

SciData Solutions

Dokumen Laporan Final Project

(dipresentasikan setiap sesi mentoring)



Stage 0

- Problem statement

- Marketing campaign yang dilakukan sebuah perusahaan tidak efisien dengan conversion rate 15,03%.

- Role

PT SciData Solution sebagai tim data scientist yang bertanggung jawab untuk memberikan solusi dari performa kampanye marketing.

- a. Nizar : Lead Data Science
- b. Alvin : Machine Learning Engineer
- c. Fahad : Data Engineer
- d. Afriansyah : Machine Learning Engineer
- e. Tsaniya : BI Analytics
- f. Rizky : Data Engineer
- g. Vida : Data Analyst
- h. Dean : Machine Learning Engineer

Stage 0

- Goal
 - Meningkatkan efisiensi marketing respon pelanggan
 - Memperkirakan tingkat respon pelanggan melalui faktor faktor yang berpengaruh dan tingkat efektivitas kampanye pemasaran online.
- Objective
 - Menghasilkan model yang dapat memprediksi customer yang memberi respon
- Metric
 - conversion rate

Stage 1

EDA, Insight & Visualization

DESCRIPTIVE STATISTICS

```
df.head()
```

```
df.info()
```

```
df.describe().T
```

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

```
df.nunique()
```

```
df.duplicated().sum()
```

dari semua metode yang digunakan menghasilkan beberapa informasi seperti :

- Type data didominasi oleh data numeric
- Terdapat nilai null pada kolom Income
- Pada nilai unique data kolom ID hanya menunjukkan nomor identitas pada setiap customer terlihat dari jumlah unique nya sama dengan jumlah row yang ada
- Terdapat data yang aneh pada kolom income dimana nilai maximum mencapai ~600,000
- Pada kolom year_birth terdapat keanehan pada nilai minimum yang ekstrem
- Terdapat beberapa kolom yang familiar yaitu pada kolom dengan nama Mnt, Num, Accepted.
- Kolom Z_ mempunyai indikasi konstanta karena tidak ada perubahan pada setiap deskripsi statistiknya
- Kolom Accepted memiliki type numeric tetapi hanya memiliki nilai 0 dan 1 sehingga merupakan data categorical binary

Stage 1

EDA, Insight & Visualization

DESCRIPTIVE STATISTICS

Homework Questions

1. Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?
 - Dt_customer tipe datanya kurang sesuai
2. Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?
 - kolom income memiliki null value
3. Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)
 - terdapat beberapa kolom yang memiliki type numeric tetapi merupakan tipe binary yaitu kolom Accepted, complain, dan response terlihat dari nilai quartiles yang 0
 - kolom Z_ memiliki nilai konstant dilihat dari min, max, dan quartiles yang memiliki nilai sama

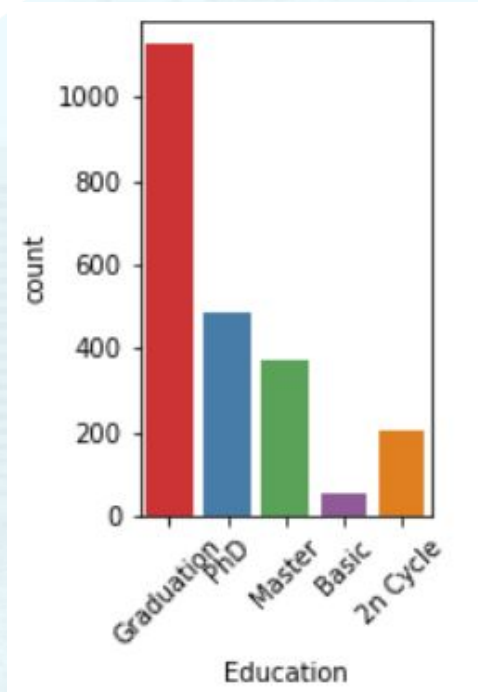
Stage 1

EDA, Insight & Visualization

UNIVARIATE ANALYSIS

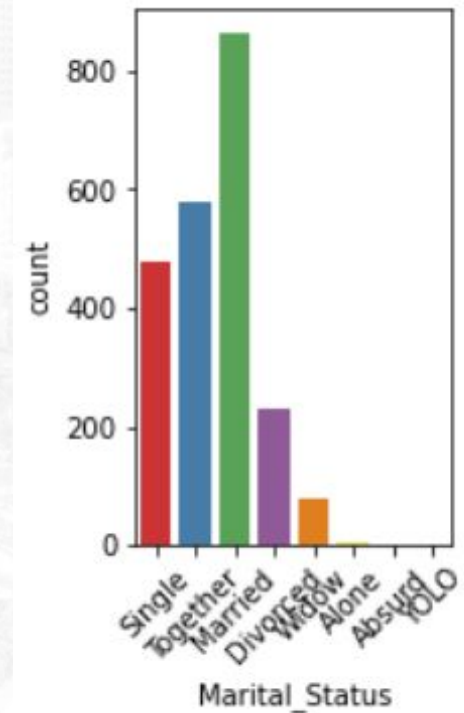
1. Education

- Kategori pendidikan terdiri dari beberapa level termasuk 'Graduation', 'PhD', 'Master', '2n Cycle', 'Basic', dan 'Unknown'.
- Kategori 'Graduation' mendominasi di antara kategori pendidikan lainnya.



2. Marital_Status

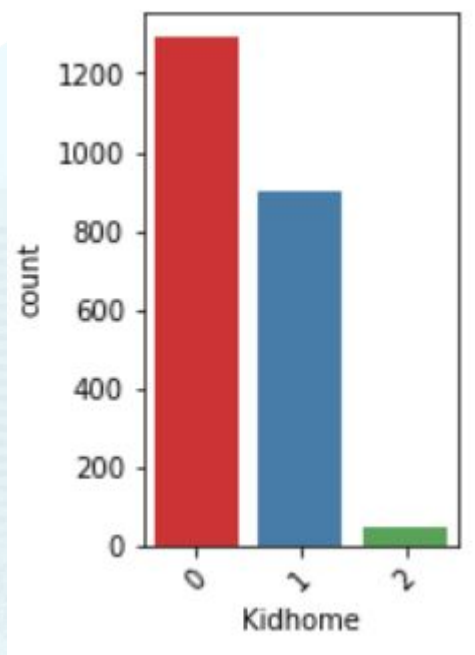
- Kategori status pernikahan terdiri dari beberapa jenis diantaranya 'Married', 'Single', 'Divorced', 'Together', 'Widow', dan 'Alone'.
- Kategori 'Married' dan 'Together' adalah yang paling umum.



Stage 1

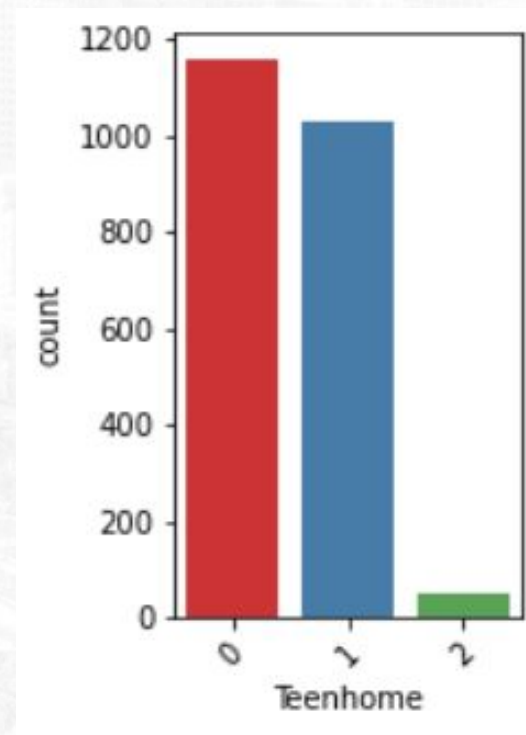
EDA, Insight & Visualization

UNIVARIATE ANALYSIS



3. Kidhome

- Pada kolom ini menunjukkan jumlah anak di dalam sebuah keluarga.
- Mayoritas suatu keluarga dalam dataset tidak memiliki anak (nilai 0).



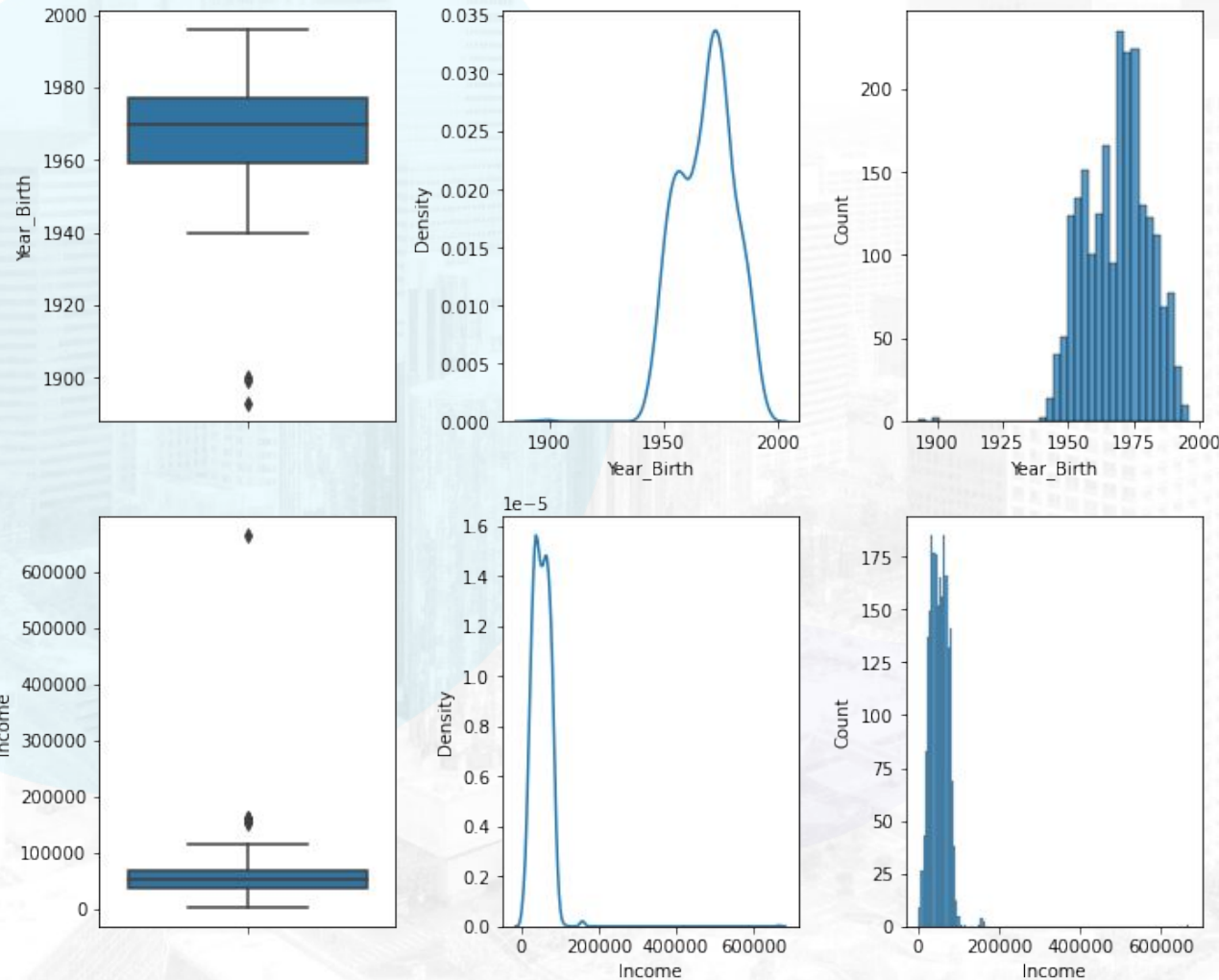
4. Teenhome

- Kolom ini menunjukkan jumlah remaja yang ada dalam suatu keluarga.
- Mayoritas suatu keluarga dalam dataset tidak memiliki remaja (nilai 0).

Stage 1

EDA, Insight & Visualization

UNIVARIATE ANALYSIS



1. Year_Birth

- Distribusi tahun kelahiran terlihat cukup normal dengan sedikit kemiringan negatif (skewed).
- Distribusi yang menjadi puncak di sekitar tahun 1970-1980

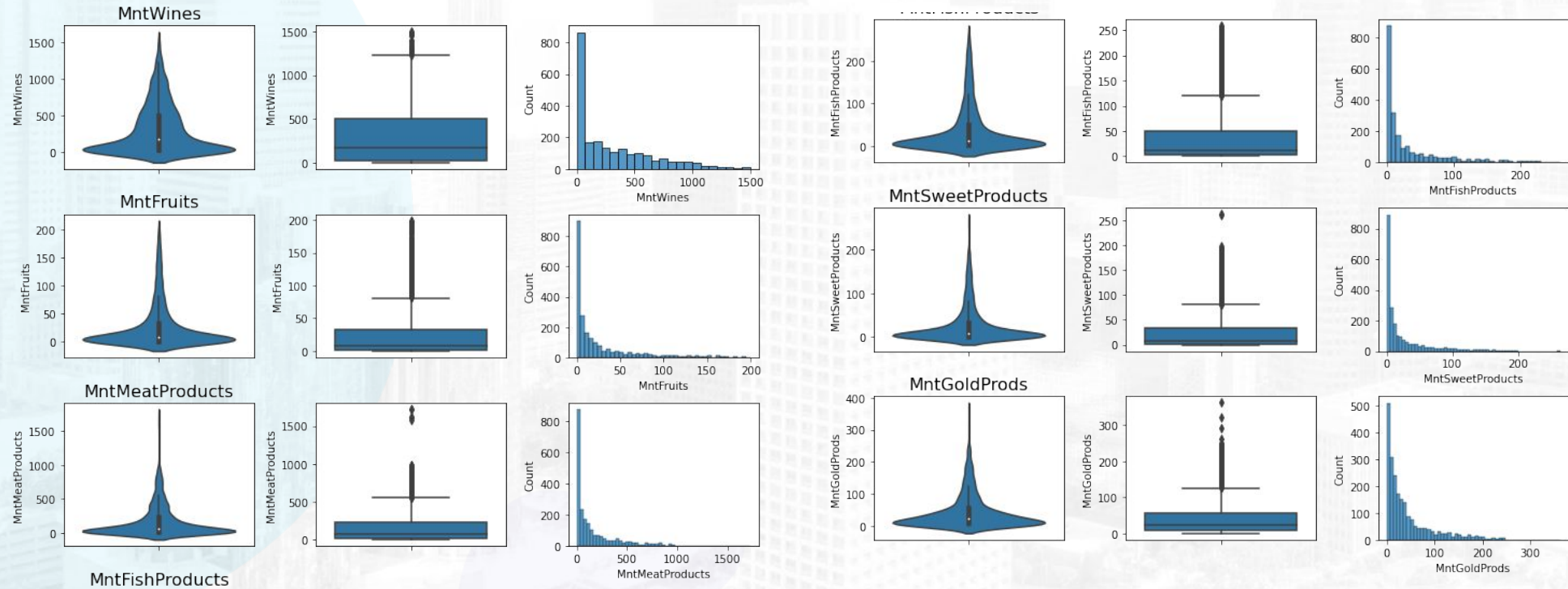
2. Income

- Distribusi pendapatan terlihat cenderung condong ke kanan (positively skewed).
- Terdapat beberapa outlier di sebelah kanan yang menunjukkan adanya pendapatan yang sangat tinggi.

Stage 1

EDA, Insight & Visualization

UNIVARIATE ANALYSIS

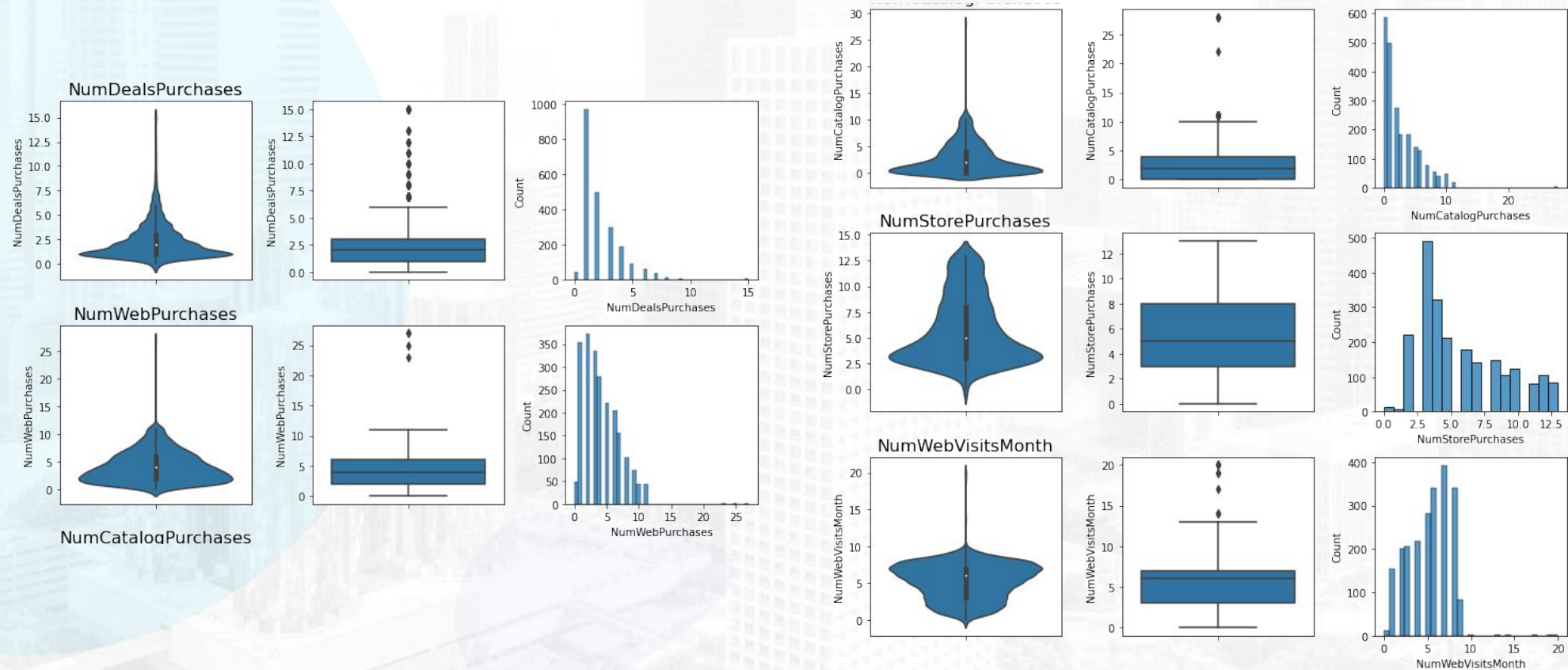


pada kelompok **kolom Mnt** dengan data visualisasi disamping memiliki distribusi yang familiar dengan positive skewed

Stage 1

EDA, Insight & Visualization

UNIVARIATE ANALYSIS

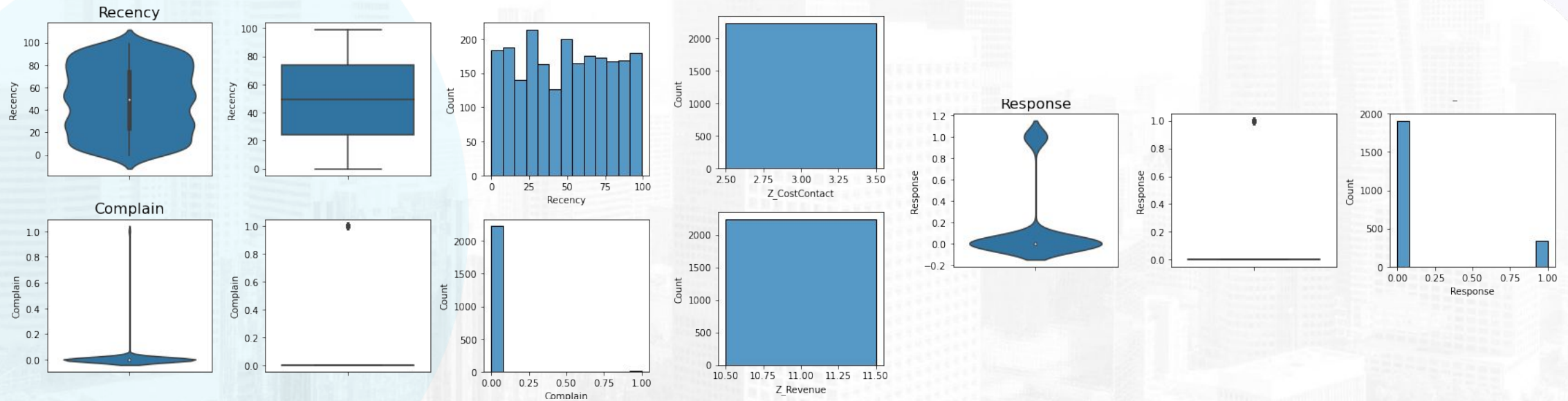


Pada kelompok **kolom Num** memiliki karakter distribusi yang variatif, sehingga dapat diketahui aktivitas pada bisnis berbeda-beda

Stage 1

EDA, Insight & Visualization

UNIVARIATE ANALYSIS



1. Recency

- Distribusi Recency jumlah hari sejak pembelian terlihat cukup normal dengan sedikit "kemiringan positif (skewed)".
- Tidak ada nilai yang mendominasi atau outlier yang mencolok pada Recency.

2. Complain

Pada kolom Complain dapat diketahui customer cenderung tidak memberikan complain

3. Z_ column

Kolom Z_ dapat terlihat dengan jelas jika kolom tersebut merupakan konstanta

4. Response

Kolom Response memiliki ketidak seimbangan pada nilainya, yang cenderung tidak memberika response(0).

Stage 1

EDA, Insight & Visualization

UNIVARIATE ANALYSIS

Tindakan pada saat pre-processing

Berikut dibawah ini adalah beberapa tindakan yang dapat diambil saat melakukan data preprocessing:

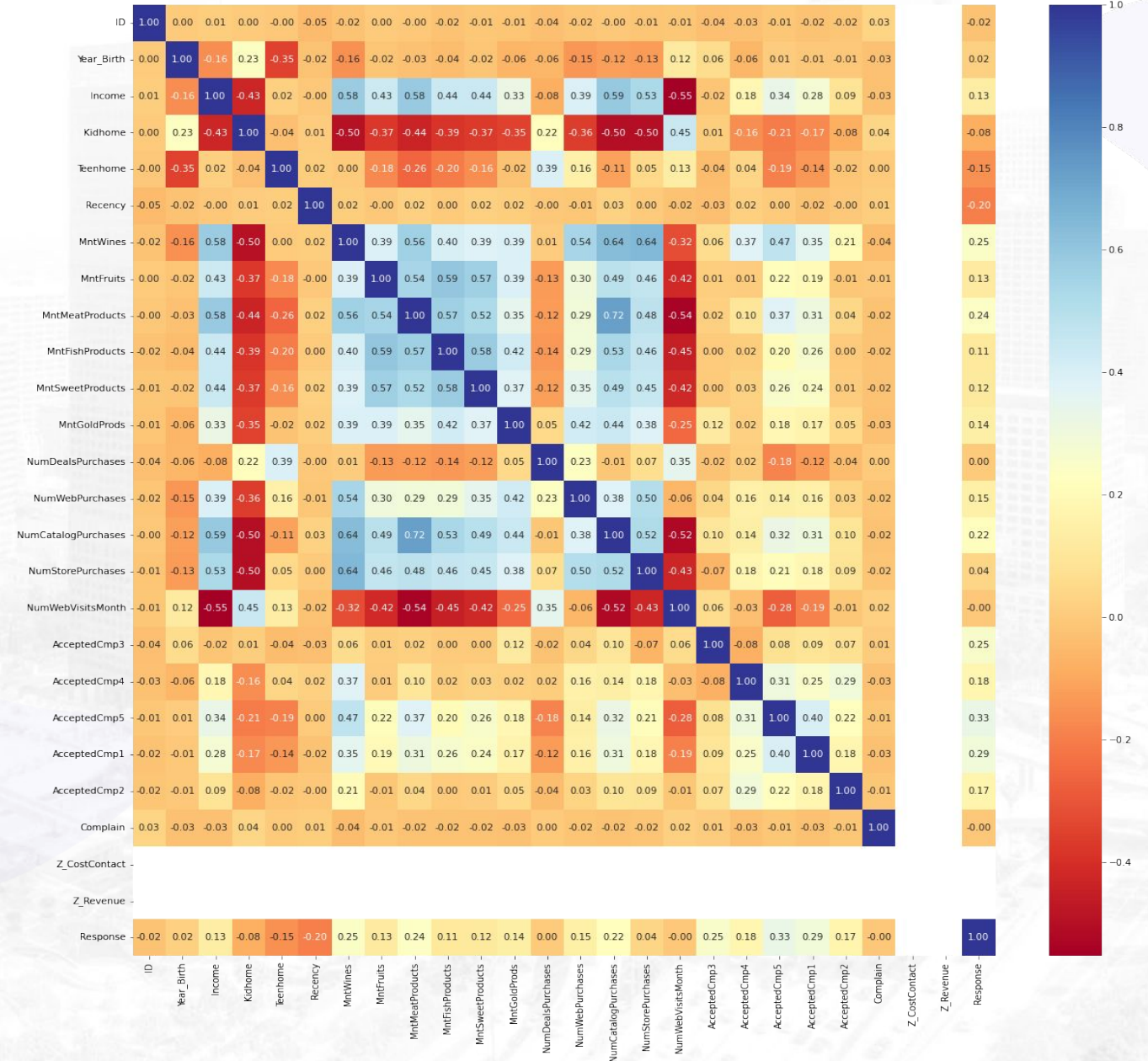
- Untuk kolom 'Education', dapat dilakukan pengelompokan kategori baru yang spesifik ke dalam kategori yang lebih umum untuk mengurangi jumlah kategori yang terlalu banyak.
- Pada kolom 'Income' dan 'MntWines', perlu dilakukan penanganan outlier dan normalisasi data agar distribusi menjadi lebih normal.
- Kolom 'Z_CostContact' dan 'Z_Revenue' dapat dihapus karena tidak terlalu memberikan informasi yang penting dalam analisis.
- Untuk kolom dengan kategori yang tidak seimbang seperti 'Marital_Status' dan 'Response', dapat dilakukan teknik pengelompokan atau pengurangan kategori agar distribusi menjadi lebih seimbang.
- Perlu dilakukan pengecekan dan penanganan terhadap missing values, duplikasi data, dan penyimpangan data lainnya yang mungkin ada dalam dataset.

Stage 1

EDA, Insight & Visualization

MULTIVARIATE ANALYSIS

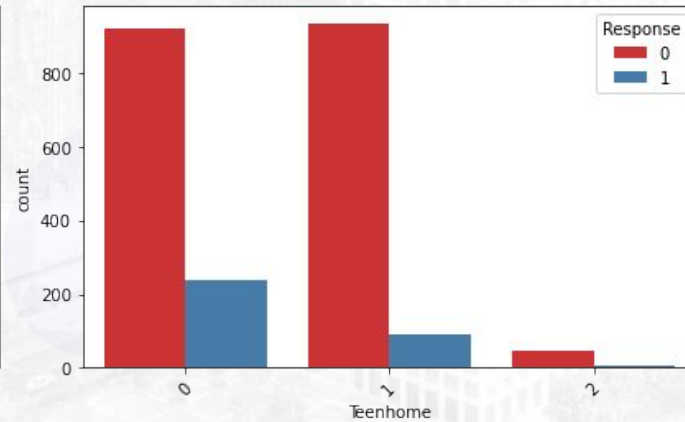
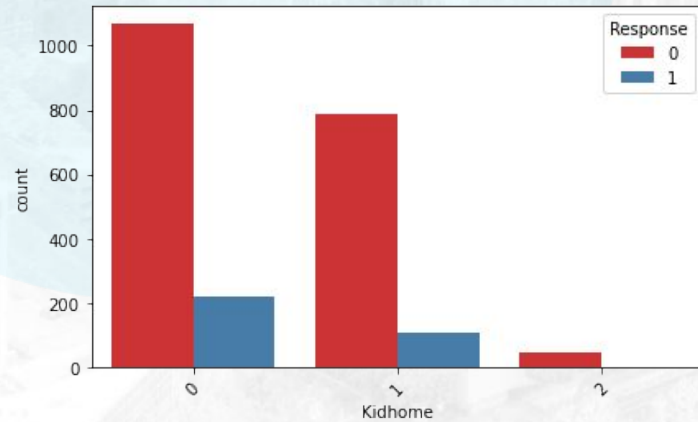
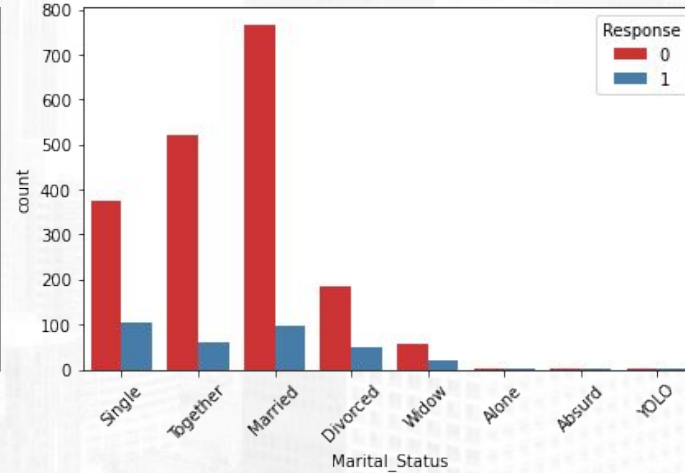
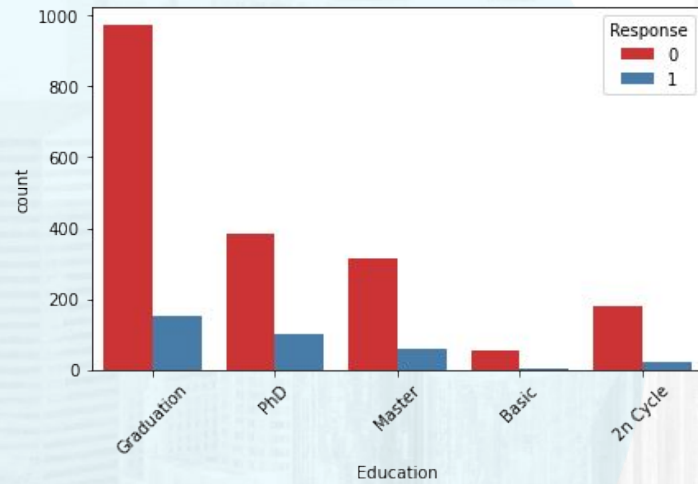
- Dari Korelasi heatmap didapatkan insight dalam kolom Z_CostContact dan Z_Revenue tidak berpengaruh dalam fitur. dan terdapat korelasi yang lebih dari 0,7 yaitu korelasi antara kolom MntMeatProducts dan NumCatalogPurchases.
- Korelasi terhadap Response mayoritas memiliki korelasi rendah



Stage 1

EDA, Insight & Visualization

MULTIVARIATE ANALYSIS



Perbandingan respon dengan sample kolom

...	Education	0	1	percent(1)
0	Graduation	975	152	6.79
1	PhD	385	101	4.51
2	Master	313	57	2.54
4	2n Cycle	181	22	0.98
3	Basic	52	2	0.09

	Marital_Status	0	1	percent(1)
0	Single	374	106	4.73
2	Married	766	98	4.38
1	Together	520	60	2.68
3	Divorced	184	48	2.14
4	Widow	58	19	0.85
5	Alone	2	1	0.04
6	Absurd	1	1	0.04
7	YOLO	1	1	0.04

	Kidhome	0	1	percent(1)
0	0	1071	222	9.91
1	1	789	110	4.91
2	2	46	2	0.09

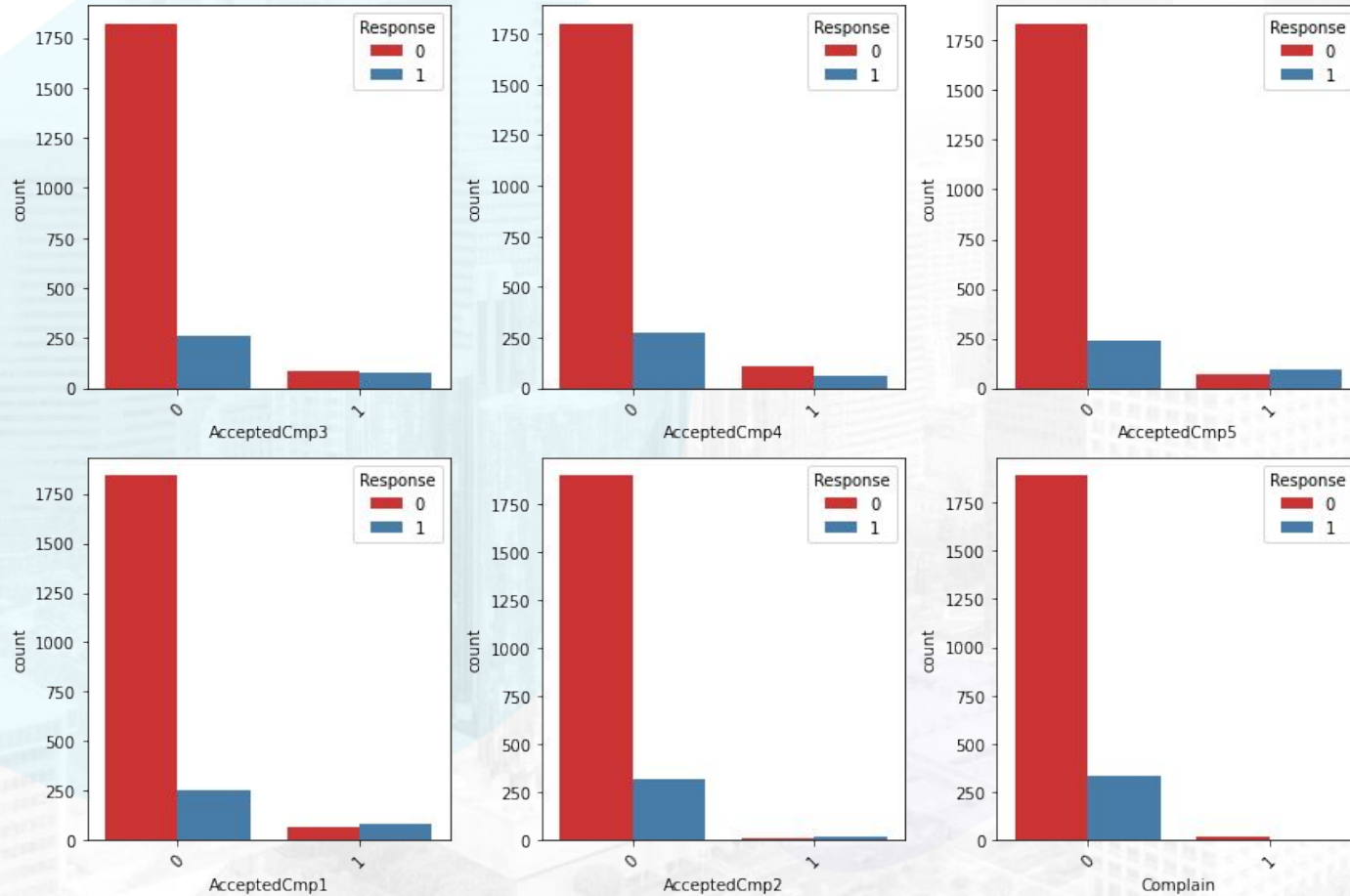
	Teenhome	0	1	percent(1)
0	0	921	237	10.58
1	1	938	92	4.11
2	2	47	5	0.22

Tabel persentase seluruh data Respon

Stage 1

EDA, Insight & Visualization

MULTIVARIATE ANALYSIS



Categorical multivariate chart and data comment

1. AcceptedCmp columns

Pada setiap campaign memiliki beragam persentase beragam pada respon, campaign dengan persentase respon tertinggi dimiliki campaign2 tetapi memiliki accept terkecil diantara yang lain. Campaign lainnya yang memiliki dampak cukup signifikan > 50% yaitu pada Campaign1 dan Campaign5.

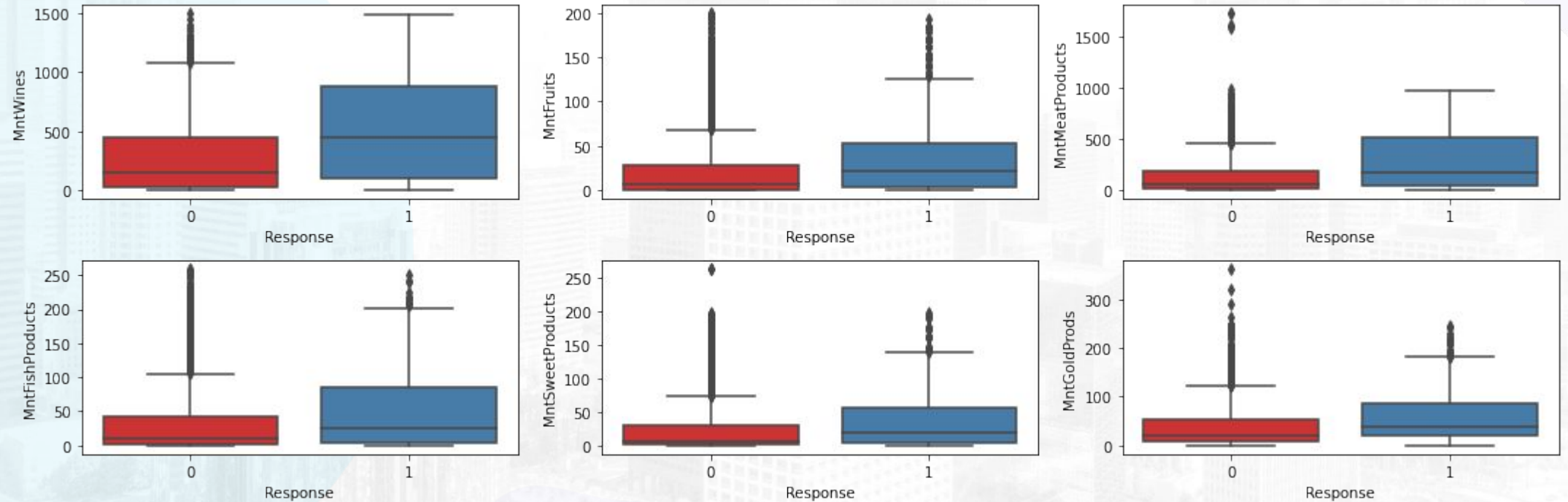
2. Complain

Pada data complain Respon terlihat jelas cenderung positif pada customer tanpa complain

Stage 1

EDA, Insight & Visualization

MULTIVARIATE ANALYSIS



Categorical multivariate chart and data comment

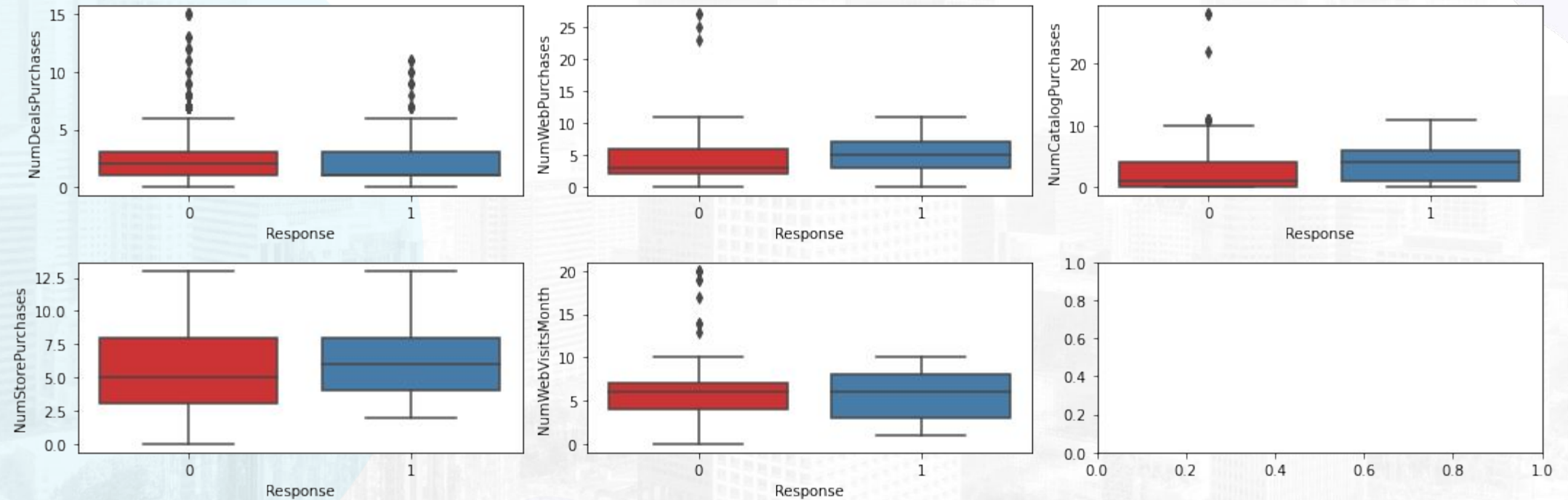
1. Mnt columns

Pada kelompok kolom ini terdapat perbedaan distribusi terhadap response dimana penerima response memiliki nilai Mnt atau spend yang lebih tinggi dari pada yang tidak merespon

Stage 1

EDA, Insight & Visualization

MULTIVARIATE ANALYSIS



Numerical multivariate chart and data comment

1. Nums columns

Pada kelompok kolom ini terdapat beberapa perbedaan pada setia kolomnya pada kolom NumWebPurchases, NumCatalogPurchases, NumWebVisitsMonth memiliki distribusi yang signifikan terhadap response, sedangkan pada kolom NumDealsPurchases, NumStorePurchases tidak ada perbedaan yang signifikan

Stage 1

EDA, Insight & Visualization

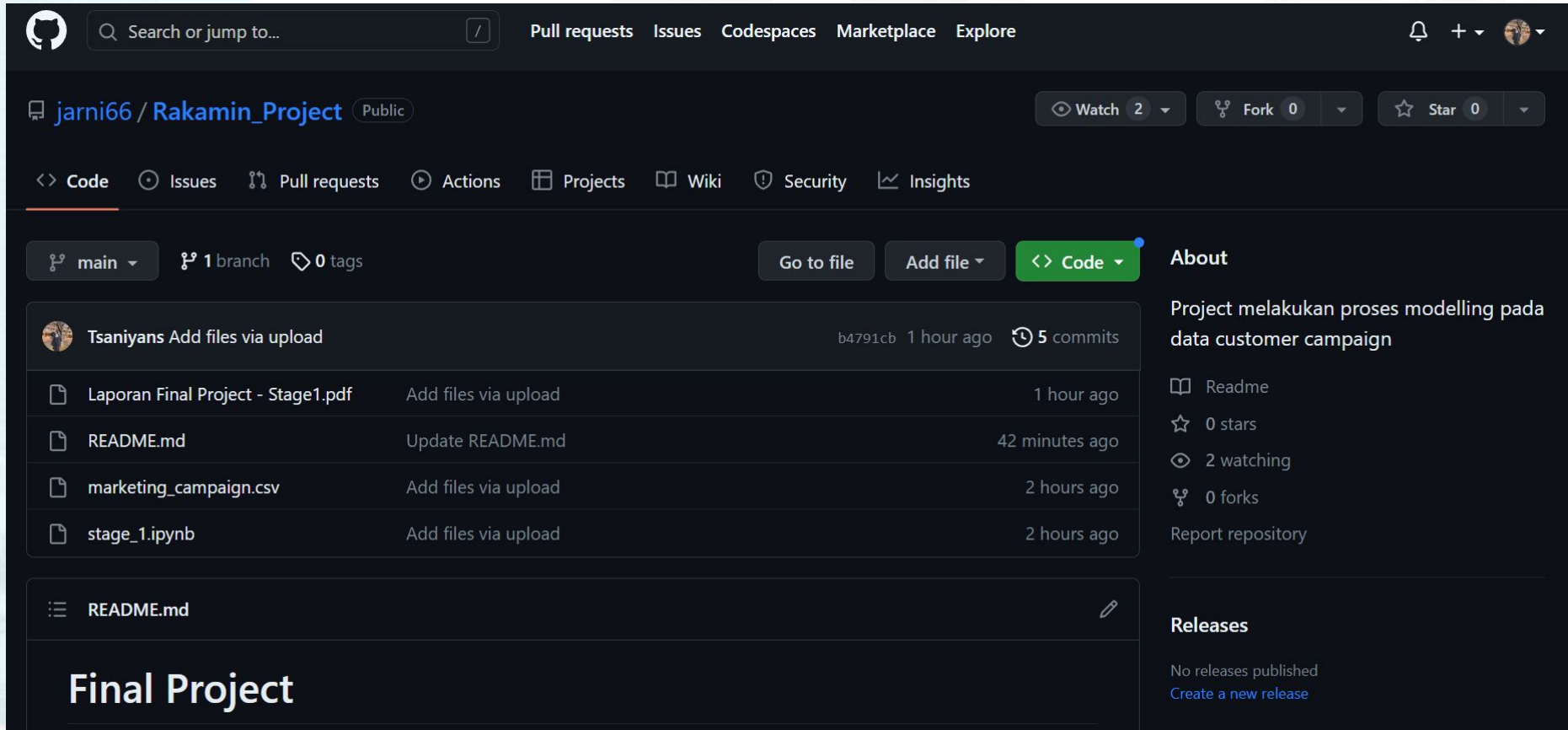
BUSINESS INSIGHT

- Pada kolom Education customer dengan background pendidikan Graduation, PhD, Master memiliki tingkat response yang tinggi sehingga dengan mentargetkan campaign pada parameter tersebut dapat meningkatkan conversion rate.
- Pada kolom Marital terdapat beberapa kelompok yang memiliki respon yang tinggi sehingga perlu disegmentasi lebih lanjut untuk mendapatkan target yang lebih terfokus untuk meningkatkan respon
- Pada kolom Kidhome dan Teenhome memiliki korelasi negatif sehingga mentargetkan customer dengan jumlah Kidhome dan Teenhome < 2 dapat meningkatkan response customer.
- Dari segi jenis campaign yang memiliki tingkat respon yang tinggi, campaign 1 dan 5 dapat menjadi opsi untuk meningkatkan conversion rate pada campaign selanjutnya
- Pada segi jumlah pembelian terdapat perbedaan yang jelas pada jenis produk Wines sehingga campaign pada target customer yang membeli produk Wines dapat meningkatkan response campaign.
- Customer yang cenderung response memiliki aktivitas yang tinggi pada pembelian Catalog sehingga dapat menjadi alternatif campaign untuk difokuskan pada aktivitas tersebut untuk meningkatkan response

Stage 1

EDA, Insight & Visualization

GIT



The screenshot shows the GitHub interface for the repository 'jarni66 / Rakamin_Project'. The repository is public and has 2 watchers, 0 forks, and 0 stars. The main branch is 'main'. The repository contains four files: 'Laporan Final Project - Stage1.pdf', 'README.md', 'marketing_campaign.csv', and 'stage_1.ipynb'. The 'README.md' file is selected, showing the title 'Final Project'.

Repository Details:

- Owner: jarni66
- Repository Name: Rakamin_Project
- Visibility: Public
- Watch: 2
- Fork: 0
- Star: 0

Files:

File Name	Action	Time
Laporan Final Project - Stage1.pdf	Add files via upload	1 hour ago
README.md	Update README.md	42 minutes ago
marketing_campaign.csv	Add files via upload	2 hours ago
stage_1.ipynb	Add files via upload	2 hours ago

About:

Project melakukan proses modelling pada data customer campaign

- Readme
- 0 stars
- 2 watching
- 0 forks

Releases:

No releases published
[Create a new release](#)

https://github.com/jarni66/Rakamin_Project

Stage 2

Data Pre-Processing

Untuk melakukan suatu tahap Data Pre-processing, maka terdapat beberapa proses seperti :

1 *Handle Missing Value*

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

```
df.shape
```

```
df.shape
```

Dengan metode disamping ini, output yang ditampilkan untuk meng-*handle missing value* dapat dianalisis seperti adanya missing value pada kolom income dengan jumlah 24, nilai ini cenderung kecil dibandingkan dengan keseluruhan data (2240) sehingga diperkirakan tidak akan berpengaruh pada model. Dengan presentase 1% missing terhadap total data maka akan didrop.

2 *Handle Duplicate*

```
df.duplicated().sum()
```

Output menunjukkan tidak adanya nilai yang duplikat pada data

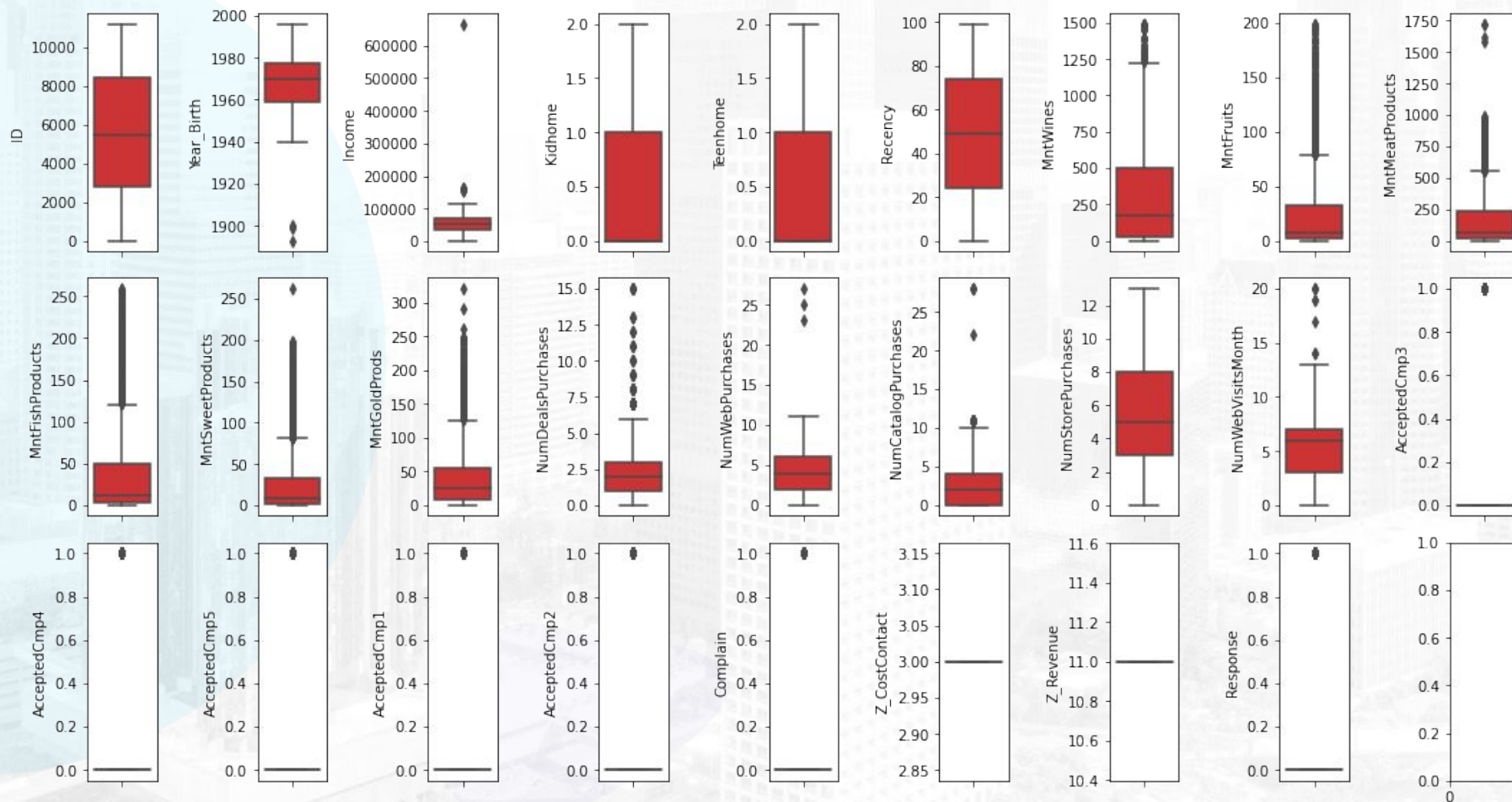
3 *Handle Outlier*

```
df.describe().T
```

selain dilihat dari deskriptif statistik dengan metode disamping, terdapat beberapa metode dengan visualisasi data sehingga dapat terlihat outliernya untuk mempermudah handling outlier.

Stage 2

Data Pre-Processing



Melihat outlier dari sisi response pelanggan

Stage 2

Data Pre-Processing

```
Presentase outlier ID: 0.0 %
Presentase outlier Year_Birth: 0.14 %
Presentase outlier Income: 0.36 %
Presentase outlier Kidhome: 0.0 %
Presentase outlier Teenhome: 0.0 %
Presentase outlier Recency: 0.0 %
Presentase outlier MntWines: 1.58 %
Presentase outlier MntFruits: 11.1 %
Presentase outlier MntMeatProducts: 7.85 %
Presentase outlier MntFishProducts: 10.02 %
Presentase outlier MntSweetProducts: 11.1 %
Presentase outlier MntGoldProds: 9.25 %
Presentase outlier NumDealsPurchases: 3.79 %
Presentase outlier NumWebPurchases: 0.14 %
Presentase outlier NumCatalogPurchases: 1.04 %
Presentase outlier NumStorePurchases: 0.0 %
Presentase outlier NumWebVisitsMonth: 0.36 %
Presentase outlier AcceptedCmp3: 7.36 %
Presentase outlier AcceptedCmp4: 7.4 %
Presentase outlier AcceptedCmp5: 7.31 %
Presentase outlier AcceptedCmp1: 6.41 %
Presentase outlier AcceptedCmp2: 1.35 %
Presentase outlier Complain: 0.95 %
Presentase outlier Z_CostContact: 0.0 %
Presentase outlier Z_Revenue: 0.0 %
Presentase outlier Response: 15.03 %
```

Dari gambar disamping, karena outliers mayoritas dibawah 1% maka kita akan menggunakan Z- Score untuk handling outliers. Kolom Response kita exclude karena merupakan target dan kolom AcceptedCmp dan Complain yang merupakan *categorical*. Berikut dibawah ini kolom *numerical* untuk *handling outlier* menggunakan Z-Score :

```
['Year_Birth', 'Income', 'MntWines', 'MntFruits', 'MntMeatProducts',
'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases',
'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases',
'NumWebVisitsMonth']
```


Stage 2

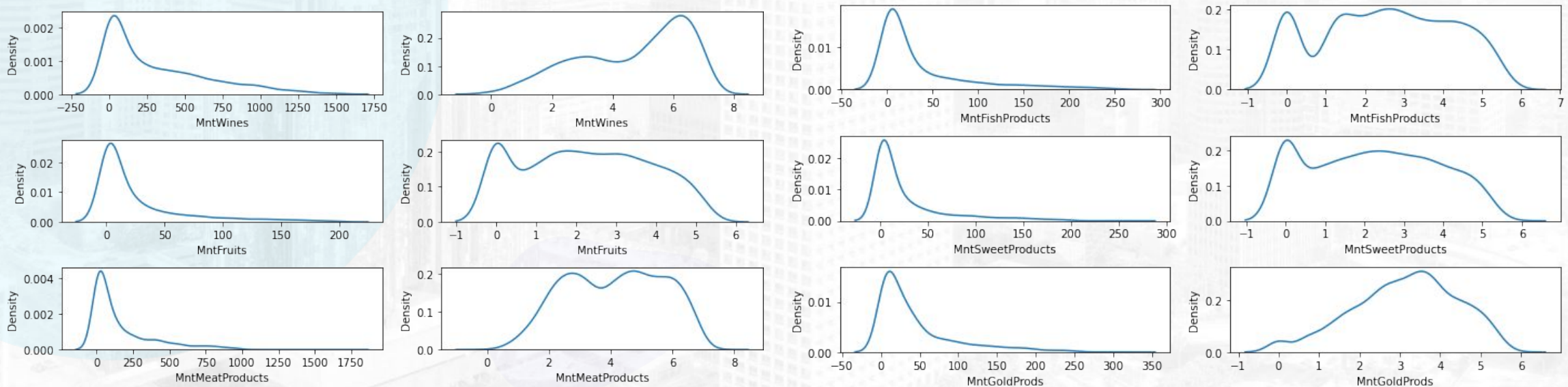
Data Pre-Processing

4 Feature Encoding

Kita lakukan label encode(LE) dan One Hot Encoding(OHE) pada Education dan OHE pada Marital_Status

5 Feature Transformation

transformasi dilakukan dengan $\log(x + 1)$, karena jika $\log(x)$ saja pada nilai values 0 akan menghasilkan nilai infinity sehingga visualisasi data sebagai berikut :



Num columns tidak kita lakukan log transformation karena cenderung tidak terlalu skewed

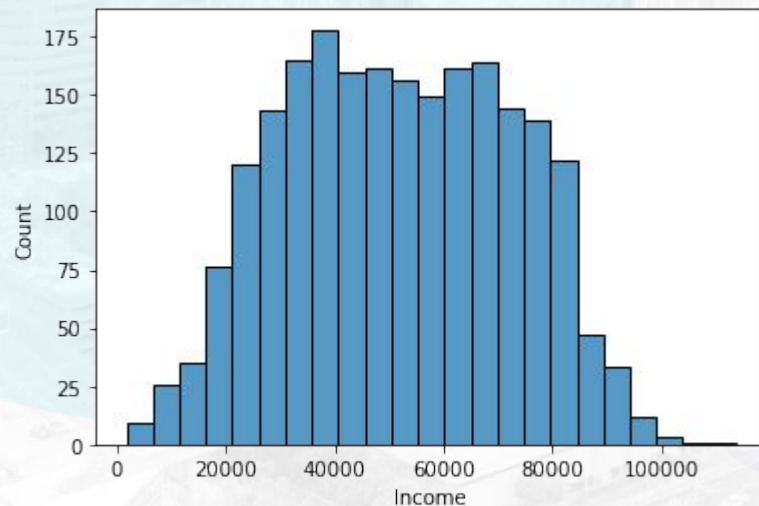
Stage 2

Feature Engineering

Kita akan mengaplikasikan Normalization kepada semua numerical kolom kecuali kolom Year_Birth, Kidhome, Teenhome, dan family Num columns yang akan kita lakukan *Feature Extraction*.

1 Feature Extraction

- Kolom Year_Birth dapat diubah menjadi kolom age untuk memudahkan pengamatan
- Kolom Kidhome dan Teenhome dapat kita transformasi menjadi kolom child karena kedua kolom ini memiliki hubungan yang sama terhadap response
- Kolom monetary dari kolom Mnt family
- Kolom frequency dari kolom Num family



dari visualisasi income disamping, kami mencoba membuat kolom baru dengan dibuat menjadi 3 kategori Low, Medium dan High

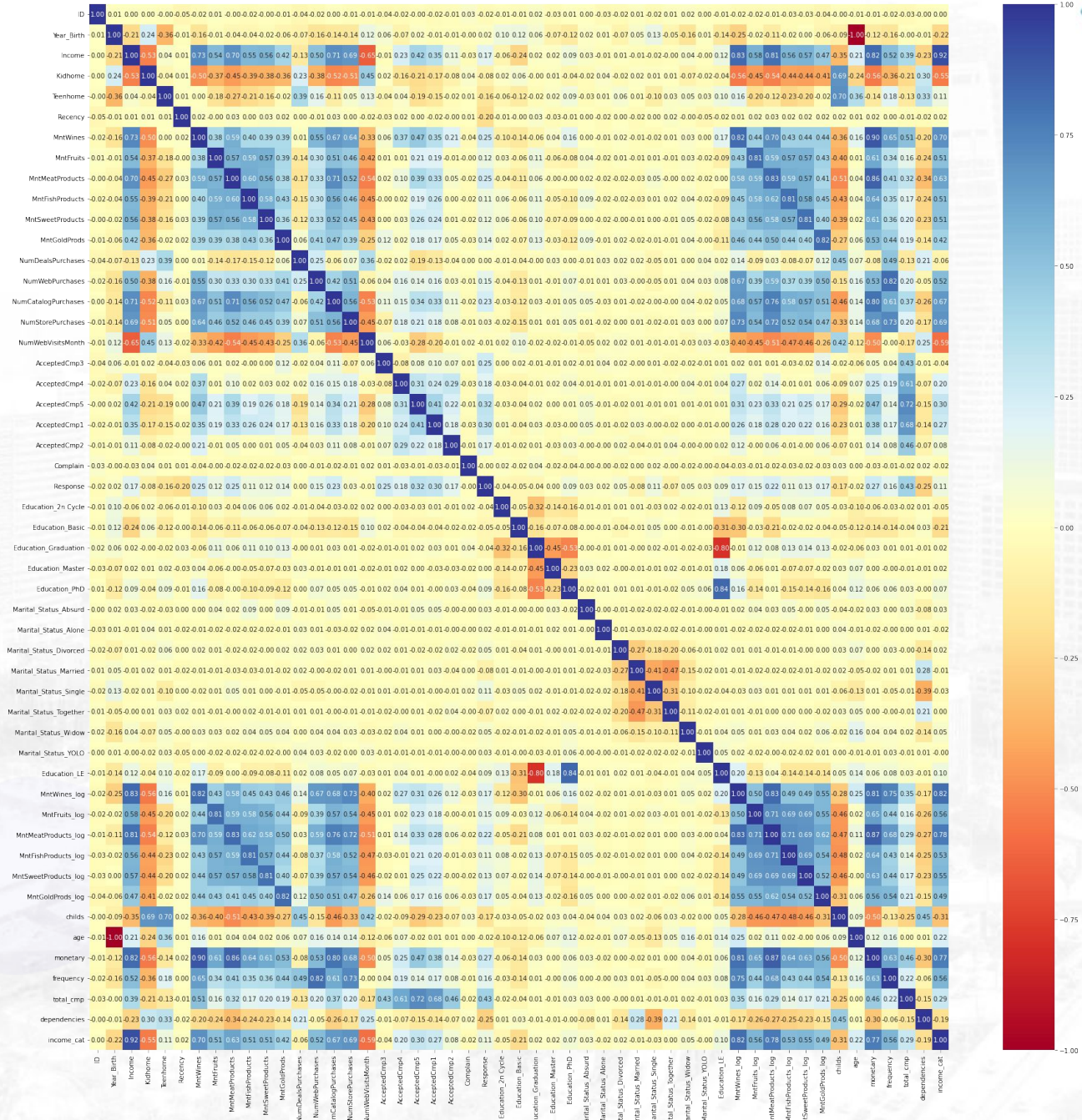
Setelah itu, untuk menentukan Feature Selection dengan melihat Correlation Check menggunakan Heatmap

Stage 2

Data Pre-Processing

Terdapat beberapa improvement pada korelasi dengan adanya feature extraction :

- Pada kolom *childs* memiliki korelasi -0.17 terhadap *response* sedangkan kolom *pembentuknya* hanya memiliki -0.15 da -0.08
- Pada kolom *monetary* memiliki korelasi 0.27 terhadap *response* sedangkan korelasi pada *famili* kolom *Mnt* memiliki korelasi tertinggi pada 0.25
- Pada kolom *frequency* memiliki korelasi 0.16 sedangkan korelasi pada *famili* kolom *Num* memiliki korelasi tertinggi pada 0.23
- Pada kolom *age* tidak terjadi perbedaan yang signifikan terhadap kolom *pembentuknya*
- Total *cmp* memiliki korelasi 0.43 terhadap *target* sedangkan *famili* kolom *AccCmp* memiliki korelasi tertinggi pada 0.32
- Kolom *income_cat* memiliki korelasi lebih rendah dari kolom *income*
- Kolom *dependencies* memiliki korelasi -0.2 lebih tinggi dari semua kolom *OHE* pada *marital status*



Stage 2

Data Pre-Processing

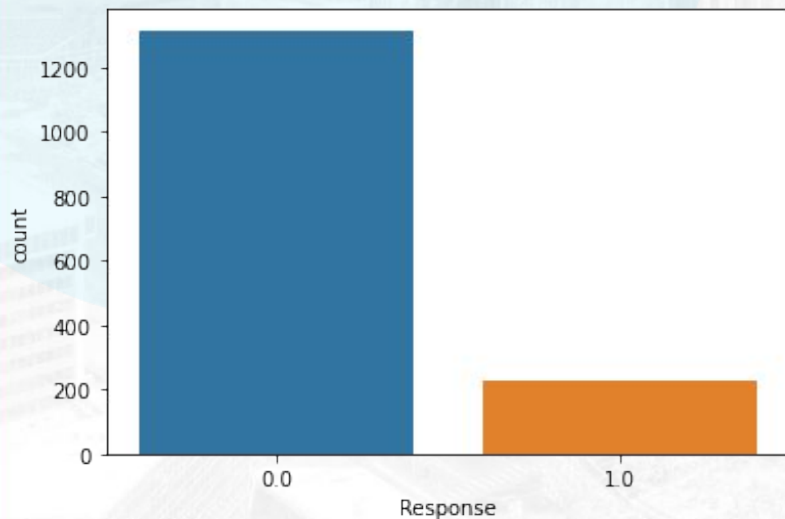
2 Feature Selection

Dari *Correlation Check* kita bisa memilih kolom mana saja yang akan menjadi *Feature Selection* sehingga terpilihlah 11 kolom

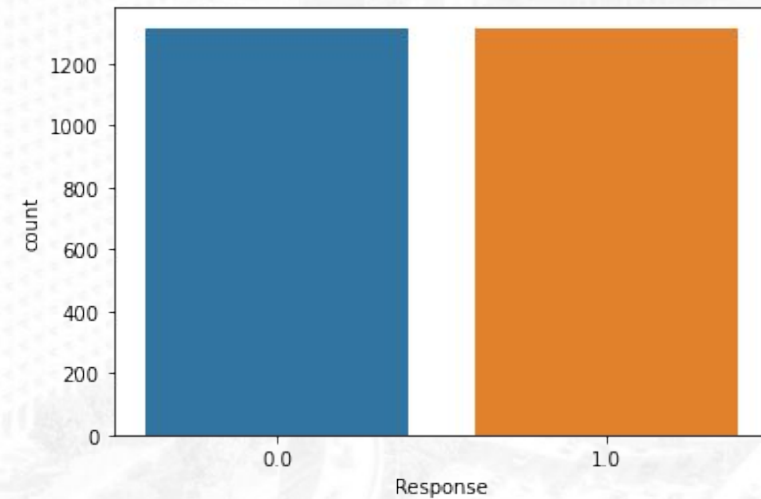
```
#picked columns
pick_col= ['age','Income','monetary','Recency','NumWebPurchases','NumCatalogPurchases','chlds','total_cmp',
           'Education_LE','dependencies','Response']
len(pick_col)
```

3 Handle Class Imbalance

Kita akan gunakan imblearn SMOTE untuk menyeimbangkan data target



Response dengan
perhitungan y_train
setelah *spliting*



Response
dengan
perhitungan
y_train
setelah
SMOTE

Stage 3

Machine Learning Modelling & Evaluation



Stage 4

Final Preparation

