EXPLORATORY DATA ANALYSIS (EDA) ON VEHICLE INSURANCE DATASET

BY-RITESH SINGH

PRESENTATION FLOW

1) Introduction & Objectives – Project background and goals
2) Dataset Overview – Structure and features of dataset
3) Methodology & Approach – Step-by-step workflow of analysis
4) Data Preprocessing & Cleaning – Handling missing values, outliers, data prep
5) Exploratory Analysis – Detailed EDA on:

- Age
- Gender
- Premiums
- Claims
- Vehicle (Age & Damage)
- Policy & Region

6) Key Insights – Summary of findings
7) Conclusion & Future Scope – Practical takeaways and future directions

INTRODUCTION TO PROJECT

❖**Importance of EDA in Data Science**

- Helps in identifying hidden patterns, trends, and correlations
- Foundation step before predictive modeling and machine learning
- Improves decision-making with data-driven insights

❖**Insurance Industry & Data-Driven Decisions**

- Premium pricing, claim settlement, and fraud detection rely on data analysis
- Customer demographics, vehicle details, and historical claims guide risk assessment

❖**Aim of This Project**

- Apply EDA techniques on a vehicle insurance dataset
- Uncover insights that support:
  - Risk Management
  - Premium Pricing Strategies
  - Claims Analysis & Fraud Prevention

PROJECT OBJECTIVES

- **Explore and Clean the Insurance Dataset**
  - Inspect dataset structure, variables, and records
  - Handle missing values, outliers, and incorrect data types
  - Ensure dataset is reliable for analysis
- **Identify Claim Patterns and Influencing Factors**
  - Study correlations between customer demographics, vehicle details, and claims
  - Detect high-risk groups (e.g., young drivers, old vehicles, damaged cars)
  - Understand frequency and distribution of claims
- **Visualize Key Trends and Insights**
  - Use charts, plots, and graphs for:
    - Customer demographics (age, gender, region)
    - Premium distribution and trends
    - Claim frequencies by different attributes
  - Make complex data easy to interpret
- **Provide Actionable Insights for Decision-Making**
  - Support insurance companies in **risk assessment**
  - Assist in designing **fair premium pricing models**
  - Contribute towards **fraud detection and prevention**
  - Strengthen **customer segmentation & targeting strategies**

ABOUT VEHICLE INSURANCE

❖**Definition & Purpose**

- Vehicle insurance provides **financial protection** against accidents, theft, or damage.
- Acts as a **risk-transfer mechanism**, reducing financial burden on policyholders.

❖**Importance of Claim Analysis**

- Claims data helps insurers evaluate:
  - **Risk exposure** of different customer segments
  - **Likelihood of fraudulent claims**
  - **Profitability of insurance products**
- Without claim analysis, insurers may face **higher losses** and poor pricing models.

❖**Impact of Vehicle Attributes on Claims**

- **Vehicle Age** – Older vehicles are more prone to accidents and mechanical failures.
- **Vehicle Damage History** – Past damage strongly predicts future claim likelihood.
- **Policy Type & Coverage** – Comprehensive vs. basic policies influence claim frequency and size.
- **Customer Demographics** (age, gender, region) also indirectly affect claim risks.

DATASET OVERVIEW

Dataset contains:

- **Demographics:** Age, Gender, Region
- **Policy Details:** Premium, Policy Type, Number of Policies
- **Vehicle Info:** Age of Vehicle, Vehicle Damage
- **Claims:** Past Claim Frequency, Claim Likelihood
- Dataset Size: (381109,12)
- Dataset-Vehicle_Insurance.csv - Google Drive

```
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
```

IMPORTANCE OF EXPLORATORY DATA ANALYSIS (EDA)

❖**Detect Hidden Patterns in Claims**

- Reveal underlying relationships between variables (e.g., vehicle age & claim frequency).
- Spot unusual trends or anomalies that may impact business outcomes.

❖**Understand Customer Risk Profiles**

- Segment customers into **low-risk** and **high-risk** categories.
- Helps insurers design **personalized policies**.
- Provides insights into demographics (age, gender, region) driving claims.

❖**Optimize Premium Pricing Models**

- Align premium amounts with actual risk exposure.
- Prevents **underpricing high-risk customers** or **overpricing low-risk customers**.
- Leads to better customer retention and profitability.

❖**Identify Fraud-Prone Cases**

- Spot unusual claim behavior through EDA.
- Example: multiple claims in short time, mismatched vehicle damage patterns.
- Early detection saves cost and maintains integrity.

❖**Build Foundation for Predictive Models**

- Prepares data for **machine learning models** (e.g., claim prediction, churn analysis).
- Validates assumptions and improves accuracy of future predictive analytics.

TOOLS & TECHNOLOGIES USED

❖**Python**

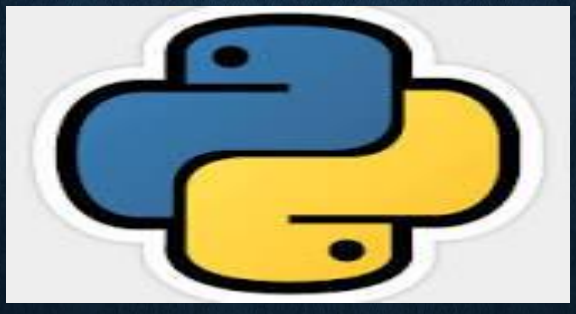
- Used for data analysis, preprocessing, and visualization.
- Provides flexibility and extensive library support for EDA.

❖**Libraries**

- Pandas – For data cleaning, manipulation, and analysis.
- NumPy – For efficient numerical computations and array operations.
- Matplotlib – For static, customizable visualizations.
- Seaborn – For advanced and aesthetic statistical plots.

❖**Jupyter Notebook**

- Interactive environment for running Python code, visualizing results, and documenting analysis in real time.



DATA PREPARATION & CLEANING

❖**Data Quality Checks**

- Checked for **missing values** and handled them using **imputation or removal** techniques.
- Identified and **removed duplicate or irrelevant records** to ensure data accuracy.

❖**Data Transformation**

- **Converted categorical variables** (e.g., Gender, Vehicle_Damage) into meaningful or numerical formats for analysis.
- Ensured data types were consistent for seamless processing.

❖ **Data Standardization**

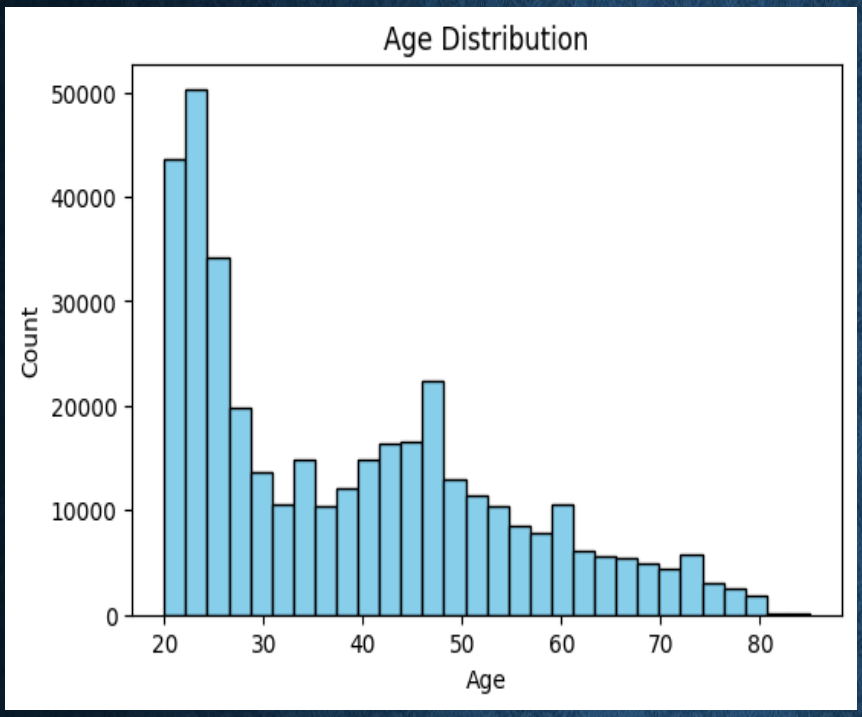
- **Standardized numerical columns** such as *Age*, *Premium*, and *Vintage* for uniform scale and better visualization.
- Ensured dataset was **clean, structured, and analysis-ready**.

```
# . Check for duplicates
duplicates = df.duplicated().sum()
print("Number of duplicate rows:", duplicates)
if duplicates > 0:
    df = df.drop_duplicates()
```

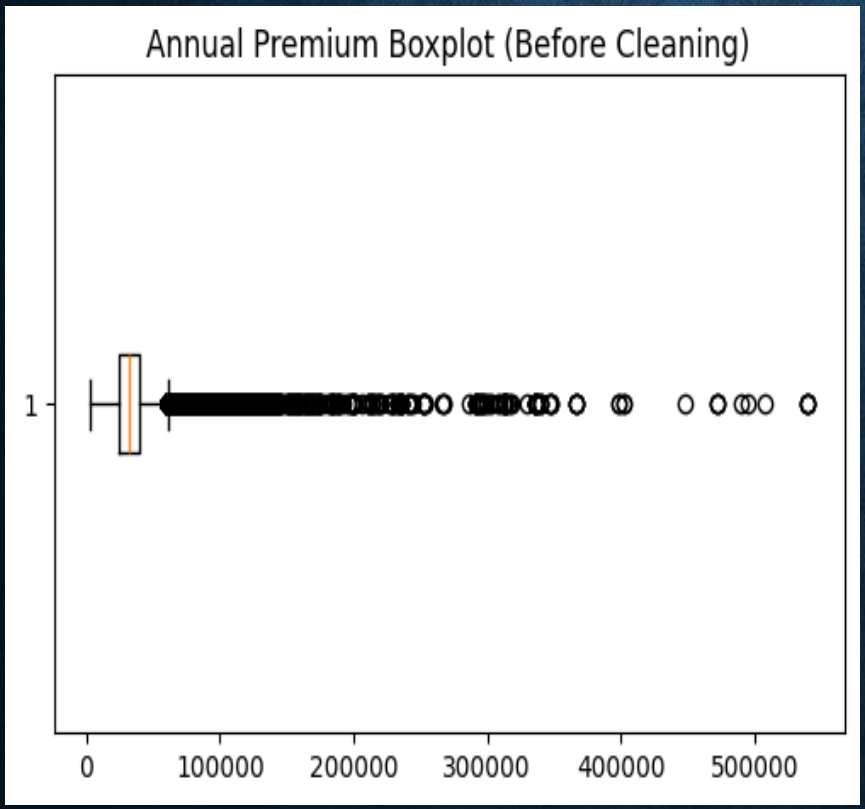
```
Number of duplicate rows: 0
```

```
#Convert float-coded categorical columns to integer
df['Region_Code'] = df['Region_Code'].astype(int)
df['Policy_Sales_Channel'] = df['Policy_Sales_Channel'].astype(int)
```

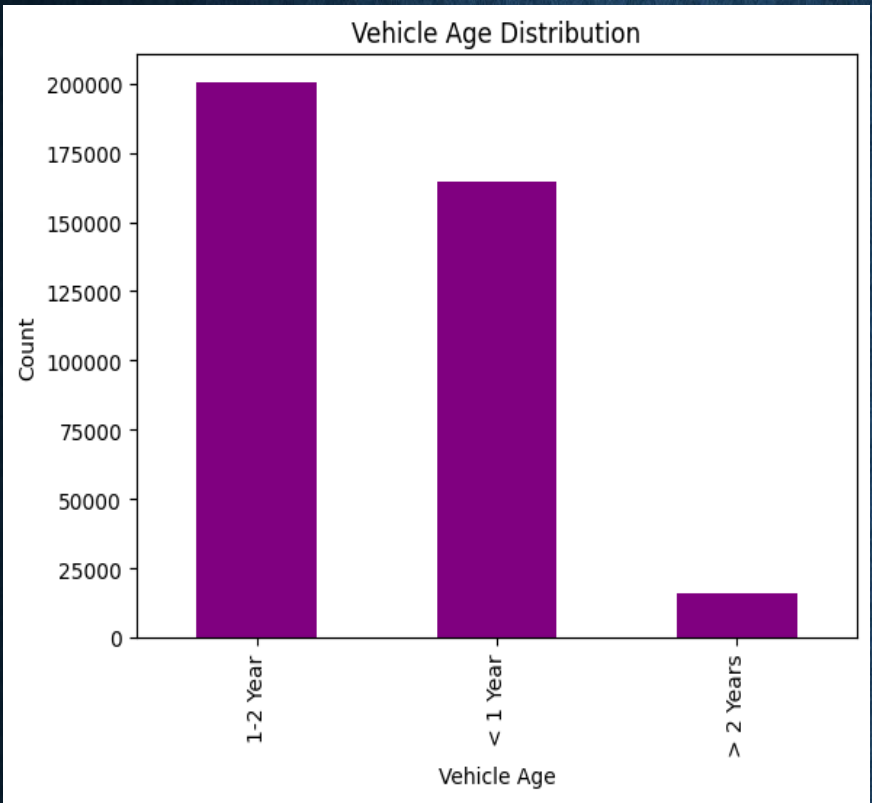
ANALYSIS



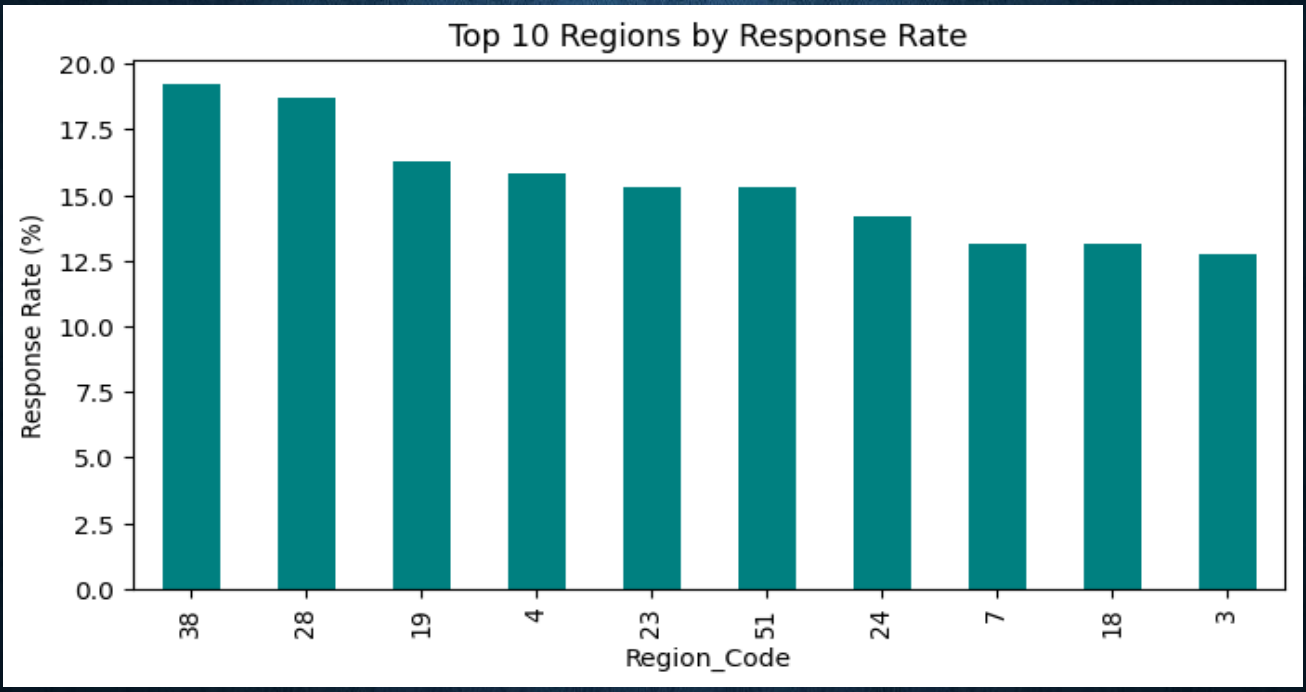
Most policyholders fall within the 30–45 age group, showing the largest customer base.
Indicates mid-career individuals are the key target segment for vehicle insurance.



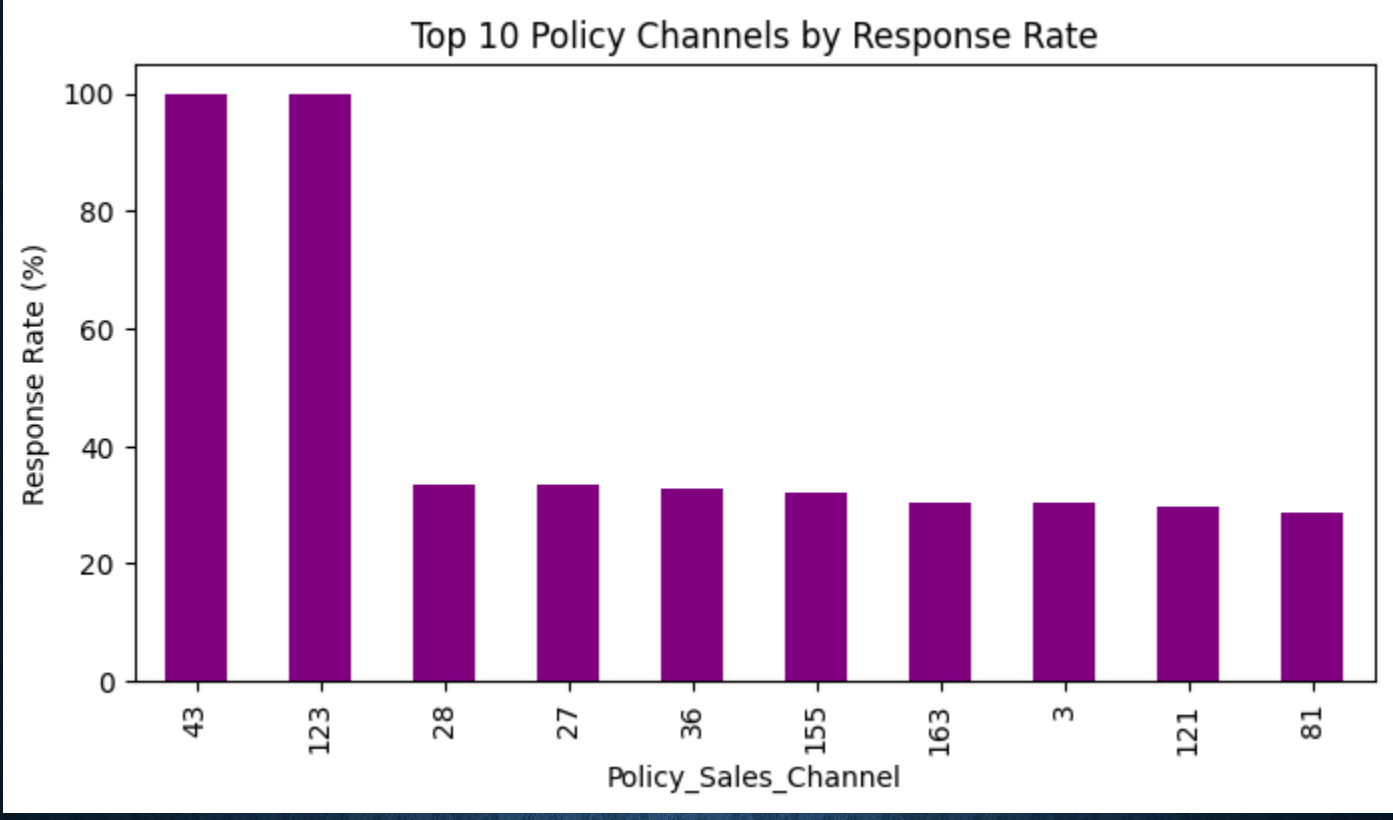
- Premium amounts are **right-skewed** with a few high-value outliers.
- Majority of policies are priced within a **moderate premium band**, showing affordability balance.



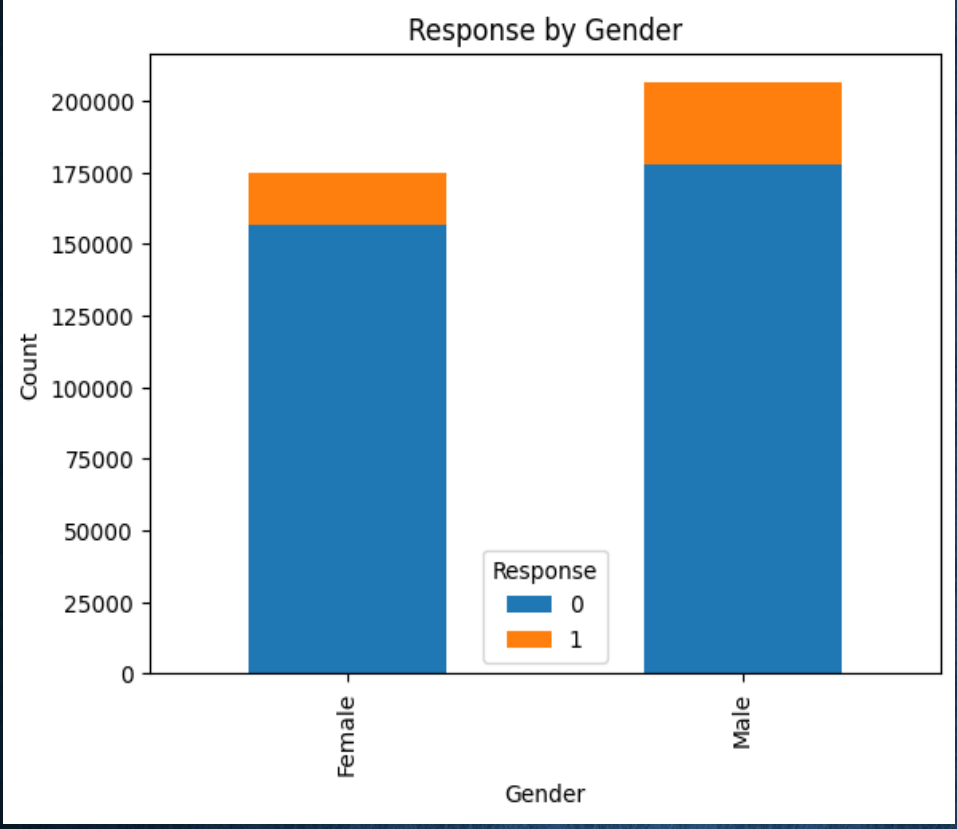
Vehicles aged 1–2 years dominate the dataset. Newer vehicles are more likely to be insured, implying customer preference for early coverage.



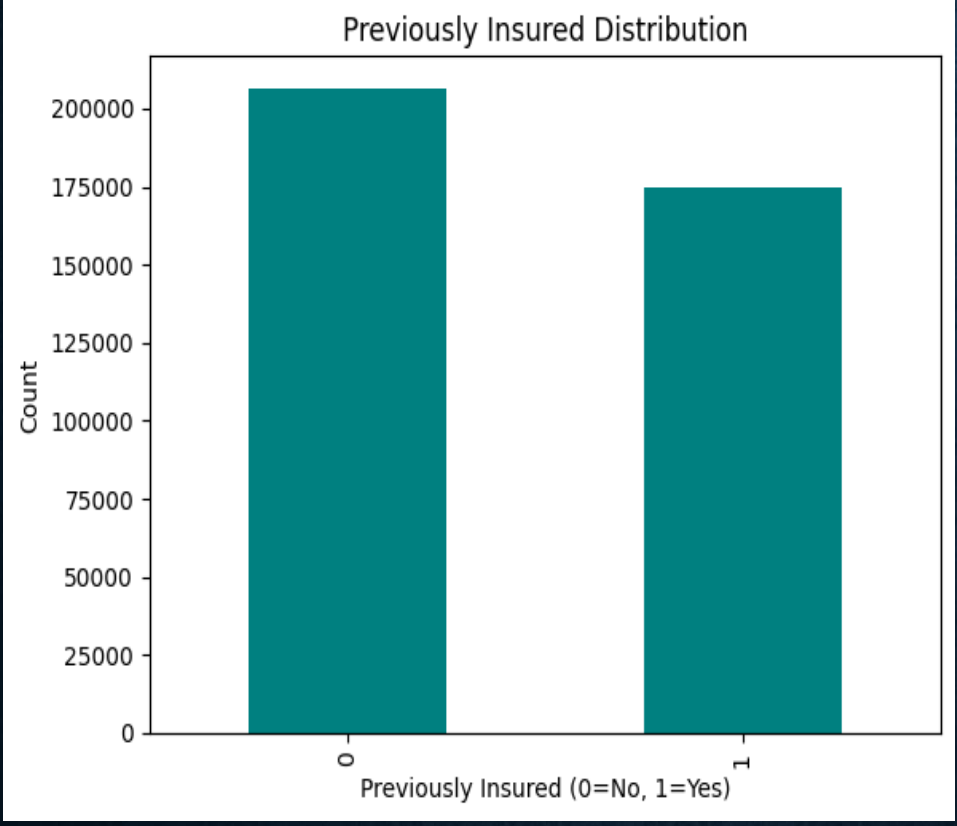
•**Region 38** shows the highest engagement, with nearly **20%** of customers responding positively, closely followed by **Region 28**. This suggests that marketing or outreach efforts might be most effective when targeted towards customers in these specific regions.



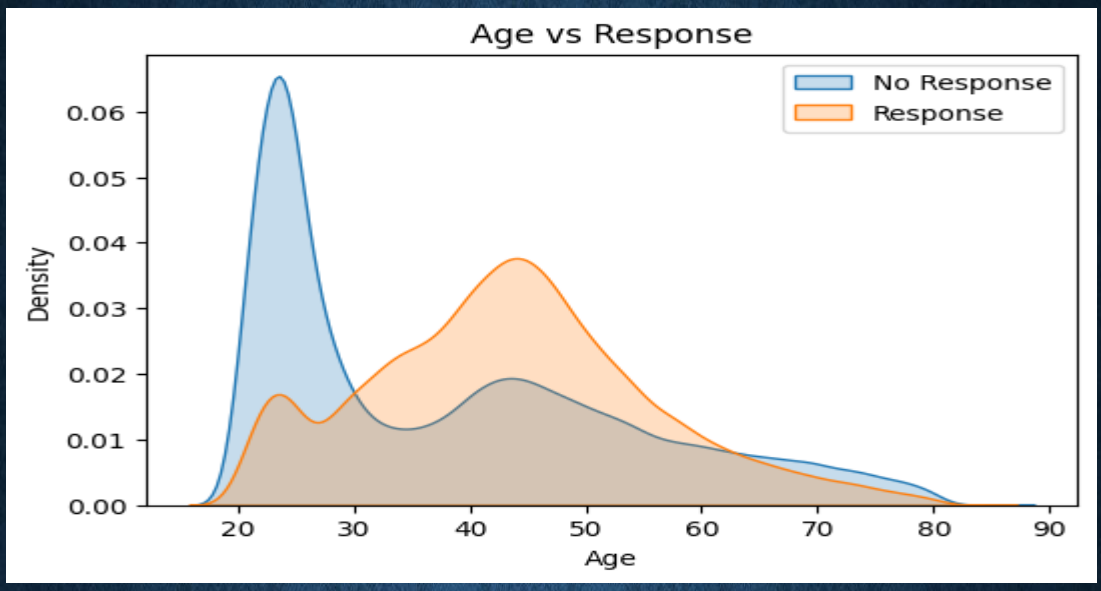
Policy Sales Channels 43 and 123 exhibit near **100% response rates**, dramatically outperforming other top channels (around 30%). This highlights their exceptional effectiveness as customer acquisition paths for this insurance offer.



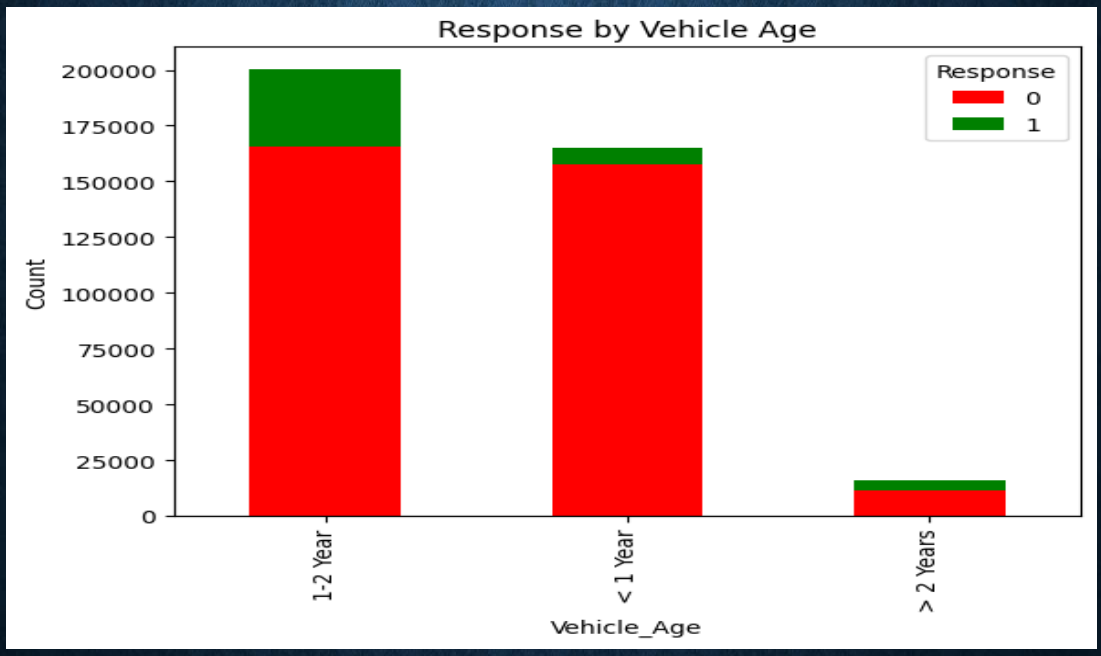
•**Male customers** show a **slightly higher claim rate** compared to females.
•Suggests possible behavioral differences or higher exposure to driving risk.



Policyholders with previous vehicle damage have a significantly higher response rate for claims. Indicates strong correlation between vehicle history and claim behavior.



The density plot shows that customers who **did not respond** are heavily concentrated in the **young adult** age group (peaking around 25). Conversely, customers who **did respond** are predominantly **middle-aged**, with the highest likelihood appearing between ages 40 and 50.



Customer response is **heavily dependent on vehicle damage**. Over **80%** of the customers who purchased the policy had previously reported **vehicle damage**, making this group the primary target market for the offer.

KEY INSIGHTS

❖Vehicle Damage & Customer Vintage

- Analysis shows that **vehicle damage history** and **customer vintage (tenure)** are the **strongest indicators** of whether a customer will file a claim.
- Customers with a **record of previous vehicle damage** are **significantly more likely** to make future claims.
- Similarly, customers who are **new or have shorter association periods** with the company show **higher claim probabilities**, indicating limited loyalty or riskier profiles.

❖Premium Variation by Region & Age

- The **annual premium amount** differs widely across **regions and age segments**.
- **Urban regions** or those with higher accident rates tend to have **higher average premiums**.
- **Younger policyholders** usually pay **lower premiums** but have **higher claim frequencies**, while **older customers** pay more stable premiums and claim less.

❖ Regional Claim Patterns

- Certain **region codes show higher claim frequencies**, identifying them as **risk-prone zones**.
- Insurers can use these insights to **adjust premium rates** and **design region-specific risk mitigation strategies**.
- Regions with **low claim rates** can be targeted for **marketing and customer acquisition** to improve profitability.

❖Gender-Based Insights

- **Gender differences** in claim rates are **minor but noticeable**.
- **Male customers** exhibit a **slightly higher claim frequency**, possibly due to **higher vehicle usage or risk exposure**.
- Gender patterns, while not dominant, can still inform **personalized offers or risk segmentation** strategies.

❖Overall Customer Behavior Trends

- Younger and newly onboarded customers are **more likely to claim** but less likely to stay long-term.
- Older and loyal customers are **less likely to claim**, offering **better stability and profitability** for insurers.
- These patterns help in building **predictive models** for customer risk profiling and retention planning.

FINDINGS & RECOMMENDATIONS

❖Key Findings

- **Customer Demographics:**
  - Majority of policyholders are **aged between 25–45 years**, with males forming a larger share.
  - **Younger customers (<25)** file more claims, while **older customers (>45)** show higher loyalty and lower claim rates.

❖Vehicle & Premium Insights:

- **Vehicle damage** and **customer vintage (tenure)** are the **most significant predictors** of claim behavior.
- **Premium values** vary notably by **region and vehicle age**, indicating potential for location-based pricing.

❖Regional & Channel Behavior:

- Certain **regions show high claim frequencies**, identifying them as **risk-prone zones**.
- Policy **sales channels** attract varied customer segments, influencing overall claim probabilities.

- **Recommendations**

❖Targeted Marketing:

- Focus marketing efforts on **low-claim, loyal customer segments** such as older policyholders.
- Design **personalized renewal offers** and **loyalty rewards** to enhance retention.

❖Incentive Programs:

- Offer **discounts or benefits** for customers with **no previous damage records** to encourage safe driving behavior.

❖Risk-Based Pricing:

- Review and adjust **premium strategies** for **high-risk regions** based on historical claim data.
- Introduce **dynamic pricing models** that factor in age, region, and vehicle condition.

❖Predictive Analytics Integration:

- Implement **AI-driven predictive models** to identify **potential high-claim customers early**.
- Use insights for **proactive risk management** and **fraud detection**.

CONCLUSION

❖Summary of the Analysis

- The **Exploratory Data Analysis (EDA)** provided a deep understanding of the **vehicle insurance dataset**, uncovering key **trends and relationships** in customer demographics, vehicle characteristics, and claim behaviors.
- The analysis highlighted **critical variables** such as **vehicle damage history**, **customer vintage**, and **region**, which play a major role in influencing claim likelihood.

❖Business Impact

- Insights derived from the analysis can help insurers make **data-driven decisions** in:
  - **Risk evaluation** – Identifying high-risk customers and regions.
  - **Fraud detection** – Spotting unusual claim behaviors early.
  - **Premium optimization** – Adjusting pricing based on customer profiles and claim probability.
- Enables **strategic planning** for targeted marketing, improved retention, and profitability.

❖ Future Scope

- Extend the project by applying **Machine Learning models** to:
  - **Predict claim likelihood** using historical data.
  - **Segment customers** based on risk profiles and policy behavior.
  - **Automate decision-making** for underwriting and renewal offers.
- Incorporate **Power BI or Tableau dashboards** for real-time claim monitoring and business insights.

GOOGLE COLABORATORY LINK

https://colab.research.google.com/drive/1D__rEA7a2zughGXdQYb6Hn6JtDOyHHi4?authuser=1#scrollTo=gWjrJlXDH5ol