

# Laporan Final Project

## Stage 0

---

### Kelompok 5 - D Avengers

- **Dinda Galuh Guminta**
- **Khaerun Nisa**
- **Teguh Ismareza**
- **Iqbal Fauzan Saputra**
- **Faldi Ramadhan**
- **Kadek Haris Dana Swara**
- **Ahya Ramdhanitasari**
- **Julius Pardamean**



(dipresentasikan setiap sesi mentoring)

# STAGE 0

## Preparation

---

- Problem Statement
- Role
- Goal
- Objective
- Business Metrics

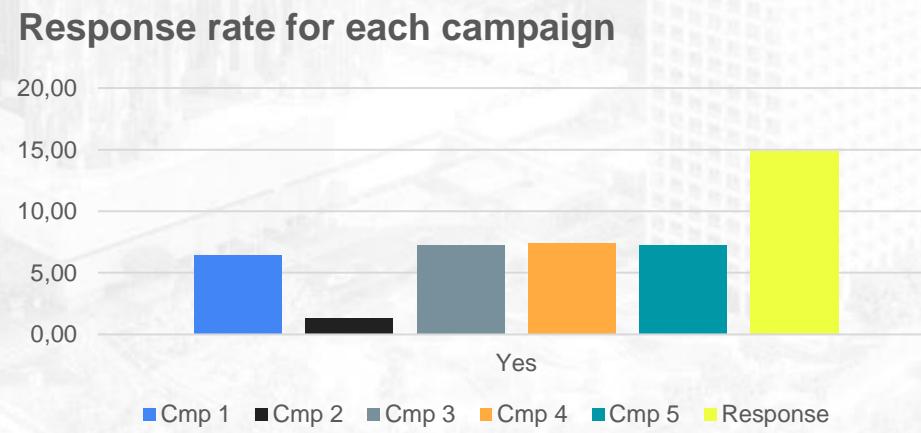
# Latar Belakang Masalah

## Problem statement:

Seiring meningkatnya minat masyarakat di negara A untuk melakukan bisnis, membuat persaingan pasar menjadi ketat pula. Perubahan pola customer yang tidak menentu menyebabkan banyak pelaku bisnis aktif terlibat dalam upaya pemasaran mereka untuk mempertahankan pelanggan atau mendapatkan pelanggan baru. Oleh karena itu diperlukan cara untuk membantu team marketing dalam meningkatkan efisiensi marketing campaign melalui peningkatan response atau pengurangan biaya campaign.



Costs	\$ 6720
Revenue	\$ 3674 ↑ 51,2 %



Meskipun terjadi peningkatan response rate tetapi diperlukan response rate min 25% yaitu minimal 1 dari 4 orang (berdasarkan cost \$3 dan revenue \$11 untuk campaign pada 1 orang)

# Latar Belakang Masalah

## Role:

Sebagai tim business analyst yang diberi tahu oleh Marketing Manager bahwa marketing campaign yang diterapkan tidak seefektif yang diharapkan. Kami bertanggung jawab untuk membantu team marketing untuk menganalisis data, memahami masalah dan memberikan solusi yang berbasis data.

## Objective:

- Prediksi kemungkinan customer memberikan respons positif
- Membagi klaster customer (segmentation)
- Mendapatkan faktor-faktor yang berkontribusi terhadap respon
- Memanfaatkan insight yang didapat untuk meningkatkan conversion rate pada campaign selanjutnya dengan biaya yang rendah

## Goal:

Efisiensi biaya campaign dengan menargetkan customer yang sesuai agar response dan keuntungan yang meningkat

## Business Metrics:

- Response rate (persentase customer yang memberikan respon positif)
- RFM metrics

# Laporan Final Project

## Stage 1

---

### Kelompok 5 - D Avengers

- **Dinda Galuh Guminta**
- **Khaerun Nisa**
- **Teguh Ismareza**
- **Iqbal Fauzan Saputra**
- **Faldi Ramadhan**
- **Kadek Haris Dana Swara**
- **Ahya Ramdhanitasari**
- **Julius Pardamean**



(dipresentasikan setiap sesi mentoring)

# STAGE 1

## EDA, Insight, & Visualization

---

- Descriptive Statistics
- Univariate Analysis
- Multivariate Analysis
- Business Insight

# Dataset

## Marketing Campaign

Variable	Description	Type
ID	customer's ID	Numerical
Year_Birth	Customer's year of birth	Numerical
DtCustomer	date of customer's enrolment with the company	Categorical
Education	customer's level of education	Categorical
Marital	customer's marital status	Categorical
Kidhome	number of small children in customer's household	Numerical
Teenhome	number of teenagers in customer's household	Numerical
Income	customer's yearly household income	Numerical
MntFishProducts	amount spent on fish products in the last 2 years	Numerical
MntMeatProducts	amount spent on meat products in the last 2 years	Numerical
MntFruits	amount spent on fruits products in the last 2 years	Numerical
MntSweetProducts	amount spent on sweet products in the last 2 years	Numerical
MntWines	amount spent on wine products in the last 2 years	Numerical
MntGoldProds	amount spent on gold products in the last 2 years	Numerical
NumDealsPurchases	number of purchases made with discount	Numerical
NumCatalogPurchases	number of purchases made using catalogue	Numerical
NumStorePurchases	number of purchases made directly in stores	Numerical
NumWebPurchases	number of purchases made through company's web site	Numerical
NumWebVisitsMonth	number of visits to company's web site in the last month	Numerical
Recency	number of days since the last purchase	Numerical
Z_CostContact	cost to contact the costumer	Numerical
Z_Revenue	revenue generated by the costumer	Numerical
AcceptedCmp1	1 if customer accepted the offer in the 1st campaign, 0 otherwise	Categorical
AcceptedCmp2	1 if customer accepted the offer in the 2nd campaign, 0 otherwise	Categorical
AcceptedCmp3	1 if customer accepted the offer in the 3rd campaign, 0 otherwise	Categorical
AcceptedCmp4	1 if customer accepted the offer in the 4th campaign, 0 otherwise	Categorical
AcceptedCmp5	1 if customer accepted the offer in the 5th campaign, 0 otherwise	Categorical
Complain	1 if customer complained in the last 2 years	Categorical
Response (target)	1 if customer accepted the offer in the last campaign, 0 otherwise	Categorical

```

<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)

```

# Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
ID	2240.0	5592.159821	3246.662198	0.0	2828.25	5458.5	8427.75	11191.0
Year_Birth	2240.0	1968.805804	11.984069	1893.0	1959.00	1970.0	1977.00	1996.0
Income	2240.0	52237.975446	25037.955891	1730.0	35538.75	51381.5	68289.75	666666.0
Kidhome	2240.0	0.444196	0.538398	0.0	0.00	0.0	1.00	2.0
Teenhome	2240.0	0.506250	0.544538	0.0	0.00	0.0	1.00	2.0
Recency	2240.0	49.109375	28.962453	0.0	24.00	49.0	74.00	99.0
MntWines	2240.0	303.935714	336.597393	0.0	23.75	173.5	504.25	1493.0
MntFruits	2240.0	26.302232	39.773434	0.0	1.00	8.0	33.00	199.0
MntMeatProducts	2240.0	166.950000	225.715373	0.0	16.00	67.0	232.00	1725.0
MntFishProducts	2240.0	37.525446	54.628979	0.0	3.00	12.0	50.00	259.0
MntSweetProducts	2240.0	27.062946	41.280498	0.0	1.00	8.0	33.00	263.0
MntGoldProd	2240.0	44.021875	52.167439	0.0	9.00	24.0	56.00	362.0
NumDealsPurchases	2240.0	2.325000	1.932238	0.0	1.00	2.0	3.00	15.0
NumWebPurchases	2240.0	4.084821	2.778714	0.0	2.00	4.0	6.00	27.0
NumCatalogPurchases	2240.0	2.662054	2.923101	0.0	0.00	2.0	4.00	28.0
NumStorePurchases	2240.0	5.790179	3.250958	0.0	3.00	5.0	8.00	13.0
NumWebVisitsMonth	2240.0	5.316518	2.426645	0.0	3.00	6.0	7.00	20.0
AcceptedCmp3	2240.0	0.072768	0.259813	0.0	0.00	0.0	0.00	1.0
AcceptedCmp4	2240.0	0.074554	0.262728	0.0	0.00	0.0	0.00	1.0
AcceptedCmp5	2240.0	0.072768	0.259813	0.0	0.00	0.0	0.00	1.0
AcceptedCmp1	2240.0	0.064286	0.245316	0.0	0.00	0.0	0.00	1.0
AcceptedCmp2	2240.0	0.013393	0.114976	0.0	0.00	0.0	0.00	1.0
Complain	2240.0	0.009375	0.096391	0.0	0.00	0.0	0.00	1.0
Z_CostContact	2240.0	3.000000	0.000000	3.0	3.00	3.0	3.00	3.0
Z_Revenue	2240.0	11.000000	0.000000	11.0	11.00	11.0	11.00	11.0
Response	2240.0	0.149107	0.356274	0.0	0.00	0.0	0.00	1.0
AcceptedCmptot	2240.0	0.297768	0.678381	0.0	0.00	0.0	0.00	4.0
Age	2240.0	45.194196	11.984069	18.0	37.00	44.0	55.00	121.0
Customer_Months	2240.0	17.194643	6.622911	6.0	11.75	17.0	23.00	29.0

- Terdapat missing value pada variabel income
- Semua tipe data sudah sesuai kecuali pada variabel Dt\_Customer sehingga diconvert menjadi date type kemudian dibuat variabel baru untuk mendapatkan lama customer join ke company
- Variabel income memiliki mean dan median yang hampir sama, namun nilai max yang sangat tinggi menunjukkan adanya indikasi outlier
- Tidak ada nilai statistika yang aneh pada variabel kidhome, teenhome, dan recency
- Pada variabel amount spent on products terlihat bahwa SEMUA PRODUK memiliki mean yang lebih besar dari median yang menandakan pola distribusi positively skewed
- Sementara itu, tidak terjadi keanehan nilai statistik pada number of purchase
- Meskipun variabel age tidak menunjukkan statistik yang aneh, namun terlihat bahwa usia customer tertua yaitu 121 tahun dimana nilai ini sangat jauh dari rata-rata
- Variabel customer\_months memiliki pola distribusi yang simetris, dapat dilihat dari nilai mean median yang hampir sama
- Kolom Z\_CostContact dan Z\_Revenue tidak dimasukkan dalam pemodelan karena mengandung nilai yang sama
- Terdapat missing value pada variabel income sebanyak 1.07 %, dilakukan imputasi menggunakan median

# Descriptive Statistics

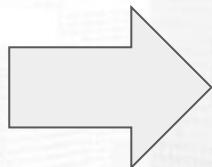
Berikut merupakan kategori pada variabel Marital\_Status dan Education

Married	864
Together	580
Single	480
Divorced	232
Widow	77
Alone	3
Absurd	2
YOLO	2

Name: Marital\_Status, dtype: int64

Graduation	1127
PhD	486
Master	370
2n Cycle	203
Basic	54

Name: Education, dtype: int64



diubah menjadi  
single

Married	864
Together	580
Single	487
Divorced	232
Widow	77

Name: Marital\_Status, dtype: int64

# Descriptive Statistics

- A. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

Semua tipe data pada tiap kolom sudah sesuai kecuali variabel DtCustomer sehingga diubah menjadi date type

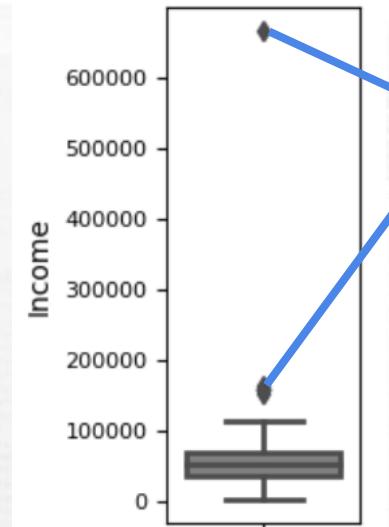
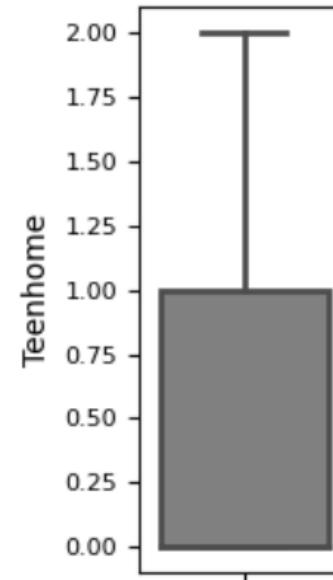
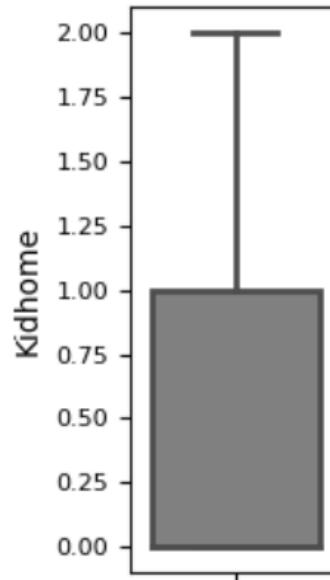
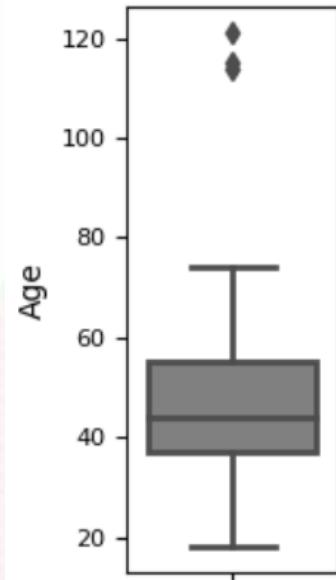
- A. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Hanya kolom income yang memiliki nilai kosong sehingga dilakukan imputasi menggunakan median

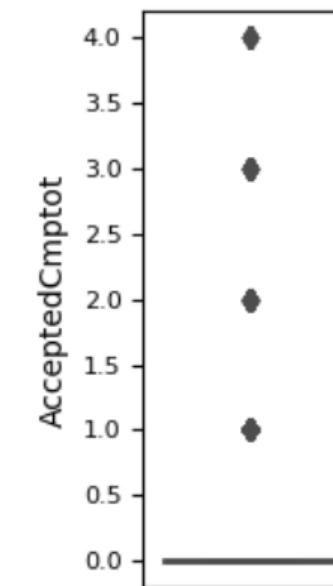
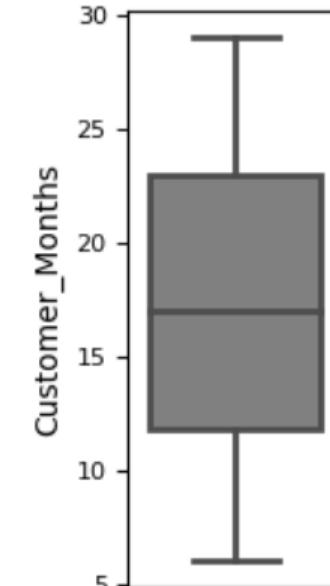
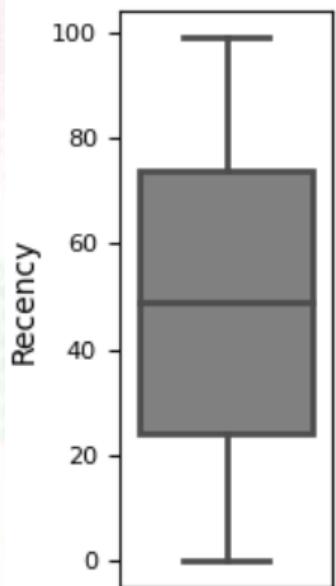
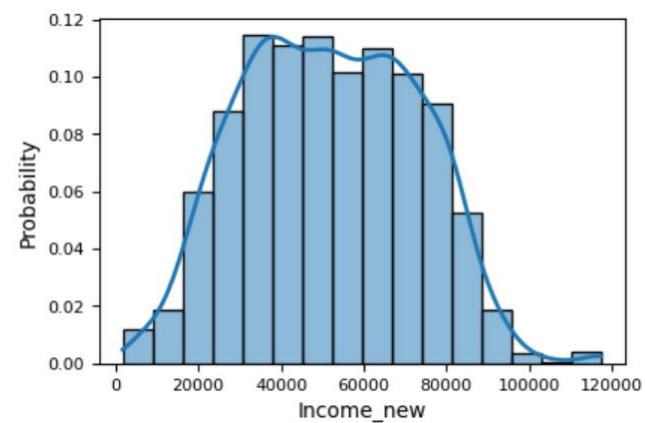
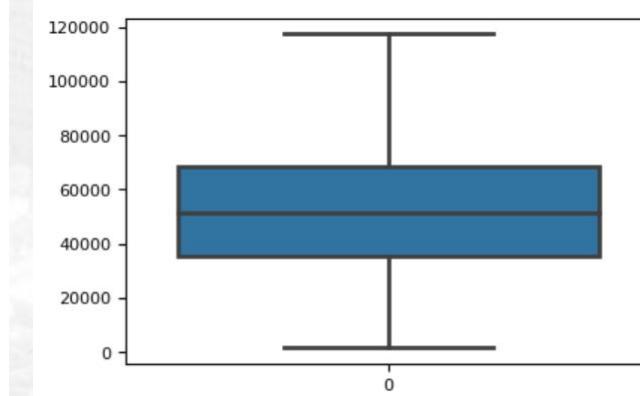
- A. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

- Variabel income dan age memiliki nilai max yang sangat tinggi
- kolom Z\_CostContact dan Z\_Revenue memiliki nilai yang sama pada tiap statistik karena kedua kolom itu merupakan bentuk nilai tetap atau perkiraan revenue jika customer memberikan response
- Selain dari kolom diatas, kolom lainnya tidak memiliki keanehan namun memiliki pola distribusi yang tidak simetris

# Univariate Analysis

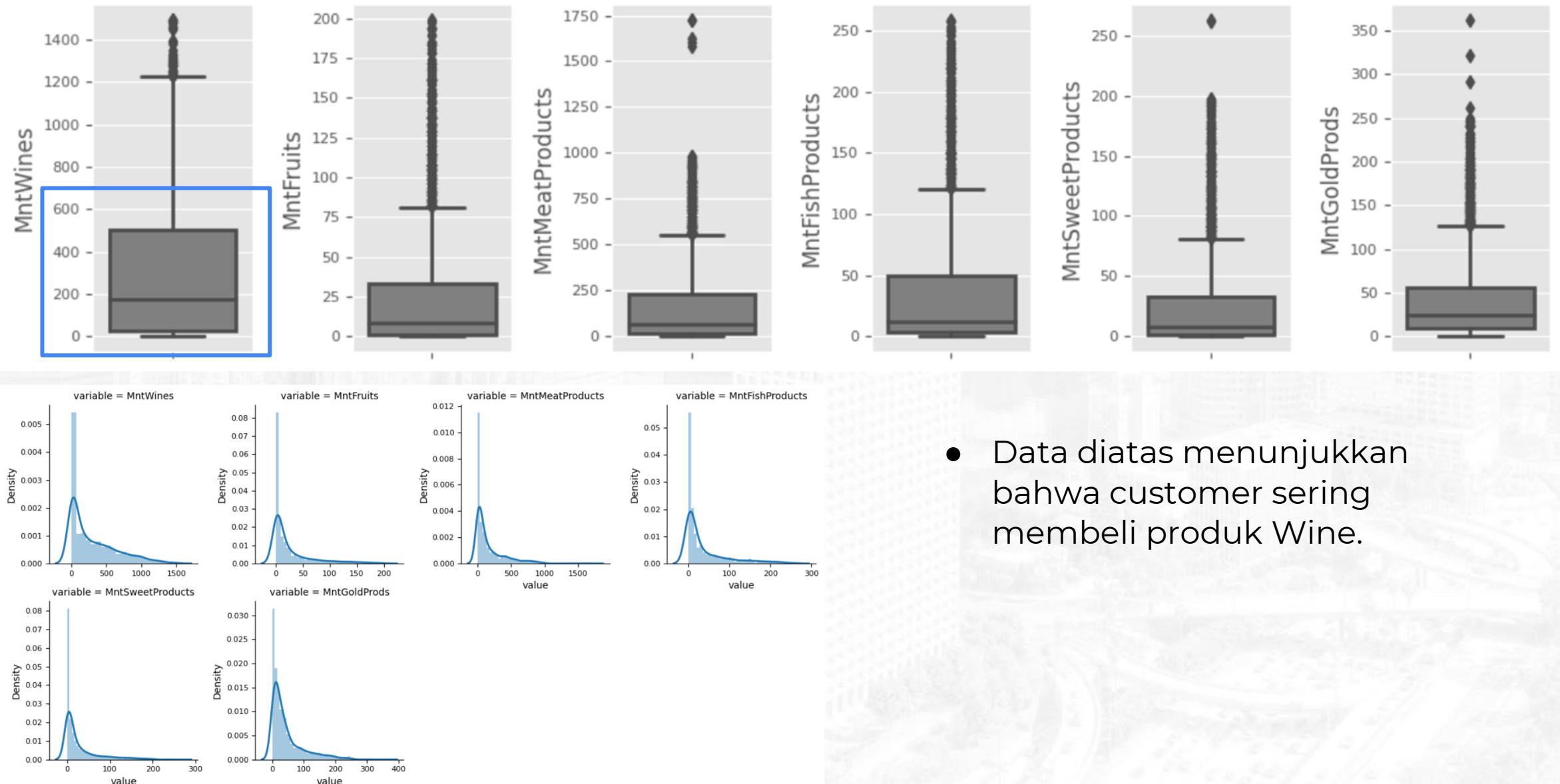


- Karena terdapat outlier yang sangat jauh.  
Diambil batas atas IQR



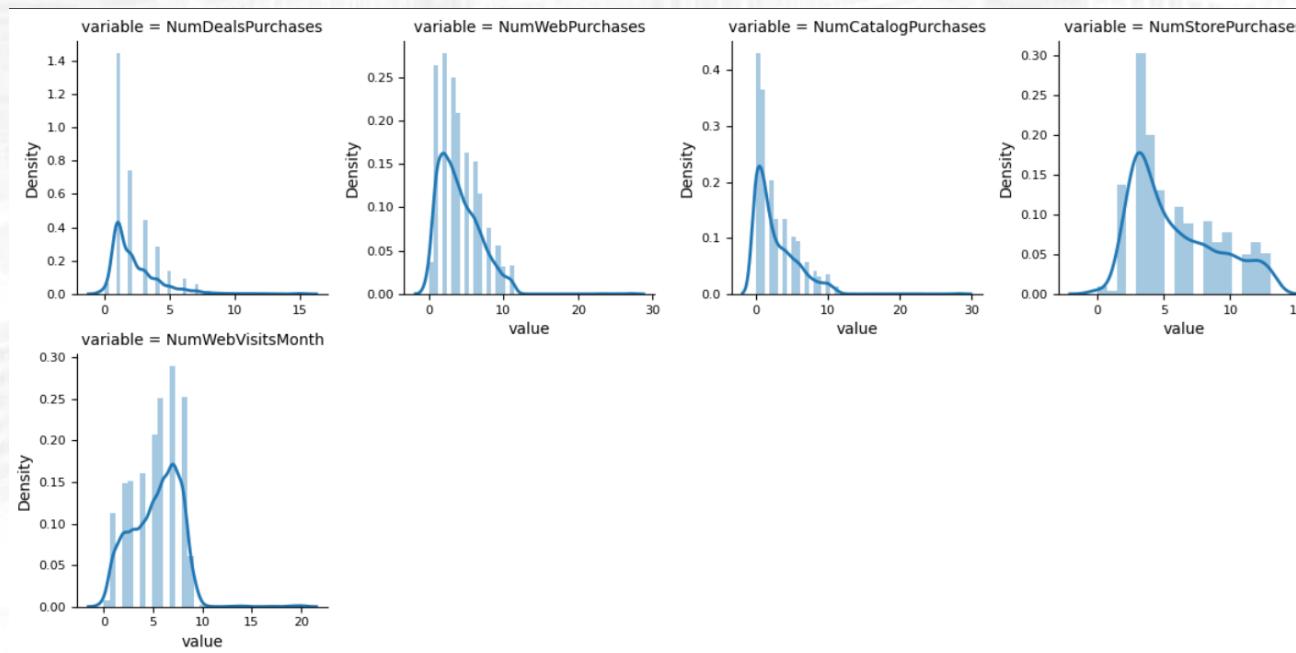
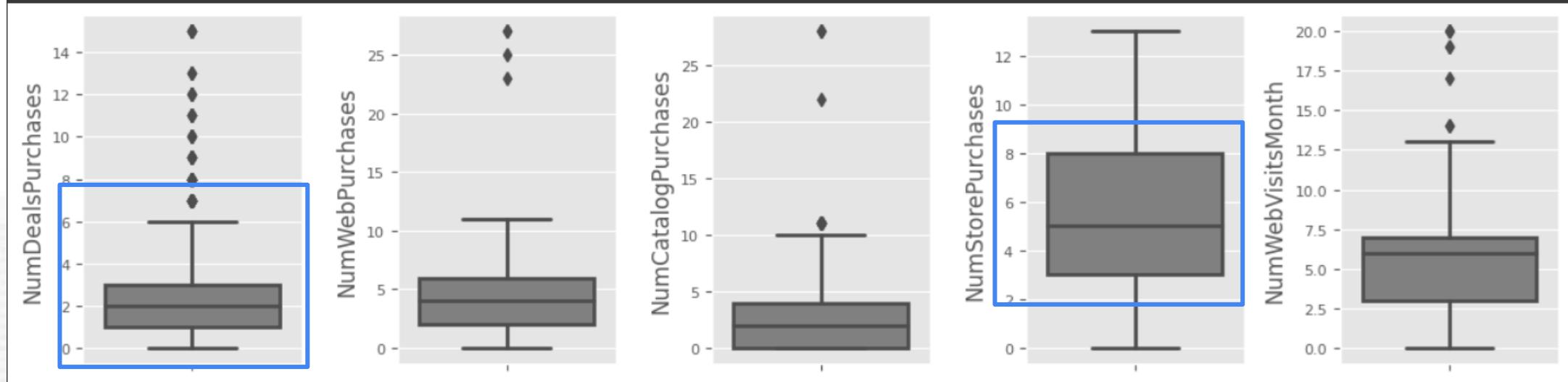
Data menjadi memiliki persebaran normal.

# Univariate Analysis



- Data diatas menunjukkan bahwa customer sering membeli produk Wine.

# Univariate Analysis

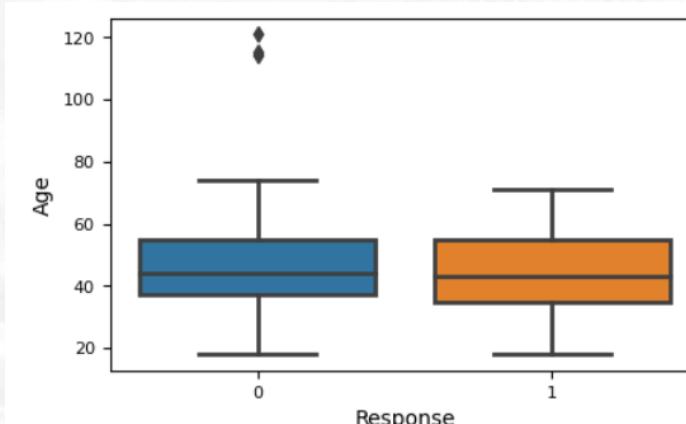
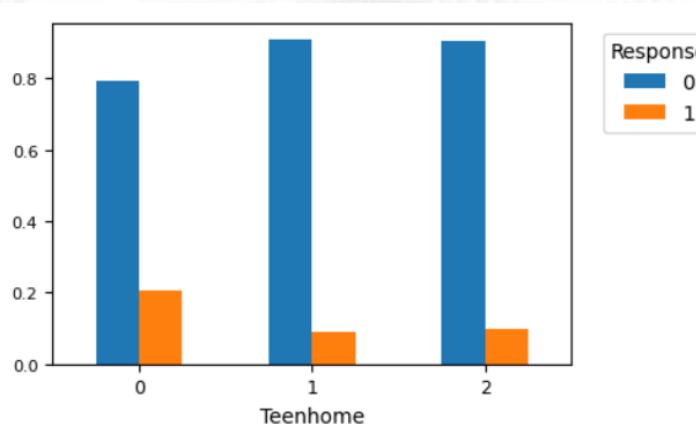
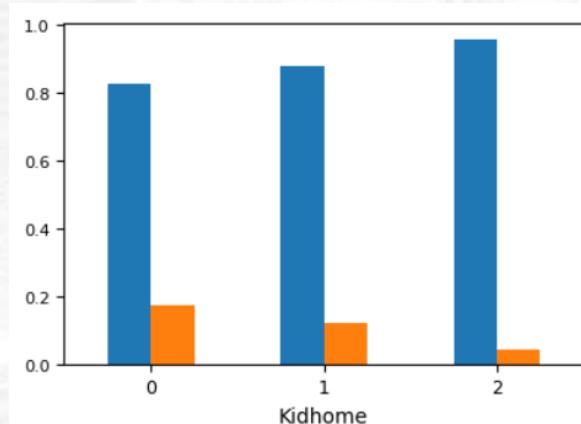
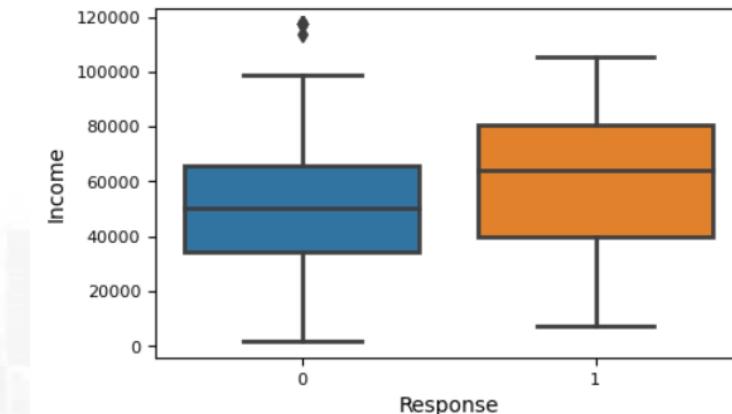
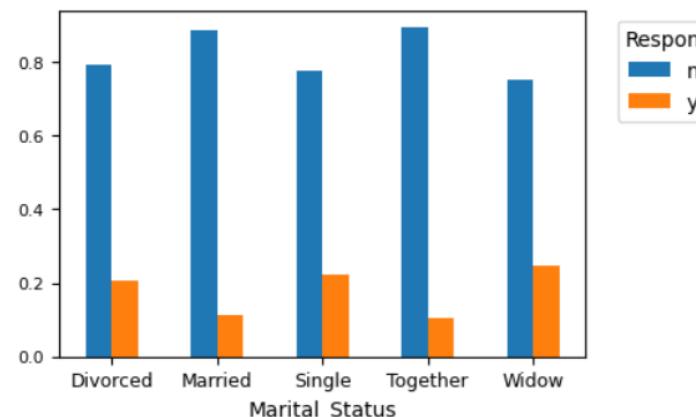
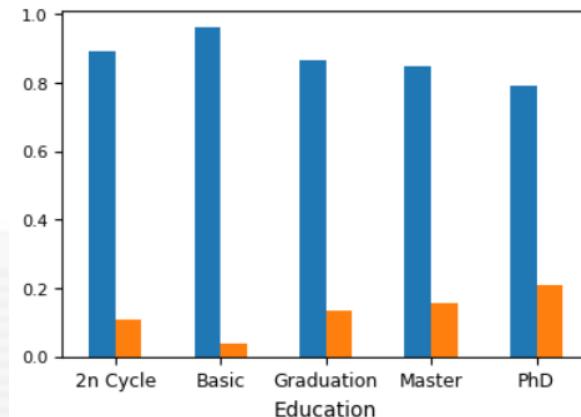


- Data diatas menunjukkan bahwa customer suka berbelanja apabila terdapat diskon
- Customer juga sering membeli produk dari store langsung.

# Univariate Analysis

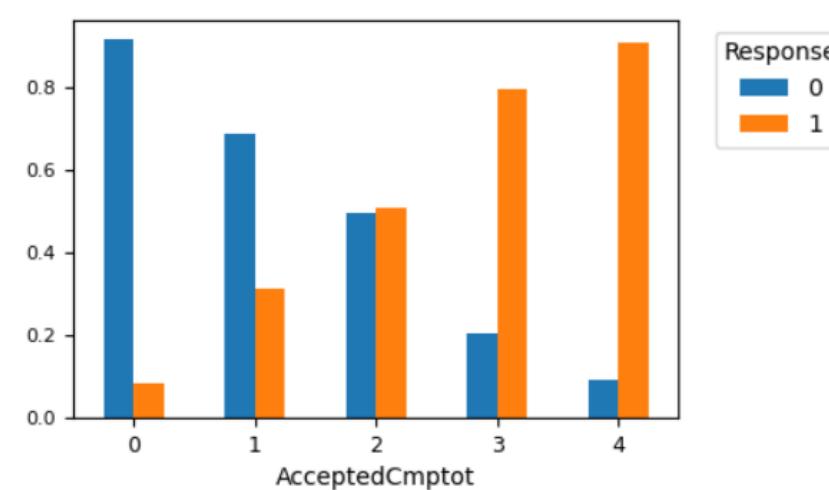
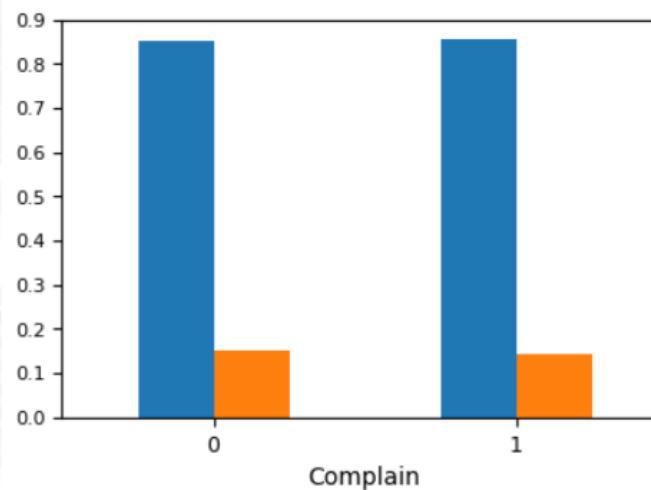
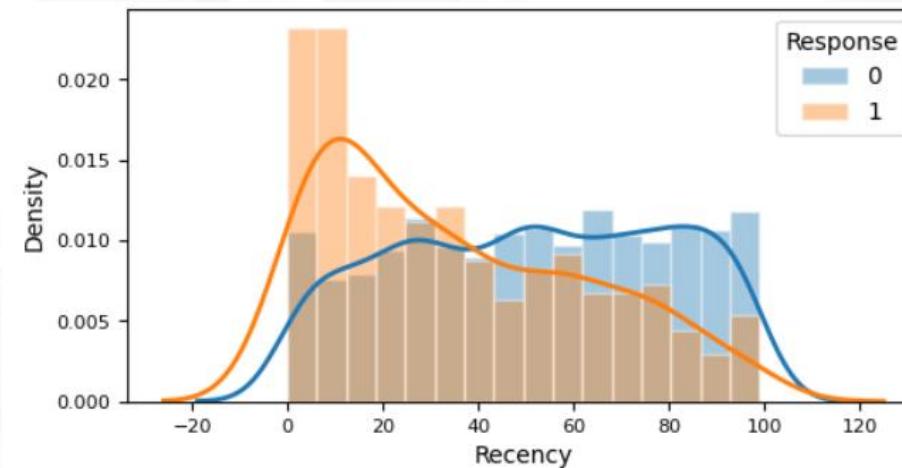
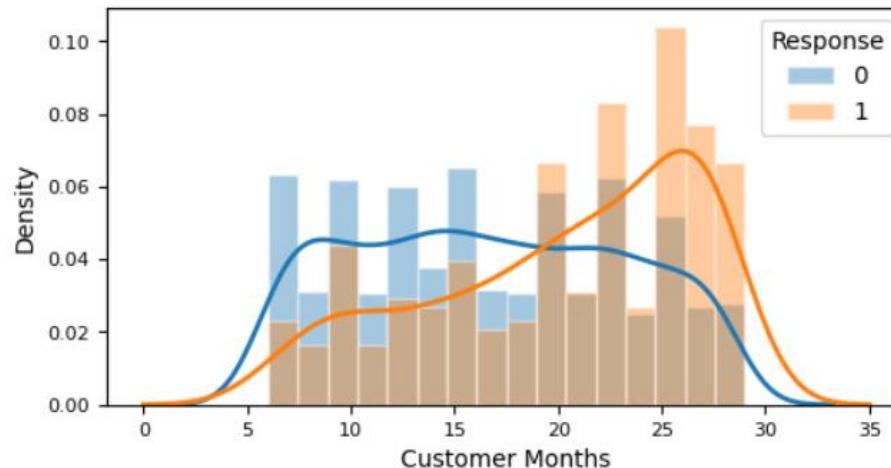
- Variabel recency dan customer\_months memiliki pola distribusi yang simetris karena nilai median berada ditengah dan garis whisker memiliki panjang sama di atas maupun di bawah kotak
- Variabel income dan age terlihat memiliki pola right skew (positively skewed)
- Variabel kidhome, teenhome, dan AcceptedCmptot memiliki bentuk boxplot yang unik karena variabel tersebut memiliki tipe data diskrit
- Amount spent pada semua jenis produk menunjukkan adanya outlier dengan pola sebaran data yang miring ke kanan (right skew)
- Pola distribusi number of purchase yaitu positively skew dimana data cenderung berkumpul di sisi kanan yang menyebabkan ekor histogram lebih panjang di sebelah kanan. Selain itu ditemukan outlier pada semua variabel kecuali NumStorePurchases.
- Handling outlier pada dataset ini hanya dilakukan pada variabel income karena memiliki rentang yang sangat jauh

# Multivariate Analysis



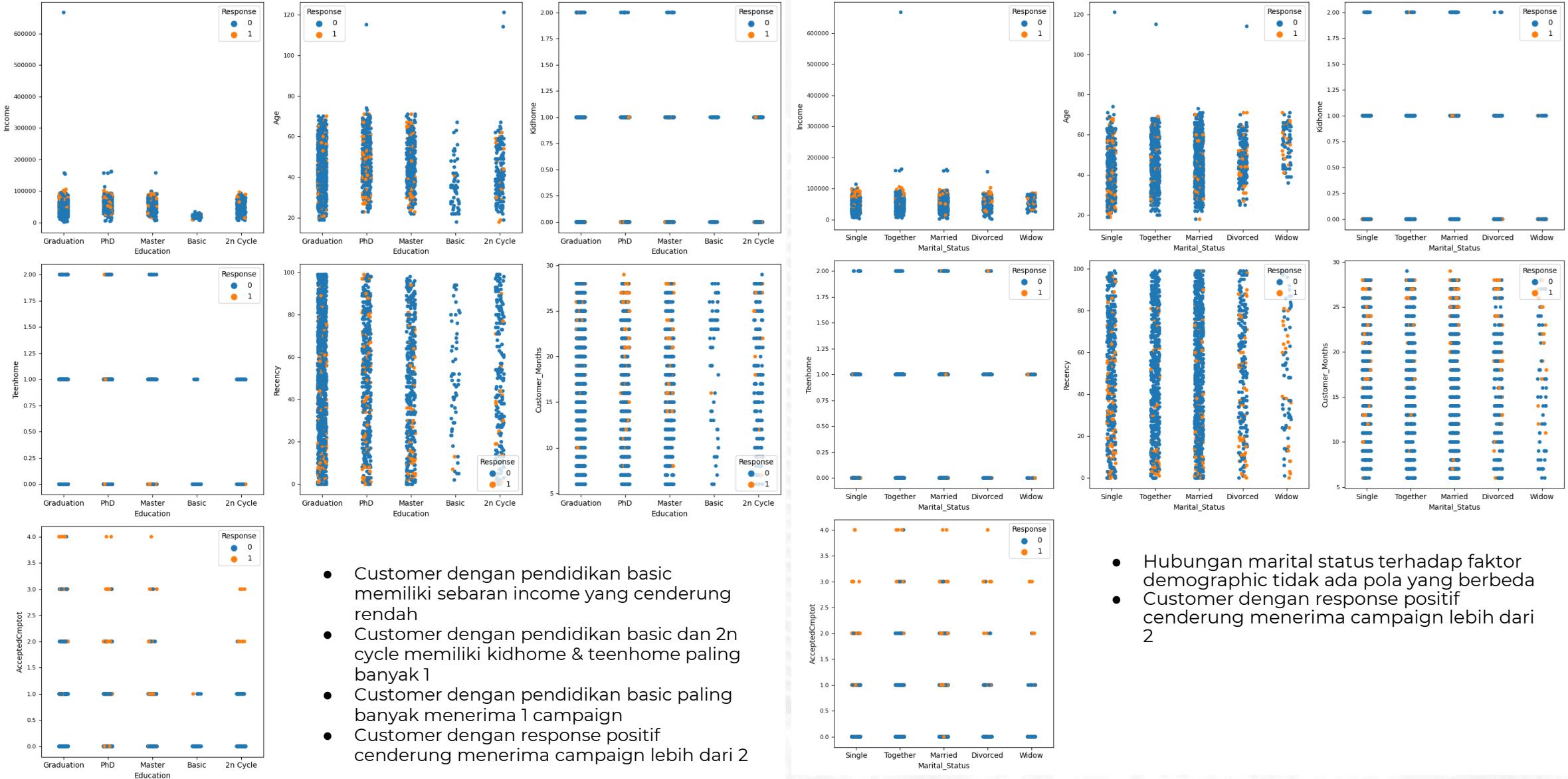
- variabel education, marital status, income, kidhome, dan teenhome berhubungan dengan response
- variabel umur memiliki pola yang sama sehingga diduga tidak memiliki hubungan dengan response

# Multivariate Analysis

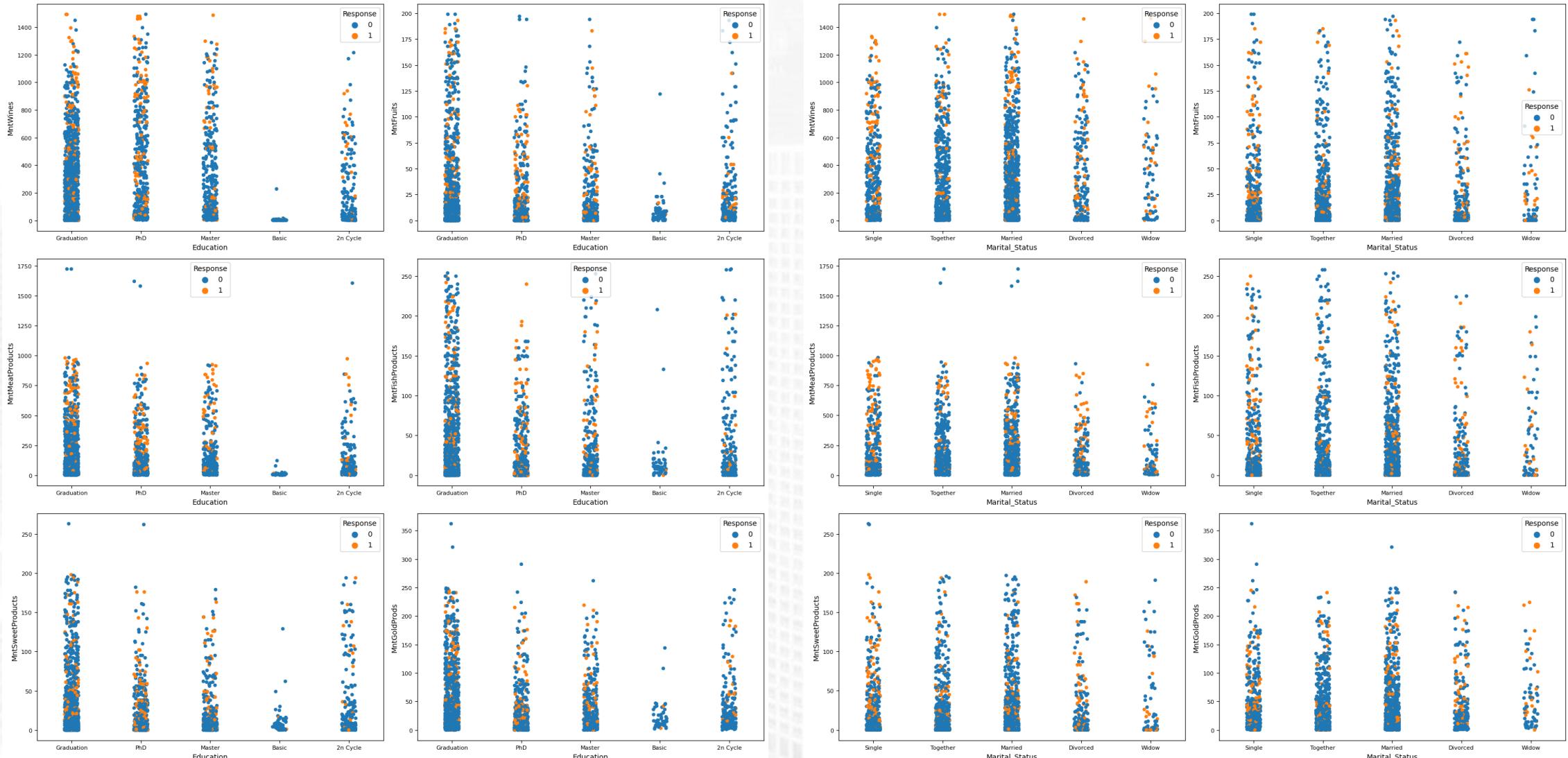


- customer month, recency, total accepted campaign berhubungan dengan response
- complain tidak memiliki hubungan dengan response

# Multivariate Analysis

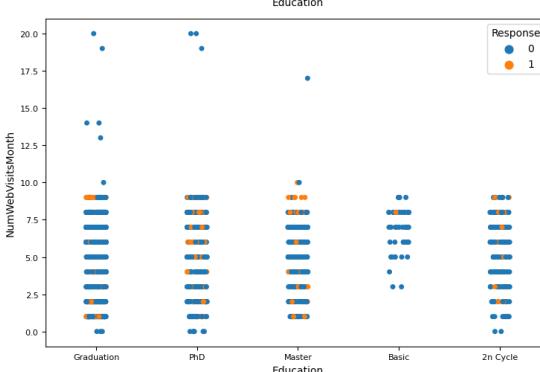
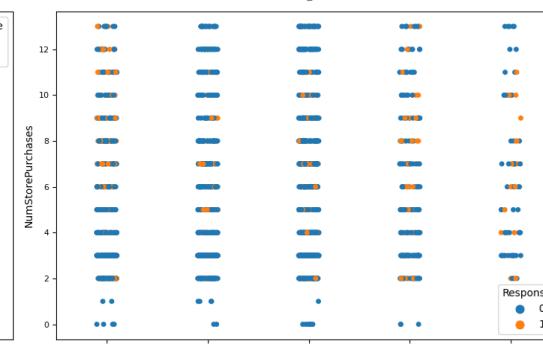
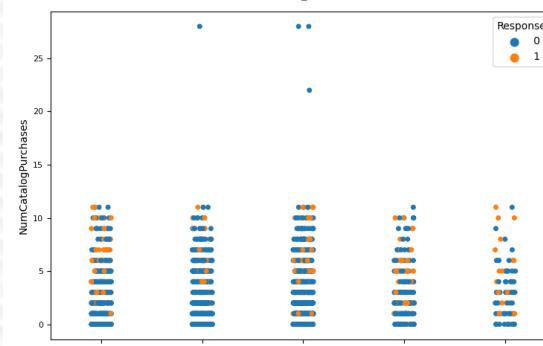
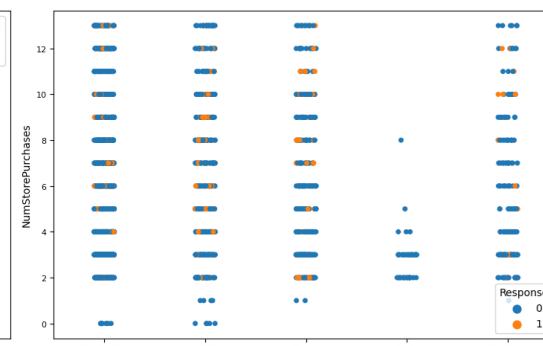
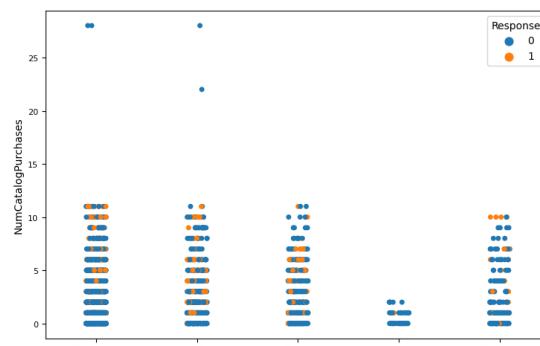
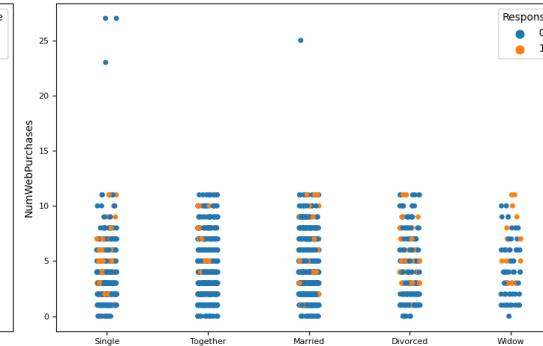
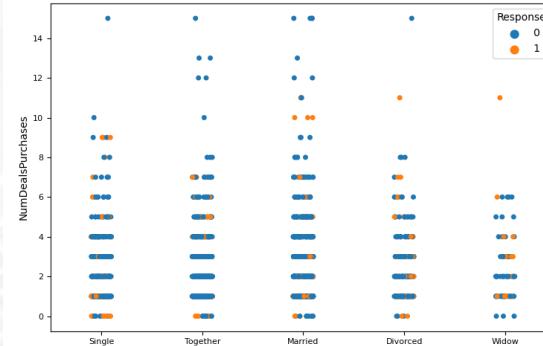
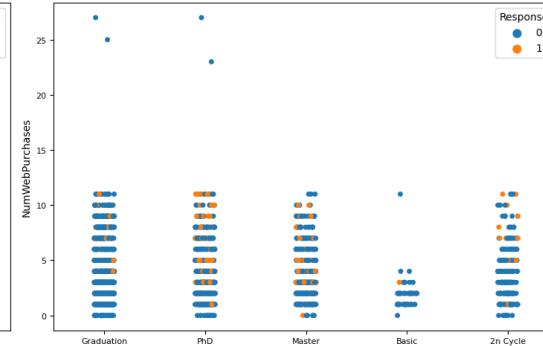
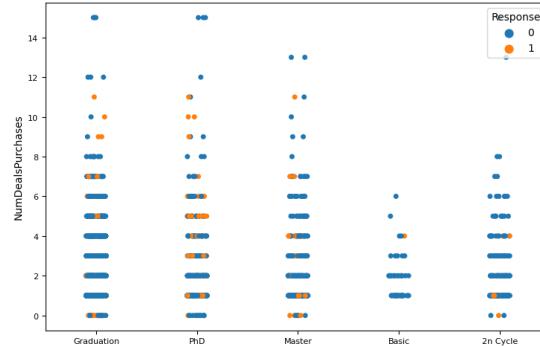


# Multivariate Analysis

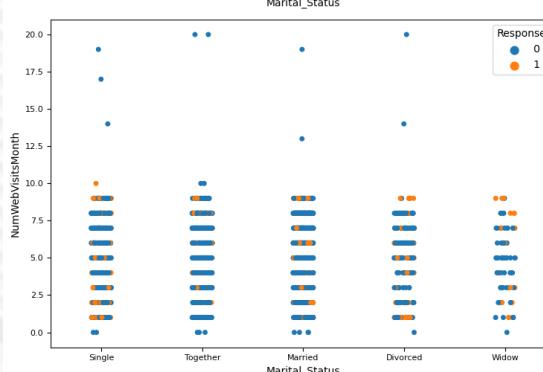


- Customer dengan pendidikan basic memiliki sebaran spent yang cenderung rendah di semua product
- Customer dengan pendidikan PhD dan Master banyak membeli product wine
- Customer dengan pendidikan graduation banyak membeli fruits, fish, sweet product, dan gold
- Tidak terdapat perbedaan pola spent customer berdasarkan status marital

# Multivariate Analysis



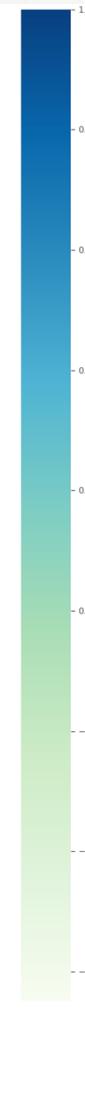
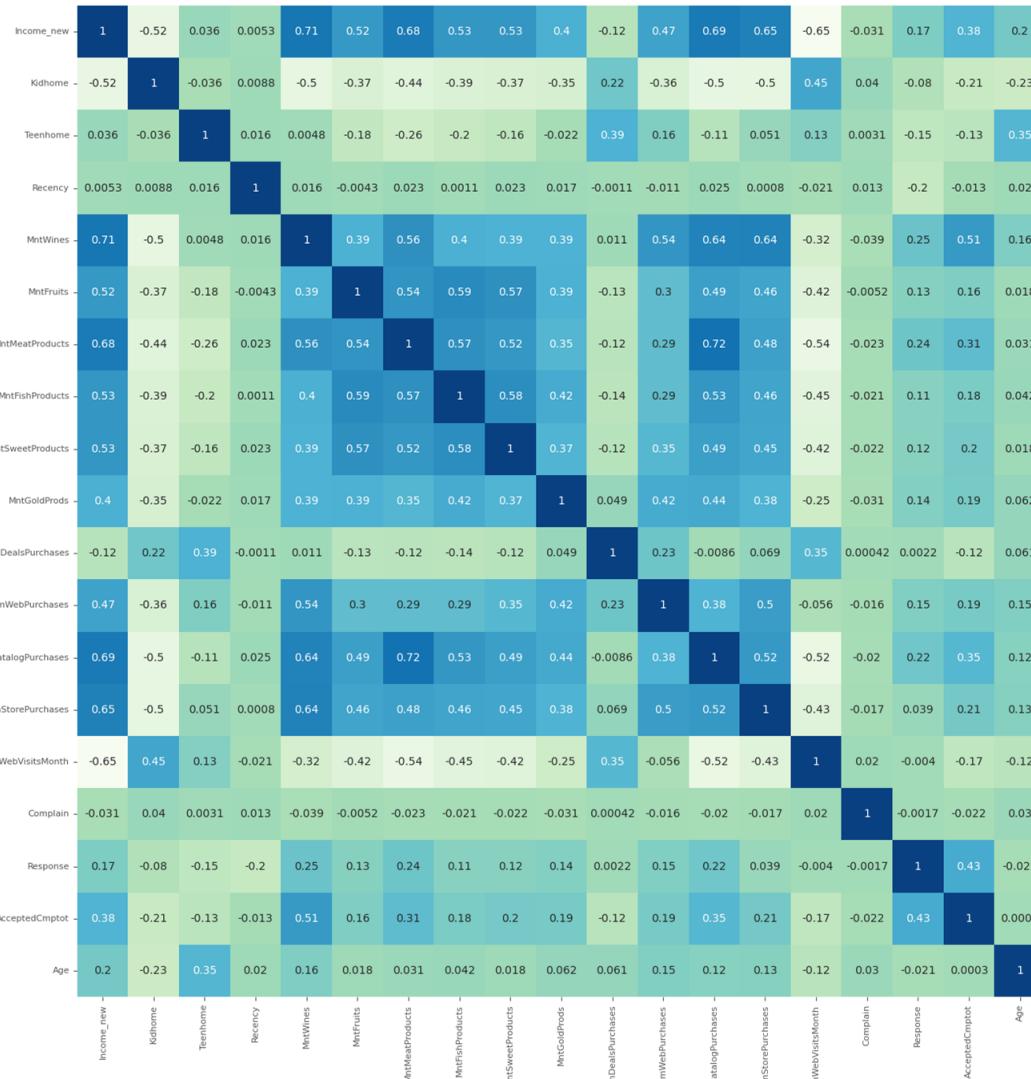
Terlihat pola yang berbeda pada variabel NumWebPurchases dan NumCatalogPurchases pada customer dengan pendidikan basic dibandingkan pendidikan lainnya



Terlihat tidak ada pola yang berbeda pada variabel purchase pada customer berdasarkan tingkat pendidikan

# Multivariate Analysis

## Korelasi tiap variabel dengan Heatmap



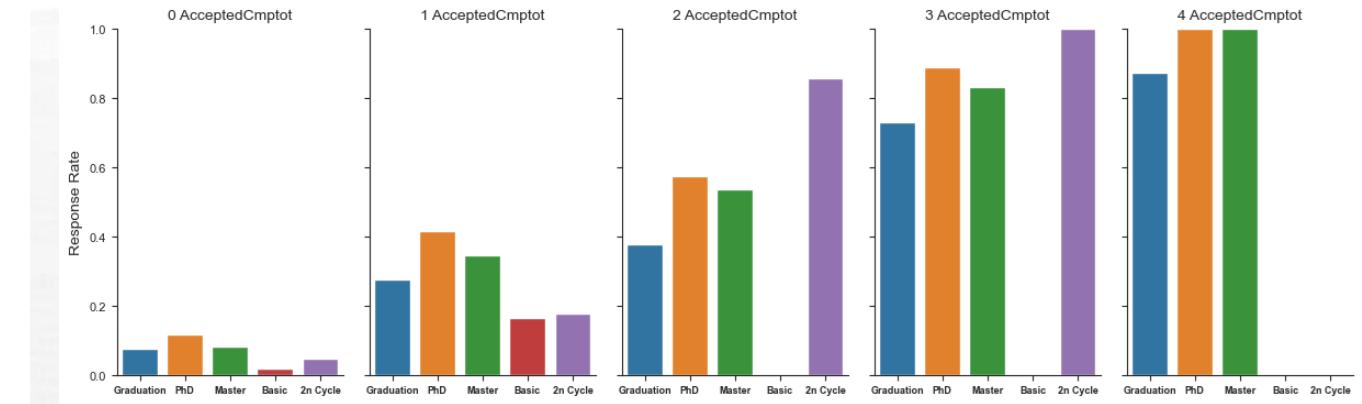
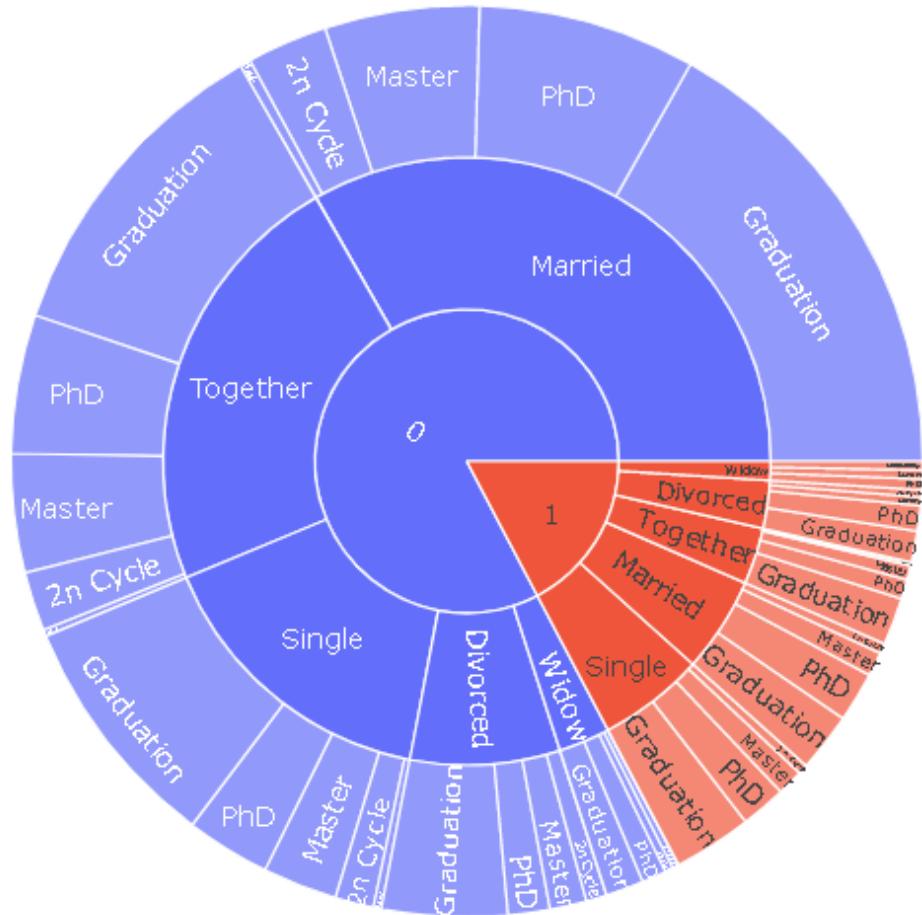
Dari Heatmap tersebut, bisa didapatkan informasi:

- Korelasi antara jumlah anak kecil dalam rumah tangga cukup besar, namun berbanding terbalik dengan jumlah pendapatan. Maka, **semakin tinggi pendapatan, semakin kecil juga jumlah anak kecil yang dimiliki**.
- Variabel lain yang memiliki korelasi yang cukup tinggi dengan pendapatan ialah jumlah konsumsi semua produk, serta jumlah pembelian melalui web, catalog, dan toko.
- Dapat diketahui pula jika variabel umur memiliki korelasi positif dengan pendapatan, namun tidak signifikan. Sehingga dapat disimpulkan, bahwa **semakin tinggi pendapatan kustomer 'umumnya' semakin tua juga umurnya**.
- Pembelian melalui toko dengan katalog memiliki korelasi yang cukup besar dan berbanding lurus dengan pembelian produk. Berbeda dengan pembelian melalui website yang berbanding terbalik, sehingga **umumnya kustomer membeli produk melalui toko dan katalog**.
- Jumlah pembelian wine dengan jumlah iklan yang diterima (accepted campaign) cukup besar dan berbanding lurus, dari hal tersebut dapat diketahui bahwa **umumnya kustomer membeli wine karena tergiur oleh iklan**.
- Serta didapatkan, bahwa **semakin banyak jumlah anak kecil dalam keluarga, maka semakin sedikit jumlah pembelian wine**.

# Multivariate Analysis

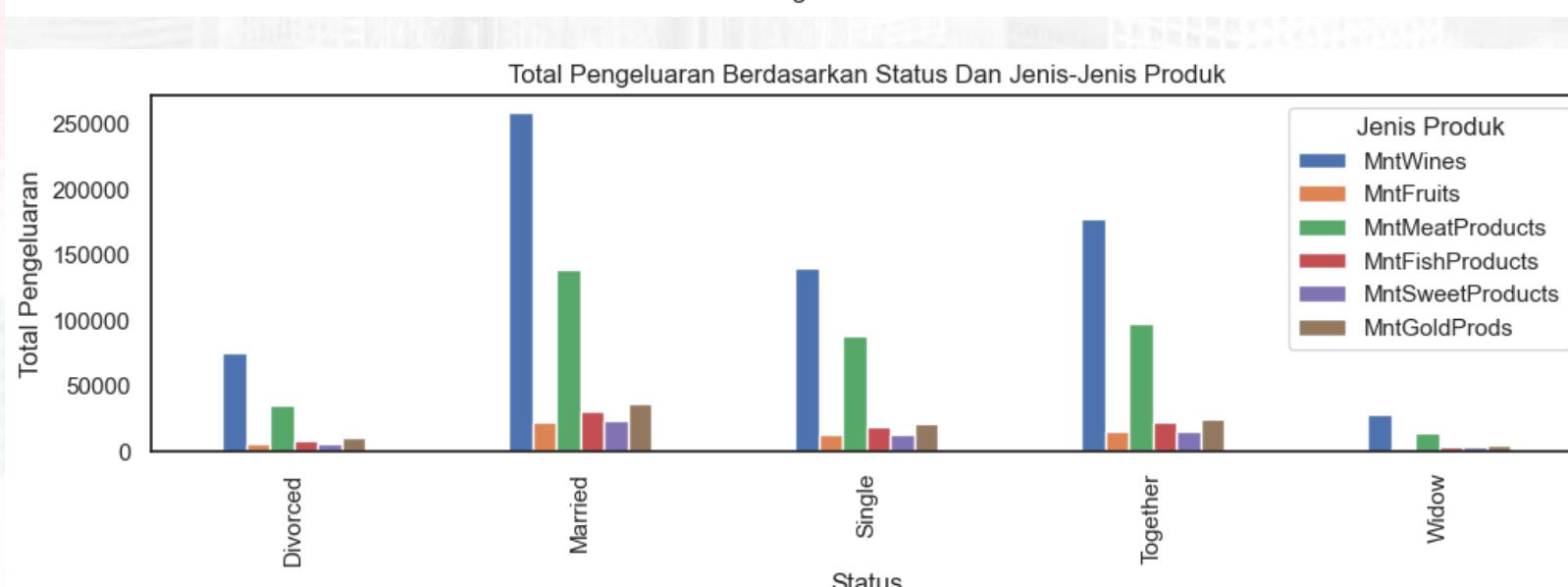
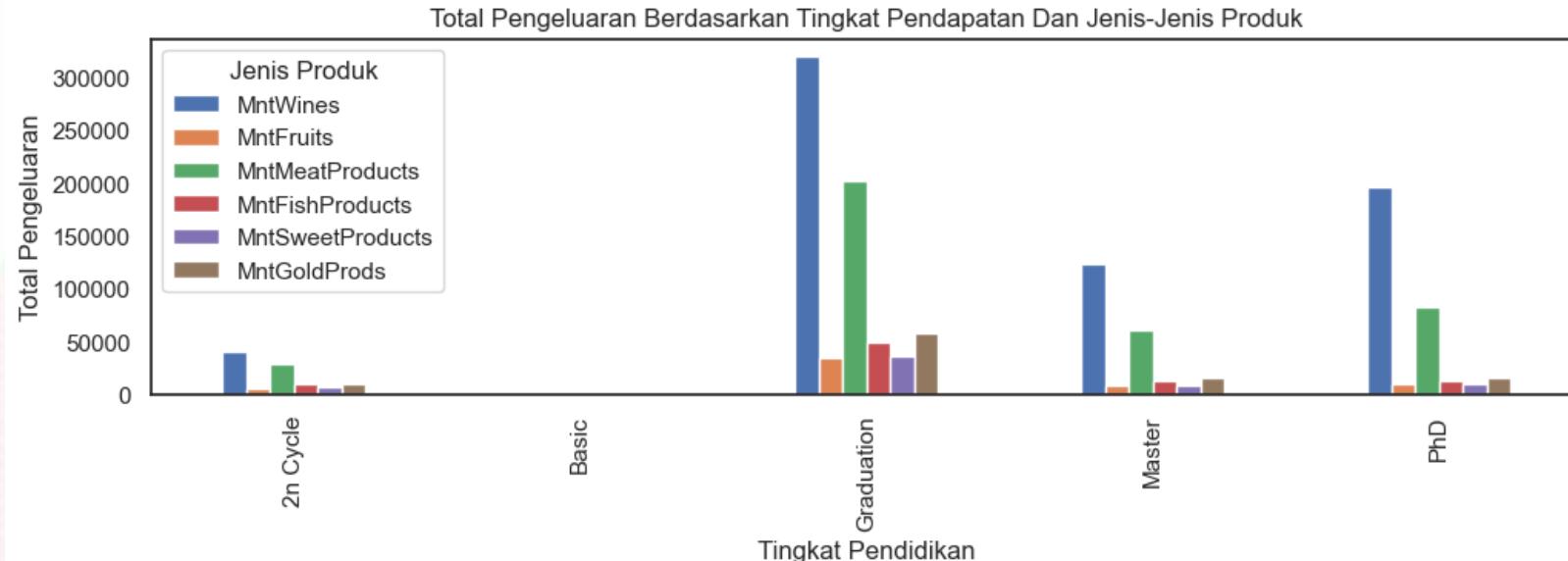
- a. Dari slide sebelumnya, dapat disimpulkan bahwa variabel yang paling relevan dan harus dipertahankan ialah variabel **income, variabel semua jenis produk yang dibeli, Jumlah anak, umur, total iklan yang direspon, serta wadah/platform pembelian (toko, katalog, website, dan sejenisnya)**.
- a. Korelasi antar fitur sudah disebutkan sebelumnya dan **yang paling menarik** perhatian adalah **variabel income** karena cukup berkorelasi dengan variabel lainnya, sehingga terdapat customer dengan ciri atau jenis tertentu yang dapat didasarkan oleh tinggi rendahnya variabel income. Dari variabel-variabel yang saling berkorelasi itulah pula **dapat dibuat segmentasi atau clustering** dengan metode *unsupervised clustering* (seperti Kmeans contohnya) untuk mendapatkan kelompok kustomer yang sejenis.

# Business Insight



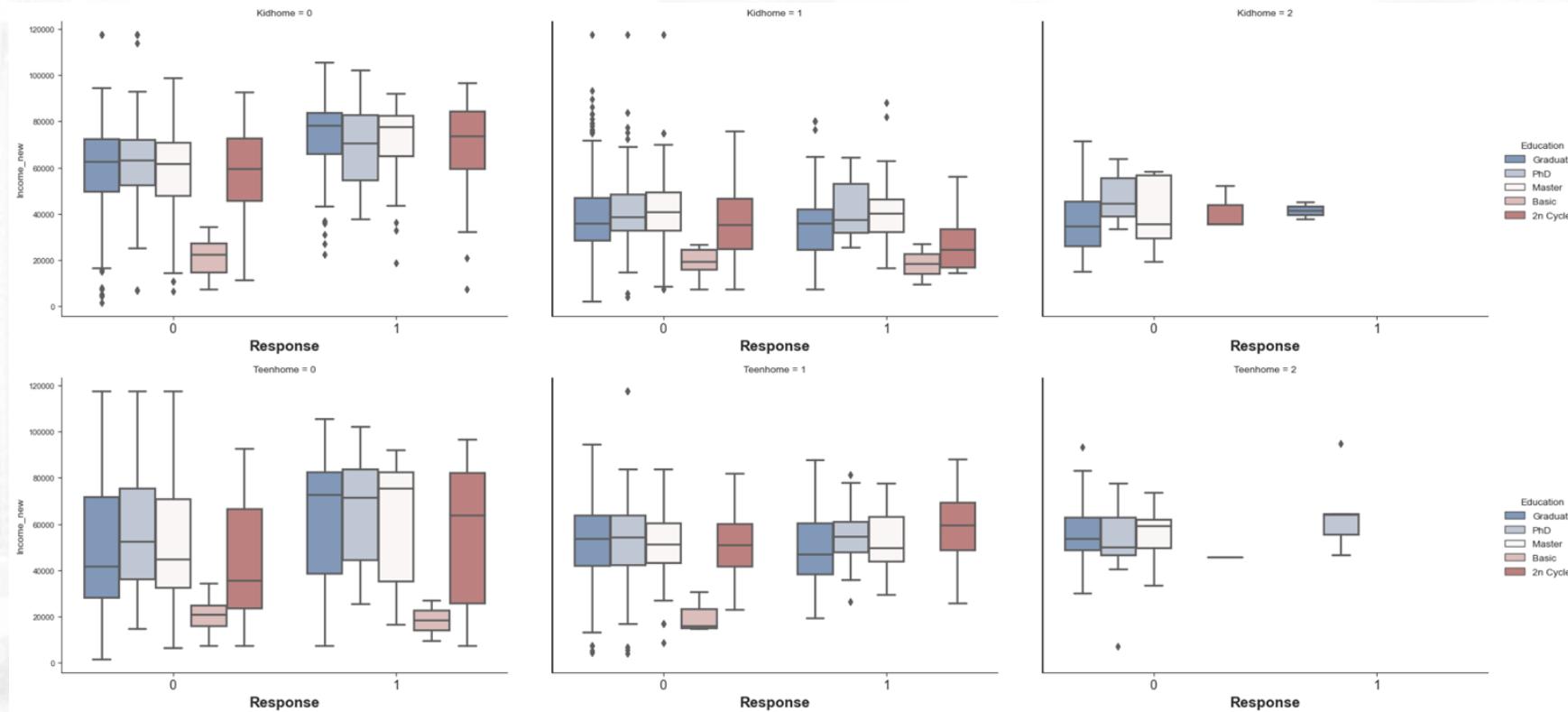
- Dari pie chart diatas dapat ditunjukkan persebaran response dari setiap customer. Selain itu, dapat kita lihat persebaran dari status pernikahan dan pendidikan masing-masing customer. Dari pie chart tersebut dapat dilihat bahwa response terbesar dimiliki oleh Customer yang Single dengan pendidikan Graduation atau sarjana. Jadi kemungkinan response tertinggi yang akan accept campaign selanjutnya adalah Single dan Sarjana.
- Bar chart menunjukkan bahwa response rate tertinggi terjadi pada customer dengan total campaign 4 dengan pendidikan minimal graduate.

# Business Insight



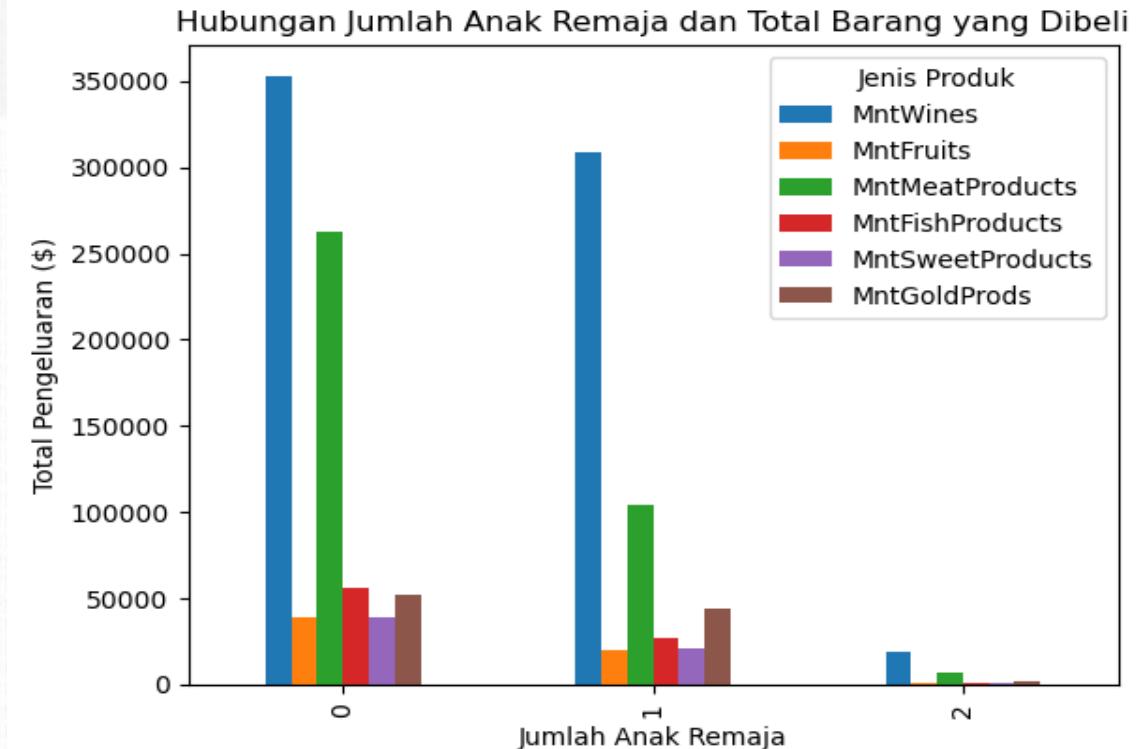
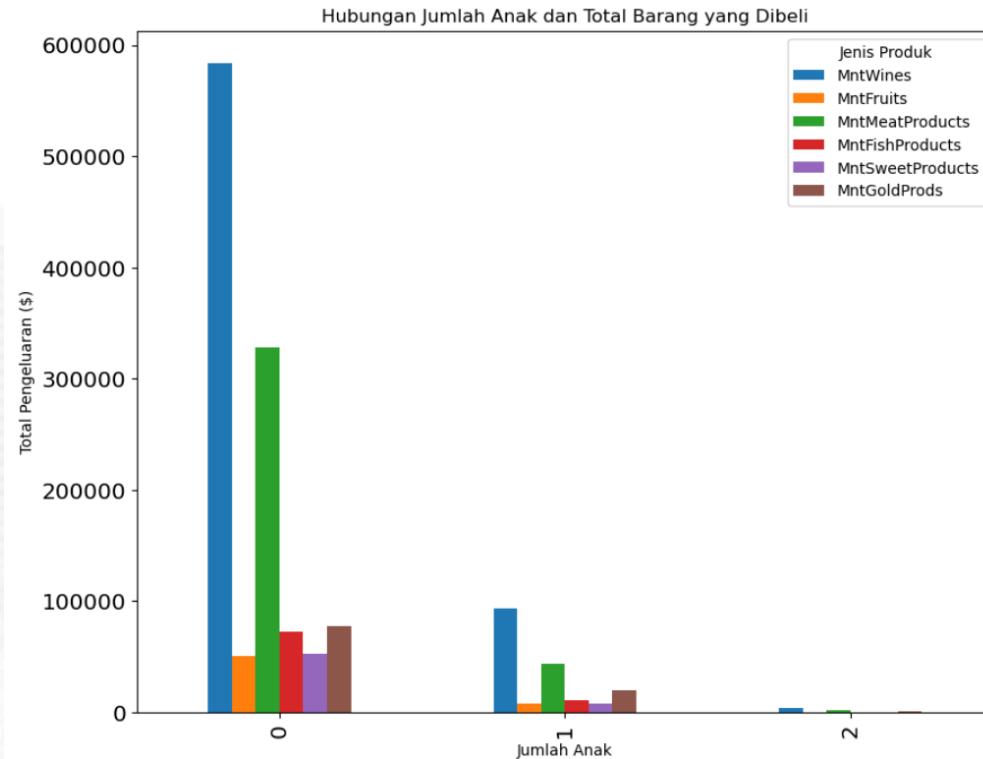
- Pada tingkat pendidikan Graduation memiliki nilai yang paling tinggi pada produk wines, begitu juga dengan tingkat pendidikan lainnya produk wines paling banyak dibeli pada kurun waktu 2 tahun terakhir ini.
- Terlihat bahwa pola pengeluaran customer pada tiap status marital memiliki tampilan yang mirip. Hal ini menunjukkan bahwa marital status kurang berpengaruh terhadap jenis pembelian produk

# Business Insight



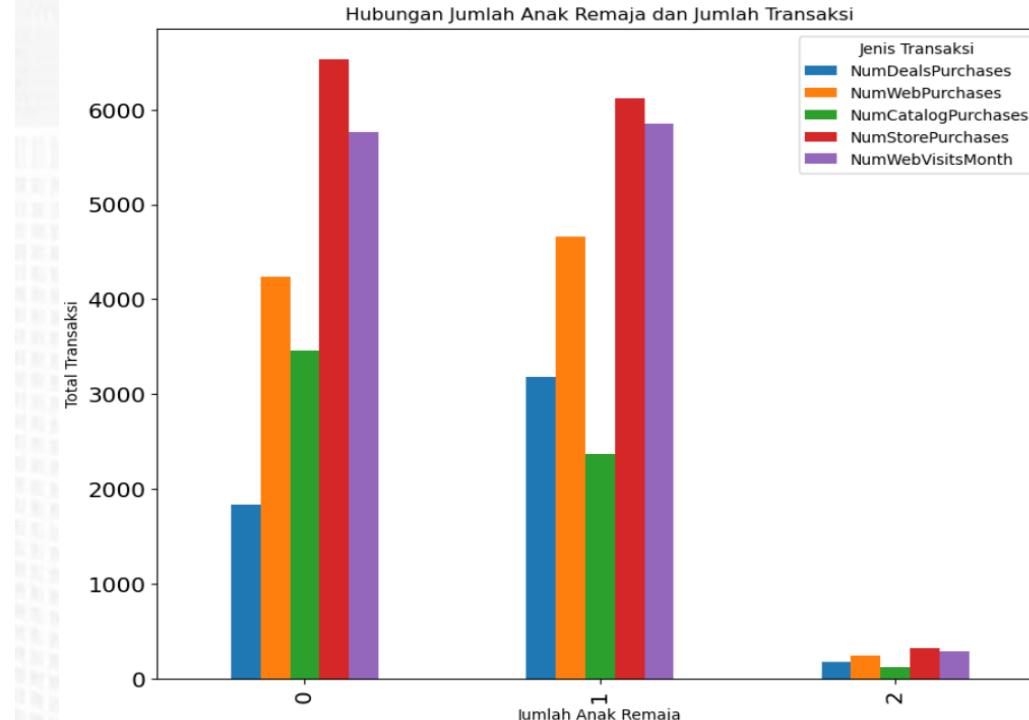
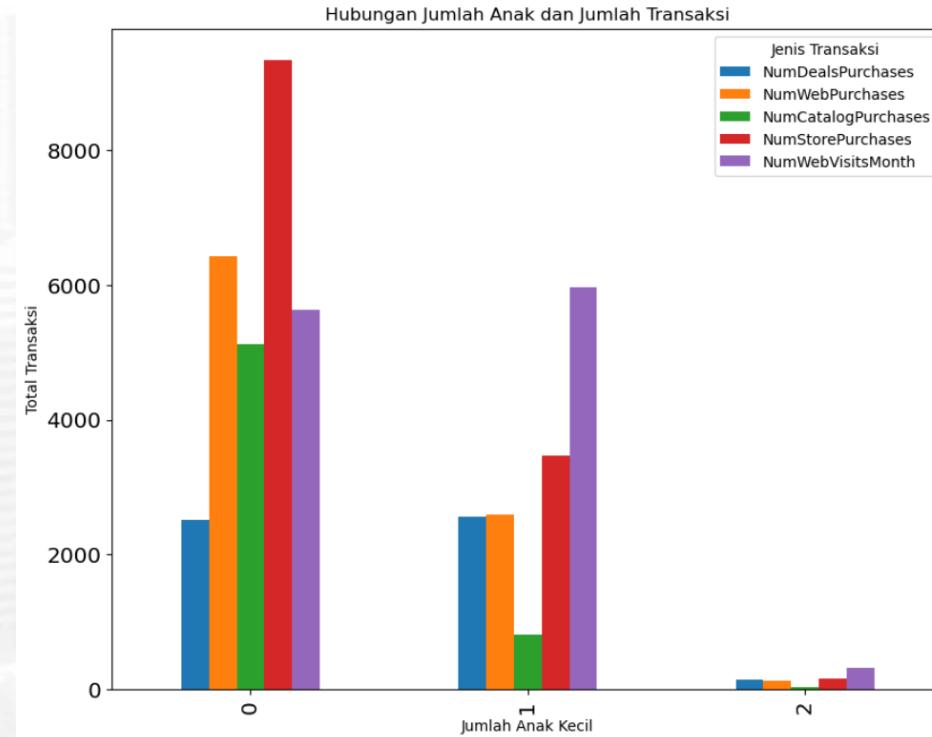
- Sebaran income customer berdasarkan education-kidhome- teenhome-response menunjukkan bahwa customer dengan response positif yang tidak memiliki kidhome cenderung memiliki income yang lebih tinggi.
- Tidak ditemukan pola income yang berbeda pada customer dengan teenhome atau tidak, namun pada customer yang tidak memiliki teenhome memiliki sebaran data yang lebih menyebar dibandingkan customer yang memiliki teenhome baik yang merespon atau tidak

# Business Insight



Dari 2 grafik di atas, dapat disimpulkan bahwa jumlah anak berpengaruh kepada jenis dan juga jumlah dari produk yang dibeli oleh customer. Akan tetapi dapat dilihat bahwa ketika memiliki anak kecil, maka pembelian wine akan menurun drastis dibandingkan dengan yang tidak memiliki anak kecil dan memiliki anak remaja

# Business Insight



Dari grafik disamping dapat disimpulkan bahwa :

- Ketika memiliki anak kecil maka customer lebih jarang untuk pergi ke toko fisik
- Ketika memiliki anak remaja, angka pembelian di toko secara fisik hanya menurun sedikit dan pembelian melalui web justru bertambah

# Business Insight

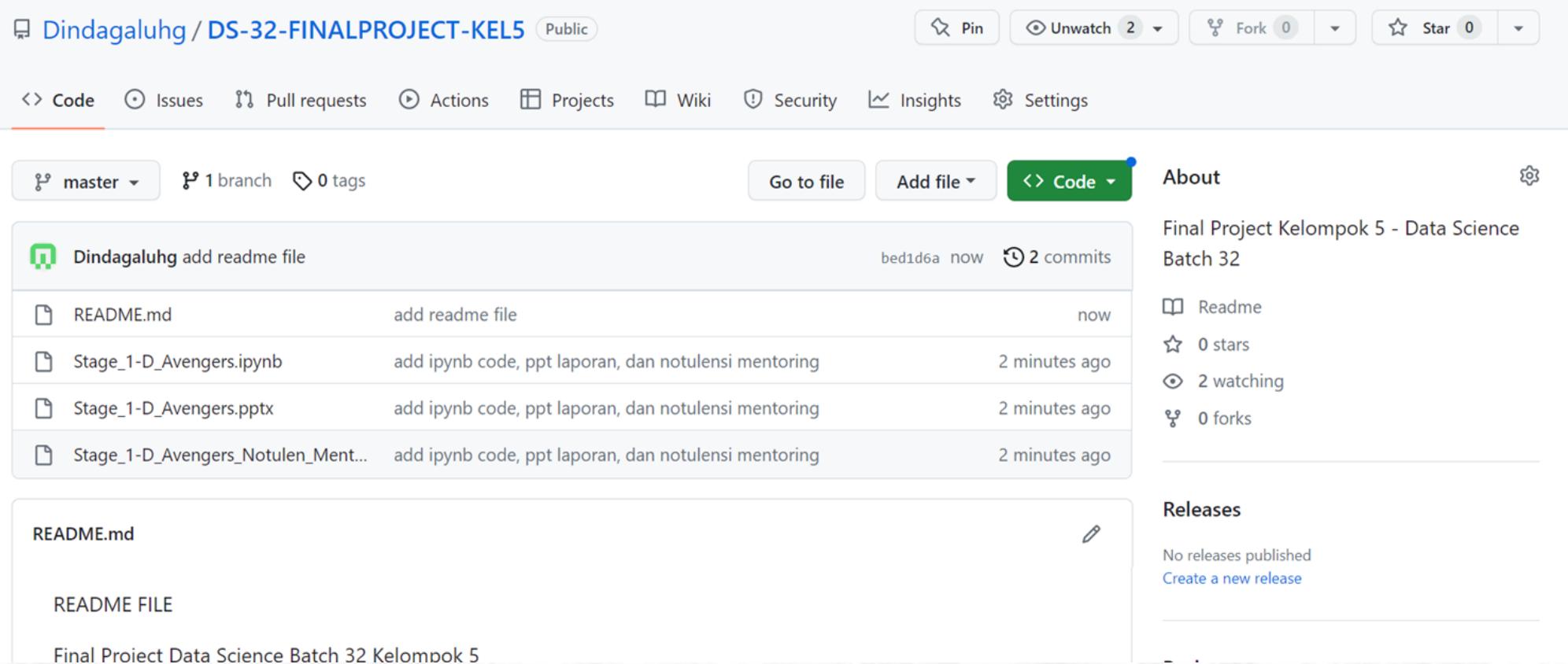
1. **Respons terbesar diberikan oleh customer dengan status single dan pendidikan minimal lulusan sarjana (graduation).** Hal ini menunjukkan bahwa kelompok ini mungkin lebih aktif dan responsif terhadap interaksi bisnis.
2. **Customer dengan pendidikan minimal lulusan sarjana** memiliki pengeluaran terbesar pada produk wine. Hal ini menunjukkan bahwa **produk wine mungkin menjadi preferensi** atau prioritas bagi kelompok pelanggan ini.
3. Perbedaan mencolok terlihat pada customer yang tidak memiliki anak dalam memberikan respons. Customer yang memberikan respons memiliki rata-rata pendapatan yang lebih tinggi dibandingkan dengan customer yang tidak memberikan respons, terlepas dari tingkat pendidikan mereka. Hal ini menunjukkan bahwa **customer dengan pendapatan tinggi cenderung lebih aktif dalam memberikan tanggapan** terhadap interaksi bisnis.
4. Total pengeluaran customer berdasarkan jenis produk menunjukkan pola yang sama, baik pada customer yang memiliki anak atau anak remaja, maupun pada customer yang tidak memiliki anak atau anak remaja. **Produk wine dan daging memiliki total pengeluaran tertinggi, terutama pada customer yang tidak memiliki anak.** Hal ini dapat memberikan wawasan penting untuk fokus pemasaran dan penjualan pada produk-produk ini.
5. Total transaksi customer juga menunjukkan pola yang sama berdasarkan jenis transaksi dan jumlah anak. **Customer yang tidak memiliki anak menghasilkan total transaksi terbesar, terutama dalam pembelian toko secara langsung.** Hal ini menunjukkan bahwa kelompok pelanggan ini mungkin merupakan segmen yang penting dalam strategi penjualan langsung.

# Suggestions

- Melakukan target marketing pada customer dengan pendidikan graduation, master, dan PhD khususnya pada produk wine
- Memberikan campaign produk wine dan daging kepada customer yang tidak memiliki anak (kid/teen) dan customer yang memiliki 1 anak remaja. Dalam kata lain, melakukan target marketing yang berbeda pada customer yang tidak memiliki dependensi (kid/teen) dan customer yang memiliki dependensi
- Semakin banyak customer dikenai campaign maka peluang customer response yang diharapkan juga semakin besar
- Melakukan personalisasi campaign: Mengetahui tingkat pendidikan pelanggan dapat memungkinkan personalisasi kampanye pemasaran. Misalnya, jika seorang pelanggan memiliki tingkat pendidikan tinggi, dapat disesuaikan konten kampanye yang lebih informatif atau fokus pada nilai-nilai intelektual. Sementara itu, bagi pelanggan dengan tingkat pendidikan yang lebih rendah, pendekatan yang lebih sederhana dan praktis mungkin lebih efektif.

# UPLOAD GITHUB

<https://github.com/Dindagaluhg/DS-32-FINALPROJECT-KEL5>



The screenshot shows a GitHub repository page. At the top, the repository name 'Dindagaluhg / DS-32-FINALPROJECT-KEL5' is displayed, along with a 'Public' badge, 'Pin' button, 'Unwatch' button (with 2 notifications), 'Fork' button (with 0 forks), and 'Star' button (with 0 stars). Below the header, there is a navigation bar with links: Code (highlighted in red), Issues, Pull requests, Actions, Projects, Wiki, Security, Insights, and Settings.

In the main content area, there is a summary bar showing 'master' branch, '1 branch', '0 tags', 'Go to file', 'Add file', and a 'Code' dropdown menu. Below this, a list of commits is shown:

- Dindagaluhg add readme file (bed1d6a now) 2 commits
- README.md add readme file now
- Stage\_1-D\_Avengers.ipynb add ipynb code, ppt laporan, dan notulensi mentoring 2 minutes ago
- Stage\_1-D\_Avengers.pptx add ipynb code, ppt laporan, dan notulensi mentoring 2 minutes ago
- Stage\_1-D\_Avengers\_Notulen\_Ment... add ipynb code, ppt laporan, dan notulensi mentoring 2 minutes ago

On the right side, there is an 'About' section with the following details:

- Final Project Kelompok 5 - Data Science Batch 32
- Readme
- 0 stars
- 2 watching
- 0 forks

Below the 'About' section, there is a 'Releases' section with the message: 'No releases published' and a link 'Create a new release'.

At the bottom left of the main content area, there is a preview of the 'README.md' file content:

```
README.md
README FILE
Final Project Data Science Batch 32 Kelompok 5
```

# Laporan Final Project

## Stage 2

---

### Kelompok 5 - D Avengers

- **Dinda Galuh Guminta**
- **Khaerun Nisa**
- **Teguh Ismareza**
- **Iqbal Fauzan Saputra**
- **Faldi Ramadhan**
- **Kadek Haris Dana Swara**
- **Ahya Ramdhanitasari**
- **Julius Pardamean**



(dipresentasikan setiap sesi mentoring)

# STAGE 2

## Data Pre-Processing

---

- Data Cleansing
- Feature Engineering

1. Missing value pada income diatas dengan median

```
df.isna().sum()
```

ID	0
Year_Birth	0
Education	0
Marital_Status	0
Income	24
Kidhome	0
Teenhome	0
Dt_Customer	0
Recency	0
MntWines	0
MntFruits	0
MntMeatProducts	0
MntFishProducts	0
MntSweetProducts	0
MntGoldProds	0
NumDealsPurchases	0
NumWebPurchases	0
NumCatalogPurchases	0
NumStorePurchases	0
NumWebVisitsMonth	0
AcceptedCmp3	0
AcceptedCmp4	0
AcceptedCmp5	0
AcceptedCmp1	0
AcceptedCmp2	0
Complain	0
Z_CostContact	0
Z_Revenue	0
Response	0
dtype:	int64

## **2. Tidak terdapat data duplikat**

```
df.duplicated(subset=['ID']).sum()
```

8

### **3. Melakukan feature extraction**

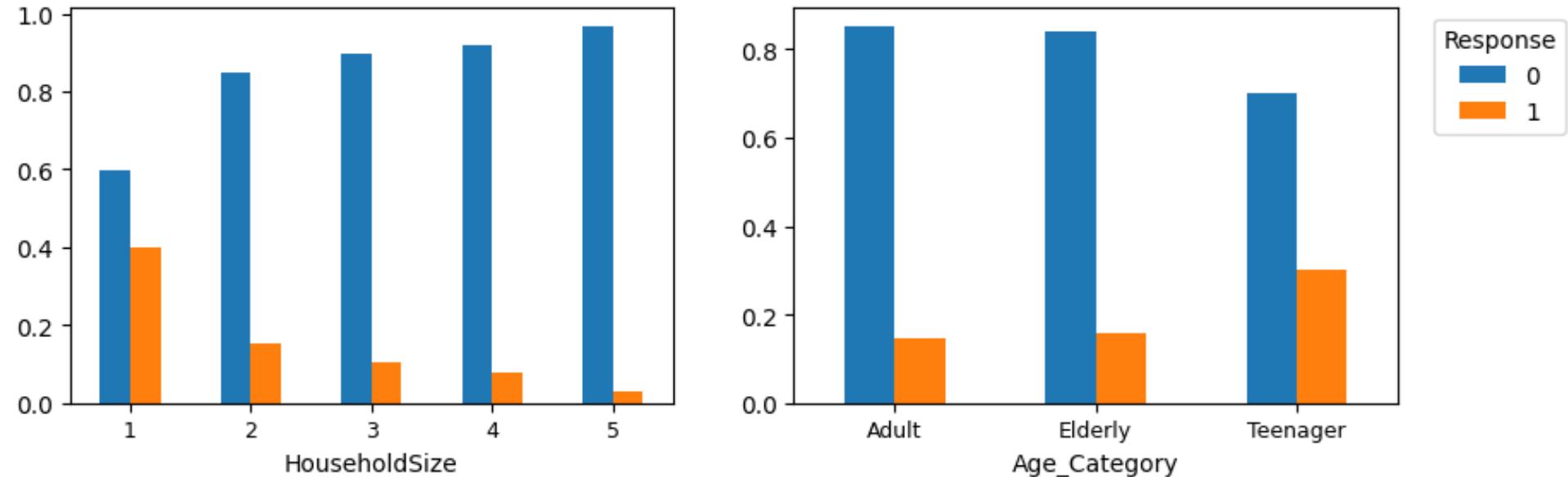
- Membuat fitur age menjadi categorical
  - Membuat fitur baru yaitu jumlah anggota keluarga

#### **4. Melakukan feature encoding pada marital status, education, dan age**

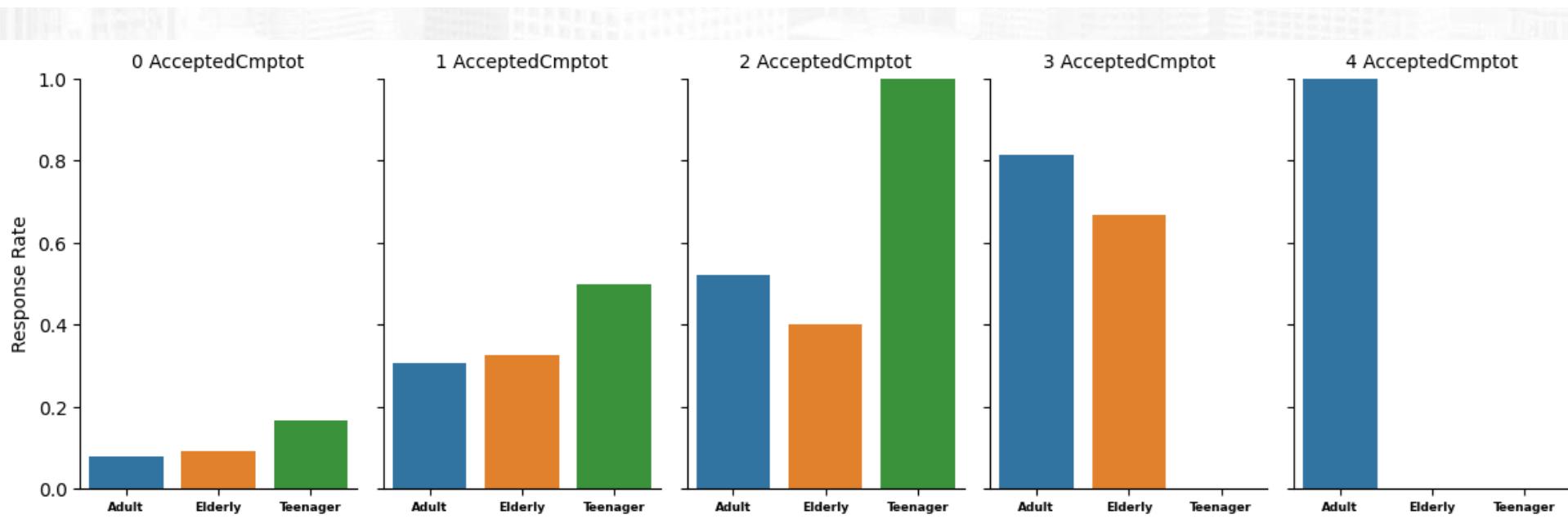
```
# feature encoding pada education dan marital status
# label encoder pada Education
df['Education'] = df['Education'].astype('category').cat.codes
df['Age_Category'] = df['Age_Category'].astype('category').cat.codes

# one hot encoding pada marital status karena tidak memiliki urutan
status_onehot = pd.get_dummies(df['Marital_Status'], prefix='Status')
df = df.join(status_onehot)
```

Hubungan jumlah anggota keluarga dan kelompok umur terhadap respon



Response rate berdasarkan total accepted campaign dan kelompok umur



## 5. Melakukan feature selection

```
: df.drop(columns=['ID', 'Year_Birth', 'Marital_Status', 'Dt_Customer', 'Z_CostContact',
                 'Z_Revenue', 'AcceptedCmp1', 'AcceptedCmp2', 'AcceptedCmp3',
                 'AcceptedCmp4', 'AcceptedCmp5', 'ParentSize', 'Age']).copy()
```

## 6. Split data menjadi training-testing (70:30)

```
y_train.value_counts()
```

Response	
0	1329
1	239
dtype: int64	

```
y_test.value_counts()
```

Response	
0	577
1	95
dtype: int64	

## 7. Mengatasi outlier pada training dan testing

```
low, high = outlier(X_train['Income'])

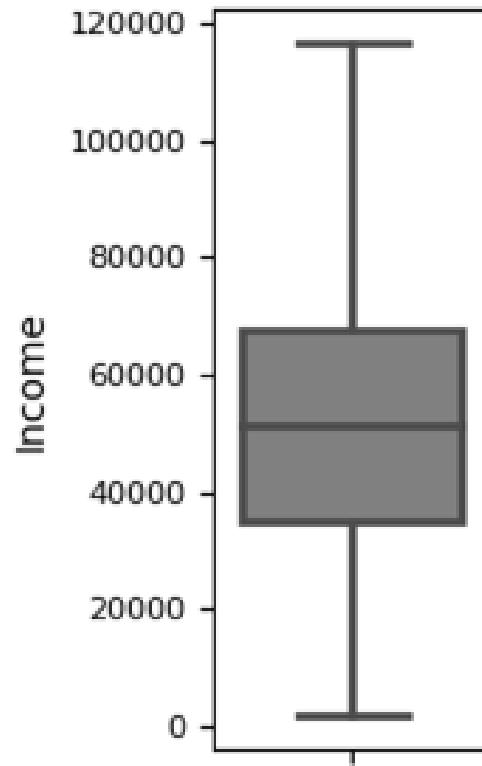
#replacing outlier with upper bound and lower bound value
X_train['Income'] = np.where(X_train['Income']>high, high, X_train['Income'])
X_train['Income'] = np.where(X_train['Income']<low, low, X_train['Income'])
X_test['Income'] = np.where(X_test['Income']>high, high, X_test['Income'])
X_test['Income'] = np.where(X_test['Income']<low, low, X_test['Income'])
```

## 8. Melakukan feature transformation dengan standardisasi

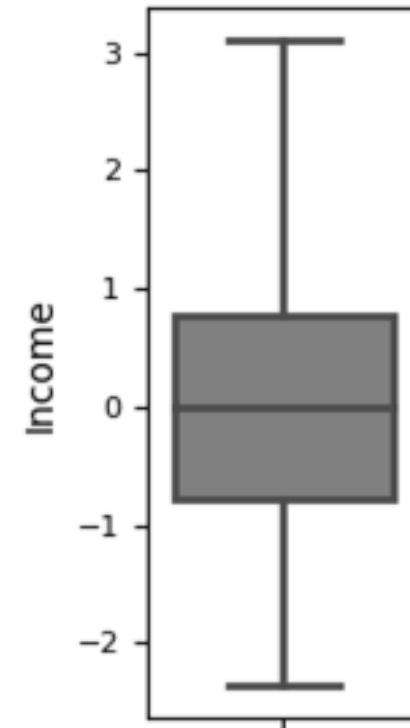
```
] from sklearn.preprocessing import StandardScaler
ss = StandardScaler()

scaler = ss.fit(X_train[['Income']])
X_train['Income'] = scaler.transform(X_train[['Income']])
X_test['Income'] = scaler.transform(X_test[['Income']])
# scaler.to_pickle(filename)
```

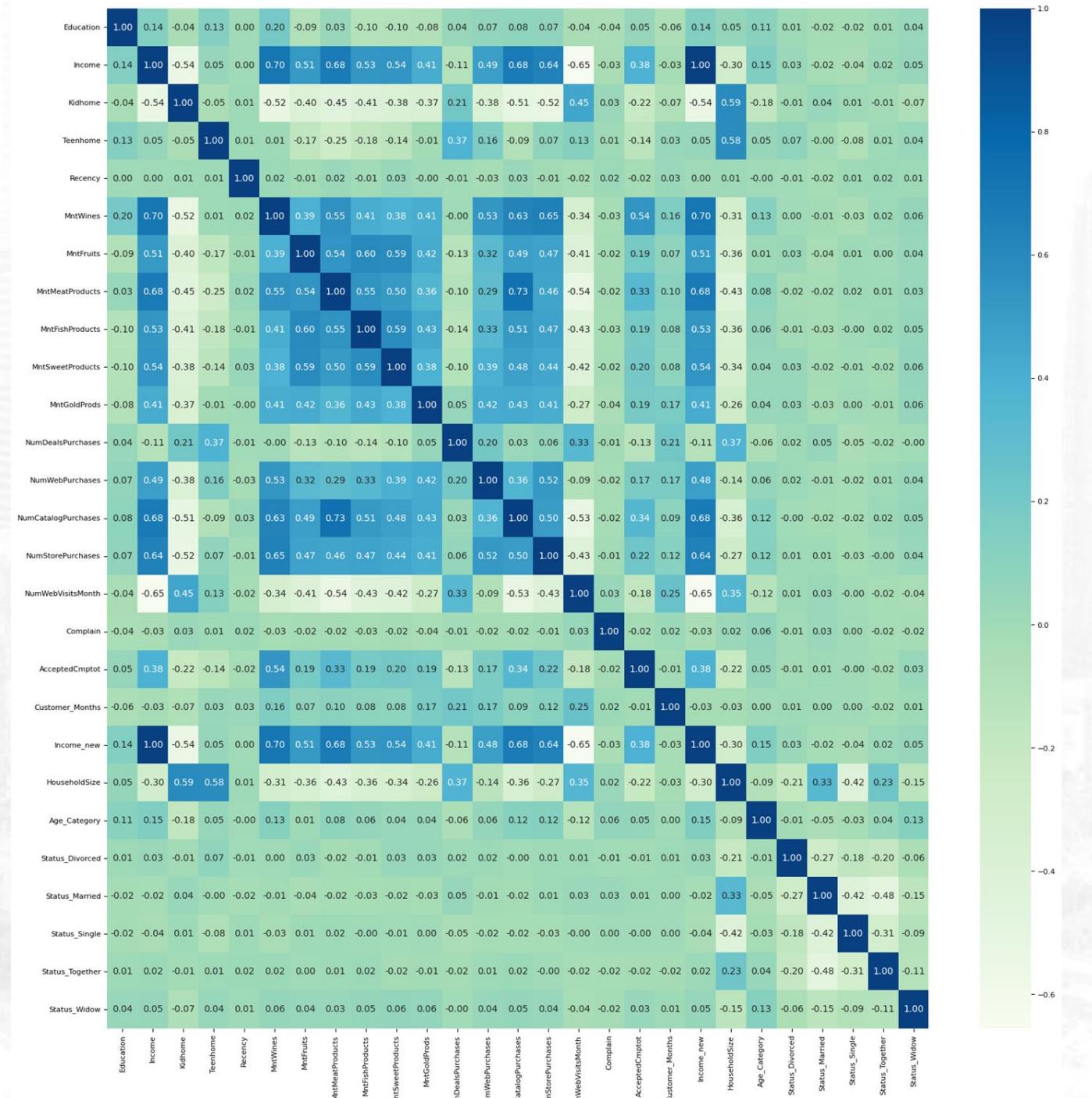
setelah handle outlier



setelah standardisasi



# Korelasi pada data training setelah preprocessing



# Handle Data Imbalance

Dipilih variabel ‘response’ karena memiliki imbalance yang cukup tinggi dan variabel ini akan dipakai untuk memprediksi ciri kustomer yang akan respons iklan.

```
# Menghitung jumlah kemunculan setiap nilai dalam variabel
value_counts = df['Response'].value_counts()

# Mengambil jumlah "yes" dan "no"
yes_count = value_counts.get('0', 1906)
no_count = value_counts.get('1', 334)

# Menghitung persentase "yes" dan "no"
total_count = len(df)
yes_percentage = (yes_count / total_count) * 100
no_percentage = (no_count / total_count) * 100

# Menampilkan hasil
print("Persentase Yes:", yes_percentage)
print("Persentase No:", no_percentage)
```

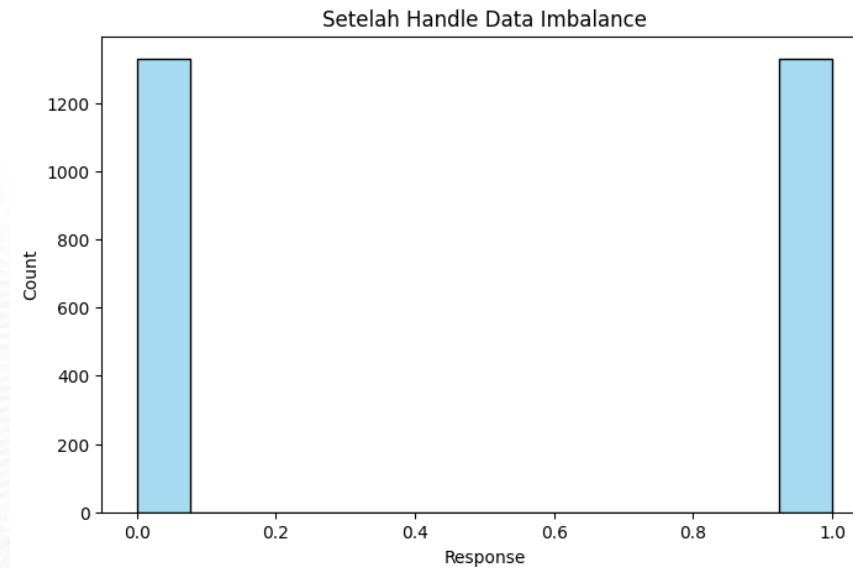
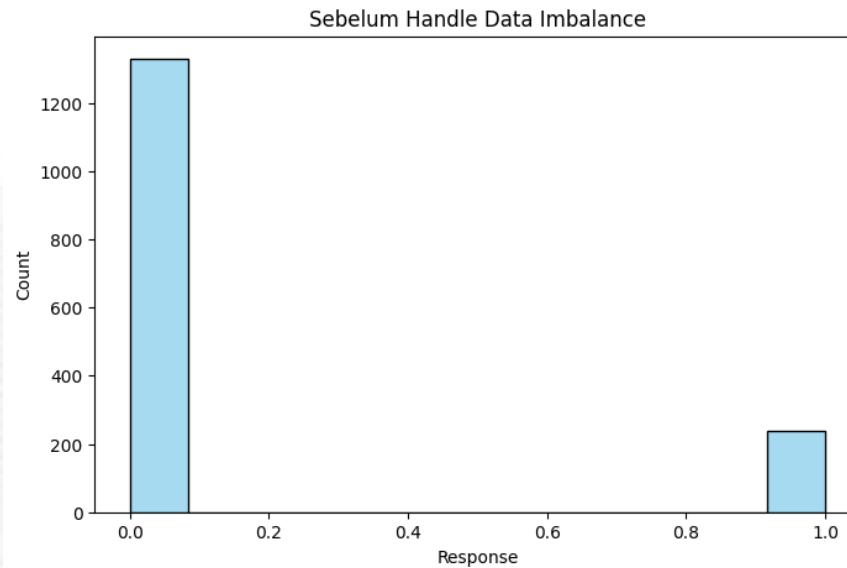
Persentase Yes: 85.08928571428571  
Persentase No: 14.910714285714285

```
# pembuatan binary label target
df['response_class'] = df['Response'] > 0.8
df['response_class'].value_counts()
```

```
False    1906
True     334
Name: response_class, dtype: int64
```

\*persentase ‘no response’ kurang dari 20%

# Handle Data Imbalance - SMOTE



```
print(x_train.shape)  
print(y_train.shape)
```

(1568, 27)  
(1568, 1)

```
print(df_over.shape)
```

(2658, 28)

- Histogram diatas menggambarkan kondisi variabel complain, response, dan total accepted campaign sebelum (kiri) dan sesudah (kanan) dilakukan proses handle data imbalance.
- Dengan adanya proses tersebut, diyakini dapat meningkatkan akurasi prediksi kustomer yang seperti apa yang akan merespons iklan.

# Feature Tambahan

1. **WebEngagement:** Menghitung tingkat interaksi pelanggan dengan situs web perusahaan. Fitur ini mencerminkan seberapa sering pelanggan mengunjungi situs web dan seberapa baru kunjungan terakhir mereka.
2. **Jumlah cicilan:** variable ini bisa membantu untuk melihat apakah customer yang memiliki cicilan bisa mempengaruhi respons campaign.
3. **Tempat tinggal customer:** untuk mengetahui apakah customer cocok dengan campaign yang akan dibuat, apakah campaign dalam bentuk web purchase atau store purchase
4. **Delivery time from web purchase:** lama pengiriman dapat mempengaruhi customer untuk merespon campaign

# STAGE 3

## Machine Learning Modeling & Evaluation

---

- Modeling
- Feature Importance

# Modeling

- A. **Split Data Train & Test** dengan rasio 70 : 30
- B. **Modeling** : logistic regression, random forest, dan adaboost
- C. **Model Evaluation** : AUC, precision, recall
- D. **Model Validation** : 5-fold cross validation
- E. **Hyperparameter Tuning**

**Dilakukan percobaan pada data tanpa smote dan data hasil smote**

# Logistic Regression

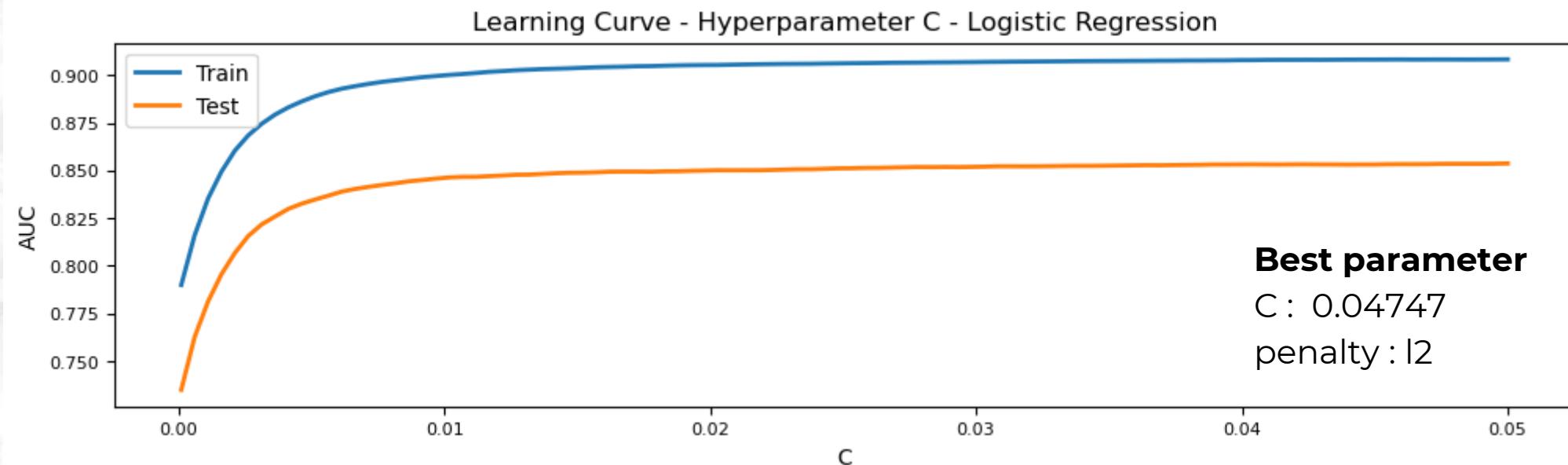
1. Dilakukan fit model
2. Validasi dengan 5-fold
3. Tuning hyperparameter dengan random search
  - penalty = ['l1', 'l2']
  - C

Model performance

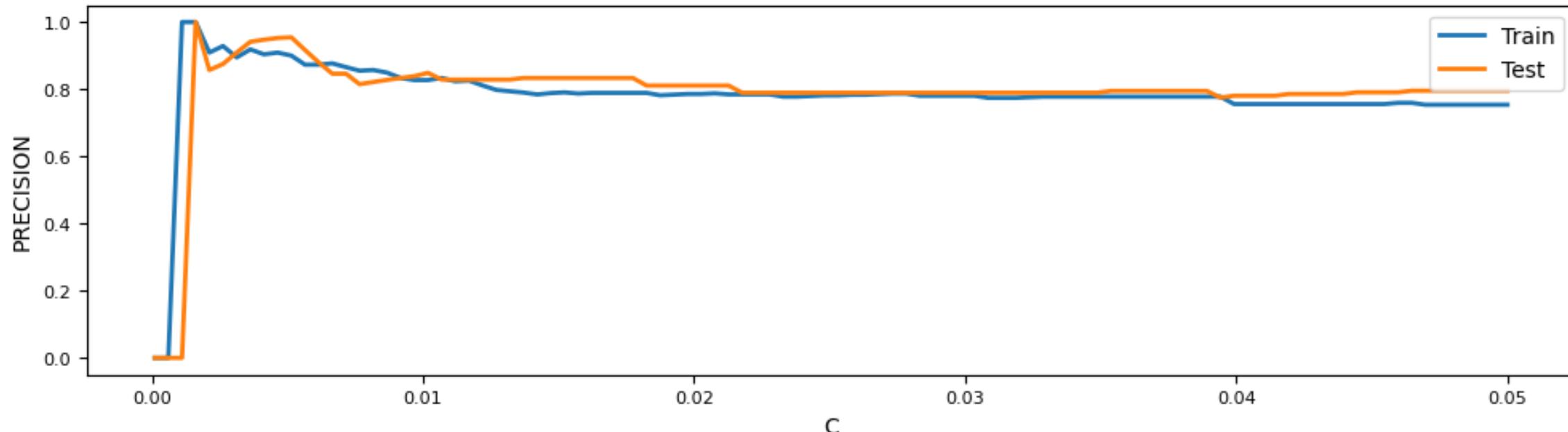
Metric	Train	Test
Accuracy	0,89	0,89
Precision	0,74	0,74
Recall	0,48	0,39
F1-Score	0,58	0,51
AUC	0,91	0,86
<b>AUC (5-fold cv)</b>	<b>0,91</b>	<b>0,90</b>

Setelah tuning parameter

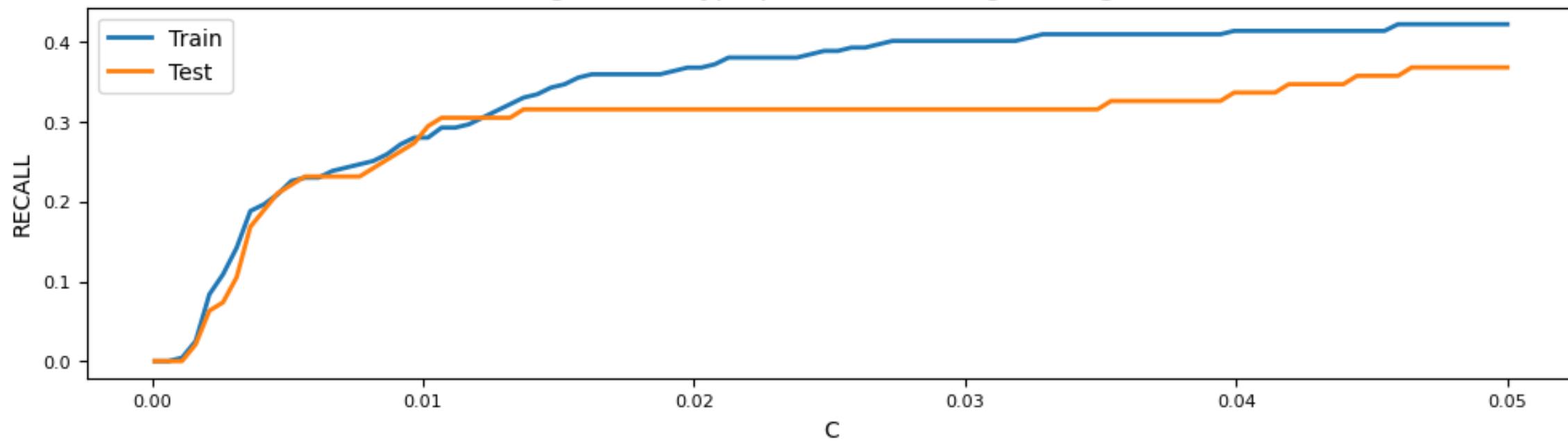
Metric	Train	Test
Accuracy	0,89	0,90
Precision	0,75	0,80
Recall	0,42	0,37
F1-Score	0,54	0,50
AUC	0,91	0,85
<b>AUC (5-fold cv)</b>	<b>0,91</b>	<b>0,90</b>



### Learning Curve - Hyperparameter C - Logistic Regression

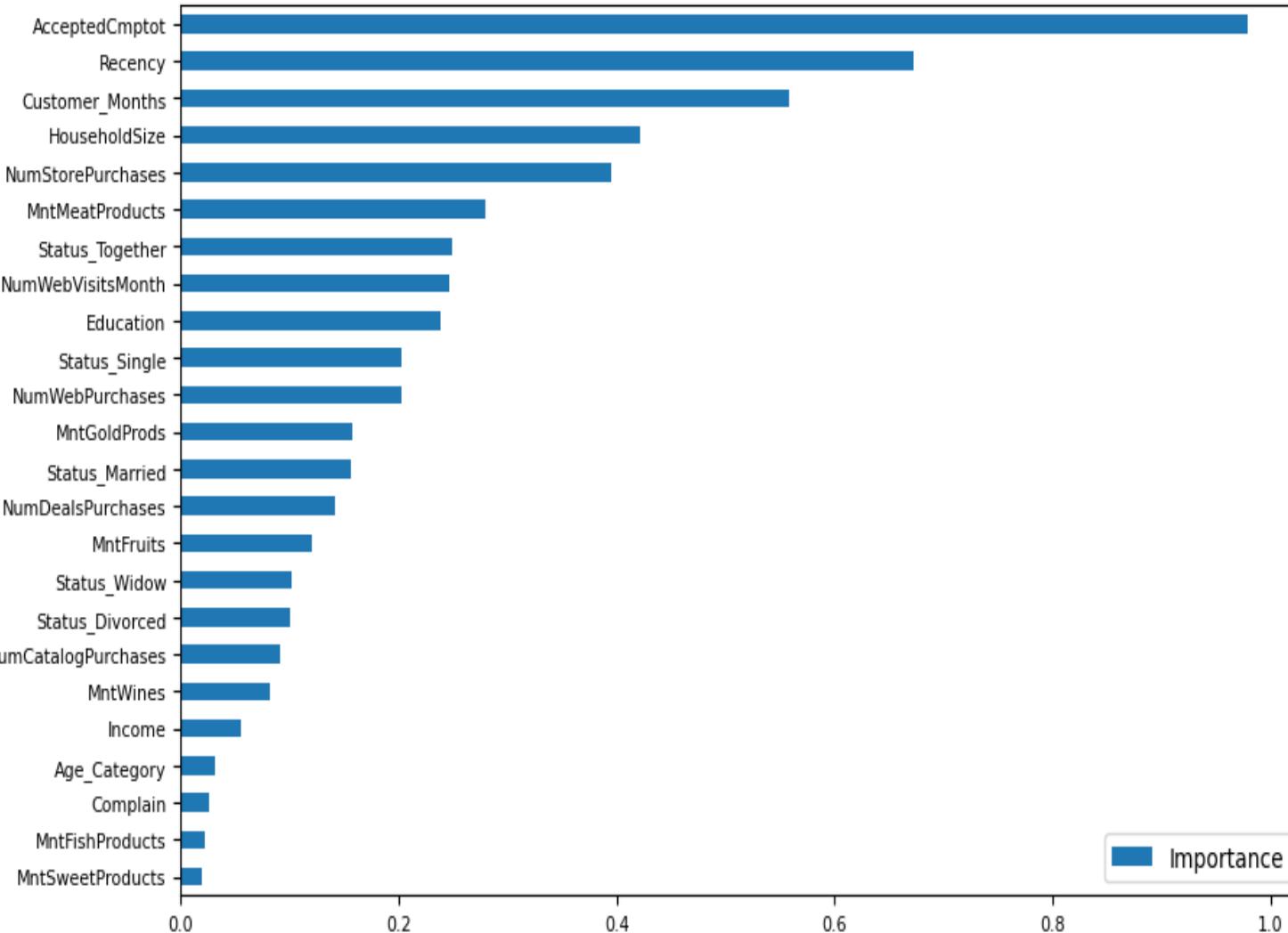


### Learning Curve - Hyperparameter C - Logistic Regression



# Logistic Regression - Feature Importance

Feature Importance dengan Koefisien Estimator



Signifikansi Parameter

	Coef.	Std.Err.	z	P> z
Education	0.3097	0.0746	4.1501	0.0000
Income	-0.0765	0.1984	-0.3858	0.6997
Recency	-0.8717	0.1023	-8.5213	0.0000
MntWines	-0.2619	0.1557	-1.6824	0.0925
MntFruits	0.1824	0.1164	1.5666	0.1172
MntMeatProducts	0.4170	0.1361	3.0647	0.0022
MntFishProducts	-0.0590	0.1189	-0.4963	0.6197
MntSweetProducts	0.0253	0.1141	0.2220	0.8243
MntGoldProds	0.1919	0.0978	1.9626	0.0497
NumDealsPurchases	0.1547	0.1072	1.4431	0.1490
NumWebPurchases	0.2811	0.1070	2.6279	0.0086
NumCatalogPurchases	0.0909	0.1355	0.6710	0.5022
NumStorePurchases	-0.5185	0.1354	-3.8290	0.0001
NumWebVisitsMonth	0.3044	0.1403	2.1699	0.0300
Complain	-1.4630	1.8813	-0.7776	0.4368
AcceptedCmptot	1.3109	0.1128	11.6251	0.0000
Customer_Months	0.7749	0.1104	7.0159	0.0000
HouseholdSize	-0.3895	0.1691	-2.3034	0.0213
Age_Category	0.1233	0.2602	0.4737	0.6357
Status_Divorced	-2.9304	0.3449	-8.4971	0.0000
Status_Married	-3.5543	0.2681	-13.2573	0.0000
Status_Single	-2.8299	0.2856	-9.9099	0.0000
Status_Together	-3.7738	0.2976	-12.6791	0.0000
Status_Widow	-2.7070	0.4750	-5.6986	0.0000

# Random Forest

1. Dilakukan fit model
2. Validasi dengan 5-fold
3. Tuning hyperparameter random search
  - n\_estimators
  - criterion : ['gini', 'entropy']
  - max\_depth
  - min\_samples\_split
  - min\_samples\_leaf

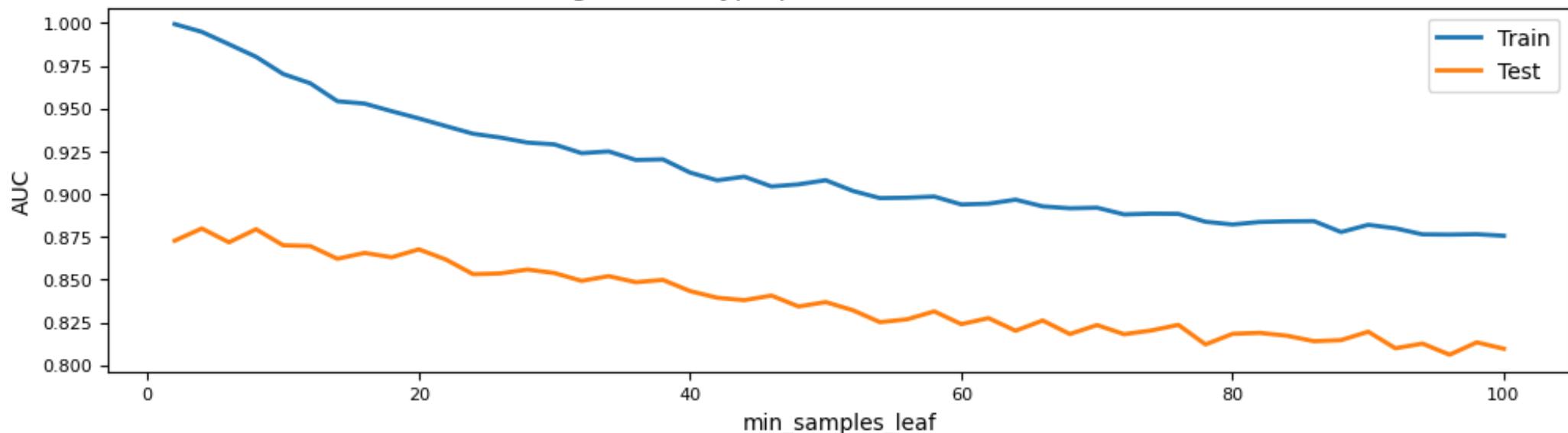
Model performance

Metric	Train	Test
<b>Accuracy</b>	0,99	0,88
<b>Precision</b>	0,98	0,68
<b>Recall</b>	0,98	0,29
<b>F1-Score</b>	0,98	0,41
<b>AUC</b>	1,00	0,87
<b>AUC (5-fold cv)</b>	1,00	0,89

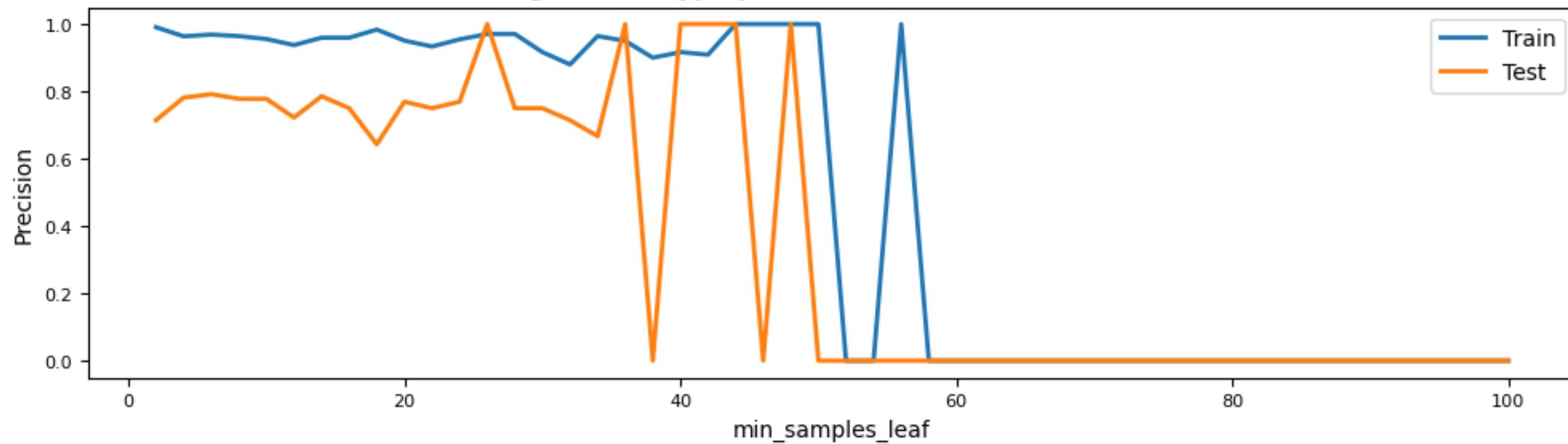
Setelah tuning parameter

Metric	Train	Test
<b>Accuracy</b>	0,90	0,87
<b>Precision</b>	0,96	0,71
<b>Recall</b>	0,36	0,13
<b>F1-Score</b>	0,52	0,21
<b>AUC</b>	0,97	0,87
<b>AUC (5-fold cv)</b>	0,99	0,89

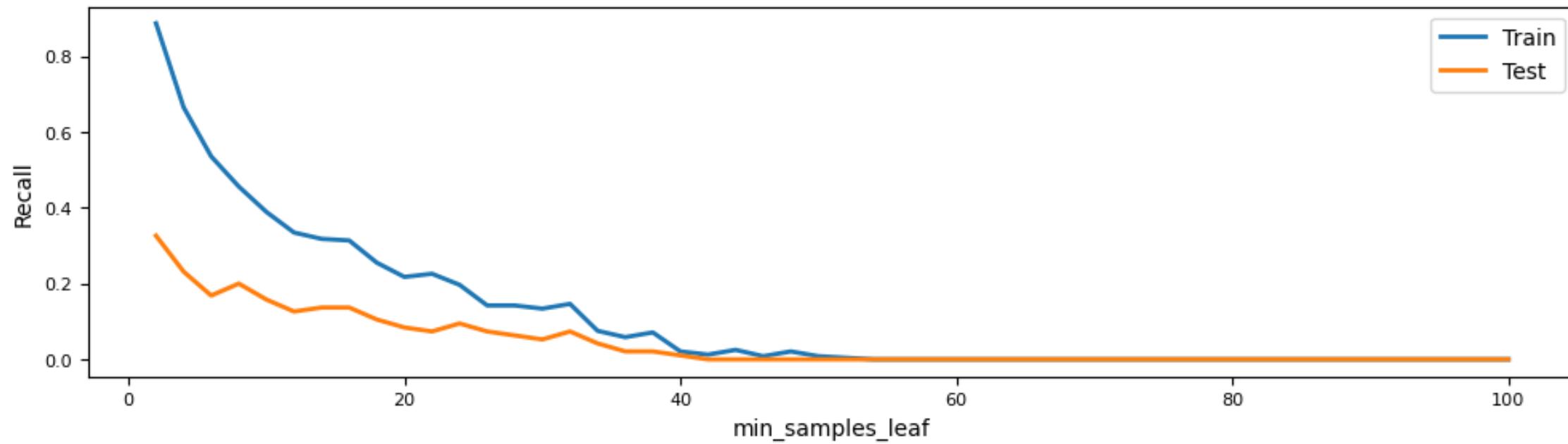
Learning Curve - Hyperparameter AUC - Random Forest



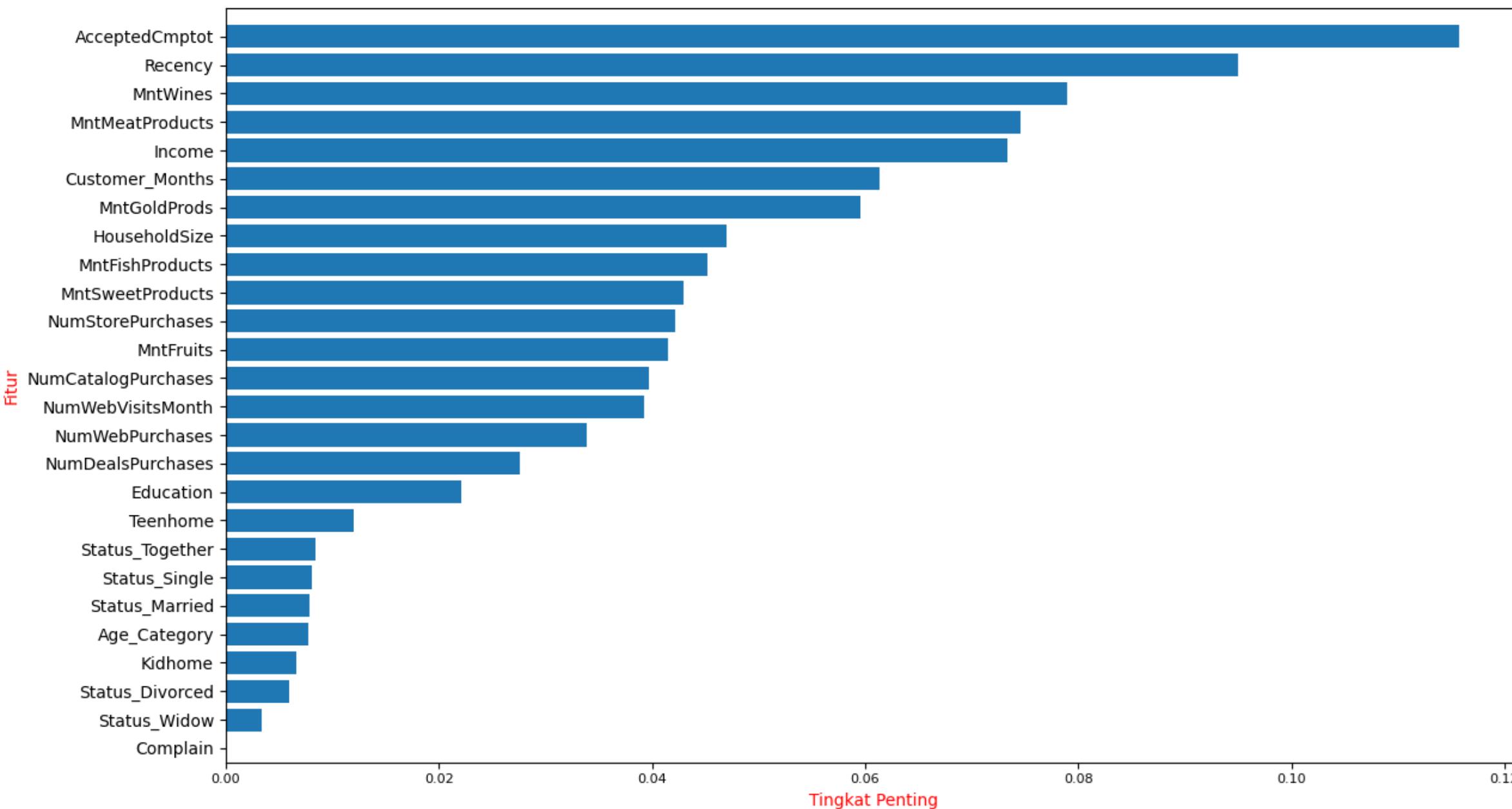
### Learning Curve - Hyperparameter Precision - Random Forest



### Learning Curve - Hyperparameter Precision - Random Forest



# Random Forest - Feature Importance



# Adaboost

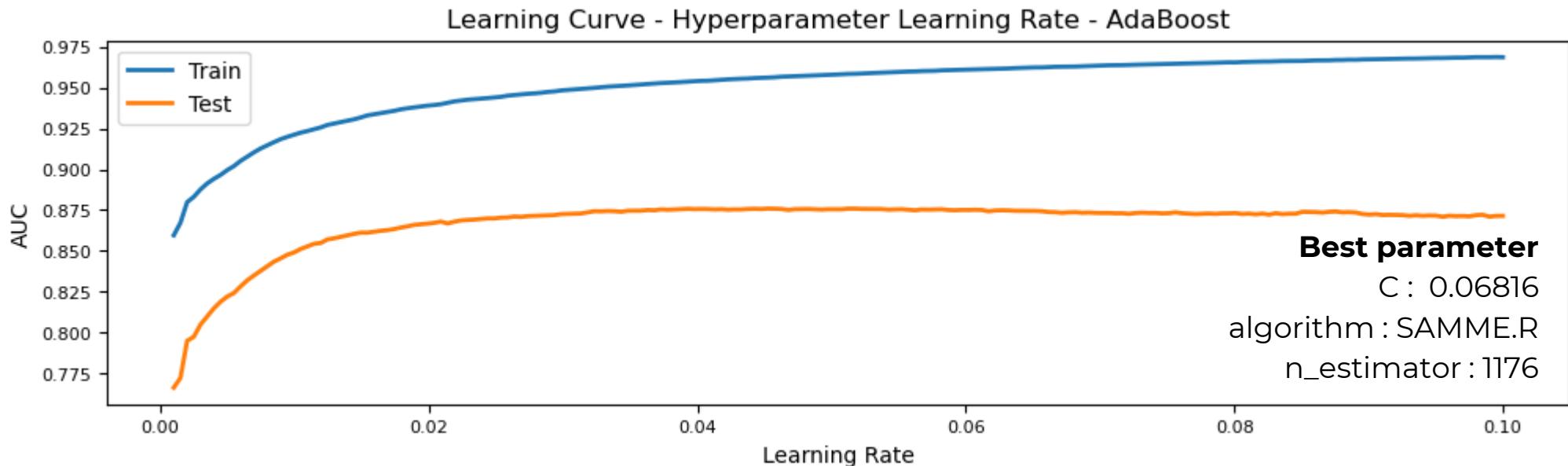
1. Dilakukan fit model
2. Validasi dengan 5-fold
3. Tuning hyperparameter random search
  - algorithm= ['SAMME', 'SAMME.R']
  - n\_estimator
  - learning\_rate

Model performance

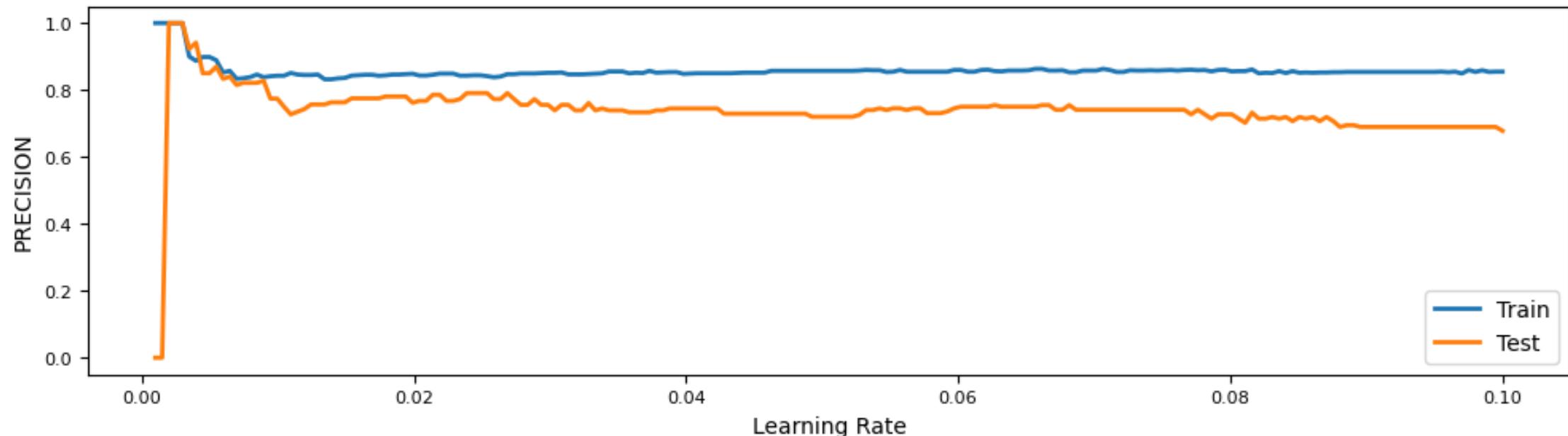
Metric	Train	Test
<b>Accuracy</b>	0,92	0,89
<b>Precision</b>	0,79	0,67
<b>Recall</b>	0,64	0,48
<b>F1-Score</b>	0,71	0,56
<b>AUC</b>	0,96	0,86
<b>AUC (5-fold cv)</b>	0,96	0,89

Setelah tuning parameter

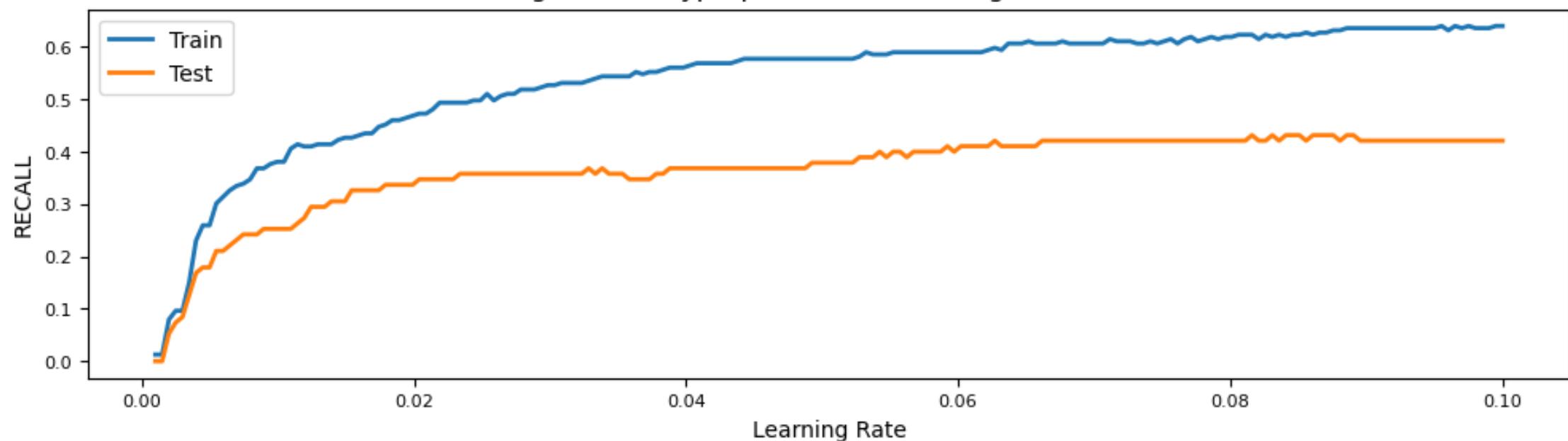
Metric	Train	Test
<b>Accuracy</b>	0,92	0,90
<b>Precision</b>	0,85	0,75
<b>Recall</b>	0,61	0,42
<b>F1-Score</b>	0,71	0,54
<b>AUC</b>	0,96	0,87
<b>AUC (5-fold cv)</b>	0,96	0,90



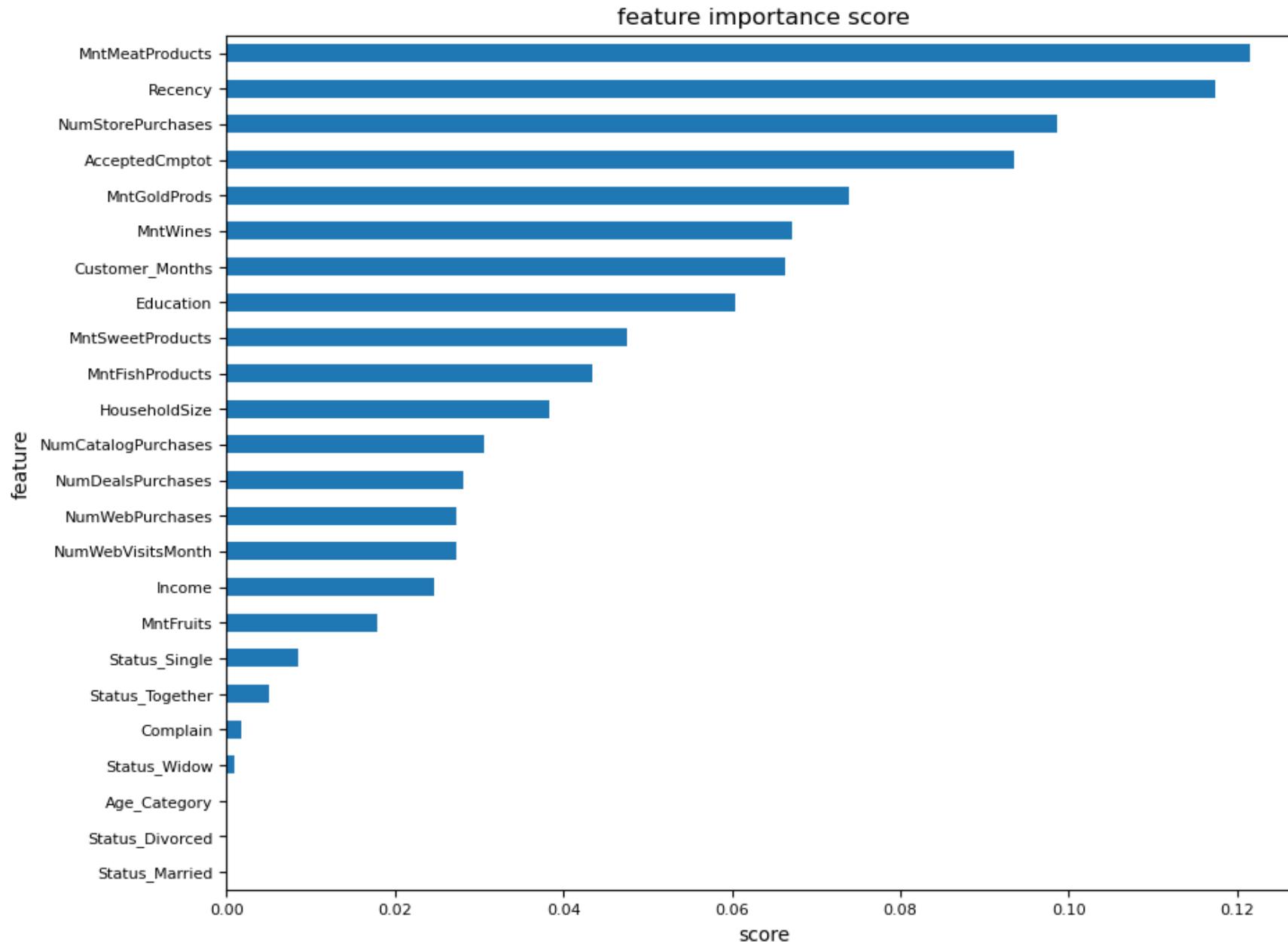
### Learning Curve - Hyperparameter Learning Rate - AdaBoost



### Learning Curve - Hyperparameter Learning Rate - AdaBoost



# AdaBoost - Feature Importance



# Perbandingan Model Klasifikasi

No	Model	Train			Test		
		AUC	Precision	Recall	AUC	Precision	Recall
1	Logistic Regression	0.91	0.74	0.48	0.86	0.74	0.39
2	Random Forest	1.00	0.98	0.98	0.87	0.68	0.29
3	Adaboost	0.96	0.79	0.64	0.86	0.67	0.48

Setelah hyperparameter tuning							
No	Model	Train			Test		
		AUC	Precision	Recall	AUC	Precision	Recall
1	Logistic Regression	0.91	0.75	0.42	0.85	0.80	0.37
2	Random Forest	0.97	0.96	0.36	0.87	0.71	0.13
3	Adaboost	0.96	0.85	0.61	0.87	0.75	0.42

# Perbandingan Model Klasifikasi - Data SMOTE

No	Model	Train			Test		
		AUC	Precision	Recall	AUC	Precision	Recall
1	Logistic Regression	0.95	0.89	0.85	0.85	0.46	0.56
2	Random Forest	1	1	1	0.87	0.56	0.51
3	Adaboost	0.86	0.92	0.92	0.98	0.51	0.58

Setelah hyperparameter tuning							
No	Model	Train			Test		
		AUC	Precision	Recall	AUC	Precision	Recall
1	Logistic Regression	0.94	0.86	0.85	0.86	0.44	0.64
2	Random Forest	1.00	0.99	0.99	0.87	0.55	0.51
3	Adaboost	0.99	0.93	0.93	0.87	0.56	0.57

# Kesimpulan

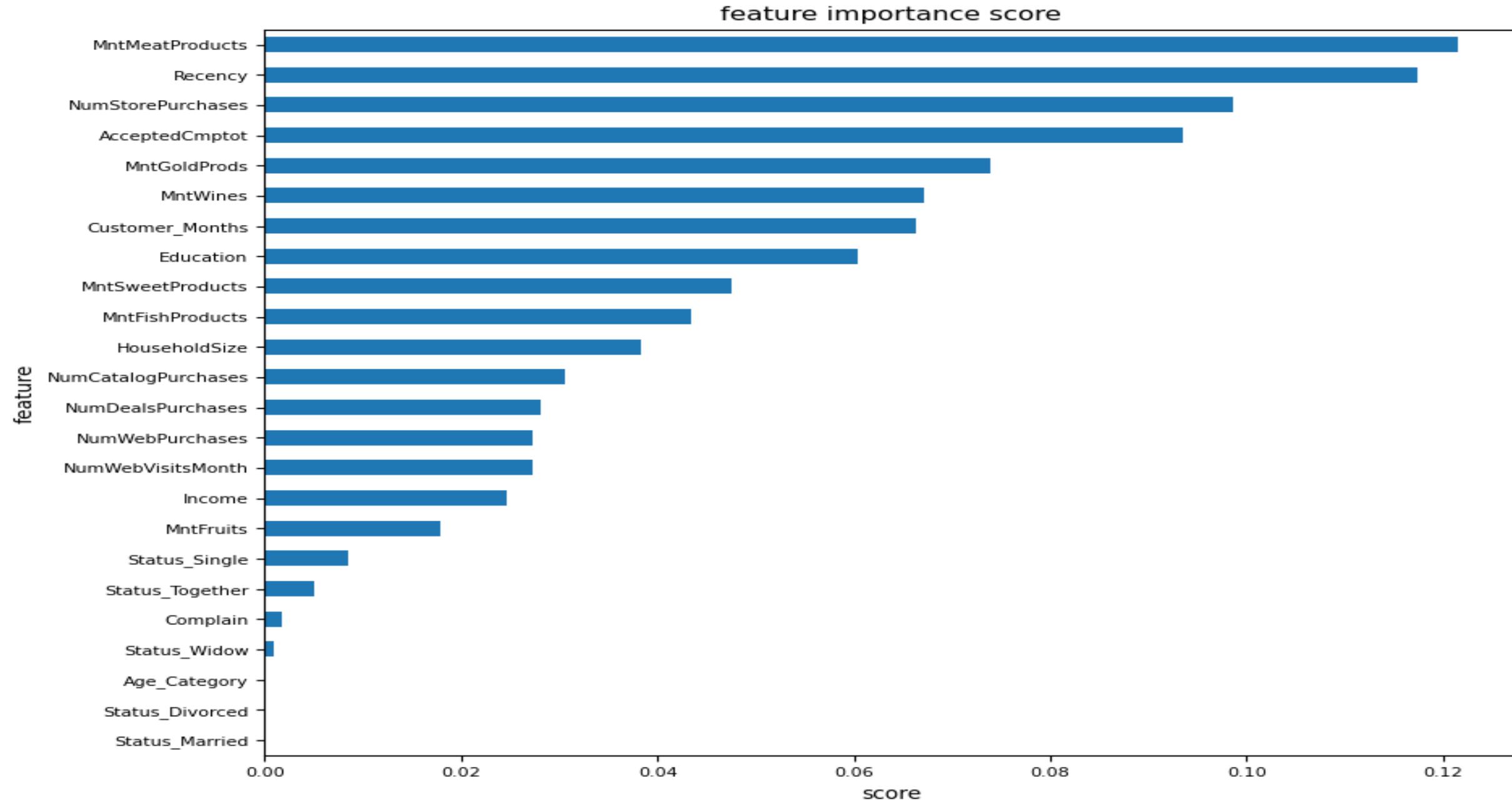
## Pada data tanpa SMOTE:

1. AdaBoost memberikan performa terbaik dibandingkan logistic regression dan random forest. Meskipun random forest memberikan precision yang paling tinggi, tetapi cenderung underfitting ketika diterapkan pada data testing. Logistic regression juga memiliki performa yang hampir sama dengan AdaBoost tetapi kurang unggul dalam hal Recall dan memiliki kecenderungan overfitting.
2. Hyperparameter tuning mampu meningkatkan performa precision pada AdaBoost sedangkan pada logistic regression dan random forest tuning parameter menyebabkan performa recall menurun.

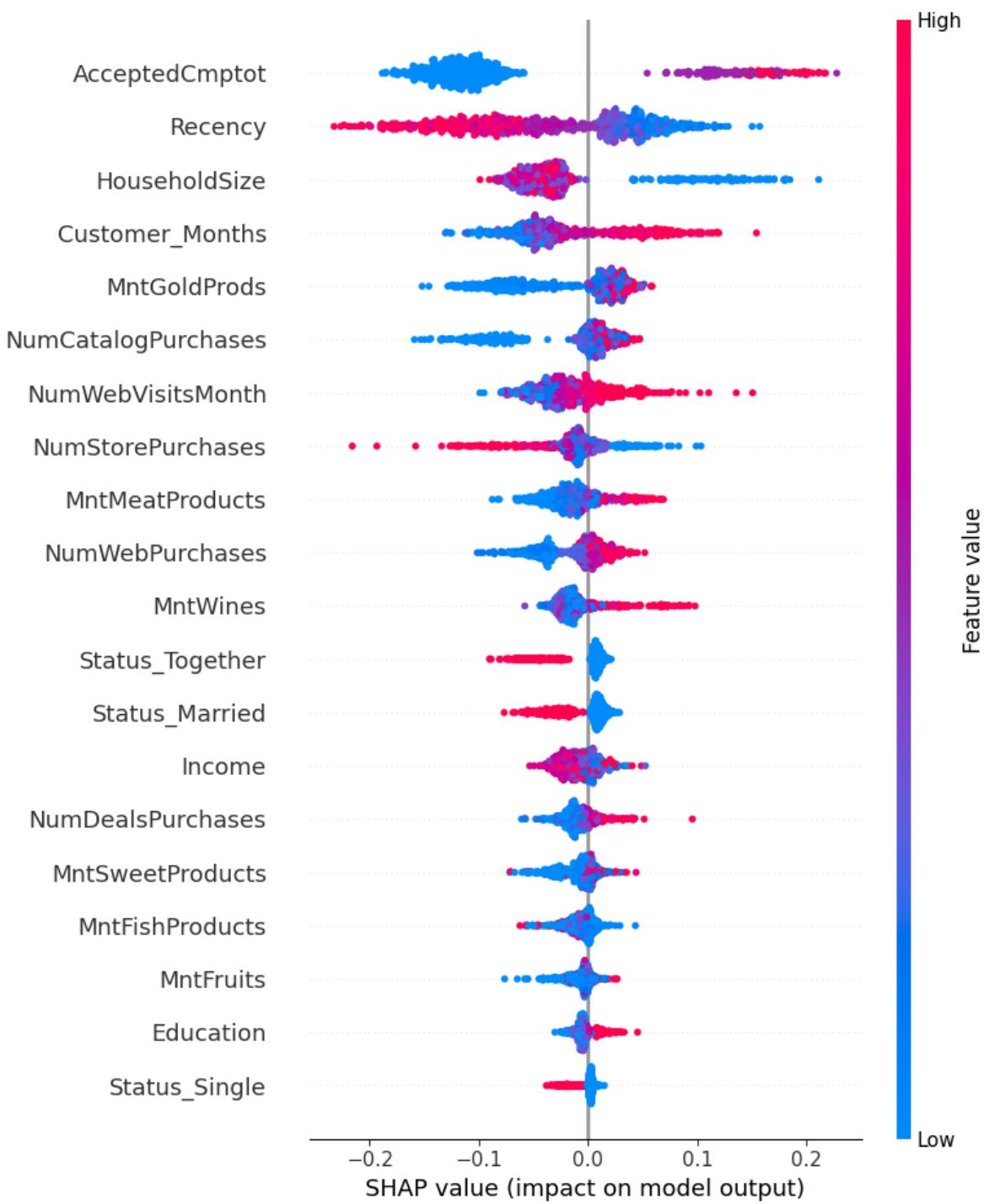
## Pada data hasil SMOTE:

1. Penerapan SMOTE pada ketiga metode menyebabkan nilai AUC menurun. Namun meningkatkan nilai precision dan recall pada data testing.
2. Metode terbaik yaitu AdaBoost berdasarkan nilai AUC dan precision pada data train dan test.
3. Hyperparameter tuning, kurang memberikan performa yang signifikan. Namun memberikan peningkatan dalam ukuran recall pada data testing.

# Feature Importance dengan Best Method (AdaBoost)



# Shap Value - Random Forest



# Business Insight

1. Variabel recency dan customer months memberikan kontribusi yang cukup besar terhadap model yang menunjukkan bahwa **semakin lama customer join ke company maka peluang response terhadap campaign juga semakin besar**. Sementara itu pada variabel recency, menunjukkan bahwa **semakin kecil nilai recency maka peluang untuk response untuk juga semakin besar** (dapat dilihat dari tanda koefisien estimator model logistic regression). Total accepted campaign memiliki pengaruh yang besar di dalam model, dimana hal ini sejalan dengan visualisasi eda yang menunjukkan bahwa **semakin besar customer dikenai campaign, maka peluang untuk respon juga semakin besar**.
2. **Daging, gold, dan wines menjadi tiga top produk yang paling sering dikonsumsi customer** yang memberikan kontribusi tinggi terhadap response campaign. **Customer yang memiliki konsumsi daging cukup tinggi memiliki peluang akan response.**
3. **Customer yang mengunjungi toko offline lebih banyak** cenderung **berpeluang untuk merespons.**
4. Faktor demografi customer seperti **education memberikan kontribusi terbesar** pelanggan dalam merespon campaign, diikuti oleh faktor jumlah anggota keluarga, dan income. Sementara itu, **status pernikahan kurang memberikan pengaruh di dalam model.**
5. MntMeatProducts, Recency, dan NumStoresPurchases merupakan Top 3 Feature.
6. Complain kurang memberikan pengaruh bagi customer untuk merespon campaign.

# Business Suggestions

1. Membuat **program keanggotaan loyalitas** yang memberikan insentif dan manfaat khusus kepada pelanggan berdasarkan variabel resensi. Semakin lama pelanggan bergabung dan aktif dalam program, semakin tinggi peluang mereka mendapatkan insentif / penawaran khusus. Memberikan perhatian khusus pada pelanggan yang telah lama bergabung untuk memastikan bahwa mereka merespon campaign dan tetap berkomunikasi dengan mereka melalui strategi pemasaran yang relevan
2. **Promosi produk daging dan wine, memberikan strategi penawaran menarik dan spesifik**, diskon khusus atau paket promo kepada pelanggan yg memiliki jumlah konsumsi daging dan wine yang tinggi.
3. Berikan program referensi : **berikan insentif kepada pelanggan yg berhasil mengajak teman/keluarganya** berbelanja. Pemberian insentif dapat memberikan respon dan mendapatkan pelanggan baru
4. Pelanggan yang sering mengunjungi toko offline memiliki peluang tinggi untuk merespon campaign. Hal ini menunjukkan bahwa interaksi langsung dengan produk dan layanan dapat mempengaruhi tingkat respons pelanggan. Oleh karena itu, penting untuk **memberikan pengalaman yang menyenangkan saat mereka mengunjungi toko fisik**. Ini dapat mencakup aspek seperti kebersihan toko, ketersediaan produk, layanan pelanggan yang baik, dan promosi khusus di toko.
5. Memperkuat strategi pemasaran untuk mendorong pelanggan untuk mengunjungi toko offline secara lebih sering, seperti dengan **menawarkan promosi eksklusif di toko atau mengadakan acara khusus yang menarik bagi pelanggan**.

# UPLOAD GITHUB

<https://github.com/Dindagaluhg/DS-32-FINALPROJECT-KEL5>

# TERIMA KASIH