EXPLORATORY ANALYSIS OF CAR INSURANCE CLAIMS PREDICTION DATA

Group 2

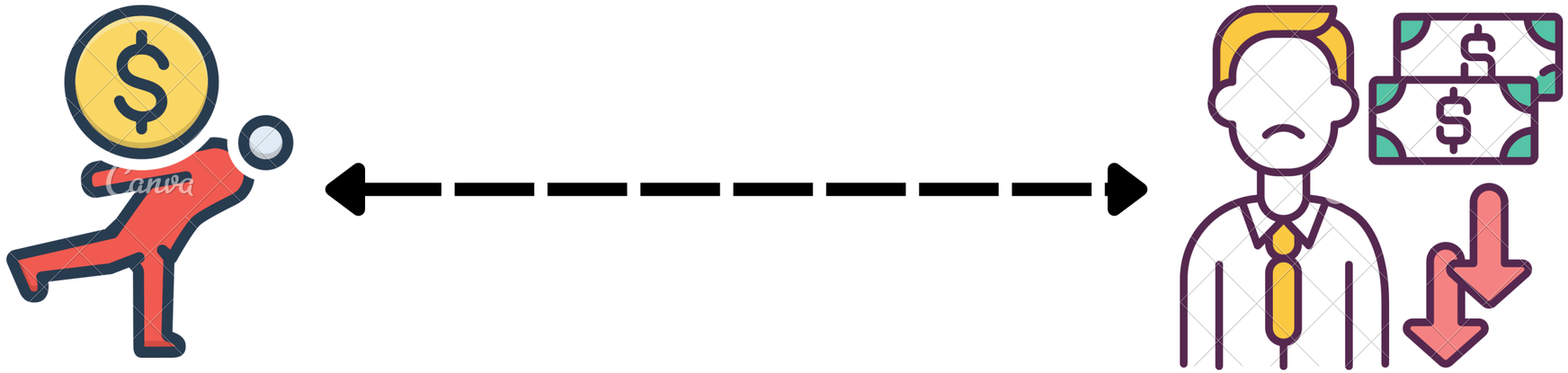
Chami Sewwandi-s16028

W.K.Hiruni Hasara-s16210

Tishani Wijekoon - s16379

S.Luxan - s16329

WHAT IS THE PURPOSE OF PREDICTING CAR INSURANCE CLAIMS?



- ✅ Offer better pricing for customers.
- ✅ Reduce financial losses.
- ✅ Detect fraud.
- ✅ Improve the overall claims process.

WHAT ARE THE QUESTIONS THAT WE HOPE TO ANSWER?

"What key factors influence the likelihood of a car insurance claim being filed within the next 6 months?"

DESCRIPTION OF THE DATASET

📊 Total Observations: 58,592
📂 Total Variables: 44 (15 numerical, 29 categorical)

Category	Key Features
Demographics	age_of_policyholder, area_cluster
Policy Details	policy_tenure, is_claim
Vehicle Information	model, segment, fuel_type, engine_type, airbags

DATA PREPROCESSING

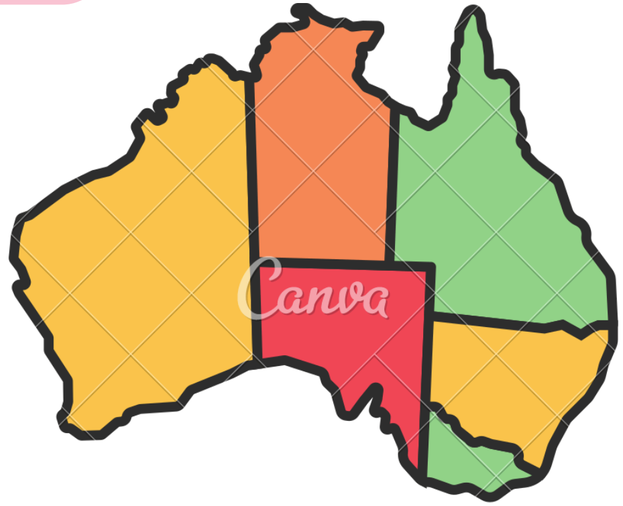
- Checked for missing values
- Encoded to categorical variables
- Handled redundant variables
- Feature selection

CHECKED FOR MISSING VALUES, DUPLICATES AND OUTLIERS

✅ Missing Values: No missing values found.
✅ Duplicates: No duplicate records detected.
✅ Outliers: Didn’t remove since they might explain some inherent variability attributed with claiming

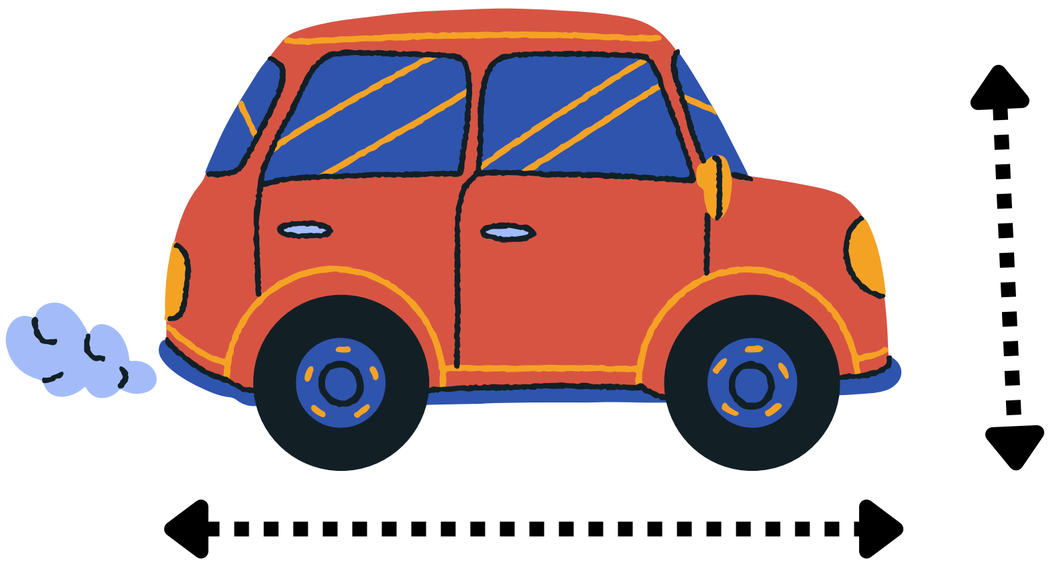
RECAREGORIZED SOME CATEGORICAL VARIABLES

area_cluster	Population density of the Policyholder city
C1, C2, ..., C20	C1 - 4990, C2-27003, ..., C20- 20905



“Low”, “Medium”, “High”

HANDLED REDUNDANT VARIABLES



“length”, “width”, “height”, “gross_weight”

“Segment of the car”

Length, width, height, and weight are already captured by the car segment

HANDLED REDUNDANT VARIABLES

- We observed each car model has a unique set of specifications.
- We used Cramér's V to measure the association between "model" and car specifications.
- With a value of 1 for each pair, the results indicate a perfect association, meaning the "model" variable contains complete information about the car specifications.

```
> # View results
> print(cramer_results)
                                  make                                segment
                                     1                                      1
                                 model                              fuel_type
                                     1                                      1
                           engine_type                                airbags
                                     1                                      1
                      rear_brakes_type                               cylinder
                                     1                                      1
                     transmission_type                               gear_box
                                     1                                      1
                         steering_type                             ncap_rating
                                     1                                      1
                            max_torque                              max_power
                                     1                                      1
                                is_esc                  is_adjustable_steering
                                     1                                      1
                               is_tpms                      is_parking_sensors
                                     1                                      1
                    is_parking_camera                      is_front_fog_lights
                                     1                                      1
                  is_rear_window_wiper                    is_rear_window_washer
                                     1                                      1
                is_rear_window_defogger                        is_brake_assist
                                     1                                      1
                  is_power_door_locks                       is_central_locking
                                     1                                      1
                is_power_steering is_driver_seat_height_adjustable
                                     1                                      1
  is_day_night_rear_view_mirror                                         is_ecw
                                     1                                      1
                    is_speed_alert                                displacement
                                     1                                      1
```

FEATURE SELECTION

✅ Kept:

- **Policyholder factors**: Age and area cluster
- **Policy details**: How long they’ve had insurance.
- **Car details**: The model and car age.

❌ Removed:

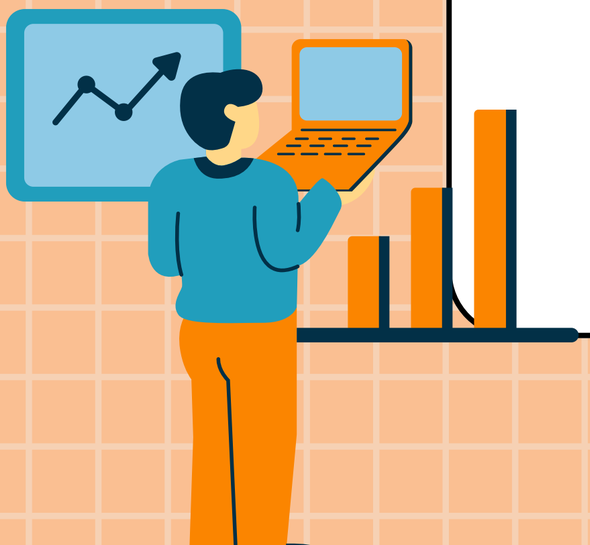
- Individual car specifications (since they are already included in the model).

EXPLORATORY DATA ANALYSIS

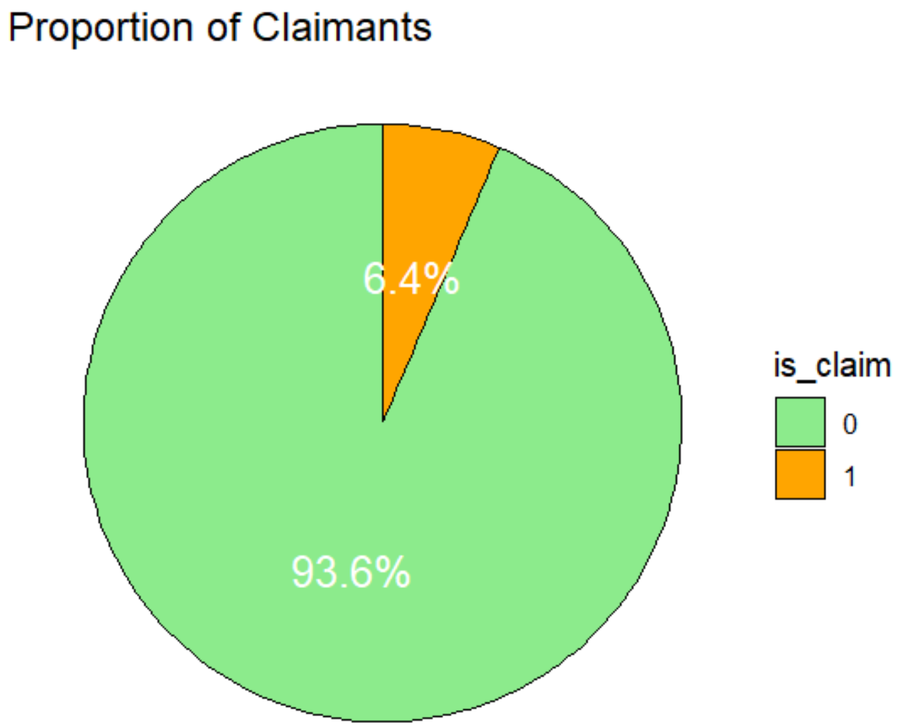
1. UNIVARIATE DATA ANALYSIS
2. BIVARIATE DATA ANALYSIS
3. FACTOR ANALYSIS FOR MIXED DATA

UNIVARIATE DATA ANALYSIS

In this section, we will examine the distribution and characteristics of significant individual variables in the dataset and identify potential issues.



RESPONSE VARIABLE-CLAIM STATUS



📌 **Clear Imbalance in Data**

📌 **Possible Reasons**

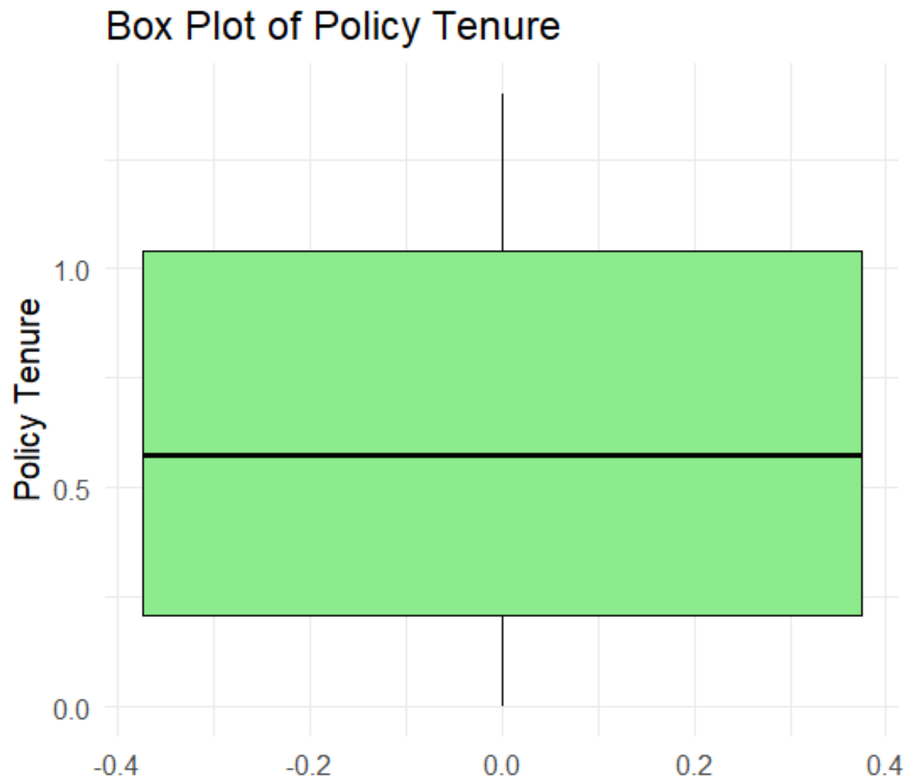
- Safe driving
- Strict policy terms
- Effective fraud prevention

📌 **Impact**

- Biases model predictions

💡 "HANDLING IMBALANCE NEEDS TAILORED MACHINE LEARNING TECHNIQUES"
BARAN AND ROLA (2022)

DISTRIBUTION OF POLICY TENURE



📌 **Average Car Insurance Policy Duration**

- Most policies last around 0.57 years ⏳
- Moderate variability observed

📌 **Interquartile Range**

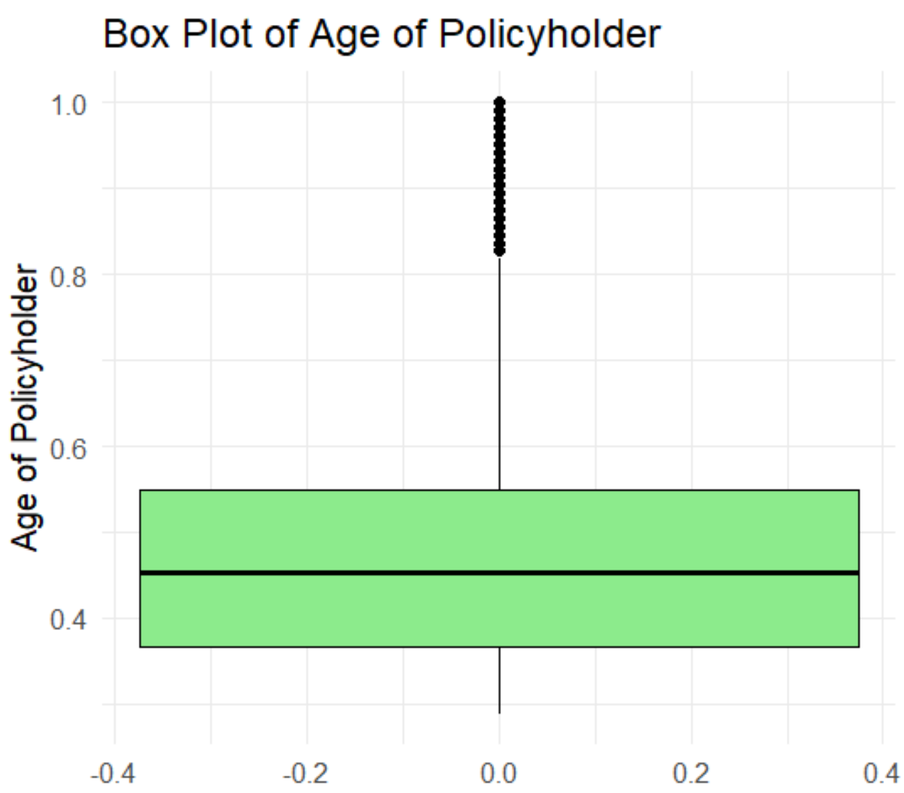
- Captures the middle 50% of policy durations

📌 **Possible Reasons**

- Standard contract lengths
- Customer retention strategies

💡 "POLICY TENURE IS A KEY FACTOR IN PREDICTING CUSTOMER CLAIM STATUS"
ACTUARIAL RESEARCH(2019)

POLICYHOLDER DEMOGRAPHICS-AGE



📌 Right-Skewed Distribution

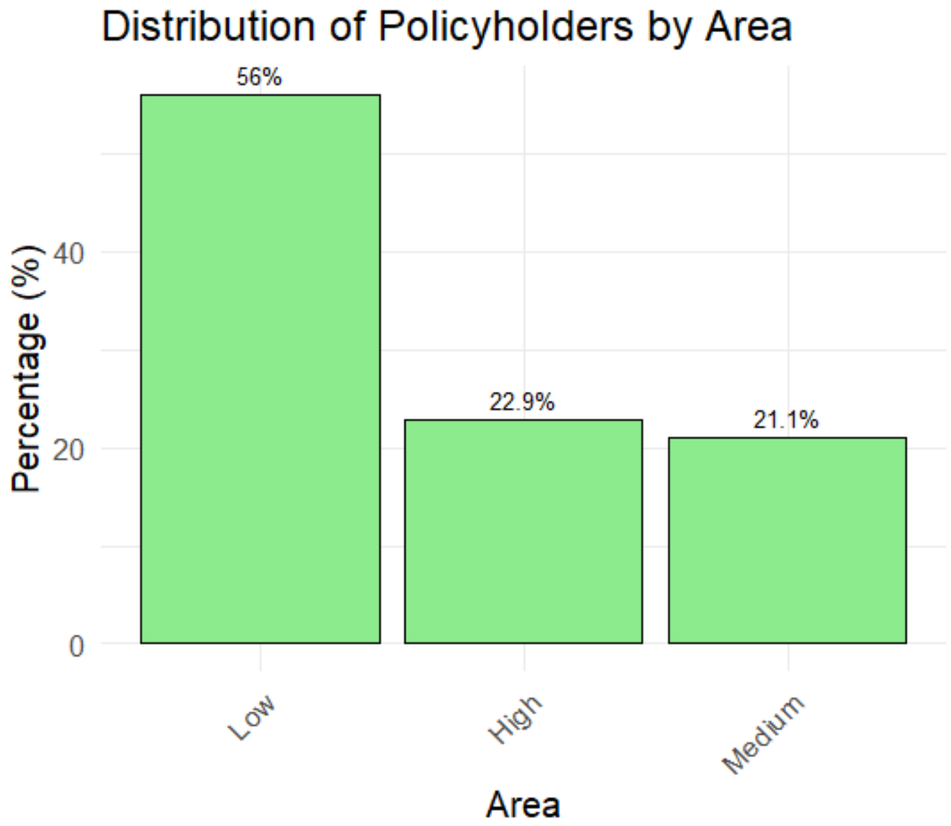
- Younger policyholders dominate
- A few older policyholders observed

📌 Possible Reasons

- Higher premiums for older policyholders 💸
- Younger individuals prefer affordable options
- Employer-sponsored coverage for younger employees

💡 "AGE IS A KEY RISK FACTOR IN PREMIUM CALCULATIONS"
ABDULLAHS, RPUBS (2021)

POLICYHOLDER DEMOGRAPHICS-AREA



📌 Policyholder Density

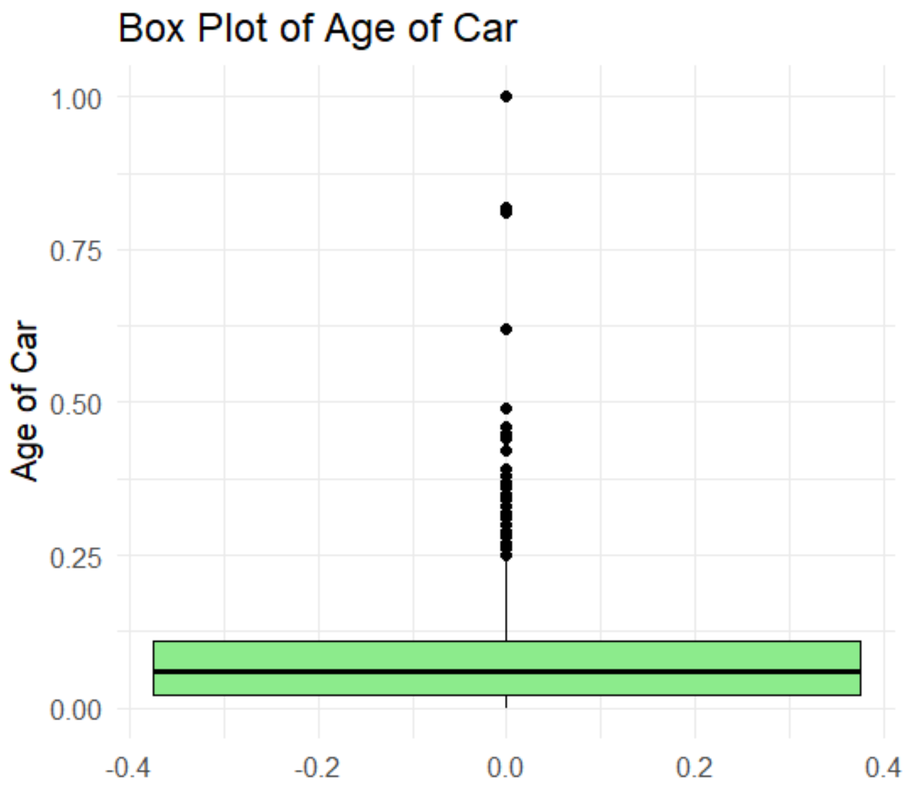
- 56% in low-density areas
- Smaller proportions in medium & high-density areas

📌 Possible Reasons

- Living costs & regional insurance access 🌍
- Targeted marketing shaping concentration 🎯

💡 "REGIONAL LOCATION PLAYS A KEY ROLE IN RISK ASSESSMENT"
HARINGA (2020)

CAR SPECIFICATIONS-AGE OF CAR



📌 Car Age Distribution

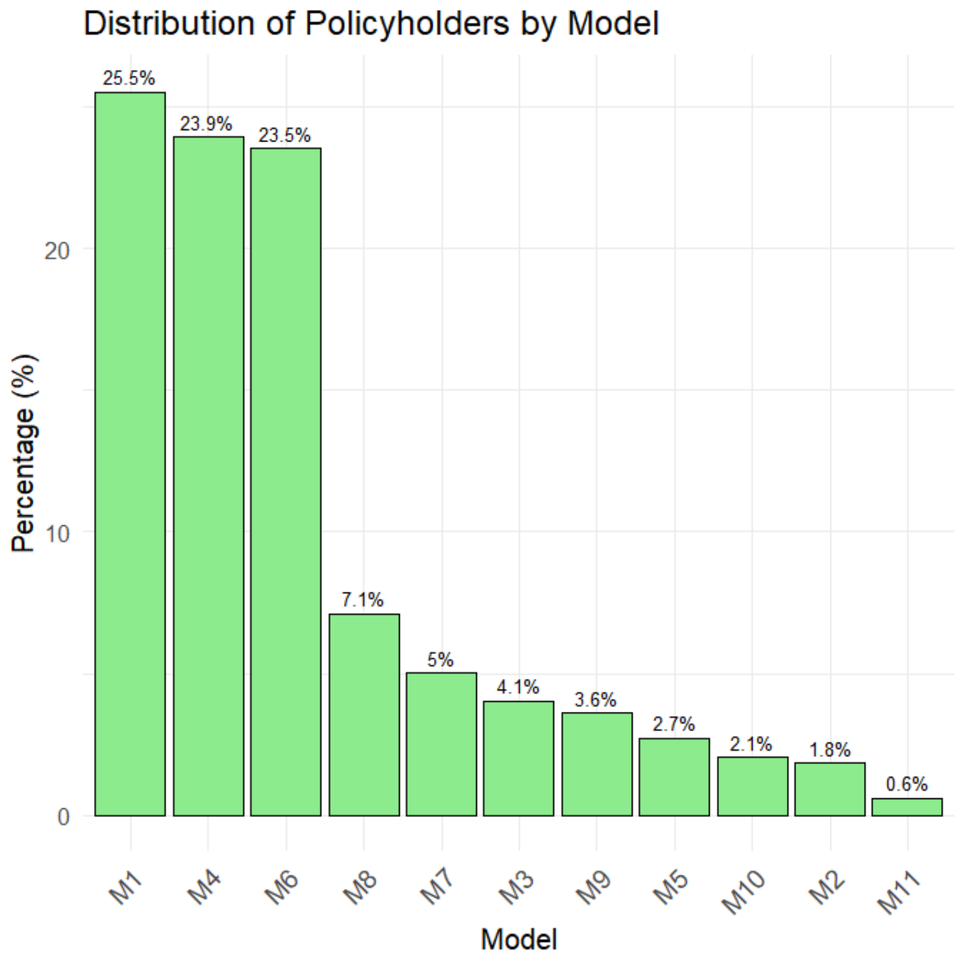
- Most cars are new, concentrated around 0-0.2 years 🚗
- Right-skewed distribution with a few older cars as outliers ⏳

📌 Possible Reasons

- New cars may not need insurance yet (warranty coverage) 🔧
- Insurers prefer newer vehicles with lower claim rates

💡 “OLDER CARS HAVE SIGNIFICANTLY HIGHER CLAIM RATES”
SCIELO BRAZIL (2018)

CAR SPECIFICATIONS-MODEL TYPE



📌 Dominant Car Models

- M1, M4, M6 make up 75% of policies 🚗
- Likely favored for affordability, safety, & brand recognition

📌 Less Popular Models

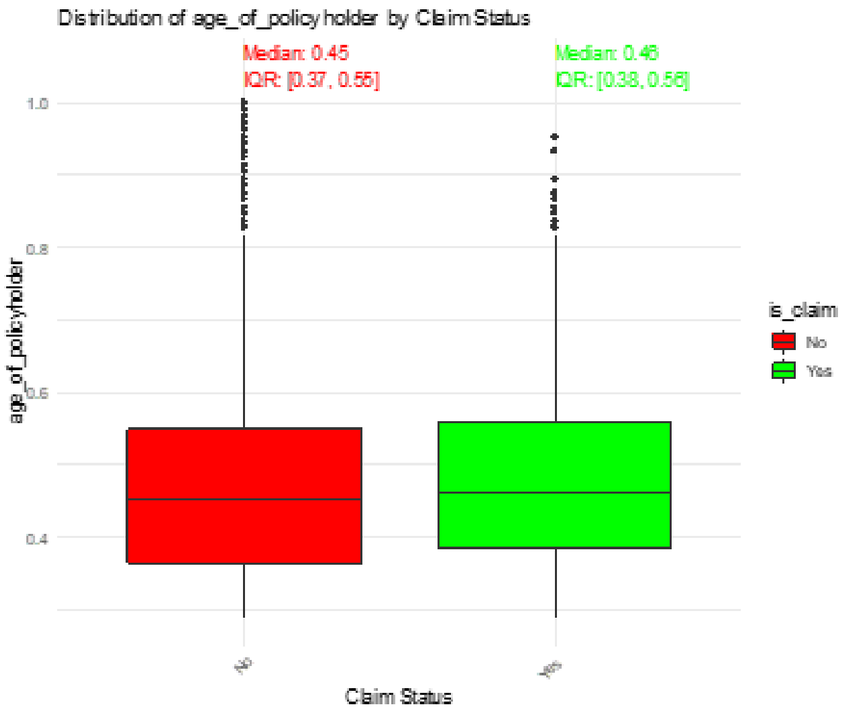
- May have higher prices, limited availability, or lower appeal

💡 “POPULAR MODELS INFLUENCE CLAIM FREQUENCY AND OVERALL RISK PROFILING” INSURANCE RISK STUDY (2020)

BIVARIATE DATA ANALYSIS

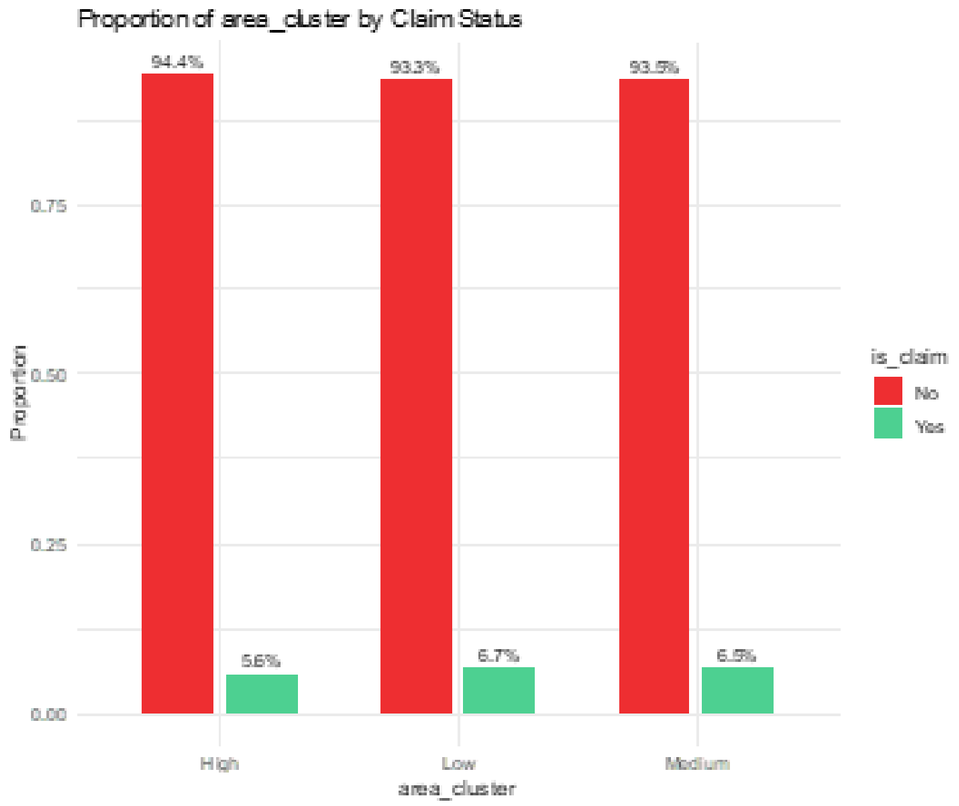
Here we explore the relationship between two variables to identify patterns, correlations, or associations.

DOES AGE MATTER? THE ANSWER MAY SURPRISE YOU!



- Boxplot comparing claim vs. non-claim age distribution.
- Median age of claimants and non claimants are nearly same.

MORE PEOPLE, MORE CLAIMS? NOT QUITE!

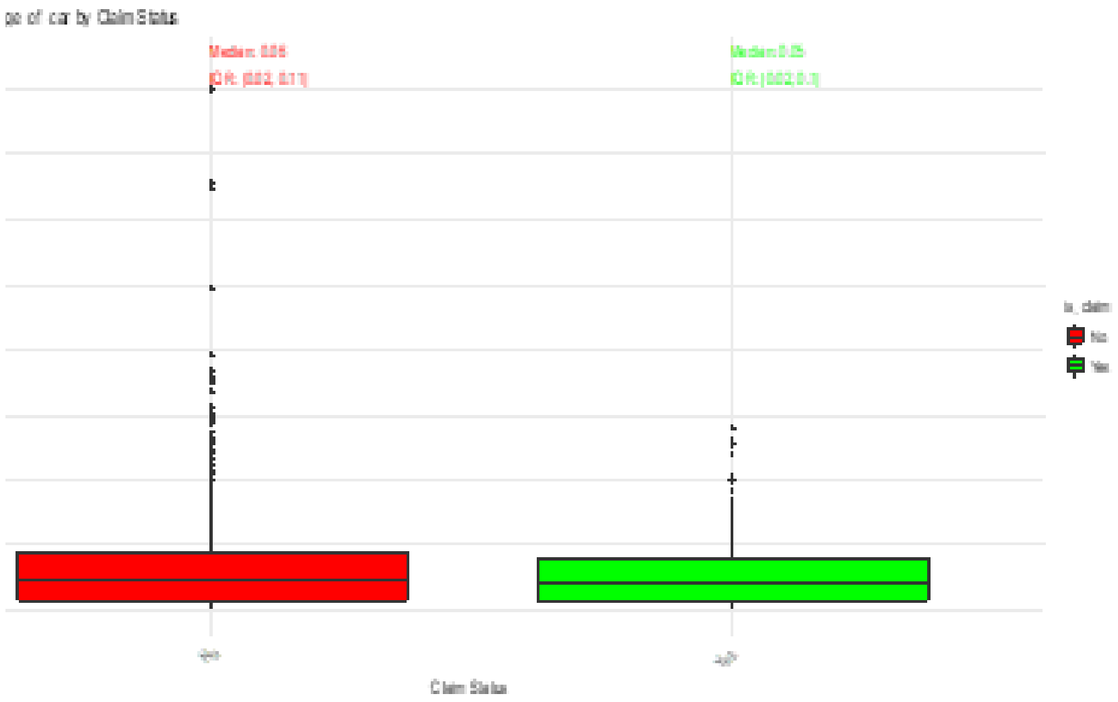


- A bar chart showing claim rates across different area densities.

- High-density areas: 5.6% claim rate
- Medium-density areas: 6.5% claim rate
- Low-density areas: 6.7% claim rate

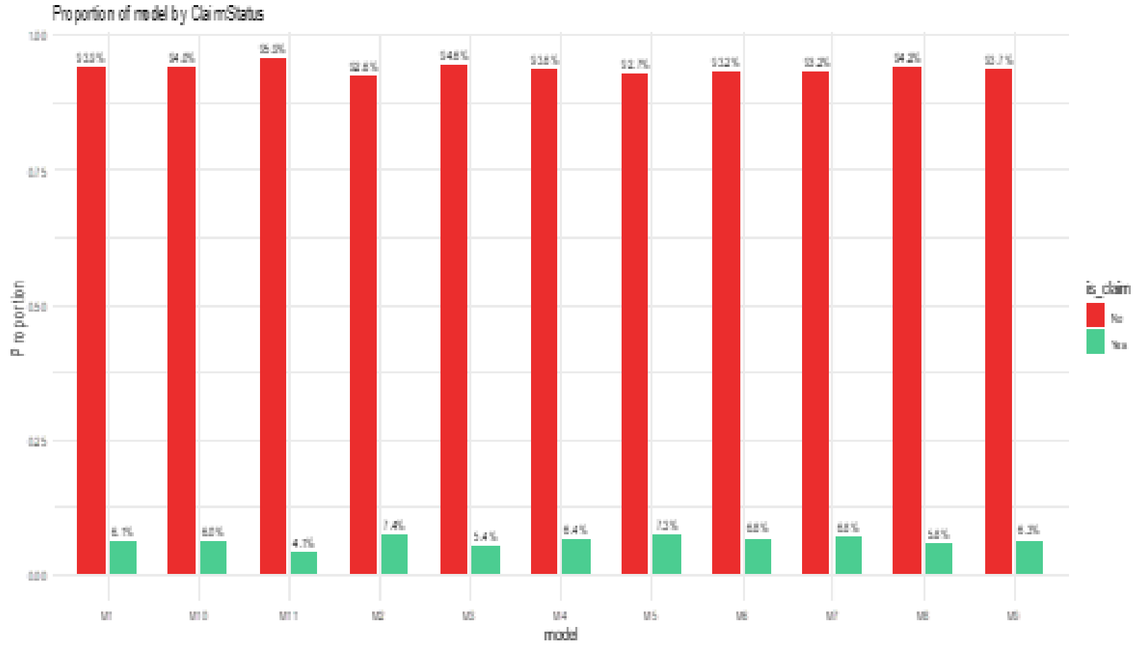
- Rural areas might lack competition, risk management, or insurance options.

DOES AN OLDER CAR MEAN MORE CLAIMS? NOPE!



- Median car age for claims and non-claims = nearly identical.
- New cars claim accidents, old cars claim repairs—it balances out!

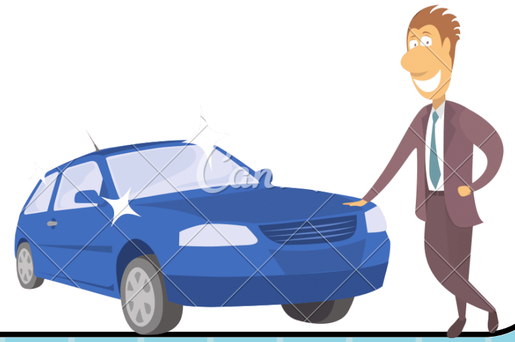
SOME CAR MODELS ARE "CLAIM MAGNETS"!



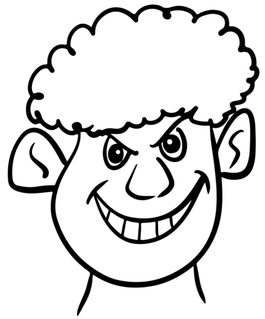
- Model 2 = highest claim rate (7.4%)
- Model 11 = lowest claim rate (4.1%)

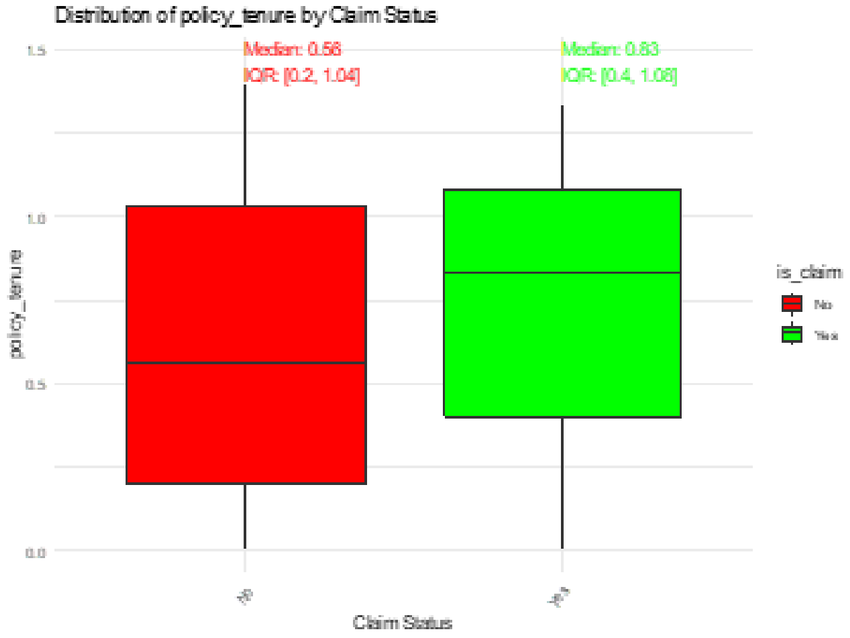
Grouping:

- <4.5% claim rate: Model 11
- 4.5%-6% claim rate: M1, M3, M8, M10
- 6%-7.5% claim rate: M4, M5, M6, M7, M9, M2



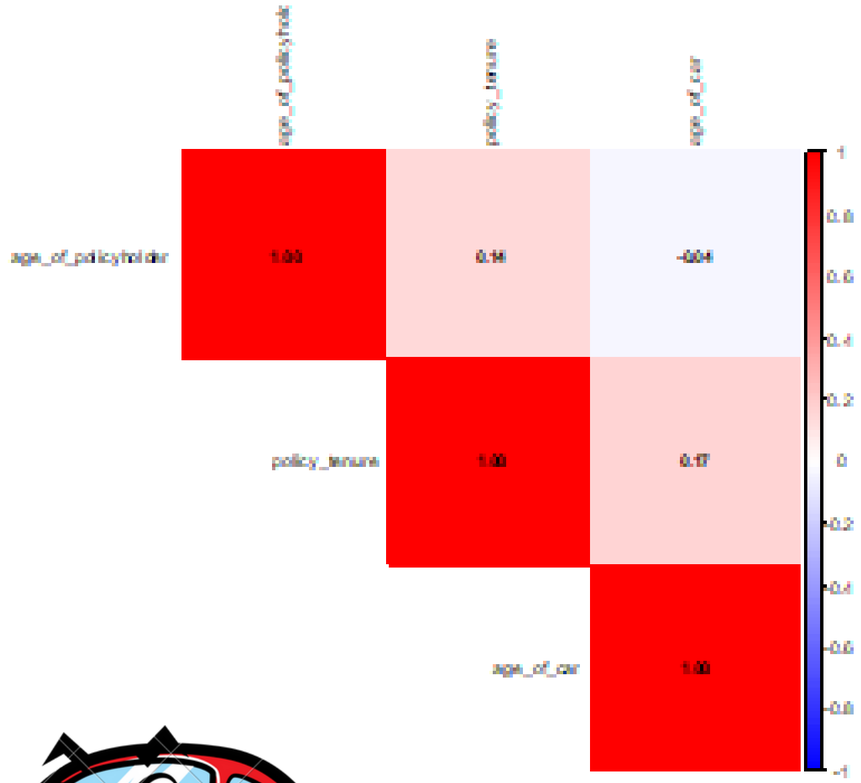
THE LONGER YOU STAY, THE MORE LIKELY YOU CLAIM!



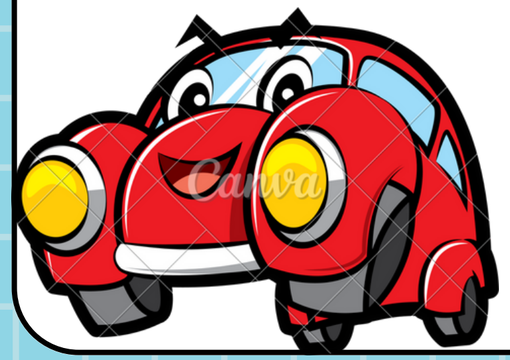


- Higher median tenure for claimants.
- Longer policy tenure may increase the likelihood of filing claims due to increased experience and incidents.
- Shorter-tenure policies less likely to result in claims, potentially due to lower-risk customers.

CORRELATION HEAT MAP OF NUMERICAL VARIABLES

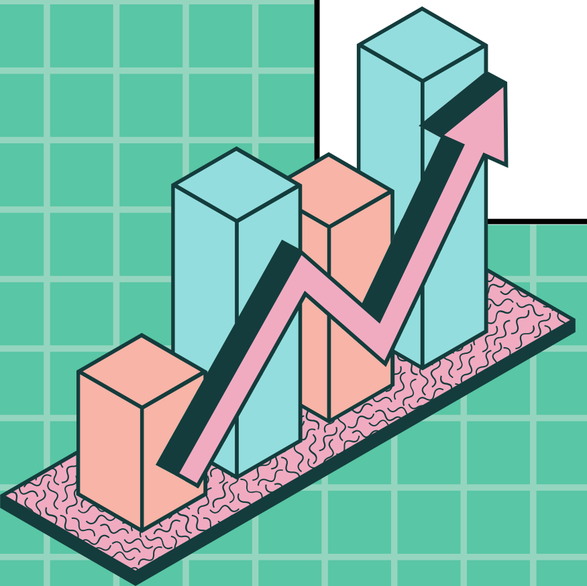


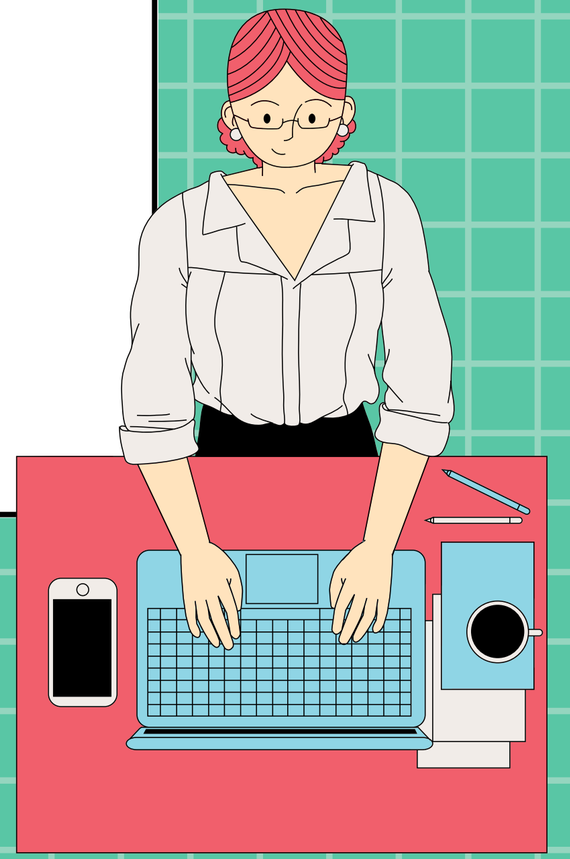
- Age of Policyholder vs. Policy Tenure: 0.14 (Weak Positive)
- Policy Tenure vs. Age of Car: 0.17 (Weak Positive)
- Age of Policyholder vs. Age of Car: -0.04 (Weak Negative)
- weak correlations indicate minimal dependency between these variables.



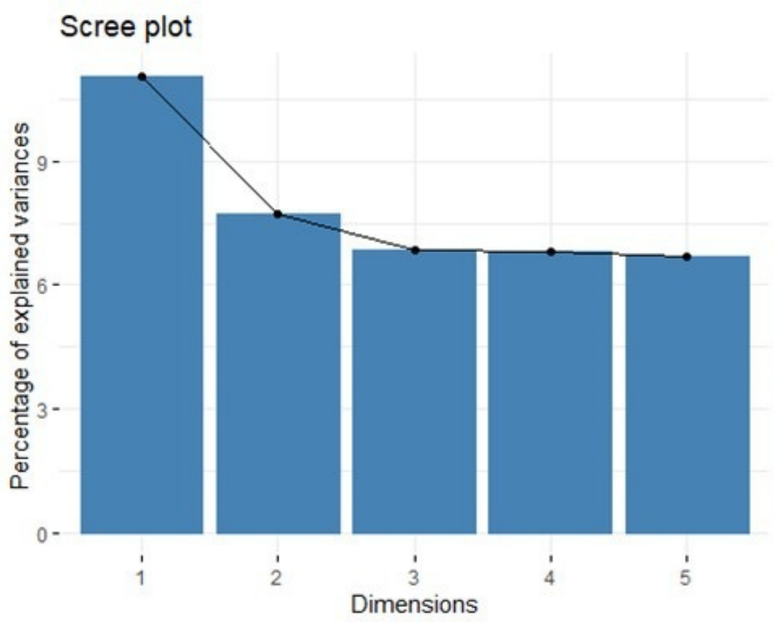
MULTIVARIATE ANALYSIS

Here we examine multiple variables simultaneously to understand their relationships and interactions.





SCREE PLOT OF THE PRICIPLE COMPONENTS



- Variability explained

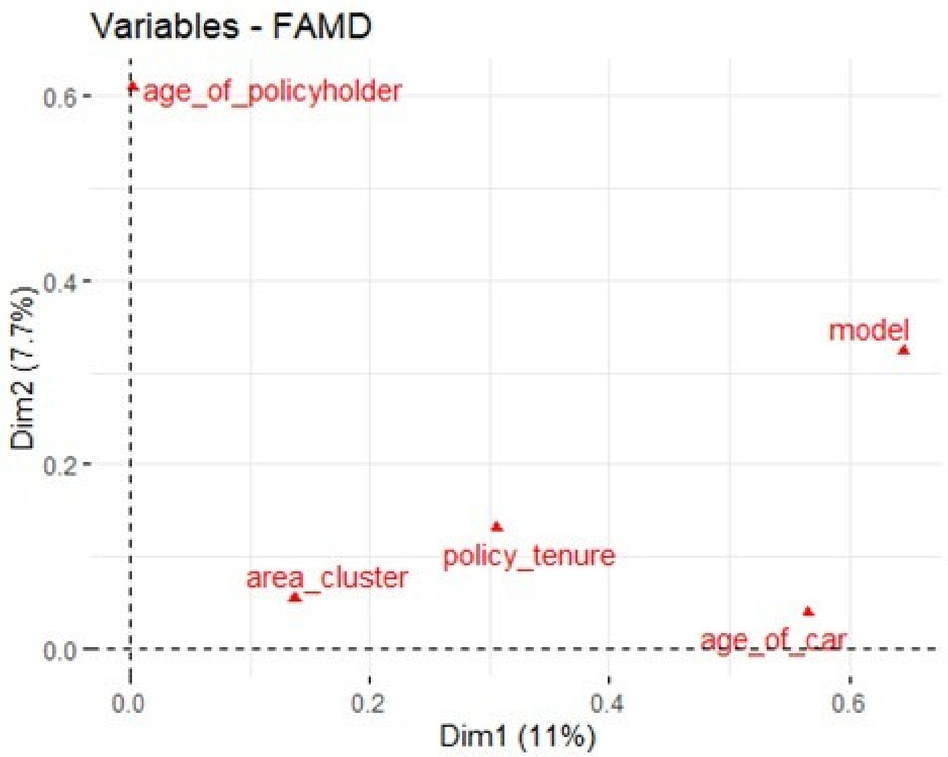
1. First component - 11.040%
2. Second component - 7.719%
3. Third component - 6.830%

- Elbow point can be observed at the third dimension.

- This indicates that the first three components capture only 25.590% of the variability.

- Captured variability is not enough for further analysis.

VARIABLE FACTOR MAP OF FAMD



- Explains how variables are correlated with first two principal components.
- Age of policy holder is highly correlated with second component
- Model is highly correlated with first component.
- Area cluster and policy tenure seems weakly correlated with both components.
- Age of car is highly correlated to first component and weakly with second component.

SUGGESTIONS FOR ADVANCED ANALYSIS TECHNIQUES

- Logistic Regression
  A baseline model to predict binary outcomes

- Decision Tree
  Handle non-linear relationships and visualize simple decision rules.

- Random Forest
  Reduce overfitting and capture complex feature interactions

- Gradient Boosting (XGBoost, LightGBM, CatBoost)
  High predictive accuracy by combining weak learners.

- Support Vector Machine(SVM)
  Effective separation of high-dimensional data

REFERENCES

- What does Cramer's V actually measure and tell us? (n.d.). Retrieved from www.reddit.com: https://www.reddit.com/r/explainlikeimfive/comments/a0igip/eli5_what_does_cramers_v_actually_measure_and/?rdt=52237
- Baran, S. (n.d.). Prediction of motor insurance claims occurrence as an imbalanced machine learning problem.
- Anwar, A. (2023). Predicting Likelihood That A Policyholder Will File A Claim In The Next Six Months. Medium. Retrieved from:
- https://medium.com/@ammasanaswar/predicting-likelihood-that-a-policyholder-will-file-a-claim-in-the-next-six-months-c5a322446505
- Reddy, Y. U. (2023). Car Insurance Claim Prediction Model. Medium. Retrieved from: https://medium.com/@yudayreddy1/car-insurance-claim-prediction-model-bff7f06a9a77
- GeeksforGeeks. (2023). What is Exploratory Data Analysis? Retrieved from: https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

THANK YOU!