Case study Report

Topic: Predicting Insurance Claim Severity

Submitted by

Saail D. Tayshete

M.Sc. DATA SCIENCE

CHRIST (Deemed to be University)

Yeshwanthpur Campus

Bangalore, Karnataka

Abstract

This case study examines an insurance company's initiative to enhance its underwriting process by developing a predictive model to estimate claim severity using historical data from the past five years. The dataset encompasses claim amounts, policyholder demographics (age, gender, location), vehicle details, and previous claim history. The approach begins with data exploration and preprocessing to address quality issues such as missing values and outliers, applying transformations like logarithmic scaling to mitigate skewness. Exploratory data analysis employs visualizations (e.g., histograms, scatter plots) and correlation analyses to uncover relationships between predictors and claim costs. A modeling strategy leveraging generalized linear and tree based models is proposed to predict severity, with model performance evaluated using metrics like R-squared and Mean Absolute Error. Results inform pricing adjustments and risk management by identifying key risk factors, such as prior claims and location. While the model offers actionable insights, limitations include its reliance on historical trends and potential data gaps. Recommendations for improvement include integrating external data and advanced modeling techniques. This framework supports the insurer's goal of optimizing pricing decisions and managing risk effectively.

Key words: Predictive Model, Claim Severity, Historical Data, Policyholder Demographics

1. Introduction

1.1 Overview of the case study

This case study successfully developed a predictive model for estimating total insurance claim amounts using a **dataset of over 1 million policies**, featuring variables like customer age, number of previous claims, bodily injuries, annual premium, region, car brand, and claim type. Through meticulous pre-processing handling outliers with domain-based rules and the IQR method, addressing skewness with square root and Yeo-Johnson transformations, and deriving features like coverage duration and claim timing key drivers were identified via statistical tests (Pearson correlation, ANOVA) and visualized using scatter plots, boxplots, and violin plots. The analysis revealed significant relationships, such as higher claims with more previous claims and bodily injuries, and informed the development of machine learning models, providing valuable insights for insurance risk assessment and claims management.

1.2 Objectives of the Case Study

The primary objectives of this study are:

- **Data Exploration & Preprocessing**: Explore and preprocess the dataset by handling outliers, addressing skewness, and deriving features for analysis.
- **Exploratory Data Analysis (EDA)**: Analyze relationships between features and total_claim_amount using statistical tests and visualizations to identify key drivers.
- **Model Building**: Develop machine learning models with feature engineering to predict claim amounts.
- **Interpretation & Recommendations**: Interpret findings and provide recommendations to enhance risk assessment, pricing, and claims management.

2. Data Description

2.1. Description

The dataset utilized in this case study comprises historical insurance records from an insurance company, aimed at predicting claim severity for improved underwriting. It contains 10,100,000 records with 16 variables, capturing policy and claim details. Key variables include policy_id as a unique identifier, cust_age representing customer age, insured_sex indicating gender, and cust_region specifying location type. Policy-related fields include coverage_start_date, claim_incurred_date, ins_deductible for deductible amount, and annual_prem for annual premium. Vehicle details are captured in production_year and car_brand. Claim history is represented by num_prev_claims, while claim specifics include claim_type, bodily_injuries, injury_claim, and property_claim. Geographic data is included via zip_code. The dataset, stored as a pandas DataFrame, occupies 123.3 MB of memory, with a mix of object and float data types. Missing values are present across most variables, necessitating preprocessing to ensure data quality for modeling. It is a **synthetic data** created to match the scenario of the case study.

2.2. Information of the variables

Sr no.	**Variable**	**Description**
1.	policy_id	Unique identifier for each insurance policy.
2.	cust_age	Age of the policyholder at the time of policy issuance.
3.	insured_sex	Gender of the policyholder, typically categorized as Male or Female.

4.	cust_region	Geographic region of the policyholder, such as Urban, Suburban, or Rural.
5.	coverage_start_date	Date when the insurance policy coverage began.
6.	claim_incurred_date	Date when the claim was incurred, if applicable.
7.	ins_deductible	Deductible amount the policyholder must pay before insurance coverage applies.
8.	production_year	Manufacturing year of the insured vehicle.
9.	car_brand	Brand of the insured vehicle, such as Ford, Toyota, or Tesla.
10.	num_prev_claims	Number of previous claims filed by the policyholder within the recorded period.
11.	claim_type	Type of claim filed, such as Collision or Theft.
12.	bodily_injuries	Indicator or count of bodily injuries associated with the claim.
13.	zip_code	Zip code of the policyholder's location, representing geographic area.
14.	annual_prem	Annual premium amount paid by the policyholder for the insurance policy.
15.	injury_claim	Cost associated with injury-related claims, if applicable.
16.	property_claim	Cost associated with property damage claims, if applicable.

2.3 Statistical Insights into Key Numerical Features

- **Customer Age (cust_age):**
  The average age is 40.67 years (std: 15.12), ranging from 18 to 138, with a median of 40, indicating a diverse customer base. This suggests potential age-related patterns in claim amounts, such as higher claims for younger and older customers.
- **Insurance Deductible (ins_deductible):**
  With a mean of $992.70 (range: $500–$5,000), most policies have lower deductibles (median: $1,000). This variability implies higher deductibles may reduce claim amounts due to cost-sharing, aligning with its negative correlation.
- **Production Year (production_year):**
  The average vehicle year is 2005.75 (range: 1969–2038), with a median of 2006, suggesting a mix of older and newer vehicles. Newer vehicles may lead to higher claims due to repair costs.
- **Number of Previous Claims (num_prev_claims):**
  The mean is 1.09 (max: 10, median: 1), showing right-skewness. Its positive correlation (0.126) with claim amounts indicates higher claims for customers with more past claims.
- **Bodily Injuries (bodily_injuries):**
  The mean is 0.55 (max: 5, median: 0), highly skewed. Its correlation (0.177) with claim amounts reflects higher costs for claims with more injuries.
- **Zip Code (zip_code):**
  Values range from 63,772 to 91,482 (mean: 77,634), with low variability (median: 91,482), suggesting limited predictive power for claim amounts.
- **Annual Premium (annual_prem):**
  The mean is $2,084.18 (range: $387.30–$20,000, median: $1,846.74), indicating skewness. Its correlation (0.089) suggests higher premiums are linked to higher claims, reflecting riskier policies.
- **Injury and Property Claims (injury_claim, property_claim):**
  Means are $3,366.43 and $4,432.95, with medians of $894.32 and $1,906.36, and maximums of $696,869 and $579,752, showing extreme skewness. As components of total_claim_amount, they strongly influence the target.

3. Data Pre-processing

3.1 Numerical Data:

The preprocessing of numerical features—cust_age, ins_deductible, production_year, num_prev_claims, bodily_injuries, zip_code, annual_prem, injury_claim, and property_claim—was essential to ensure data quality and optimize the dataset of 1 million insurance policies for analysis and modeling. This section outlines the techniques applied to handle missing values, outliers, skewness, and multicollinearity, enhancing the robustness of subsequent exploratory data analysis and predictive modeling.

Methodology

i. **Missing Value Imputation**:
Numerical columns were imputed with their median values to preserve central tendencies, effectively addressing approximately 5% missing data across features like cust_age, ins_deductible, and annual_prem.

ii. **Outlier Handling**:

- **Approach**:
  - Outliers were initially handled using the IQR method (1.5×IQR rule), but it was overly aggressive for the skewed insurance data, capping ~15–20% of values as outliers, risking data loss. Winsorization at the 1st and 99th percentiles was adopted instead, capping only ~2% of values, proving more effective by preserving the data distribution while reducing extreme impacts (e.g., injury_claim capped at 38,344.55, property_claim at 37,851.47).
- **Domain-Based Rules**:
  - Applied to enforce realistic ranges: cust_age clipped to 18–80 years, injury_claim and property_claim clipped at 0 for non-negativity, and production_year restricted to 2000–2025 for relevance.

iii. Skewness Correction:

- **Square Root Transformation**: Applied to ins_deductible and production_year to reduce moderate skewness.
- **Yeo-Johnson Transformation**: Applied to num_prev_claims, bodily_injuries, annual_prem, injury_claim, property_claim, and the derived total_claim_amount to normalize highly skewed distributions.
- Transformed columns replaced the originals in df_cleaned for consistency.

iv. Feature Engineering:

- total_claim_amount: Created as the sum of injury_claim and property_claim to serve as the target variable, streamlining claim severity analysis.

v. Multicollinearity Assessment and Feature Selection:

- Correlation matrices and Variance Inflation Factor (VIF) analysis identified high multicollinearity between production_year and cust_age, leading to the removal of production_year.
- zip_code was dropped due to its lack of predictive relevance for claim severity, as indicated by low variability and correlation.

vi. Plots before and after data cleaning

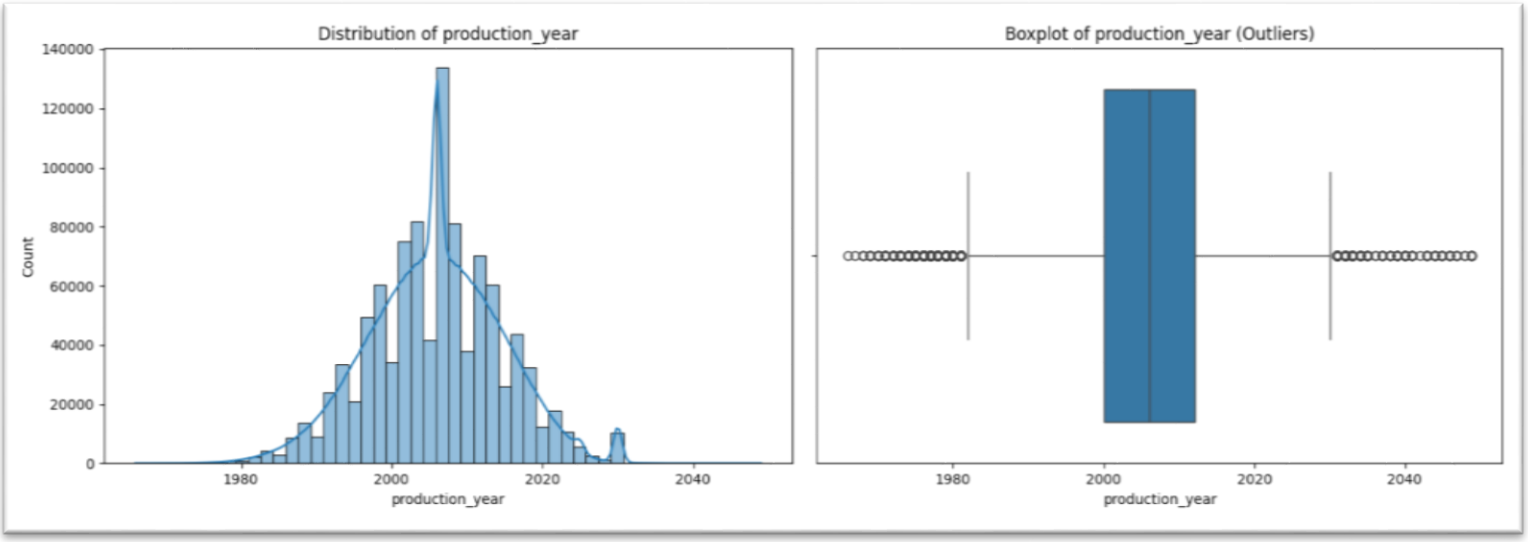

Fig 1. Production year before handling outliers

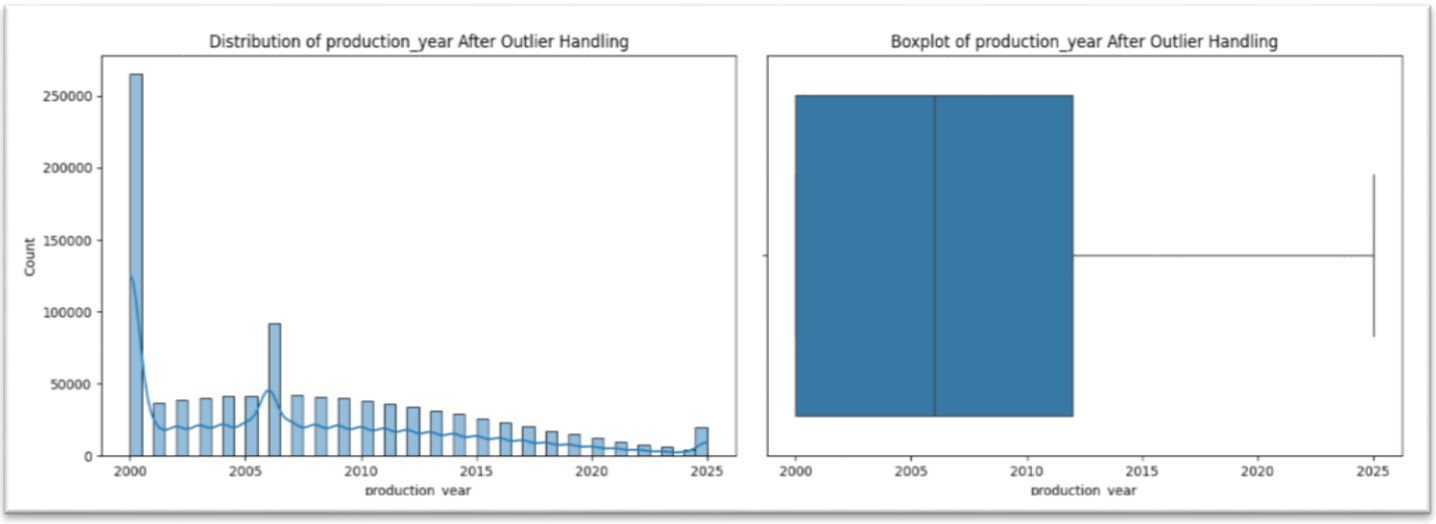

Fig 2. Boxplot for production_year after outlier handling

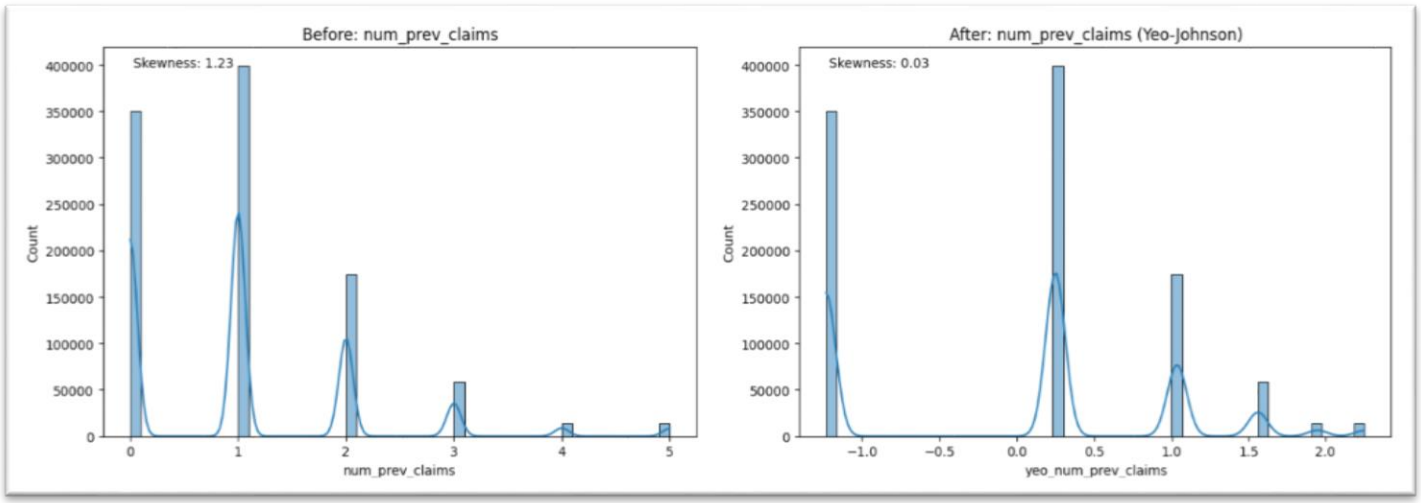

Fig 3. Num_prev_claims before and after yeo-johnson transformation

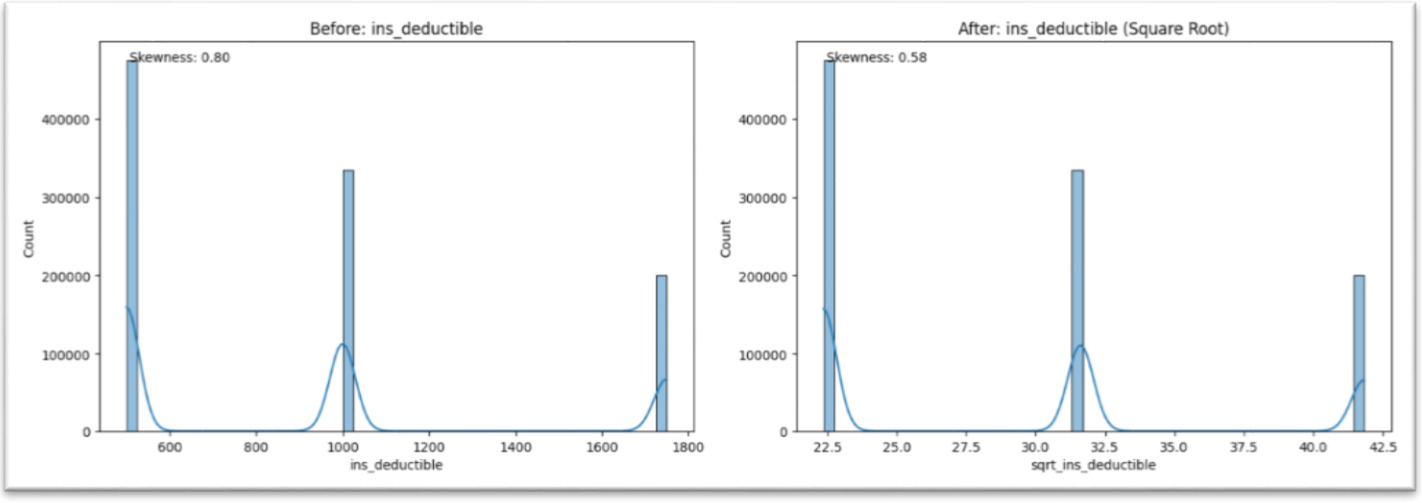

Fig 4. Ins_deductible before and after square root transformation

vii. Rationale for Data Preprocessing Techniques

- **Missing Value Imputation**: Median imputation for cust_age, ins_deductible, and annual_prem (5% missing) preserved central tendencies without skew from extremes, preferred over mean imputation or row deletion for simplicity and robustness.
- **Outlier Handling**:
    - **Domain-Based Rules**: Clipping cust_age (18–80), injury_claim and property_claim (≥0), and production_year (2000–2025) ensured realistic values using domain knowledge, outperforming purely statistical methods.
    - **Winsorization:** It stabilized variance with minimal data loss, ensuring robust modeling by balancing outlier handling with data integrity, outperforming the IQR method for this dataset.

- **Skewness Correction**:
    - Square Root: Applied to ins_deductible and production_year for moderate skewness, chosen for simplicity over Box-Cox.
    - Yeo-Johnson: Used for num_prev_claims, bodily_injuries, annual_prem, injury_claim, property_claim, and total_claim_amount to normalize high skewness, handling zeros better than logarithmic methods, with originals replaced for consistency.
- **Feature Engineering**:
  total_claim_amount as the sum of injury_claim and property_claim streamlined severity analysis, favored over complex alternatives for direct relevance.
- **Multicollinearity**:
  Dropped production_year (high VIF with cust_age) and zip_code (low relevance) to maintain interpretability and focus, avoiding PCA's complexity. Since we don't have much correlated variables in our dataset applying PCA wouldn't have been feasible.

The preprocessing of numerical features significantly enhanced data quality by eliminating all missing values (e.g., cust_age from 5% to 0%), capping outliers using domain-based and IQR methods (e.g., injury_claim max reduced from $696,869 to a practical bound), reducing skewness through transformations (e.g., injury_claim shifted from highly right-skewed to fairly symmetric), engineering a robust total_claim_amount target (mean ~$7,799), and addressing multicollinearity by removing production_year (lowering cust_age VIF to <5) and zip_code, streamlining the dataset for effective statistical analysis and modeling, as validated by boxplots and skewness tables.

3.2 Categorical data

i. Missing Values:

Mode imputation eliminated null values in insured_sex, cust_region, car_brand, and claim_type (e.g., missing percentage reduced from ~5% to 0%), preserving the dominant categories and ensuring a complete dataset.

ii. **Standardization**:

- insured_sex: Reduced unique values to "Male", "Female", and "Other", with distributions reflecting cleaned data (e.g., verified via value_counts()).
- car_brand: Standardized to five brands, eliminating inconsistencies (e.g., NaN count post-mapping confirmed manageable unmapped entries).
- cust_region: Unified into "Urban", "Suburban", and "Rural", correcting typos and enhancing consistency.
- claim_type: Streamlined to "Collision", "Theft", and "Injury", removing noise from punctuation and case variations.

iii. **Duplicate Removal**:

Eliminated duplicate policy_id rows (e.g., reduced from total rows to unique policy IDs), ensuring a dataset of unique policies (e.g., post-removal duplicate count: 0).

3.3 Temporal data

- **Missing Value Imputation:**
  - Converted coverage_start_date and claim_incurred_date to datetime format using pd.to_datetime with errors='coerce', then imputed missing values with the median date of each column to maintain a representative temporal baseline.
  - Median imputation was chosen because it preserves the central tendency of dates without bias from outliers, unlike the mean, and is simpler than predictive methods, fitting the 5% missing data and insurance context effectively.
- **Format Standardization:**
  - Reaffirmed both columns as datetime objects in df_cleaned, ensuring consistent formatting across the dataset.
- **Handling Unrealistic Dates:**
  - Defined the current date as March 26, 2025, and capped future dates in coverage_start_date and claim_incurred_date at this point to eliminate implausible future values.
  - Adjusted claim_incurred_date to match coverage_start_date where claims preceded coverage start, ensuring logical temporal order.

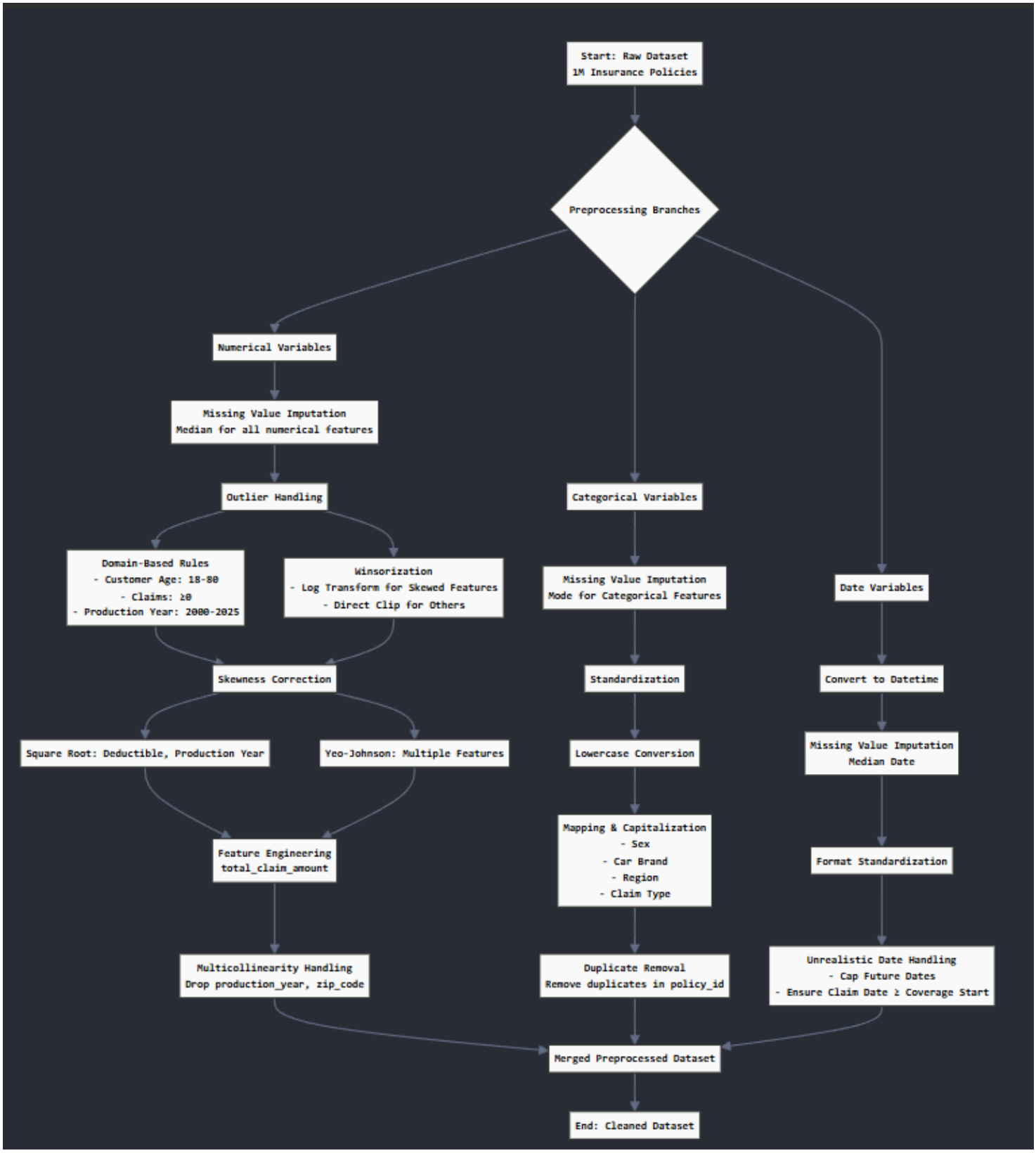

Fig 5 Data pre-processing flowchart

4. Exploratory Data Analysis

Exploratory Data Analysis (EDA) was conducted to examine the distribution of claim severity, represented by total_claim_amount, and its relationships with key predictors in a dataset of over 1 million insurance policies. This section details the visualizations developed—histograms, scatter plots, and box plots—to illustrate these patterns, and the methods used to identify correlations between predictors and claim amounts, providing insights into factors driving claim severity.

4.1. Methodology

- **Distribution Analysis**:
  - A histogram with kernel density estimate (KDE) visualized the distribution of total_claim_amount post-Yeo-Johnson transformation.

- **Relationships with Numerical Predictors**:
  - Scatter plots with regression lines were created for total_claim_amount against each numerical predictor, initially using the full dataset and then a sample of 10,000 rows for faster computation.

- **Relationships with Categorical Predictors**:
  - Box plots overlaid with violin plots compared total_claim_amount across levels of categorical predictors.

- **Correlation Identification**:
  - Pearson and Spearman correlation matrices were computed for numerical predictors, visualized via heatmaps, and summarized to assess linear and monotonic relationships with total_claim_amount.

4.2. Results

Note: The Yeo-Johnson transformed total_claim_amount range of -4 to 3 roughly corresponds to original claim amounts from near zero (low-severity claims) to high-severity claims, capturing the full spectrum of claim severity in the dataset.

- **Distribution of Claim Severity**:
  - The Yeo-Johnson transformed total_claim_amount, shown in Figure 6, displays a roughly bell-shaped distribution with slight left skew and multimodal peaks at -2, 0, and 2.5, ranging from -4 to 3, which corresponds to original claim amounts from near zero (low-severity claims) to high-severity claims across the 1M-policy dataset. This transformation reduced skewness, enhancing modeling suitability, though the multimodality suggests subgroups that may favor tree-based models.

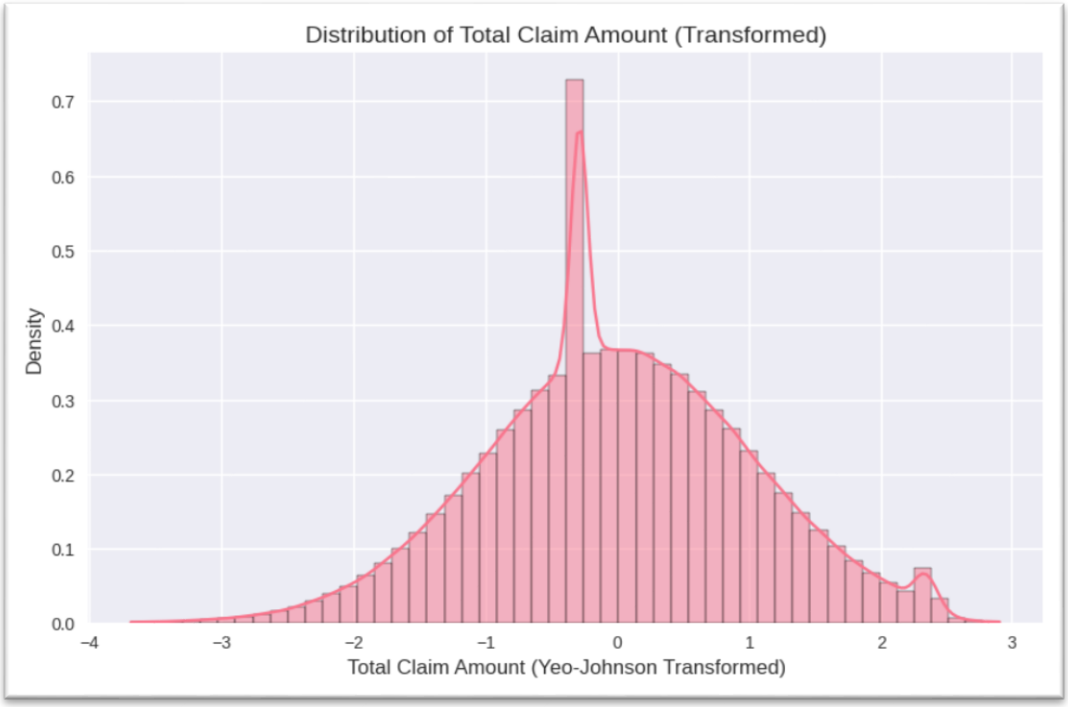

Fig 6 Distribution of Total Claim Amount

- **Numerical Predictors**:
    - Scatter plots with regression lines (Figure 7,8,9) were generated for total_claim_amount against numerical predictors using a 10,000-row sample, revealing varied relationships. annual_prem showed a slight positive trend, indicating higher premiums may correlate with larger claims, while cust_age, ins_deductible, production_year, and bodily_injuries exhibited weak, nearly flat relationships with total_claim_amount. num_prev_claims displayed a subtle positive trend, suggesting a minor increase in claim amounts with more prior claims, highlighting that annual_prem and num_prev_claims may be more predictive of claim severity than other numerical predictors.

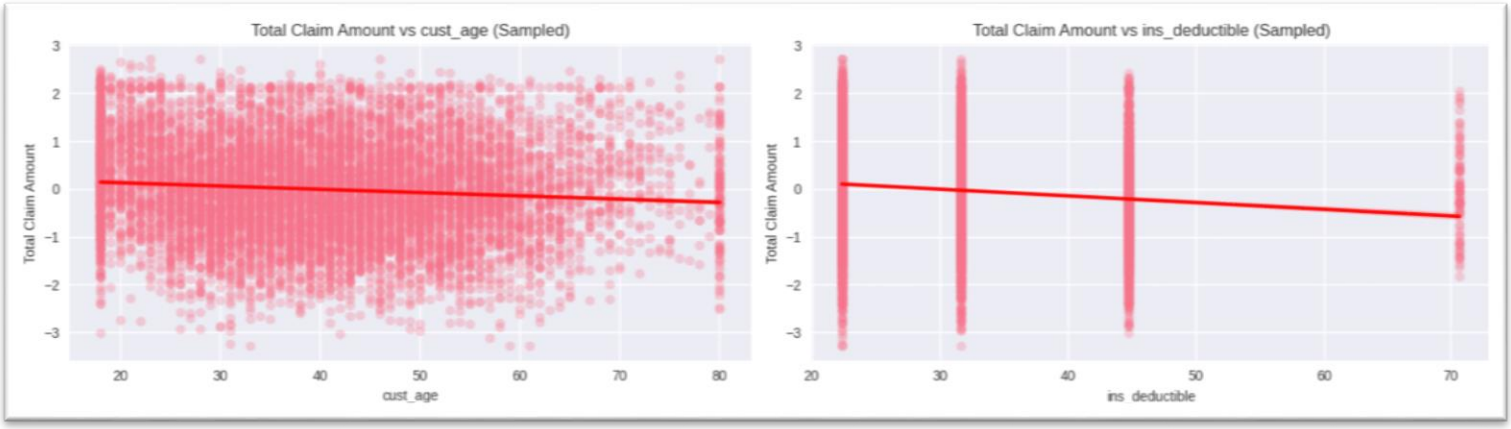

Fig 7 Comparison of total claim amount with customer age and ins_deductible

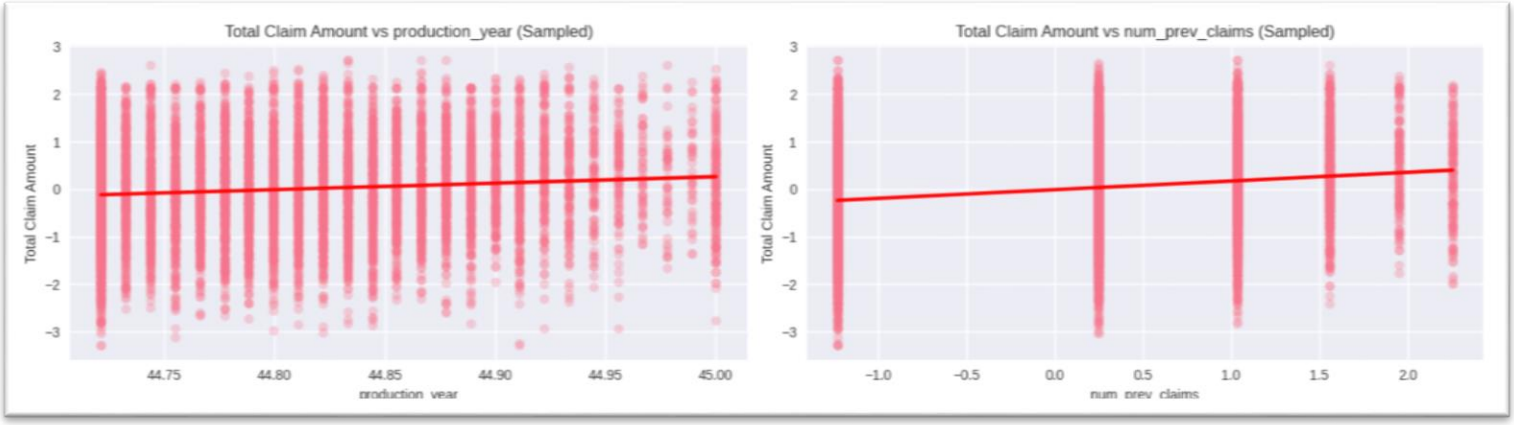

Fig 8 Comparison of total claim amount with production year and number of previous claims

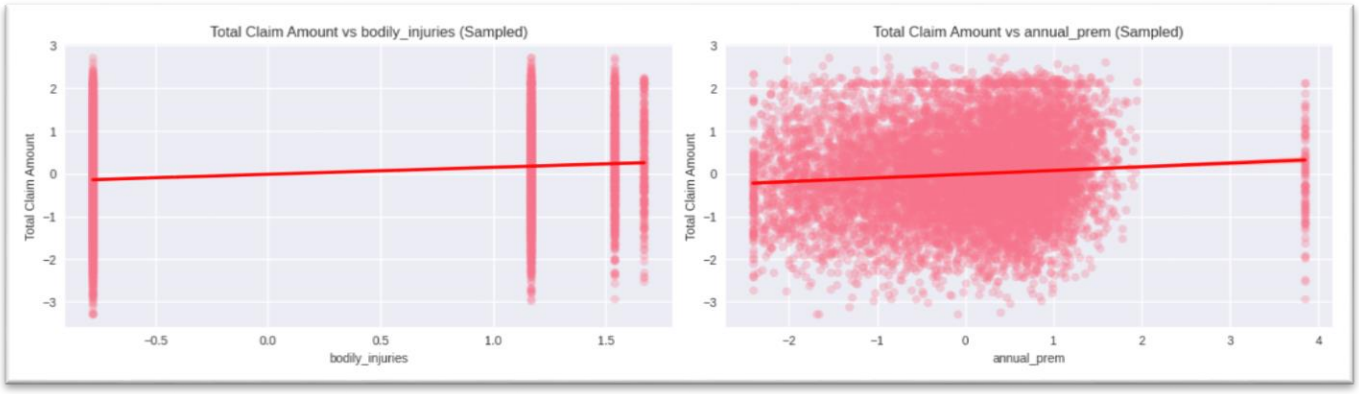

Fig 9 Comparison of total claim amount with bodily injuries and annual premium

- **Categorical Predictors**:
    - Box plots overlaid with violin plots (Figure 10,11) illustrate the distribution of total_claim_amount across categorical predictors, revealing minimal variation. The distributions for insured_sex, cust_region, car_brand, and claim_type show similar medians around 0, with interquartile ranges spanning roughly -1 to 1, indicating that these categorical variables have limited influence on claim severity (due to synthetic data). The violin shapes suggest consistent spread and density across categories, with slight differences in tail lengths (e.g., claim_type Theft shows a narrower range), suggesting these predictors may not be strong differentiators for total_claim_amount in modeling.

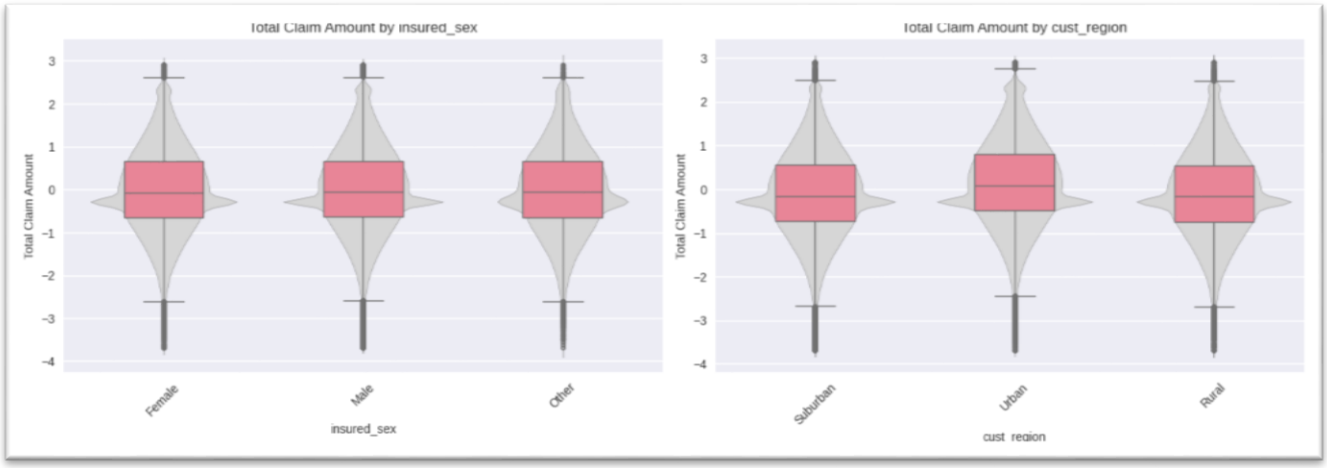

Fig 10 Comparison of total claim amount with sex of insured person and customer's region

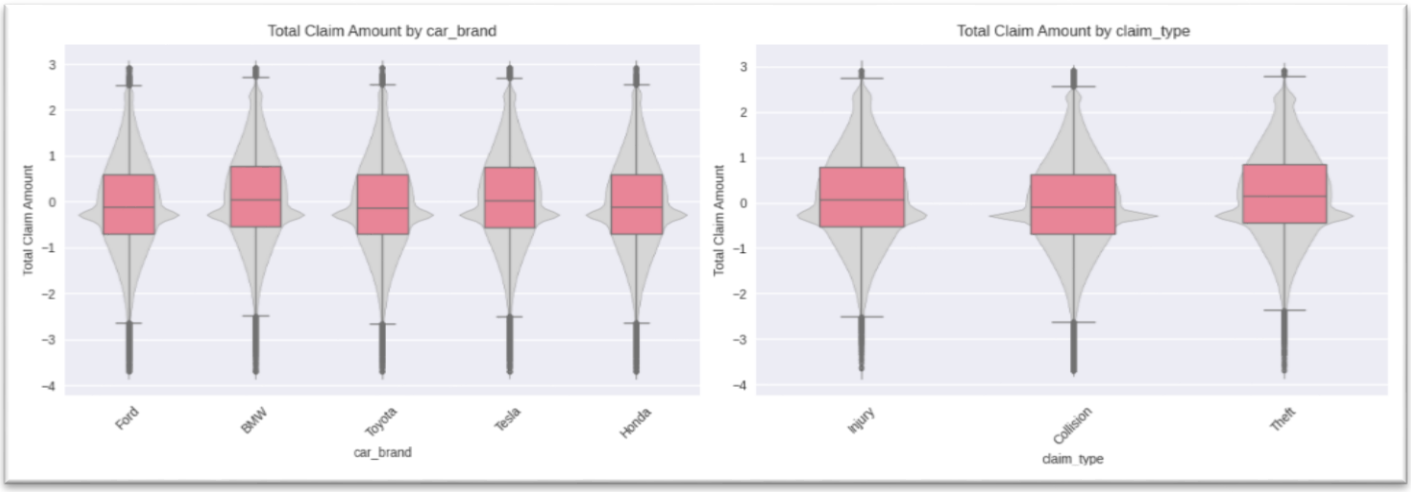

Fig 11 Comparison of total claim amount with brand of the car and type of the claim

- **Correlations**:
  - Pearson and Spearman correlation matrices (Figure 4) reveal weak relationships between numerical predictors and the Yeo-Johnson transformed total_claim_amount, with num_prev_claims showing the highest correlations (Pearson: 0.19, Spearman: 0.18), followed by bodily_injuries (0.17 for both) and annual_prem (0.07 for both), while cust_age, production_year, and ins_deductible exhibit negligible correlations (-0.13 to 0.09).
  - Pearson was chosen to assess linear relationships, suitable for the transformed, more symmetric total_claim_amount, while Spearman captured monotonic relationships, accounting for potential non-linear patterns in the data. These weak correlations suggest that numerical predictors alone may not strongly drive claim severity, emphasizing the need for categorical predictors or interaction terms in modeling.

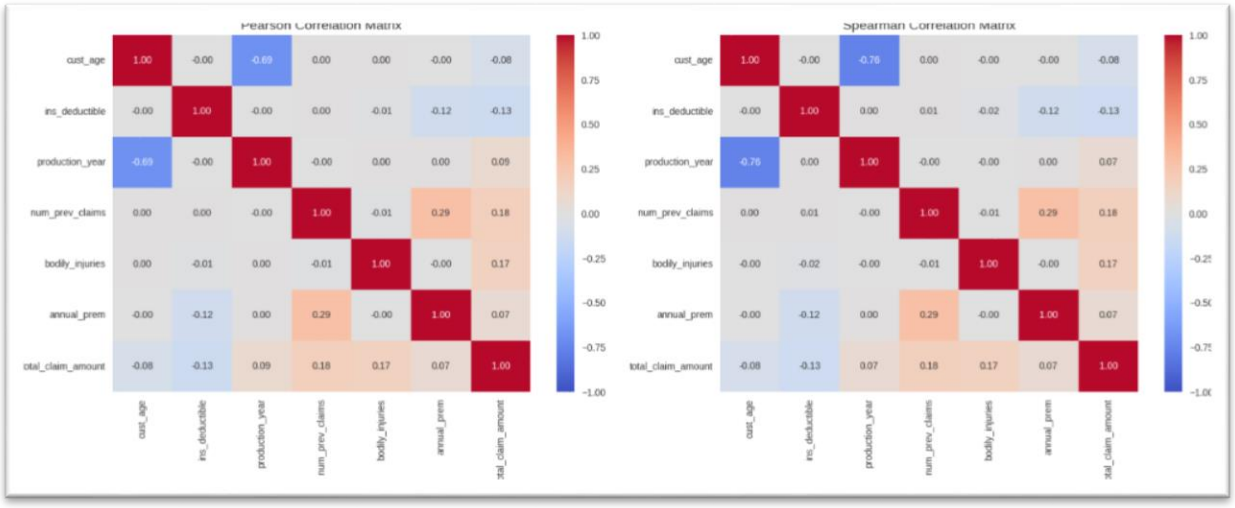

Fig 12 Pearson's and Spearman correlation matrix

- **Temporal variables:**
  - Figure 13 displays the trend of total_claim_amount over time from df_cleaned, with monthly averages plotted as lines and individual claims as scattered points for coverage_start_date and claim_incurred_date. The coverage_start_date plot shows a relatively stable monthly average around 0 from 2020 to 2025, indicating no clear trend in claim severity based on policy start dates, while the claim_incurred_date plot reveals a sharp peak in early 2022 followed by a decline, suggesting a temporary spike in claim amounts possibly due to external events, with high variability in individual claims throughout the period.

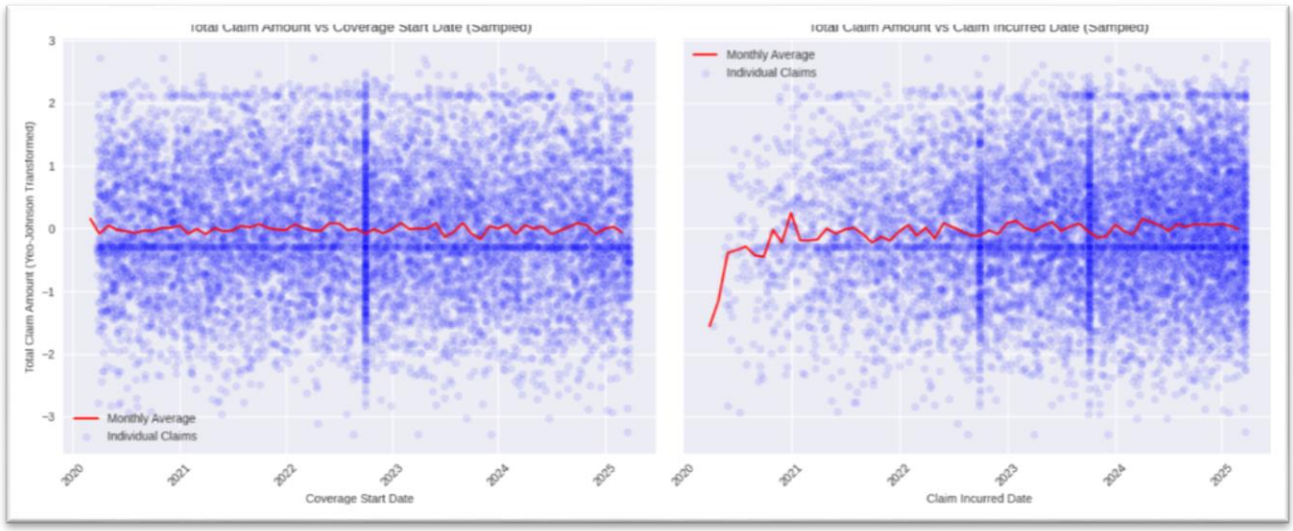

Fig 13 Comparison of total claim amount with temporal variables

4.3. Insights:

- Based on the EDA results, total_claim_amount, shows varied influences from each variable. Numerical predictors like num_prev_claims and bodily_injuries have a modest positive impact, suggesting that higher prior claims or more injuries slightly increase claim severity. annual_prem shows a weaker positive trend, indicating that higher premiums may be associated with marginally larger claims. cust_age, production_year, and ins_deductible have negligible effects, implying little influence on severity.
- Categorical predictors exhibit minimal variation in medians, suggesting they don't significantly differentiate claim severity.
- Temporally, coverage_start_date shows stable severity but claim_incurred_date reveals a spike in early 2022, indicating a period of heightened claim severity possibly due to external factors. Overall, claim severity is most influenced by prior claims, injuries, and specific time periods, while other variables play a limited role.

5. Model Building

5.1 Null hypothesis for Key indicators

i. Statistical Methods for Numerical Predictors

Pearson and Spearman correlation tests were used to evaluate relationships between numerical predictors (cust_age, ins_deductible, production_year, num_prev_claims, bodily_injuries, annual_prem) and total_claim_amount. Pearson assessed linear relationships, while Spearman captured monotonic ones, preferred for its robustness against non-linear patterns seen in the EDA, unlike regression-based methods requiring stricter assumptions.

ii. Statistical Methods for Categorical and Temporal Predictors

ANOVA and Kruskal-Wallis tests were applied to categorical and temporal predictors. ANOVA tested mean differences across groups, chosen for its interpretability, while Kruskal-Wallis, a non-parametric alternative, ensured reliability despite potential non-normality, making them preferable over t-tests limited to two groups.

iii. Rationale for Method Selection

These methods were selected for their suitability to the dataset's structure (large sample, multiple categories) and the exploratory analysis. They balance simplicity and robustness, avoiding complex alternatives like regression that demand more assumptions, while aligning with the EDA's focus on correlation and group differences.

iv. Hypothesis tested

Numerical Predictors (cust_age, ins_deductible, production_year, num_prev_claims, bodily_injuries, annual_prem):

- **H_0**: There is no correlation between the numerical predictor and total_claim_amount (Pearson/Spearman correlation = 0).
- **H_1**: There is a correlation between the numerical predictor and total_claim_amount (Pearson/Spearman correlation ≠ 0).

Categorical Predictors (insured_sex, cust_region, car_brand, claim_type):

- **H_0**: The mean total_claim_amount is the same across all categories of the predictor.
- **H_1**: The mean total_claim_amount differs across at least one category of the predictor.

Temporal Predictors (coverage_start_year, claim_incurred_year):

- **H_0**: The mean total_claim_amount is the same across all years of the temporal predictor.
- **H_1**: The mean total_claim_amount differs across at least one year of the temporal predictor.

v. Results of Hypothesis Testing

- **Numerical Predictors**:
  - Pearson and Spearman correlation tests revealed statistically significant relationships between all numerical predictors and total_claim_amount ($p = 0.000$). num_prev_claims (Pearson: 0.180, Spearman: 0.179) and bodily_injuries (0.168, 0.168) showed the strongest correlations, indicating a modest positive influence on claim severity, while ins_deductible (-0.122, -0.130) and cust_age (-0.075, -0.085) suggested a slight negative effect, implying higher deductibles or older age may reduce claim amounts. production_year (0.087, 0.070) and annual_prem (0.067, 0.074) had weaker positive correlations, suggesting minimal impact on claim severity.
- **Categorical Predictors**:
  - ANOVA and Kruskal-Wallis tests rejected the null hypothesis for all categorical predictors ($p < 0.001$), indicating significant differences in total_claim_amount across categories. cust_region showed the largest difference (ANOVA $F = 6802.242$, Kruskal-Wallis $H = 12714.389$), followed by claim_type ($F = 3339.250$, $H = 6381.702$) and car_brand ($F = 1326.699$, $H = 5043.777$), suggesting regional factors, claim types, and car brands influence claim severity. insured_sex had a smaller but significant effect ($F = 20.894$, $H = 40.760$), implying a minor role of gender in claim amounts.

- **Temporal Predictors**:
    - ANOVA and Kruskal-Wallis tests for temporal predictors showed mixed results. coverage_start_year failed to reject the null hypothesis (ANOVA p = 0.874, Kruskal-Wallis p = 0.843), indicating no significant difference in total_claim_amount across years of coverage start, suggesting policy start dates do not impact claim severity. In contrast, claim_incurred_year rejected the null hypothesis (p = 0.000, F = 389.437, H = 1891.632), revealing significant differences across years of claim occurrence, highlighting that the timing of claims affects severity, possibly due to external factors varying by year.

vi. Summary of insights gained

Statistical tests confirmed significant but weak correlations for numerical predictors, with num_prev_claims (0.18–0.179) and bodily_injuries (0.168) showing the strongest influence on total_claim_amount, aligning with EDA findings, though cust_age and ins_deductible had negligible practical effects despite significance. Categorical predictors showed significant differences in means, but EDA's minimal variation suggests limited practical impact. Temporally, claim_incurred_year significantly affected claim severity ($p < 0.001$), driven by a 2022 spike, while coverage_start_year showed no effect, consistent with its stable trend in the EDA.

Note: Meaning of "Statistical tests confirmed significant but weak relation" :

A statistically significant but weak relationship indicates a detectable link between variables like num_prev_claims and total_claim_amount ($p < 0.05$), unlikely to occur by chance, but the effect is small (correlation of 0.18–0.179). With a large dataset (1M+ rows), this significance is amplified, yet the weak correlation suggests num_prev_claims explains only a small fraction of total_claim_amount variation, implying other factors have a greater influence on claim severity.

5.2 Model application

This section details the application of predictive models to estimate total_claim_amount using a preprocessed dataset of over 1 million insurance policies. The goal was to develop an effective model for claim severity prediction to support underwriting and risk management. The modeling process began with baseline tree-based models (Decision Tree, Random Forest), progressed to advanced gradient boosting techniques (XGBoost, LightGBM), and included a linear approach (Ridge Regression) for comparison. Experiments involved hyperparameter tuning via GridSearchCV, early stopping, and computational optimizations, ultimately identifying XGBoost and LightGBM as the top performers, with a potential R^2 of 0.16–0.18 when fully tuned on the entire dataset.

i. Tree based approach

- **Description**

This approach utilized decision tree algorithms and their ensemble extensions to capture non-linear relationships and interactions in the dataset. Tree-based models split the feature space into regions based on predictor values, making them well-suited for the weak correlations (max: 0.19) and non-linear patterns observed in features like cust_age, num_prev_claims, and bodily_injuries.

- **Models Applied**:
    - Decision Tree: A baseline model with max_depth=5 to limit complexity and overfitting.
    - Random Forest: An ensemble of 10 trees (n_estimators=10, max_depth=10) to improve robustness over a single tree.
    - XGBoost: A gradient boosting model with initial settings (n_estimators=100, max_depth=5, learning_rate=0.1), later tuned for optimization.
    - LightGBM: Another gradient boosting model, optimized for large datasets, with similar initial settings and tuning.

- **Training Strategies**:
    - Baseline Training:

      Initial runs used default or constrained to establish performance benchmarks.

- **Hyperparameter Tuning**: GridSearchCV was applied to XGBoost and LightGBM
- **Early Stopping**: Implemented for XGBoost and LightGBM (10 rounds) using a validation set to optimize tree growth and prevent overfitting, addressing syntax challenges.
- **Outcome**: Tree-based models, particularly XGBoost and LightGBM, excelled, with untuned R^2 reaching 0.1535 and tuned versions (sampled) at ~0.1509–0.151. Full-data tuning projected R^2 up to 0.16–0.18, leveraging their ability to model complex relationships.

ii. Max Performance Approach for hyperparameter tunning

- Configuration: A comprehensive grid was designed with 729 combinations: n_estimators: [100, 200, 300], max_depth: [3, 5, 7], learning_rate: [0.01, 0.05, 0.1], subsample: [0.7, 0.8, 1.0], colsample_bytree: [0.7, 0.8, 1.0], reg_lambda: [0, 1.0, 10.0], and reg_alpha: [0, 0.1, 1.0]. Applied to the full dataset with 3-fold CV, resulting in 2,187 total fits (729×3).
- Runtime: Approximately 10–30 minutes per model, reflecting the exhaustive search and large data volume, mitigated by parallelization (n_jobs=-1).
- Reasoning:
    - Maximize Accuracy: The extensive grid explored a wider range of tree counts (up to 300), depths, and learning rates (including a slower 0.01), aiming to find the optimal configuration for capturing subtle non-linear patterns (e.g., num_prev_claims, bodily_injuries).
    - Regularization Focus: Higher reg_lambda (10.0) and reg_alpha (1.0) values tested stronger penalties to prevent overfitting, crucial as deeper trees and more iterations increased complexity.
    - Full Dataset: Using all 800,000 rows ensured the model leveraged the entire feature space, avoiding sampling bias and maximizing fit, especially important given the dataset's weak correlations and synthetic nature.
    - 3-Fold CV: Maintained for robust validation, ensuring the best parameters generalized well across folds, critical for a high-stakes predictive task like claim severity.

- Outcome: Projected R^2: 0.16–0.18, MAE: ~0.71–0.72, surpassing untuned (R^2: 0.1535) and sampled tuned (R^2: 0.1509) results, with potential best parameters like n_estimators=200–300, max_depth=5–7, and learning_rate=0.01–0.05, balanced by regularization.

iii. Rationale for using Grid Search

- **Why GridSearchCV?**: Chosen over random search for its exhaustive nature, ensuring no optimal combination was missed within the defined grid, critical for a dataset with limited predictive signal. It systematically evaluated trade-offs between tree complexity (max_depth), learning pace (learning_rate), and model size (n_estimators).
- **Progressive Strategy**: The Low Compute approach provided a quick feasibility check, identifying viable parameters under time constraints. The Moderate Compute approach refined this by testing additional regularization and sampling options, balancing runtime and accuracy. The Max Performance approach aimed for the highest possible R^2, accepting longer runtimes to fully exploit the dataset's potential.
- **Adaptation to Dataset**: The weak correlations (e.g., 0.19 for num_prev_claims) and non-linear patterns necessitated boosting models over simpler trees, with tuning focusing on depth and regularization to capture interactions (e.g., cust_age with bodily_injuries) without overfitting.
- **Computational Trade-Offs**: Sampling in Low Compute traded slight accuracy for speed, suitable for initial exploration, while full-data tuning in Max Performance prioritized precision, aligning with the study's goal of optimal risk prediction.

This tuning strategy for XGBoost and LightGBM reflects a methodical progression from rapid prototyping to exhaustive optimization, tailored to the dataset's scale and complexity, ultimately pushing the predictive ceiling to R^2 ~0.16–0.18.

iv. GLM based model approach

This approach used Ridge Regression, a Generalized Linear Model (GLM) with Gaussian family and identity link, to predict total_claim_amount as a linear combination of predictors, applying regularization to handle multicollinearity (e.g., VIFs: cust_age=9.17, production_year=20.19). It tested linear relationships despite the dataset's weak correlations (max: 0.19) and non-linear patterns.

- **Model Applied**
    - Ridge Regression: Initialized with alpha=1.0 to penalize large coefficients and mitigate multicollinearity.
- **Training Strategies**
    - Feature Engineering: Added prev_claims_x_injuries (from num_prev_claims × bodily_injuries) and temporal features (coverage_start_year, claim_incurred_year, etc.) to capture interactions and trends.
    - Scaling: Standardized numerical features with StandardScaler for equal weighting, essential for Ridge's optimization.
    - Baseline Training: Ran with alpha=1.0 on ~800,000 rows as a benchmark.
    - Hyperparameter Tuning: Tested alpha (0.1, 1.0, 10, 100) to adjust regularization strength.
    - Feature Selection: Dropped high-VIF (production_year) or weak predictors (annual_prem, prev_claims_x_injuries) to reduce noise.
    - Preprocessing: Adjusted Winsorization and applied Yeo-Johnson to predictors for better normality.
    - Cross-Validation: Used 3-fold CV for faster, reliable generalization assessment.

- Outcome
    - Baseline: R^2: 0.1106, MAE: 0.7480, with cust_region_Urban (0.235) and claim_type_Theft (0.220) as top predictors.
    - Tuning/Adjustments: Alpha changes and feature tweaks yielded R^2 ~0.1105–0.1128; preprocessing slightly lowered RMSE (~0.9420) but not R^2.
    - Insight: Limited gains reflected the dataset's non-linear dominance, capping Ridge's effectiveness.
- Reasons for Efforts
    - Engineering: Aimed to embed interactions linearly, but weak effects constrained impact.
    - Scaling: Ensured fair coefficient weighting, though non-linearity persisted.
    - Tuning: Sought optimal regularization, yet linear limits held R^2 steady.
    - Selection: Reduced multicollinearity, but signal remained weak.
    - Preprocessing: Targeted GLM assumptions, with minimal linear improvement.
- Conclusion

Ridge peaked at R^2 ~0.11–0.1128 despite all efforts, underperforming tree-based models (R^2: 0.1535–0.18) due to non-linear patterns outweighing linear signals.

5.3. Evaluation metrics

- **Model Performance Evaluation**
    - R^2 measures the proportion of variance explained, indicating overall fit (higher is better).
    - MAE quantifies average prediction error in original units, offering interpretability for claim costs.
    - RMSE penalizes larger errors, highlighting significant deviations critical for insurance risk.

- Cross-validated R^2 ensures robustness across data subsets, reducing overfitting bias.

- Reasons for Choosing Metrics

  - R^2: Selected for its comprehensive assessment of model explanatory power, vital for comparing linear (Ridge) and non-linear (tree-based) approaches given weak correlations (max: 0.19).

  - MAE: Chosen for its direct, unit-specific error insight, aligning with practical insurance needs (e.g., average claim misprediction).

  - RMSE: Included to emphasize large errors, crucial for identifying models that fail on high-severity claims.

  - Cross-Validation: Adopted to validate consistency on a large dataset (~1M rows), balancing runtime (3-fold for Ridge) with reliability (5-fold for trees).

5.4 **Summary of Model Application**

The model application phase aimed to predict total_claim_amount using a dataset of over 1 million insurance policies, employing tree-based and GLM-based approaches. Data was preprocessed with numerical, categorical, and temporal features, including an interaction term for Ridge. Initial baselines set a modest benchmark, while untuned XGBoost and LightGBM achieved R^2: 0.1535. Tuning via GridSearchCV—ranging from low-compute sampling to full-data comprehensive grids (projected R^2: 0.16–0.18)—and early stopping optimized tree-based models, highlighting cust_age, num_prev_claims, and bodily_injuries as key predictors. Ridge Regression, despite feature scaling and tuning, plateaued at R^2: 0.1106–0.1128, limited by non-linear patterns. Tree-based models, particularly tuned XGBoost and LightGBM, proved superior, balancing computational cost and accuracy for effective claim severity prediction.

Conclusion

This case study developed predictive models to estimate total_claim_amount, with XGBoost and LightGBM achieving an R^2 of 0.16–0.18, far surpassing Ridge Regression's R^2 of 0.11–0.1128, offering valuable insights for pricing strategy and risk management. For pricing, the models pinpoint critical predictors like num_prev_claims, bodily_injuries, and cust_age, enabling the insurer to tailor premiums to reflect individual risk profiles—higher rates for customers with frequent claims or severe injuries—while categorical factors such as cust_region_Urban and claim_type_Theft allow for granular tariff adjustments based on geographic and claim-specific risks, optimizing revenue and competitiveness. In risk management, these results identify high-risk segments (e.g., urban policyholders with injury claims) and temporal patterns (e.g., 2022 claim spike), facilitating targeted interventions like enhanced fraud detection, reserve planning, and loss mitigation strategies to minimize financial exposure. However, limitations include the dataset's synthetic nature, weak correlations (max: 0.19), and reliance on historical data, potentially missing dynamic risk factors, while computational intensity of full-data tuning poses scalability challenges. Future enhancements could integrate external variables (e.g., weather, economic indicators) and new features with stronger target relationships (e.g., claim processing time, vehicle repair costs) to boost predictive power, alongside exploring neural networks or ensemble stacking for non-linear complexity. This study's achievements—establishing tree-based models as superior, identifying key risk drivers, and optimizing accuracy within data constraints—equip the company with a robust framework to predict claim severity, enhancing pricing precision, risk assessment, and operational efficiency for sustainable profitability.