EXPLORATORY DATA ANALYSIS
Vehicle Insurance

Project by
KUSHAGR SAXENA

Introduction:

Vehicle insurance is a contract between a vehicle owner and an insurance company that provides financial protection against losses related to the vehicle. This can include coverage for damage to the vehicle itself, liability for injuries or damages to other people and their property, and protection against theft or vandalism.

The dataset is sourced from 'Skill Circle' & contains comprehensive information about vehicle insurance, including details on insured individuals, their vehicles, and associated claims. It offers insights into various factors such as Age, Gender, Regional code, Annual premiums, and Policy types.

By analyzing these variables in depth, we aim to uncover patterns, trends, and correlations that shed light on the factors affecting insurance claims.

This will involve steps such as loading the data, generating descriptive statistics, profiling the data, identifying outliers, and using visualization techniques.

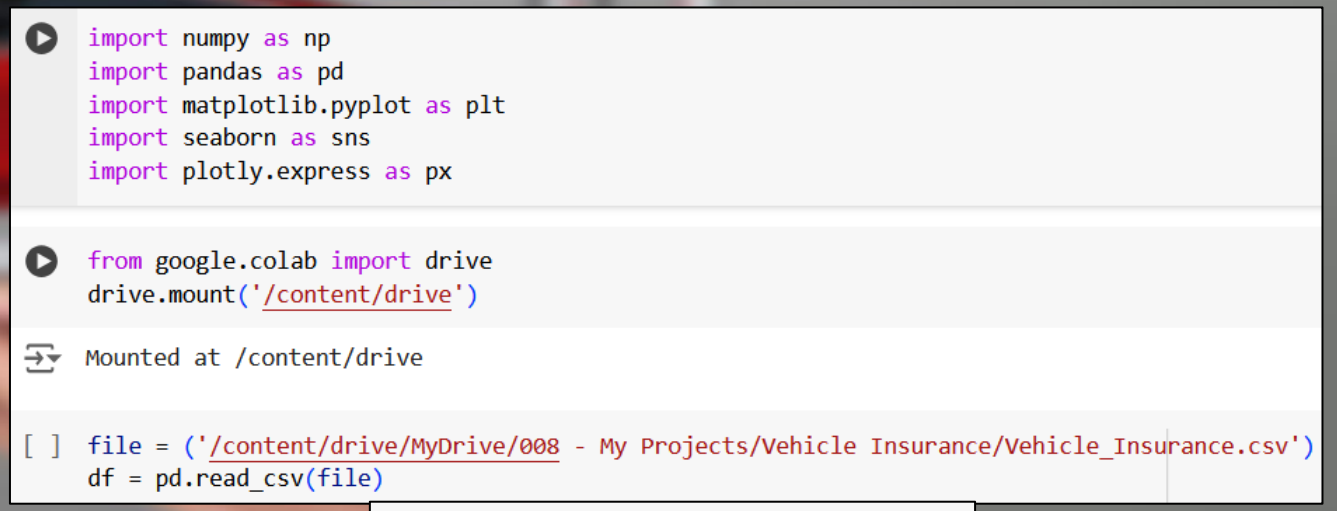
Objectives of the project:

The goals of this assessment is to –

I. ***Data Visualization:*** *Utilize various visualization techniques to explore the distribution of key variables.*
II. ***Feature Analysis:*** *Examine the relationship between features and the target variable.*
III. ***Age Distribution:*** *Analyze the age distribution within the dataset and its impact on insurance claims.*
IV. ***Premium Analysis:*** *Investigate the distribution of insurance premiums and their correlation with claim frequencies.*
V. ***Claim Frequencies:*** *Explore factors contributing to higher claim frequencies.*
VI. ***Gender Analysis:*** *Investigate the role of gender in insurance claims.*
VII. ***Vehicle Age and Claims:*** *Examine the impact of vehicle age on the likelihood of a claim.*
VIII. ***Claim Frequency by Vehicle Damage:*** *Investigate the relationship between vehicle damage and claim frequencies.*

Description of Dataset:

- *The Dataset has been imported from Google Drive.*
- *I have performed my work using Google Colaboratory Notebook.*
- *As we begin our Exploratory Data Analysis (EDA), I've named the dataset 'df'.*
- *The dataset comprises of 3,81,109 Rows and 12 Columns.*
- *For Data cleaning/visualization, I have utilized libraries like Numpy, Pandas, Seaborn, Matplotlib & Plotly.*
- *Any duplicate entries that were found have also been removed.*



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

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

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

```
[ ] file = ('/content/drive/MyDrive/008 - My Projects/Vehicle Insurance/Vehicle_Insurance.csv')
df = pd.read_csv(file)
```

```
'''Let's drop any duplicate entries

df.drop_duplicates()
print(f"Dataset Shape:", df.shape)
```

```
Dataset Shape: (381109, 12)
```

Description of Dataset:

The dataset being analyzed contains a thorough collection of information about vehicle insurance, encompassing various characteristics of insured individuals, their vehicles, and related insurance claims. Here's a detailed overview of the key components and variables in the dataset:

Key features include:

- ***ID:*** *A unique identifier for each record in the dataset.*
- ***Gender:*** *Indicates the gender of the insured individual (e.g., male or female).*
- ***Age:*** *Represents the age of the insured individual.*
- ***Driving License:*** *A binary indicator showing whether the individual holds a valid driving license (0 for no and 1 for yes). This is crucial for evaluating driving risk.*
- ***Region Code:*** *A code representing the geographic location of the insured individual.*

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 381109 entries, 0 to 381108
Data columns (total 12 columns):
 #   Column                Non-Null Count   Dtype  
---  ------                --------------   -----  
 0   id                    381109 non-null  int64  
 1   Gender                381109 non-null  object 
 2   Age                   381109 non-null  int64  
 3   Driving_License       381109 non-null  int64  
 4   Region_Code           381109 non-null  float64
 5   Previously_Insured    381109 non-null  int64  
 6   Vehicle_Age           381109 non-null  object 
 7   Vehicle_Damage        381109 non-null  object 
 8   Annual_Premium        381109 non-null  float64
 9   Policy_Sales_Channel  381109 non-null  float64
 10  Vintage               381109 non-null  int64  
 11  Response              381109 non-null  int64  
dtypes: float64(3), int64(6), object(3)
memory usage: 34.9+ MB
```

Description of Dataset:

Key features include:

- ***Previously Insured:*** *Indicates whether the individual has had insurance coverage before (0 for no, 1 for yes).*
- ***Vehicle Age:*** *Specifies the age of the vehicle being insured.*
- ***Vehicle Damage:*** *A binary variable indicating whether the vehicle has previously suffered damage (0 for no, 1 for yes).*
- ***Annual Premium:*** *The amount paid annually for the insurance policy.*
- ***Policy Sales Channel:*** *Indicates the channel through which the insurance policy was sold (e.g., online, agent, etc.).*
- ***Vintage:*** *Refers to the number of days the customer has been associated with the insurance company.*
- ***Response:*** *The target variable indicating whether the insured individual has made a claim (usually binary: 0 for no claim, 1 for claim).*

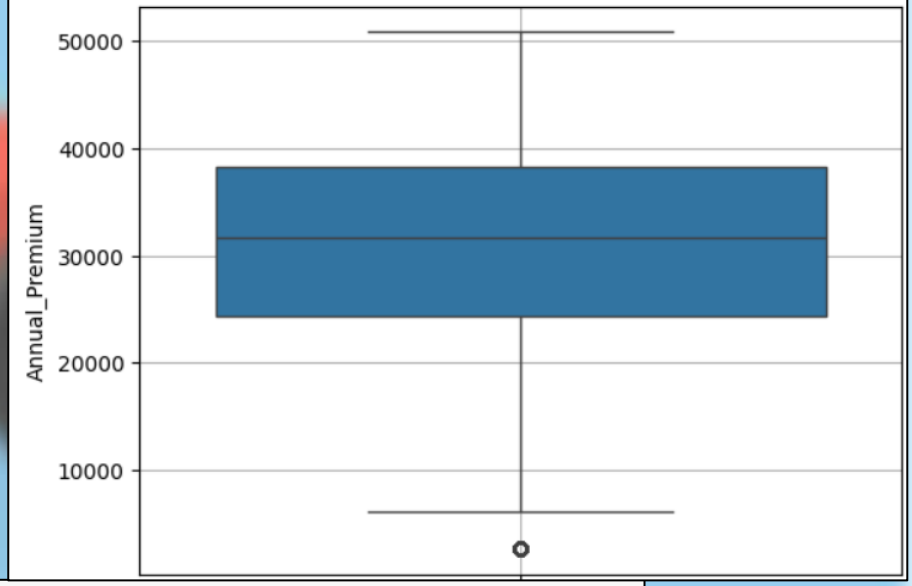
Data Cleaning & Pre-Processing:

The Dataset contains a total of 0 Null Values but we still have one Numerical feature ('Annual Premium') left to check at least for Outliers to maintain data equilibrium for better insights.

Annual Premium: This feature did contain a few Outliers which has been identified by using the Inter-Quartile Range (IQR) Method. Then those Outliers has been replaced with 'Median' to ensure data accuracy.

- *Since 64,877 values of 'Annual Premium' are same, which is 2,630.0, it will not be appropriate to replace or remove them. As it will impact accuracy of the dataset, therefore considering it as cleaned and proceeding further for EDA.*

```
sns.boxplot(df['Annual_Premium'])
plt.grid(True)
plt.show()
```




```
# Checking Median.

median = df['Annual_Premium'].median()
median
```

```
31669.0
```

```
[ ]  # Replacing Outliers in 'Annual_Premium' with Median

     df['Annual_Premium'] = np.where((df['Annual_Premium'] < lower_bound) | (df['Annual_Premium'] > upper_bound), median, df['Annual_Premium'])
```

Data Cleaning & Pre-Processing:

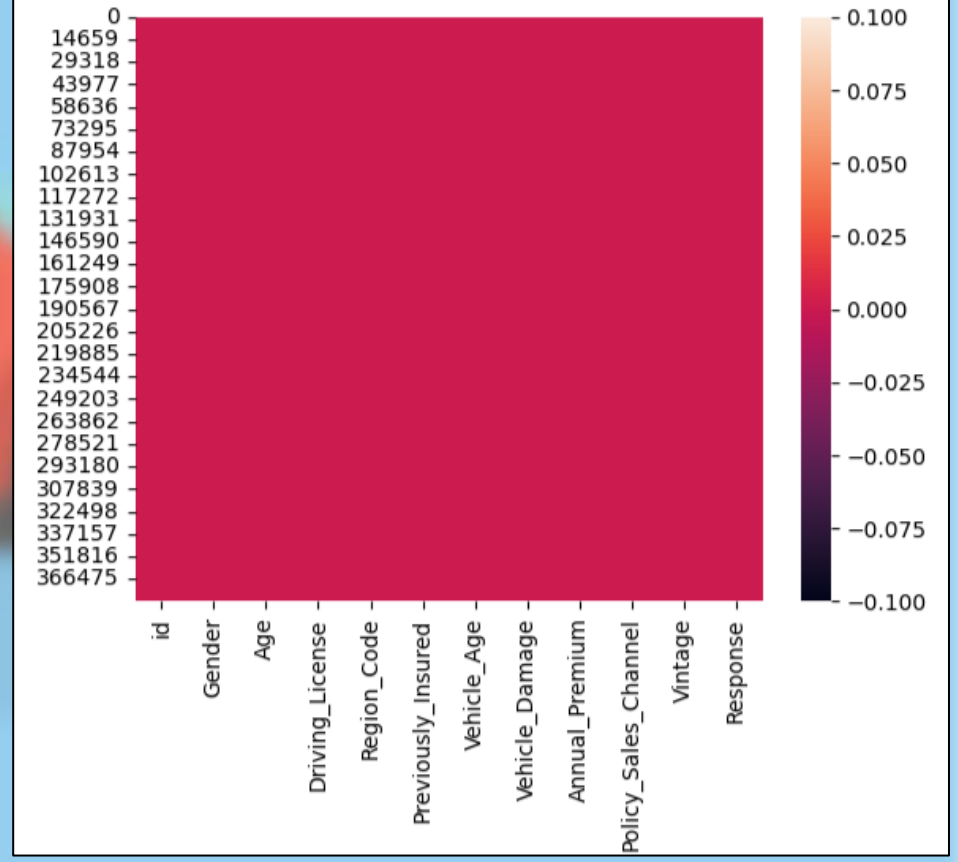
Summary *- To summarize, addressing Null values and Outliers necessitates a methodical approach tailored to the data's characteristics and specific attributes. By applying the outlined strategies, we can efficiently manage and fill in the missing values, thereby ensuring the dataset's completeness, integrity, and reliability for future analysis and insights.*

*Outliers were present in '**Annual Premium**'. These were addressed using the IQR method and capping to boundary values.*

With these Null, Missing, and Invalid values appropriately addressed, we are now ready to move forward with analyzing the dataset.

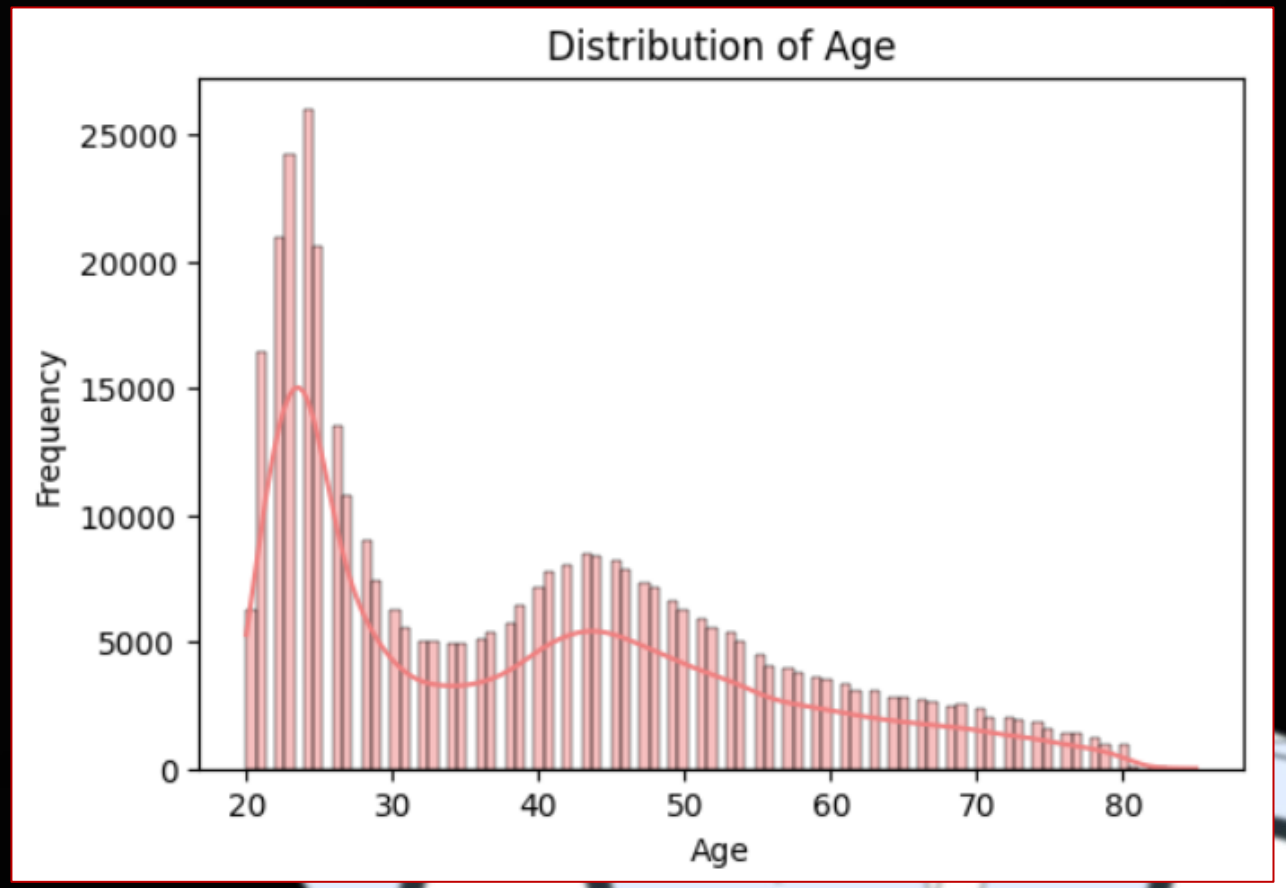
```
'''Heatmap of Null values'''

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



Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Age.



Data Visualization and Insights

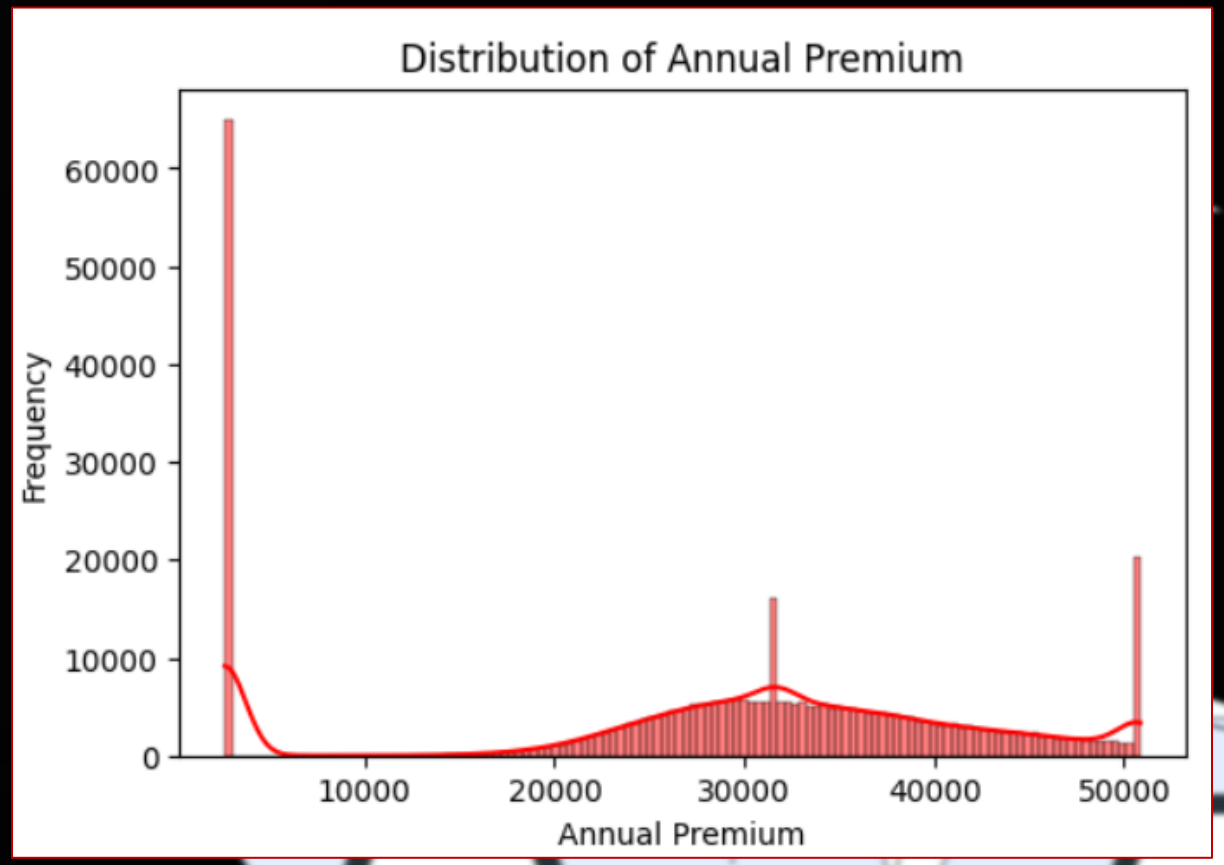
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Age.

Key Insights

- ***The age of policyholders is mostly distributed between 20 and 60 years old.***
- ***It shows a right-skewed distribution, indicating a higher concentration of younger individuals compared to older individuals.***
- ***There's a peak around the age of 25-30, suggesting a significant portion of the policyholders belong to that age group.***

Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Annual premium.



Data Visualization and Insights

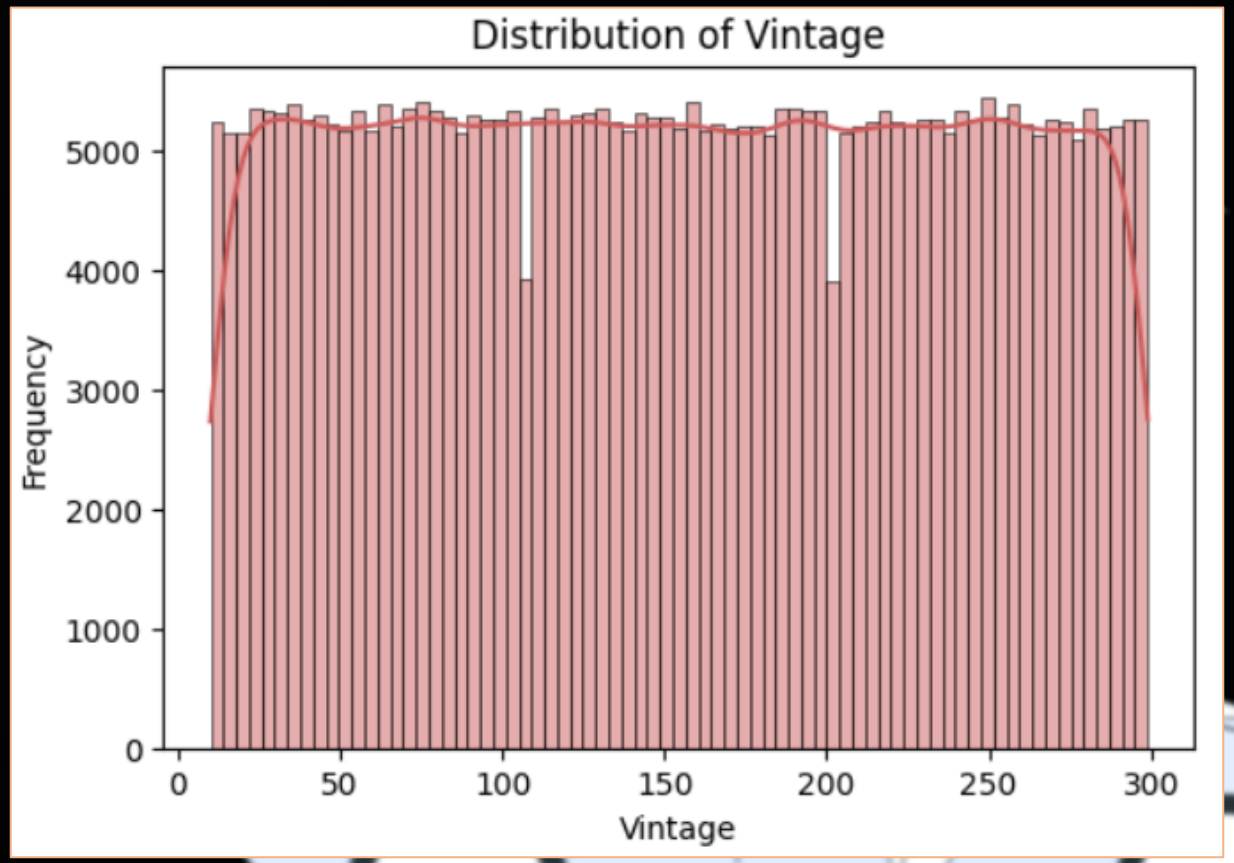
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Annual premium.

Key Insights

- ***The majority of policyholders have an annual premium that falls within a specific range.***
- ***The distribution of annual premium seems to be right-skewed, indicating a potential concentration of individuals with lower premium amounts and fewer with higher premium amounts.***
- ***The presence of a peak suggests a common annual premium amount or a particular segment of individuals paying around that amount.***

Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Vintage.



Data Visualization and Insights

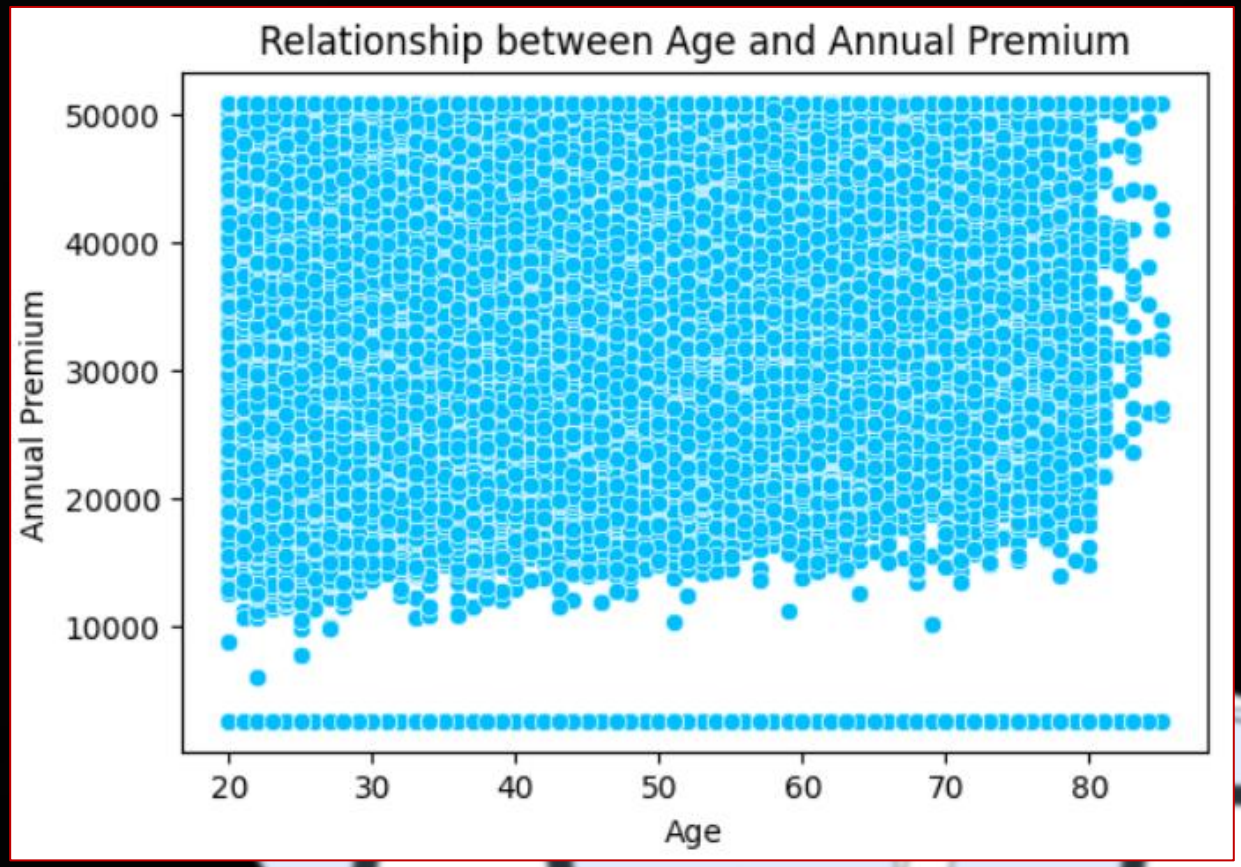
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Vintage.

Key Insights

- ***The vintage of policyholders is mostly concentrated towards the lower range, indicating that a significant number of policyholders have been associated with the company for a relatively shorter duration.***
- ***It seems to be a right-skewed distribution with a longer tail towards the higher vintage values. This could imply that fewer customers have been with the company for a longer period of time compared to those with shorter tenure.***
- ***The presence of a peak at the lower end of the vintage range reinforces the observation that most policyholders are relatively new customers.***

Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution between Age & Annual Premium.



Data Visualization and Insights

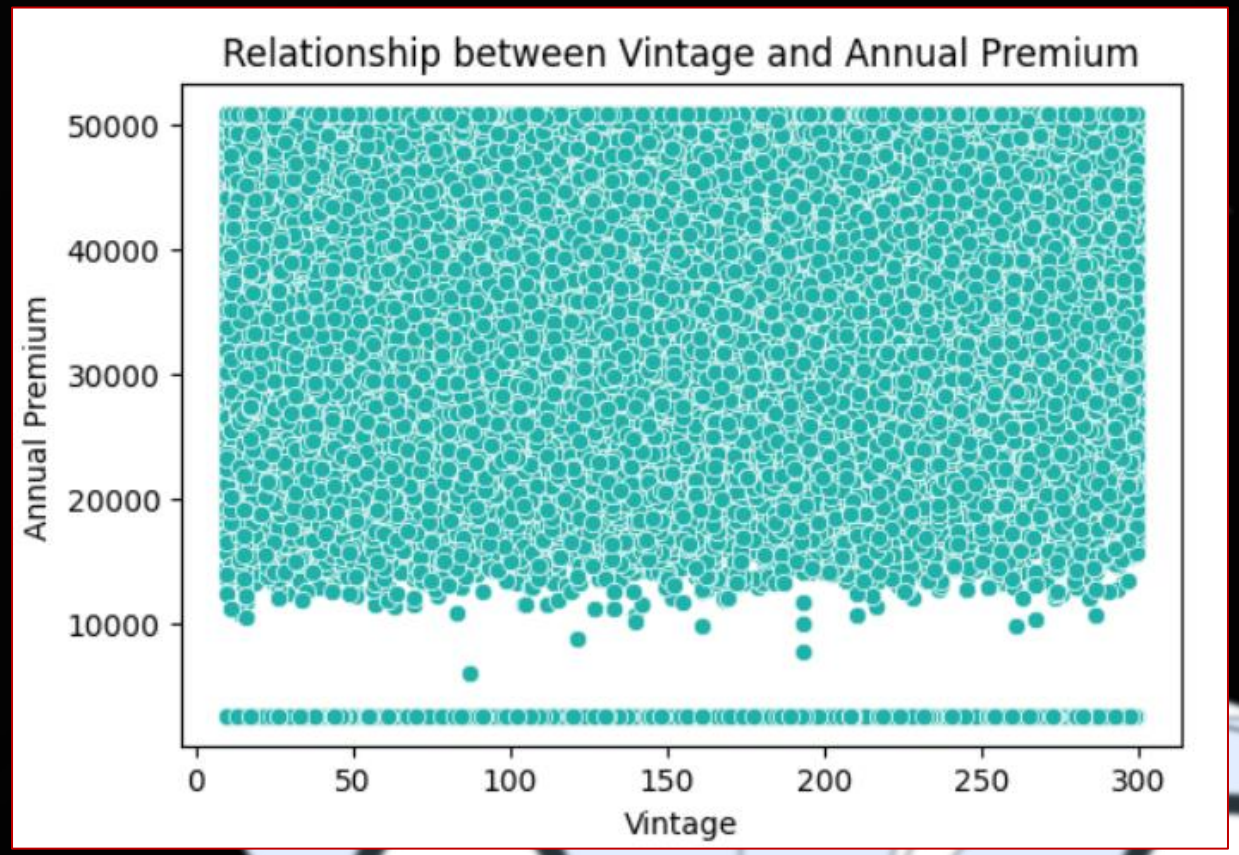
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution between Age & Annual Premium.

Key Insights

- *There's no clear linear relationship between the age of policyholders and their annual premium amount.*
- *We can see a large cluster of points within the lower range of annual premiums and various ages, suggesting that several policyholders have similar annual premiums irrespective of age.*
- *Though there's no strict correlation, we can observe some policyholders with higher annual premiums tend to be older. This could indicate that age might be a contributing factor to higher premium amounts.*
- *It's essential to consider other variables that influence premium amounts, such as vehicle type or driving history, for a more comprehensive understanding of the relationship between age and premium costs.*

Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution between Vintage & Annual Premium.



Data Visualization and Insights

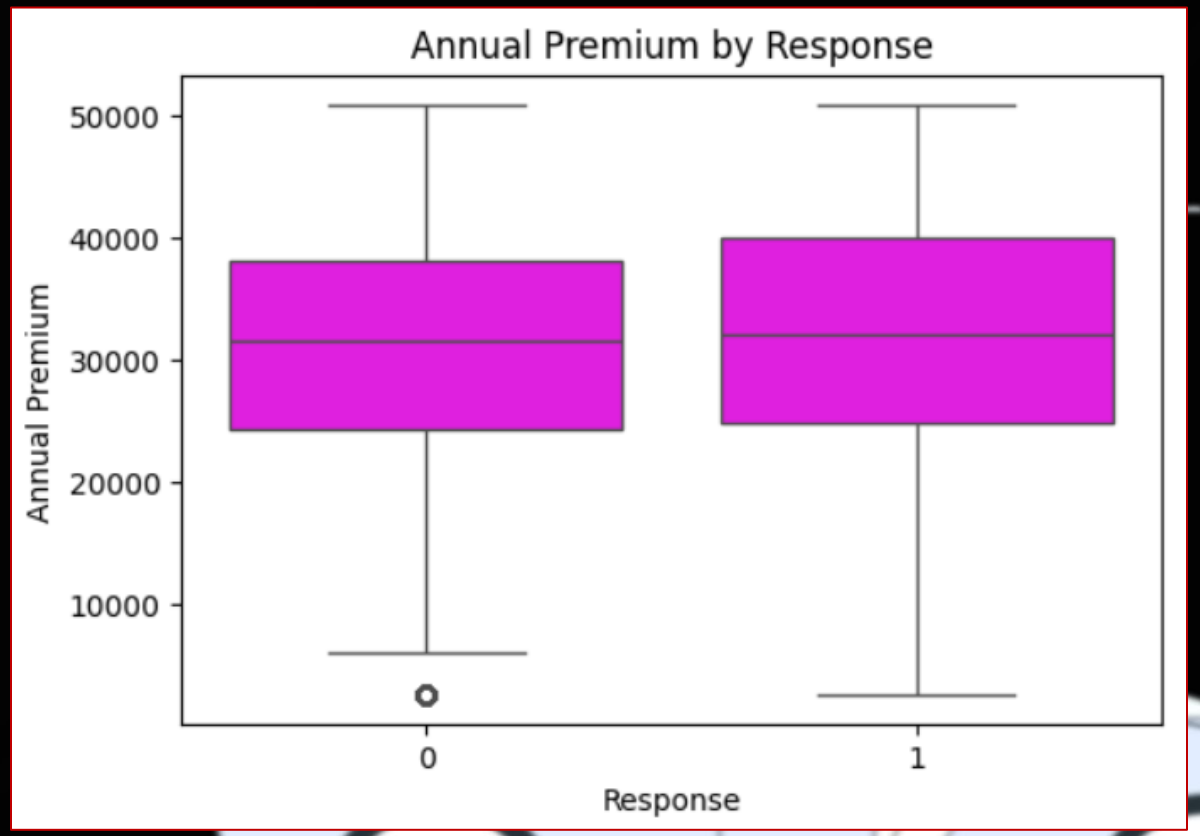
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution between Vintage & Annual Premium.

Key Insights

- *There is no clear linear relationship between the vintage of policyholders and their annual premium amount.*
- *We can see a large cluster of points within the lower range of annual premiums and various vintage, suggesting that several policyholders have similar annual premiums irrespective of the time they have been with the company.*
- *There doesn't seem to be a strong correlation between vintage and premium amount. This implies that the duration for which a customer has been with the company might not be a significant determinant of the premium they pay.*
- *The scatterplot indicates that the relationship is not linear or monotonic. It's possible that other factors, such as the type of policy or vehicle characteristics, play a more significant role in determining the premium amount than simply how long a customer has been with the company.*

Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution between Annual Premium & Response.



Data Visualization and Insights

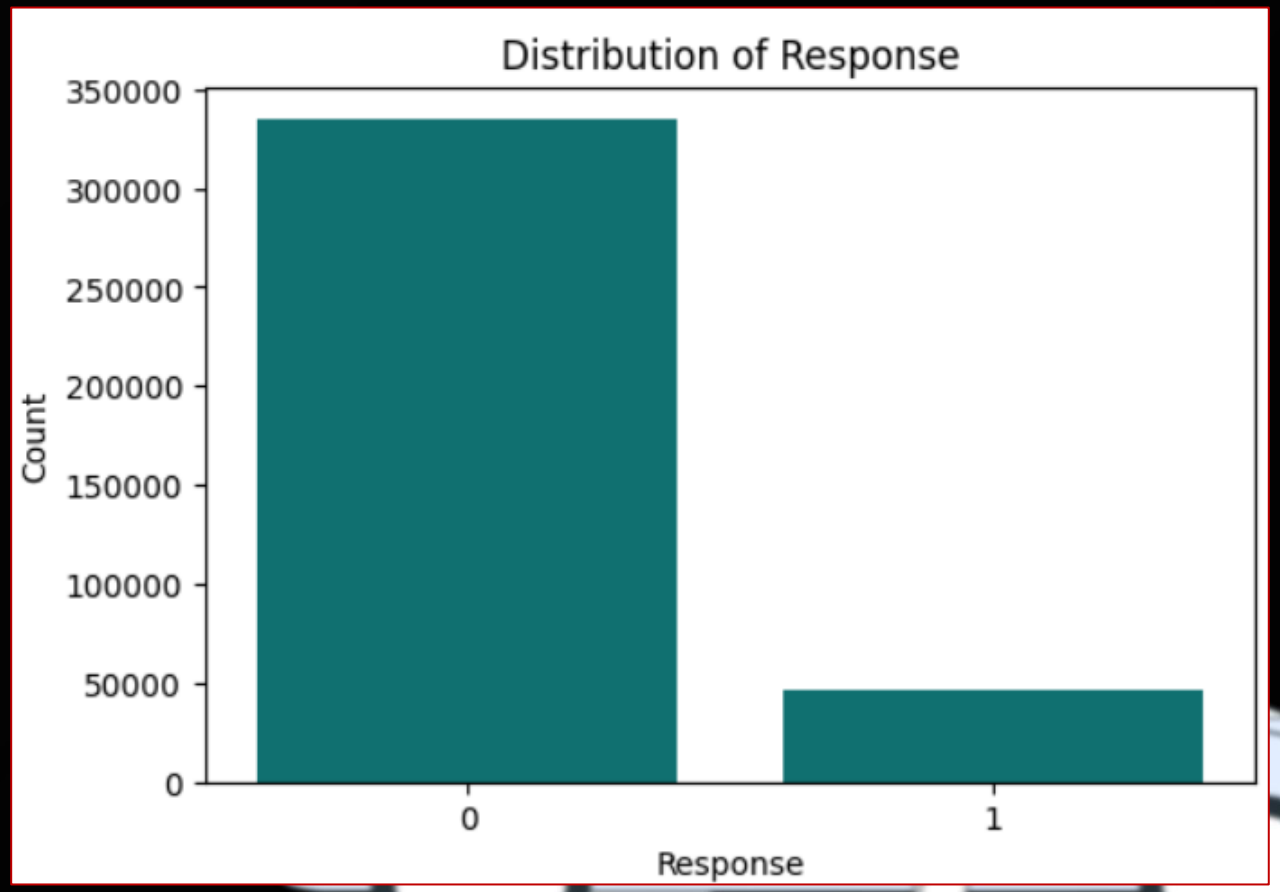
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution between Annual Premium & Response.

Key Insights

- ***The distribution of annual premium seems to be relatively similar for both response categories (0 and 1).***
- ***There's a slight tendency for policyholders who responded positively (Response = 1) to have a slightly higher median annual premium compared to those who didn't respond positively (Response = 0).***
- ***The interquartile ranges (IQR) are also comparable for both response categories, suggesting that the variability in annual premiums is similar for both groups.***
- ***However, the presence of outliers in both categories might influence the comparison of median annual premiums between response categories.***

Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Response.



Data Visualization and Insights

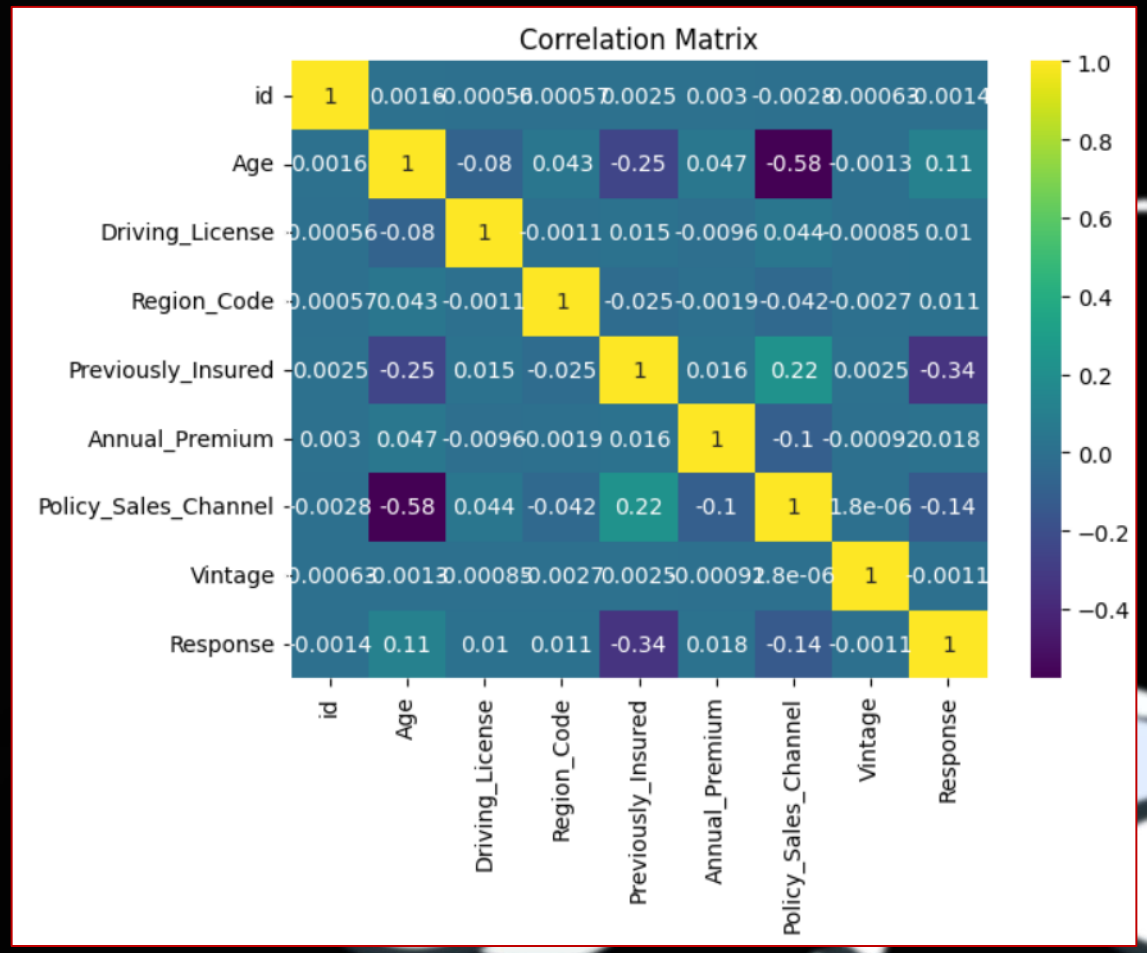
Utilize various visualization techniques to explore the distribution of key variables: Show the Distribution of Response.

Key Insights

- ***The majority of policyholders did not respond positively to the campaign.***
- ***The number of policyholders who responded positively is significantly lower than those who did not respond.***
- ***This indicates that the campaign may have been less effective in converting potential customers.***
- ***Further analysis is needed to understand the reasons for the low response rate and to improve the effectiveness of future campaigns.***

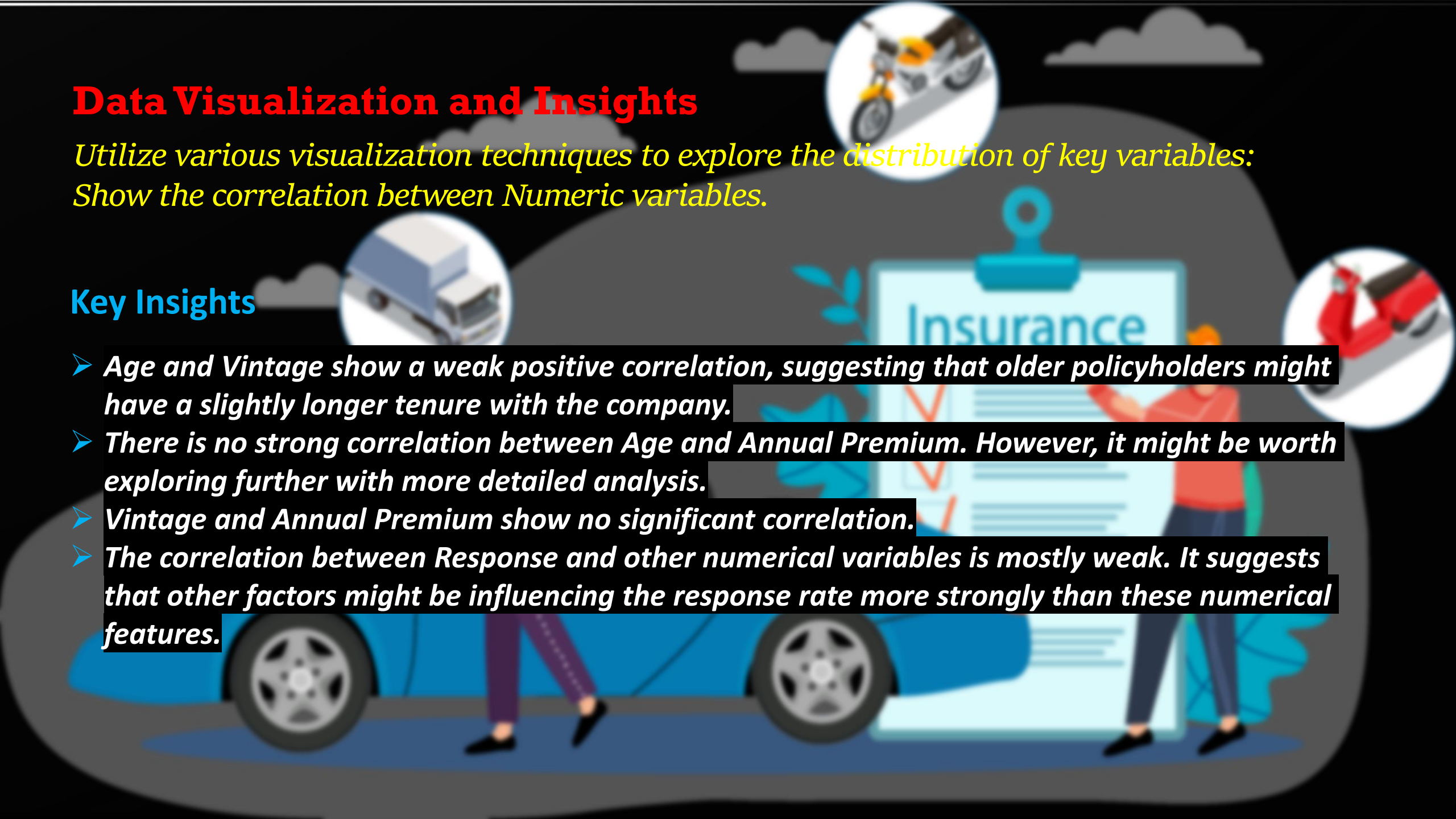
Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the correlation between Numeric variables.



Data Visualization and Insights

Utilize various visualization techniques to explore the distribution of key variables: Show the correlation between Numeric variables.

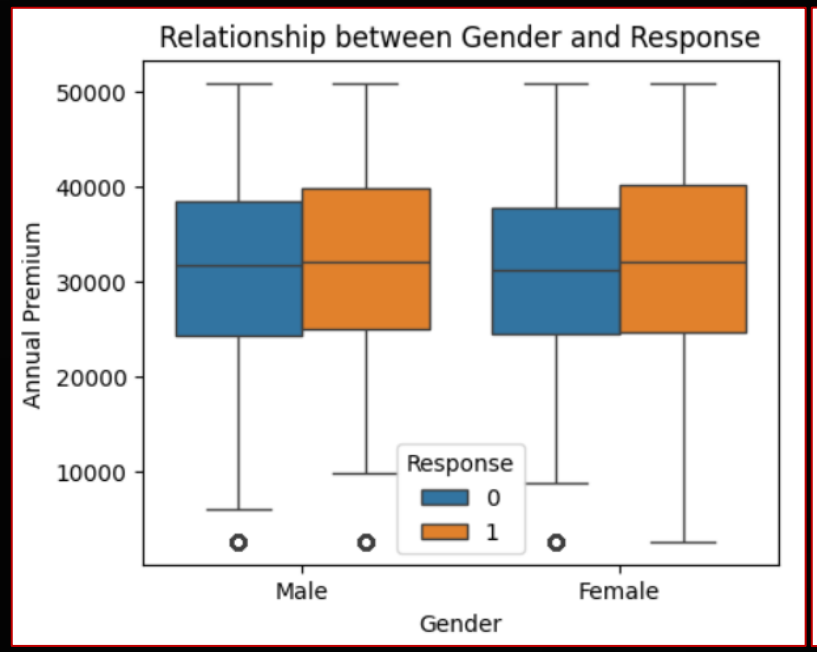


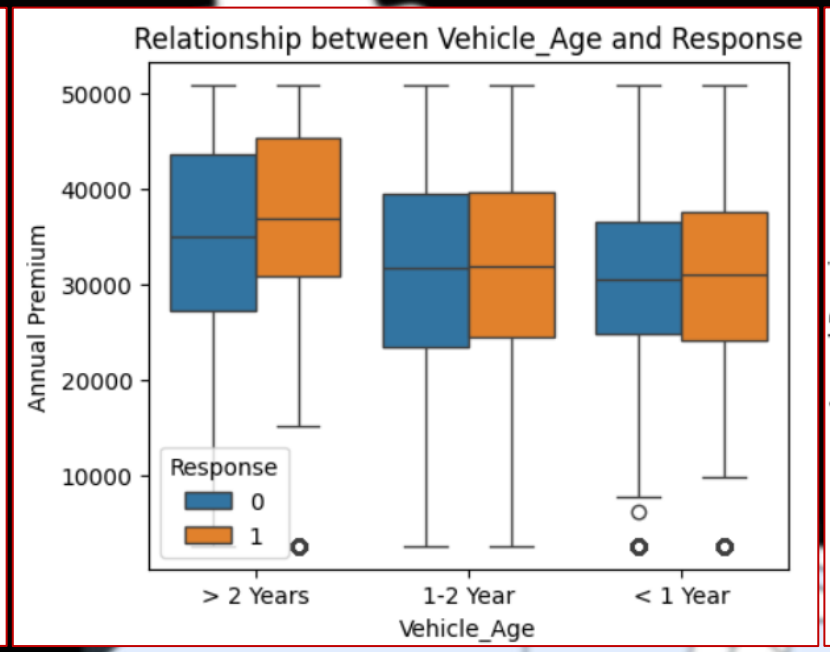
Key Insights

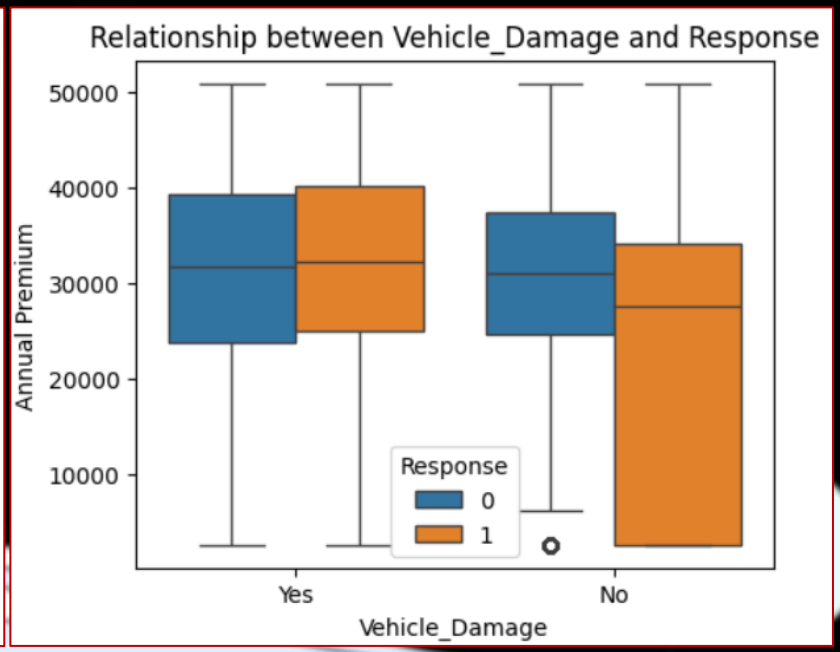
- ***Age and Vintage show a weak positive correlation, suggesting that older policyholders might have a slightly longer tenure with the company.***
- ***There is no strong correlation between Age and Annual Premium. However, it might be worth exploring further with more detailed analysis.***
- ***Vintage and Annual Premium show no significant correlation.***
- ***The correlation between Response and other numerical variables is mostly weak. It suggests that other factors might be influencing the response rate more strongly than these numerical features.***

Feature Analysis

Examine the relationship between Features and the Target variable (Insurance claims): Show the Distribution of Categorical features vs. Response.







Feature Analysis

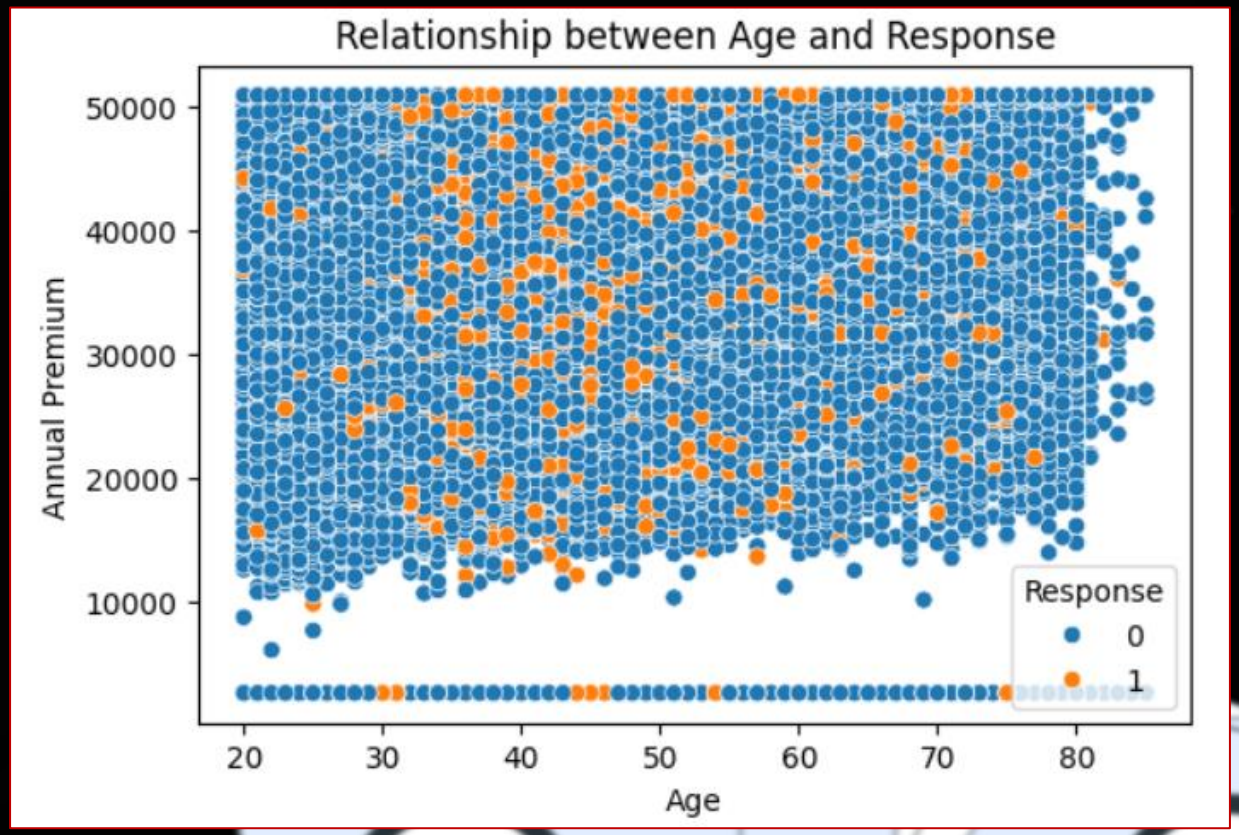
Examine the relationship between Features and the Target variable (Insurance claims): Show the Distribution of Categorical features vs. Response.

Key Insights

- ***If there's a difference in the median or distribution of premiums for specific vehicle age categories and their response, it might suggest that vehicle age influences the likelihood of a customer responding to the campaign.***
- ***If there's a difference in premium distributions between genders and their responses, it could imply a gender-based pattern in campaign responses.***
- ***If customers with a history of vehicle damage show a different pattern in responses, it could indicate that previous claims or damage might influence their likelihood of responding positively.***

Feature Analysis

Examine the relationship between Features and the Target variable (Insurance claims): Show the Distribution of Numerical features vs. Response.



Feature Analysis

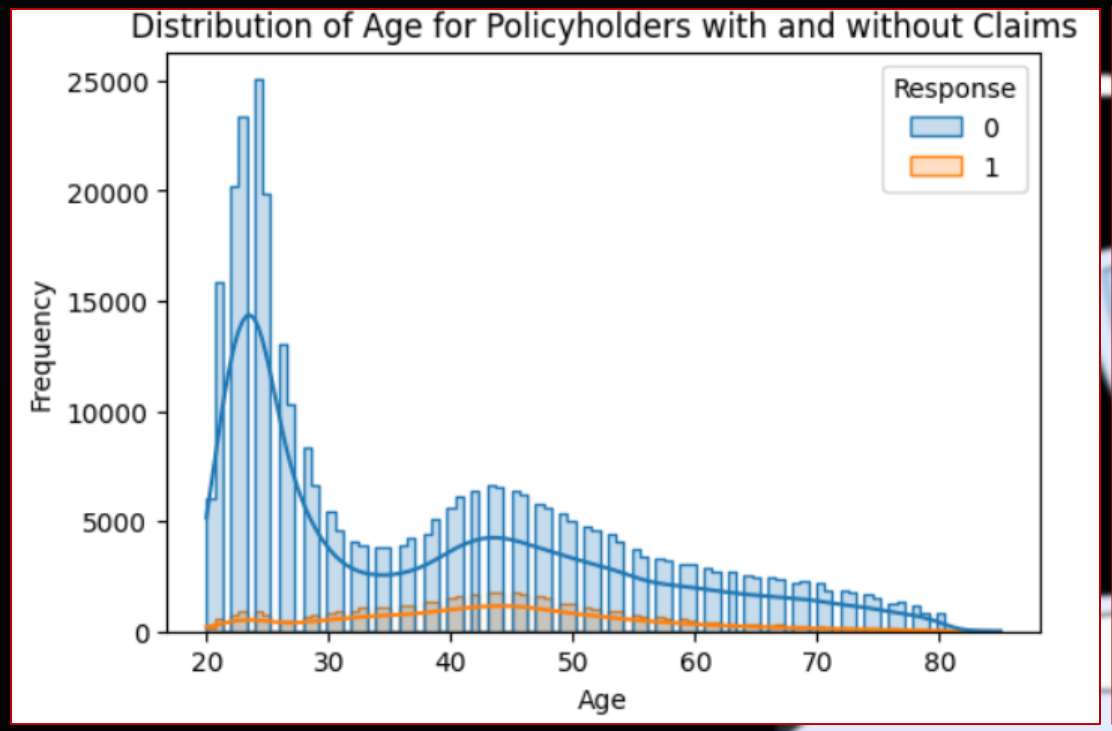
Examine the relationship between Features and the Target variable (Insurance claims): Show the Distribution of Numerical features vs. Response.

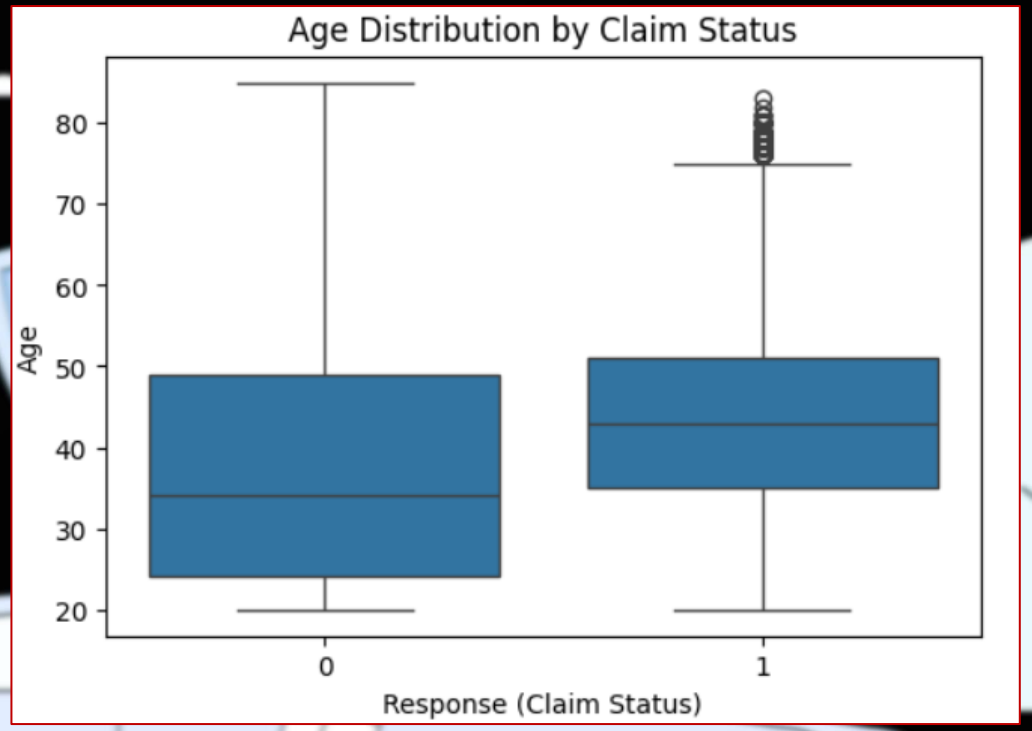
Key Insights

- ***The scatter plot visualizes the relationship between age, annual premium, and whether a customer responded positively to the campaign.***
- ***By examining the distribution of points for different response categories, we can understand if there's a difference in the relationship between age and annual premium for customers who responded positively versus those who did not.***
- ***For example, if we see a higher concentration of positive responses among a particular age group (e.g., younger individuals) with a certain range of annual premiums, it could suggest that the campaign is more effective for that specific segment of customers.***
- ***We might also observe if older customers with higher premiums are more or less likely to respond positively.***

Age Distribution

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





Age Distribution

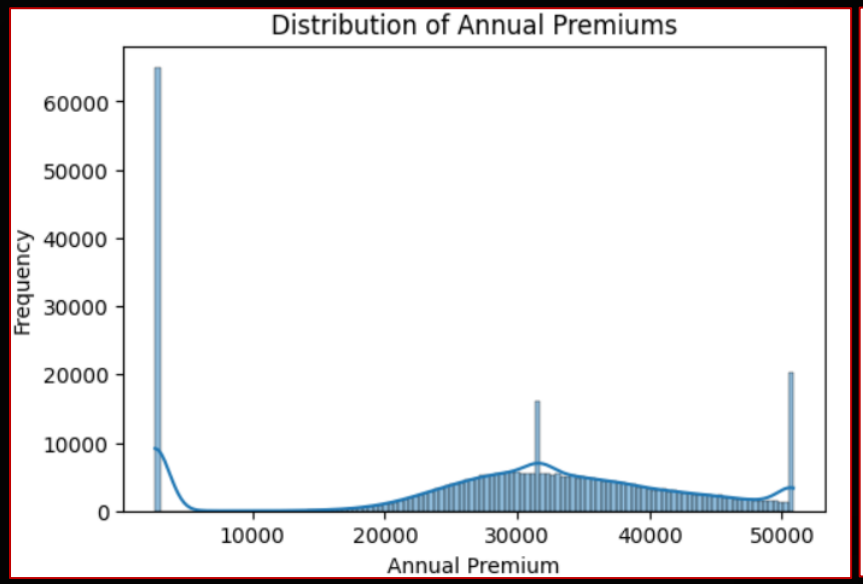
Analyze the age distribution within the dataset and its impact on insurance claims.

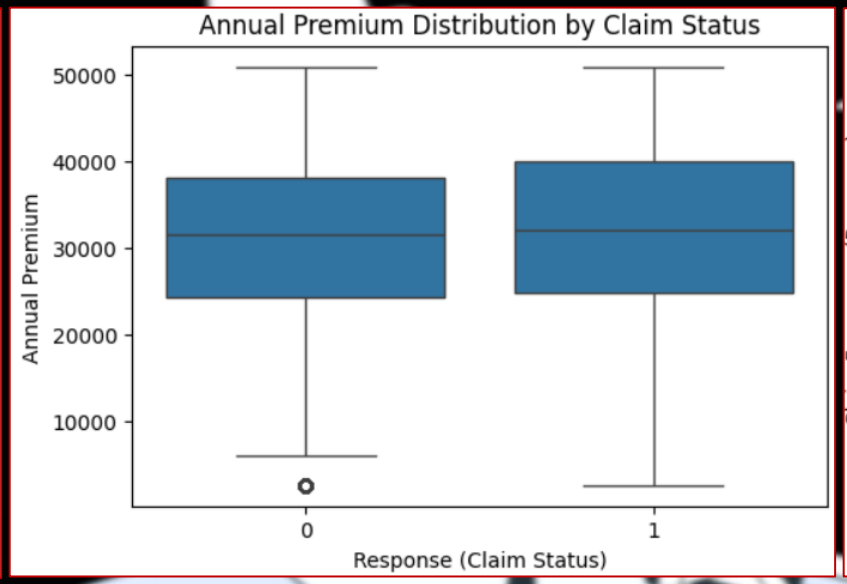
Key Insights

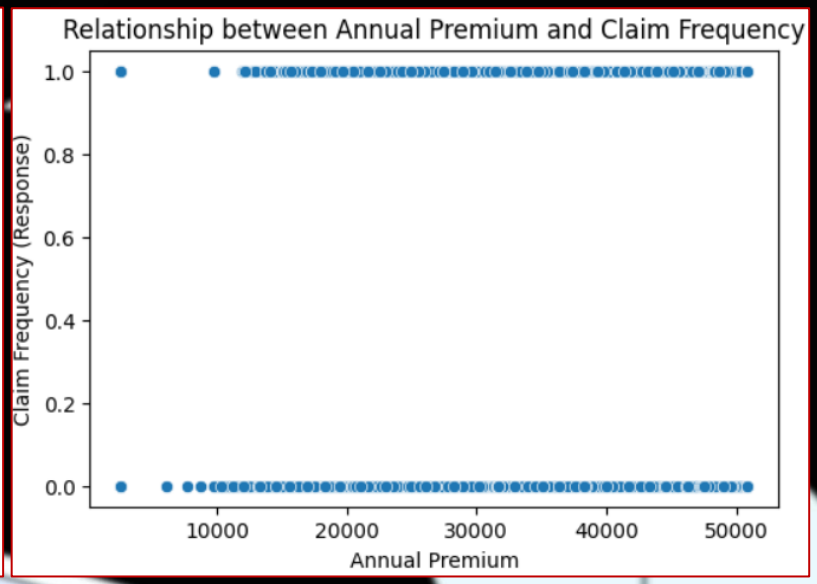
- *The histograms/box plots will show how age is distributed across policyholders with and without claims.*
- *If there's a significant difference in the mean or median age between the two groups, it suggests a potential relationship between age and claim likelihood.*
- *If certain age ranges show a higher proportion of claims, it indicates a specific age group is more prone to claiming insurance.*
- *The analysis can help identify age-related risk factors and inform strategies for targeted risk assessment and pricing.*

Premium Analysis

Investigate the distribution of insurance premiums and their correlation with claim frequencies.







Premium Analysis

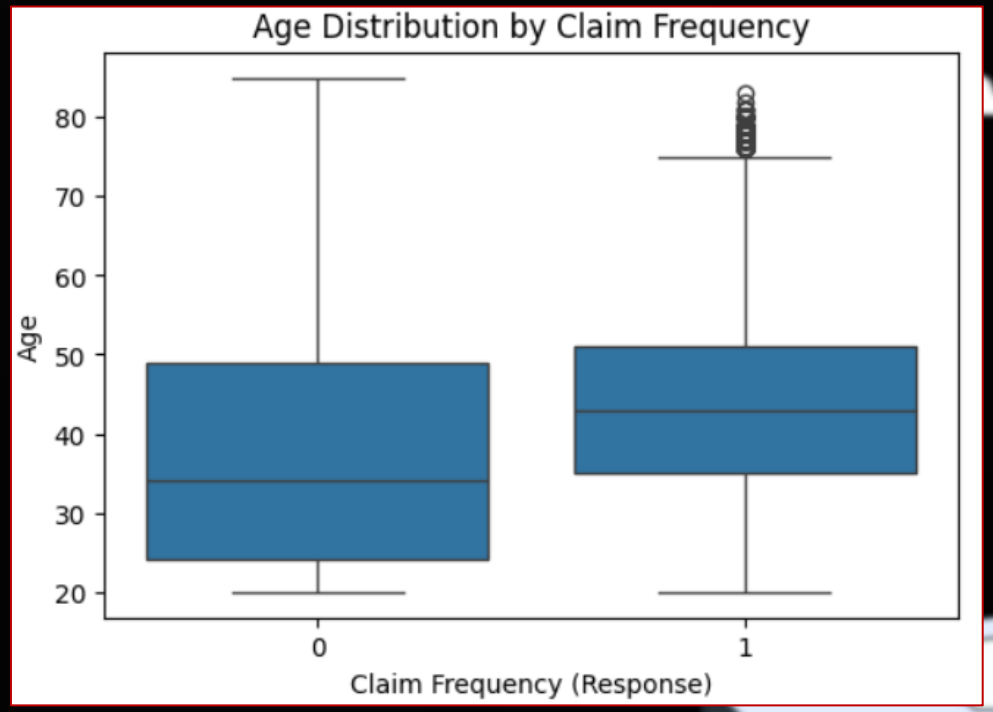
Investigate the distribution of insurance premiums and their correlation with claim frequencies.

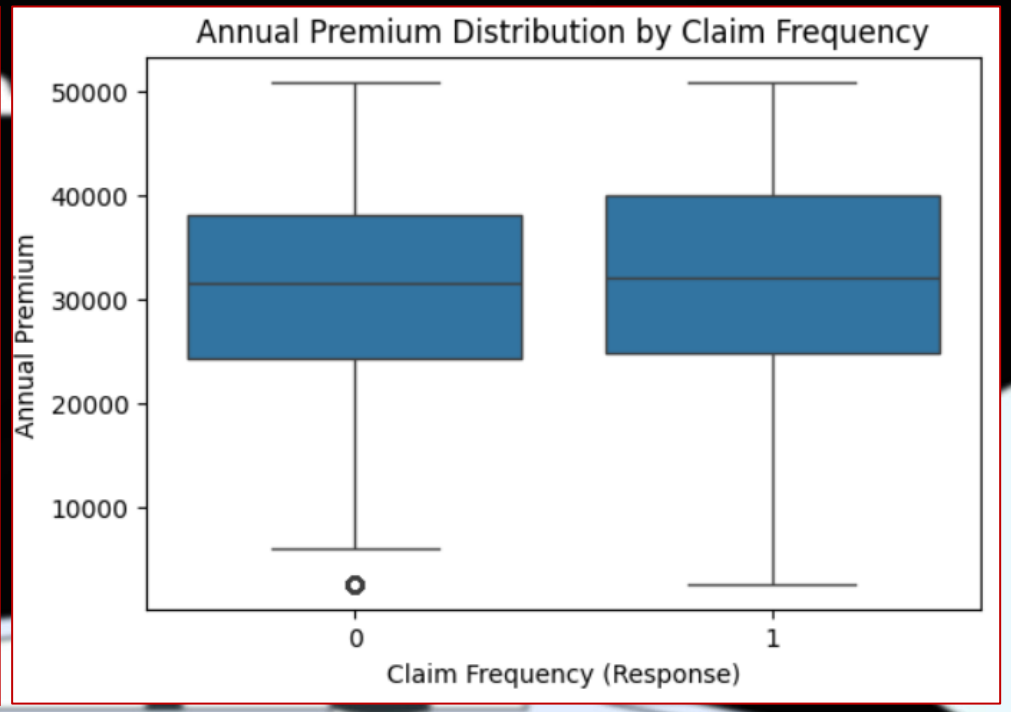
Key Insights

- ***The histogram will show the overall distribution of annual premiums within the dataset.***
- ***The box plots will visualize the distribution of premiums for policyholders with and without claims (Response = 1 or 0).***
- ***The scatter plot will reveal the relationship between annual premiums and the likelihood of claiming insurance.***
- ***Comparing the mean annual premiums for policyholders with and without claims can help identify potential differences in premium levels between these two groups.***
- ***If the mean premium for those with claims is significantly higher or lower than those without claims, it could suggest that premium level plays a role in claim frequency.***

Claim Frequencies

Explore factors contributing to higher claim frequencies.





Claim Frequencies

Explore factors contributing to higher claim frequencies.

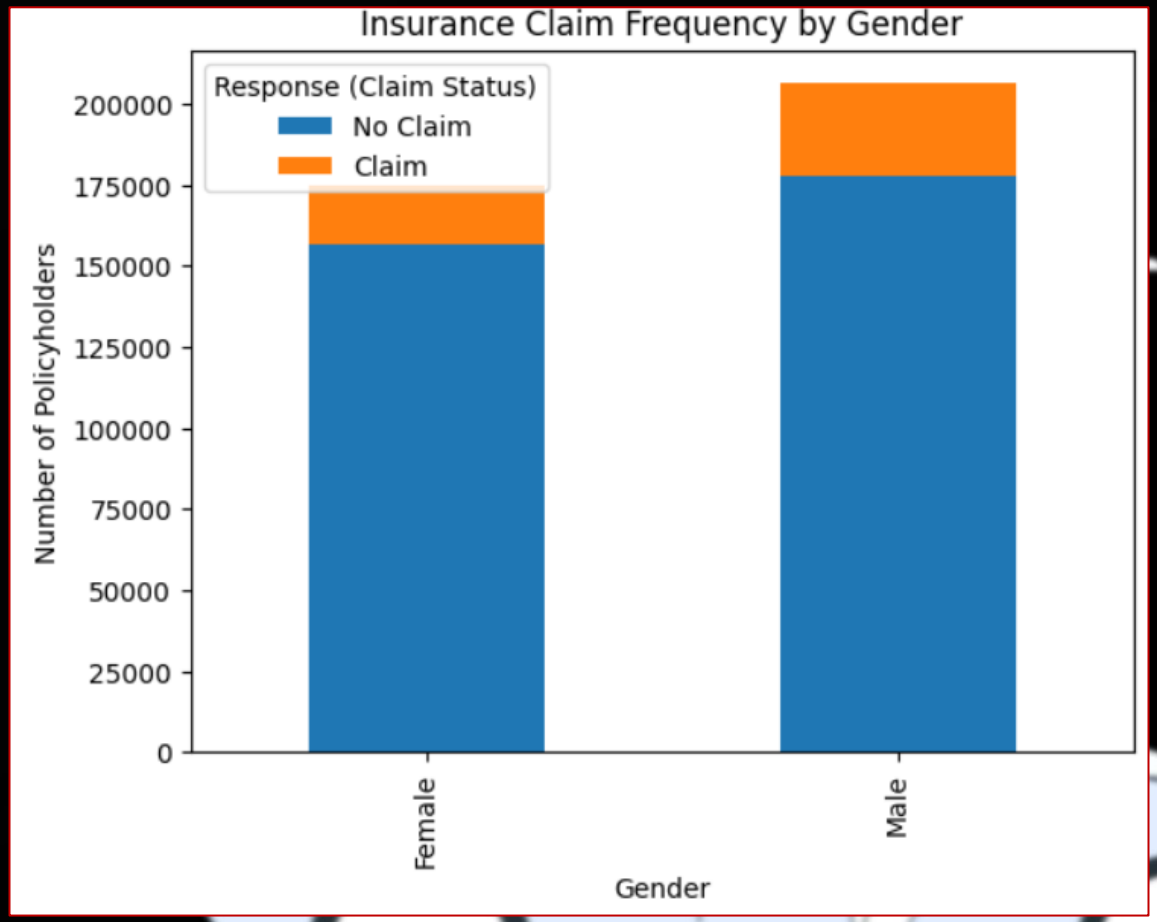
Key Insights

- ***Age and Claim Frequency:***
  - ***The box plot comparing 'Age' and 'Response' (Claim Frequency) can reveal potential relationships between age and the likelihood of claiming insurance.***
  - ***If the median age of policyholders who filed claims is significantly different from those who didn't, it indicates that age might be a factor influencing claim frequency.***
  - ***We might observe that a specific age range (e.g., younger or older drivers) is associated with a higher likelihood of claims.***

- ***Annual Premium and Claim Frequency:***
  - ***The box plot comparing 'Annual Premium' and 'Response' can provide insights into whether premium levels are linked to claim frequency.***
  - ***If the median annual premium for policyholders who filed claims is higher or lower than those who didn't, it suggests that premium amounts might be a contributing factor to claims.***

Gender Analysis

Investigate the role of gender in insurance claims.



Gender Analysis

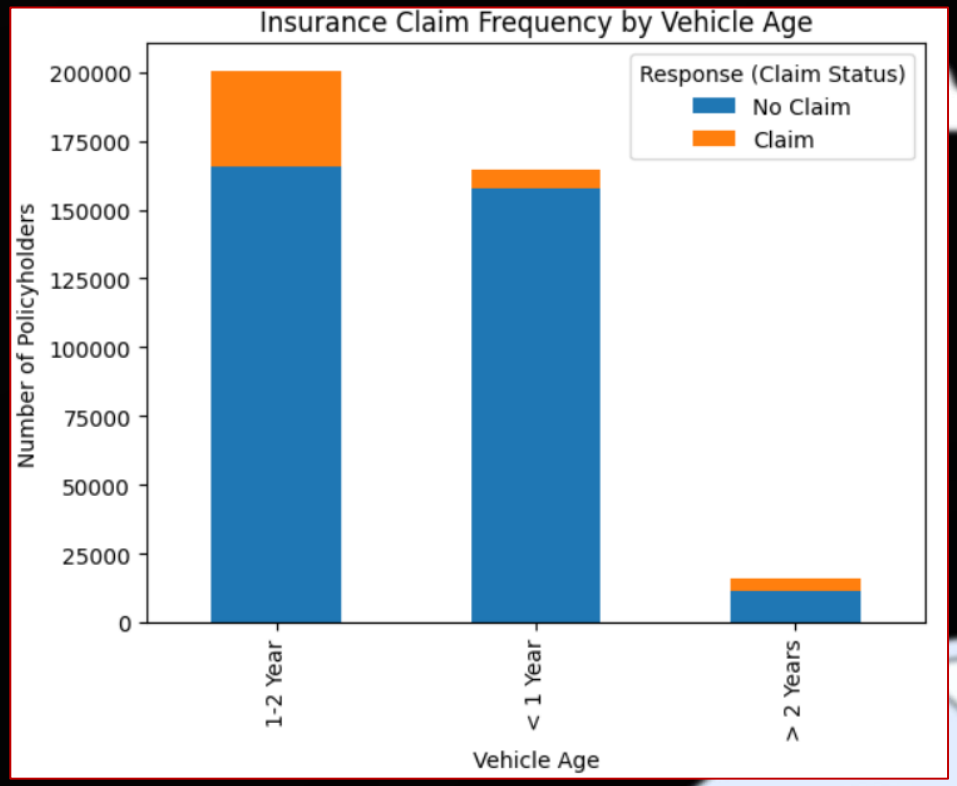
Investigate the role of gender in insurance claims.

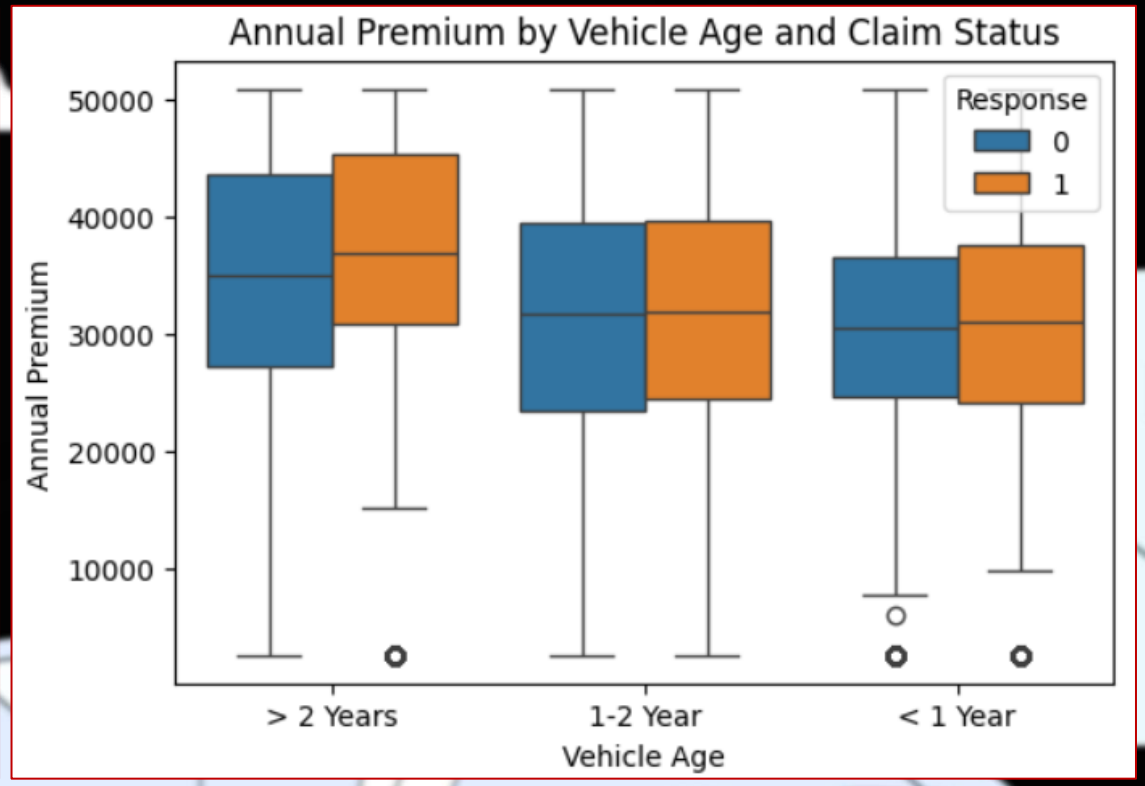
Key Insights

- ***By analyzing the relationship between 'Gender' and 'Response' (claim status), we can understand if there's a difference in claim frequency between male and female policyholders.***
- ***The crosstabulation and bar chart can reveal the proportion of claims filed by each gender.***
- ***If one gender has a significantly higher claim rate than the other, it might suggest that gender plays a role in risk assessment and claim likelihood.***
- ***For example, if men have a higher claim rate than women, it could indicate that male drivers might be involved in more accidents or have different driving behaviors.***

Vehicle Age & Claims

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





Vehicle Age & Claims

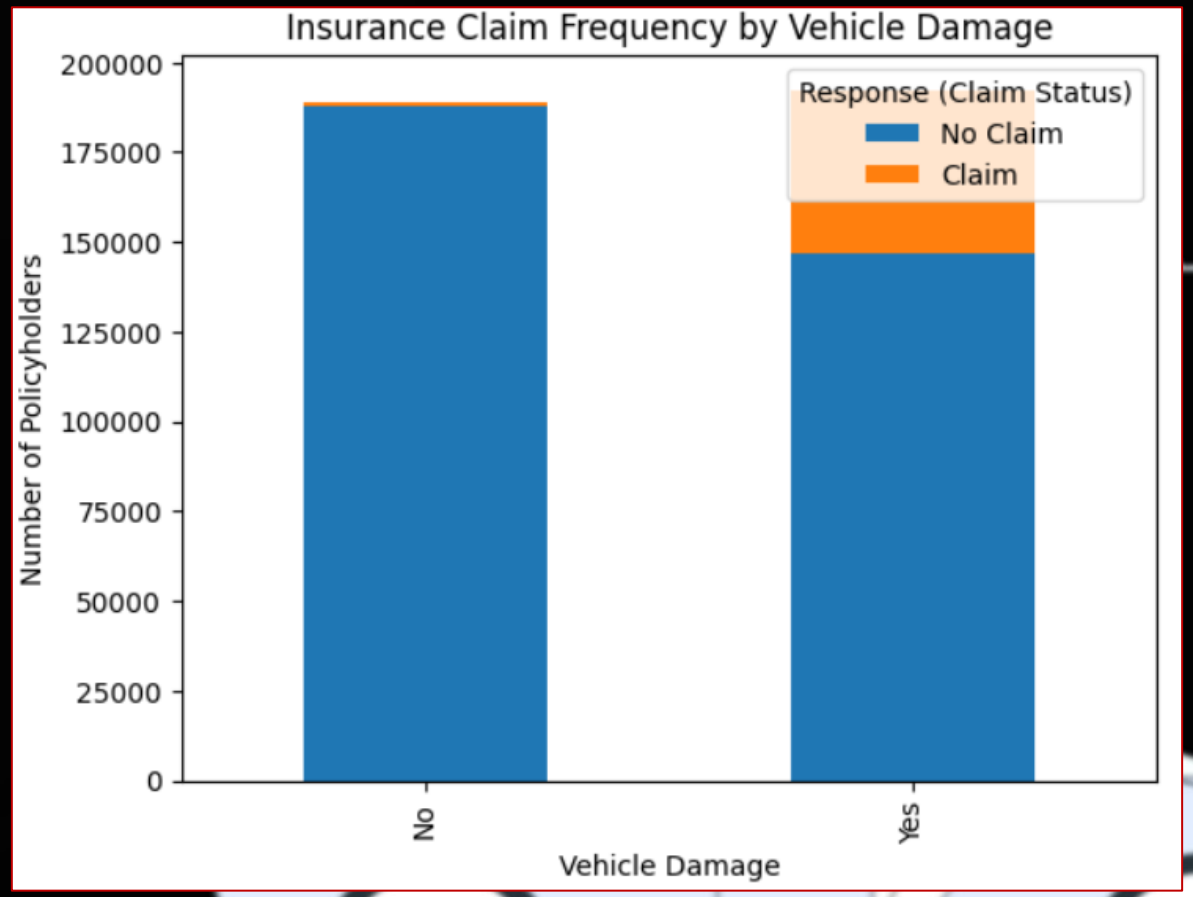
Examine the impact of vehicle age on the likelihood of a claim.

Key Insights

- ***By analyzing the relationship between 'Vehicle Age' and 'Response' (claim status), we can understand if there's a correlation between the age of a vehicle and the likelihood of filing a claim.***
- ***The crosstabulation and bar chart can reveal the proportion of claims filed for different vehicle age categories (e.g., < 1 Year, 1-2 Year, > 2 Years).***
- ***If older vehicles have a significantly higher claim rate compared to newer vehicles, it might suggest that older vehicles are more prone to mechanical issues or accidents.***

Claim frequency by vehicle damage

Investigate the relationship between vehicle damage and claim frequencies.



Claim frequency by vehicle damage

Investigate the relationship between vehicle damage and claim frequencies.

Key Insights

- ***By analyzing the relationship between 'Vehicle Damage' and 'Response' (claim status), we can understand if there's a correlation between having a history of vehicle damage and the likelihood of filing a claim.***
- ***The crosstabulation and bar chart can reveal the proportion of claims filed by policyholders who have experienced vehicle damage versus those who haven't.***
- ***If policyholders with a history of vehicle damage have a significantly higher claim rate, it suggests that prior damage might be an indicator of increased risk and a higher likelihood of future claims.***

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis:

KEY FINDINGS:

✓ Claim Frequency:

- *Age, Annual Premium, Vehicle Damage, and Vehicle Age are potential factors influencing claim frequencies.*
- *Policyholders with a history of vehicle damage are more likely to file claims.*
- *There might be slight differences in claim frequencies between genders.*
- *Older vehicles tend to have higher claim rates compared to newer ones.*

✓ Gender:

- *There appears to be a slight difference in claim frequency between genders, but it's not a major factor.*

✓ Vehicle Age:

- *Older vehicles have a significantly higher likelihood of claims than newer vehicles, likely due to increased wear and tear or safety concerns.*

✓ Vehicle Damage:

- *A history of vehicle damage is strongly associated with an increased risk of future claims.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis:

Conclusions:

- *Insurance claim frequency is influenced by a combination of factors, including age, annual premium, vehicle age, and vehicle damage history.*
- *Vehicle age and vehicle damage history appear to be the most prominent factors influencing claim frequency.*
- *While there might be slight differences in claim frequency between genders, it's not a major determining factor compared to other variables like vehicle age or damage history.*
- *These findings highlight the importance of considering these factors in risk assessment and pricing strategies.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis:

Recommendations:

✓ Risk Assessment and Pricing:

- *Develop more sophisticated risk assessment models that incorporate factors like vehicle age, vehicle damage history, and driver age.*
- *Adjust premium calculations based on these risk factors to reflect the likelihood of claims more accurately.*
- *Consider implementing tiered pricing structures that differentiate premiums based on vehicle age and damage history.*

✓ Targeted Interventions:

- *Offer targeted driver education programs and safety awareness campaigns to reduce accidents and claims, especially for drivers with a history of vehicle damage or those with older vehicles.*
- *Promote regular vehicle maintenance and safety inspections, especially for older vehicles, to minimize the risk of breakdowns and accidents.*
- *Provide incentives for policyholders to replace older vehicles with newer, safer models.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis:

Recommendations:

✓ Claims Management:

- *Implement strategies to enhance claims processing efficiency, especially for claims related to vehicles with prior damage.*
- *Consider developing dedicated fraud detection systems that monitor claims involving vehicles with a history of damage.*
- *Implement robust data analysis procedures to identify patterns and trends in claims data and improve risk prediction models continuously.*

✓ Continuous Monitoring and Improvement:

- *Regularly monitor claim trends and patterns to identify emerging risks and adjust pricing strategies and intervention programs accordingly.*
- *Analyze data on accident types, claim costs, and driver behaviors to continually refine risk assessment models and predictive analytics.*

FINAL REPORT

Summarizing Key findings, drawing conclusions & providing recommendations based on the insights gained from the analysis:

Recommendations:

✓ Ethical Considerations:

- *Ensure that pricing and risk assessment strategies are implemented ethically and avoid discrimination based on gender or other protected characteristics.*
- *Maintain transparency in pricing and risk assessment methodologies to build trust with policyholders.*

❖ *By following these recommendations, insurance companies can improve their risk management practices, develop more accurate pricing models, and implement targeted interventions to reduce claim frequency.*

❖ *This can ultimately lead to a more efficient and sustainable insurance system that benefits both policyholders and insurers.*

THANK YOU FOR READING

For coding part, kindly refer to below link:-

https://github.com/SaxenaKushagr/Vehicle-Insurance-Analysis