

# Group 9

## Batch 28

---

### Dokumen Laporan Final Project

(dipresentasikan setiap sesi mentoring)



# Latar Belakang Masalah

Sebuah perusahaan diketahui memiliki tingkat acceptance marketing campaign sekitar 14.91%, yang mana hal tersebut masih dianggap kurang oleh manajemen dalam menghadapi persaingan bisnis.

Sehingga pihak manajemen meminta tim marketing untuk meningkatkan lagi tingkat acceptance marketing campaign tersebut agar cost yang dikeluarkan perusahaan dalam melakukan marketing campaign lebih efisien dan jumlah revenue yang mereka raih pada tahun-tahun berikutnya meningkat.

Maka dari itu, tim marketing berencana menerapkan strategi Targeted Marketing dengan bantuan tim data science untuk mengolah data historis penjualan yang telah mereka rekap sebelumnya dan mengelompokkan user ke dalam sebuah kategori tertentu sesuai dengan karakteristiknya masing-masing, sehingga dapat dipilah antara yang layak mendapatkan campaign dengan yang tidak mendapatkan campaign

# Peran

Sebagai tim data scientist, kami bertanggungjawab untuk memberikan rekomendasi untuk meningkatkan efektivitas marketing campaign perusahaan berdasarkan pola data yang tersedia.

# Goal

Meningkatkan response rate atau tingkat acceptance rate dari marketing campaign yang dilakukan oleh perusahaan, sehingga profit perusahaan serta efisiensi marketing cost dapat lebih optimal.



# Objective

Membuat sistem prediksi model klasifikasi/clustering yang dapat menentukan targeting user yang tepat. Dengan ini tentu akan memperbesar nilai business metrics yang telah ditentukan seperti traffic dan sales performance. Sistem sudah menentukan mana user yang memang sedang tertarik atau bagian dari market untuk campaign yang akan dijalankan.

## Business Metrics

- Response rate : Rasio jumlah customer yang merespon dibandingkan dengan total impresi campaign  
$$\text{Jumlah Customer yang Merespon} / \text{Total Campaign}$$
- Revenue rate : Rasio jumlah keuntungan (profit) yang diperoleh perusahaan berdasarkan total response customer  
$$[(\text{Revenue} * \text{Total Response}) - (\text{Cost} * \text{Total Campaign})] / \text{Total Revenue}$$

# Descriptive Analysis

1. Pada data terdapat tipe data tidak sesuai, yaitu kolom Dt\_Customer
2. Terdapat kolom yang memiliki Missing Value yaitu Income sebanyak 24 value

```
# Check Missing Value
```

```
df.info()
```

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

```
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)
```

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

```
[ ] # Check Duplicate
```

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

```
0
```



# Descriptive Analysis

3. Terdapat kolom dengan value yang perlu diperhatikan, yaitu kolom Marital\_Status dan Education karena memiliki kategori terlalu bervariasi, selain itu pada kolom Year\_Birth kurang valid untuk data minnya yang terlalu jauh.

▶ # Check Value for each Columns -- Check Invalid Data

```
for x in df.columns.to_list():
    print(x + '=')
    print(df[x].sort_values().unique())
    print('')
```

```
ID=
[ 0      1      9 ... 11187 11188 11191]
```

```
Year_Birth=
[1893 1899 1900 1940 1941 1943 1944 1945 1946 1947 1948 1949 1950 1951
 1952 1953 1954 1955 1956 1957 1958 1959 1960 1961 1962 1963 1964 1965
 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976 1977 1978 1979
 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993
 1994 1995 1996]
```

```
Education=
['2n Cycle' 'Basic' 'Graduation' 'Master' 'PhD']
```

```
Marital_Status=
['Absurd' 'Alone' 'Divorced' 'Married' 'Single' 'Together' 'Widow' 'YOLO']
```

```
Income=
[ 1730.   2447.   3502. ... 160803. 162397. 666666.]
```

```
[ ] #categorized each column based on num or cat data type, column that has only 1 unique value (Z_CostContact and Z_Revenue) is not categorized
cats = ["Education", "Marital_Status", "Kidhome", "Teenhome", "AcceptedCmp3", "AcceptedCmp4", "AcceptedCmp5",
        "AcceptedCmp1", "AcceptedCmp2", "Complain", "Response"]
num = ['ID', 'Year_Birth', 'Income', 'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts',
        'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases',
        'NumStorePurchases', 'NumWebVisitsMonth']
product = ['MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds']
purchase = ['NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases']
```

▶ df[num].describe()

	ID	Year_Birth	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDe
count	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	2216.000000	
mean	5588.353339	1968.820397	52247.251354	49.012635	305.091606	26.356047	166.995939	37.637635	27.028881	43.965253	
std	3249.376275	11.985554	25173.076661	28.948352	337.327920	39.793917	224.283273	54.752082	41.072046	51.815414	
min	0.000000	1893.000000	1730.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	
25%	2814.750000	1959.000000	35303.000000	24.000000	24.000000	2.000000	16.000000	3.000000	1.000000	9.000000	
50%	5458.500000	1970.000000	51381.500000	49.000000	174.500000	8.000000	68.000000	12.000000	8.000000	24.500000	
75%	8421.750000	1977.000000	68522.000000	74.000000	505.000000	33.000000	232.250000	50.000000	33.000000	56.000000	
max	11191.000000	1996.000000	666666.000000	99.000000	1493.000000	199.000000	1725.000000	259.000000	262.000000	321.000000	

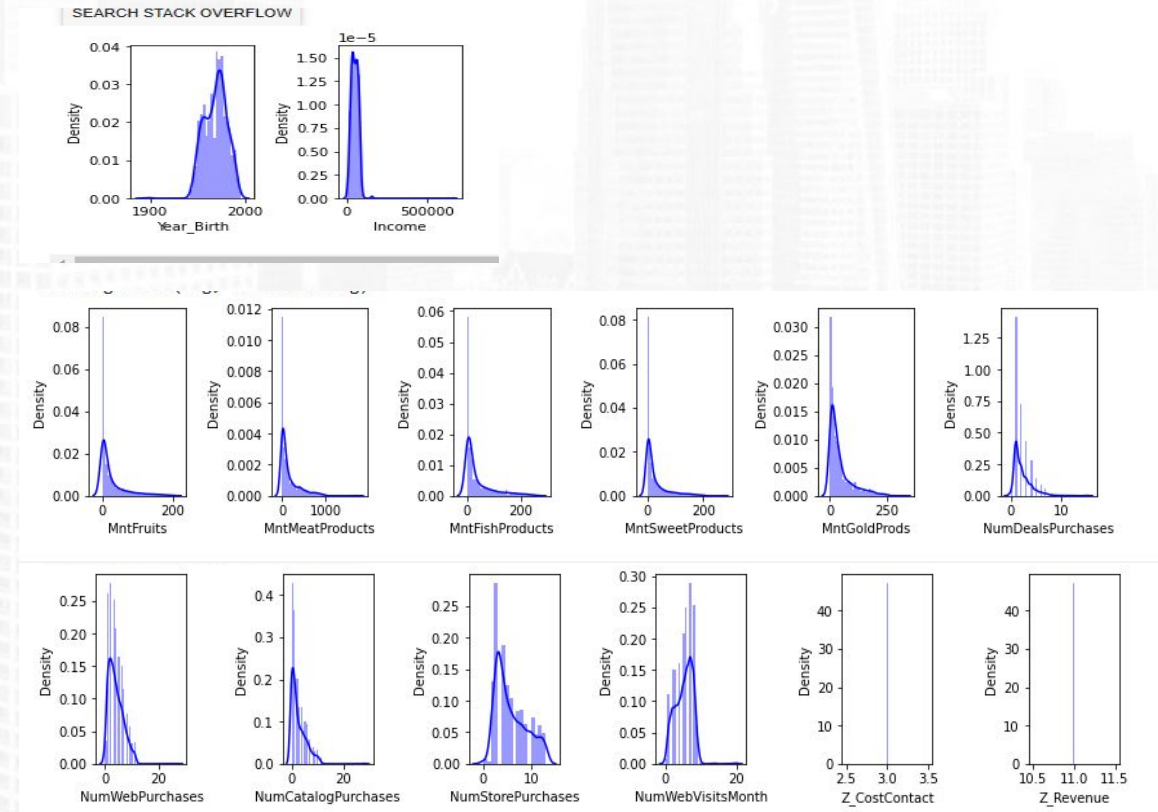
▶ df[cats].describe()

	Education	Marital_Status	Kidhome	Teenhome	AcceptedCmp3	AcceptedCmp4	AcceptedCmp5	AcceptedCmp1	AcceptedCmp2	Complain	Response
count	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216	2216
unique	5	8	3	3	2	2	2	2	2	2	2
top	Graduation	Married	no kid	no teen	didn't accept	didn't accept	didn't accept	didn't accept	didn't accept	no complain	no respond
freq	1116	857	1283	1147	2053	2052	2054	2074	2186	2195	1883

# Univariate Analysis

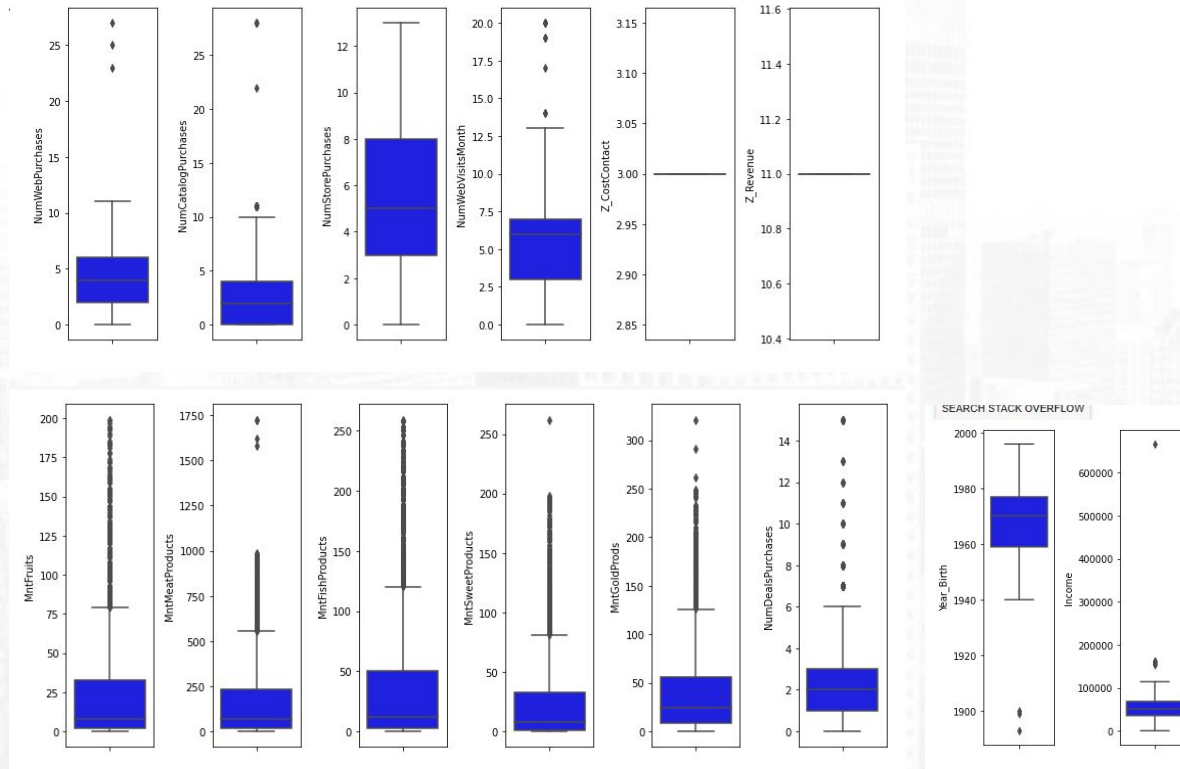
Berikut merupakan beberapa jenis distribusi data yang terdapat pada dataset:

1. Negative Skew : Year\_Birth
2. Positive Skew : Income, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases, NumWebPurchases, NumCatalogPurchases, NumStorePurchases, NumWebVisitsMonth,
3. Bimodal : Teenhome, Kidhome
4. Uniform : Recency



Untuk distribusi positive skew, di proses data pre-processing akan dilakukan log transformation. Untuk distribusi negative skew, di proses data pre-processing akan dilakukan standardization. Untuk distribusi

# Univariate Analysis

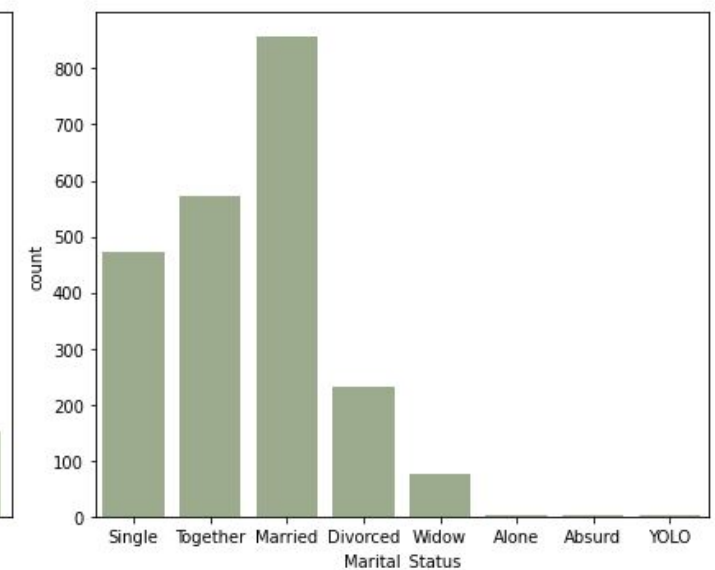
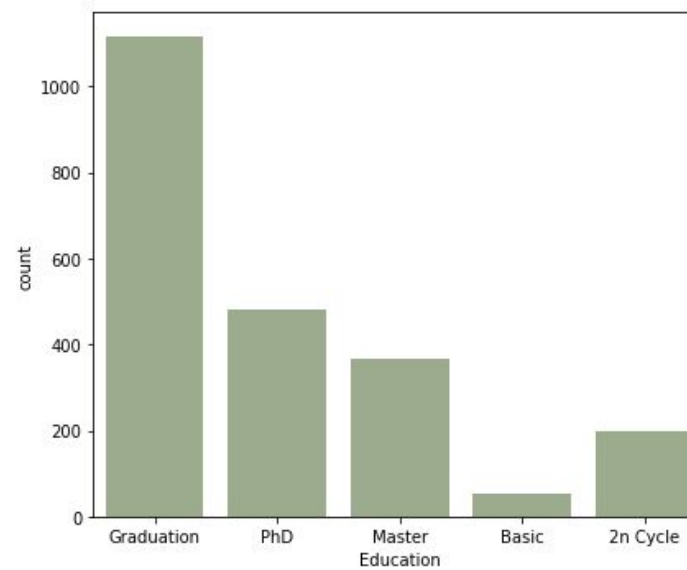
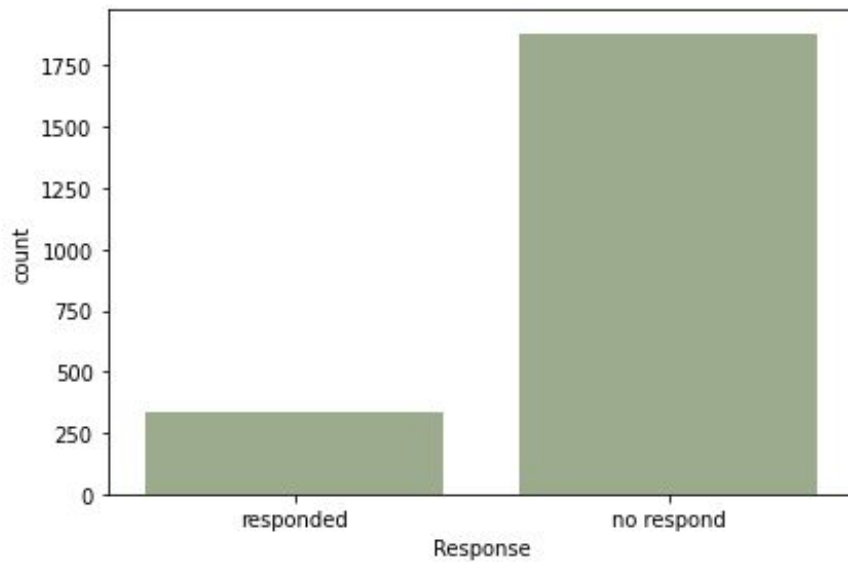


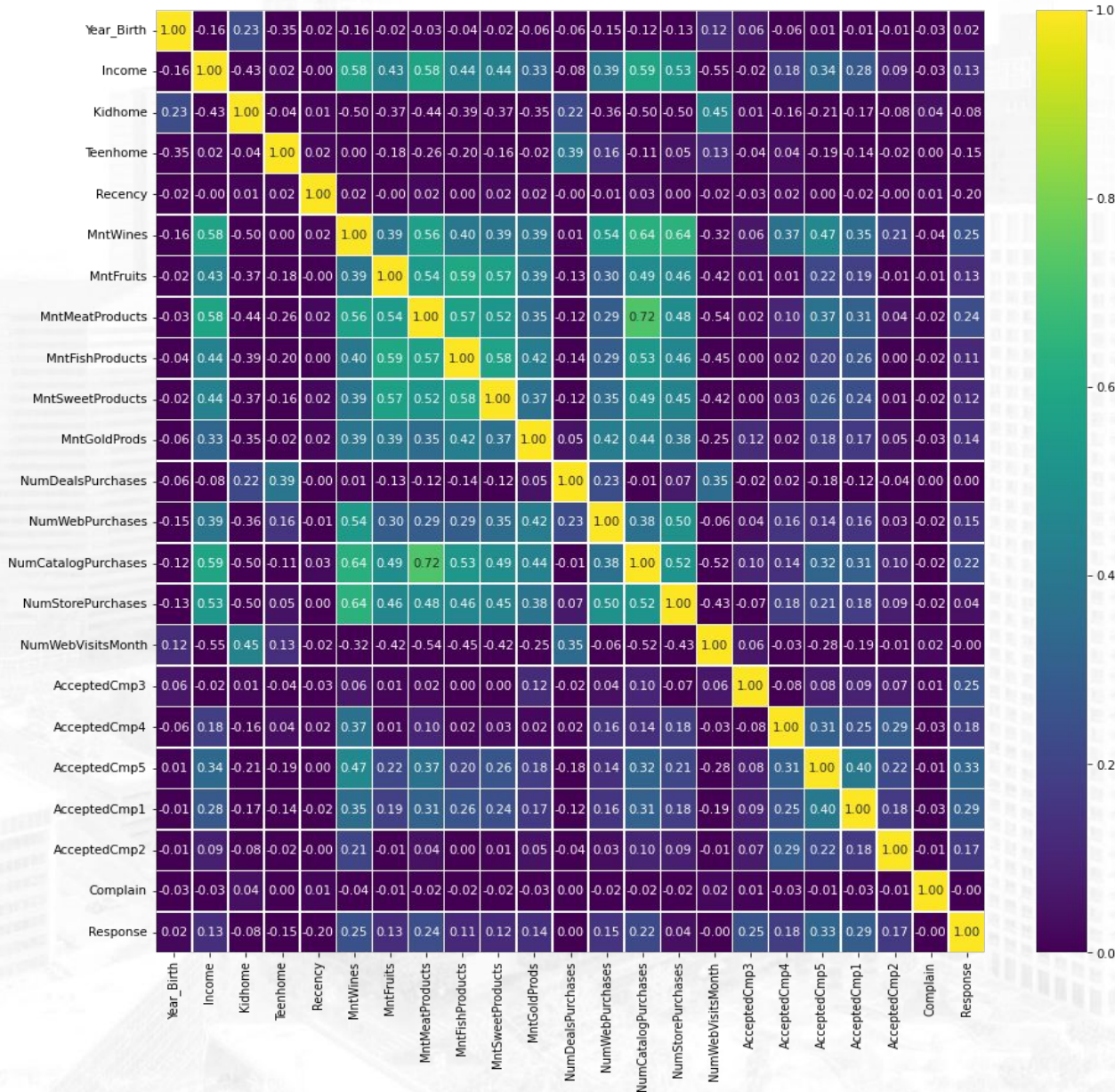
Kemudian, ditemukan outlier pada kolom Year\_Birth, Income, Recency, MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases, NumWebPurchases, NumWebVisitsMonth. Sehingga pada proses pre-processing, kolom yang memiliki outlier dapat dihapus berdasarkan z-score atau memakai IQR.



# Univariate Analysis

Kemudian, untuk data target ditemukan bahwa terdapat ketimpangan sehingga pada pre-processing data dapat dilakukan over/under sampling.





# Multivariate Analysis

Dari Heatmap diatas diketahui bahwa fitur yang memiliki korelasi yang cukup tinggi dengan response adalah income, mntwines, mnfruit, mntmeatproducts, mntfishproducts, mntsweetproducts, mntgoldprods, numwebpurchases, numcatalogpurchases dan acceptedcmp1-5

Dari seluruh korelasi antara feature-target, seluruhnya berada di range 0.00 sampai 0.33. Oleh karena itu, kami memutuskan untuk membuat nilai threshold di angka 0.20. Feature-feature di atas yang kami pertahankan adalah feature yang memiliki nilai korelasi >0.20.



# Multivariate Analysis

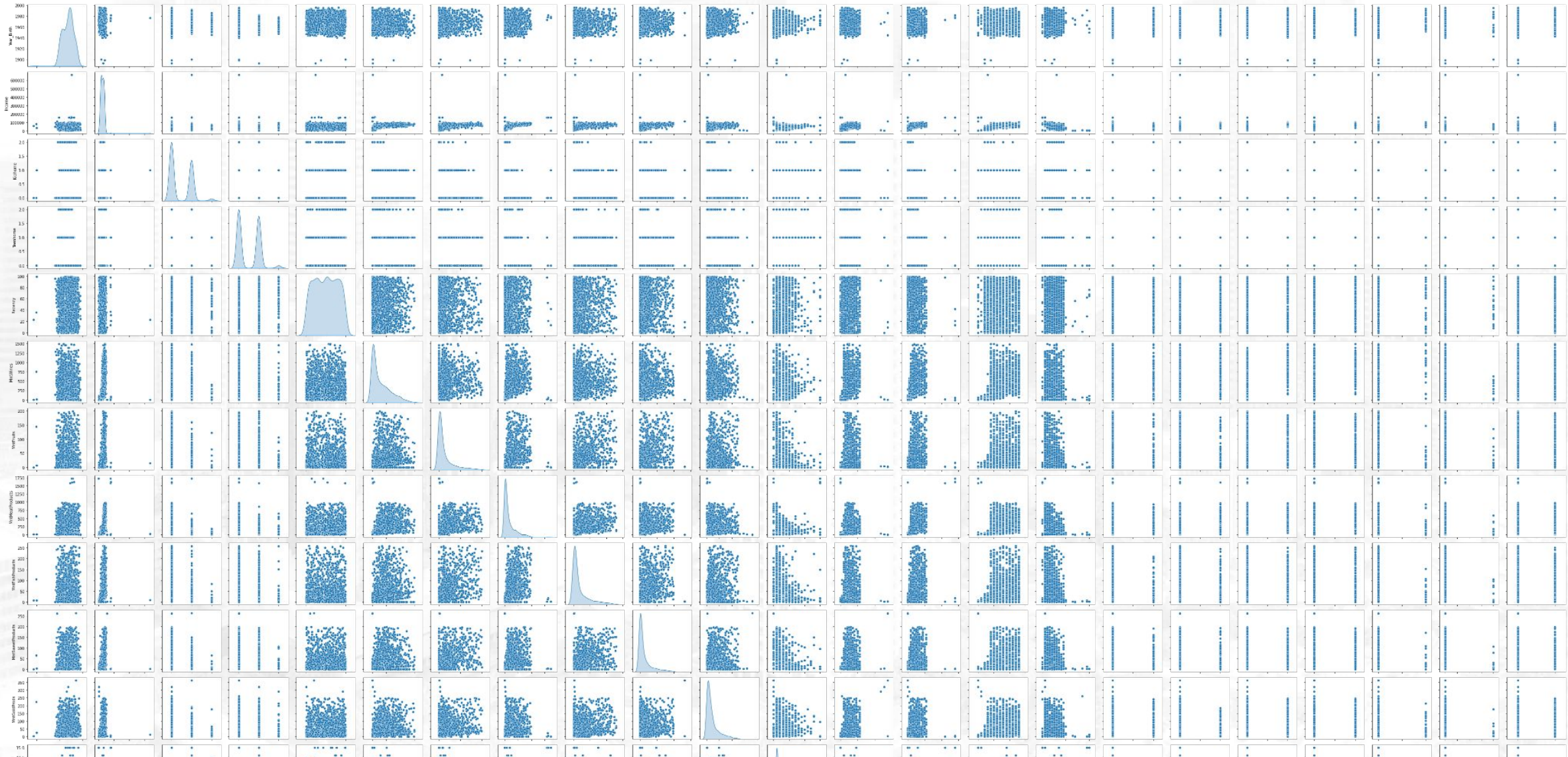
Berdasarkan analisa awal antar fitur yang kami lakukan terhadap fitur yang memiliki korelasi lebih tinggi dengan target, didapatkan hasil sebagai berikut:

- [Recency]: Nilai korelasi Recency dengan feature lainnya memiliki range 0.00 sampai 0.05
- [MntWines]: Berikut adalah feature yang berkorelasi dengan MntWines: Income(0.58), NumCatalogPurchases(0.64), NumStorePurchases(0.64)
- [MntMeatProducts]: Berikut adalah feature yang berkorelasi dengan MntMeatProducts: NumCatalogPurchases(0.72), Income(0.58), MntWines(0.56)
- [NumCatalogPurchases]: Berikut adalah feature yang berkorelasi dengan NumCatalogPurchases: MntMeatProducts(0.72), MntWines(0.64), Income(0.59)
- [AcceptedCmp3]: Berikut adalah feature yang berkorelasi dengan AcceptedCmp3: MntGoldProducts(0.12)
- [AcceptedCmp5]: Berikut adalah feature yang berkorelasi dengan AcceptedCmp5: MntWines(0.47), MntMeatProducts(0.37), Income(0.34)
- [AcceptedCmp1]: Berikut adalah feature yang berkorelasi dengan AcceptedCmp1: AcceptedCmp5(0.40), MntWines(0.35), MntMeatProducts(0.31), NumCatalogPurchases(0.31)

Dari hasil tersebut, kemungkinan besar nantinya akan kami gunakan sebagai fitur prioritas dalam keputusan penentuan indikator pendukung untuk pengkategorisasian customer mana yang layak diberikan campaign.

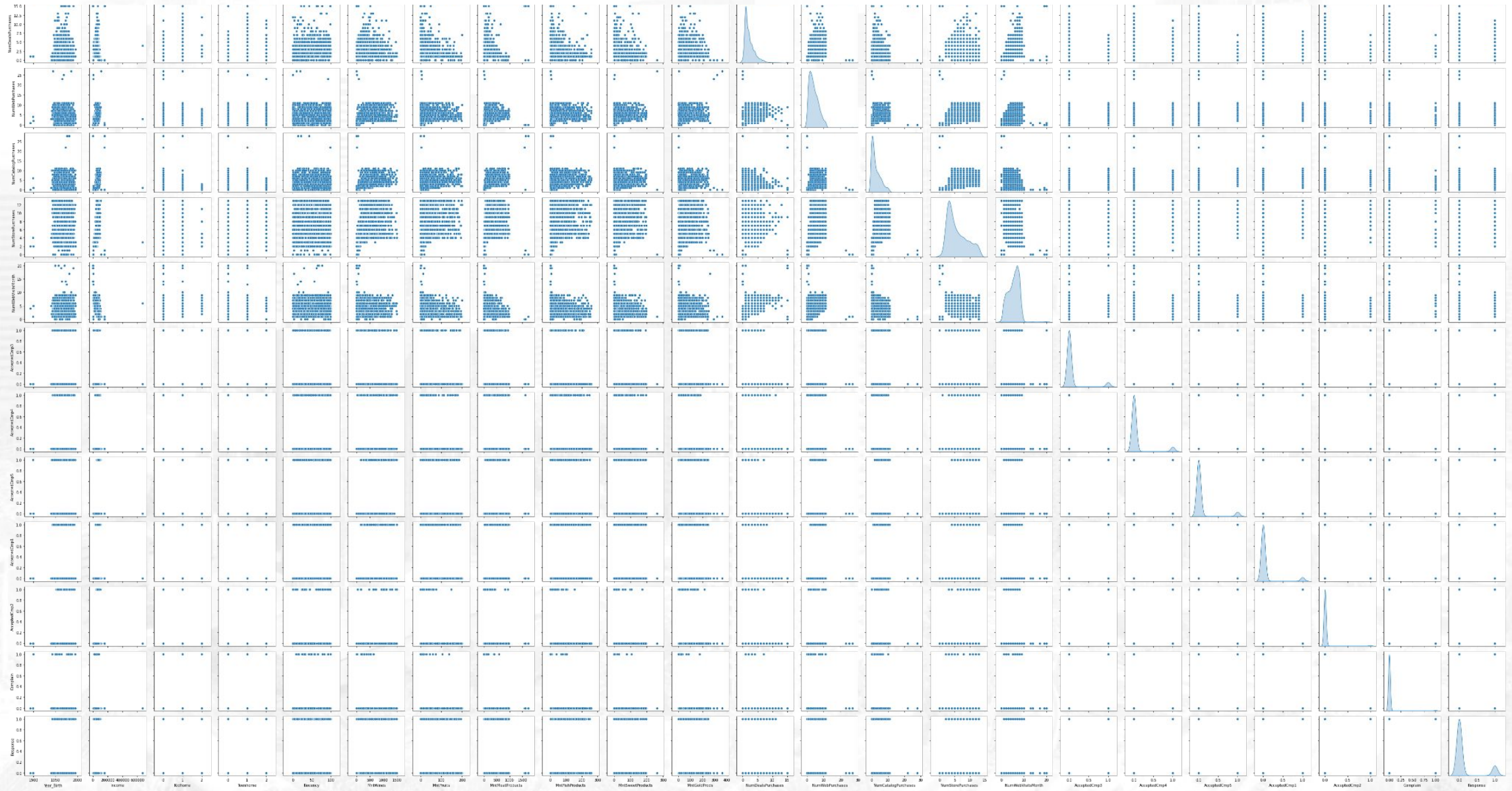


# Multivariate Analysis

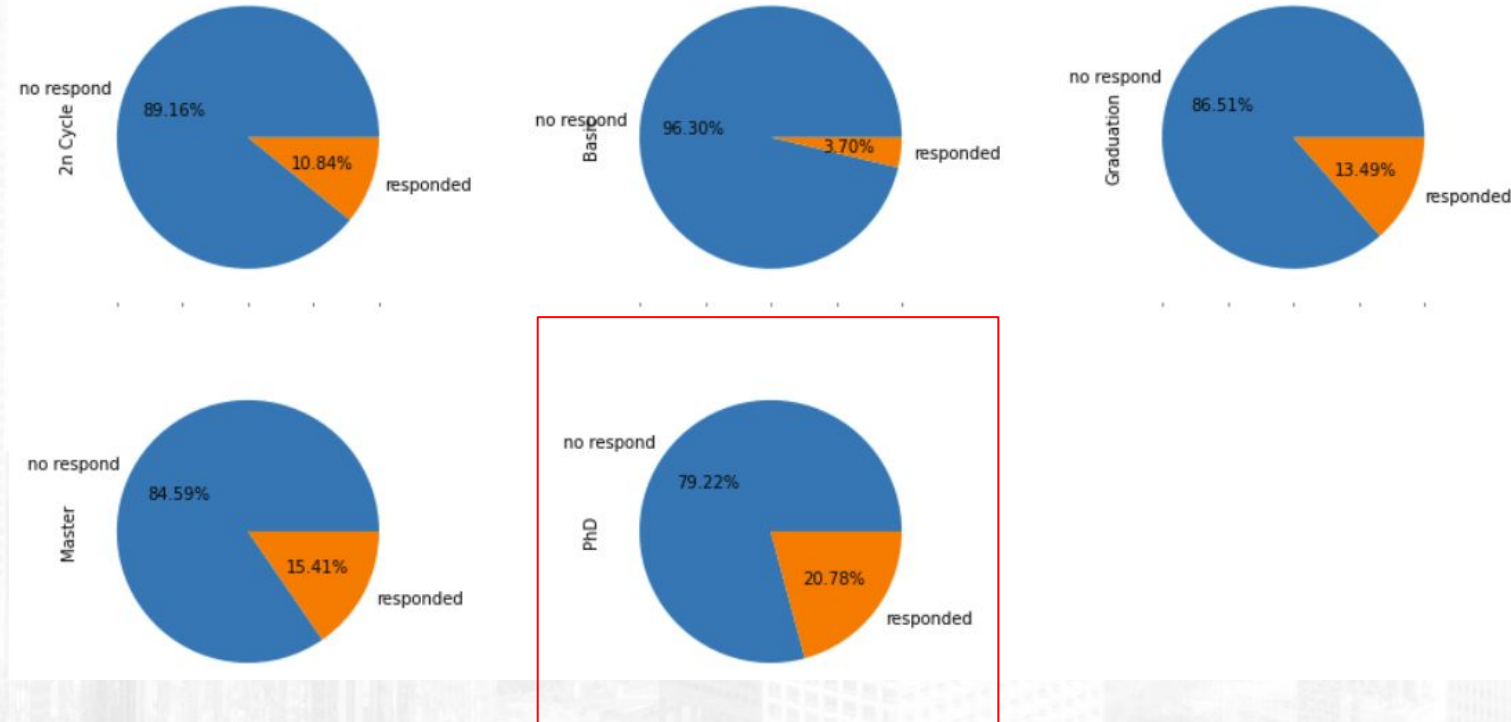




# Multivariate Analysis



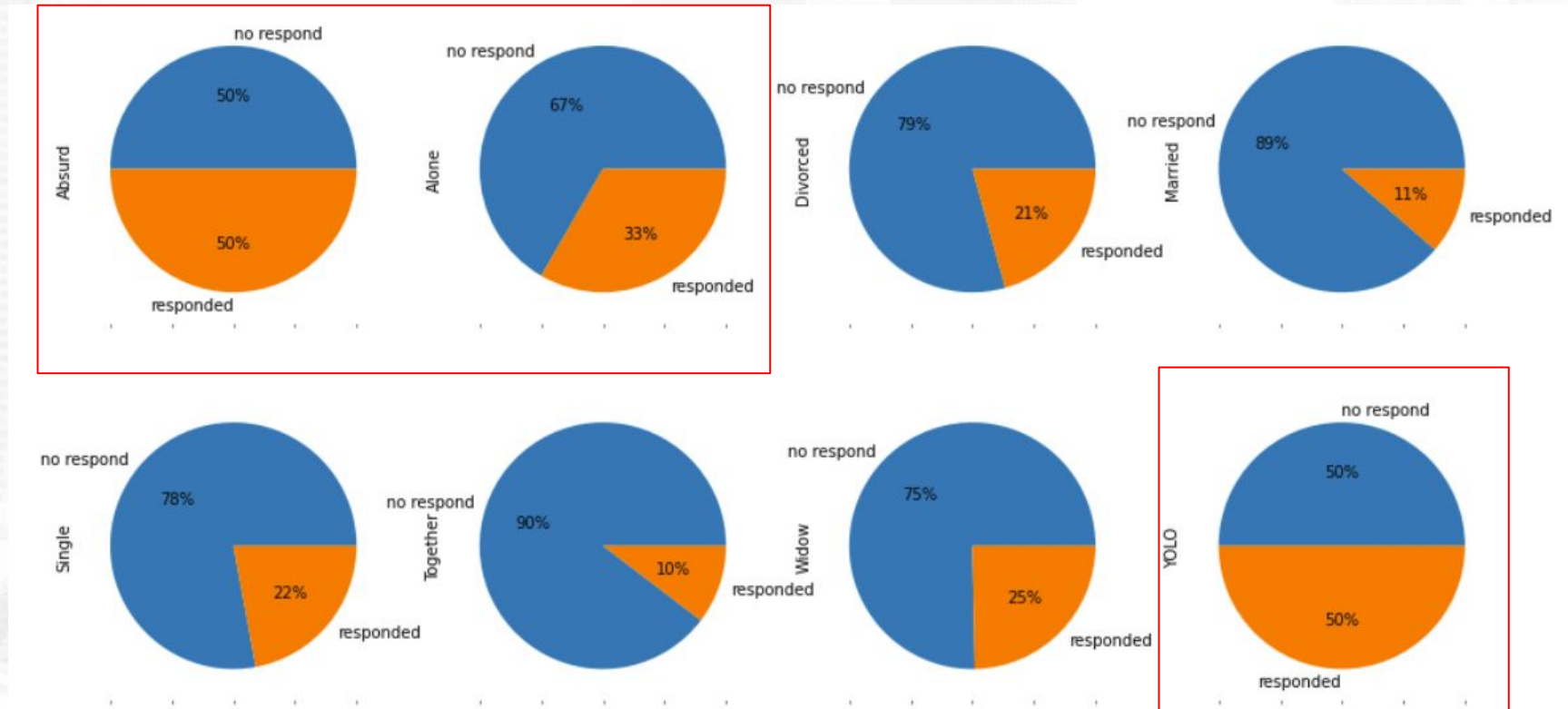
# Business Insight



Dari visualisasi Juga dapat dilihat bahwa customer yang merespon terbanyak berasal dari customer yang memiliki edukasi PhD (20.78%), disusul dengan Master



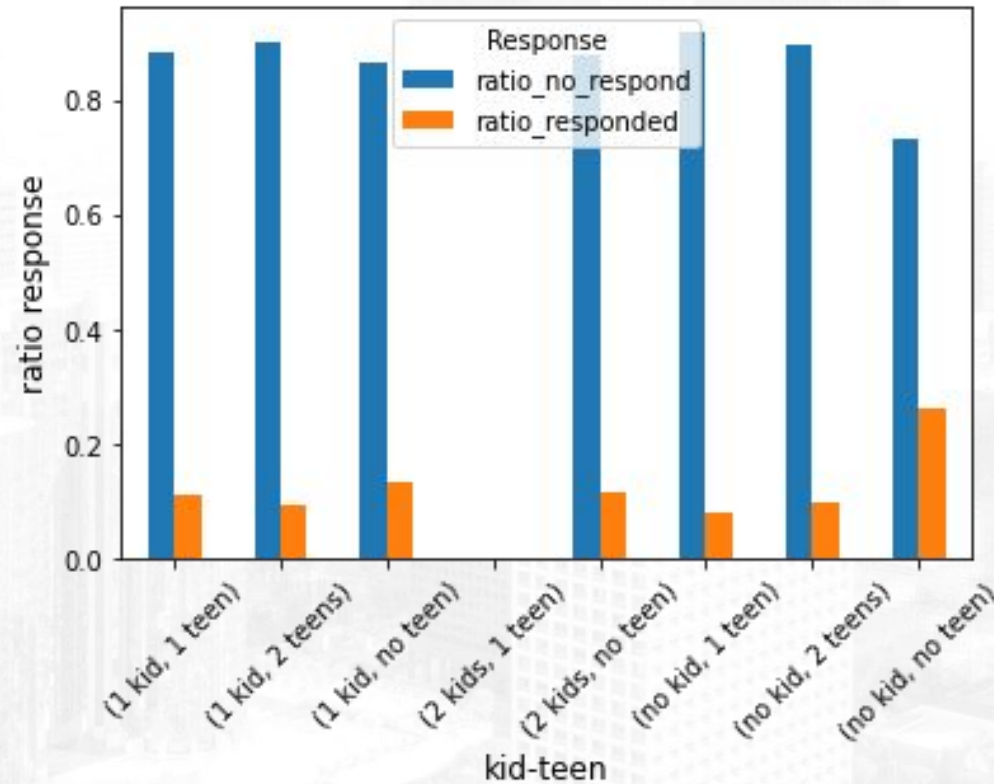
# Business Insight



Dari visualisasi piechart Marital Status, dapat dilihat bahwa customer yang merespon terbanyak berasal dari customer yang berstatus Absurd (50%) dan Yolo(50%), disusul dengan Alone(33%) dan Widow (25%), sehingga marketing team dapat memfokuskan campaign ke customer "Absurd" dan "Yolo".

# Business Insight

Perbandingan ratio response dengan Customer yang memiliki Kid-Teen



Dari visualisasi kidhome dan teenhome, dapat dilihat bahwa customer yang merespon terbanyak berasal dari customer yang tidak mempunyai anak dan tidak mempunyai remaja (0.265403), sehingga marketing team dapat memfokuskan campaign ke customer yang tidak mempunyai anak dan tidak mempunyai remaja.

## Next Improvement

Selain dari tiga business insight tersebut, kami juga memiliki satu buah insight lagi berupa sebuah trend suatu product yang memiliki korelasi kuat (Gold, Meat, dan Wines) terhadap campaign 1 sampai dengan campaign 5. Kemudian, hasil dari visualisasi insight tersebut nantinya dapat digunakan oleh perusahaan untuk memprioritaskan produk mana yang akan dijual atau dipromosikan guna menarik jumlah customer. Sehingga diharapkan dengan adanya kenaikan jumlah customer tersebut, jumlah revenue perusahaan pun dapat bertambah

Namun, dikarenakan keterbatasan waktu dan skill, kami belum sempat membuat visualisasi insight terakhir tersebut dan berencana untuk menjadikannya sebagai salah satu next improvement.



**[Link GDrive - Halcyon](#)**

**[Link Github - Halcyon](#)**