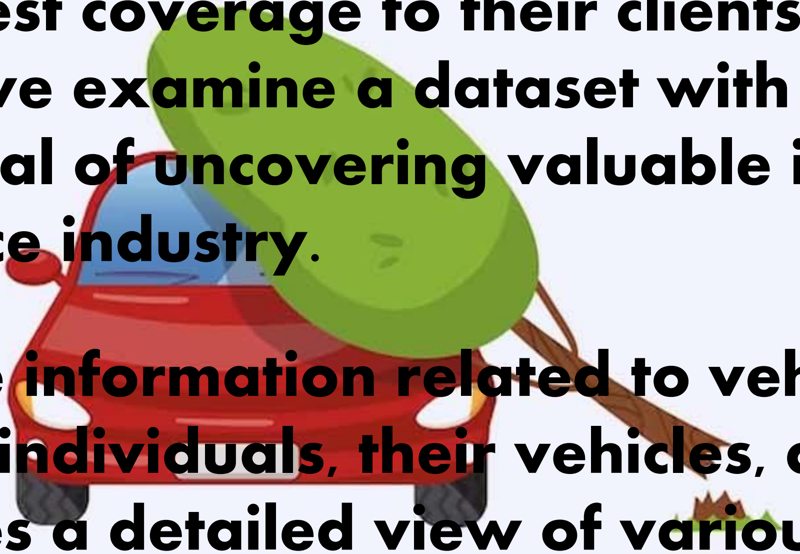
Exploratory Data Analysis

VEHICLE INSURANCE

INTRODUCTION

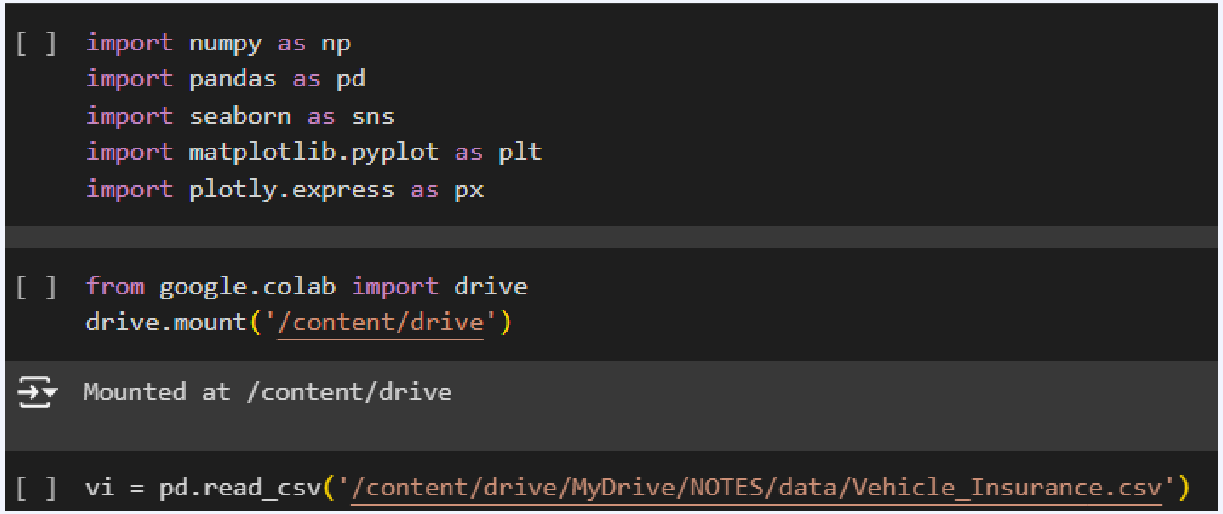
- **Vehicle insurance plays a vital role in contemporary life, offering financial security and peace of mind to both individuals and businesses. Grasping the factors that impact insurance claims is crucial for insurance companies to manage risk effectively and deliver the best coverage to their clients. In this exploratory data analysis (EDA) project, we examine a dataset with detailed vehicle insurance information, with the goal of uncovering valuable insights to guide decision-making within the insurance industry.**



- **The dataset in question contains extensive information related to vehicle insurance, including details about insured individuals, their vehicles, and the corresponding insurance claims. It provides a detailed view of various factors such as age, gender, geographic location, insurance premiums, policy types, and more. By thoroughly analysing these variables, our goal is to identify patterns, trends, and correlations that reveal insights into the factors influencing insurance claims.**

DESCRIPTION

- I have conducted my work using Google Colab Notebook.
- The dataset has been imported from Google Drive.
- As we begin our Exploratory Data Analysis (EDA), I've named the dataset 'vi' .
- The dataset comprises of 381109 rows and 12 columns.
- For data cleaning, I have utilized libraries Numpy , Pandas , Matplotlib , Plotly and Seaborn .
- Any duplicate entries that were found have also been removed.

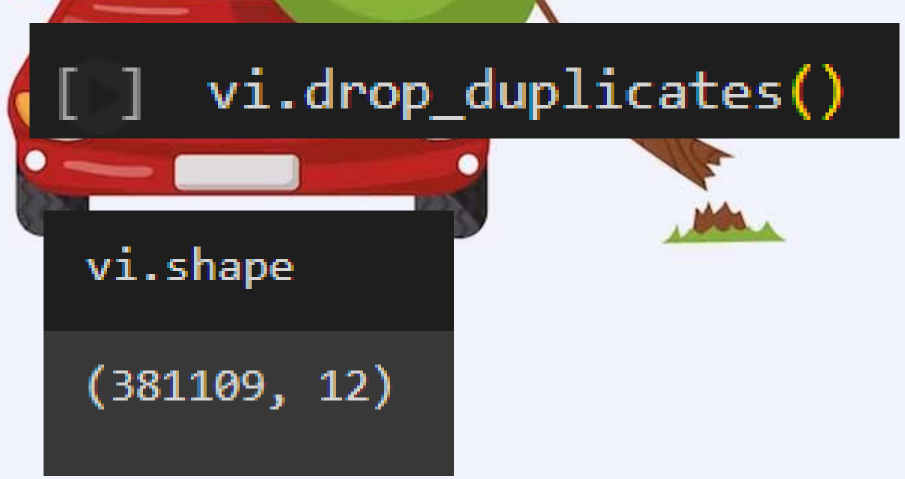


```
[ ] import numpy as np
    import pandas as pd
    import seaborn as sns
    import matplotlib.pyplot as plt
    import plotly.express as px
```

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

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

```
[ ] vi = pd.read_csv('/content/drive/MyDrive/NOTES/data/Vehicle_Insurance.csv')
```



```
[ ] vi.drop_duplicates()
```

```
vi.shape
```

```
(381109, 12)
```

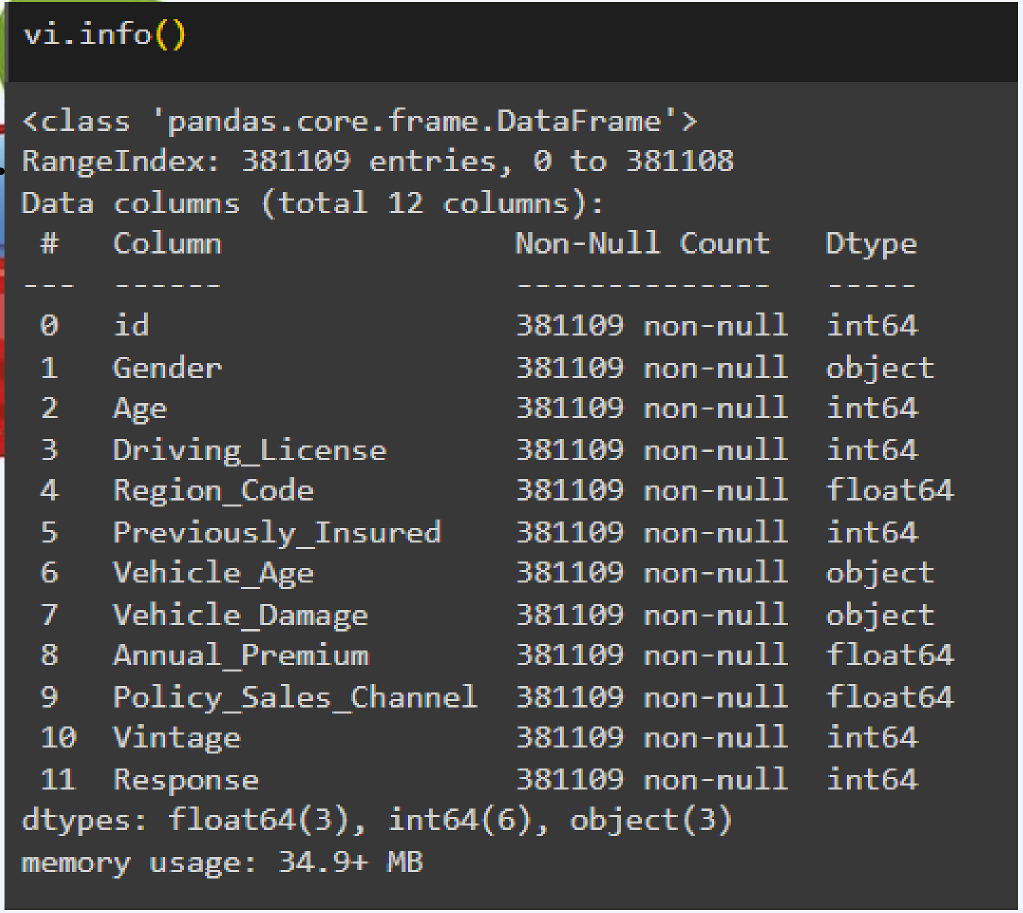
DESCRIPTION

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.**



```
vi.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

- **Region Code:** The geographic location of the insured individual, reflecting regional differences in risk factors and claim rates.

- **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.

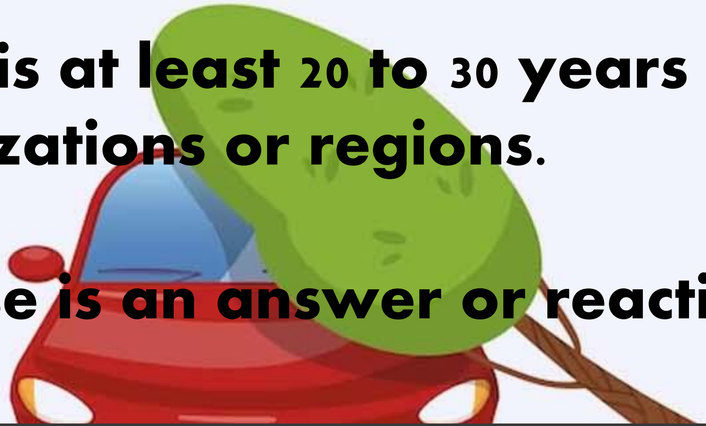
- **Annual Premium:** The total amount of money paid by the insured individual for their vehicle insurance policy over the course of a year.

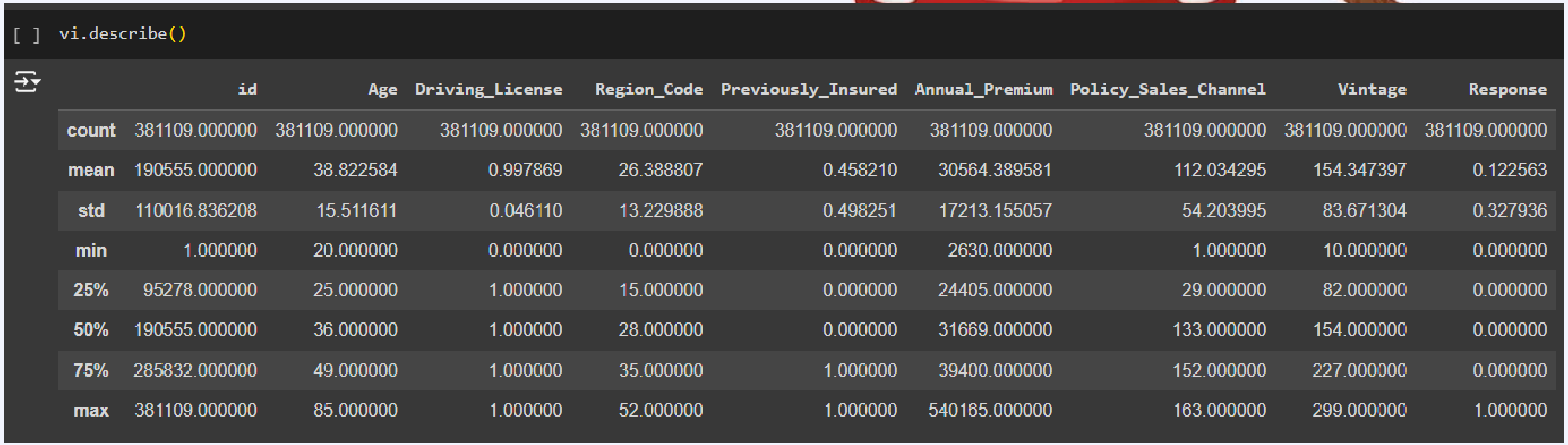
DESCRIPTION

- **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.



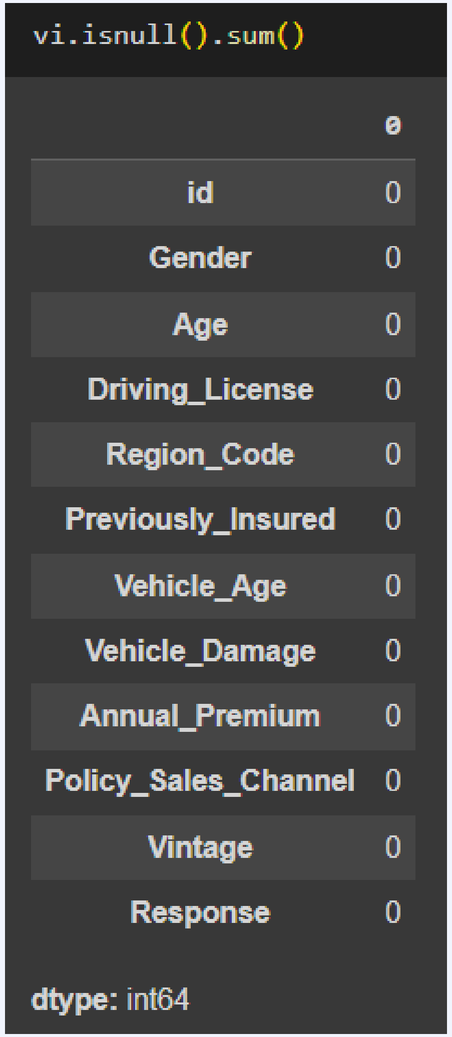


```
[ ]  vi.describe()
```

```
                  id            Age  Driving_License  Region_Code  Previously_Insured  Annual_Premium  Policy_Sales_Channel        Vintage       Response
count  381109.000000  381109.000000    381109.000000  381109.000000       381109.000000   381109.000000         381109.000000  381109.000000  381109.000000
mean   190555.000000      38.822584         0.997869      26.388807            0.458210    30564.389581            112.034295     154.347397       0.122563
std    110016.836208      15.511611         0.046110      13.229888            0.498251    17213.155057             54.203995      83.671304       0.327936
min         1.000000      20.000000         0.000000       0.000000            0.000000     2630.000000              1.000000      10.000000       0.000000
25%     95278.000000      25.000000         1.000000      15.000000            0.000000    24405.000000             29.000000      82.000000       0.000000
50%    190555.000000      36.000000         1.000000      28.000000            0.000000    31669.000000            133.000000     154.000000       0.000000
75%    285832.000000      49.000000         1.000000      35.000000            1.000000    39400.000000            152.000000     227.000000       0.000000
max    381109.000000      85.000000         1.000000      52.000000            1.000000   540165.000000            163.000000     299.000000       1.000000
```

Data Cleaning & Pre-Processing:

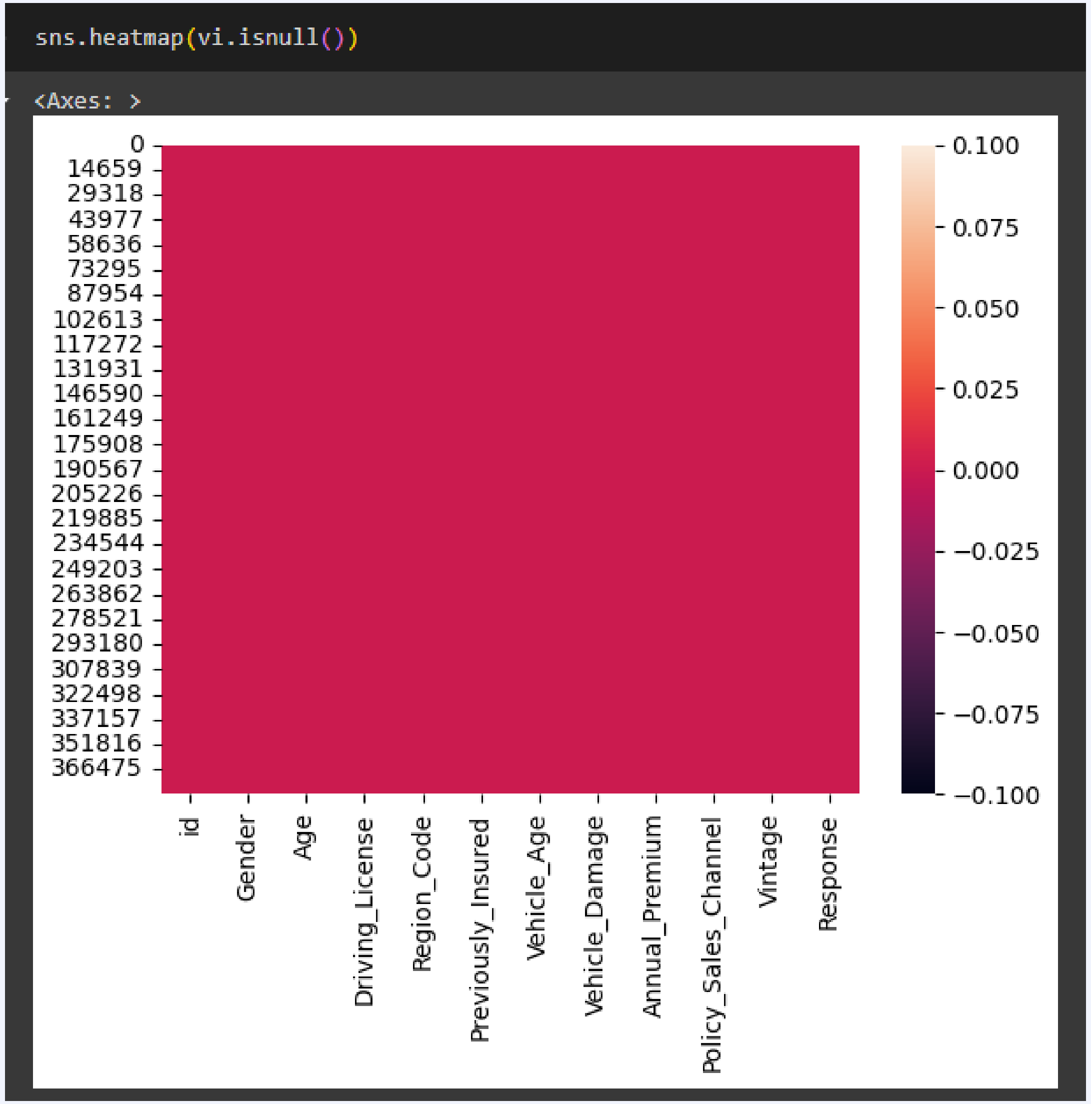
As we have seen that this data set has `0` null values. So we can move to the next step



```python
vi.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



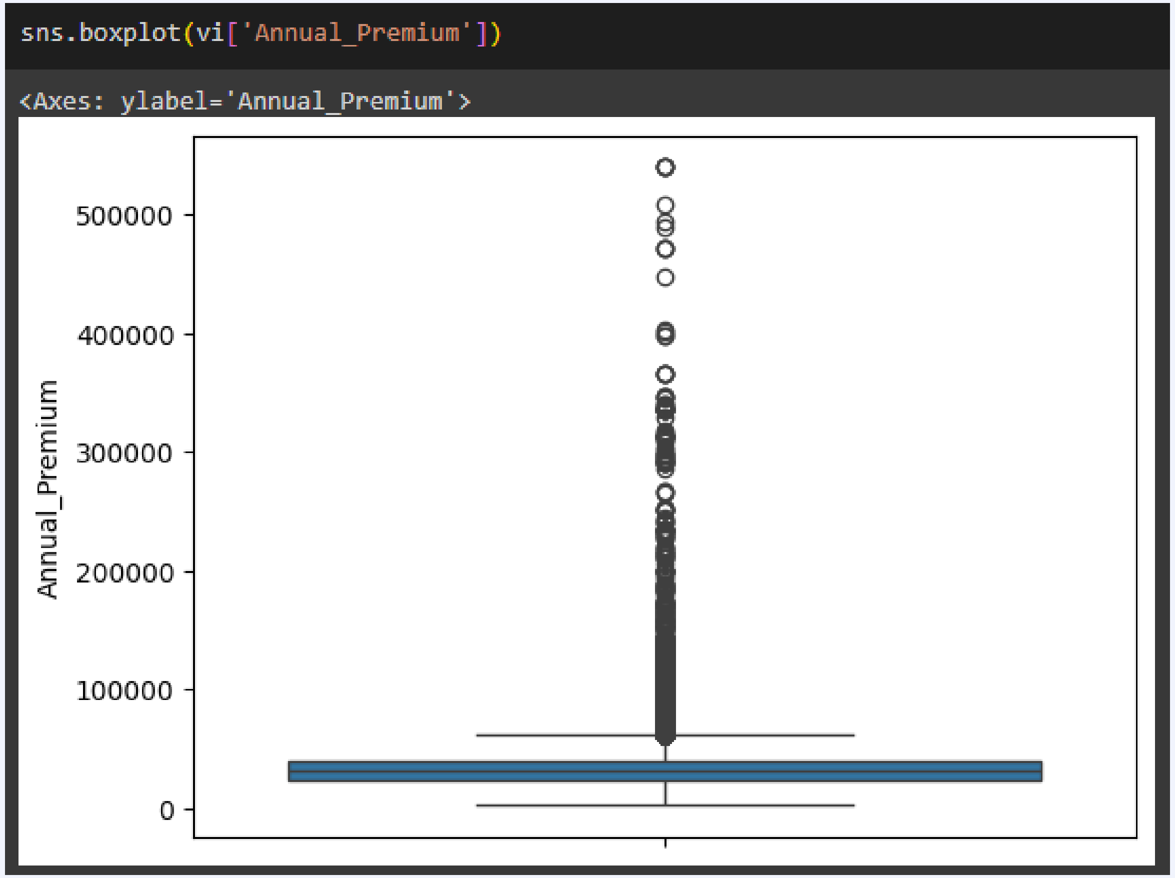
```python
sns.heatmap(vi.isnull())
```

```
<Axes: >
```

Data Cleaning & Pre-Processing:

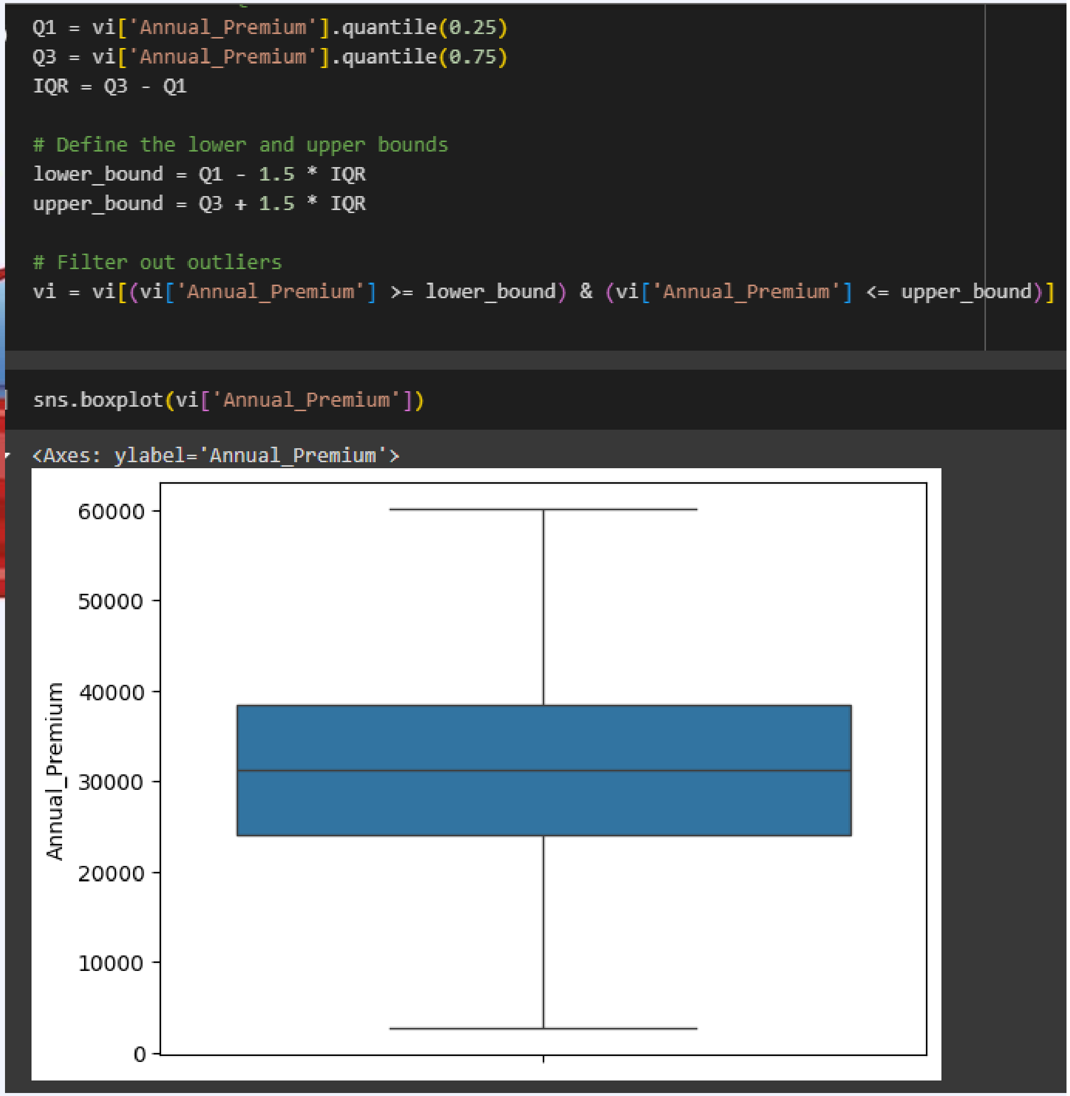
- Key consideration : Although we have 0 null values, we still need to examine numerical features- 'Annual Premium' for outliers to ensure data balance and improve the quality of insights .

- Using IQR method for removing these outliers.



```
sns.boxplot(vi['Annual_Premium'])
```

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

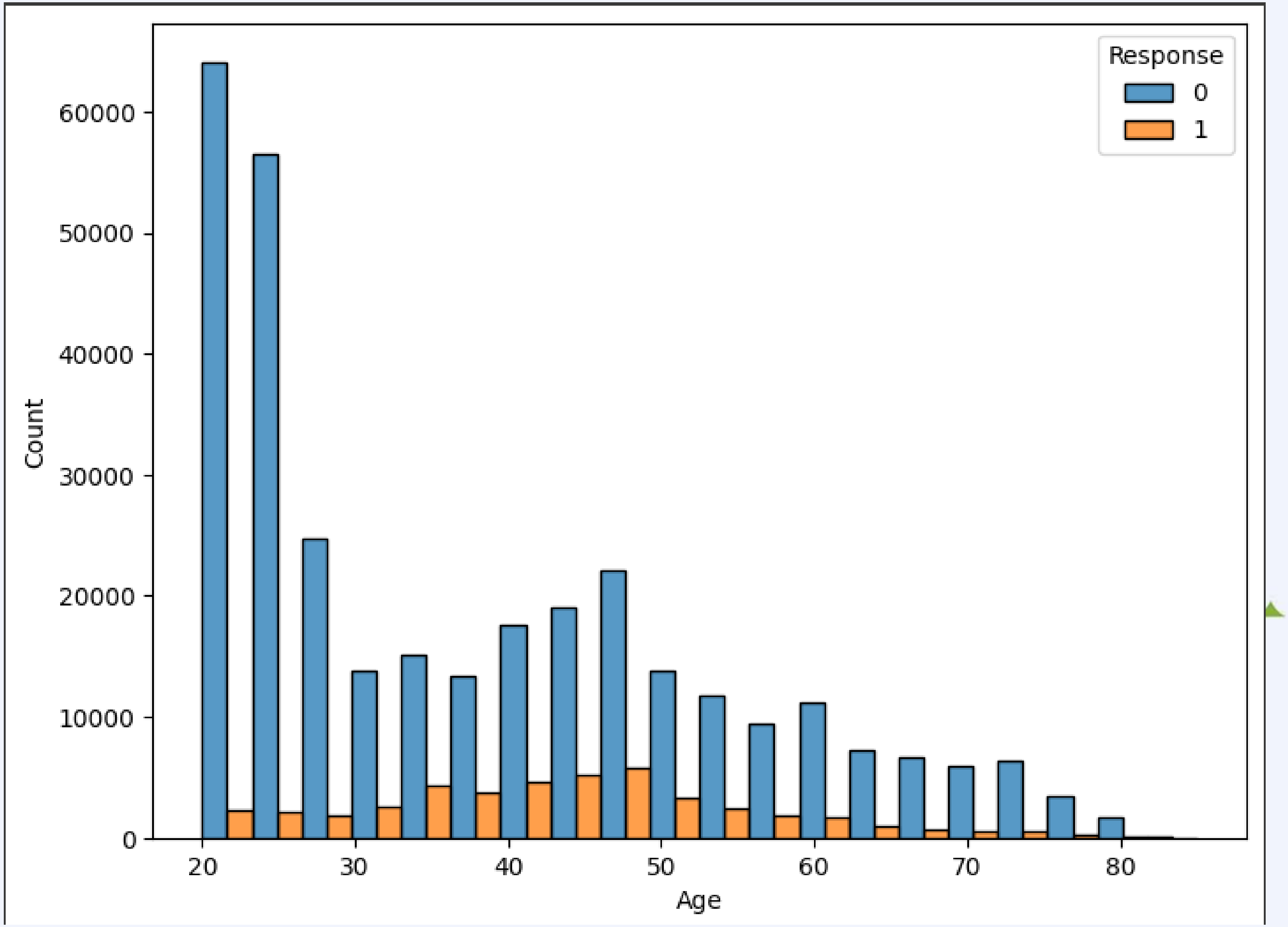


```
Q1 = vi['Annual_Premium'].quantile(0.25)
Q3 = vi['Annual_Premium'].quantile(0.75)
IQR = Q3 - Q1

# Define the lower and upper bounds
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR

# Filter out outliers
vi = vi[(vi['Annual_Premium'] >= lower_bound) & (vi['Annual_Premium'] <= upper_bound)]
```

```
sns.boxplot(vi['Annual_Premium'])
```

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

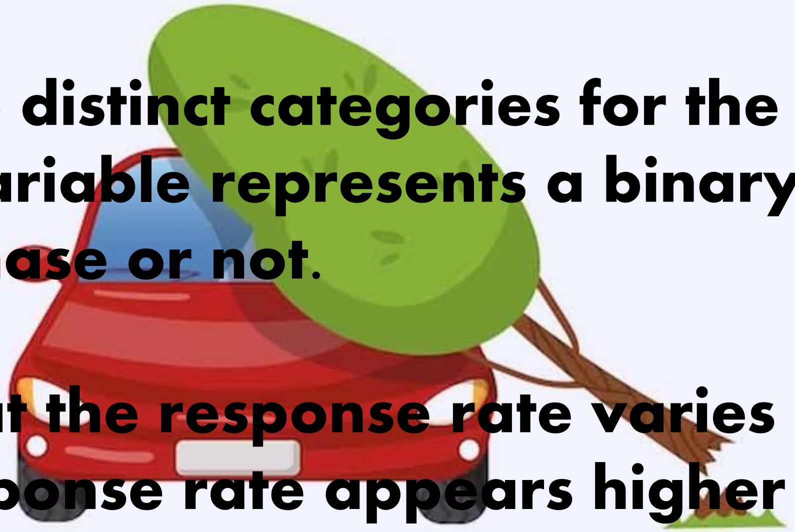
Data Visualization and Insights:



Key Insights:

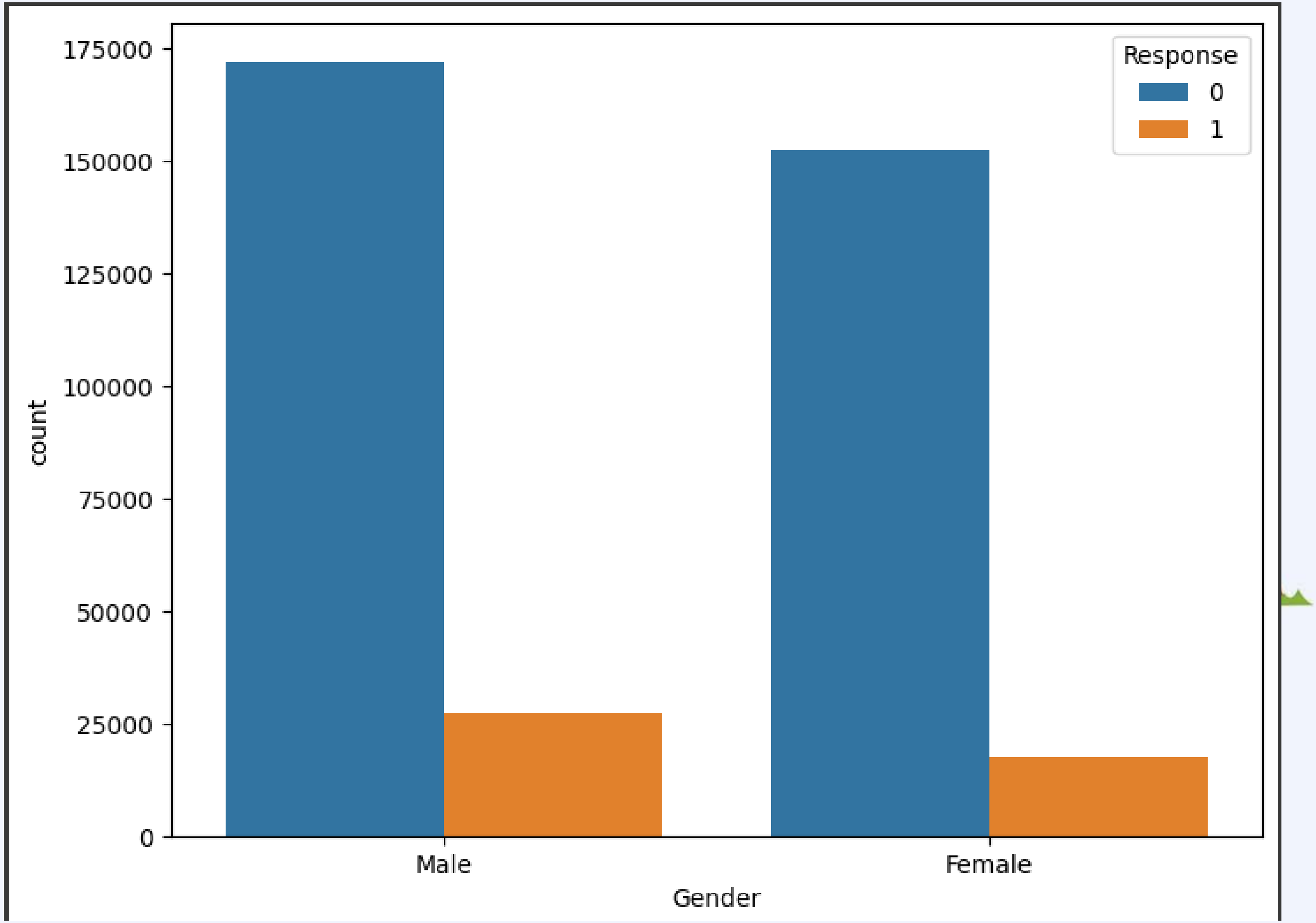
- **Age Distribution :** The dataset has a wide age range, from around 20 to 80 years old. The majority of individuals fall within the 20-50 age group, with the peak around the 30-40 age range.

- **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 Visualization and Insights:

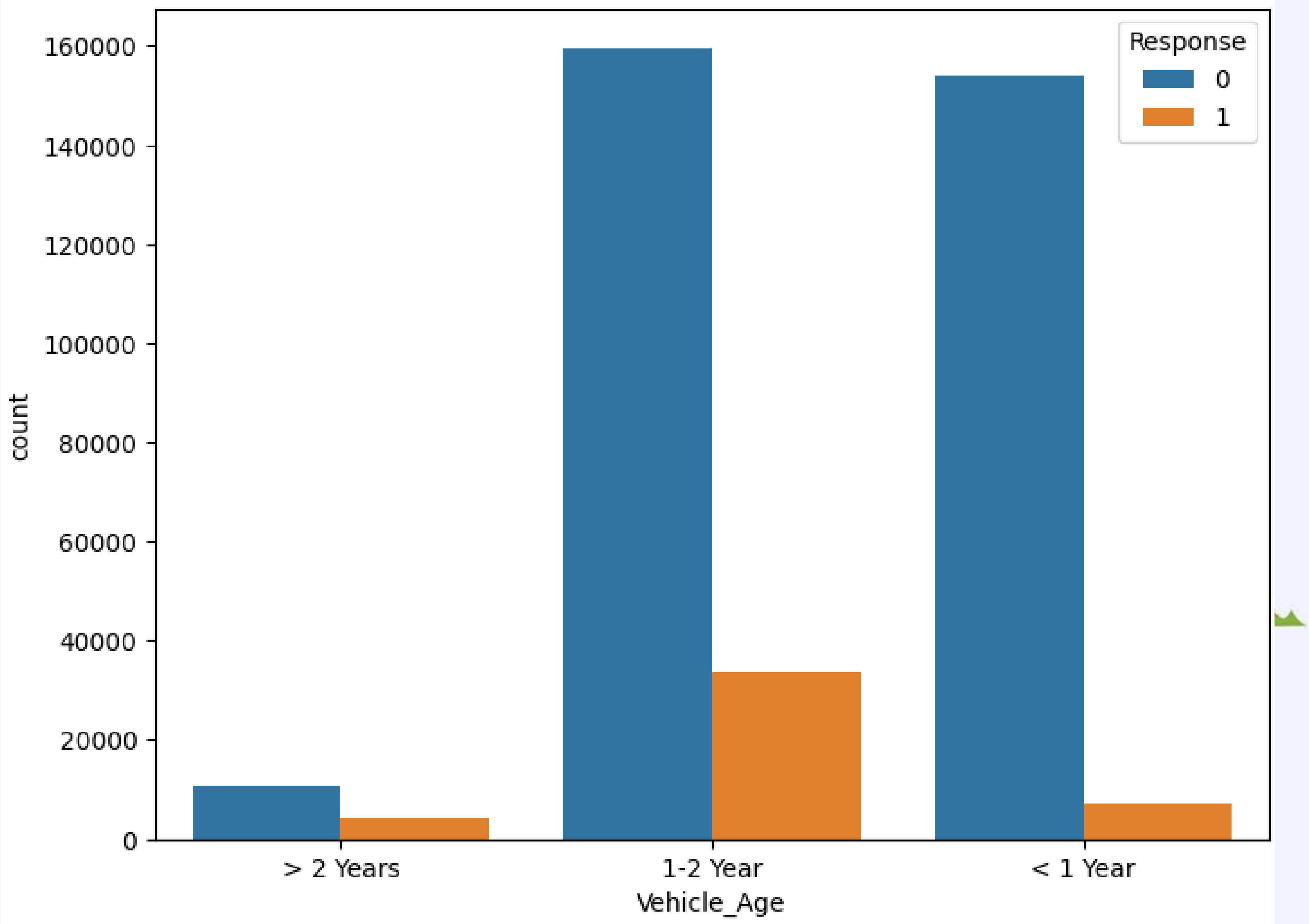


Key Insights:

- **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 Visualization and Insights:

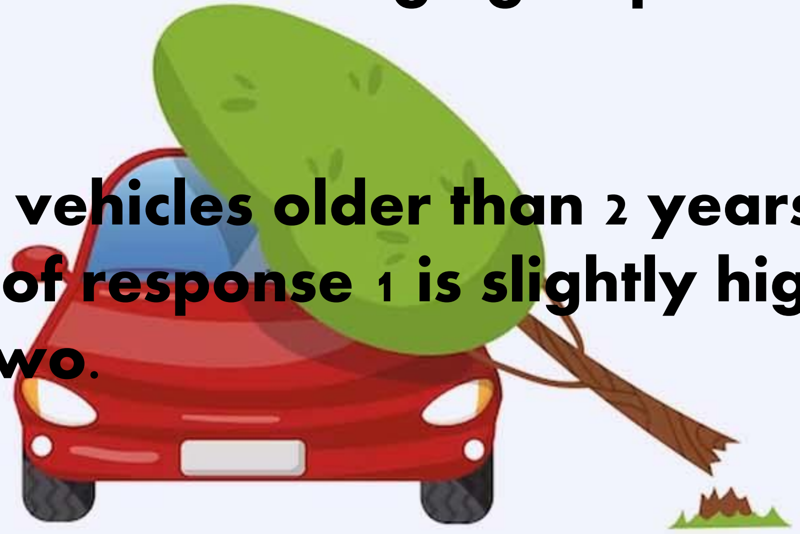


Key Insights:

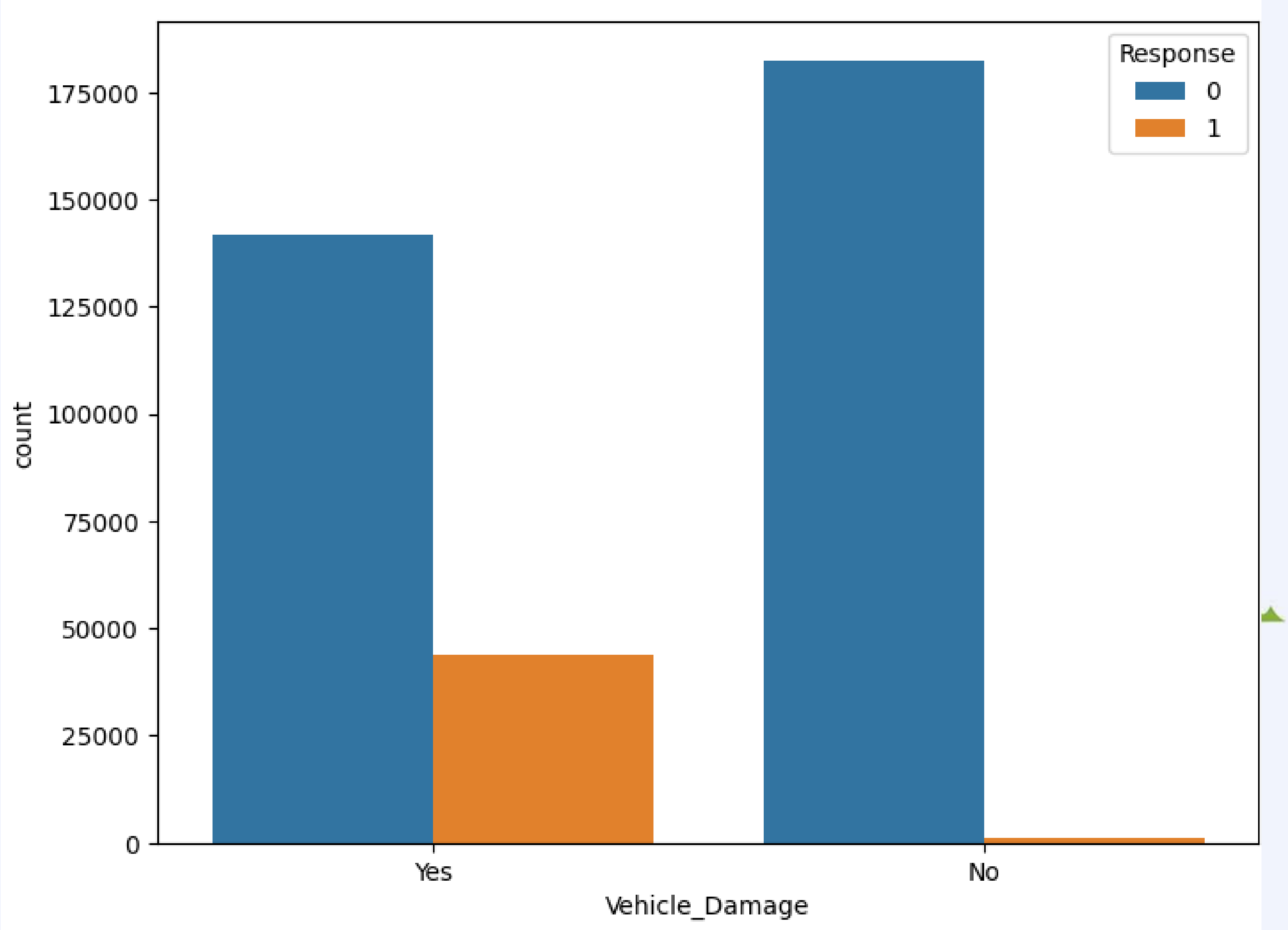
- **Most vehicles are less than 1 year old. This category has the highest count for both response values.**

- **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 Visualization and Insights:



Key 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.

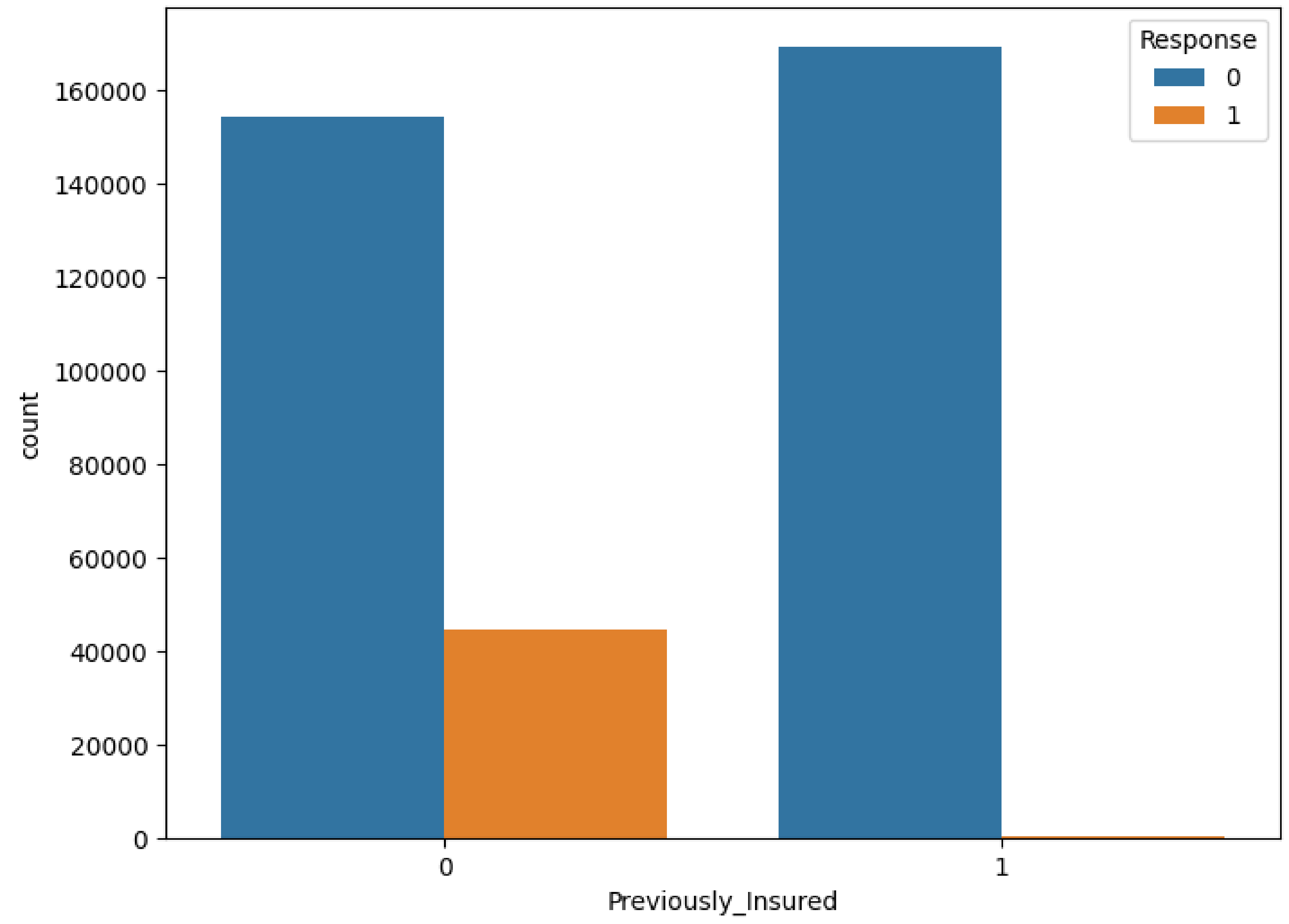
- Vehicle damage negatively impacts loan approval .

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.

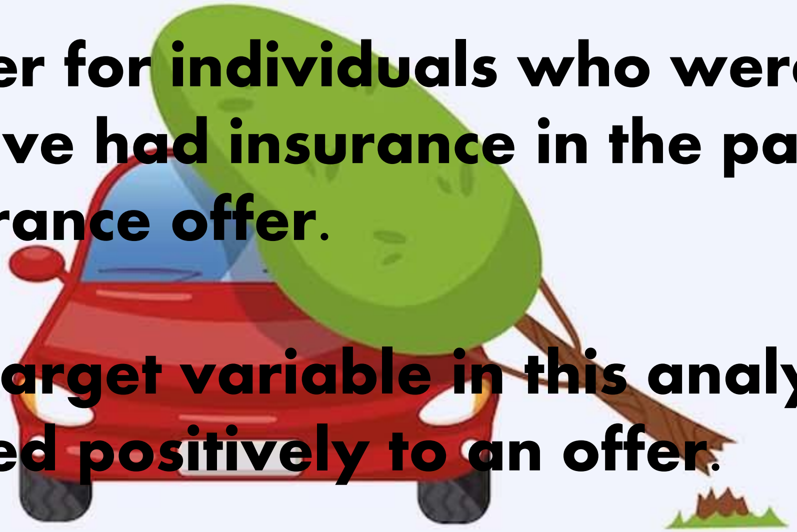
Data Visualization and Insights:



Key Insights:

- **Class Imbalance :** The dataset is imbalanced, with a significantly higher number of individuals who were previously insured (value 1) compared to those who were not (value 0).

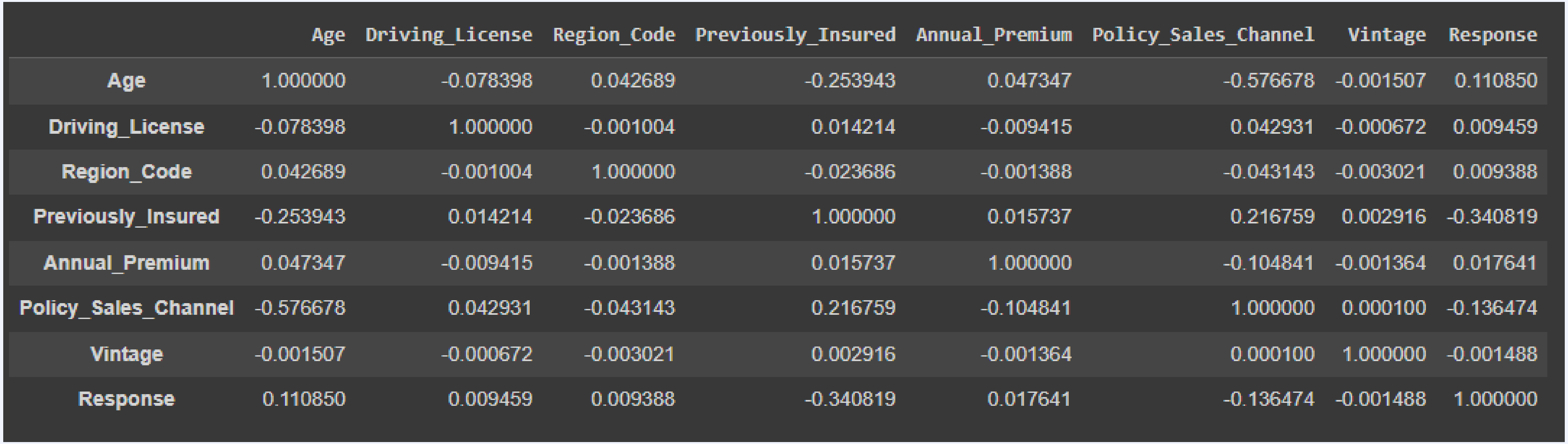
- **Response Rate :** The response rate is higher for individuals who were previously insured. This indicates that people who have had insurance in the past are more likely to respond positively to a new insurance offer.



- **Target Variable :** "Response" is likely the target variable in this analysis, indicating whether an individual responded positively to an offer.

- **Predictive Power :** "Previously Insured" could be a valuable predictor for the "Response" variable, given the difference in response rates

Data Visualization and Insights:

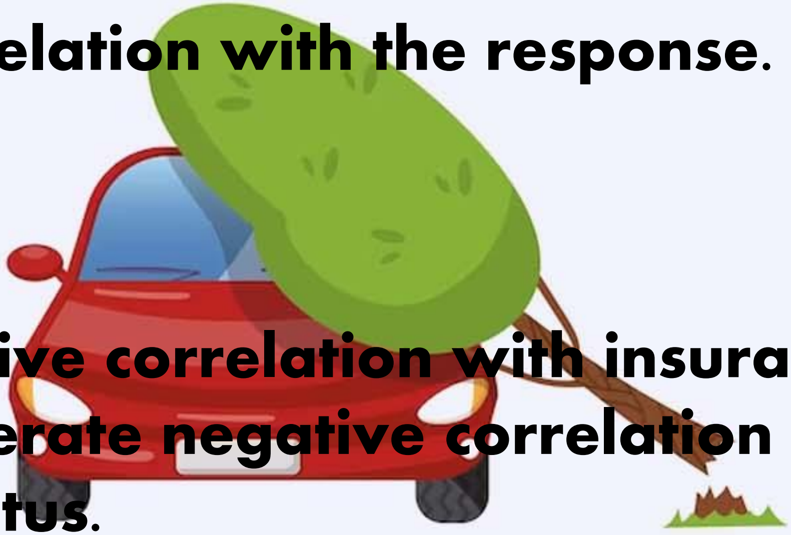


	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
Age	1.000000	-0.078398	0.042689	-0.253943	0.047347	-0.576678	-0.001507	0.110850
Driving_License	-0.078398	1.000000	-0.001004	0.014214	-0.009415	0.042931	-0.000672	0.009459
Region_Code	0.042689	-0.001004	1.000000	-0.023686	-0.001388	-0.043143	-0.003021	0.009388
Previously_Insured	-0.253943	0.014214	-0.023686	1.000000	0.015737	0.216759	0.002916	-0.340819
Annual_Premium	0.047347	-0.009415	-0.001388	0.015737	1.000000	-0.104841	-0.001364	0.017641
Policy_Sales_Channel	-0.576678	0.042931	-0.043143	0.216759	-0.104841	1.000000	0.000100	-0.136474
Vintage	-0.001507	-0.000672	-0.003021	0.002916	-0.001364	0.000100	1.000000	-0.001488
Response	0.110850	0.009459	0.009388	-0.340819	0.017641	-0.136474	-0.001488	1.000000

Key Insights:

- **Positive Correlations :**

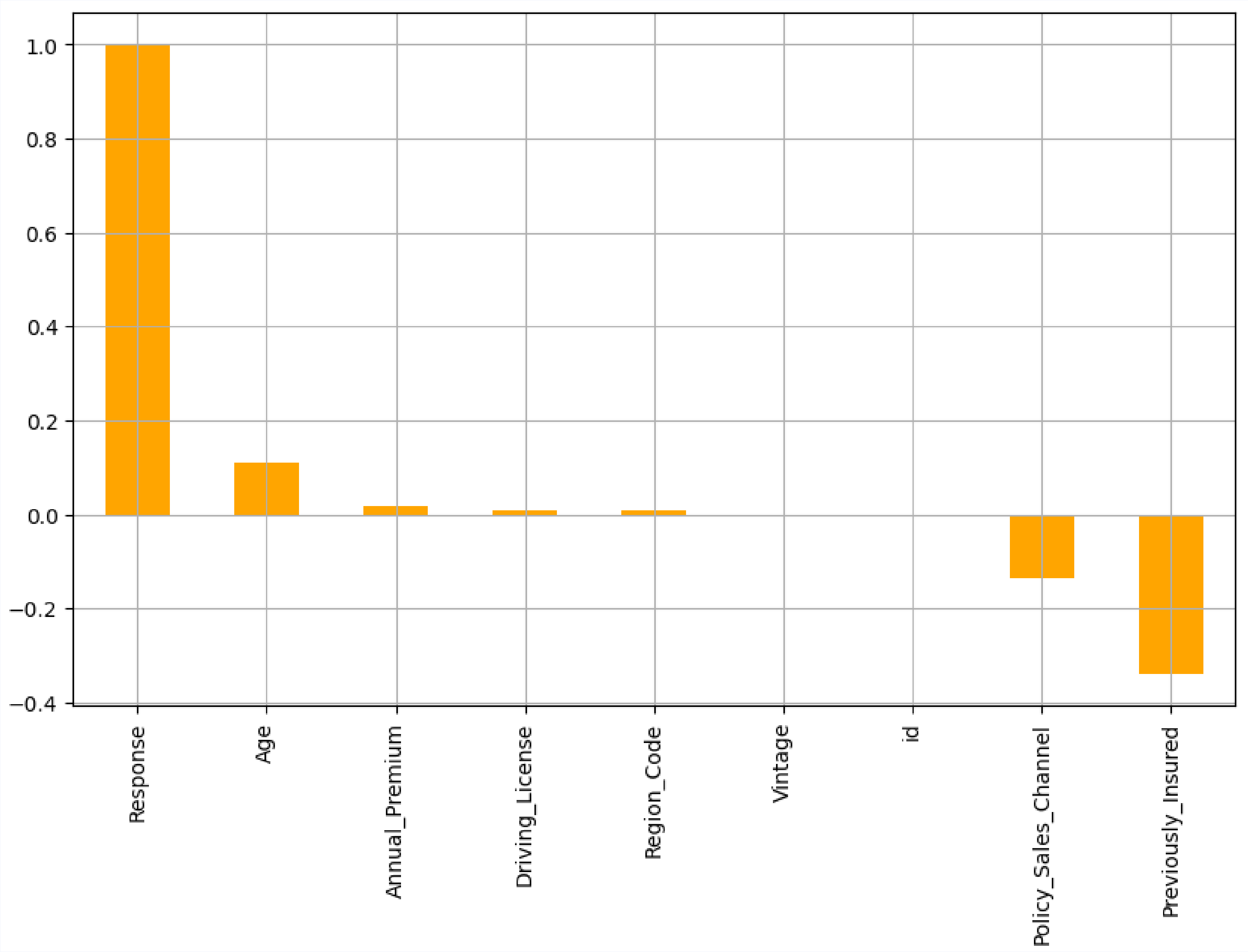
Age : There is a moderate positive correlation with insurance claims.
Response : Shows a moderate positive correlation with insurance claims.
Insurance Claim : Has a strong positive correlation with the response.



- **Negative Correlations :**

Previously Insured : Exhibits a strong negative correlation with insurance claims.
Policy Sales Channel : Demonstrates a moderate negative correlation with both insurance claims and previously insured status.
Age : Displays a slight negative correlation with the Policy Sales Channel.
Annual Premium : Shows minor negative correlations with both the Policy Sales Channel and insurance claims.

Data Visualization and Insights:



Key Insights:

- **Strong Positive Correlation:**

The ″Response″ variable has a strong positive correlation with the ″Response″ feature, which is expected as it′s likely the same variable.

- **Moderate Positive Correlation:**

The ″Age″ feature has a moderate positive correlation with the ″Response″ variable. This suggests that as age increases, the response variable tends to increase as well.

- **Weak Positive Correlation:**

The ″Annual_Premium″ feature shows a weak positive correlation.

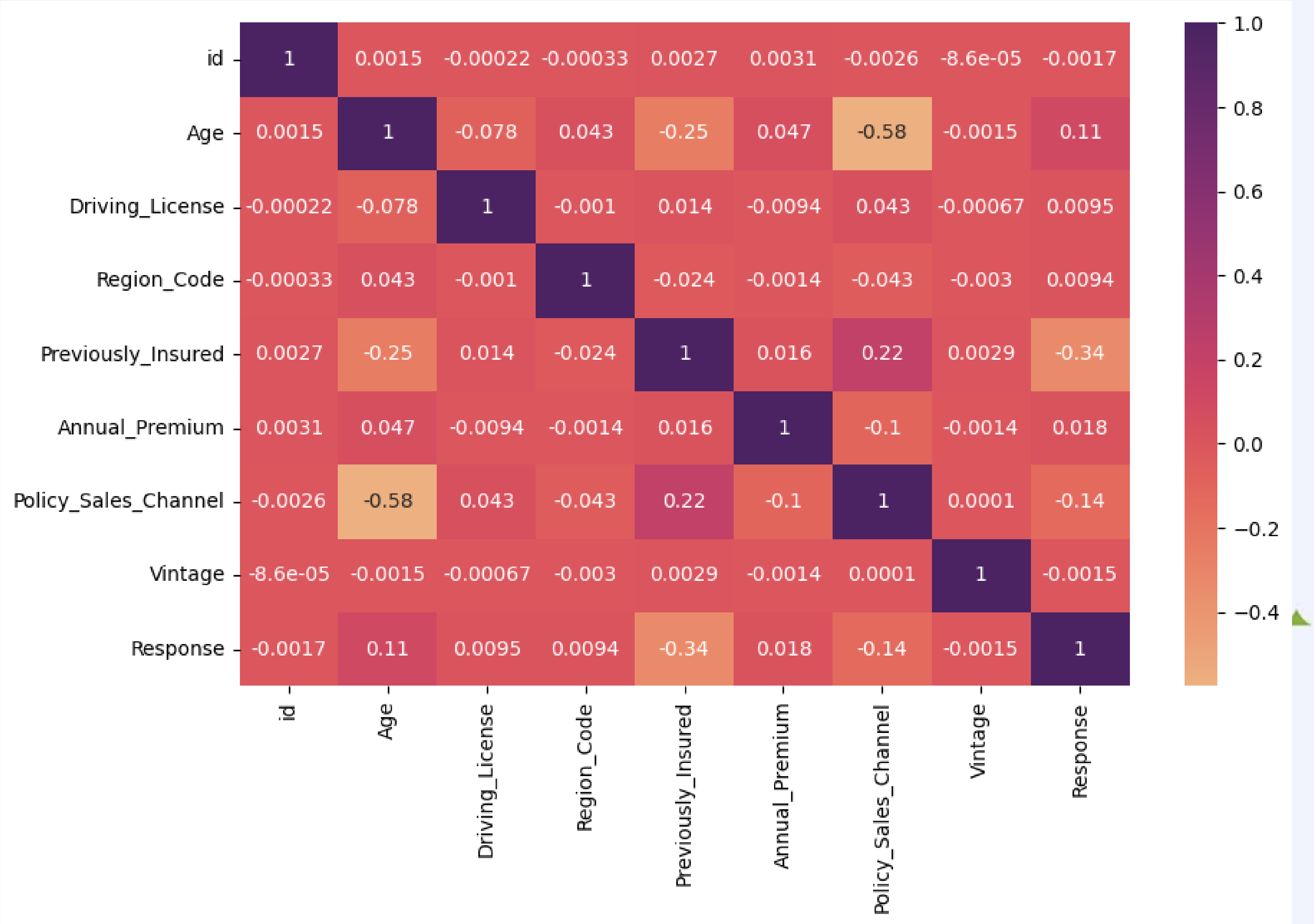
- **Very Weak or No Correlation:**

The features ″Driving_License,″ ″Region_Code,″ ″Vintage,″ and ″id″ show very weak or almost no correlation with the ″Response″ variable. This suggests these features do not significantly influence the target variable.

- **Moderate Negative Correlation:**

The ″Policy_Sales_Channel″ and ″Previously_Insured″ features have a moderate negative correlation with the ″Response″ variable. This means as these features increase, the response variable tends to decrease

Data Visualization and Insights:



	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sales_Channel	Vintage	Response
id	1	0.0015	-0.00022	-0.00033	0.0027	0.0031	-0.0026	-8.6e-05	-0.0017
Age	0.0015	1	-0.078	0.043	-0.25	0.047	-0.58	-0.0015	0.11
Driving_License	-0.00022	-0.078	1	-0.001	0.014	-0.0094	0.043	-0.00067	0.0095
Region_Code	-0.00033	0.043	-0.001	1	-0.024	-0.0014	-0.043	-0.003	0.0094
Previously_Insured	0.0027	-0.25	0.014	-0.024	1	0.016	0.22	0.0029	-0.34
Annual_Premium	0.0031	0.047	-0.0094	-0.0014	0.016	1	-0.1	-0.0014	0.018
Policy_Sales_Channel	-0.0026	-0.58	0.043	-0.043	0.22	-0.1	1	0.0001	-0.14
Vintage	-8.6e-05	-0.0015	-0.00067	-0.003	0.0029	-0.0014	0.0001	1	-0.0015
Response	-0.0017	0.11	0.0095	0.0094	-0.34	0.018	-0.14	-0.0015	1

Key Insights:

- This correlation matrix provides a quick visual representation of the relationships between variables.

- The strongest positive correlation observed is between Previously_Insured and Response, which could be a valuable insight for marketing strategies.

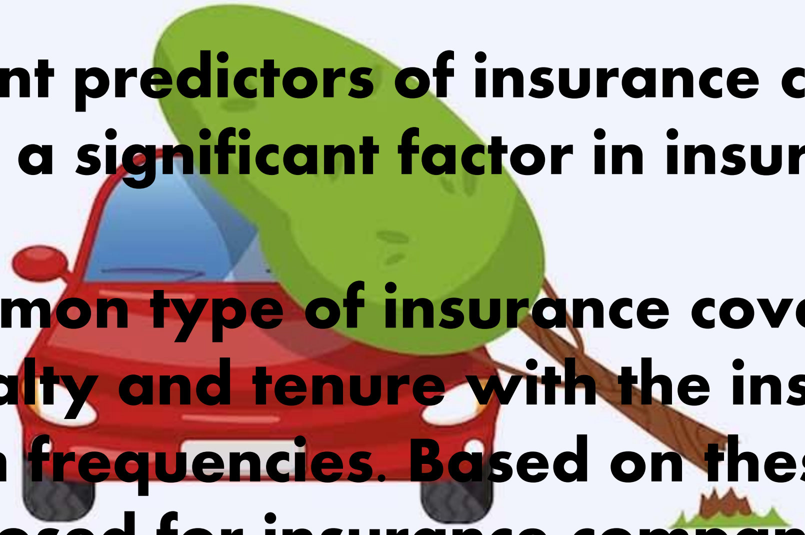
- The lack of strong correlations suggests that the variables are relatively independent of each other.

- For further analysis, it is recommended to delve into individual variable relationships using scatter plots or other visualization techniques.

Final Report

- 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.

Final Report

- 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.

Final Report

- **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 decision-making 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://github.com/sahilyadav7i/Vehicle-Insurence