HOMEWORK

FINAL PROJECT

Theofilus Arifin

Christofer Bryan N. K.

Ramlan Apriyansyah

Muhammad Iqbal

Hanifah Arrasyidah

Christopher Stephen

Muhammad Rizq N. A.

Ujang Pian

1. Data Cleansing

a. Handle Missing Values

Pertama kita akan melakukan pengecekan apakah terdapat missing value dalam data atau tidak.

```
[ ]  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
```

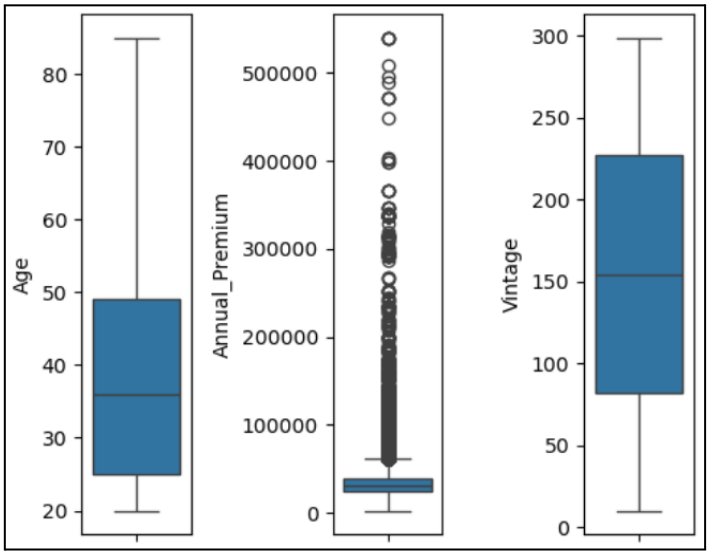
Dapat dilihat bahwa dalam dataset ini tidak ada missing values, sehingga proses handle missing value tidak perlu dilakukan.

b. Handel Duplicated Data

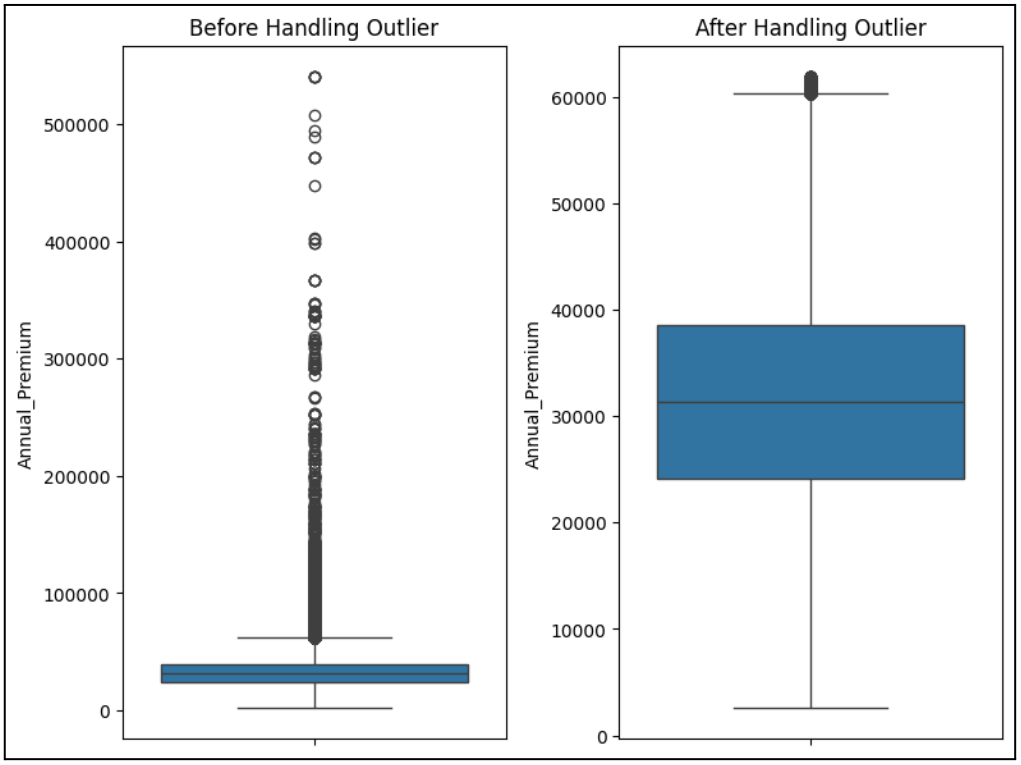
```
[7]: df.duplicated(subset=['id']).sum()

[7]: 0
```

Dapat dilihat bahwa pada kolom ‘id’ tidak ada data duplikat yang ditemukan, sehingga dapat disimpulkan bahwa tidak ada data duplikat dalam dataset ini.

c. Handle Outliers



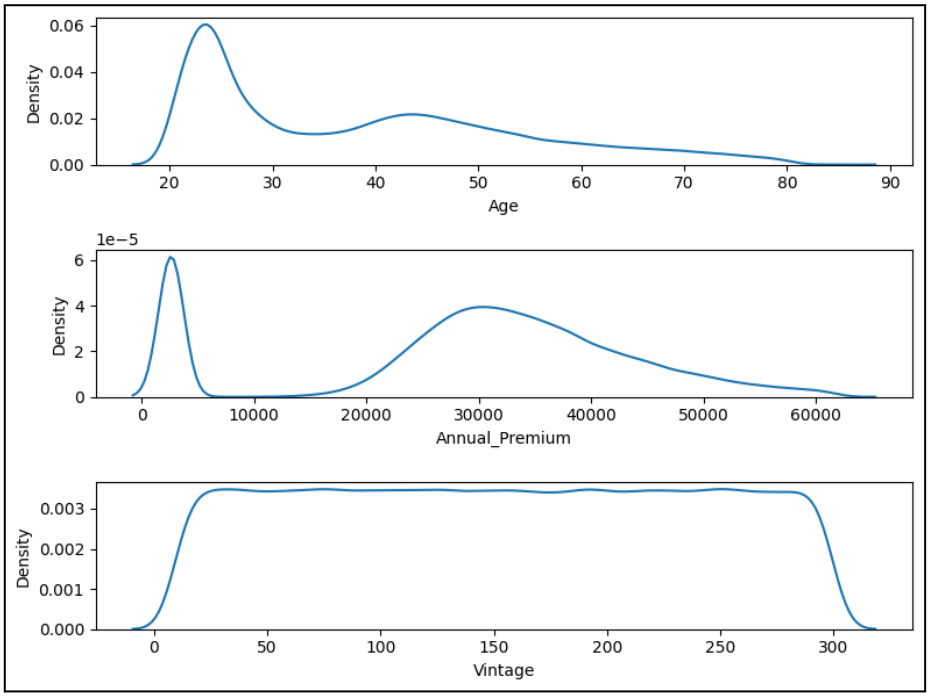
Setelah dilakukan pengecekan outlier pada kolom numerikal dengan menggunakan boxplot didapat bahwa kolom 'Age' dan 'Vintage' tidak memiliki outlier, sedangakan pada kolom 'Annual_Premium' terdapat outliers.Karena visualisasi outlier dilakukan menggunakan boxplot, maka penghapusan outlier akan menggunakan metode IQR. Berukut adalah hasil dari penghapusan outlier.



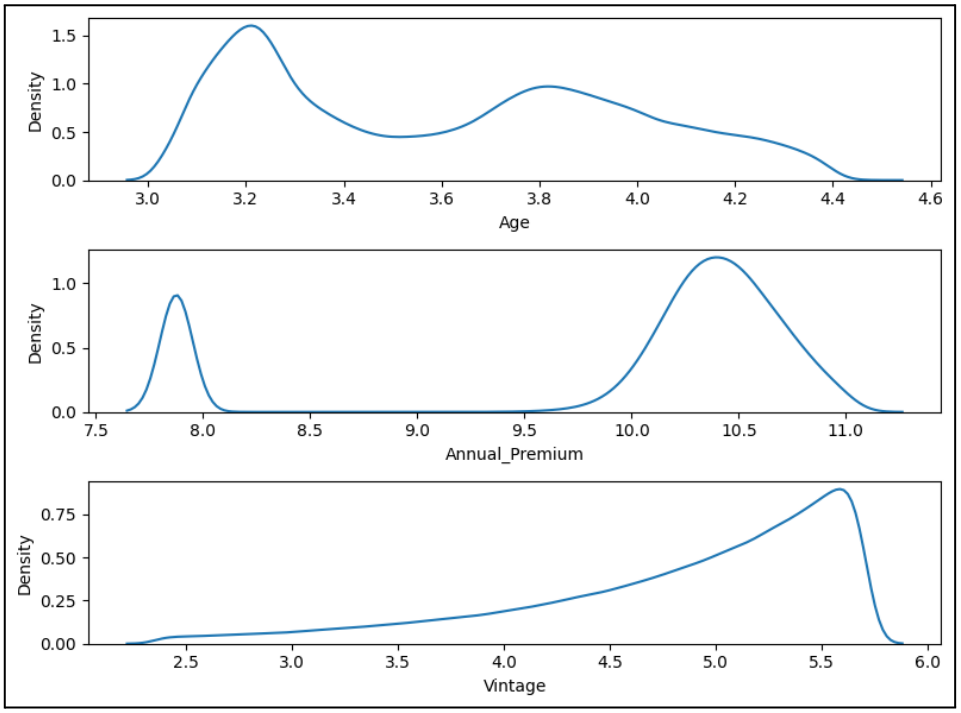
Dapat dilihat outlier sudah terhapus sebanyak 1320, dapat dilihat pada boxplot bahwa range dari data annual premium sudah masuk ke dalam range boxplot yang lama

d. Feature Transformation

Transformasi fitur akan dilakukan pada data yang bertipe numerik.



Dapat dilihat bahwa persebaran fitur numerik pada daset ini belum normal. Selanjutnya, pada annual_premium dapat dilihat bahwa range yang dimiliki sangat berbeda jauh dari kedua fitur numerik lainnya. Maka dari itu, Transformasi fitur akan menggunakan Log Tranformation untuk membuat distribusi fitur lebih simetrik.



Hasil persebaran fitur setelah transformasi log dapat dilihat pada gambar di atas. Dapat dilihat bahwa annual premium memiliki range yang lebih kecil dibandingkan sebelumnya.

e. Feature Encoding

Feature encoding akan dilakukan pada fitur kategorikal yaitu gender, vehicle age, dan vehicle damage. Di bawah ini adalah label encoding untuk fitur kategorikal yang ada

```
mapping_gender = {
    'Male' : 0,
    'Female' : 1
}

mapping_vehicle_age = {
    '< 1 Year' : 0,
    '1-2 Year' : 1,
    '> 2 Years' : 2
}

mapping_vehicle_damage = {
    'No' : 0,
    'Yes' : 1
}
```

Label encoding akan dilakukan pada train dan test agar fitur yang dimasukkan

ke dalam model machine learning dapat direpresentasikan sebagai nilai numerik.

f. Handle Class Imbalance

```
Response
0    325634
1     45155
Name: count, dtype: int64
```

Dapat dilihat, hasil response menunjukkan bahwa terjadi class imbalance antara response 0 dan 1. Pada penelitian ini, kami akan menggunakan undersampling karena ingin menangani ketidakseimbangan kelas dalam dataset. Undersampling membantu mengurangi jumlah instansi dari kelas mayoritas sehingga setiap kelas memiliki jumlah observasi yang lebih seimbang.

Kami melakukan undersampling karena banyak data kelas minoritas masih cukup banyak dan kami beropini bahwa menggunakan data real lebih baik daripada membuat data sintetis menggunakan oversampling

```
Response
0    45155
1    45155
Name: count, dtype: int64
```

Setelah melalui undersampling class response menjadi seimbang.

2. Feature Engineering

a. Feature Extraction

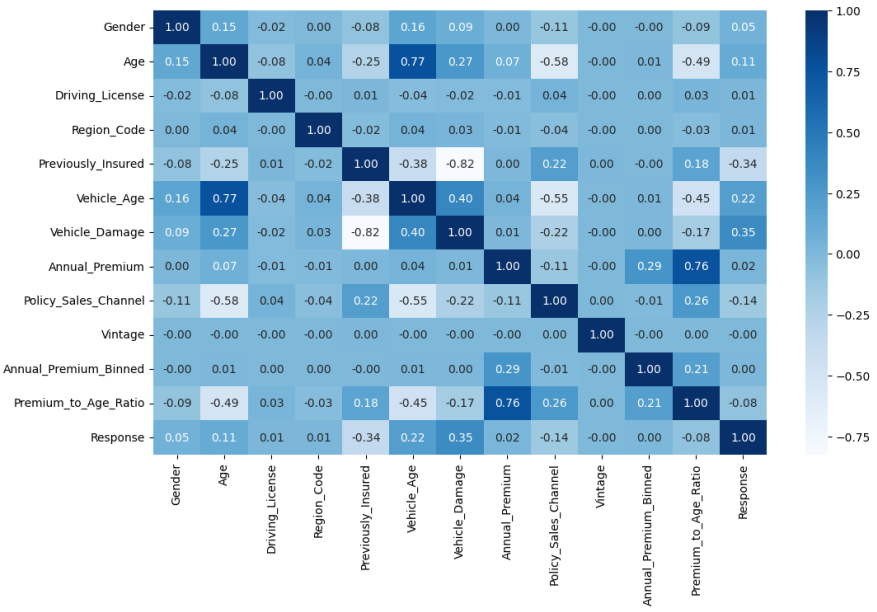
Feature extraction akan dilakukan dengan membuat fitur baru dari fitur yang sudah ada. Feature Extraaxtion akan dilakukan pada data train dan data test. Beberapa fitur baru yang dibuat adalah sebagai berikut ini.

- Vintage Bin
  Fitur ini dilakukan dengan melakukan labelling sesuai dengan quartil dari vintage
- Age Bin
  Fitur ini dilakukan dengan melakukan labelling sesuai dengan quartil dari age
- Premium to Age Ratio
  Fitur ini dibuat dengan membagi annual premium dengan age
- Age and Vehicle Age Interaction
  Fitur ini dibuat dengan mengalikan age bin dengan vehicle age
- Vehicle Damage dan Age Interaction
  Fitur ini dibuat dengan mengalikan age bin dengan vehicle damage

b. Feature Selection

Seleksi fitur akan dilakukan melalui beberapa tahap yaitu correlation matrix, mutual information, dan business insight.

1. **Correlation Matrix**
   Correlation matrix kami gunakan untuk mengetahui korelasi masing-masing fitur terhadap target dan juga korelasi masing-masing fitur dengan fitur yang lainnya. Menggunakan correlation matrix kami dapat menentukan fitur mana yang merupakan multicolinearity. Fitur tersebut adalah fitur yang memiliki korelasi di atas 85%. Fitur ini akan kami hapus karena merupakan fitur yang redundan.



- Age dengan Age Bin (Age Di hapus)

Age dengan Age Bin, bernilai Redundant yaitu 0,92 maka salah satu tersebut harus dihapus dengan pertimbangan diantara kedua feature tersebut mana yang memiliki korelasi paling tinggi terhadap Response

Age terhadap target/ Response = 0.11
Age Bin terhadap target/ Response = 0.29

Maka yang di hapus adalah feature Age

- Vehicle Age dengan Age vehicle Age Interaction (Age vehicle Age Interaction di hapus)

Vehicle Age dengan Age vehicle Age Interaction bernilai Redundant yaitu 0,92 maka salah satu tersebut harus dihapus dengan pertimbangan diantara kedua feature tersebut mana yang memiliki korelasi paling tinggi terhadap Response

Vehicle Age terhadap target/ Response = 0.34
Age vehicle Age Interactionterhadap target/ Response = 0.28

Maka Age vehicle Age Interaction di hapus dari data set

- Vintage dengan Vintage Binned (vintage Binned)

Vintage dengan Vintage Binbernilai Redundant yaitu 0,90 maka salah satu tersebut harus dihapus dengan pertimbangan diantara kedua feature tersebut mana yang memiliki korelasi paling tinggi terhadap Response

Vintage terhadap target/ Response = 0.00
Vintage Binned hadap target/ Response = 0.00

Karena kedua nilai tersebut sama, maka salah satunya harus tetap di hapus yaitu Vintage Binned

- Annual_Premium dengan Premium to Age Ratio (Annual Premium di hapus dari data set)

Premium_to_Age_Ratio dengan Annual_Premium bernilai redundant yaitu 0.76, maka salah satunya harus di hapus dengan pertimbangan diantara kedua feature tersebut mana yang memiliki korelasi paling tinggi terhadap target

Premium_to_Age_Ratio terhadap target/ Response = 0.02
Annual_Premium terhadap target/ Response = -0.08

Maka Annual Premium di hapus dari data set.

2. **Mutual Information**

   Mutual information adalah suatu metrik yang mengukur seberapa banyak informasi tentang satu variabel yang dapat diperoleh dari variabel lainnya. Dalam konteks penentuan fitur (feature selection), mutual information dapat digunakan untuk mengukur seberapa informatif atau relevan suatu fitur terhadap variabel target. Jika mutual information antara suatu fitur dan variabel target tinggi, itu menunjukkan bahwa fitur tersebut memberikan banyak informasi tentang variabel target begitu juga sebaliknya.

   Berdasarkan Mutual Information Fitur yang dibuang adalah:

1. Nilai Mutual Information tidak dilakukan label endoding:
   - Gender: 0.003123553119994682
   - Driving_License: 0.06975209749103684
   - Region_Code: 0.0
   - Previously_Insured: 0.026838975259560804
   - Vehicle_Age: 0.22319677938428217
   - Vehicle_Damage: 0.06264364544182577
   - Policy_Sales_Channel: 0.21121699020381346
   - Vintage: 0.0046138914641638
   - Age_Bin: 0.09602409525249112
   - Premium_to_Age_Ratio: 0.002209415337581788
   - Vehicle_Damage_Age_Interaction: 0.0026198310921099477

2. Nilai Mutual Information dilakukan label endoding:
   - Gender: 0.009850142892536384
   - Driving_License: 0.00888954380142426
   - Region_Code: 0.02342921450209956
   - Previously_Insured: 0.22469129243372254
   - Vehicle_Age: 0.0682555100059814
   - Vehicle_Damage: 0.21959586490041993
   - Policy_Sales_Channel: 0.10064080852039092
   - Vintage: 0.0
   - Age_Bin: 0.0670335023289077
   - Premium_to_Age_Ratio: 0.06730438969765062
   - Vehicle_Damage_Age_Interaction: 0.17833157319911885

   Dapat dilihat dari nilai mutual_info_values dari kedua pengujian tersebut bahwa :

   - feature Vintage memiliki nilai < dari 0.02 di kedua pengujian, yang berarti feature tersebut tidak bisa digunakan untuk prediksi.
   - Feature Premium_to_Age_Ratio memiliki nilai < 0.02 di kedua pengujian, yang berarti feature tersebut tidak bisa digunakan untuk prediksi.

- Feature Gender memiliki nilai < dari 0.02 di kedua pengujian, yang berarti feature tersebut tidak bisa digunakan untuk prediksi.

3. **Business Insight**
   Business Insight merupakan penentu terakhir apakah suatu fitur dibuang atau tidak. Hal ini kami gunakan karena mungkin saja terdapat beberapa fitur yang korelasinya rendah secara data namun secara bisnis seharusnya fitur tersebut memiliki korelasi dengan target. Hal lain yang terjadi mungkin saja fitur tersebut tidak memiliki korelasi yang linear sehingga nilai korelasi dengan target sangat kecil. Maka dari itu business insight penting dalam penentu terakhir selain data yang ada apakah suatu fitur dibuang atau tidak.

- Previously_Insured dengan Vehicle_Damage benilai Redundant yaitu -0.82, akan tetapi di pertahankan dengan pertimbangan :

  kedua feature bernilai Redundant yaitu -0.82 akan tetapi jika berasumsi terhadap kolom tersebut penting dalam menentukan response 0 atau 1 , kalau dilihat dari distribusinya Previously_Insured (customer sudah memiliki asuransi kendaraan atau belum) banyak berpengaruh pada response 1, begitu juga dengan Vehicle_Damage (customer pernah mengalami kerusakan kendaraan atau belum) terhadap response bernilai 1 distribusinya pada data set cukup banyak, bisa jadi karena pernah mengalami kerusakan kendaraan jadi customer tertarik untuk berlangganan asuransi kendaraan.

c. Feature Tambahan

- **Feature Income** diperlukan untuk melihat penghasilan pelanggan yang dapat membantu mengidentifikasi kategori pelanggan yang lebih cocok untuk ditawarkan produk baru, seperti Asuransi Kendaraan Bermotor. Feature ini bertipe numerical
- **Feature Retirement_Savings** diperlukan untuk melihat Pelanggan yang sudah menyimpan uang untuk pensiun dapat dianggap sebagai kandidat yang tepat untuk menawarkan produk asuransi. Feature ini bertipe boolean
- **Feature Vehicle** adalah feature yang berisi berapa banyak pelanggan memiliki kendaraan. Kendaraan yang dimiliki pelanggan dapat memberikan gambaran tentang budget mereka untuk pembelian asuransi. Feature ini bertipe numerical
- **Feature Home_Ownership** adalah feature yang berisi apakah pelanggan memiliki rumah pribadi atau tidak. Status kepemilikan rumah pelanggan dapat memberikan gambaran tentang status keuangan mereka. Misalnya, pemilik rumah dapat dianggap lebih stabil finansial untuk mengambil alat asuransi yang mengandalkan pemilik rumah. Feature ini bertipe boolean

- **Feature Employment_Status** adalah feature yang berisi tentang status pekerjaan yang dimiliki nasabah, seperti contohnya Full-time, Intern, dll. Status pekerjaan pelanggan dapat memberikan informasi tambahan tentang preferensi mereka saat memilih asuransi. Feature bertipe Object