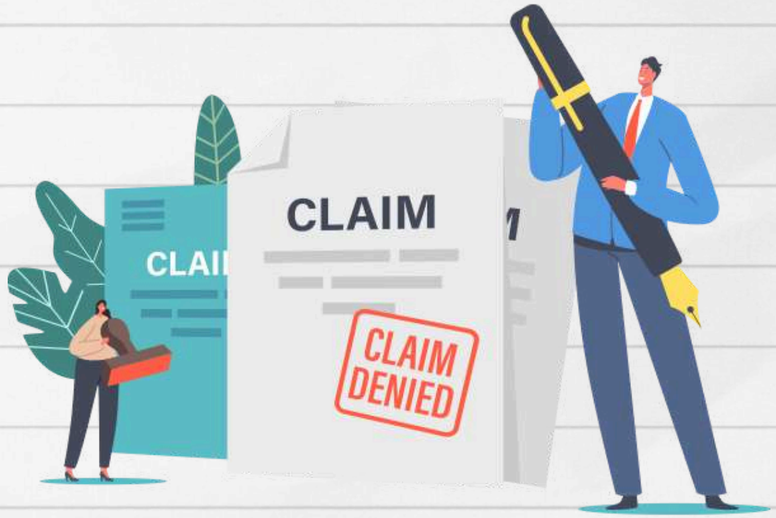
Group 3

Exploring Patterns in fraudulence Household Insurance claims

ACCORDING TO:



FBI

The total cost of insurance fraud (non-health insurance) is estimated to be more than **$40 billion per year.**

Insurance Fraud costs the average U.S. family between **$400 and $700 per year** in the form of increased premiums.

IFB

Loss due to insurance fraud in the **United Kingdom** is about **£1.5 billion** , causing a **5% increase** in insurance premiums.

OBJECTIVE

IDENTIFY THE FEATURES THAT DISTINGUISH FRAUDULENT CLAIMS FROM LEGITIMATE ONES.

HOW WE ARE GOING TO ACHIVE IT :

Using ***insurance_claims.csv***
consists of ***4415 records*** across ***21 variables.***

Through the application of,

- Descriptive statistics
- Data preprocessing
- Advanced analytical techniques.

Data Pre-Processing

PRE PROCESSING

Step 1

- remove the duplicate rows using claim id
- check whether if there any mising values

Step 2

- remove unnecessary identifiers
- Date columns are standardized

Step 3

- create new variables like age, policy duration.
- and remove dob and policy date columns.

Step 4

- all fraudulent claims (463) are retained while 500 non-fraudulent claims are randomly selected.
- Split the data set

Descriptive Analysis

Pearson's Correlation between numerical predictors

	income	Claim_amount	coverage	deductible	age	Policy_duration
income	1	0.57051	0.27046	0.24177	0.07235	0.07427
claim amount	0.57051	1	0.42337	0.53226	0.10899	0.15195
coverage	0.27046	0.42337	1	0.20418	0.05993	0.07259
deductible	0.24177	0.53226	0.20418	1	0.23565	0.24199
age	0.07235	0.10899	0.05993	0.23565	1	0.82866
Policy_duration	0.07427	0.15195	0.07259	0.24199	0.82866	1

- Age and policy duration (0.83) show the strongest association, while claim amount and Income (0.57) also have a notable link.
- Moderate correlations between claim amount, coverage (0.42), and deductible (0.53) suggest financial influences on claims.

Spearman's Rank Correlation for Response with numerical variables

	income	claim amount	coverage	deductible	age	policy duration
Fraudulent Claims	-0.00429	-0.0223	-0.02394	0.03708	-0.06972	-0.09885

Response Variable → Fraudulent claim

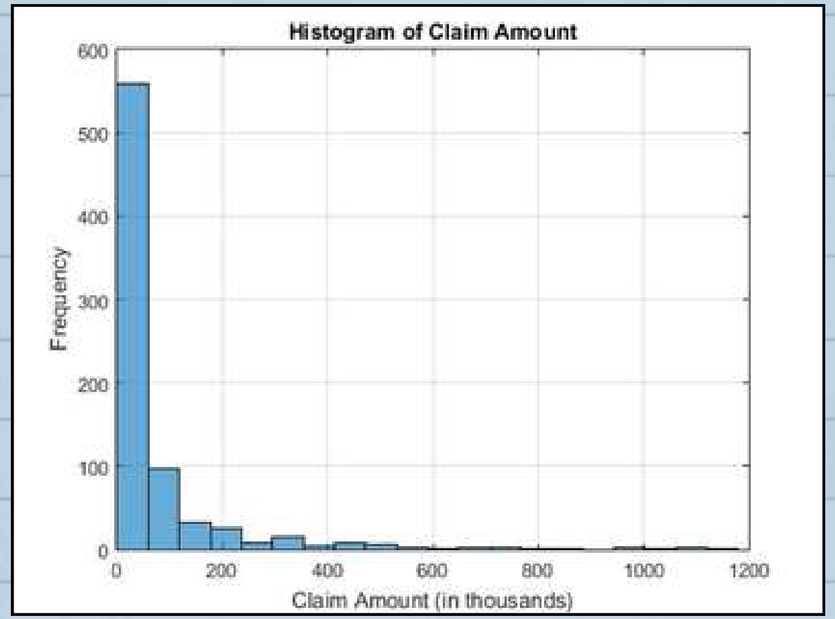
The numerical predictors show negligible correlation with fraudulence, suggesting weak associations.

Pearson's Chi squared test for Response with categorical variables.

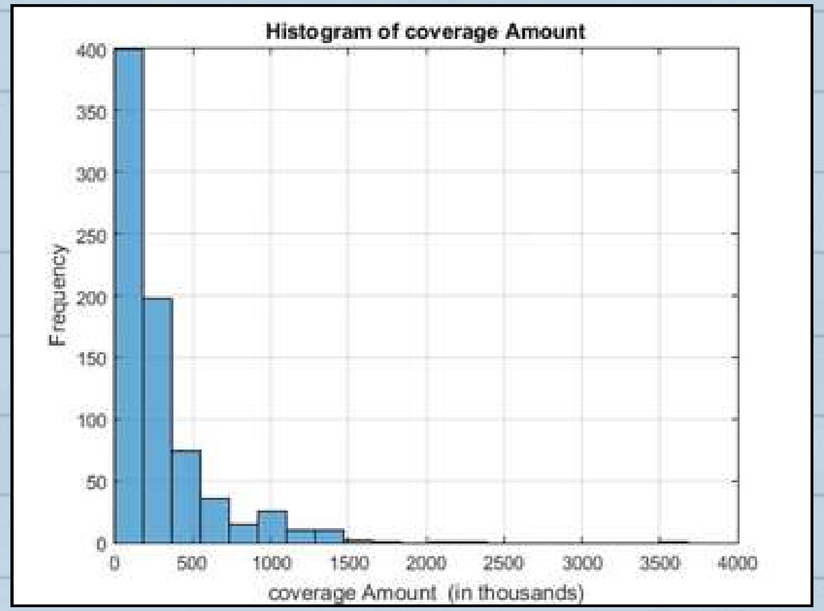
Categorical Variable	claim type	Uninhabitable	town size	gender	edcat	retire	marital	reside	primary residence
P Value	0.0607	0.0207	0.4363	0.8422	0.6076	0.0580	0.9606	0.0764	0.2177

There was a significant relationship only between the fraudulent response variable and the uninhabitable predictor variable.

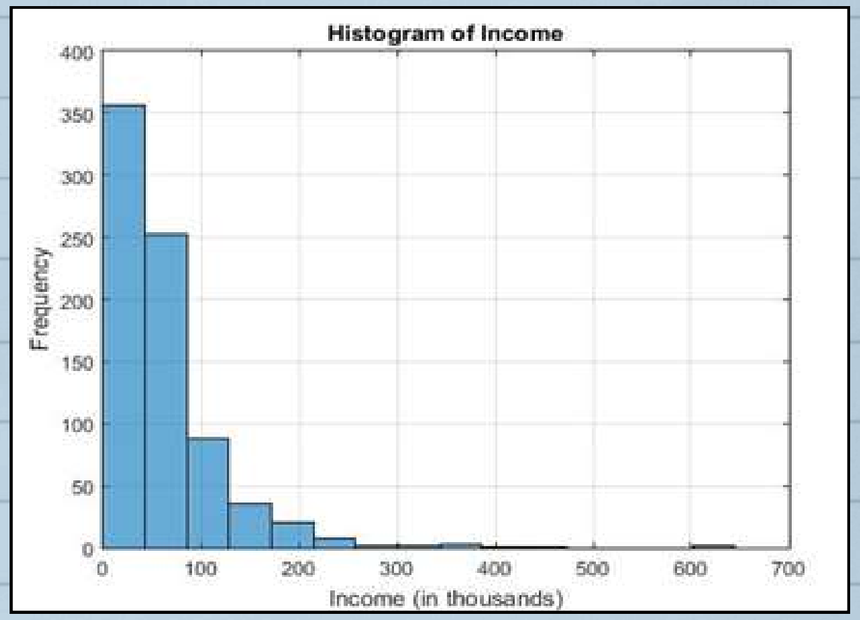
Univariate Analysis



Claim Amount



Coverage Amount



Income

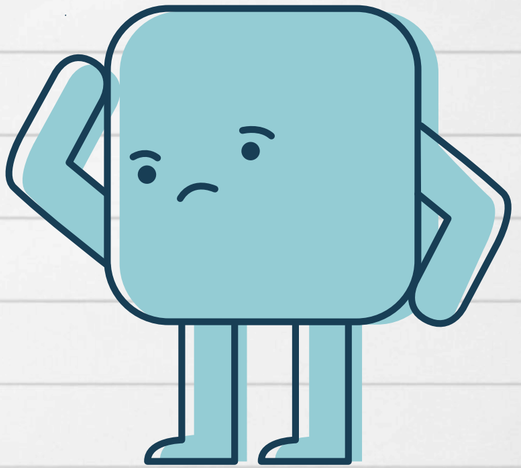
The income, claim amount, and coverage distributions exhibit negative skewness.

ARE NUMERICAL PREDICTORS SIGNIFICANT WITH FRAUDULENCE STATUS?

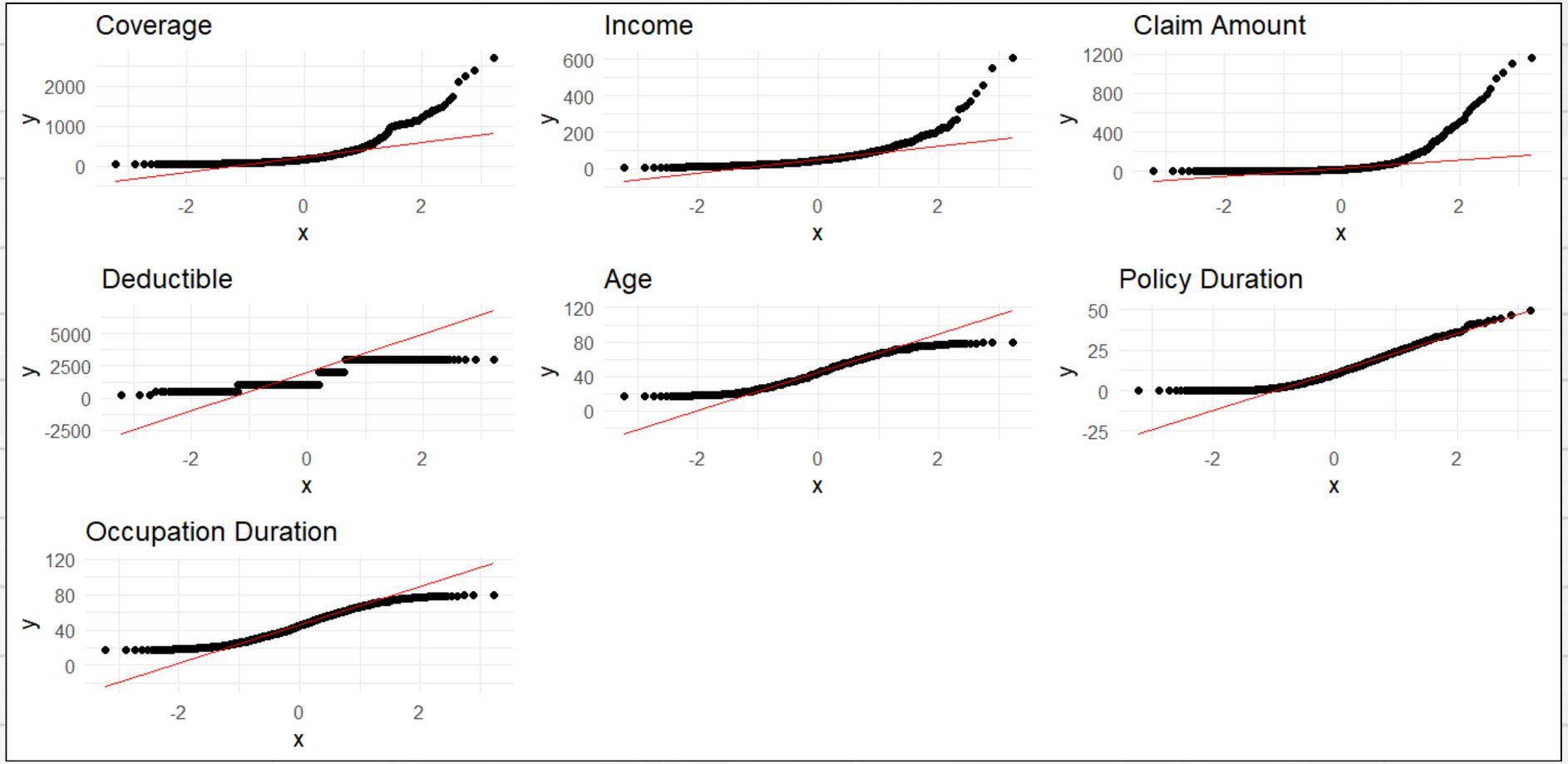
What Tests Can We Apply? — T test or Mann-Whitney Test

How To Decide Which One? — Check Normality Assumption

How To Check Normality? — Visual Methods or Statistical Tests



CHECKING NORMALITY WITH QQ PLOTS...



NORMALITY ASSUMPTION IS VIOLATED!

APPLYING MANN-WHITNEY U TEST...

Variable	p-value
Coverage	0.50663
Income	0.90537
Claim Amount	0.53615
Deductible	0.30366
Age	0.05307
Policy Duration	0.00609
Job Duration	0.05138

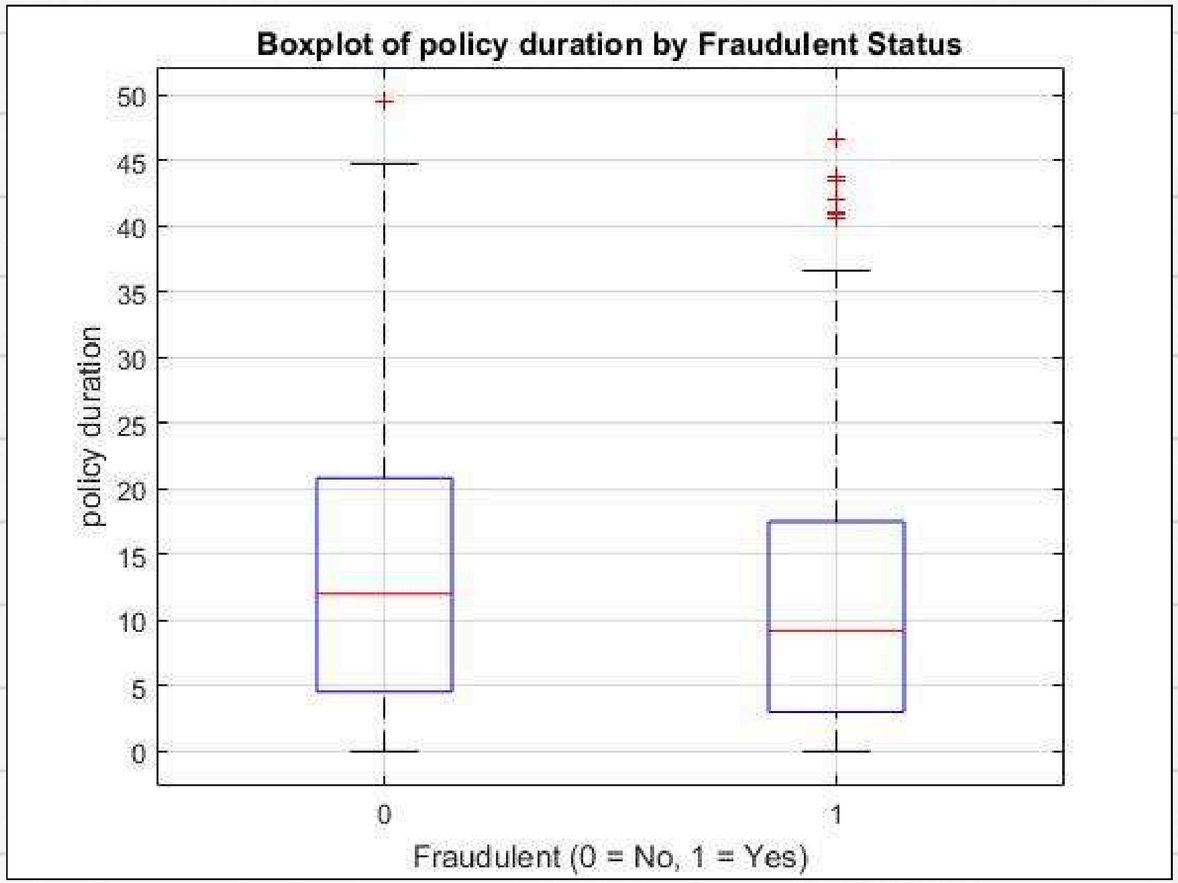
Null Hypothesis: There is no difference between the two groups

Alternative Hypothesis: There is a difference between the two groups

Reject Null Hypothesis if p-value < alpha level

POLICY DURATION IS SIGNIFICANT

FRAUDULENT STATUS WITH POLICY DURATION



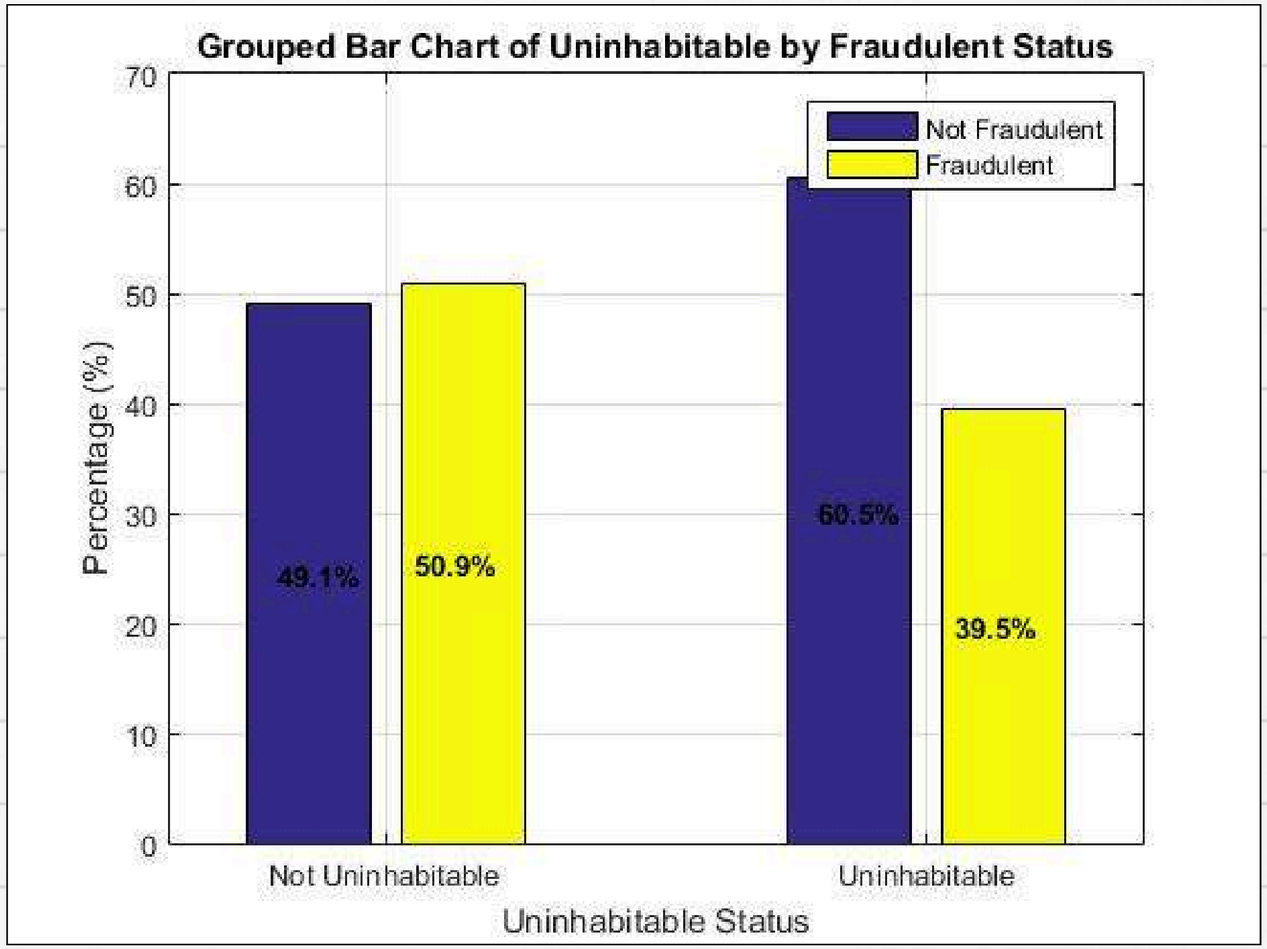
Observations:

- Significant Difference between Distributions.
- Presence of Outliers.

“ Fraudulent Claims can occur across various policy durations, affecting most short-term insurers.”

- Just Money website (2017) -

FRAUDULENT STATUS WITH HABITAT STATUS



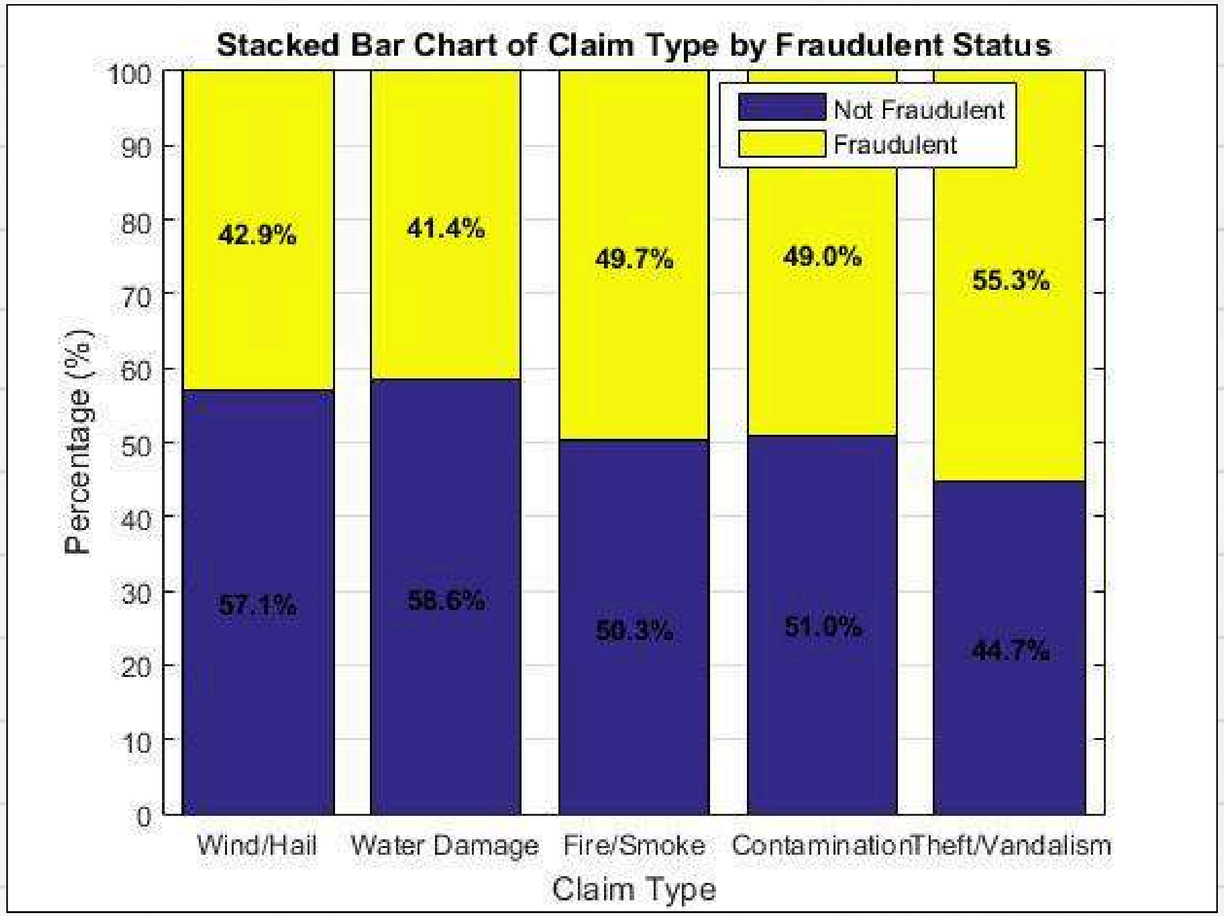
Observations:

- Significant Difference between Percentages of Fraudulent and Non-Fraudulent Claims.
- Chi-Square Test Results Confirm Significance.

"Vacant Properties are often targeted for fraud because they are Unoccupied."

- Core Title Website (2023) -

FRAUDULENT STATUS WITH CLAIM TYPE



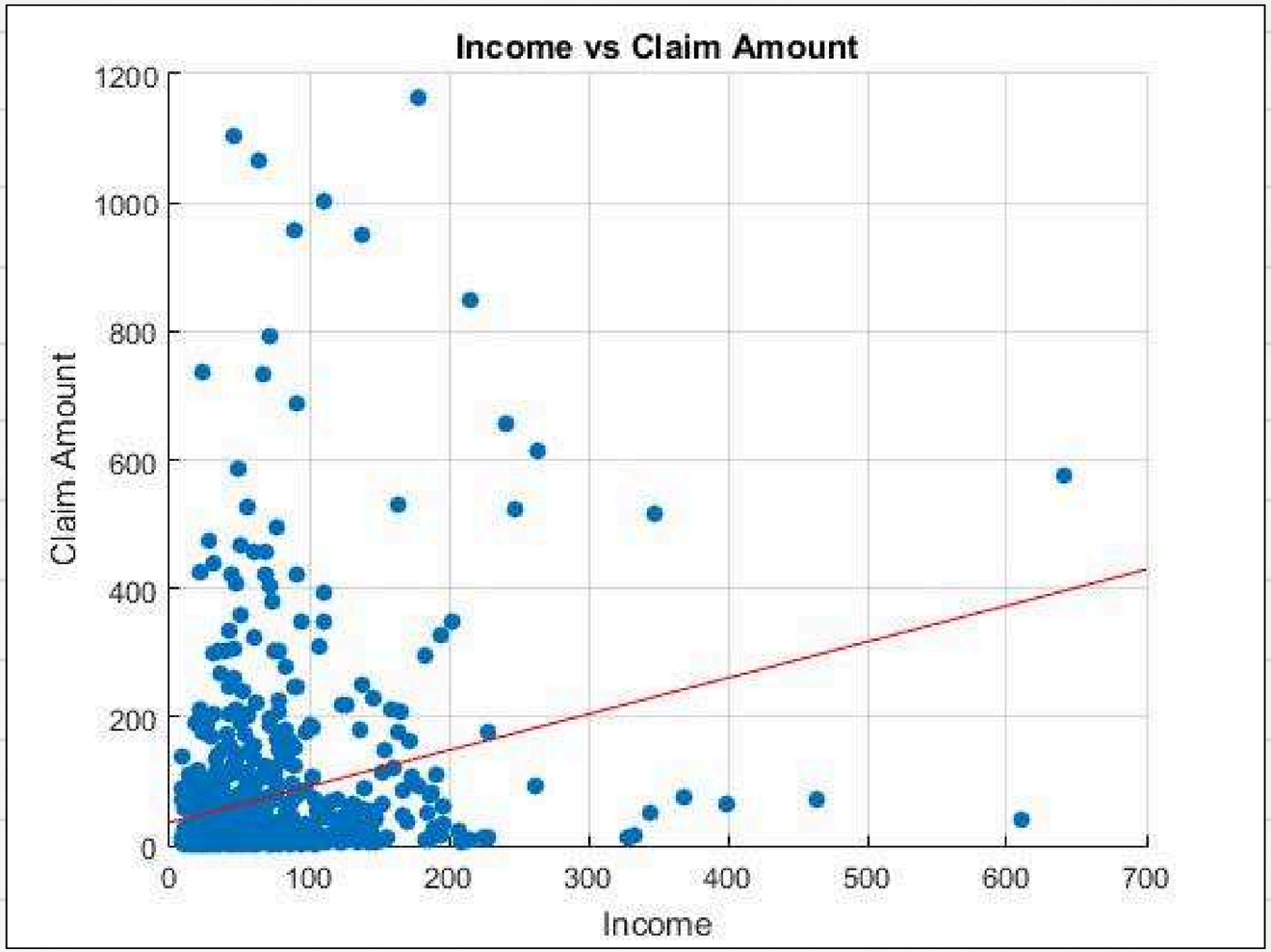
Observations:

- More than 50% of Theft/ Vandalism insurance results in Fraudulent Claims.
- Insurance associated with Natural Disasters result in less Fraudulent Claims.

"Fire and theft/vandalism are two of the most common types of insurance fraud."

- BURTON COPELAND website (2023) -

ASSOCIATION BETWEEN NUMERICAL PREDICTORS

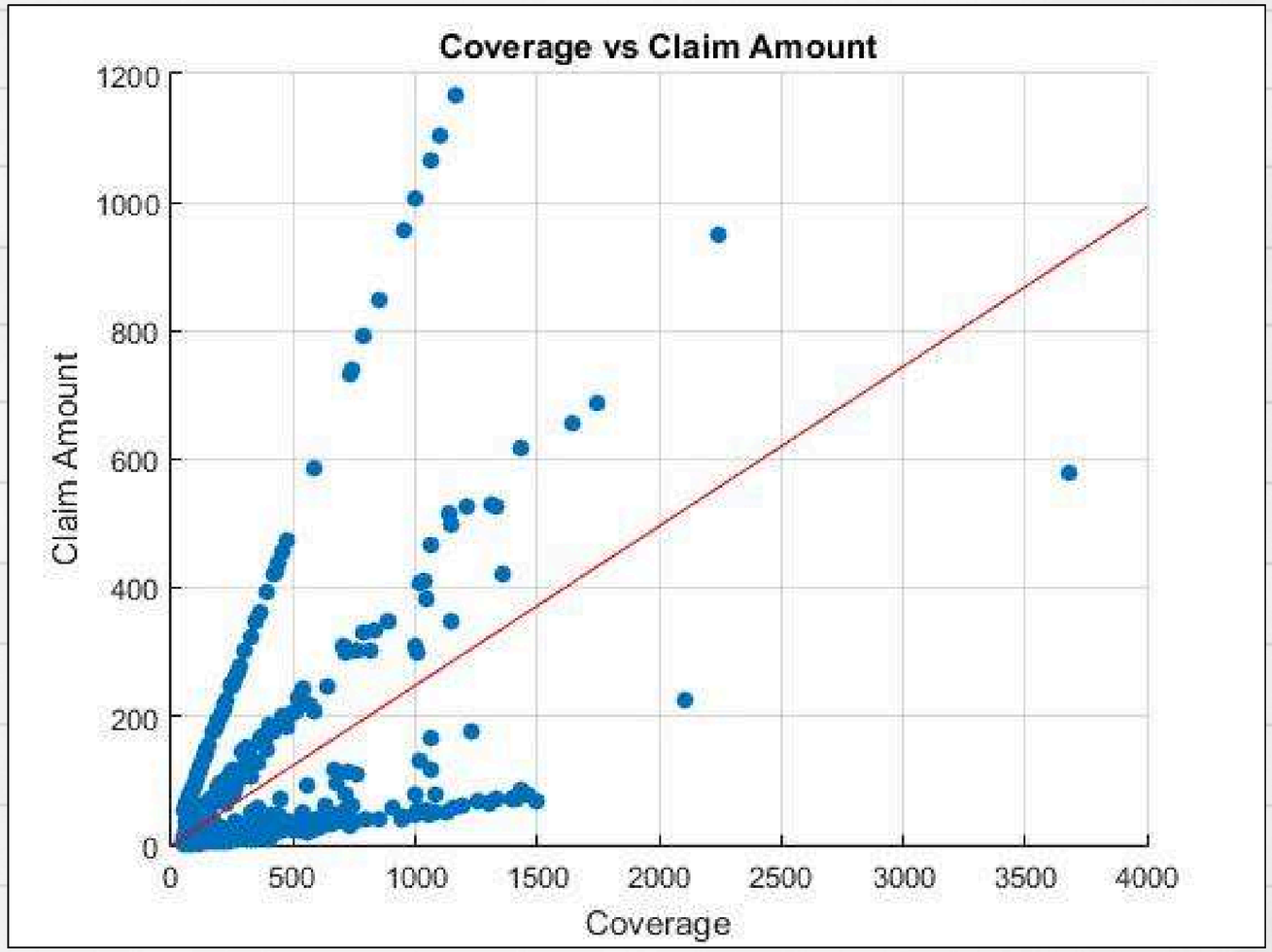


- Positive Correlation between Income and Claim Amount
- Presence of Outliers

ASSOCIATION BETWEEN NUMERICAL PREDICTORS

- Positive Correlation between Coverage and Claim Amount
- Presence of Groups

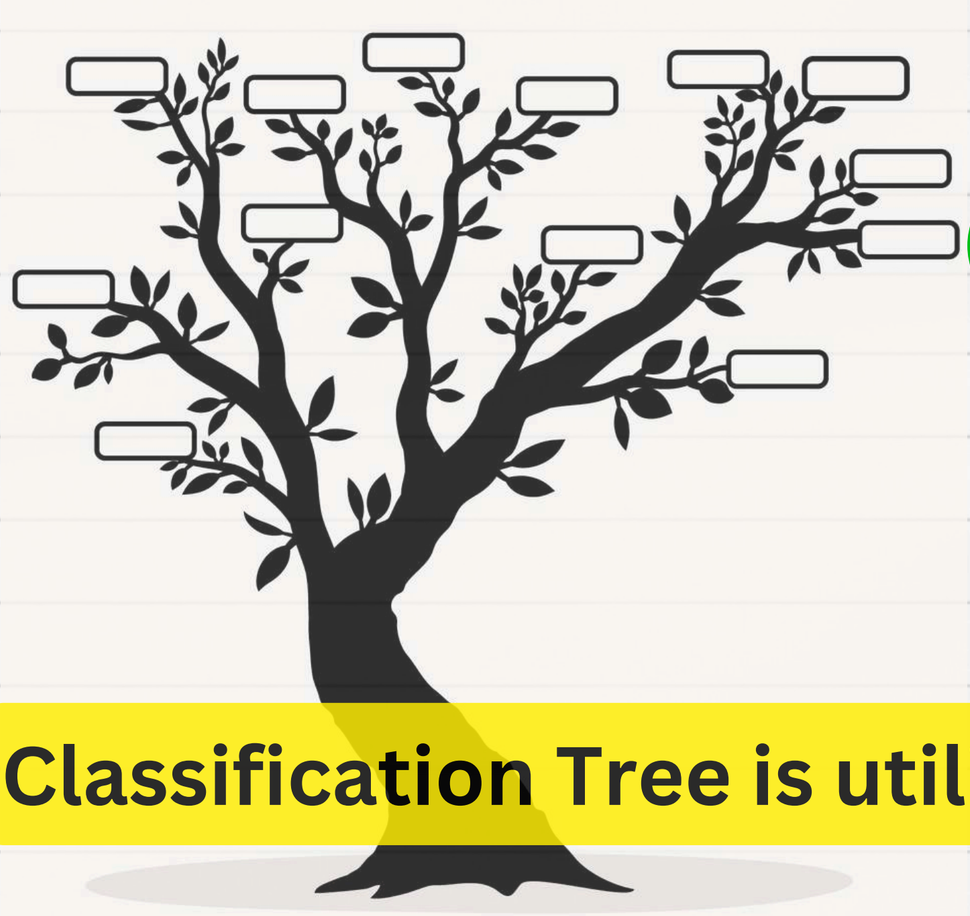
Further Discussed Under Cluster Analysis



Advanced Analysis

Decision Tree

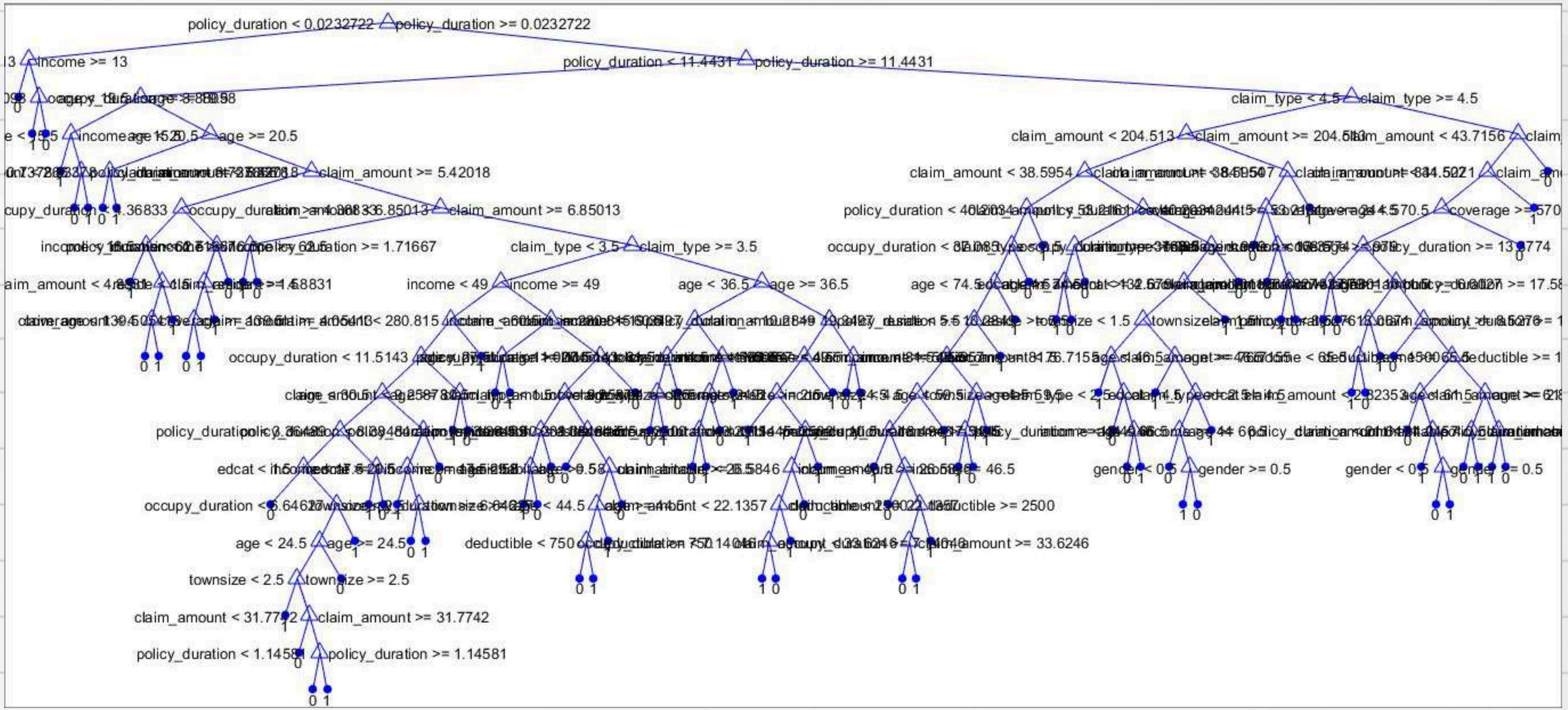
Why Used Decision Tree Approach



To Identify the variables which helps to determine a claim is fraudulence or not.

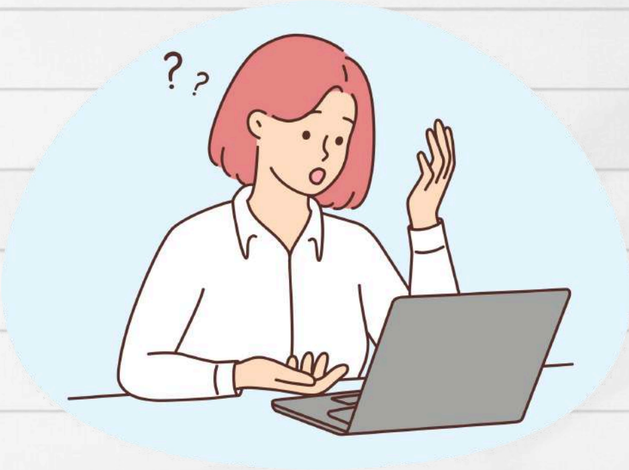
Classification Tree is utilized since the claim is binary.

Classification Tree



Classification Tree

The Overfitting Problem



What's Happening?

Classification tree can grow too deep, memorizing noise in the data instead of learning patterns.

Result: Poor performance on new, unseen data

Overfitting is like studying only past exam questions—you'll fail when faced with new problems!

Classification Tree

The Solution: Prune to Perfection

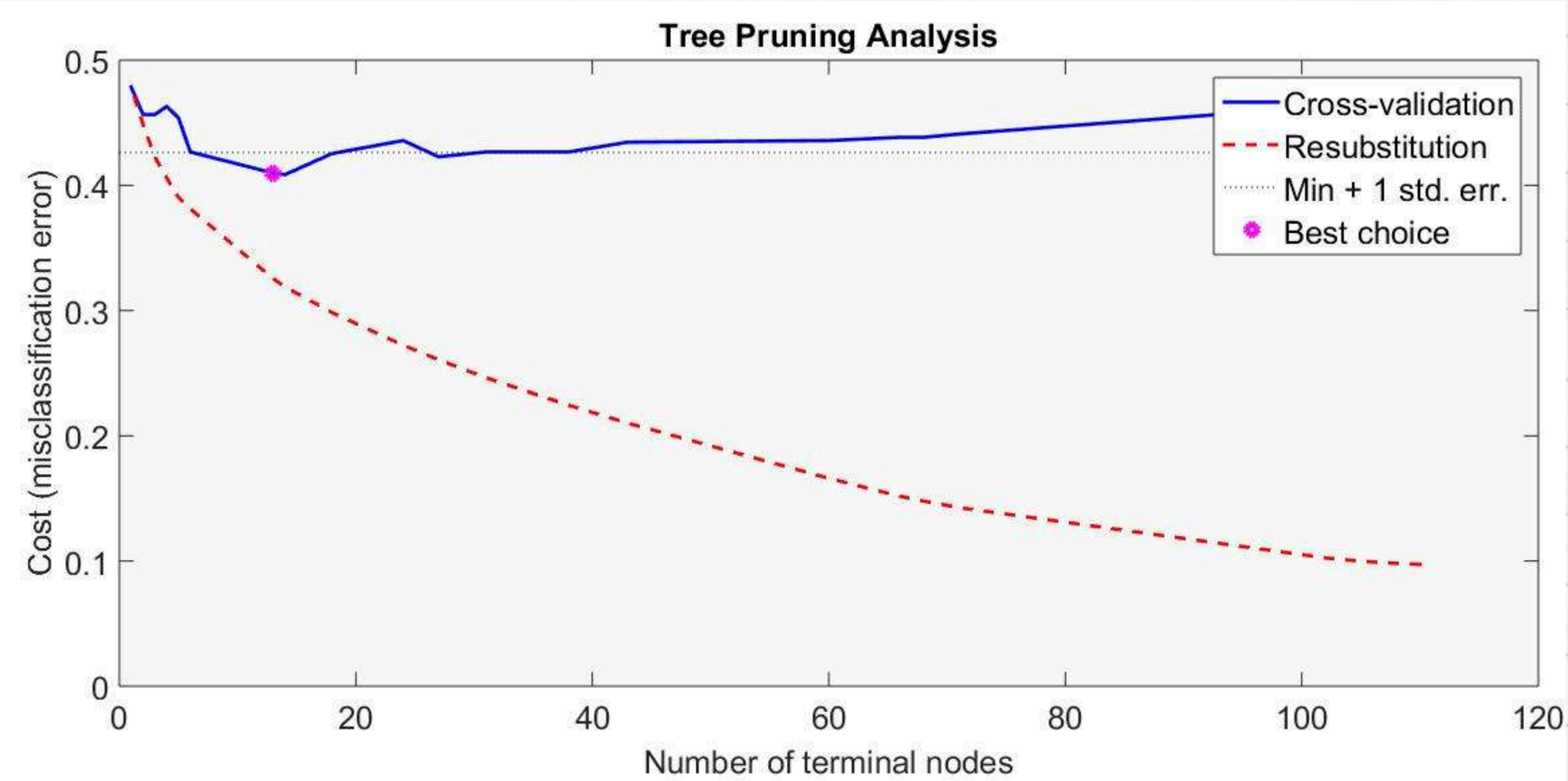
Pruning: Trim the tree to remove unnecessary splits.

- Pre-Pruning: Stop growing the tree early (e.g., limit depth, set minimum samples per leaf).
- Post-Pruning: Grow the tree fully, then cut back branches that add little value

Benefit

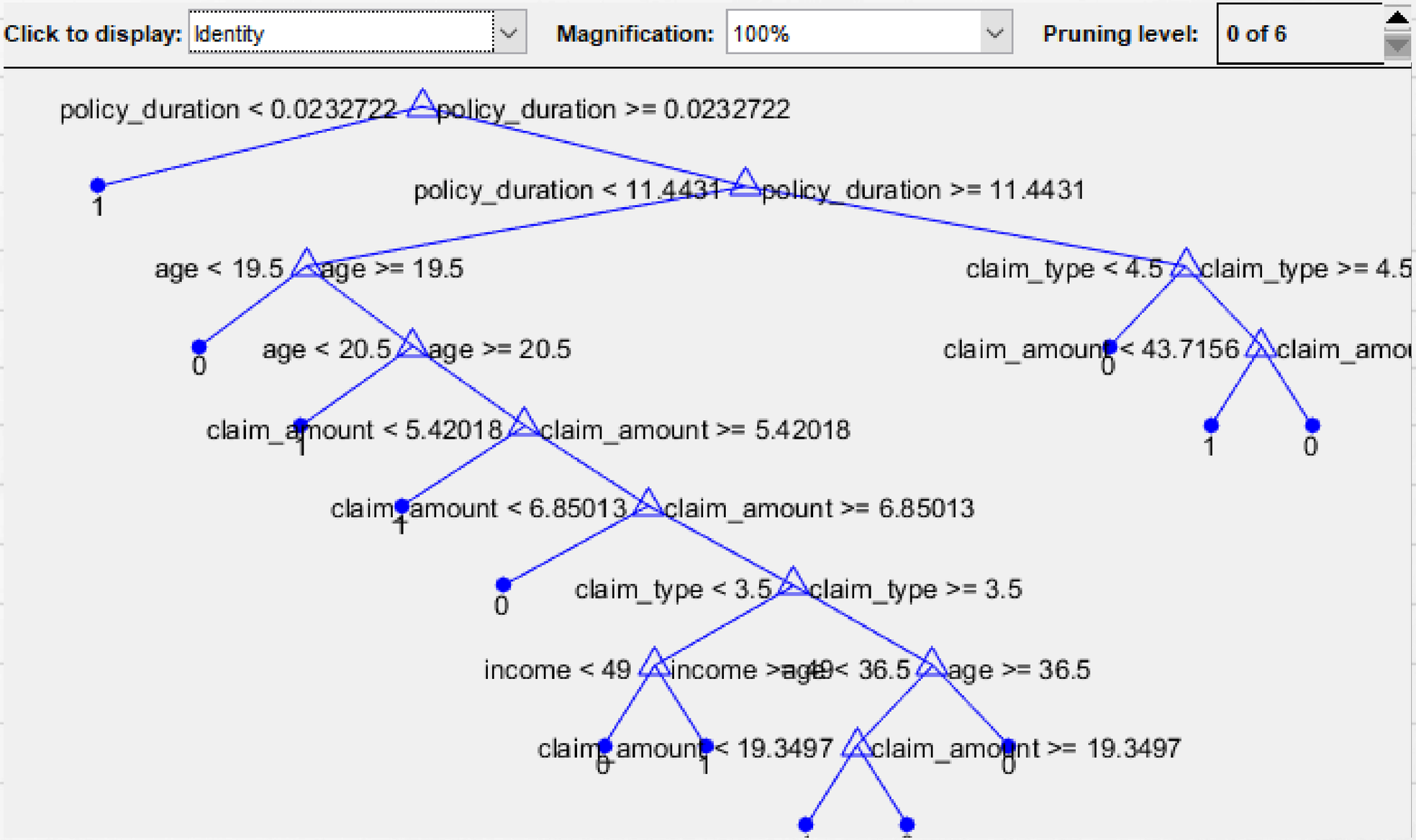
- Simpler Trees: Fewer levels, easier to interpret.
- Better Generalization: Performs well on new data

Tree Pruning Analysis

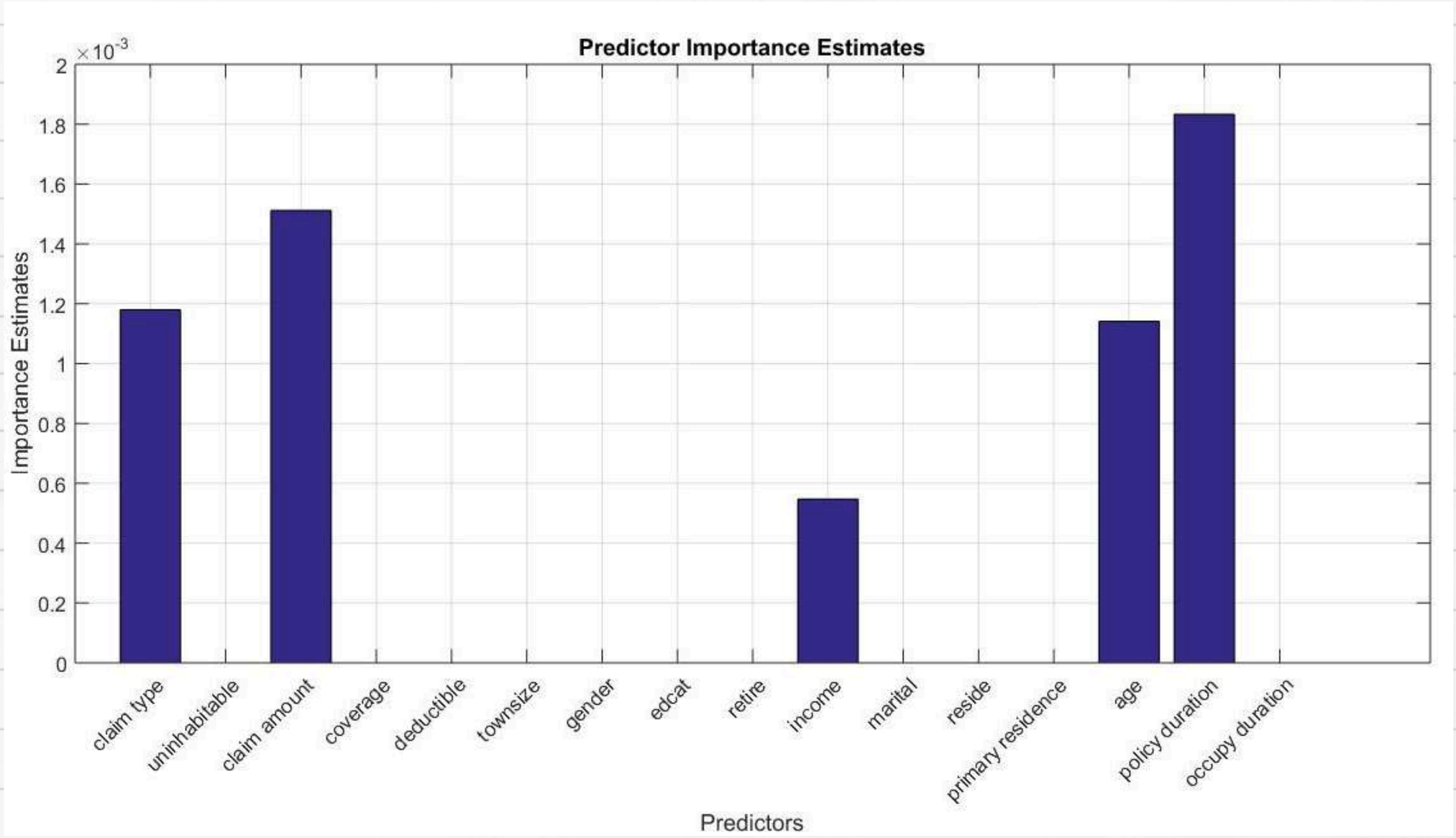


It helps balance the trade-off between model complexity and predictive accuracy

Pruned Tree


Click to display: Identity
Magnification: 100%
Pruning level: 0 of 6
policy_duration < 0.0232722
policy_duration >= 0.0232722
1
policy_duration < 11.4431
policy_duration >= 11.4431
age < 19.5
age >= 19.5
claim_type < 4.5
claim_type >= 4.5
0
age < 20.5
age >= 20.5
claim_amount < 43.7156
1
0
claim_amount < 5.42018
claim_amount >= 5.42018
1
claim_amount < 6.85013
claim_amount >= 6.85013
1
0
claim_type < 3.5
claim_type >= 3.5
income < 49
income >= 49
age < 36.5
age >= 36.5
claim_amount < 19.3497
claim_amount >= 19.3497
1
0

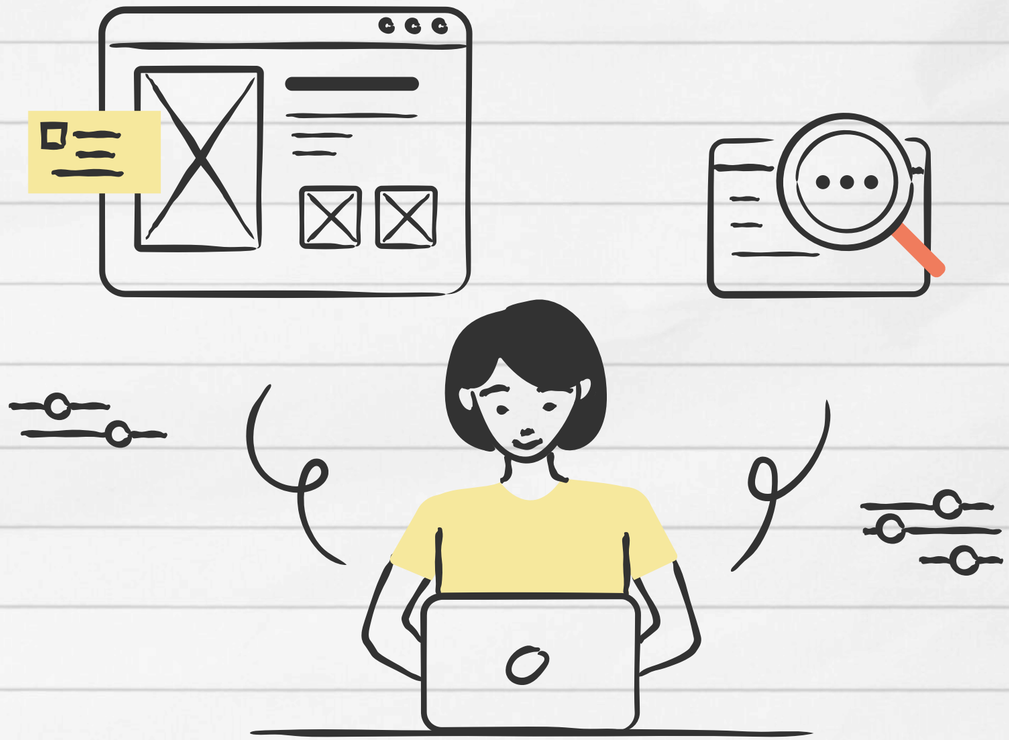
Identifying Key Variables for Fraud Detection



Identifying Key Variables for Fraud Detection

Key Variables Identified:

- **Claim Type**
- **Claim Amount**
- **Income**
- **Age**
- **Policy Duration**

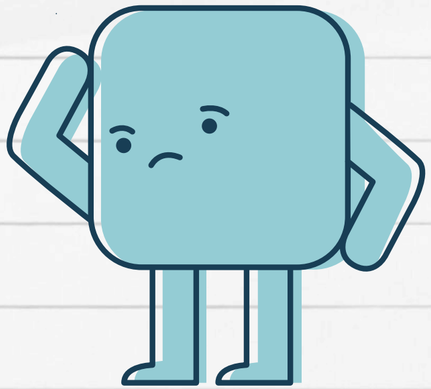


The Classification tree identified reliable and interpretable key variables for fraud detection, validated by multiple models.

CLUSTER ANALYSIS

WHY CLUSTER ANALYSIS?

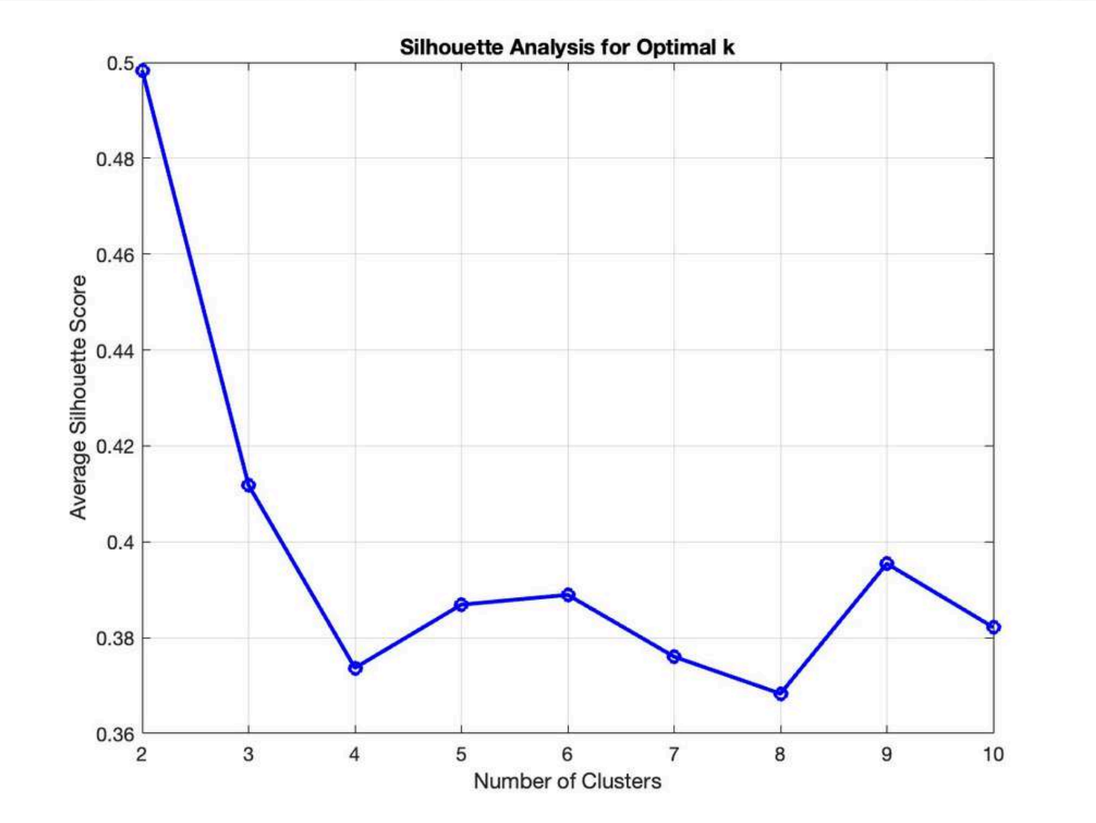
BY DIVIDING CLAIMS INTO SIMILAR GROUPS CAN FIND COMMON PATTERNS AMONG THEM .

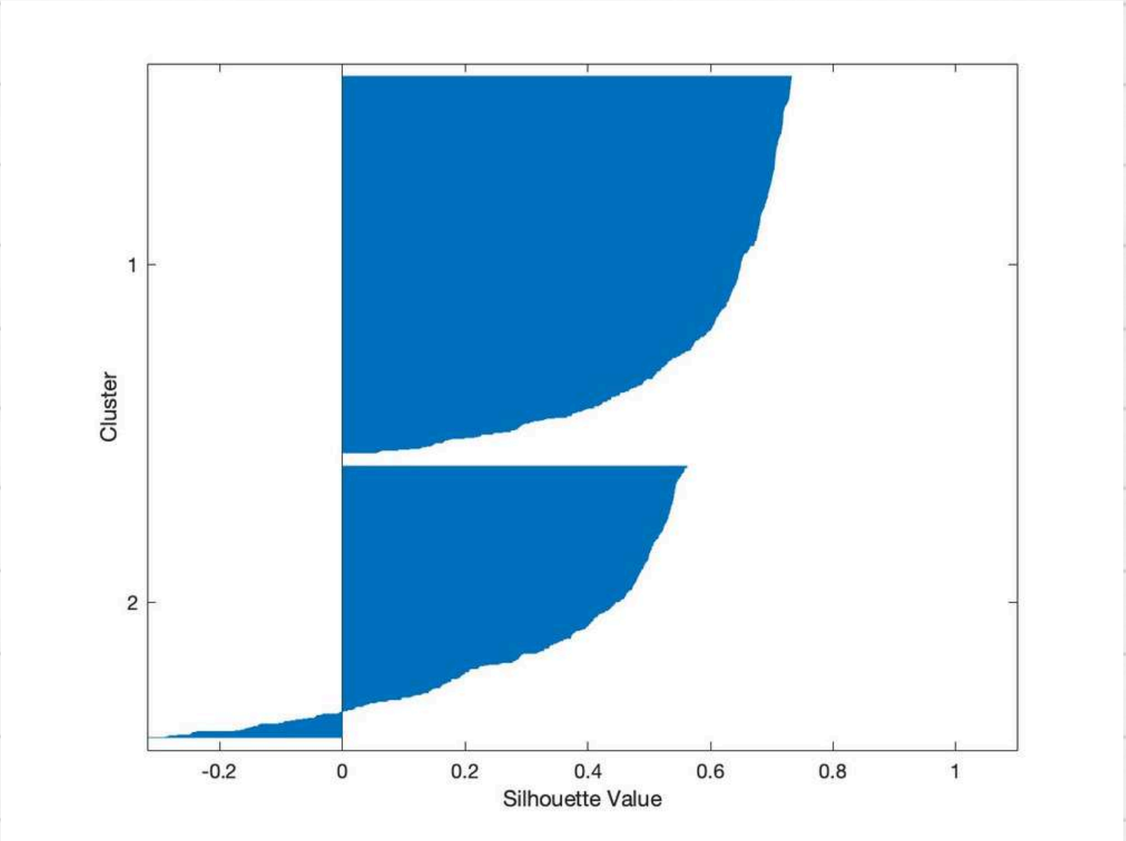


AGE GROUP
INCOME LEVEL
POLICY DURATION
CLAIM AMOUNT

HAVE THE HIGHEST NUMBER OF **FRAUDULENT CLAIMS.**

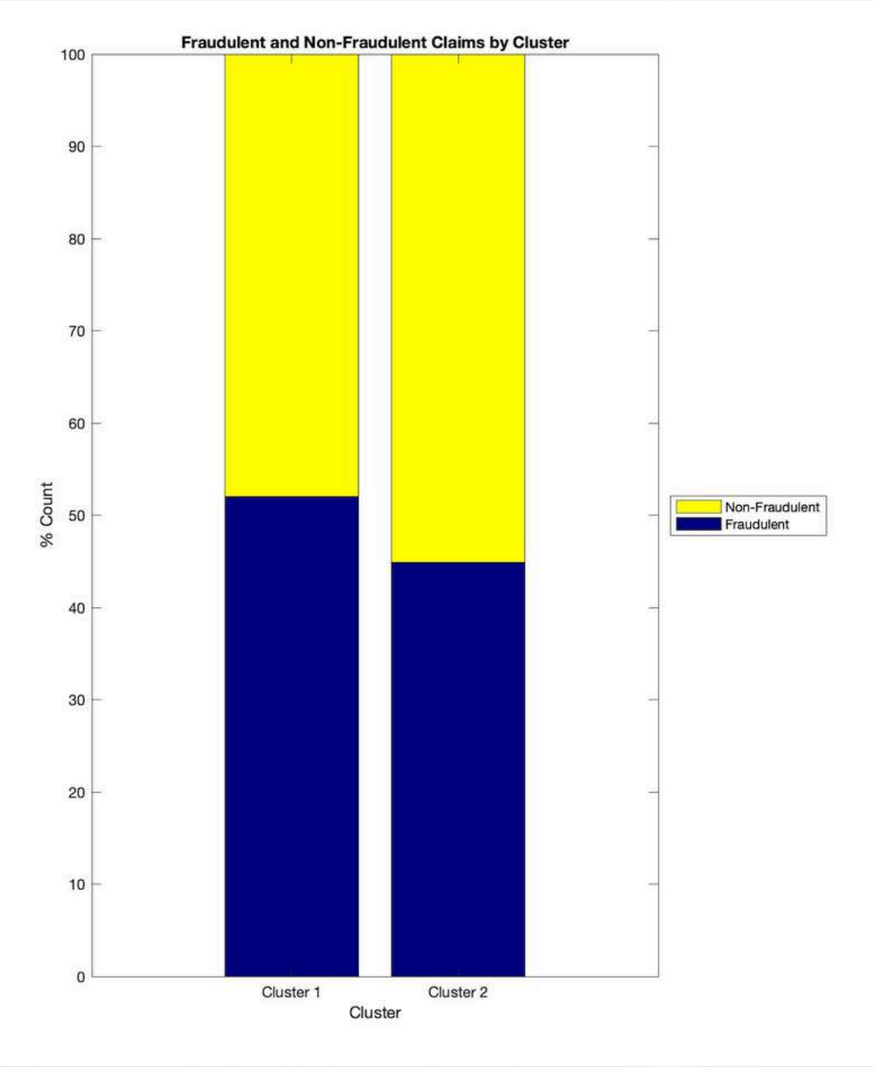
DIVIDE INTO CLUSTERS



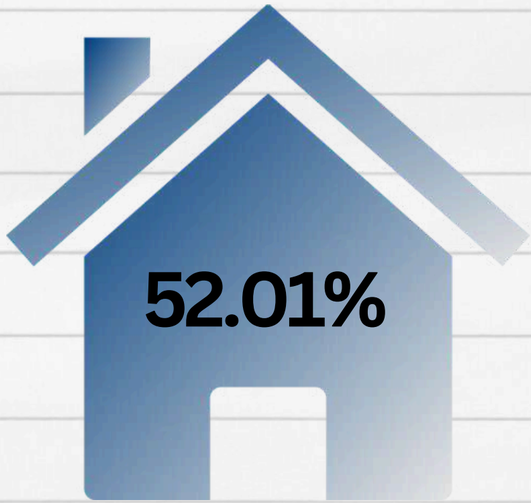


TWO CLUSTERS

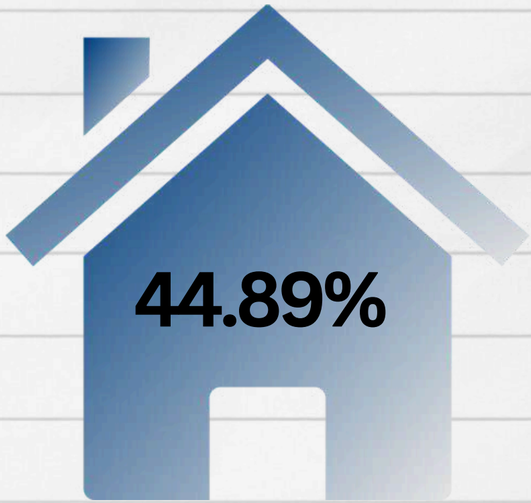
FRAUDULENT CLAIMS PERCENTAGE



Fraudulent Claims Percentage



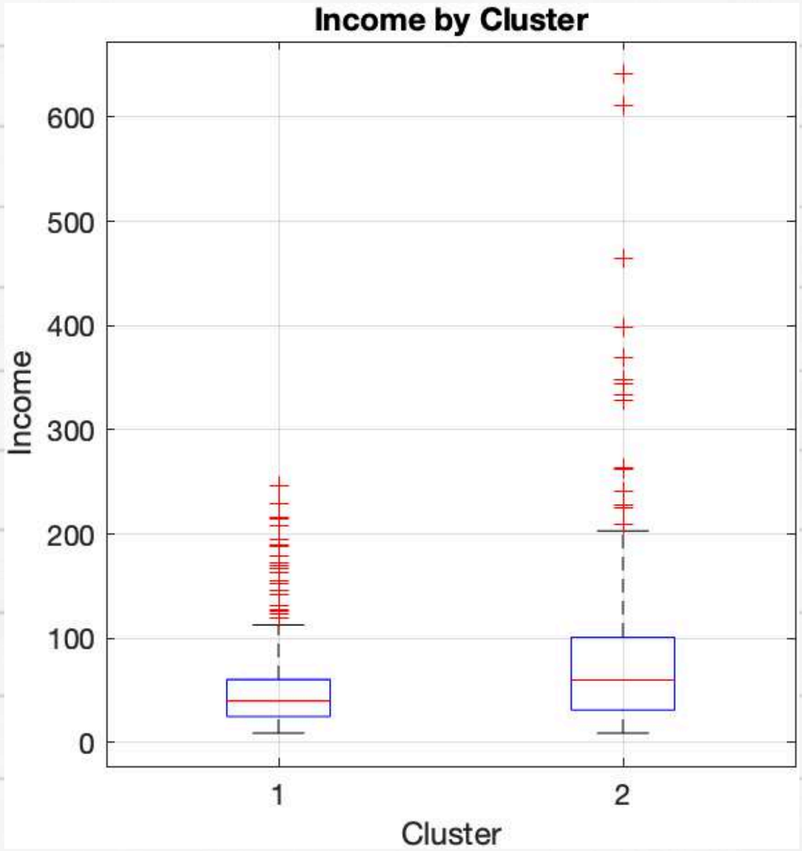
Cluster 01

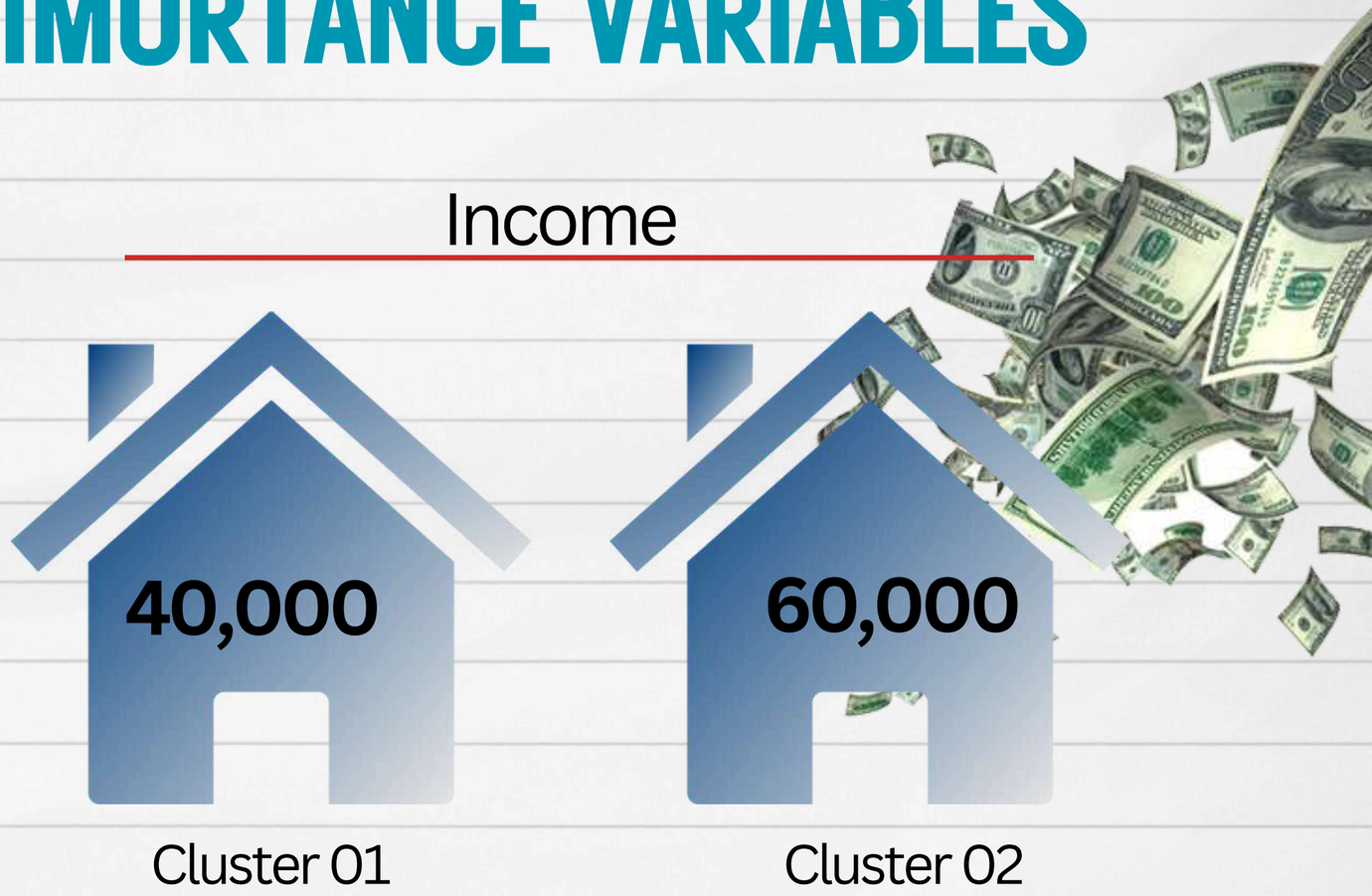


Cluster 02

HAVE TO CONCERN MORE ON CLUSTER 1 !

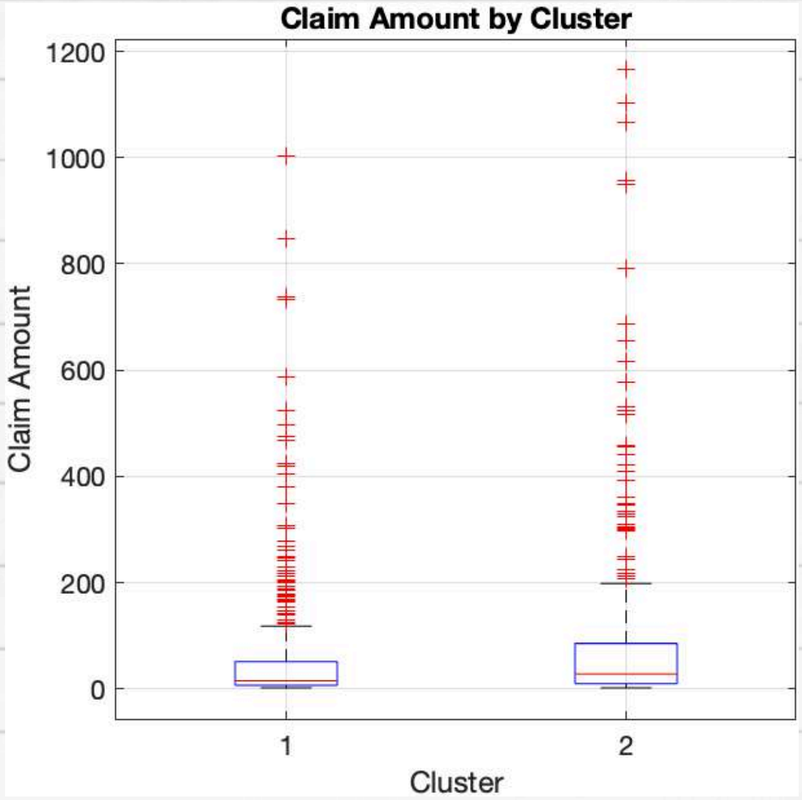
MORE FINDING ON IMORTANCE VARIABLES





In **Cluster 1**, income levels are closely grouped together, while in Cluster 2, incomes vary more and include higher values. This suggests that income level may be linked to claim behavior.

MORE FINDING ON IMPORTANCE VARIABLES



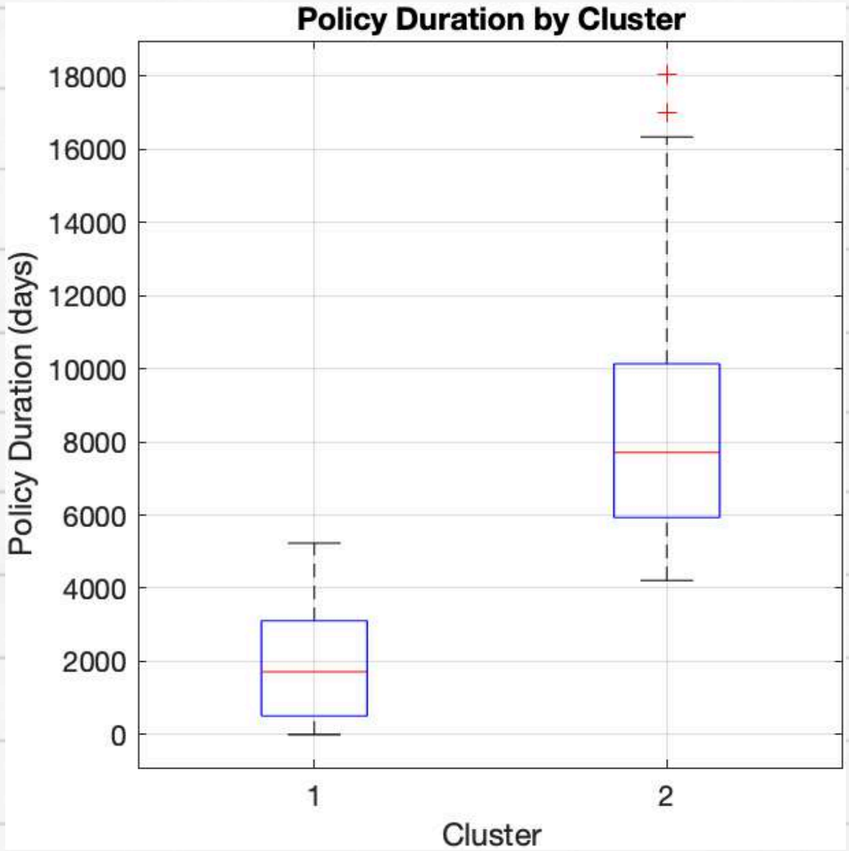
Claim Amount

Cluster 01

Cluster 02

The boxplot shows that Cluster 1 has claim amounts that are closely grouped together (smaller box). Cluster 2 has claim amounts that vary more, with some very high values (outliers).

MORE FINDING ON IMPORTANCE VARIABLES



Policy Duration

1714.5

Cluster 01

Short-term Policyholders

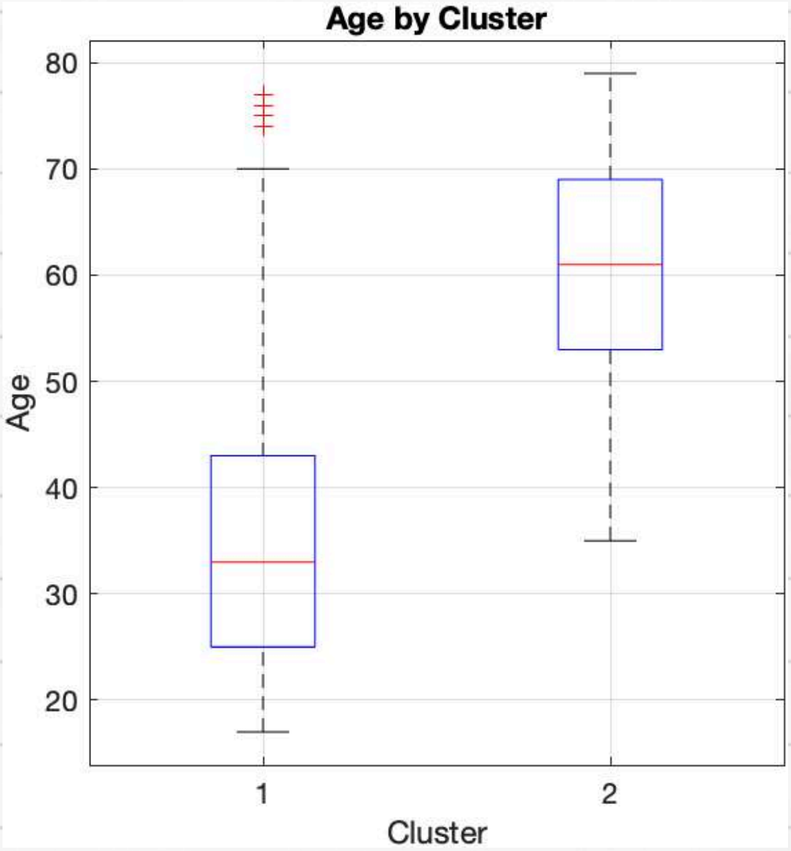
7711

Cluster 02

Long-term Policyholders

The dramatic difference suggests **Cluster 2** represents **long-term policyholders** (over 21 years on average).
Cluster 1 contains **newer customers with less established** relationships

MORE FINDING ON IMPORTANCE VARIABLE



Age

Cluster 01

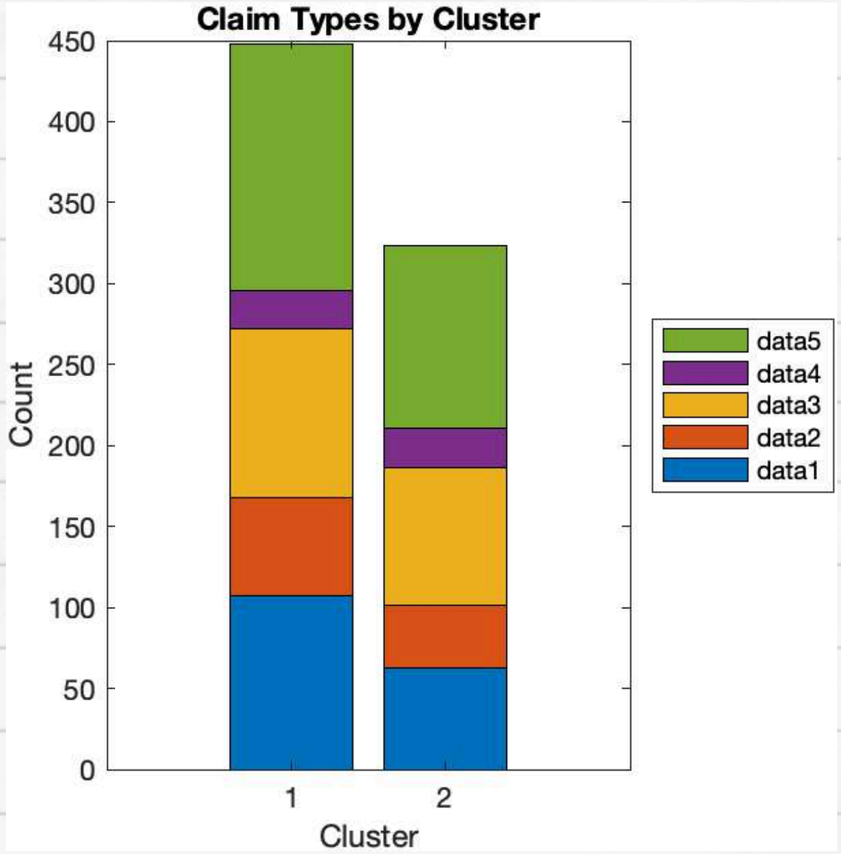
Younger Policyholders

Cluster 02

Older Policyholders

This represents one of **the most distinctive separating features** between the clusters.

MORE FINDING ON IMPORTANCE VARIABLES



Claim Type

Cluster 1 has more total claims, the relative proportions of claim types are somewhat similar between clusters.

Type 5 (green)/Theft represents the largest proportion in both clusters.

REFERENCES

- Vacant Property Fraud
  - https://coretitle.com/vacant-property-fraud-the-latest-real-estate-scam-to-look-out-for/
- 5 most common examples of insurance fraud
  - https://www.burtoncopeland.com/news/5-most-common-examples-insurance-fraud/
- Fraud in the Short-term Insurance Sector
  - https://www.justmoney.co.za/articles/fraud-in-the-short-term-insurance-sector/
- K-Means Clustering
  - https://www.geeksforgeeks.org/k-means-clustering-in-matlab/
- How to Build Decision Tree in MATLAB?
  - https://www.geeksforgeeks.org/how-to-build-decision-tree-in-matlab/

Thank you!