Vehicle Insurance Data Analysis Report

2025

Prepared by

APOORVA SHARMA

Data Analysis File Link:

Vehicle Insurance Mini Project

Table of Content

OVERVIEW OF DATASET

AIM AND OBJECTIVES

TOOLS AND FORMULAS

STEPS TO FOLLOW

LOADING AND CLEANING

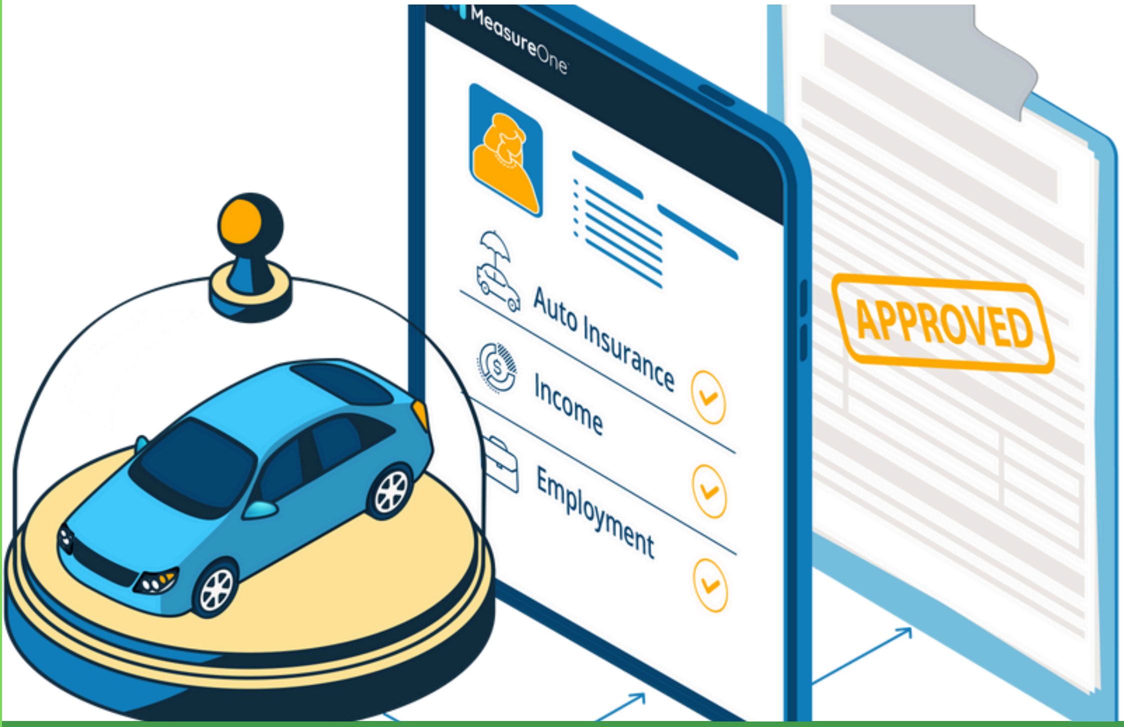
GRAPHS AND INSIGHTS

CONCLUSION

ACKNOWLEDGEMENT

OVERVIEW OF DATASET

The dataset contains information related to vehicle insurance, including details about insured individuals, their vehicles, and insurance claims. Students will explore columns such as age, gender, region, insurance premiums, policy types, and more. The ultimate goal is to derive meaningful insights that can inform decision-making processes within the insurance domain



AIM AND OBJECTIVE

AIM

To analyze and interpret meaningful patterns from the given dataset to draw insights, support decision-making, and enhance understanding of the underlying trends

OBJECTIVE

The primary objective of this project is to conduct an in-depth Exploratory Data Analysis (EDA) on a dataset related to vehicle insurance. Through this analysis, students will gain valuable insights into the patterns, trends, and factors influencing insurance claims. The project encompasses various aspects of data preprocessing, visualization, and statistical analysis.

TOOLS AND GRAPHS

TOOLS USED

NumPy

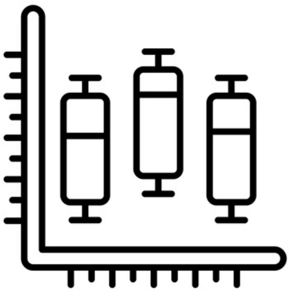
pandas

Matplotlib

seaborn

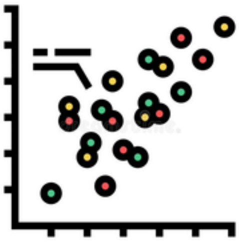
plotly

GRAPHS USED

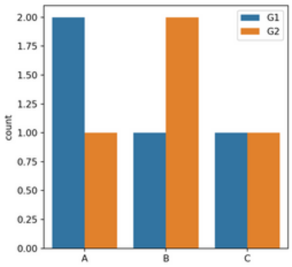


BOXPLOT

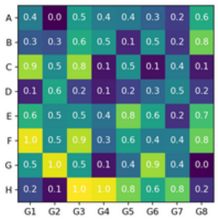
BAR GRAPH



SCATTERPLOT



COUNTPLOT

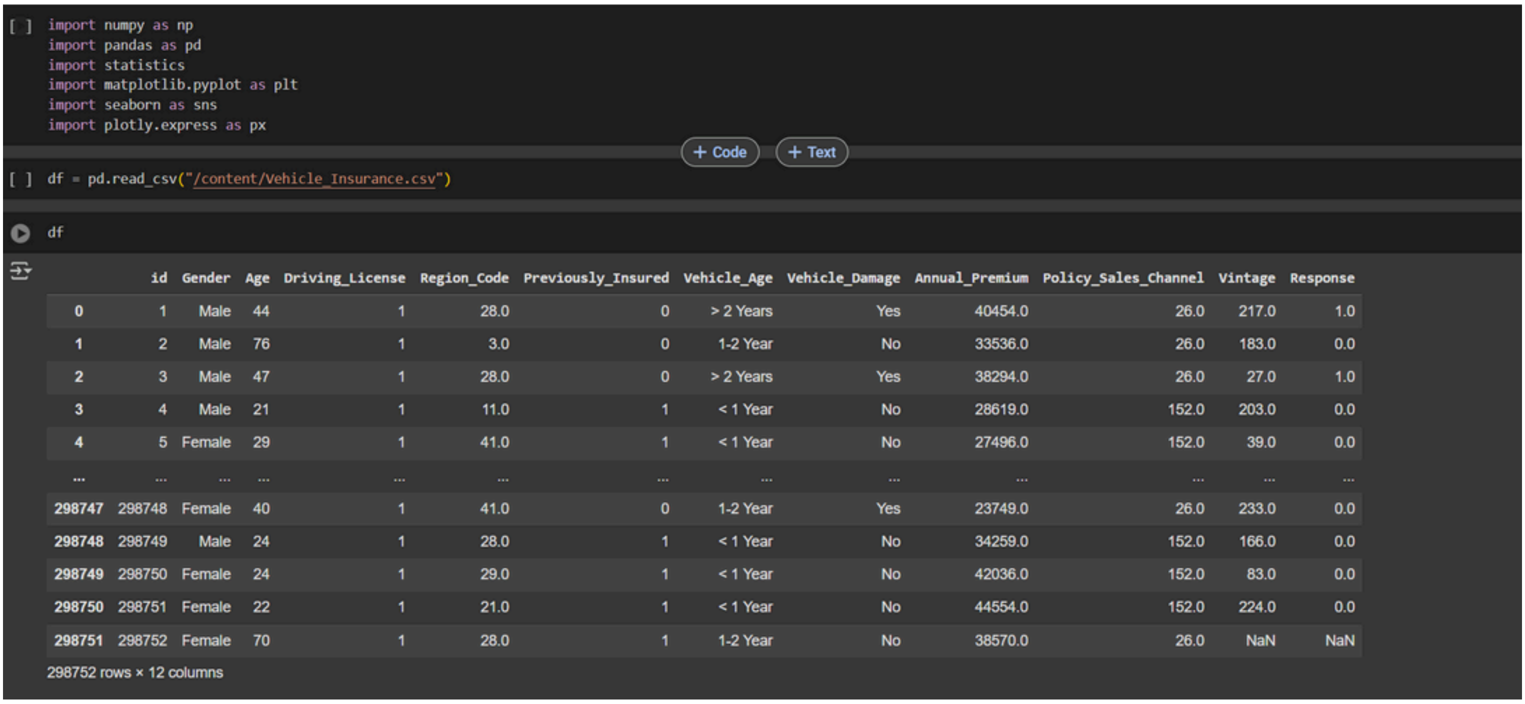


HEATMAP

STEPS TO FOLLOW

- Data Loading and Inspection: Understand the structure of the dataset. Identify the types of information available.
- Data Cleaning: Handle missing values and outliers appropriately.
- Data Visualization: Utilize various visualization techniques to explore the distribution of key variables.
- Feature Analysis: Examine the relationship between features and the target variable (insurance claims).
- Age Distribution: Analyze the age distribution within the dataset and its impact on insurance claims.
- Premium Analysis: Investigate the distribution of insurance premiums and their correlation with claim frequencies.
- Claim Frequencies: Explore factors contributing to higher claim frequencies.
- Gender Analysis: Investigate the role of gender in insurance claims.
- Vehicle Age and Claims: Examine the impact of vehicle age on the likelihood of a claim.
- Region-wise Analysis: Analyze regional patterns in insurance claims.
- Policy Analysis: Explore the distribution and impact of different insurance policy types.
- Claim Frequency by Vehicle Damage: Investigate the relationship between vehicle damage and claim frequencies.
- Customer Loyalty: Analyze if the number of policies held by a customer influences claim likelihood.
- Time Analysis: If applicable, explore temporal patterns in insurance claims

Loading Dataset



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

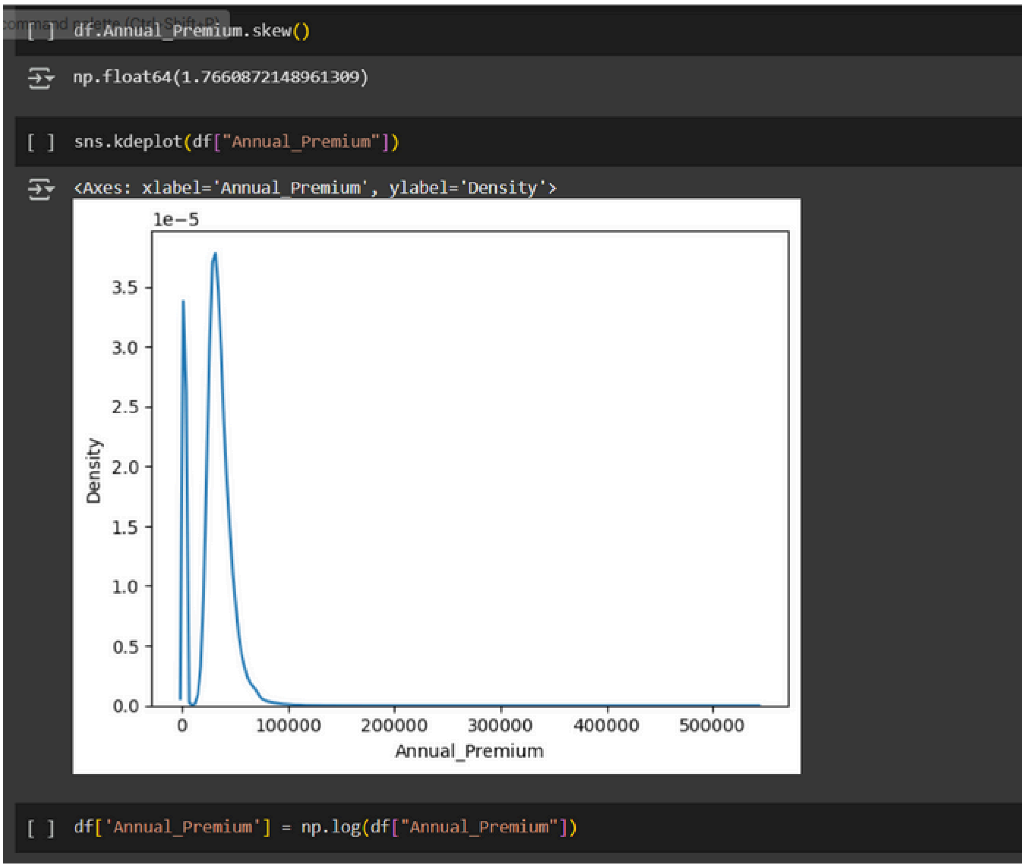
```
[ ] df = pd.read_csv("/content/Vehicle_Insurance.csv")
```

```
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	217.0	1.0
1	2	Male	76	1	3.0	0	1-2 Year	No	33536.0	26.0	183.0	0.0
2	3	Male	47	1	28.0	0	> 2 Years	Yes	38294.0	26.0	27.0	1.0
3	4	Male	21	1	11.0	1	< 1 Year	No	28619.0	152.0	203.0	0.0
4	5	Female	29	1	41.0	1	< 1 Year	No	27496.0	152.0	39.0	0.0
...	...	...	...	...	...	...	...	...	...	...	...	...
298747	298748	Female	40	1	41.0	0	1-2 Year	Yes	23749.0	26.0	233.0	0.0
298748	298749	Male	24	1	28.0	1	< 1 Year	No	34259.0	152.0	166.0	0.0
298749	298750	Female	24	1	29.0	1	< 1 Year	No	42036.0	152.0	83.0	0.0
298750	298751	Female	22	1	21.0	1	< 1 Year	No	44554.0	152.0	224.0	0.0
298751	298752	Female	70	1	28.0	1	1-2 Year	No	38570.0	26.0	NaN	NaN

298752 rows × 12 columns

Data Cleaning



command palette (Ctrl+Shift+P)

```
[ ] df.Annual_Premium.skew()
```

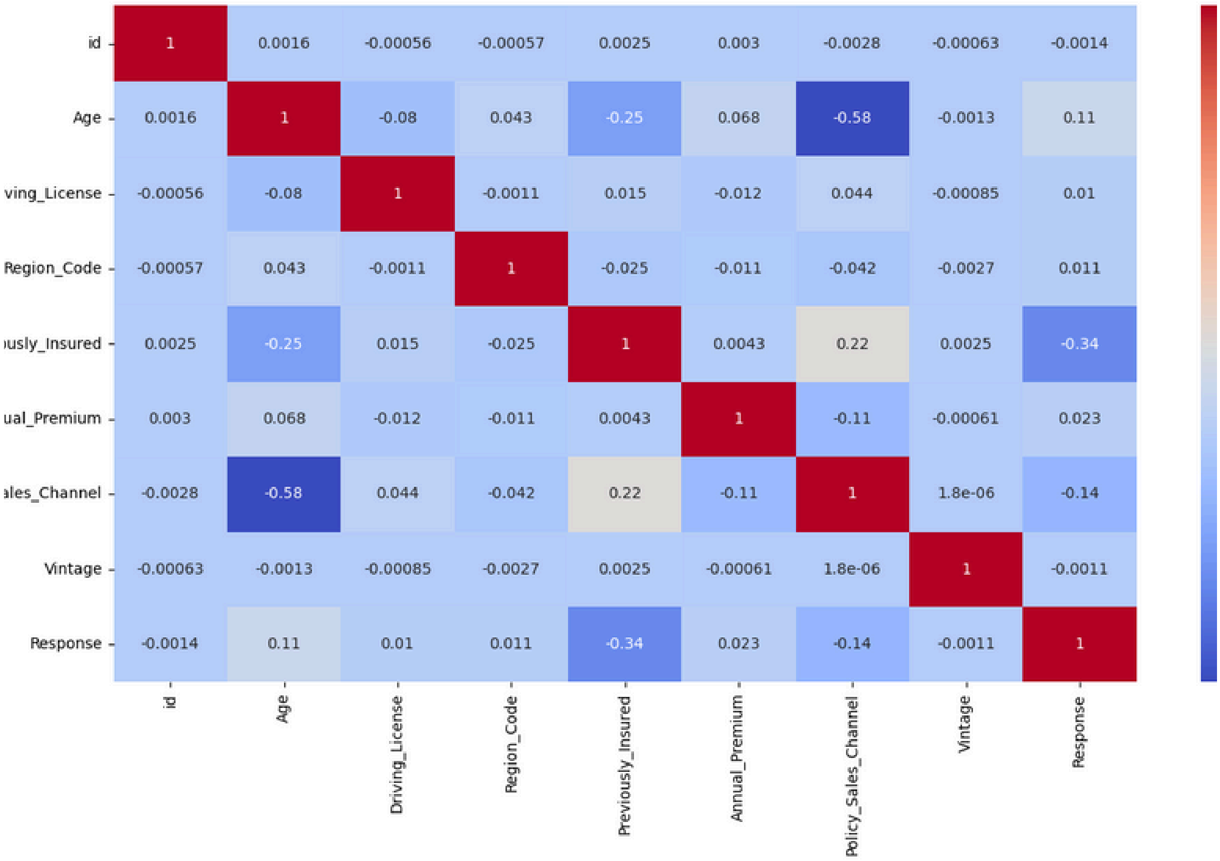
```
np.float64(1.7660872148961309)
```

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

```
<Axes: xlabel='Annual_Premium', ylabel='Density'>
```

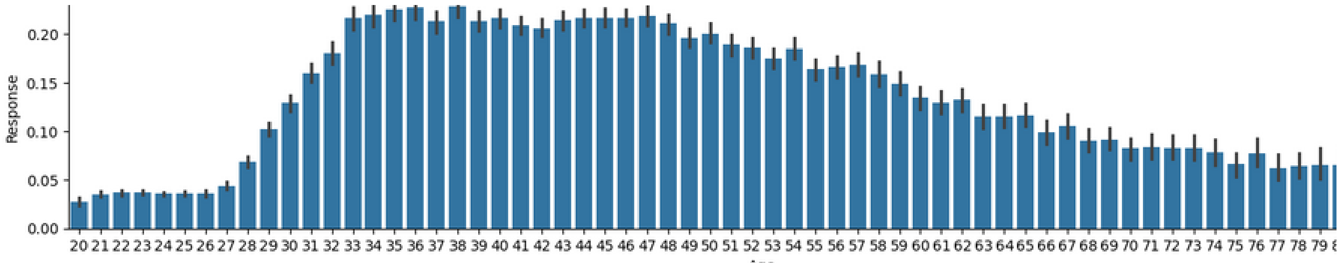
```
[ ] df['Annual_Premium'] = np.log(df["Annual_Premium"])
```

Data Visualization



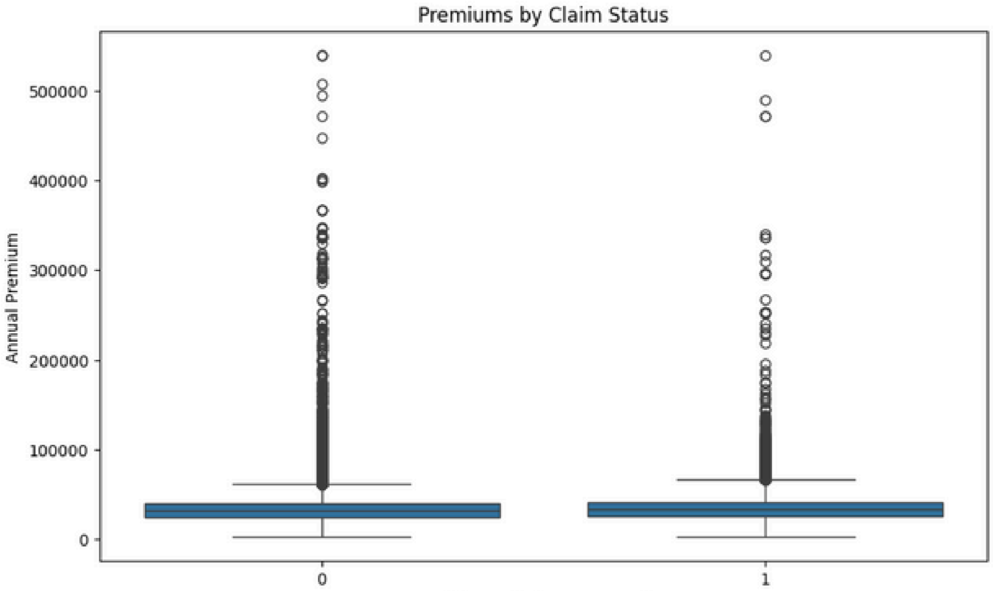
FEATURE ANALYSIS

The correlation matrix heatmap shows how different variables in the dataset are connected. Most variables don't affect each other much—there's very little relationship between them. However, two variables, "Previously_Insured" and "Policy_Sales_Channel," have some connection to "Response. This means these two might help predict the outcome more than others. Overall, the data features are mostly independent, which is good for building machine learning models and means it will have less problems with variables being too similar or "multicollinear."



AGE DISTRIBUTION

This bar graph shows how “Response” varies with “Age.” The response value rises sharply in the early 30s and stays high through the late 30s and early 40s, then slowly declines as age increases. There are more fluctuations in older ages, probably due to fewer people in those age groups. In short: Middle-aged people (roughly 30–45) have the highest response, and response drops for both younger and older people



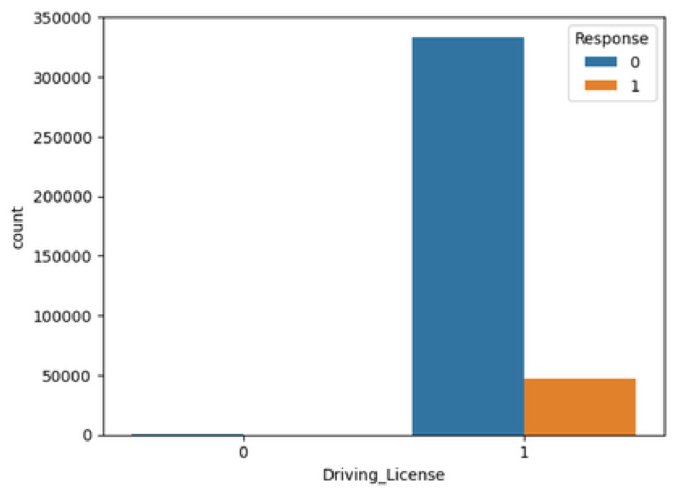
This chart compares car insurance premiums for people who made a claim (1 = Yes) versus those who did not (0 = No).

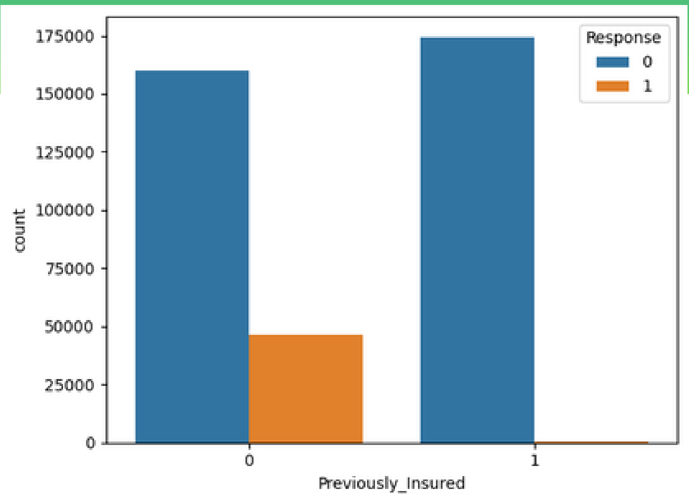
The distribution of annual premiums looks very similar for both groups: most people pay around the same amount, regardless of whether they made a claim or not.

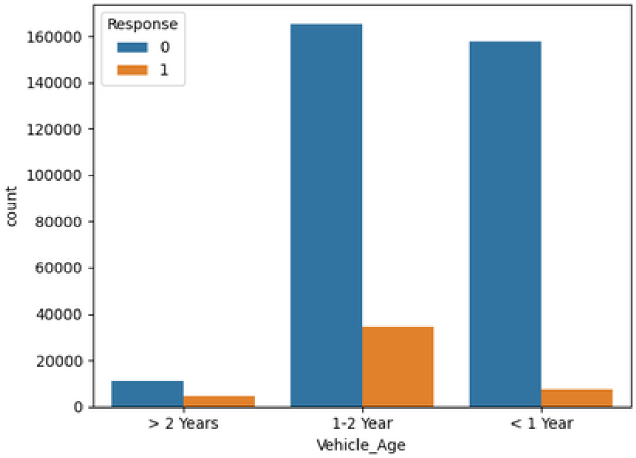
There are a few very high premiums (outliers) in both groups, shown as circles above the main “box.” These are unusual cases, but the bulk of the data is concentrated in the lower ranges for both groups.

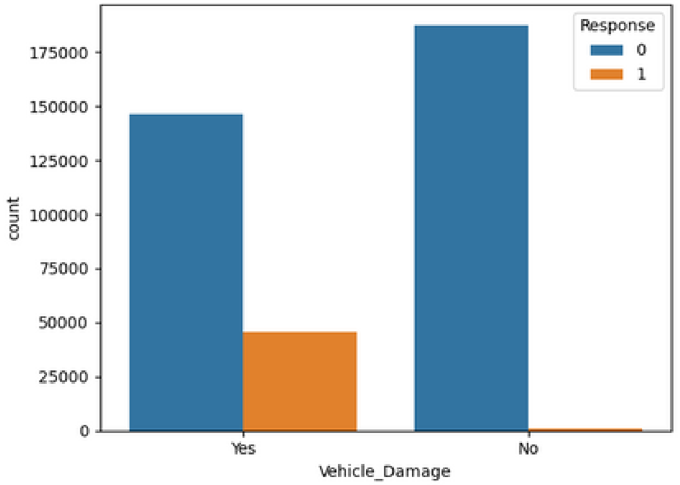
The “boxes” (showing where most data falls) and the median lines inside each box are almost at the same level, suggesting that the typical (median) premium does not change much based on claim status.

In short: Whether or not someone made a claim, their car insurance premium is usually similar. Only a small number of individuals pay unusually high premiums, but these outliers happen in both groups.





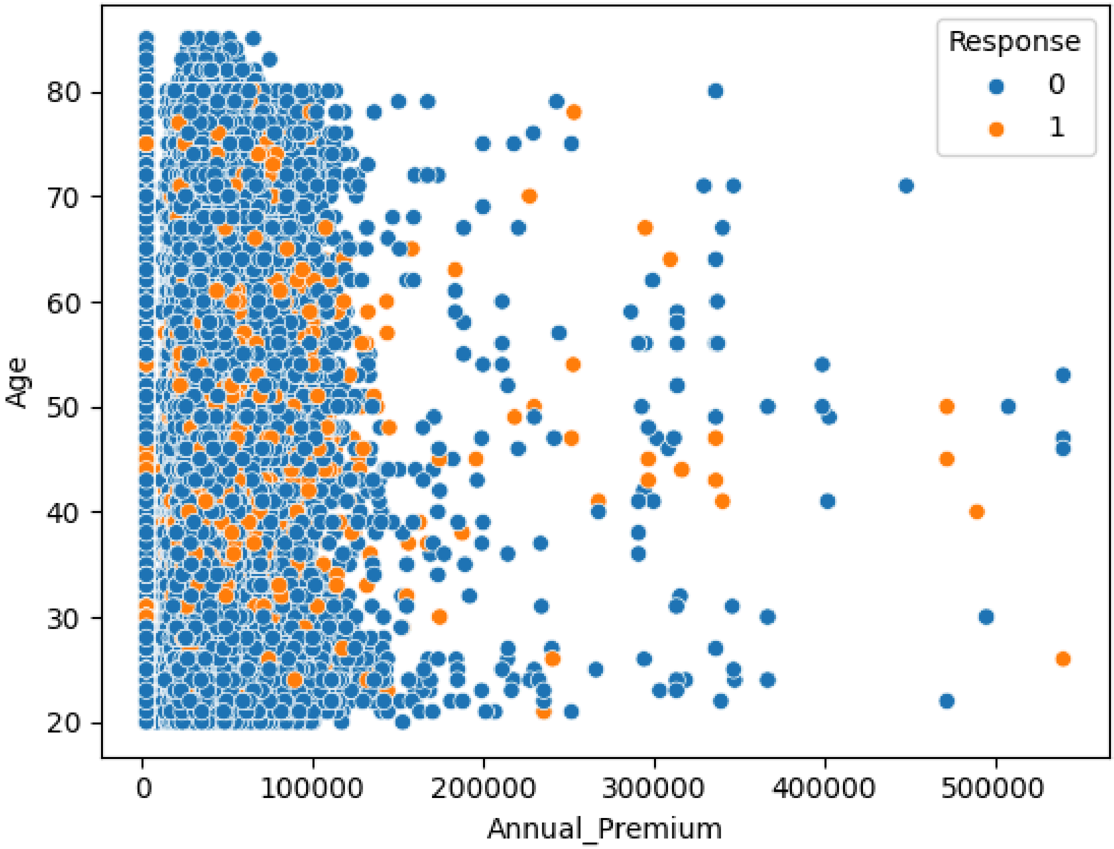




- Previously_Insured: People who were not previously insured (0) are much more likely to say yes (orange bar) to buying car insurance now than those who were previously insured (1), where almost everyone says no (blue bar).
- Driving_License: Nearly everyone in the data has a driving license (1), but having a license doesn't make a big difference—most of these people say no to the insurance, and only a smaller group says yes. Very few people without a driving license.
- Vehicle_Damage: Customers whose vehicles were previously damaged ("Yes") are more likely to say yes to insurance, while those with no past damage almost always say no.
- Vehicle_Age: People with cars aged 1–2 years are the most likely to say yes to insurance. Very few people with cars older than 2 years or less than 1 year say yes—the vast majority in those groups say no.

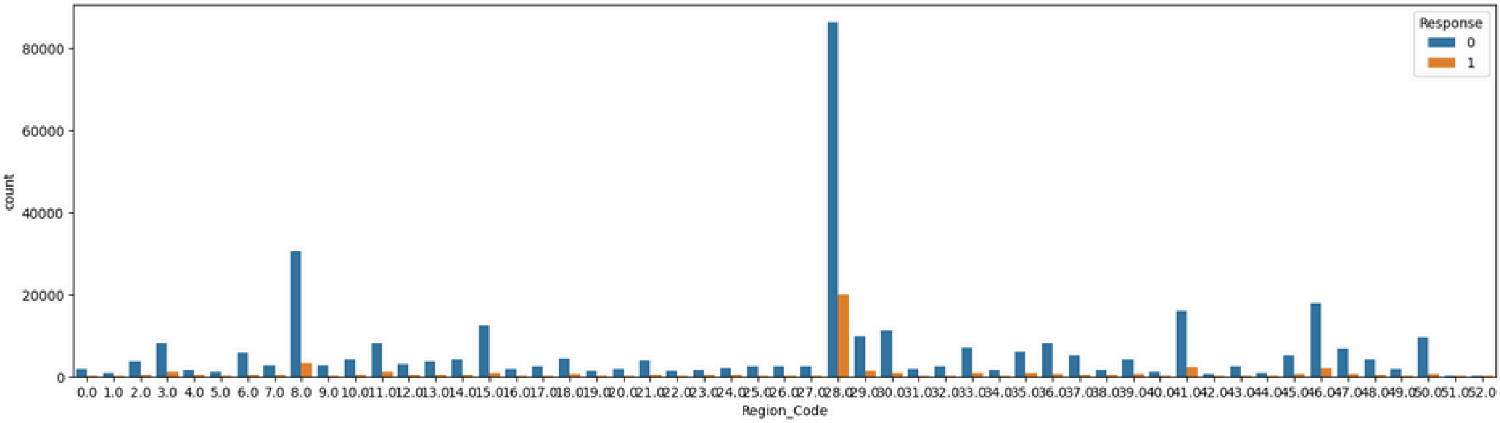
In short:

- Not being previously insured, having prior vehicle damage, and car age (especially 1–2 years) are linked to a higher chance the customer will accept insurance.
- Driving license status isn't a strong differentiator, since almost everyone has one, and response is low regardless.



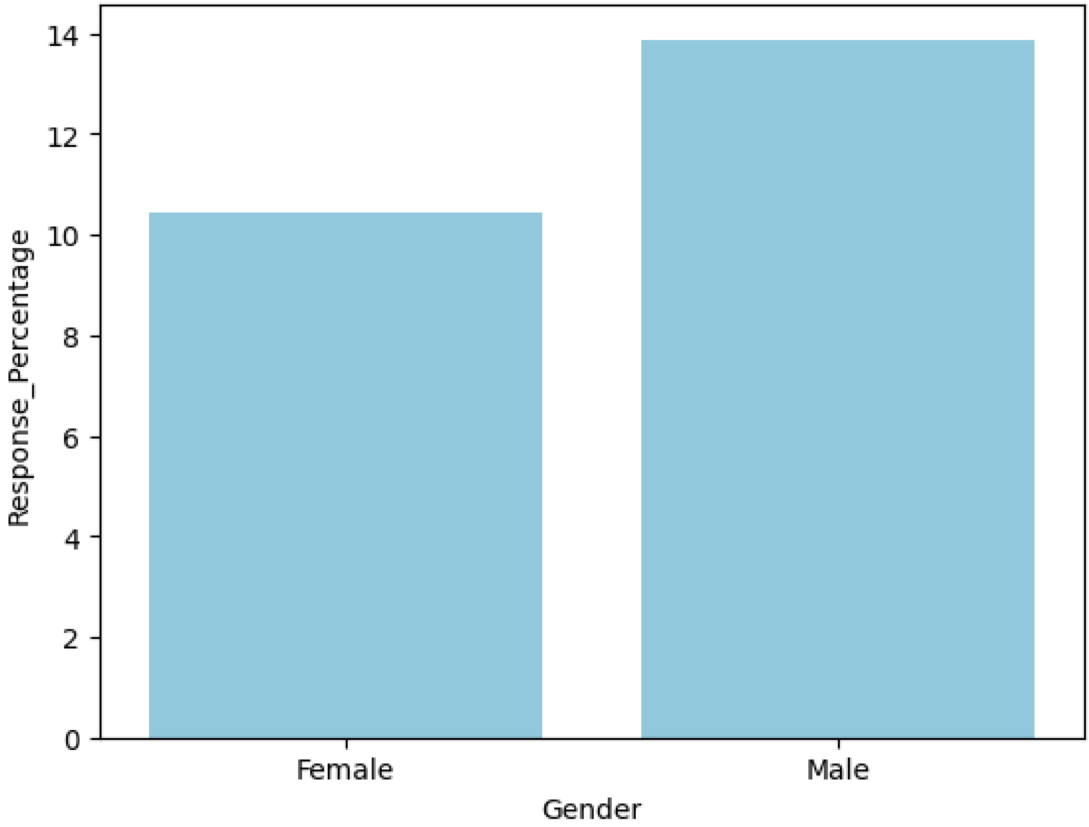
Age vs Annual_Premium by Response

This scatter plot shows how “Age” and “Annual_Premium” relate for both groups—those who accepted insurance and those who didn’t. Both “yes” (orange) and “no” (blue) responses are spread across all ages and premium amounts, but most points cluster in lower premium ranges (below about 150,000). There isn’t a strong difference in premium or age between those who said yes or no, indicating people across all ages and premium ranges behave similarly. A small number pay very high premiums, but those are exceptions.

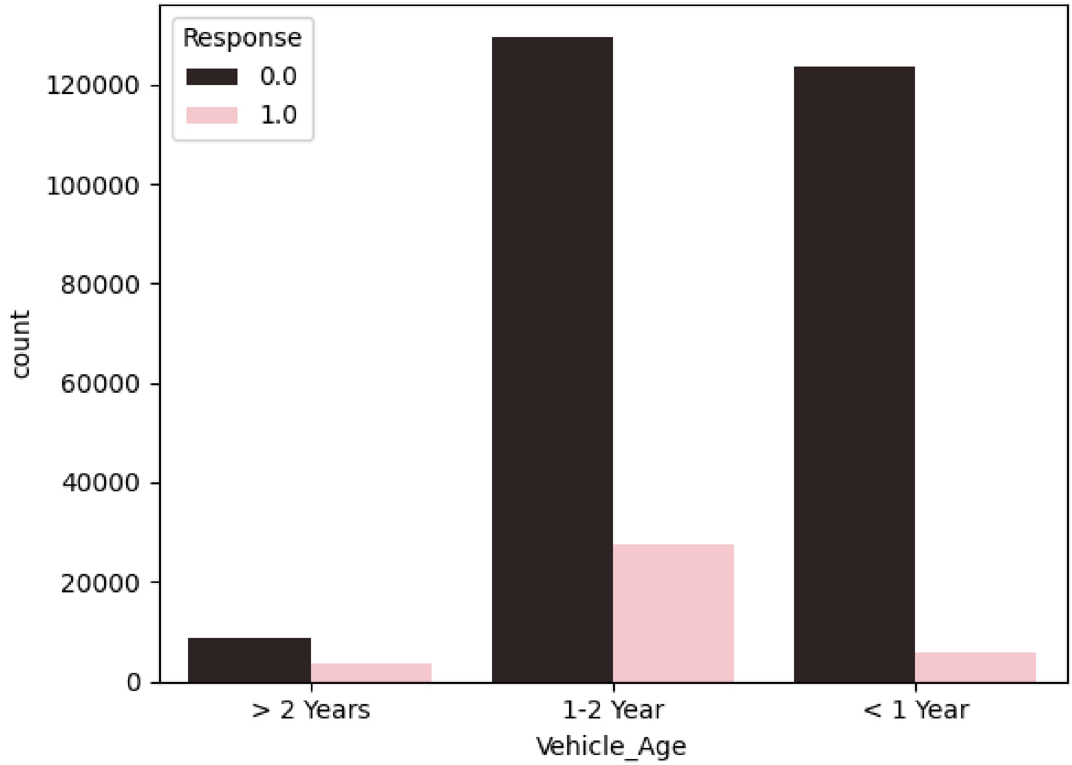


Region_Code vs Response

This bar chart shows the number of people in each Region_Code and how many said yes ("Response" = 1) or no ("Response" = 0) to car insurance. Most customers are clustered in a few specific regions, with some Region_Codes having many more policyholders than others. In every region, far more people say no than yes, but a few regions stand out, showing comparatively higher counts of yes responses. This suggests that where a person lives may influence their likelihood of accepting car insurance, likely due to demographic or local market factors.

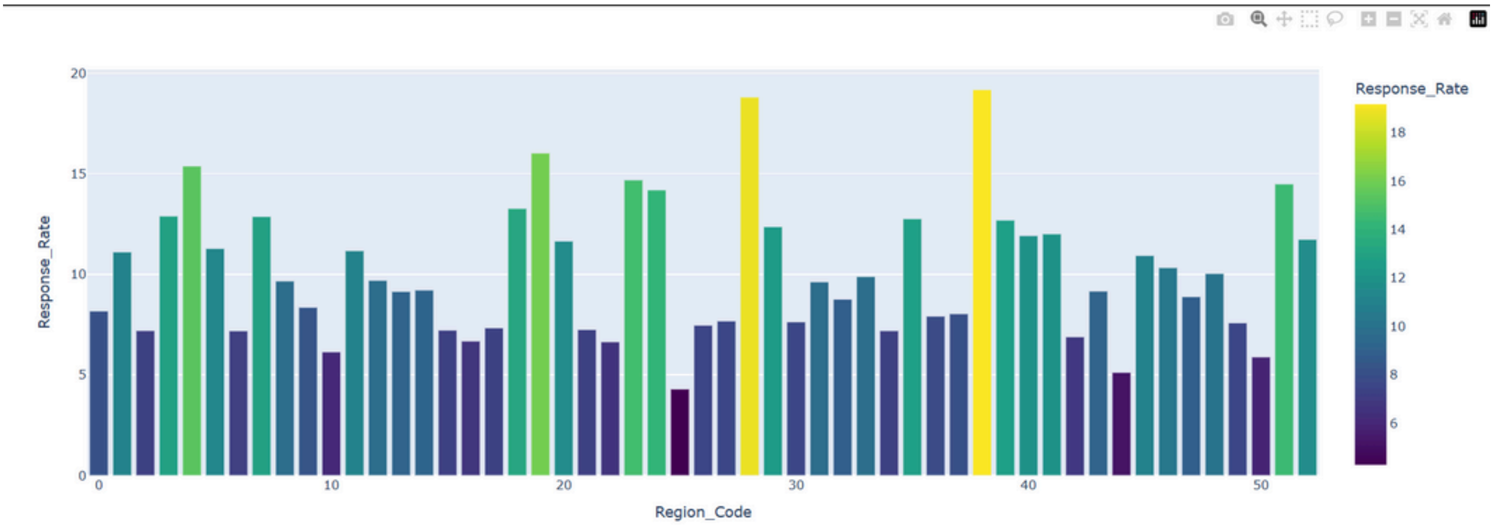


Gender vs Response Percentage
This bar chart compares the percentage of males and females who responded "yes" to car insurance. Males have a noticeably higher acceptance rate (~14%) compared to females (~10%). This means that, all else equal, men are more likely to say yes to buying car insurance than women. While the difference isn’t massive, it is consistent and might be useful for tailoring marketing strategies or understanding gender-based preferences in insurance.

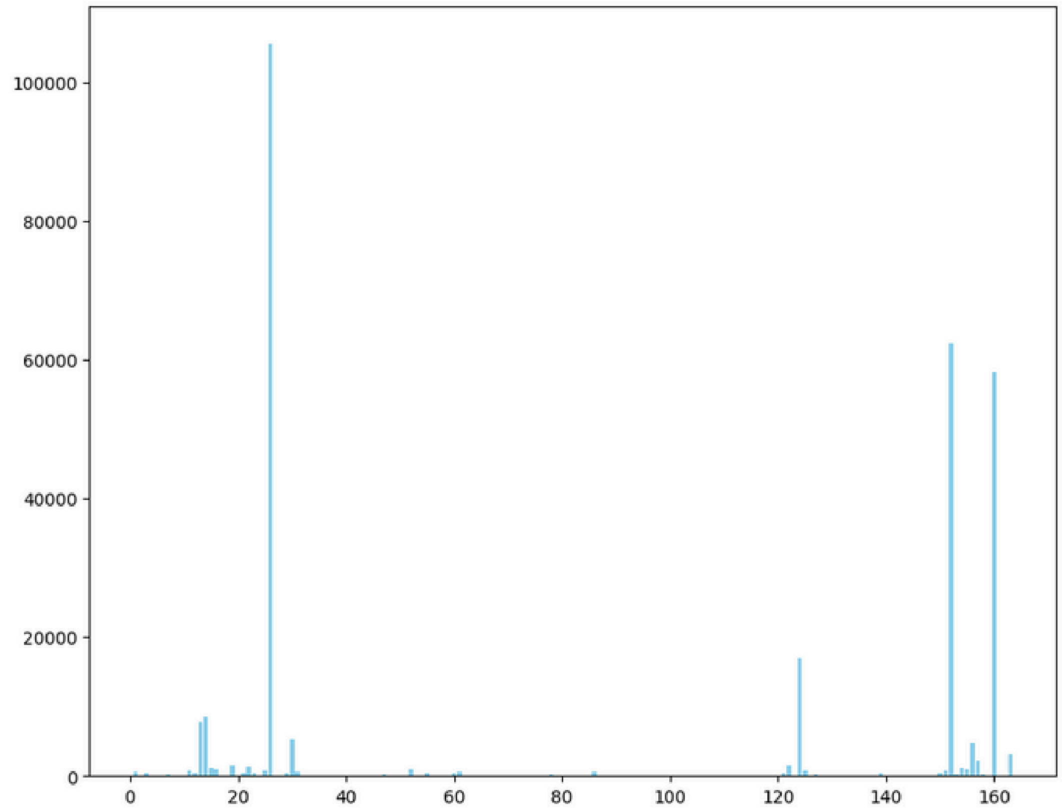


Vehicle_Age vs Response

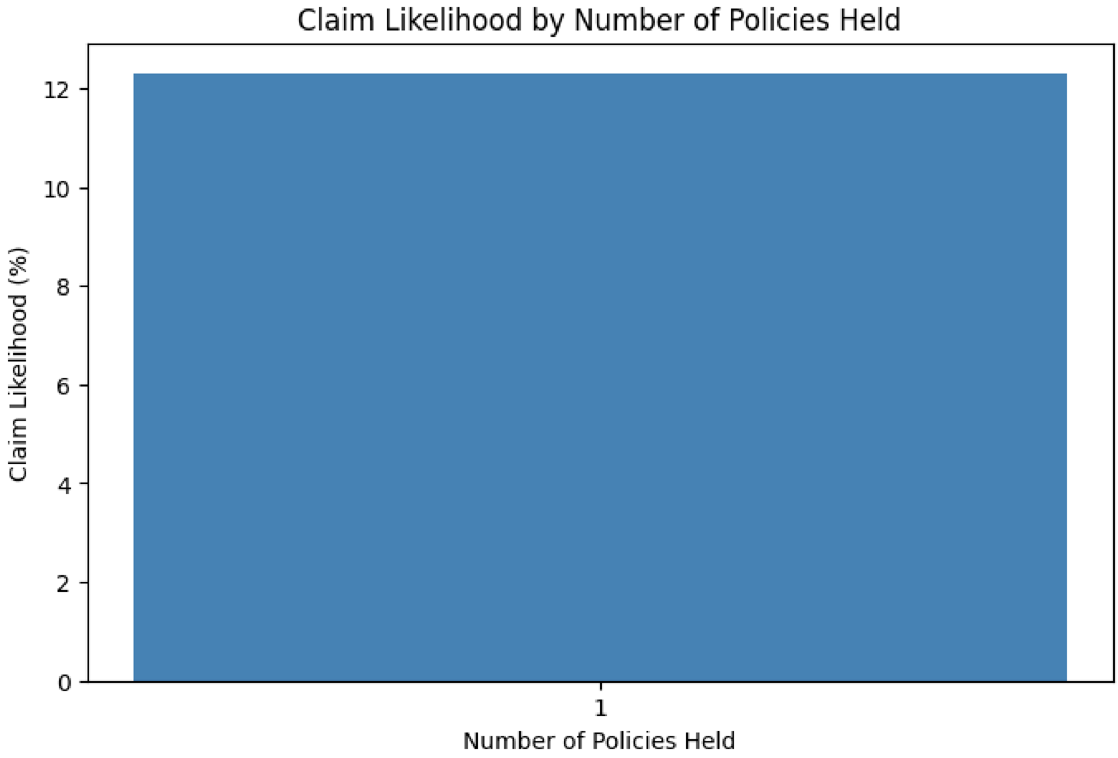
This chart shows how car age affects the likelihood of saying yes (“Response” = 1) to insurance. People with cars aged 1–2 years are much more likely to accept insurance, with a significantly higher count of “yes” responses compared to cars that are less than 1 year or older than 2 years. In the “<1 Year” and “>2 Years” groups, nearly everyone says no. This suggests that customers are most open to buying insurance for cars that are neither very new nor very old.



This bar chart illustrates how the acceptance rate of car insurance (“Response” = yes or no) varies across different geographic regions, identified by their Region_Code. Most regions show a similar pattern where the majority of potential customers decline the insurance (say no), but there are notable exceptions. Some specific Region_Codes have a significantly higher proportion and count of people saying yes to car insurance compared to others. This suggests that regional factors such as local demographics, economic conditions, cultural attitudes, or even regional marketing efforts might influence how likely residents are to accept car insurance offers. Understanding these differences can help insurance providers tailor their strategies to target regions with lower acceptance rates or expand successful approaches from regions with higher uptake. Overall, while most regions have relatively low response rates, the variation seen here highlights the importance of geography in customer behavior related to car insurance decisions.



The Policy_Sales_Channel variable represents the various methods through which car insurance policies are sold, such as agents, banks, online platforms, or call centers. The distribution graph shows that most policies are sold through just a few key channels, which appear as tall spikes in the graph. This concentration indicates that a small number of sales channels dominate the market. Understanding which channels are most effective helps insurance companies focus their marketing and sales strategies on those routes, improving customer reach and business performance while also identifying areas to expand or diversify their sales efforts.



This bar chart shows the relationship between how many insurance policies a person holds and their likelihood of making a claim (shown as a percentage). In this chart, we see that customers who hold just one policy have a claim likelihood of about 12%. Since only one bar appears, it suggests either almost all customers are single-policy holders, or the data is heavily focused on those with just one policy. This means, for most people in the data, having a single policy corresponds with a moderate chance of making a claim.

CONCLUSION

The dataset reveals that customer acceptance of car insurance ("Response") is influenced by a few important factors but lacks strong linear relationships between most features, suggesting more complex dynamics at play.

Key takeaways include:

- Previously_Insured and Vehicle_Damage have the strongest connections to whether customers buy insurance. Those not previously insured and those with prior vehicle damage are more likely to say yes.
- Car Age matters: Customers with cars 1–2 years old show the highest acceptance rates, while very new or older cars have lower acceptance.
- Demographics and Geography: Middle-aged people (30–45 years) respond more positively, and some regions show higher acceptance rates than others, indicating local factors affect buying behavior.
- Policy Sales Channels: Insurance policies are mostly purchased through a few dominant sales channels, which insurance companies can target to maximize sales efficiency.
- Low Multicollinearity: Most features are fairly independent, making the data suitable for modeling but requiring more advanced techniques beyond simple correlation to capture patterns.
- Gender and License: Males have slightly higher acceptance rates than females, and driving license status doesn't strongly influence response since most customers have one.
- Premiums and Claims: Premium amounts tend to be similar regardless of claim history, indicating premiums aren't drastically different based on claim likelihood.

Overall, the insights highlight that customer behavior around car insurance is shaped by specific personal and vehicle factors, regional differences, and sales channels.

ACKNOWLEDGEMENT

I WOULD LIKE TO TAKE THIS OPPORTUNITY TO EXPRESS MY GRATITUDE FOR THE RESOURCES, TOOLS, AND KNOWLEDGE THAT ENABLED ME TO SUCCESSFULLY COMPLETE THIS PROJECT.

AS THIS WAS AN INDEPENDENT EFFORT, I AM PROUD TO HAVE CARRIED OUT ALL ASPECTS OF THE PROJECT—FROM DATA CLEANING AND VISUALIZATION TO ANALYSIS AND DOCUMENTATION—ON MY OWN. THIS EXPERIENCE ALLOWED ME TO APPLY THEORETICAL KNOWLEDGE TO A REAL-WORLD DATASET AND DEVELOP A DEEPER UNDERSTANDING OF DATA SCIENCE CONCEPTS AND PRACTICAL TOOLS SUCH AS PYTHON, PANDAS, MATPLOTLIB, AND SEABORN.

THIS PROJECT HAS BEEN A VALUABLE LEARNING JOURNEY, AND I AM GRATEFUL FOR THE OPPORTUNITY TO EXPLORE AND GROW THROUGH SELF-DRIVEN WORK.