette="pastel")
          plt.show()
```

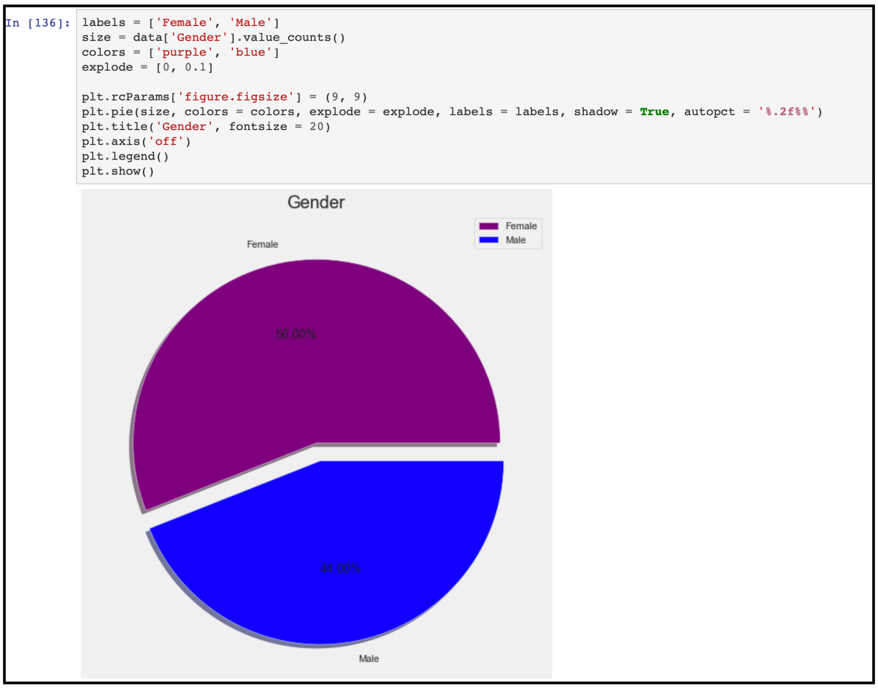


```
In [138]: import warnings
          warnings.filterwarnings('ignore')

          plt.rcParams['figure.figsize'] = (18, 8)

          plt.subplot(1, 2, 1)
          sns.set(style = 'whitegrid')
          sns.distplot(data['Annual Income (k$)'])
          plt.title('Distribution of Annual Income', fontsize = 20)
          plt.xlabel('Range of Annual Income')
          plt.ylabel('Count')

          plt.subplot(1, 2, 2)
          sns.set(style = 'whitegrid')
          sns.distplot(data['Age'], color = 'red')
          plt.title('Distribution of Age', fontsize = 20)
          plt.xlabel('Range of Age')
          plt.ylabel('Count')
          plt.show()
```

In the above Plots we can see the **Distribution pattern of Annual Income ($) and Age**, By looking at the plots, we can infer one thing that There are few people who earn more than 100 US Dollars. Most of the people have an earning of around 50-75 US Dollars. Also, we can say that the least Income is around 20 US Dollars.

Taking inferences about the Customers.

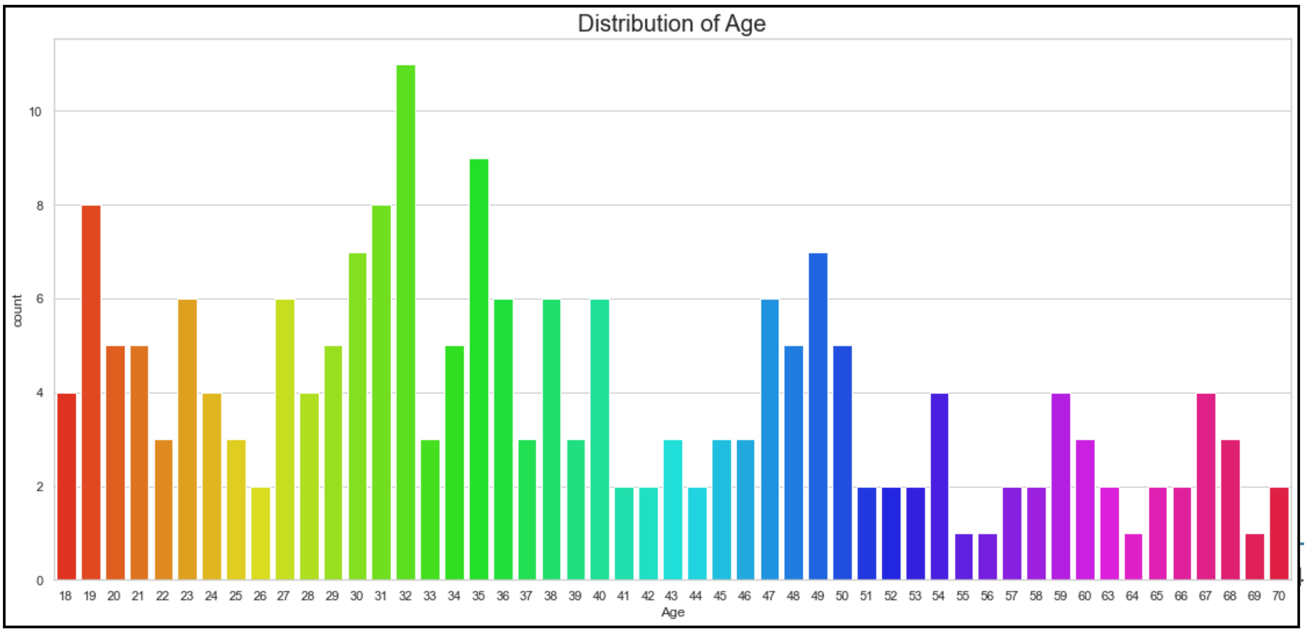
The most regular customers for the Mall has age around 30-35 years of age. Whereas the the senior citizens age group is the least frequent visitor in the Mall. Youngsters are lesser in number as compared to the Middle aged people.



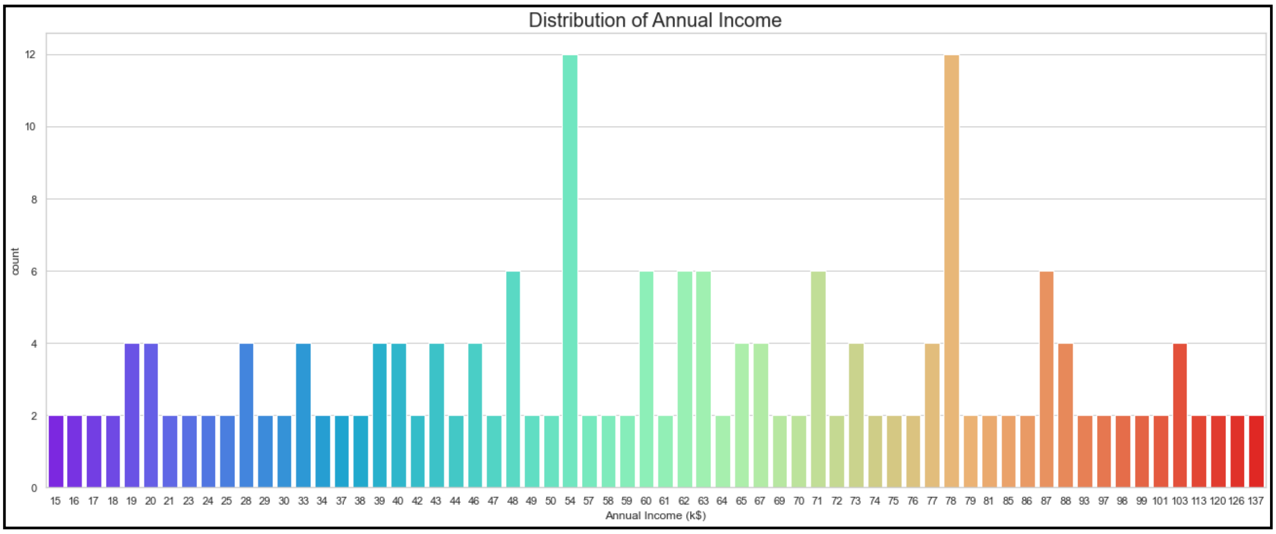
```
In [136]: labels = ['Female', 'Male']
          size = data['Gender'].value_counts()
          colors = ['purple', 'blue']
          explode = [0, 0.1]

          plt.rcParams['figure.figsize'] = (9, 9)
          plt.pie(size, colors = colors, explode = explode, labels = labels, shadow = True, autopct = '%.2f%%')
          plt.title('Gender', fontsize = 20)
          plt.axis('off')
          plt.legend()
          plt.show()
```

Here, Females are in the lead with a share of 56% whereas the Males have a share of 44%, that's a huge gap specially when the population of Males is comparatively higher than Females.

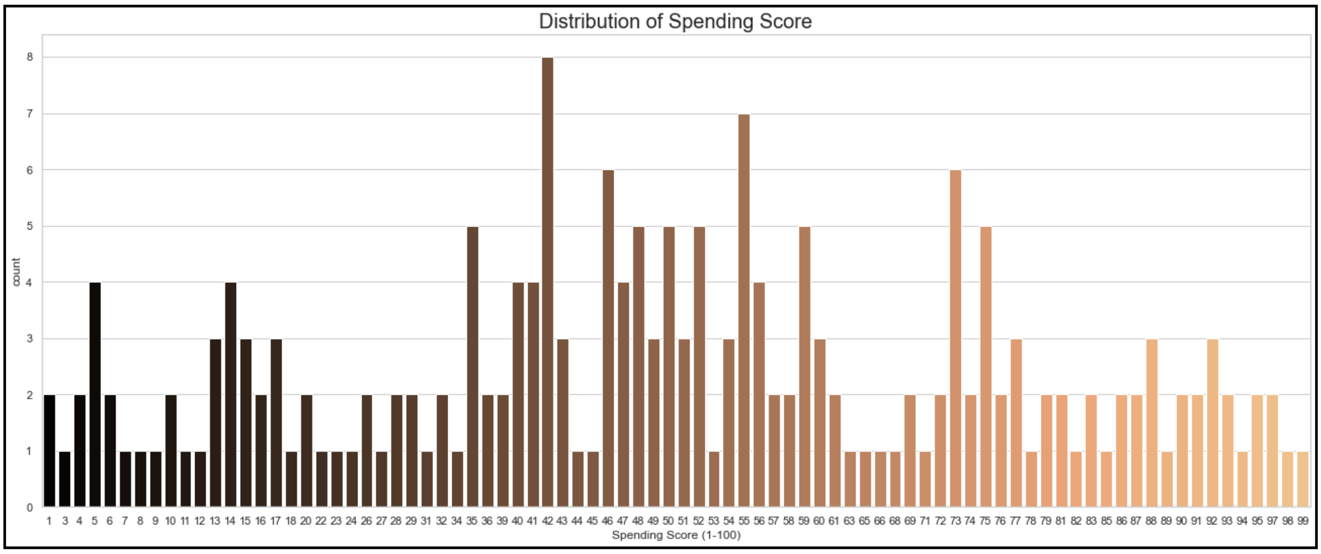
In Visualisation, the size, shape, colour can be modified to the user's choice.



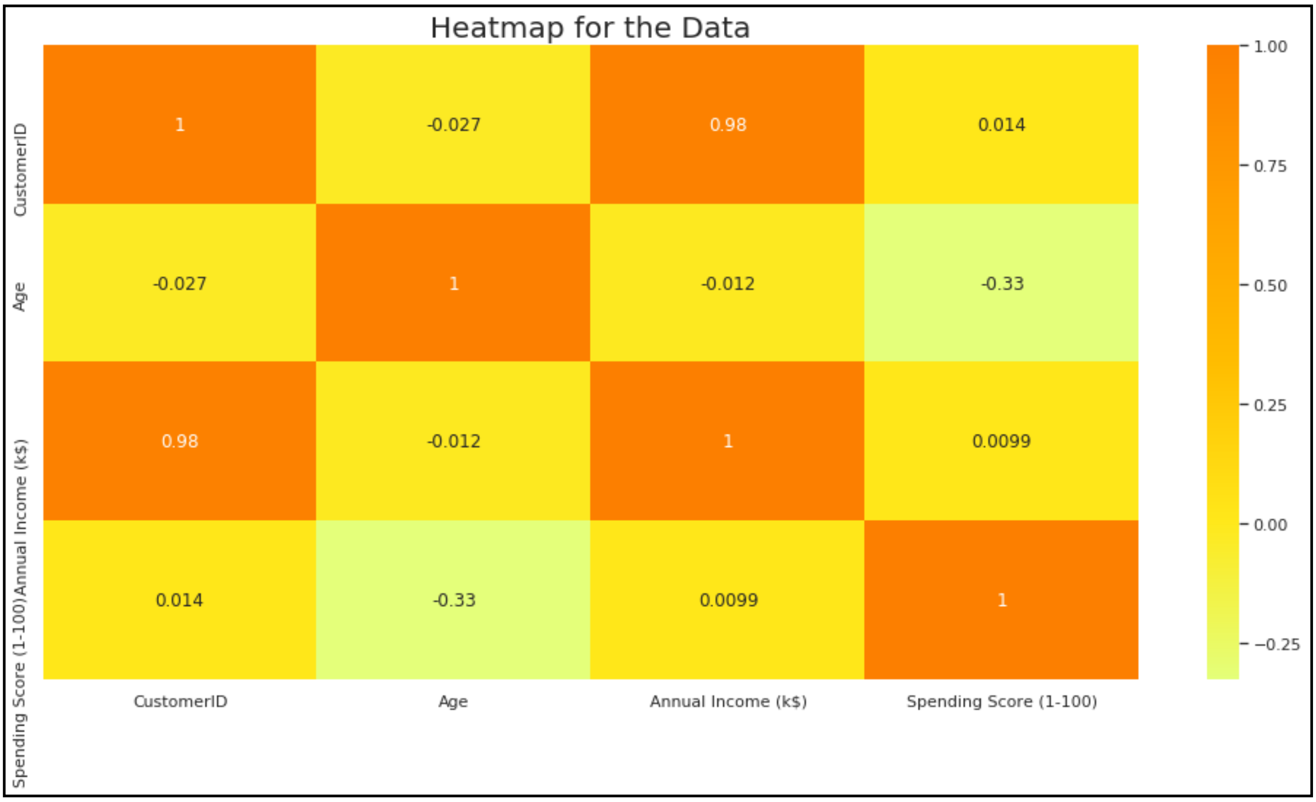
- This Graph shows a more Interactive Chart about the **Distribution of each Age Group** in the Mall for more clarity about the Visitor's Age Group in the Mall. By looking at the above graph-, It can be seen that the Ages from 27 to 39 are very much frequent but there is no clear pattern, we can only find some group wise patterns such as the the older age groups are lesser frequent in comparison. Interesting Fact, There are equal no. of Visitors in the Mall for the Agee 18 and 67. People of Age 55, 56, 69, 64 are very less frequent in the Malls. People at Age 32 are the Most Frequent Visitors in the Mall.



A chart to better explain the **Distribution of Each Income level**, Interesting there are customers in the mall with a very much comparable frequency with their Annual Income ranging from 15 US Dollars to 137K US Dollars. There are more Customers in the Mall who have their Annual Income as 54k US Dollars or 78 US Dollars.

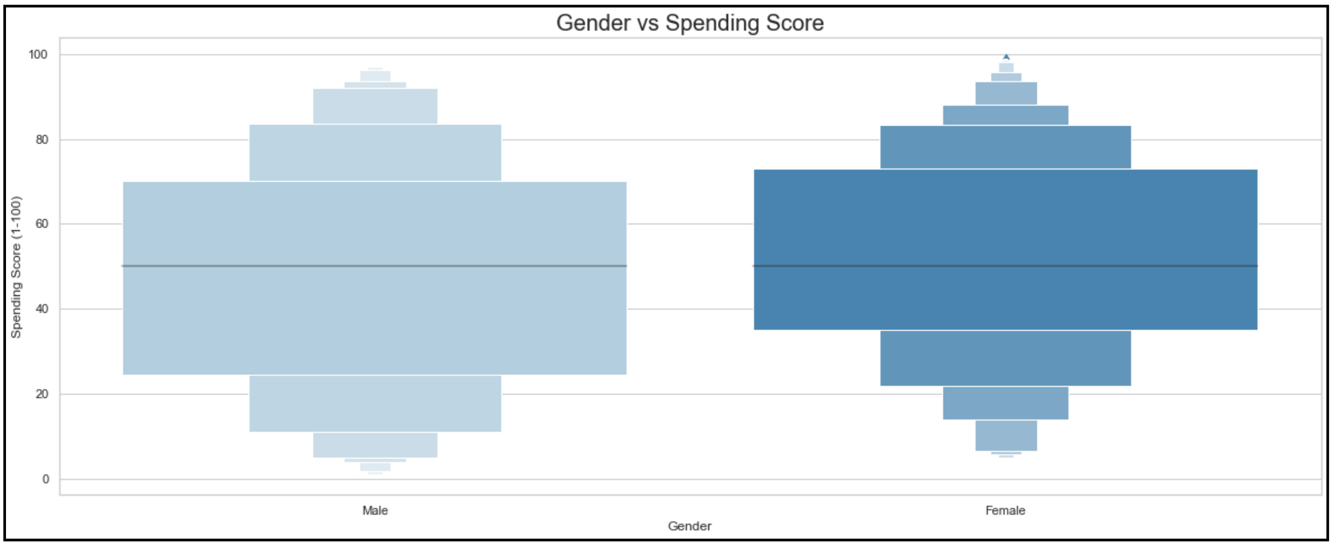


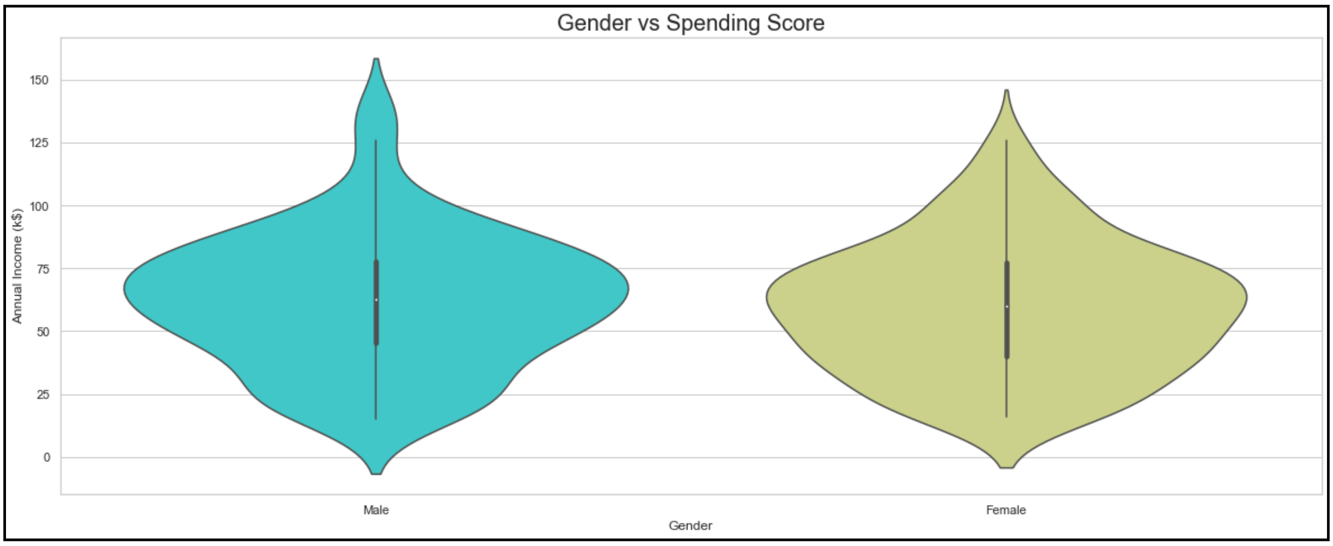
- On a general level, we may conclude that most of the Customers have their **Spending Score** in the range of 35-60. Interesting there are customers having I spending score also, and 99 Spending score also, Which shows that the mall caters to the variety of Customers with Varying needs and requirements available in the Mall.



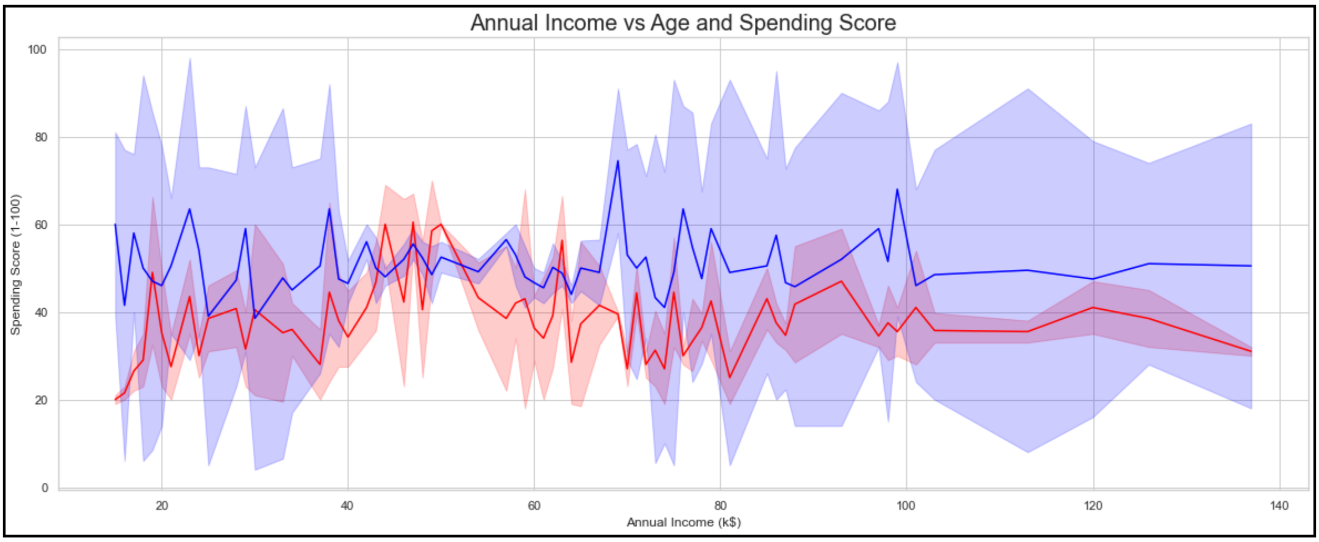
- The Above Graph for Showing the correlation between the different attributes of the Mall Customer Segmentation Dataset, This **Heat map** reflects the most correlated features with Orange Colour and least correlated features with yellow colour.
- We can clearly see that these attributes do not have good correlation among them, that's why we will proceed with all of the features.

- **Bi-variate Analysis** between **Gender** **and** **Spending Score,**
- It is clearly visible that the most of the males have a Spending Score of around 25k US Dollars to 70k US Dollars whereas the Females have a spending score of around 35k US Dollars to 75k US Dollars. which again points to the fact that women are Shopping Leaders.





- **Bi-variate Analysis between the Gender and the Annual Income,**
- There are more number of males who get paid more than females. But, The number of males and females are equal in number when it comes to low annual income.



- The above Plot Between **Annual Income and Age** represented by a **blue** colour line, and a plot between **Annual Income and the Spending Score** represented by a **pink** colour. shows how Age and Spending Varies with Annual Income.

5.METHODOLGY

5.1 UNSUPERVISED LEARNING TECHNIQUE:

- K-means algorithm explores for a preplanned number of clusters in an unlabelled multidimensional dataset, it concludes this via an easy interpretation of how an optimised cluster can be expressed.

- Primarily the concept would be in two steps, **firstly,** the cluster centre is the arithmetic mean (AM) of all the data points associated with the cluster. **Secondly**, each point is adjoint to its cluster centre in comparison to other cluster centres. These two interpretations are the foundation of the k-means clustering model.

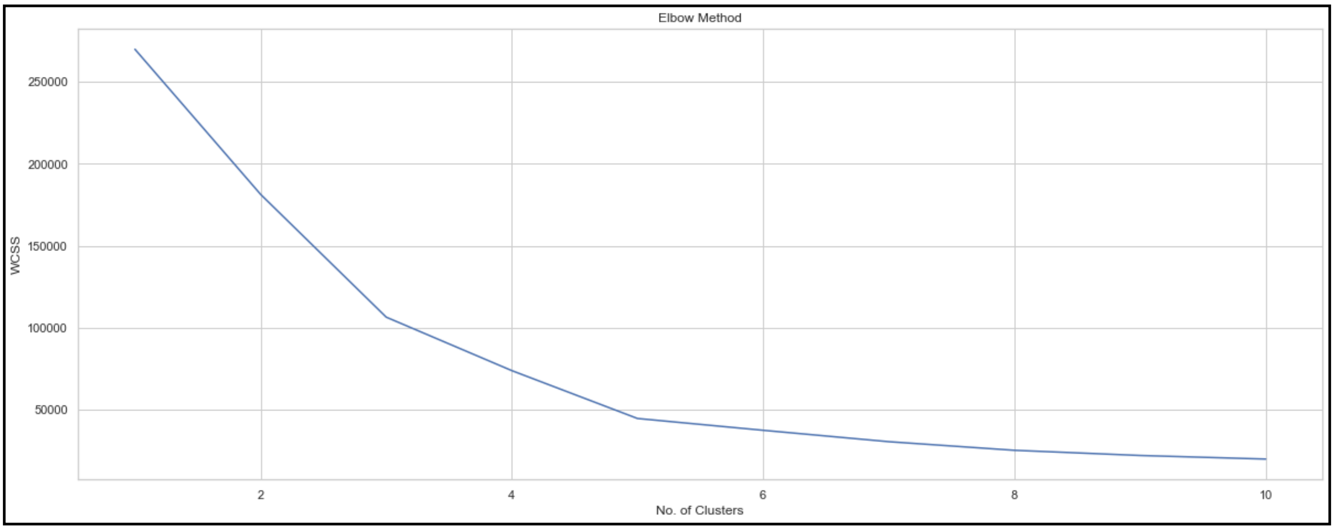
- You can take the centre as a data point that outlines the means of the cluster, also it might not possibly be a member of the dataset.

- K-means Clustering is applied in the **Call Detail Record (CDR) Analysis**. It gives in-depth vision about customer requirements and satisfaction on the basis of call-traffic during the time of the day and demographic of a particular location.

5.2 K-MEANS CLUSTERING(ELBOW METHOD):

- I will use the K-Means Clustering algorithm to cluster the data.
- To implement K-Means clustering, we need to look at the **Elbow Method**.

The Elbow method is a method of interpretation and validation of consistency within-cluster analysis designed to help to find the appropriate number of clusters in a dataset.



- ***Step 1***: Taking annual income and spending score in x to make clusters

```
x=data.iloc[:,[3,4]]
```

- ***Step 2***: Using elbow method in order to find the optimal number of clusters

```
from sklearn.cluster import KMeans
wcss=[]
for i in range (1,11):
    kmeans=KMeans(n_clusters=i, init=“k-means+
+”,max_iter=300,n_init=10, random_state=0)
    kmeans.fit(x)
    wcss.append(kmeans.inertia_)
plt.plot(range(1,11),wcss)
plt.xlabel("No. of Clusters")
plt.ylabel("WCSS")
plt.title("Elbow Method")
plt.show()
```

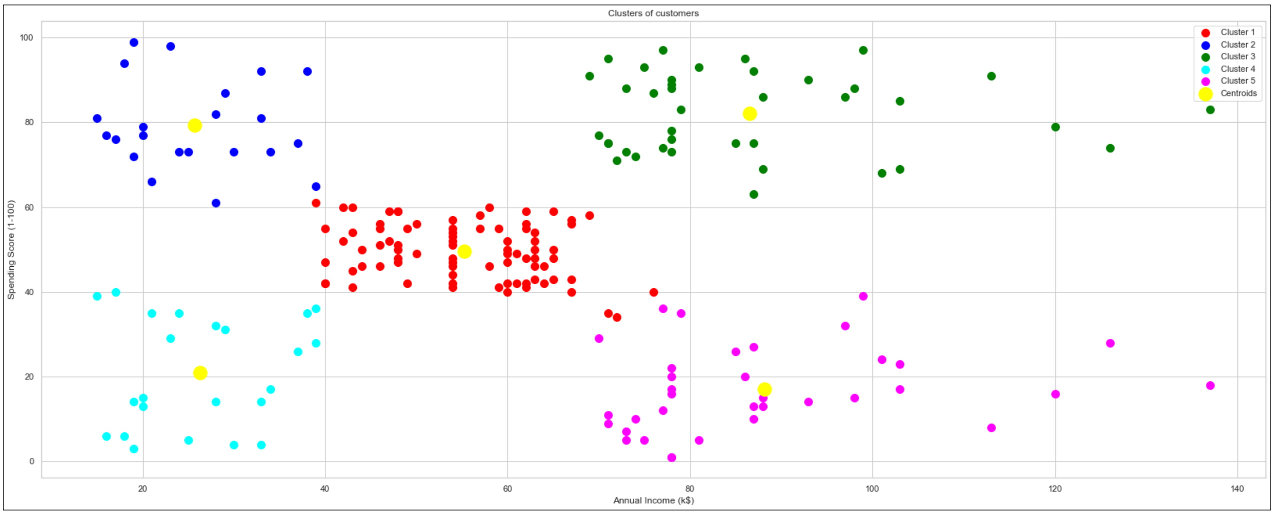
It is clear from the figure that we should take the number of clusters equal to 5, as the slope of the curve is not steep enough after it.

- The labels property of the K-means clustering example dataset that is, how the data points are categorised.

```
array([3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1,
       3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 1, 3, 0,
       3, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 4, 2, 0, 2, 4, 2, 4, 2,
       0, 2, 4, 2, 4, 2, 4, 2, 4, 2, 0, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2,
       4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2,
       4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2, 4, 2,
       4, 2], dtype=int32)
```

- **After applying visualisation, Here is how the data is displayed**

The data(clusters) are plotted on a **Spending score V/S Annual income curve**.

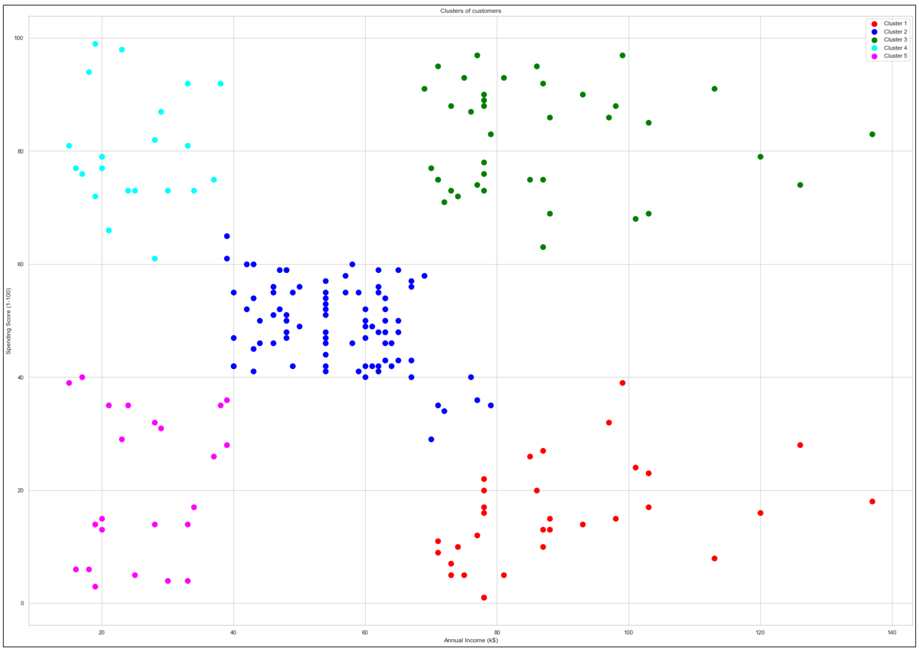


5.3 HIERARCHAL CLUSTERING:

- **Hierarchical clustering**, also known as **hierarchical cluster** analysis, is an algorithm that groups similar objects into groups called **clusters**. The endpoint is a set of **clusters**, where each **cluster** is distinct from each other **cluster**, and the objects within each **cluster** are broadly similar to each other.

- There are two types of hierarchical clustering, Divisive and Agglomerative.
- The optimal number of clusters = 5
- Now, I imported AgglomerativeClustering and created a object hc class AgglomerativeClustering()

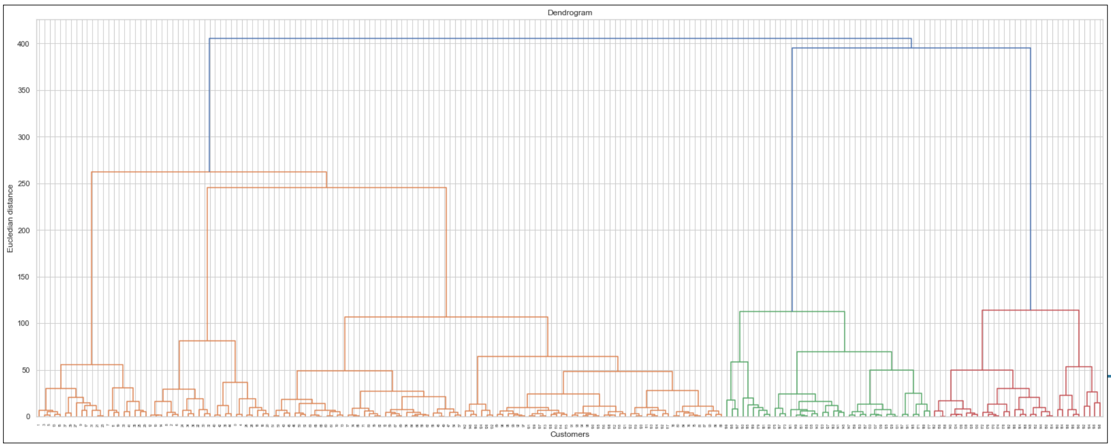
- hc=AgglomerativeClustering(n_clusters=5, affinity=“euclidean”, linkage="ward")



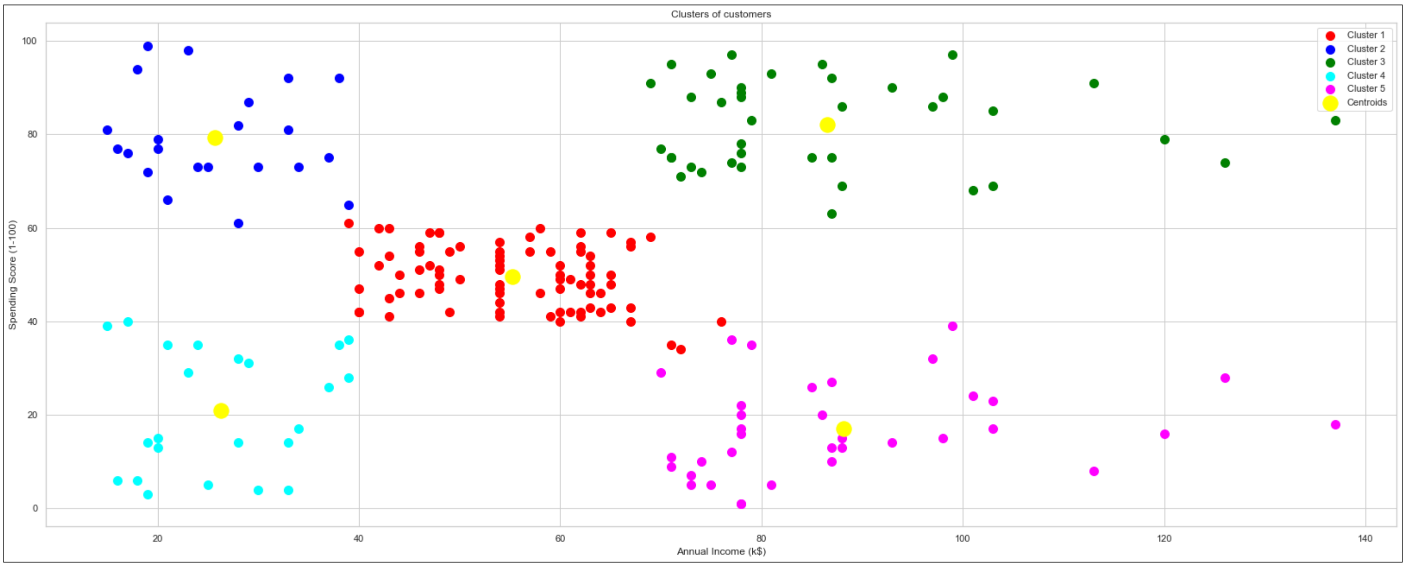
- y_hc=hc.fit_predict(x) to give an array which tells as to which data point belongs to which cluster (0,1,2,3,4).

- And now I visualise the clusters by plotting and labelling.

We need to create a **Dendrogram** it is a plot between Euclidean distance (y-axis) and data points (x-axis) by importing scipy.cluster.hierarchy as sch and we then assign size and plot it to get the following:



6. FINAL ANALYSIS:



The data(clusters) are plotted on a **Spending score Vs Annual income** curve.
Let us now analyse the results of the model.

Analysing the Results:-

We can see that the mall customers can be broadly grouped into 5 groups based on their purchases made in the mall.

- **In Cluster 4 (LIGHT BLUE)** we can see people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, in fact, these are the wise people who know how to spend and save money. The shops/mall will be least interested in people belonging to this cluster.
- **In Cluster 2 (NAVY BLUE)** we can see that people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. Maybe it's because these people are more than satisfied with the mall services. The shops/malls might not target these people that effectively but still will not lose them.
- **In Cluster 5 (PINK)** we see that people have average income and an average spending score, these people again will not be the prime targets of the shops or mall, but again they will be considered and other data analysis techniques may be used to increase their spending score.
- **In Cluster 1 (RED)** we see that people have medium income and medium spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. These people might be the regular customers of the mall and are convinced by the mall's facilities.
- **In Cluster 3 (GREEN)** we see that people have high income, high spending scores, this is interesting. Maybe these are the people who are unsatisfied or unhappy by the mall's services. These can be the prime targets of the mall, as they have the potential to spend money.

7. CONCLUSION

- So, in conclusion the mall authorities will try to add new facilities so that they can attract these people and can meet their needs.we may deduce that to increase the profits of the mall, the mall authorities should target people belonging to cluster 3 and cluster 5 and should also maintain its standards to keep the people belonging to cluster 1 and cluster 2 for profits.

- Having a better understanding of the customers segments, a company could make better and more informed decisions. An example, there are customers with high annual income but low spending score. A more strategic and targeted marketing approach could lift their interest and make them become higher spenders. The focus should also be on the "loyal" customers and maintain their satisfaction.

- We have thus seen, how we could arrive at meaningful insights and recommendations by using clustering algorithms to generate customer segments. For the sake of simplicity, the dataset used only 2 variables— **Annual income and Annual spending($)**.

- In a typical business scenario, there could be several variables which could possibly generate much more realistic and extreme business-specific insights.

REFERENCES:

- *hr@exposysdata.com*
- *https://www.geeksforgeeks.org/supervised-unsupervised-learning/*
- *https://www.geeksforgeeks.org/k-means-clustering-introduction/?ref=lbp*
- *http://www.mit.edu/~9.54/fall14/slides/Class13.pdf*
- *https://en.wikipedia.org/wiki/Unsupervised_learning*
- *https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/*

THANK YOU