Vehicle Insurance

Exploratory Data Analysis (EDA)

Skill
Circle
Dataset

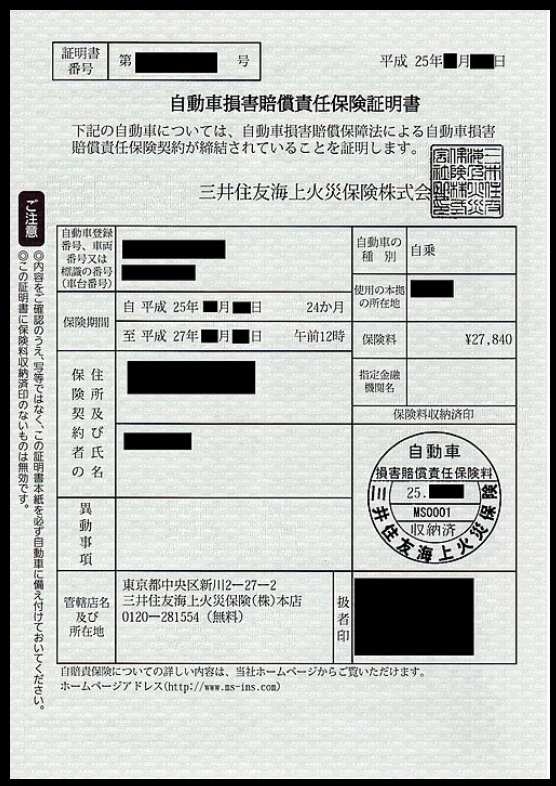
- By Akshay Gujare

Mob:7741841945
Gmail: 77418anslyst@gmsil.com

About Vehicle Insurance

Vehicle insurance (also known as car insurance, motor insurance, or auto insurance) is insurance for cars, trucks, motorcycles, and other road vehicles. Its primary use is to provide financial protection against physical damage or bodily injury resulting from traffic collisions and against liability that could also arise from incidents in a vehicle. Vehicle insurance may additionally offer financial protection against theft of the vehicle, and a

Widespread use of the motor car began after the First World War in urban areas. Cars were relatively fast and dangerous by that stage, yet there was still no compulsory form of car insurance anywhere in the world. This meant that injured victims would rarely get any compensation in a crash, and drivers often faced considerable costs for damage to their car and property. A compulsory car insurance scheme was introduced in the United Kingdom with the Road Traffic Act 1930. This ensured that all vehicle owners and drivers had to be insured for their liability for injury or death to third parties while their vehicle was being used on a public road.Ireland replicated the obligation via the Road Traffic Act, 1933.Germany enacted similar



証明書番号 第 [redacted] 号　　平成 25年[redacted]月[redacted]日

自動車損害賠償責任保険証明書

下記の自動車については、自動車損害賠償保障法による自動車損害賠償責任保険契約が締結されていることを証明します。

三井住友海上火災保険株式会社

ご注意
◎内容をご確認のうえ、写等ではなく、この証明書本紙を必ず自動車に備え付けておいてください。
◎この証明書に保険料収納済印のないものは無効です。

自動車登録番号、車両番号又は標識の番号（車台番号）	[redacted]	自動車の種別	自乗
保険期間	自 平成 25年[redacted]月[redacted]日　24か月	使用の本拠の所在地	[redacted]
	至 平成 27年[redacted]月[redacted]日　午前12時	保険料	¥27,840
保険契約者の 住所及び氏名	[redacted]	指定金融機関名	
		保険料収納済印	自動車 損害賠償責任保険料 25.[redacted] MS0001 収納済 三井住友海上火災
異動事項			
官幅店名及び所在地	東京都中央区新川2−27−2 三井住友海上火災保険(株)本店 0120−281554（無料）	扱者印	[redacted]

自賠責保険についての詳しい内容は、当社ホームページからご覧いただけます。
ホームページアドレス(http://www.ms-ins.com)

A Japanese vehicle insurance policy issued by the Mitsui Sumitomo Insurance company.

legislation in 1939 called the "Act on the Implementation of Compulsory Insurance for Motor Vehicle Owners".The EU (then EEC) required mandatory insurance cover be mandated by all member states, from 1973.

Insurance Claim Process for Accident

Process:

1. The collision occurs.

2. You gather information and evidence from the scene.

3. The police take information.

4. Inform your insurance company.

5. This is a good time to contact a lawyer.

6. A claims adjuster inspects your car's damage.

7. Your insurance authorizes your claim.

8. A body shop repairs your car.

9. You keep the records of your settlement.

Types of Insurance

1. LIABILITY

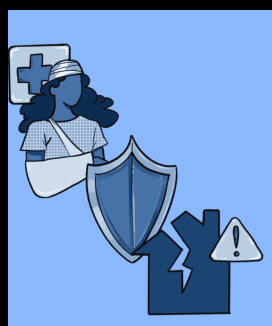
Protects you from property and injury costs to the **other party**. Required in 49 states.

2. COLLISION

Covers the cost of damage to **your own car** in an accident. Required if you have a loan or lease.

3. COMPREHENSIVE

Covers **non-collision damage** like theft or natural disaster. Lenders may also require this.



Liability Insurance

[ˌlī-ə-ˈbi-lə-tē in-ˈshur-ən(t)s]

An insurance product that provides protection against claims resulting from injuries and damage to other people or property.

Comprehensive Insurance

[ˌkäm-pri-ˈhen(t)-siv in-ˈshur-ən(t)s]

Automobile insurance that covers damage to your car from causes other than a collision.

Collision Damage Waiver (CDW)

[kə-ˈli-zhən ˈda-mij ˈwā-vər]

Additional insurance coverage offered to an individual renting an automobile.

Objective

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

In simple terms:

1.Data Preprocessing: Cleaning and preparing the raw data by handling missing values, correcting data types, and removing inconsistencies.

2.Visualization: Using charts and graphs to explore relationships between variables such as vehicle type, age, claim amount, or accident history.

3.Statistical Analysis: Applying descriptive and inferential statistics to identify trends, correlations, and insights that help understand what influences insurance claims.

Overall, the project hepls to learn how to analyze real-world data, discover meaningful insights, and support data-driven decision-making in the vehicle insurance sector.

TOOLS

1. Google Colab
Full form: Google Collaboratory
Use: A cloud-based platform to write and run Python code in your browser.
Features:
No installation needed.
Free access to CPUs, GPUs, and TPUs.
Ideal for data analysis, machine learning, and Python practice.

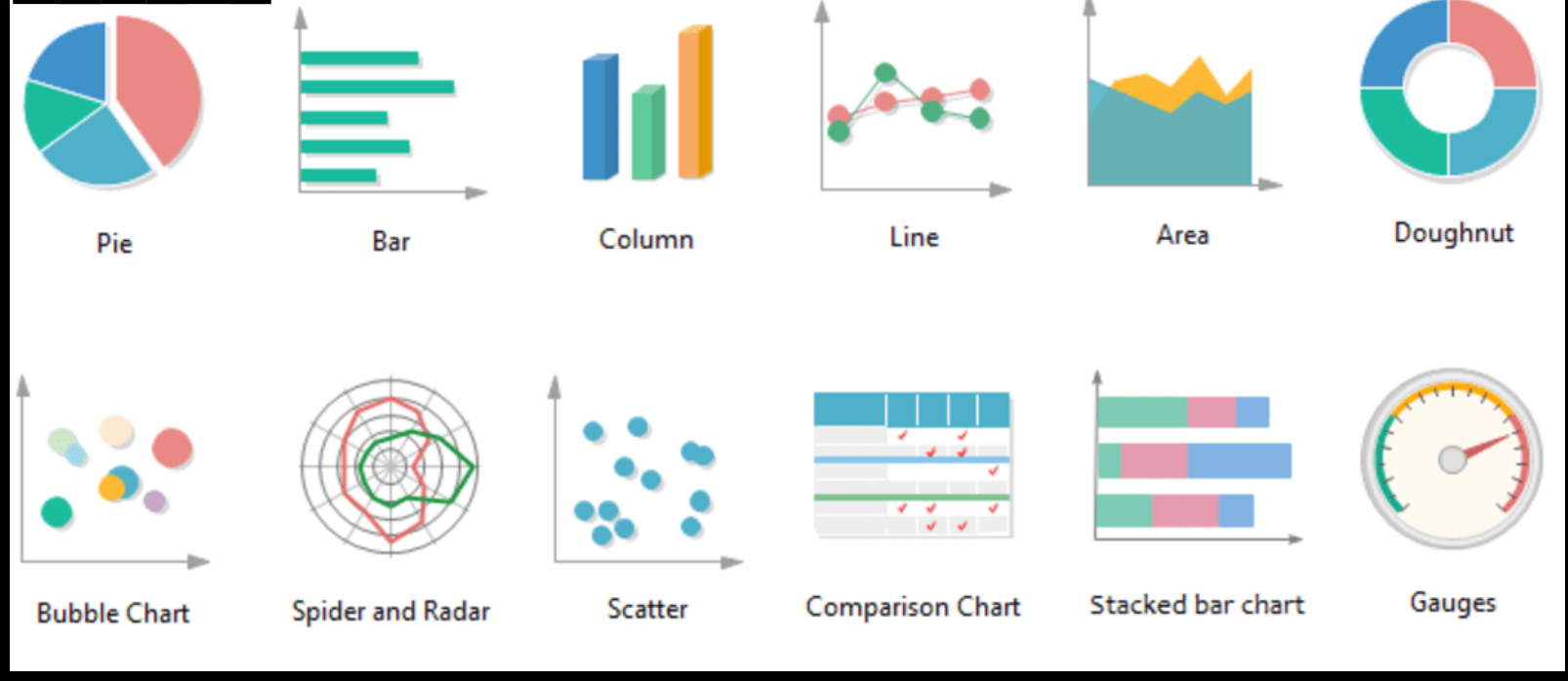
2. Python
Use: A high-level programming language widely used for data analysis, automation, and AI.
Features:
Easy to learn and read.
Huge library support for data, web, and AI.
Backbone for libraries like Pandas, NumPy, Matplotlib, and Seaborn.

3.Pandas
Use: A Python library for data manipulation and analysis.
Features:
Provides DataFrame and Series structures to handle tabular data.
Used for data cleaning, transformation, filtering, and summarization.
Example: reading Excel/CSV files and analyzing data efficiently.

4. Matplotlib
Use: A data visualization library in Python.
Features:
Used to create basic plots like line, bar, pie, histogram, scatter, etc.
Highly customizable and integrates well with Pandas and NumPy.
Example: plt.plot() for line graphs.

5. Seaborn
Use: An advanced data visualization library built on top of Matplotlib.
Features:
Provides beautiful and statistical plots with fewer lines of code.
Supports heatmaps, pair plots, violin plots, etc.
Works seamlessly with Pandas DataFrames.

GRAPHS

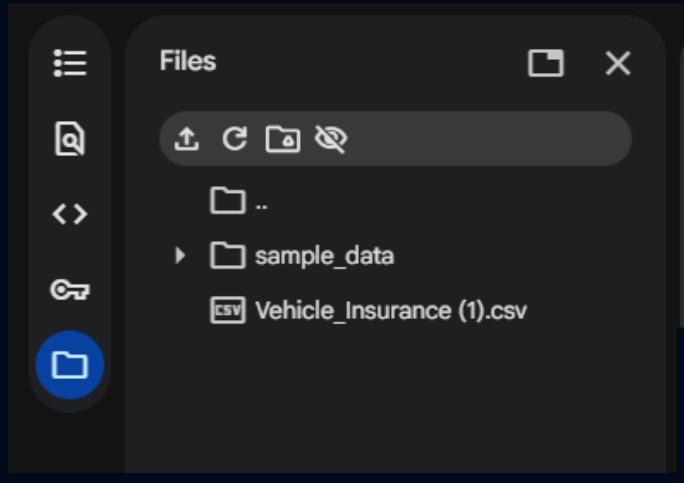


Data overview

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

Data loading and initial structure inception

Data Imported & Check Shape (Dimensions)




```
[3]
✓ 1s
1 import pandas as pd
2 import numpy as np
3 import matplotlib.pyplot as plt
4 import seaborn as sns

[4]
✓ 0s
1 df=pd.read_csv("/content/Vehicle_Insurance (1).csv")
```

Outcome:

The pandas,numpy,Matplotlib & seaborn libraries were imported, and the datasets named Vehicle_Insurance.csv was load into the Dataframe named "df"

Check Shape of Dataset

```
[ ]
1 df.shape

(56842, 12)
```

Outcome:

Dataset contain 56842 rows and 12 columns

Check Columns Name

```
[ ]  1 df.columns
```

```
Index(['id', 'Gender', 'Age', 'Driving_License', 'Region_Code',
       'Previously_Insured', 'Vehicle_Age', 'Vehicle_Damage', 'Annual_Premium',
       'Policy_Sales_Channel', 'Vintage', 'Response'],
      dtype='object')
```

Outcome:

Using df.columns we found that there are 13 columns in our given dataset.

View Sample Data

```
[ ]  1 df.head()
```

	id	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Age	Vehicle_Damage	Annual_Premi
0	1	Male	44	1	28.0	0	> 2 Years	Yes	4045
1	2	Male	76	1	3.0	0	1-2 Year	No	3353
2	3	Male	47	1	28.0	0	> 2 Years	Yes	3829
3	4	Male	21	1	11.0	1	< 1 Year	No	2861
4	5	Female	29	1	41.0	1	< 1 Year	No	2749

Outcome:

Here we can be able to see few starting rows of the dataset using df.head method

Data Summarization and Missing Values Identification

Check Datatype and Non-Null Counts

```
[ ]
```

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56842 entries, 0 to 56841
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   id                    56842 non-null  int64
 1   Gender                56842 non-null  object
 2   Age                   56842 non-null  int64
 3   Driving_License       56842 non-null  int64
 4   Region_Code           56842 non-null  float64
 5   Previously_Insured    56842 non-null  int64
 6   Vehicle_Age           56842 non-null  object
 7   Vehicle_Damage        56842 non-null  object
 8   Annual_Premium        56842 non-null  float64
 9   Policy_Sales_Channel  56842 non-null  float64
 10  Vintage               56842 non-null  int64
 11  Response              56842 non-null  int64
dtypes: float64(3), int64(6), object(3)
memory usage: 5.2+ MB
```

Outcome:

The dataset contains 56,842 entries and 12 columns.
No Missing Values: All 12 columns show 56,842 non-null counts, which means there are no missing values in this dataset based on this initial check.
Data Types: The columns are a mix of numeric and categorical types:

Numeric (mostly \text{int}64 or \text{float}64): ID, Age, Driving_License, Region_Code, Previously_Insured, Annual_Premium, Response, Vintage.

Categorical/Object (\text{object}): Gender, Vehicle_Age, Vehicle_Damage, Policy_Sales_Channel.

Descriptive Statistics

```
[ ]
```

```
1 df.describe()
```

	id	Age	Driving_License	Region_Code	Previously_Insured	Annual_Premium	Policy_Sal
count	56842.000000	56842.000000	56842.000000	56842.000000	56842.000000	56842.000000	5
mean	28421.500000	38.766546	0.997783	26.407586	0.458446	30537.745646	
std	16409.016338	15.505450	0.047030	13.188777	0.498275	17123.414502	
min	1.000000	20.000000	0.000000	0.000000	0.000000	2630.000000	
25%	14211.250000	25.000000	1.000000	15.000000	0.000000	24373.000000	
50%	28421.500000	36.000000	1.000000	28.000000	0.000000	31706.000000	
75%	42631.750000	49.000000	1.000000	35.000000	1.000000	39532.750000	
max	56842.000000	85.000000	1.000000	52.000000	1.000000	540165.000000	

Outcome:

Outcome summary for a numerical column were generated

Initial Null Value Summation

```
[ ]
1 df.isnull().sum()
```

	0
id	0
Gender	0
Age	0
Driving_License	0
Region_Code	0
Previously_Insured	0
Vehicle_Age	0
Vehicle_Damage	0
Annual_Premium	0
Policy_Sales_Channel	0
Vintage	0
Response	0

dtype: int64

Outcome:

Using df.isnull().sum(), we can calculates the total count of missing (null) values for every column in the DataFrame.
The outcome shows that every column in the DataFrame has a count of 0 null values, indicating that the dataset currently has no missing data in any of its features.
The columns checked are:
Id
Gender
Age
Driving_License
Region_Code
Previously_Insured
Vehicle_Age
Vehicle_Damage
Annual_Premium
Policy_Sales_Channel
Vintage
Response

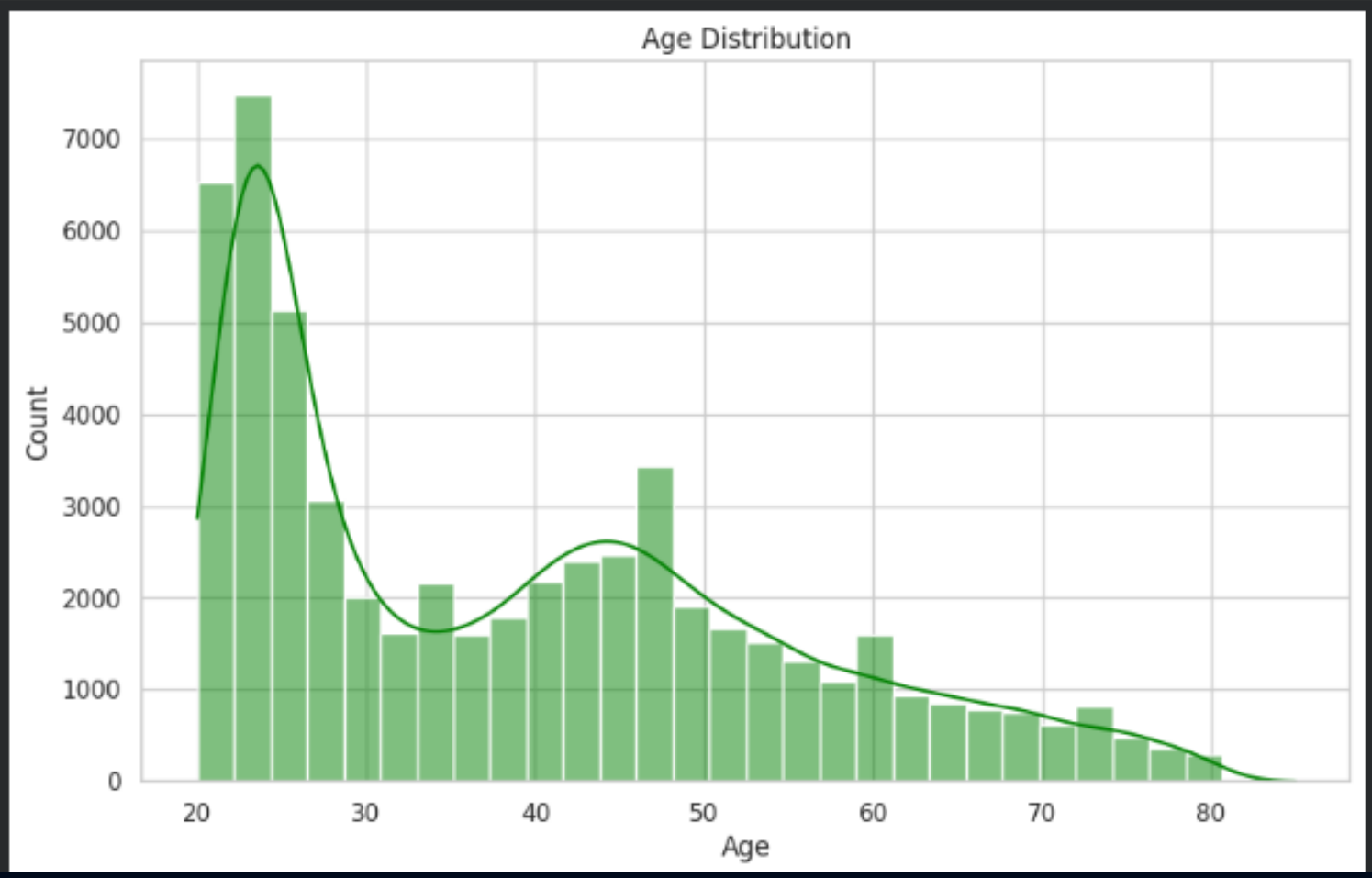
The data type of the result (the counts) is dtype: int64.After imputation of all analysed columns showed '0'missing value

Analysis and Insights

Univariate Analysis (Distribution Exploration)

Age Distribution

```
[ ]
1 sns.set(style="whitegrid",palette="pastel")
2 plt.rcParams['figure.figsize']=(10,6)
3 sns.histplot(df['Age'],kde=True,bins=30,color="green")
4 plt.title("Age Distribution")
5 plt.show()
```



Analysis Outcome :

1.The majority of individuals fall within the 20-30 years age group, indicating a younger population dominates the dataset.

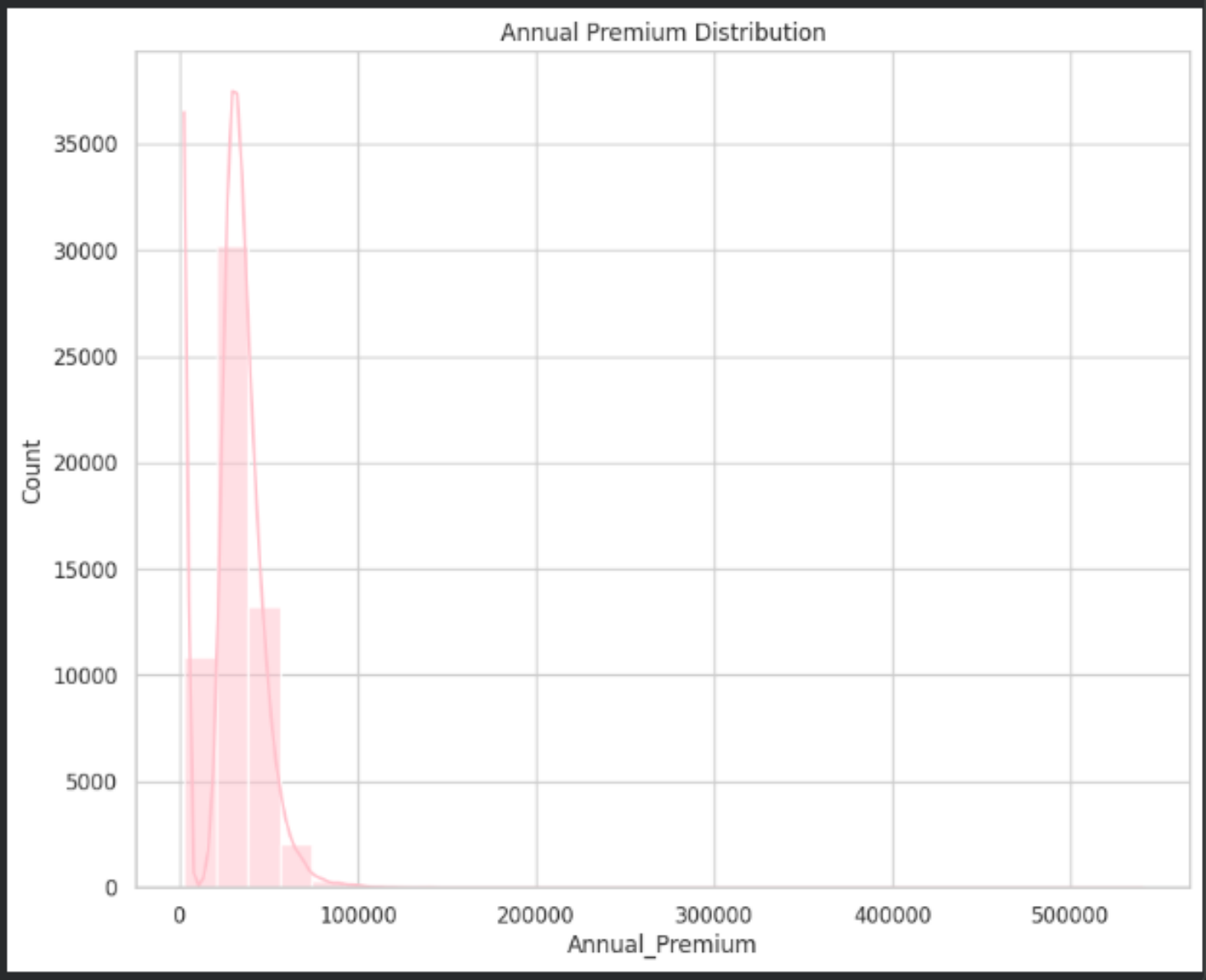
2.There is a secondary smaller peak around 45-50 years, suggesting another concentration of middle-aged individuals.

3.Beyond 60 years, the frequency of individuals gradually decreases, showing fewer older participants.

4.The distribution is right-skewed, meaning there are more younger individuals than older ones.

Annual Premium Distribution

```
[ ]
1 plt.figure(figsize=(10,8))
2 sns.histplot(df['Annual_Premium'],kde=True,bins=30,color="pink")
3 plt.title("Annual Premium Distribution")
4 plt.show()
```



Analysis Outcome :

1.The majority of annual premiums are concentrated in the lower range (below ₹50,000)

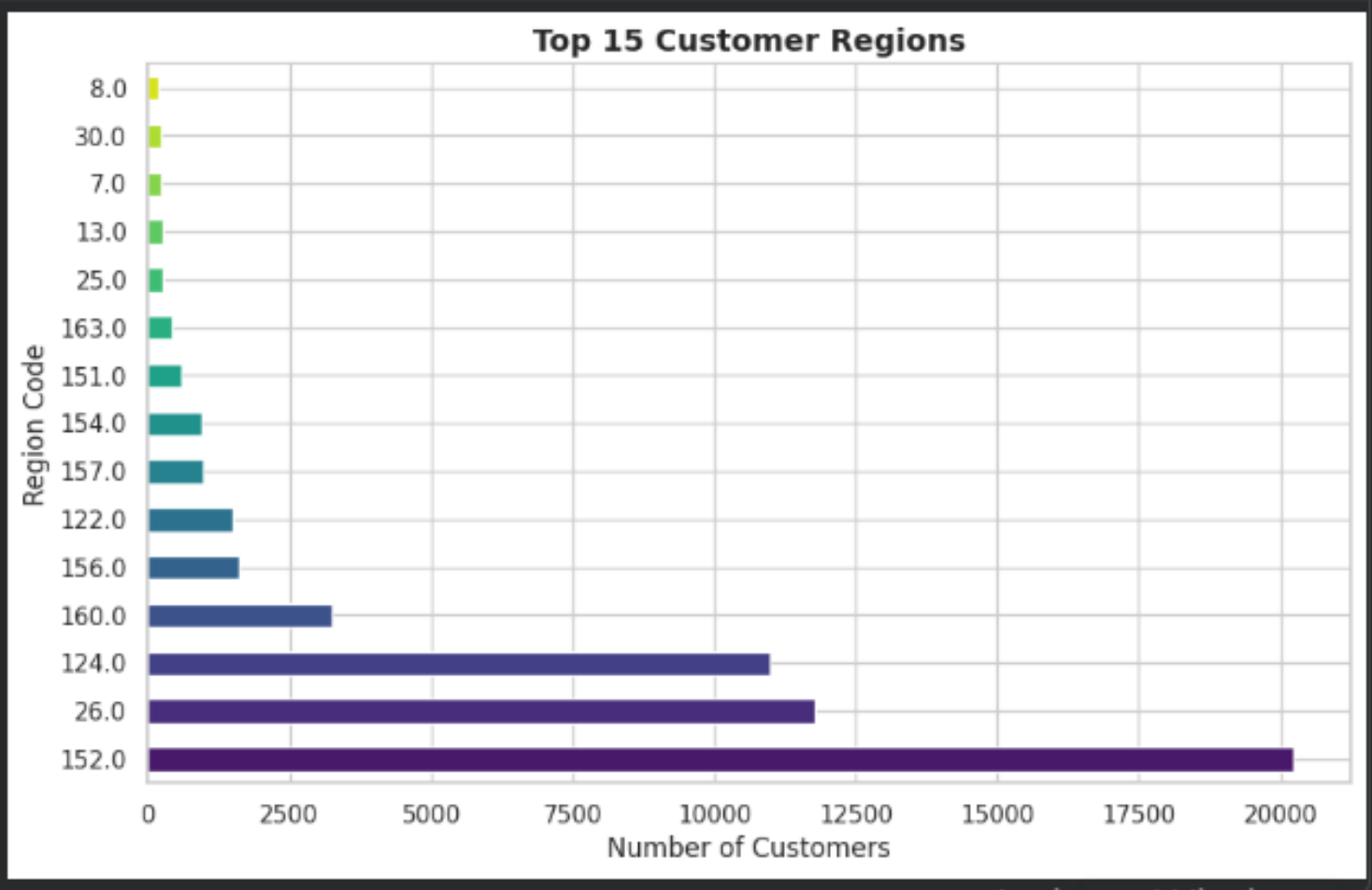
2.The distribution is highly right-skewed, indicating that most customers pay relatively low premiums, while a few customers pay very high premiums.

3.The long tail extending towards higher premium values suggests the presence of outliers or a small number of high-value policies.

4.The data shows that premium payments are unevenly distributed, with a large concentration of low to moderate payers.

Region Wise Policy Sales Count

```
[ ]
1 plt.figure(figsize=(10,6))
2 df['Policy_Sales_Channel'].value_counts().head(15).plot.barh(color=sns.color_palette("viridis", 15))
3 plt.title("Top 15 Customer Regions", fontsize=14, fontweight="bold")
4 plt.xlabel("Number of Customers")
5 plt.ylabel("Region Code")
6 plt.show()
```



Analysis Outcome: Region-Wise Policy Sales Count

1.The chart displays the number of customers (policyholders) across the top 15 regions.

2.Region Code 152.0 has the highest number of customers, nearing 20,000, making it the dominant region in policy sales.

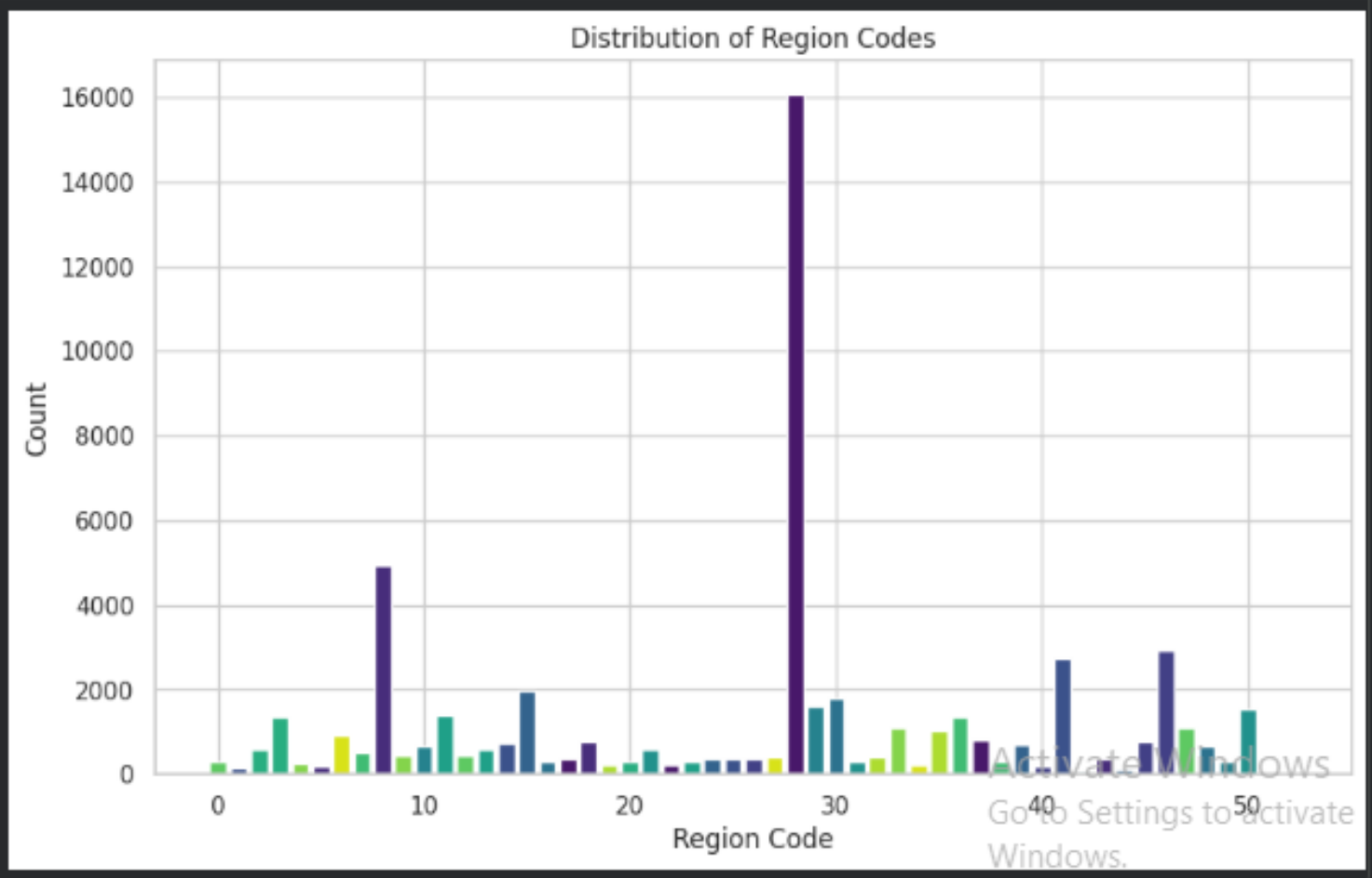
3.Regions 26.0 and 124.0 also have significant customer bases, with counts between 10,000–13,000.

4.Remaining regions show comparatively lower policy sales, indicating uneven regional distribution.

5.A few regions contribute very little to total sales, suggesting limited market reach or customer presence there.

Region Wise Code Distribution

```
[ ]
1 region_counts = df['Region_Code'].value_counts()
2 plt.bar(region_counts.index,region_counts.values,
3         color=sns.color_palette("viridis",15))
4 plt.xlabel("Region Code")
5 plt.ylabel("Count")
6 plt.title("Distribution of Region Codes")
7 plt.show()
```



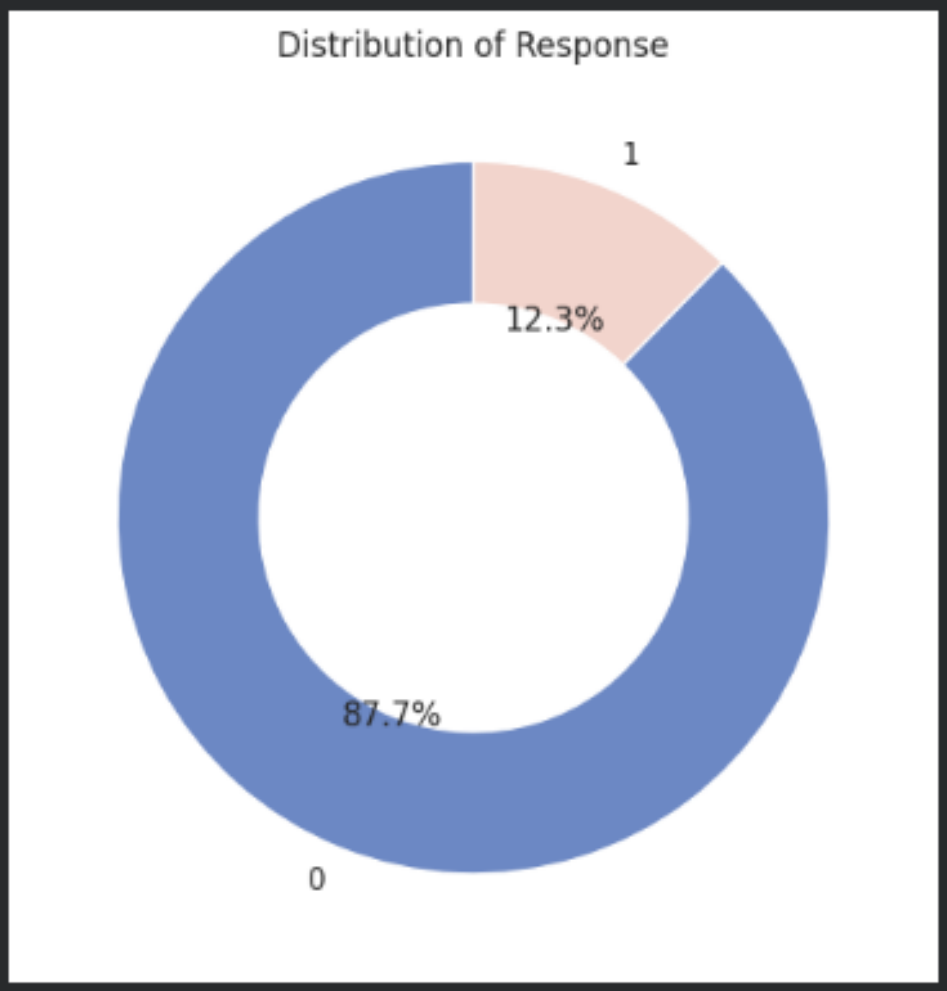
Analysis Outcome:

The analysis outcome for the "Region Wise Code Distribution" shows a bar chart illustrating the frequency of different region codes within the dataset. Notably, there are significant spikes at region codes around 30 and 10, indicating these regions have the highest counts, with the code near 30 being the most prominent. The color palette uses "viridis" to distinguish different counts visually. This distribution provides insight into the concentration of data points across various regions, highlighting regions with the most activity or entries in the dataset.

Bivariate Analysis (Distribution Exploration)

Distribution of Response

```
[ ]
1 response_counts = df['Response'].value_counts()
2 plt.pie(response_counts,labels=response_counts.index,
3 autopct="%1.1f%%",colors=["#6C88C4","#F2D4CC"],startangle=90,wedgeprops=dict(width=0.4))
4 plt.title('Distribution of Response')
5 plt.show()
```



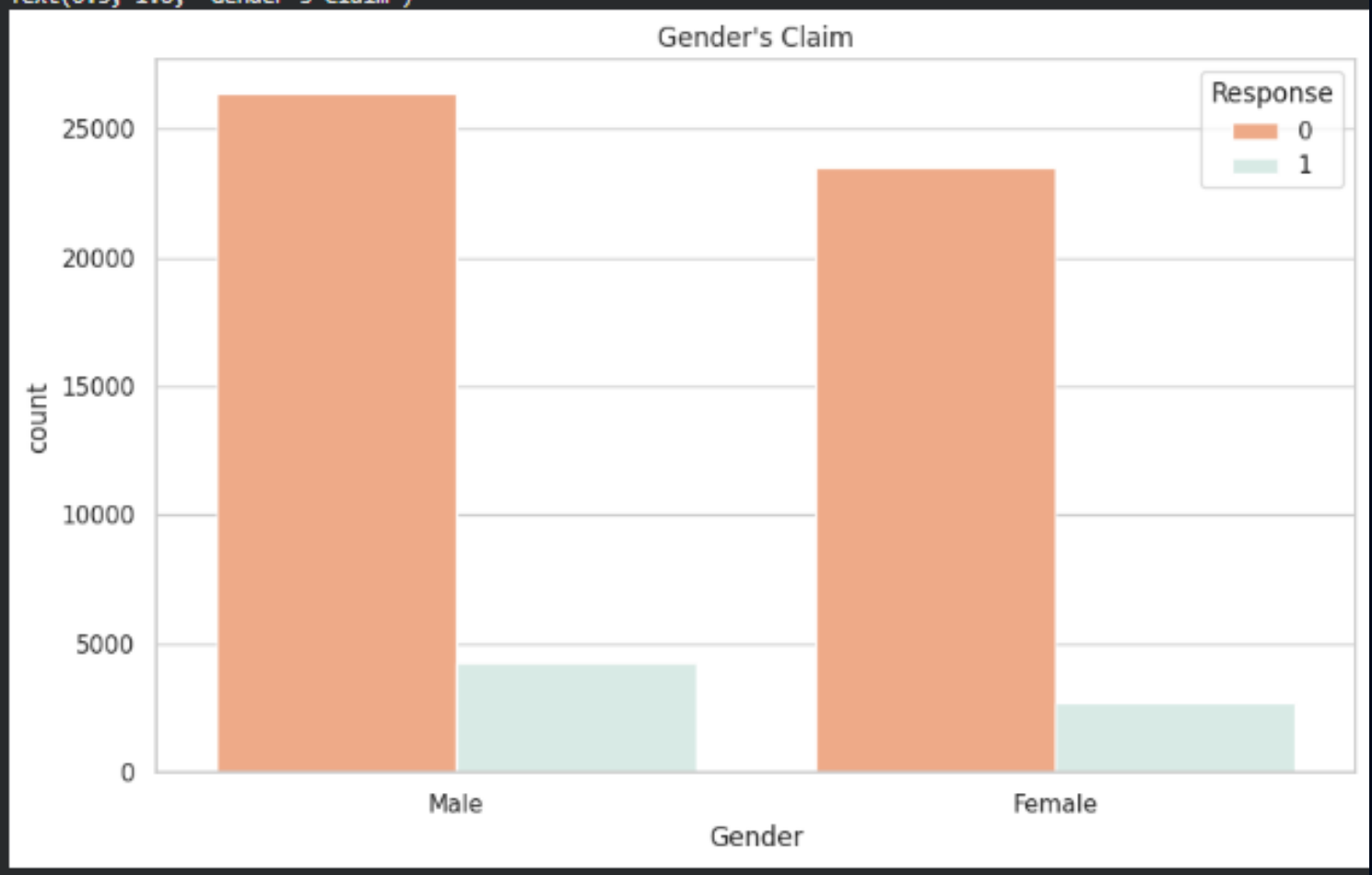
Analysis Outcome:

The analysis outcome depicted in the donut chart illustrates the distribution of responses within a dataset, highlighting that the majority of responses fall into one category, accounting for 87.7%, while a smaller portion, 12.3%, represents another response category, and an even smaller segment, possibly zero responses, is also shown. This suggests a significant skew toward one response type, indicating a dominant trend or preference in the data.

Gender Wise Claim Mapping

```
[ ]
1 sns.countplot(data=df,x="Gender",hue="Response",palette=["#FFA477","#D5EDE7"])
2 plt.title("Gender's Claim")
```

```
Text(0.5, 1.0, "Gender's Claim")
```



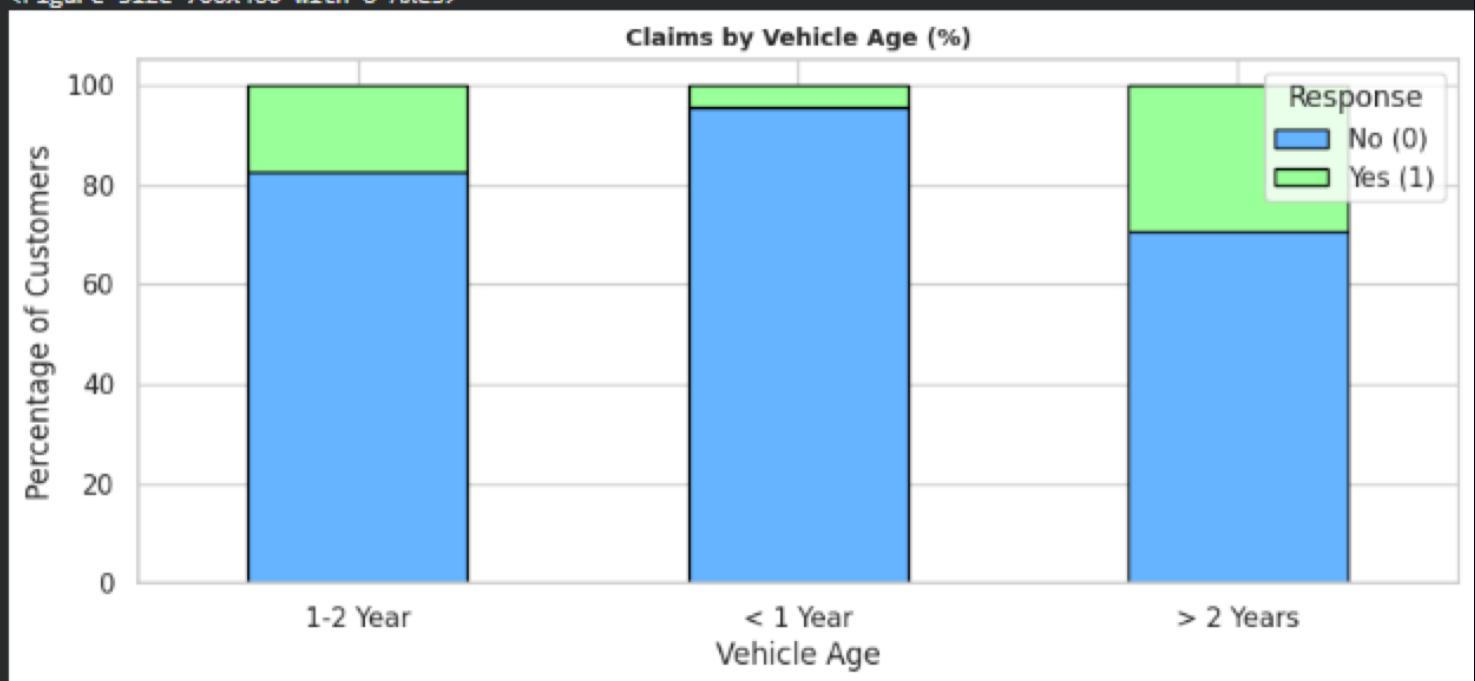
Analysis Outcome:

The analysis of the "Gender Wise Claim Mapping" chart indicates that males have a significantly higher claim count compared to females, with males showing the highest response count of around 25,000, while females have a slightly lower count. The data suggests that males are more likely to make claims than females in this dataset, pointing to a gender disparity in claim submissions.

Claim by Vehicle Age (In terms of percentage)

```
[ ]
1 cross_tab = pd.crosstab(df['Vehicle_Age'], df['Response'], normalize="index")*100
2 plt.figure(figsize=(7,4))
3 cross_tab.plot(kind="bar", stacked=True, color=["#66b3ff","#99ff99"], figsize=(10,4), edgecolor="black
4 plt.title("Claims by Vehicle Age (%)", fontsize=10, fontweight="bold")
5 plt.xlabel("Vehicle Age")
6 plt.ylabel("Percentage of Customers")
7 plt.legend(title="Response", labels=["No (0)", "Yes (1)"])
8 plt.xticks(rotation=0)
9 plt.show()
```

```
<Figure size 700x400 with 0 Axes>
```



Analysis Outcome:

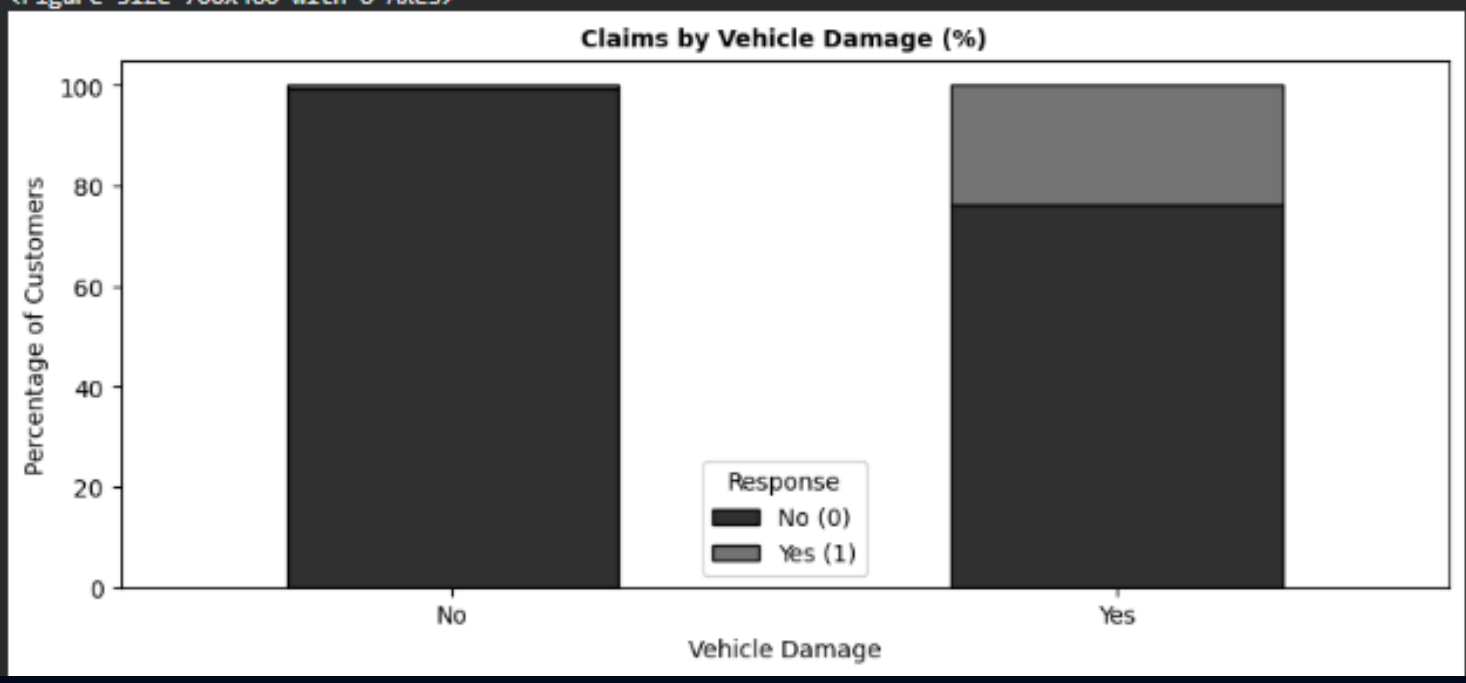
The chart illustrates that the percentage of customers filing claims increases with the age of the vehicle. Customers with vehicles aged more than 2 years have the highest claim percentage, with a notable portion responding "Yes" (claiming), while those with 1-2 years or less than 1 year show a lower claim rate, predominantly "No" (not claiming). This suggests that older vehicles are more likely to be associated with claims.

Claim by vehicle Damage (In terms of percentage)

[7] ✓ 0s

```
1 import matplotlib.pyplot as plt
2 plt.figure(figsize=(7,4))
3 cross_tab = pd.crosstab(df['Vehicle_Damage'], df['Response'], normalize="index")*100
4 cross_tab.plot(kind="bar", stacked=True, color=["#303030","#737373"], figsize=(10,4), edgecolor="black
5 plt.title("Claims by Vehicle Damage (%)", fontsize=10, fontweight="bold")
6 plt.xlabel("Vehicle Damage")
7 plt.ylabel("Percentage of Customers")
8 plt.legend(title="Response", labels=["No (0)", "Yes (1)"])
9 plt.xticks(rotation=0)
10 plt.show()
```

```
<Figure size 700x400 with 0 Axes>
```



Analysis Outcome:

1. Customers with damaged vehicles are more likely to file claims ("Yes").

2. The percentage of claims increases with the presence of vehicle damage.

3. Customers without damage tend to file fewer claims ("No").

4. Overall, vehicle damage is a key factor influencing claim frequency.

Conclusion

The comprehensive analysis of the dataset reveals several key insights across demographic, financial, regional, and claim-related dimensions. Firstly, the age distribution indicates a predominantly young population, with most individuals falling within the 20-30 years age bracket, followed by a smaller peak around 45-50 years, and a decline beyond 60 years. This right-skewed distribution highlights a youthful user base with decreasing participation among older individuals. Financially, the annual premium distribution is heavily concentrated below ₹50,000, with a significant skew towards lower premiums, though a long tail suggests the presence of high-value outliers. Region-wise analysis shows that certain regions, notably region code 152.0, dominate with nearly 20,000 policyholders, indicating uneven market penetration. Other regions have moderate to low policy counts, underscoring regional disparities in market reach. The distribution of region codes further emphasizes concentrated activity around specific codes like 30 and 10. Claim analysis indicates a gender disparity, with males submitting notably more claims than females, and a clear correlation between vehicle age and claim frequency — older vehicles (beyond 2 years) have a higher likelihood of claims. Additionally, vehicle damage significantly influences claim behavior, with damaged vehicles leading to a higher probability of claims. The distribution of responses in the dataset is heavily skewed, with one category dominating at 87.7%, suggesting a prevalent trend or preference. Overall, these insights highlight demographic tendencies, regional disparities, financial behaviors, and claim patterns, which can inform targeted marketing strategies, risk assessment, and policy customization. The findings underscore the importance of regional and demographic factors in shaping the insurance landscape and suggest avenues for optimizing product offerings and improving market penetration across diverse customer segments.

Acknowledgement

I wish to express my heartfelt thanks to everyone involved in my eighth-grade computer project.

First and foremost, I am grateful to Mr. Tushar Bhadouria sir for providing the necessary resources and guidance for undertaking this project. Their patience and encouragement made the entire process enjoyable.

I would also like to thank for advocating technology education at our institution. The computer lab facilities provided me with the opportunity to learn new skills through hands-on practice.

Additionally, I must extend my gratitude to my parents, whose unwavering support was evident at every step. From accompanying me to the store for supplies to motivating me during challenging moments, their support proved to be invaluable.

Finally, I appreciate the constructive feedback and suggestions from my classmates, which played a crucial role in refining my project at various stages.

The successful completion of this project was a collective effort, and I am deeply appreciative of the time and care invested by everyone. This learning experience has been incredibly fruitful, and I will always remember the collaborative efforts that made it possible.

Thank you