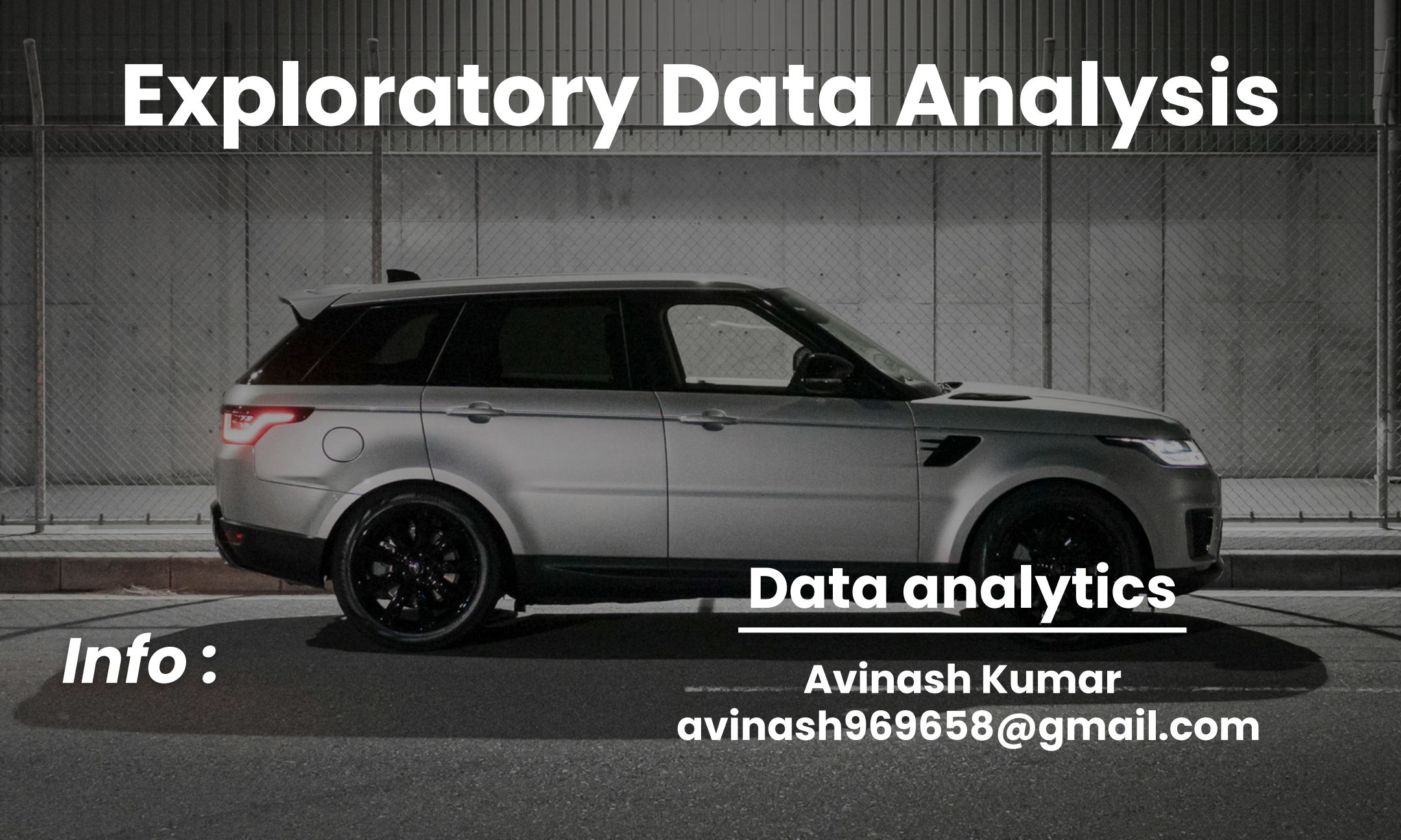
VEHICLE INSURANCE

Exploratory Data Analysis



Data analytics

Info :

Avinash Kumar
avinash969658@gmail.com

Vehicle Insurance

Tools

NumPy

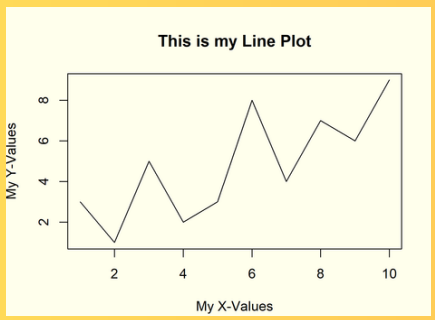
pandas

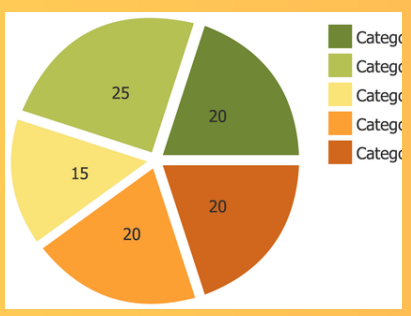
seaborn

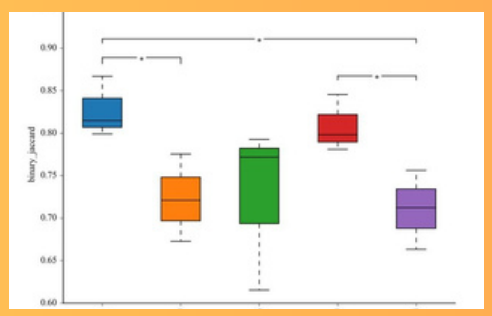
plotly

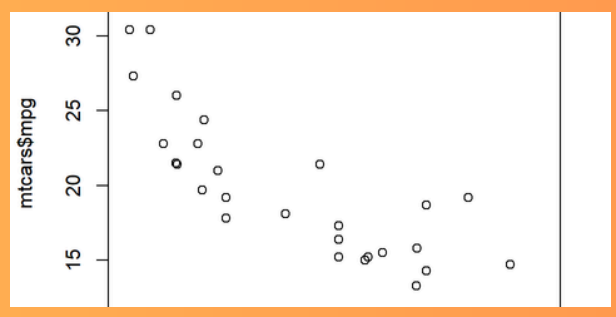
matplotlib

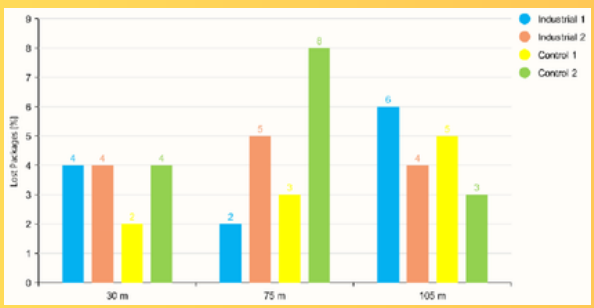
Graph

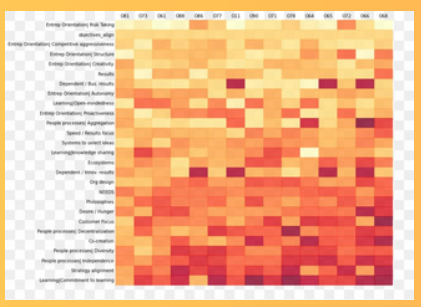


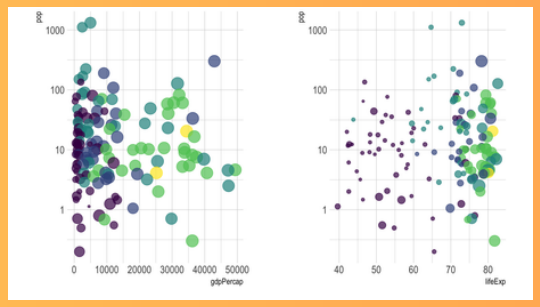












About the Insurance

Vehicle insurance is a financial protection plan that provides coverage against losses or damages to a vehicle (such as a car, bike, or truck) due to accidents, theft, or other unforeseen events. It also protects the vehicle owner and third parties (other people or property) involved in road accidents.

Vehicle insurance is a contract between the vehicle owner and the insurance company. Under this agreement, the insurer promises to compensate the policyholder for losses or damages to the vehicle during the policy period, as per the terms and conditions.

- **This is mandatory by law in most countries (including India).**
- **It covers the damage or injury caused to another person or property due to your vehicle.**
- **It does not cover damage to your own vehicle.**

- **This policy provides complete protection.**
- **It includes third-party coverage plus coverage for your own vehicle.**
- **It protects against accidents, theft, fire, and natural disasters.**

- **Covers only the damage to your own vehicle.**
- **Useful if you already have third-party coverage.**

Vehicle insurance is not just a legal requirement — it's a financial safety net that protects you from unexpected expenses due to accidents, theft, or natural disasters. Choosing the right insurance policy ensures both financial security and peace of mind for every vehicle owner.

Description

- Loading and Preparing vehicle insaurance Dataset for Analysis in Google Colab

```
[ ]  from google.colab import drive
     drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

- **in this analysis , we import Pandas for data manipulation ,Numpy for numerical operations ,Matplotlib for creating visualization ,and Seaborn and plotly for enhanced statistical graphic.**

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import plotly.express as px
```

- We import the Google drive module to mount Google `Drive`, enabiling access **to files and dataset in the for our analysis.**

```
df=pd.read_csv("/content/drive/MyDrive/Car vehicle/Vehicle_Insurance.csv")
```

- **We use pandas to Read the CSV file containing big vehicle insaurance from Google Drive, loading it into a DATAFRAME Name 'df' for analysis.**

Data overview

```
df
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response
0	1	Male	44	1	28.0	0	> 2 Years	Yes	40454.0	26.0	217	1
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183	0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27	1
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203	0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39	0
...	...	...	...	...	...	...	...	...	...	...	...	...
381104	381105	Male	74	1	26.0	1	1-2 Year	No	30170.0	26.0	88	0
381105	381106	Male	30	1	37.0	1	< 1 Year	No	40016.0	152.0	131	0
381106	381107	Male	21	1	30.0	1	< 1 Year	No	35118.0	160.0	161	0
381107	381108	Female	68	1	14.0	0	> 2 Years	Yes	44617.0	124.0	74	0

- **The dataset comprises of 381109 rows and 12 columns.**

```
8]
0s
f"columns:{df.shape[0]},Rows:{df.shape[1]}"
```

```
'columns:381109,Rows:12'
```

```
df.drop_duplicates()
```

Description

The dataset being analysed offers an extensive array of information about vehicle insurance, including various details about insured individuals, their vehicles, and the insurance claims associated with them. The following is a thorough overview of the dataset's main components and variables:

- Age: The age of the insured person, which reflects their life stage and possible risk profile.
- Gender: The gender of the insured person, which could impact insurance premiums and the frequency of claims.
- Driving License: The status or type of driving license held by the insured individual, which may include factors such as the license's validity, class, and any endorsements or restrictions.
- Region Code: The geographic location of the insured individual, reflecting regional differences in risk factors and claim rates.

```
df.info()
```

```
<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
```

Description

- **Previously Insured: This refers to whether the insured individual had prior vehicle insurance coverage before their current policy. It can indicate their previous insurance history and may influence risk assessment and policy terms.**

- **Vehicle Age: The age or model year of the insured vehicle, indicating its condition and the potential risk of accidents or damage.**

- **Vehicle Damage: Details about the vehicle's damage status, which can influence insurance premiums and the frequency of claims.**

- **Annual Premium: The total amount of money paid by the insured individual for their vehicle insurance policy over the course of a year.**

- **Policy Sales Channel: The method or platform through which the insurance policy was sold to the insured individual. This could include channels such as direct sales, brokers, online platforms, or agents.**

- **Vintage: Typically refers to a vehicle that is at least 20 to 30 years old, depending on the definition used by different organizations or regions.**

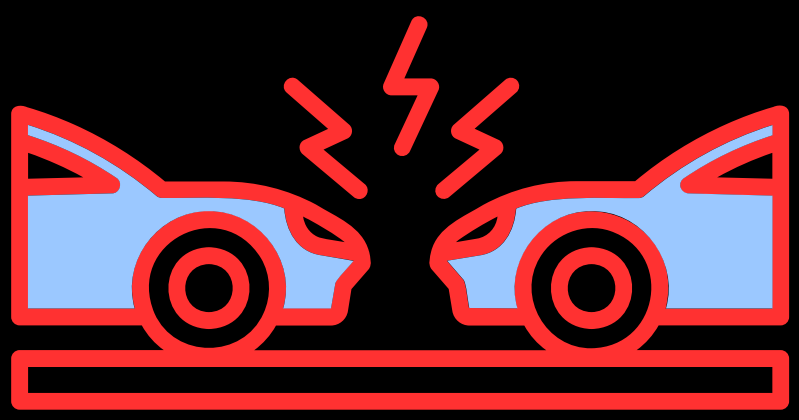
- **Response: In a general context, a response is an answer or reaction to a question, request, or stimulus.**

Describe

```
df.describe()
```

	id	Age	Region_Code	Annual_Premium	Policy_Sales_Channel	Vintage
count	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000	381109.000000
mean	190555.000000	38.822584	26.388807	30564.389581	112.034295	154.347397
std	110016.836208	15.511611	13.229888	17213.155057	54.203995	83.671304
min	1.000000	20.000000	0.000000	2630.000000	1.000000	10.000000
25%	95278.000000	25.000000	15.000000	24405.000000	29.000000	82.000000
50%	190555.000000	36.000000	28.000000	31669.000000	133.000000	154.000000
75%	285832.000000	49.000000	35.000000	39400.000000	152.000000	227.000000
max	381109.000000	85.000000	52.000000	540165.000000	163.000000	299.000000

- Dataset has 381,109 total records and no missing values in numeric columns.
- Age distribution shows most customers are middle-aged.
- Annual_Premium has a wide range, so scaling or outlier handling may be needed.
- Region_Code and Policy_Sales_Channel are likely categorical features represented numerically.
- Vintage is fairly spread — indicates different customer tenures.



Data cleaning & pre_processing

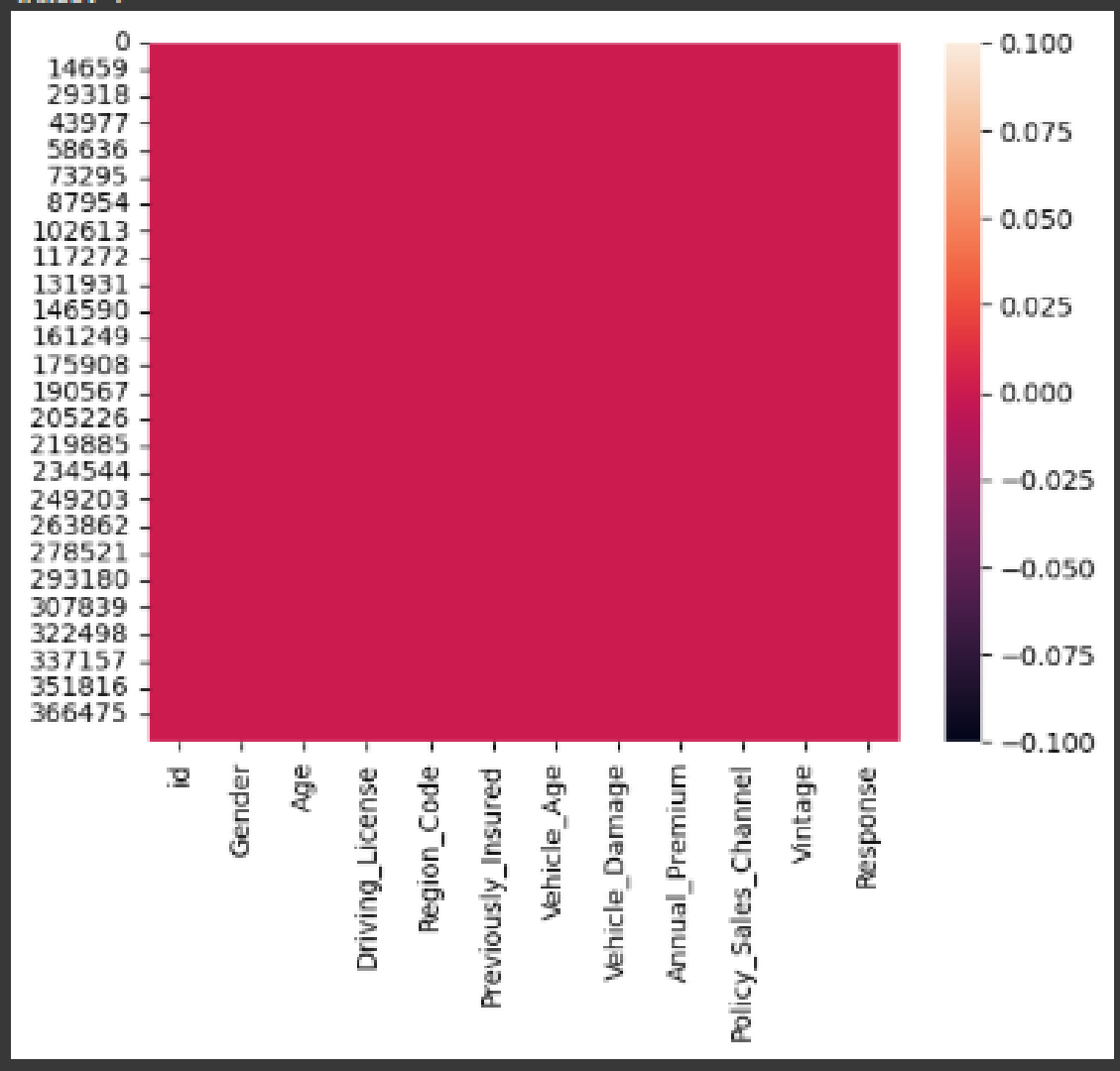
- **As we have seen that this data set has 0 null values. So we can move to the next step**

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

	0
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

```
sns.heatmap(df.isnull())
```



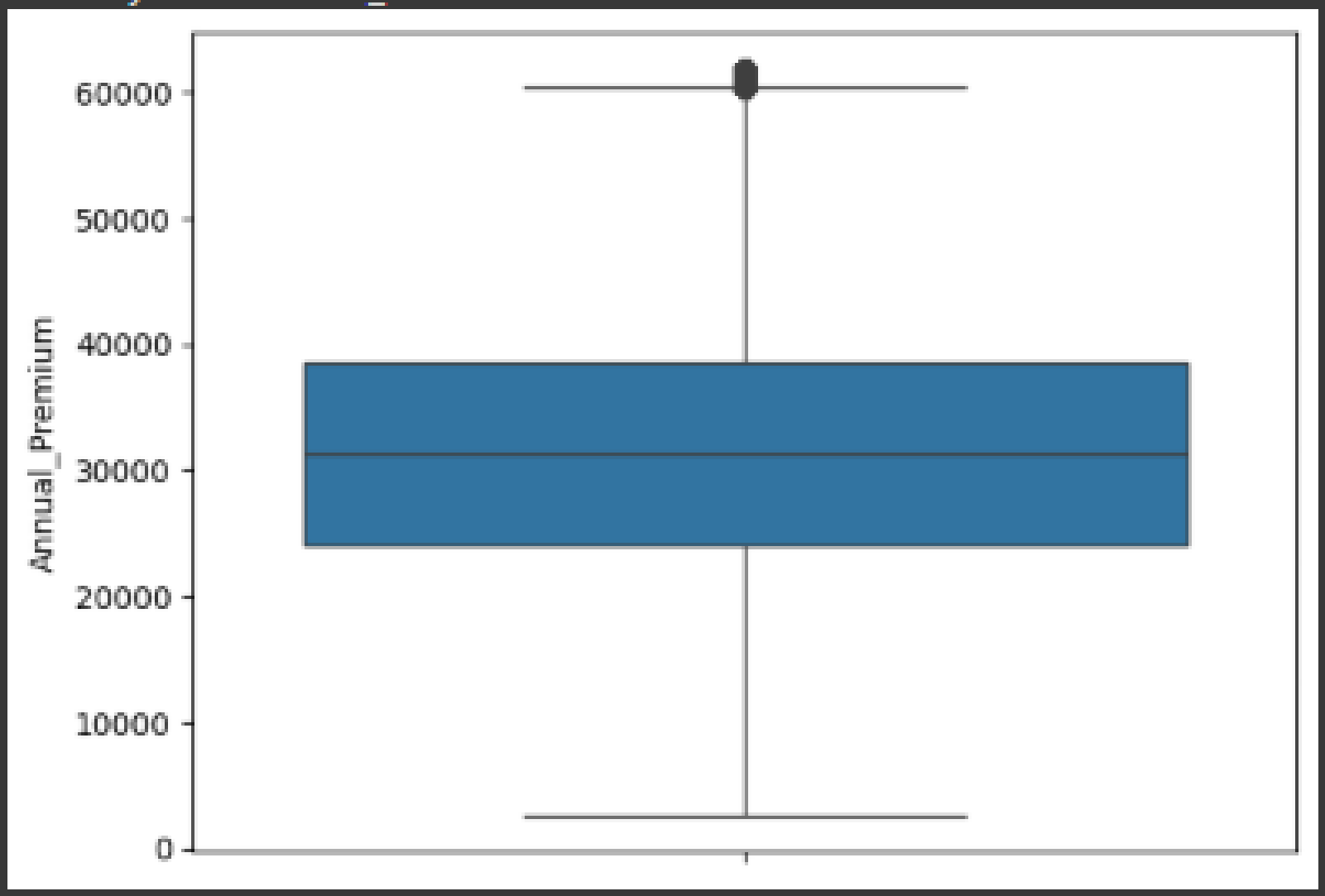
Data cleaning & pre_processing

- **Key consideration : Although we have 0 null values, we still need to examine numerical features- 'Annual Premium' for outliers to ensure data balance and improve the quality of insights . ▪ Using IQR method**

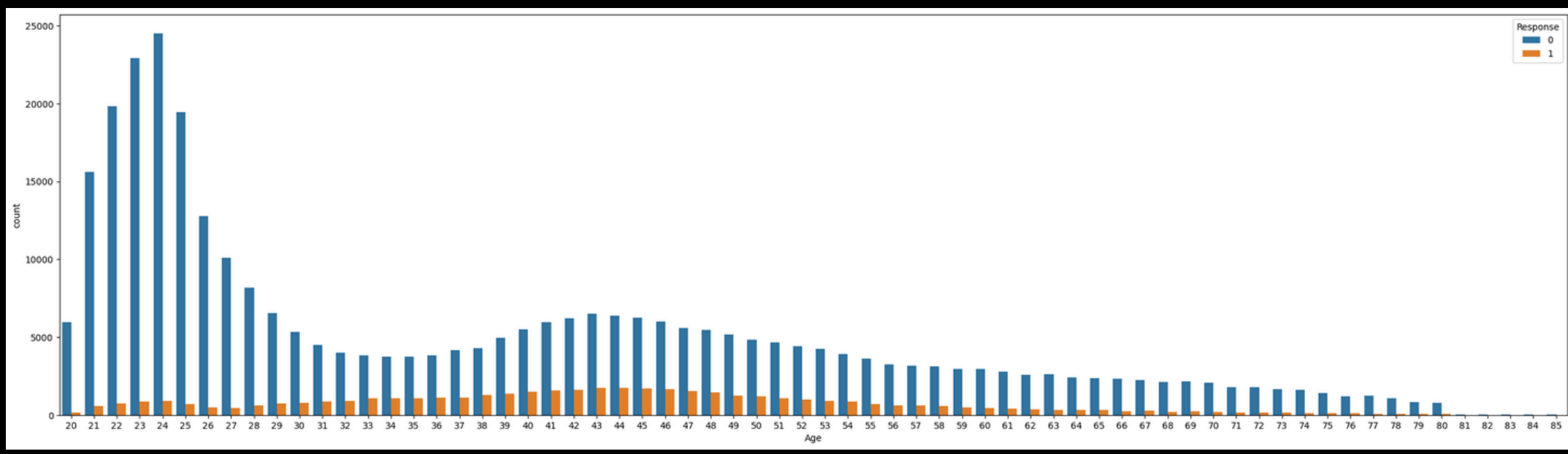
```
Q1=df["Annual_Premium"].quantile(0.25)
Q3=df["Annual_Premium"].quantile(0.75)
IQR=Q3-Q1
# define the lower & uper bounds
lower_bond=Q1-1.5*IQR
upper_bond=Q3+1.5*IQR
# fillter outlier
df=df[(df["Annual_Premium"]>=lower_bond) & (df["Annual_Premium"]<=upper_bond)]
```

```
sns.boxplot(df["Annual_Premium"])
```

```
<Axes: ylabel='Annual_Premium'>
```



Data visulization and insights

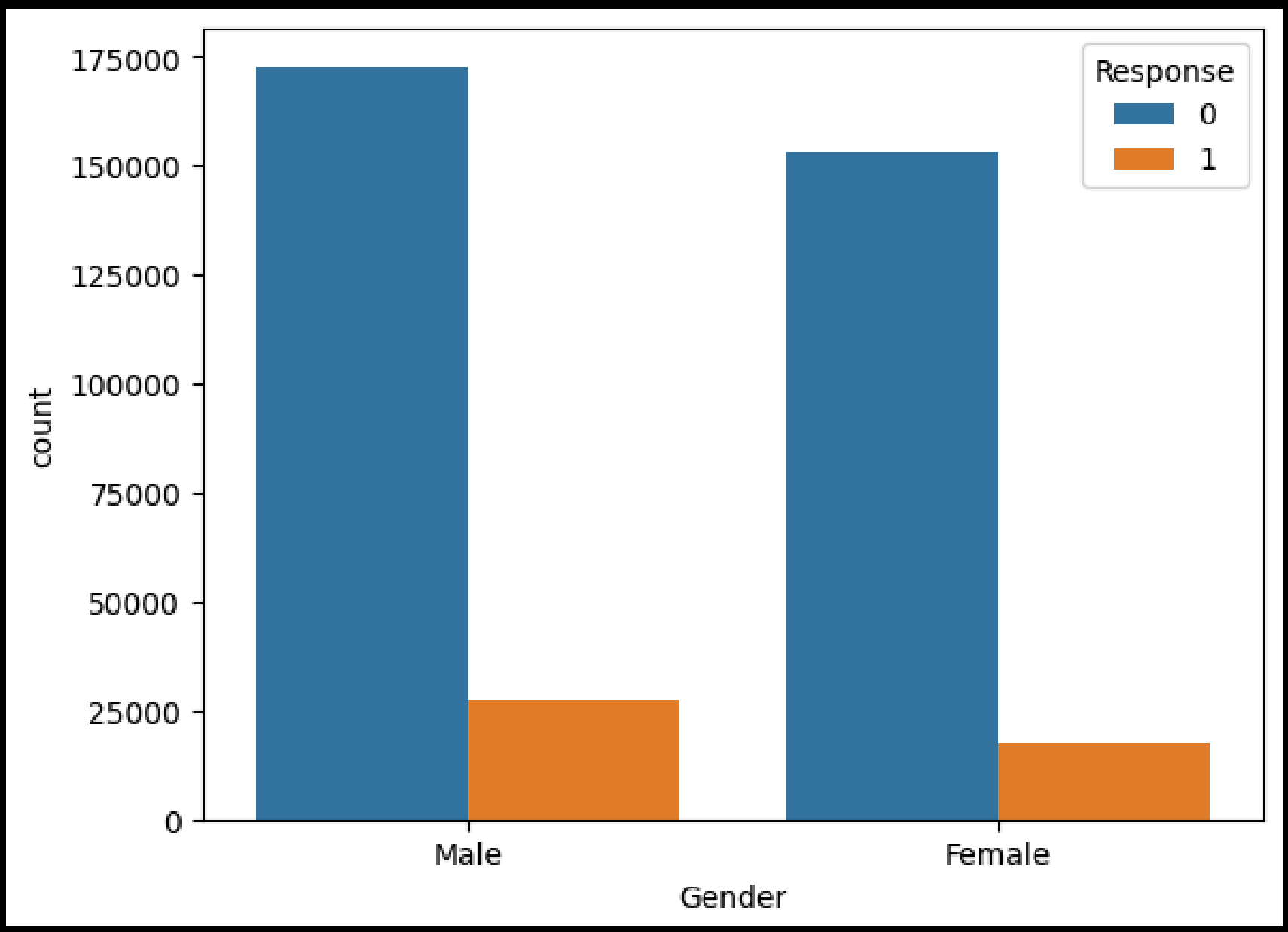


- **Age Distribution** : The dataset has a wide age range, from around 20 to 80 years old. The majority of individuals fall within the 20-50 age group, with the peak around the 30-40 age range.

- **Response Variable** : The graph shows two distinct categories for the "Response" variable (0 and 1). We can infer that this variable represents a binary outcome, such as whether a customer made a purchase or not.

- **Response by Age** : The graph suggests that the response rate varies across different age groups. For instance, the response rate appears higher in the 30-50 age range compared to the younger and older age groups

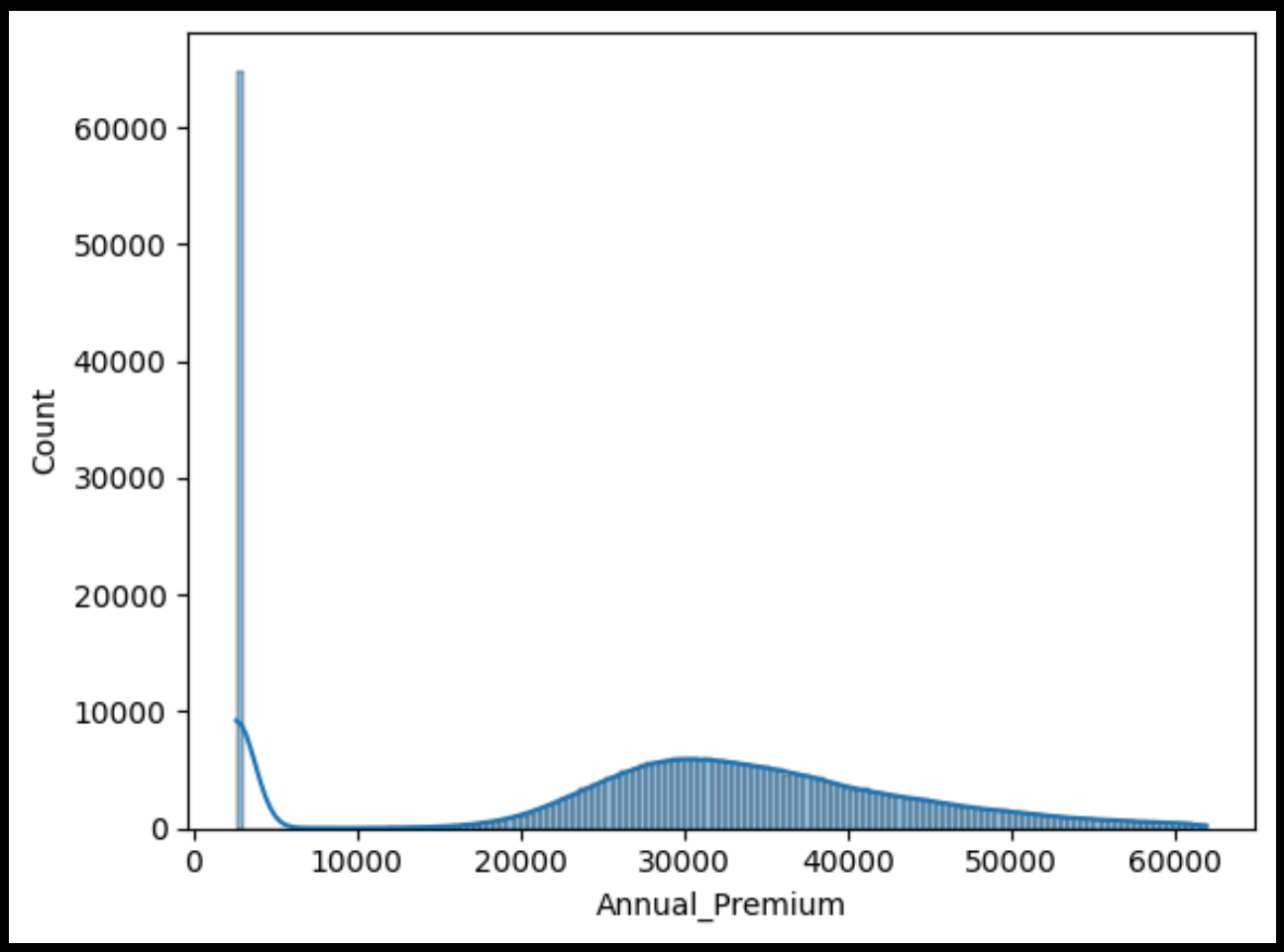
Data visulization and insights



- **More males responded** : The blue bars, representing response '0', are significantly higher for males than females, indicating that more males participated in the survey or provided this response.

- **Similar response patterns** : The proportion of '1' responses (orange bars) to '0' responses seems consistent across both genders, suggesting that there might not be a significant difference in the way males and females responded to the question or survey.

Data visulization and insights



Right-Skewed Distribution (Positive Skewness)

- Most of the data is concentrated on the left side (lower premium values).
- A long tail extends to the right, showing a few customers with very high premiums.

High Frequency of Low Premiums

- A large spike near 0–5000 indicates that many customers are paying very low annual premiums.
- This may represent entry-level or basic insurance plans.

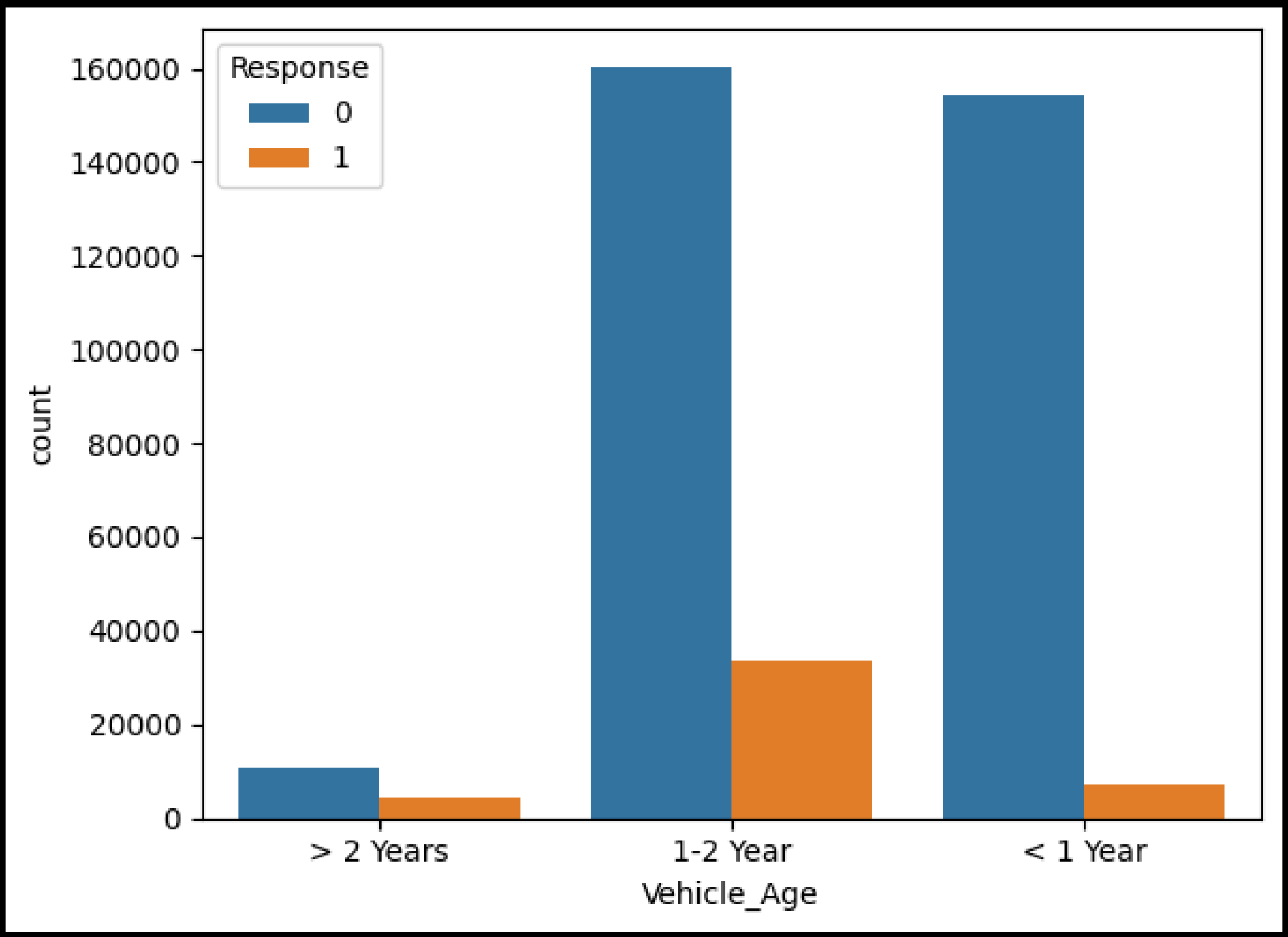
Majority of Premiums Cluster Around 25,000–35,000

- There is a smooth hump (peak) around this range — suggesting most customers pay moderate premiums.
- This range likely represents the average or standard insurance product.

Presence of Outliers

- A few extreme values go beyond ₹60,000 (and some even higher if not capped).
- Indicates outliers that could distort model training — might need scaling or capping (winsorization).

Data visulization and insights

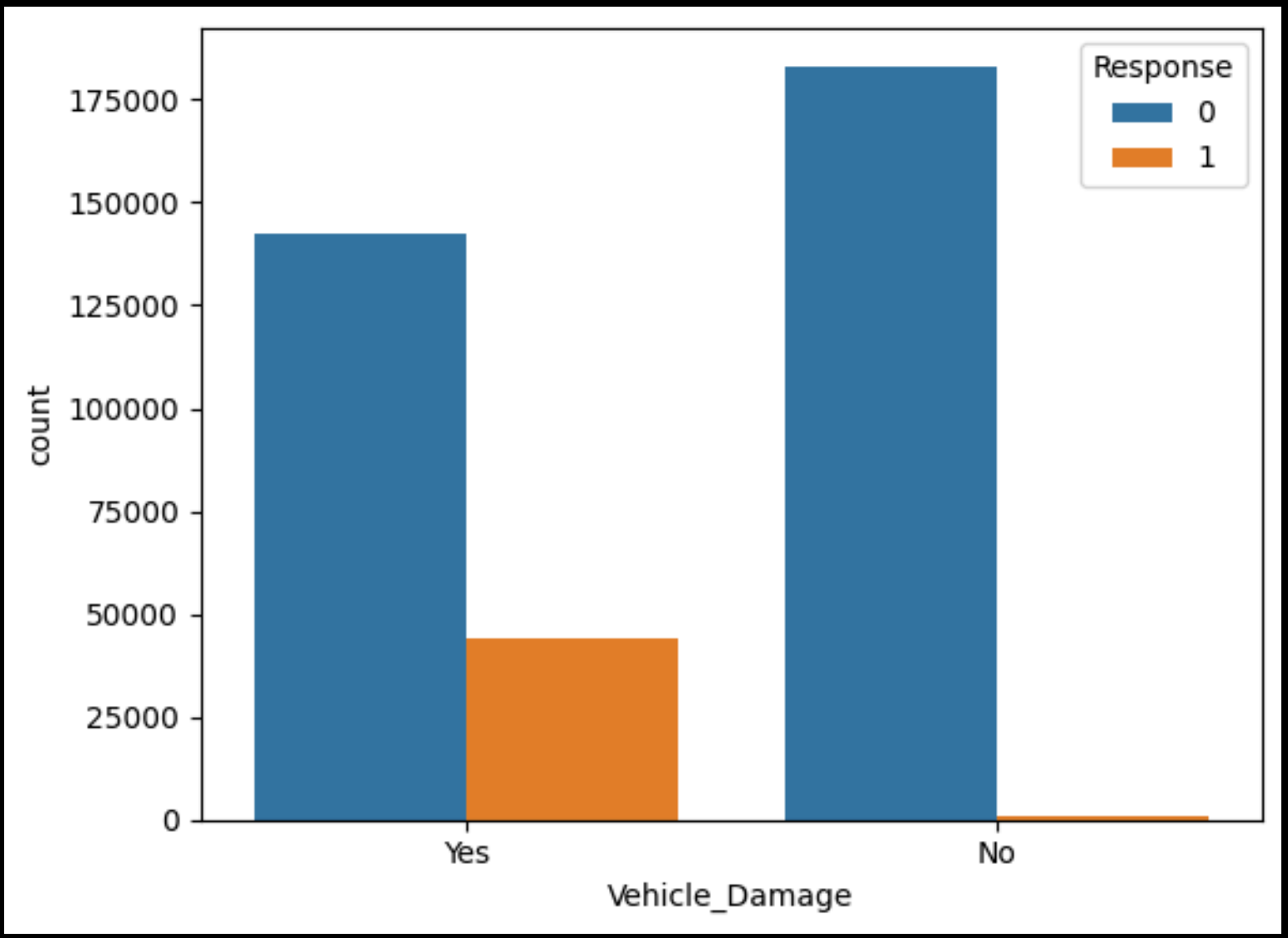


- **Most vehicles are less than 1 year old. This category has the highest count for both response values.**

- **Response 0 is more frequent overall. For each vehicle age group, the count for response 0 is higher than response 1.**

- **Response 1 is relatively more common for vehicles older than 2 years. While response 0 still dominates, the proportion of response 1 is slightly higher in the oldest age group compared to the other two.**

Data visulization and insights



- **More loans are approved for those without vehicle damage.**

The "No" category under "Vehicle Damage" has significantly higher counts for both approved (Response = 1) and rejected (Response = 0) loans.

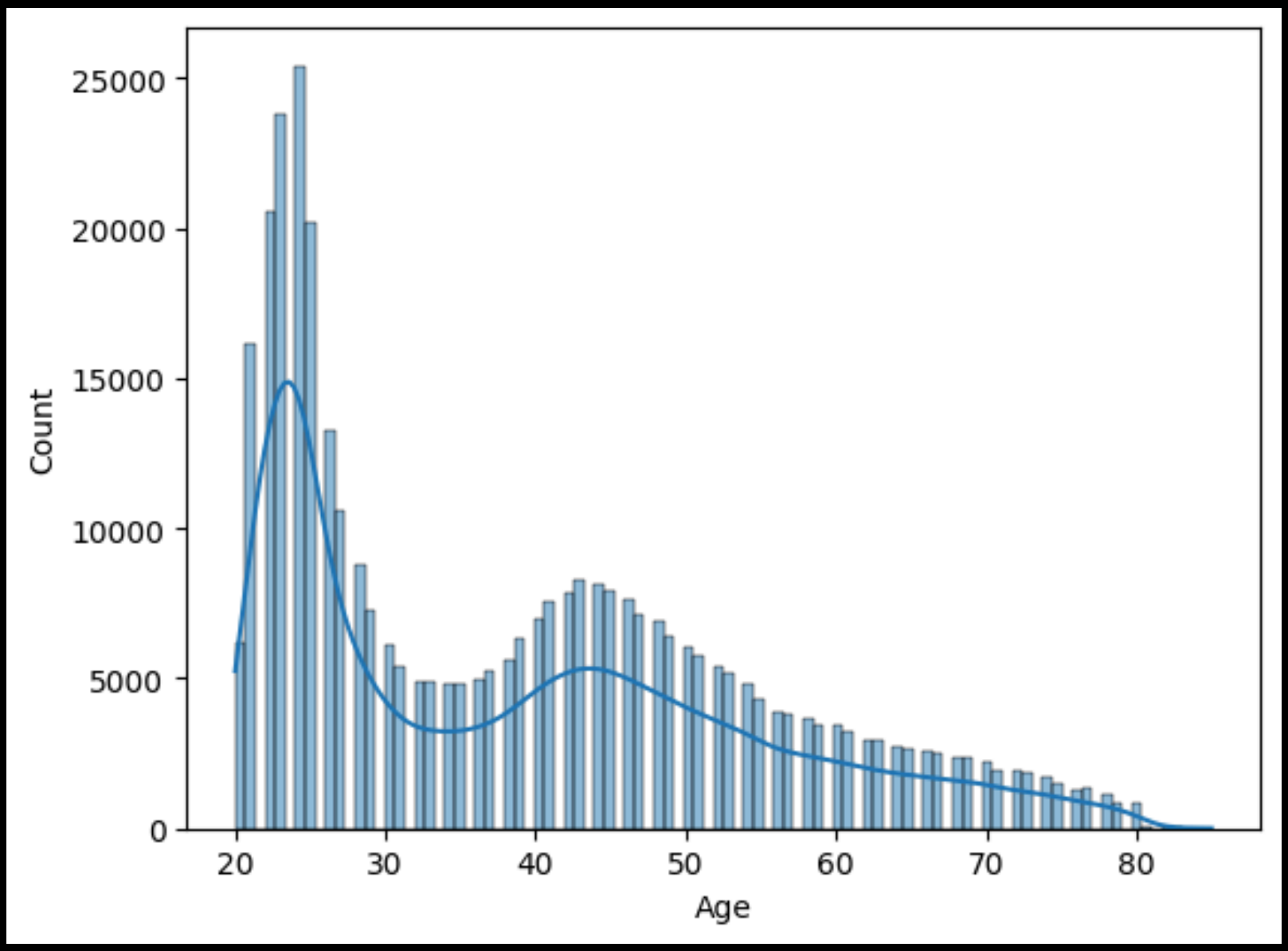
- **Vehicle damage negatively impacts loan approval .**

While many loans are still approved for those with vehicle damage, the proportion of approvals is lower compared to those without vehicle damage.

- **The majority of loan applicants do not have vehicle damage .**

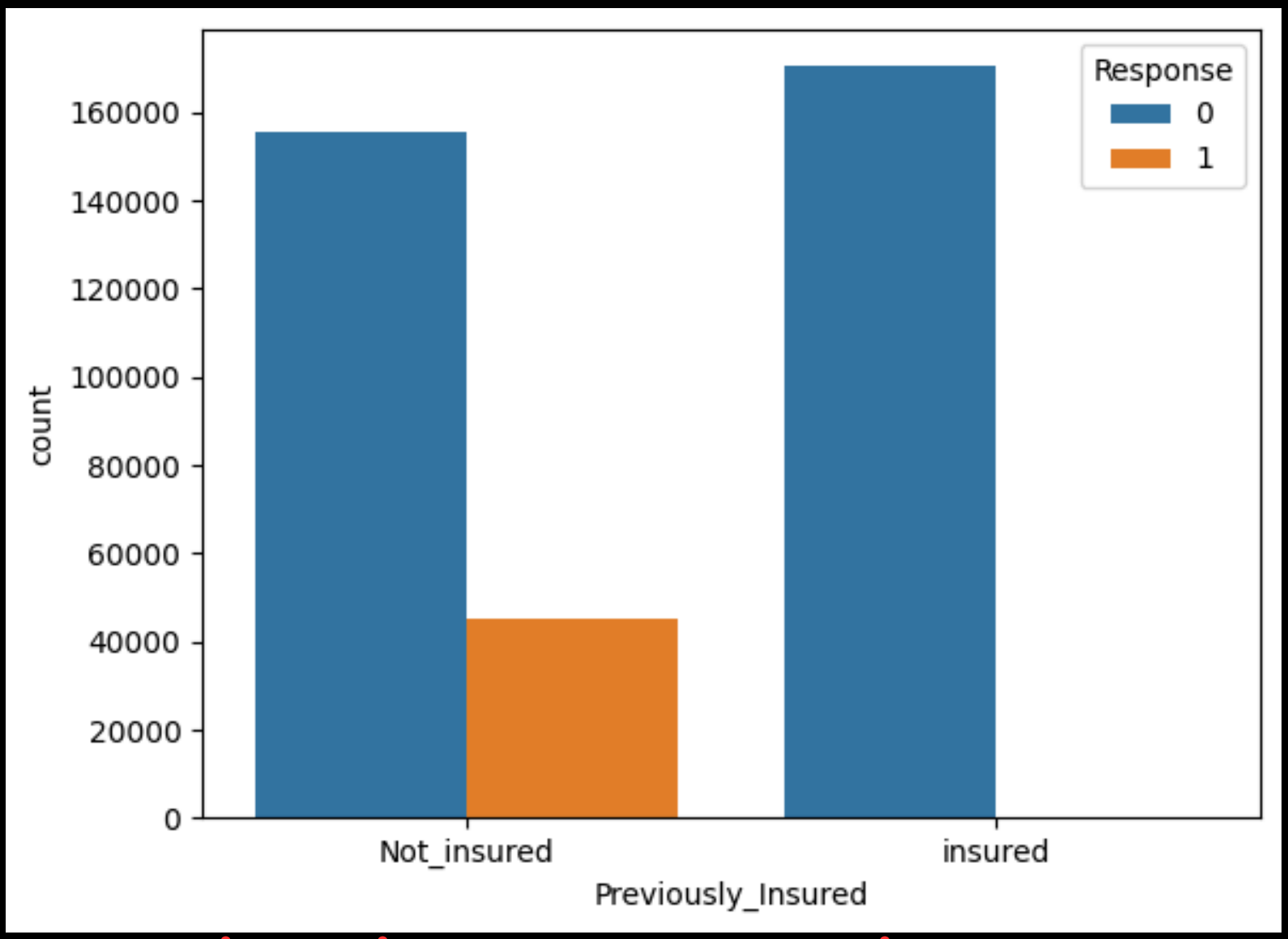
This is evident from the much higher counts in the "No" category compared to the "Yes" category.

Data visulization and insights



- **Age and Insurance Count: The graph shows the number of people ("Count") in different age groups ("Age") who have vehicle insurance.**
- **Distribution: The highest count of vehicle insurance holders is among the younger age groups, specifically around the 20s. The count gradually decreases as age increases**

Data visulization and insights



Most previously insured customers did not respond positively

- The “insured” group (already have insurance) has almost no orange bars (Response = 1).
- That means customers who already have insurance are not interested in purchasing another policy.

Not insured customers show higher interest

- In the “Not_insured” group, the orange bar (Response = 1) is significant, meaning:
- These customers are more likely to respond positively and buy insurance.
- This makes logical sense — those who don’t already have insurance are more open to it.

Response imbalance

- Overall, Response = 0 (blue bars) dominates across both groups.
- Indicates class imbalance in the dataset — far more people are not interested than those who are.

Final reports

- EDA project provided valuable insights into the dynamics of vehicle insurance, highlighting the factors that influence insurance claims, premiums, and customer behavior. Key findings include.
- Age, vehicle age, and region are significant predictors of insurance claims and premiums. Gender does not appear to be a significant factor in insurance claims.
- Comprehensive policies are the most common type of insurance coverage among insured individuals. Customer loyalty and tenure with the insurance company are associated with lower claim frequencies. Based on these insights, the following recommendations are proposed for insurance companies.
- Develop personalized insurance products and pricing strategies tailored to different age groups, vehicle types, and regional risk factors.
- Enhance customer loyalty programs and retention strategies to incentivize long-term relationships and reduce claim frequencies.

- Monitor and analyze claim data regularly to identify emerging trends, patterns, and risk factors, enabling proactive risk management and strategic decision- making.

- This dataset offers a rich and diverse set of variables that provide valuable insights into the dynamics of vehicle insurance. Through meticulous data pre processing, visualization, and statistical analysis, the project successfully identified key factors influencing insurance claims, premiums, and customer behaviour . The findings from this EDA project can inform strategic decisionmaking processes within the insurance domain, aiding in the development of personalized insurance products, pricing strategies, and customer retention initiatives tailored to different demographic groups, vehicle types, and regional risk factors.

Thanks for reading

for coding part.....

https://colab.research.google.com/drive/1IBVCa8W0i7v2746ejz2cA0yJoKrVoa2p#scrollTo=QlkTiC0G_dZv