
# Vehicle Insurance Dataset - Exploratory Data Analysis

## Objective :

Perform an in-depth Exploratory Data Analysis (EDA) to derive insights about insurance claims.

## Dataset Overview :

The dataset includes information about insured individuals, their vehicles, and claims.

# Step 1 : Import Necessary Libraries and Load the Dataset :->

In [1]:
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("C:/Users/Administrator/Downloads/Vehicle_Insurance.csv")

# Configure plots
```

```
%matplotlib inline
sns.set(style="whitegrid")
```

# Step 2 : Inspect Data :->

## 2.1) Inspecting the First and Last ten rows :

Using df.head() and df.tail() to have a preview of data. To Understand its Structure

In [3]:
```
df.head(10)
```

Out[3]:

| | id | Gender | Age | Driving_License | Region_Code | Previously_Insured | Vehicle_Age | Vehicle_Damage | Annual_Premium | Policy_Sal |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | Male | 44 | 1 | 28.0 | 0 | > 2 Years | Yes | 40454.0 | |
| **1** | 2 | Male | 76 | 1 | 3.0 | 0 | 1-2 Year | No | 33536.0 | |
| **2** | 3 | Male | 47 | 1 | 28.0 | 0 | > 2 Years | Yes | 38294.0 | |
| **3** | 4 | Male | 21 | 1 | 11.0 | 1 | < 1 Year | No | 28619.0 | |
| **4** | 5 | Female | 29 | 1 | 41.0 | 1 | < 1 Year | No | 27496.0 | |
| **5** | 6 | Female | 24 | 1 | 33.0 | 0 | < 1 Year | Yes | 2630.0 | |
| **6** | 7 | Male | 23 | 1 | 11.0 | 0 | < 1 Year | Yes | 23367.0 | |
| **7** | 8 | Female | 56 | 1 | 28.0 | 0 | 1-2 Year | Yes | 32031.0 | |
| **8** | 9 | Female | 24 | 1 | 3.0 | 1 | < 1 Year | No | 27619.0 | |
| **9** | 10 | Female | 32 | 1 | 6.0 | 1 | < 1 Year | No | 28771.0 | |

In [5]:
```
df.tail(10)
```

Out[5]:

| | id | Gender | Age | Driving_License | Region_Code | Previously_Insured | Vehicle_Age | Vehicle_Damage | Annual_Premium |
|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **381099** | 381100 | Female | 51 | 1 | 28.0 | 0 | 1-2 Year | Yes | 44504.0 |
| **381100** | 381101 | Female | 29 | 1 | 28.0 | 0 | < 1 Year | Yes | 49007.0 |
| **381101** | 381102 | Female | 70 | 1 | 28.0 | 0 | > 2 Years | Yes | 50904.0 |
| **381102** | 381103 | Female | 25 | 1 | 41.0 | 1 | < 1 Year | Yes | 2630.0 |
| **381103** | 381104 | Male | 47 | 1 | 50.0 | 0 | 1-2 Year | Yes | 39831.0 |
| **381104** | 381105 | Male | 74 | 1 | 26.0 | 1 | 1-2 Year | No | 30170.0 |
| **381105** | 381106 | Male | 30 | 1 | 37.0 | 1 | < 1 Year | No | 40016.0 |
| **381106** | 381107 | Male | 21 | 1 | 30.0 | 1 | < 1 Year | No | 35118.0 |
| **381107** | 381108 | Female | 68 | 1 | 14.0 | 0 | > 2 Years | Yes | 44617.0 |
| **381108** | 381109 | Male | 46 | 1 | 29.0 | 0 | 1-2 Year | No | 41777.0 |

## 2.2) Understanding Data types :

In [7]:
```
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
```

## 2.3) Summarizing the Data :

In [9]:
```
df.describe()
```

Out[9]:

| | id | Age | Driving_License | Region_Code | Previously_Insured | Annual_Premium | Policy_Sales_Channel |
|---|---|---|---|---|---|---|---|
| **count** | 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 |
| **std** | 110016.836208 | 15.511611 | 0.046110 | 13.229888 | 0.498251 | 17213.155057 | 54.203995 |
| **min** | 1.000000 | 20.000000 | 0.000000 | 0.000000 | 0.000000 | 2630.000000 | 1.000000 |
| **25%** | 95278.000000 | 25.000000 | 1.000000 | 15.000000 | 0.000000 | 24405.000000 | 29.000000 |
| **50%** | 190555.000000 | 36.000000 | 1.000000 | 28.000000 | 0.000000 | 31669.000000 | 133.000000 |
| **75%** | 285832.000000 | 49.000000 | 1.000000 | 35.000000 | 1.000000 | 39400.000000 | 152.000000 |
| **max** | 381109.000000 | 85.000000 | 1.000000 | 52.000000 | 1.000000 | 540165.000000 | 163.000000 |

## 2.4) Shape and size of the dataset :

In [11]:
```
df.shape
```

Out[11]: (381109, 12)

In [13]:
```
df.size
```

Out[13]: 4573308

# Step 3 : Data Cleaning :->

## 3.1) Handling Missing Values :

In [15]:
```
df.isnull().sum()
```

Out[15]:
```
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
```

**In this case we don't have any missing value so we can procede to next step , otherwise we would have to treat those values first.**

## 3.2) Checking uniqe values :

In [17]:
```
for column in df.columns:
    unique_count = df[column].nunique()
    print(f"'{column}' has {unique_count} unique values")
```

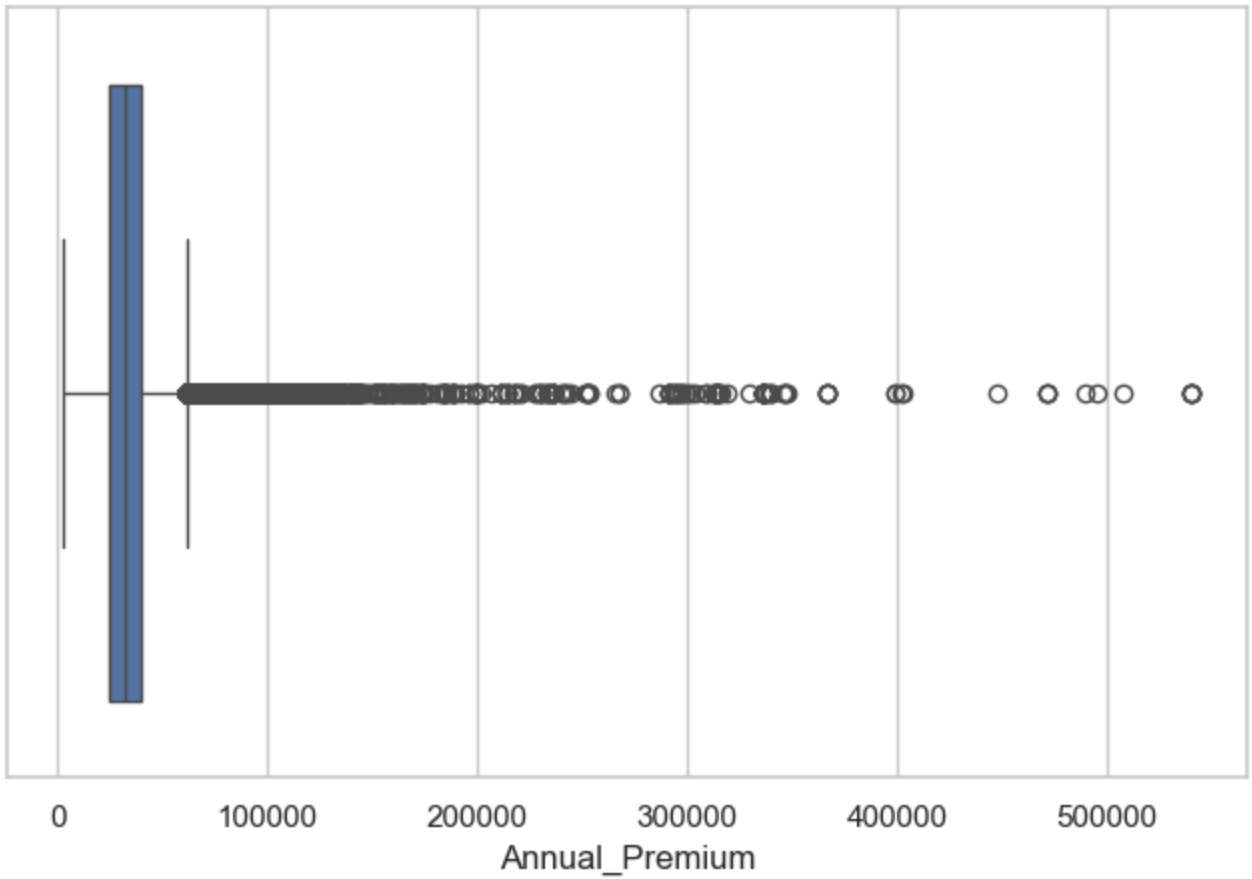
```
'id' has 381109 unique values
'Gender' has 2 unique values
'Age' has 66 unique values
'Driving_License' has 2 unique values
'Region_Code' has 53 unique values
'Previously_Insured' has 2 unique values
'Vehicle_Age' has 3 unique values
'Vehicle_Damage' has 2 unique values
'Annual_Premium' has 48838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

```
'Vintage' has 290 unique values
'Response' has 2 unique values
```

## 3.3) Handling Outliers :

**Identify outliers in numerical columns like Age and Annual_Premium using box plots.**

In [19]:
```
# Visualize outliers in data
plt.figure(figsize=(8, 5))
sns.boxplot(x=df['Annual_Premium'])
plt.show()
```




Key Takeaways of Outliers Using Boxplot :

- **Significant outliers present**, with values exceeding 500,000.
- **Distribution is heavily right-skewed.**
- **Majority of `Annual_Premium` values** are clustered below 100,000.
- **Recommendation:** Investigate outliers and apply data transformation to reduce skewness.

## 3.4) Checking outliers in a column using the IQR method with count :

In [21]:
```
def count_outliers(series):
    Q1 = series.quantile(0.25)
    Q3 = series.quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    outliers = series[(series < lower_bound) | (series > upper_bound)]
    return outliers.count()

numerical_columns = ['Age', 'Region_Code', 'Annual_Premium', 'Policy_Sales_Channel', 'Vintage']
outlier_counts = {col: count_outliers(df[col]) for col in numerical_columns}

outlier_counts
```

Out[21]:
```
{'Age': 0,
 'Region_Code': 0,
 'Annual_Premium': 10320,
 'Policy_Sales_Channel': 0,
 'Vintage': 0}
```

## 3.5) Filling outliers with mean :

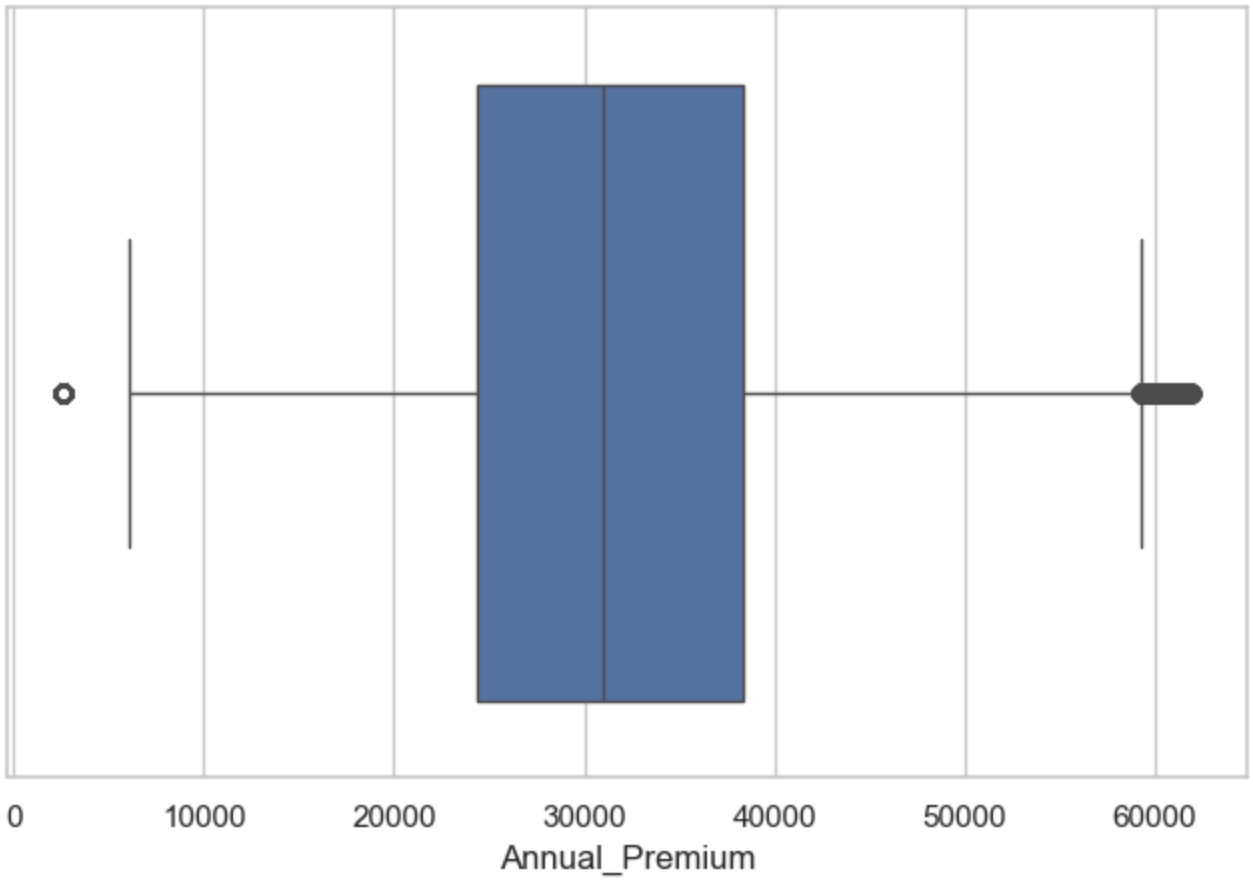
In [23]:
```
def replace_outliers_with_mean(df, column_name):

  # Calculate the Interquartile Range (IQR)
  Q1 = df[column_name].quantile(0.25)
  Q3 = df[column_name].quantile(0.75)
  IQR = Q3 - Q1

  # Define upper and lower bounds for outliers
  lower_bound = Q1 - 1.5 * IQR
```

```
    upper_bound = Q3 + 1.5 * IQR

    # Replace outliers with the mean
    df.loc[(df[column_name] < lower_bound) | (df[column_name] > upper_bound), column_name] = df[column_name].

    return df
```

In [25]:
```
df1=replace_outliers_with_mean(df,'Annual_Premium')
```

## 3.6) Checking if outliers are treated or not :

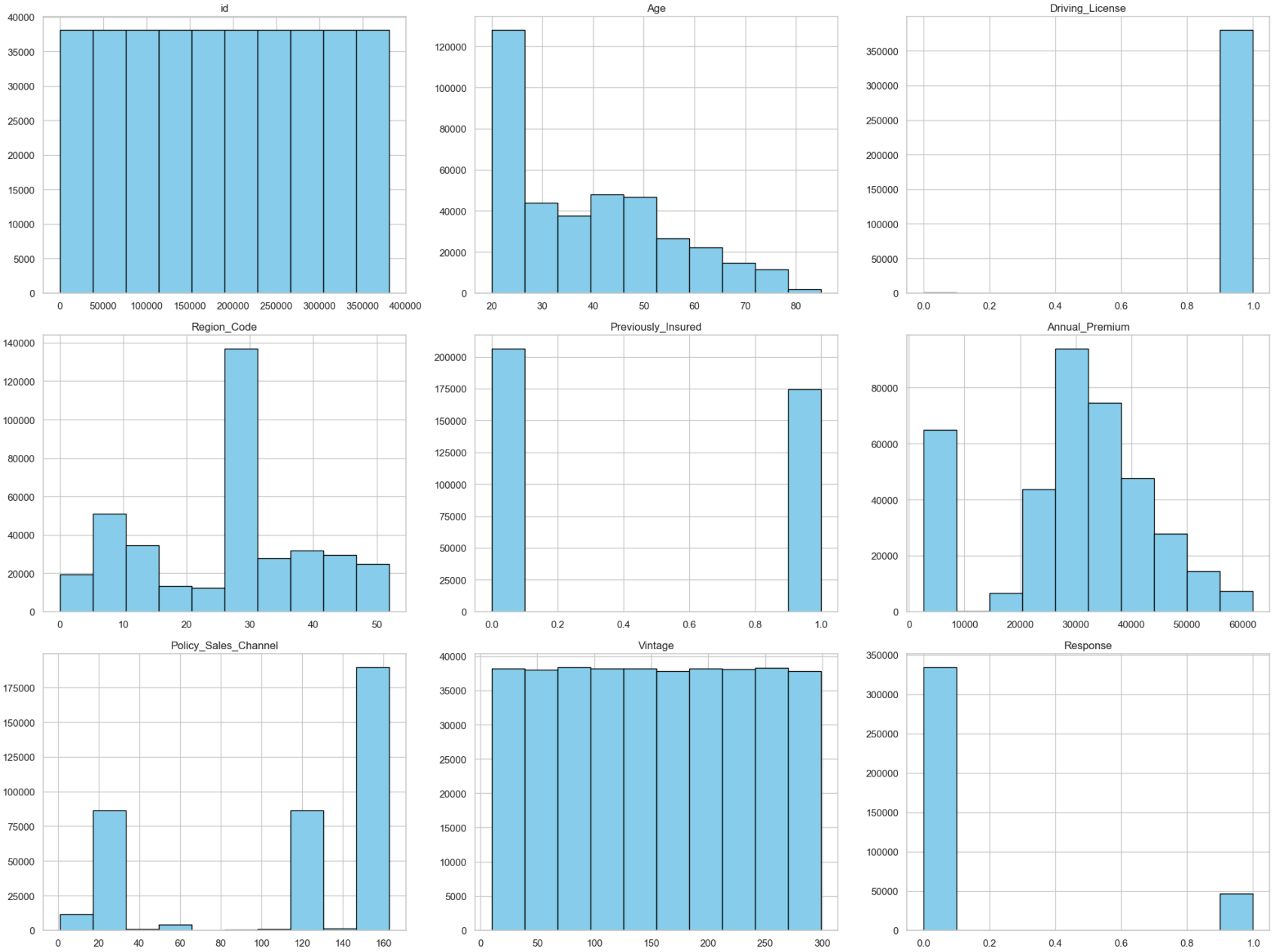
In [27]:
```
plt.figure(figsize=(8, 5))
sns.boxplot(x=df1['Annual_Premium'])
plt.show()
```




# Step 4: Data Vizualisation :->

## 4.1) Ploting graphs (histograms) :

In [29]:
```
df1.hist(figsize=(20,15),color='skyblue',edgecolor='black')
plt.gcf().set_facecolor('white')
plt.tight_layout()
plt.show()
```




Key Takeaways :

- **Age:** Most of the policyholders are younger adults, especially in their 20s and 30s, with fewer older customers.
- **Region:** There's one region where a lot more people have policies, indicating a significant presence or popularity there.
- **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.
- **Sales Channels:** One sales channel dominates, likely serving as the primary method for reaching new customers.
- **Annual Premiums:** Most customers are paying mid-range premiums, suggesting the company targets everyday, middle-income drivers.
- **Previous Insurance:** Many customers didn't have prior insurance, reflecting the company's success in reaching first-time insurance buyers.

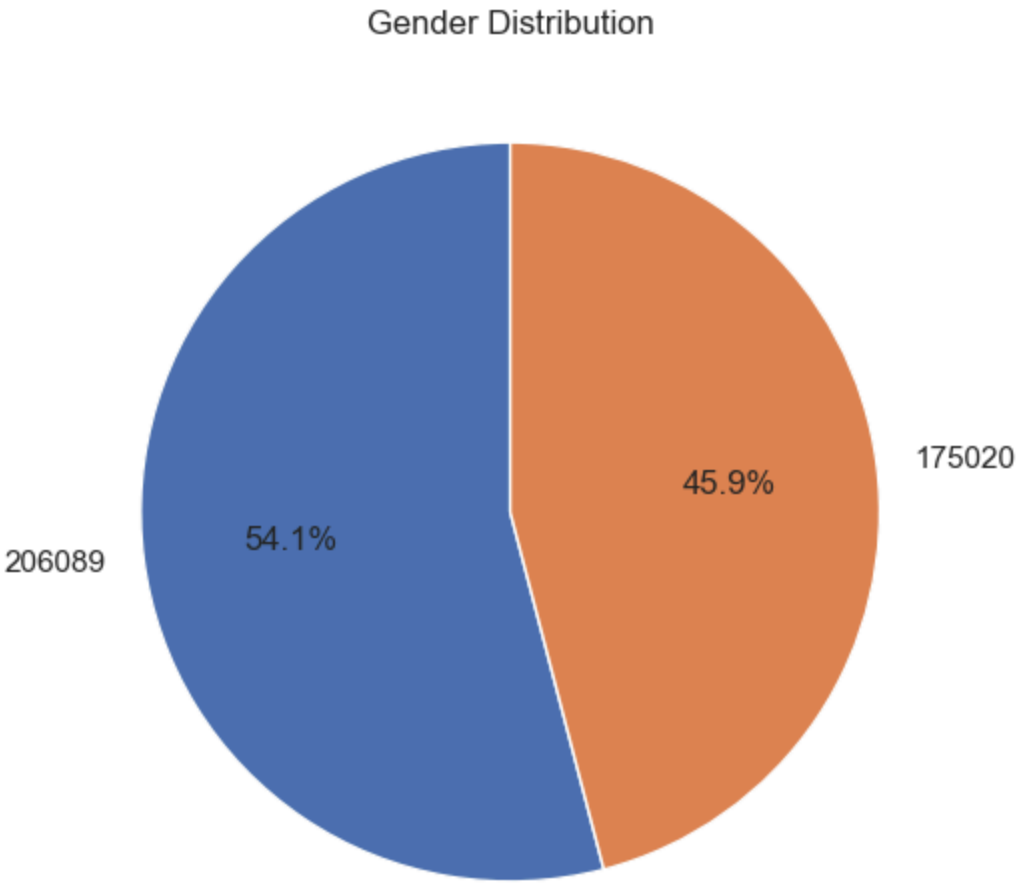
## 4.2) Gender Distribution :

Visualizing the gender distribution to see if there is any imbalance.

In [31]:
```
df.Gender.value_counts()
```

Out[31]:
```
Gender
Male      206089
Female    175020
Name: count, dtype: int64
```

In [33]:
```
gender_counts=df['Gender'].value_counts()
plt.figure(figsize=(6,6))
plt.pie(gender_counts,labels=gender_counts,autopct='%1.1f%%',startangle=90)
plt.title('Gender Distribution')
plt.show()
```




## Gender Distribution

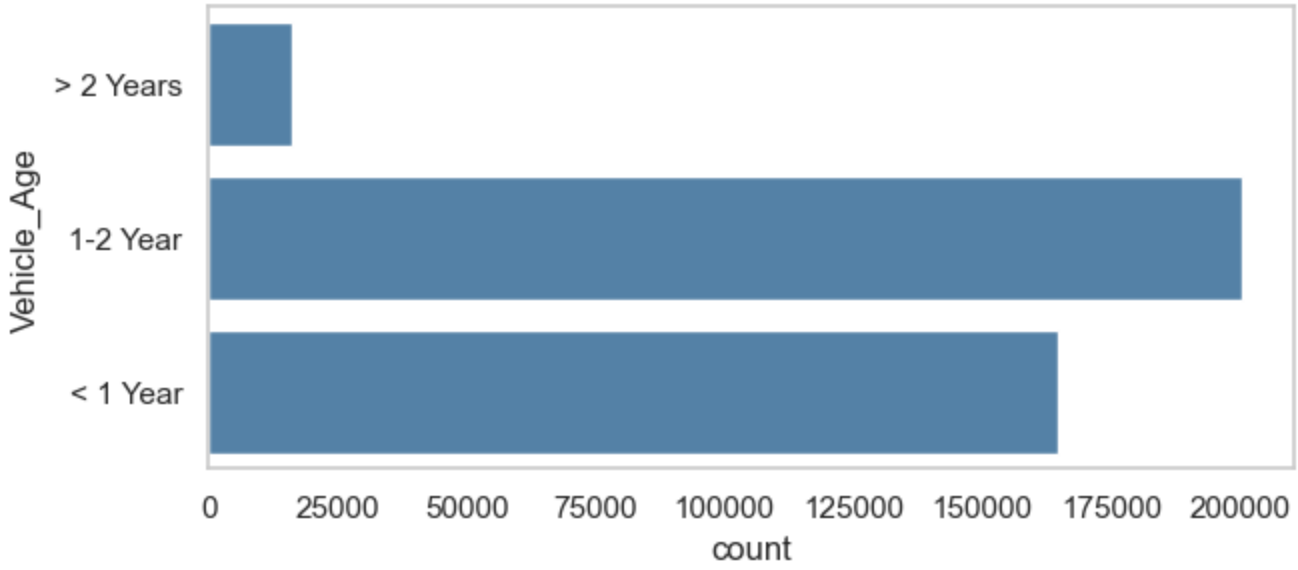
- **Male:** The count is slightly over 200,000.
- **Female:** The count is somewhat around 175,000.

**Findings:**

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

## 4.3) Vehicle Age analysis :

In [35]:
```
plt.figure(figsize=(7, 3))
sns.countplot(data=df1, y='Vehicle_Age', color='steelblue')
plt.grid(False)
plt.show()
```




## Count Distribution Across Time Categories

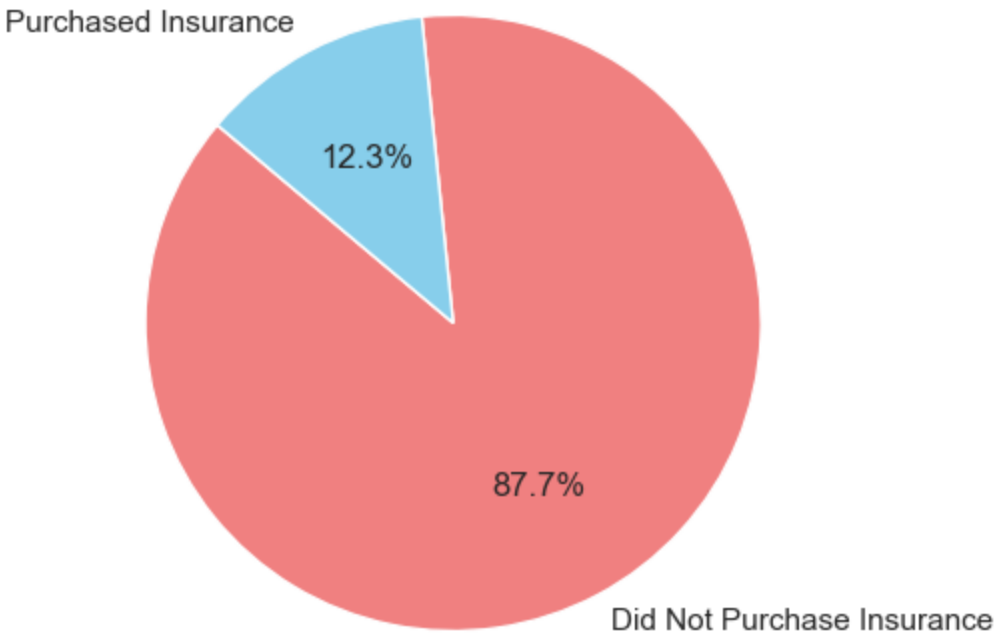
- **Categories:**
  - **1-2 Year:**
    - Has the highest count, exceeding **200,000**.
  - **< 1 Year:**
    - Lower than the 1-2 Year category but still substantial at approximately **175,000**.
  - **> 2 Years:**
    - Significantly lower count, well below **50,000**.

- **Findings:**

- **Most Common Time Category:**
  - The **1-2 Year** category is the most frequent, indicating a large proportion of the dataset belongs here.
- **Least Common Time Category:**
  - The **> 2 Years** category is the least common, representing a smaller subset.
- **Distribution Pattern:**
  - There is a sharp decline in counts from the **1-2 Year** to the **> 2 Years** category.

## Customer Sentiment on Purchasing Vehicle Insurance :

In [37]:
```
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=['lightcoral', 'skyblue'])
plt.show()
```



- **Insights :**
  - **87.7%** of customers did not purchase vehicle insurance.
  - Only **12.3%** of customers opted to purchase vehicle insurance.
  - The overwhelming majority chose not to buy insurance, highlighting a substantial disparity.
  - The chart emphasizes a **low conversion rate** for vehicle insurance among the surveyed group.

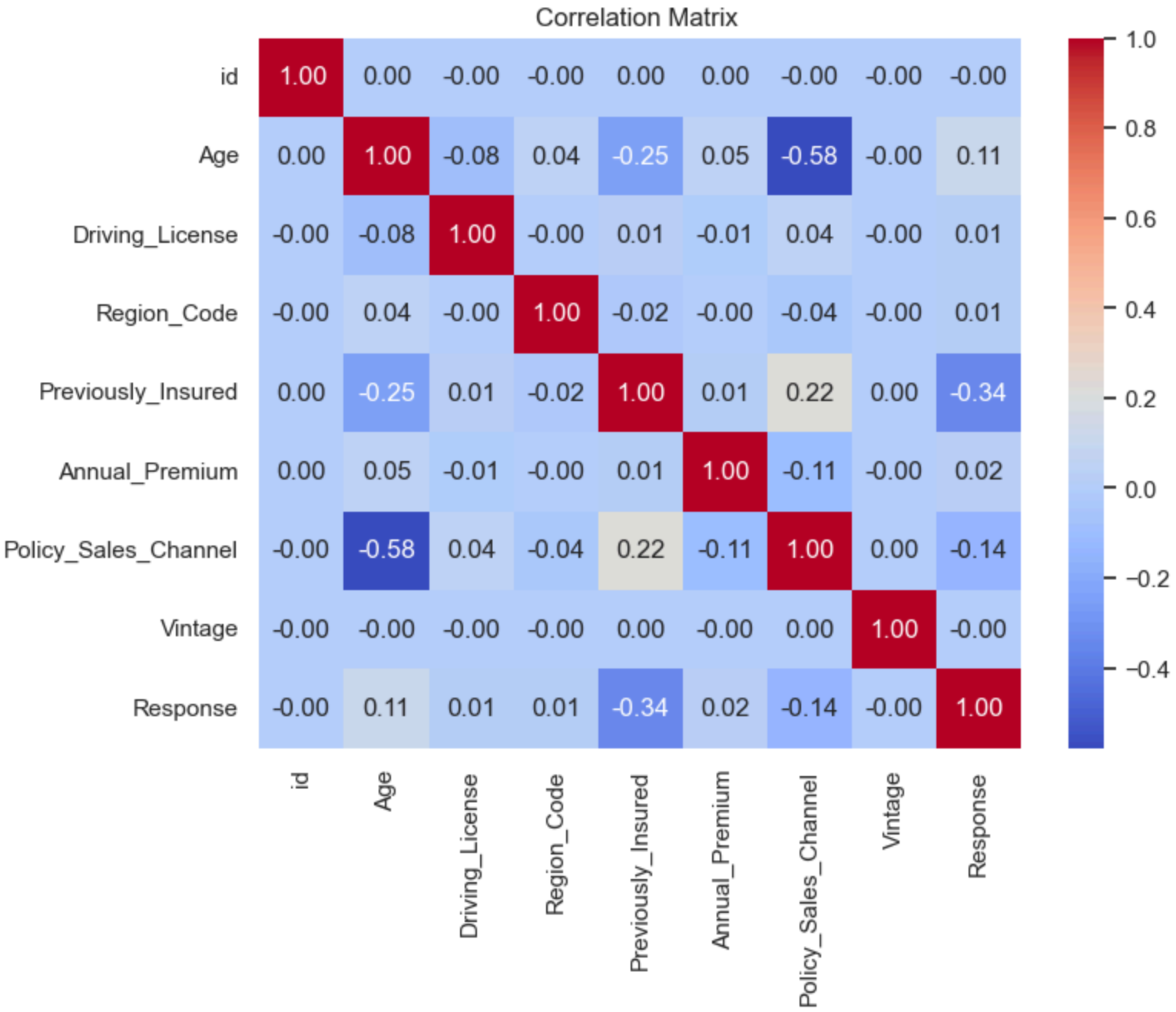
# Step 5 : Feature Analysis :->

## 5.1) Target Variable (Response)

Visualizing the distribution of the Response Variables, Which indicates if a customer made an insurance claim.

In [39]:
```
numeric_df = df1.select_dtypes(include=['float64', 'int64'])
correlation_matrix = numeric_df.corr()
```

In [41]:
```
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Correlation Matrix ')
plt.show()
```




Key Findings :

- **Key Negative Relationships:**
  - **Sales channels are less used by older individuals**: There is a strong negative correlation between age and policy sales channel usage (**-0.58**).
  - **Already insured individuals are less likely to respond**: Those who were previously insured tend to show lower response rates to offers (**-0.34**).

- **Positive but Weak Connections:**
  - **Sales channel usage increases slightly with prior insurance**: There is a weak positive correlation (**0.22**) between prior insurance and sales channel usage.
  - **Older individuals are slightly more likely to respond**: A weak positive relationship (**0.11**) exists between age and response rates.

- **Most Features Don't Strongly Relate to Each Other or the Target (Response):**
  - Variables such as **Region_Code**, **Driving_License**, and **Vintage** show minimal impact or strong correlation with the response rate, indicating that most features don't strongly relate to the target variable.

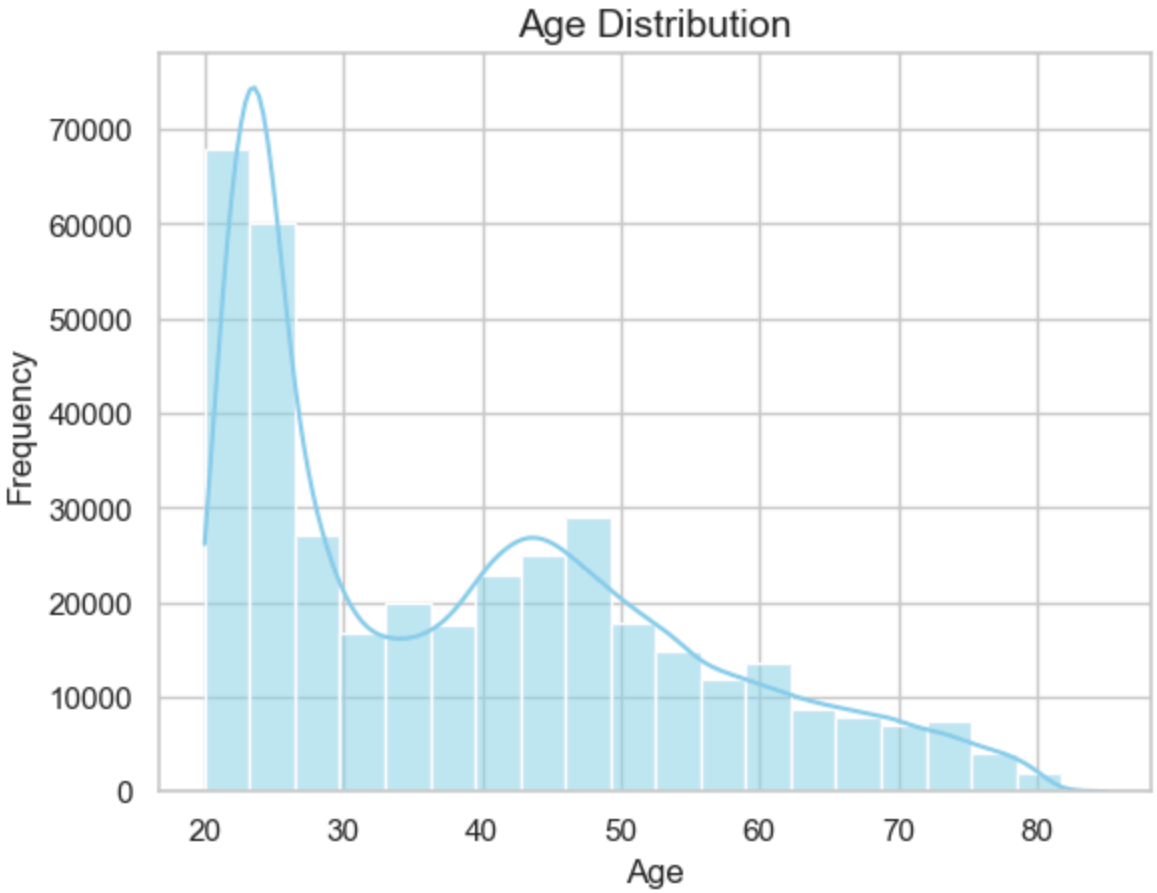
# Step 6 : Age Distribution :->

## Objective:

Analyze the age distribution of insured individuals and examine its impact on the likelihood of claims.

## 6.1) Age Distribution by Customers :

In [43]:
```
sns.histplot(df['Age'],bins=20, kde=True, color='skyblue')
plt.title('Age Distribution', fontsize=14)
plt.xlabel('Age', fontsize=12)
plt.ylabel('Frequency', fontsize=12)
plt.show()
```

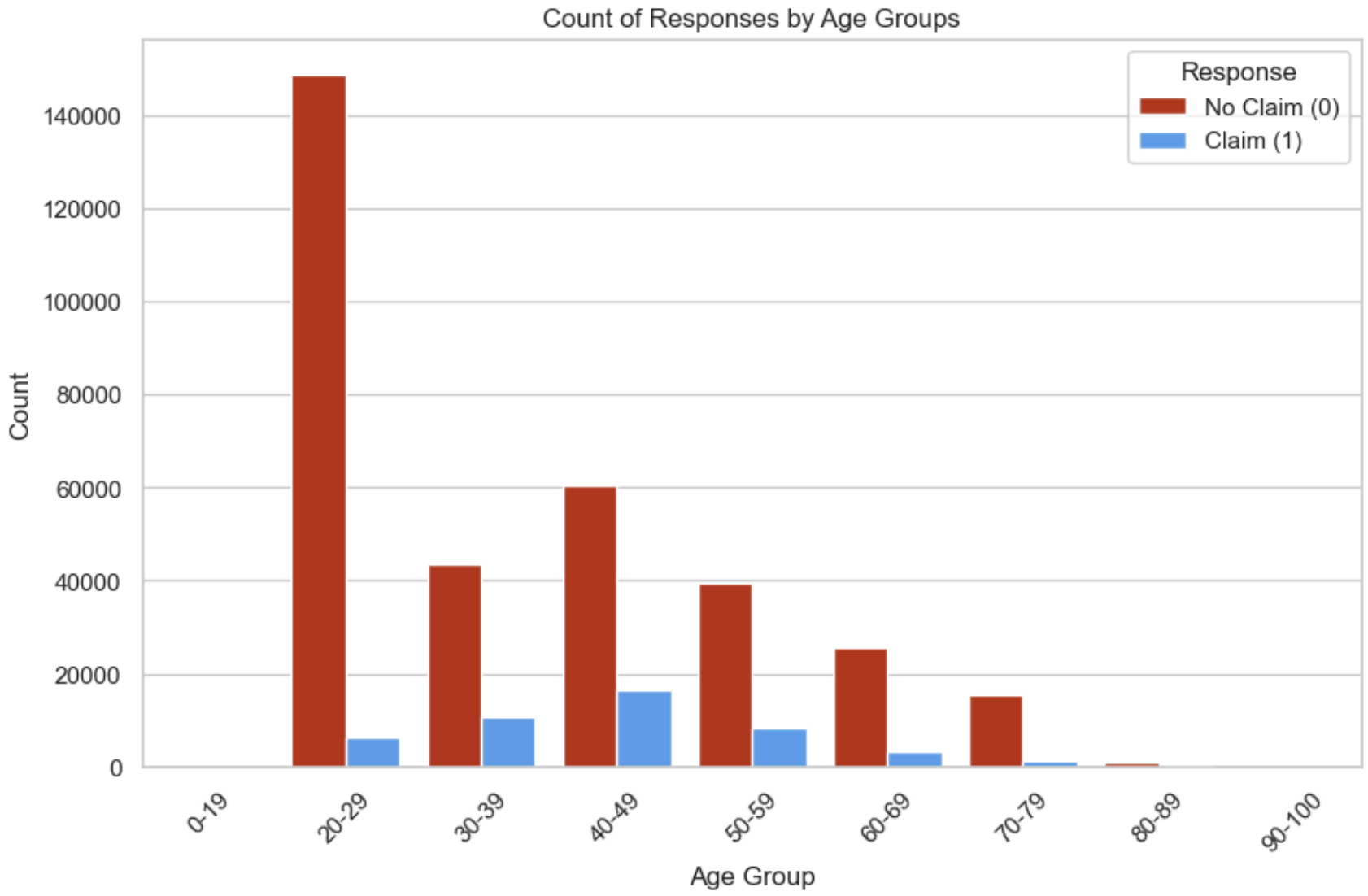



## Key Findings :

- **Highest Frequency (Ages 20-25)**: The dataset has the highest number of individuals in the 20-25 age range.
- **Decreasing Frequency After Age 30**: The number of individuals steadily decreases as age increases past 30.
- **Few Older Individuals (Above 70)**: The dataset contains very few individuals over the age of 70.
- **Skewed Distribution**: The age distribution is heavily skewed toward younger demographics, which may affect age-related analysis or modeing.

## 6.2) Age impact on Response (claims) :

In [45]:
```
# Define age bins and labels
bins = [0, 20, 30, 40, 50, 60, 70, 80, 90, 100]  # Adjust bins as needed
labels = ['0-19', '20-29', '30-39', '40-49', '50-59', '60-69', '70-79', '80-89', '90-100']
df1['Age Group'] = pd.cut(df1['Age'], bins=bins, labels=labels, right=False)

# Create a count plot for responses by age groups with two distinct colors for hue categories
plt.figure(figsize=(10, 6))
sns.countplot(x='Age Group', hue='Response', data=df1, palette=['#c82b09', '#4c9aff'])  # Provide two color
plt.title('Count of Responses by Age Groups')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.xticks(rotation=45)  # Rotate x-axis labels for better visibility
plt.legend(title='Response', loc='upper right', labels=['No Claim (0)', 'Claim (1)'])
plt.show()
```




## Findings from the Countplot

- **Age 20-29**: This age group has the highest count of "No Claim" responses.
- **Older Age Groups (40-49, 50-59)**: These groups show moderate "No Claim" counts, with a slight increase in "Claim" responses compared to younger groups.
- **Very Old Groups (80+)**: These groups have significantly lower counts for both "Claim" and "No Claim" responses.

**Overall Trend** : Claims are less frequent than no claims, particularly among younger populations.

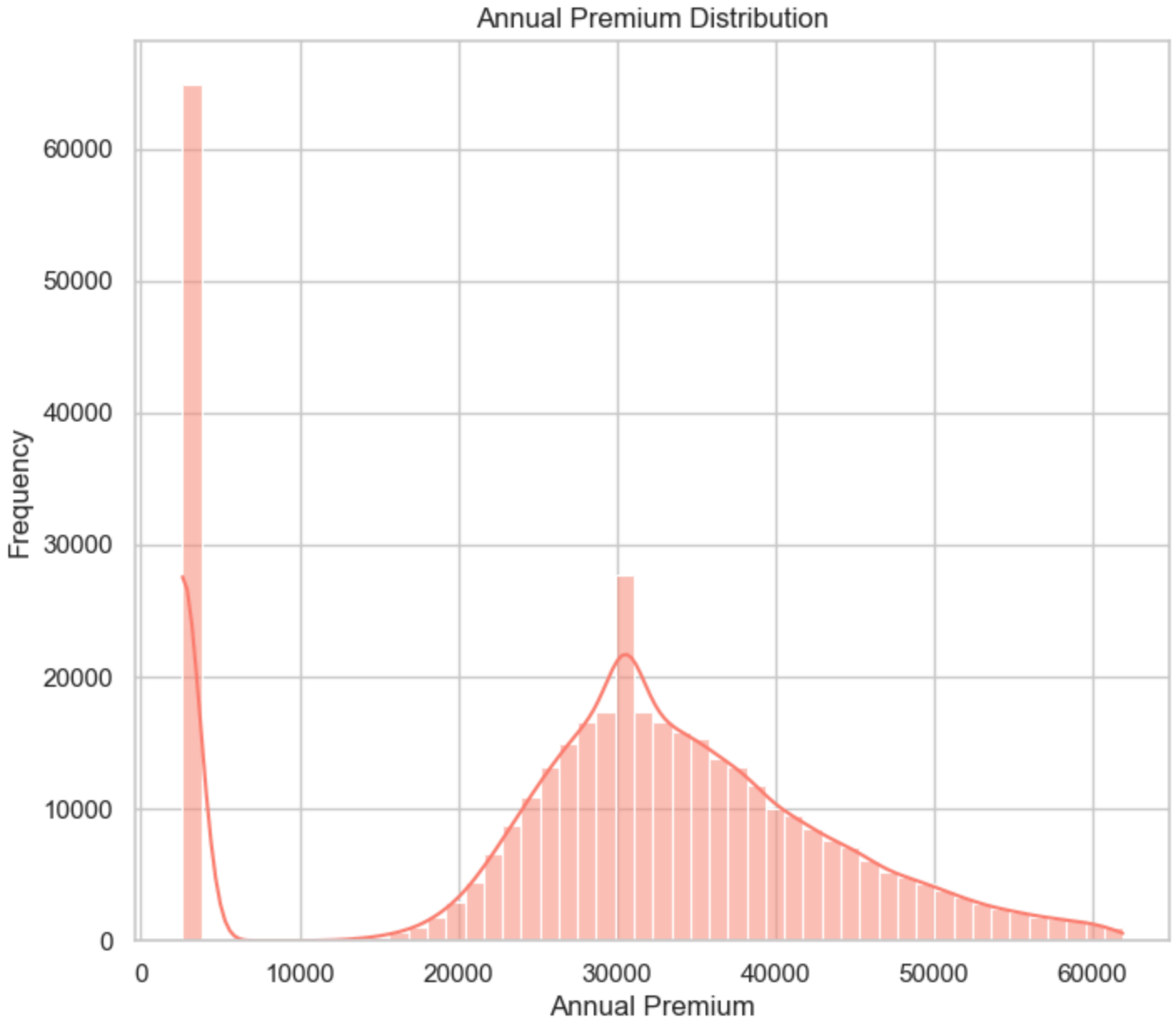
# Step 7: Premium Analysis :->

Objective:

Investigate the distribution of annual premiums to understand how insurance costs vary and whether premiums influence claim frequencies.

## 7.1) The distribution of Annual Premium :

In [47]:
```
plt.figure(figsize=(8, 7))
sns.histplot(df1['Annual_Premium'], kde=True, bins=50, color="salmon")
plt.title('Annual Premium Distribution')
plt.xlabel('Annual Premium')
plt.ylabel('Frequency')
plt.show()
```




## Findings from the Histogram: Annual Premium Distribution

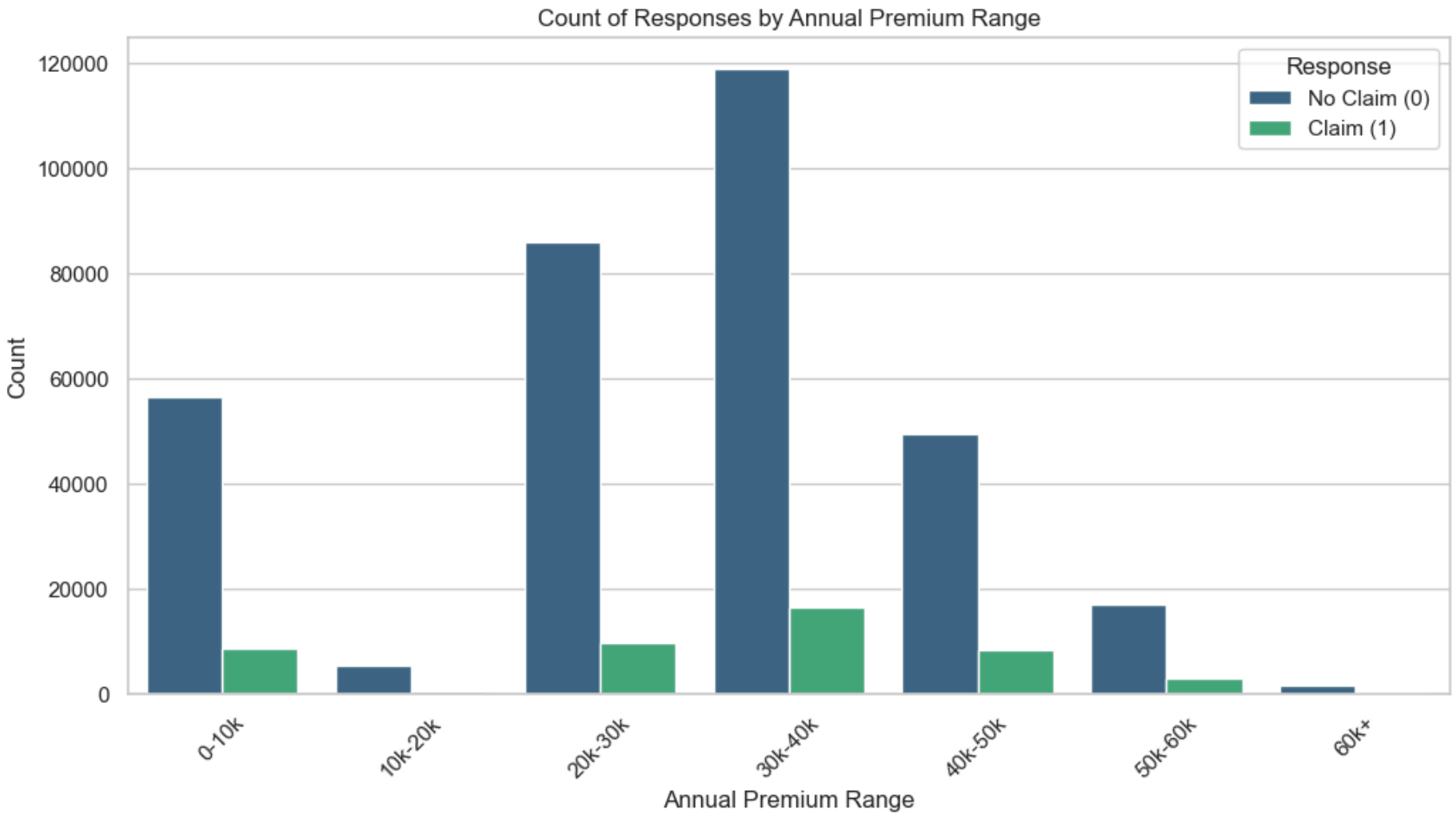
- **High Frequency at Low Premiums**: A significant spike near 0 indicates a high frequency of low premium values.

- **Bell-Shaped Distribution**: Beyond the initial spike, the distribution exhibits a normal-like shape, peaking around 30,000.
- **Right Skewness**: The tail extends toward higher premium values (above 30,000), indicating fewer high premium policies.

**Conclusion**: Most customers pay moderate to low premiums, with a few outliers paying very high amounts.

## 7.2) Impact of Annual Premium on Response :

In [49]:
```
# Bin Annual_Premium into ranges
bins = [0, 10000, 20000, 30000, 40000, 50000, 60000, 70000]
labels = ['0-10k', '10k-20k', '20k-30k', '30k-40k', '40k-50k', '50k-60k', '60k+']
df1['Premium_Range'] = pd.cut(df1['Annual_Premium'], bins=bins, labels=labels, right=False)

# Create the countplot
plt.figure(figsize=(12, 6))
sns.countplot(x='Premium_Range', hue='Response', data=df1, 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()
```




## Findings from the Bar Chart: Response by Annual Premium Range

1. **Premium Range 0-10k**:
   - Higher count of "No Claim" responses compared to "Claim" responses.

2. **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.

3. **Premium Range 40k-50k**:
    - The count of "No Claim" responses decreases but remains significant.
    - "Claim" responses slightly increase compared to lower ranges.

4. **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:

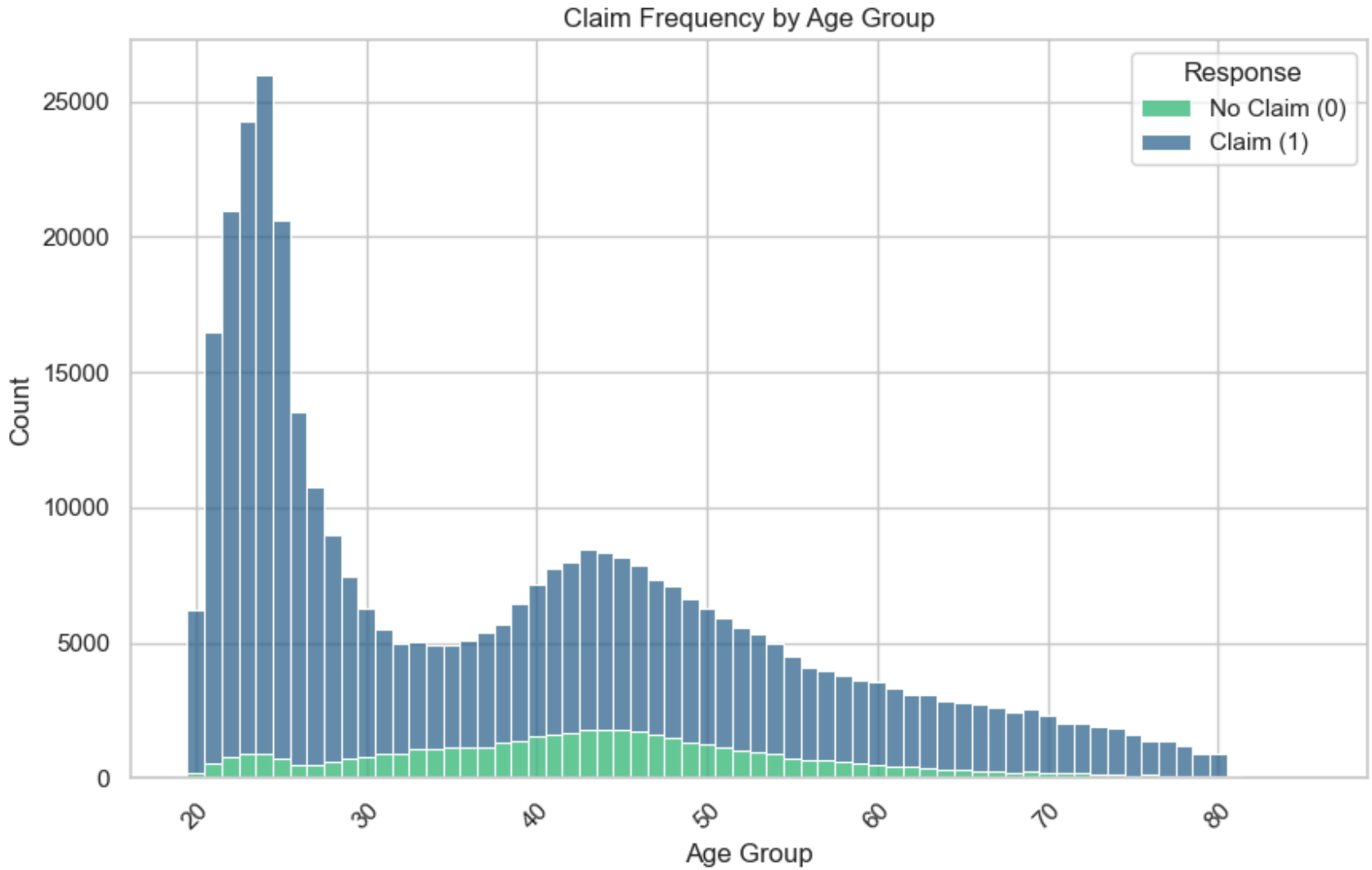
- Lower and moderate premium ranges (0-40k) dominate the dataset.
- "No Claim" responses significantly outnumber "Claim" responses across all premium ranges.
- Claims are relatively more frequent in the 30k-50k premium ranges compared to other ranges.

# Step 8: Claim Frequency Analysis :->

**Explore factors contributing to higher claim frequencies**

## 8.1) Claim Frequency by Age :

In [51]:
```
# Plotting the histogram with claim frequencies by age group
plt.figure(figsize=(10, 6))
sns.histplot(data=df1, x='Age', hue='Response', multiple='stack', palette='viridis', discrete=True)
plt.title('Claim Frequency by Age Group')
plt.xlabel('Age Group')
plt.ylabel('Count')
plt.legend(title='Response', labels=['No Claim (0)', 'Claim (1)'])
plt.xticks(rotation=45)
plt.show()
```




## Findings from the Bar Chart: Claim Frequency by Region

- **Variation Across Regions**:
  - Claim frequencies vary significantly, with some regions having higher proportions than others.

- **No Clear Trend**:

    - There is no apparent sequential pattern; high and low frequencies are scattered across region codes.

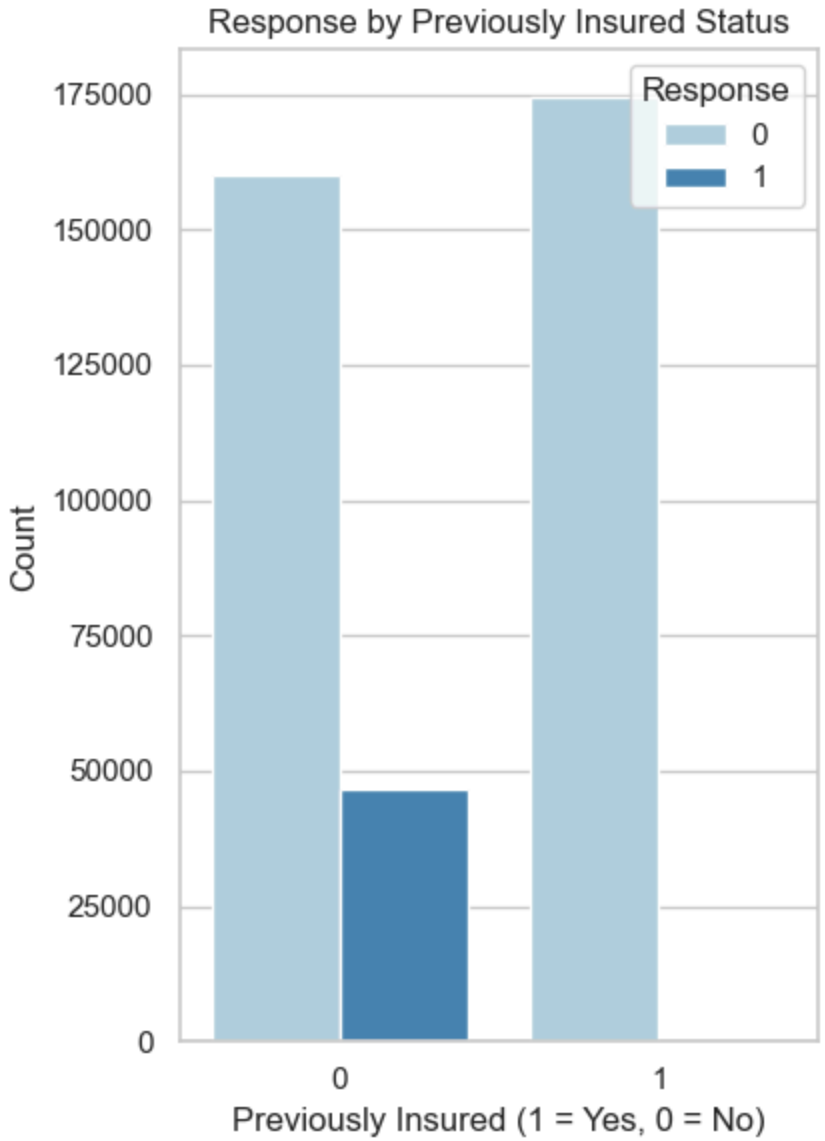
- **Highest Frequencies**:
    - Peaks occur for certain regions (e.g., around 10, 30, and 40), indicating they have the highest claim rates.

### Insight:

- **Claim frequency is region-dependent but not uniform or linear.**

## 8.2) Claim Frequency by Previously Insured :

In [53]:
```
plt.figure(figsize=(12, 6))  # Adjust the width and height as needed
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()
```




## Findings from the Chart: Responses Based on Insurance Status

- **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.

### Key Insight:

- Uninsured individuals are slightly more likely to respond positively, though negative responses dominate both groups.

# Step 9 : Gender Analysis :->

### Objective:

Evaluate the role of gender in claim frequencies to identify any significant differences in claim behavior between male and female policyholders.

## 9.1) Claim Frequency by Gender :

In [55]:
```
sns.countplot(data=df1, x='Gender', hue='Response', palette='Set2')
plt.title('Claims by Gender', fontsize=14)
plt.xlabel('Gender', fontsize=12)
plt.ylabel('Count', fontsize=12)
plt.show()
```


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

## Observations :

- **No Claims (0)**:
  - Dominate for both genders, represented by the larger blue bars.
- **Claims (1)**:
  - Relatively low for both males and females, with orange bars much shorter than blue bars.
  - The distribution of claims and no claims appears similar for both genders, with no significant difference in the pattern.

## Key Insight:

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

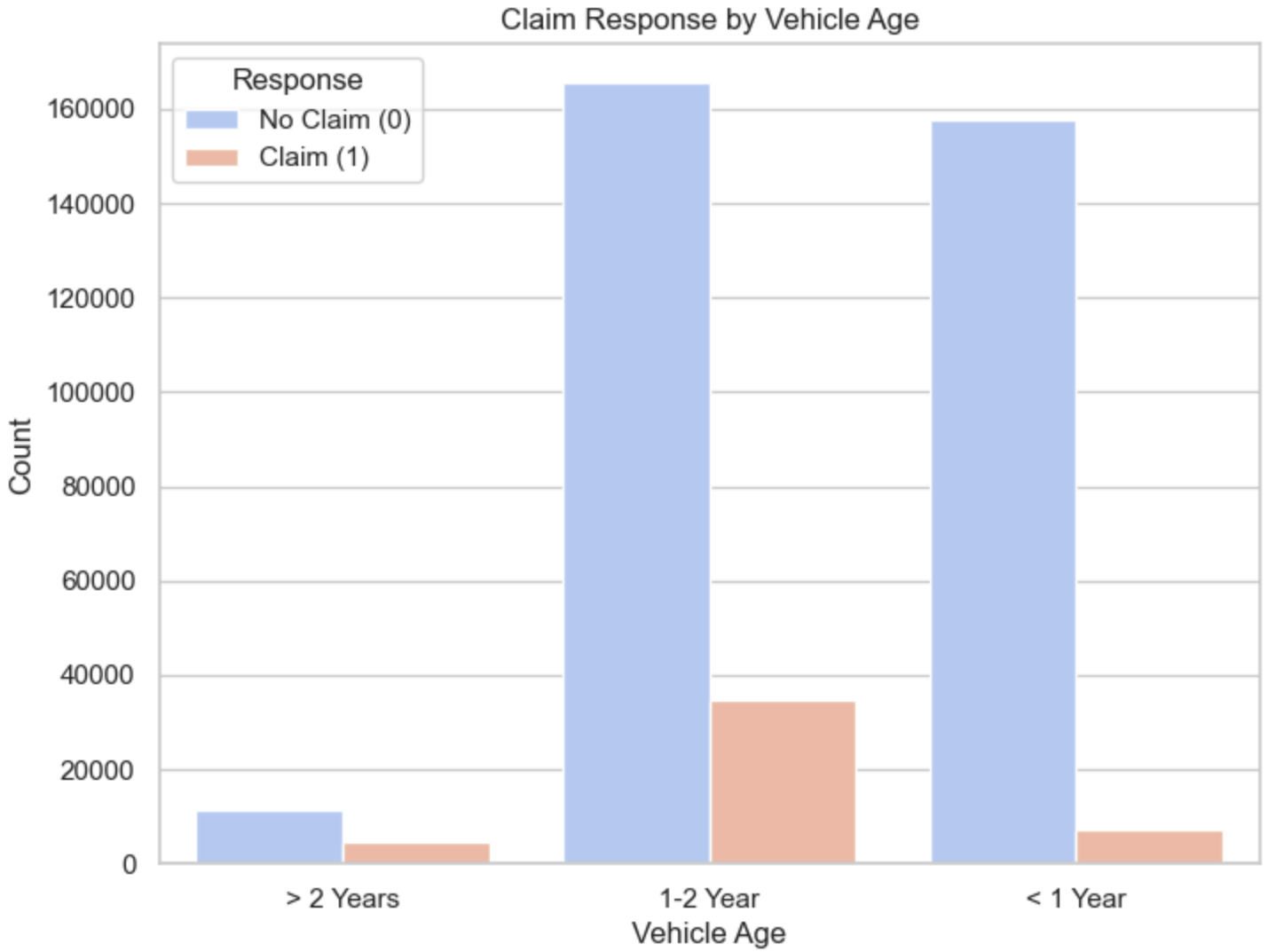
# Step 10 : Vehicle Age and Claim Frequency :->

## Objective:

Explore how the age of the vehicle influences the likelihood of a claim being filed.

### 10.1) Claim Response by Vehicle Age :->

In [57]:
```
plt.figure(figsize=(8, 6))
sns.countplot(x='Vehicle_Age', hue='Response', data=df1, 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()
```




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

## Observations :

- Most Vehicles :
  - The majority of vehicles are less than 1 year old.

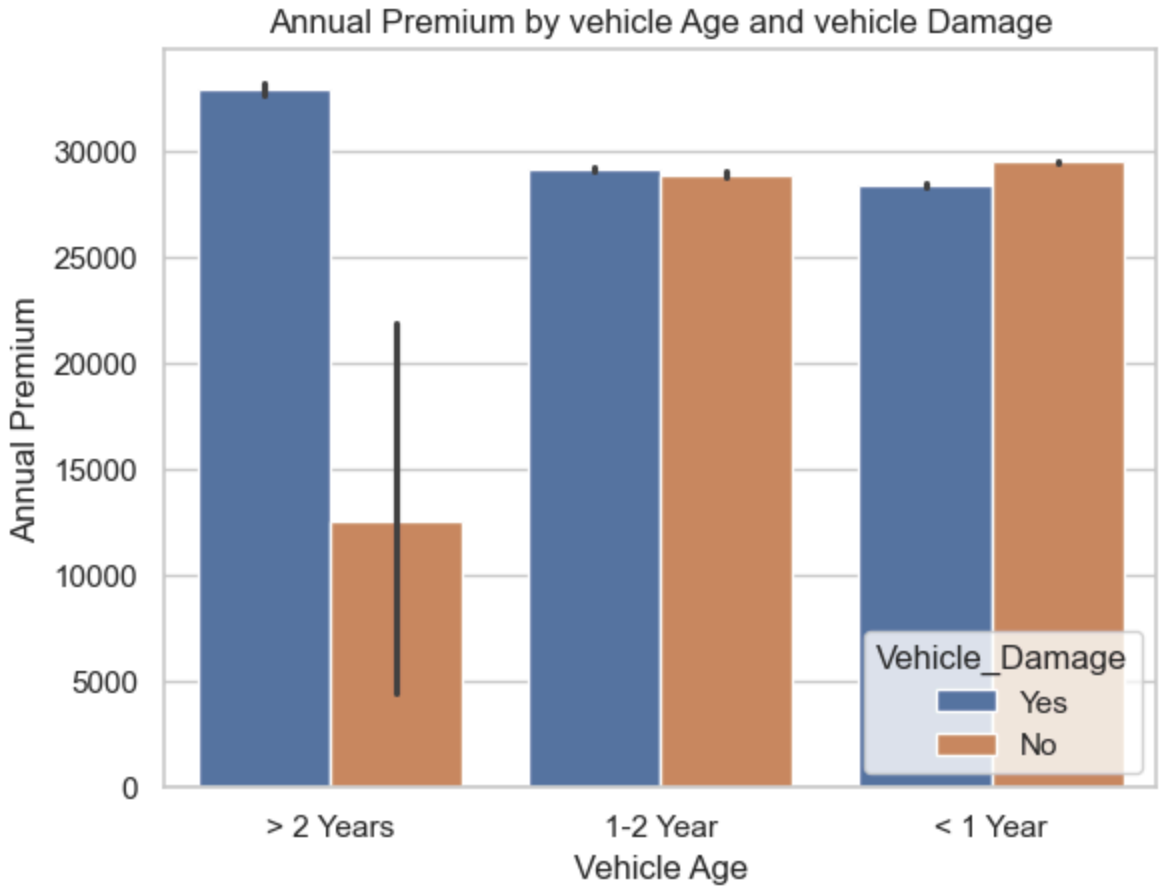
- Claims :
  - Newer vehicles (<1 year old) have a higher number of claims compared to older vehicles.
- Older Vehicles :
  - Vehicles older than 2 years are fewer in number and also have fewer claims.

**Key Insight** :

- The data indicates that newer vehicles are more likely to have claims, while older vehicles are both less frequent and less likely to file claims.

## 10.2) Annual Premium by vehicle Age and vehicle Damage :

In [59]:
```
# Examine the impact of vehicle age on insuance claims
sns.barplot(data=df,x='Vehicle_Age',y='Annual_Premium',hue='Vehicle_Damage')
plt.title("Annual Premium by vehicle Age and vehicle Damage")
plt.xlabel("Vehicle Age")
plt.ylabel("Annual Premium")
plt.show()
```




Observations:

1. **Vehicles Older than 2 Years**:
   - Vehicles that have been damaged have significantly higher annual premiums compared to undamaged vehicles.
   - There is a noticeable gap in premiums between damaged (higher) and undamaged (lower) vehicles in this category.
2. **Vehicles Aged 1-2 Years**:
   - The annual premiums for damaged and undamaged vehicles are very similar, showing minimal variance.
   - This indicates a consistent premium structure regardless of vehicle damage status in this age group.
3. **Vehicles Less than 1 Year Old**:

- Similar to the 1-2 year category, the annual premiums for damaged and undamaged vehicles are almost equal.
- Premiums appear to be uniformly distributed, unaffected by vehicle damage status.

### Key Insights:

- **Premium Differentiation**:
  - Older vehicles (>2 years) have a clear differentiation in premiums based on damage status, while newer vehicles (<1 or 1-2 years) do not.
- **Damage Impact**:
  - Vehicle damage has a significant impact on premiums for older vehicles but little to no impact for newer vehicles.
- **Premium Uniformity**:
  - The premium structure is more uniform for newer vehicles, regardless of their damage status.

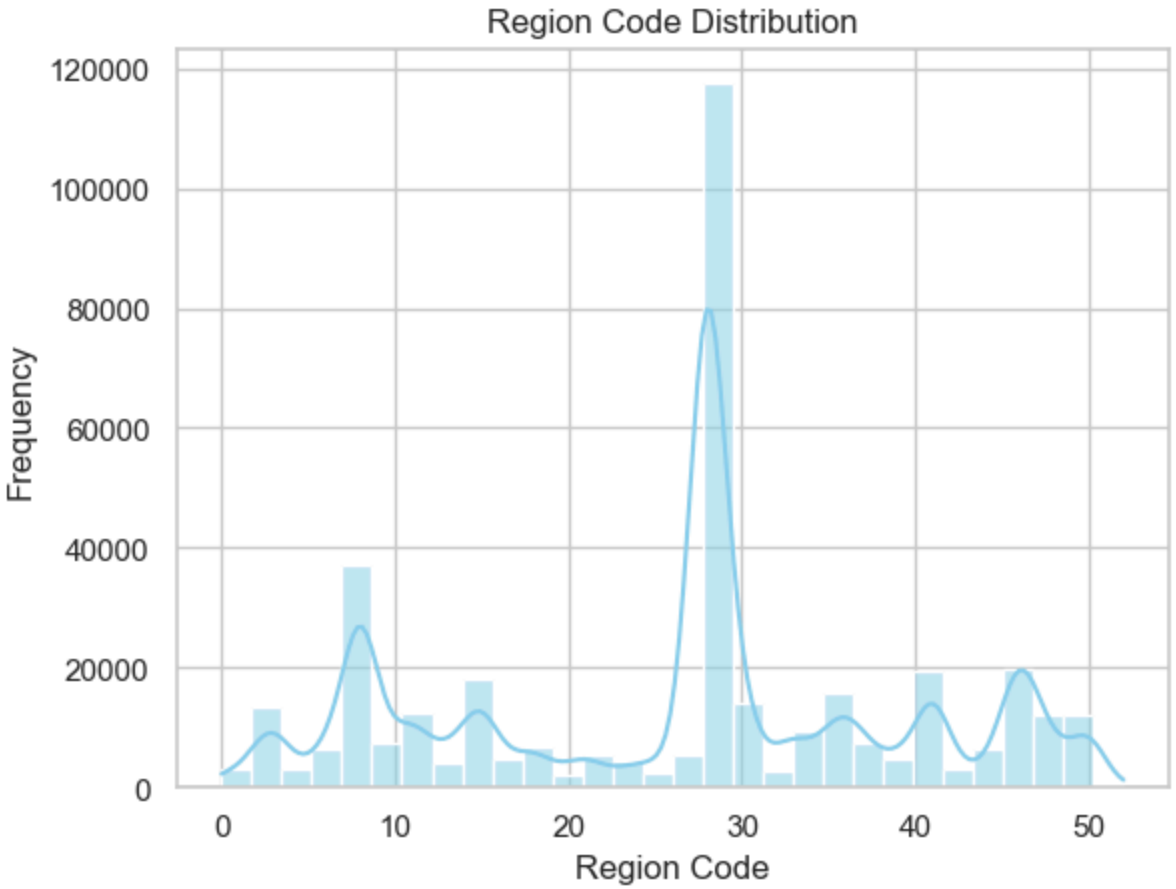
# Step 11 : Region-wise Analysis :->

### Objective:

Analyze claims distribution across different regions to identify regional patterns or trends in claim behavior.

## 11.1) Region wise Distribution of Claims :

In [61]:
```
sns.histplot(df['Region_Code'],bins=30,kde=True,color='skyblue')
plt.title("Region Code Distribution")
plt.xlabel("Region Code")
plt.ylabel("Frequency")
plt.show()
```




Observations :

1. **High Concentration in Specific Regions**:
    - The region with the highest frequency of claims is around Region Code 30, with a peak frequency exceeding 120,000.
    - This region stands out significantly compared to others.
2. **Moderate Claim Frequency**:
    - Regions near Region Code 10 show a moderate frequency of claims, but they are much lower compared to the peak around Region Code 30.
3. **Low Claim Frequency**:

- Most other regions (codes beyond 30 and below 10) have relatively low frequencies of claims, distributed sporadically across the graph.

4. **Asymmetry in Distribution**:
   - The claim frequency is highly skewed, with a sharp peak around Region Code 30 and a gradual decline across other regions.

### Key Insights :

- **Dominant Region**:
  - Claims are heavily concentrated in a specific region (Region Code 30), suggesting this area has the highest activity or policyholders.
- **Skewed Distribution**:
  - The claim frequency is not evenly distributed across regions, indicating region-specific factors may influence claims.
- **Target Analysis**:
  - Regions with high claim frequencies (e.g., Region Code 30 and near Region Code 10) might require deeper investigation into the factors driving this trend, such as population density or risk exposure.

# Step 12 : Policy Analysis :->

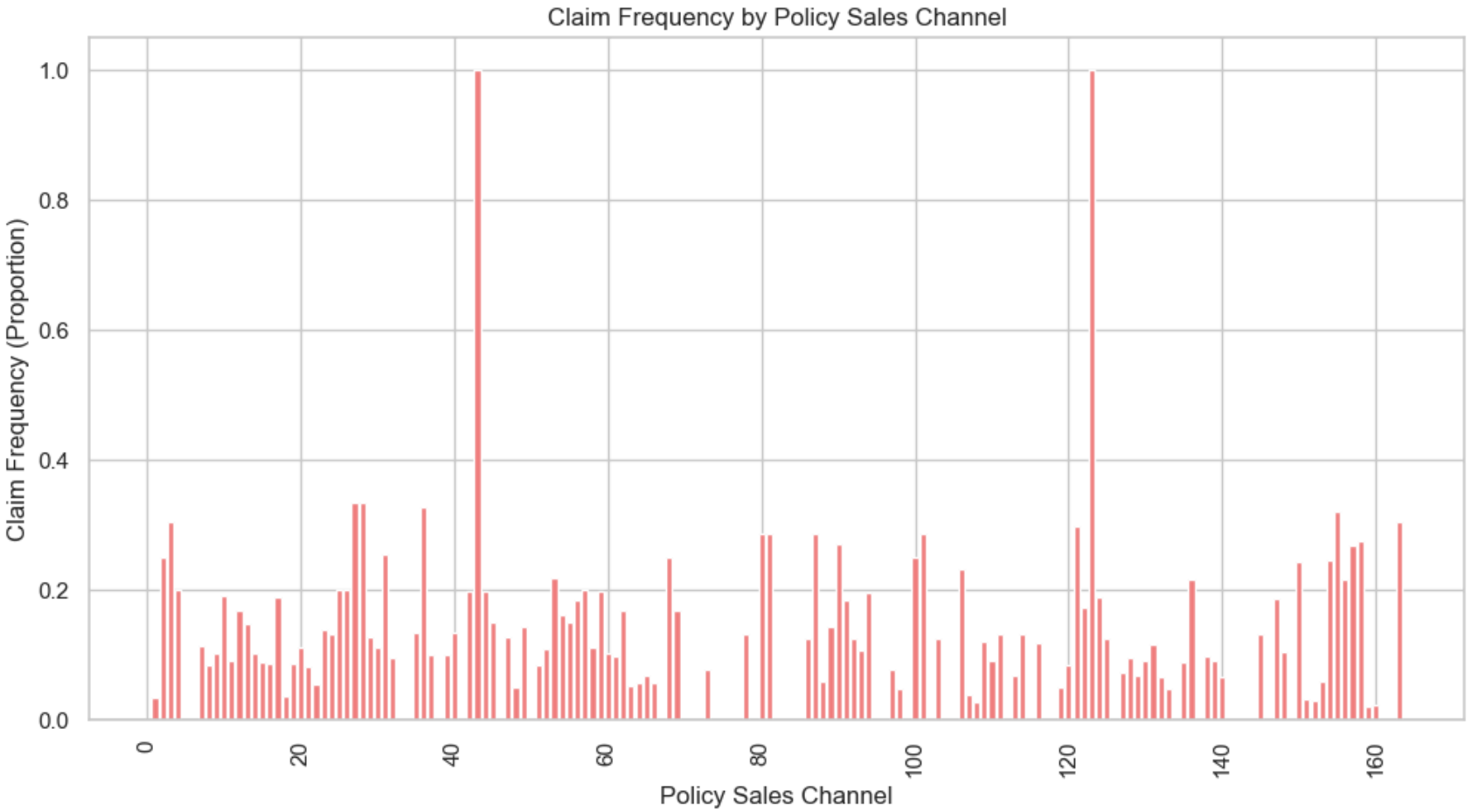
Objective:

Investigate different policy types (Policy_Sales_Channel) and their influence on claims.

## 12.1) Claim Frequency by Policy Sales Channel :

In [63]:
```
sales_channel_frequency = df1.groupby('Policy_Sales_Channel')['Response'].mean()
plt.figure(figsize=(12, 6))
plt.bar(sales_channel_frequency.index, sales_channel_frequency.values, color="lightcoral")
plt.title('Claim Frequency by Policy Sales Channel')
plt.xlabel('Policy Sales Channel')
```

```
plt.ylabel('Claim Frequency (Proportion)')
plt.xticks(rotation=90)
plt.show()
```




## Findings :

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

1. **Low Claim Frequencies**:
   Most policy sales channels have relatively low claim frequencies, generally below 0.2.

2. **High Claim Frequencies in Specific Channels**:
   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. **Significant Variability**:
   There is significant variability in claim frequencies across the sales channels, with some clusters showing moderate frequencies and others near-zero.

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

**This variability indicates a potential need for tailored risk management by sales channel.**

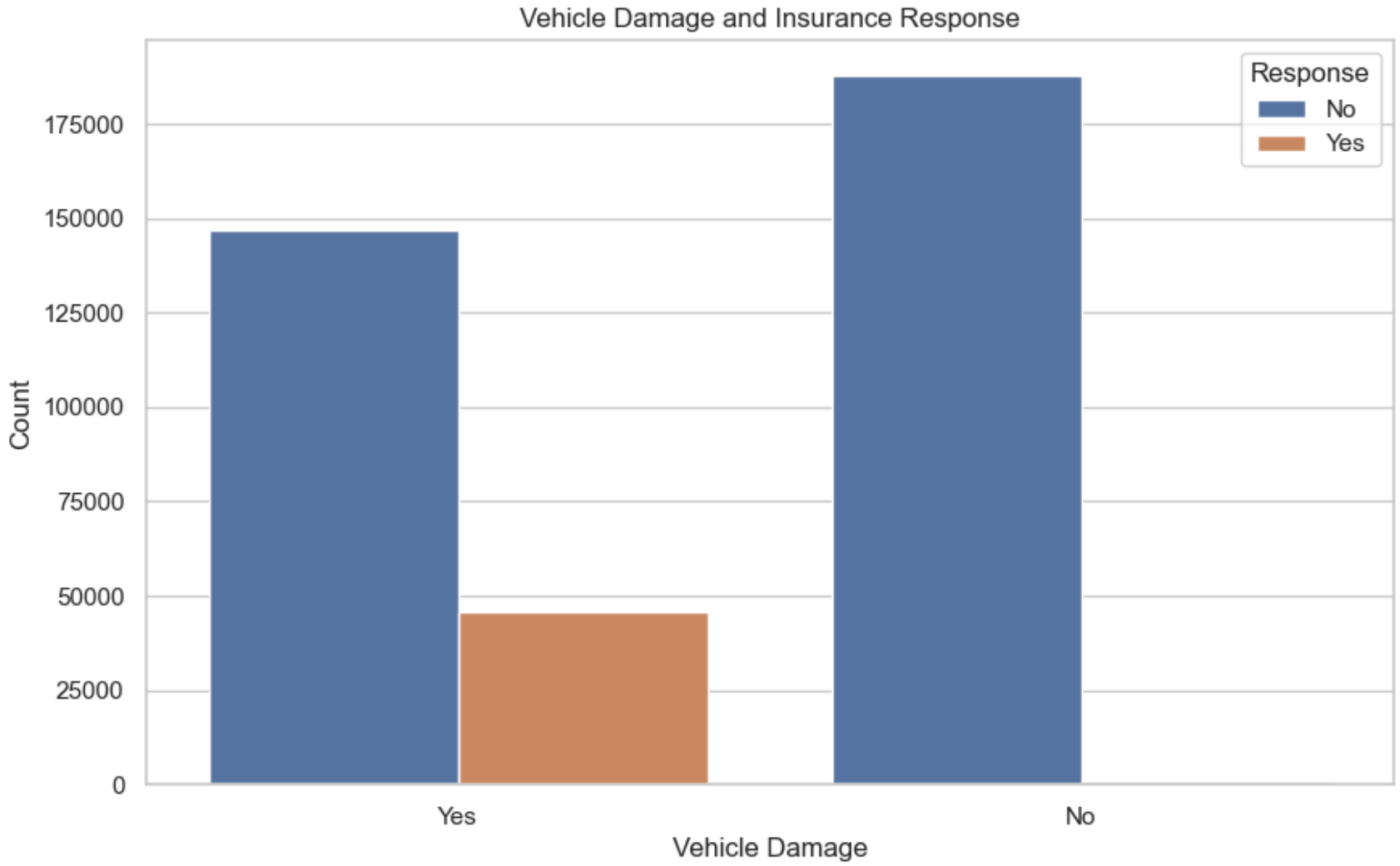
# Step 13 : Claim Frequency by Vehicle Damage :->

Objective:

Assess how Vehicle_Damage affects claim frequency.

## 13.1) Relationship between vehicle damage and claim frequencies :

In [65]:
```
plt.figure(figsize=(10, 6))
sns.countplot(data=df1, 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()
```




## Key Findings from the Chart :

1. **Overall Observations**:
   - 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.

2. **Specifics**:

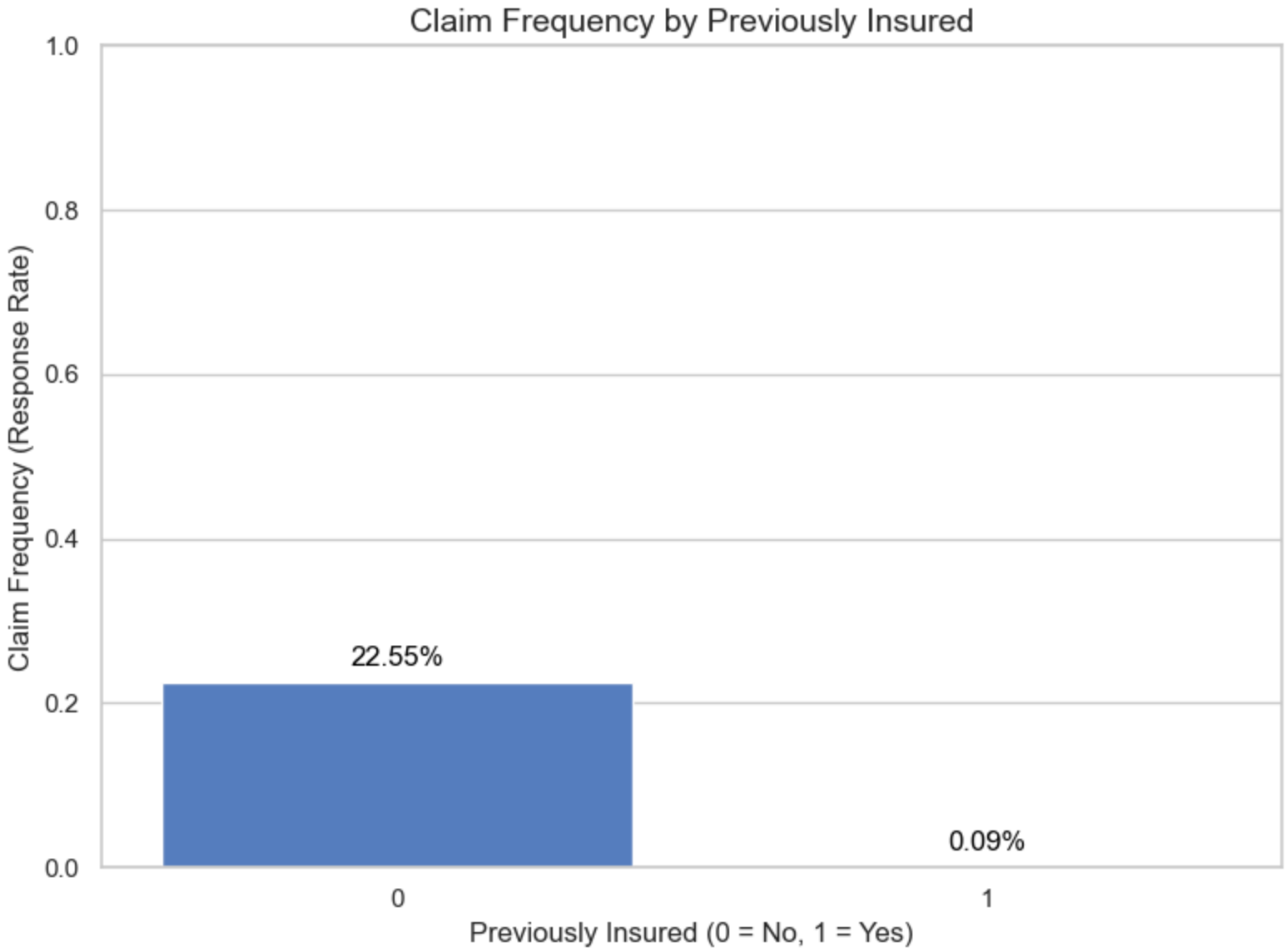
- In cases of no vehicle damage, the vast majority did not file an insurance claim.
- In cases of vehicle damage, a significant proportion filed an insurance claim.

3. **Additional Notes**:
   - 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 14: Customer Loyalty Analysis :->

## 14.1) Claim Frequency by Previously Insured :

In [67]:
```
insured_response_rate = df1.groupby('Previously_Insured')['Response'].mean().reset_index()

# Bar Plot for Claim Frequency
plt.figure(figsize=(8, 6))
bar_plot = sns.barplot(
    x='Previously_Insured', y='Response', data=insured_response_rate, hue='Previously_Insured', palette="mu
)

# Annotate bars with percentages
for index, row in insured_response_rate.iterrows():
    bar_plot.text(
        x=index, y=row['Response'] + 0.02,
        s=f"{row['Response']:.2%}",
        ha='center', fontsize=12, color='black'
    )

plt.title("Claim Frequency by Previously Insured", fontsize=14)
plt.xlabel("Previously Insured (0 = No, 1 = Yes)", fontsize=12)
plt.ylabel("Claim Frequency (Response Rate)", fontsize=12)
plt.ylim(0, 1)
plt.legend([], [], frameon=False)  # Removes legend if not needed
plt.tight_layout()
plt.show()
```




## Findings:

1. **Previously Insured (Value = 1)**:
   - Claim frequency (response rate) is **0.09%**sured.

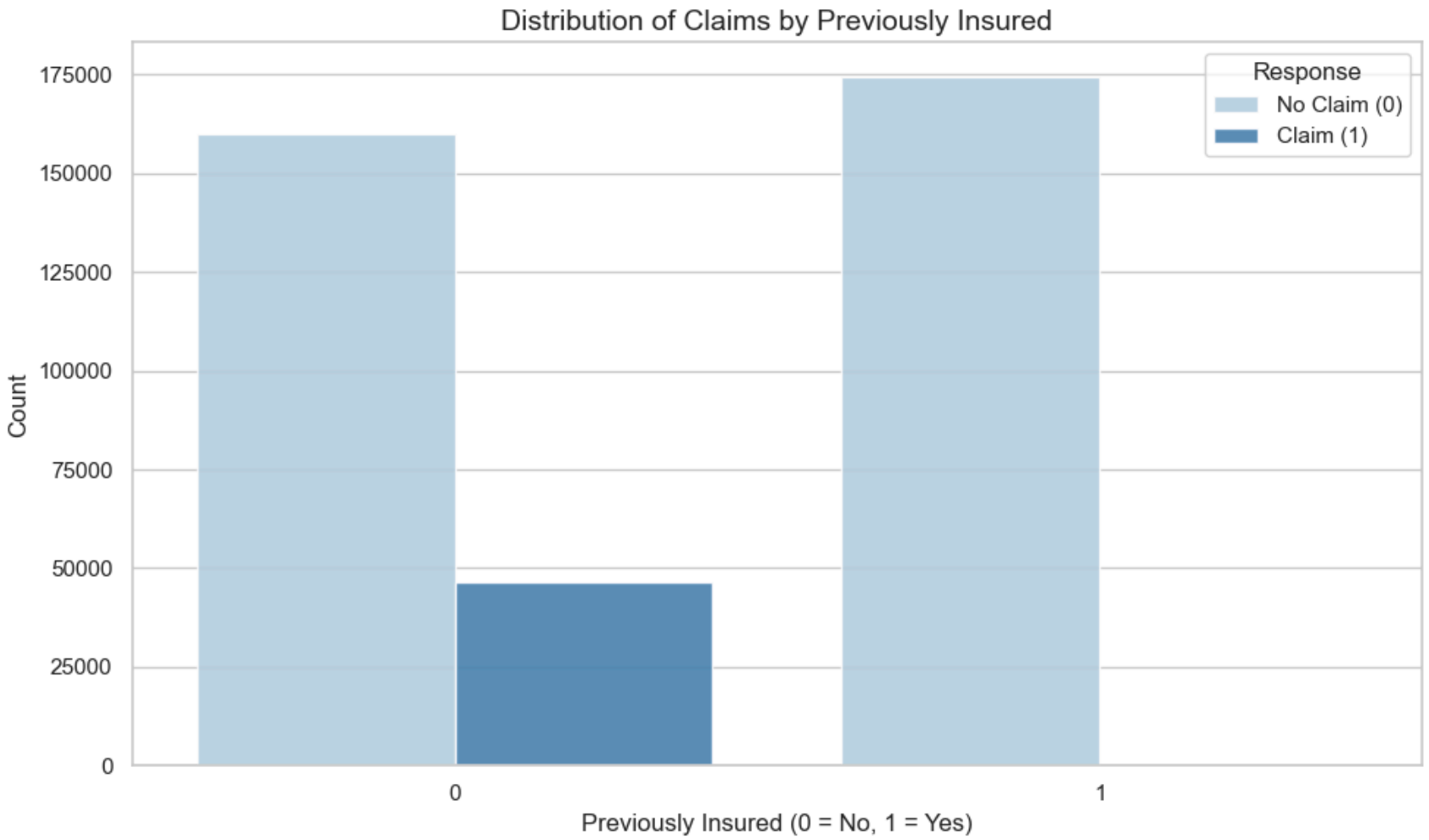
2. **Not Previously Insured (Value = 0)**:
    - Claim frequency (response rate) is **22.55%**.

### Insight:

- Individuals not previously insured have a **significantly higher claim frequency** compared to those who were previously insured.

## 14.2) Claims by Previously Insured :

In [69]:
```
plt.figure(figsize=(10, 6))
sns.countplot(
    x='Previously_Insured', hue='Response', data=df1, palette="Blues", alpha=0.9
)
plt.title("Distribution of Claims by Previously Insured", fontsize=14)
plt.xlabel("Previously Insured (0 = No, 1 = Yes)", fontsize=12)
plt.ylabel("Count", fontsize=12)
plt.legend(
    title="Response", loc="upper right", labels=["No Claim (0)", "Claim (1)"]
)
plt.tight_layout()
plt.show()
```




## Findings :

1. **Claim Likelihood**: People previously insured (value = 1) are **more likely to file a claim** than those who were not previously insured (value = 0).

2. **Non-Claiming Majority**: The bar chart shows that the **number of people who were not previously insured (value = 0) and did not file a claim (value = 0)** is much larger than all other combinations.

3. **Skewed Distribution**: The distribution of claims is heavily **skewed towards people who were not previously insured (value = 0) and did not file a claim (value =0)**.

# Project Summary: Vehicle Insurance Analysis

This project aimed to explore key aspects of vehicle insurance data, providing insights into customer behavior, claim patterns, and premium distributions. Here's a summary of the findings:

## 1. Data Preprocessing and Cleaning

- Missing values were identified and addressed.
- Data types were standardized for analysis.

## 2. Exploratory Data Analysis (EDA)

- **Premium Analysis**: Premiums are heavily skewed towards lower values, with a majority of customers paying relatively low premiums.
- **Vehicle Age and Claims**: Older vehicles showed a higher claim frequency, suggesting a potential correlation between vehicle age and risk.
- **Region-wise Claims**: Claims were disproportionately higher in certain regions, indicating potential regional risk factors.

## 3. Customer Demographics

- **Gender and Age Distribution**: No significant difference in claim response between genders. However, specific age groups exhibited higher claim frequencies.
- **Customer Segmentation**: Younger customers with newer vehicles tended to have lower claims, whereas older vehicles and customers had higher claim frequencies.

## 4. Policy and Claim Analysis

- **Claim Frequencies**: Factors such as vehicle damage and previously insured status were strong indicators of claim likelihood.
- **Time Analysis**: Certain months or periods had higher claims, suggesting seasonal trends.

## 5. Insights and Recommendations

- Insurance premiums could be better optimized based on regional and vehicle-specific risk factors.
- Tailored policies could be designed for high-risk demographics and regions to mitigate potential losses.
- Further investigation into seasonal trends can help in better resource allocation during peak periods.

# Conclusion

This project provided a comprehensive overview of the dataset, uncovering key patterns and relationships that could guide better decision-making in the insurance industry. These insights can be leveraged to refine pricing strategies, enhance customer targeting, and improve overall risk management.