ABOUT DATASET

This dataset is taken from **KAGGLE**.

Context

Customer Personality Analysis is a detailed analysis of a company’s ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors and concerns of different types of customers. Customer personality analysis helps a business to modify its product based on its target customers from different types of customer segments. For example, instead of spending money to market a new product to every customer in the company’s database, a company can analyze which customer segment is most likely to buy the product and then market the product only on that particular segment.

Link of the dataset

https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis?select=marketing_campaign.csv

Content

People

* ID: Customer's unique identifier

* Year_Birth: Customer's birth year

* Education: Customer's education level

* Marital_Status: Customer's marital status

* Income: Customer's yearly household income

* Kidhome: Number of children in customer's household

* Teenhome: Number of teenagers in customer's household

* Dt_Customer: Date of customer's enrollment with the company

* Recency: Number of days since customer's last purchase

* Complain: 1 if the customer complained in the last 2 years, 0 otherwise

Products

* MntWines: Amount spent on wine in last 2 years

* MntFruits: Amount spent on fruits in last 2 years

* MntMeatProducts: Amount spent on meat in last 2 years

* MntFishProducts: Amount spent on fish in last 2 years

* MntSweetProducts: Amount spent on sweets in last 2 years

* MntGoldProds: Amount spent on gold in last 2 years

Promotion

* NumDealsPurchases: Number of purchases made with a discount

* AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise

* AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise

* AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise

* AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise

* AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise

* Response: 1 if customer accepted the offer in the last campaign, 0 otherwise

Place

* NumWebPurchases: Number of purchases made through the company’s website

* NumCatalogPurchases: Number of purchases made using a catalogue

* NumStorePurchases: Number of purchases made directly in stores

* NumWebVisitsMonth: Number of visits to company's website in the last month

MAIN OBEJCTIVE

* Will perform clustering to summarize customer segments.

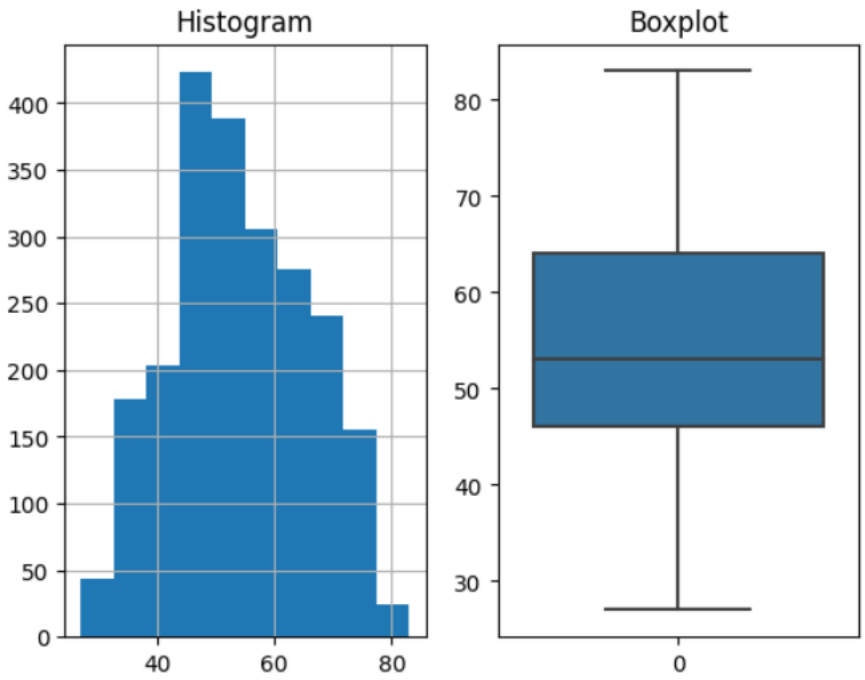
DATA CLEANING AND EXPLORATION

1. Categorical variables encoded with the help of label encoder. Unsupervised algorithms are distance based algorithms and data should in numeric. Marital_status and Education column
2. Data types corrected for date column
3. Age is extracted for year_birth column
4. Unique Values in all the columns

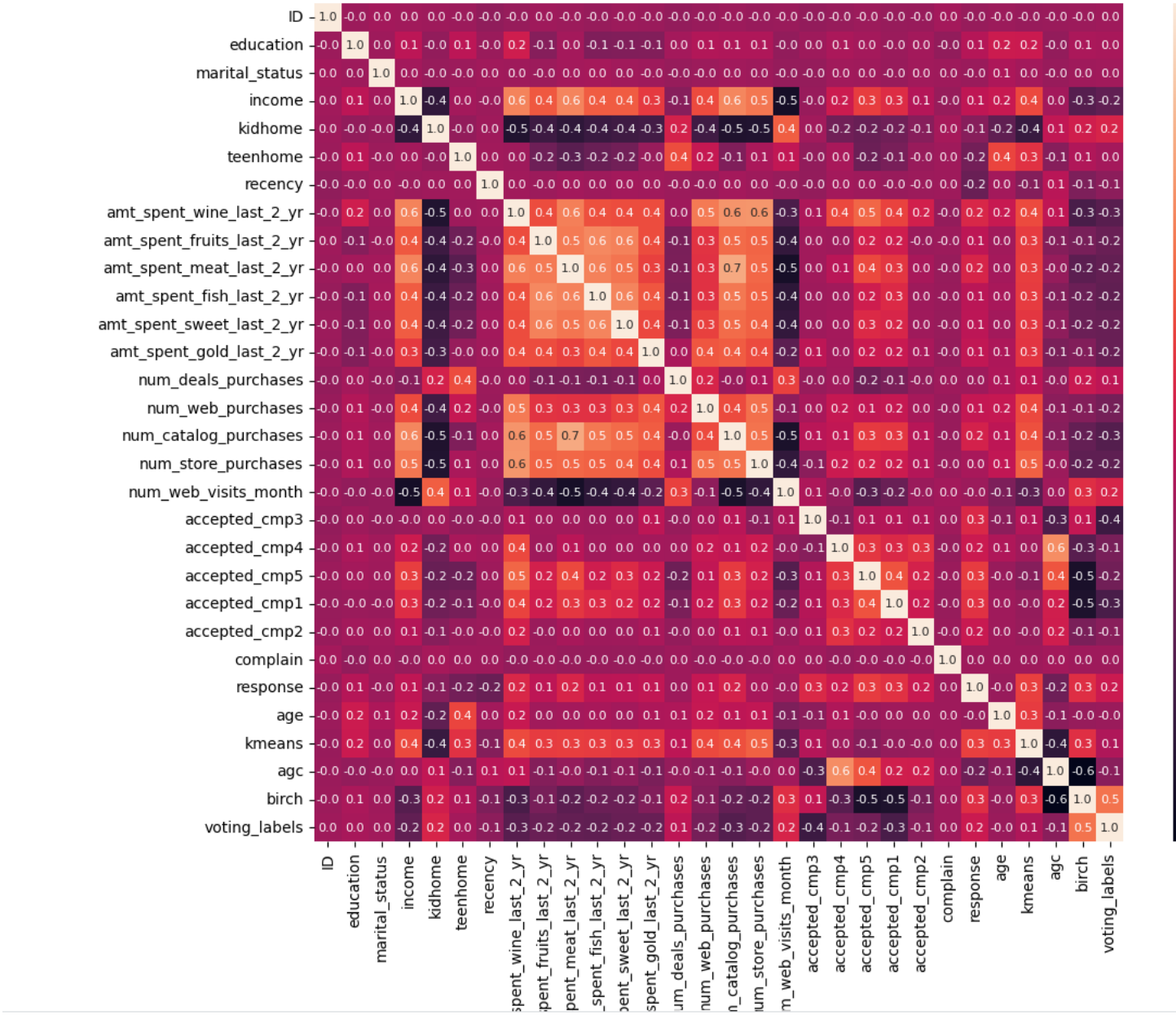
Variable	Unique Values
ID	2240
education	5
marital_status	8
income	1975
kidhome	3
teenhome	3
cust_enrol_dt	663
recency	100
amt_spent_wine_last_2_yr	776
amt_spent_fruits_last_2_yr	158
amt_spent_meat_last_2_yr	558
amt_spent_fish_last_2_yr	182
amt_spent_sweet_last_2_yr	177
amt_spent_gold_last_2_yr	213
num_deals_purchases	15
num_web_purchases	15
num_catalog_purchases	14

num_store_purchases	14
num_web_visits_month	16
accepted_cmp3	2
accepted_cmp4	2
accepted_cmp5	2
accepted_cmp1	2
accepted_cmp2	2
complain	2
Z_cost_contact	1
Z_revenue	1
response	2
age	5

5. Income column contained some missing values. Dealt with imputation.
6. No duplicate user ids were found.
7. Histogram and Boxplot of age column. Some age values were greater than 100. So, those values were removed from the analysis.



8. Correlation plot of all the variables.



9. Data standardized with the help of MinMaxScaler.

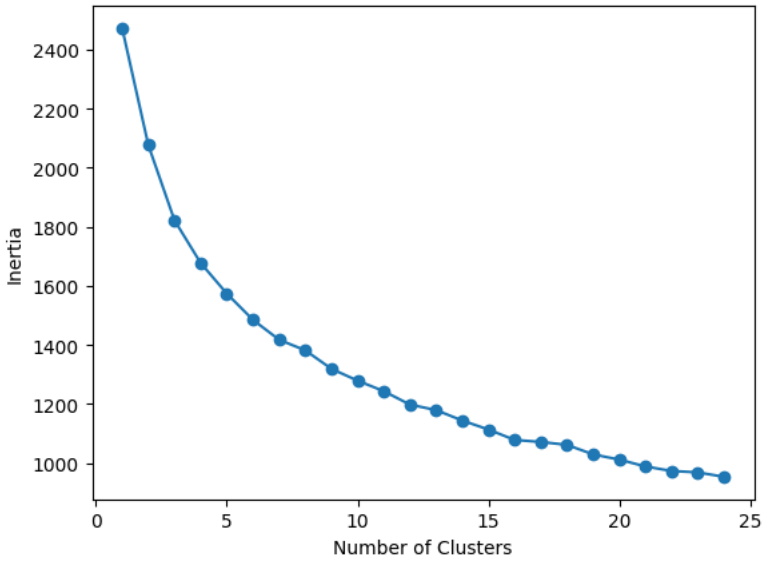
MODEL TRAINING AND PREDICTIONS

Total of three models were trained KMEANS, BIRCH & AGGLOMERATIVE CLUSTERING

Summary of training three variations of the unsupervised model. Here is a table to identify the distribution of data with respect to different clusters

CLUSTERS	VOTING_LABELS	BIRCH	AGC	KMEANS
0	229	156	277	578
1	611	163	784	145
2	1181	1629	902	673
3	57	132	164	620
4	159	157	110	221

Selection of Kmeans



K=5 clusters were created with three different algorithms, but the distribution keeps on shifting from one cluster to another.

And for my analysis I have found kmeans to be the right choice as it is able to identify clusters and distribute data somewhat equally.

ANALYSIS

- Cluster averages were calculated for expressive features/columns. Table for the same given below
- Maximum and Minimum values of features are highlighted in the table

					Amount Spent						Purchased				
Clusters	**rows**	**Age**	**Income**	**Receny**	**Wine**	**Fruits**	**Meat**	**Fish**	**Sweets**	**Gold**	**Deals**	**Web**	**Catalog**	**Store**	**Webvisits**
1	578	45.65	32931.46	51.89	39.40	7.18	37.95	10.41	7.30	18.61	1.94	2.30	0.63	3.27	6.63
2	145	52.51	80589.58	45.06	867.98	54.95	457.37	81.81	60.81	77.44	1.14	5.49	6.12	8.08	3.26
3	673	59.66	47853.22	50.23	188.30	6.22	51.63	8.68	7.15	27.68	2.96	3.73	1.37	4.64	5.90
4	620	56.28	69259.22	52.23	536.25	58.24	331.81	84.29	58.97	74.93	2.04	5.56	4.97	9.08	3.61
5	221	54.19	49718.03	32.28	326.72	28.76	202.12	36.02	27.79	51.09	2.97	4.77	3.14	5.17	6.29

SUMMARY & INSIGHTS

From the above table, we can be inferred that

- Cluster 1 – **Lower Earners and Youngest Age group**
  - The average income is around **33K** which is low compared to other groups and they spend most of their money on **WINE**, **MEAT and GOLD**
  - Visits the store is every **52 days**
  - Purchased most of their items through **webvisits**
- Cluster 2 – **Highest Earners**
  - The average income is around **81K** which is highest compared to all and they spend most of the their money on **WINE**, **MEAT and FISH**
  - Visits the store is every **45 days**
  - Purchased most of their items by **visiting store**

- Cluster 3 – **Highest Age group**
    - They spend most of their money on **WINE**, **MEAT and GOLD**
    - Visits the store is every **50 days**
    - Purchased most of their items by **webvisits**
- Cluster 2 and 3 have somewhat same behaviour and can be combined.
- Cluster 4 – **Second Highest Earners and Second Highest Age group**
    - They spend most of their money on **WINE**, **MEAT and FISH**
    - Visits the store is every **52 days**
    - Purchased most of their items by **visiting store**
- Cluster 1 and 4 have somewhat same behaviour and can be combined.
- Cluster 5 – **Frequent Visitors**
    - They spend most of their money on **WINE**, **MEAT and GOLD**
    - Visits the store is every **32 days**
    - Purchased most of the items through **Webvisits**

SUGGESTIONS AND NEXT STEPS FOR REVISITING THE MODEL

We could further optimize these models

1. By incorporating Dimensionality Reduction techniques to understand the model well but that will force the model to lose its explanatory power.

2. We could also change our model based on the inputs received from our stakeholders about the business.

3. We could also use density-based algorithms like GMM, DBSCAN etc.