HOMEWORK

FINAL PROJECT

Theofilus Arifin

Christofer Bryan N. K.

Ramlan Apriyansyah

Muhammad Iqbal

Hanifah Arrasyidah

Christopher Stephen

Muhammad Rizq N. A.

Ujang Pian

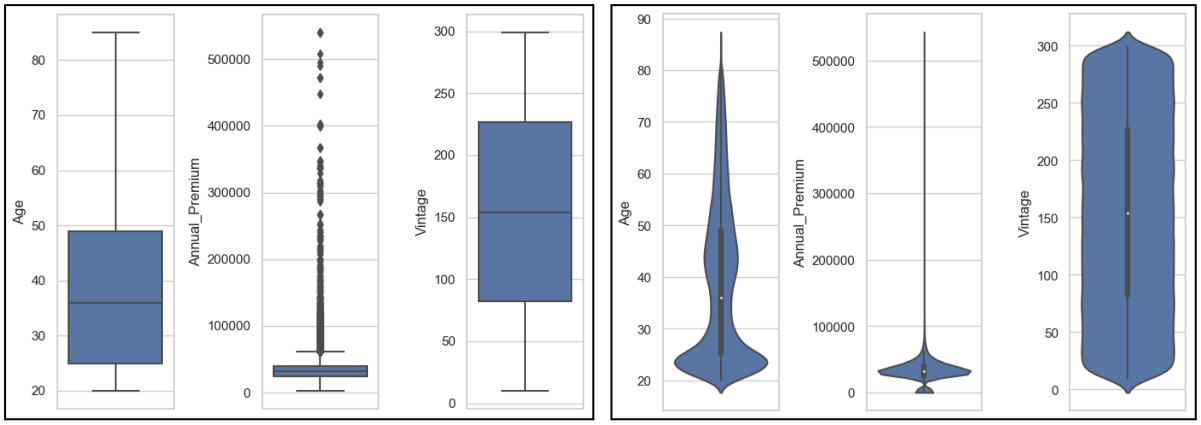
I. Descriptive Statistics

1. Semua tipe kolom pada dataset sudah sesuai
2. Tidak ada kolom yang memiliki nilai kosong
3. Berikut merupakan beberapa penjelasan summary dari masing-masing kolom:
    - Pada kolom Age dan Vintage nilai mean dan median tidak memiliki gap yang signifikan, menunjukkan bahwa distribusi cenderung normal
    - Pada kolom Annual_Premium nilai mean dan median memiliki gap yang signifikan, menunjukkan adanya outlier atau skewed distribution
    - Pada kolom driving license nilai "1" yang sangat dominan dibandingkan nilai "0"
    - Pada kolom response nilai "0" yang sangat dominan dibandingkan nilai "1"

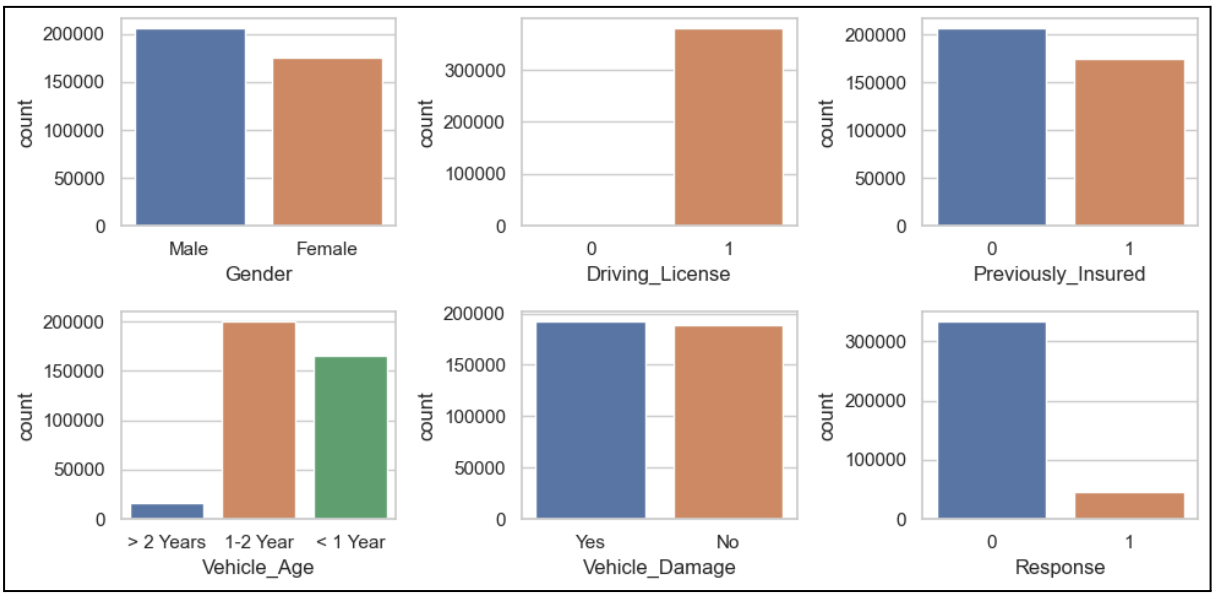
II. Univariate Analysis

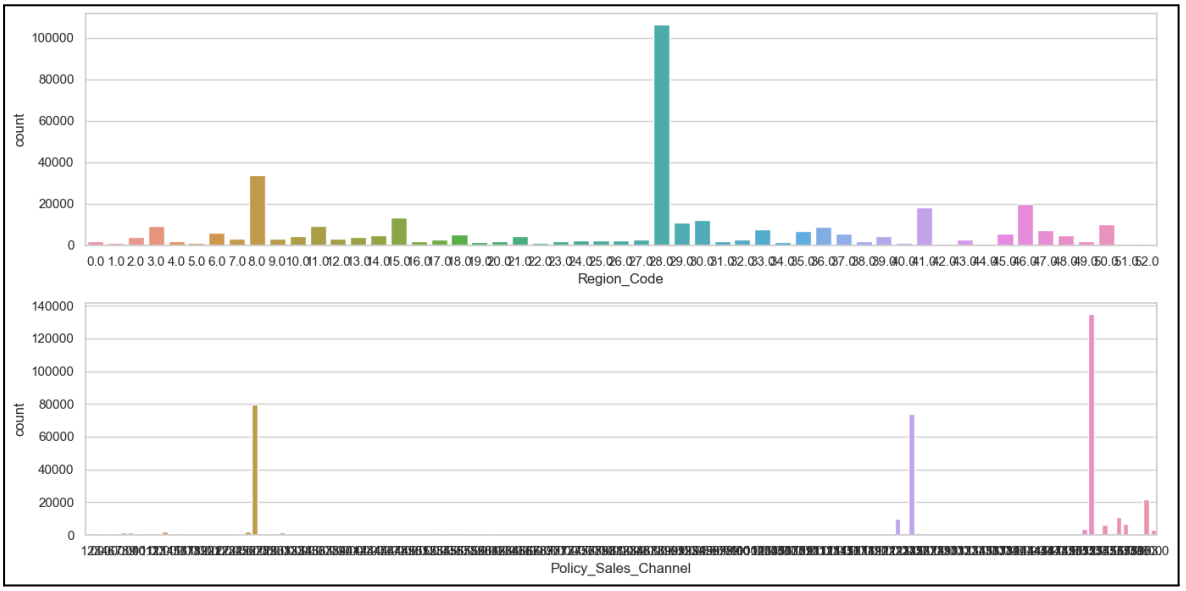
Visualisasi

1. Data Numerik



2. Data Kategorikal





Penjelasan Distribusi

1. Age

Age adalah usia pelanggan dalam tahun. Nilai minimum dan maksimum tampak masuk akal (20 hingga 85 tahun). Persebaran paling dominan pada umur 20-30 tahun. Tidak ada outlier yang jelas dalam data ini.

2. Annual Premium

Annual Premiumm adalah Premi Tahunan yang dibayar oleh nasabah. Ada variasi yang cukup besar dalam data ini. Dari boxplot, kita dapat melihat bahwa ada beberapa nilai yang jauh dari kuartil atas, yang menunjukkan adanya outlier.

3. Vintage

Vintage adalah banyak hari nasabah telah terasosiasi dengan perusahaan. Nilai minimum adalah 10 hari dan Nilai maksimum adalah 299 hari, yang berarti ada pelanggan yang baru bergabung dan ada yang telah hampir satu tahun. Dari histogram, kita dapat melihat bahwa distribusi 'Vintage' hampir seragam, yang berarti setiap durasi asosiasi memiliki jumlah pelanggan yang hampir sama. Tidak ada durasi asosiasi tertentu yang jauh lebih umum daripada yang lain.

4. Driving License

Driving License menunjukkan apakah pelanggan memiliki surat izin mengemudi atau tidak, dengan 1 berarti 'Ya' dan 0 berarti 'Tidak'. Sebagian besar pelanggan (380297 dari 381109) yaitu 99.79 memiliki surat izin mengemudi

5. Region Code

Region Code adalah Kode wilayah yang merupakan variabel kategorikal yang telah diencode sebagai numerik. Data paling banyak berada pada region code 28.

6. Previously Insured

Previously Insured menunjukkan apakah pelanggan sudah memiliki asuransi kendaraan atau tidak. Dapat dilihat bahwa persebaran antara pelanggan yang telah memiliki dan belum mmemiliki asuransi kendaraan tidak jauh berbeda.

7. Policy Sales Channel
Policy Sales Channel adalah Kode untuk suatu channel untuk menghubungi nasabah yang merupakan variabel kategorikal yang telah diencode sebagai numerik. Data paling banyak berada pada Policy Sales Channel 152.

8. Gender
Gender merupakan jenis kelamin nasabah. Dapat dilihat pada countplot persebaran gender nasabah sudah cukup seimbang.

9. Vehicle Age
Vehicle age merupakan umur kendaraan nasabah. Dapat dilihat bahwa kendaraan yang berumur di atas 2 tahun sangat sedikit jika dibangingkan dengan kendaraan yang berumur 1 tahun atau 1-2 tahun

10. Vehicle Damage
Vehiclle damage menunjukkan apakah kendaraan suatu nasabah pernah kecelakaan di masa lalu. Dapat dilihat pada countplot persebaran vehicle damage sudah cukup seimbang.

11. Response
Response adalah target label yang menunjukan apakah suatu nasabah tertarik atau tidak untuk membeli asuransi kendaraan. Persebaran response di dominasi oleh nilai 0 dan sangat sedikit nilai 1 pada data

Preprocessing Follow-Up

1. Age
Kita bisa mempertimbangkan untuk melakukan binning atau pengelompokan usia ke dalam beberapa kategori (misalnya: 'Muda', 'Dewasa', 'Lansia') untuk memudahkan interpretasi.

2. Annual Premium
Kita mungkin perlu melakukan transformasi, seperti log transform, untuk mengurangi skewness. Selanjutnya penghapusan outlier menggunakan IQR juga bisa dilakukan.

3. Vintage
Kita bisa mengubah 'Vintage' menjadi fitur kategorikal dengan membaginya ke dalam beberapa 'bin' atau kategori. Misalnya, kita bisa membuat kategori 'kurang dari 50 hari', '50-100 hari', '100-150 hari', dan seterusnya.

4. Driving License
Mengingat sebagian besar pelanggan memiliki surat izin mengemudi, fitur ini mungkin tidak memberikan banyak informasi untuk memprediksi respons sehingga mungkin bisa dihapus.

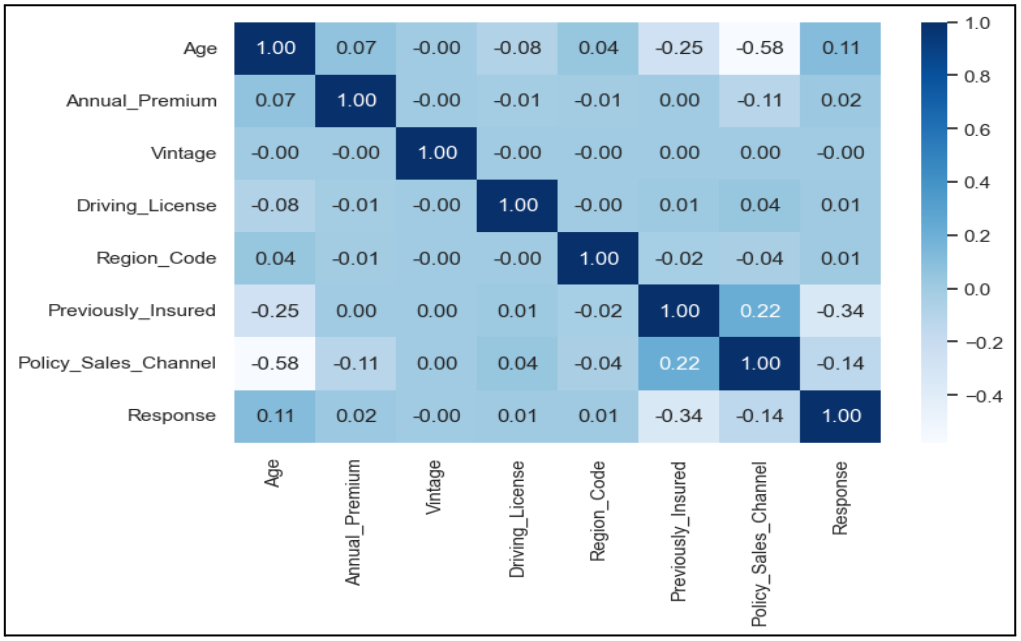
5. Region Code
kita bisa mempertimbangkan untuk mengelompokkan wilayah berdasarkan frekuensi. Misalnya, kita bisa memiliki kategori 'wilayah berfrekuensi tinggi', 'wilayah berfrekuensi sedang', dan 'wilayah berfrekuensi rendah'.

6. Policy Sales Channel
kita bisa mempertimbangkan untuk mengelompokkan channel berdasarkan frekuensi. Misalnya, kita bisa memiliki kategori 'channel berfrekuensi tinggi', 'channel berfrekuensi sedang', dan 'channel berfrekuensi rendah'.

7. Response
Perlu dilakukan over sampling untuk mengatasi class imbalance pada response.

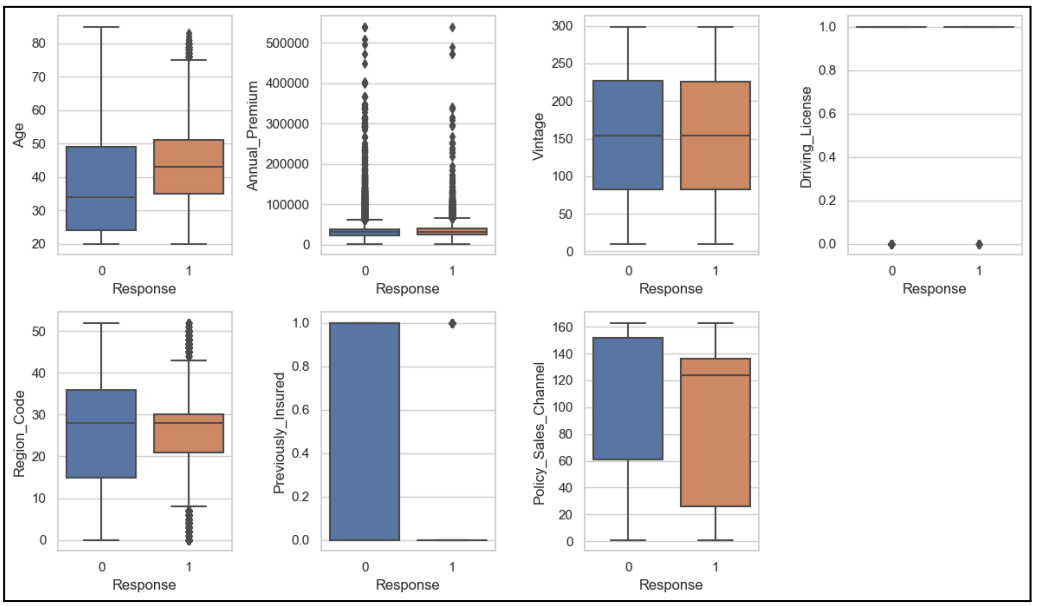
III. Multivariate Analysis

A. Correlation Matrix



Berdasarkan correlation matrix, belum ada korelasi yang sangat kuat antara Response dengan kolom lainnya. Namun previously insured, policy sales channel, dan juga age sedikit berpengaruh pada Response.

B. Boxplot



- Age dan Response

  Pada boxplot diatas nilai response 0 dan 1 berdasarkan usia, Response 0 memiliki persebaran di rentang usia 23 - 49 tahun, sedangkan Response 1 memiliki persebaran di rentang 35 -52 tahun, dan pada kedua response tersebut ada beberapa usia yang berada pada garis outlier, sehingga dapat disimpulkan respone 1 berdasarkan usia memiliki rata-rata umur 35-52 tahun (dewasa -lans

- Annual_Premium dan Response

  Pada boxplot diatas Respone 0 dan 1 berdasarkan Annual_Premium memiliki nilai yang cukup seimbang yaitu di memiliki Annual_Premium di bawah 100000 dan bebrapa outlier pada kedua Response tersebut.

- Vintage dan Response

  Pada Boxplot diatas Response 0 dan 1 Berdasarkan Vintage, lamanya nasabah yang terasosiasi dengan perusahaan memiliki jangka waktu yang seimbang yaitu berada pada rentang nilai 230 hari

- Driving License dan Response

  Pada Boxplot diatas, tidak banyak informasi yang di dapat, antara Driving license dan Response tidak terlihat jelas persebaran nya .

- Region Code dan Response

  Pada Boxplot diatas, response 1 berdasarkan Region Code berada pada rentang code 21 - 30 dan 50% ada pada rentang code 21-28 serta ada beberapa nilai outlier.

- Previously_Insured dan Response

  Pada Boxplot diatas, Response 0 berdasarkan Previously_Insured itu mendominasi dibanding dengan response 1, hal tersebut menandakan bahwa yang response 1 atau pelanggan yang tertarik sebelumnya tidak memiliki asuransidaraan sebelumnya.

- Policy_Sales_Channel dan Response

  Pada Boxplot diatas, Response 1 berdasarkan Policy_Sales_Channel memiliki persebaran code yang cukup beragam yaitu code 25-135, tetapi 50% pada Response 1 tersebut memiliki rentang code 25-125.

IV. Business Insight

- Customer di dominasi oleh laki-laki, baik yang tertarik ataupun yang tidak.
- Seluruh customer memiliki Driving License
- Seluruh customer yang tertarik dengan asuransi kendaraan adalah customer yang belum memiliki asuransi kendaraan sebelumnya. Namun jumlah tersebut hanya 23% dari total customer yang belum memiliki asuransi kendaraan.
- Kelompok Customer yang memiliki kendaraan dengan umur 1 - 2 tahun adalah kelompok customer terbanyak.
- Jumlah Customer yang tidak pernah mengalami kecelakaan melebih jumlah customer yang pernah mengalami kecelakaan. Dan hampir seluruh customer yang tertarik dengan asuransi kendaraan adalah customer yang pernah mengalami kecelakaan.
- Sebagian besar customers yang tertarik dengan asuransi kendaraan berasal dari Region Code 28, yaitu sebanyak 19.917 customers, atau 18.7% dari total customers yang tinggal di Region tersebut.
- Dari segi proporsi, Region Code 38 adalah yang tertinggi, yaitu sebesar 19,2%. Namun jumlah customer yang tertarik sangat sedikit. Hal tersebut dikarenakan total customer di Region tersebut jauh lebih sedikit dibanding total customer di Region Code 28.
- Policy Sales Channel yang memperoleh customer asuransi kendaraan terbanyak adalah Channel 26, yaitu sekitar 16.000 customers, atau 19,9% dari total customer yang ditawarkan melalui Channel tersebut. Namun dari segi proporsi, Channel 157 memiliki tingkat konversi yang paling tinggi, yaitu sebesar 26,8%