CUSTOMER PROFILE

GROUP 6

Aiswarya Prabhalan
Bini Abraham
Darling Oscanoa
Shreyash Banawala
Suruchi Pokhrel

AGENDA

1) INTRODUCTION

2) METHODS

3) RESULTS

4) CONCLUSION AND FUTURE WORK

5) REFERENCES

1) INTRODUCTION

The company's targeting and retention efforts are informed by analysis of consumer data on demographics, purchase patterns, and engagement.

The objectives include identifying new segments for expensive products, reliable current clients, assessing new clients' long-term potential, and segmenting the consumer base by spending and preferences.

2) METHODS

In this project, our "Online retail company" provided us with a Dataset where they recorded all their customer information stored under traditional Marketing from 1.0 format (The 4Ps of Marketing), which is Product Oriented. Now, our online retail company wants to move forward by adding some customer features when defining new target products called "Marketing Mix."

2.1) DATASET USED

This comprehensive dataset allows for a detailed analysis of customer demographics, purchasing behavior, and engagement, enabling the company to develop targeted strategies for customer acquisition and retention. The Dataset has 2240 records.

Field in 4Ps	Features	Description
Customers	ID	Customer ID
	Year_Birth	Year of birth of the customer
	Education	Education level of the customer
	Marital_Status	Marital status of the customer
	Income	Income level of the customer
	Kidhome	Number of children in the customer's household
	Teenhome	Number of teenagers in the customer's household
	Dt_Customer	Date the customer became a customer
	Recency	Number of days since the customer's last purchase
Product	MntWines	Amount spent on wine in the last 2 years
	MntFruits	Amount spent on fruits in the last 2 years
	MntMeatProducts	Amount spent on meat products in the last 2 years
	MntFishProducts	Amount spent on fish products in the last 2 years
	MntSweetProducts	Amount spent on sweet products in the last 2 years
	MntGoldProds	Amount spent on gold products in the last 2 years
Place	NumDealsPurchases	Number of purchases made with discount
	NumWebPurchases	Number of purchases made through the company's website
	NumCatalogPurchas	Number of purchases made using a catalog
	NumStorePurchases	Number of purchases made directly in stores
	NumWebVisitsMont	Average number of visits to the company's website in the last month
Promotion	AcceptedCmp3	1 if the customer accepted the offer in the 3rd campaign, 0 otherwise
	AcceptedCmp4	1 if the customer accepted the offer in the 4th campaign, 0 otherwise
	AcceptedCmp5	1 if the customer accepted the offer in the 5th campaign, 0 otherwise
	AcceptedCmp1	1 if the customer accepted the offer in the 1st campaign, 0 otherwise
	AcceptedCmp2	1 if the customer accepted the offer in the 2nd campaign, 0 otherwise
Customers	Complain	1 if the customer complained, 0 otherwise
	Z_CostContact	Cost of contacting the customer
	Z_Revenue	Revenue generated by the customer
Promotion	Response	1 if the customer responded to the last campaign, 0 otherwise

```
[12]: df.head(10)
```

[12]:

	Birth	Education	Marital_Status	Income	Kidhome	Teenhome	Dt_Customer	Recency	MntWines	...	NumWebVisitsMonth	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Z_CostContact	Z_Revenue	Response
	1957	Graduation	Single	58138.0	0	0	2012-09-04	58	635	...	7	0	0	0	0	0	0	3	11	1
	1954	Graduation	Single	46344.0	1	1	2014-03-08	38	11	...	5	0	0	0	0	0	0	3	11	0
	1965	Graduation	Together	71613.0	0	0	2013-08-21	26	426	...	4	0	0	0	0	0	0	3	11	0
	1984	Graduation	Together	26646.0	1	0	2014-02-10	26	11	...	6	0	0	0	0	0	0	3	11	0
	1981	PhD	Married	58293.0	1	0	2014-01-19	94	173	...	5	0	0	0	0	0	0	3	11	0
	1967	Master	Together	62513.0	0	1	2013-09-09	16	520	...	6	0	0	0	0	0	0	3	11	0
	1971	Graduation	Divorced	55635.0	0	1	2012-11-13	34	235	...	6	0	0	0	0	0	0	3	11	0
	1985	PhD	Married	33454.0	1	0	2013-05-08	32	76	...	8	0	0	0	0	0	0	3	11	0
	1974	PhD	Together	30351.0	1	0	2013-06-06	19	14	...	9	0	0	0	0	0	0	3	11	1
	1950	PhD	Together	5648.0	1	1	2014-03-13	68	28	...	20	1	0	0	0	0	0	3	11	0

2.2 DATA PREPROCESSING

Missing Values

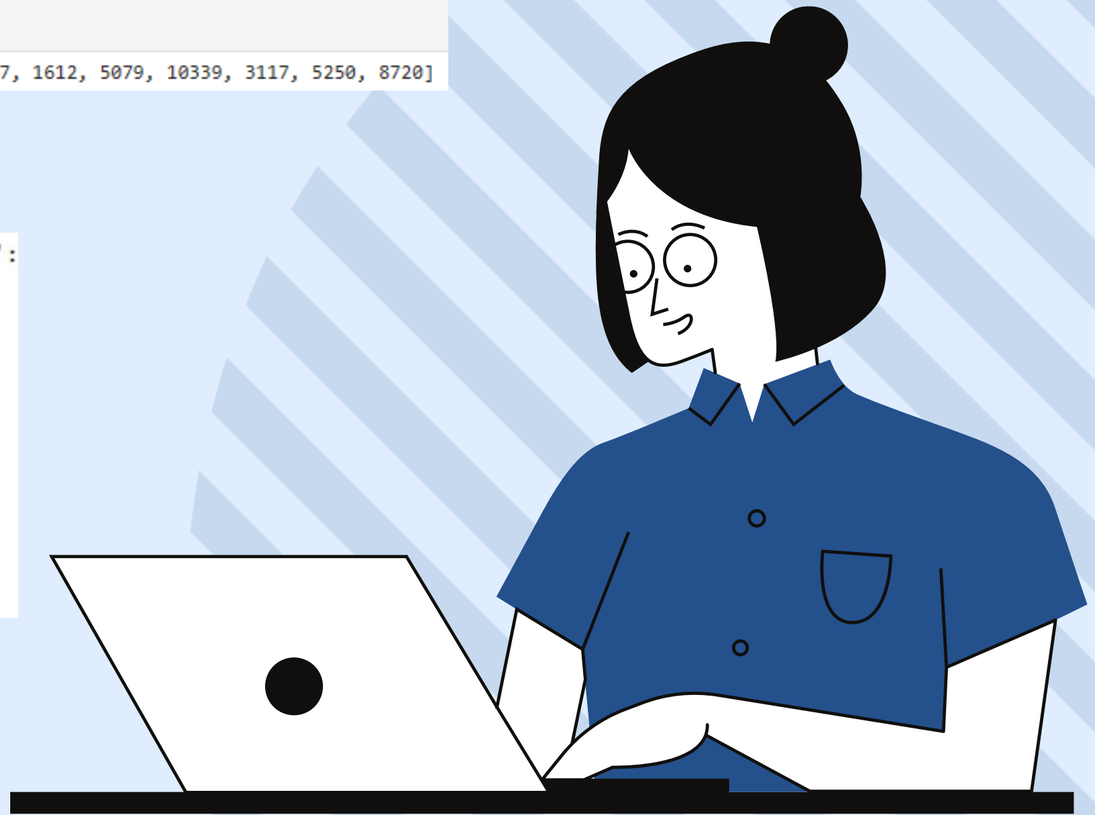
```
list_imputed = df.loc[df['Income'].isna(), 'ID'].tolist()
imputer = KNNImputer(n_neighbors=5, weights='uniform')
df['Income'] = imputer.fit_transform(df[['Income']])

print(list_imputed)
```

```
[1994, 5255, 7281, 7244, 8557, 10629, 8996, 9235, 5798, 8268, 1295, 2437, 2863, 10475, 2902, 4345, 3769, 7187, 1612, 5079, 10339, 3117, 5250, 8720]
```

Uniques Values

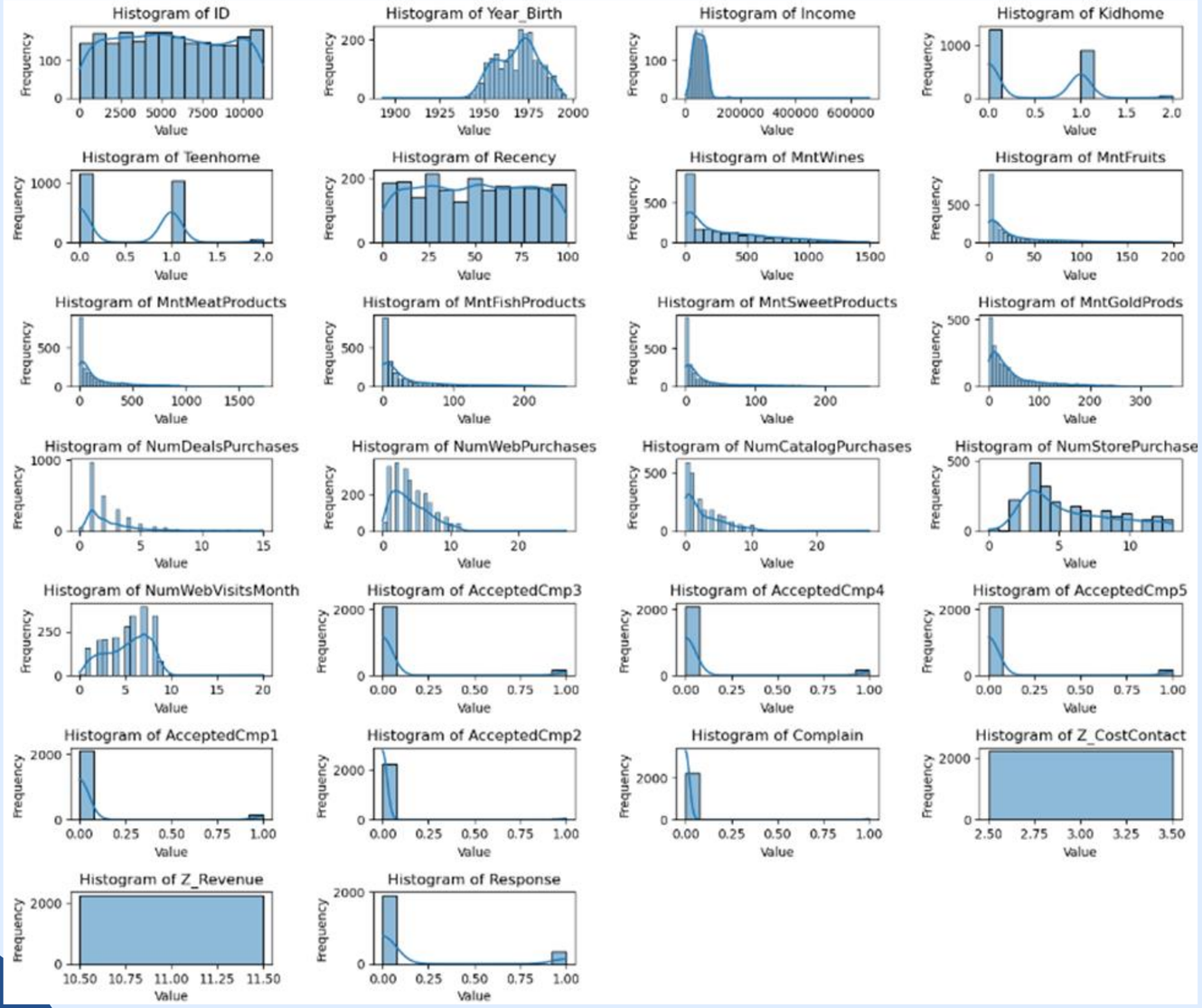
```
Frequency of values in 'Education':
Education
Graduation    1127
PhD            486
Master         370
2n Cycle       203
Basic           54
Name: count, dtype: int64
```

```
Frequency of values in 'Marital_Status':
Marital_Status
Married     864
Together    580
Single      480
Divorced    232
Widow        77
Alone         3
Absurd        2
YOLO          2
Name: count, dtype: int64
```

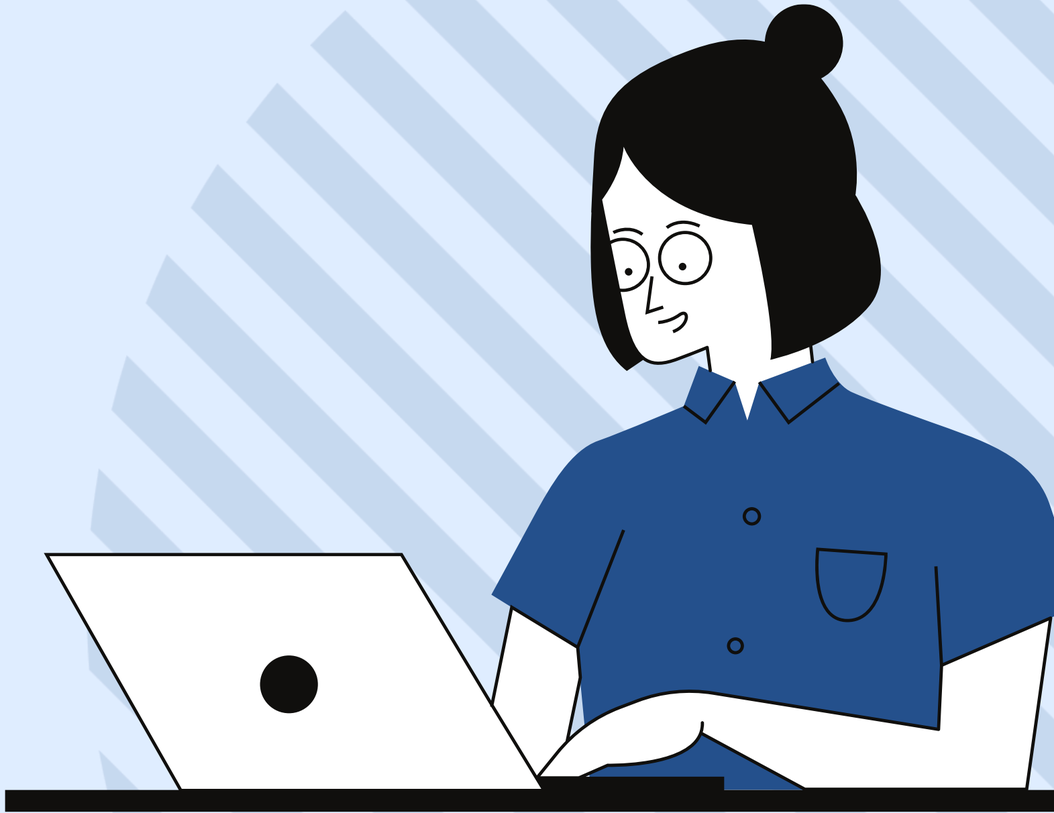


2.2 DATA PREPROCESSING (CONT)

EDA, Histogram

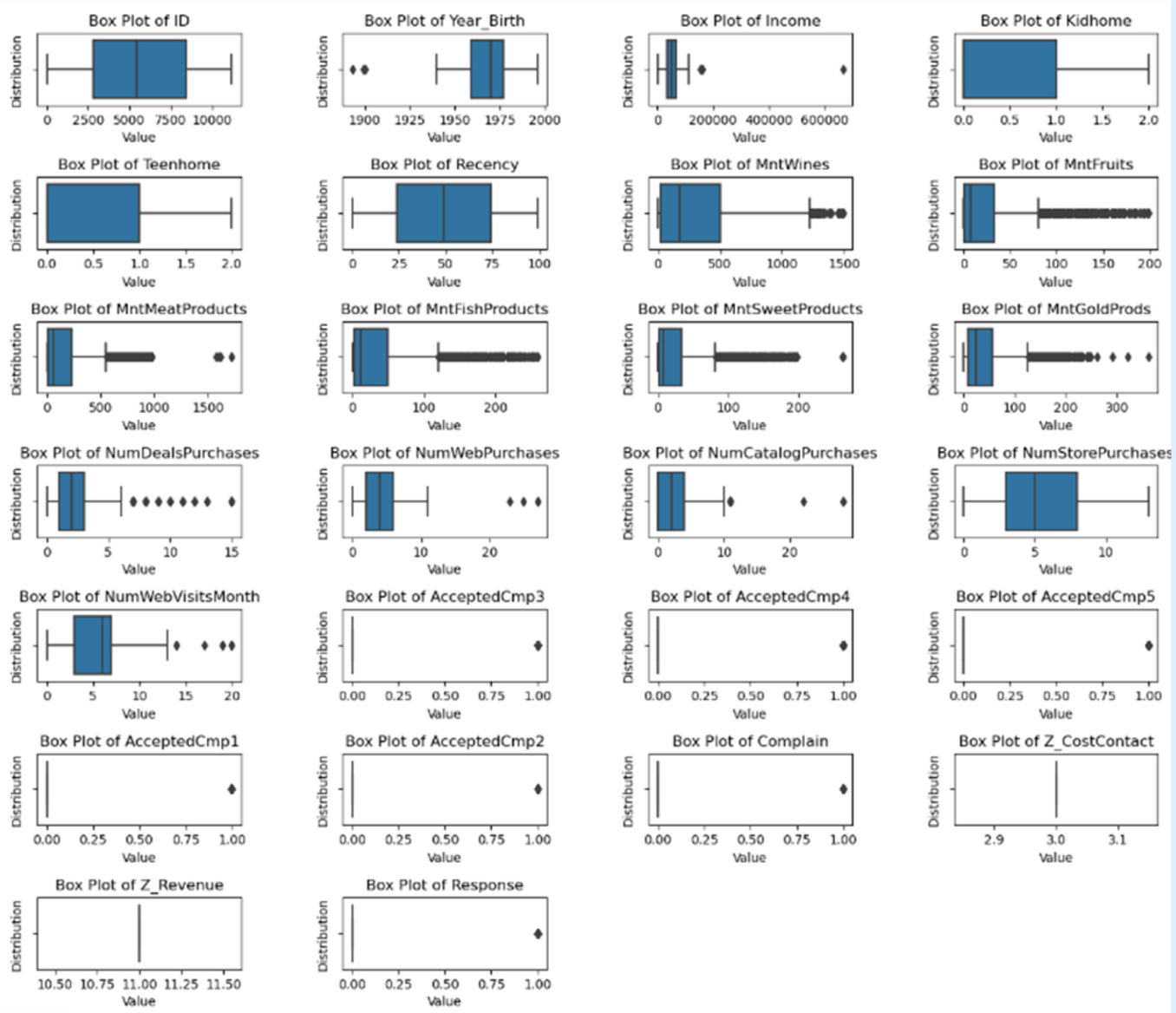


THE MOST AMOUNT SPENT IS ON WINE AND GOLDPRODUCTS, NOT SURPRISINGLY ALSO MEATPRODUCTS



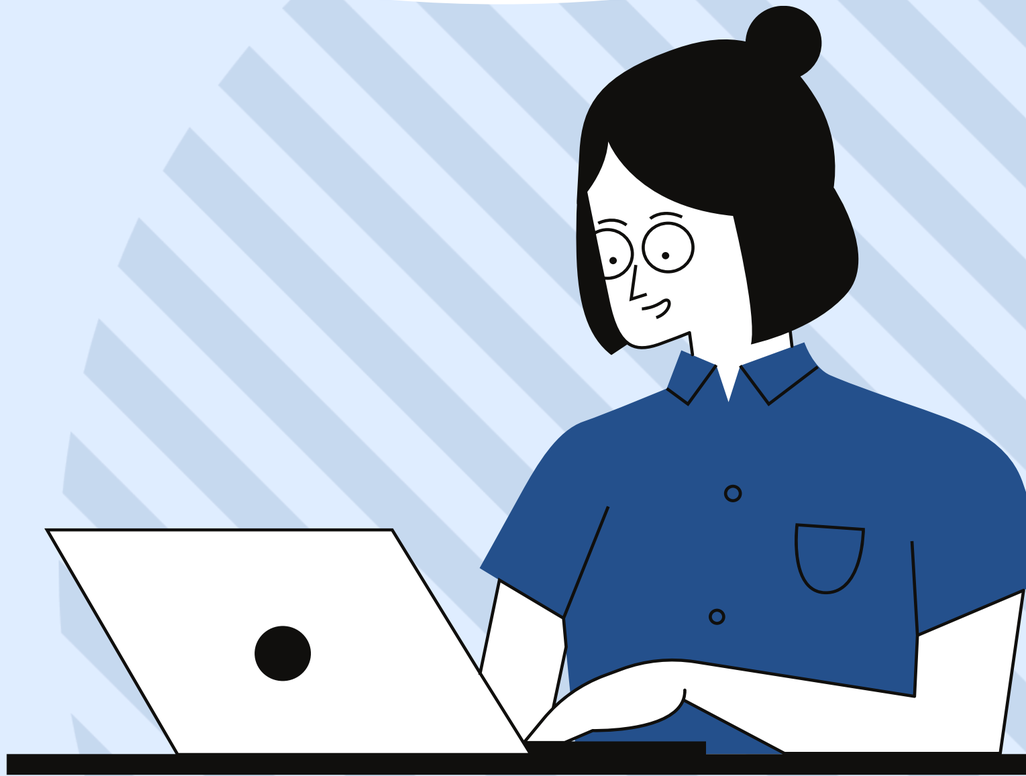
2.2 DATA PREPROCESSING (CONT)

EDA, Boxplot



-WE SEE MANY OUTLIERS FOR YEAR_BIRTH, INCOME, AND ALL AMOUNTS IN PRODUCTS AND PURCHASES BUT AGAIN WE ARE NOT GOING TO DEAL WITH THEM AT THIS STAGE

-Z_CostContact and Z_Revenue are constant



2.3 FEATURE ENGINEERING

Other transforms: For “Education”: As there are 5 level of education and are categorical, we are going to encode those
For “Marital_Status”: There are 7 values, then we used LabelEncoder from scikit-learn library.
For “Dt_Customer”: As Dt_Customer is the date when a customer becomes a customer, the number of days being a customer rather than the date is most important.

Transform Year_Birth to Age Ranges

```
import datetime
today = datetime.date(2014, 12, 31)
df['Age'] = today.year - df['Year_Birth']

df['Age_Range'] = df['Age'].apply(lambda x:
                                  0 if x <= 18 else  # Child
                                  1 if x <= 24 else  # Youth
                                  2 if x <= 34 else  # Young Adult
                                  3 if x <= 44 else  # Adult
                                  4 if x <= 54 else  # Middle-Aged
                                  5 if x <= 64 else  # Pre-Senior
                                  6)                 # Senior

df = df.drop(['Year_Birth', 'Age'], axis=1)
df.head(10)
```

3) RESULTS

a) Which new customers can the company target for high-cost products?

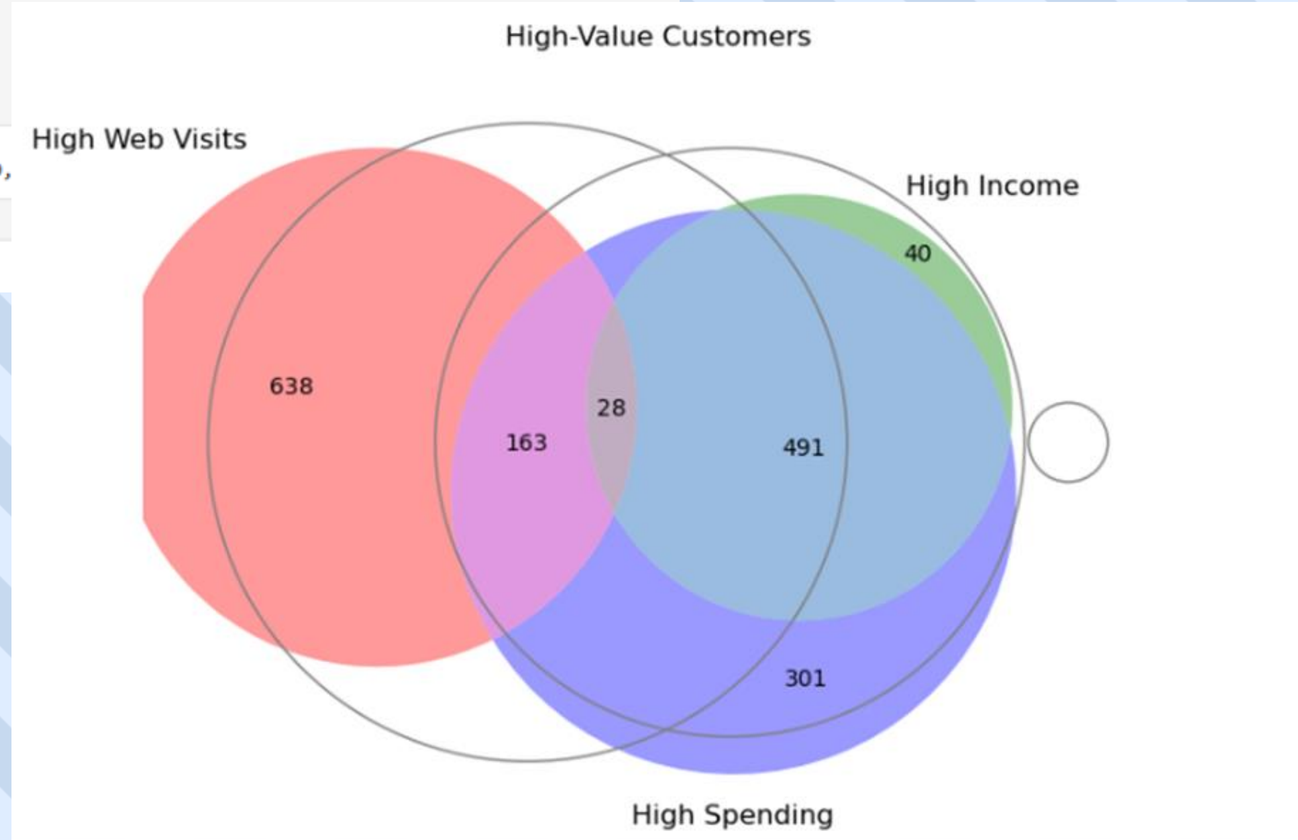
```
high_web_visits_customers = high_web_visits_customers['ID'].tolist()
high_income_customers = high_income_customers['ID'].tolist()
high_spending_customers = high_spending_customers['ID'].tolist()

high_value_customers = list(set(high_web_visits_customers) & set(high_income_customers) & set(high_spending_customers))

#We will remove any high-value customers that coincide with the "list_imputed"
high_value_customers = [cid for cid in high_value_customers if cid not in list_imputed]

print("Customer IDs of High-Value Customers:")
print(high_value_customers)
```

```
Customer IDs of High-Value Customers:
[7441, 6566, 4910, 8755, 5299, 10678, 1079, 5831, 10057, 2379, 4299, 6606, 3667, 7899, 6749, 5341, 3426, 5989,
```

```
len(high_value_customers)
```

```
28
```

1ST APPROACH: WE FOUND 28 CUSTOMERS WITH THEIR CUSTOMER ”IDS" THAT, AFTER COMPARING WITH THE IMPUTED VALUES FOR INCOME, DON’T HAVE THE SAME VALUE, SO WE CAN CONCLUDE THAT THIS IS A GOOD LIST TO SHOW TO THE BUSINESS



3) RESULTS

a) Which new customers can the company target for high-cost products?

```
income_75 = df['Income'].quantile(0.75)
web_visits_75 = df['NumWebVisitsMonth'].quantile(0.75)
accepted_cmp_75 = df[['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5']].sum(axis=1).quantile(0.75)    #Now we are considering the previous Accepted marketing campaign
recency_25 = df['Recency'].quantile(0.25)                                                                                            #Now we are taking only the most recent buyers (the lower quartile)

#Now combining all of them:
high_value_new_customers_raft = df[(df['Income'] >= income_75) &
                                   (df['NumWebVisitsMonth'] >= web_visits_75) &
                                   (df[['AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5']].sum(axis=1) >= 2) &
                                   (df['Recency'] <= recency_25)]
```

```
# Filtering the imputed
target_customer_ids_raft = high_value_new_customers_raft['ID'].tolist()
target_customer_ids_raft = [cid for cid in target_customer_ids_raft if cid not in list_imputed]
target_customer_ids_raft
```

```
[3667, 2535]
```

A 2ND APPROACH OR TECHNIQUE IS WELL-KNOWN IN THE MARKETING FIELD CALLED RAFT (RECENCY, AMOUNT, FREQUENCY, TENURE) TO FIND HIGHLY VALUABLE CUSTOMERS. THIS TECHNIQUE ASSESSES CLIENTS' MONETARY VALUE (HOW MUCH MONEY THEY SPEND), FREQUENCY (HOW OFTEN THEY MAKE PURCHASES), AND RECENCY (HOW LONG AGO THEY MADE A PURCHASE). COMBINING ALL TOGETHER. WE FOUND ONLY 2 IDS

3) RESULTS

b) Which old customers has the company consistently relied on?

Some key characteristics of the "old customers" that companies should target include:

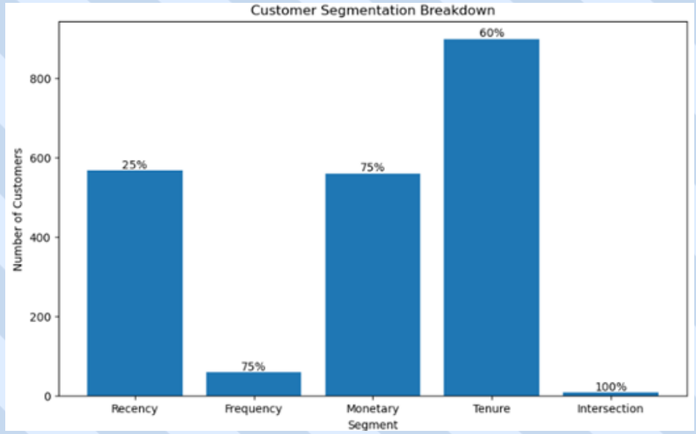
-High Recency, Frequency, and Monetary (RFM) values

-Customers who are part of the company's loyalty program

-Customers with a history of repeat purchases and high total spending

```python
# Recency (R)
high_recency_customers = df[df['Recency'] <= df['Recency'].quantile(0.25)]
print("R:", len(high_recency_customers))

# Frequency (F)
high_frequency_customers = df[(df['NumDealsPurchases'] >= df['NumDealsPurchases'].quantile(0.75)) &
                              (df['NumWebPurchases'] >= df['NumWebPurchases'].quantile(0.75)) &
                              (df['NumCatalogPurchases'] >= df['NumCatalogPurchases'].quantile(0.75)) &
                              (df['NumStorePurchases'] >= df['NumStorePurchases'].quantile(0.75))]
print("F:", len(high_frequency_customers))

# Monetary Value (M)
high_monetary_customers = df[df['Z_Revenue'] >= df['Z_Revenue'].quantile(0.75)]
print("M:", len(high_monetary_customers))

# Tenure (T) #We ran first with 75% percentile but the result was only 3 records, so then iterate until convergence got that with 60% percentile
high_tenure_customers = df[df['Dt_Customer_Days'] >= df['Dt_Customer_Days'].quantile(0.6)]
print("T:", len(high_tenure_customers))

# The intersection
target_customers_b = high_recency_customers.merge(high_frequency_customers, on='ID', how='inner', suffixes=('_r', '_f'))
target_customers_b = target_customers_b.merge(high_monetary_customers, on='ID', how='inner', suffixes=('', '_m'))
target_customers_b = target_customers_b.merge(high_tenure_customers, on='ID', how='inner', suffixes=('', '_t'))
print("Intersection (R, F, M, T):", len(target_customers_b))
target_customers_b

#Loyalty program
loyalty_customers = df[df['AcceptedCmp1'] == 1]
```

```
R: 567
F: 58
M: 560
T: 899
Intersection (R, F, M, T): 7
```



3) RESULTS

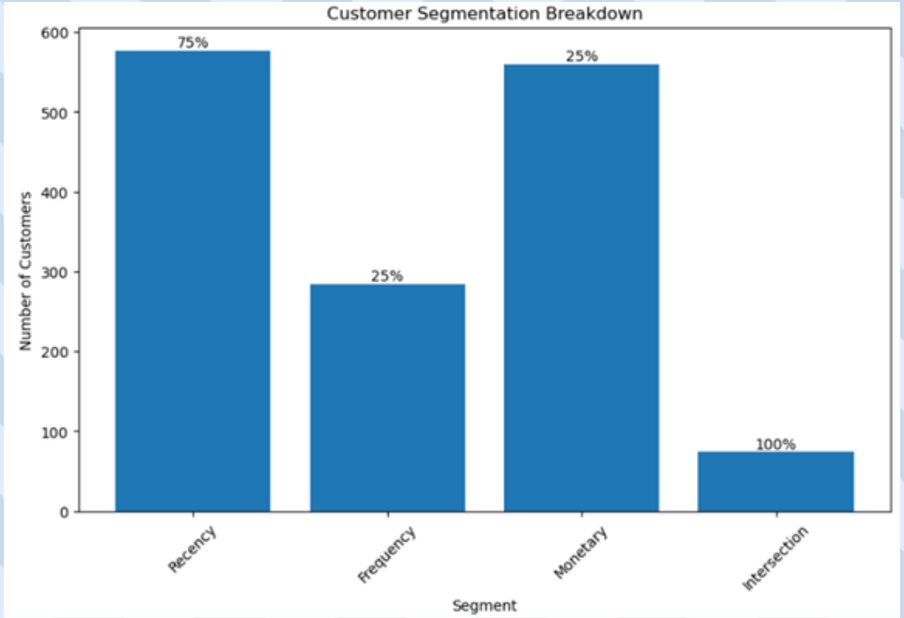
c) For which old customers are targeting efforts, not going to yield much benefit?

With the same concept from Clientbook but opposite values to the previous question, we need to look for the following:

-Customers with Low Recency, Frequency, and Monetary (RFM) values

-Customers who have not purchased in a long time (low Recency)

-Customers with low total spending and purchase frequency



3) RESULTS

d) Which new customers have the danger of going into territory-c?

Another concern from the business is detecting new clients who need to carefully assess their long-term potential to avoid them becoming unproductive or with low value and take action immediately:

```
# Recency (R)
high_recency_customers = df[df['Recency'] <= df['Recency'].quantile(0.25)]
print("R:", len(high_recency_customers))

# Frequency (F)
low_frequency_customers = df[(df['NumWebVisitsMonth'] <= df['NumWebVisitsMonth'].quantile(0.3)) &
                             (df['AcceptedCmp5'] == 0) &
                             (df['AcceptedCmp1'] == 0) &
                             (df['AcceptedCmp2'] == 0)]
print("F:", len(low_frequency_customers))

# Monetary Value (M)
low_monetary_customers = df[df['Z_Revenue'] <= df['Z_Revenue'].quantile(0.25)]
print("M:", len(low_monetary_customers))


target_customers_c = high_recency_customers.merge(low_frequency_customers, on='ID', how='inner', suffixes=('_r', '_f'))
target_customers_c = target_customers_c.merge(low_monetary_customers, on='ID', how='inner', suffixes=('', '_m'))
print("Intersection (R, F, M):", len(target_customers_c))
```

```
R: 567
F: 607
M: 560
Intersection (R, F, M): 17
```

3) RESULTS

e) Using ‘spending’ behaviour, classify the customers into at least 4 categories

For this step we made more feature engineering and outliers treatment

Finally, we used DBSCAN to make our clustering and after some interaction, we see the best names for our clusters are:

Outliers

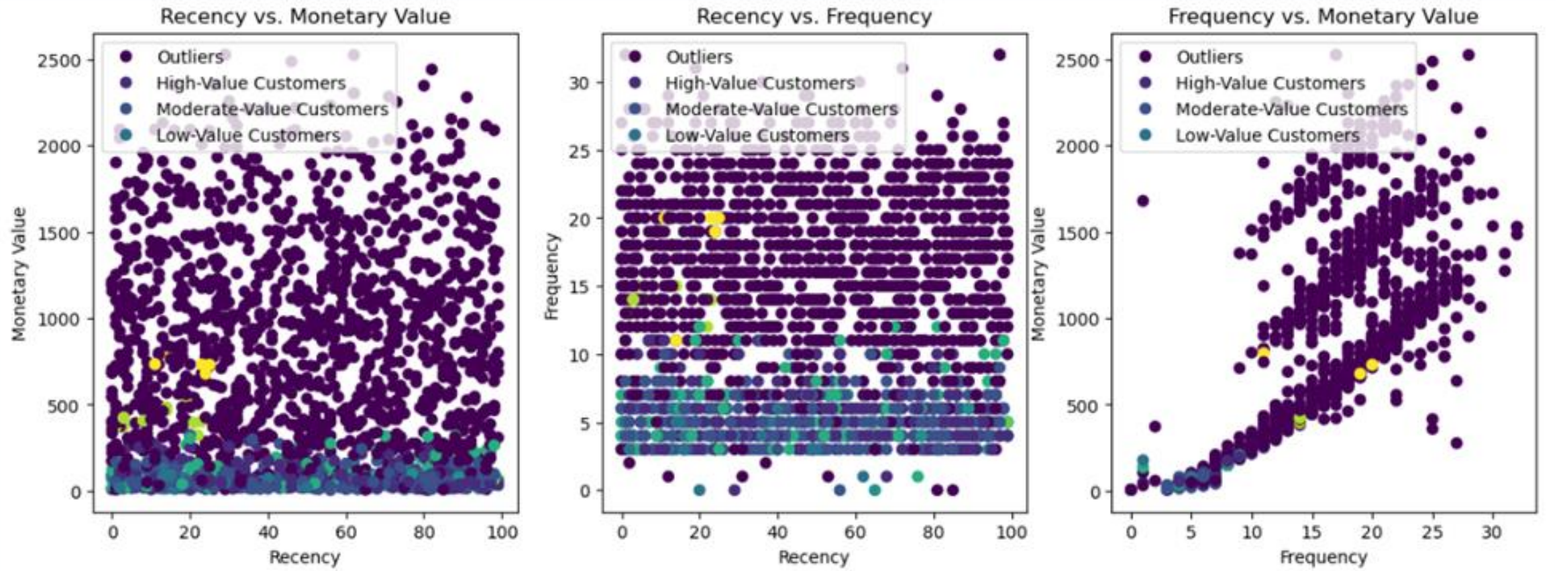
High-Value Customers

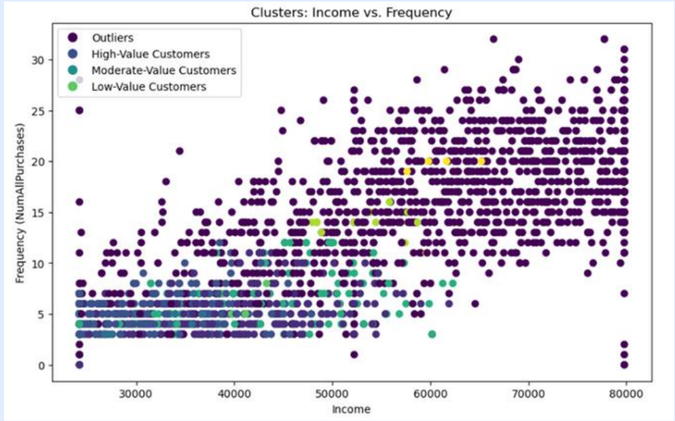
Moderate-Value Customers

Low-Value Customers

3) RESULTS

e) Using ‘spending’ behaviour, classify the customers into at least 4 categories (cont)





As expected, low value customers are located by low income. However, there are some High value customers in the low Income too.

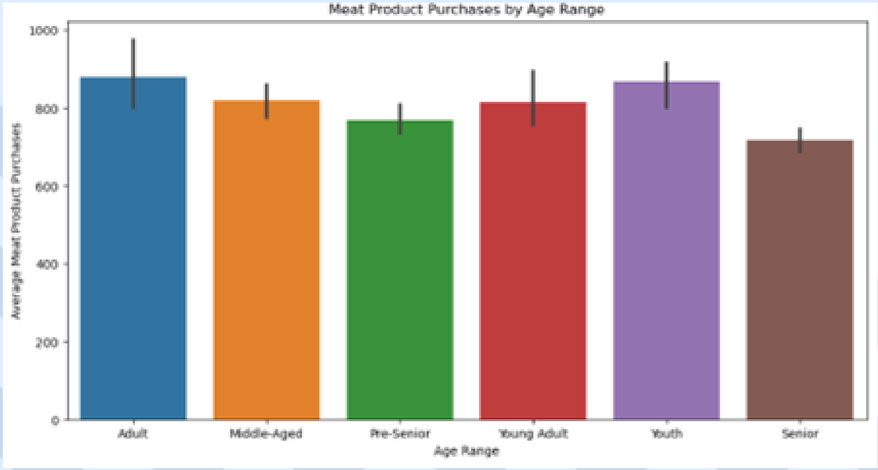
3) RESULTS

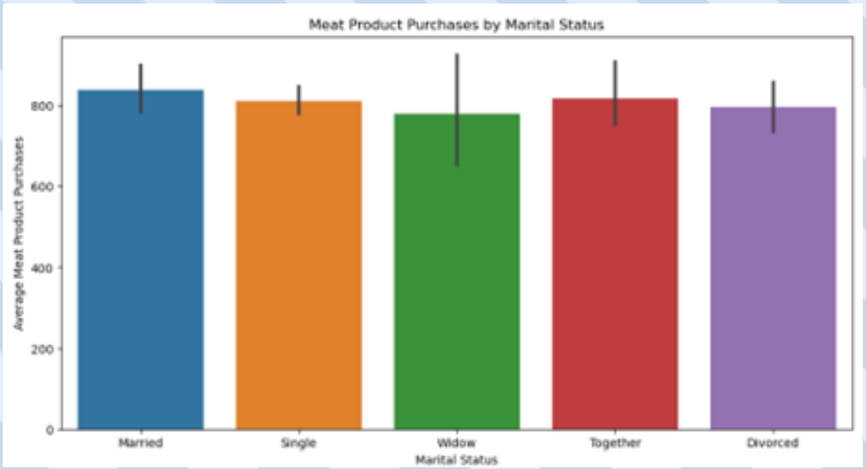
f) What are the characteristics of customers who are the highest buyers of Meat products?

We will take a rank, let's say for 100th top highest buyers (we are not considering the outliers). And then, we are going to take the mean as a single value to represent how those high meat buyers are.

The biggest consumer are Adult, Married or youth in a relationship

```
Characteristics of the highest buyers of meat products:
Average Income: 76614.59577964456
Average Kidhome: 0.015384615384615385
Average Teenhome: 0.03076923076923077
Average Recency: 52.207692307692305
Average MntWines: 674.1769230769231
Average MntFruits: 72.13846153846154
Average MntFishProducts: 95.84615384615384
Average MntSweetProducts: 70.4
Average MntGoldProds: 71.46923076923076
Average Response: 0.4153846153846154
Age Range Distribution:
Age_Range_Mapped
Adult          40
Young Adult    26
Middle-Aged    22
Pre-Senior     22
Senior         12
Youth           8
Name: count, dtype: int64
Education Level Distribution:
Education_Level
3    68
5    27
4    27
2     8
Name: count, dtype: int64
Marital Status Distribution:
Marital_Status_Encoded_Mapped
Married     49
Single      38
Together    33
Divorced     7
Widow        3
Name: count, dtype: int64
```





4) CONCLUSIONS

a)The business ought to go after new "Premium Buyer" with history of buying expensive goods like meat, gold, and wines, as well as high salaries. Premium offerings are probably going to be accepted by these clients.

b)The company can consistently rely on its existing "Loyal Diversified Buyers" - customers .

c)Targeting efforts are unlikely to yield significant benefits for low-income and inadequate education customers

d)The company should carefully evaluate the lifetime value potential of "Emerging Buyer" customers - new customers with mixed purchasing behavior - before investing heavily in acquisition, as they pose a risk of falling into the unproductive category described

4) FUTURE WORK

Now the challenge is using all information to create a closer and more agile relationship with customers. We can focus on 2 of the 5 main components of Marketing 5.0 that we can cover in our future work:

Predictive marketing, uses data and analytics to forecast the most effective marketing actions and strategies to drive business

Contextual marketing, is about using the various interfaces to analyze the physical environment as experienced by the user

5) REFERENCES

- Business.org. (2024, March 30). What is the 4P marketing matrix? https://www.business.org/marketing/sales/marketing-101-4p-matrix/
- Clientbook. (2024, March 29). How retail customer segmentation works. https://www.clientbook.com/blog/how-retail-customer-segmentation-works
- Digital Commerce 30. (2023, June 8). The retailer speaks: Digital marketing 2023. https://www.digitalcommerce30.com/2023/06/08/the-retailer-speaks-digital-marketing-2023/
- Hurra, T. (n.d.). DBSCAN — Make density-based clusters by hand. Towards Data Science. https://towardsdatascience.com/dbscan-make-density-based-clusters-by-hand-2689dc335120
- Perez-Font, R. (n.d.). What is marketing 5.0? LinkedIn. https://www.linkedin.com/pulse/what-marketing-50-ricardo-perez-font/
- Sthda. (n.d.). DBSCAN: Density-based clustering for discovering clusters in large datasets with noise. http://www.sthda.com/english/wiki/wiki.php?id_contents=7940
- Taneja, A., & Gupta, S. (2021). Marketing 5.0: The era of technology and the challenges faced by it. International Journal of Advanced Engineering and Management, 6(1), 1-6. https://ijaem.net/issue_dcp/Marketing%205.0%20The%20Era%20of%20Technology%20and%20the%20Challenges%20Faced%20By%20It.pdf

THANK YOU