



Final Report Marketing Campaign Response Prediction

Marketective
Rakamin Data Science
Batch 38





Anggota Kelompok:

1. Candraditya Dwaya Putra
2. Muhammad Ayuvi Laksana Putra
3. Muhammad Rido
4. Putriarrum Kusuma Wardani
5. Radithya Arif Pambudi
6. Rizandhi Aulia Adipradana
7. Salwa Salsabila
8. Taufik Yasir Sukarda

...

Table of Contents

1
...

Background

2
...

EDA & Insight

3
...

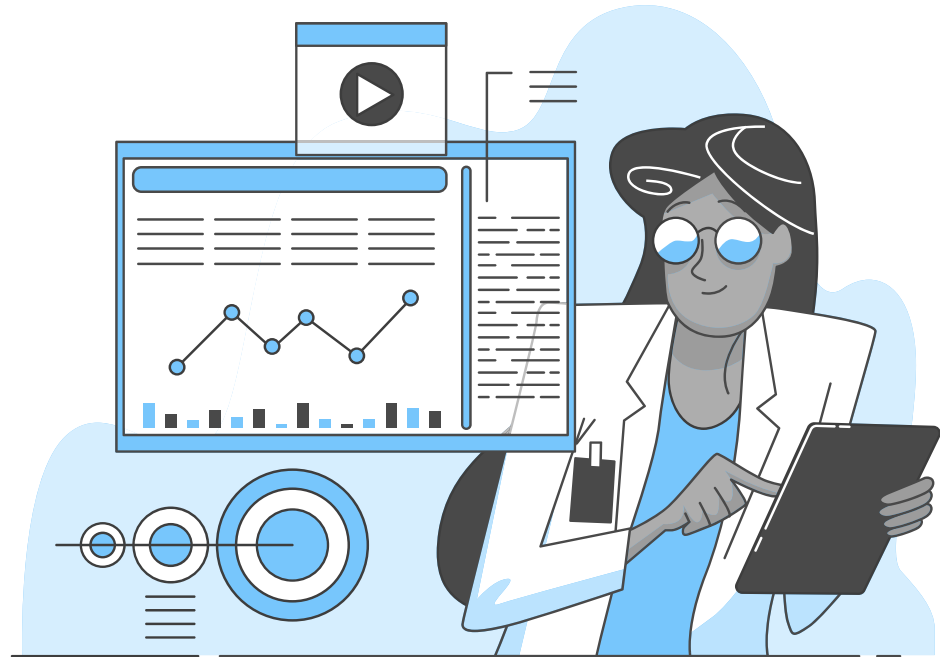
Pre-Processing

4
...

Modeling

5
...

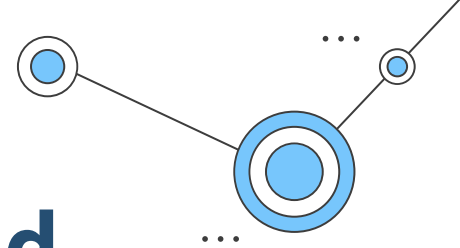
Recommendation





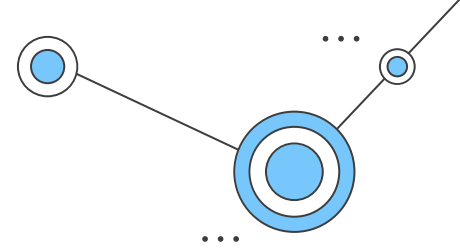
01

Background



Background Company

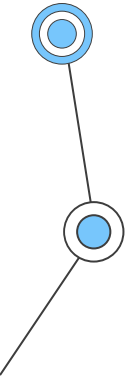
Marketective, part of the marketing division of Marketfood.

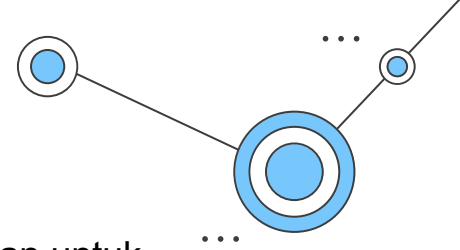


Problem

Rendahnya jumlah konsumen yang merespon marketing campaign.

Tim bisnis Marketfood mengindikasi ketidak maksimalan campaign yang dilakukan sebelumnya, sehingga meminta Marketectives memberikan insight bisnis. Hanya 334 dari total 2,240 konsumen yang merespon marketing campaign yang diberikan.





Goals

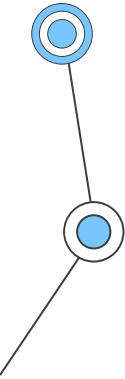
Meningkatkan Respon konsumen atas marketing campaign yang diberikan untuk mendapatkan profit semaksimal mungkin.

Objectives

- Initial selection of potential customers.
- Building a machine learning model that can predict consumers likely to respond to the campaign.

Business Metrics

- Marketing Provit
- Respon Rate



02

EDA & Insight

#	Column	Non-Null Count	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	Z_CostContact	2240 non-null	int64
27	Z_Revenue	2240 non-null	int64
28	Response	2240 non-null	int64

dtypes: float64(1), int64(25), object(3)
memory usage: 507.6+ KB

Descriptive Statistics

1. Dataset terdiri dari 29 columns dan 2240 rows data
2. Variasi tipe data yang ada pada dataset adalah int64, object, float64

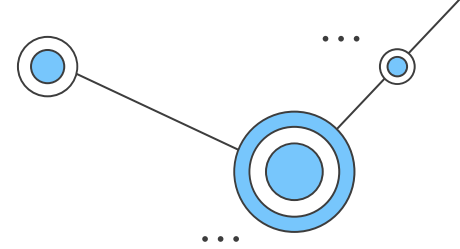
Pada kolom Income memiliki 2216 rows nilai non-null dari keseluruhan 2240 rows yang berarti terdapat 24 nilai null / missing values sedangkan untuk kolom lainnya jumlah rows non-null sama dengan keseluruhan rows data yang ada

3. Type data keseluruhan sudah tidak mengalami masalah hanya sebaiknya pada data Dt_customer yang isinya berkaitan dengan waktu dan tanggal ada baiknya dijadikan ke tipe Datetime saja

	null_counts	%
Income	24	1.07
ID	0	0.00
NumDealsPurchases	0	0.00
Z_Revenue	0	0.00
Z_CostContact	0	0.00
Complain	0	0.00
AcceptedCmp2	0	0.00
AcceptedCmp1	0	0.00
AcceptedCmp5	0	0.00
AcceptedCmp4	0	0.00
AcceptedCmp3	0	0.00
NumWebVisitsMonth	0	0.00
NumStorePurchases	0	0.00
NumCatalogPurchases	0	0.00
NumWebPurchases	0	0.00
MntGoldProds	0	0.00
Year_Birth	0	0.00
MntSweetProducts	0	0.00
MntFishProducts	0	0.00
MntMeatProducts	0	0.00
MntFruits	0	0.00
MntWines	0	0.00
Recency	0	0.00
Dt_Customer	0	0.00
Teenhome	0	0.00
Kidhome	0	0.00
Marital_Status	0	0.00
Education	0	0.00
Response	0	0.00

Missing Value

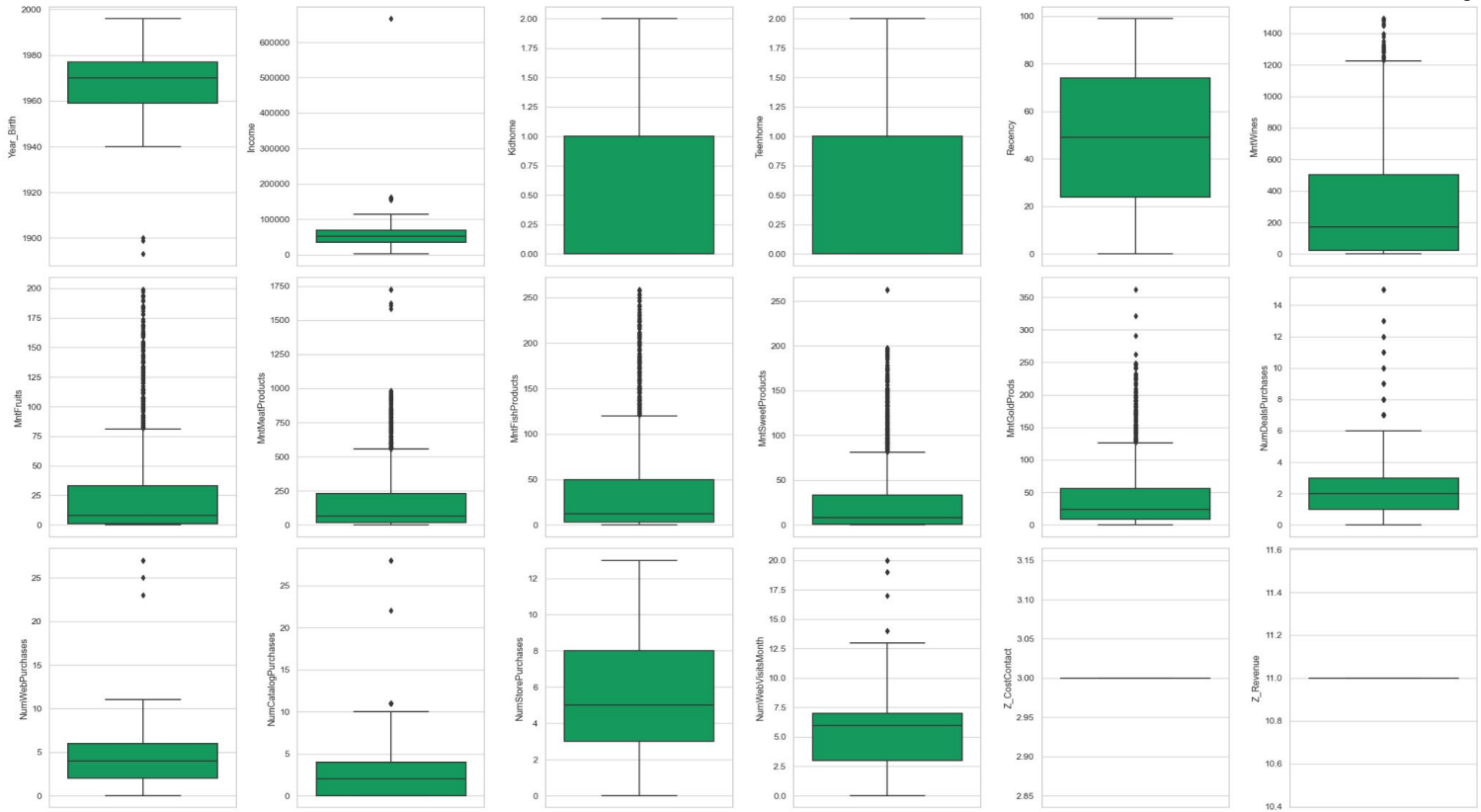
Pada column Income terdapat 24 nilai Null atau sekitar 1%, karena nilainya $< 10\%$ maka dilakukan handling nilai Null



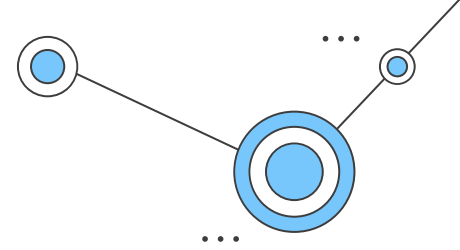
	count	unique	top	freq		
Education	2240	4	Graduation	1127	% Graduation	: 50.31
Marital_Status	2240	5	Married	864	% Married	: 38.57
Dt_Customer	2240	663	2012-08-31	12	% 2012-08-31	: 0.54

Pada data categorical, terdapat 3 features yaitu 'Education', 'Marital_Status' dan 'Dt_Customer'.

1. Pada kolom 'Education' terdapat 4 unique values yang di dominasi oleh Graduation dengan nilai 50.31%.
2. Pada kolom 'Marital_Status' terdapat 5 unique values yang di dominasi oleh Married dengan 38.57%
3. Pada kolom 'Dt_Customer' terdapat 663 unique values yang di dominasi oleh tanggal 2012-08-31 dengan 0,54%



UNIVARIATE ANALYSIS



Terdapat **outliers** pada beberapa kolom di dalam dataset, diantaranya:

- Year_Birth, outlier terjauh adalah di bawah 1900
- Income, outlier terjauh adalah di atas \$600,000
- MntWines, outlier berada pada angka 1200 keatas
- MntFruits, outlier berada di sekitar angka 80 sampai 200
- MntMeatProducts, outlier terjauh ada di sekitar angka 1,750
- MntFishProducts, outlier berada di sekitar angka 125 sampai diatas 250
- MntSweetProducts, outlier terjauh berada di sekitar angka 250
- MntGoldProds, outlier terjauh berada di sekitar angka 350 baris data.
- NumDealsPurchases, outlier terjauh berada di angka 15
- NumWebPurchases, outlier berada di sekitar angka 25
- NumCatalogPurchases, outlier terjauh berada di atas angka 25
- NumWebVisitMonth, outlier terjauh berada di angka 20

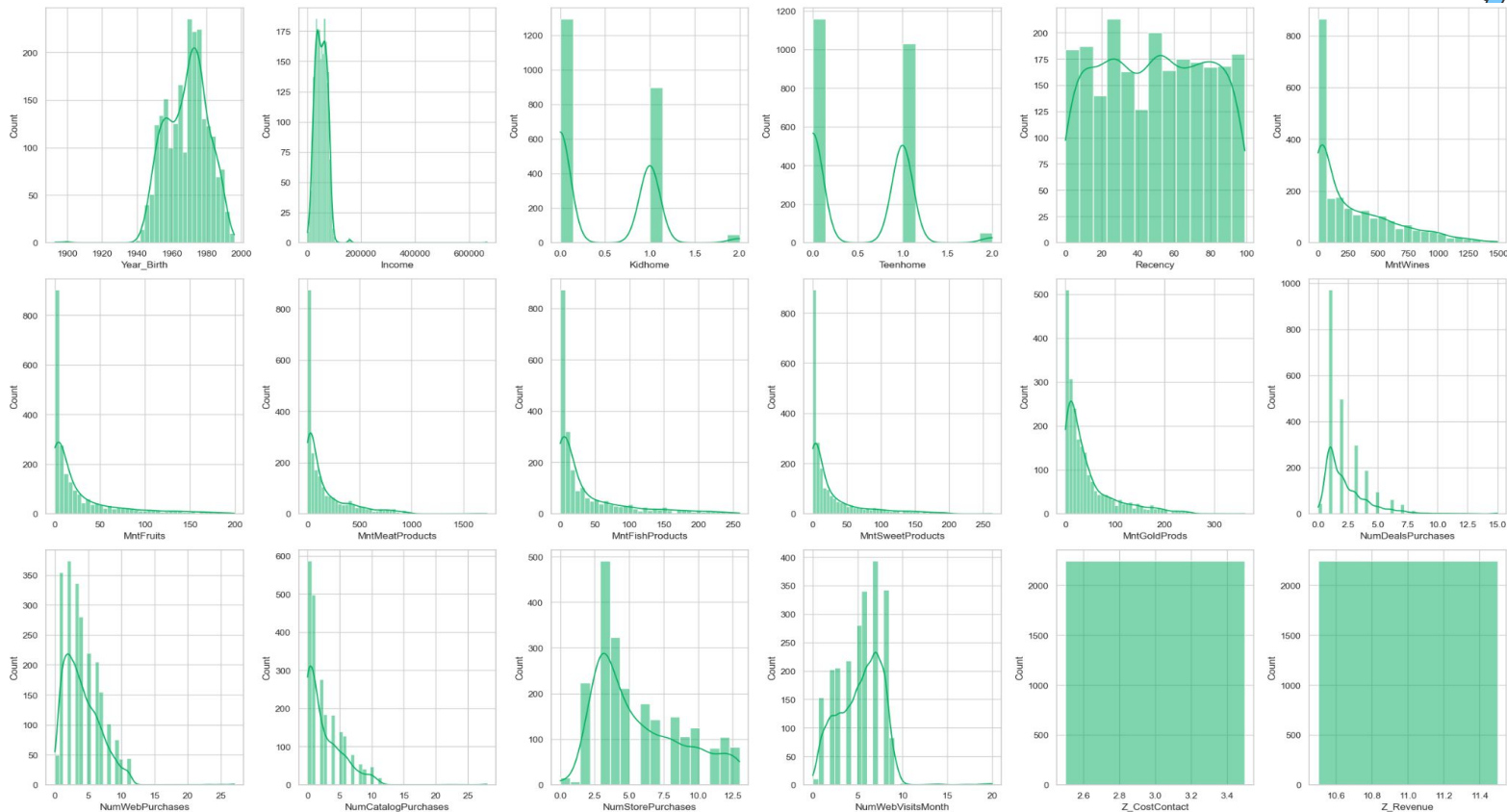
Hal yang harus dilakukan pada saat **Data Pre-Processing** adalah:

- karena jumlah data yang terbatas yaitu sebanyak **2240 baris** data saja, ada baiknya untuk dipilih handling yang sebisa mungkin tidak mengurangi jumlah data, atau dapat dipilih model yang robust terhadap outlier.
- **Sebagai opsi alternatif**, membersihkan data dengan **menghapus outlier** menggunakan **metode IQR** atau **Z-score** juga merupakan pilihan, namun **tindakan ini akan mengurangi jumlah data yang tersedia**.



Univariate Analysis

DISTRIBUTION AND TABLE OF NUMERICAL VALUES





Berdasarkan pola distribusi data yang terlihat, dapat disimpulkan bahwa beberapa variabel mengandung outlier dan beberapa di antaranya menunjukkan distribusi yang miring (Skewed Distribution). Berikut adalah daftar variabel berdasarkan jenis distribusinya:

A. Distribusi Normal

- Recency Normal Distribution (Simetris)
- Year_Birth Moderately Normal Distribution (Simetris)
- NumWebVisitsMonth Moderately Normal Distribution (Simetris)

B. Distribusi Seragam (Uniform)

- Z_CostContact - Distribusi Seragam dengan hanya satu nilai
- Z_Revenue - Distribusi Seragam dengan hanya satu nilai

C. Distribusi Positive Skewed (Miring Positif)

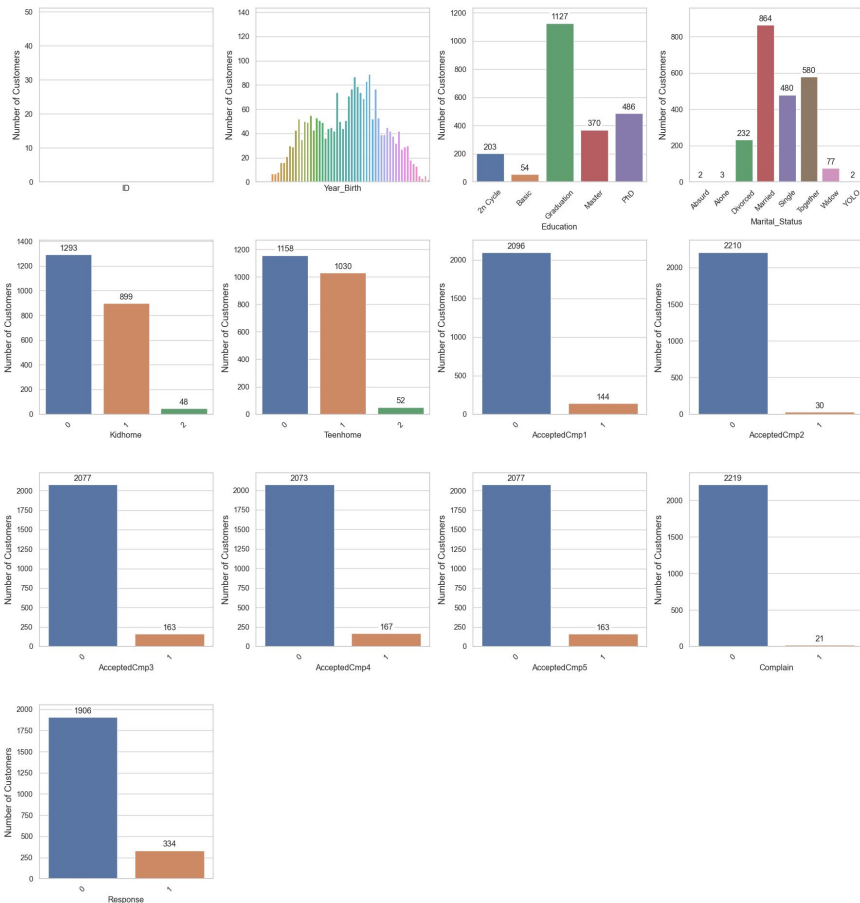
Pendapatan (Income)

- Jumlah Pembelian Produk Wine (MntWines)
- Jumlah Pembelian Produk Buah (MntFruits)
- Jumlah Pembelian Produk Daging (MntMeatProducts)
- Jumlah Pembelian Produk Ikan (MntFishProducts)
- Jumlah Pembelian Produk Manis (MntSweetProducts)
- Jumlah Pembelian Produk Emas (MntGoldProds)
- Jumlah Pembelian Paket Spesial (NumDealsPurchases)
- Jumlah Pembelian melalui Web (NumWebPurchases)
- Jumlah Pembelian melalui Katalog (NumCatalogPurchases)
- Jumlah Pembelian melalui Toko (NumStorePurchases)

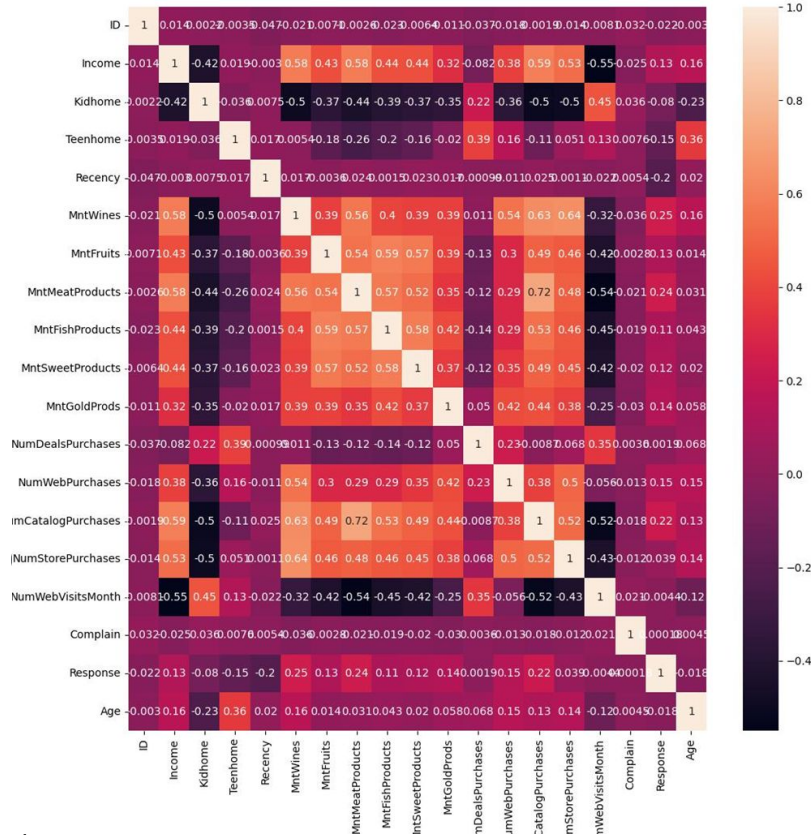
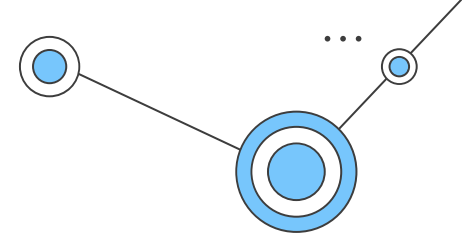
D. Distribusi Bimodal

- Jumlah Anak Kecil dalam Rumah Tangga Pelanggan (Kidhome)
- Jumlah Remaja dalam Rumah Tangga Pelanggan (Teenhome)

	Column Name	Skewness	Kurtosis	Type of Distribution
0	Year_Birth	-0.350	0.713	Moderately Normal Distribution (Symmetric)
1	Income	6.759	159.274	Highly Positively Skewed
2	Kidhome	0.635	-0.781	Bimodal Distribution
3	Teenhome	0.407	-0.987	Bimodal Distribution
4	Recency	-0.002	-1.202	Normal Distribution (Symmetric)
5	MntWines	1.175	0.595	Highly Positively Skewed
6	MntFruits	2.101	4.039	Highly Positively Skewed
7	MntMeatProducts	2.082	5.502	Highly Positively Skewed
8	MntFishProducts	1.918	3.087	Highly Positively Skewed
9	MntSweetProducts	2.135	4.364	Highly Positively Skewed
10	MntGoldProds	1.885	3.541	Highly Positively Skewed
11	NumDealsPurchases	2.417	8.914	Highly Positively Skewed
12	NumWebPurchases	1.382	5.688	Highly Positively Skewed
13	NumCatalogPurchases	1.880	8.027	Highly Positively Skewed
14	NumStorePurchases	0.702	-0.623	Moderately Positively Skewed
15	NumWebVisitsMonth	0.208	1.815	Moderately Normal Distribution (Symmetric)
16	Z_CostContact	NaN	NaN	Uniform Distribution
17	Z_Revenue	NaN	NaN	Uniform Distribution

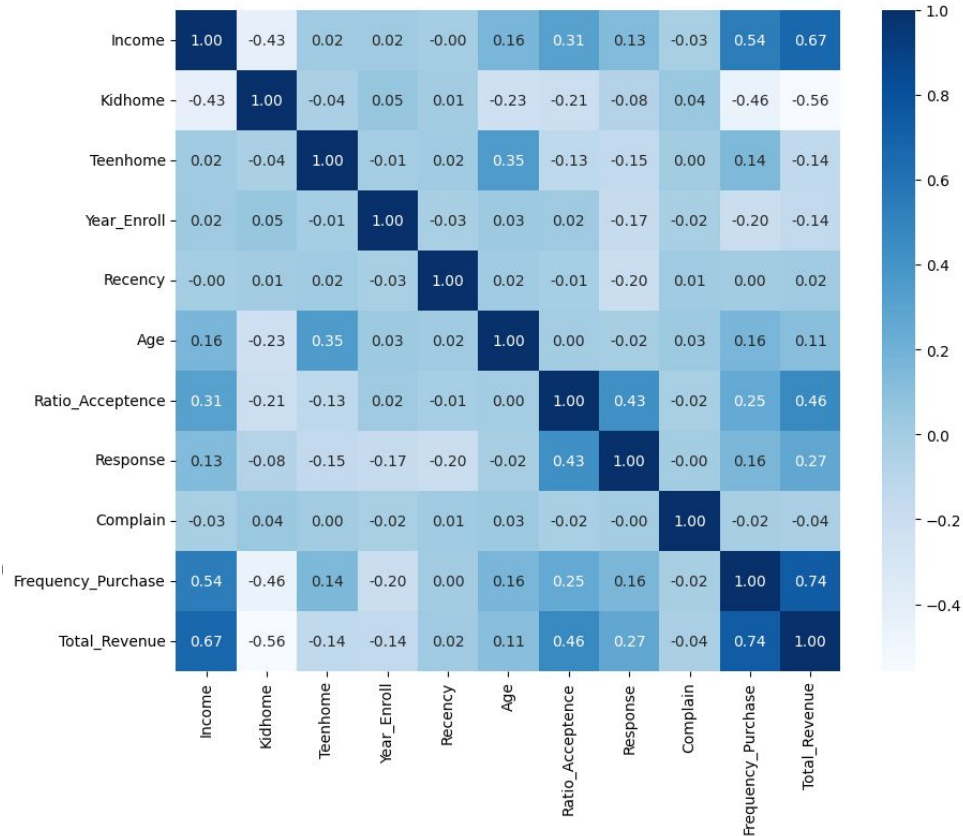
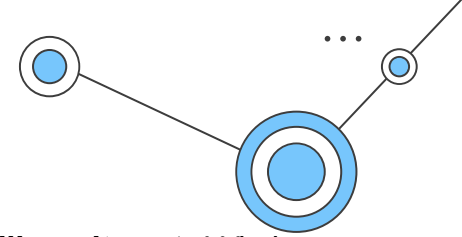


GENERAL MULTIVARIATE ANALYSIS



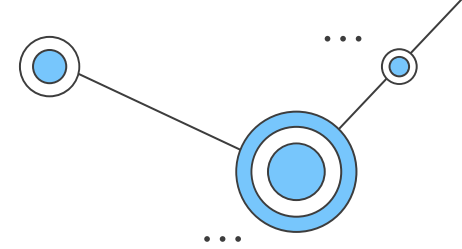
- **Income** memiliki korelasi baik **dengan **NumCatalogPurchase & NumStorePurchase** tetapi berbanding terbalik dengan **NumWebVisitMonth**
- **Income** memiliki korelasi baik dengan **Jumlah barang yang dibeli (Mnt..Products)**
- Banyaknya jumlah anak memiliki korelasi yang berbanding terbalik dengan jumlah barang yang dibeli(Mnt..Products), pesanan yang dibeli menggunakan web, catalog & store akan tetapi memiliki korelasi yang baik dengan web visit
- Begitu pula dengan teenhome akan tetapi dengan nilai korelasi yang lebih kecil.
- Beberapa produk memiliki korelasi yang lebih baik antara satu dan lainnya, hal ini mungkin dapat dijadikan acuan untuk mengkombinasi penawaran produk (ex : buah dan lainnya , daging dan wine)

SPESIFIC MULTIVARIATE ANALYSIS

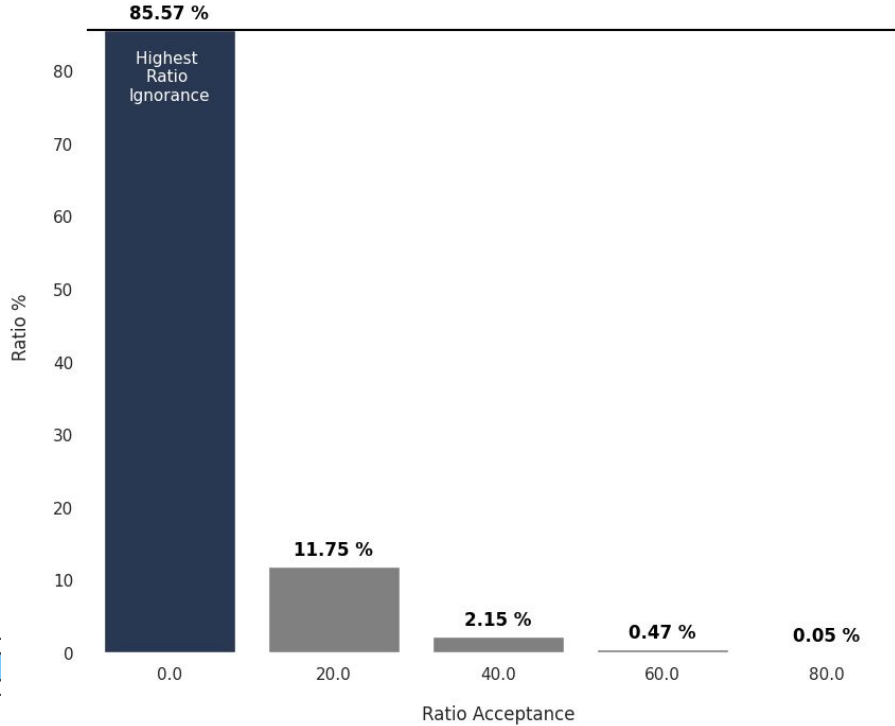


Multivariate Numeric:

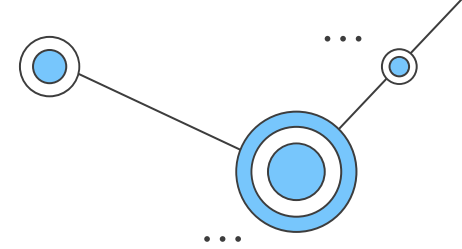
- Tidak terdapat multicollinearity antara kolom sehingga **semua kolom dapat dipertahankan** atau tidak perlu dieliminasi.
- Terdapat beberapa kolom numerik yang berkorelasi cukup kuat dengan target (`Response`) baik berkorelasi positif maupun negatif.
- Kolom numerik yang memiliki **korelasi positif** cukup kuat dengan **target (`Response`)** yaitu **Ratio_Acceptance (0.43)**, **Total_Revenue (0.27)**, **Total_Purchase (0.16)**, dan **Income (0.13)**.
- Kolom numerik yang memiliki **korelasi negatif** cukup kuat dengan **target (`Response`)** yaitu **Recency (-0.20)**, **Year_Enroll (-0.17)**, dan **Teenhome (-0.15)**.
- Target (`Response`) memiliki korelasi yang terkuat dengan **Total_Campaign**.



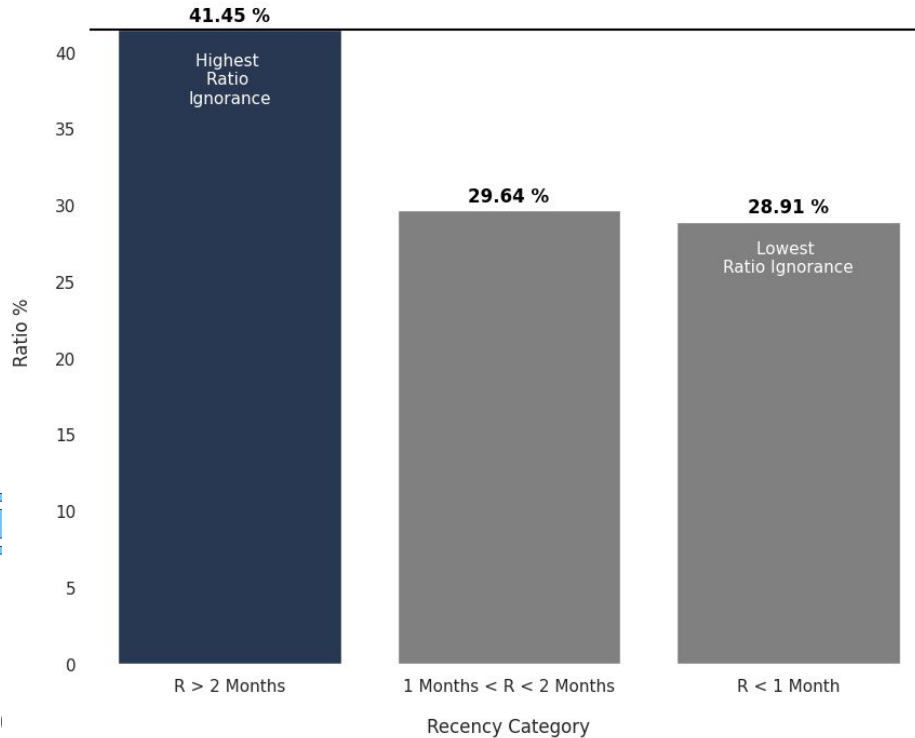
Ratio Ignorance by Ratio_Acceptance



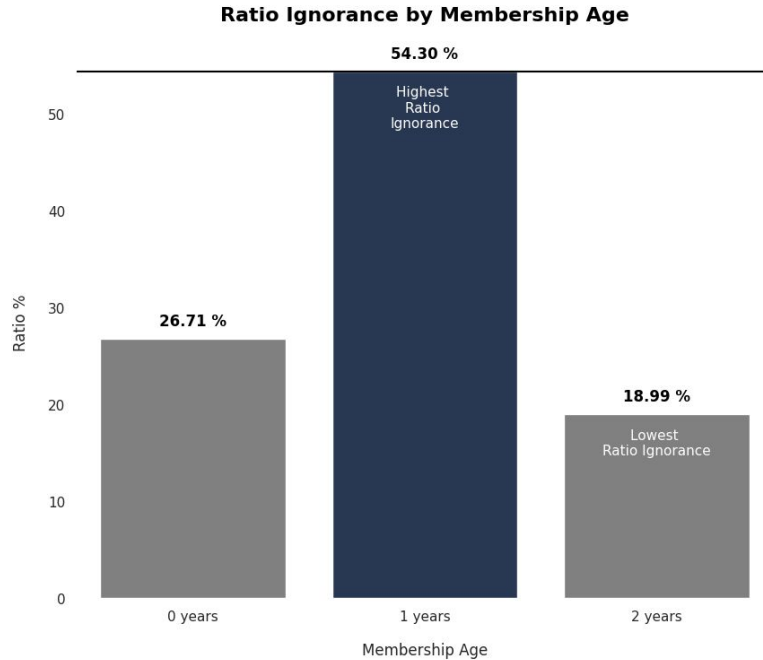
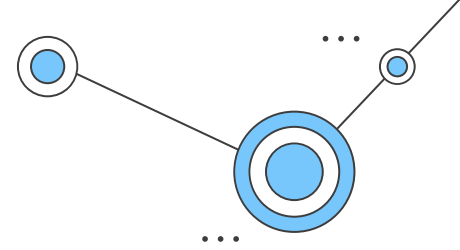
Sebagian besar pelanggan yang menolak campaign merupakan pelanggan yang belum pernah menerima campaign sebelumnya yaitu sebanyak (1,631).



Ratio Ignorance by Recency Category



Hampir setengah dari seluruh pelanggan yang menolak campaign merupakan pelanggan yang sudah tidak bertransaksi selama 2 bulan terakhir sebanyak 551.



Sebagian besar pelanggan yang menolak campaign merupakan pelanggan yang sudah berlangganan selama 1 tahun sebanyak 1,035.

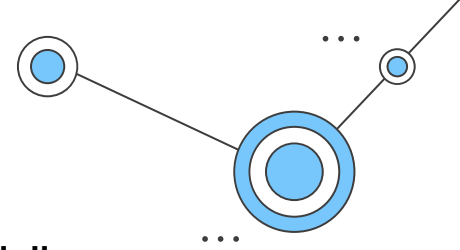
03

Pre-Processing

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2240 entries, 0 to 2239
Data columns (total 29 columns):
#   Column                Non-Null Count  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  Z_CostContact        2240 non-null   int64
27  Z_Revenue            2240 non-null   int64
28  Response             2240 non-null   int64
dtypes: float64(1), int64(25), object(3)
```

Summary

1. Dataset terdiri dari **29 columns** dan **2240 rows** data
2. Variasi tipe data yang ada pada dataset adalah **int64, object, float64**
3. Pada kolom **Income** memiliki **2216 rows** nilai **non-null** dari keseluruhan **2240 rows** yang berarti terdapat **24 nilai null / missing values** sedangkan untuk kolom lainya jumlah rows non-null sama dengan keseluruhan rows data yang ada
4. Type data keseluruhan sudah tidak mengalami masalah hanya sebaiknya pada data Dt_customer yang isinya berkaitan dengan waktu dan tanggal ada baiknya dijadikan ke tipe Datetime saja

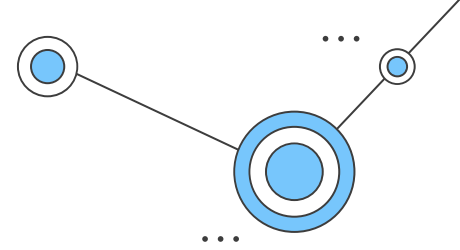


Presentase nilai null sebelum dropna:

ID	0.000000
Year_Birth	0.000000
Education	0.000000
Marital_Status	0.000000
Income	1.071429
Kidhome	0.000000
Teenhome	0.000000
Dt_Customer	0.000000
Recency	0.000000
MntWines	0.000000
MntFruits	0.000000
MntMeatProducts	0.000000
MntFishProducts	0.000000
MntSweetProducts	0.000000
MntGoldProds	0.000000
NumDealsPurchases	0.000000
NumWebPurchases	0.000000
NumCatalogPurchases	0.000000
NumStorePurchases	0.000000
NumWebVisitsMonth	0.000000
AcceptedCmp3	0.000000
AcceptedCmp4	0.000000
AcceptedCmp5	0.000000
AcceptedCmp1	0.000000
AcceptedCmp2	0.000000

Handling Missing Value

1. Pada column Income terdapat **24** nilai **Null** atau sekitar **1%**, karena nilainya **< 10%** maka **dilakukan handling nilai Null**.
2. Setelah dilakukan handling Null jumlah data **sekarang berjumlah 2216**.
3. Pada column Marital_Status terdapat value **Alone, YOLO, dan Absurd** yang diasumsikan sama dengan **Single** maka untuk value tersebut dilakukan perubahan menjadi Single.
4. Pada column Education terdapat **2n Cycle dan Master** yang sebenarnya setara maka untuk value **2n Cycle** diubah menjadi **Master**.



Handling Duplicate

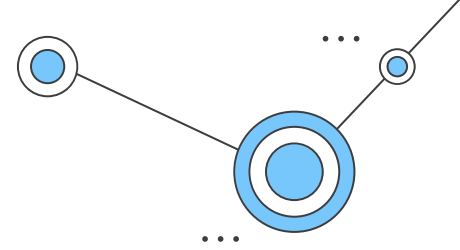
✓ Data Duplicate

```
[ ] df_clean.duplicated('ID').sum()
```

0

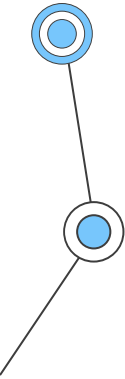
1. **Pengecekan duplicate dilakukan berdasarkan ID** untuk mengetahui apakah data ganda berdasarkan ID customer, setelah dilakukan pengecekan didapatkan hasil bahwa **tidak ada duplicate sehingga tidak dilakukan handling duplicate.**

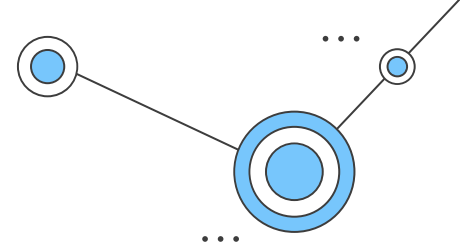




Handling Outlier

karena jumlah data yang terbatas yaitu sebanyak **2240 baris** data saja, ada baiknya untuk dipilih handling yang sebisa mungkin tidak mengurangi jumlah data, atau dapat dipilih model yang robust terhadap outlier.





Future Extraction

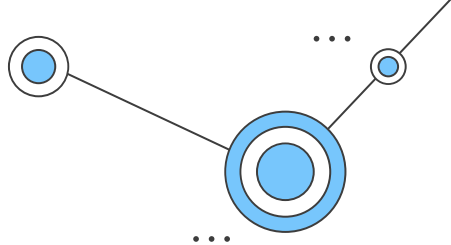
1. Customer_Age
2. Membership_Age
3. Ratio Acceptance
4. Frequency_Online_Purchase
5. Frequency_Offline_Purchase
6. Frequency_Promo_Purchase
7. Consumable_Expenses
8. Non_Consumable_Expenses
9. Favorite_product
10. 2nd_Favorite_product

Int64Index: 2216 entries, 0 to 2239

Data columns (total 21 columns):

#	Column	Non-Null Count	Dtype
0	Education	2216 non-null	object
1	Marital_Status	2216 non-null	object
2	Income	2216 non-null	float64
3	Kidhome	2216 non-null	int64
4	Teenhome	2216 non-null	int64
5	Recency	2216 non-null	int64
6	NumWebVisitsMonth	2216 non-null	int64
7	Complain	2216 non-null	int64
8	Z_CostContact	2216 non-null	int64
9	Z_Revenue	2216 non-null	int64
10	Response	2216 non-null	int64
11	Membership_Age	2216 non-null	int64
12	Customer_Age	2216 non-null	int64
13	Ratio_Acceptance	2216 non-null	float64
14	Frequency_Online_Purchase	2216 non-null	int64
15	Frequency_Offline_Purchase	2216 non-null	int64
16	Frequency_Promo_Purchase	2216 non-null	int64
17	Consumable_Expenses	2216 non-null	int64
18	Non_Consumable_Expenses	2216 non-null	int64
19	Favorite_Product	2216 non-null	object
20	2nd_Favorite_Product	2216 non-null	object

dtypes: float64(2), int64(15), object(4)



Future Selection

Int64Index: 2216 entries, 0 to 2239

Data columns (total 21 columns):

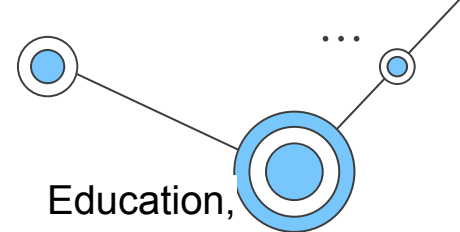
#	Column	Non-Null Count	Dtype
0	Education	2216 non-null	object
1	Marital_Status	2216 non-null	object
2	Income	2216 non-null	float64
3	Kidhome	2216 non-null	int64
4	Teenhome	2216 non-null	int64
5	Recency	2216 non-null	int64
6	NumWebVisitsMonth	2216 non-null	int64
7	Complain	2216 non-null	int64
8	Z_CostContact	2216 non-null	int64
9	Z_Revenue	2216 non-null	int64
10	Response	2216 non-null	int64
11	Membership_Age	2216 non-null	int64
12	Customer_Age	2216 non-null	int64
13	Ratio_Acceptance	2216 non-null	float64
14	Frequency_Online_Purchase	2216 non-null	int64
15	Frequency_Offline_Purchase	2216 non-null	int64
16	Frequency_Promo_Purchase	2216 non-null	int64
17	Consumable_Expenses	2216 non-null	int64
18	Non_Consumable_Expenses	2216 non-null	int64
19	Favorite_Product	2216 non-null	object
20	2nd_Favorite_Product	2216 non-null	object

dtypes: float64(2), int64(15), object(4)

1. Features **Z_CostContact** dan **Z_Revenue** tidak dijadikan sebagai features nantinya melainkan dijadikan acuan penghitungan karena std = 0
2. Features yang perlu dieliminasi

ID, Dt_Customer, Year_Birth, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, NumCatalogPurchases, NumWebPurchases, NumStorePurchases, NumDealsPurchases, MntWines, MntFishProducts, MntMeatProducts, MntFruits, MntSweetProducts, dan MntGoldProds.

Features Encoding



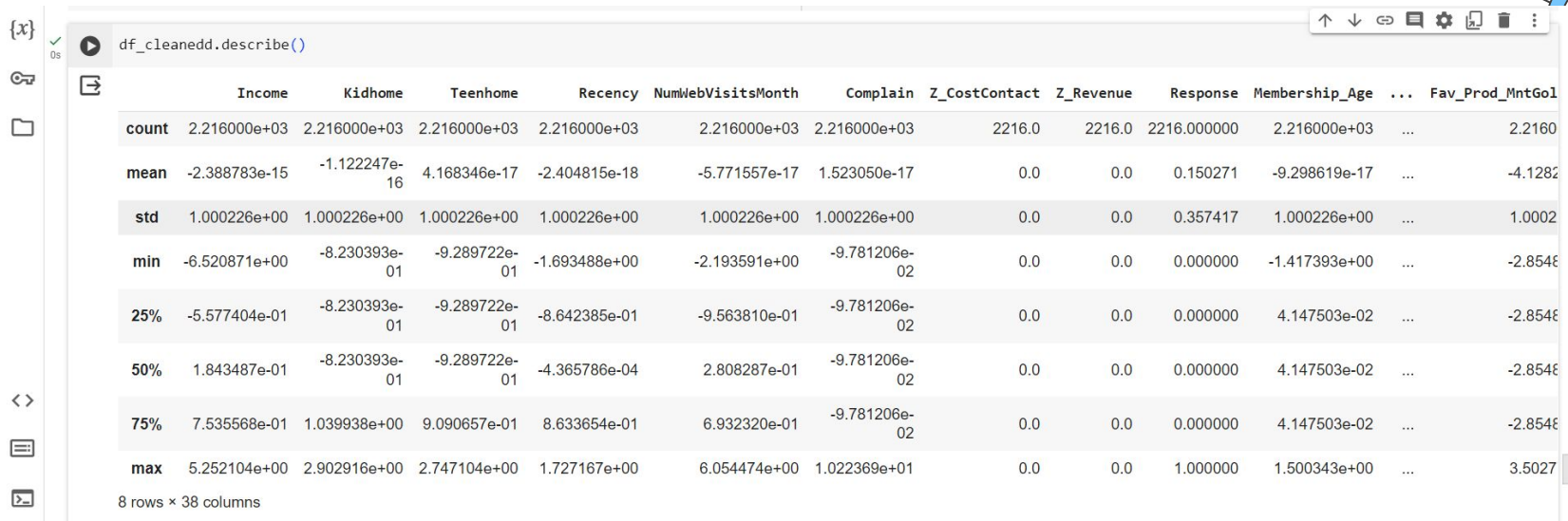
```

13 Frequency_Offline_Purchase 2216 non-null int64
14 Frequency_Promo_Purchase 2216 non-null int64
15 Consumable_Expenses 2216 non-null int64
16 Non_Consumable_Expenses 2216 non-null int64
17 Edu_Basic 2216 non-null uint8
18 Edu_Graduation 2216 non-null uint8
19 Edu_Master 2216 non-null uint8
20 Edu_PhD 2216 non-null uint8
21 MS_Divorced 2216 non-null uint8
22 MS_Married 2216 non-null uint8
23 MS_Single 2216 non-null uint8
24 MS_Together 2216 non-null uint8
25 MS_Widow 2216 non-null uint8
26 Fav_Prod_MntFishProducts 2216 non-null uint8
27 Fav_Prod_MntFruits 2216 non-null uint8
28 Fav_Prod_MntGoldProds 2216 non-null uint8
29 Fav_Prod_MntMeatProducts 2216 non-null uint8
30 Fav_Prod_MntSweetProducts 2216 non-null uint8
31 Fav_Prod_MntWines 2216 non-null uint8
32 ndFav_Prod_MntFishProducts 2216 non-null uint8
33 ndFav_Prod_MntFruits 2216 non-null uint8
34 ndFav_Prod_MntGoldProds 2216 non-null uint8
35 ndFav_Prod_MntMeatProducts 2216 non-null uint8
36 ndFav_Prod_MntSweetProducts 2216 non-null uint8
37 ndFav_Prod_MntWines 2216 non-null uint8
dtypes: float64(2), int64(15), uint8(21)
memory usage: 257.1 KB

```

1. Column kategori (Marital_Status, Education, Favorite_Product, 2nd_Favorite_Product) dilakukan pengecekan korelasi terhadap target(Response) untuk mengetahui apakah kategori tersebut perlu dilakukan features encoding atau tidak.
2. Algoritma pengecekan korelasi yang digunakan adalah **chi2_contingency**, berdasarkan hasil pengecekan ternyata **data kategori memiliki korelasi yang kuat terhadap target (Response) sehingga perlu dilakukan features Encoding.**
3. Features encoding dilakukan supaya pemodelan machine learning nantinya dapat belajar data yang lebih banyak, **features encoding yang dilakukan menggunakan metode One Hot Encoding.**

Standardization



```
df_cleanedd.describe()
```

	Income	Kidhome	Teenhome	Recency	NumWebVisitsMonth	Complain	Z_CostContact	Z_Revenue	Response	Membership_Age	...	Fav_Prod_MntGol
count	2.216000e+03	2.216000e+03	2.216000e+03	2.216000e+03	2.216000e+03	2.216000e+03	2216.0	2216.0	2216.000000	2.216000e+03	...	2.2160
mean	-2.388783e-15	-1.122247e-16	4.168346e-17	-2.404815e-18	-5.771557e-17	1.523050e-17	0.0	0.0	0.150271	-9.298619e-17	...	-4.1282
std	1.000226e+00	1.000226e+00	1.000226e+00	1.000226e+00	1.000226e+00	1.000226e+00	0.0	0.0	0.357417	1.000226e+00	...	1.0002
min	-6.520871e+00	-8.230393e-01	-9.289722e-01	-1.693488e+00	-2.193591e+00	-9.781206e-02	0.0	0.0	0.000000	-1.417393e+00	...	-2.8548
25%	-5.577404e-01	-8.230393e-01	-9.289722e-01	-8.642385e-01	-9.563810e-01	-9.781206e-02	0.0	0.0	0.000000	4.147503e-02	...	-2.8548
50%	1.843487e-01	-8.230393e-01	-9.289722e-01	-4.365786e-04	2.808287e-01	-9.781206e-02	0.0	0.0	0.000000	4.147503e-02	...	-2.8548
75%	7.535568e-01	1.039938e+00	9.090657e-01	8.633654e-01	6.932320e-01	-9.781206e-02	0.0	0.0	0.000000	4.147503e-02	...	-2.8548
max	5.252104e+00	2.902916e+00	2.747104e+00	1.727167e+00	6.054474e+00	1.022369e+01	0.0	0.0	1.000000	1.500343e+00	...	3.5027

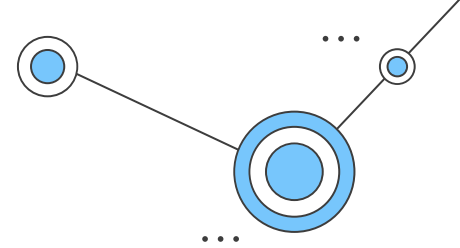
8 rows x 38 columns

Summary

1. **Standardization dilakukan dengan menggunakan StandardScaler, setelah dilakukan standardization didapatkan hasil bahwa sekarang std data numeric adalah 1.000226e+00.**

04

Modeling



Strateginya membagi data menjadi dua bagian: Training Data (70%) dan Data Testing (30%). Dari Training Data, kita akan membaginya lagi menjadi dua bagian: Training Data dan Validasi Data. Data Validasi akan digunakan untuk menilai kemampuan prediksi model.

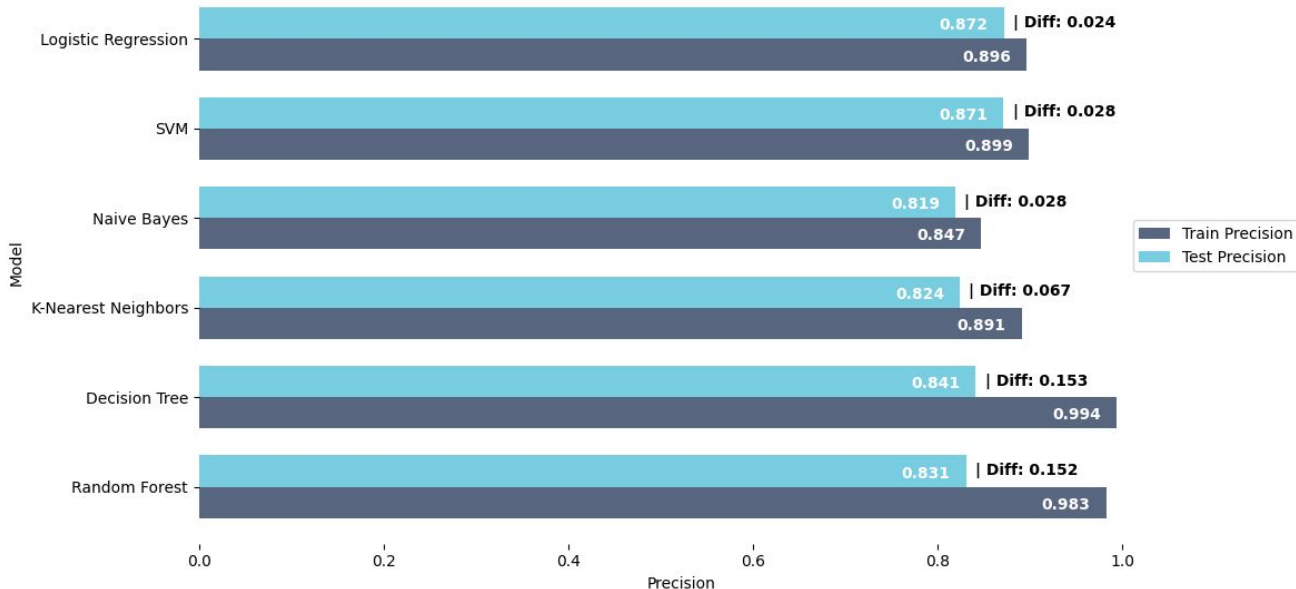
	0	1	
0	TN	FP	Actual
1	FN	TP	
	Prediction		

Precision

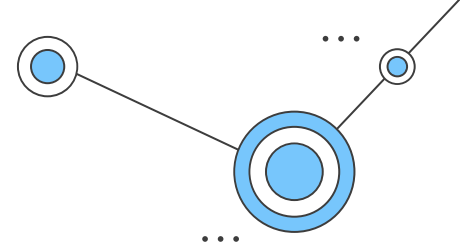
$$\frac{TP}{TP + FP}$$

TP : Pelanggan yang diberikan marketing, kemudian merespon marketing tersebut
FP : Pelanggan yang tidak merespon marketing, namun tetap diberi marketing

Train and Test Precision Scores for Different Models



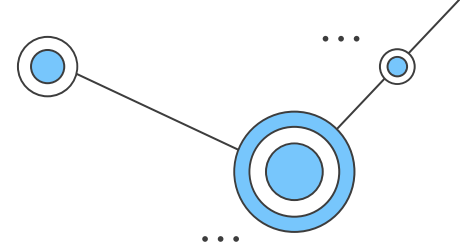
Kami akan berusaha untuk mengoptimalkan nilai Precision. Sejalan dengan tujuan utama kami yaitu untuk meminimalkan *Profit loss* dari kampanye yang dilakukan. Jadi, kami akan memilih model dengan presisi yang tinggi dan memiliki perbedaan presisi yang kecil antara *test* dan *train*. (Perbedaan presisi yang besar mengidentifikasikan bahwa model tidak sesuai, sehingga perlu dilakukan tuning hyperparameter)



Hyperparameter Tuning



	Model	Train Precision Before Tuning	Test Precision Before Tuning	Precision Difference Before Tuning	Train Precision After Tuning	Test Precision After Tuning	Precision Difference After Tuning
0	Logistic Regression	0.896011	0.871669	0.024	0.894998	0.870256	0.025
1	SVM	0.898520	0.870971	0.028	0.896673	0.870971	0.026

Kami akan memilih model dengan presisi yang tinggi dan memiliki perbedaan presisi yang kecil antara test dan train. Setelah dilakukan tuning hyperparameter, kami memilih model *Logistic Regression* sebelum dilakukan tuning hyperparameter, karena model ini memiliki perbedaan presisi yang kecil antara test dan train serta memiliki Test Precision yang paling tinggi.

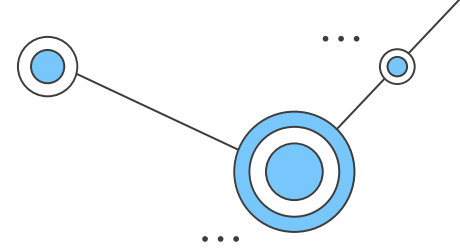


Cross Validation

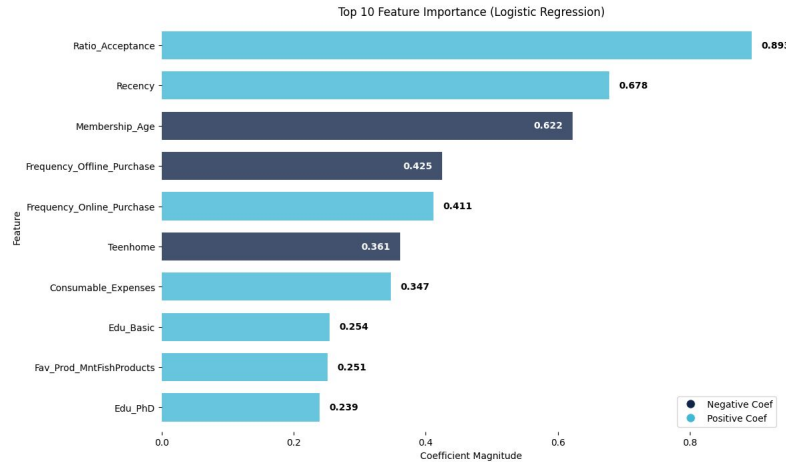
Number of samples in each class for cross-validation: 5

	Model	Precision	
0	Logistic Regression	0.837517	
1	Logistic Regression	0.840189	
2	Logistic Regression	0.856706	
3	Logistic Regression	0.852565	
4	Logistic Regression	0.878781	

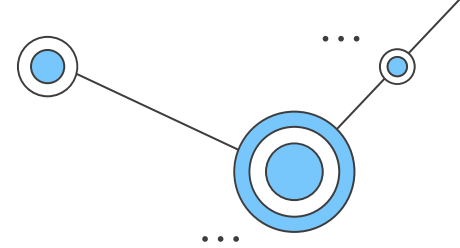
Kami melakukan cross-validation pada model Regresi Logistik untuk memberikan estimasi yang lebih stabil terhadap kinerja model dengan membagi dataset menjadi beberapa subset yang terpisah. Dari hasil cross-validation, kami dapat mengamati bahwa model Regresi Logistik menunjukkan variasi presisi yang relatif stabil dengan nilai yang berkisar antara 0.837 hingga 0.879 pada setiap iterasi validasi silang. Dengan presisi data uji sebesar 0,872, kita dapat menyimpulkan bahwa model secara konsisten berkinerja baik



Feature Importance



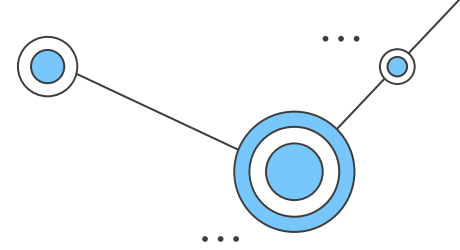
Berdasarkan Feature Importance Scores yang diberikan, dapat dilihat bahwa fitur-fitur seperti Ratio_Acceptance (0.89), Recency (0.68), dan Membership_Age (0.62) merupakan fitur-fitur yang paling penting dan dapat dijadikan fokus untuk mendapatkan lebih banyak insight. Fitur-fitur ini dapat menjadi dasar untuk pemilihan fitur pada iterasi selanjutnya. Fitur Frequency_Offline_Purchase, Frequency_Online_Purchase, Teenhome dan fitur lainya memiliki nilai Feature Importance yang lebih rendah.



Precision

87.2%

We have successfully maximized the **Precision score to 87,2%**. However, what does this mean for the business?



Without Models

14.91%
rate

With Models

87.2 %
rate

Diff

72.9 %

Tujuan dari penggunaan **model machine learning** pada dataset adalah untuk membantu menemukan pelanggan yang akan dijadikan target kampanye pemasaran, hal ini akan memberikan perbedaan pada proses pemilihan target pemasaran dimana sebelum menggunakan model pemasaran dilakukan pada seluruh pelanggan atau, sedangkan setelah menggunakan model akan dipilih pelanggan yang berpotensi untuk diberikan kampanye pemasaran.

Nilai Precision dari model tersebut adalah 87.2%, yang berarti dari seluruh pelanggan yang terpilih (dengan model) yang memberikan respon adalah sebesar 87.2%.

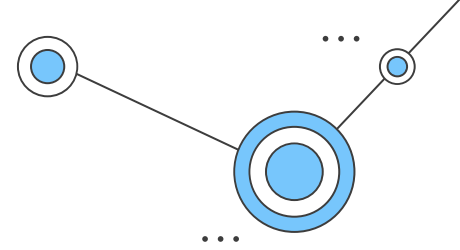
Sedangkan jika kampanye pemasaran dilakukan kepada seluruh pelanggan tanpa seleksi, maka pelanggan yang memberikan respon adalah sebesar 14.91%.

yang berarti bahwa model berhasil meningkatkan Response Rate

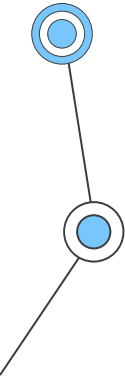


05

Recommendation



- Mengutamakan pemberian campaign kepada pelanggan yang pernah merespon campaign minimal 1x supaya potensi penerimaan terhadap campaign yang akan diberikan lebih tinggi.
- Mengutamakan pemberian campaign kepada pelanggan yang bertransaksi selama < 2 bulan terakhir supaya potensi penerimaan terhadap campaign yang akan diberikan lebih tinggi.
- Mengutamakan pemberian campaign kepada pelanggan yang berlangganan < 1 tahun supaya potensi penerimaan terhadap campaign yang akan diberikan lebih tinggi.





Thank You!

