SI 370 GROUP PROJECT: FINAL REPORT

"Product Analysis based on Consumer Behavior Preferences"

Group Members: Anthony Ho, Ashley Anderson, Nina Yang, Tiara Amadia, Jason Liang

Statement of Purpose

Objective: This project aims to predict, with the highest accuracy rate achieved, the highest spending consumer demographic and other behavioral patterns they have in common (ie. education level, number of children, etc.) We will use an external Kaggle dataset depicting different customers and their purchasing preferences, as well as biographical and background information.

Significance: Understanding spending patterns is essential for a company's sales and marketing efforts. Narrowing in on higher spending demographics gives companies their biggest chance at targeting potential customers that might spend a lot of money on their products, thus increasing profits. In this dataset specifically, knowing which products within a grocery store customers spend more on allows owners to change where they store their products and how to promote them for maximum effect.

Data Description: Our dataset is divided into 4 main categories: People, Products, Promotion, and Place. "People" looks at a customer's varying biographical information. "Products" look at their purchase behavior. "Promotion" looks at the number of purchases made with a discount. "Place" refers to purchase history. We have attached a table of all column names (Appendix A).

Hypothesis: We predict that the highest spending consumer type will be married families with 0 or 1 child or teen in the home. Households with fewer kids typically have more disposable income and will be used to a certain standard of living. We hypothesize that they will spend most of their groceries on meat products, averaging around $250/month. Couples with less children also have more money set aside for themselves, so we also hypothesize that they will be in the 90th percentile of wine spending (>$600/month).

Use of Analytic Techniques

Data Cleaning: When exploring different aspects of the data, it was necessary to look into data points that were not filled in our dataset. For this, we cleaned the data by filling them in with averages of the column. Additionally, we reconfigured data that would be hard to process to something that would be easier to visualize. For example, we modified the date of the customer's enrollment to simply show how many years the customer has been enrolled. Changes like this are important since it can simplify how we can present data.

Exploratory Data Analysis: To explore patterns in the data and also help determine the direction of the project, we started off by making simple plots to observe trends in specific background information variables in the dataset, such as “Education Level” or “Income”. Afterward, we looked into whether any of these variables correlated with each other. It is important to take a deep dive into the background information as these serve as explanatory variables that influence their spending patterns. Finally, further analysis was conducted to study the breakdown of spending, based on the types of goods (i.e. fruits, meat, wine, etc.) and the method of purchasing (i.e. Number of web purchases, number of store purchases, etc.) to help identify any further patterns.

Machine Learning: Our goal in this project is to create a model that can accurately predict whether a customer is a “high-value customer”. The necessary steps for creating a machine learning model include training and testing data. In our case, we decided to do an 80-20 split, meaning we used 80% of our data to train our machine learning model and 20% to test the accuracy of our model. Since our data had both numerical and categorical data, we used a pipeline that used “StandardScaler” and “OneHotEncoder”, with “StandardScaler” standardizing the numerical data and “OneHotEncoder” encoding the categorical data. After this, we used different methods of fitting our data including a simple linear regression model and different classifiers. Finally, we tested our data’s accuracy using root-mean-square deviation and classifier scores.

Clustering: While investigating the data, it is important to look at the grouping of subjects in the dataset to determine the similarity and dissimilarity of those in in-groups and out-groups respectively. More specifically, we used the K-means clustering to better understand how many clusters were present in our dataset and the spread of the data. In addition, the PCA was used to look at customers in relation to whether they were high-value or low-value spenders. The data was scaled in both cases and pipelines were used in the K-means analysis.

Use of Visualization Techniques

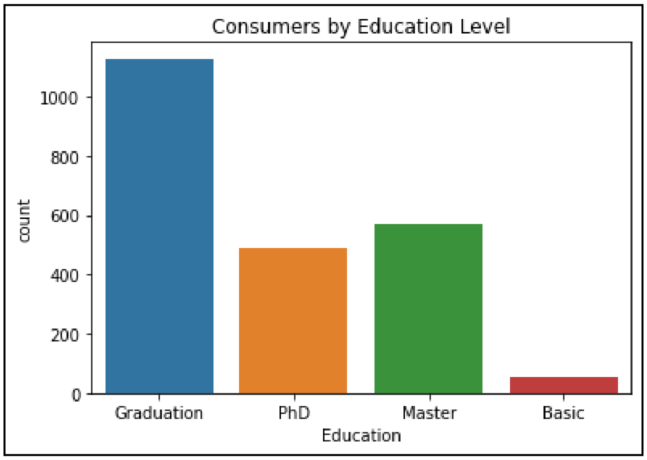
Data Cleaning: During the data cleaning process, we plotted several quick graphs (i.e. histograms) to look for any anomalies that are not representative of the rest of the data. For example, through investigating the “Income” column, it was found that one customer had an income significantly higher than the rest of the customers and we decided to remove them from the dataset.

Exploratory Data Analysis: To show the distribution of customers for quantitative variables, histograms were used (i.e. distribution of income). For qualitative variables, such as education level, count plots were used to count the frequency. Next, using a heatmap, we can investigate whether the background information and whether some of the variables correlated with each other and had an influence on the amount spent at the store. Finally, stacked bar graphs were used to show the breakdown of spending by type of goods and method of purchase, to better visually display the proportions.

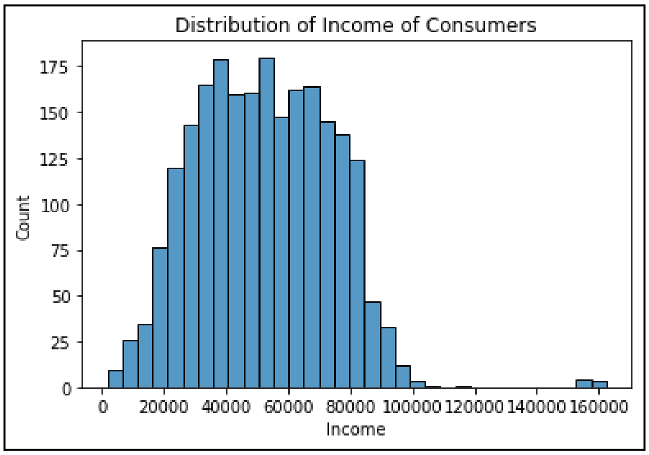
Clustering: To determine the clusters in the dataset, K-means clustering and PCA were both performed. For the K-means clustering, a silhouette analysis was generated with two plots, a silhouette plot and a visualization of the clustered data. For PCA, a visualization of the clustered data was generated with data points shaded based on whether the customer was high value or low value. Cluster plots allow us to better discern the distribution and separation of the groups.

Explanations of Main Findings

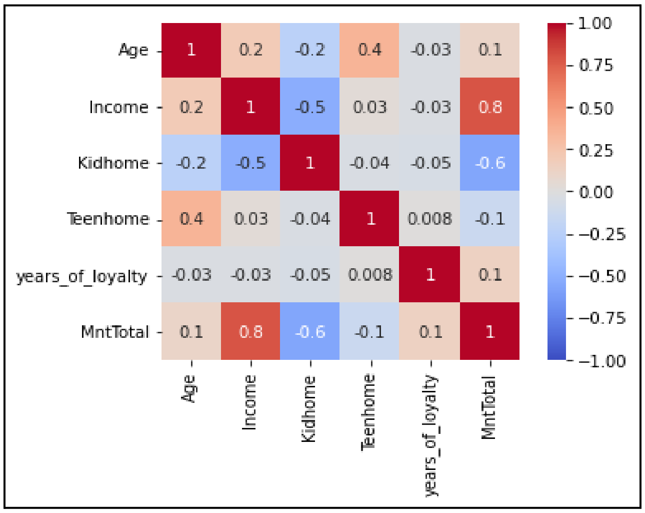
Exploratory Data Analysis:



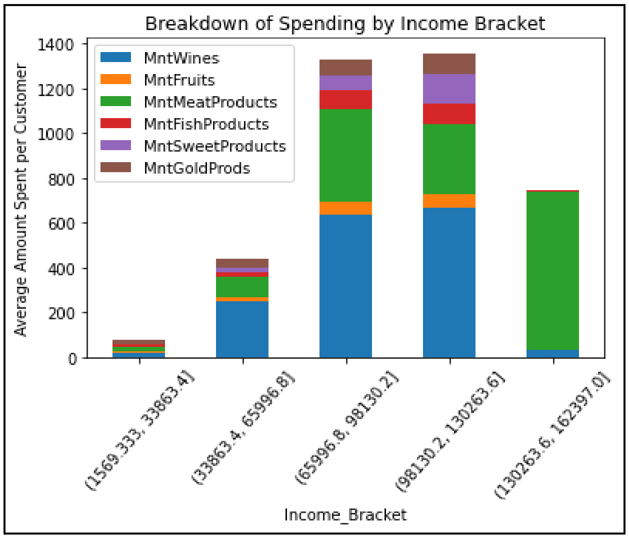
If we examine the relationship between Education and Consumers by Education Level, it is evident that most consumers have at least a college education. This aids in determining whether or not the level of education of a consumer will affect how said consumer will shop/spend.



Looking at the relationship between Income Level and Consumers, it appears that the majority of consumers fall in the income bracket of $25,000 – $75,000, with some high-income consumers with over $100,000.

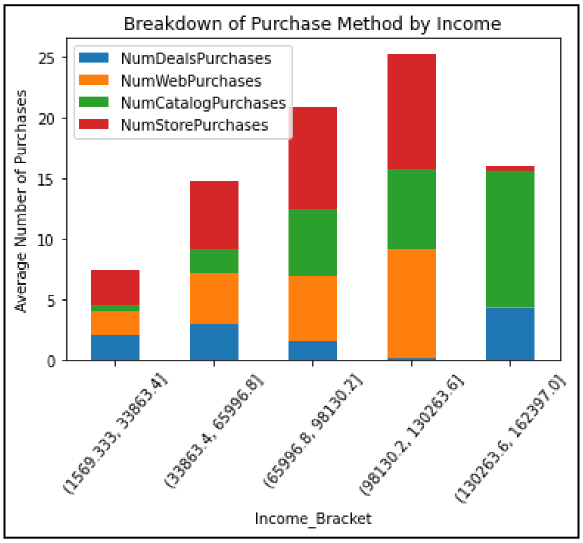


Looking at the correlation of the variables related to customers' background information and how it correlates to the amount spent, it appears that the background information variables are generally uncorrelated with each other. When looking at how the variables are correlated with the amount spent, it appears that income exhibits a strong positive correlation, and having kids at home has a moderate negative correlation. It is not a surprise that income is correlated strongly with consumption or the amount spent as these consumers can consume more goods and spend on more expensive products. Conversely, it is surprising to see that having households with children at home exhibit a negative correlation with the amount spent, as these families usually require more spending with more people to satisfy.



With the consumer's income having a strong correlation with the amount spent, it is useful to break down the data into income brackets and look into how much more consumers spend as their income increases and the breakdown of their consumption at the store. It appears that along with the previously stated observation that the amount spent is positively correlated with income, there is a significant jump in consumption once consumers reach around $66,000/year in income, but flattens out afterward and even falls after approximately $130,000/year even when they earn more. This is an interesting observation as this reflects that although income is correlated strongly with consumption, it becomes less correlated past a certain level of income, and extra income is spent possibly in other types of stores. As for the amount spent on each type of good, it can be seen that consumers spend a much higher proportion of their income on wine and gold products as their income increases, while a relatively smaller increase in proportion can also be observed for meat products.

On the other hand, other goods such as fruits and sweets stay relatively the same and in proportion to increases in income. Meat, gold, and wine tend to be more expensive at grocery stores, this reflects that consumers not only increase their spending when they earn more but also spend a higher proportion of their income on goods that are more expensive or luxury goods.



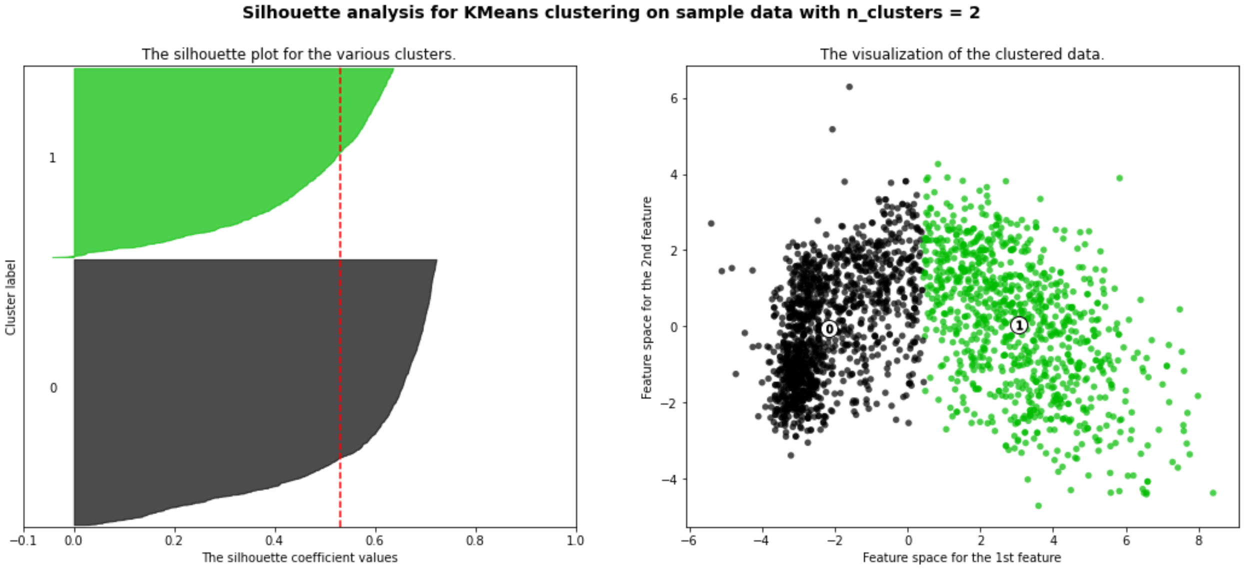
Looking at the breakdown of purchase methods by income bracket, it appears that as income increases, the number of purchases also increases. This makes sense as this follows the pattern that higher earners spend more at the store. The most significant increase in the proportion of total purchases is web purchases and catalog purchases (delivery by mail). This makes sense as these methods are more often used by those who want to make larger purchases or higher quantities of products where it is difficult to do so buying in-person at the store.

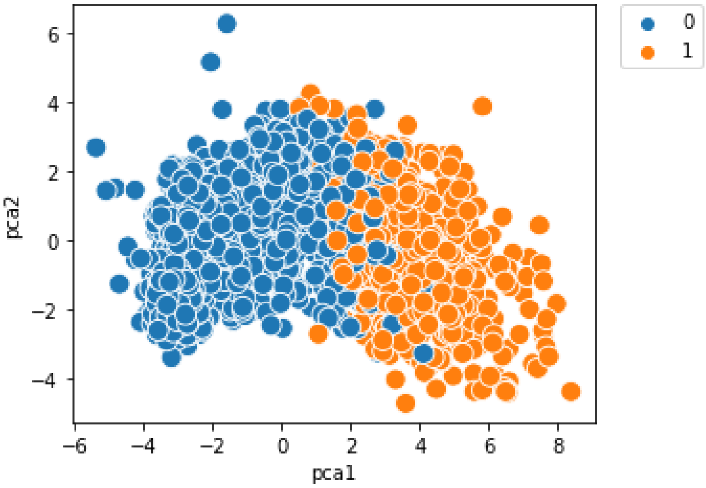
Machine Learning:

When testing our model, we saw that the root-mean-square deviation of a simple linear regression fit was 0.201, which depicts that a linear regression model can relatively predict the data accurately. However, when training and testing with different classifiers, we found models that predicted data much better. We found that both the Decision Tree classifier and the AdaBoost classifier had an accuracy of 100% and an RMSE score of 0. Other extremely accurate classifiers included Linear SVM, Gaussian Process Accuracy, and Neural Net Accuracy, which all had accuracies above 98% and RMSE scores under 0.150.

Clustering:

```
For n_clusters = 2 The average silhouette_score is: 0.5298815348398656
```





To visualize the clustering in the dataset, we tested K-means with various clusters. Out of all the ones tested, the silhouette score for when n_clusters = 2 was the highest, where the subjects matched well to their clusters and poorly to neighboring clusters. This makes sense since there are likely two distinct groups that can be separated where customers are either high-value or low-value. While it is also possible that there are more clusters, such as n_clusters = 3, there is also a larger gray area where it is difficult to separate the clusters from each other, which could lead to a lower average silhouette score.

After the K-means clustering, we also performed a PCA where the data points were shaded in different colors based on whether the customer was high value (1) or low value (0). It is important to note that there does not seem to be a clear break between the two groups. Once again, while there are two general groups of high-value and low-value customers, it will never be completely accurate as there will always be some overlap.

Statement of Limitations, Challenges Encountered, and Future Work

Dataset: Our dataset seems to be published sometime around the last 20 years. This could mean that their behavioral patterns and pricing are a bit outdated, given that the economy has changed tremendously during the Information Revolution. As proof, our mean birth year for this dataset is 1968, with the youngest being born in 1977 and the oldest being born in 1893. So, a good portion of this dataset records the spending patterns of a generation (Gen X) that no longer accurately represents today's working class (Millennials).

Future Work: If we obtain more current data that reflects the current generation's working and spending habits, it could benefit marketing efforts much more. Due to the nature of the modern digital economy, this new population body influences a significant portion of consumer behavior drastically (e.g. Buy Now, Pay Later has affected this generation's spending habits).

Constricted income range: Our dataset's mean income range is $52,000/year, with the 75% percentile having a mean of $68,000/year. During the 1980s, it might correspond to the middle class or upper middle class. However, this generally doesn't give much information on the spending habits of different income classes (low income--1%). This disturbs our analysis because we cannot accurately predict the highest spending consumer type if most of the dataset fits into one socioeconomic class.

*Future Work***:** Due to that challenge, it would be logical to consider excluding high-end, expensive-branded products and instead assume that the products they buy do not have outrageous prices. The outcome this would have on the analysis is that it would eliminate the outliers that may extrapolate the conclusions of our data for the general population of consumers. Therefore instead, by excluding some extremely high-earning consumers as well, we maintain the status quo of having not much variation in income to normalize the data for prediction purposes.

Inaccurate reflection of web history: These days, marketing efforts are focused entirely on online promotion. In the 1980s, the World Wide Web wasn't that popular so the column "NumWebVisitsMonth" produces a much lower amount than we expected, 5/month. Thus, this dataset doesn't tell us about the online traffic or spending patterns of today's consumers because the state of today’s economy is very different from when the dataset was published. This restricts our analysis to only look at real-life spending patterns and not look deeper into online marketing effects, which is essential in today’s society.

Future Work: Similar to how we proposed using current data to reflect modern spending patterns, having relevant data (e.g., SEO metrics) for today’s online environment where digital shopping is practically the norm would be more beneficial.

Lack of inflation-adjusted prices: The prices of this dataset do not reflect current prices in today’s economy. Without inflation-adjusted prices, it would be really hard to predict spending patterns in terms of objective quantity accurately. For example, the mean amount spent on wine per month is $300, which would equate to a lot in the 1990s but would only equate to 2-5 bottles today, depending on the region. This problem affects our objective because we have a more vague hold of predicting how much higher spending consumers spend on certain products if we do not know exactly how many products those represent.

Future Work: Certain adjustments could be made to account for inflation to make the data agnostic to market economics. Such as, if we normalized it using the Consumer Price Index (CPI), we would not encounter issues with using potentially outdated data for a predictive model, and this dataset could still be relevant today. Further implications of this change would allow us to measure consumer behavior over time, uncovering key insights such as seasonality and highlighting cyclical patterns of consumption that may help marketing teams for long-term outlooks.

No information on product type: We have no information regarding the brand type of products in our dataset. For example, we have no idea if the number of wine products consists of cheap or expensive bottles and if the fish products are expensive cuts like salmon or discount cuts like tilapia, etc. This hinders our project objective because it doesn’t allow for a more in-depth prediction of what a high-spending consumer is. Recognizing the price tag and product type of a good is essential in diagnosing the nature of consumer behavior, as a high purchase does not necessarily mean big spending (i.e., a large family relying on coupons can have the same grocery bill as a couple who prefers expensive things).

Future Work: Obtaining context behind the sourcing of products used in our analysis would help explain our findings and some of the anomalies we encountered and possibly provide a reason as to certain outliers or what types of products are purchased more for different consumer segments/clusters of traits. Ultimately, gathering more information would open the door for multiple analyses, as elaborated further below.

Beyond the Scope

Considering all of the possible changes above, as for future work, in summary, we could take a more open-ended approach to this analysis as we limited the number of attributes for this project. The amount of traits that can potentially affect the sale of various items is extensive. Variables that could be looked into further include but are not limited to the source of the purchase of wine, the number of people per household, or the purchase of other items affecting each other.

We would like to explore all of these traits and look more into the depth of the various correlations with not only the sale of wine but all of the other products included in the dataset as well and more, incorporating different algorithms, clustering techniques, and additional plots.

Additionally, to further standardize the implications of creating a predictive model for consumer behavior, it would be appropriate to replicate our analysis using a more varied sample of respondents. Moreover, it has not demonstrated the differentiating character of retailer personality as it took only one retailer into account, which is admittedly unknown as the author stated that he obtained "this dataset while exploring some Machine Learning projects… [without] any idea about [its] source". Future studies on retailer personality should include other, specified types of retailers, such as department markets or food markets. Other consequences of retailer personality, such as trust, attachment, and commitment, could also be studied.

Appendices

Appendix A - Dataset Description

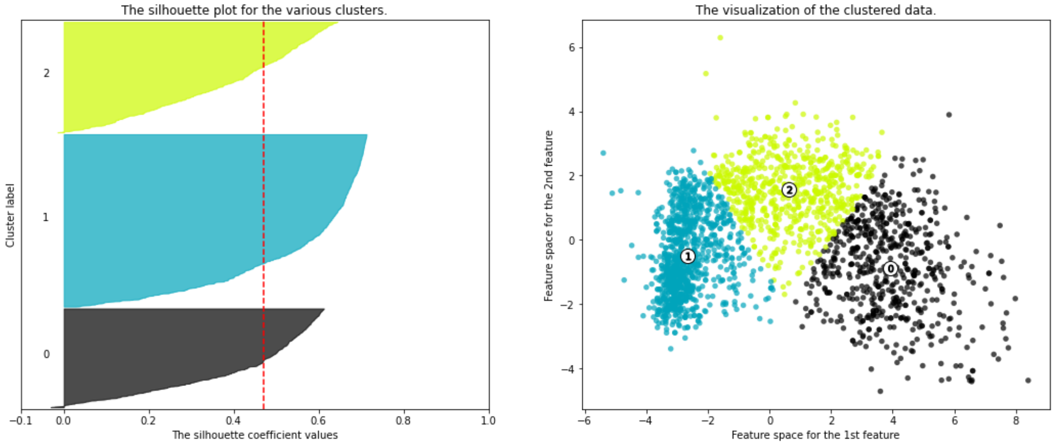
PEOPLE	PRODUCTS	PROMOTION	PLACE
ID: Unique Identifier	MntWines: Amount spent on wine in last 2 years	NumDealsPurchases: Number of purchases made with a discount	NumWebPurchases: Number of purchases made through the company's website
Year_Birth: Customer's birth year	MntFruits: Amount spent on fruits in last 2 years	AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise	NumCatalogPurchases: Number of purchases made using a catalogue
Education: Customer's education level	MntMeatProducts: Amount spent on meat in last 2 years	AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise	NumStorePurchases: Number of purchases made directly in stores
Marital_Status: Customer's marital status	MntFishProducts: Amount spent on fish in last 2 years	AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise	NumWebVisitsMonth: Number of visits to company's website in the last month

Income: Customer’s yearly household income	MntSweetProducts: Amount spent on sweets in last 2 years	AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise	
Kidhome: No. of children in household	MntGoldProds: Amount spent on gold in last 2 years	Response: 1 if customer accepted the offer in the last campaign, 0 otherwise	
Dt_Customer: Date of customer's enrollment with the company			
Teenhome: No. of teenager in household			
Recency: no. of days since last purchase			
Complain: Boolean if customer complained within 2 years			

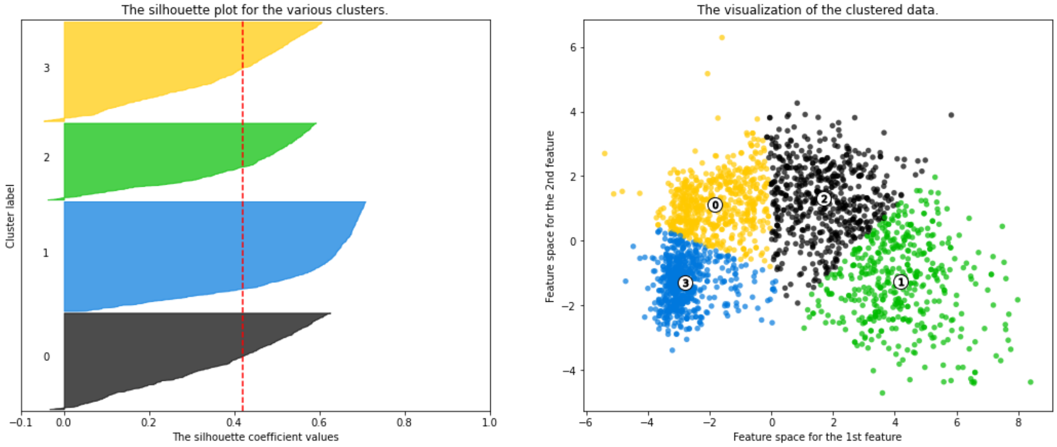
Appendix B - K-Means Clustering

```
For n_clusters = 3 The average silhouette_score is : 0.4710618698191774
For n_clusters = 4 The average silhouette_score is : 0.4197049856416666
For n_clusters = 5 The average silhouette_score is : 0.393383273574781
```

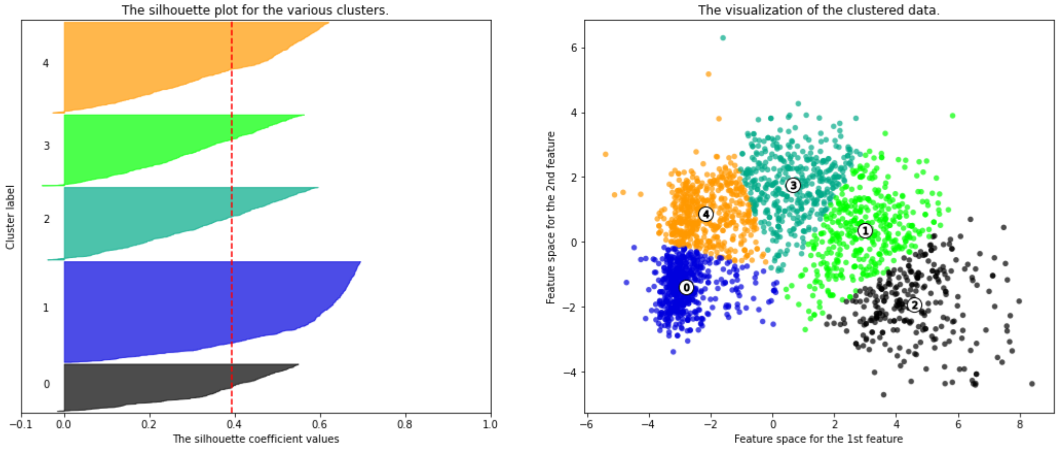
Silhouette analysis for KMeans clustering on sample data with n_clusters = 3



Silhouette analysis for KMeans clustering on sample data with n_clusters = 4



Silhouette analysis for KMeans clustering on sample data with n_clusters = 5


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature