Mall Customers Clustering Analysis

Installing the Libraries

In [1]:

```
# for basic mathematics operation
import numpy as np

# for dataframe manipulations
import pandas as pd

# for Data Visualizations
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
plt.style.use('fivethirtyeight')
import plotly
import plotly.express as px

# For Statistics
import statistics
```

Examining Data

In [2]:

```
# importing the dataset
data = pd.read_csv('Customers.csv')
```

In [3]:

```
# lets check the shape of the dataset
print("Shape of the dataset :", data.shape)
```

```
Shape of the dataset : (200, 5)
```

In [4]:

```
# lets check the head of the data
data.head()
```

Out[4]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	34	Male	18.0	33	92
1	66	Male	18.0	48	59
2	92	Male	18.0	59	41
3	115	Female	18.0	65	48
4	1	Male	19.0	15	39

In [5]:

```
# lets check the tail of the data
data.tail()
```

Out[5]:

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
195	91	Female	68.0	59	55
196	109	Male	68.0	63	43
197	58	Male	69.0	44	46
198	61	Male	70.0	46	56
199	71	Male	70.0	49	55

Descriptive Statistics

In [6]:

```
# describing the data
data.describe()
```

Out[6]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
count	200.000000	196.000000	200.000000	200.000000
mean	100.500000	38.734694	60.560000	50.200000
std	57.879185	13.964094	26.264721	25.823522
min	1.000000	18.000000	15.000000	1.000000
25%	50.750000	28.000000	41.500000	34.750000
50%	100.500000	35.500000	61.500000	50.000000
75%	150.250000	49.000000	78.000000	73.000000
max	200.000000	70.000000	137.000000	99.000000

In [7]:

```
data.isnull().sum()
```

Out[7]:

```
CustomerID              0
Gender                  3
Age                     4
Annual Income (k$)      0
Spending Score (1-100)  0
dtype: int64
```

Two types of variable missing

Gender(Categorical varibles) used Mode to replace null values.

Age() used Mean to replace null values.

In [8]:

```
# fill mode value of age at the place of missing values(Female)
data["Gender"].fillna(statistics.mode(data["Gender"]), inplace = True)
```

In [9]:

```
# fill Mean value of age at the place of missing values(38)
data["Age"].fillna(np.mean(data["Age"]), inplace = True)
```

In [10]:

```
data.isnull().sum()
```

Out[10]:

```
CustomerID                0
Gender                    0
Age                       0
Annual Income (k$)        0
Spending Score (1-100)    0
dtype: int64
```

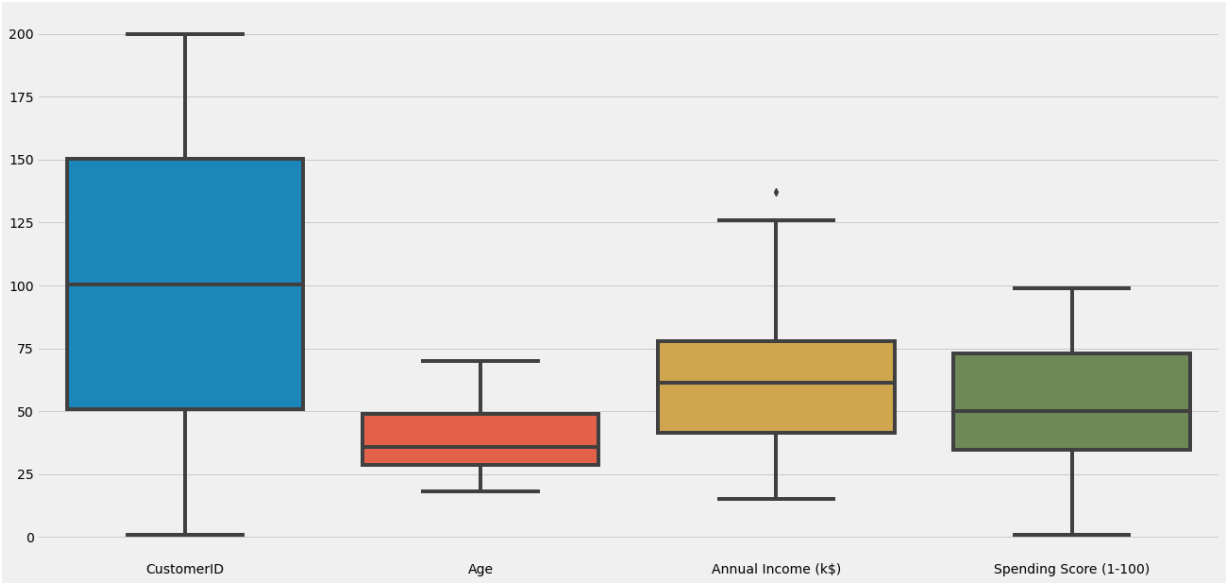
Outliers

In [11]:

```
# Finding outliers using Box Plot
plt.figure(figsize=(20,10))
sns.boxplot(data = data)
```

Out[11]:

<AxesSubplot:>



Correlation

In [12]:

```
data.corr().style.background_gradient(axis=None)
```

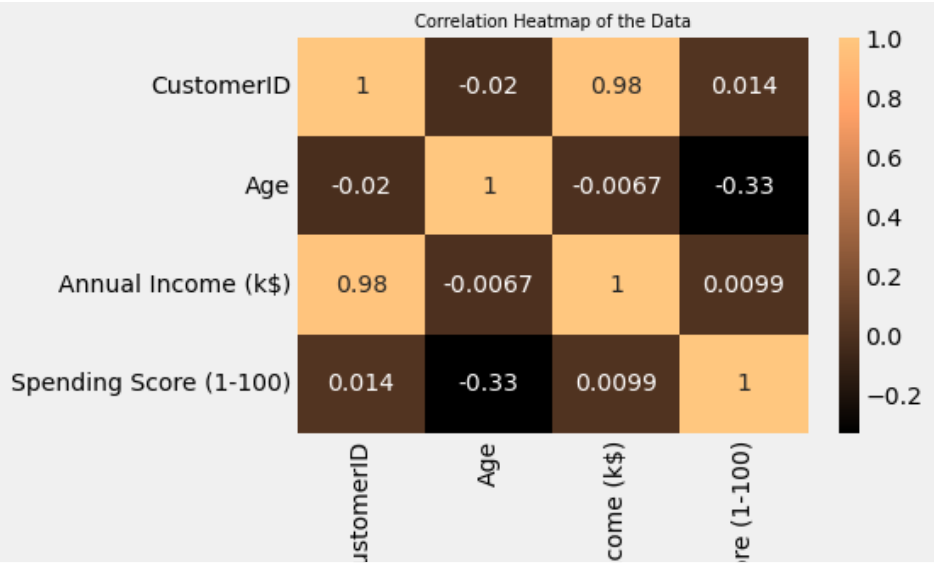
Out[12]:

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)
CustomerID	1.000000	-0.020136	0.977548	0.013835
Age	-0.020136	1.000000	-0.006708	-0.332353
Annual Income (k$)	0.977548	-0.006708	1.000000	0.009903
Spending Score (1-100)	0.013835	-0.332353	0.009903	1.000000

In [13]:

```
# lets check the Correlation Heat Map of the Data

sns.heatmap(data.corr(), annot = True,  cmap = 'copper')
plt.title('Correlation Heatmap of the Data', fontsize = 10)
plt.show()
```



The Above Graph for Showing the correlation between the different attributes of the Mall Customer Segementation Dataset, This Heat map reflects the most correlated features with dark Orange Color and least correlated features with yellow color.

> 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.

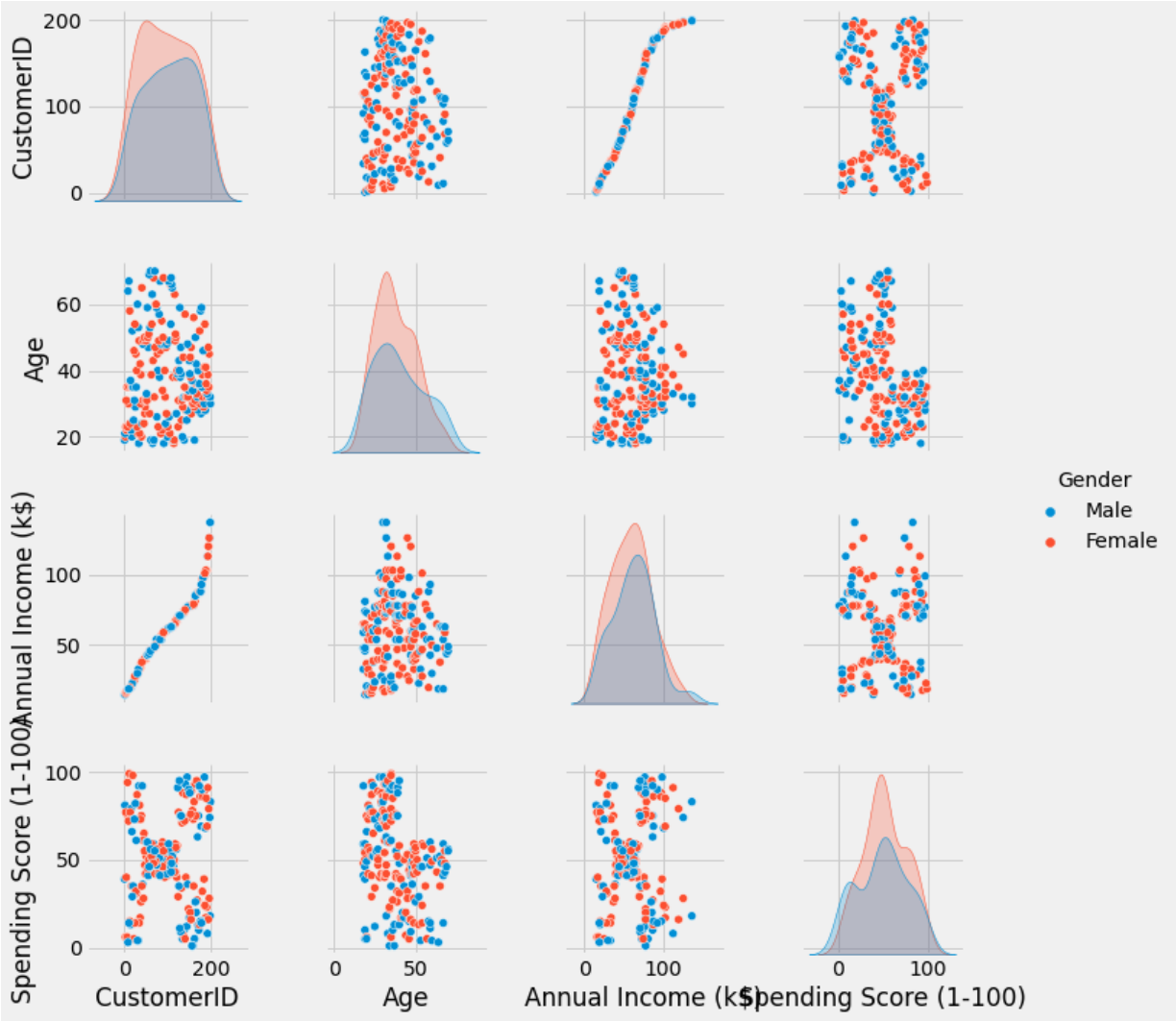
Data visualization

In [14]:

```
1 sns.pairplot(data, hue = "Gender")
```

Out[14]:

```
<seaborn.axisgrid.PairGrid at 0x1a07a9e13c0>
```

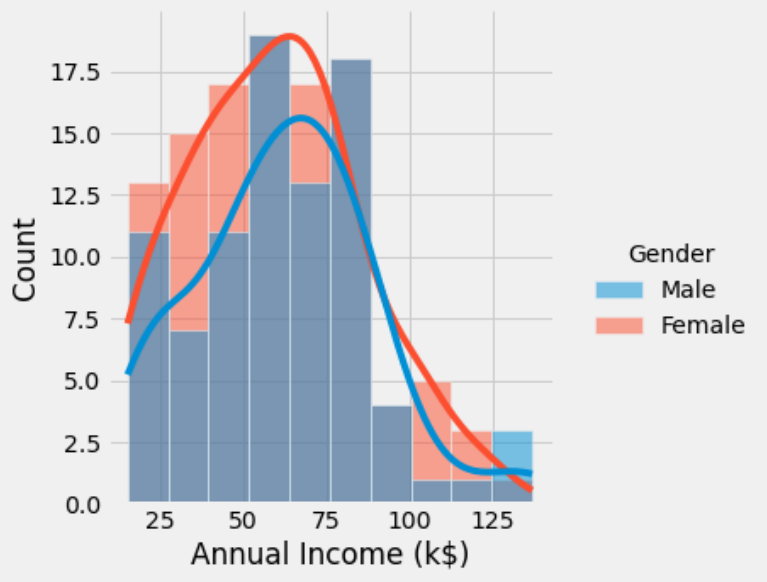


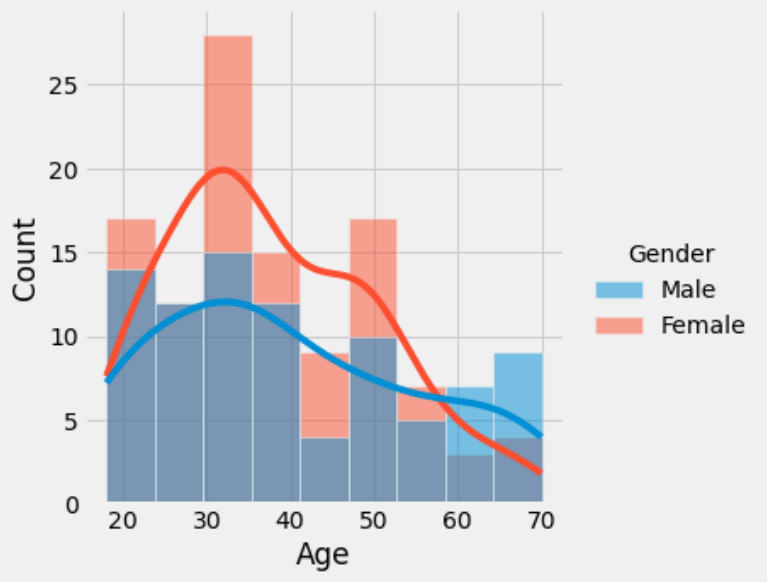
In [15]:

```
sns.displot(x = "Annual Income (k$)", data = data, hue = "Gender", kde = True)
sns.displot(x = "Age", data = data, hue = "Gender", kde = True)
```

Out[15]:

```
<seaborn.axisgrid.FacetGrid at 0x1a07d6b3220>
```





Here, 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 umber as compared to the Middle aged people.

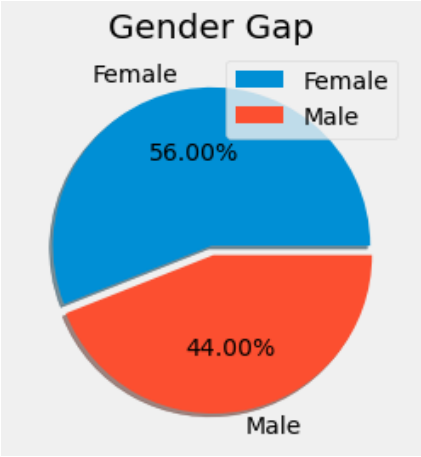
In [16]:

```
labels = ['Female', 'Male']
size = data['Gender'].value_counts()
explode = [0, 0.06]

plt.pie(size, explode = explode, labels = labels, shadow = True, autopct = '%0.2f%%')
plt.title('Gender Gap', fontsize = 20)
plt.legend()
```

Out[16]:

```
<matplotlib.legend.Legend at 0x1a07d761690>
```



By looking at the above pie chart which explains about the distribution of Gender in the Mall

> Interestingly, The 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.

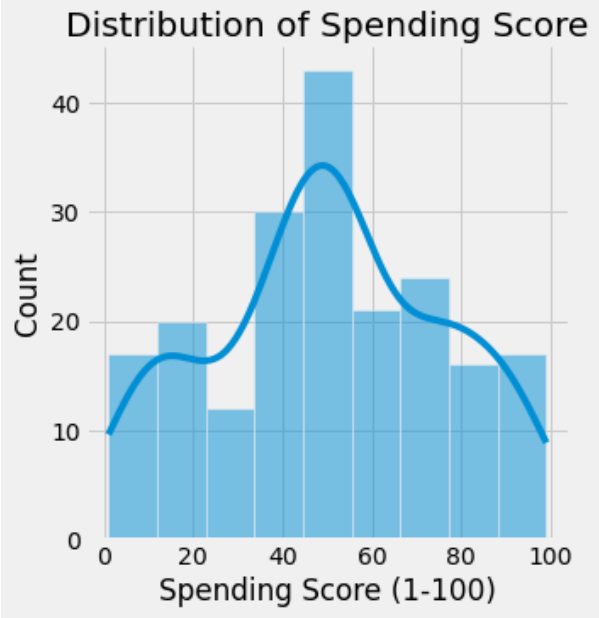
In [17]:

```
# lets check the distribution of Spending Score

sns.displot(data['Spending Score (1-100)'], kde = True)
plt.title("Distribution of Spending Score")
```

Out[17]:

```
Text(0.5, 1.0, 'Distribution of Spending Score')
```



This is the Most Important Chart in the perspective of Mall, as It is very Important to have some intuition and idea about the Spending Score of the Customers Visiting the Mall.

> On a general level, we may conclude that most of the Customers have their Spending Score in the range of 40-60. 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.

In [18]:

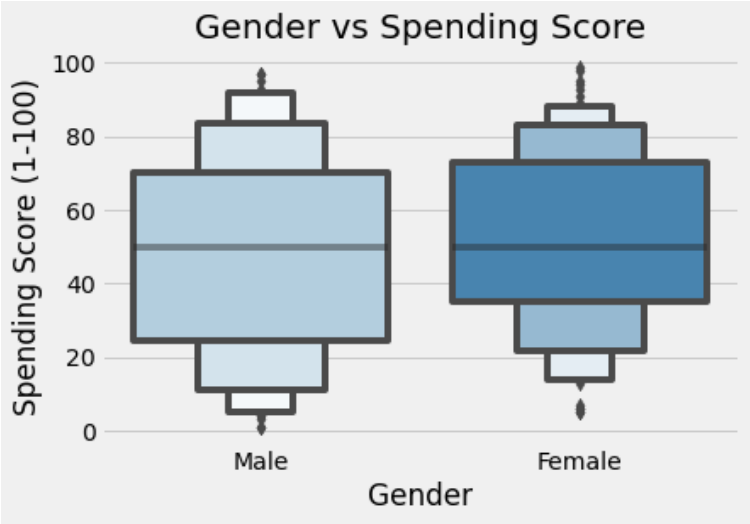
```
#  Gender vs Spendscore

sns.boxenplot(data['Gender'], data['Spending Score (1-100)'], palette = 'Blues')
plt.title('Gender vs Spending Score', fontsize = 20)
```

```
C:\Users\SANKET ARGADE\AppData\Local\Programs\Python\Python310\lib\site-packages\seaborn\_dec
orators.py:36: FutureWarning: Pass the following variables as keyword args: x, y. From versio
n 0.12, the only valid positional argument will be `data`, and passing other arguments withou
t an explicit keyword will result in an error or misinterpretation.
  warnings.warn(
```

Out[18]:

```
Text(0.5, 1.0, 'Gender vs Spending Score')
```



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.

Clustering Analysis

In [19]:

```
# we want to perform clusters of Customers who share similar behaviour for that lets select the columns
# Spending score, Annual Income, and Age

# Lets select the Spending score, Annual Income and Age Columns from the Data
cluster_table = data.iloc[:, 2:]

print(cluster_table.shape)
```

```
(200, 3)
```

In [20]:

```
1 cluster_table
```

Out[20]:

	Age	Annual Income (k$)	Spending Score (1-100)
0	18.0	33	92
1	18.0	48	59
2	18.0	59	41
3	18.0	65	48
4	19.0	15	39
...	...	...	...
195	68.0	59	55
196	68.0	63	43
197	69.0	44	46
198	70.0	46	56
199	70.0	49	55

200 rows × 3 columns

Kmeans Algorithm

The Elbow Method to find the No. of Optimal Clusters

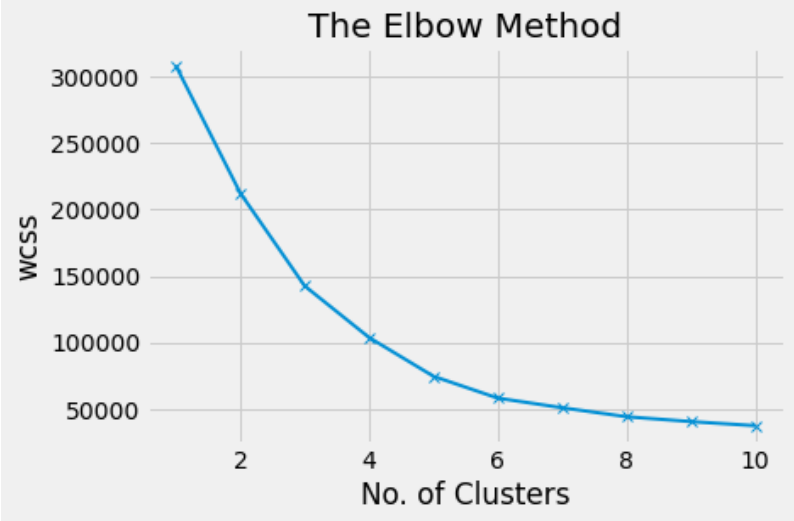
In [21]:

```
from sklearn.cluster import KMeans
wcss = []
for i in range(1, 11):
    km = KMeans(i)
    km.fit(cluster_table)
    wcss.append(km.inertia_)            # inertia_: Sum of squared distances of samples to their closest
                                        #           (wcss value)

plt.plot(range(1, 11), wcss, "x-", linewidth = 2)
plt.title('The Elbow Method', fontsize = 20)
plt.xlabel('No. of Clusters')
plt.ylabel('wcss')
```

Out[21]:

```
Text(0, 0.5, 'wcss')
```



In [22]:

```
pd.DataFrame({"No. of Clusters":np.arange(1,11),"wcss":wcss})
```

Out[22]:

	No. of Clusters	wcss
0	1	308005.484082
1	2	212393.290026
2	3	142611.736870
3	4	103874.215066
4	5	74743.137432
5	6	58322.868797
6	7	51103.333731
7	8	44355.194258
8	9	40710.939620
9	10	37634.839314

| As the number of clusters increase, the WCSS decreases.

In [23]:

```
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score

scorelist = []
for i in range (2,11):
    model = KMeans(i)
    resultlist = model.fit_predict(cluster_table)
    score = silhouette_score(cluster_table,resultlist)
    scorelist.append(score)

print(pd.DataFrame({"Clusters":np.arange(2,11),"silhouette_score":scorelist}))
K_number = scorelist.index(max(scorelist))+2

print("Final K: ",K_number)
```

```
   Clusters  silhouette_score
0         2          0.291942
1         3          0.385101
2         4          0.405384
3         5          0.445329
4         6          0.448469
5         7          0.434612
6         8          0.428719
7         9          0.392438
8        10          0.385472
Final K:  6
```

In [24]:

```
finalmodel = KMeans(K_number)                        # k_number = 6 clusters
resultlist = finalmodel.fit_predict(cluster_table)

cluster_table = pd.concat([cluster_table,pd.Series(resultlist, name = "cluster")], axis = 1)
cluster_table
```

Out[24]:

	Age	Annual Income (k$)	Spending Score (1-100)	cluster
0	18.0	33	92	2
1	18.0	48	59	4
2	18.0	59	41	4
3	18.0	65	48	4
4	19.0	15	39	5
...	...	...	...	...
195	68.0	59	55	3
196	68.0	63	43	3
197	69.0	44	46	3
198	70.0	46	56	3
199	70.0	49	55	3

200 rows × 4 columns

In [25]:

```
cluster_mean = cluster_table.groupby(cluster_table["cluster"]).mean()
cluster_mean
```

Out[25]:

cluster	Age	Annual Income (k$)	Spending Score (1-100)
0	41.735277	88.228571	17.285714
1	32.711146	86.538462	82.128205
2	25.272727	25.727273	79.363636
3	55.909091	53.522727	49.022727
4	27.300890	56.410256	49.205128
5	44.273081	25.142857	19.523810

Visualizaing the Clusters

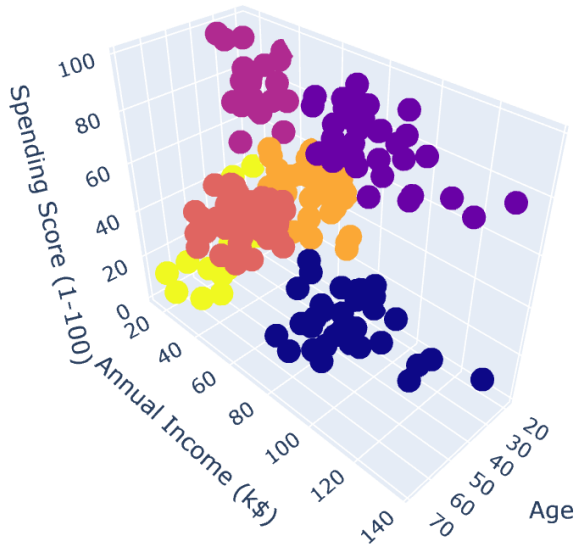
In [26]:

```
px.bar(cluster_mean,barmode="group",title="Mean of customer data by cluster")
```

In [31]:

```
import plotly.express as px

px.scatter_3d(cluster_table,x='Age', y='Annual Income (k$)', z='Spending Score (1-100)',color='cluster')
```



This Clustering Analysis gives us a very clear insight about the different segments of the customers in the Mall. There are clearly Six segments of Customers namely as there color code.

> Pink, Yellow, Blue, Orange, Red and Violet Based on their Annual Income, Spending Score and Age which are reportedly the best factors/attributes to determine the segments of a customer in a Mall.

🟣 CLUSTER PURPLE – TARGET CUSTOMERS: Earning high and also spending high Target Customers. Annual Income High as well as Spending Score is high, so a target consumer. These people might be the regular customers of the mall and are convinced by the mall's facilities.

🔵CLUSTER BLUE – PITCH PENNY CUSTOMERS: Earning high and spending less.These can be the prime targets of the mall, as they have the potential to spend money. So, the mall authorities will try to add new facilities so that they can attract these people and can meet their need

🟡Cluster Yellow – less spenders: They earn less and spend less. We can see people have low annual income and low spending scores, 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.

⚫Cluster Pink - SPENDERS: This type of customers earns less but spends more Annual Income is less but spending high, so can also be treated as potential target customer. The shops/malls might not target these people that effectively but still will not lose them.

🟠CLUSTER Orange – NORMAL CUSTOMERS(Under fourty): Customers are average in terms of earning and spending, 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.

🔴CLUSTER RED – NORMAL CUSTOMERS(Above fourty): Customers are average in terms of earning and spending, 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.

Outcome

- It will help to Prioritized advertising boardings and digital signages.
- It will help to developing in-mall infrastructures.
- To Prioritized the customer segment who belongs from cluster number 1 and 2.
- And discuss different marketing strategies and policies to attract the cutomers who belongs from cluster number 0, 2, and 3.