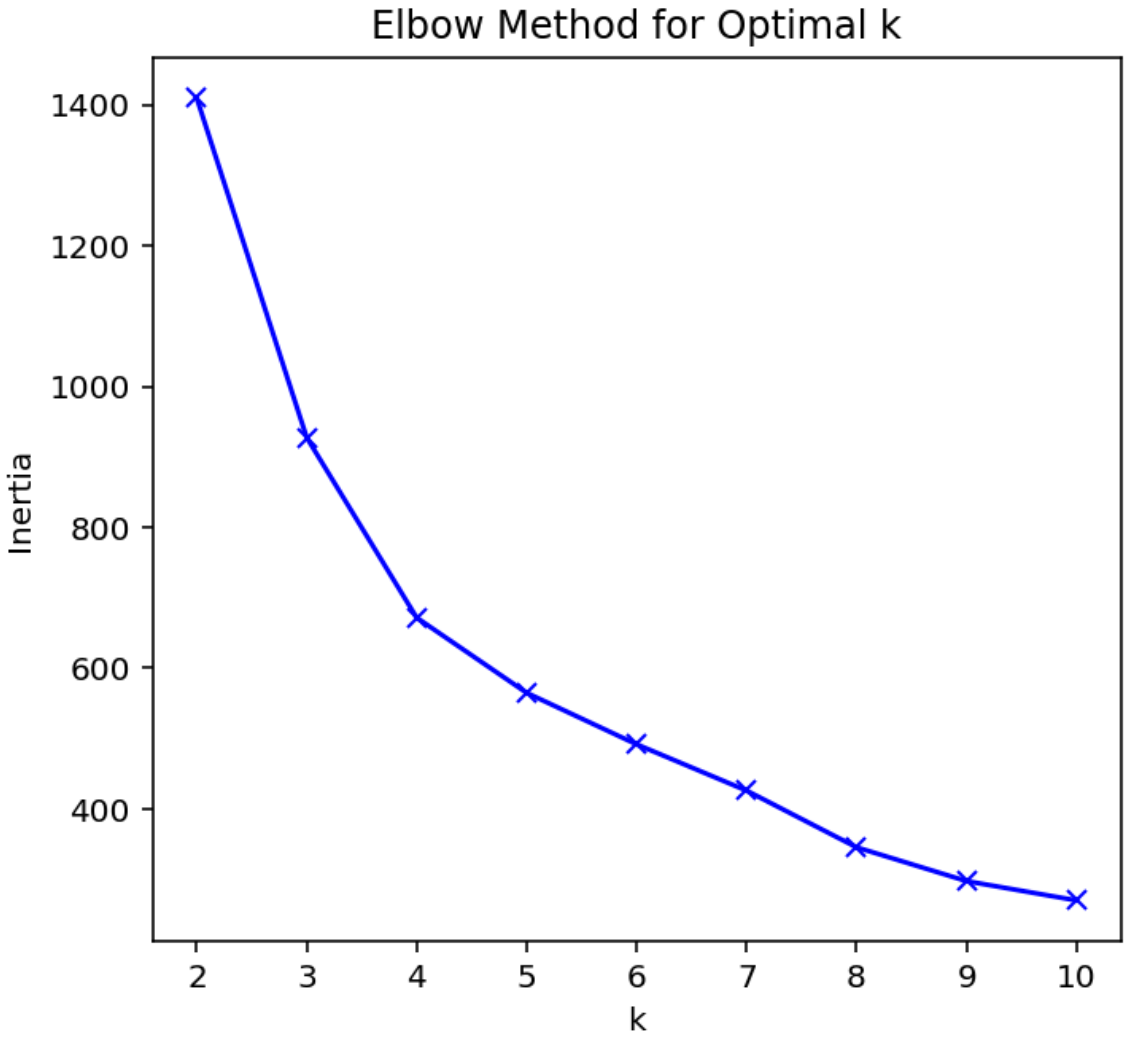
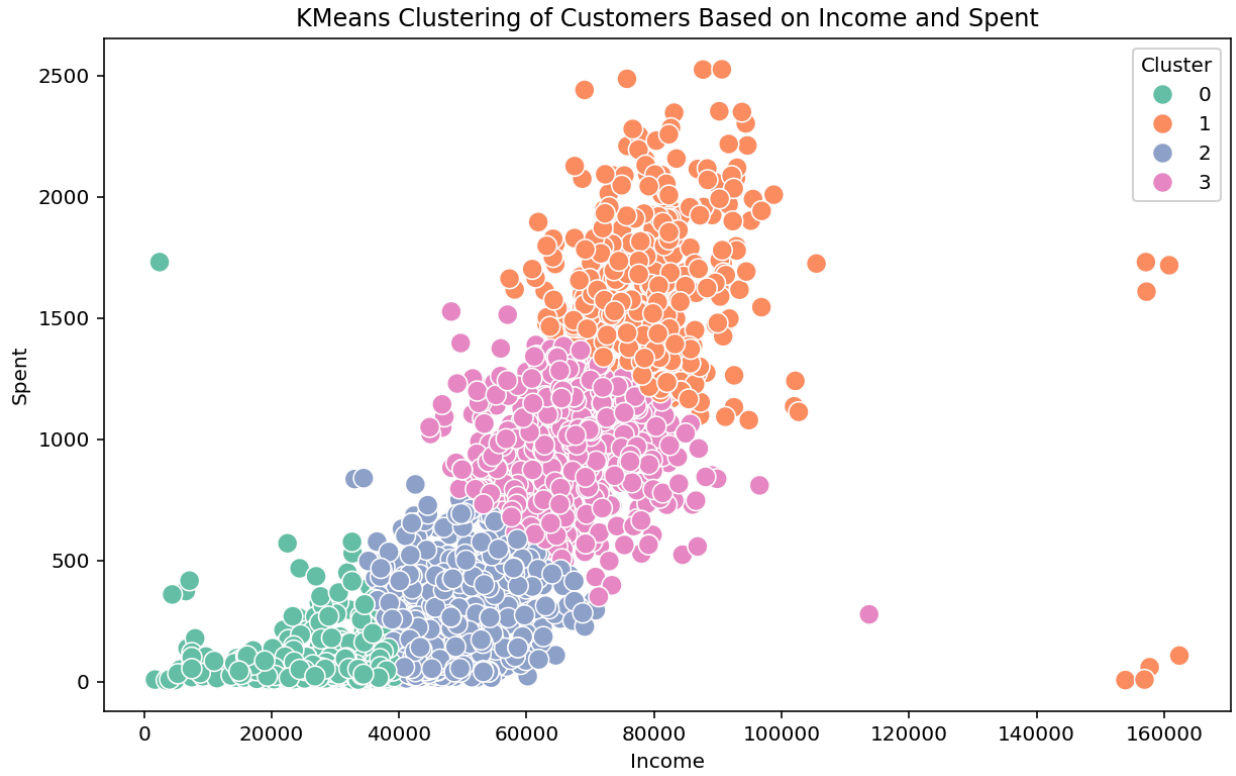
Machine Learning Dataset Analysis

The dataset that I chose is a customer marketing campaign which contains information about customers ranging from their date of birth, income and shows their spending habits. It helps businesses better understand their customers and makes it easier for them to modify products according to specific needs and behaviours. Our main goal is to cluster the data with respect to their income and expenditure (1).

So for this dataset at first we calculated the total expenditure of them in various products and created a new column based on that. We also removed some outliers from the data such as income over 600000. Then we scaled the data to normalize which is both the income and spent column. The clustering methodology that we decided to use is Kmeans clustering which is an unsupervised learning algorithm. So before clustering the data we created an elbow plot to find out the optimum number of clusters required to classify the data.



From the above plot it is clear that the optimal number of clusters is 4. So now we utilize the clustering algorithm and create a scatter plot with income in X axis and Spent in Y axis and we get this:



Results:

The data is divided into four clusters based on Income and Spent:

- **Cluster 0 (green)**: Represents customers with lower income and lower spending.
- **Cluster 1 (orange)**: Represents customers with higher income and higher spending.
- **Cluster 2 (blue)**: Represents customers with moderate income and moderate spending.
- **Cluster 3 (purple/pink)**: Represents customers with moderate-to-high income and spending.

So from here we can see that there is a moderate correlation between income and spending. Generally as customers' income increases their total spending also increases.
Each cluster here roughly follows a diagonal pattern which also indicates high earners tend to spend more.
The plot also has some outlier in Cluster 0 (green) where a customer with very low income has high expenditure. This could indicate external financial support.

Based on these clusters, the business can tailor its marketing and sales strategies:

- **Cluster 1 (orange)**: High-income, high-spending customers could be targeted with premium products and personalized offers.
- **Cluster 0 (green)**: Low-income, low-spending customers may prefer budget-friendly options and discount promotions.
- **Cluster 2 and Cluster 3 (blue and pink)**: These middle segments could be offered mid-tier products and bundled deals to encourage increased spending.

Moreover, Cluster 3 (pink) might represent customers who are on the verge of transitioning into higher spending. So they could be moved to that category by providing loyalty programs and promotions.

So to conclude it could be said that the clustering provides few insights on how to take actions based on the customer category. We can see the positive trend between income and spending is visible from the scatter plot, confirming that higher-income customers tend to spend more. The presence of few outliers highlights potential areas for further business exploration or specialized marketing. The separation of customers into these groups provides a framework for creating differentiated marketing strategies aimed at maximizing engagement and revenue for each segment.

References:
https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis