MINI PROJECT -Exploratory Data Analysis of Vehicle Insurance Trends

Step 1. Data Loading and Inspection:

Load the Dataset

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
df = pd.read_csv("C:/Users/Dell/Downloads/Vehicle_Insurance.csv")
```

Basic Structure using head() function:

```
df.head()
```

```
   id  Gender  Age  Driving_License  Region_Code
Previously_Insured  \
0   1    Male   44               1         28.0                    0

1   2    Male   76               1          3.0                    0

2   3    Male   47               1         28.0                    0

3   4    Male   21               1         11.0                    1

4   5  Female   29               1         41.0                    1


  Vehicle_Age Vehicle_Damage  Annual_Premium  Policy_Sales_Channel
Vintage  \
0   > 2 Years            Yes         40454.0                   26.0
217
1    1-2 Year             No         33536.0                   26.0
183
2   > 2 Years            Yes         38294.0                   26.0
27
3    < 1 Year             No         28619.0                  152.0
203
4    < 1 Year             No         27496.0                  152.0
39

   Response
0         1
1         0
2         1
```

```
3         0
4         0

df.tail()

            id  Gender  Age  Driving_License  Region_Code
Previously_Insured  \
381104  381105    Male   74                1         26.0
1
381105  381106    Male   30                1         37.0
1
381106  381107    Male   21                1         30.0
1
381107  381108  Female   68                1         14.0
0
381108  381109    Male   46                1         29.0
0

       Vehicle_Age Vehicle_Damage  Annual_Premium
Policy_Sales_Channel  \
381104    1-2 Year             No         30170.0
26.0
381105    < 1 Year             No         40016.0
152.0
381106    < 1 Year             No         35118.0
160.0
381107   > 2 Years            Yes         44617.0
124.0
381108    1-2 Year             No         41777.0
26.0

        Vintage  Response
381104       88         0
381105      131         0
381106      161         0
381107       74         0
381108      237         0
```

Overview of Column Information using info() method:

```
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
```

Summary Statistics using describe() method:

```
df.describe()
```

```
                  id            Age  Driving_License    Region_Code  \
count  381109.000000  381109.000000    381109.000000  381109.000000
mean   190555.000000      38.822584         0.997869      26.388807
std    110016.836208      15.511611         0.046110      13.229888
min         1.000000      20.000000         0.000000       0.000000
25%     95278.000000      25.000000         1.000000      15.000000
50%    190555.000000      36.000000         1.000000      28.000000
75%    285832.000000      49.000000         1.000000      35.000000
max    381109.000000      85.000000         1.000000      52.000000

       Previously_Insured  Annual_Premium  Policy_Sales_Channel  \
count       381109.000000   381109.000000         381109.000000
mean             0.458210    30564.389581            112.034295
std              0.498251    17213.155057             54.203995
min              0.000000     2630.000000              1.000000
25%              0.000000    24405.000000             29.000000
50%              0.000000    31669.000000            133.000000
75%              1.000000    39400.000000            152.000000
max              1.000000   540165.000000            163.000000

             Vintage       Response
count  381109.000000  381109.000000
mean      154.347397       0.122563
std        83.671304       0.327936
min        10.000000       0.000000
25%        82.000000       0.000000
50%       154.000000       0.000000
75%       227.000000       0.000000
max       299.000000       1.000000
```

Shape and size of the dataset :

```
df.shape
```

```
(381109, 12)
```

```
df.size
```

```
4573308
```

Step 2. Data Cleaning:

Identify and Handle Missing Values

```
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
```

Checking unique values

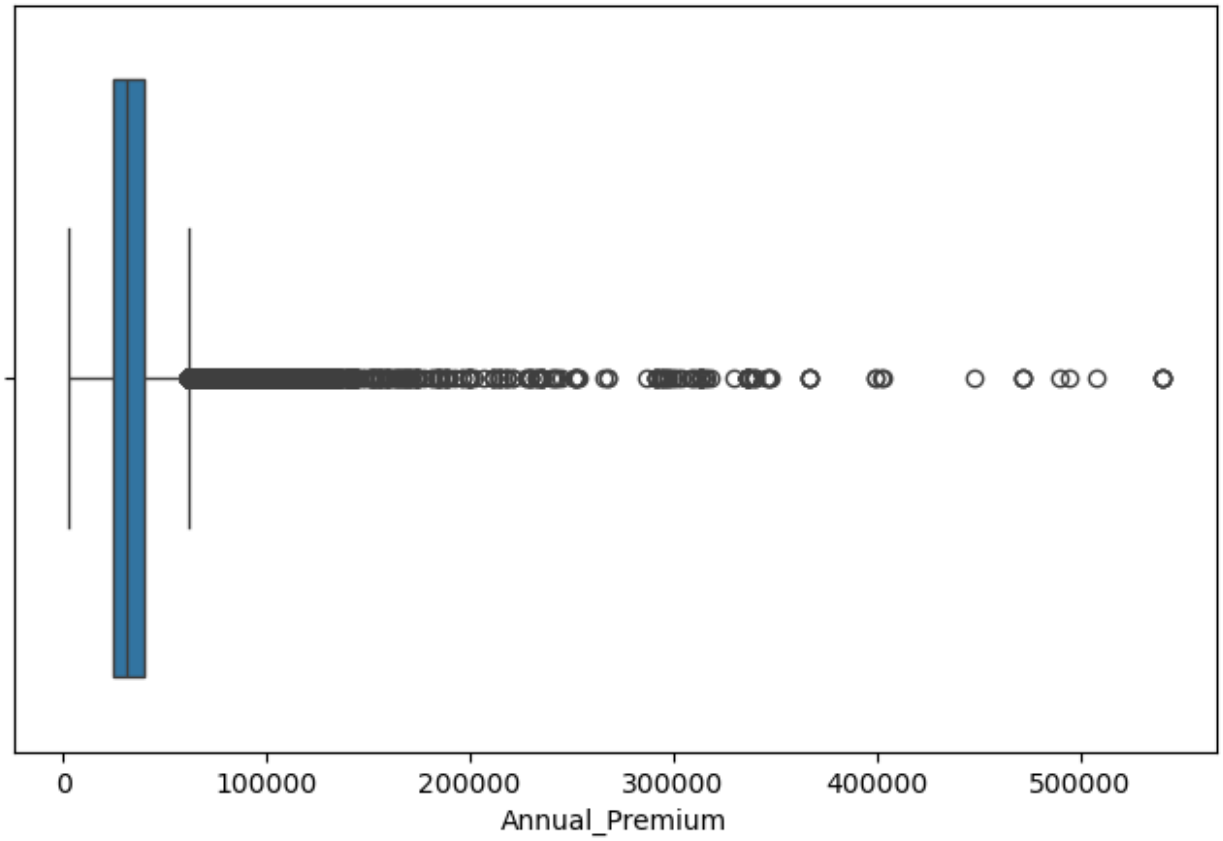
```
for column in df.columns:
    unique_count = df[column].nunique()
    print(f"'{column}' has {unique_count} unique values")
```

```
'id' has 381109 unique values
'Gender' has 2 unique values
'Age' has 66 unique values
'Driving_License' has 2 unique values
'Region_Code' has 53 unique values
'Previously_Insured' has 2 unique values
'Vehicle_Age' has 3 unique values
'Vehicle_Damage' has 2 unique values
'Annual_Premium' has 48838 unique values
'Policy_Sales_Channel' has 155 unique values
'Vintage' has 290 unique values
'Response' has 2 unique values
```

Checking outliers with boxplot

```
plt.figure(figsize=(8, 5))
sns.boxplot(x=df['Annual_Premium'])
```

```
plt.show()
```



Findings outlieries using boxplot

1.Significant outliers present, with values exceeding 500,000.

2.Distribution is heavily right-skewed.

3.Majority of Annual_Premium values are clustered below 100,000.

4.Suggest investigating outliers and applying data transformation to reduce skewness.

Checking outliers in a column using the IQR method with count

```
def find_outliers(d,col):
 q1=d[col].quantile(0.25)
 q3=d[col].quantile(0.75)
 IQR=q3-q1
 lower_limit= q1-1.5*IQR
 upper_limit= q3+1.5*IQR
 return lower_limit,upper_limit

a= ['Age', 'Region_Code', 'Annual_Premium', 'Policy_Sales_Channel',
```

```
'Vintage']
for i in a:
 print(find_outliers(df,i))
```

```
(-11.0, 85.0)
(-15.0, 65.0)
(3493.5, 59257.5)
(-155.5, 336.5)
(-135.5, 444.5)
```

Filling outliers with mean

```
df["Age"]=df["Age"].apply(lambda x: df["Age"].mean() if(x<-11.0 or
x>85.0) else x)
df["Age"].head(5)
```

```
0    44
1    76
2    47
3    21
4    29
Name: Age, dtype: int64
```

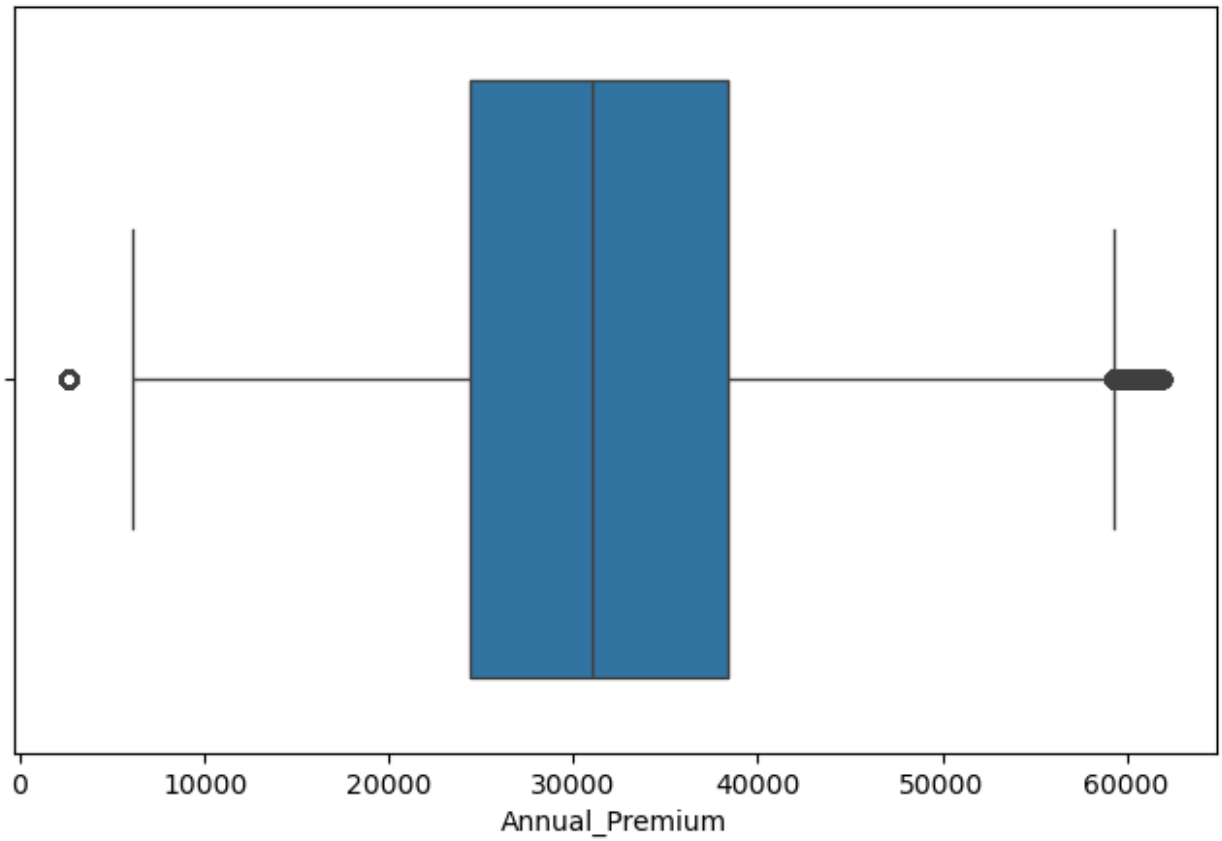
```
df["Region_Code"]= df["Region_Code"].apply(lambda x:
df["Region_Code"].mean() if(x<-11.0 or x>65.0) else x)
df["Region_Code"].head(5)
```

```
0    28.0
1     3.0
2    28.0
3    11.0
4    41.0
Name: Region_Code, dtype: float64
```

```
df["Annual_Premium"]=df["Annual_Premium"].apply(lambda x:
df["Annual_Premium"].mean() if(x<1912.5 or x>61892.5) else x)
df["Annual_Premium"].head(5)
```

```
0    40454.0
1    33536.0
2    38294.0
3    28619.0
4    27496.0
Name: Annual_Premium, dtype: float64
```

```
plt.figure  (figsize=(8, 5))
sns.boxplot(x=df['Annual_Premium'])

plt.show()
```

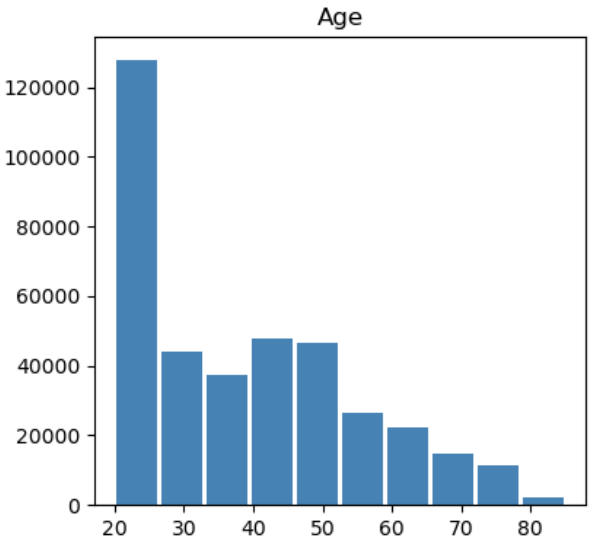


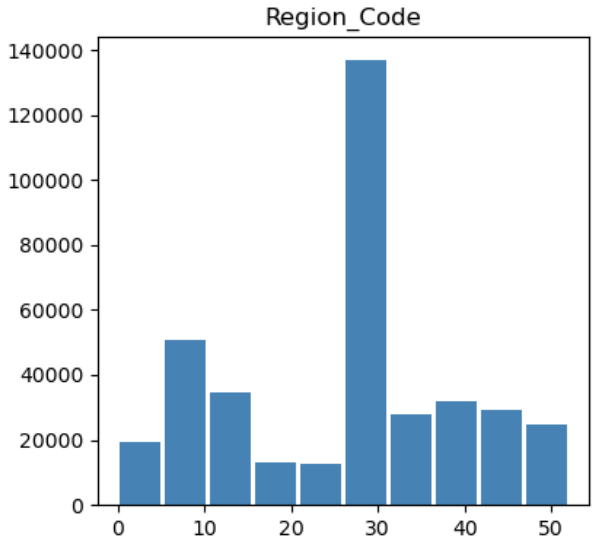
Step 3: Data Visulization

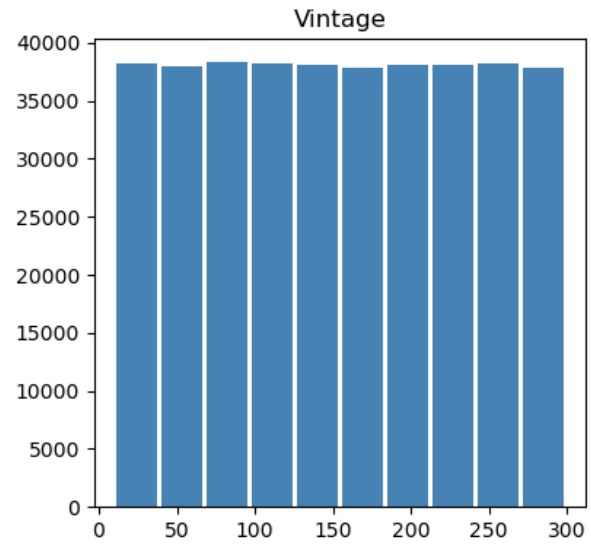
Univariate Analysis

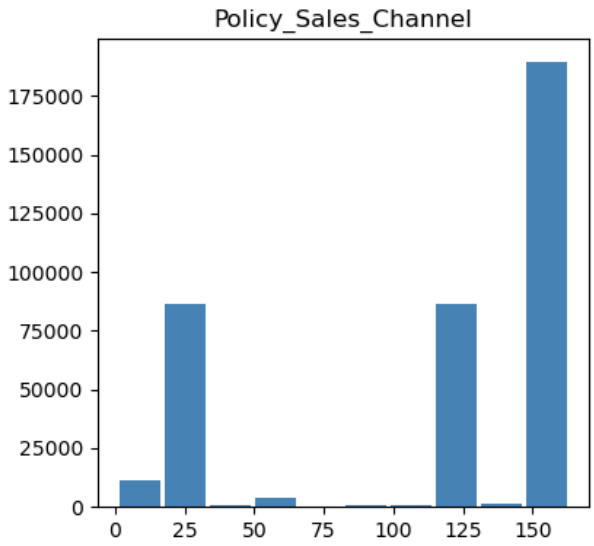
3.1 ploting graphs (histograms)

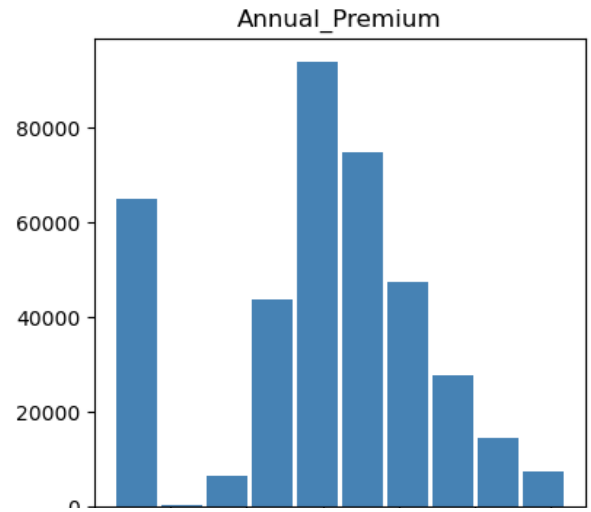
```
df.hist(column=['Age','Region_Code',
                        'Vintage','Policy_Sales_Channel',
                        'Annual_Premium', 'Previously_Insured'],
                color='steelblue',

                figsize=(10,15),
                bins=10,
                rwidth=0.9,
                grid=False #
               )
plt.show()
```

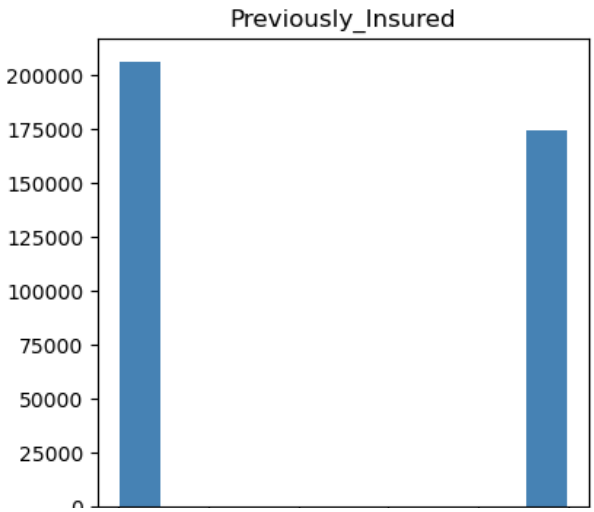

Age
120000
100000
80000
60000
40000
20000
0
20
30
40
50
60
70
80


Region_Code
140000
120000
100000
80000
60000
40000
20000
0
0
10
20
30
40
50


Vintage
40000
35000
30000
25000
20000
15000
10000
5000
0
0
50
100
150
200
250
300


Policy_Sales_Channel
175000
150000
125000
100000
75000
50000
25000
0
0
25
50
75
100
125
150


Annual_Premium
80000
60000
40000
20000


Previously_Insured
200000
175000
150000
125000
100000
75000
50000
25000

Histogram Findings:

1. Age: Most of the policyholders are younger adults, especially in their 20s and 30s, with fewer older customers.There is a noticeable decline of 50% in the number of insured individuals as the age increases beyond 50, compared to the 20-25 age group.

2. Region: There's one region where a lot more people have policies, meaning the company has a big presence or popularity there.

3. Policy Duration: Policies are spread out evenly in terms of how long customers have had them, showing a steady stream of new customers over time.

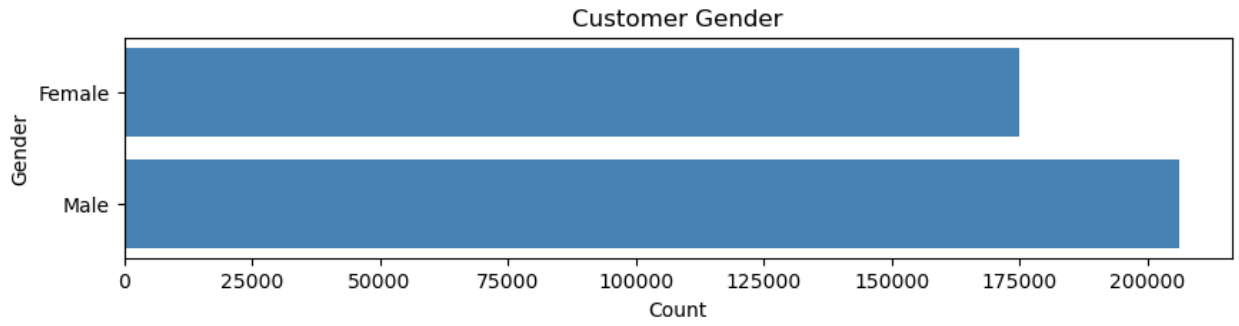
4. Sales Channels: One sales channel is doing most of the work, so it's likely the go-to method for reaching new customers.

5. Annual Premiums: Most people are paying mid-range premiums, suggesting that the company is focused on attracting everyday, middle-income drivers.The majority of the data points are clustered towards the lower end of the distribution, while a smaller number of data points exhibit higher premium values

6. Previous Insurance: A lot of customers didn't have insurance before, so the company is doing well at reaching people who are new to insurance altogether.

3.2 Customer gender distribution

```
gender_counts = df['Gender'].value_counts()
plt.figure(figsize=(10, 2))
plt.barh(gender_counts.index, gender_counts.values, color='steelblue')
plt.title('Customer Gender')
plt.xlabel('Count')
plt.ylabel('Gender')

plt.show()
```



Findings:

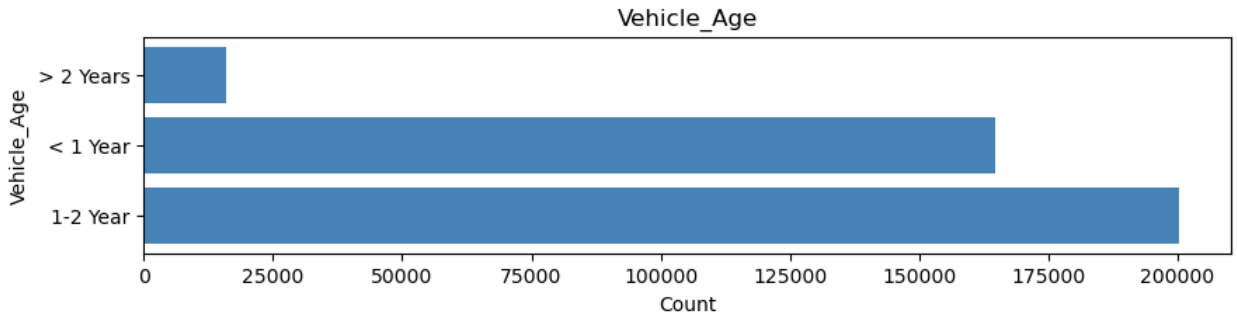
There are more males than females in the dataset. The difference between male and female counts is noticeable but not extremely large.

Male: The count is slightly over 200,000

Female: The count is somewhat below 200,000.

3.3 Vehicle Age Analysis

```
age_counts = df['Vehicle_Age'].value_counts()
plt.figure(figsize=(10, 2))
plt.barh(age_counts.index, age_counts.values, color='steelblue')
plt.title('Vehicle_Age')
plt.xlabel('Count')
plt.ylabel('Vehicle_Age')

plt.show()
```



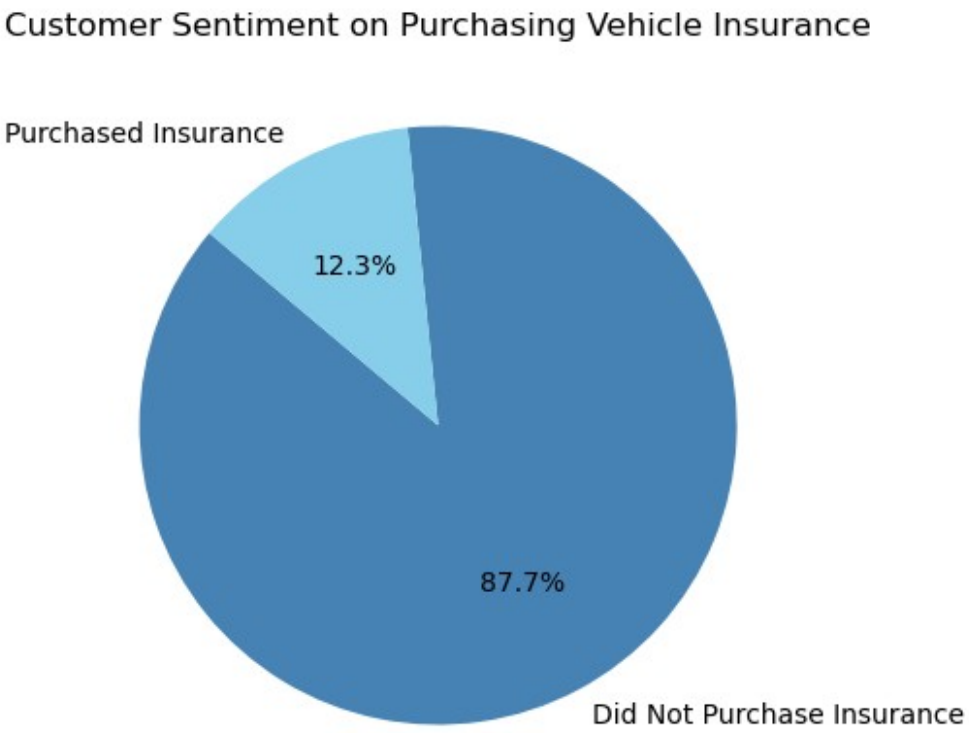
The count distribution across different time categories:

1-2 Year: This category has the highest count, reaching over 200,000.

< 1 Year: The count for this category is lower than the 1-2 Year category but still quite high, around 175,000.

> 2 Years: This category has a significantly lower count, much less than 50,000.

```
counts = df['Response'].value_counts()
labels = ['Did Not Purchase Insurance', 'Purchased Insurance']
plt.figure(figsize=(5, 5))
plt.pie(counts, labels=labels, autopct='%1.1f%%', startangle=140,
colors=['steelblue', 'skyblue'])
plt.title('Customer Sentiment on Purchasing Vehicle Insurance')
plt.show()
```



The pie chart illustrates customer sentiment regarding vehicle insurance purchases:

1) 87.7% of customers did not purchase vehicle insurance.

2) Only 12.3% of customers purchased vehicle insurance.

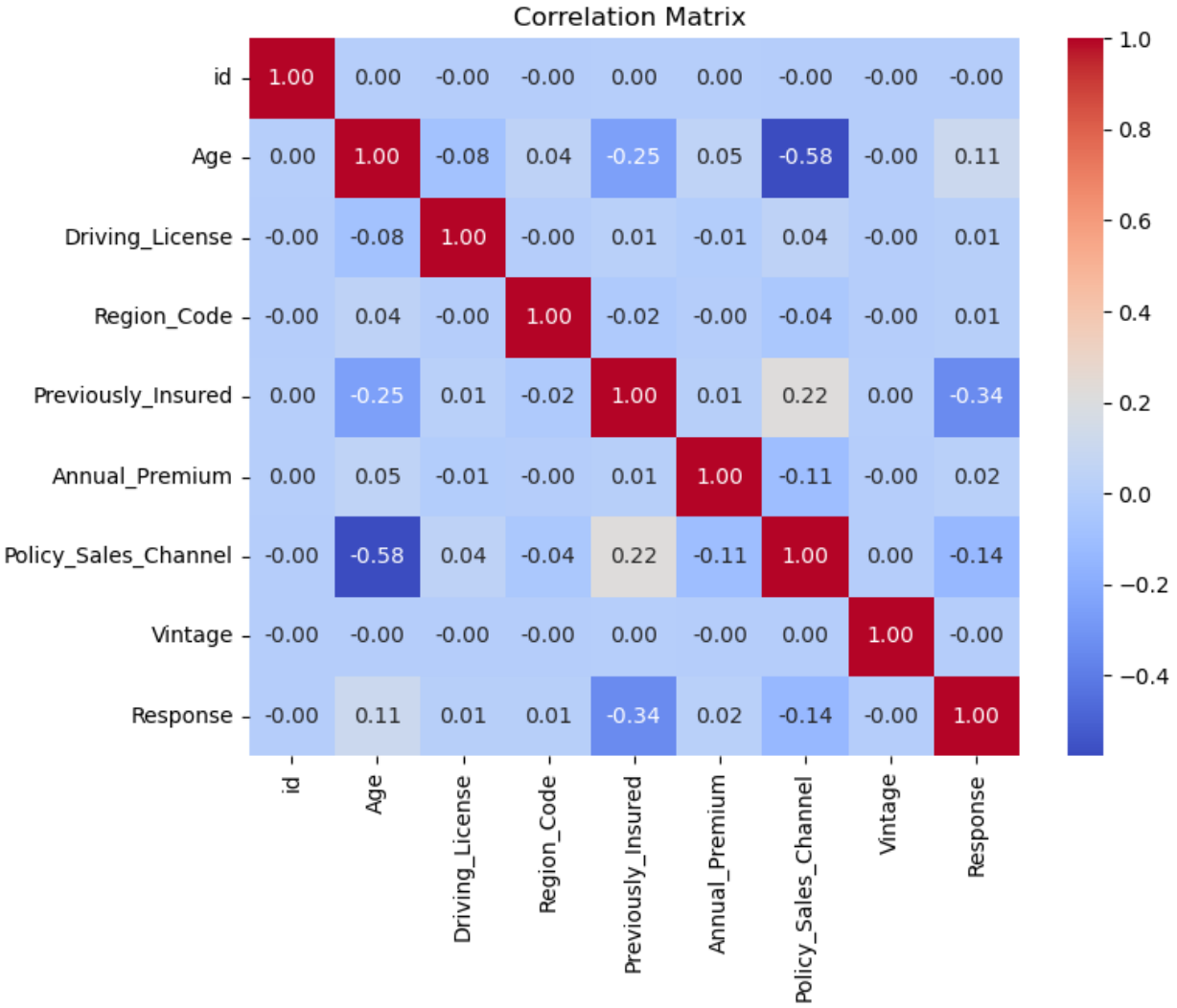
The overwhelming majority of customers chose not to buy insurance. This indicates a low conversion rate for vehicle insurance among the surveyed group.

Step 4: Features Analysis

4.1 Heatmap Analysis-

Analyse the relation between Target and Feature variable

```
numeric_df = df.select_dtypes(include=['float64', 'int64'])
correlation_matrix = numeric_df.corr()

plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm',
fmt=".2f")
plt.title('Correlation Matrix ')
plt.show()
```


Correlation Matrix
id
Age
Driving_License
Region_Code
Previously_Insured
Annual_Premium
Policy_Sales_Channel
Vintage
Response
1.00 0.00 -0.00 -0.00 0.00 0.00 -0.00 -0.00 -0.00
0.00 1.00 -0.08 0.04 -0.25 0.05 -0.58 -0.00 0.11
-0.00 -0.08 1.00 -0.00 0.01 -0.01 0.04 -0.00 0.01
-0.00 0.04 -0.00 1.00 -0.02 -0.00 -0.04 -0.00 0.01
0.00 -0.25 0.01 -0.02 1.00 0.01 0.22 0.00 -0.34
0.00 0.05 -0.01 -0.00 0.01 1.00 -0.11 -0.00 0.02
-0.00 -0.58 0.04 -0.04 0.22 -0.11 1.00 0.00 -0.14
-0.00 -0.00 -0.00 -0.00 0.00 -0.00 0.00 1.00 -0.00
-0.00 0.11 0.01 0.01 -0.34 0.02 -0.14 -0.00 1.00
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Findings:

Key Negative Relationships:

Sales channels are less used by older people (Policy_Sales_Channel vs. Age: -0.58).People who were already insured tend not to respond to offers (Previously_Insured vs. Response: -0.34).

Positive but Weak Connections:

Sales channel use slightly increases with prior insurance (Policy_Sales_Channel vs. Previously_Insured: 0.22).Older people are a bit more likely to respond (Age vs. Response: 0.11).

Most Features Don't Strongly Relate to Each Other or the Target (Response):

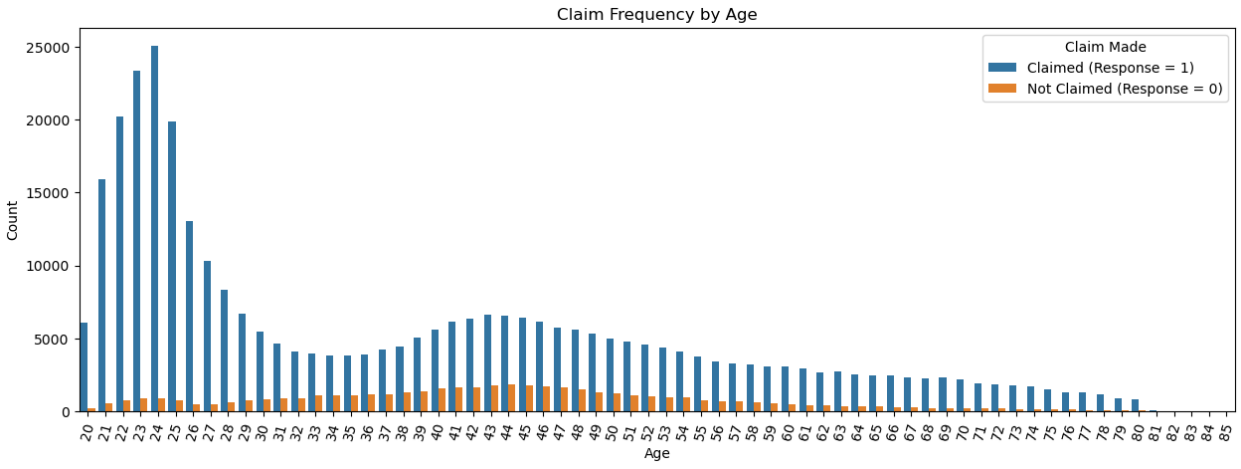
Variables like Region_Code, Driving_License, and Vintage aren't making much of an impact.

Step 5 : Age Distribution

5.1 Age Distribution Analysis

Analyze the age distribution within the dataset and its impact on insurance claims.

```
plt.figure(figsize=(15,5))
sns.countplot(data=df,x="Age",hue="Response")
plt.title('Claim Frequency by Age')
plt.xlabel('Age')
plt.ylabel('Count')
plt.xticks(rotation=75)
plt.legend(title='Claim Made', labels=['Claimed (Response = 1)', 'Not Claimed (Response = 0)'])
plt.show()
```



Findings from the countplot:

The dataset has the highest number of individuals aged 20-25. The frequency of individuals steadily decreases after age 30. Older individuals (above 70) are very few in the dataset.

Age 20-29: Dominates with the highest count of "No Claim" responses.

Older age groups (40-49, 50-59): Show moderate "No Claim" counts, with a slight increase in "Claim" responses compared to younger groups.

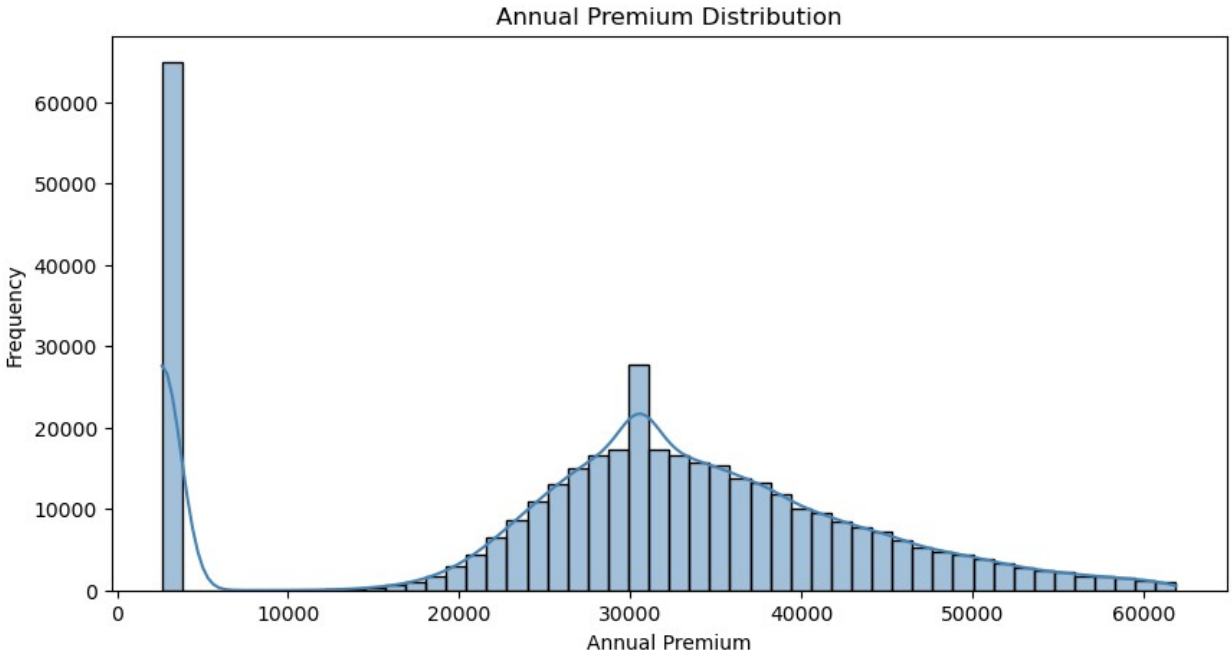
Very old groups (80+): Have significantly lower counts for both "Claim" and "No Claim" responses.

The overall trend suggests that claims are less frequent than no claims, especially in younger populations.

Step 6: Premium Analysis

6.1 The distribution of Annual Premium

```
plt.figure(figsize=(10, 5))
sns.histplot(df['Annual_Premium'], kde=True, bins=50,
color="steelblue")
plt.title('Annual Premium Distribution')
plt.xlabel('Annual Premium')
plt.ylabel('Frequency')
plt.show()
```



This histogram depicts the distribution of annual premiums:

1) High Frequency at Low Premiums: A significant spike near 0 indicates a high frequency of low premium values.

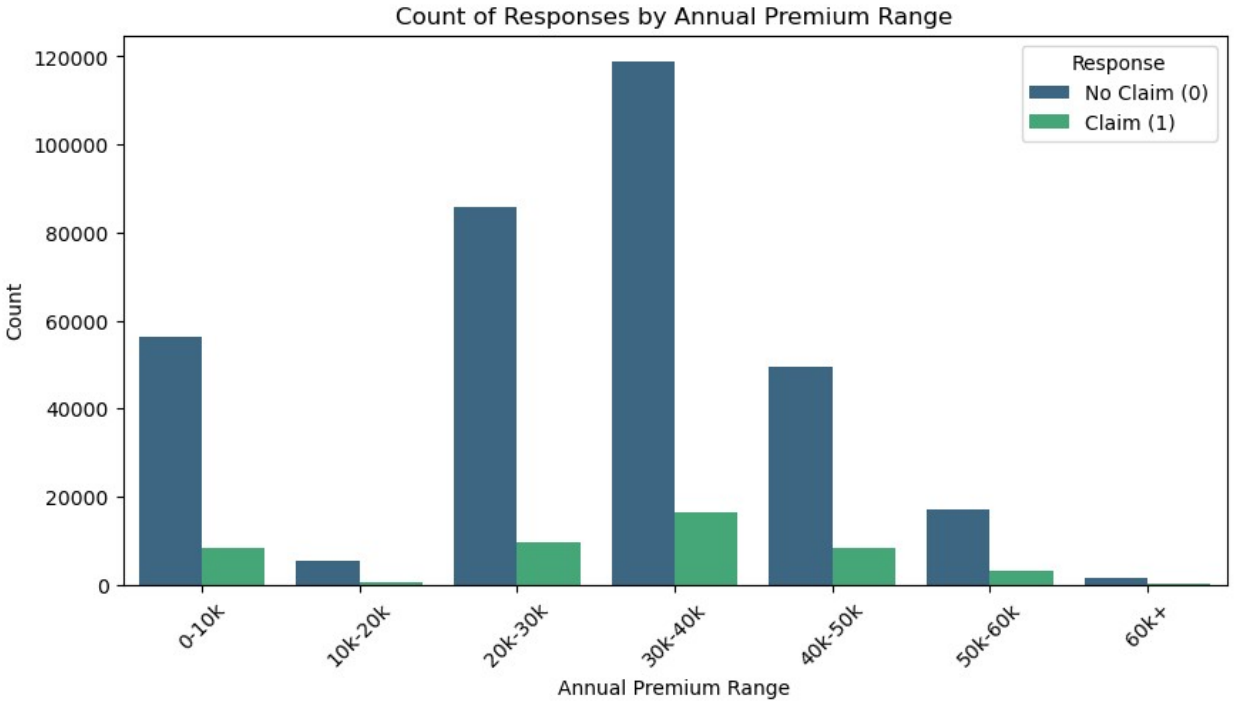
2) Bell-Shaped Distribution: Beyond the initial spike, the distribution exhibits a normal-like shape, peaking around 30,000.

3)Right Skewness: The tail extends toward higher premium values (above 30,000), indicating fewer high premium policies.

This suggests most customers pay moderate to low premiums, with a few outliers paying very high amounts.

6.2 Impact of Annual Premium on Response

```
bins = [0, 10000, 20000, 30000, 40000, 50000, 60000, 70000]
labels = ['0-10k', '10k-20k', '20k-30k', '30k-40k', '40k-50k', '50k-
60k', '60k+']
df['Premium_Range'] = pd.cut(df['Annual_Premium'], bins=bins,
labels=labels, right=False)

# Create the countplot
plt.figure(figsize=(10, 5))
sns.countplot(x='Premium_Range', hue='Response', data=df,
palette='viridis')
plt.title('Count of Responses by Annual Premium Range')
plt.xlabel('Annual Premium Range')
plt.ylabel('Count')
plt.xticks(rotation=45)
plt.legend(title='Response', labels=['No Claim (0)', 'Claim (1)'])
plt.show()
```


Count of Responses by Annual Premium Range
Response
No Claim (0)
Claim (1)
Count
120000
100000
80000
60000
40000
20000
0
0-10k
10k-20k
20k-30k
30k-40k
40k-50k
50k-60k
60k+
Annual Premium Range

This bar chart presents the count of Response (No Claim vs. Claim) grouped by ranges of Annual Premium. Here are the key findings:

Premium Range 0-10k:

Higher count of "No Claim" responses compared to "Claim" responses.

Premium Range 20k-30k and 30k-40k:

These ranges have the highest overall counts, dominated by "No Claim" responses. A noticeable but smaller proportion of "Claim" responses appears in these ranges.

Premium Range 40k-50k:

The count of "No Claim" responses decreases but remains significant."Claim" responses slightly increase compared to lower ranges.

Premium Range 50k-60k and 60k+:

Counts drop sharply for both "No Claim" and "Claim" responses, indicating fewer policies in higher premium ranges.

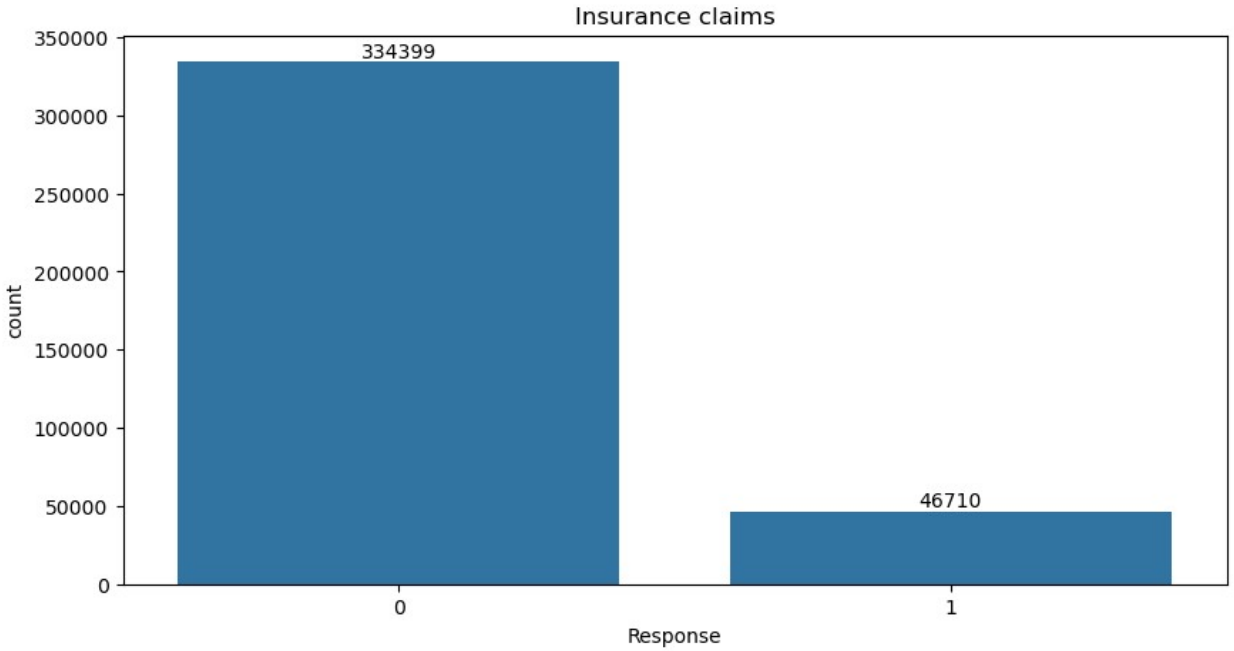
General Insight:

Lower and moderate premium ranges (0-40k) dominate the dataset."No Claim" responses significantly outnumber "Claim" responses across all premium ranges.Claims are relatively more frequent in the 30k-50k premium ranges compared to other ranges.

Step 7: Claim Frequency Analysis

7.1 Claim Frequency by Age

```
plt.figure(figsize=(10, 5))
ax=sns.countplot(data=df,x="Response")
plt.title("Insurance claims")
plt.bar_label(ax.containers[0])
plt.show()
```

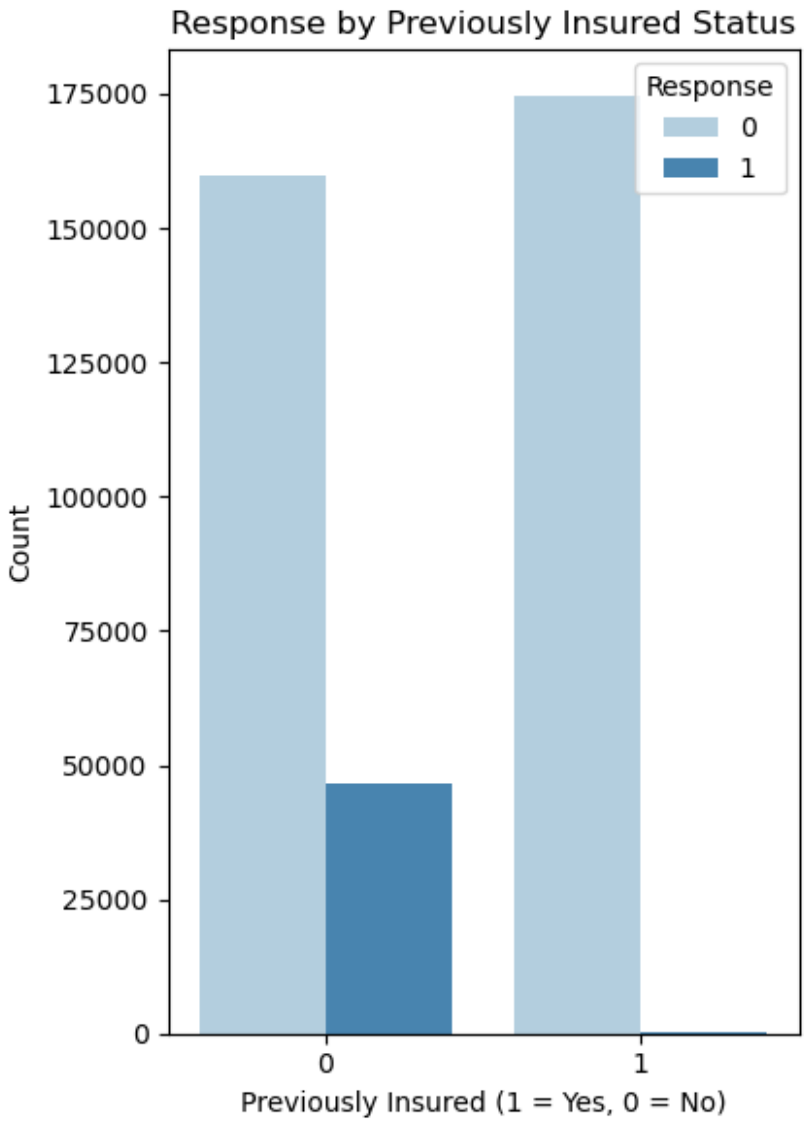


Findings:

Not Previously Insured (0): Many did not respond positively, though positive responses are higher than for the insured. Previously Insured (1): Very few responded positively, with most showing no response.

7.2 Claim Frequency by Previously Insured

```
plt.figure(figsize=(12, 6))
plt.subplot(1, 3, 2)
sns.countplot(data=df, x='Previously_Insured', hue='Response',
palette='Blues')
plt.title('Response by Previously Insured Status')
plt.xlabel('Previously Insured (1 = Yes, 0 = No)')
plt.ylabel('Count')

plt.tight_layout()
plt.show()
```



The chart shows responses based on insurance status: Uninsured individuals are slightly more likely to respond positively, though negative responses dominate both groups.

Not Previously Insured (0): Many did not respond positively, though positive responses are higher than for the insured.

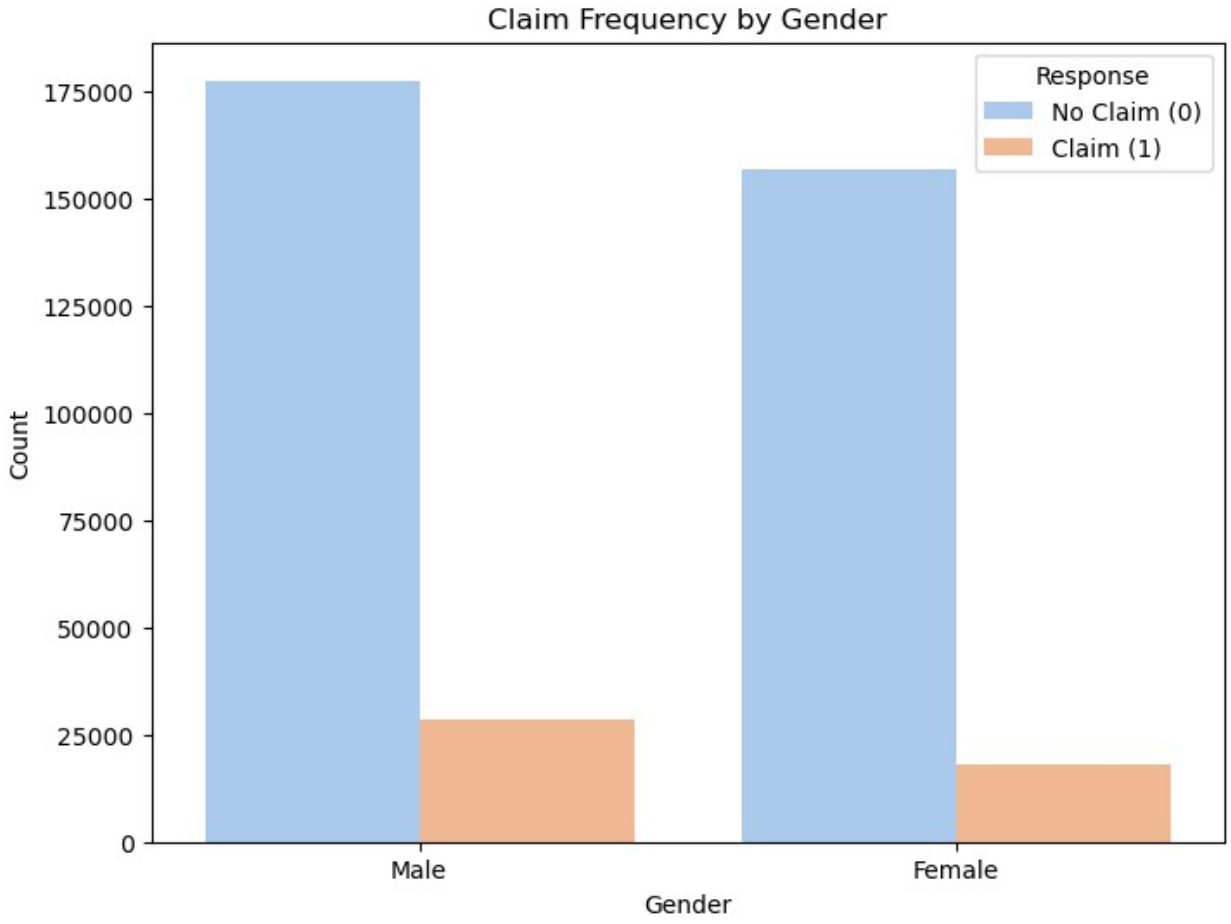
Previously Insured (1): Very few responded positively, with most showing no response.

Step 8: Gender Analysis

8.1 Claim Frequency by Gender:

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender', hue='Response', data=df, palette="pastel")
```

```
plt.title('Claim Frequency by Gender')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.legend(title='Response', labels=['No Claim (0)', 'Claim (1)'])
plt.show()
```



This bar chart shows the frequency of insurance claims based on gender:

Findings : Most individuals, regardless of gender, do not file claims, indicating no notable gender-based variation in claim behavior.

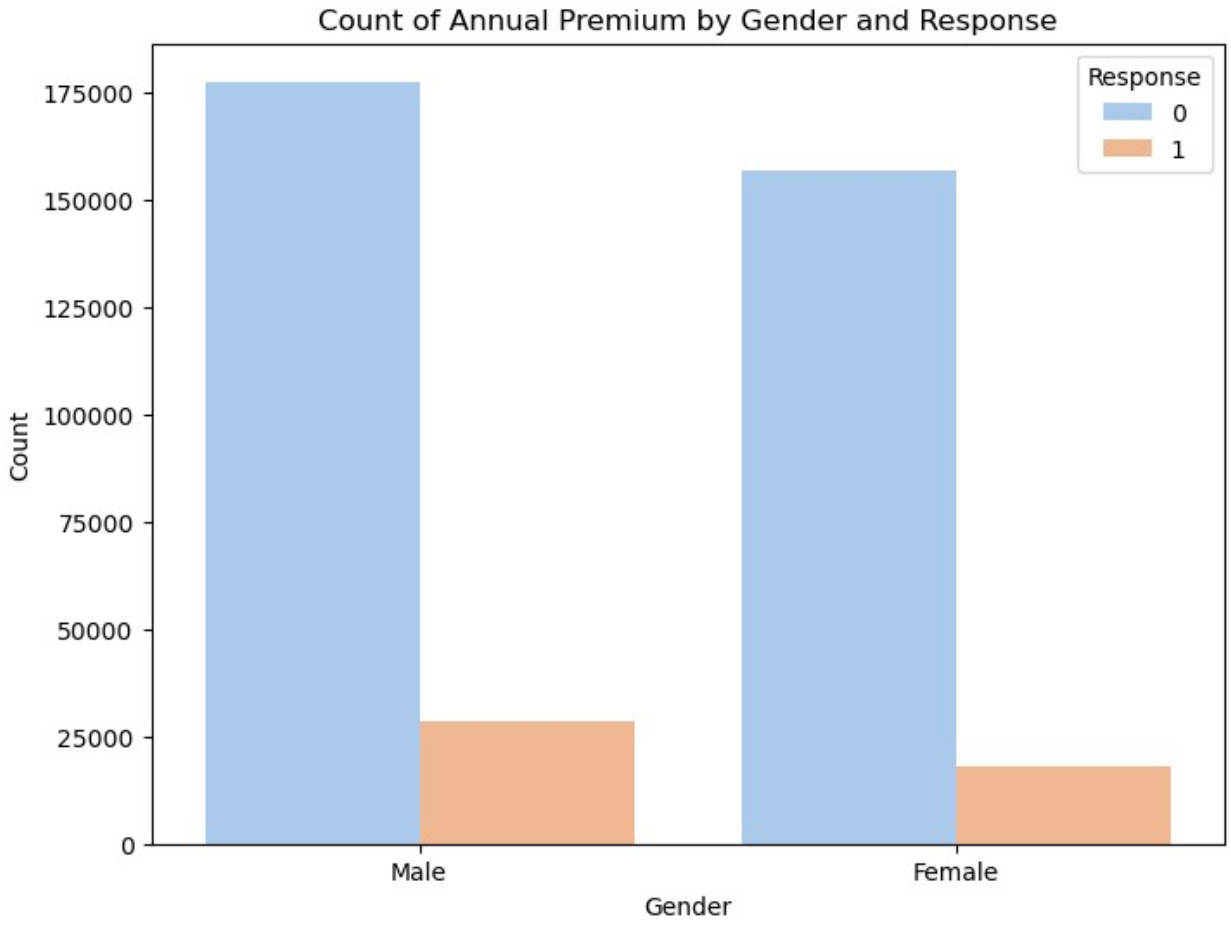
No Claims (0) dominate for both genders, represented by the larger blue bars.

Claims (1) are relatively low for both males and females, with orange bars much shorter than blue bars.

8.2 Annual Premium by Gender:

```
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender', data=df, hue='Response', palette="pastel")
plt.title('Count of Annual Premium by Gender and Response')
plt.xlabel('Gender')
```

```
plt.ylabel('Count')
plt.legend(title='Response')
plt.show()
```



This bar chart shows the frequency of annual premiums based on gender:

Findings : Most individuals, regardless of gender, do not pay annual premium, indicating no notable gender-based variation in annual premium behavior.

No annual premium (0) dominate for both genders, represented by the larger blue bars.

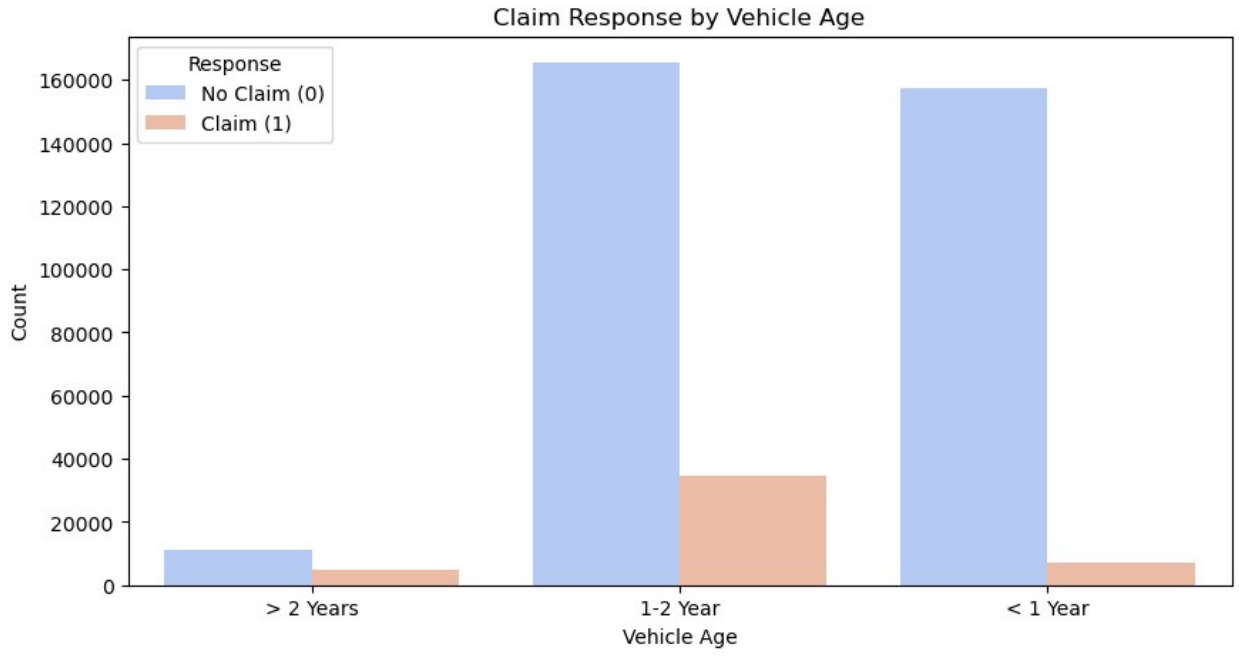
annual premium(1) are relatively low for both males and females, with orange bars much shorter than blue bars.

Step 9: Vehicle Age and Claims

9.1 Claim Response by Vehicle Age

Examine the impact of vehicle age on the likelihood of a claim.

```
plt.figure(figsize=(10, 5))
sns.countplot(x='Vehicle_Age', hue='Response', data=df,
palette="coolwarm")
plt.title('Claim Response by Vehicle Age')
plt.xlabel('Vehicle Age')
plt.ylabel('Count')
plt.legend(title='Response', labels=['No Claim (0)', 'Claim (1)'])
plt.show()
```



Findings from the graph: The graph shows the claim response by vehicle age.

Most Vehicles: The most vehicles are under 1 year old.

Claims: The number of claims are higher in newer vehicles.

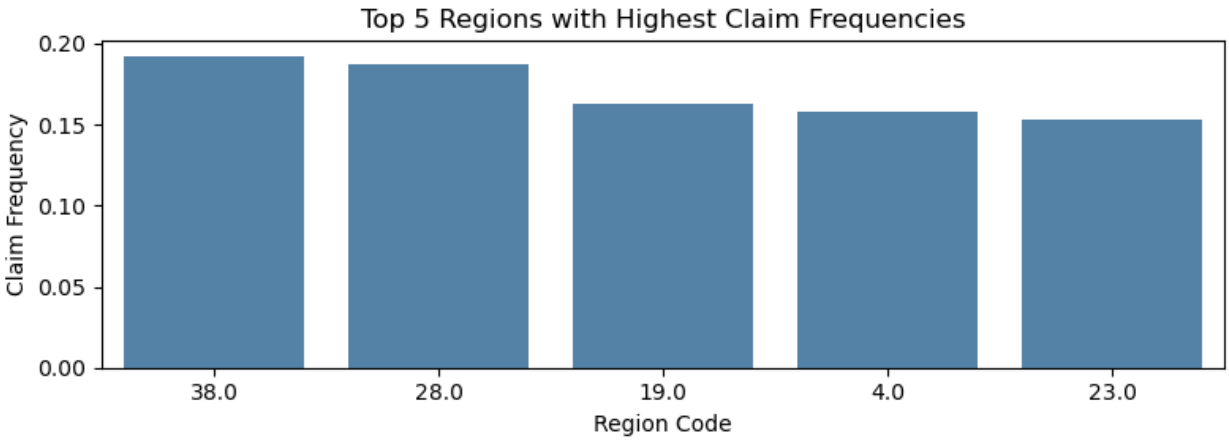
Older Vehicles: There are fewer older vehicles (>2 years old) and also fewer claims.

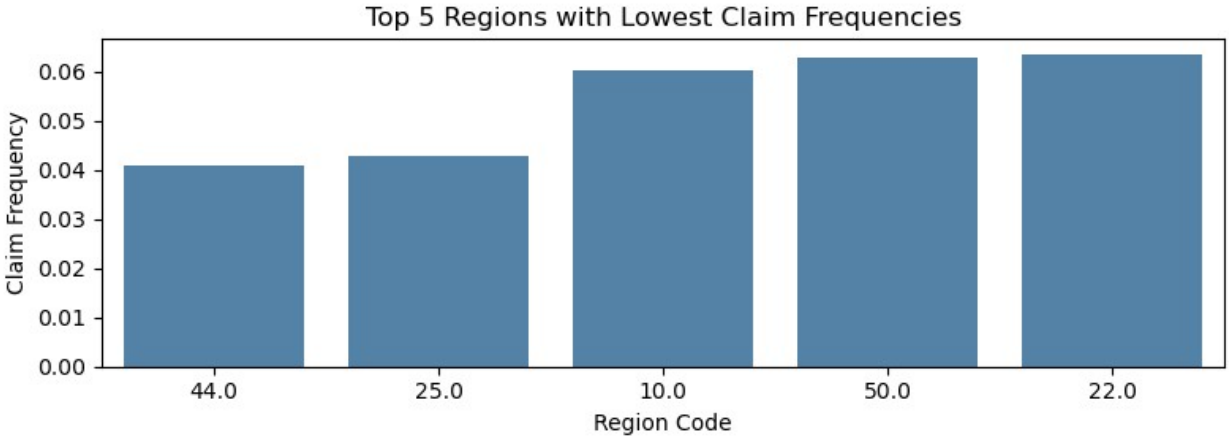
Step 10: Region-wise Analysis

10.1: claim frequency by top 5 and 5 lowest region analysis.

Analyze regional patterns in insurance claims.

```
# Calculate claim frequency by region
region_claim_frequency = df.groupby('Region_Code')['Response'].mean()

# Find the top 5 regions with the highest claim frequencies
top_regions = region_claim_frequency.nlargest(5)

# Find the top 5 regions with the lowest claim frequencies
bottom_regions = region_claim_frequency.nsmallest(5)

# Plot top 5 regions with highest claim frequencies
plt.figure(figsize=(8, 3))
sns.barplot(x=top_regions.index.astype(str), y=top_regions.values,
color='steelblue')
plt.title("Top 5 Regions with Highest Claim Frequencies")
plt.xlabel("Region Code")
plt.ylabel("Claim Frequency")
plt.tight_layout()
plt.show()

# Plot top 5 regions with lowest claim frequencies
plt.figure(figsize=(8, 3))
sns.barplot(x=bottom_regions.index.astype(str),
y=bottom_regions.values, color="steelblue")
plt.title("Top 5 Regions with Lowest Claim Frequencies")
plt.xlabel("Region Code")
plt.ylabel("Claim Frequency")
plt.tight_layout()
plt.show()
```


Top 5 Regions with Highest Claim Frequencies
Claim Frequency
0.20
0.15
0.10
0.05
0.00
38.0
28.0
19.0
4.0
23.0
Region Code


Top 5 Regions with Lowest Claim Frequencies
Claim Frequency
0.06
0.05
0.04
0.03
0.02
0.01
0.00
44.0
25.0
10.0
50.0
22.0
Region Code

Findings from the Bar Chart

(Top 5 Regions with Highest Claim Frequencies)

1.The chart shows the top 5 regions with the highest claim frequencies.

2.Region 38.0 has the highest claim frequency, followed by Region 28.0.

3.The claim frequency for Region 38.0 and Region 28.0 are almost identical, slightly above 0.18.

4.Regions 19.0, 4.0, and 23.0 have claim frequencies closer to 0.16.

5.The claim frequencies for all regions are above 0.15, indicating a higher than average likelihood of claims in these regions.

(Top 5 Regions with Lowest Claim Frequencies)

1. The chart displays the top 5 regions with the lowest claim frequencies.

2. Region Code 44.0 has the lowest claim frequency among the five regions.

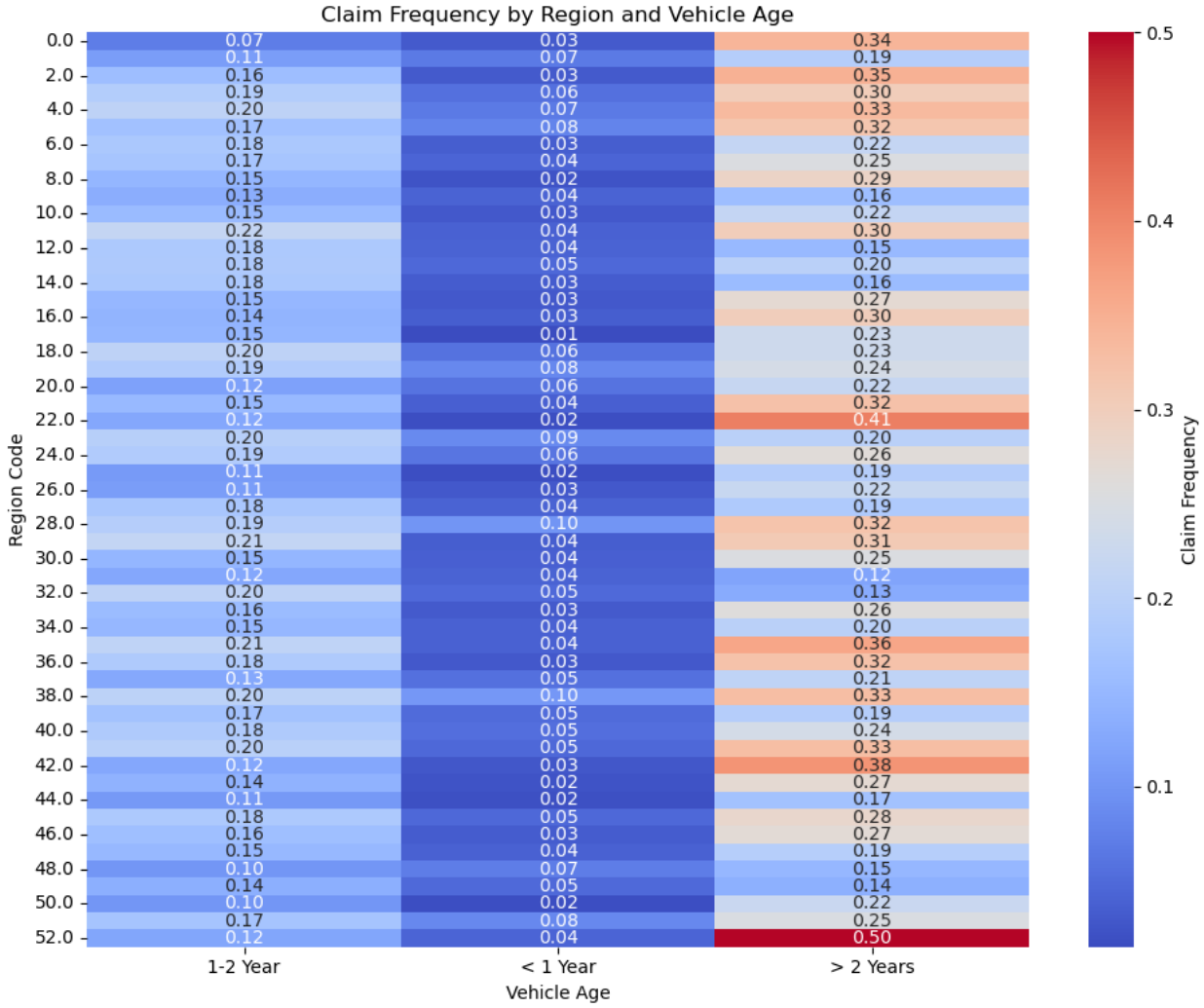
3. Region Code 50.0 has the highest claim frequency within this low-claim group.

4. The claim frequencies gradually increase from Region 44.0 to Region 50.0.

5. The data indicates significant variability even among regions with relatively low claim frequencies.

10.2 Claim Frequency by Region and Vehicle Age

```
# Calculate claim frequency by Region and Vehicle Age
region_vehicle_age = df.groupby(['Region_Code', 'Vehicle_Age'])
['Response'].mean().unstack()

# Visualize the relationship
plt.figure(figsize=(10, 8))
sns.heatmap(region_vehicle_age, annot=True, fmt=".2f",
cmap="coolwarm", cbar_kws={'label': 'Claim Frequency'})
plt.title("Claim Frequency by Region and Vehicle Age")
plt.xlabel("Vehicle Age")
plt.ylabel("Region Code")
plt.tight_layout()
plt.show()
```


Claim Frequency by Region and Vehicle Age
Region Code
Vehicle Age
1-2 Year
< 1 Year
> 2 Years
Claim Frequency
0.5
0.4
0.3
0.2
0.1
0.0 0.07 0.03 0.34
0.11 0.07 0.19
2.0 0.16 0.03 0.35
0.19 0.06 0.30
4.0 0.20 0.07 0.33
0.17 0.08 0.32
6.0 0.18 0.03 0.22
0.17 0.04 0.25
8.0 0.15 0.02 0.29
0.13 0.04 0.16
10.0 0.15 0.03 0.22
0.22 0.04 0.30
12.0 0.18 0.04 0.15
0.18 0.05 0.20
14.0 0.18 0.03 0.16
0.15 0.03 0.27
16.0 0.14 0.03 0.30
0.15 0.01 0.23
18.0 0.20 0.06 0.23
0.19 0.08 0.24
20.0 0.12 0.06 0.22
0.15 0.04 0.32
22.0 0.12 0.02 0.41
0.20 0.09 0.20
24.0 0.19 0.06 0.26
0.11 0.02 0.19
26.0 0.11 0.03 0.22
0.18 0.04 0.19
28.0 0.19 0.10 0.32
0.21 0.04 0.31
30.0 0.15 0.04 0.25
0.12 0.04 0.12
32.0 0.20 0.05 0.13
0.16 0.03 0.26
34.0 0.15 0.04 0.20
0.21 0.04 0.36
36.0 0.18 0.03 0.32
0.13 0.05 0.21
38.0 0.20 0.10 0.33
0.17 0.05 0.19
40.0 0.18 0.05 0.24
0.20 0.05 0.33
42.0 0.12 0.03 0.38
0.14 0.02 0.27
44.0 0.11 0.02 0.17
0.18 0.05 0.28
46.0 0.16 0.03 0.27
0.15 0.04 0.19
48.0 0.10 0.07 0.15
0.14 0.05 0.14
50.0 0.10 0.02 0.22
0.17 0.08 0.25
52.0 0.12 0.04 0.50

The heatmap provides insights into claim frequency by region and vehicle age. Here are the key findings:

1) Claim frequency is highest for vehicles older than 2 years across several regions, with the darkest red areas indicating claim frequencies up to 0.50.

2) Regions like 20.0 and 22.0 show particularly high claim frequencies (e.g., 0.41) for vehicles older than 2 years.

3) For vehicles less than 1 year old, claim frequencies are consistently low across all regions, typically below 0.10.

4) Vehicles aged 1–2 years have moderate claim frequencies, falling between the frequencies for newer and older vehicles, with some regional variations.

4) The heatmap highlights that age of the vehicle significantly impacts claim frequency, with older vehicles (>2 years) being more prone to claims.

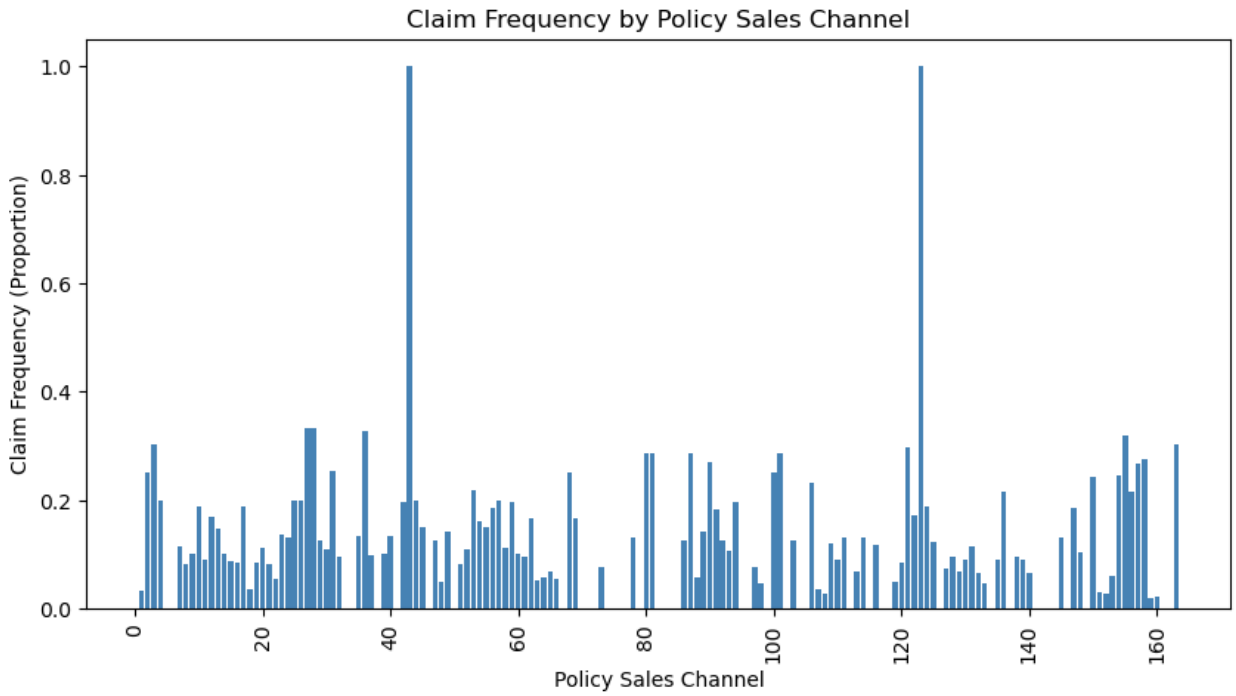
Regions and vehicle age clearly interact, with older vehicles contributing most to claim risks, especially in specific regions like 20.0 and 22.0.

Step 11: Policy Analysis

11.1 Claim Frequency by Policy Sales Channel:

Explore the distribution and impact of different insurance policy types.

```
sales_channel_frequency = df.groupby('Policy_Sales_Channel')
['Response'].mean()
plt.figure(figsize=(10, 5))
plt.bar(sales_channel_frequency.index, sales_channel_frequency.values,
color="steelblue")
plt.title('Claim Frequency by Policy Sales Channel')
plt.xlabel('Policy Sales Channel')
plt.ylabel('Claim Frequency (Proportion)')
plt.xticks(rotation=90)
plt.show()
```



The bar chart shows claim frequency by policy sales channel. Key findings include:

1) Most policy sales channels have relatively low claim frequencies, generally below 0.2.

2) A few specific sales channels (notably around 40 and 120) exhibit extremely high claim frequencies, reaching 1.0, indicating that all policies sold through these channels resulted in claims.

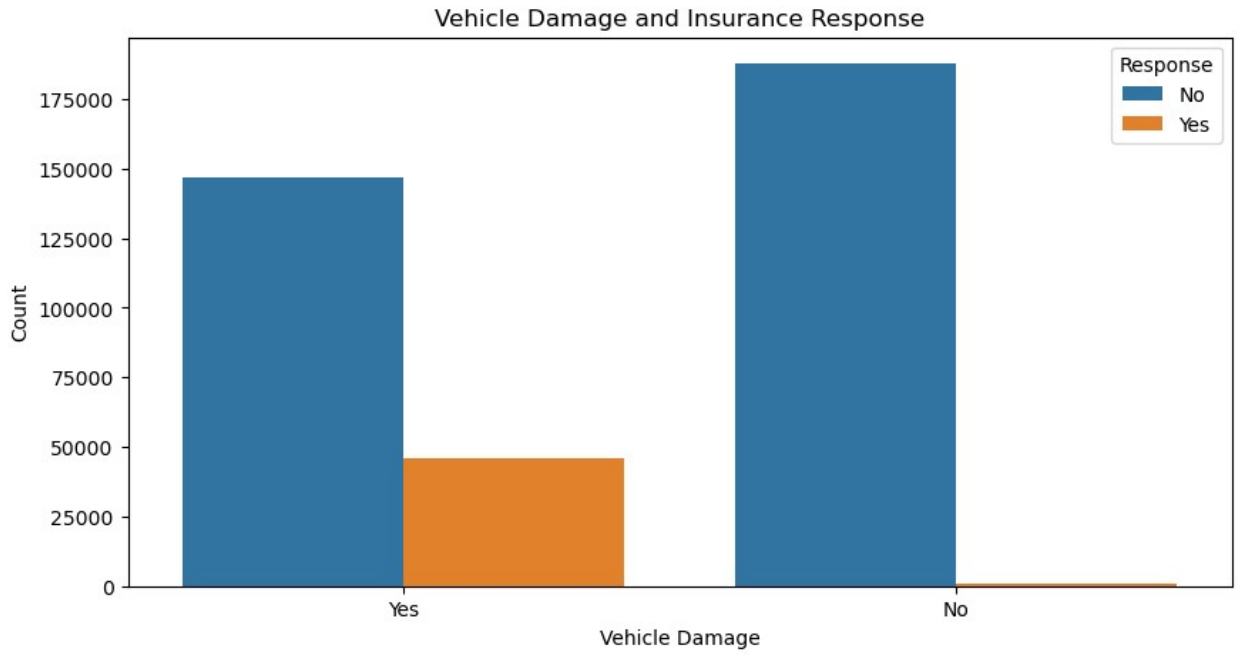
3) There is significant variability in claim frequencies across the sales channels, with some clusters showing moderate frequencies and others near-zero.

4) This suggests that certain sales channels are highly associated with claims, likely requiring further investigation into their policyholder demographics or practices.

Step 12: Claim Frequency by Vehicle Damage

Investigate the relationship between vehicle damage and claim frequencies.

```
plt.figure(figsize=(10, 5))
sns.countplot(data=df, x='Vehicle_Damage', hue='Response')
plt.title('Vehicle Damage and Insurance Response')
plt.xlabel('Vehicle Damage')
plt.ylabel('Count')
plt.legend(title='Response', labels=['No', 'Yes'])
plt.show()
```



Findings from the chart:

There are significantly more instances of no vehicle damage than vehicle damage. The majority of cases with vehicle damage resulted in an insurance claim being made.

Specifics:

In cases of no vehicle damage, the vast majority did not file an insurance claim.

In cases of vehicle damage, a significant proportion filed an insurance claim.

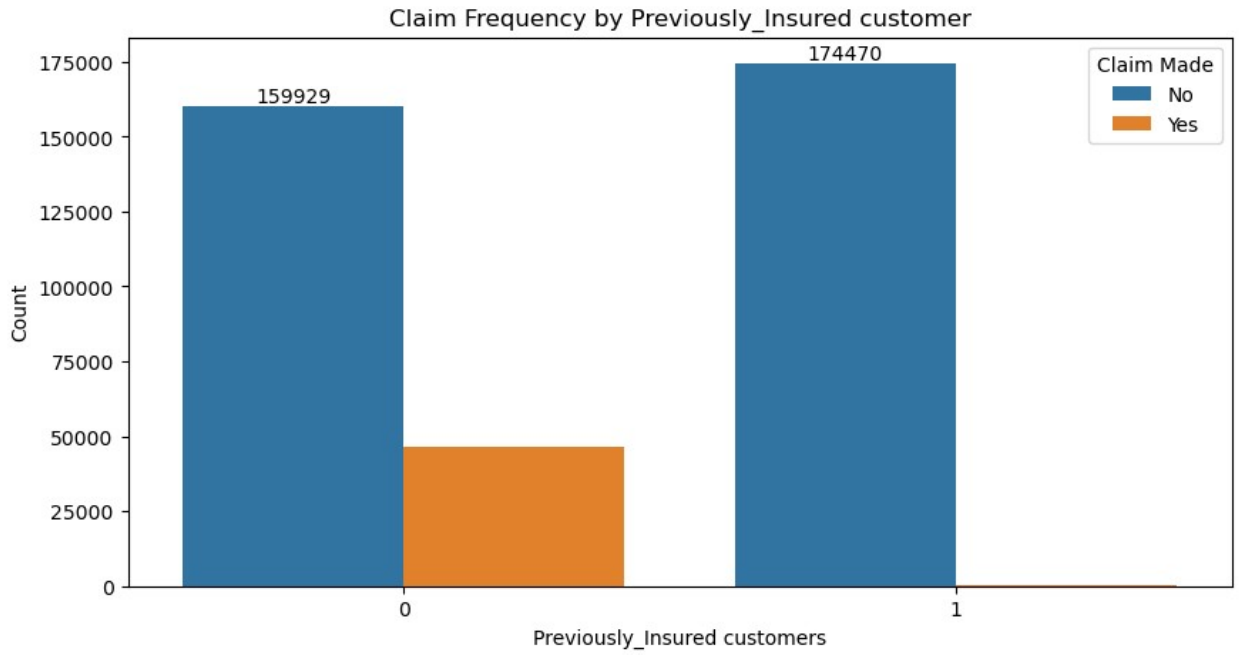
The chart does not show the total number of insurance claims made.

The chart does not show the total number of vehicles involved in accidents.

Step 13: Customer Loyalty Analysis

Analyze if the number of policies held by a customer influences claim likelihood.

```
plt.figure(figsize=(10, 5))
ax=sns.countplot(x='Previously_Insured', hue='Response', data=df)
plt.bar_label(ax.containers[0])
plt.title('Claim Frequency by Previously_Insured customer')
plt.xlabel('Previously_Insured customers')
plt.ylabel('Count')
plt.legend(title='Claim Made', labels=['No', 'Yes'])
plt.show()
```



Findings : Distribution of Claims by Previously Insured

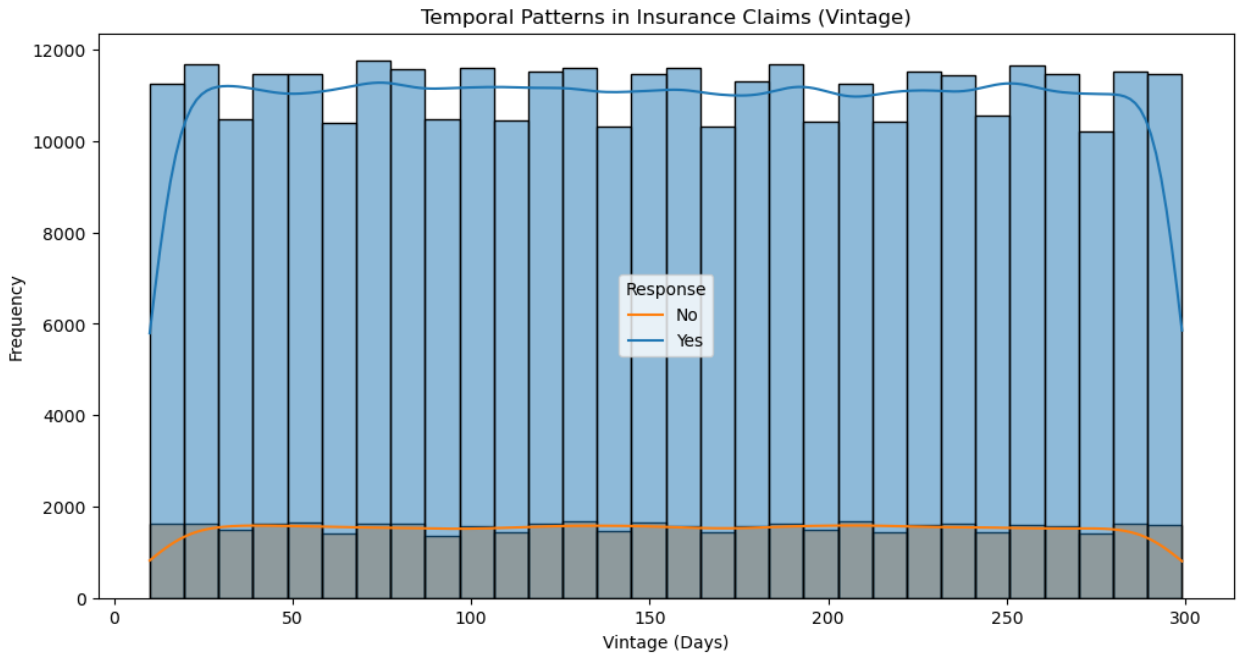
People previously insured (1) are more likely to file a claim than people who were not previously insured (0).

The bar chart shows that the number of people who were not previously insured (0) and did not file a claim (0) is much larger than all other combinations.

The distribution of claims is heavily skewed towards people who were not previously insured (0) and did not file a claim (0).

Step 14: Time Analysis

```
plt.figure(figsize=(12, 6))
sns.histplot(data=df, x='Vintage', hue='Response', kde=True, bins=30)
plt.title('Temporal Patterns in Insurance Claims (Vintage)')
plt.xlabel('Vintage (Days)')
plt.ylabel('Frequency')
plt.legend(title='Response', labels=['No', 'Yes'])
plt.show()
```



Findings:

1) Stable claim patterns over time: The frequency of claims remains relatively consistent across the "vintage" period (0–300 days), with only minor fluctuations.

2) Low response for claims: Very few claims are marked "Yes," while the majority are "No," indicating a low claim response rate throughout the period.

Conclusion

Through the EDA of the vehicle insurance dataset, we uncovered actionable insights into policyholder demographics, driving behavior, and insurance claim trends. These insights can be utilized to design data-driven strategies, enhance customer segmentation, and mitigate risks effectively. This analysis provides a foundation for further predictive modeling and decision-making processes within the organization.