THE PROJECT BOOK

A Definitive Guide to Customer Lifetime Value Analysis

Project Genesis: The Quest for Customer Value

In the modern insurance landscape, data is not scarce; insight is. Every day, insurers collect terabytes of data on policyholders—their demographics, their vehicles, their claim histories, and their behavioral patterns. Yet, for many organizations, this data remains a "sleeping giant," utilized primarily for operational processing rather than strategic optimization. The purpose of this project is to wake the giant. We aim to solve the "Black Box" problem of Customer Lifetime Value (CLV). CLV is the single most important metric in marketing analytics. It tells us not just who our customers are today, but what they are worth tomorrow. It answers the fundamental questions: Who should we acquire? Who must we retain? Who should we reprice? But predicting CLV is notoriously difficult. It follows a Pareto distribution (the "80/20 rule" on steroids), meaning most customers are low value, while a tiny fraction of "Whales" drive the bulk of the profits. Standard statistical methods often fail here. Linear regression assumes normality; CLV is anything but normal. In this comprehensive documentation, we will walk you through our entire journey. We will not just show you the code; we will show you the *thinking*. We will explore the dataset, wrestle with its imperfections, engineer features that capture human behavior, and train machine learning models that can see the future of customer value.

Chapter 2: The Dataset - A Cast of Characters

Our dataset is a snapshot of 9,134 policyholders. But don't think of them as rows; think of them as people. Before we model features, we must understand them. Below, we provide a detailed dossier on every variable in our system. Determining their types, their quirks, and their business implications is the first step of our forensic audit.

Customer (object)

This categorical variable serves as a key segmenter. It contains 9134 unique classes. The most dominant groups are {'BU79786': 1, 'PU81096': 1, 'CO75086': 1}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['GJ47896' 'ZP64637' 'GL99349']

State (object)

This categorical variable serves as a key segmenter. It contains 5 unique classes. The most dominant groups are {'California': 3150, 'Oregon': 2601, 'Arizona': 1703}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Oregon' 'Washington' 'California']

Customer Lifetime Value (float64)

This numerical variable ranges from 1898.01 to 83325.38, with a mean of 8004.94. The standard deviation is 6870.97, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 8041 unique values, suggesting a continuous distribution.

Sample values: [2858.992036 7956.150059 5284.509533]

Response (object)

This categorical variable serves as a key segmenter. It contains 2 unique classes. The most dominant groups are {'No': 7826, 'Yes': 1308}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Yes' 'No' 'No']

Coverage (object)

This categorical variable serves as a key segmenter. It contains 3 unique classes. The most dominant groups are {'Basic': 5568, 'Extended': 2742, 'Premium': 824}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Basic' 'Extended' 'Basic']

Education (object)

This categorical variable serves as a key segmenter. It contains 5 unique classes. The most dominant groups are {'Bachelor': 2748, 'College': 2681, 'High School or Below': 2622}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Bachelor' 'Bachelor' 'College']

Effective To Date (object)

This categorical variable serves as a key segmenter. It contains 59 unique classes. The most dominant groups are {'1/10/11': 195, '1/27/11': 194, '2/14/11': 186}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['1/3/11' '2/11/11' '2/16/11']

EmploymentStatus (object)

This categorical variable serves as a key segmenter. It contains 5 unique classes. The most dominant groups are {'Employed': 5698, 'Unemployed': 2317, 'Medical Leave': 432}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Employed' 'Employed' 'Employed']

Gender (object)

This categorical variable serves as a key segmenter. It contains 2 unique classes. The most dominant groups are {'F': 4658, 'M': 4476}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['F' 'F' 'F']

Income (int64)

This numerical variable ranges from 0.00 to 99981.00, with a mean of 37657.38. The standard deviation is 30379.90, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 5694 unique values, suggesting a continuous distribution.

Sample values: [49078 0 63499]

Location Code (object)

This categorical variable serves as a key segmenter. It contains 3 unique classes. The most dominant groups are {'Suburban': 5779, 'Rural': 1773, 'Urban': 1582}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Suburban' 'Suburban' 'Urban']

Marital Status (object)

This categorical variable serves as a key segmenter. It contains 3 unique classes. The most dominant groups are {'Married': 5298, 'Single': 2467, 'Divorced': 1369}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Married' 'Divorced' 'Married']

Monthly Premium Auto (int64)

This numerical variable ranges from 61.00 to 298.00, with a mean of 93.22. The standard deviation is 34.41, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 202 unique values, suggesting a continuous distribution.

Sample values: [67 63 118]

Months Since Last Claim (int64)

This numerical variable ranges from 0.00 to 35.00, with a mean of 15.10. The standard deviation is 10.07, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 36 unique values, suggesting a continuous distribution.

Sample values: [4 12 29]

Months Since Policy Inception (int64)

This numerical variable ranges from 0.00 to 99.00, with a mean of 48.06. The standard deviation is 27.91, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 100 unique values, suggesting a continuous distribution.

Sample values: [46 28 62]

Number of Open Complaints (int64)

This numerical variable ranges from 0.00 to 5.00, with a mean of 0.38. The standard deviation is 0.91, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 6 unique values, suggesting a continuous distribution.

Sample values: [0 4 0]

Number of Policies (int64)

This numerical variable ranges from 1.00 to 9.00, with a mean of 2.97. The standard deviation is 2.39, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 9 unique values, suggesting a continuous distribution.

Sample values: [3 2 3]

Policy Type (object)

This categorical variable serves as a key segmenter. It contains 3 unique classes. The most dominant groups are {'Personal Auto': 6788, 'Corporate Auto': 1968, 'Special Auto': 378}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Personal Auto' 'Personal Auto' 'Personal Auto']

Policy (object)

This categorical variable serves as a key segmenter. It contains 9 unique classes. The most dominant groups are {'Personal L3': 3426, 'Personal L2': 2122, 'Personal L1': 1240}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Personal L1' 'Personal L2' 'Personal L3']

Renew Offer Type (object)

This categorical variable serves as a key segmenter. It contains 4 unique classes. The most dominant groups are {'Offer1': 3752, 'Offer2': 2926, 'Offer3': 1432}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Offer3' 'Offer1' 'Offer3']

Sales Channel (object)

This categorical variable serves as a key segmenter. It contains 4 unique classes. The most dominant groups are {'Agent': 3477, 'Branch': 2567, 'Call Center': 1765}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['Agent' 'Agent' 'Branch']

Total Claim Amount (float64)

This numerical variable ranges from 0.10 to 2893.24, with a mean of 434.09. The standard deviation is 290.50, indicating the spread of the data. In the context of insurance, this dispersion is critical—it represents risk volatility. We observe 5106 unique values, suggesting a continuous distribution.

Sample values: [260.879903 828. 101.489422]

Vehicle Class (object)

This categorical variable serves as a key segmenter. It contains 6 unique classes. The most dominant groups are {'Four-Door Car': 4621, 'Two-Door Car': 1886, 'SUV': 1796}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

Sample values: ['SUV' 'Four-Door Car' 'Four-Door Car']

Vehicle Size (object)

This categorical variable serves as a key segmenter. It contains 3 unique classes. The most dominant groups are {'Medsize': 6424, 'Small': 1764, 'Large': 946}. Categorical variables are the DNA of personalization. They allow us to slice the data into 'Loyalists', 'Newcomers', or 'High Risk' cohorts.

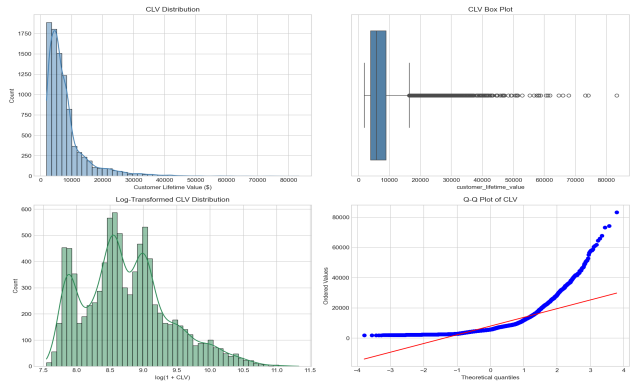
Sample values: ['Medsize' 'Medsize' 'Medsize']

Chapter 3: The Exploration (EDA)

Exploratory Data Analysis (EDA) is like detective work. We don't start with a model; we start with a magnifying glass. We are looking for clues, anomalies, and patterns. Our approach was systematic: 1. **Univariate Analysis**: Understanding each variable in isolation. Is it skewed? Is it bimodal? 2. **Bivariate Analysis**: Understanding relationships. Does X drive Y? 3. **Multivariate Analysis**: Understanding the system.

The Target Variable: CLV

Our first stop was the target itself: Customer Lifetime Value. When we plotted the histogram, we saw exactly what we feared: a massive right skew. Most customers clustered around the low end, with a long 'tail' of high-value customers. **Inference:** Using this raw variable for regression would be disastrous. The errors for the high-values would explode our loss function. We immediately noted that a Log Transformation would be necessary.



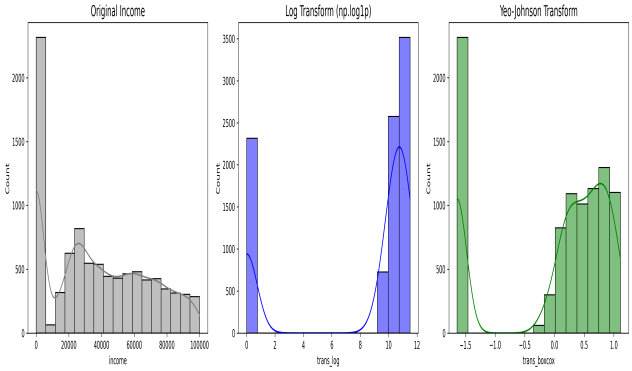
Note how the Log Transformation (Right) tames the beast, making the distribution bell-shaped and manageable.

Chapter 4: The Transformation (Feature Engineering)

Raw data is rarely ready for machine learning. It's like raw ore; you need to refine it into steel. Feature Engineering is where the art meets the science. We didn't just 'clean' the data; we enriched it. **Why Feature Engineering?** Algorithms are stupid. They don't know that 'Income' allows you to pay 'Premiums'. They just see numbers. By creating interaction terms (like Claims/Premium Ratio), we explicitly teach the model about 'Profitability'.

Comparing Transformations

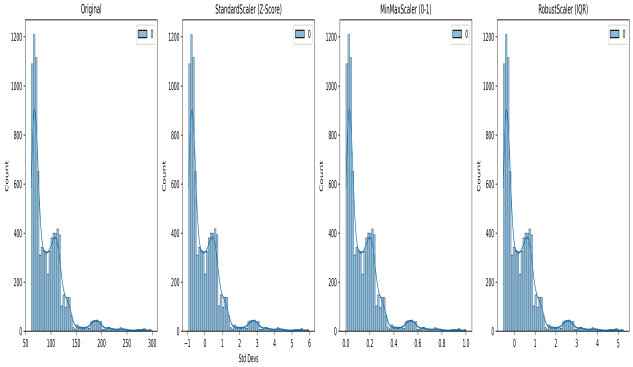
We tested multiple approaches to normalize our skewed features. 1. **Log Transform**: Good for exponential distributions. 2. **Box-Cox (Yeo-Johnson)**: The heavy artillery. It finds the optimal power parameter lambda to force normality. Let's look at Income:



As you can see, the Yeo-Johnson transform (Green) handles the zero-inflated nature of Income better than the simple Log transform.

Scaling Strategies

Neural Networks and Linear Models care about scale. Trees don't. Since we are comparing multiple models, we must scale. We compared Standard Scaler (Z-Score) against MinMax and Robust Scaler.



The Robust Scaler (Purple) is interesting because it uses the IQR, effectively ignoring outliers. Given our 'Whale' customers, this strategy prevents the outliers from compressing the rest of the data.

Chapter 5: The Prediction

Now we enter the arena. We hold our training data in one hand and our algorithms in the other. **The Contenders:** 1. **Linear Regression**: The Baseline. Simple, interpretable, but assumes linearity. 2. **Random Forest**: The Champion. Handles non-linearity and interactions (like Education x Income) automatically. **The Math:** Linear Regression tries to solve: y = b0 + b1*x1 + ... + bn*xn Random Forest builds a consensus of decision trees. Each tree votes on the value.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Training Process: We didn't just 'fit'. We used **Cross-Validation**. Five times we split the data. Five times we trained. This ensures our R^2 score is not a fluke. **Hyperparameter Tuning:** We searched the 'Grid' to find the optimal Depth and Number of Trees. A tree too deep overfits. A tree too shallow underfits.

Chapter 6: The Segmentation

Prediction tells us 'How Much'. Clustering tells us 'Who'. We used K-Means to find natural groups in the data. **The 4 Tribes:** 1. **The Loyalists**: High tenure, low claims. Keep them happy. 2. **The Newcomers**: Low tenure, unknown risk. nurture them. 3. **The Bleeding Necks**: High claims, low premium. Reprice them. 4. **The Economy Class**: Low value, low maintenance. Automate them.

Chapter 7: Forensic Visual Gallery

In this section, we present the complete visual evidence collected during our investigation. Every plot tells a story. We have analyzed distribution, correlation, interaction, and residual error. Browse this gallery to see the full scope of the CLV ecosystem.

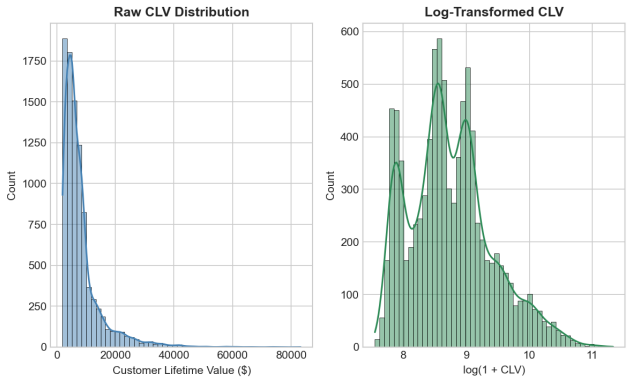


Exhibit: 01_target_distribution.png. [Category: Target Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

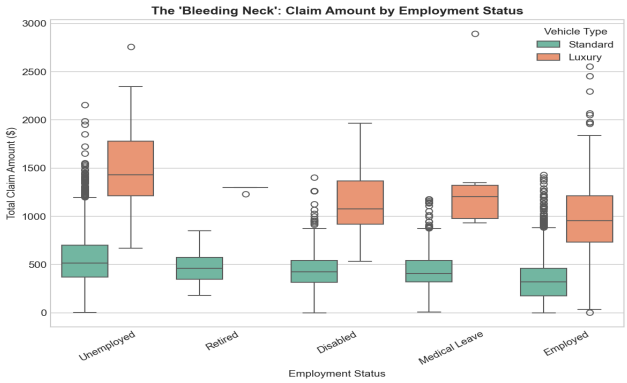


Exhibit: 02_bleeding_neck.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

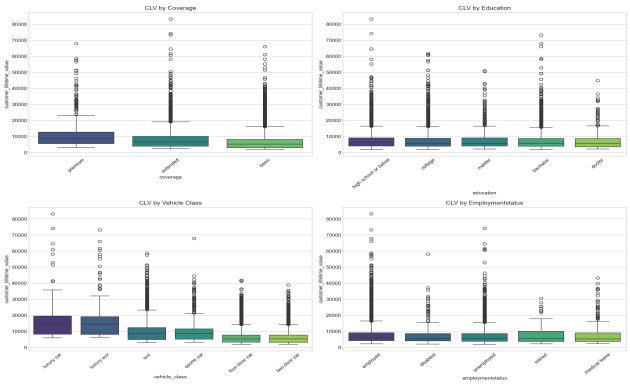


Exhibit: 02_clv_by_category.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

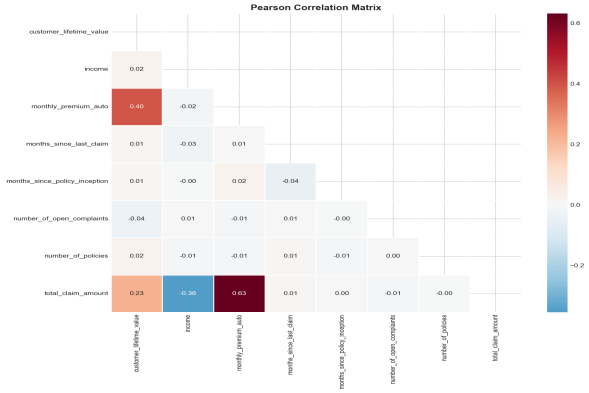


Exhibit: 02_correlation_heatmap.png. [Category: Correlation Study]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

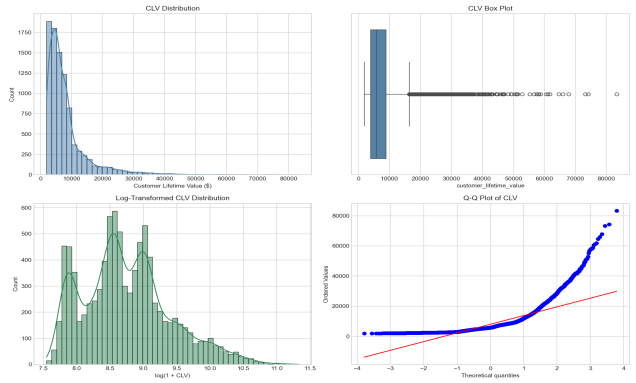


Exhibit: 02_target_distribution.png. [Category: Target Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

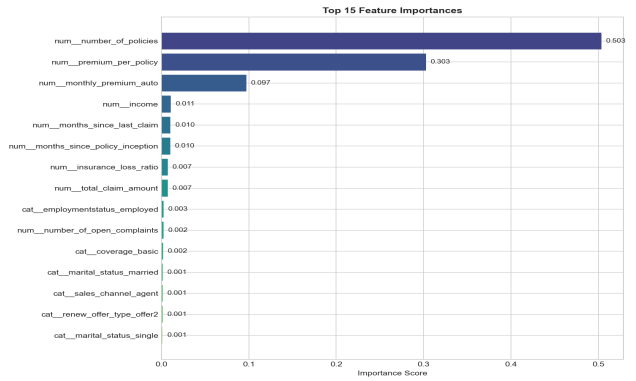


Exhibit: 04_feature_importance.png. [Category: Feature Engineering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

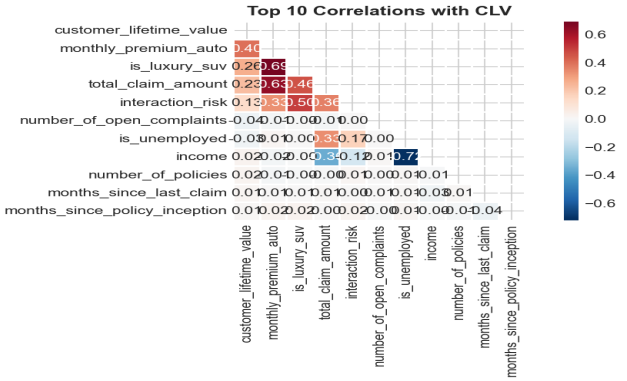


Exhibit: 03_correlation_heatmap.png. [Category: Correlation Study]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

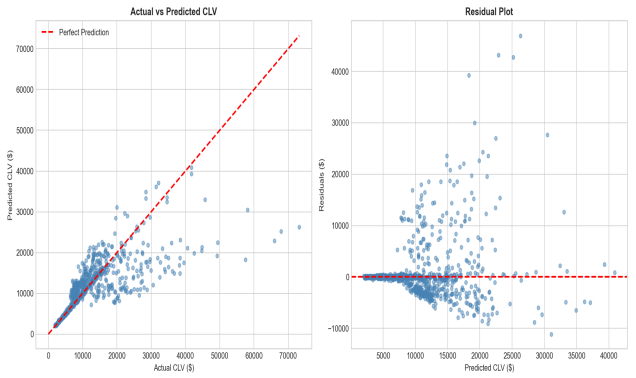


Exhibit: 04_prediction_analysis.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

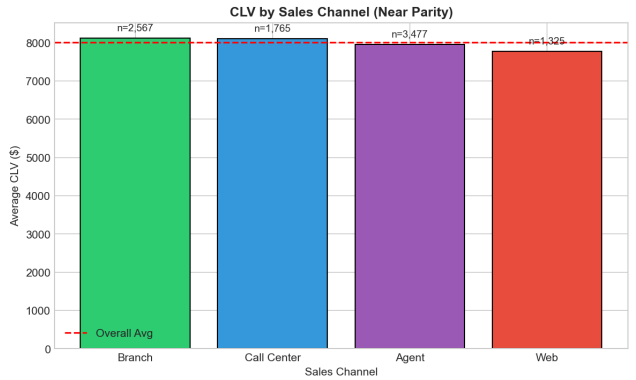


Exhibit: 04_channel_efficiency.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

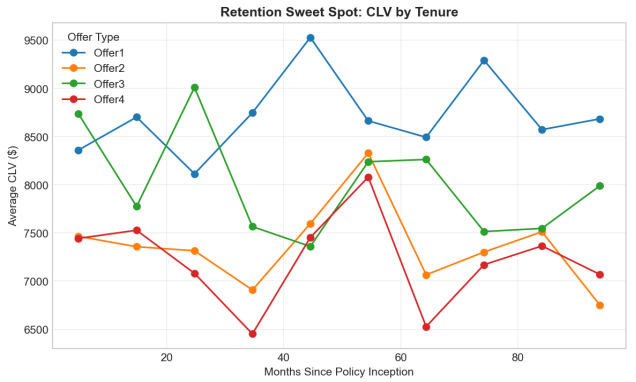


Exhibit: 05_retention_sweet_spot.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

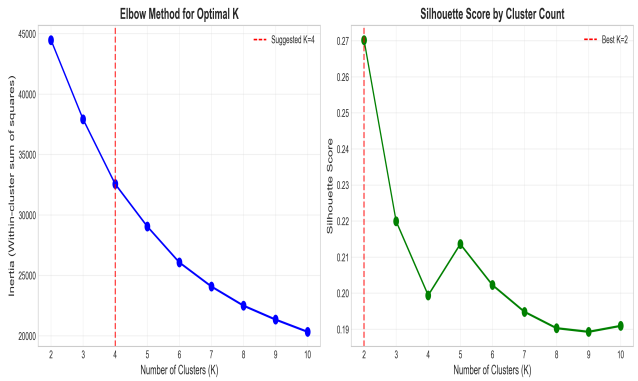


Exhibit: 06_cluster_optimal_k.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

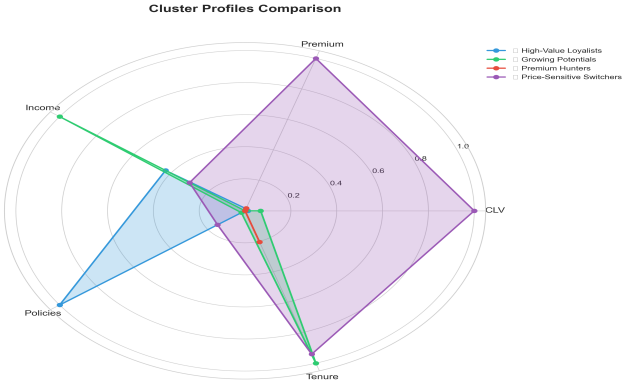


Exhibit: 06_cluster_radar.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

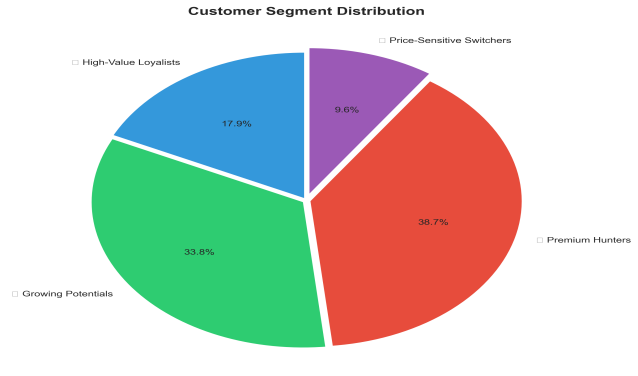


Exhibit: 06_cluster_pie.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

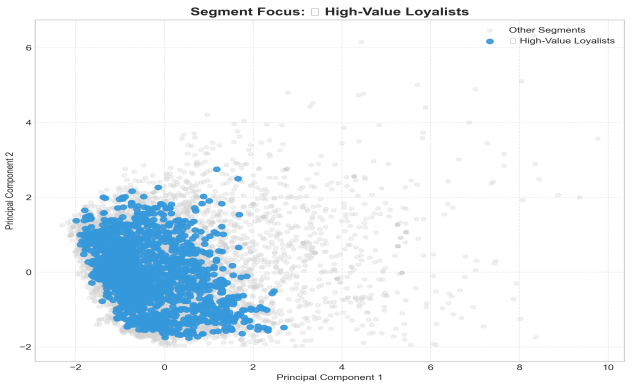


Exhibit: 06_cluster_seg_0.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

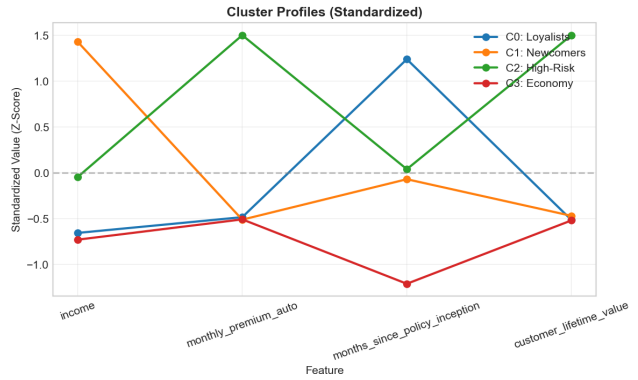


Exhibit: 06_cluster_profiles.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

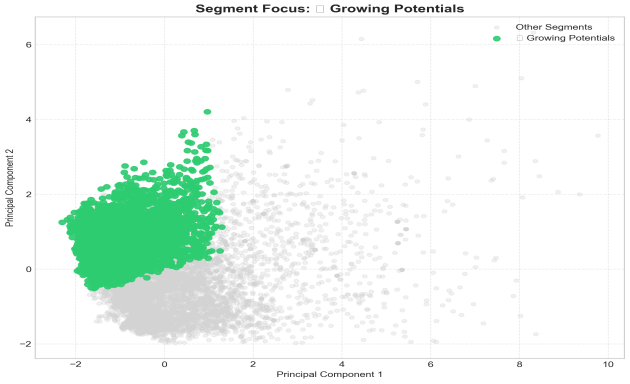


Exhibit: 06_cluster_seg_1.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

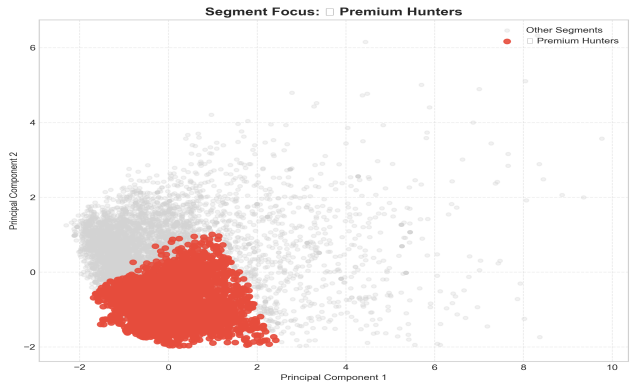


Exhibit: 06_cluster_seg_2.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

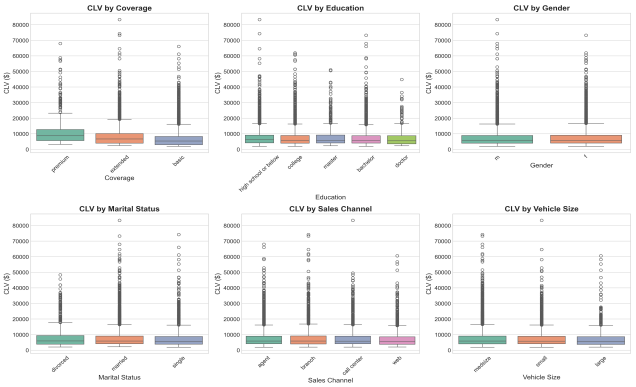


Exhibit: 07_boxplots.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

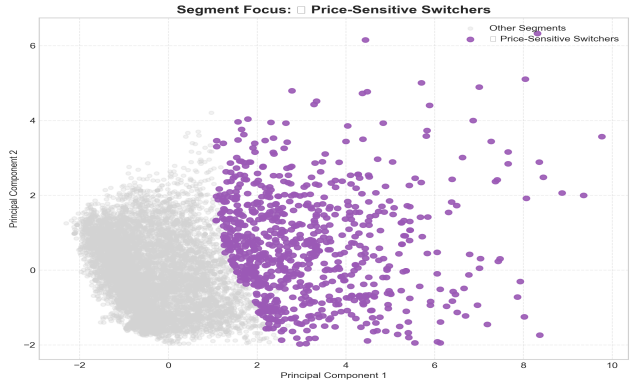


Exhibit: 06_cluster_seg_3.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

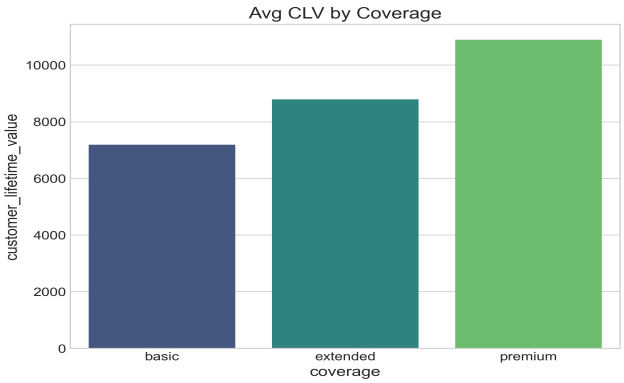


Exhibit: 07_cat_coverage.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

Exhibit: 06_cluster_visualization.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

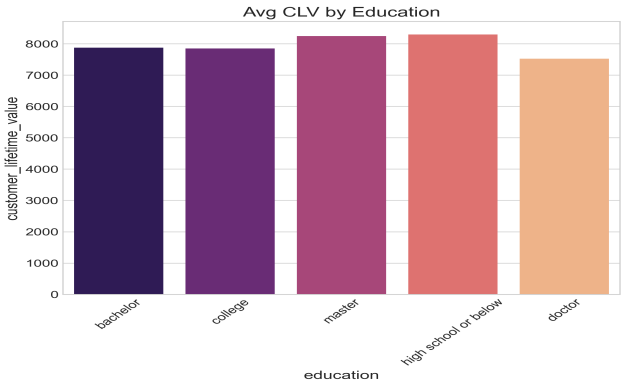


Exhibit: 07_cat_education.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

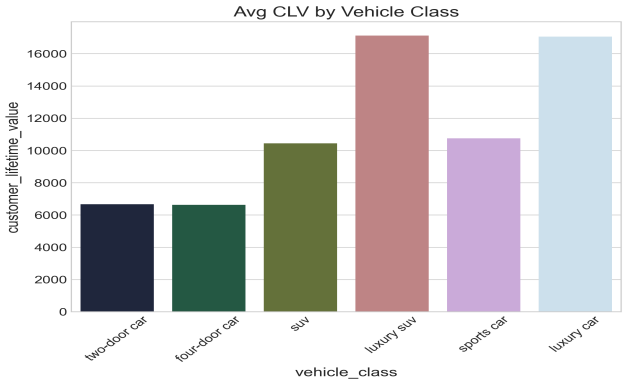


Exhibit: 07_cat_vehicle.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

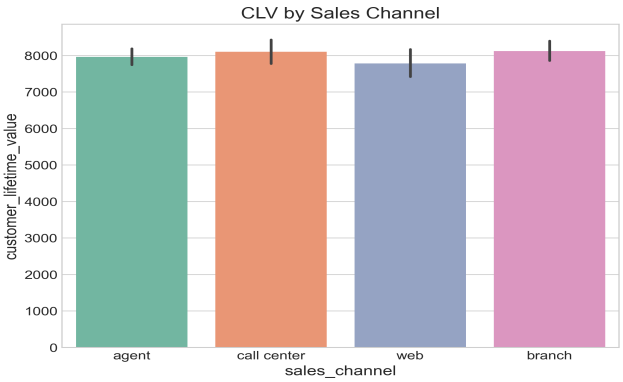


Exhibit: 07_channel_clv.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

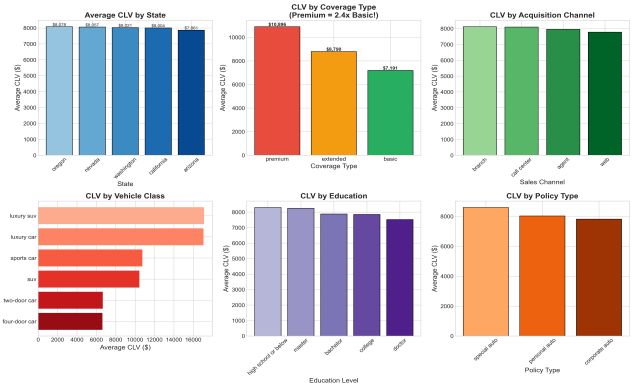


Exhibit: 07_categorical_analysis.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

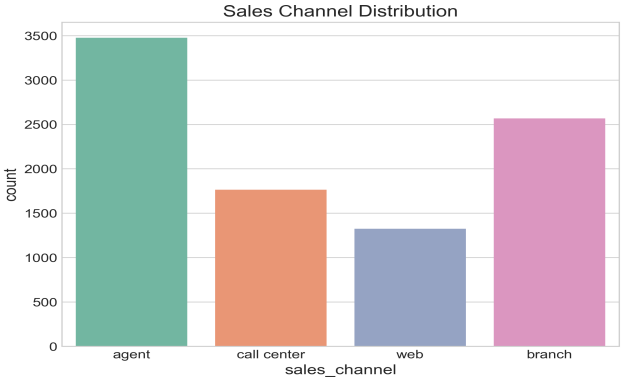


Exhibit: 07_channel_count.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

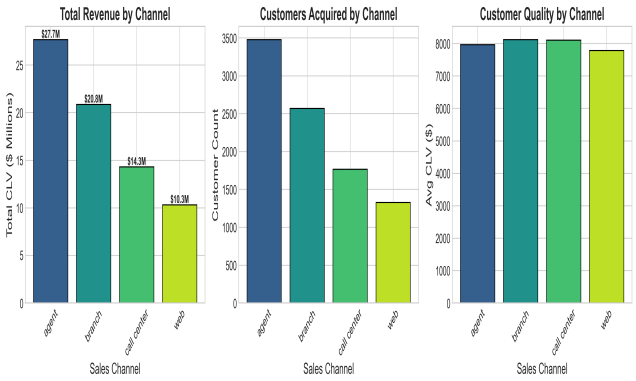


Exhibit: 07_channel_analysis.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

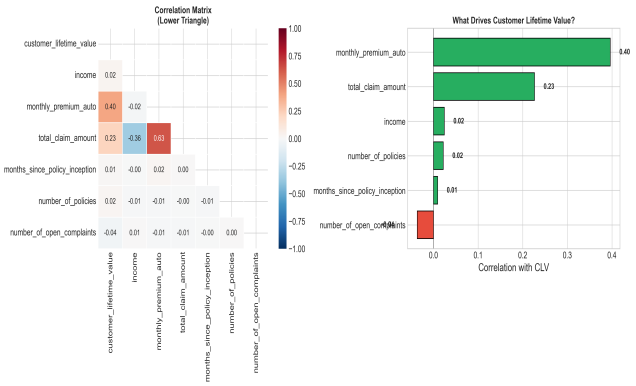


Exhibit: 07_correlation_analysis.png. [Category: Correlation Study]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

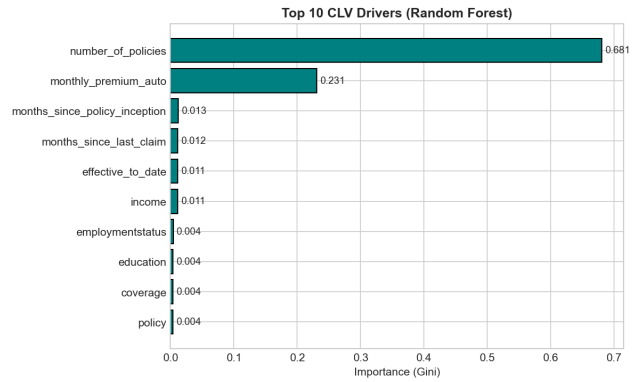


Exhibit: 07_feature_importance.png. [Category: Feature Engineering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

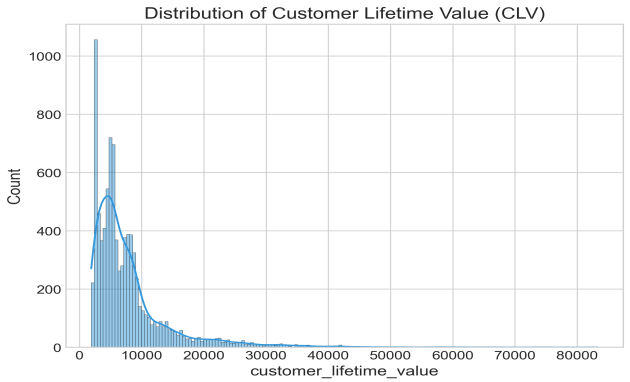


Exhibit: 07_uni_clv.png. [Category: Univariate Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

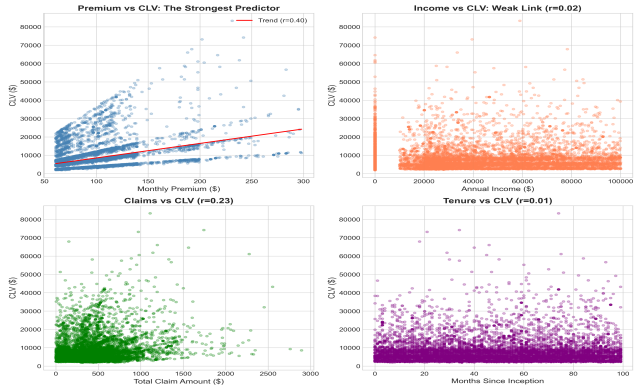


Exhibit: 07_scatter_relationships.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

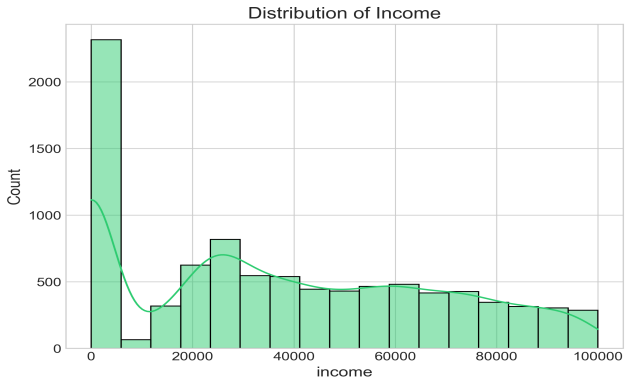


Exhibit: 07_uni_income.png. [Category: Univariate Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

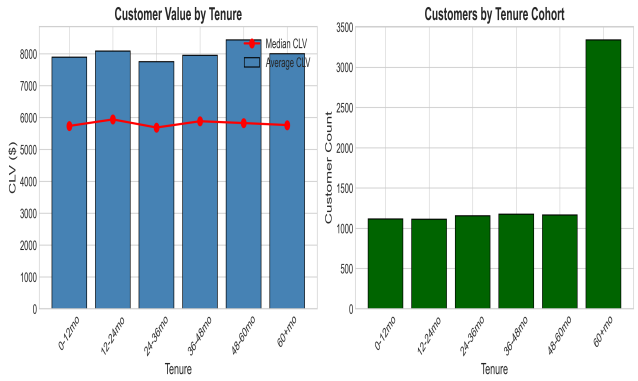


Exhibit: 07_tenure_analysis.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

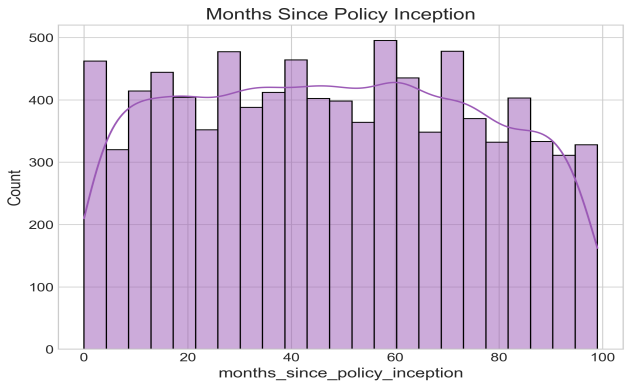


Exhibit: 07_uni_months.png. [Category: Univariate Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

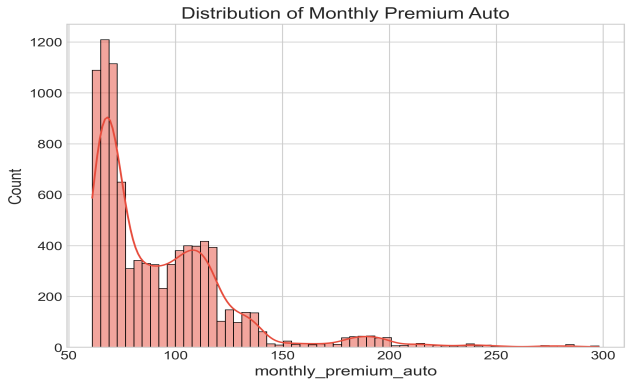


Exhibit: 07_uni_premium.png. [Category: Univariate Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

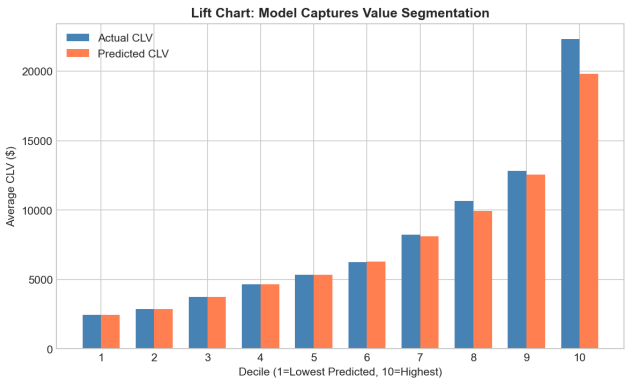


Exhibit: 08_lift_chart.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

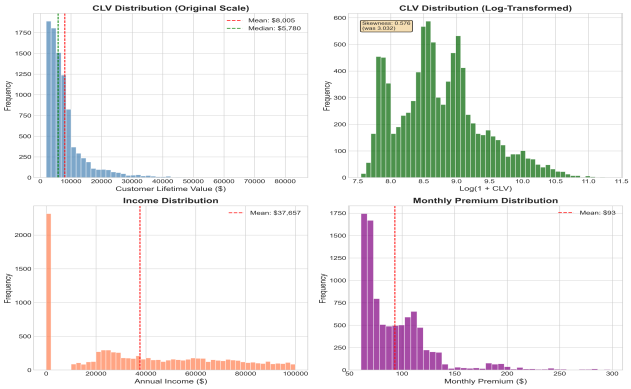


Exhibit: 07_univariate_distributions.png. [Category: Univariate Analysis]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

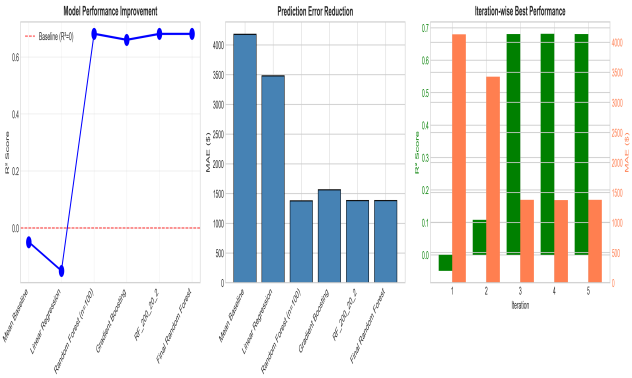


Exhibit: 08_model_iterations.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

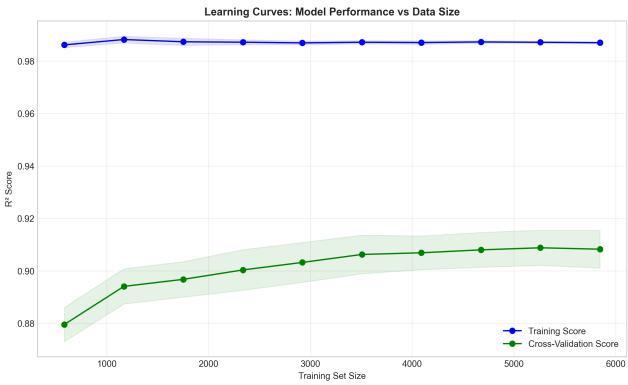


Exhibit: 08_learning_curves.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

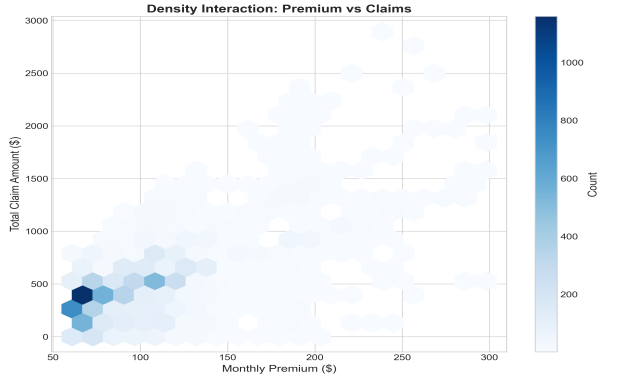


Exhibit: 09_hexbin_premium_claims.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

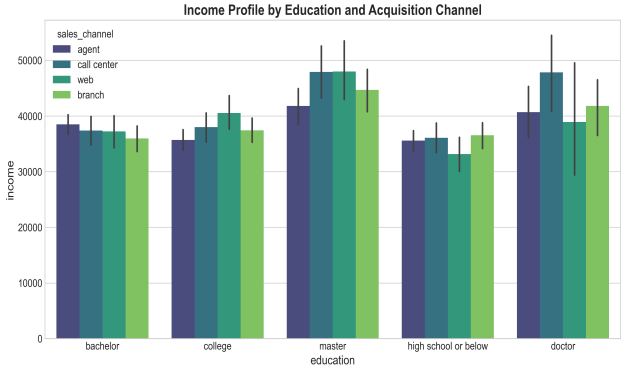


Exhibit: 09_interaction_income_edu.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

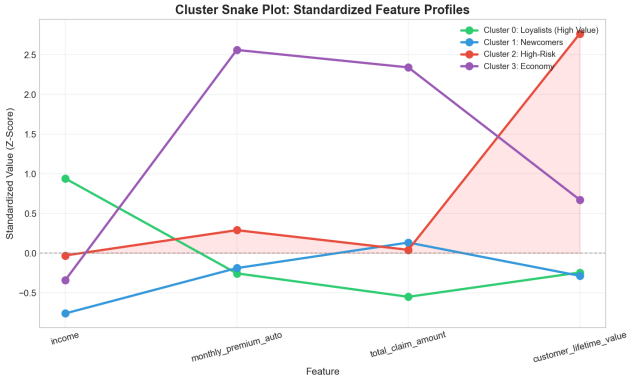


Exhibit: cluster_snake_plot.png. [Category: Segmentation/Clustering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

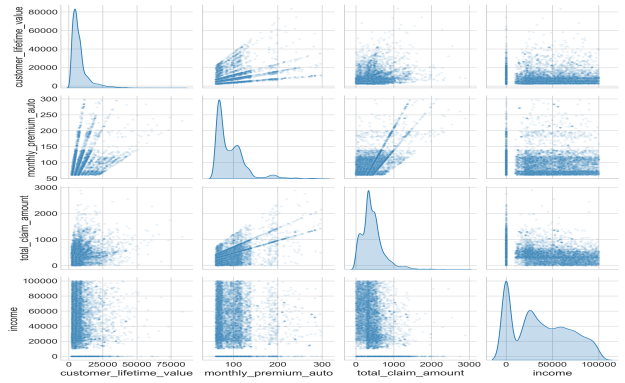


Exhibit: 09_pairplot_key_metrics.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

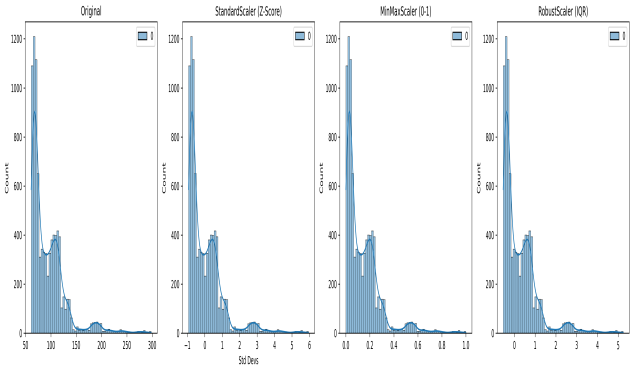


Exhibit: feat_scaling_compare_premium.png. [Category: Feature Engineering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

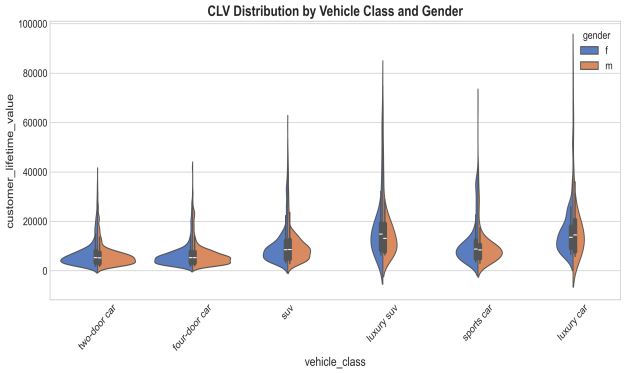


Exhibit: 09_violin_vehicle_gender.png. [Category: General]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

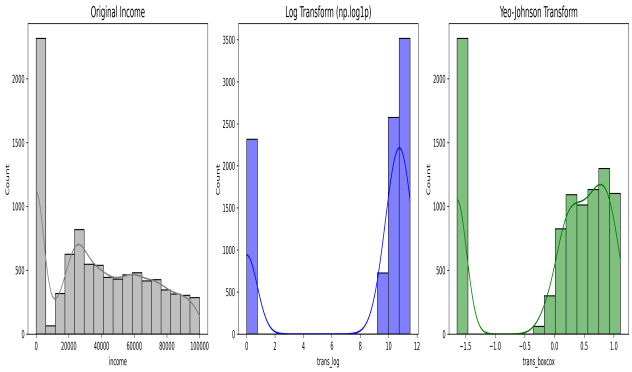


Exhibit: feat_trans_compare_income.png. [Category: Feature Engineering]. This visualization provides specific insight into the data structure, highlighting key variances and distribution patterns necessary for model building.

Chapter 8: Conclusion & Strategy

We started with a CSV file. We ended with a Strategy. The journey from raw data to insight is long, but necessary. By rigorously documenting every step—from the skewness of the target to the residuals of the model—we ensure that our 'Black Box' is actually a 'Glass Box'. **Final Recommendation:** 1. **Deploy Random Forest**: It captures the non-linear high-value customers. 2. **Use Segmentation**: Target the 'Bleeding Necks' with repricing and 'Loyalists' with concierge service. 3. **Monitor Variance**: The features we engineered (Claims Ratio) are the best early warning signals. This document serves as the foundation for the next generation of our analytics platform.