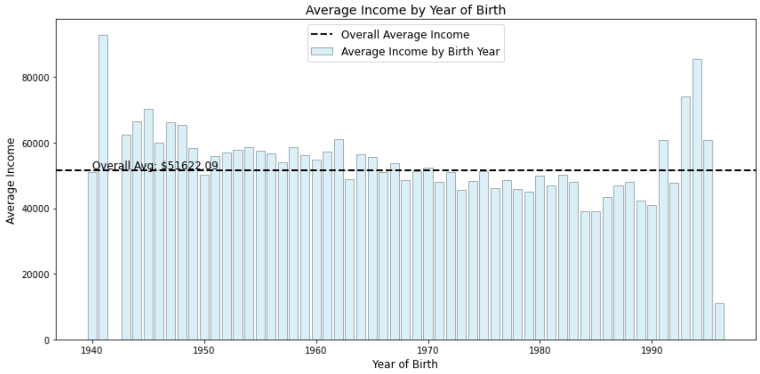
Clustering Assignment

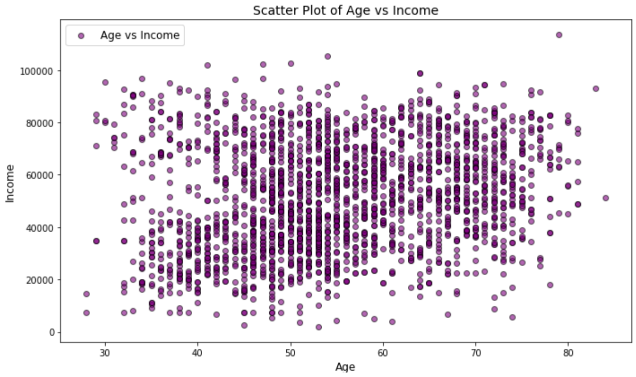
Name: Shanmukha Karthik Gundlapalli Subramanyam Student ID: 23121764

To find different segments, this clustering assignment employs a dataset emphasising on consumer behaviour and preferences. It offers information on the consumer's personal past by including several demographic and transactional elements like age (by birth year), education level, marital status, and income. With information on the amount spent across many product categories—e.g., wines, fruits, meats—and purchase channels—e.g., internet, physical shops, catalogues—the dataset also provides thorough buying behaviours. Customer involvement is also recorded using factors such recency, frequency of online visits, and marketing campaign participation. Because it helps companies classify consumers according on behaviour, tastes, and reactions to prior marketing campaigns, this comprehensive collection of characteristics is perfect for clustering. By means of segment analysis, businesses may create more focused marketing plans, thereby optimising resource allocation and raising the success of their efforts.

I have performed various data preprocessing tasks before creating the visualizations. Using df.isnull(), first I looked for any missing values in the dataset.sum() and, using df.dropna(), discarded any rows with missing data. Since they didn't provide any helpful information for this study, I then deleted columns such "Response," "Z_Cost Contact," and "Z_Revenue." Using the interquartile range (IQR), I additionally eliminated outliers from the "Income" and "Year_ Birth" columns to guarantee data integrity. This approach removes outliers that can skew the results.

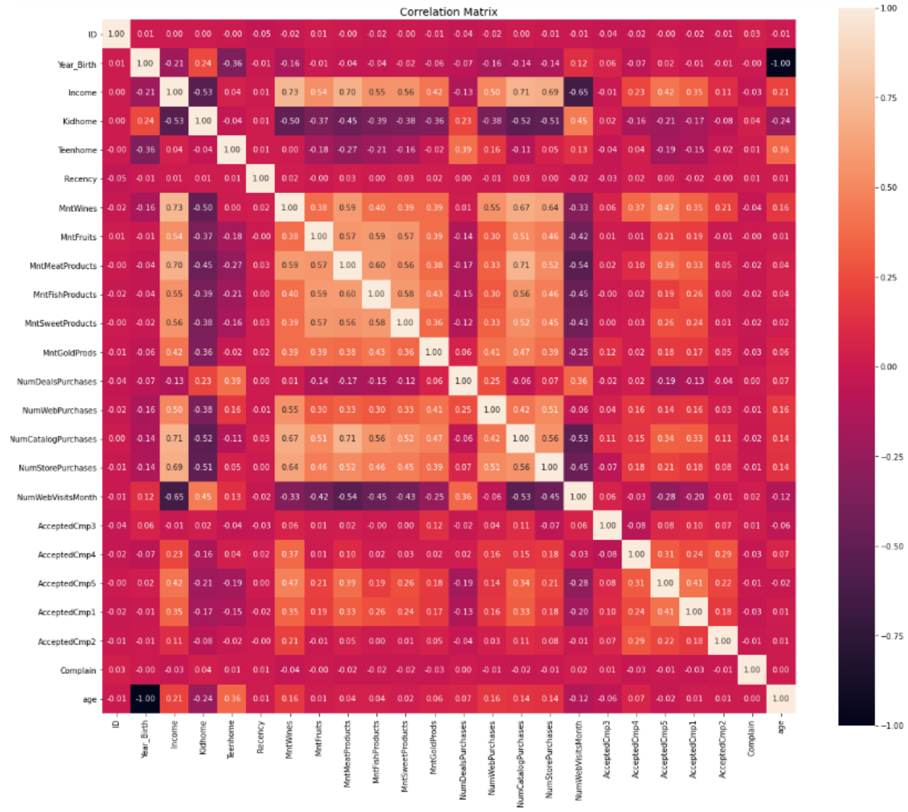


From the customer segmentation statistics, I have visualized the average income by year of birth. Two main elements define the plot: a dashed black line showing the general average income across all the records in the dataset and bars denoting the average income for every year of birth. With varied heights for every bar, the plot itself depicts the average salary for every year of birth. About $51,622.09, the dashed line shows the general average income across all of the clients. From the plot, it is evident that certain years—those born in the 1940s and 1990s—have very high average wages while other years show more steady or declining values. Based on consumer birth years, this lets me spot patterns and anomalies that could direct more tailored marketing plans.

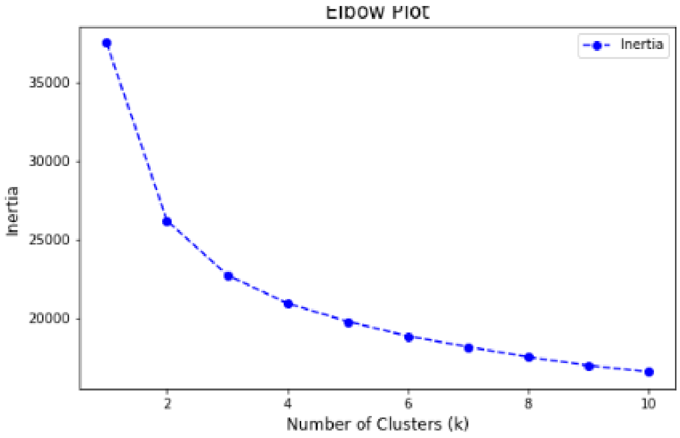


I have seen from my customer dataset the association between age and income in this scatter plot. I initially computed the age of every client by subtracting their birth year—from the "Year_Birth" column—from the current year, 2024. Income from the "Income" column is plotted on the y-axis; the data points on the x-axis were then displayed using this new "age" column. Every customer is shown in the plot as a dot; the position on the x-axis indicates their age and the position on the y-axis their income. While the black boundaries provide clarity surrounding every data point, the purple colour and the semi-transparent dots (alpha = 0.6) let one better examine overlapping points. Based on the graphic, income seems to have a weak, unclear trend with age; dispersion across all age groups is somewhat evident. Nonetheless, I can see that the vast spectrum of income distribution among consumers of various ages indicates that both younger and older ones have different income levels.

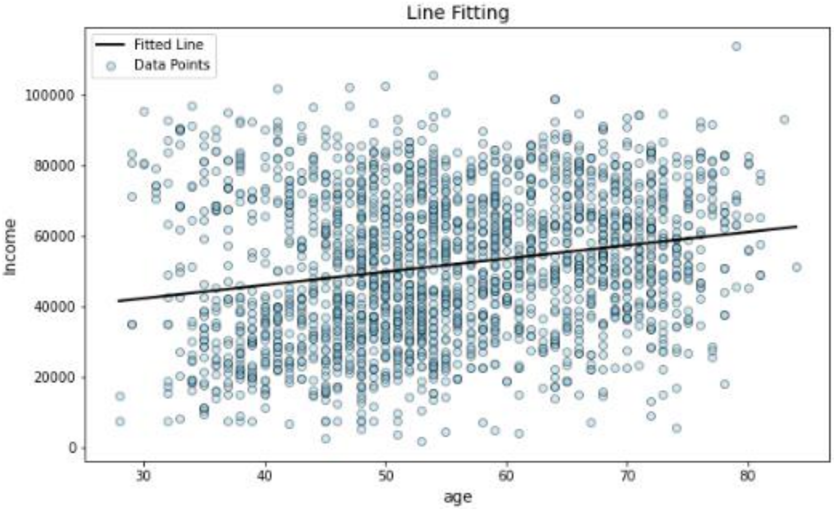
Further, I saw the interactions across many aspects of my client dataset in this correlation matrix. Ranging from -1 to +1, the values in the matrix provide Pearson correlation coefficients, therefore demonstrating the strength and direction of linear correlations between pairs of variables. With a strong positive correlation of 0.59, "Mnt Meat Products" and "Mnt Fish Products" for instance indicate that consumers who spend on one are probably going to spend on the other. On the other hand, "Year_Birth" and "Income" show a correlation of 0.56, suggesting that older consumers often have greater earnings. This matrix enables the identification of correlations between characteristics that could direct more research or client segmentation.



Further, Standard Scaler let me scale the data to normalise it so it would be ready for K-means clustering. I computed the inertia for clusters ranging in count from 1 to 10 and displayed them to show the "elbow," or point of declining returns. The Elbow Plot shows how inertia falls off as cluster count rises. From the plot, the elbow is clear at k=3, indicating that three clusters best balance simplicity and compactness. This led me to decide on k=3 and use the K-means technique to assign each data point to a cluster, therefore adding a "Cluster" column into the dataset for further examination.



The scatter figure shows clear segmentation of the clusters I found depending on age and wealth. Cluster 2 consists of people with high salaries across many age groups, therefore demonstrating their buying power and potential as valuable consumers. While Cluster 1 includes low-income people, Cluster 0 stands for moderate-income people with mainly middle tier income values. The way the dataset is segmented by the clustering allows focused study. To understand how these two factors affect consumer segmentation, I set age and income as the axes. This method clarifies the link between financial capacity and age in the framework of consumer behaviour, therefore laying groundwork for customised marketing plans or unique product offers.



With a linear regression model, the plot shows how age and income relate. Examining if a linear trend exists between these two variables, I fit a straight line to the data points. Every blue scatter point stands for a single data point; the black line is the fitted regression line, therefore projecting income depending on age. Although the line exhibits a little upward slope, demonstrating a positive relationship between age and income, the dispersion of points refers to somewhat high variability in income that age by itself cannot explain. This stage allows me to assess the degree of age effect on income, therefore revealing the structure of the dataset and if other factors should be taken into account for more strong modelling.