

$$h \sum_{i=1}^9 \text{llo Sigma}_{\text{Tech.}}$$

**Dokumen
Laporan Final
Project – Stage 1**





Descriptive Statistic

□ Info Data

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 381109 entries, 0 to 381108
Data columns (total 12 columns):
#   Column                Non-Null Count  Dtype
---  -
0   id                    381109 non-null int64
1   Gender                381109 non-null object
2   Age                  381109 non-null int64
3   Driving_License       381109 non-null int64
4   Region_Code          381109 non-null float64
5   Previously_Insured    381109 non-null int64
6   Vehicle_Age          381109 non-null object
7   Vehicle_Damage       381109 non-null object
8   Annual_Premium       381109 non-null float64
9   Policy_Sales_Channel  381109 non-null float64
10  Vintage              381109 non-null int64
11  Response              381109 non-null int64
dtypes: float64(3), int64(6), object(3)
memory usage: 34.9+ MB
```

```
df.isnull().sum()
```

```
id                    0
Gender                0
Age                  0
Driving_License       0
Region_Code          0
Previously_Insured    0
Vehicle_Age          0
Vehicle_Damage       0
Annual_Premium       0
Policy_Sales_Channel  0
Vintage              0
Response              0
dtype: int64
```

- Tidak ada data yang memiliki nilai null
- Semua tipe data sudah sesuai, namun perlu penyesuaian EDA dan pada Pre-Processing untuk Modelling
- Nilai unik pada kolom juga tidak memiliki suatu kejanggalan



Descriptive Statistic

□ Data Numerik

	Age	Annual_Premium	Vintage
count	381109.000000	381109.000000	381109.000000
mean	38.822584	30564.389581	154.347397
std	15.511611	17213.155057	83.671304
min	20.000000	2630.000000	10.000000
25%	25.000000	24405.000000	82.000000
50%	36.000000	31669.000000	154.000000
75%	49.000000	39400.000000	227.000000
max	85.000000	540165.000000	299.000000

- Tidak ada dominasi yang berlebih di antara tiap unique nilai pada kolom "Gender" dan "Vehicle_Damage".
- Sedangkan pada kolom "Vehicle_Age" dominasi ada pada nilai "1-2 Year" dan "< 1 Year" dibandingkan dengan "> 2 Years" dengan perbedaan yang cukup signifikan

```
1-2 Year      200316
< 1 Year      164786
> 2 Years     16007
Name: Vehicle_Age, dtype: int64
```

□ Data Kategorik

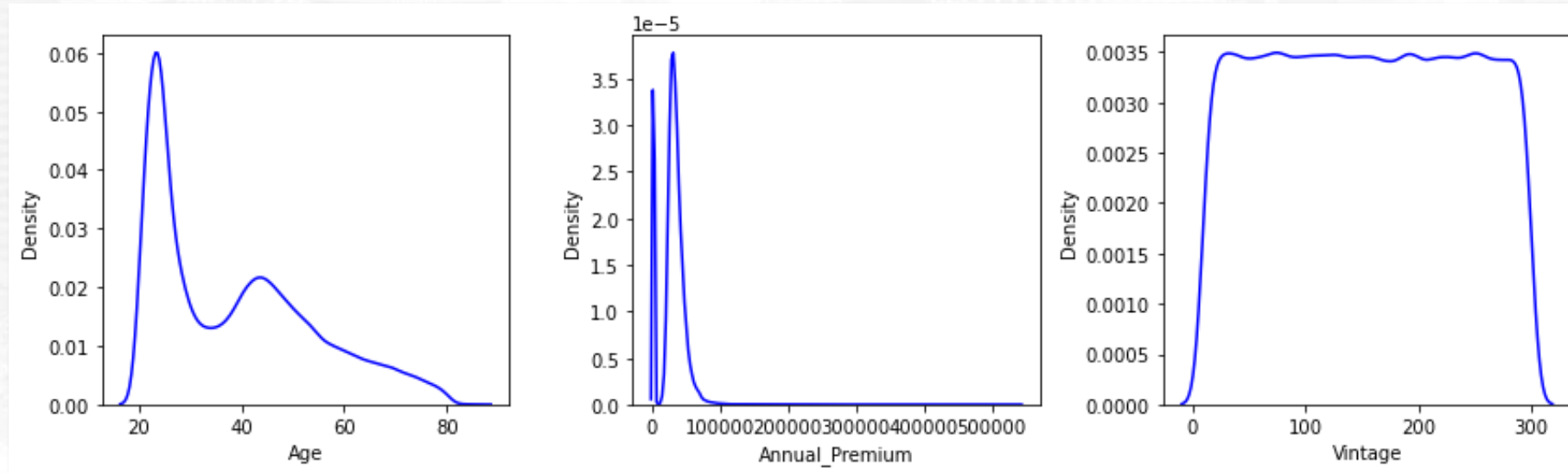
	id	Gender	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Policy_Sales_Channel	Response
count	381109	381109	381109	381109.0	381109	381109	381109	381109.0	381109
unique	381109	2	2	53.0	2	3	2	155.0	2
top	1	Male	1	28.0	0	1-2 Year	Yes	152.0	0
freq	1	206089	380297	106415.0	206481	200316	192413	134784.0	334399



Univariate Analysis

□ Distribusi tiap variabel data numerical

KDE Plot



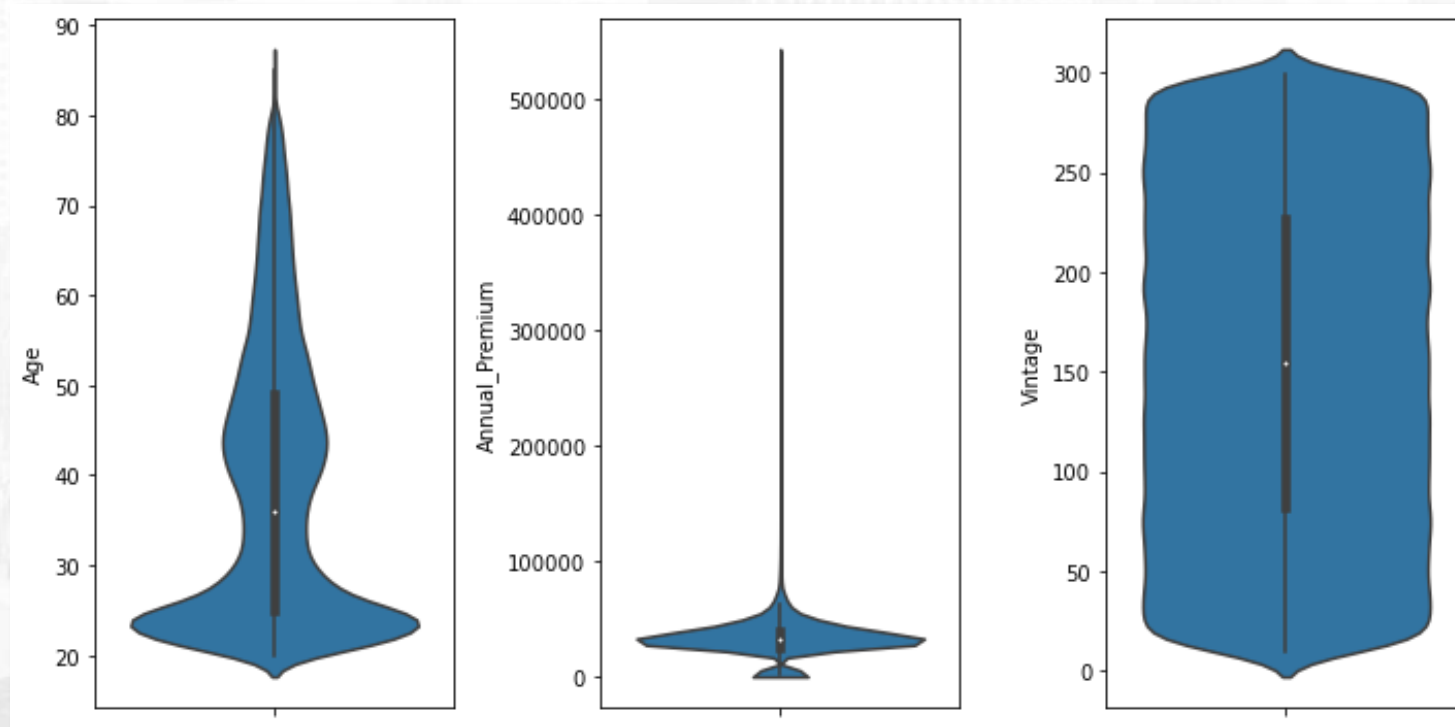
- Distribusi pada variable Age adalah skewness positif
- Pada Annual_Premium, variabel tersebut memiliki distribusi bimodal dan ekornya cenderung ke arah kanan.



Univariate Analysis

- ❑ Distribusi tiap variabel data numerical

Violin Plot

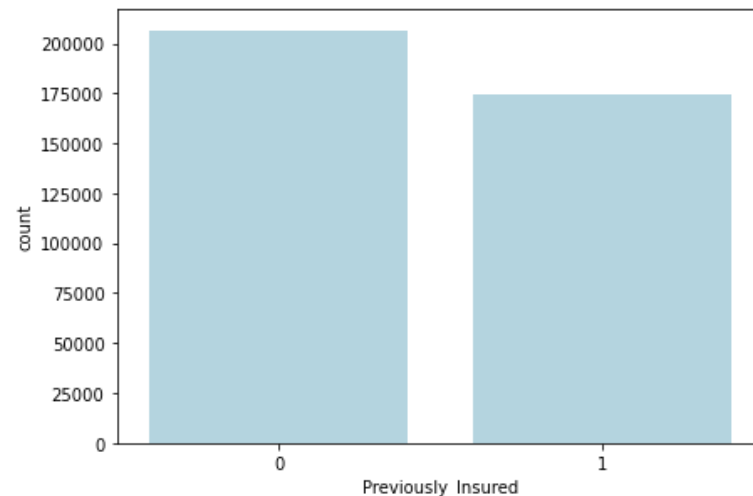
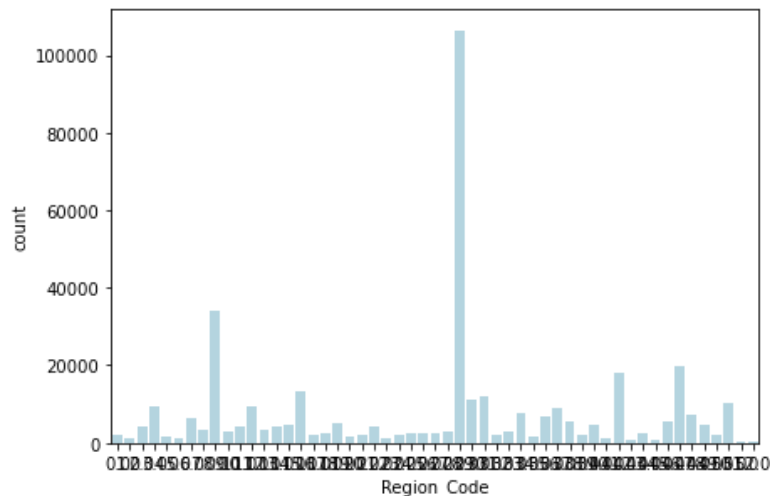
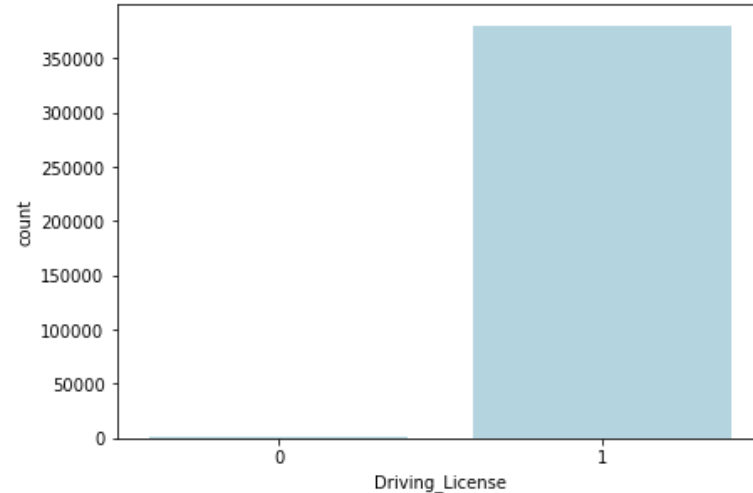
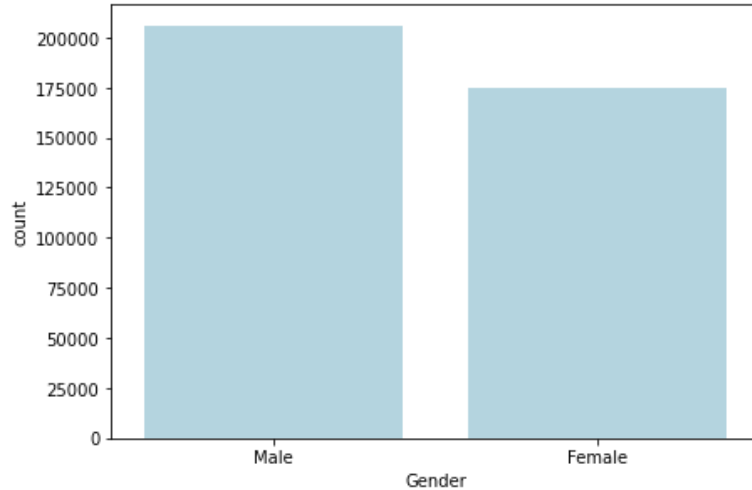


- Pada violinplot nilai dengan outlier terbanyak ada pada kolom "Annual_Premium dengan jumlah outlier 10320 baris data



Univariate Analysis

Countplot Data Kategorik

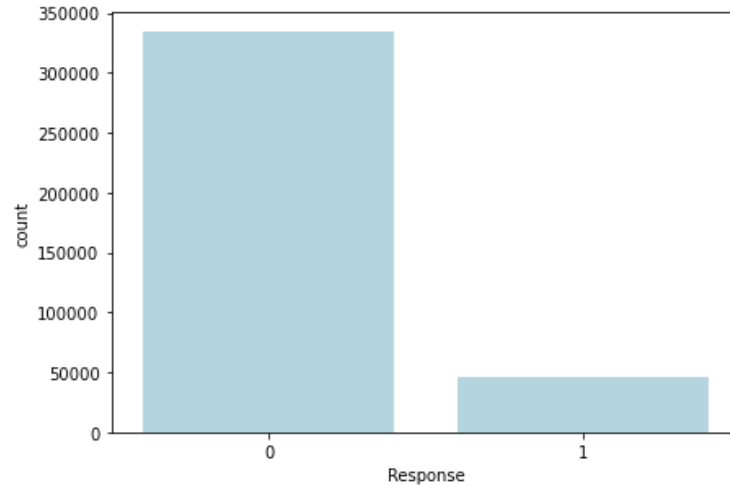
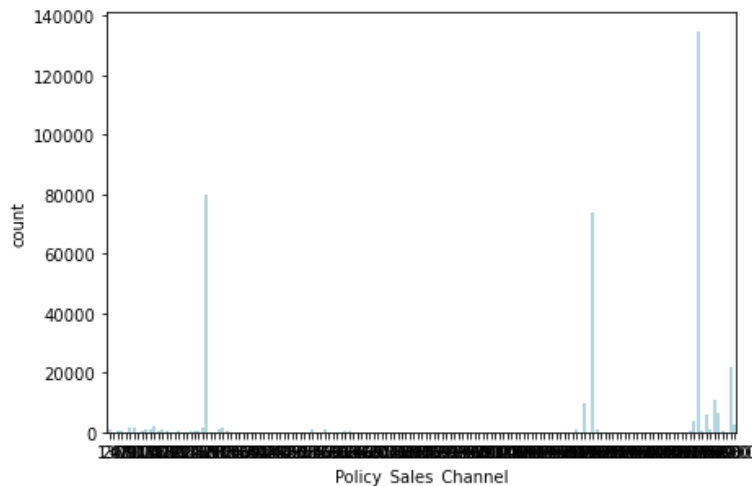
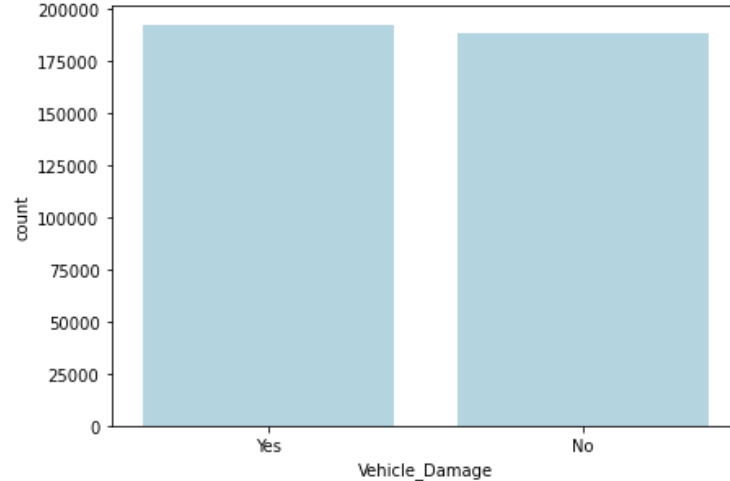
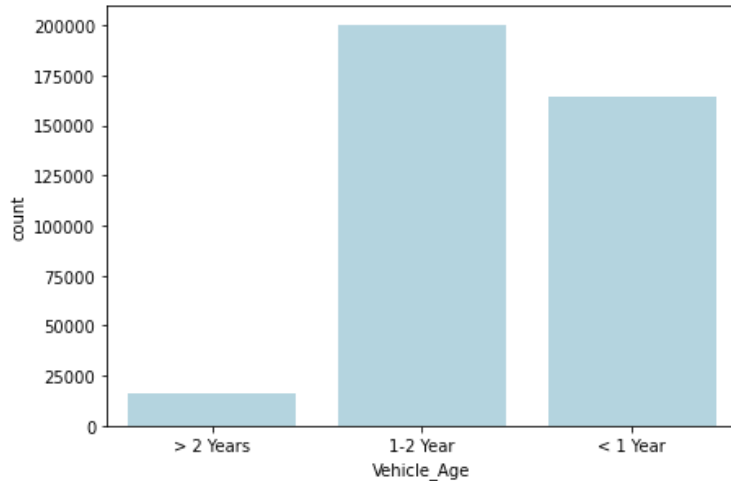


- Tidak ada nilai yang terlalu mendominasi pada kolom Gender dan Previously_Insured
- Pada Kolom Region_Code, kode Region terbanyak yang terdata adalah kode 28 yang sangat mendominasi dibandingkan yang lainnya.
- Variabel Driving_License memiliki satu kelompok yang mendominasi.



Univariate Analysis

Countplot Data Kategorik



- Tidak ada nilai yang terlalu mendominasi pada variable Vehicle_Damage
- Pada kolom Vehicle_Age terdapat dua kelompok yang mendominasi yaitu Kolom Vehicle_Age terdapat dua nilai yang mendominasi yaitu "1-2 Year", "<2 Year".
- Pada kolom Policy_Sales_Channel atau dapat diartikan Tipe Sales yang diterima pelanggan, tipe dengan kode 152, 26, dan 124 adalah tipe Sales yang menduduki 3 besar terbanyak digunakan.
- Kolom Response sebagai Target memiliki satu nilai yang mendominasi.



Univariate Analysis

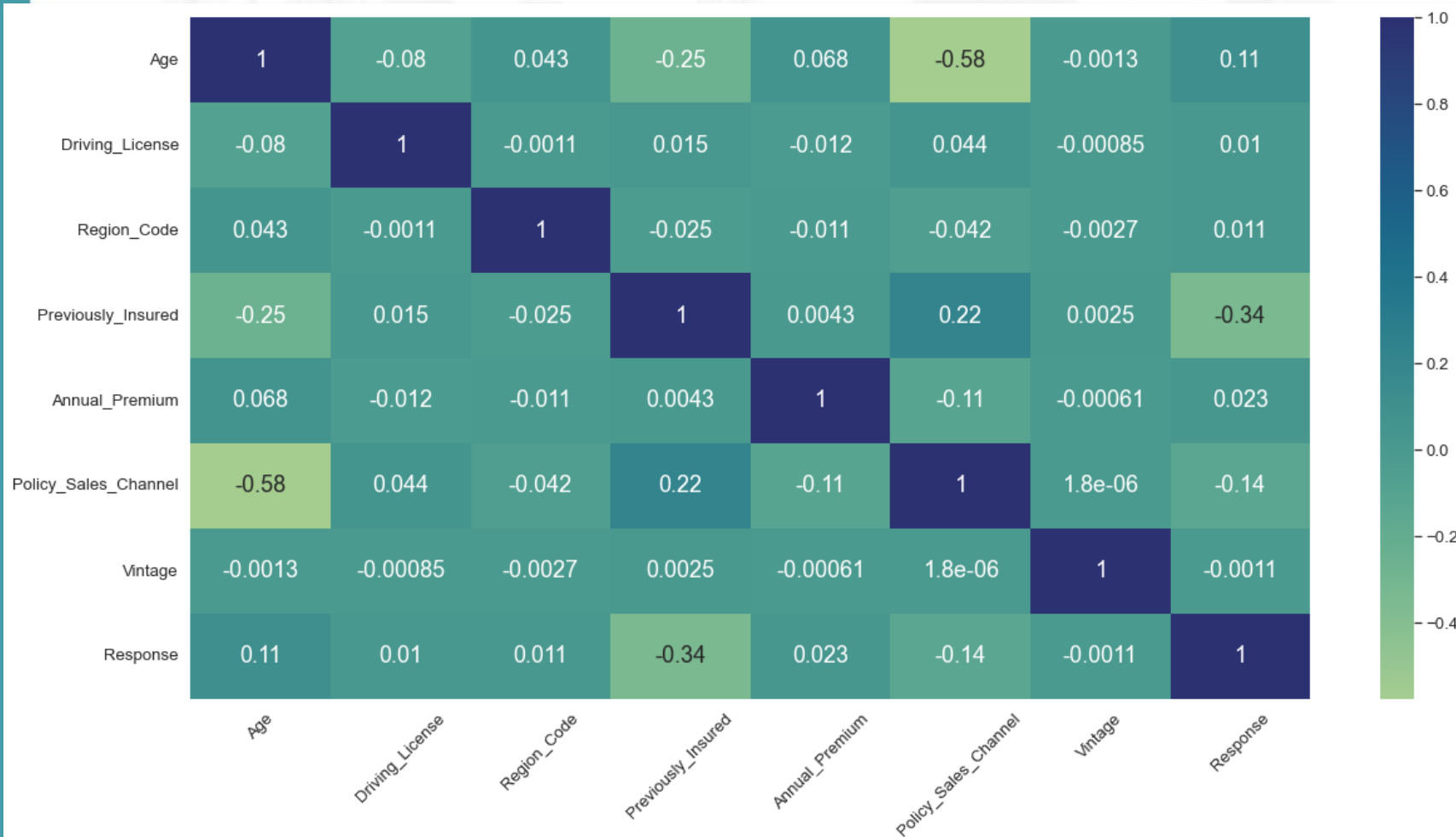
❑ Follow up untuk Pre-Processing

- Melakukan handling outlier
- Melakukan standarisasi pada kolom-kolom yang jauh dari distribusi normal
- Melakukan feature encoding pada kolom yang berisi data categorical
- Melakukan class imbalance pada kolom yang memiliki nilai dominasi pada kolom Response sebagai Target pada case ini



Multivariate Analysis

❑ Korelasi antar kolom

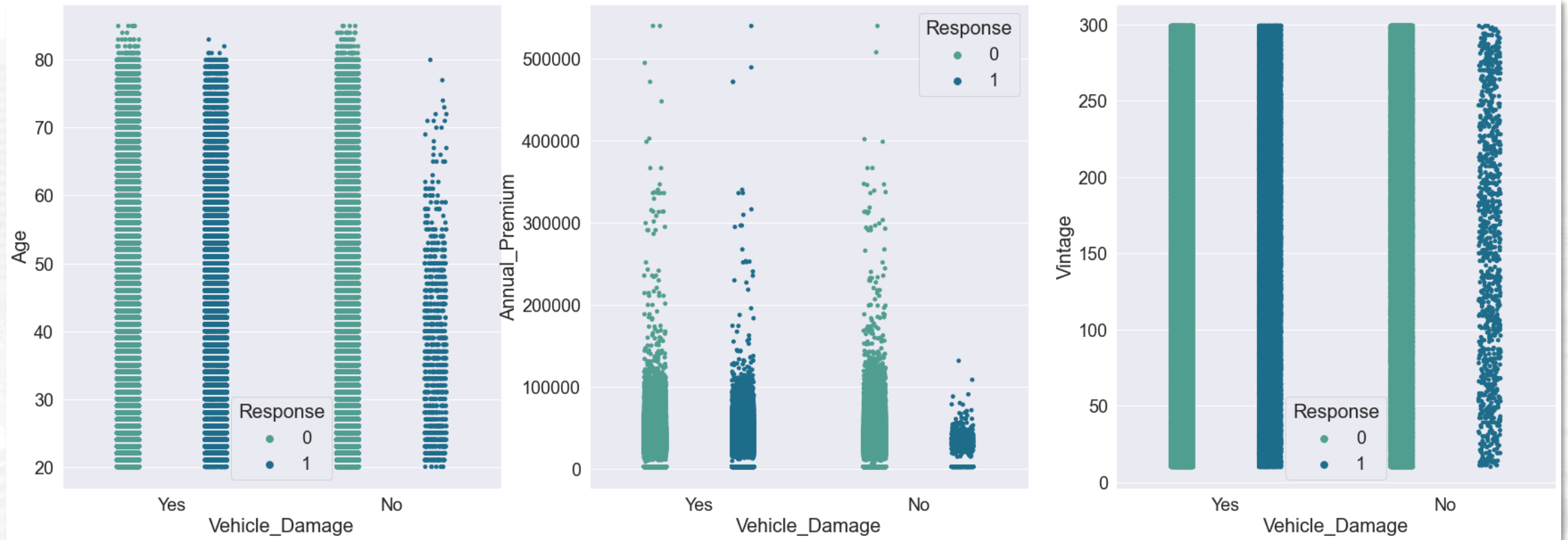


- Tidak ada variabel yang memiliki korelasi kuat atau $r \geq 0.7$ atau $r \leq -0.7$
- Hanya variabel Age dengan Policy_Sales_Channel memiliki korelasi yang cukup kuat yaitu -0.58



Multivariate Analysis

Category Plots

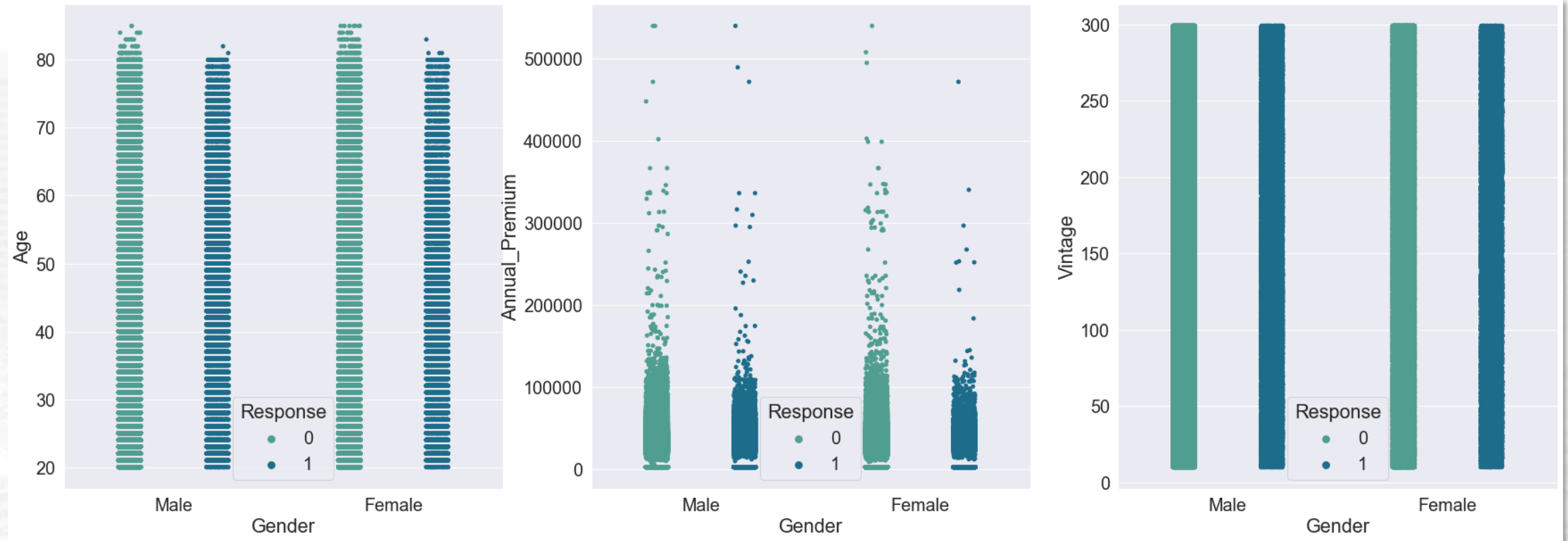


Dapat dilihat dari grafik di atas beberapa variabel terhadap Vehicle_damage dengan hue Response, pada variable Age dan vintage distribusinya tidak terlihat perbedaan yang signifikan.



Multivariate Analysis

Category Plots

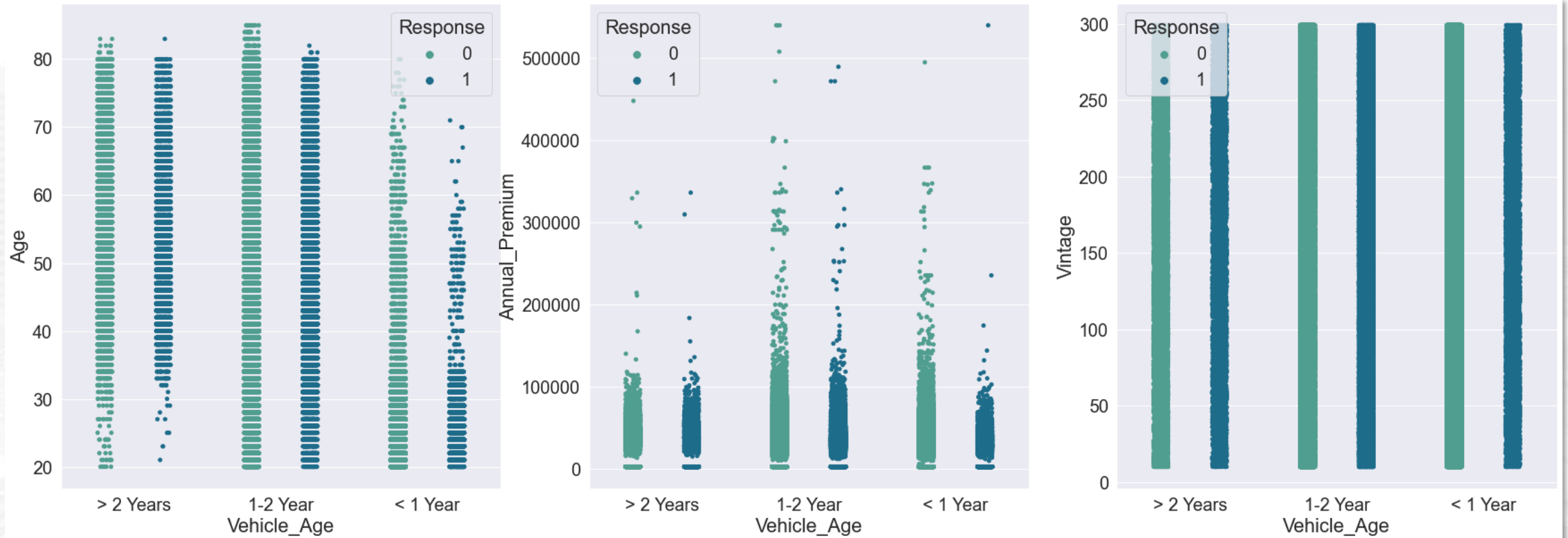


Dapat dilihat dari grafik di atas beberapa variabel terhadap Gender dengan hue Response, pada variable Age dan vintage distribusinya tidak terlihat perbedaan yang signifikan.



Multivariate Analysis

Category Plots

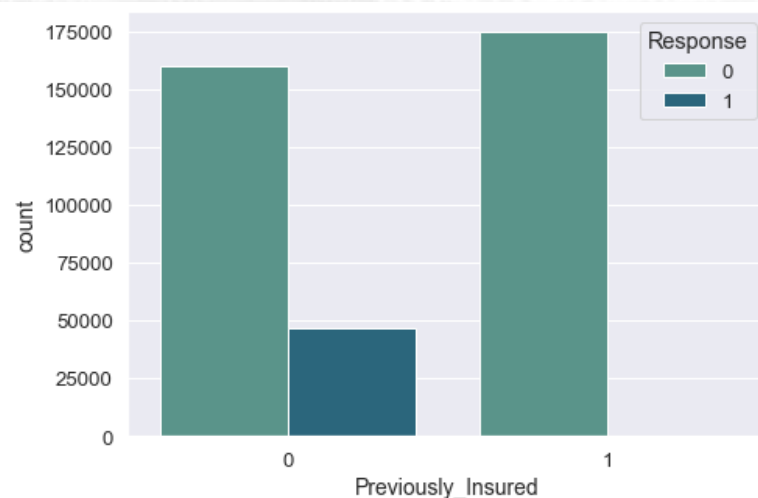
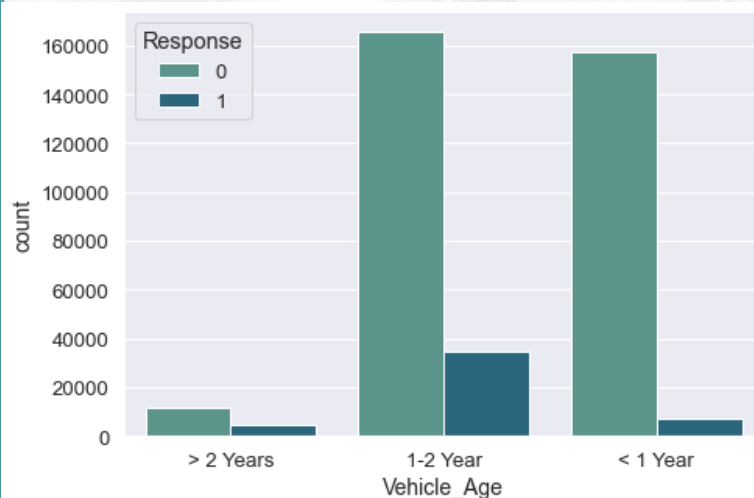
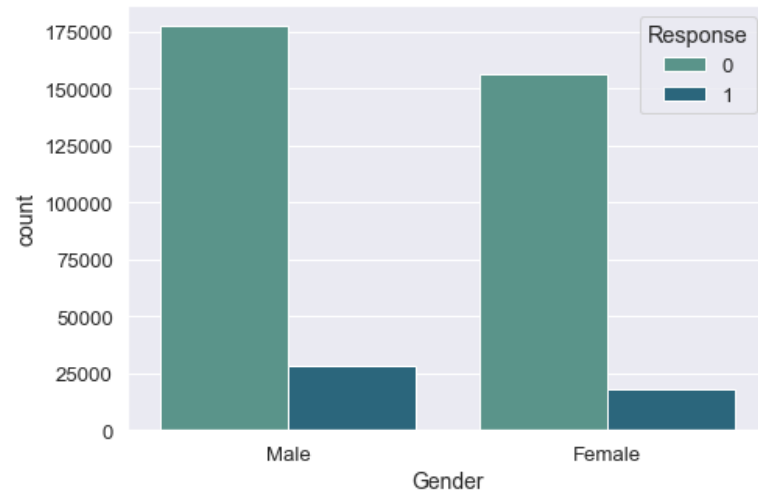
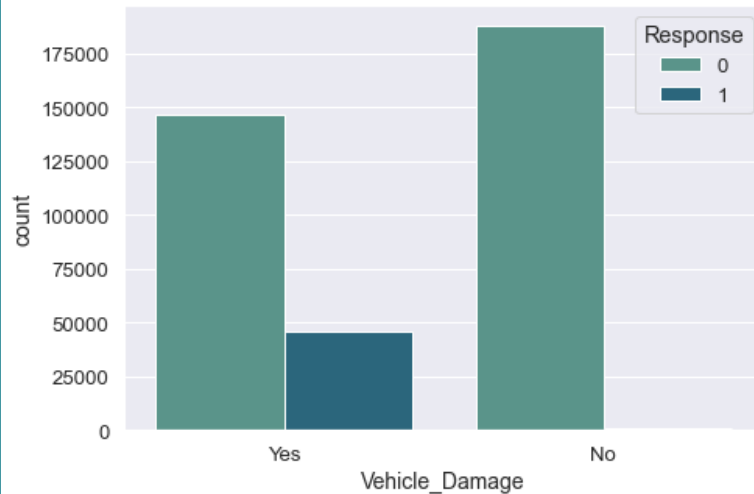


Dapat dilihat dari grafik di atas beberapa variabel terhadap `Vehicle_Age` dengan hue `Response`. Pada variabel `vintage` distribusinya tidak terlihat perbedaan yang signifikan bahkan distribusinya cenderung rapat. Kemudian pada `Vehicle_Age` dengan `Age`, hanya pada kelompok `1-2 Year` distribusinya tidak terlihat perbedaan yang signifikan. Pada variabel `Vehicle_Age` dengan `Annual_Premium`, distribusinya tidak terlihat perbedaan signifikan saat `Annual_Premium` di sekitar 2.630 hingga 90.000



Insight Business

❑ Kecenderungan variabel Response terhadap beberapa variabel lain

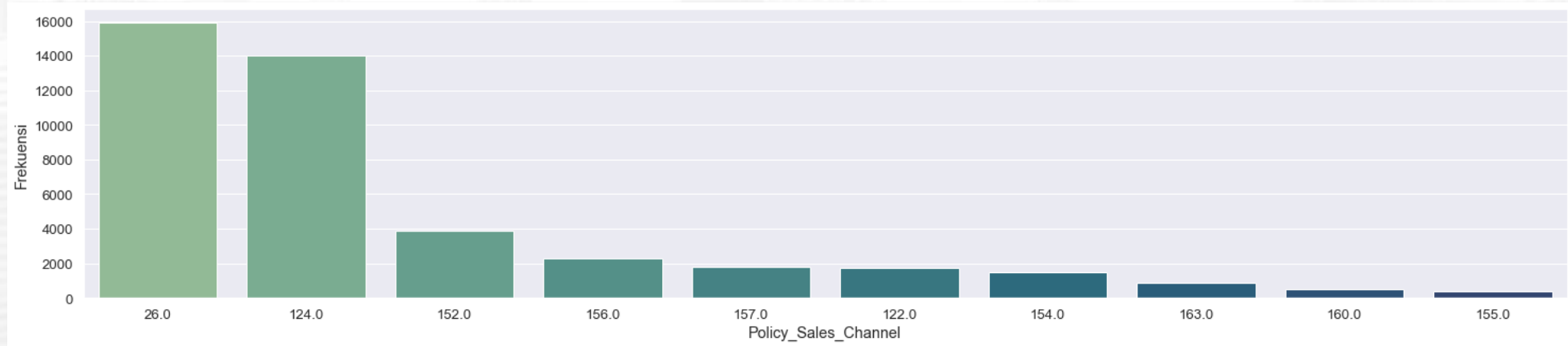


- Pada nasabah yang tertarik menggunakan Asuransi Kendaraan, kondisi kendaraan sebelumnya lebih banyak yang telah mengalami kerusakan dibandingkan yang belum.
- Nasabah yang tertarik menggunakan Asuransi Kendaraan baik Male maupun Female tidak terdapat perbedaan yang signifikan.
- Pada kelompok usia kendaraan 1-2 tahun lebih banyak yang tertarik menggunakan asuransi Kendaraan dibandingkan kelompok lain seperti kelompok usia kendaraan <1 tahun dan usia kendaraan >2 tahun
- Nasabah yang tertarik menggunakan Asuransi kendaraan, sebelumnya banyak yang belum menggunakan Asuransi Kendaraan.



Insight Business

❑ Kecenderungan variabel Response terhadap beberapa varibel lain



- Ada 10 kode jenis Policy_Sales_Channel yang paling banyak dalam memberikan kontribusi ketertarikan nasabah menggunakan Asuransi Kendaraan.



Insight Business

❑ Distribusi Age berdasarkan variable Response

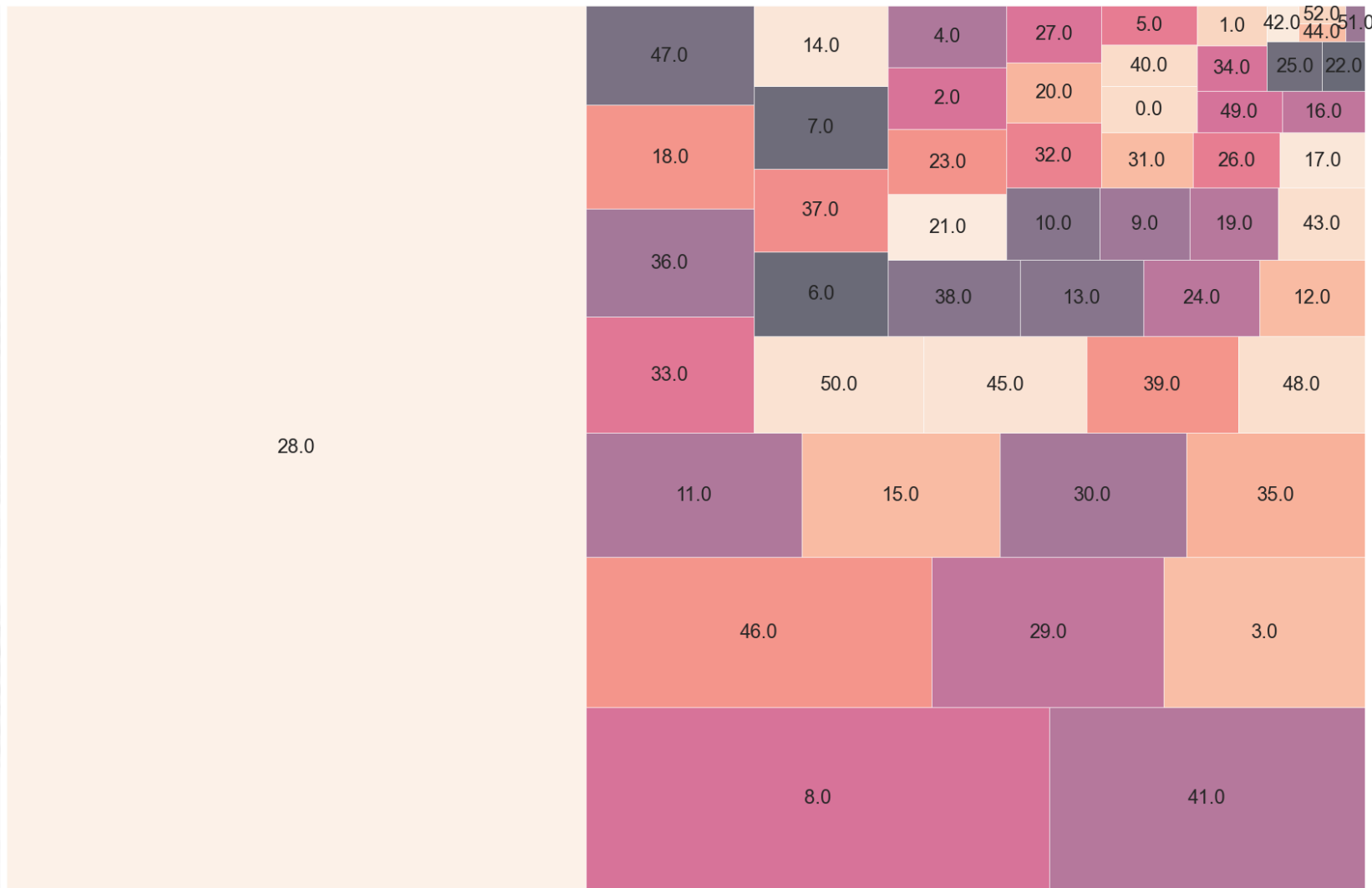


- Usia Nasabah dari 33 – 52 merupakan kelompok umur yang paling banyak tertarik menggunakan Asuransi Kendaraan.



Insight Business

❑ Kecenderungan variabel Response terhadap beberapa varibel lain



- Ada beberapa Region_Code dengan code 28.0, 41.0, 8.0, 46.0, 29.0, dimana wilayah tersebut nasabah yang tertarik menggunakan Asuransi Kendaraan lebih banyak dibandingkan Region lain.



Business Recommendation

❑ Rekomendasi Bisnis untuk Perusahaan Asuransi:



Menggunakan top 10 jenis Policy_Sales_Channel sebagai media promosinya



Perusahaan dapat memfokuskan pada beberapa Region yang memiliki nasabah dengan ketertarikan Asuransi Kendaraan paling banyak dibandingkan Region lain.



Memfokuskan promosi ke user dengan rentang umur 33-52 tahun dan pernah mengalami kerusakan pada kendaraannya serta umur kendaraannya berada pada 1-2 tahun.