# Marketing Segmentation

## Data Analysis and Visualization Business Report

_____

## Unsupervised Learning

## Contents / Agenda

- Executive Summary
- Data Overview
- Exploratory Data Analysis and Data Preprocessing
- Observations
- Model Building
- Conclusion
- Recommendations
- Appendix

## Executive Summary

In order for our marketing efforts to be successful and have a better return on investment dollars spent, we must have an understanding of certain aspects of who our customer is.

This information will have a direct consequence on the type of sales and marketing strategies we need to implement.

The goal of this analysis is to understand our existing customers purchasing habits and define the most profitable customer segments based on these purchasing habits.

The information extracted from the data will tell about the customers' behaviors, characteristics, and unique needs which will be used to craft an effective marketing campaign and product strategies.

This analysis revealed 2 sets of strongly correlated categories which hold the most monetary value.

• food and wine

• deals and online

It is my recommendation to first focus our marketing campaign on these 2 areas by offering a combination of the two and also a cross combination.

As an example, the data shows food and wine is a winning combination. I also suggest a cross combination offering of food deals with online wine sales in order to bringer a strong segment of one market into another area of focus.

Promoting highly-correlated products could potentially lead to increased overall sales.

There is enough data to use this information for multiple marketing campaigns across various customer groups.

# Data Overview

**All data types are correct**

24 numerical columns (all discrete, except income = continuous)
(some numerical columns are categorical)

3 object columns (2 = categorical, 1 = date )

**Missing values are not significant.**

numerical = income has 1%

categorical = no missing customers in the categorical data
2240/2240

DATA COLLECTION YEAR = 2016

**Data Shape**

- Original data shape: (2240, 27)

- Filtered data shape: (2240, 27)
    This is a copy created to look at outliers.

**MADE SOME ADJUSTMENTS TO THE FILTERED DATA SET**

Condensed some categories

    Education:
            Placed "2n Cycle" in the "Master" category

    Marital_Status:
            Placed "Alone", "Absurd, "YOLO" in "Single" category

    Added
            "quarter" category to look at enrollment dates

    Removed
            income outliers

**NEW Filtered data shape: (2182, 28)**

        Separated into 2 data sets due to significant outliers

**No outlier data shape: (1336, 28)**

- separated significant outliers in monetary food purchase categories

**Outlier only data shape: (846, 28)**

- created a new data set with significant outliers to use as a segmentation category

# Exploratory Data Analysis and Data Preprocessing

**<u>Univariate Analysis of Income  (histogram and box plot)</u>**
- Range approx 35 - 80k
- Histogram and box plot showed outliers
- income for those were 118k
- Were 8 customers in that group
- Checked to see what income percentage 99.5% of customers were under.
    - 99.5% are below $102,145 so I am removing these 8 from the data set.

**<u>Univariate Analysis of Amout Spent on Food Items  (histogram)</u>**
- A lot of people spent a very small amount where few people spent a much lesser amount

**<u>Univariate Analysis of Categorical Data  (bar plot)</u>**
- 64.4% customers are in a relationship
- Education is nearly 50/50 degree vs no degree
- Most without children or teens, although not an overwhelming marginal difference
-  If they do have a child, it's only 1

## FEATURE ENGINEERING TECHNIQUES
- Combined kids in the home and teens in the home into one category = Kids
- Combined marital status into 2 categories of single or relationship and is now = Relationship Status

- Created a new category which combines the Relationship Status + Kids to = Family Size
- Combined the amounts spent on all food categories to = Expenses which = 1435
- Combined all the modes of purchases into one category = Number of Total Purchases = 137
- Took the enrollment dates to create a number of days a customer has been with the company to = 1089
- Dropped some columns not needed in order to create models.
- Created AVG INCOME IN RANGES

  - 0.0 -  25,000
    - 12,500

  - 25,000 - 50,000
    - 37,500

  - 50,000 - 75,000
    - 62,500

  - 75,000 - 100,000
    - 87,500

## FEATURE ENGINEERING OBSERVATION

> A larger family size contributes to an increase in income.
> - larger households are consistent with those in a relationship
> - more than one adult is contributing to the household income.

# Observations
## Upper and Lower 5 Observations

**AGE RANGE**
- birth years between 1946 & 1984

**EDUCATION**
- Grad - Phd - Master

**FAMILY DYNAMICS**
- married - single -together - divorced
- 0 to 2 kids home - 2 = 1946 (caring for grandkids)
- teens home 0 - 1 (54,67,46,56,54 - caring for grandkids)
- income between 26,645 (1984) & 71,613

**ENROLLMENT DATE WITH COMPANY**
- 2012 - 2014
- 1st Q = 5, 2nd Q = 2, 3 Q = 1, 4 Q = 1

**DAYS SINCE LAST PURCHASE** (recency)
- 8  = 1 teen home, 1956, master, 69k, $428 in wine
- 94 = 81, phd, 1 child, 58k, $173 wine

**MODE OF PURCHASE**
- catalog purchase (most all on low end)
  - 0 = 1984, 1 child, lowest income
  - 10 = 1957, 58k, single, 635 wine, 58 days since last buy
- store purchase 2 - 13

**FREQUENCY**
- visit /mo 3 - 7

**CAMPAIGN OPT INS**
- 1st campaign = 1 accept
- 2nd campaign = 0 accepts
- 3rd campaign = 0 accepts
- 4th campaign = 1 accept
- 5th campaign = 0 accepts
- response (campaign) (accepted) = 0

*last campaign (assuming would be 6) is missing from data.
- Head and tail showed 2 accepted "last" marketing
- 5th campaign = 0 accepts
- is the last one accounted for in my data

**COMPLAINTS** = 0

_____

## OVERALL CATEGORICAL OBSERVATIONS
HISTOGRAM & BOX PLOT ANALYSIS OBSERVATION:
- 2 groups of customers have emerged
    - 1. Typical (majority) who spent small amount
    - 2. Outlier (minority) who spent nearly as much as the typical grouping


STRATEGY:
1. segment customers
    - A. typical purchaser group
    - B. outlier purchaser group


OBSERVATION:
Filtered data shape - has outliers: (2182, 28)
No outlier data shape: (1336, 28)
Outlier only data shape: (846, 28)
- Total observations in the 2 new data sets = 2182

- 28 original columns

2. Use clustering on these 2 to better understand the relationship with the other categories

_____

## INCOME OBSERVATIONS
- Outliers are beyond the 118k whisker
- Majority is under 102k
- count and % are insignificant

STRATEGY: Filter out outliers
CONCLUSION:
- Histogram data is now evenly distributed

### *FIRST CATEGORY TO FILTER OUTLIERS*

**Income**   1974  unique    (use this in range form)
HISTOGRAM & BOX PLOT OBSERVATION:
Upper Whisker Limit: 118350.5
99.5% of observations in 'Income' are under: 102145.75
Number of outliers in 'Income': 8 out of 2240
Outlier Percentage in 'Income': 0.36%

_____

## WINE PRODUCT OBSERVATIONS
- Performed 2 paces of outlier removal and data is still skewed

STRATEGY: Will need to use other method of standardization

### *SECOND CATEGORY TO FILTER OUTLIERS*

MntWines 776                 (use this for an avg)
HISTOGRAM & BOX PLOT OBSERVATION:
Upper Whisker Limit: 1226.5
99.5% of observations in 'MntWines' are under: 1374.35
Number of outliers in 'MntWines': 35 out of 2232
Outlier Percentage in 'MntWines': 1.57%
- OBSERVATION:
  - Outliers make up very small percentage of data
- DECISION:
  - Filter out outliers
- CONCLUSION:
  - Filtering out the outliers didn't evenly distribute the data.
RECHECKED DUE TO OUTLIERS REMAINING IN THE NEW BOXPLOT. DID A
SECOND ITERATION OF FILTERING BASED ON THE NUMBERS BELOW
Upper Whisker Limit: 1173.0
Number of outliers in 'MntWines': 15 out of 2197

Outlier Percentage in 'MntWines': 0.68%
- OBSERVATION:
  - This took care of most of the outliers
  CONCLUSION:
  - Data is still skewed.

---

## FRUIT PRODUCT OBSERVATIONS

- 10% of the data are outliers
- I am not filtering the outliers

STRATEGY: Create 2 categories from this group

*THIRD CATEGORY TO FILTER OUT OUTLIERS*

MntFruits 158            (use this for an avg)
HISTOGRAM & BOX PLOT OBSERVATION:
      Upper Whisker Limit: 80
      99.5% of observations in 'MntFruits' are under: 185.0
      Number of outliers in 'MntFruits': 224 out of 2182
      Outlier Percentage in 'MntFruits': 10.27%
- OBSERVATION:
  - Upper whisker is 80 and box plot shows concentration is in the lower range
  - However the percentage observations indicate a large amount was spent outside the upper limit.
  - Removing second run of outliers in the Wine category didn't have a huge effect on the stats for the next category of fruit.
    - The whisker limit and % observations were nearly unaffected. It decreased number of outliers by 20 and the outlier percentage decrease was approx. 1%

---

## MEAT PRODUCT OBSERVATIONS

- Not filtering outliers

STRATEGY: Create 2 categories from this group

*FOURTH CATEGORY TO FILTER OUT OUTLIERS*

<u>MntMeatProducts 558</u>        <u>(use this for an avg)</u>
HISTOGRAM & BOX PLOT OBSERVATION:
    Upper Whisker Limit: 520.0
    99.5% of observations in 'MntMeatProducts' are under: 936.38
    Number of outliers in 'MntMeatProducts': 191 out of 2182
    Outlier Percentage in 'MntMeatProducts': 8.75%
- OBSERVATION:
  - Upper whisker @ 520
  - Difference in upper whisker limit and 99.5% = 416
  - Need to look at this group

_____

## FISH PRODUCT OBSERVATIONS

- Not filtering outliers

STRATEGY: Create 2 categories from this group

*FIFTH CATEGORY TO FILTER OUT OUTLIERS*

<u>MntFishProducts 182</u>        <u>(use this for an avg)</u>
HISTOGRAM & BOX PLOT:
    Upper Whisker Limit: 118.0
    99.5% of observations in 'MntFishProducts' are under: 242.38
    Number of outliers in 'MntFishProducts': 231 out of 2182
    Outlier Percentage in 'MntFishProducts': 10.59%
- OBSERVATION:
  - upper whisker @118
  - Difference in upper whisker limit and 99.5% = 124
  - Need to look at this group

_____

## SWEET PRODUCT OBSERVATIONS

- Not filtering the outliers

STRATEGY: Create 2 categories from this group

*SIXTH CATEGORY TO FILTER OUT OUTLIERS*

MntSweetProducts 177       (use this for an avg)
HISTOGRAM & BOX PLOT:
     Upper Whisker Limit: 78.5
     99.5% of observations in 'MntSweetProducts' are under: 192.0
     Number of outliers in 'MntSweetProducts': 252 out of 2182
     Outlier Percentage in 'MntSweetProducts': 11.55%
- OBSERVATION:
  - upper whisker @ 78.5
  - Difference in upper whisker limit and 99.5% = 113.5
  - Need to look at this group

_____

# GOLD PRODUCT OBSERVATIONS
- Not filtering the outliers

STRATEGY: Create 2 categories from this group

*SEVENTH CATEGORY TO FILTER OUT OUTLIERS*

MntGoldProds 213           (use this for an avg)
HISTOGRAM & BOX PLOT:
Upper Whisker Limit: 126.5
99.5% of observations in 'MntGoldProds' are under: 241.10
Number of outliers in 'MntGoldProds': 197 out of 2182
Outlier Percentage in 'MntGoldProds': 9.03%
- OBSERVATION:
  - upper whisker @ 126.5
  - Difference in upper whisker limit and 99.5% = 114.60
  - Need to look at this group

## FAMILY DYNAMICS OBSERVATIONS

- 64.4% customers are in a relationship/have a family grouping
- Education is nearly 50/50 degree vs no degree
- Standard characteristic - children
  - Outlier characteristic - no children

- Education & Marital Status —> Income
  - Outliers have more income than the Standard group by up to 30k more
  - Outlier income same for PhD, Master, Graduation & all marital statuses

STRATEGY: grouping = children / no children

higher income = no children

75k+ - no children

50 - 75k - 1-2 children

less than 50k - 2-3 children

**Education**   ~~5~~  4  unique  (not relevant to looking at habits)(decision tree)

UNIQUE DATA & BAR PLOT OBSERVATION

UNFILTERED DATA

- Graduation = 50.4% = graduation
- Master = 25.6% = masters
- PhD = 21.6% = Phd
- Basic = 2.4%

NO OUTLIERS

- Same order
- Percentage difference not significant

OUTLIERS ONLY

- Same order
- Percentage difference not significant

CONCLUSION:

- Outliers not a factor for this grouping

CATEGORICAL CORRELATION OBSERVATION

Education & Income - Earning from greatest to least

NO OUTLIERS

- PhD - 45k

- Master - 40k
- Graduation - 38k
- Basic - 20k

OUTLIERS
- PhD -69k
- Graduation - 68,5k
- Master - 68k
- Basic - 31k

———————————————————————

*Combined Together + Married to = Relationship
  and Single, Divorced, and Widow to = Single*

**Marital_Status**   ~~8~~ 5  unique   (not relevant to looking at habits)(decision tree)
UNIQUE DATA & BAR PLOT OBSERVATION
UNFILTERED DATA
- Married = 38.6%
- Together = 25.8%
- Single = 21.8%
- Divorced = 10.3%
- Widow = 3.4%

NO OUTLIERS
- Same order
- Percentage difference not significant

OUTLIERS ONLY
- Same order
- Percentage difference not significant

CONCLUSION:
- Outliers not a factor for this grouping

CATEGORICAL CORRELATION OBSERVATION
Marital Status & Income - Earning from greatest to least
NO OUTLIERS
- Widow - 45k
- Married - 40k
- Divorced - 40k
- Together - 38k

- Single - 37k

OUTLIERS
- Married - 69k
- Together - 68,5k
- Divorced - 68,5k
- Single - 68k
- Widow - 67k

_____

*Combined Kidhome and Teenhome to = Kids*

**Kidhome**     3 unique     (0 - 2 range) (could be relevant to looking at habits) (decision tree)
UNIQUE DATA & BAR PLOT OBSERVATION
UNFILTERED DATA
- 0 = 57.7%
- 1 = 40.1%
- 2 = 2.2%
NO OUTLIERS
- 0 = 36.8%
- 1 = 59.8%
- 2 = 3.4%
OUTLIERS ONLY
- 0 = 88.9%
- 1 = 10.8%
- 2 = .4%
CONCLUSION: dominant
- 0 = Outlier group
- 1 = Standard group
- 2 = Standard group - but small %
- Outlier Characteristic
    - no children
- Standard Characteristic
-       children

**Teenhome**   3  unique  (0 - 2 range) (could be relevant to looking at habits) (decision tree)
UNIQUE DATA & BAR PLOT OBSERVATION
UNFILTERED DATA
- 0 = 51.6%

- 1 = 46.1%
- 2 = 2.3%

NO OUTLIERS
- 0 = 45.7%
- 1 = 51.8%
- 2 = 2.5%

OUTLIERS ONLY
- 0 = 59.9%
- 1 = 38.2%
- 2 = 1.9%

CONCLUSION: dominant
- 0 = Outlier group
- 1 = Standard group
- 2 = Standard group - but small %
- Outlier Characteristic
    - no children
- Standard Characteristic
    - children

_____

*Assigned Relationship Status a numerical value, then combined Relationship Status + Kids to = Family Size.*

_____

## CUSTOMER ENROLLMENT DATE OBSERVATIONS
- small degree of variance amongst the quarters
- 4th Q had more enrollments
- 3rd Q had the least
- 1st enrollment = 1089 days
- last enrollment = 26 days

STRATEGY: Although there is a 2 approaches: to raise 3rd Q enrollment and/or to capitalize on 4th Q when customers enrollment has historically been higher.

**Dt_Customer**   663   unique
*(Already know there are 69 unique observations. Want to know the enrollment count in <u>each quarter to see if there is an pattern we can capitalize on.</u>)*

UNIQUE DATA & BAR PLOT OBSERVATION
UNFILTERED DATA
- 4 = 26.7%     174
- 1 = 25.9%     163
- 2 = 24.3%     170
- 3 = 23.1%     156

NO OUTLIERS
- 1 = 25.7%
- 2 = 25.4%
- 3 = 22.2%
- 4 = 26.6%

OUTLIERS ONLY
- 1 = 26.5%
- 2 = 22.7%
- 3 = 24%
- 4 = 26.8%

CONCLUSION:
- Overall well balanced

_____

## COMPLAINT OBSERVATIONS
- Basically no complaints in the past 2 years
- Happy customers

STRATEGY: Capitalize on this in the marketing campaign

**Complain**     2   unique   (0 or 1 - not relevant to looking at habits)(decision tree)
UNIQUE DATA & BAR PLOT OBSERVATION
UNFILTERED DATA
- 99.1 % complaint free in past 2 years.

NO OUTLIERS
- NO Percentage difference

OUTLIERS ONLY
- NO Percentage difference

CONCLUSION:
- Outliers are not a  factor for this grouping

## AGE OBSERVATIONS

- Customers caring for grandchildren

STRATEGY:  Capitalize on the pain points of older customers with younger children to care for
- easy, diverse, healthy meals
- items kids can help themselves with

Year_Birth     59 unique
HEAD & TAIL OBSERVATION:
- Some older customers have children and/or teens in the home
- Check age and children in the home. Could be a market segment.

UNIQUE OBSERVATION:
- Broad customer age range.
- Look at habits of older customers vs habits of younger customers.

_____

## RECENTLY VISITED OBSERVATIONS

- The recency of visits to the store didn't show to have an impact on their purchasing habits

Recency     100 unique
HEAD & TAIL OBSERVATION:
- Do frequent shoppers spend more money on indulgences, such as wine?

UNIQUE OBSERVATION:
- 100 unique amounts of time
  - out of the 2240 customers
- Segment into what these ranges are
- Get an idea of numbers on possibly who shops often vs who doesn't shop often
- Use this to compare to
  - where they shop
  - when they shop

- ex: shop weekly & mostly in store

_____

## MODE OF PURCHASE OBSERVATIONS
- Balanced across modes
- Number of products purchased by the customers = 137 avg

STRATEGY: Craft marketing offers to the shopping mode for the age group based on the driving factor for that segment

The  total.

Maximum Amount Per Purchase: 137.0


NumDealsPurchases 15     (use this for avg / range / in ratio to)

NumCatalogPurchases 14 (use this for avg / range / in ratio to)
NumStorePurchases 14     (use this for avg / range / in ratio to)
NumWebPurchases 15       (use this for avg / range / in ratio to)

NumWebVisitsMonth 16    (use this for avg / range / in ratio to)
HEAD & TAIL OBSERVATION:
- See if there is a correlation to customers who have been enrolled a long time and how often they are shopping.


_____

## CAMPAIGN OPT IN OBSERVATIONS
*last campaign (assuming would be 6) is missing from data.
- Head and tail showed 2 accepted "last" marketing

- 5th campaign = 0 accepts
- is the last one accounted for in my data


STRATEGY: The last campaign had more opt ins than others so mirror similarities

AcceptedCmp3   2 (0 or 1 - not relevant to looking at habits)(decision tree)
AcceptedCmp4   2 (0 or 1 - not relevant to looking at habits)(decision tree)
AcceptedCmp5   2 (0 or 1 - not relevant to looking at habits)(decision tree)
AcceptedCmp1   2 (0 or 1 - not relevant to looking at habits)(decision tree)
AcceptedCmp2   2 (0 or 1 - not relevant to looking at habits)(decision tree)
Response          2 (0 or 1 - not relevant to looking at habits)(decision tree)

## HEAD & TAIL OBSERVATION:

- 1st campaign = 1 accept

- 2nd campaign = 0 accepts

- 3rd campaign = 0 accepts

- 4th campaign = 1 accept

- 5th campaign = 0 accepts

- response (campaign) (accepted) = 0

## OVERALL OBSERVATION

- Maximum amount of products purchased per purchase = 137

- Max expenses for all categories past 2 yrs = 1435

- Amount / purchase, distributed over time = more frequent shoppers average spending less

- Higher earners spend more.

- Expenses, distributed over time = 2-4 wks for majority

- Larger family size = increase in income

- Max days customer has been with company = 1035

- Max time with no engagement = 1089 days

_____

# Model Building

## Comparison of Techniques and their Performances

I used the filtering method for outliers because it creates a view of the original DataFrame, not a new DataFrame. The whole of the data frame is needed for running certain models so I wanted to maintain the integrity of the true dataset.

Even though some categories remain highly skewed after filtering out the outliers,  I opted to not standardize the data in that category because doing so would significantly affect the relationship with the other categories. I need to see their relationship with the other categories to accurately segment customers.

I ran comparison in categories without outliers and with outliers, documented the findings, then filtered outliers and removed unnecessary variables in order to run models.

## SCATTER PLOT OBSERVATION

- There is a positive correlation between income and expenses. As income increases, so does the amount of expenses the customer has in the store.

## HEATMAP OBSERVATION

- Grouped and iterated over and over based on the heat map findings before irrelevant variables were removed

STRATEGY: The last campaign had more opt ins than others so mirror similarities. Use the common connections within the groups to craft personal and custom targeted offers.

Complain Least
      Yr Birth
      Money Spent on Gold Prods
      Wine

Kids in Home
Teens in Home
Income

Number of Store Purchases
Money Spent on Meat
Kids in Home
Money Spent on Gold Prod
Number of Web Purchases

Money Spent on Gold Prod
Money Spent on Meat
Number of Purchases Catalog

Number of Web Visits / Mo
Income
Number of Deals Purchased

Money Spent on Wine
Income
Number Web Purchases
Number of Store Purchases
Number of Catalog Purchases
Accepted Campaign 3

Response to Last Campaign
Accepted Campaign 3
Recency Since Last Purchase

Money Spent on Fruit
Money Spent on Sweets
Money Spent on Fish

Accepted Campaign 5
Accepted Campaign 1
Accepted Campaign 2

_____

- Heat map findings after irrelevant variables were removed

## Strong Positive Correlations:

- Wine & Overall Total:
  - Customers who spend more on wine have higher total spending across all product categories.
  - Wine is likely a significant part of the total spending.

- Fruit, Meat Products, Fish Products, Sweet Products, GoldProds & Overall Total:
  - Customers who purchase more of these products tend to spend more overall.

## Moderate Positive Correlations:

- Number of Deals Purchased & Number of Web Purchases:
  - Moderate correlation
  - Customers who make more purchases using deals also make more purchases online.

- Number of Web Visits per Month &  Number of Web Purchases:
  - Customers who visit the website more often tend to make more purchases online.

## Weak or No Correlations:

- Most other feature combinations show weak or no significant correlations, implying that they are relatively independent of each other. For example, Recency (number of days since last purchase) has little correlation with most other features.

## OVERALL HEATMAP OBSERVATION:

The heatmap indicates overall spending is mostly on wine and other product categories. Customers who engage more with deals and online purchases tend to have higher overall spending. Recency of purchases doesn't seem to have a strong relationship with spending habits in this dataset.

---

## CLUSTERING OBSERVATION

- T-SNE clustering showed organized groups tracking along a predictable pattern.
- PCA was performed to organize the data along an even scale for all categories
- 3 categories emerged enabling the perfect environment for K Means clustering.

---

## CLUSTERING & BOX PLOT PROFILING OBSERVATION

- As a whole, this method didn't have any stand out categories as to why this method gave any insight into segmentation.

## Proposal for the Final Solution Design

Heatmap and K Means clustering are the best models to use for segmentation.

The heat map gave clear correlations between categories.

K Means clearly showed 3 saturated categories of customers.

## Conclusions

By using past information to learn about customer engagement with various marketing activities, above the line activities and below the line campaigns, and targeted personalized offers, a successful marketing campaign can be built around what resonates with the customer base.

A main marketing campaign can be developed and slightly modified for each of the saturated categories of customers

## Recommendations

There are 2 sets of strongly correlated categories which hold the most monetary value.
• food and wine
• deals and online

Focusing on promoting highly-correlated products could potentially lead to increased overall sales.

It is my recommendation to begin with these 2 areas by offering a combination of the two and also a cross combination.

As an example, the data shows food and wine is a winning combination. I also suggest a cross combination offering of food deals with online wine sales in order to bringer a strong segment of one market into another area of focus.

# Appendix

## CHECKING ORIGINAL DATA SET AND FILTERED DATA SET BEFORE ANY CHANGES

**original_data**

| # | Column | Count | Non-Null | Dtype |
|---|--------|-------|----------|-------|
| 0 | ID | 2240 | non-null | int64 |
| 1 | Year_Birth | 2240 | non-null | int64 |
| 2 | Education | 2240 | non-null | object |
| 3 | Marital_Status | 2240 | non-null | object |
| 4 | Income | 2216 | non-null | float64 |
| 5 | Kidhome | 2240 | non-null | int64 |
| 6 | Teenhome | 2240 | non-null | int64 |
| 7 | Dt_Customer | 2240 | non-null | object |
| 8 | Recency | 2240 | non-null | int64 |
| 9 | MntWines | 2240 | non-null | int64 |
| 10 | MntFruits | 2240 | non-null | int64 |
| 11 | MntMeatProducts | 2240 | non-null | int64 |
| 12 | MntFishProducts | 2240 | non-null | int64 |
| 13 | MntSweetProducts | 2240 | non-null | int64 |
| 14 | MntGoldProds | 2240 | non-null | int64 |
| 15 | NumDealsPurchases | 2240 | non-null | int64 |
| 16 | NumWebPurchases | 2240 | non-null | int64 |
| 17 | NumCatalogPurchas | 2240 | non-null | int64 |
| 18 | NumStorePurchases | 2240 | non-null | int64 |
| 19 | NumWebVisitsMonth | 2240 | non-null | int64 |
| 20 | AcceptedCmp3 | 2240 | non-null | int64 |
| 21 | AcceptedCmp4 | 2240 | non-null | int64 |
| 22 | AcceptedCmp5 | 2240 | non-null | int64 |
| 23 | AcceptedCmp1 | 2240 | non-null | int64 |
| 24 | AcceptedCmp2 | 2240 | non-null | int64 |
| 25 | Complain | 2240 | non-null | int64 |
| 26 | Response | 2240 | non-null | int64 |

dtypes:float64(1),    int64(23),    object(3)

**filtered_data_copy**

| # | Column | Count | Non-Null | Dtype |
|---|--------|-------|----------|-------|
| 0 | ID | 2240 | non-null | int64 |
| 1 | Year_Birth | 2240 | non-null | int64 |
| 2 | Education | 2240 | non-null | object |
| 3 | Marital_Status | 2240 | non-null | object |
| 4 | Income | 2216 | non-null | float64 |
| 5 | Kidhome | 2240 | non-null | int64 |
| 6 | Teenhome | 2240 | non-null | int64 |
| 7 | Dt_Customer | 2240 | non-null | object |
| 8 | Recency | 2240 | non-null | int64 |
| 9 | MntWines | 2240 | non-null | int64 |
| 10 | MntFruits | 2240 | non-null | int64 |
| 11 | MntMeatProducts | 2240 | non-null | int64 |
| 12 | MntFishProducts | 2240 | non-null | int64 |
| 13 | MntSweetProducts | 2240 | non-null | int64 |
| 14 | MntGoldProds | 2240 | non-null | int64 |
| 15 | NumDealsPurchases | 2240 | non-null | int64 |
| 16 | NumWebPurchases | 2240 | non-null | int64 |
| 17 | NumCatalogPurchases | 2240 | non-null | int64 |
| 18 | NumStorePurchases | 2240 | non-null | int64 |
| 19 | NumWebVisitsMonth | 2240 | non-null | int64 |
| 20 | AcceptedCmp3 | 2240 | non-null | int64 |
| 21 | AcceptedCmp4 | 2240 | non-null | int64 |
| 22 | AcceptedCmp5 | 2240 | non-null | int64 |
| 23 | AcceptedCmp1 | 2240 | non-null | int64 |
| 24 | AcceptedCmp2 | 2240 | non-null | int64 |
| 25 | Complain | 2240 | non-null | int64 |
| 26 | Response | 2240 | non-null | int64 |

dtypes:float64(1),    int64(23),    object(3)

## NUMERIC SUMMARY

```
ID  Year_Birth     Income  Kidhome  Teenhome  Recency  MntWines  \
count    2240.00    2240.00    2216.00   2240.00    2240.00   2240.00    2240.00
mean     5592.16    1968.81   52247.25      0.44       0.51     49.11     303.94
std      3246.66      11.98   25173.08      0.54       0.54     28.96     336.60
min         0.00    1893.00    1730.00      0.00       0.00      0.00       0.00
25%      2828.25    1959.00   35303.00      0.00       0.00     24.00      23.75
50%      5458.50    1970.00   51381.50      0.00       0.00     49.00     173.50
75%      8427.75    1977.00   68522.00      1.00       1.00     74.00     504.25
max     11191.00    1996.00  666666.00      2.00       2.00     99.00    1493.00

        MntFruits  MntMeatProducts  MntFishProducts  ...  NumCatalogPurchases
\
count    2240.00          2240.00          2240.00   ...              2240.00
mean       26.30           166.95            37.53   ...                 2.66
std        39.77           225.72            54.63   ...                 2.92
min         0.00             0.00             0.00   ...                 0.00
25%         1.00            16.00             3.00   ...                 0.00
50%         8.00            67.00            12.00   ...                 2.00
75%        33.00           232.00            50.00   ...                 4.00
max       199.00          1725.00           259.00  ...                28.00

        NumStorePurchases  NumWebVisitsMonth  AcceptedCmp3  AcceptedCmp4  \
count            2240.00            2240.00       2240.00       2240.00
mean                5.79               5.32          0.07          0.07
std                 3.25               2.43          0.26          0.26
min                 0.00               0.00          0.00          0.00
25%                 3.00               3.00          0.00          0.00
50%                 5.00               6.00          0.00          0.00
75%                 8.00               7.00          0.00          0.00
max                13.00              20.00          1.00          1.00

        AcceptedCmp5  AcceptedCmp1  AcceptedCmp2  Complain  Response
count        2240.00       2240.00       2240.00   2240.00   2240.00
mean            0.07          0.06          0.01      0.01      0.15
std             0.26          0.25          0.11      0.10      0.36
min             0.00          0.00          0.00      0.00      0.00
25%             0.00          0.00          0.00      0.00      0.00
50%             0.00          0.00          0.00      0.00      0.00
75%             0.00          0.00          0.00      0.00      0.00
max             1.00          1.00          1.00      1.00      1.00

[8 rows x 24 columns]
```

## CHECKING ORIGINAL DATA SET AND FILTERED DATA SET AFTER CHANGES

- Red shows 2 categories which were condensed and one category added for quarter of enrollment
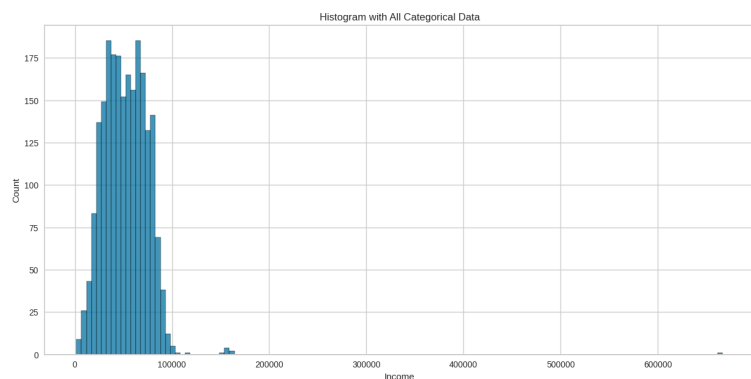
| original_data | | | filtered_data | |
|---|---|---|---|---|
| **Column** | **Unique Values** | | **Column** | **Unique Values** |
| ID | 2240 | | ID | 2240 |
| Year_Birth | 59 | | Year_Birth | 59 |
| Education | 5 | | Education | 4 |
| Marital_Status | 8 | | Marital_Status | 5 |
| Income | 1974 | | Income | 1974 |
| Kidhome | 3 | | Kidhome | 3 |
| Teenhome | 3 | | Teenhome | 3 |
| Dt_Customer | 663 | | Dt_Customer | 663 |
| Recency | 100 | | Recency | 100 |
| MntWines | 776 | | MntWines | 776 |
| MntFruits | 158 | | MntFruits | 158 |
| MntMeatProducts | 558 | | MntMeatProducts | 558 |
| MntFishProducts | 182 | | MntFishProducts | 182 |
| MntSweetProducts | 177 | | MntSweetProducts | 177 |
| MntGoldProds | 213 | | MntGoldProds | 213 |
| NumDealsPurchases | 15 | | NumDealsPurchases | 15 |
| NumWebPurchases | 15 | | NumWebPurchases | 15 |
| NumCatalogPurchases | 14 | | NumCatalogPurchases | 14 |
| NumStorePurchases | 14 | | NumStorePurchases | 14 |
| NumWebVisitsMonth | 16 | | NumWebVisitsMonth | 16 |
| AcceptedCmp3 | 2 | | AcceptedCmp3 | 2 |
| AcceptedCmp4 | 2 | | AcceptedCmp4 | 2 |
| AcceptedCmp5 | 2 | | AcceptedCmp5 | 2 |
| AcceptedCmp1 | 2 | | AcceptedCmp1 | 2 |
| AcceptedCmp2 | 2 | | AcceptedCmp2 | 2 |
| Complain | 2 | | | |
| Response | 2 | | | |

# INCOME



Histogram with All Categorical Data



Boxplot with All Categorical Data



Histogram with Outliers Highlighted
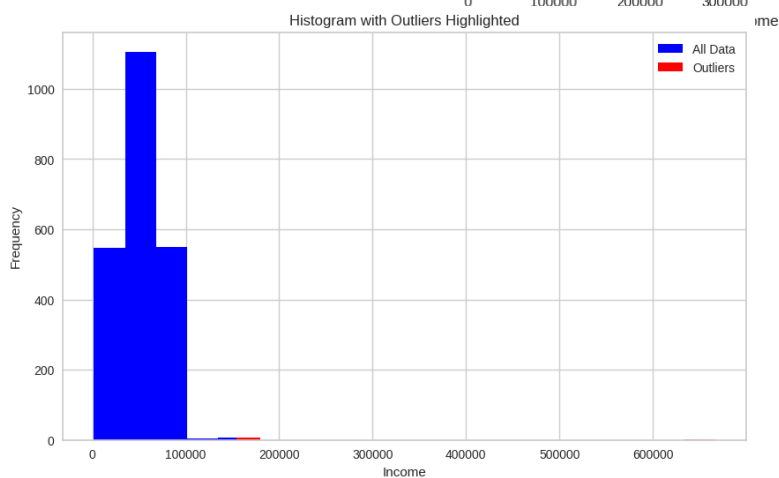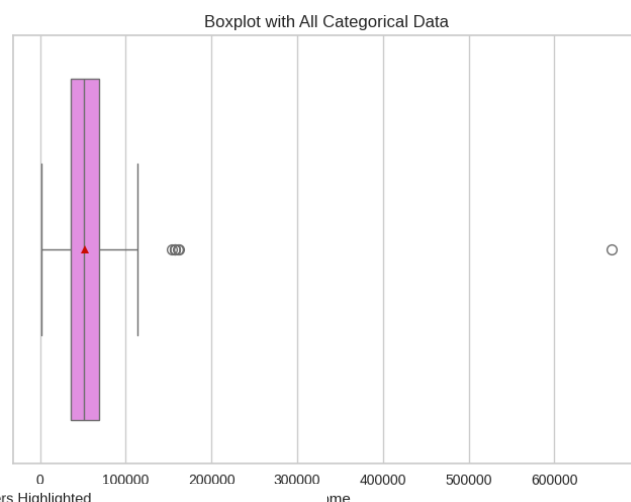
# AFTER OUTLIER FILTERING



Histogram with Outliers Filtered Out



Boxplot with Outliers Filtered Out
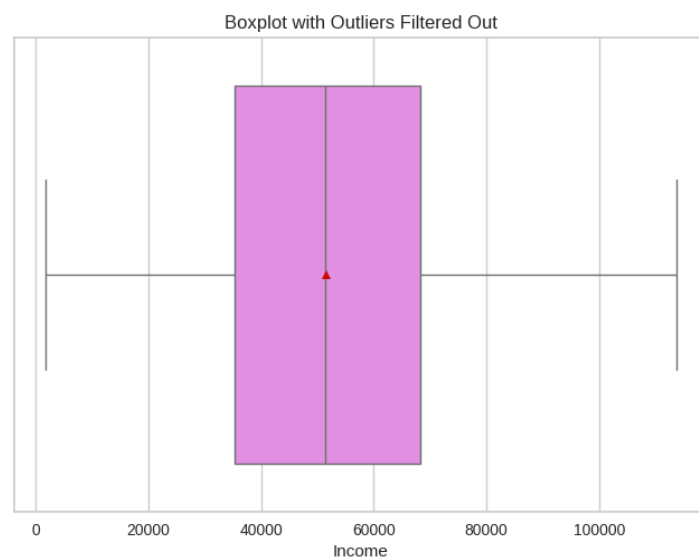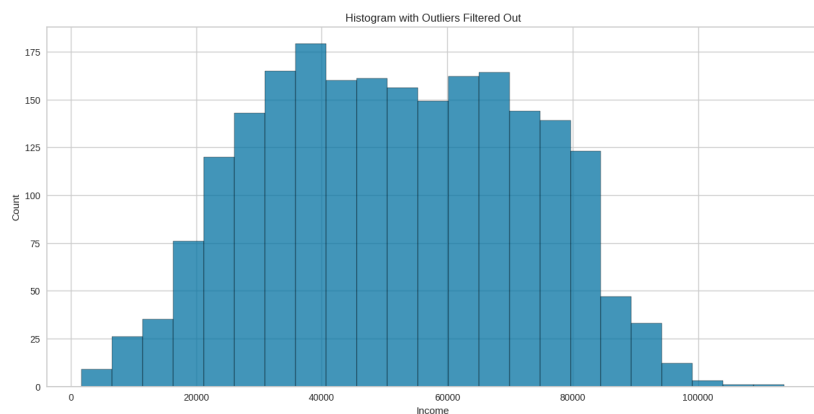
```
Original data shape: (2240, 27)
        (removed 58 outliers)

Filtered data shape: (2232, 28)
```
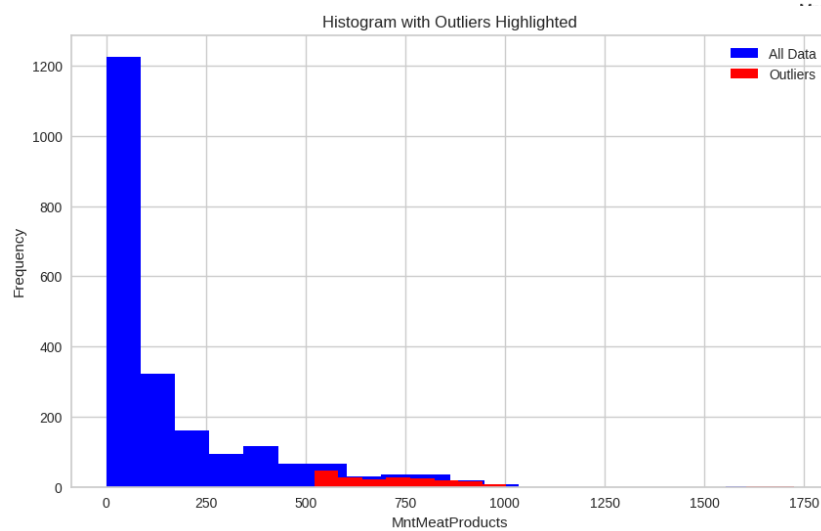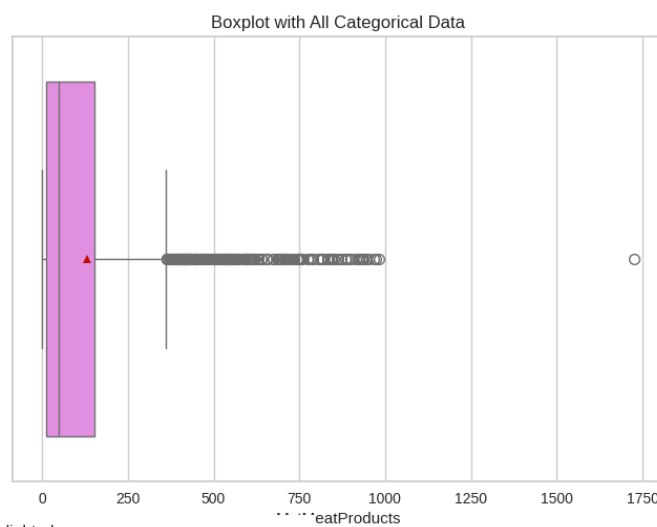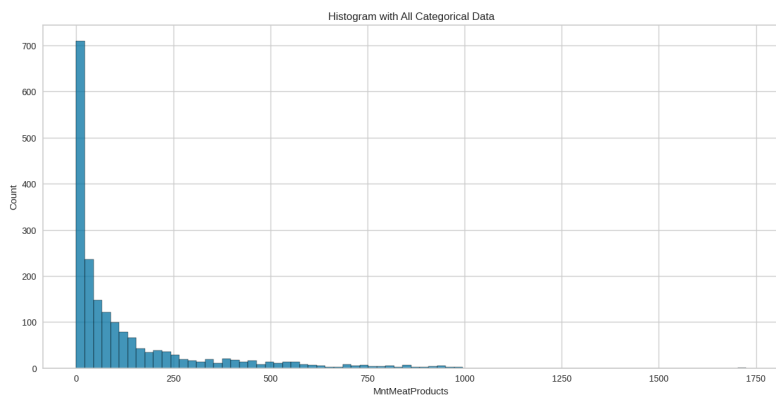
# MONEY SPENT ON FRUIT

Histogram with All Categorical Data

Boxplot with All Categorical Data

Histogram with Outliers Highlighted

Original data shape: (2240, 27)
Filtered data shape: (2182, 28)

# MONEY SPENT ON MEAT

Histogram with All Categorical Data

Boxplot with All Categorical Data

Histogram with Outliers Highlighted

```
Original data shape: (2240, 27)
Filtered data shape: (2182, 28)
```

# MONEY SPENT OF FISH

Histogram with All Categorical Data

Boxplot with All Categorical Data

Histogram with Outliers Highlighted

- All Data
- Outliers

```
Original data shape: (2240, 27)
Filtered data shape: (2182, 28)
```

## MONEY SPENT ON SWEETS



Histogram with All Categorical Data



Boxplot with All Categorical Data



Histogram with Outliers Highlighted
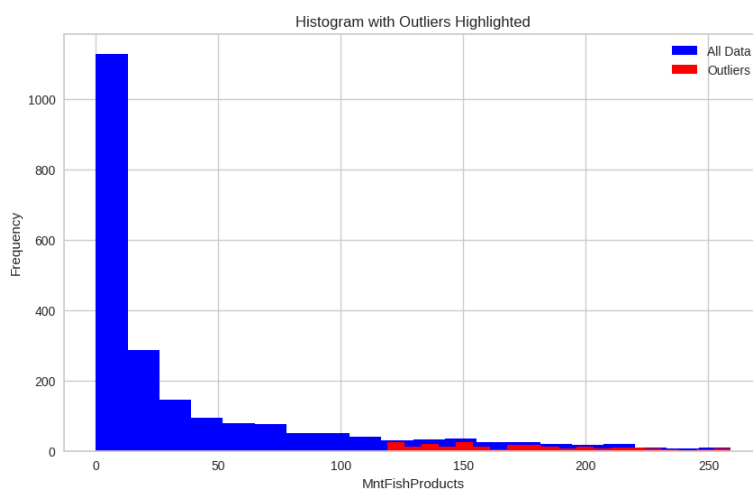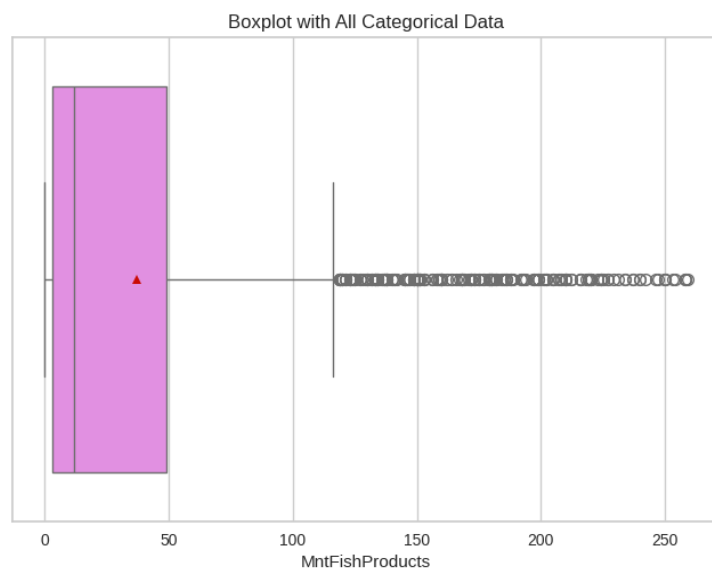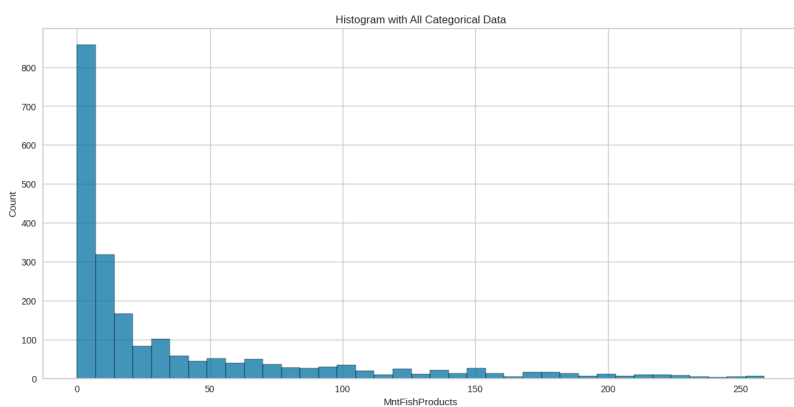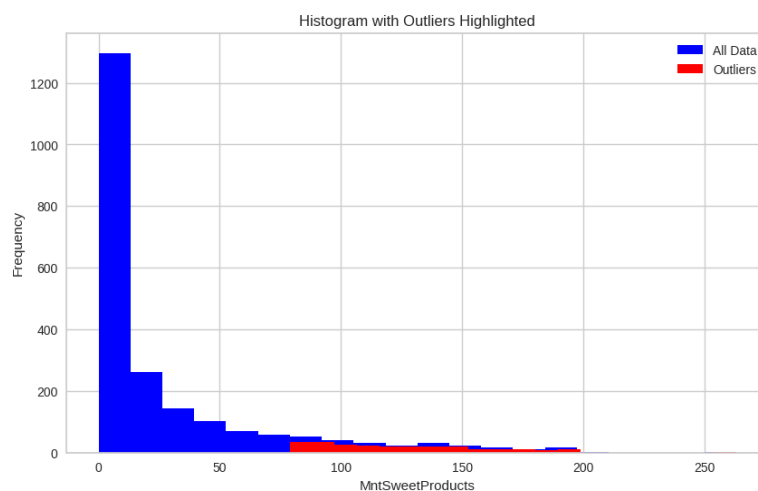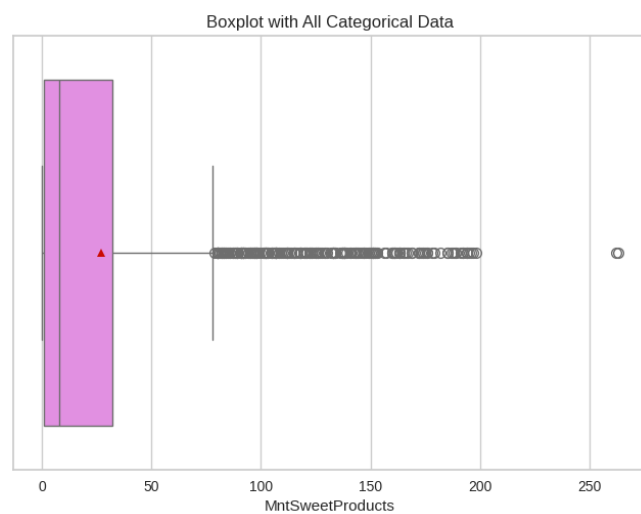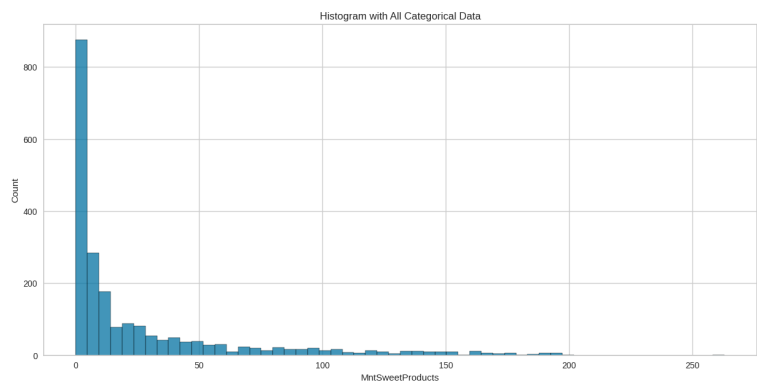
```
Original data shape: (2240, 27)
Filtered data shape: (2182, 28)
```

## MONEY SPENT ON GOLD PRODUCTS

Histogram with All Categorical Data

Boxplot with All Categorical Data

Histogram with Outliers Highlighted

All Data
Outliers
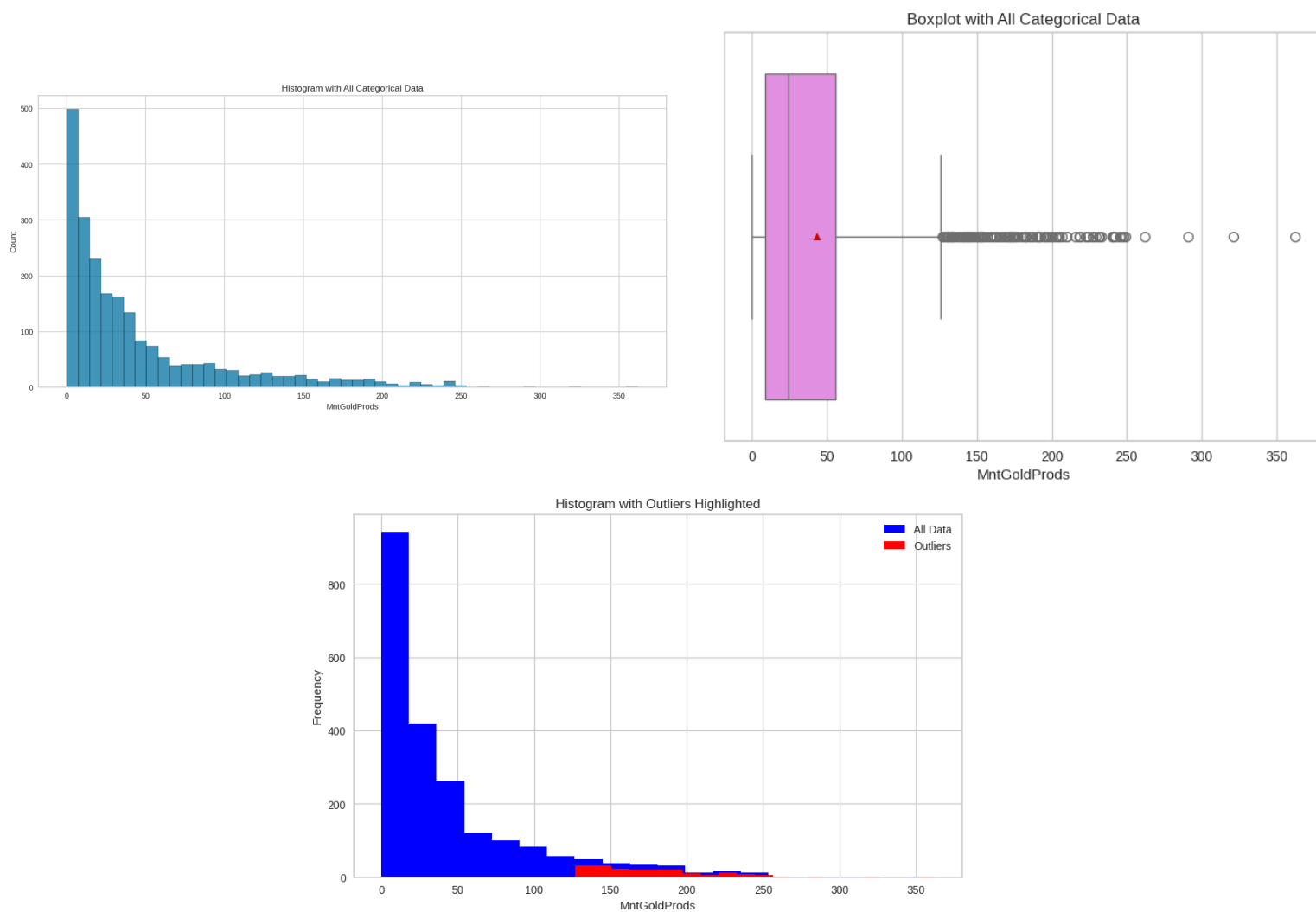
Original data shape: (2240, 27)
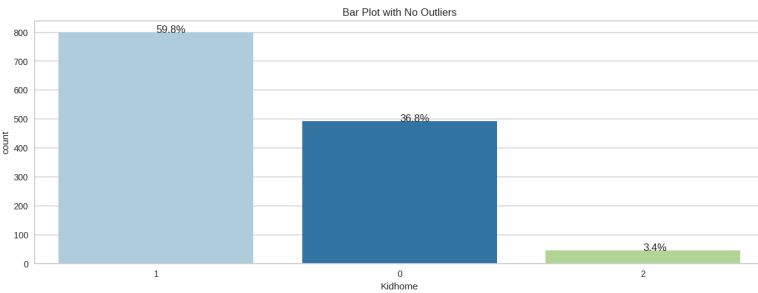Filtered data shape: (2182, 28)

# CHECKING :
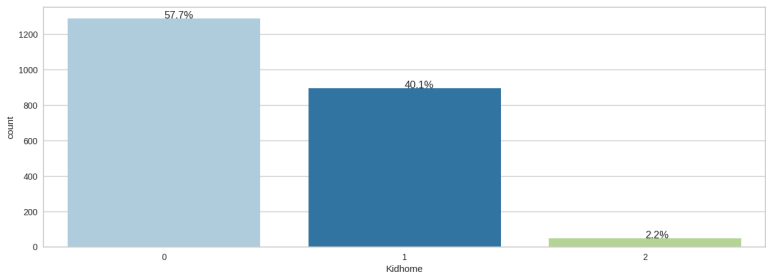# NO OUTLIERS DATA SET   ~ AND ~   ONLY OUTLIERS DATA SET
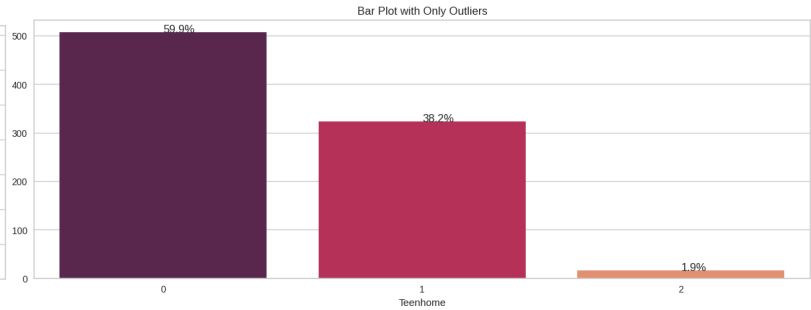
## Bar plot for 'Marital_Status'
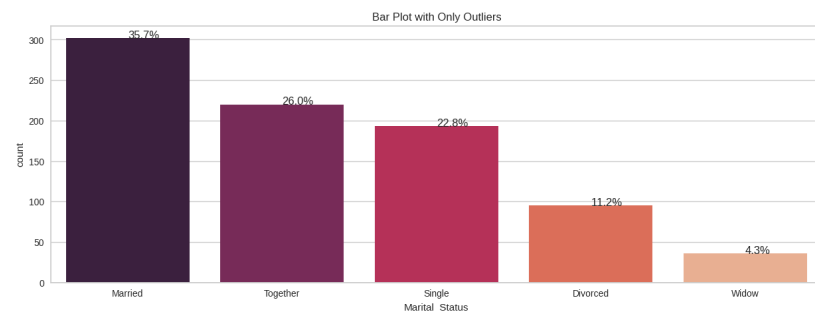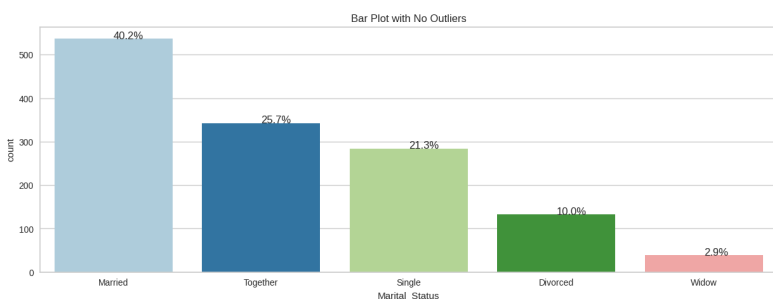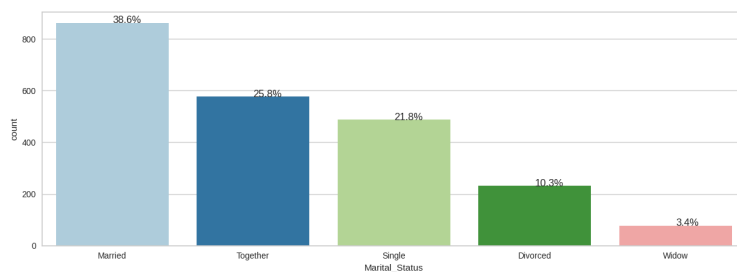


## Bar plot for 'Education'
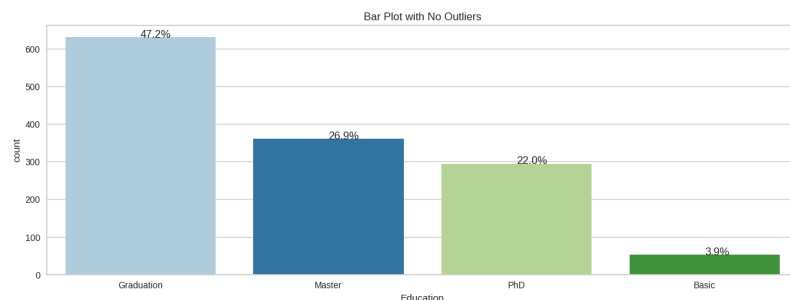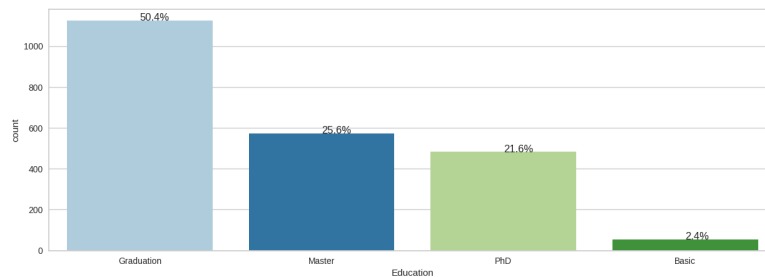
## Bar plot for 'Kidhome'
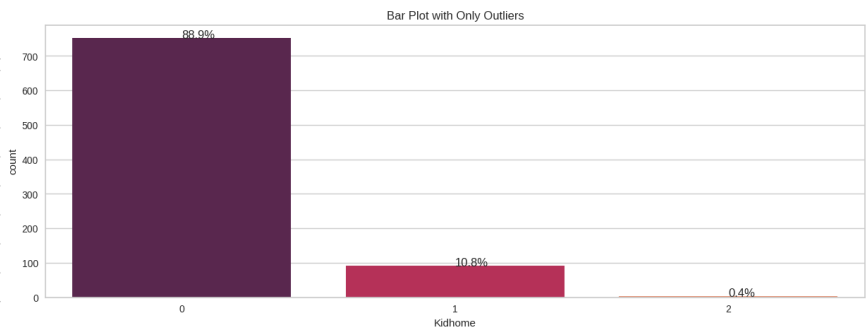


## Bar plot for 'Teenhome'
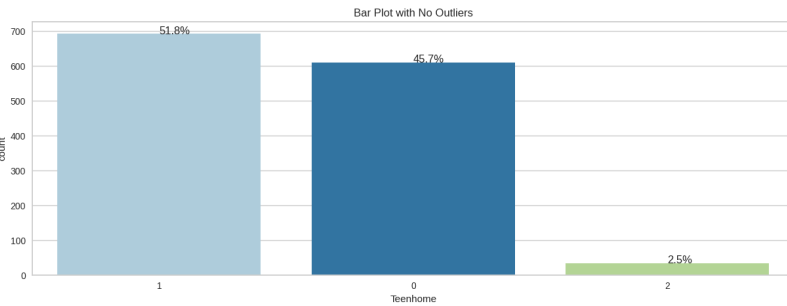
# Bar plot for 'Complain'
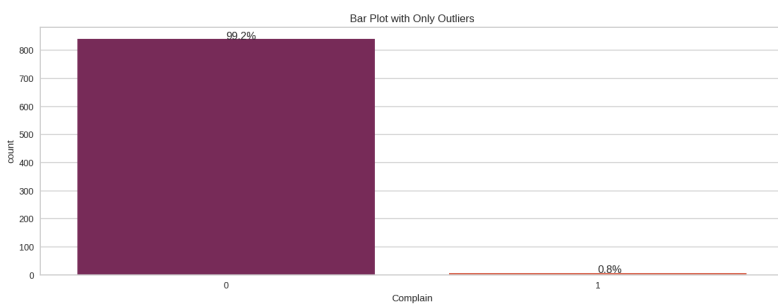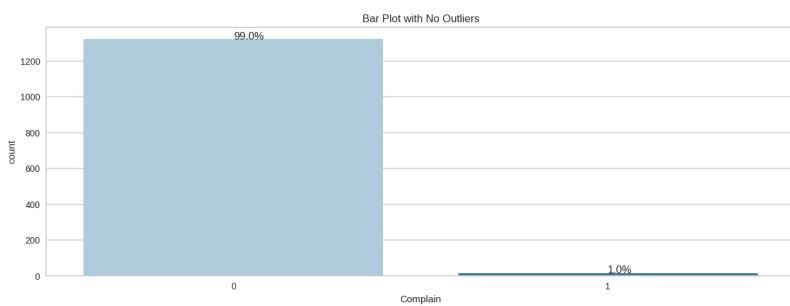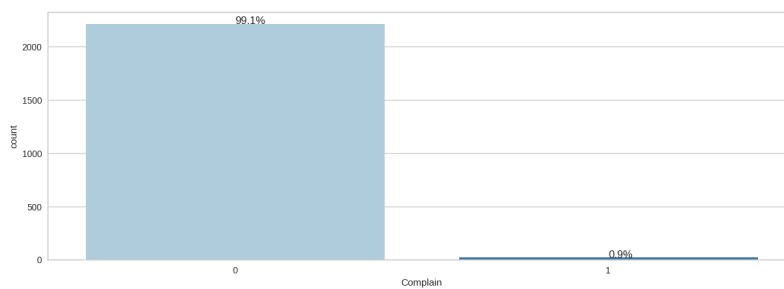




# Bar plot for 'Dt_Customer by quarter'
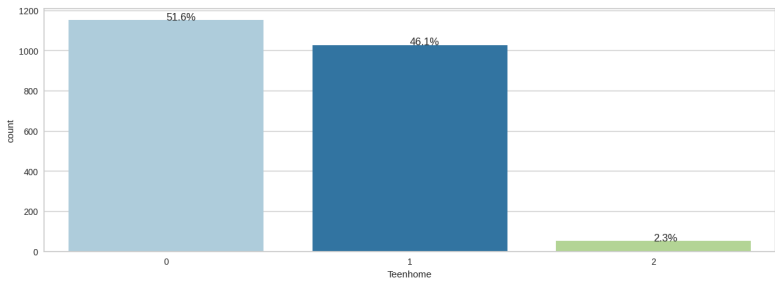
Histogram with Outliers Filtered Out



Bar Plot with No Outliers

### Customer Age Distribution No Outliers



### Customer Age Distribution Only Outliers

**Bar Plot with No Outliers**



**Distribution of Expenses (Left Tail)**

Scatter Plot of Expenses vs. Income



Bar Plot of Income vs. Family_Size
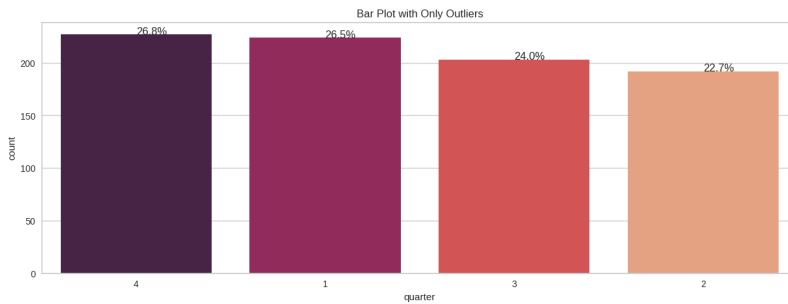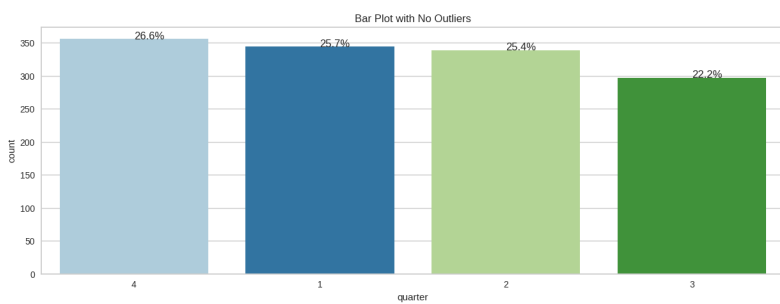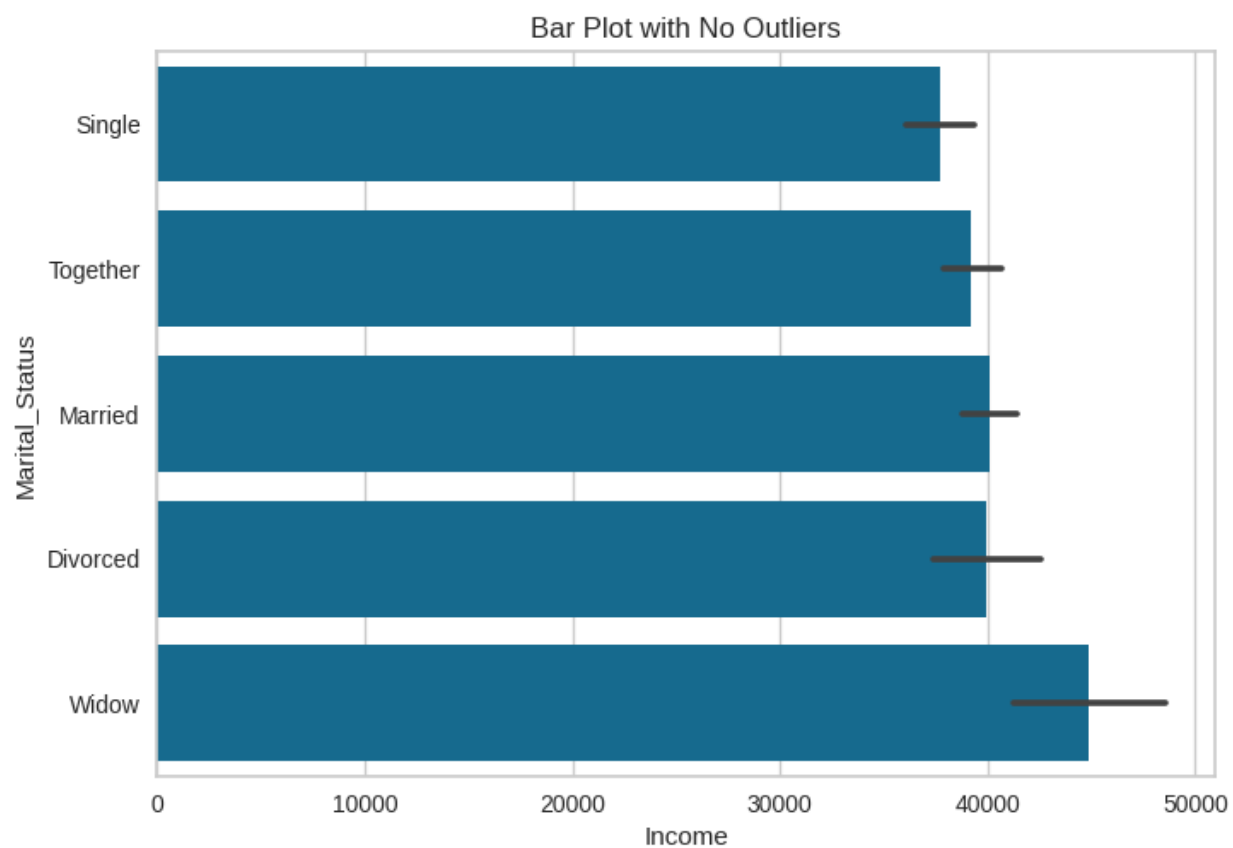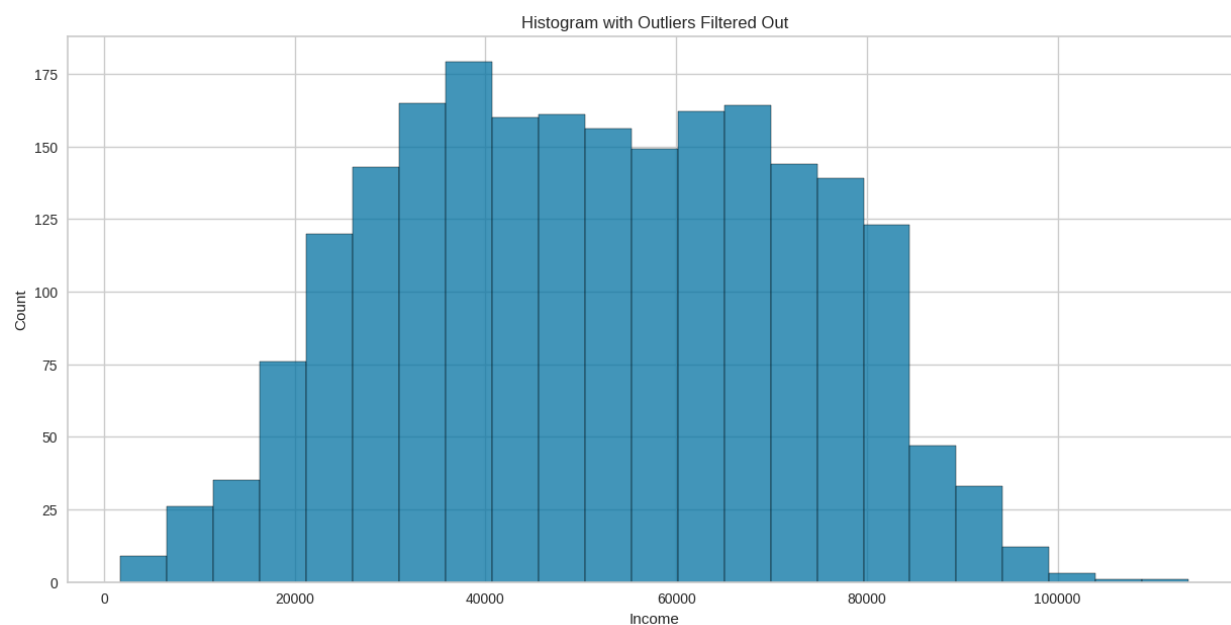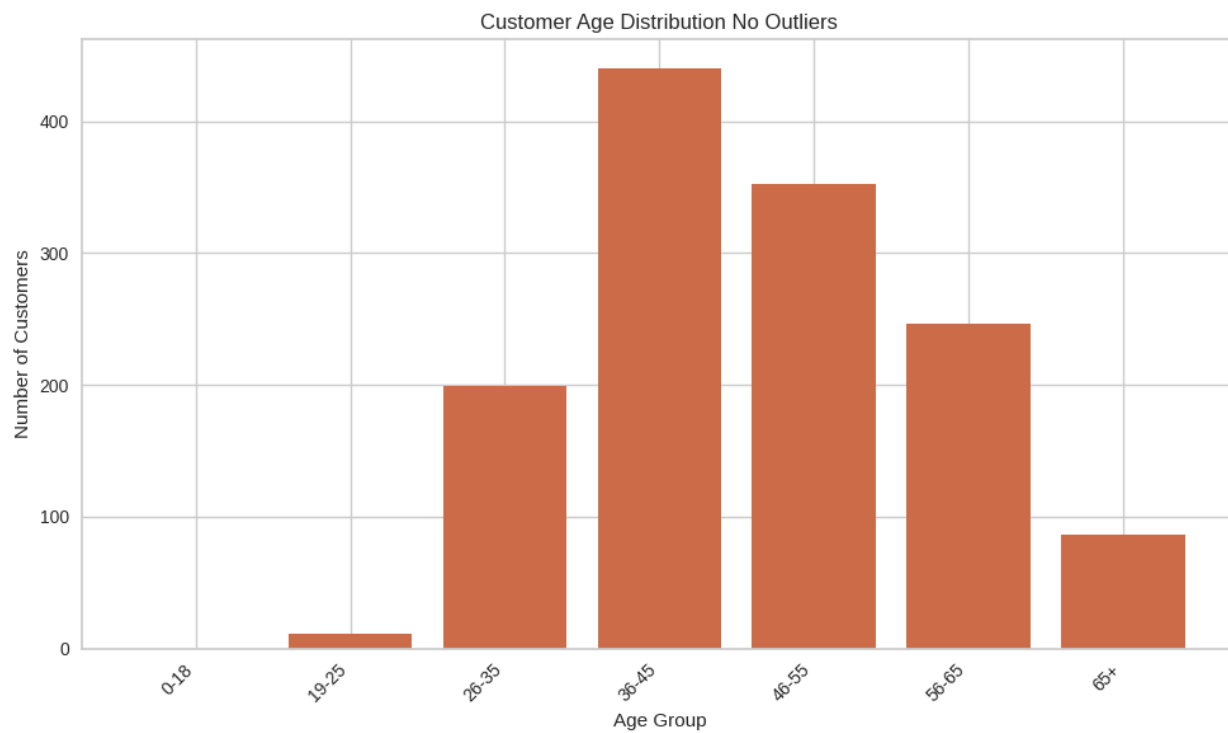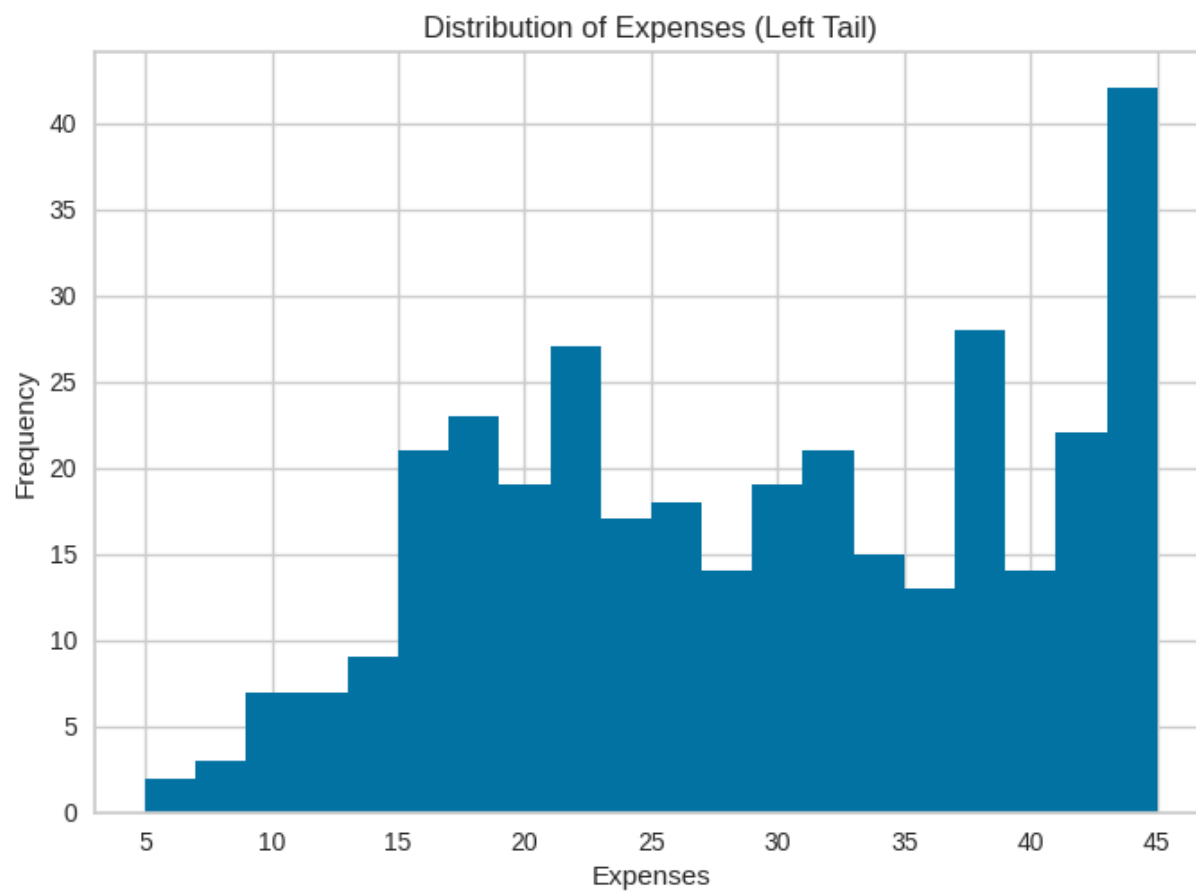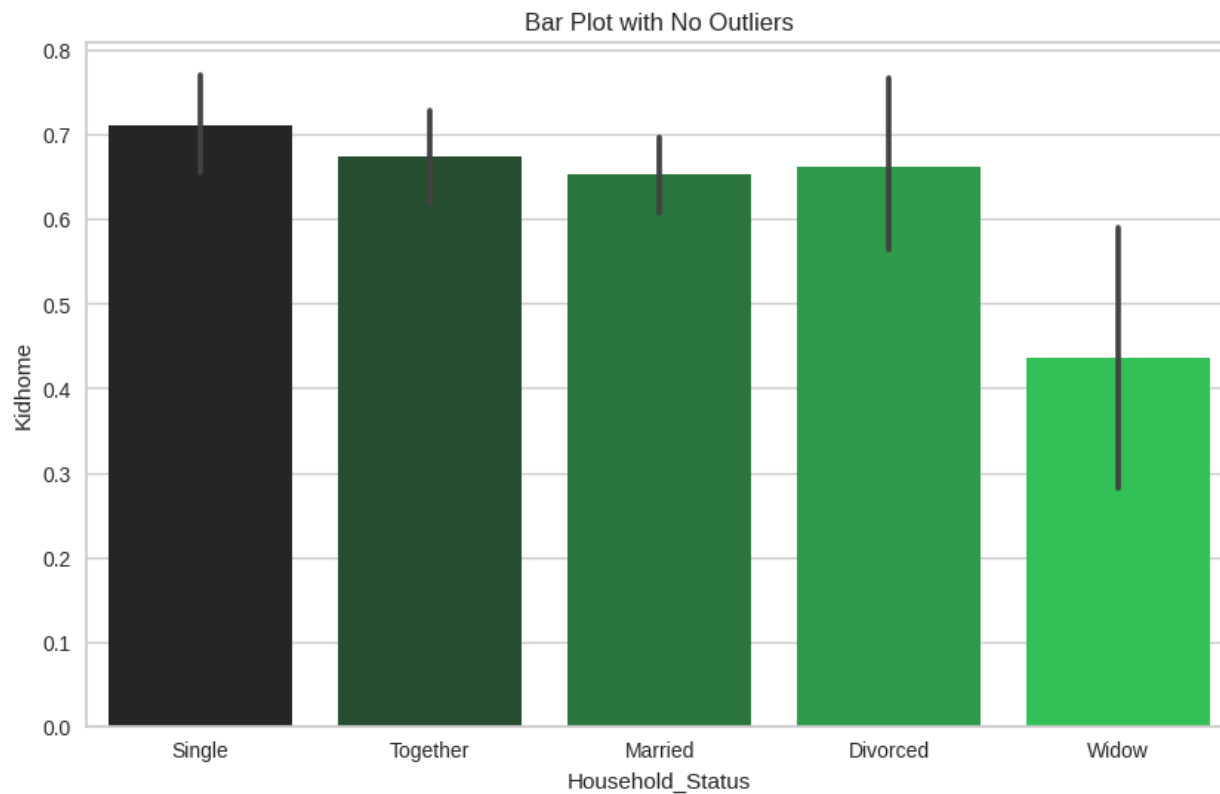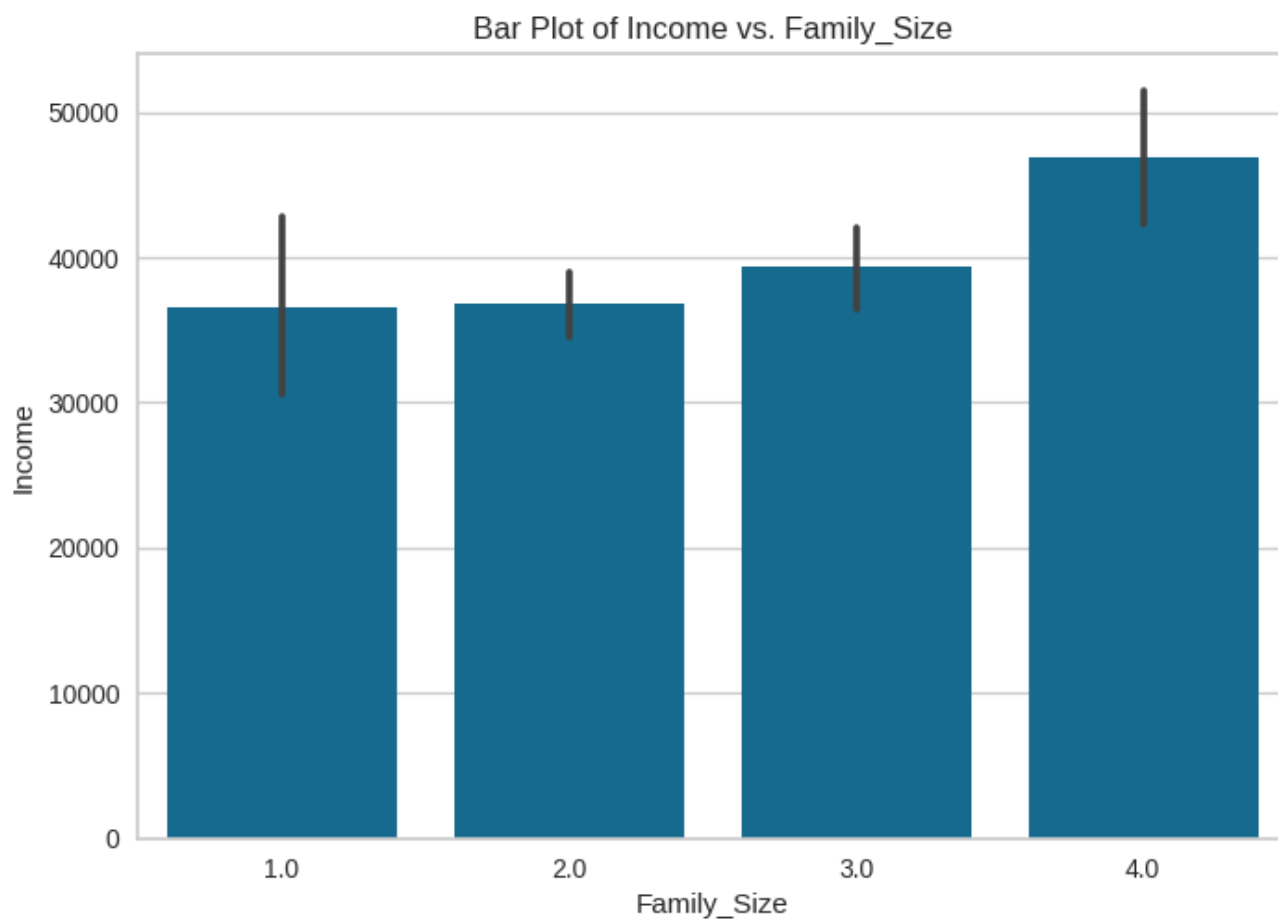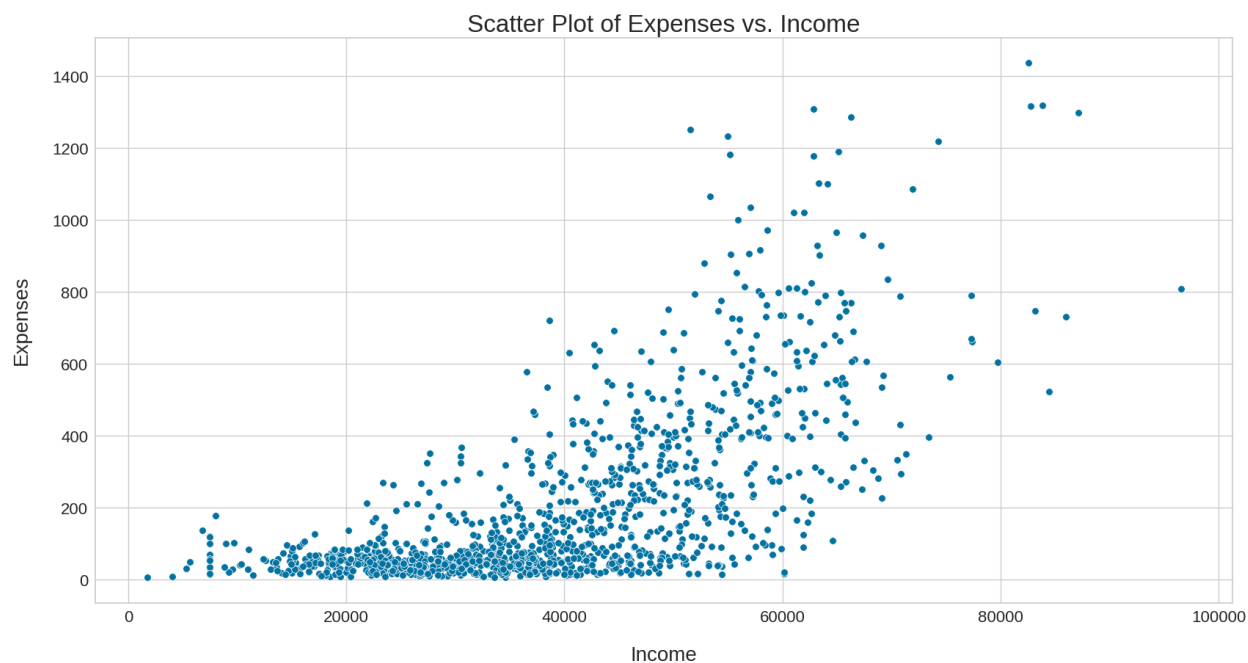
## data_no_outliers

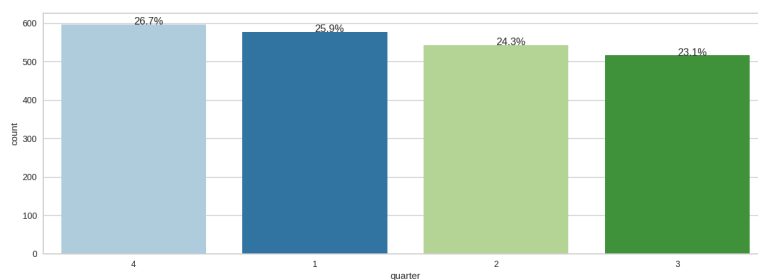| # | Column | Count | Non-Null | Dtype |
|---|--------|-------|----------|-------|
| 0 | ID | 1336 | non-null | int64 |
| 1 | Year_Birth | 1336 | non-null | int64 |
| 2 | Education | 1336 | non-null | object |
| 3 | Marital_Status | 1336 | non-null | object |
| 4 | Income | 1321 | non-null | float64 |
| 5 | Kidhome | 1336 | non-null | int64 |
| 6 | Teenhome | 1336 | non-null | int64 |
| 7 | Dt_Customer | 1336 | non-null | datetime64[ns] |
| 8 | Recency | 1336 | non-null | int64 |
| 9 | MntWines | 1336 | non-null | int64 |
| 10 | MntFruits | 1336 | non-null | int64 |
| 11 | MntMeatProducts | 1336 | non-null | int64 |
| 12 | MntFishProducts | 1336 | non-null | int64 |
| 13 | MntSweetProducts | 1336 | non-null | int64 |
| 14 | MntGoldProds | 1336 | non-null | int64 |
| 15 | NumDealsPurchases | 1336 | non-null | int64 |
| 16 | NumWebPurchases | 1336 | non-null | int64 |
| 17 | NumCatalogPurchas | 1336 | non-null | int64 |
| 18 | NumStorePurchases | 1336 | non-null | int64 |
| 19 | NumWebVisitsMont | 1336 | non-null | int64 |
| 20 | AcceptedCmp3 | 1336 | non-null | int64 |
| 21 | AcceptedCmp4 | 1336 | non-null | int64 |
| 22 | AcceptedCmp5 | 1336 | non-null | int64 |
| 23 | AcceptedCmp1 | 1336 | non-null | int64 |
| 24 | AcceptedCmp2 | 1336 | non-null | int64 |
| 25 | Complain | 1336 | non-null | int64 |
| 26 | Response | 1336 | non-null | int64 |
| 27 | quarter | 1336 | non-null | int32 |

dtypes: datetime64[ns](1), float64(1), int32(1), int64(23), object(2)

## data_only_outliers

| # | Column | Count | Non-Null | Dtype |
|---|--------|-------|----------|-------|
| 0 | ID | 846 | non-null | int64 |
| 1 | Year_Birth | 846 | non-null | int64 |
| 2 | Education | 846 | non-null | object |
| 3 | Marital_Status | 846 | non-null | object |
| 4 | Income | 837 | non-null | float64 |
| 5 | Kidhome | 846 | non-null | int64 |
| 6 | Teenhome | 846 | non-null | int64 |
| 7 | Dt_Customer | 846 | non-null | datetime64[ns] |
| 8 | Recency | 846 | non-null | int64 |
| 9 | MntWines | 846 | non-null | int64 |
| 10 | MntFruits | 846 | non-null | int64 |
| 11 | MntMeatProducts | 846 | non-null | int64 |
| 12 | MntFishProducts | 846 | non-null | int64 |
| 13 | MntSweetProducts | 846 | non-null | int64 |
| 14 | MntGoldProds | 846 | non-null | int64 |
| 15 | NumDealsPurchases | 846 | non-null | int64 |
| 16 | NumWebPurchases | 846 | non-null | int64 |
| 17 | NumCatalogPurchas | 846 | non-null | int64 |
| 18 | NumStorePurchases | 846 | non-null | int64 |
| 19 | NumWebVisitsMont | 846 | non-null | int64 |
| 20 | AcceptedCmp3 | 846 | non-null | int64 |
| 21 | AcceptedCmp4 | 846 | non-null | int64 |
| 22 | AcceptedCmp5 | 846 | non-null | int64 |
| 23 | AcceptedCmp1 | 846 | non-null | int64 |
| 24 | AcceptedCmp2 | 846 | non-null | int64 |
| 25 | Complain | 846 | non-null | int64 |
| 26 | Response | 846 | non-null | int64 |
| 27 | quarter | 846 | non-null | int32 |

dtypes: datetime64[ns](1), float64(1), int32(1), int64(23), object(2)

Heat Map with No Outliers



Bar Plot with No Outliers

## Bar Plot with No Outliers



## Bar Plot with Only Outliers

## Bar Plot with Only Outliers



## Bar Plot with Only Outliers

## Distribution of Amount Per Purchase
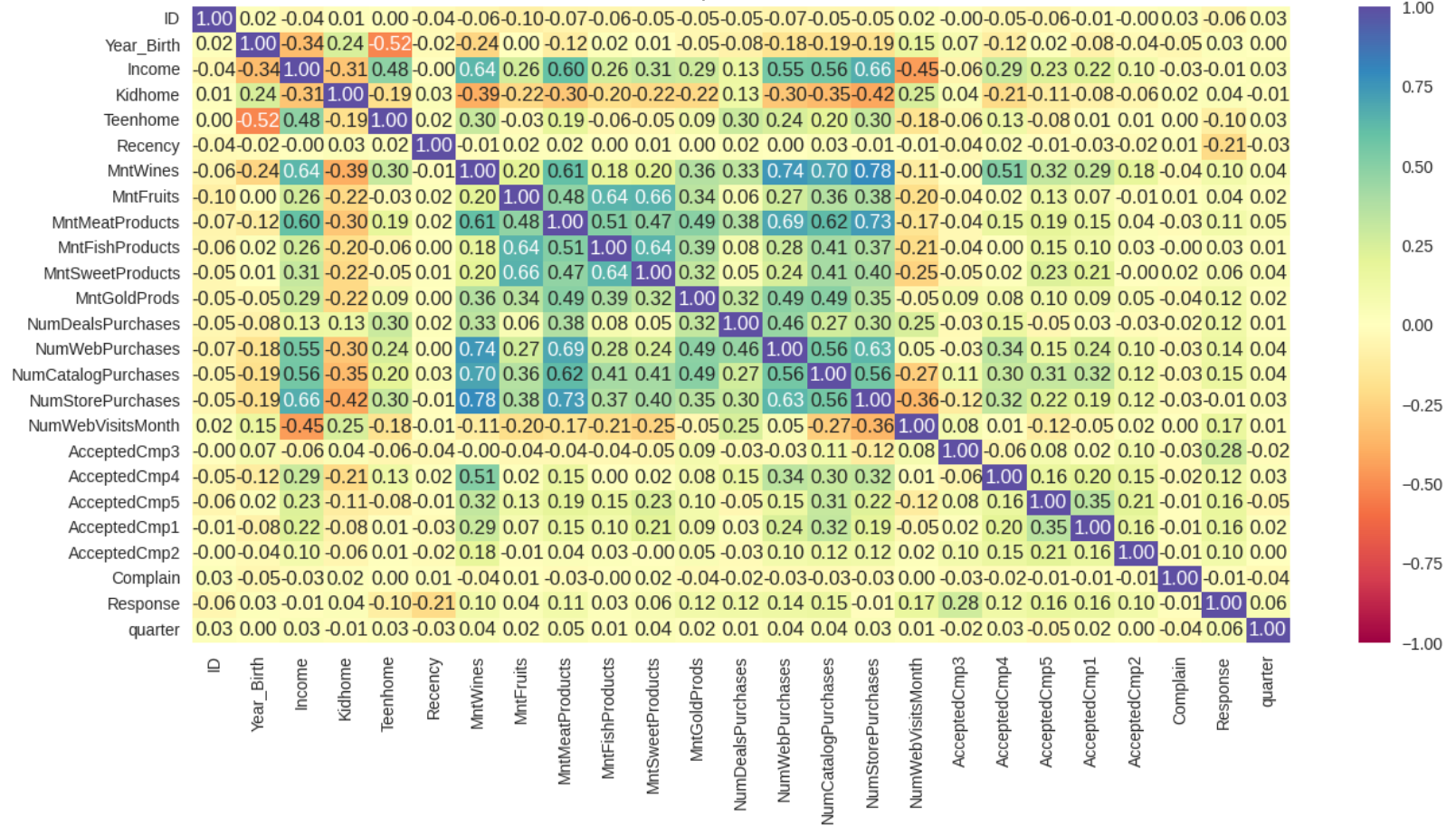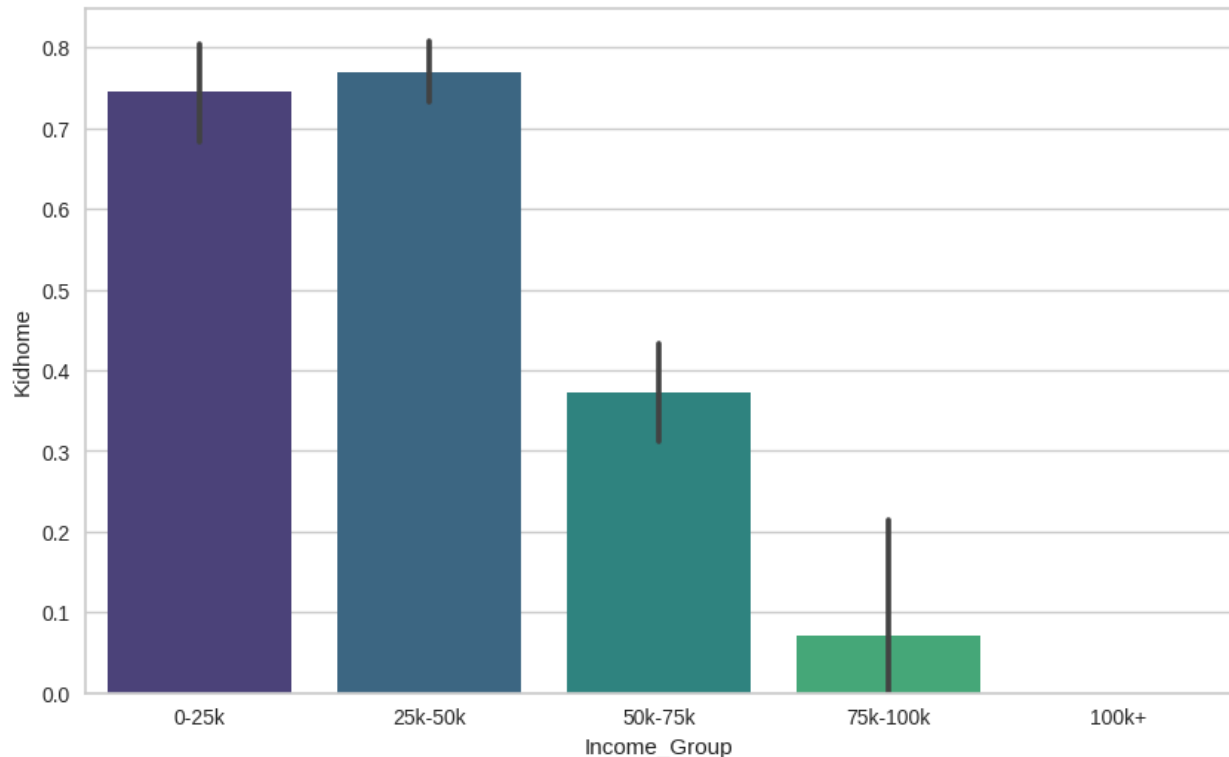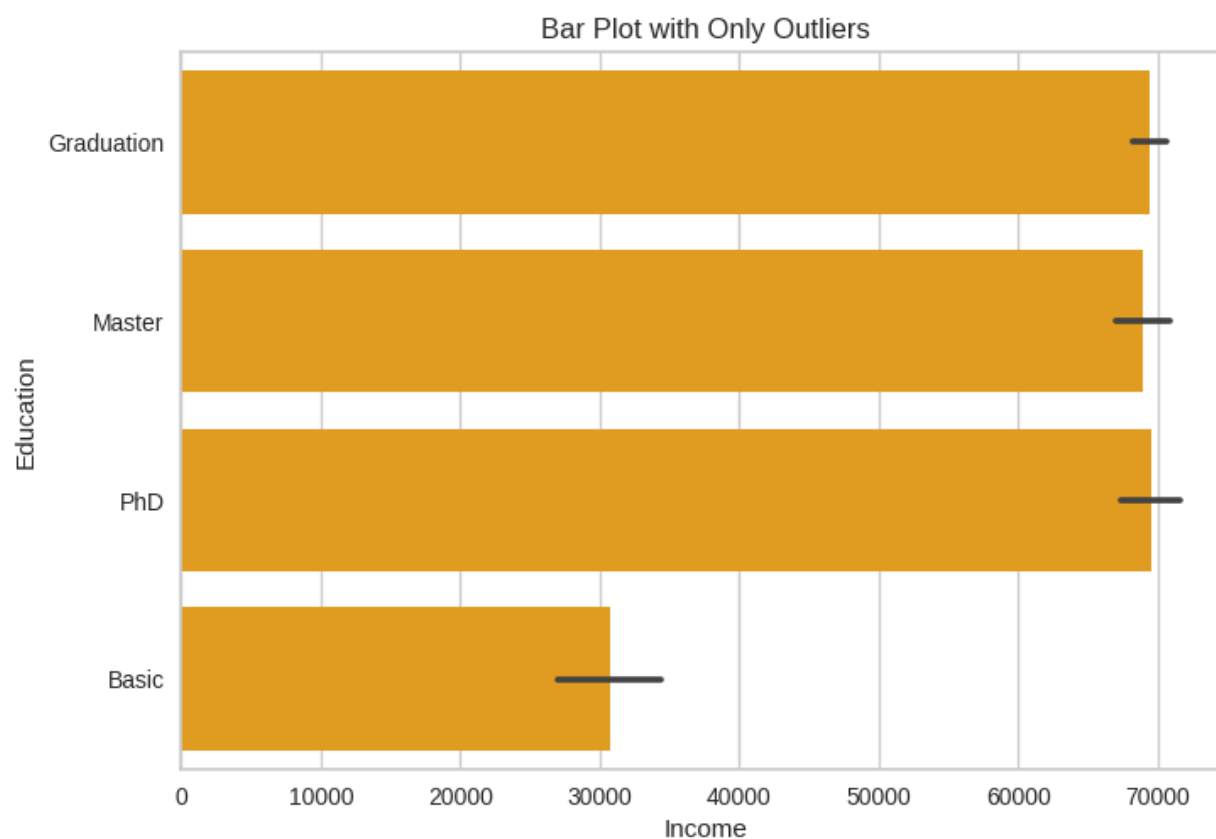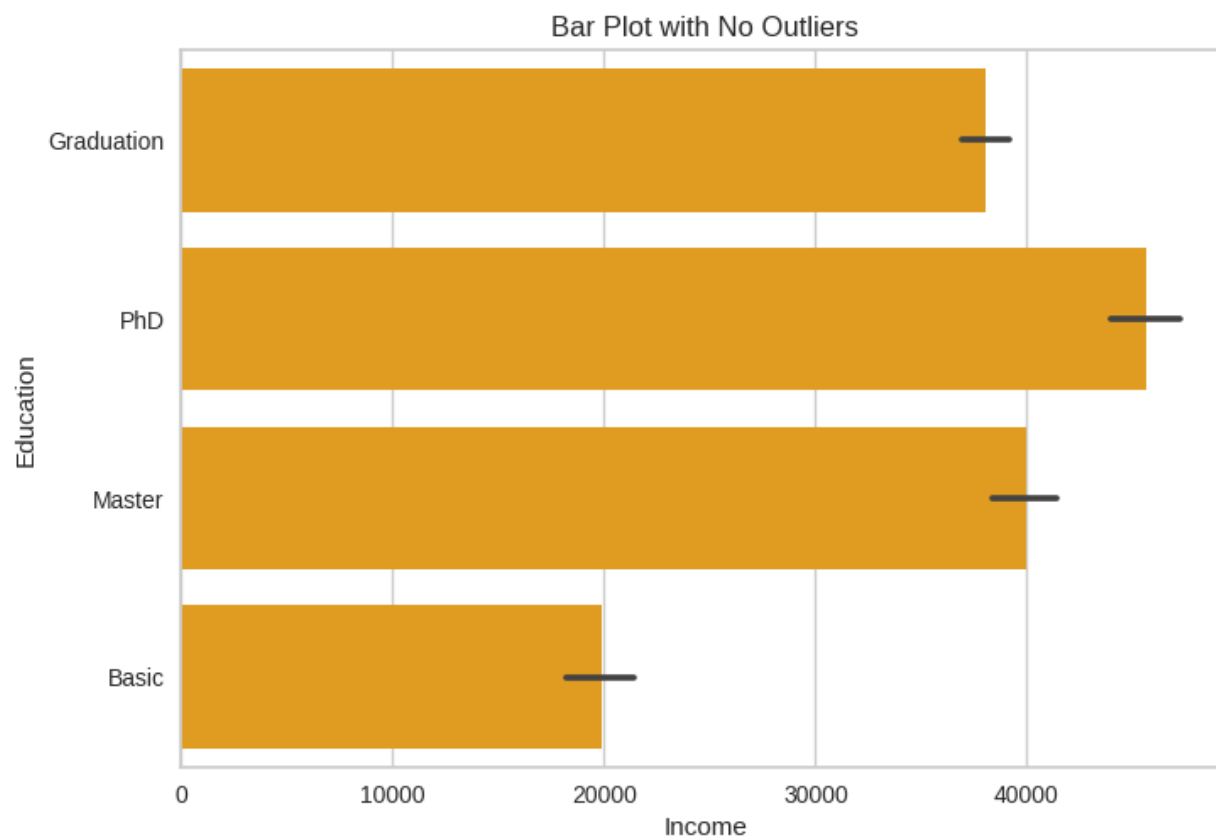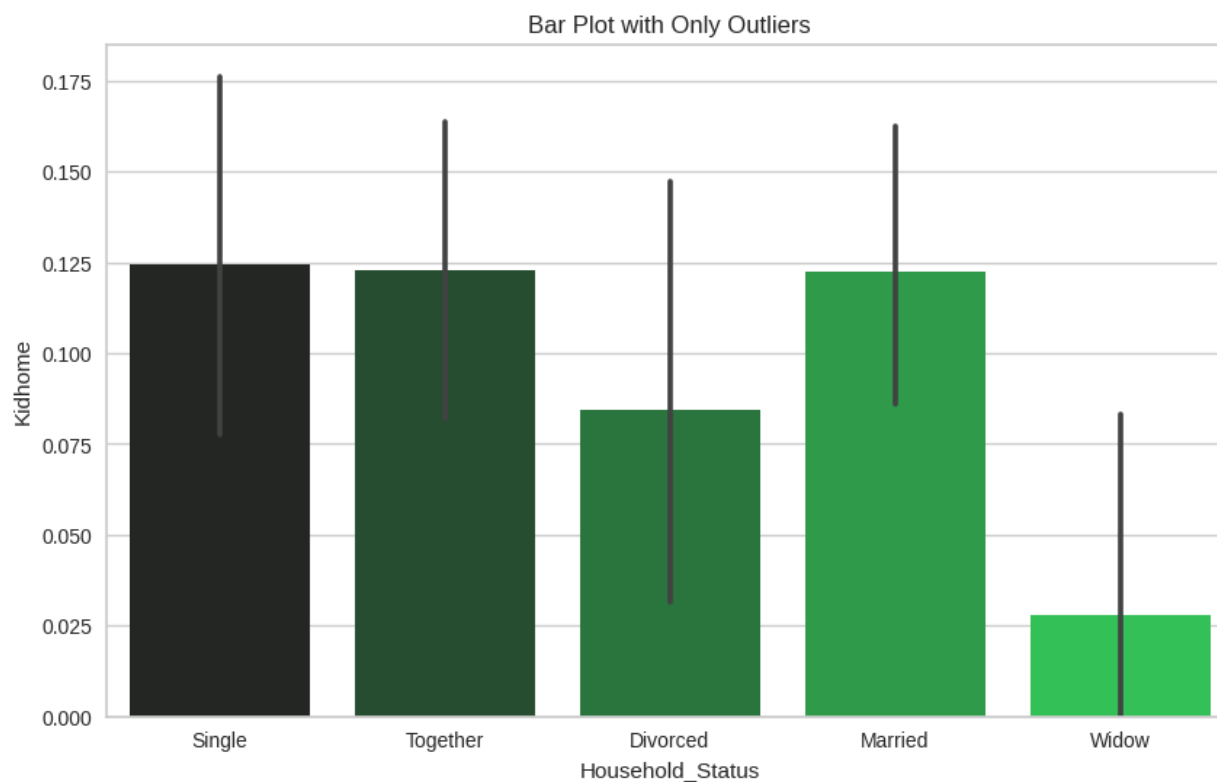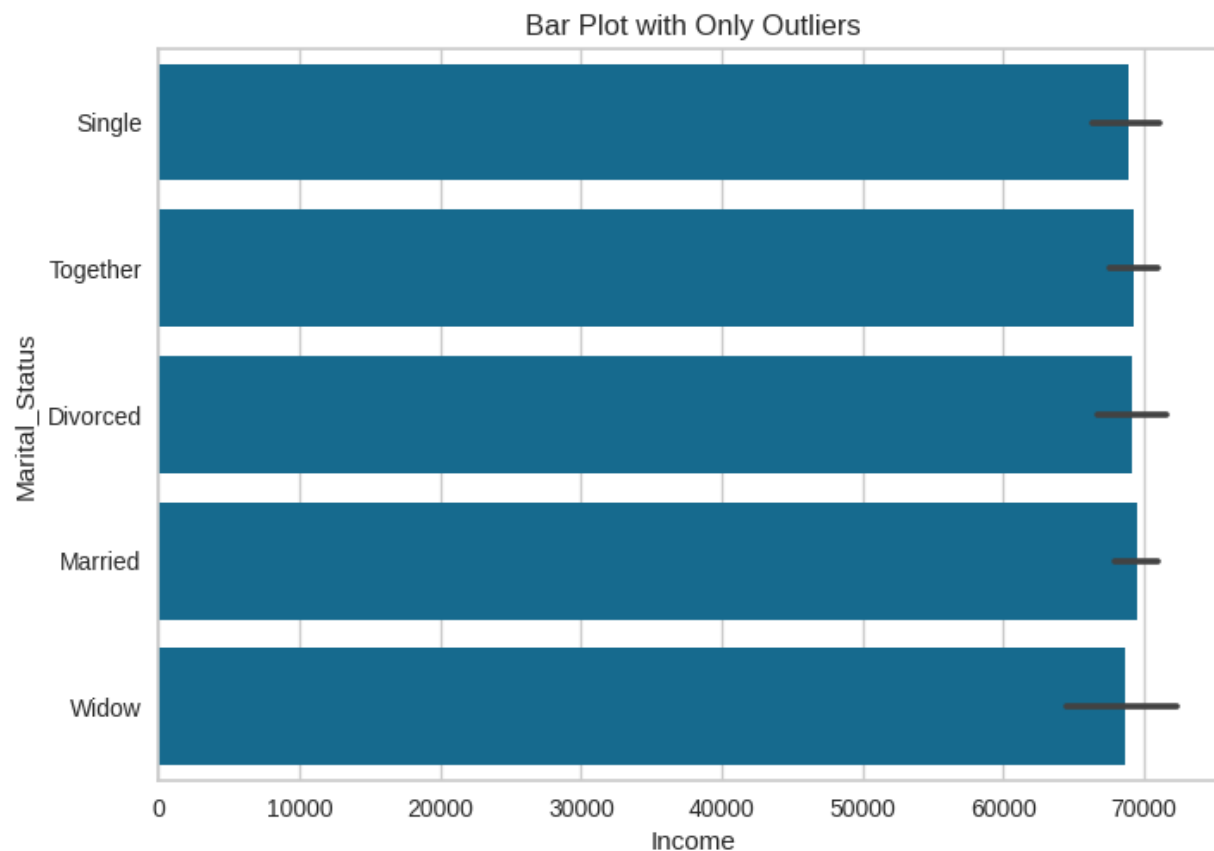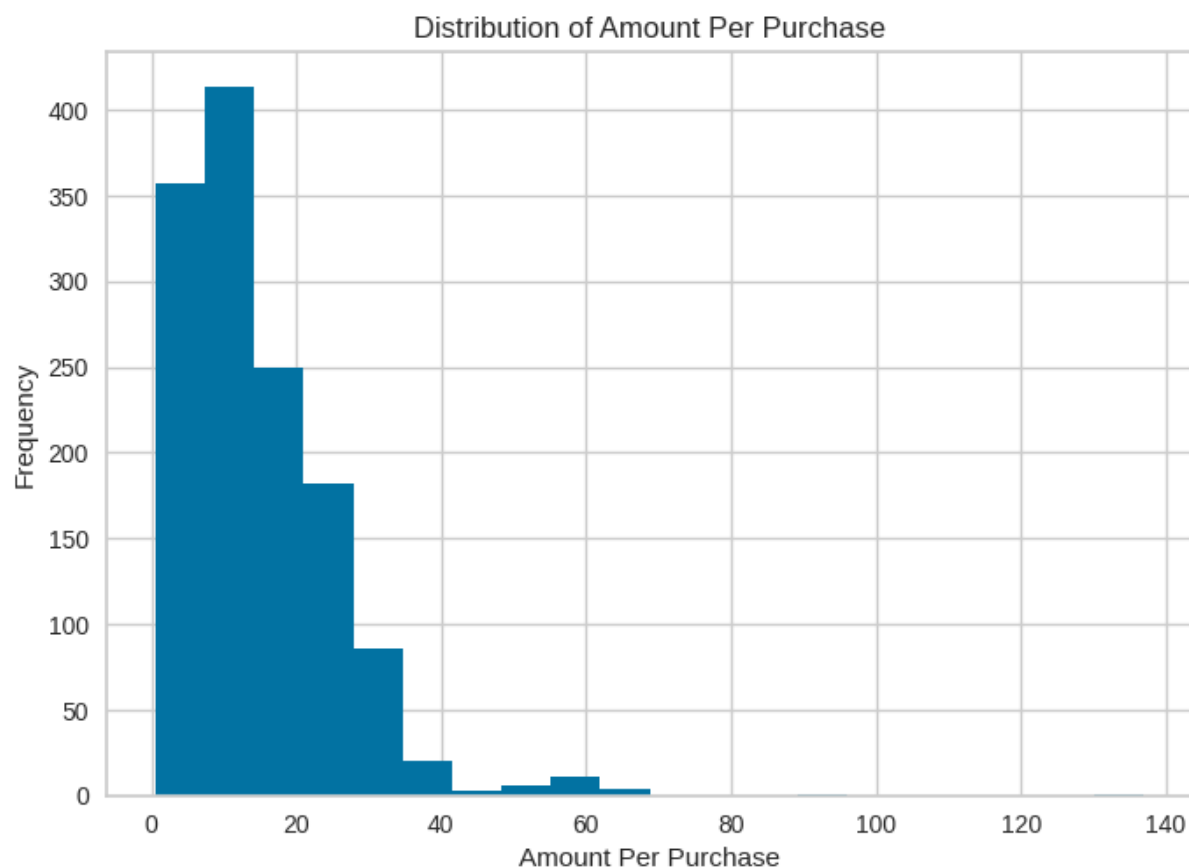


## A summary of days with the company:

|  | Dt_Customer | DaysWithCompany |
|---|---|---|
| count | 1336 | 1336.000000 |
| mean | 2013-08-03 13:10:03.592814336 | 515.451347 |
| min | 2012-01-08 00:00:00 | 26.000000 |
| 25% | 2013-02-13 00:00:00 | 338.000000 |
| 50% | 2013-08-07 00:00:00 | 512.000000 |
| 75% | 2014-01-28 00:00:00 | 687.000000 |
| 90% | 2014-05-25 12:00:00 | 827.000000 |
| 95% | 2014-08-03 06:00:00 | 875.250000 |
| 99% | 2014-12-01 00:00:00 | 1027.300000 |
| max | 2014-12-06 00:00:00 | 1089.000000 |
| std | NaN | 231.922648 |

Amount Per Purchase

## Income Vs Expenses



Scatter Plot of Expenses vs. Income

# Family Size Vs Income

Correlation Matrix of Data Model with Irrelevant Variables Removed

| | ID | Recency | MntWines | MntFruits | MntMeatProducts | MntFishProducts | MntSweetProducts | MntGoldProds | NumDealsPurchases | NumWebPurchases | NumCatalogPurchases | NumStorePurchases | NumWebVisitsMonth | Family_Size | Expenses | NumTotalPurchases | AmountPerPurchase | DaysWithCompany | TotalAcceptedCmp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | 1.00 | -0.04 | -0.06 | -0.10 | -0.06 | -0.06 | -0.05 | -0.05 | -0.05 | -0.07 | -0.05 | -0.05 | 0.02 | 0.03 | -0.07 | -0.07 | -0.08 | -0.02 | -0.06 |
| Recency | -0.04 | 1.00 | -0.02 | 0.02 | 0.02 | 0.00 | 0.01 | 0.00 | 0.02 | 0.00 | 0.03 | -0.02 | -0.00 | 0.04 | -0.01 | 0.01 | -0.00 | 0.03 | -0.13 |
| MntWines | -0.06 | -0.02 | 1.00 | 0.20 | 0.61 | 0.17 | 0.20 | 0.36 | 0.33 | 0.74 | 0.70 | 0.78 | -0.11 | -0.04 | 0.96 | 0.83 | 0.83 | 0.11 | 0.38 |
| MntFruits | -0.10 | 0.02 | 0.20 | 1.00 | 0.48 | 0.64 | 0.66 | 0.34 | 0.06 | 0.27 | 0.36 | 0.38 | -0.20 | -0.16 | 0.39 | 0.34 | 0.41 | 0.12 | 0.04 |
| MntMeatProducts | -0.06 | 0.02 | 0.61 | 0.48 | 1.00 | 0.51 | 0.47 | 0.49 | 0.37 | 0.69 | 0.62 | 0.73 | -0.17 | -0.05 | 0.76 | 0.79 | 0.71 | 0.15 | 0.16 |
| MntFishProducts | -0.06 | 0.00 | 0.17 | 0.64 | 0.51 | 1.00 | 0.64 | 0.39 | 0.08 | 0.28 | 0.41 | 0.37 | -0.21 | -0.14 | 0.38 | 0.36 | 0.39 | 0.14 | 0.04 |
| MntSweetProducts | -0.05 | 0.01 | 0.20 | 0.66 | 0.47 | 0.64 | 1.00 | 0.32 | 0.05 | 0.24 | 0.41 | 0.40 | -0.25 | -0.17 | 0.39 | 0.35 | 0.39 | 0.06 | 0.09 |
| MntGoldProds | -0.05 | 0.00 | 0.36 | 0.34 | 0.49 | 0.39 | 0.32 | 1.00 | 0.32 | 0.49 | 0.49 | 0.34 | -0.04 | -0.07 | 0.52 | 0.52 | 0.53 | 0.21 | 0.17 |
| NumDealsPurchases | -0.05 | 0.02 | 0.33 | 0.06 | 0.37 | 0.08 | 0.05 | 0.32 | 1.00 | 0.46 | 0.27 | 0.30 | 0.26 | 0.27 | 0.36 | 0.65 | 0.19 | 0.19 | 0.10 |
| NumWebPurchases | -0.07 | 0.00 | 0.74 | 0.27 | 0.69 | 0.28 | 0.24 | 0.49 | 0.46 | 1.00 | 0.56 | 0.63 | 0.06 | -0.03 | 0.79 | 0.88 | 0.67 | 0.12 | 0.27 |
| NumCatalogPurchases | -0.05 | 0.03 | 0.70 | 0.36 | 0.62 | 0.41 | 0.41 | 0.49 | 0.27 | 0.56 | 1.00 | 0.56 | -0.27 | -0.10 | 0.77 | 0.73 | 0.68 | 0.11 | 0.37 |
| NumStorePurchases | -0.05 | -0.02 | 0.78 | 0.38 | 0.73 | 0.37 | 0.40 | 0.34 | 0.30 | 0.63 | 0.56 | 1.00 | -0.36 | -0.05 | 0.83 | 0.83 | 0.71 | 0.05 | 0.15 |
| NumWebVisitsMonth | 0.02 | -0.00 | -0.11 | -0.20 | -0.17 | -0.21 | -0.25 | -0.04 | 0.26 | 0.06 | -0.27 | -0.36 | 1.00 | 0.05 | -0.15 | -0.09 | -0.12 | 0.33 | 0.10 |
| Family_Size | 0.03 | 0.04 | -0.04 | -0.16 | -0.05 | -0.14 | -0.17 | -0.07 | 0.27 | -0.03 | -0.10 | -0.05 | 0.05 | 1.00 | -0.08 | 0.03 | -0.14 | -0.03 | -0.11 |
| Expenses | -0.07 | -0.01 | 0.96 | 0.39 | 0.76 | 0.38 | 0.39 | 0.52 | 0.36 | 0.79 | 0.77 | 0.83 | -0.15 | -0.08 | 1.00 | 0.89 | 0.89 | 0.15 | 0.35 |
| NumTotalPurchases | -0.07 | 0.01 | 0.83 | 0.34 | 0.79 | 0.36 | 0.35 | 0.52 | 0.65 | 0.88 | 0.73 | 0.83 | -0.09 | 0.03 | 0.89 | 1.00 | 0.73 | 0.14 | 0.27 |
| AmountPerPurchase | -0.08 | -0.00 | 0.83 | 0.41 | 0.71 | 0.39 | 0.39 | 0.53 | 0.19 | 0.67 | 0.68 | 0.71 | -0.12 | -0.14 | 0.89 | 0.73 | 1.00 | 0.15 | 0.30 |
| DaysWithCompany | -0.02 | 0.03 | 0.11 | 0.12 | 0.15 | 0.14 | 0.06 | 0.21 | 0.19 | 0.12 | 0.11 | 0.05 | 0.33 | -0.03 | 0.15 | 0.14 | 0.15 | 1.00 | 0.09 |
| TotalAcceptedCmp | -0.06 | -0.13 | 0.38 | 0.04 | 0.16 | 0.04 | 0.09 | 0.17 | 0.10 | 0.27 | 0.37 | 0.15 | 0.10 | -0.11 | 0.35 | 0.27 | 0.30 | 0.09 | 1.00 |

T-SNE Plot Showing Similarities in Features



Scatter Plot of First Two Principal Components
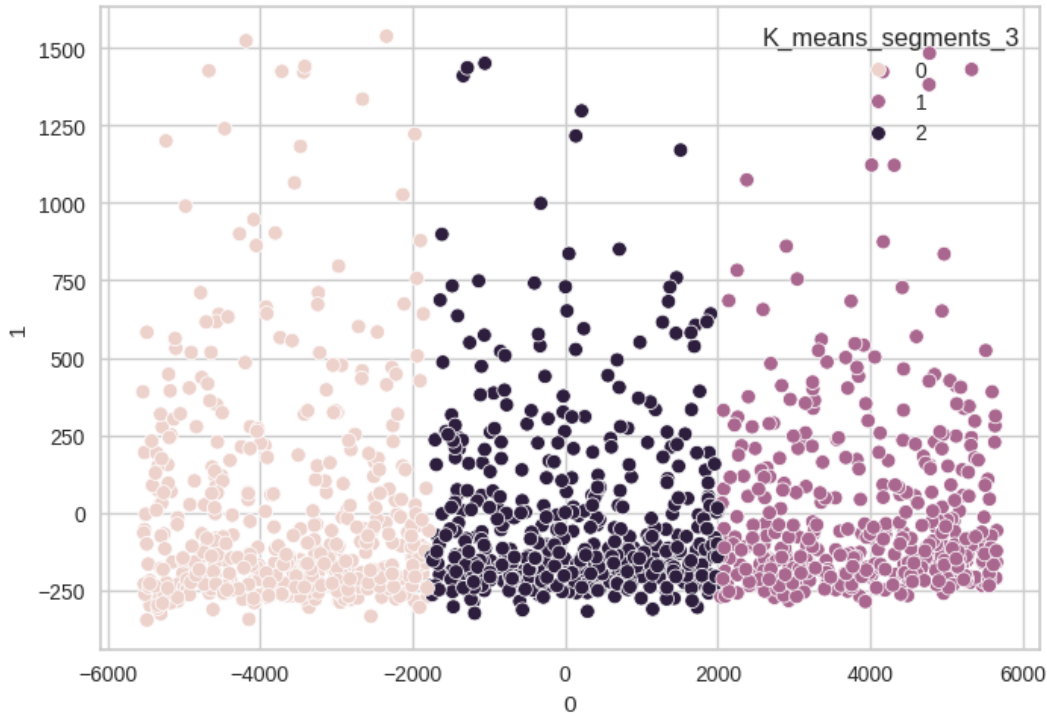
Characteristic Summary

```
0     Cluster 0: Characterized by high 1, 2, 0
1     Cluster 1: Characterized by high 1, 0, 2
2     Cluster 2: Characterized by high 0, 2, 1
3     Cluster 3: Characterized by high 2, 0, 1
4     Cluster 4: Characterized by high 2, 1, 0
5     Cluster 5: Characterized by high 2, 1, 0
6     Cluster 6: Characterized by high 1, 0, 2
7     Cluster 7: Characterized by high 2, 1, 0
8     Cluster 8: Characterized by high 0, 1, 2
9     Cluster 9: Characterized by high 2, 1, 0
10   Cluster 10: Characterized by high 2, 0, 1
11   Cluster 11: Characterized by high 0, 1, 2
12   Cluster 12: Characterized by high 2, 1, 0
13   Cluster 13: Characterized by high 0, 2, 1
14   Cluster 14: Characterized by high 0, 1, 2
15   Cluster 15: Characterized by high 2, 1, 0
16   Cluster 16: Characterized by high 1, 0, 2
17   Cluster 17: Characterized by high 2, 0, 1
18   Cluster 18: Characterized by high 2, 1,
```