Problem Statement

Global Insure, a leading insurance company, processes thousands of claims annually. However, a significant percentage of these claims turn out to be fraudulent, resulting in considerable financial losses. The company's current process for identifying fraudulent claims involves manual inspections, which is time-consuming and inefficient. Fraudulent claims are often detected too late in the process, after the company has already paid out significant amounts. Global Insure wants to improve its fraud detection process using data-driven insights to classify claims as fraudulent or legitimate early in the approval process. This would minimise financial losses and optimise the overall claims handling process.

Business Objective

Global Insure wants to build a model to classify insurance claims as either fraudulent or legitimate based on historical claim details and customer profiles. By using features like claim amounts, customer profiles and claim types, the company aims to predict which claims are likely to be fraudulent before they are approved.

Based on this assignment, you have to answer the following questions:

- How can we analyze historical claim data to detect patterns that indicate fraudulent claims?
- Which features are most predictive of fraudulent behavior?
- Can we predict the likelihood of fraud for an incoming claim, based on past data?
- What insights can be drawn from the model that can help in improving the fraud detection process?

Key Findings and Insights

1. **Target Variable (fraud_reported)**

Imbalance: The dataset is imbalanced, with 75.3% of claims being non-fraudulent ('N') and 24.7% being fraudulent ('Y'). This imbalance needs to be considered during model building (e.g., using oversampling, undersampling, or SMOTE).

2. **Policyholder Information**

- **months_as_customer and age:**
    - The distributions of months_as_customer and age are similar for both fraudulent and non-fraudulent claims. This suggests that neither customer tenure nor age is a strong individual predictor of fraud.
- **policy_state:**
    - The proportion of fraudulent claims is slightly higher in 'OH' (25.9%) and 'IN' (25.5%) compared to 'IL' (22.8%). While the difference isn't huge, it might be worth investigating regional factors.
- **policy_csl (Combined Single Limit):**
    - Claims with '250/500' CSL have a slightly higher fraud rate (26.2%) compared to '100/300' (25.8%) and '500/1000' (21.7%). Higher coverage limits might be associated with a slightly increased risk of fraud.
- **policy_deductable and policy_annual_premium:**
    - The distributions of these features are very similar for both classes, suggesting they are not strong differentiators.
- **umbrella_limit:**
    - A notable difference is observed here. Fraudulent claims appear to have a higher mean umbrella limit compared to non-fraudulent claims. This could indicate that individuals with higher coverage are more likely to engage in fraudulent activities.
- **insured_sex:**
    - Males have a slightly higher fraud rate (26.1%) than females (23.5%).
- **insured_education_level:**
    - Fraud rates are fairly consistent across education levels, ranging from 22.4% to 26.4%. 'PhD' and 'College' show the highest fraud percentages.
- **insured_occupation:**

- This feature shows more significant variation. Occupations like 'exec-managerial' (36.8%), 'farming-fishing' (30.2%), and 'craft-repair' (29.7%) have considerably higher fraud rates. 'Other-service' and 'adm-clerical' have the lowest (16.9%).
- **insured_hobbies:**
- Certain hobbies show higher fraud rates: 'yachting' (30.2%), 'board-games' (29.2%), 'base-jumping' (26.5%). 'Camping' and 'kayaking' have very low fraud rates (around 9%).
- **insured_relationship:**
- 'Other-relative' (29.4%) has the highest fraud rate, while 'husband' (20.6%) has the lowest.

3. **Incident Details**

- **incident_type:**
- 'Single Vehicle Collision' (29.0%) and 'Multi-vehicle Collision' (27.2%) have much higher fraud rates than 'Parked Car' (9.5%) or 'Vehicle Theft' (8.5%).
- **collision_type:**
- Claims where the collision type is missing ('?') have a very low fraud rate (9.0%). 'Rear Collision' has the highest (31.2%).
- **incident_severity:**
- This is a critical feature. 'Major Damage' has an alarming 60.5% fraud rate, strongly indicating that severe incidents are much more likely to be fraudulent.
- **authorities_contacted:**
- Claims where 'None' were contacted have a significantly higher fraud rate (44.1%).
- **incident_state and incident_city:**
- These features can provide regional insights, but the analysis needs to be detailed to find significant differences.
- **incident_hour_of_the_day:**
- The distribution appears similar across fraud categories.
- **number_of_vehicles_involved:**

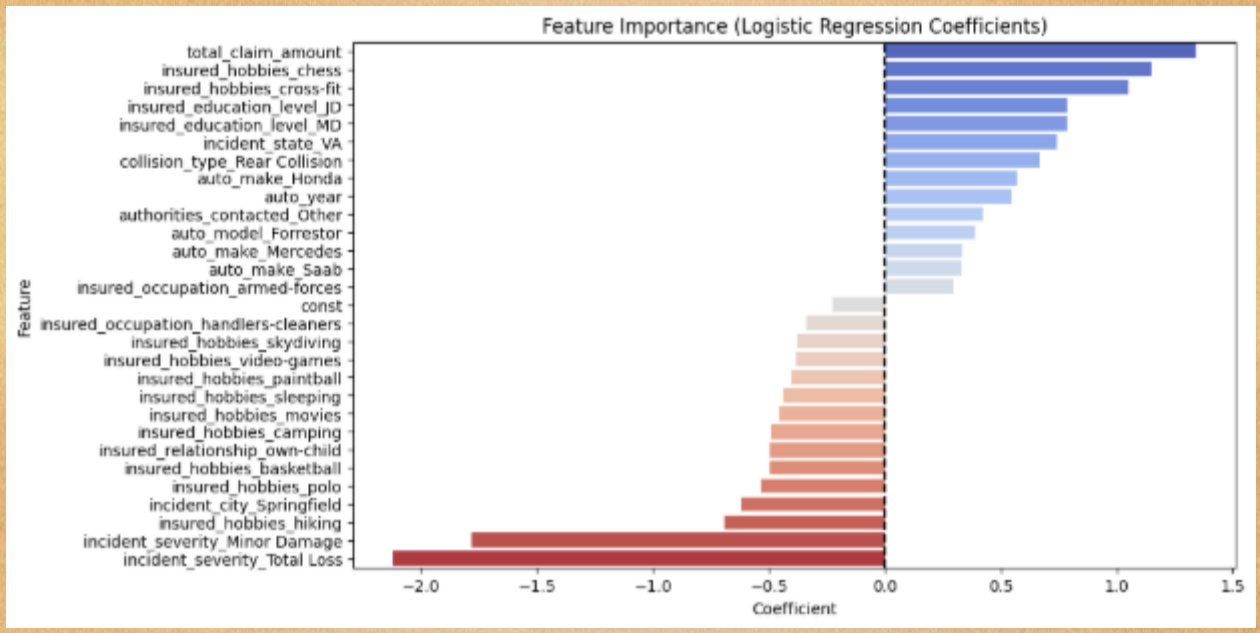
- Higher fraud rates are seen in incidents involving 1 or 3 vehicles.
- **property_damage and police_report_available:**
- Missing values ('?') in these columns need to be handled carefully. The presence or absence of this information might be indicative of fraud.
- **bodily_injuries and witnesses:**
- Higher numbers of bodily injuries and witnesses might correlate with lower fraud rates, but the effect isn't very strong.

4. **Claim Information**

- **total_claim_amount, injury_claim, property_claim, vehicle_claim:**
- Fraudulent claims tend to have higher claim amounts overall. The distributions of these features differ noticeably between the two classes.
- **auto_make, auto_model, auto_year:**
- Certain auto makes and models might be more prone to fraud. Older vehicles also tend to be associated with higher fraud.

5. **Correlation Analysis**

- There's a strong positive correlation between total_claim_amount and its components (injury_claim, property_claim, vehicle_claim). This is expected.
- No extremely high correlations between other independent variables, so multicollinearity might not be a major concern.


Feature Importance (Logistic Regression Coefficients)
total_claim_amount
insured_hobbies_chess
insured_hobbies_cross-fit
insured_education_level_JD
insured_education_level_MD
incident_state_VA
collision_type_Rear Collision
auto_make_Honda
auto_year
authorities_contacted_Other
auto_model_Forrestor
auto_make_Mercedes
auto_make_Saab
insured_occupation_armed-forces
const
insured_occupation_handlers-cleaners
insured_hobbies_skydiving
insured_hobbies_video-games
insured_hobbies_paintball
insured_hobbies_sleeping
insured_hobbies_movies
insured_hobbies_camping
insured_relationship_own-child
insured_hobbies_basketball
insured_hobbies_polo
incident_city_Springfield
insured_hobbies_hiking
incident_severity_Minor Damage
incident_severity_Total Loss
Feature
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
Coefficient

Final Conclusion is

- Key drivers increasing fraud likelihood according to Logistic Regression
- total_claim_amount: A higher total claim amount is strongly associated with a higher likelihood of the target event occurring.
- insured_hobbies_chess: Individuals who enjoy chess as a hobby are more likely to experience the target event.
- insured_hobbies_cross-fit: Similarly, those who engage in cross-fit are more prone to the target event.
- insured_education_level_JD, insured_education_level_MD, insured_education_level_VA: Higher levels of education (Juris Doctor, Doctor of Medicine, and potentially a specific vocational associate degree) appear to be positively correlated with the target variable.
- incident_state_VA: Incidents occurring in the state of Virginia show a positive association with the target variable.
- collision_type_Rear Collision: Rear-end collisions are linked to a higher probability of the target event.
- auto_make_Honda, auto_year, authorities_contacted_Other, auto_model_Forrestor, auto_make_Mercedes, auto_make_Saab: Certain car makes, the year of the car, contacting authorities other than the police, and specific car models (Forrestor, Mercedes, Saab) show a positive correlation.

- These features decrease the odds of a fraudulent claim.
- Strong negative influence (less likely to be fraud):

○Incident_severity_Total Loss: Incidents classified as a "Total Loss" are strongly associated with a lower likelihood of the target event. This might seem counterintuitive and warrants further investigation into what the target variable represents.
○incident_severity_Minor Damage: Similarly, incidents with "Minor Damage" are negatively correlated with the target variable.
○insured_hobbies_hiking, incident_city_Springfield, insured_hobbies_polo,
○insured_hobbies_basketball, insured_relationship_own-child,
○insured_hobbies_camping, insured_hobbies_movies, insured_hobbies_sleeping,
○insured_hobbies_paintball, insured_hobbies_video-games,
○insured_hobbies_skydiving, insured_occupation_handlers-cleaners: Engaging in these hobbies, living in Springfield, having a relationship of "own-child" with the insured, and being in the "handlers-cleaners" occupation are associated with a lower probability of the target event.
○insured_occupation_armed-forces: Individuals in the armed forces as their occupation are less likely to experience the target event.

- Model Results
- Accuracy on Train set by Logistic Regression after optimum cut-off point is **0.9011**

- Sensitivity: 0.932806324110672
  Specificity: 0.8695652173913043
  Precision: 0.8773234200743495
  Recall: 0.932806324110672
  F1 Score: 0.9042145593869733

- Accuracy on Train set by Random Forest after hyper parameter tuning using grid-search is **0.915**
- Sensitivity (Tuned RF): 0.9525691699604744
  Specificity (Tuned RF): 0.8774703557312253
  Precision (Tuned RF): 0.8860294117647058

- Accuracy on Test set by Logistic Regression after optimum cut-off point is **0.748**
- Sensitivity (Test LR): 0.6470588235294118
  Specificity (Test LR): 0.7798165137614679
  Precision (Test LR): 0.4782608695652174
- Accuracy on Test set by Random Forest is **0.7902**

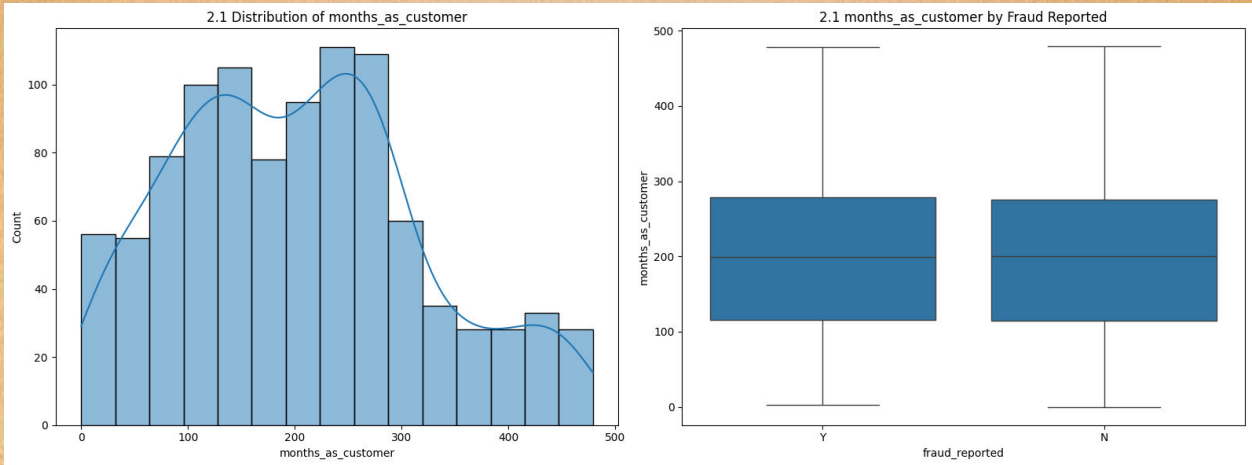
- Sensitivity (Test RF): 0.6470588235294118
  Specificity (Test RF): 0.7798165137614679
  Precision (Test RF): 0.4782608695652174

- The Logistic Regression model performs quite well on the training data with a good balance between **precision** and **recall**, leading to a strong **F1 score**. However, high recall indicates it catches most fraud cases, but specificity is slightly lower, meaning it allows more false positives.
- The tuned Random Forest model **outperforms Logistic Regression** on all major metrics. It's especially better at identifying fraud cases (**higher recall**) while also maintaining high precision. This suggests the model generalizes better to the nuances in training data.

1. Target Variable Analysis ---

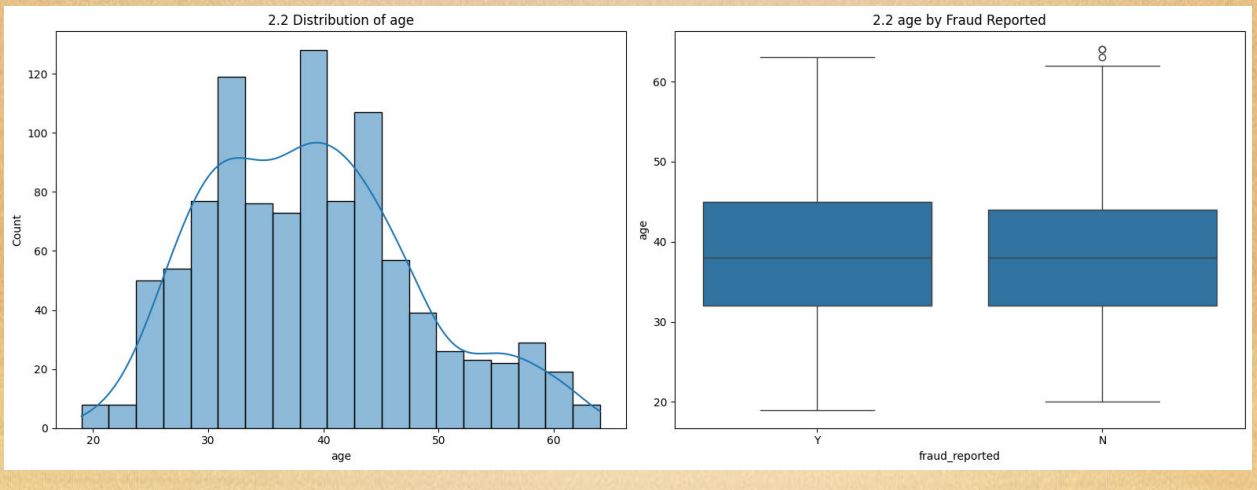
- Observation: The target variable is categorical with two classes: 'Y' (Fraudulent) and 'N' (Legitimate).
- Inference: There is a significant class imbalance, with 'N' being much more frequent than 'Y'.
- Summary: The majority of claims are legitimate, which is typical in fraud detection scenarios.

This imbalance needs to be considered during model training.



2.1 Numerical Feature Analysis: months_as_customer ---

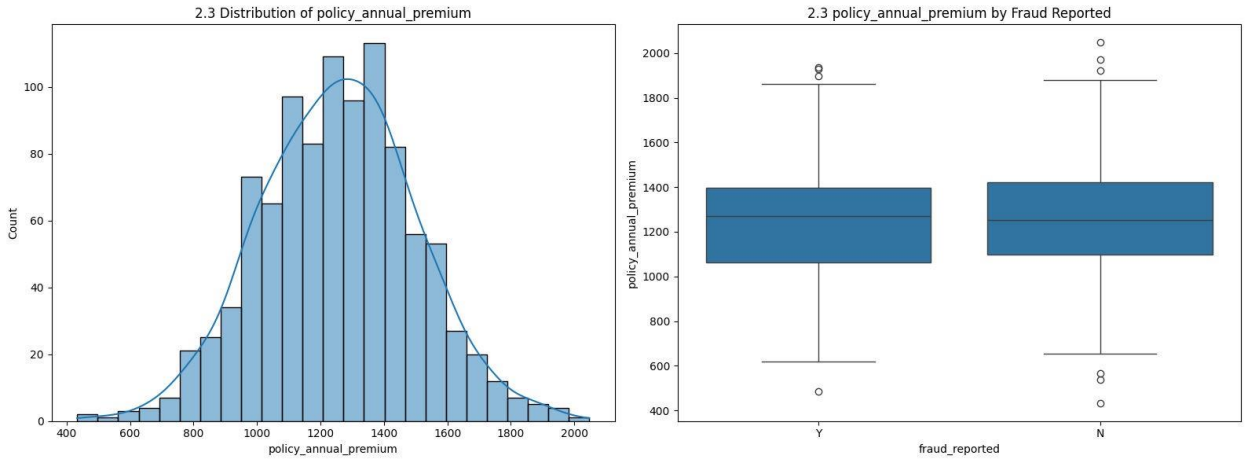
- Observation: The distribution is right-skewed.
- Inference: Most customers are relatively new. There's no clear distinction in the distribution of 'months_as_customer' between fraudulent and non-fraudulent claims.
- Summary: Most customers are relatively new, and customer tenure doesn't strongly differentiate fraudulent claims.



2.2 Numerical Feature Analysis: age ---

- Observation: The age distribution is fairly normal.

- Inference: The age of the insured person does not appear to be a strong indicator of fraud.
- Summary: Age is normally distributed and doesn't show a clear difference between fraudulent and non-fraudulent claims.

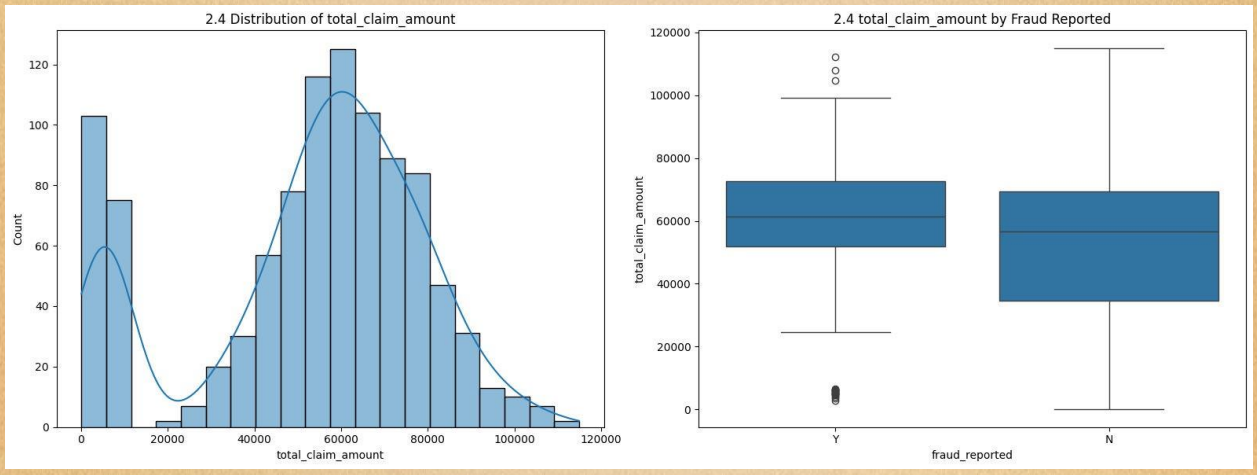


2.3 Numerical Feature Analysis: policy_annual_premium ---

•Observation: 'policy_annual_premium' shows a wide range of values with potential outliers.

•Inference: Some policies are significantly more expensive than others, which could be due to various factors (coverage, vehicle type, etc.). No significant difference between fraudulent and non-fraudulent claims.

•Summary: Policy premiums vary widely, but don't clearly distinguish between fraudulent and non-fraudulent claims.

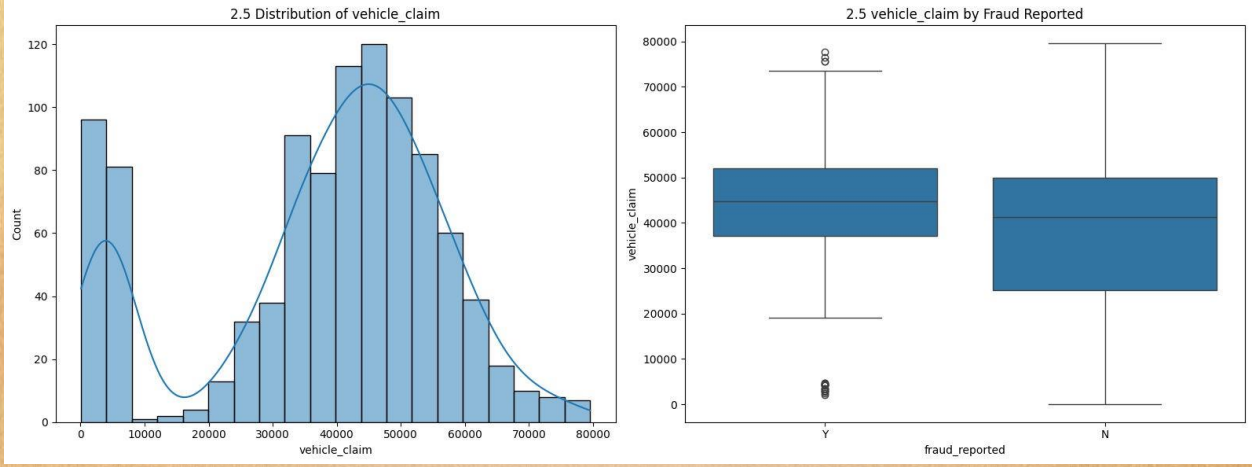


2.4 Numerical Feature Analysis: total_claim_amount ---

•Observation: 'total_claim_amount' is right-skewed.

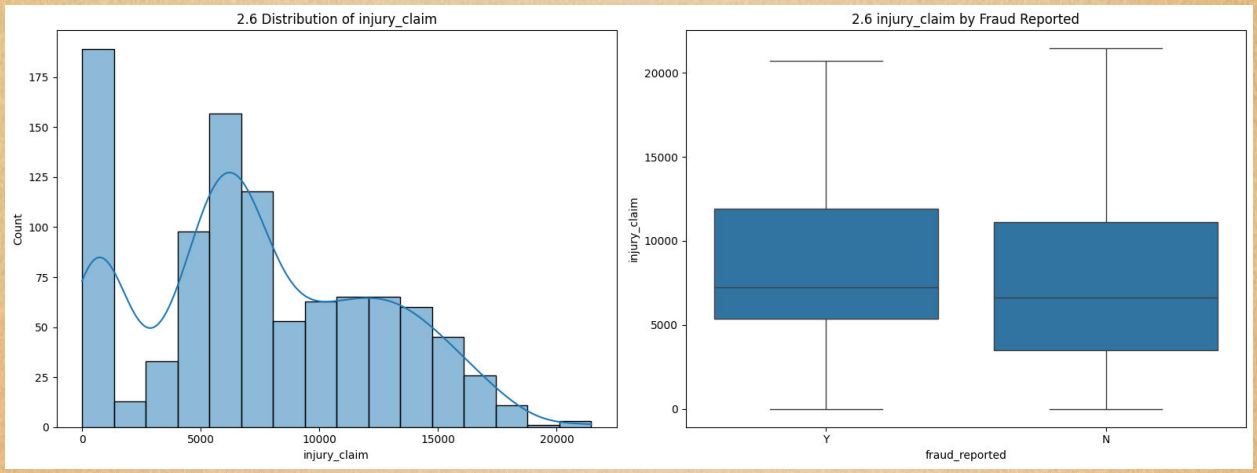
•Inference: Many claims are of smaller amounts, with a few claims of very high amounts. Fraudulent claims tend to have a higher total claim amount.

•Summary: Claim amounts are typically right-skewed, with fraudulent claims showing a tendency towards higher values.



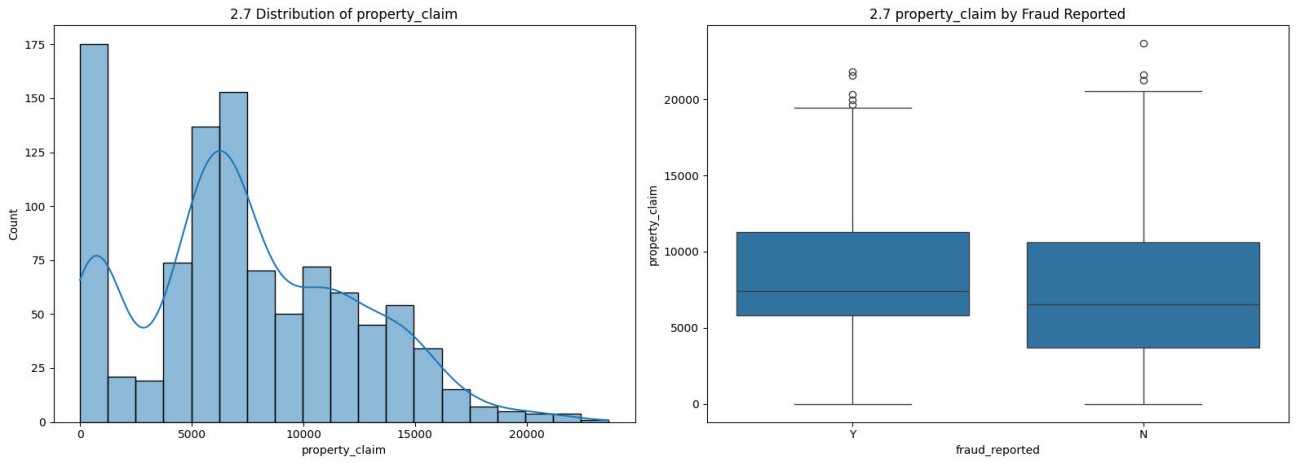
2.5 Numerical Feature Analysis: vehicle_claim ---

- Observation: Right-skewed distribution, similar to total_claim_amount.
- Inference: Higher vehicle claim amounts are observed in fraudulent cases.
- Summary: Vehicle claim amounts show a right-skewed distribution, with fraudulent claims tending to have higher amounts.



2.6 Numerical Feature Analysis: injury_claim ---

- Observation: Right-skewed distribution.
- Inference: Fraudulent claims tend to have higher injury claim amounts.
- Summary: Injury claim amounts are right-skewed, with fraudulent claims often involving larger sums.



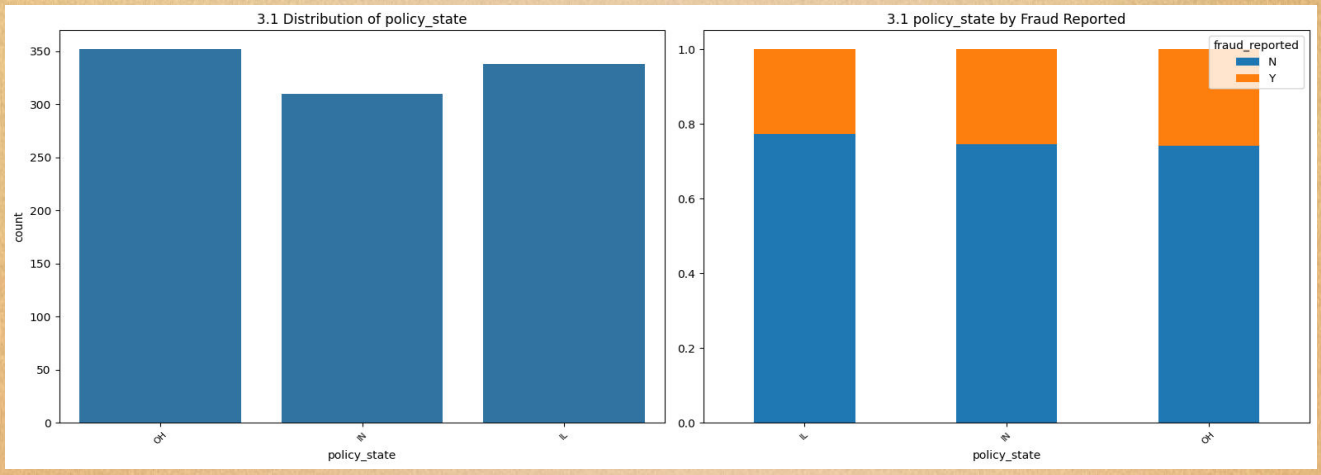
2.7 Numerical Feature Analysis: property_claim ---

- Observation: Right-skewed distribution.
- Inference: Fraudulent claims may involve higher property claim amounts.
- Summary: Property claim amounts are right-skewed, and fraudulent claims may be associated with larger claims.

Summary for Numerical Features Analysis

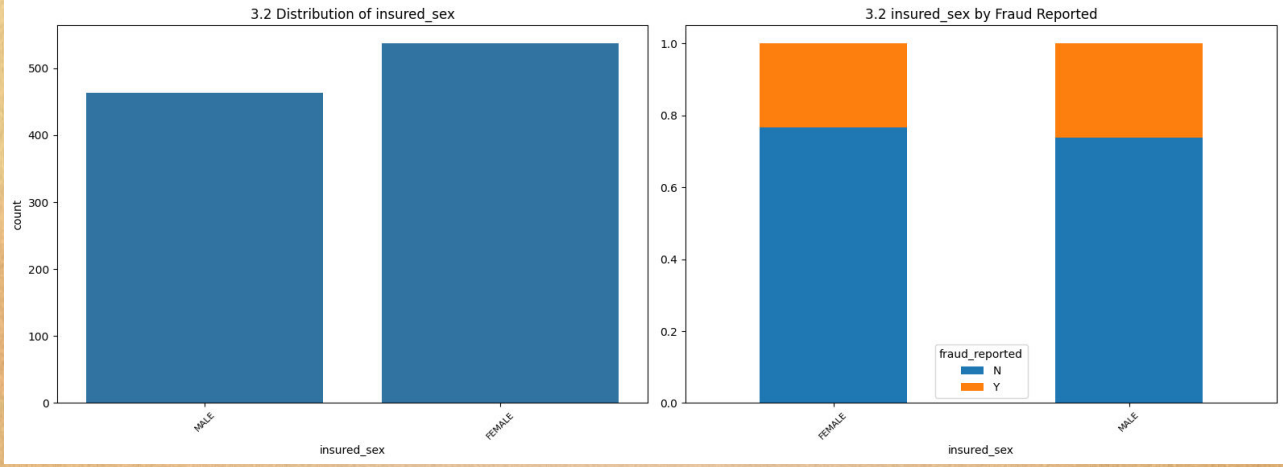
- Summary: Numerical features exhibit different distributions, with claim amounts being skewed.

Boxplot help visualize potential differences in these distributions between fraudulent and legitimate claims.



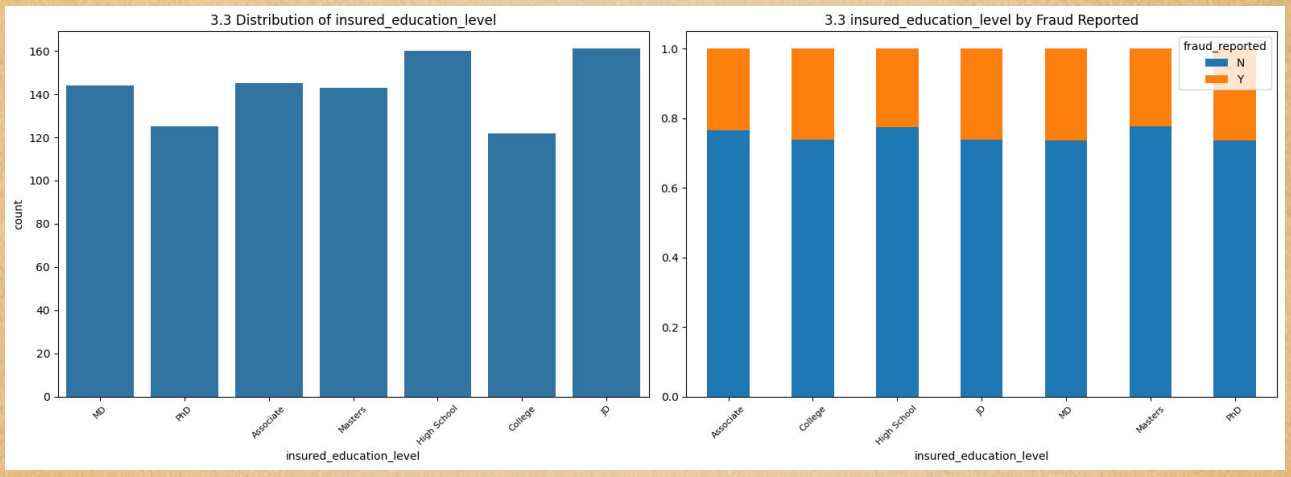
3.1 Categorical Feature Analysis: policy_state ---

- Observation: Uneven distribution of policies across states.
- Inference: Some states have a higher volume of claims. Fraud proportion varies by state.
- Summary: Policy state distribution and fraud proportion varies.



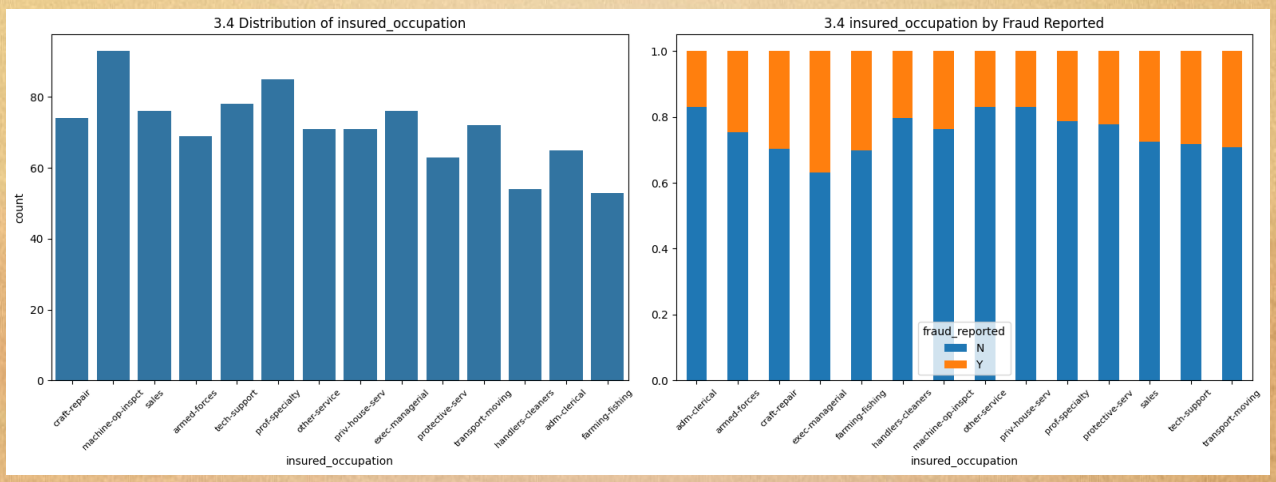
3.2 Categorical Feature Analysis: insured_sex ---

- Observation: More males than females.
- Inference: Gender distribution in the dataset is skewed towards males. Fraud rates are roughly similar.
- Summary: Gender distribution is imbalanced, but fraud rates are similar.



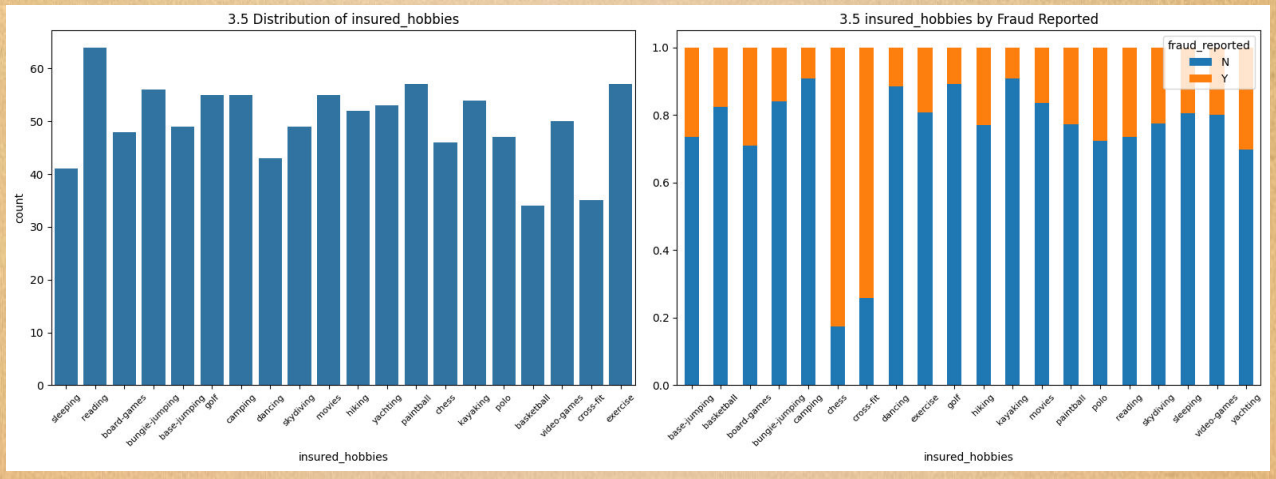
3.3 Categorical Feature Analysis: insured_education_level ---

- Observation: Distribution across education levels.
- Inference: Education level distribution varies. Fraud rates are somewhat similar across education levels.
- Summary: Education level distribution and fraud rates.



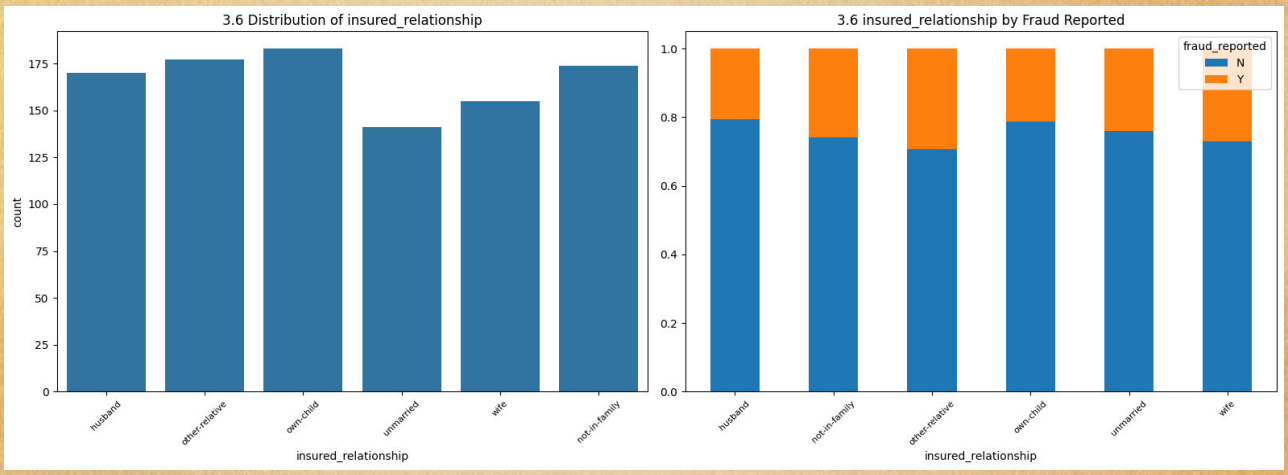
3.4 Categorical Feature Analysis: insured_occupation ---

- Observation: Diverse occupations.
- Inference: Occupational distribution varies widely. Fraud rates vary by occupation, with some occupations showing slightly higher fraud proportions.
- Summary: Occupation distribution and fraud rates.



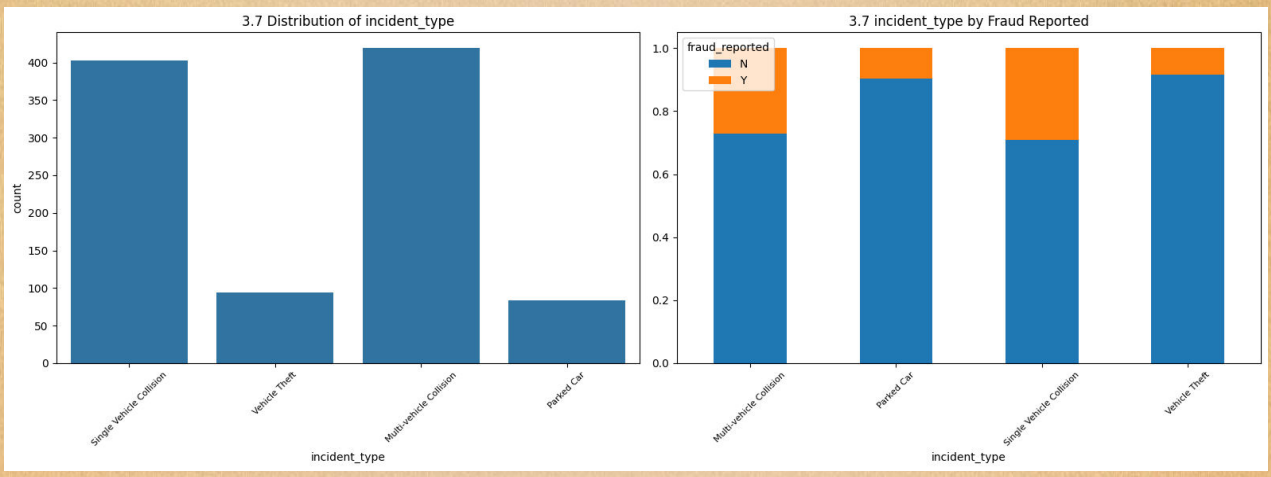
3.5 Categorical Feature Analysis: insured_hobbies ---

- Observation: Distribution of hobbies.
- Inference: Hobbies are distributed differently. Fraud rates vary slightly across hobbies.
- Summary: Hobby distribution and fraud rates.

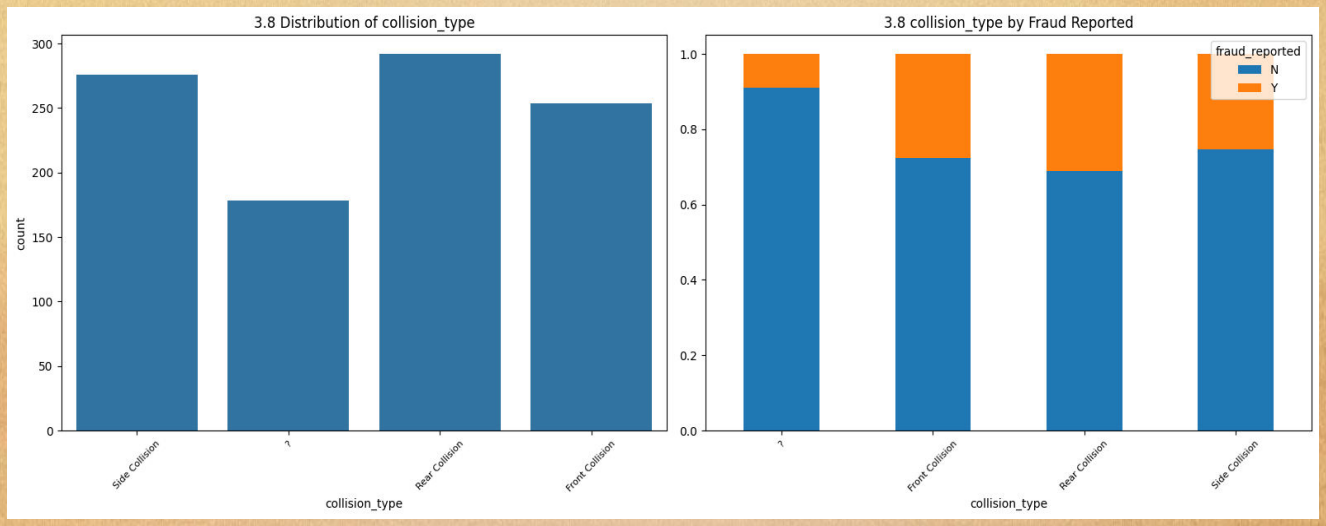


3.6 Categorical Feature Analysis: insured_relationship ---

- Observation: Distribution of relationships.
- Inference: Relationship to insured varies. Fraud rates vary by relationship type.
- Summary: Relationship distribution and fraud rates.

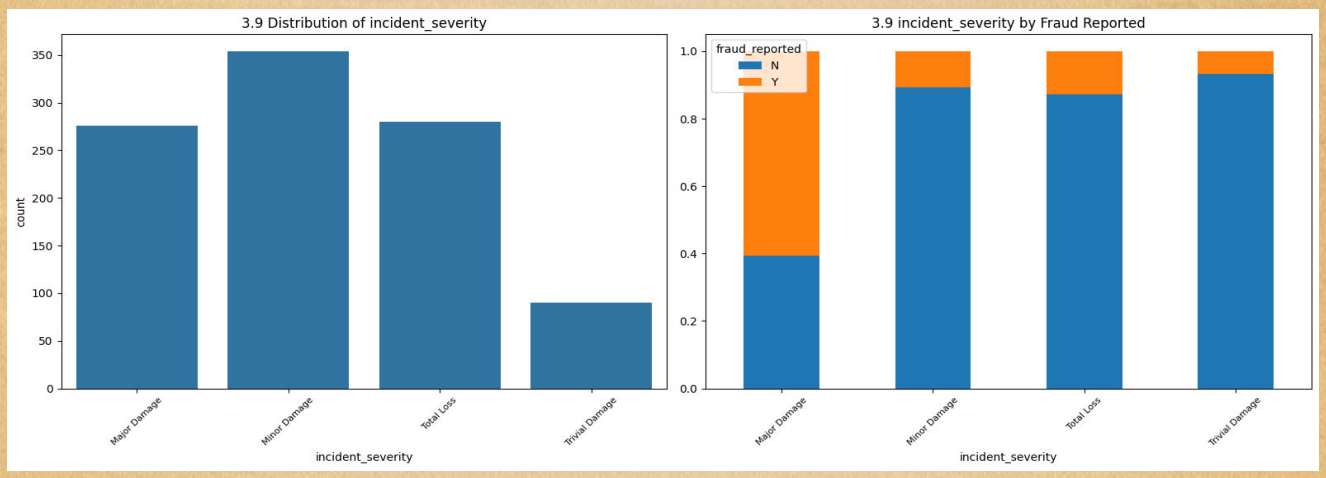


- **3.7 Categorical Feature Analysis: incident_type ---**
- Observation: Distribution of incident types.
- Inference: Some incident types have higher fraud rates.
- Summary: Incident type is a strong indicator of fraud.
-

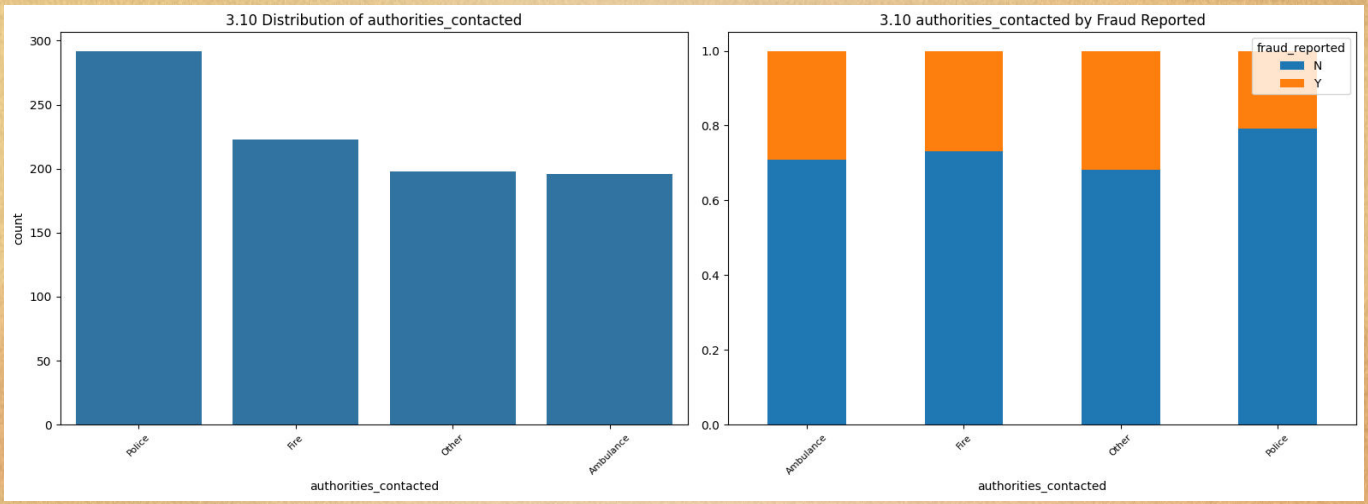


3.8 Categorical Feature Analysis: collision_type ---

- Observation: Distribution of collision types.
- Inference: Fraud rates vary by collision type.
- Summary: Collision type influences fraud rates.

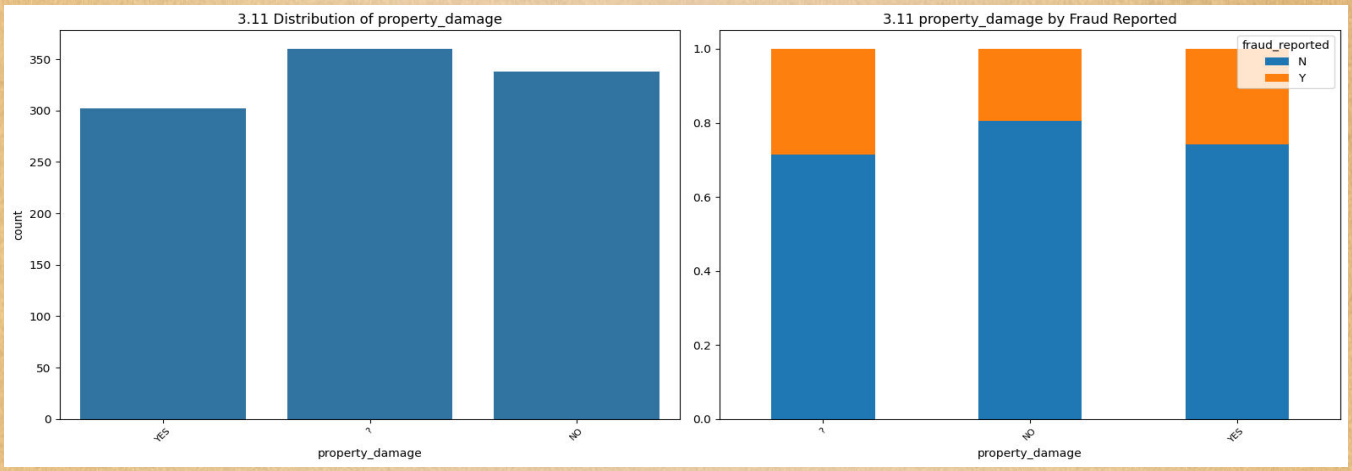


- **3.9 Categorical Feature Analysis: incident_severity ---**
- Observation: Distribution of incident severity.
- Inference: More severe incidents have higher fraud rates.
- Summary: Incident severity is a strong predictor of fraud.

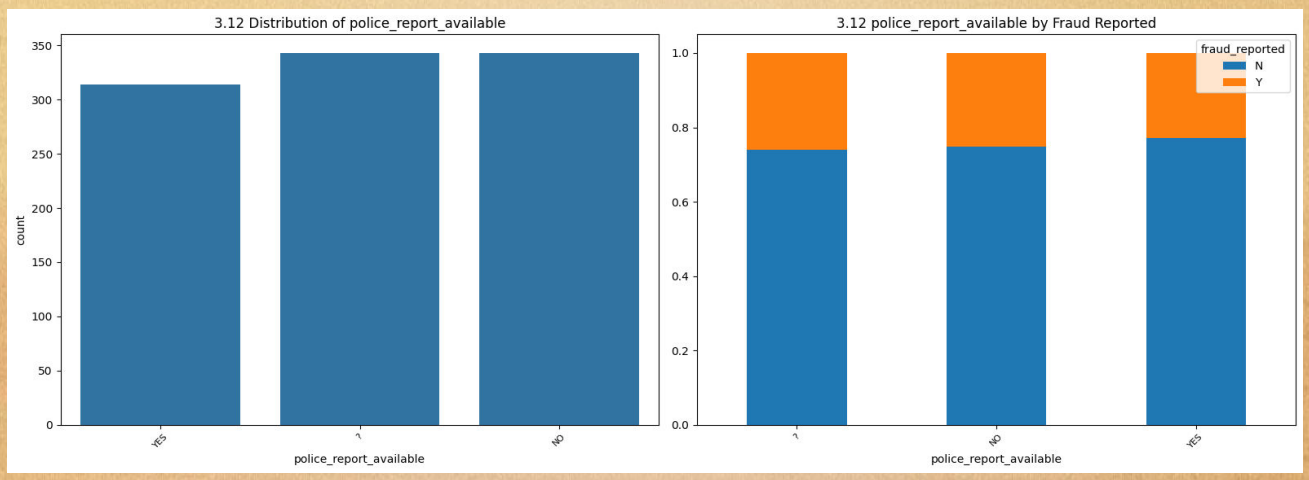


3.10 Categorical Feature Analysis: authorities_contacted ---

- Observation: Distribution of authorities contacted.
- Inference: Fraud rates vary by authority contacted.
- Summary: Authority contacted may relate to fraud.



- **3.11 Categorical Feature Analysis: property_damage ---**
- Observation: Distribution of property damage.
- Inference: Fraud rates vary based on property damage.
- Summary: Property damage is associated with fraud.
-

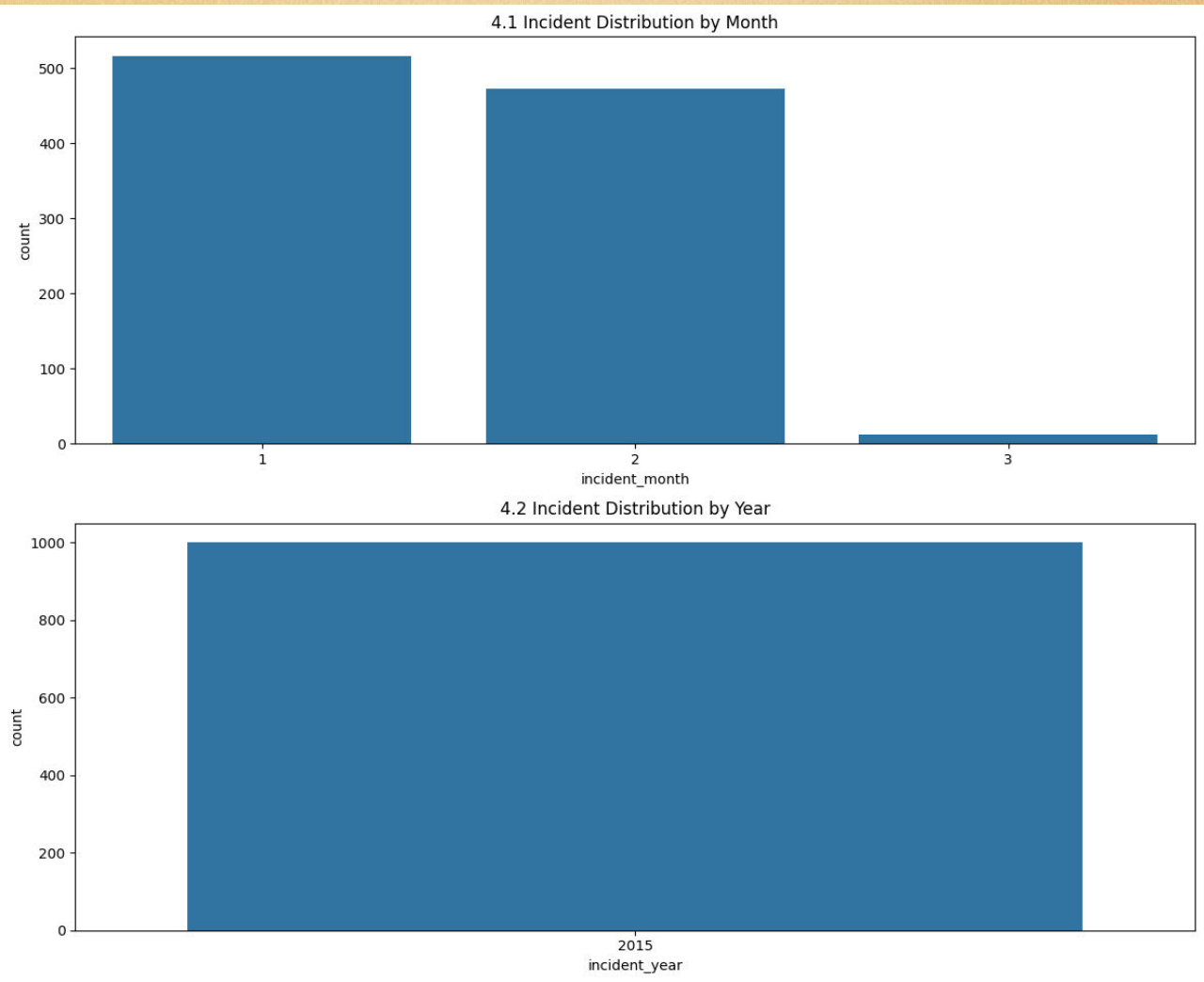


3.12 Categorical Feature Analysis: police_report_available ---

- Observation: Distribution of police report availability.
- Inference: Fraud rates differ based on police report availability.
- Summary: Police report availability is linked to fraud.

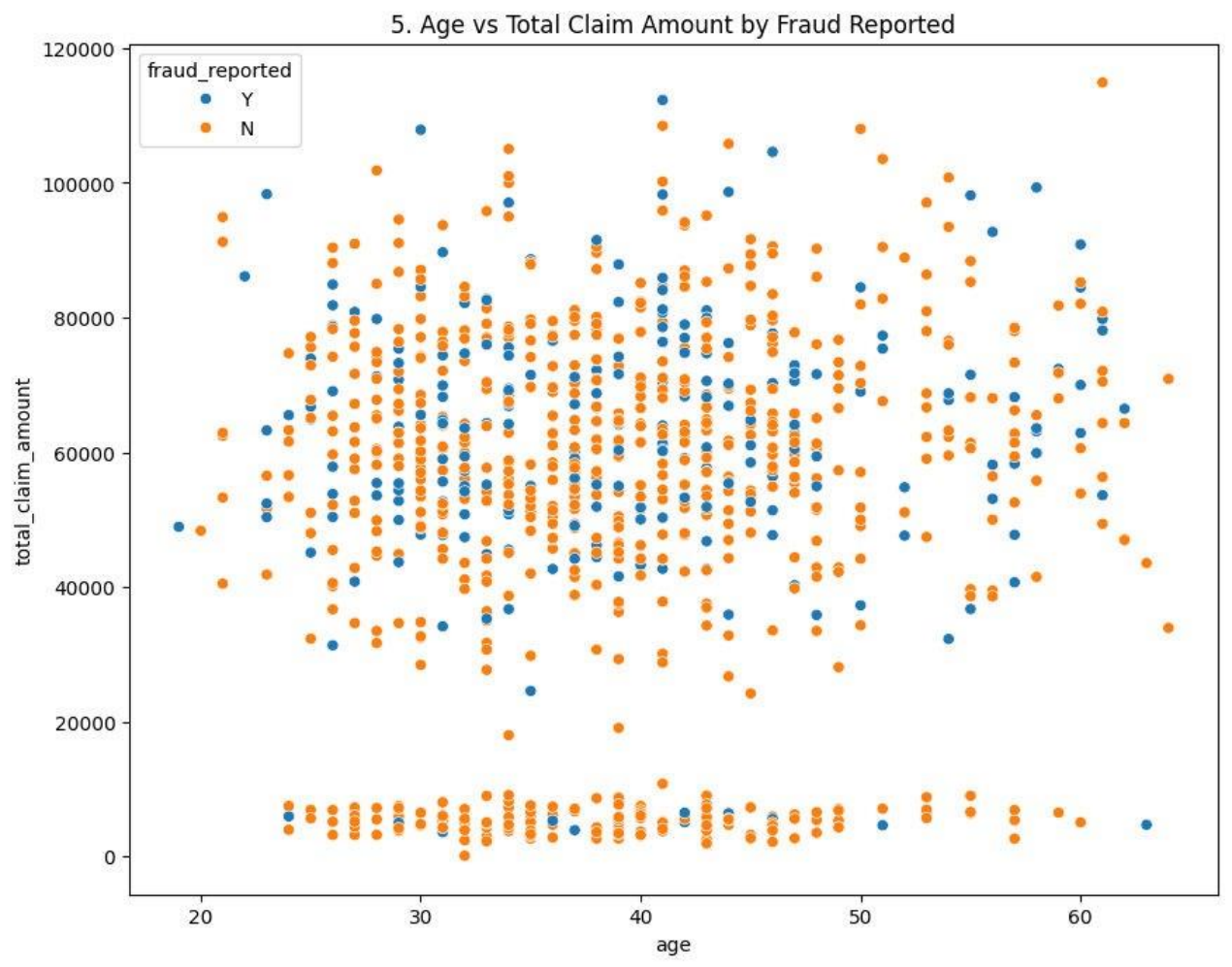
- Summary for Categorical Features Analysis
- Summary: Categorical features show varying distributions. Stacked bar plots visualize how fraud is distributed across different categories,

  helping identify potential risk factors.



4. Date/Time Feature Analysis ---

- Observation: The dataset contains incident dates primarily within a specific year (2015).
- Inference: Analysis of longer time periods might reveal seasonal or yearly trends in fraud.
- Observation: Incident counts vary by month.
- Inference: Some months might have a higher occurrence of incidents (and potentially fraud).
- Summary: Date features help understand the temporal aspect of the data.



5. Feature Interactions Analysis ---

- Observation: No clear pattern of fraud based solely on the interaction between age and total claim amount.
- Inference: Fraud is not strongly correlated with specific age groups or claim amounts in this simple bivariate analysis.
- Summary: Feature interactions can reveal more complex relationships, but in this case, age and total claim amount don't show a straightforward link to fraud.