**Q1**

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

**Q2**

* Clustering on two variables (**Annual income** and **spending score**)

Chart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generated

* Clustering on two variables (**Age** and **spending score**)

Chart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generated

* Clustering on two variables (**Age** and **Annual Income**)

Chart, bubble chart

Description automatically generated

Chart, bubble chart

Description automatically generated

**Q3**

Chart, line chart

Description automatically generated

To find which setting (K=3 or K=5) provides better result, we could find the optimal number of clusters by using the Elbow Method. We want to find a value of k such that the Total within sum of square is low. By visualizing the plot, we observe that K=5 would be an appropriate number of clusters as this value bended the plot. K=5 has smaller total WSS compared to k=3. The slope of the line dramatically decreases from 1 to 4. However, we can see that the slope decreases consistently after K=5.

We can compare the graphs in Q2. When k=3, we can see some overlapping between clusters. When k=5, we can clearly see some better separation between clusters compared to k=3. K=5 is better to group similar data points into the same group.

Hence, it suggests that 5 clusters would generate better result.

**Q4**

Cluster plot:

**Annual income** and **spending score**

Chart, bubble chart

Description automatically generated

We can see that the mall customers can be grouped into 5 groups. We focus on two variables or attributes (Annual Income and spending score) to run the k-means clustering. All clusters seem to be well-separated when we have 5 clusters, although there are some tiny portions of overlapping in the middle part. We could also see two outliers on cluster 4 (point 199) and cluster 5 (point 200). We can also see that the clusters represent the following 5 customers’ groups:

The 1st cluster indicates customers with high annual income and low spending score. It describes the people with high annual income prefer to buy less.

The 2nd cluster indicates customers with high annual income and high spending score. It describes the people with high annual income prefer to buy more.

The 3rd cluster indicates customers with low annual income and high spending score. It describes the people with low annual income prefer to buy more. The size of this group is larger compared to all others. Large portions of customers fall into this group.

The 4th cluster indicates customers with medium annual income and medium spending score. It describes the people with average annual income have average spending score.

The 5th cluster indicates customers with low annual income and low spending score. It describes the people with low annual income prefer to buy less.

**Age** and **spending score:**

Chart, bubble chart

Description automatically generated

All clusters seem to be well-separated when we have 5 clusters, although there are some small portions of overlapping between clusters. We can also see that the clusters represent the following 5 customers’ groups:

The 1st cluster shows all different ages (from younger to older people) with the lowest spending score in the group.

The 2nd cluster indicates that this younger customer group have the highest spending score. This group tends to have best shopping behavior.

The middle parts (cluster 3,4 and 5) indicates that different age groups have the medium or average spending scores.

**Age** and **Annual Income:**

Chart, bubble chart

Description automatically generated

All clusters seem to be fairly separated. We can see some overlapping between clusters. We can also see that the clusters represent the following 5 customers’ groups:

The 1st cluster indicates that the average age (around 40) group has the highest annual income compared with all other groups.

The 2nd cluster shows younger people with average annual income. The size of this group is relatively higher than other groups.

The 3rd cluster shows younger people with low annual income.

The 4th cluster shows older people with average annual income.

The 5th cluster shows older people with low annual income.

We can see the density of overall plot. The average age group is more likely to have a higher annual income. Younger or average age groups tends to have higher annual income. Older people tend to have medium annual income or below the average.

Businesses could implement different marketing strategies to desired target groups to increase their profit. Businesses could apply k-means clustering to make more informed decisions by having better understanding of these specified customers’ groups.