Rakamin Academy

Predict Customer Personality to boost marketing campaign by using Machine Learning

Created by:
Tsabitah Karimah
tsabitahkarimah@gmail.com
LinkedIn

“Recent dietetics graduate with a newfound interest for data science, seeking to leverage analytical skills and healthcare knowledge in a data scientist role. Completed an intensive 5-month data science course, gaining proficiency in key programming languages and machine learning techniques. Eager to apply newly acquired skills in data manipulation, visualization, and predictive modeling to extract meaningful insights from complex datasets. Combines technical acumen with strong problem-solving abilities to drive data-informed decision-making in a data scientist role.”

Supported by:
Rakamin Academy
Career Acceleration School
www.rakamin.com

Overview

“A company can experience rapid growth when it understands customer personality behavior, allowing it to provide better services and benefits to customers who have the potential to become loyal customers. By analyzing historical marketing campaign data to improve performance and target the right customers, the goal is to create a predictive clustering model that will help the company make decisions more easily.”

Conversion Rate Analysis Based on Income, Spending and Age

Exploratory Data Analysis (EDA)

- The dataset have 2240 rows dan 30 columns.
- Column 'Income' appear to have missing value and there is a need to fill in the missing value.
  - The missing value filled with the mean of Income column
- The data type seems to match with the value of each columns, except the Dt_Customer, where the data type should be datetime.
  - It is then changed to date time.
- There is no duplicate data in this dataset.

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Feature Engineering

- Column ‘conversion_rate’ defined by response divided by visits.
  - There are missing value and after further analysis, the missing value is resulted from 0 value for both column response and NumWebVisitsMonth. Hence, the missing value will be replaced with 0.
- Column ‘Age’ defined by current year (2024) subtracted with column Year_Birth
  - Prior to that, few datas from column year_birth seems to be unsettling as the birth year is dated before 1901, resulting the age to be more than 100 years. I assume this to be a mistake, hence, the data was replaced with the mean of the birth year column.
  - Then, new column ‘Age’ engineered by subtracting 2024 with column Year_Birth
- New column named ‘Age_Group’ was engineered into 6 age group based on life stages (Childhood 0-12, Teen 13-19, Young adult 20-35, Middle Aged Adult 36-50, Senior Adult 51-65, Elderly 66 and above)
- Column ‘jumlah_anak’ defined by adding the value of ‘Kidhome’ and ‘Teenhome’
- Column ‘total_pengeluaran’ defined by the adding the value of 'MntCoke', 'MntFruits', ‘MntMeatProducts', 'MntFishProducts','MntSweetProducts', and 'MntGoldProds'
- Column ‘total_transaksi’ defined by adding the value of 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', and 'NumStorePurchases'.
- Column ‘total_acc_campaign’ defined by the total of campaign that the customer accepted.

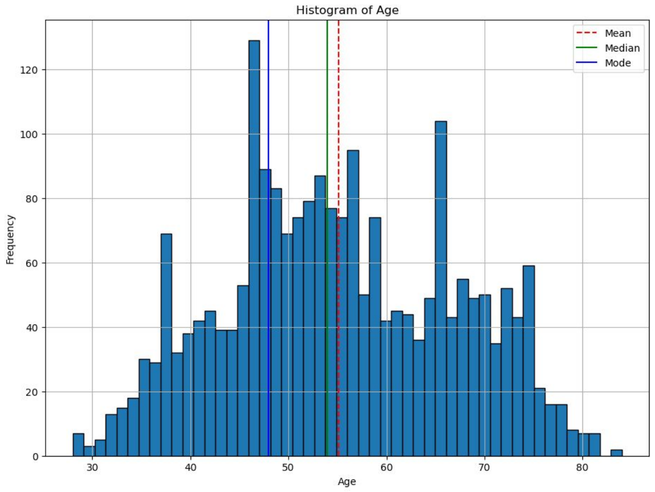
- After adding 7 new columns, the total columns of data is 36 with 2240 rows

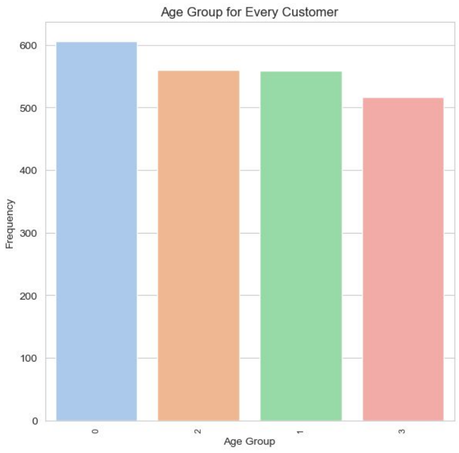
For further details: Source Code

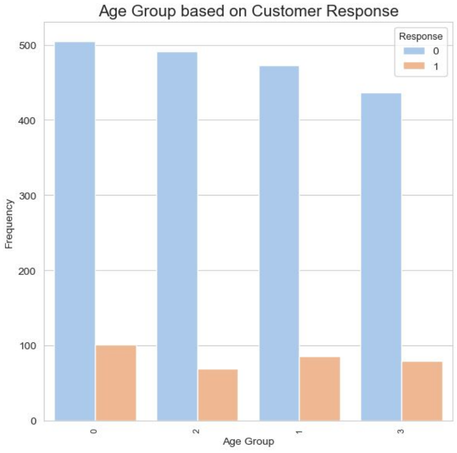
Conversion Rate Analysis Based on Income, Spending and Age

Age

- Age = near normal distribution, mean (55.1) > med(54), min: 28, max: 84
- Customer’s mean age is 55 years olds and median of 54 years old where the data is slightly positive skewed. The youngest and the oldest is 28 and 84, respectively.
- Customers who responded has lower age range, aged 47 years old and below



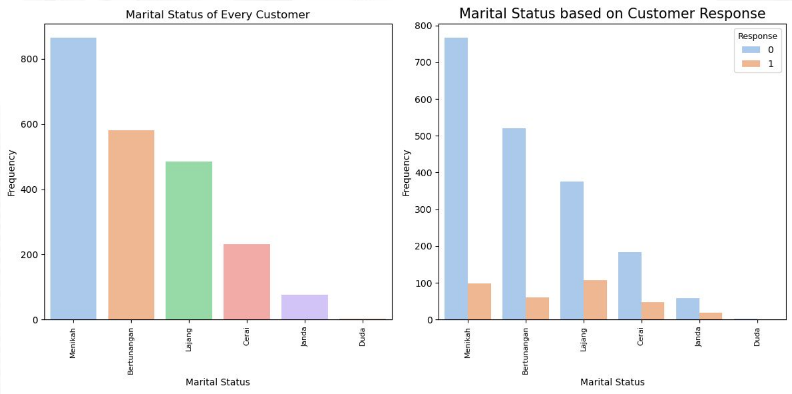


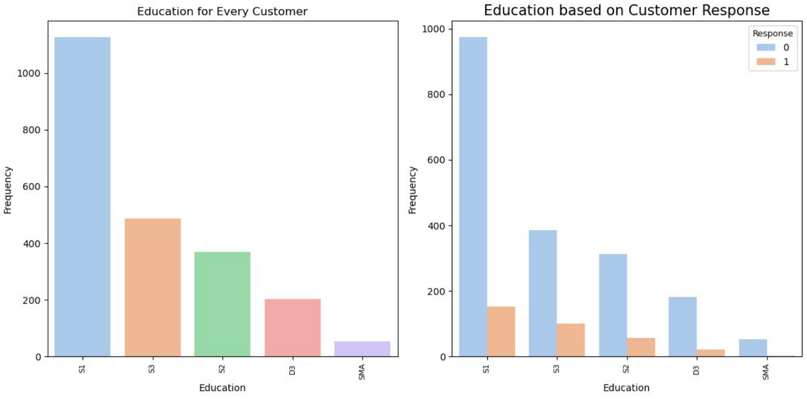


For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Education & Marital Status





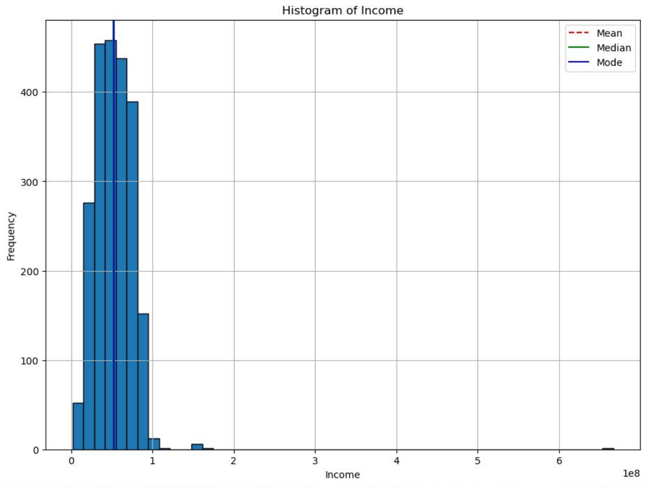
- Most of the customers are bachelor graduate from university and married
- Most of the customers who responded yes are those graduated as S1, followed by phd holder (S3), master holder (S2), diploma holder (D3) and high school graduate
- Customers who responded yes are mostly single, followed by married, engaged, divorced, and widowed

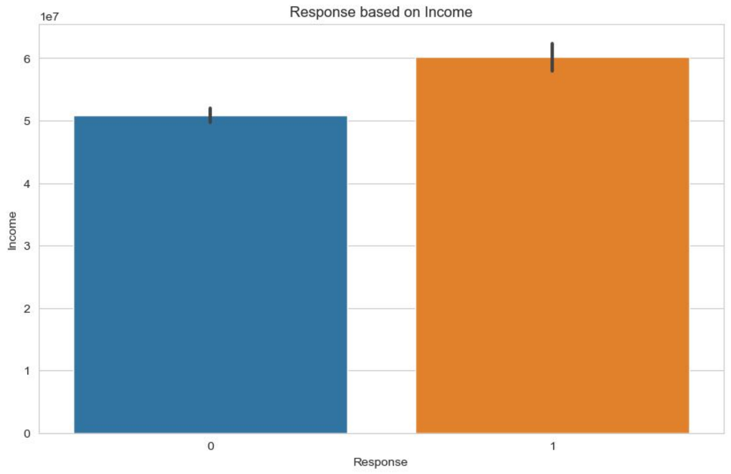
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Income

- The data for income is slightly positive skewed as the mean (52.247.251) is higher than the med(51.741.500) with minimum income of 1.730.000 and maximum income of 666.666.000
- Customers who responded are those with higher income

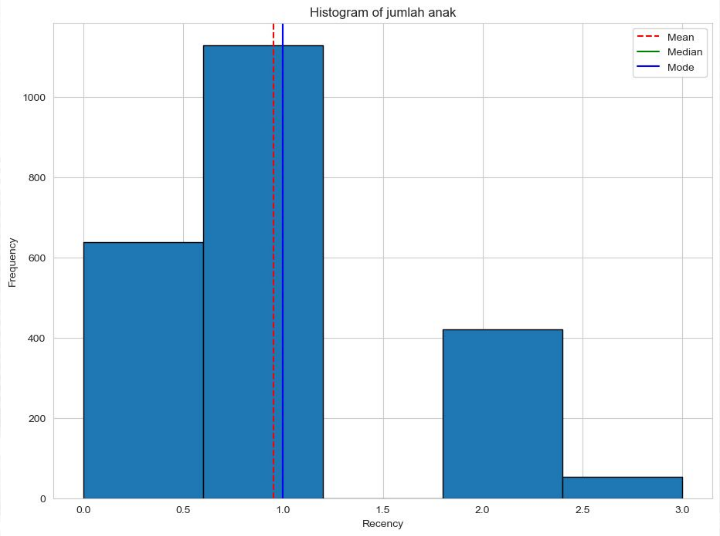


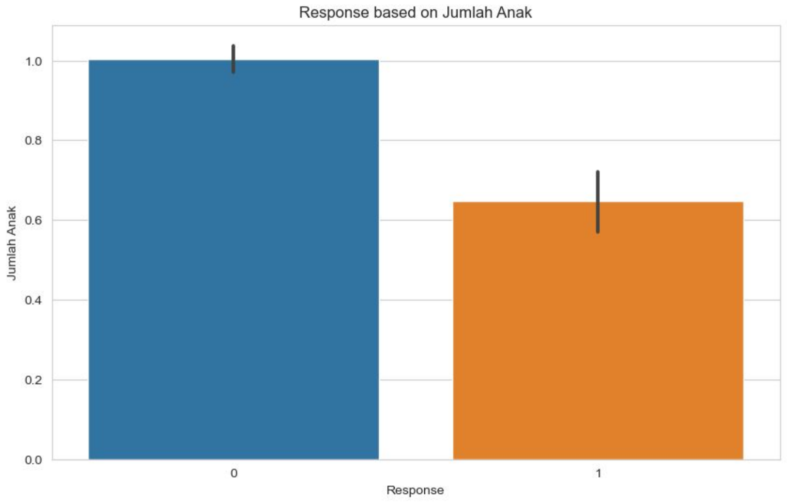


For further details: [Source Code]

Conversion Rate Analysis Based on Income, Spending and Age

Number of Children



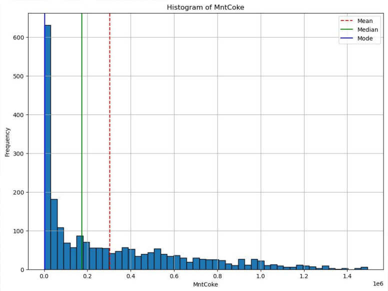


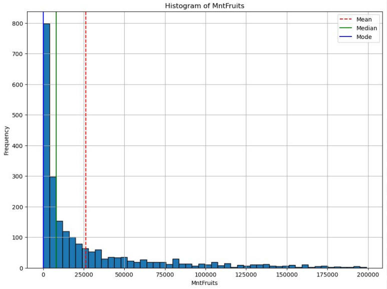
- The distribution for total children is normal as the mean (1) is equal to the median(0) with minimum number of 0 and maximum of 3
- Those who responded yes has lower number of kids and teens at home

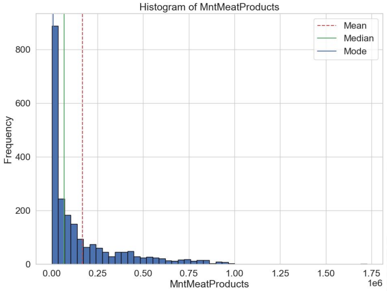
For further details: Source Code

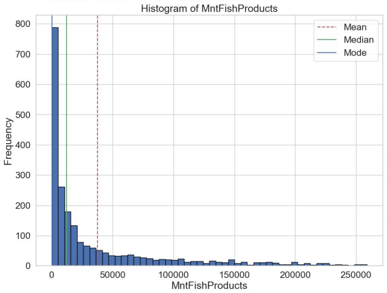
Conversion Rate Analysis Based on Income, Spending and Age

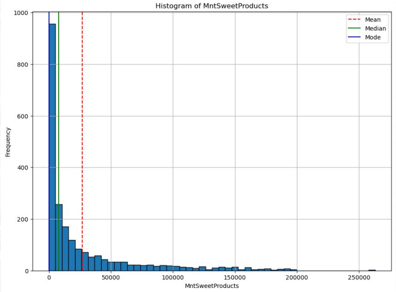
Total Expenditure for Each Category

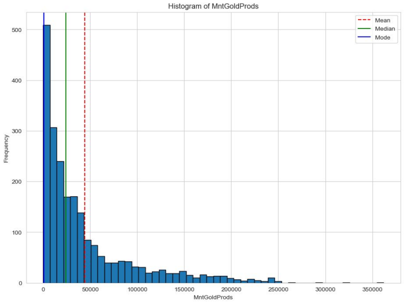










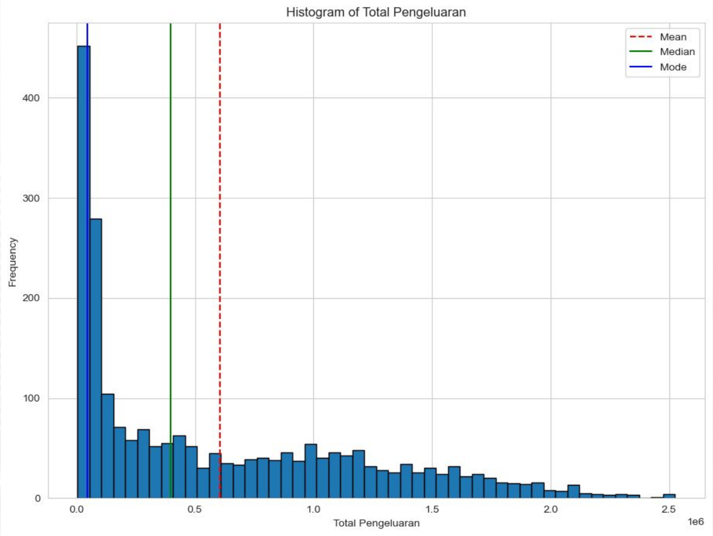


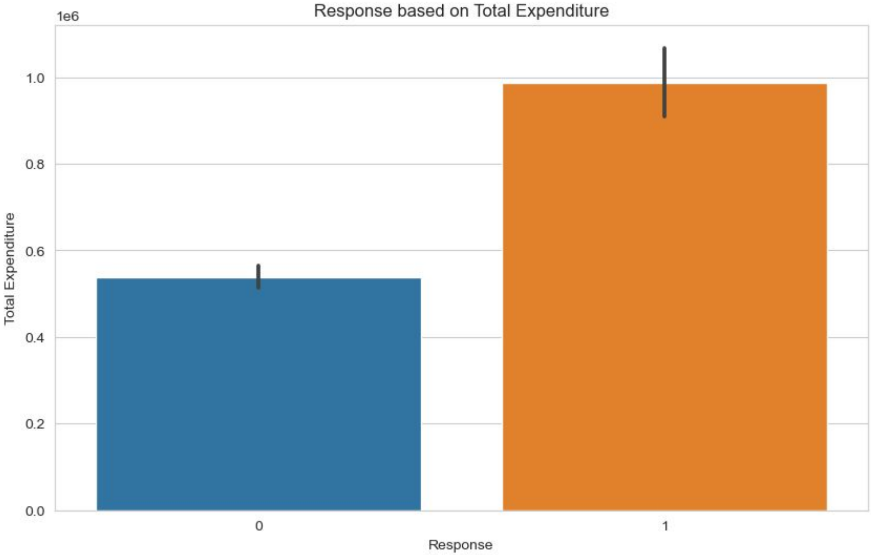
- MntCoke = the data is positive skewed, mean(3039) > med(1735), min:0, max:14930
- MntFruits = the data is positive skewed, mean(26302) > med(8000), min: 0, max: 199000
- MntMeatProducts = the data is positive skewed, mean(166950) > med(67000.00), min:0, max:1725000
- MntFishProducts = the data is positive skewed, mean(37525.4) > med(12000.0), min: 0, max: 259000.0
- MntSweetProducts = the data is positive skewed, mean(27062.9) > med(8000.0), min: 0, max: 263000.0
- MntGoldProds = the data is positive skewed, mean(44021.9) > med(24000), min: 0, max: 362000.0
- Those who responded yes bought more coke, fruits, mean products, fish products, sweet products, gold products

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Total Expenditure



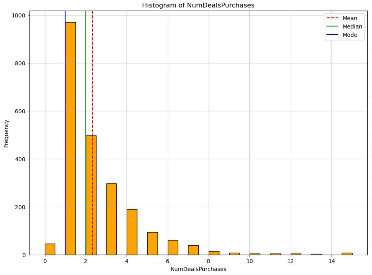


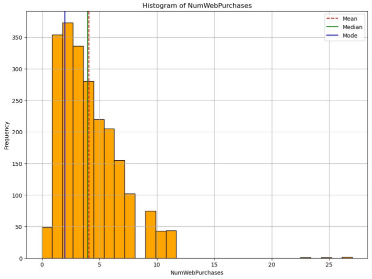
- The data for total expenditure is positive skewed with the mean (605.798) that is higher than the median (396.000).
- The minimum expenditure is 5000 while the maximum expenditure is 2525000
- Those who responded yes bought more coke, fruits, mean products, fish products, sweet products, gold products

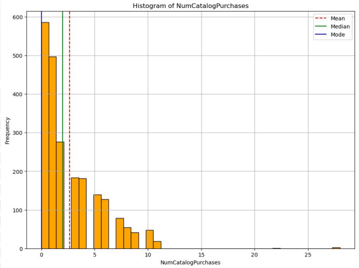
For further details: Source Code

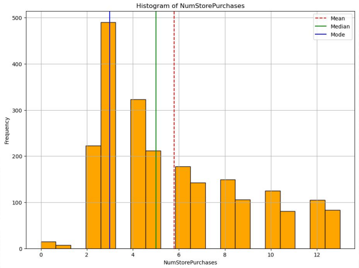
Conversion Rate Analysis Based on Income, Spending and Age

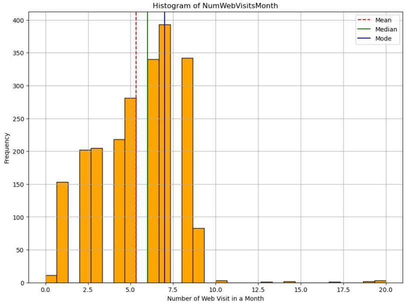
Number of Purchases









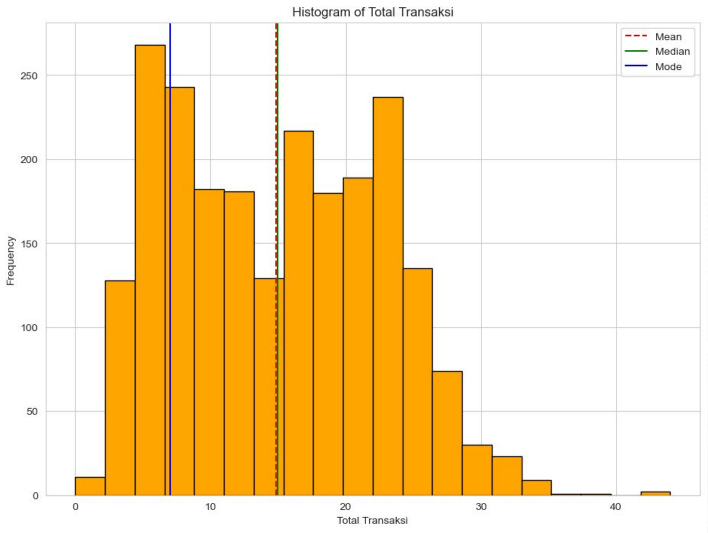


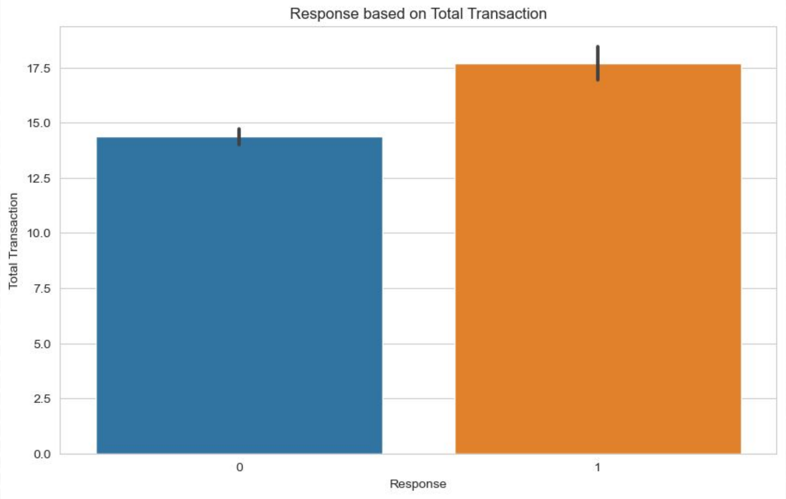
- NumDealsPurchases = near normal distribution, mean (2.3) > med (2.0), min:0, max:15
- NumWebPurchases = near normal distribution, mean (4.1) > med (4.0), min:0, max:27
- NumCatalogPurchases = near normal distribution, mean (2.7) > med (2.0), min:0, max:28
- NumStorePurchases = near normal distribution, mean (5.8) > med (5.0), min:0, max:13
- NumWebVisitsMonth = near normal distribution, med(6.0) > mean (5.3), min: 1.7m, max: 20
- Those responded yes has higher number of web, catalog, store puchases

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Total Transaction





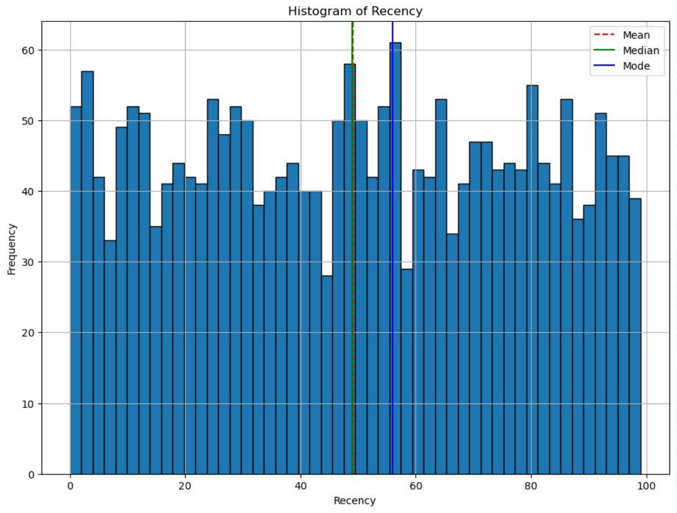
- The data for total transaction is slightly negative skewed with the mean (14.9) that is higher than the median (15).
- The minimum transaction is 0 while the maximum transaction is 44
- Customer who responded yes has higher total transaction

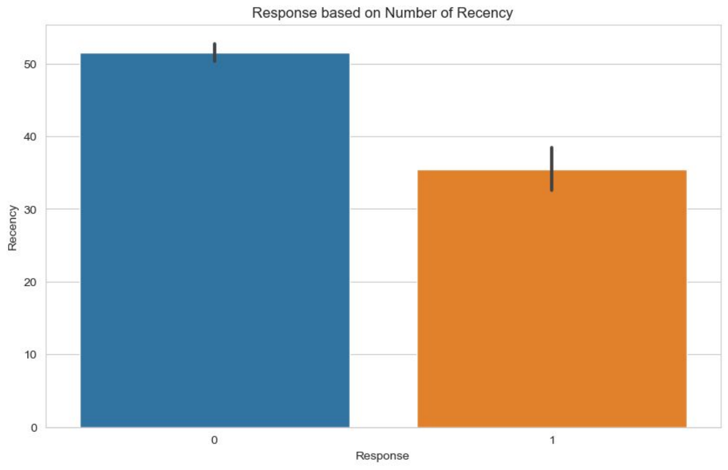
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Rakamin Academy

Recency



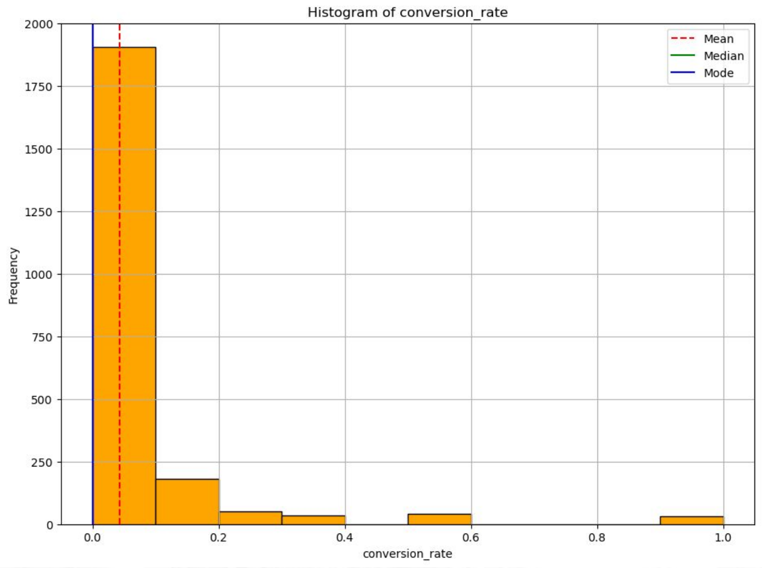


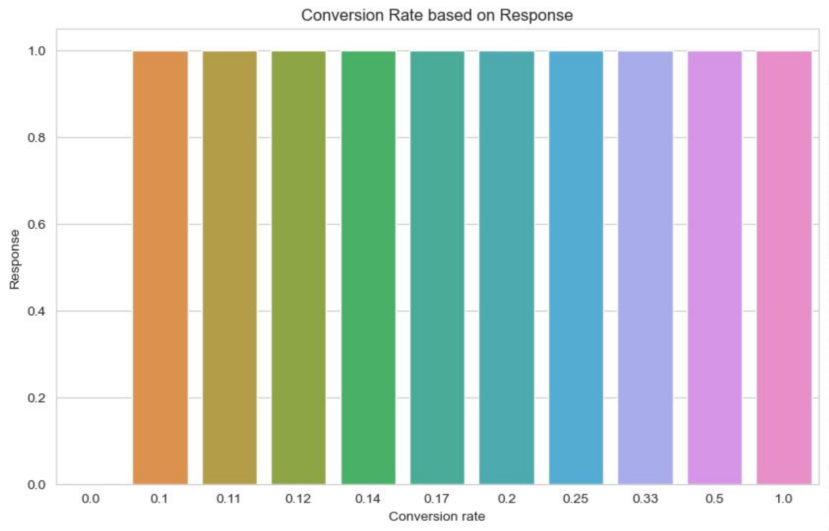
- The recency data has normal distribution whereby the mean is equal to the median (49), with minimum number of 0 and maximum of 99
- Customer who responded yes has lower recency

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Conversion Rate



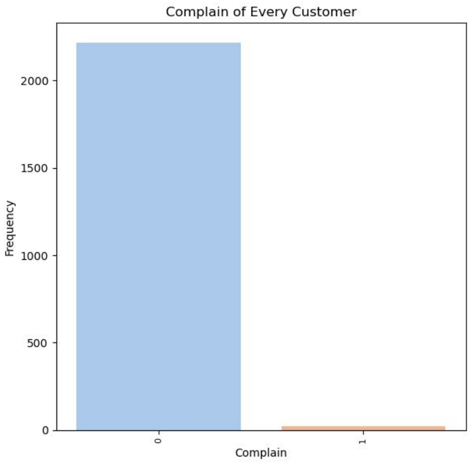


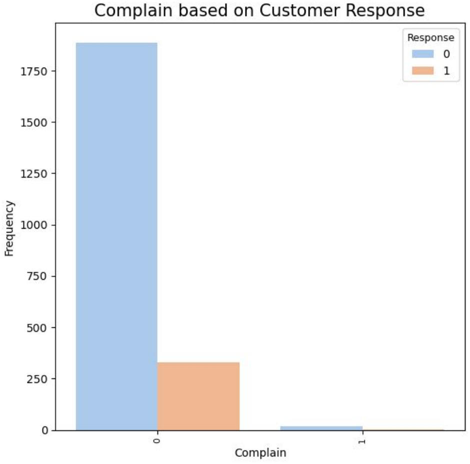
- The conversion rate data has normal distribution, whereby the mean is equal to the median (0)
- Customer who responded yes are those who are converted

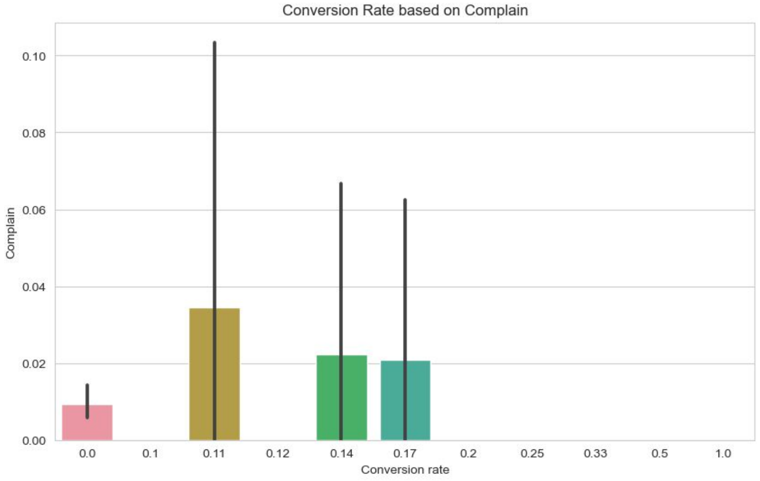
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Complain



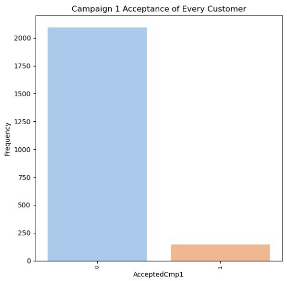


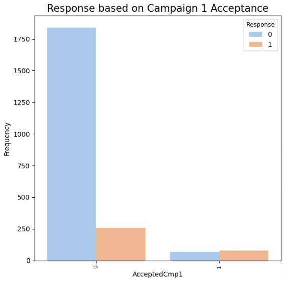


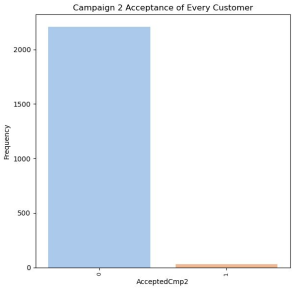
- Most of the customers did not file complain and did not respond.
- Those who responded yes did not file any complaining

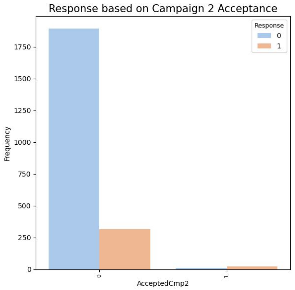
For further details: Source Code

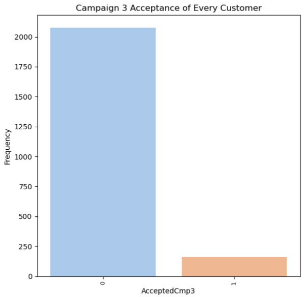
Conversion Rate Analysis Based on Income, Spending and Age

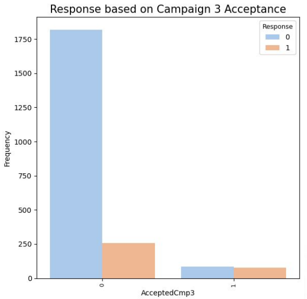


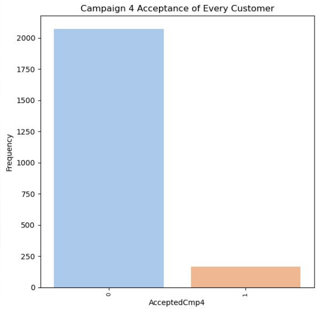


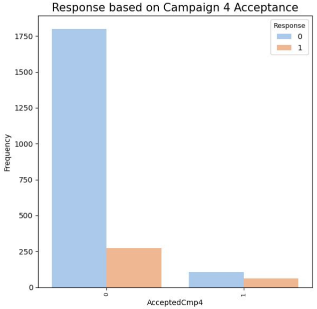


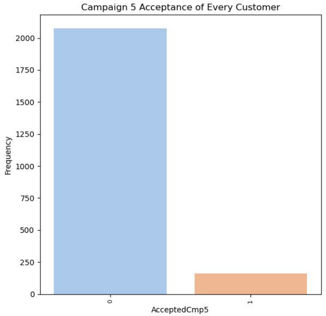


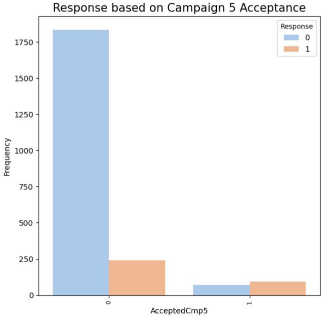












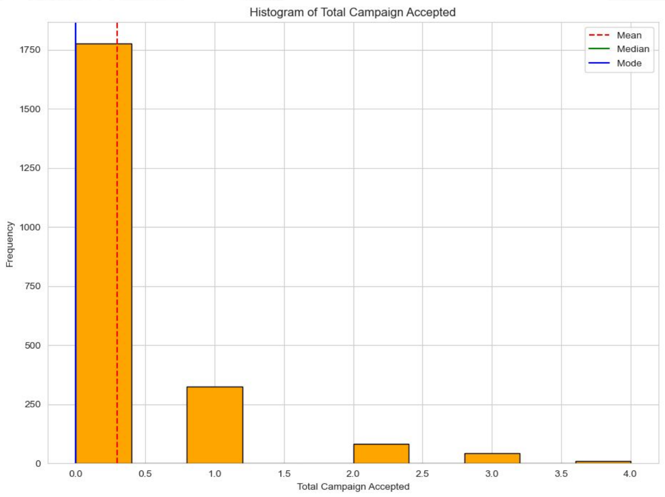
- Most of the customers did not accept campaign 1-5
- Campaign 1: Interestingly, most of customers who responded yes are among those who didnt accept the first campaign. Among those who accept the campaign, more people responded yes, however, there are still customemers who responded no. the pattern is similar with campaign 2, campaign 5
- Campaign 3: despite of accepting the campaign, more customers responded no than yes. for campaign 4, the pattern is similar with campaign 3, however, the gap is bigger between those who respond no (higher) and yes (lower) among those who accept the campaign

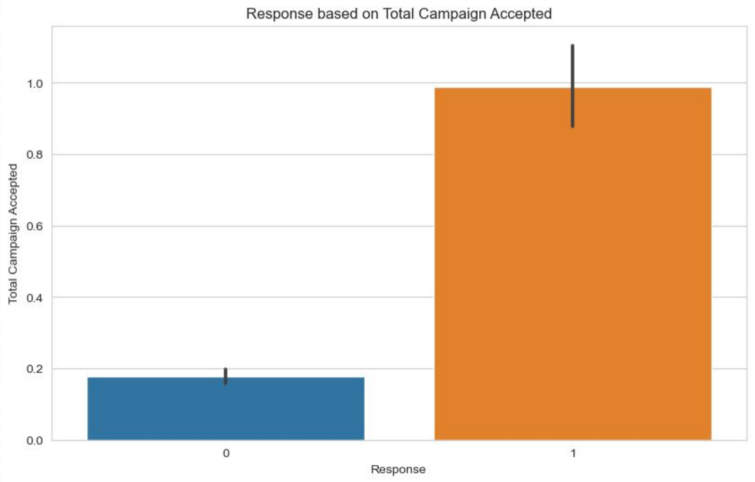
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Rakamin Academy

Total Campaign Accepted



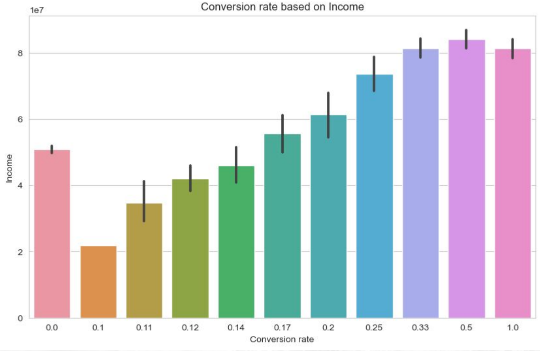


- The data distribution for the total of campaign accepted is normal with mean equal to median (0)
- The minimum number is 0 and maximum of 4 which means that not one customer accepted all 5 campaigns provided.
- Customers who responded yes has been accepting the campaigns.

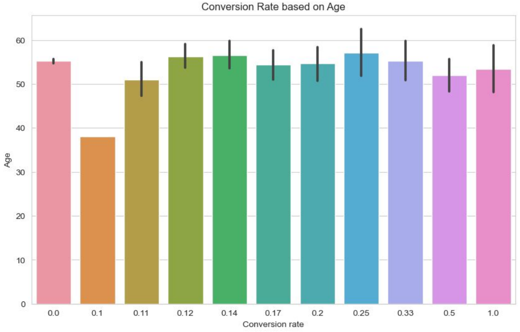
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

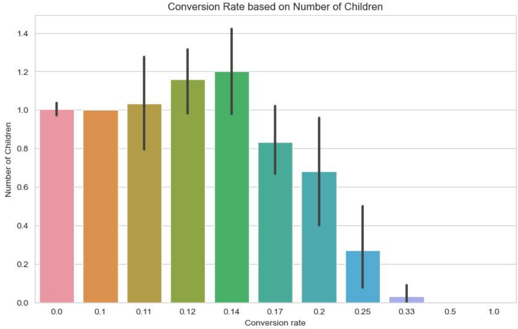
Conversion Rate Analysis



- As the income increase, the conversion rate is also increasing



- There is no specific pattern in term of age and conversion analysis. Customers who has higher conversion rate. Group 0, aged 47 and below have the highest conversion rate.



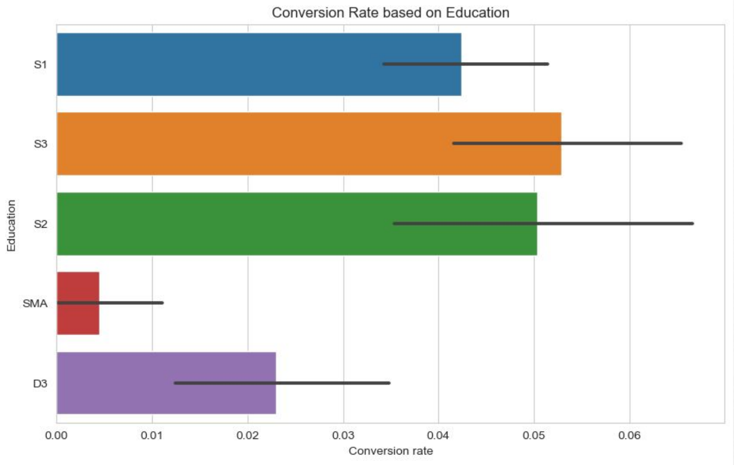
- Customers who have children may increase the conversion rate up until 14%, however, conversion rate increase as the customer did not have a children

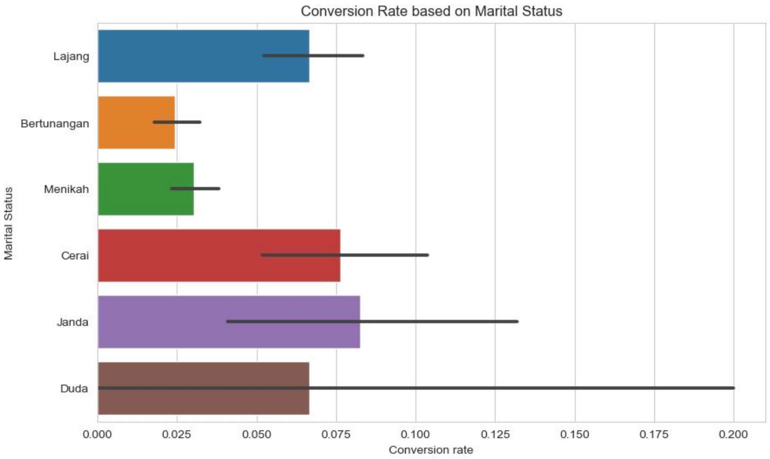
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Rakamin Academy

Conversion Rate Analysis





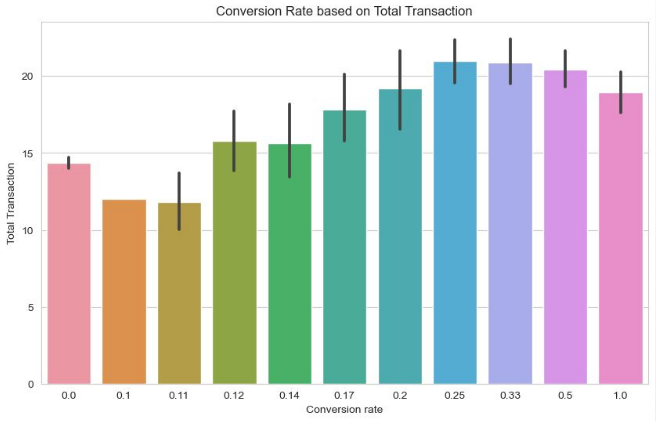
- Phd holder seems to have higher conversion rate, followed by master holder, bachelor graduate, diploma graduate. While high school graduate has the lowest conversion rate.
- Janda has the highest conversion rate, followed by cerai, lajang dan duda. Meanwhile bertunangan dan menikah have lower conversion rate

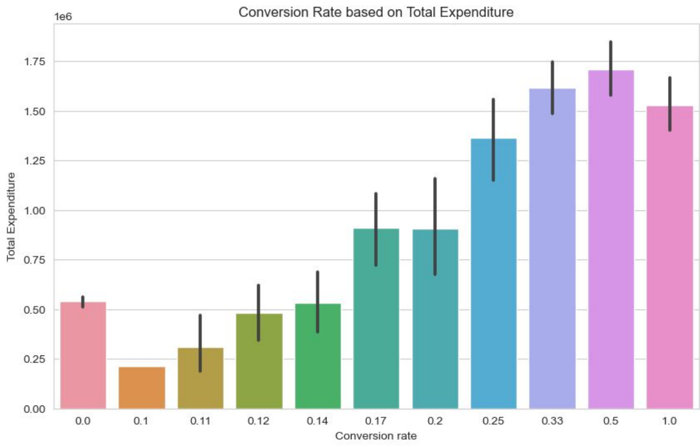
For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Rakamin Academy

Conversion Rate Analysis





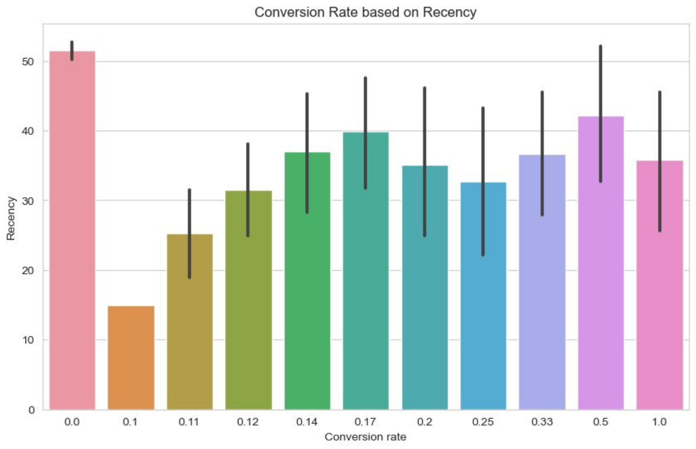
- Conversion rate increases as total transaction increase up until ~30% where there is more than 20 transactions.

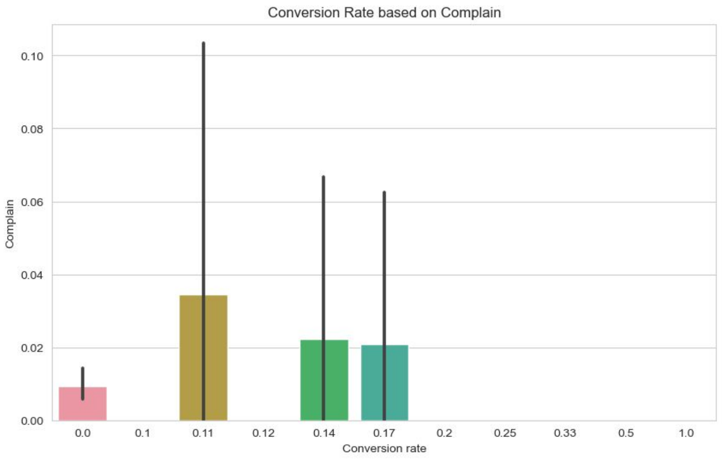
- As total expenditure increase, the conversion rate increase. It peaked in between 1.5m to 1.75m.

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Conversion Rate Analysis





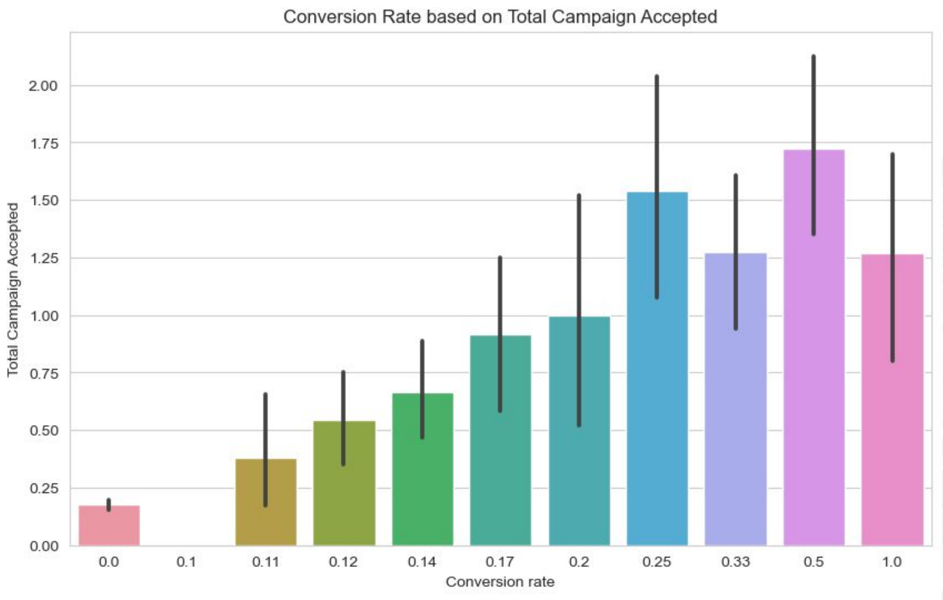
- Recency: as recency increase, conversion rate increase up until 17%, then it decreases

- Customer who had no complain have higher conversion rate

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Conversion Rate Analysis

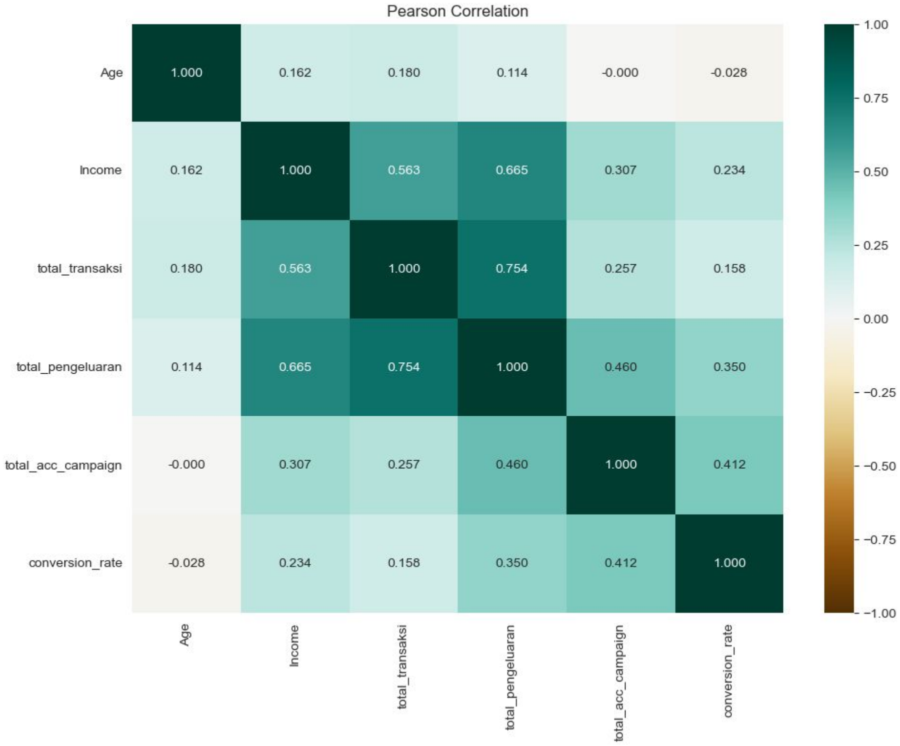


- More campaign accepted, the higher the conversion rate.
- Converted customers have at least accepted 1 campaign

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Multivariate Analysis



- Conversion rate have positive association with Income (r=0.23), total transaksi (r=0.15), total pengeluaran (r=0.35), and total_acc_campaign (r=0.41)
- Meanwhile, conversion rate have negative association with age (r=-0.028). The r value implies that the association is not significant
- With total_acc_campaign has the highest r value, it may implies that accepting more campaign have been associated with higher conversion rate

For further details: Source Code

Conversion Rate Analysis Based on Income, Spending and Age

Summary of EDA

- Customers who responded are those who are 47 years old and below, graduated as S1, single, has higher income, lower number of children and has lower recency.
- Customers with higher total transaction, expedition, and accepted campaign are those who responded yes.
- The conversion rate is higher among customers with higher income, younger age, phd holder, less number of children, no complain, responded yes, a widow, higher recency, and higher total transaction, expenditure and accepted campaign.
- There is no significant association between Age and Conversion Rate

For further details: Source Code

Data Cleaning & Preprocessing

- Missing value can be seen from two column, Income and conversion rate
  - Missing value for income is filled with median of income considering that the data distribution is skewed.
  - After further analysis, the missing value is resulted from 0 value for both column response and NumWebVisitsMonth. Hence, the missing value will be replaced with 0.

- There is no duplicate data in this data set.

For further details: Source Code

Data Cleaning & Preprocessing

Rakamin Academy

Feature Encoding

- Ordinal encoder is used for Education and Age group
  - There are in total of 5 categories which are 0=SMA (High school graduate), 1=D3 (Diploma graduate), 2=S1 (Bachelor holder), 3=S2 (Master holder), and 4=S3 (phd holder)
  - For Age Group the order is as follow, 'Childhood', 'Teen', 'Young adult', 'Middle aged adult', 'Senior adult', 'Elderly'
- One hot encoding for Marital Status
  - As there is 6 categories for marital status, hence, the 6 columns is added
- Column Education and Marital Status are dropped after being encoded
  - Hence, the total column is 42

For further details: Source Code

Data Cleaning & Preprocessing

Feature Transformation

- Standardization for column ‘Income’ was done as it has a much larger scale than other features, which can dominate the model's learning process and skew results. Standardizing this column ensures it contributes proportionally to the model, improving performance and interpretability.

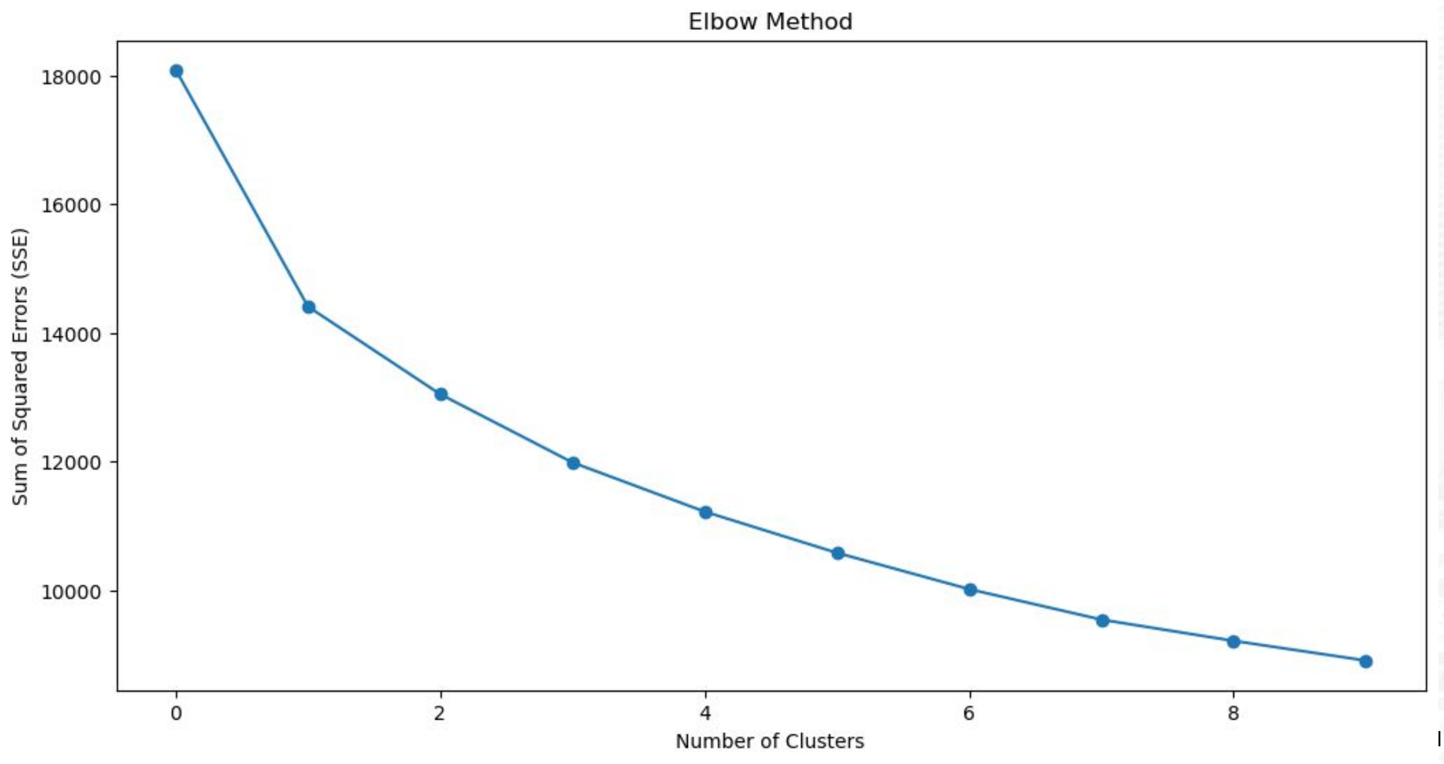
Feature Selection

- Columns that are highly correlated are dropped to reduce redundancy

For further details: Source Code

Data Modeling

- The figure below shows the visualization of **Elbow Method** by using ***K-Means Clustering.***
- It can be seen that the optimal number of cluster is 3.

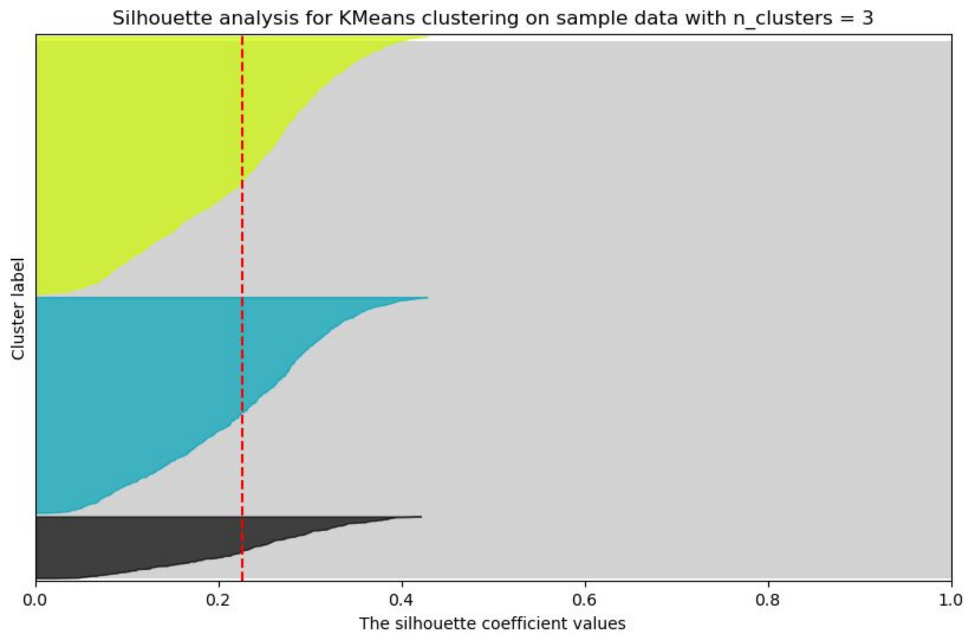


ırther details: Source Code

Data Modeling

Rakamin Academy

The figure below shows the visualization of Silhouette Score



- The clustering shows some structure, but it's not very strong as seen by the average score of approximately 0.2

- The top cluster (light green) has the widest silhouette plot, indicating it contains the most samples. It also has a large portion of its samples with silhouette coefficients higher than the average, suggesting good cluster cohesion.

- The middle cluster (teal) has a narrower plot but still shows decent cohesion with many samples above the average silhouette score.

- The bottom cluster (dark gray) is the smallest and has more samples below the average silhouette score, indicating it might be less well-defined than the others.

For further details: [Source Code]()

Customer Personality Analysis for Marketing Retargeting

Cluster 0

Characteristics (11.52% of customers)

- Highest average income
- Lowest web visits
- Highest conversion rate
- Lowest number of children
- Highest total accepted campaigns
- Medium education level
- Youngest age group

Business Recommendations:

1. Premium Targeting: These are high-value customers with high conversion rates. Target them with premium products and exclusive offers.
2. Mobile Optimization: Despite low web visits, ensure mobile platforms are optimized as they might prefer mobile shopping.
3. Loyalty Program: Implement a high-tier loyalty program to retain these valuable customers.
4. Personalized Marketing: Use their high campaign acceptance rate to send personalized, exclusive offers.

For further details: Source Code

Customer Personality Analysis for Marketing Retargeting

Cluster 1
Characteristics (40.40% of customers)
- Medium income
- Medium web visits
- Low conversion rate
- Medium number of children
- Lowest campaign acceptance
- Highest education level
- Oldest age group

Business Recommendations:
1. Educational Content: Leverage their high education level by providing in-depth product information and educational content.
2. Family-Oriented Products: Offer products suitable for families with children.
3. Senior-Friendly Interface: Ensure the website is easy to navigate for older customers.

For further details: Source Code

Customer Personality Analysis for Marketing Retargeting

Cluster 2
Characteristics (48.08% of customers):
- Lowest income
- Highest web visits
- Very low conversion rate
- Highest number of children
- Low campaign acceptance
- Lowest education level
- Medium age group

Business Recommendations:
1. Budget-Friendly Options: Offer more affordable product lines or payment plans to cater to their lower income.
2. Family Bundles: Create cost-effective family bundles or bulk purchase options.
3. Simplified Messaging: Given the lower education level, ensure marketing messages and product information are clear and straightforward.
4. Engagement to Conversion: Implement strategies to turn their high engagement (web visits) into sales, such as limited-time offers or personalized recommendations.
5. Educational Marketing: Provide value through educational content about products to build trust and encourage purchases.

For further details: Source Code