Insurances claim data

Maryem, Mikel, Helena and Gurdit
Group 2

Explanation and Goals of the Analysis

- We collaborated with a vehicle insurance company to support decision-making in their membership insurance programs.
- The goal of this Exploratory Data Analysis (EDA) is to identify which vehicle, customer, and policy characteristics most strongly influence the likelihood of filing an insurance claim.
- By gaining a deeper understanding of these risk factors across customers, vehicles, and regions we can help the company:
    a. Develop more accurate risk profiles
    b. Optimize pricing strategies
    c. Increase overall profitability and revenue

Data check

Objective:
Ensure data quality, consistency, and accuracy before analysis.https://www.kaggle.com/datasets/litvinenko630/insurance-claims?utm_source=chatgpt.com

Main Steps:

- Loaded dataset and reviewed structure (shape, types, summary)
- Checked for missing values and duplicates (policy_id)
- Identified invalid or extreme values (e.g., vehicle_age = 0, subscription_length = 0)
- Verified categorical vs. numerical columns

Data cleaning

Refinement Actions:

- Standardized and cleaned inconsistent records
- Ensured unique policies and valid data ranges

Final Output:

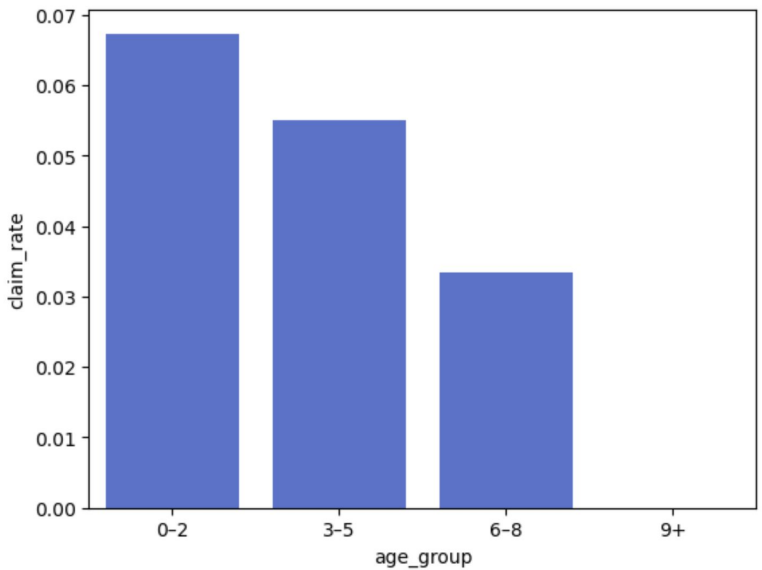
- Clean and reliable dataset → cleaned_data.csv
- Ready for EDA and risk profiling analysis

Analysis Focus and Approach

- The target variable of our study is Claim Status, indicating whether a customer filed an insurance claim.
- The dataset includes multiple features describing the customer, their demographics, and vehicle characteristics.
- In the Univariate EDA, we explored individual features to understand their distribution and their potential relationship with claim behavior.

Univariable EDA

- **Claim Rate by Vehicle Age**



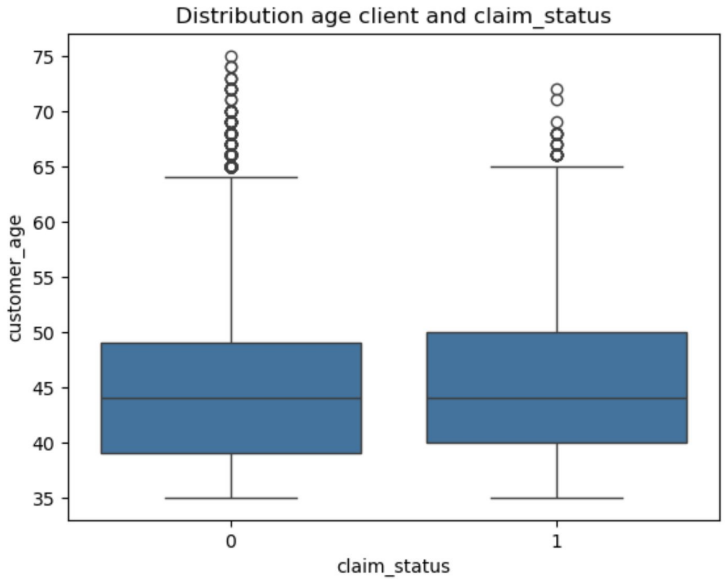
Observation:

- Claim rate **decreases as vehicles get older**.
- **0–2 year-old vehicles:** Highest claim rate (~6.7%)
- **6+ year-old vehicles:** Almost **no claims** reported

Insights:

- **Newer cars** → More claims
  - Owners of new vehicles often have **comprehensive coverage**
  - More likely to **report even minor damages**
- **Older cars** → Fewer claims
  - Owners may **downgrade to basic insurance**
  - Less inclined to file for **small incidents**

Univariable EDA



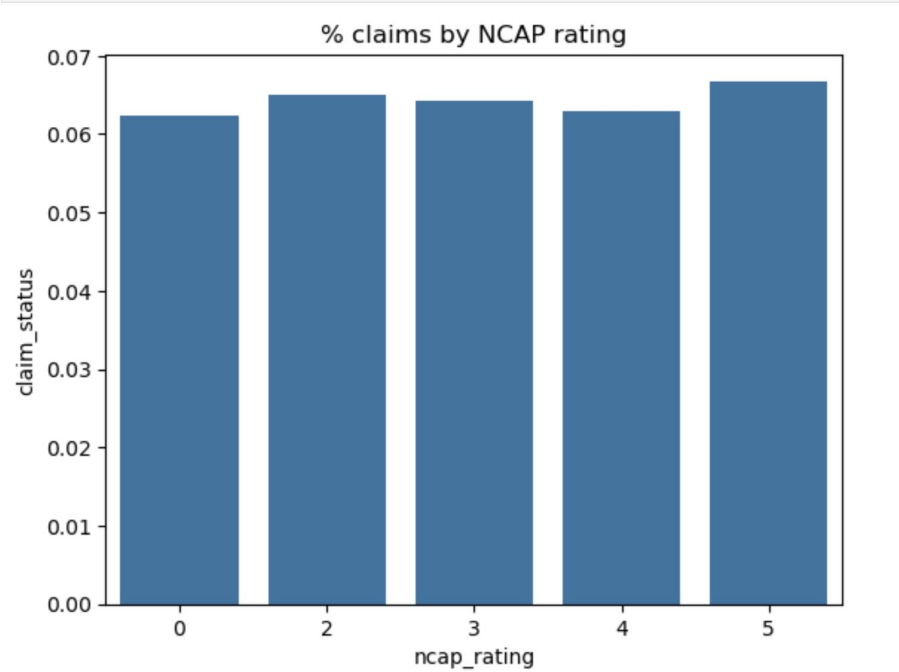
Observation:

- The **boxplot** shows that the **age distributions** of customers **with and without claims** are quite similar.
- The **claiming group** tends to be **slightly older**, with a **higher median age**.

Insight:

- This pattern is consistent with the previous finding:
  - Customers **over 55 years old** have **higher claim rates**.
- Overall, **customer age has a mild effect** —
  - **Older customers** claim **slightly more often**,
  - but the **difference is not significant**.

Univariable EDA



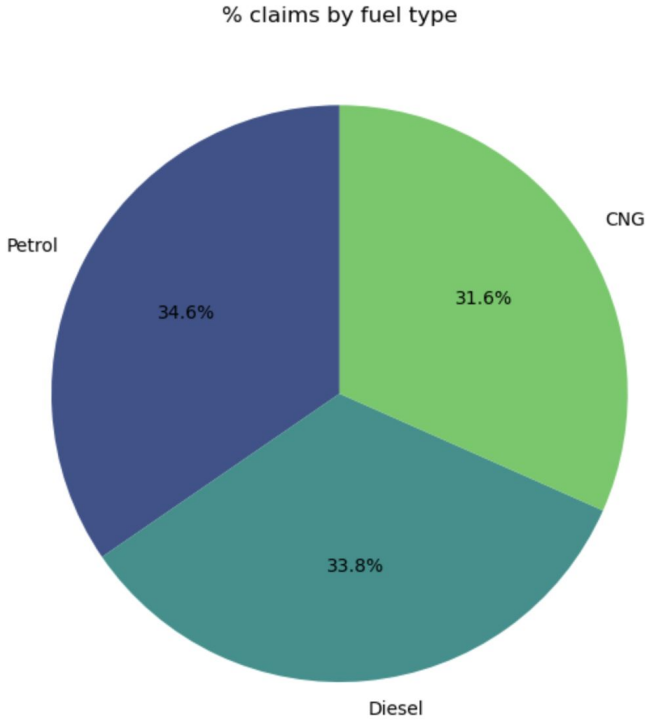
Observation:

- Claim rates are **very similar** across all **NCAP values (0–5)**.
- **No clear trend** is observed — neither increasing nor decreasing

Insight:

- The **NCAP safety rating** is **not significantly related** to claim frequency.
- This may be because **NCAP focuses on passive safety** (crash protection),
  rather than the **likelihood of an accident** or a **driver's claim behavior**.

Univariable EDA



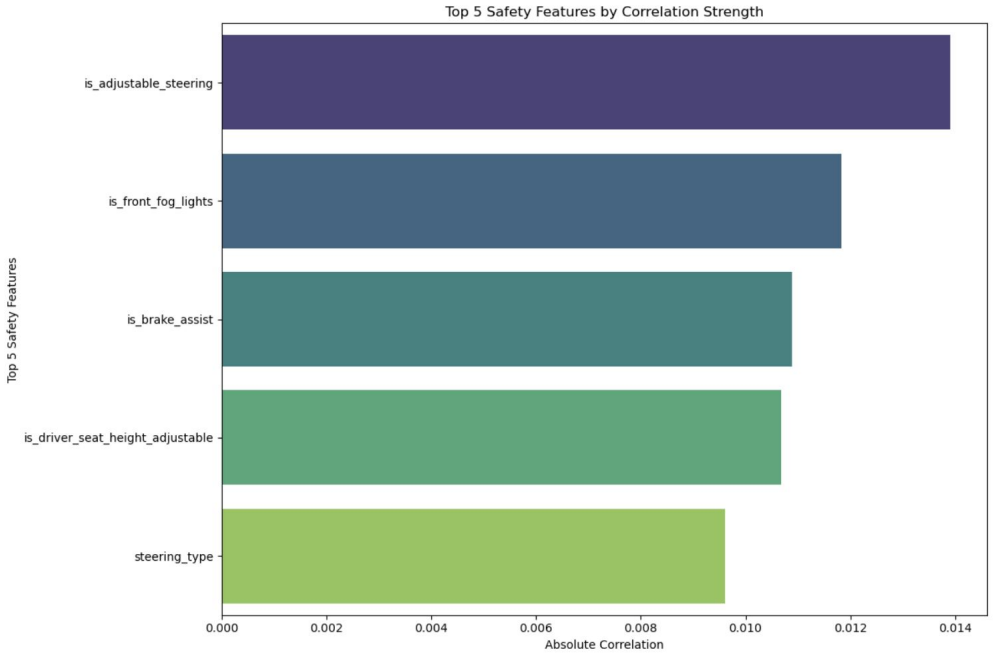
Observation:

- Claim rates are **fairly similar** across fuel types:
  - **Petrol:** 34%
  - **Diesel:** 33%
  - **CNG:** 31%

Insight:

- **Petrol vehicles** show a **slightly higher claim rate**, but the difference is **minimal**.
- Overall, **fuel type has only a modest impact** on claim risk.

Bivariate EDA (Two Variables at a Time)



Observation:

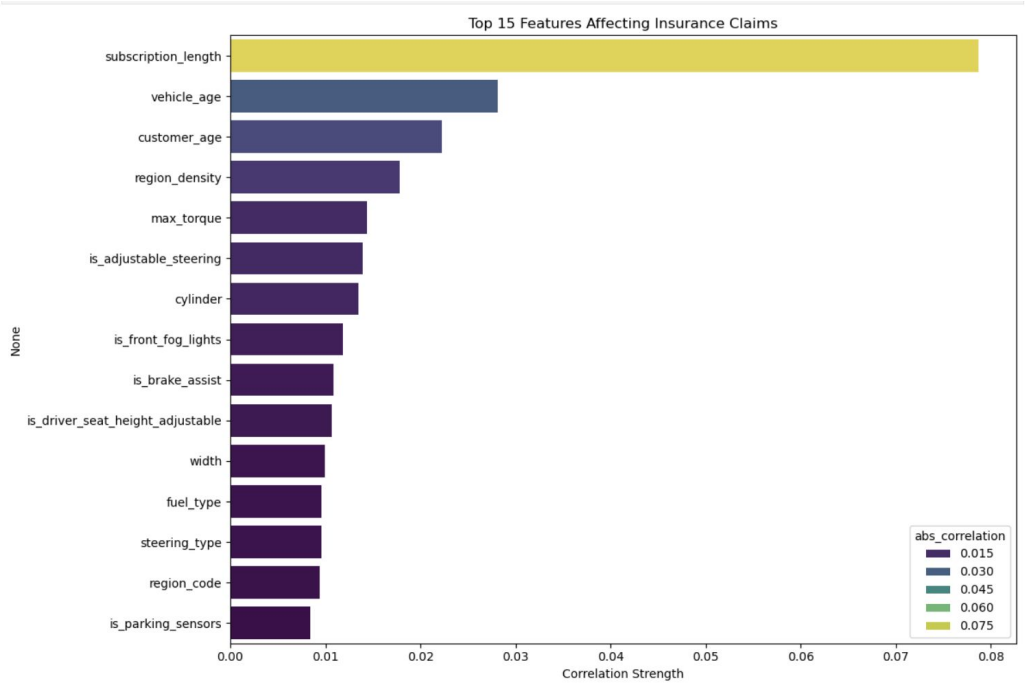
- Our dataset includes **multiple safety features**, such as **parking cameras**, **sensors**, and others.

Approach:

- To analyze their impact on **claim frequency**, we first examined the **correlation** between each safety feature and the **claim status**.

- Then, we **filtered the top five features** that showed the **strongest relationship** with claim behavior.

Bivariate EDA (all features analysis)



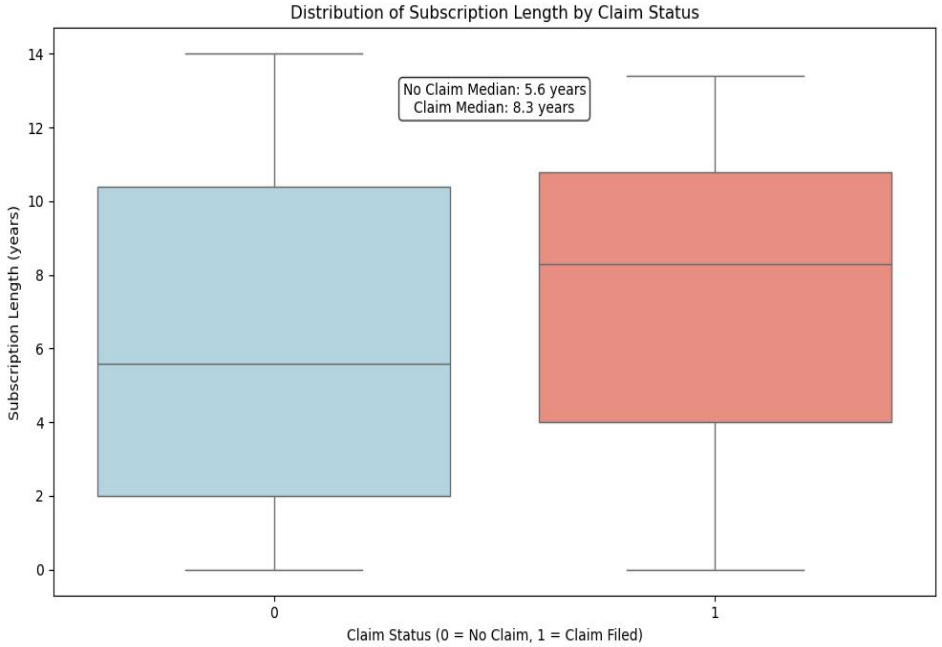
Observation:

- This analysis identifies which **features** have the **strongest influence** on **claim rates**.
- A **clear dominance** is observed for **subscription length**, showing the **highest correlation** with the target variable.

Insight:

- The **length of a customer's subscription** is **significantly more relevant** than other features.
- However, **vehicle age**, **customer age**, and **regional density** also show **notable correlations** with claim behavior.

Bivariate EDA (all features analysis)



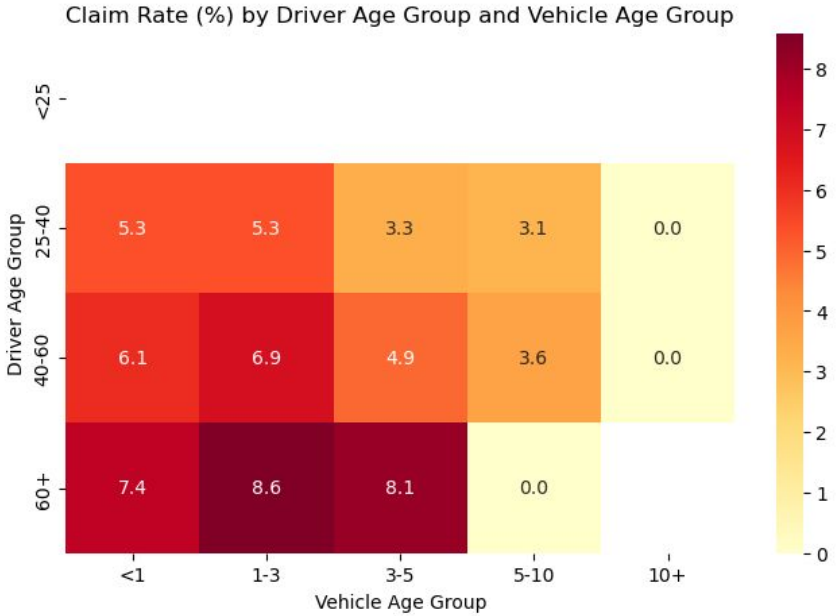
Median Comparison:

- Customers who **did not file a claim** have a **shorter median subscription length**.
- Customers who **filed a claim** tend to have a **longer median subscription length**.
- This suggests that the **likelihood of filing a claim increases** the longer a customer stays with the company.

Distribution Insight:

- For **no-claim customers**, the **boxplot is lower and narrower**, indicating that most **new or short-term subscribers** have **not yet filed claims**.

Bivariate EDA (all features analysis)



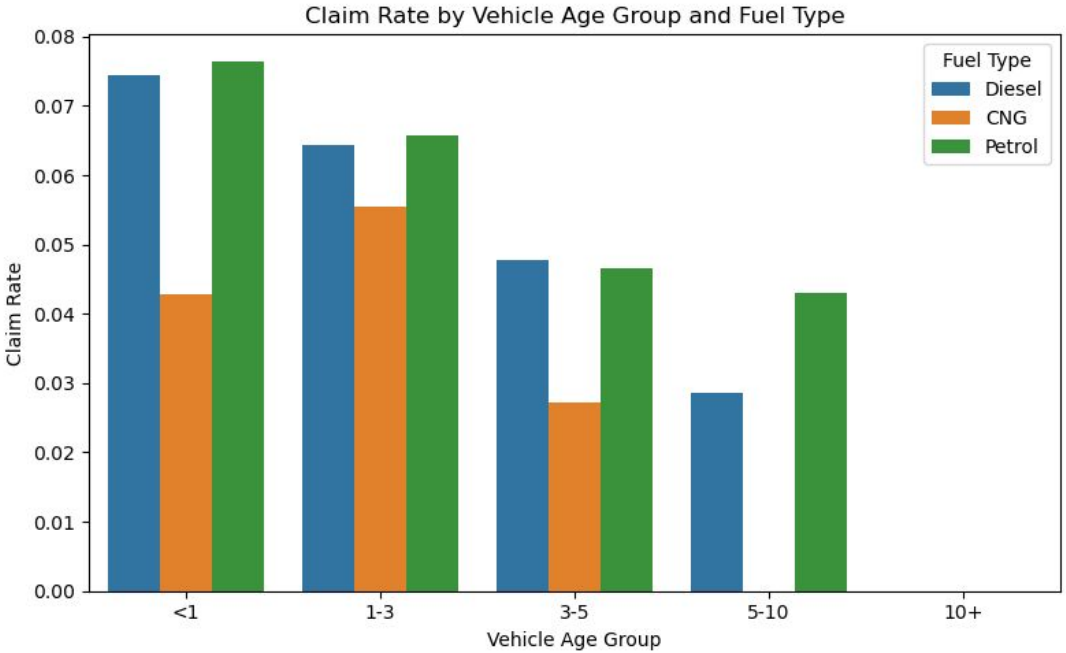
Observation:

- **Older drivers (60+)** show the **highest claim rates**, even when driving **newer vehicles (1–3 years old)**.
- **Younger and middle-aged drivers** maintain **lower claim rates**, even with **older vehicles (5–10 years old)**.

Insight:

- **Driver age** appears to be a **stronger predictor** of claim probability than **vehicle age**.

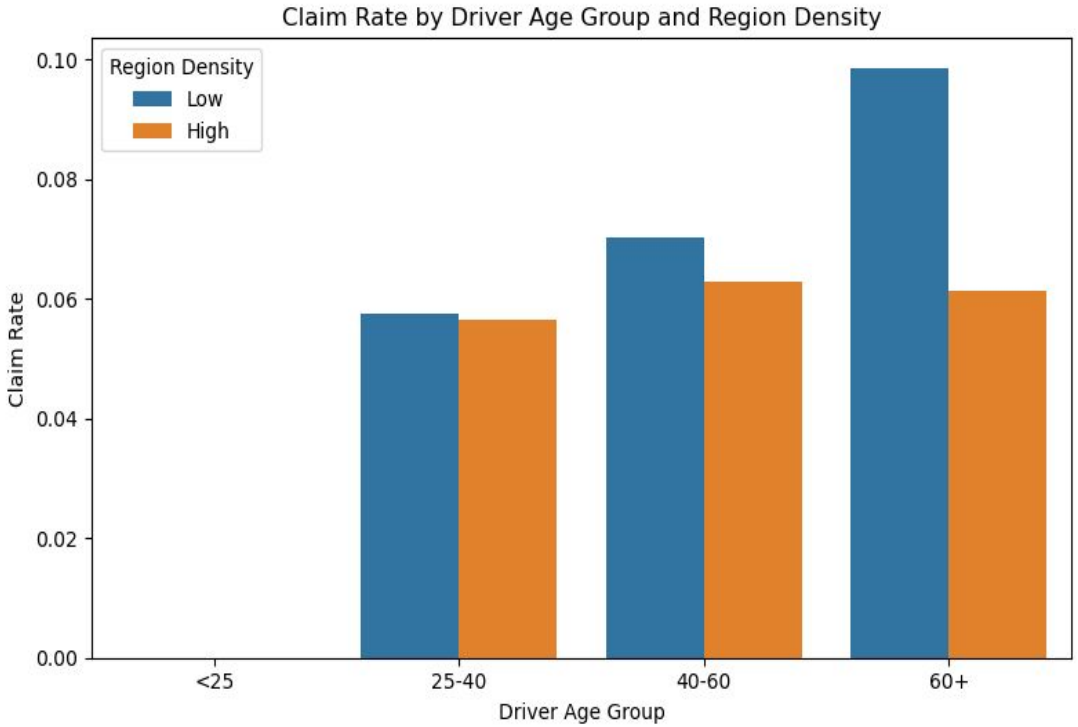
Bivariate EDA (all features analysis)



Insights:

- clearly **petrol vehicles** have the **highest claim rate** regardless of the vehicule age

Bivariate EDA (all features analysis)



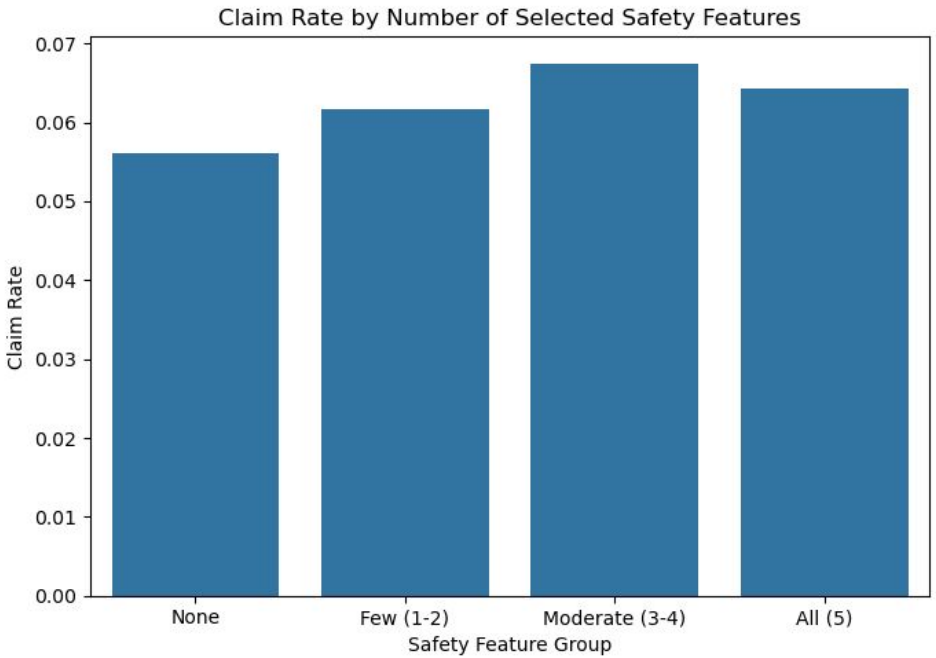
Observation:

- **Low-density regions** (rural or suburban) often have **higher claim severity** due to road conditions and driving patterns.

Insights:

- **Higher speeds** on open roads can make accidents **more severe**.
- **Rural roads** may have **poor lighting**, **fewer signs**, increasing accident risk.
- **Drivers in rural areas** may be **less cautious**, unlike in **urban zones** where **congested traffic** results in **less incidents**

Bivariate EDA (all features analysis)



safety features : Adjustable steering, brake assist, front fog lights ,is driver seat height adjustable, steering type

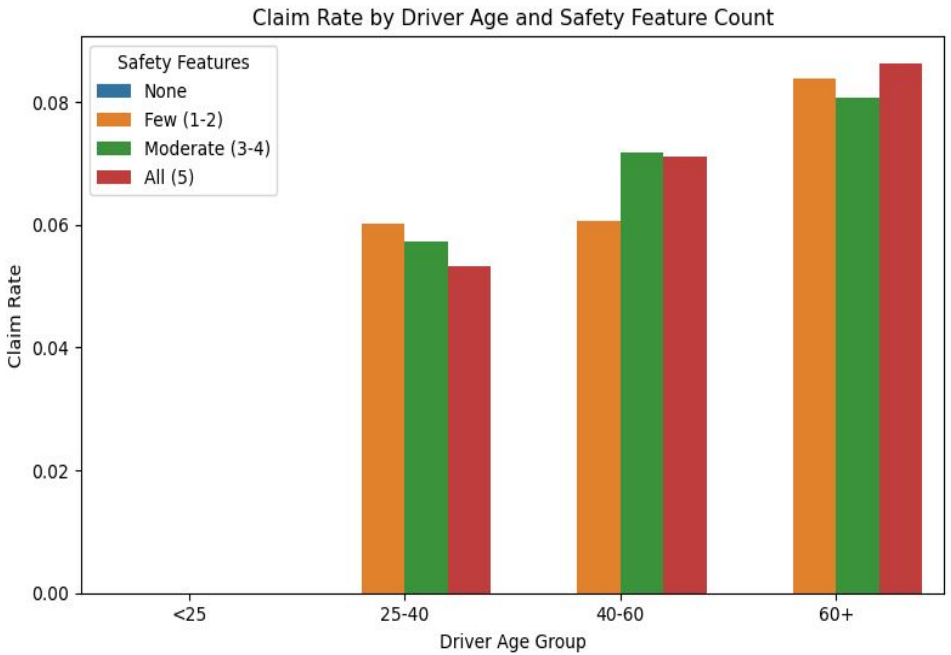
Observation:

- At first glance, the **claim rate by safety feature group** appears **inconsistent**.
- Surprisingly, vehicles equipped with **all available safety features** show **higher claim rates** than those with **fewer features**.

Next Step:

- To understand this pattern better, we analyze the relationship between **customer age** and **safety options**
  to see whether **age differences** explain the **unexpected trend**.

Bivariate EDA (all features analysis)



Observation:

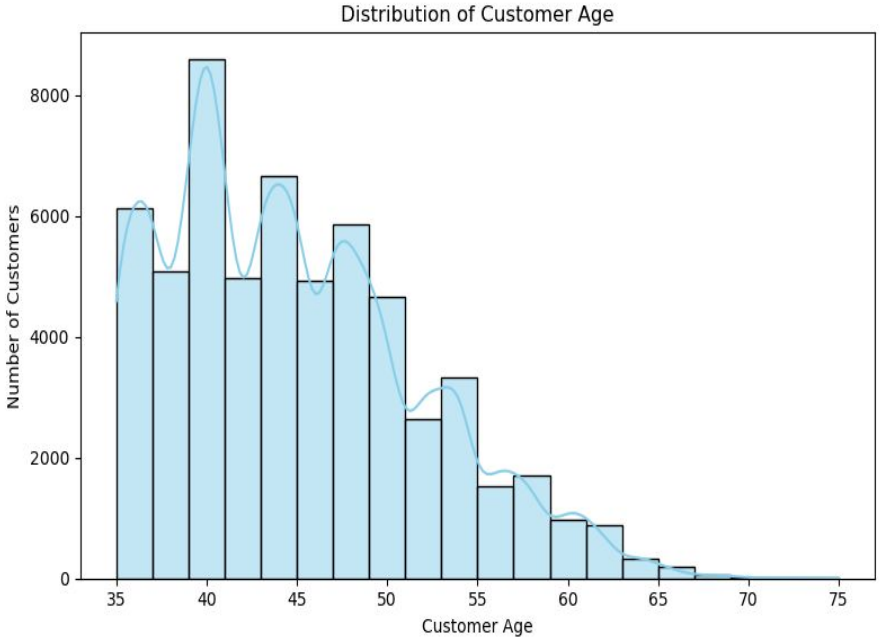
- **Young drivers** with **few safety features** tend to **file more claims**, likely due to **inexperience** and **riskier driving behavior**.
- When these drivers use vehicles with **advanced safety systems**, their **claim rates drop significantly**, showing that **technology helps reduce risk**.

Insight:

- For **older drivers**, even vehicles with **full safety packages** show **higher claim rates**.
- This confirms that **driver age** is a **stronger determinant** of claim likelihood than **safety features**.

Data distribution

Checking if the conclusions are consistent by getting a graph with the customer's age data description.



Mean Age: 44.82

Median Age: 44.00

Mode Age: 40.00

Skewness: 0.66

Data distribution insight (customer’s age)

· The average policyholder is about 45 years old.

· Half of all customers are younger than 44, half older indicating a fairly centered age spread.

· The most common age group in the dataset is around 40 years old.

· The distribution is positively skewed, meaning there are more younger and middle-aged customers, with a longer tail of older ages.

· The dataset is not dominated by older clients instead, it’s relatively balanced with a slight tilt toward younger drivers.

· This means the age–risk relationship is real, not just an artifact of who’s in our dataset.

· Even though there are fewer older policyholders, they still account for a disproportionate share of claims, confirming that age is a genuine behavioral risk factor

Summary

Highest-Risk Profiles

- Older drivers (40+) living in low-density regions
- Driving petrol vehicles, even with full safety features
- Young drivers without safety options

Summary

Lower-Risk Profiles

- Younger drivers with modern vehicles
- Equipped with multiple safety systems

Summary

Implications for the Insurer

- Price policies more accurately:
    a. Apply higher premiums for new or short-term customers
- Reward loyalty:
    a. Offer discounts to long-term policyholders, who represent lower risk