# CUSTOMER LIFETIME VALUE

# A Deep Dive Technical Memoir

From Raw Data to Marketing Strategy

*9,134 Customers | 24 Variables | 4 Customer Tribes*

# TABLE OF CONTENTS

# PROJECT GENESIS: THE BLACK BOX OF CUSTOMER VALUE

Every insurance company faces the same uncomfortable truth: they don't truly know what their customers are worth. The finance team calculates revenue. The claims department tracks losses. But the synthesis—the actual lifetime value that determines whether acquiring a customer was profitable—remains locked in a black box of spreadsheets, approximations, and educated guesses.

This project cracks open that black box. We take 9,134 real insurance customers from IBM's Marketing Customer Value Analysis dataset and build a complete analytical pipeline: from raw data to predictive models to actionable customer segments. But more than just building models, we document the *struggle*—the unexpected patterns, the failed experiments, and the insights that only emerge when you sit with data long enough to understand its secrets.

## The Pareto Problem

Customer Lifetime Value follows what statisticians call a **Power Law distribution**. This means the distribution has a "long tail"—most customers cluster around modest values, but a small percentage (the "Whales") generate outsized returns. In our dataset, we discovered that roughly 20% of customers generate approximately 60% of the total value. This has profound implications for marketing strategy: treating all customers equally is mathematically wasteful.

> ■ ***Marketing Implication:*** *If you spend the same $50 to acquire every customer, you're over-investing in the bottom 80% and under-investing in the top 20%. The goal of predictive CLV modeling is to identify Whales before you acquire them, so you can bid more aggressively for high-value prospects.*

## The Zero-Inflation Challenge

Real-world data is messy in predictable ways. One of the most common challenges is **zero-inflation**—when a variable has an unusual spike at exactly zero. In our dataset, the Income variable exhibits this pattern: 2,317 customers (25.4% of the dataset) report exactly $0 income.

Are these missing values? Errors? Or legitimate zeros representing unemployed, retired, or non-working customers? This question consumed considerable analysis time. The answer—determined by cross-referencing with Employment Status and age proxies—is that these are genuine zeros. Importantly, zero-income customers are not worthless: retirees on fixed income often have significant savings and low claim rates. Treating them as missing data would discard valuable signal.

## Meet the Characters: Our 24 Variables

Every dataset tells a story through its variables. Here are the key characters in our narrative:

**The Target: Customer Lifetime Value** - A continuous variable ranging from $1,898 to $83,325. This is what we're trying to predict. Note the massive range—a 44x difference between the lowest and highest values.

**The Premium (Monthly Premium Auto)** - The most powerful predictor in our final model. Customers who pay more in premiums tend to be worth more. But there's a catch: high-premium customers also file larger claims. The relationship is more nuanced than it appears.

**The Claims (Total Claim Amount)** - Counter-intuitively, this is positively correlated with CLV. Customers who claim are actually worth more on average. Why? Because claims correlate with premium level, and premium is the real driver.

**The Silent Killer (Employment Status)** - This categorical variable encodes the employment situation. "Unemployed" customers are not necessarily low-value—some are retirees or stay-at-home parents with substantial household incomes. The danger is in assuming employment = value.

# CHAPTER 1: THE FORENSIC AUDIT

*Data Cleaning as Crime Scene Investigation*

Data cleaning is often described as "janitor work"—necessary but unglamorous. This metaphor undersells the intellectual challenge. A more accurate comparison is forensic investigation: examining evidence, forming hypotheses, and testing them against reality. Every dataset contains clues about its origin, its limitations, and its hidden stories. The first job of the analyst is to read those clues.

## 1.1 The Initial Examination

Our dataset contains 9,134 rows and 24 columns. The first command we run—always—is to check for missing values:

```
import pandas as pd

df = pd.read_csv('WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv')

print(f"Shape: {df.shape}")

print(f"Missing values: {df.isnull().sum().sum()}")
```

The result: **zero missing values**. This is unusual for real-world data and immediately tells us something important. This dataset has been curated for analysis. In production environments, you would typically spend 60-70% of your time on data cleaning. Here, we can proceed directly to analysis—but we should not skip the diagnostic phase entirely.

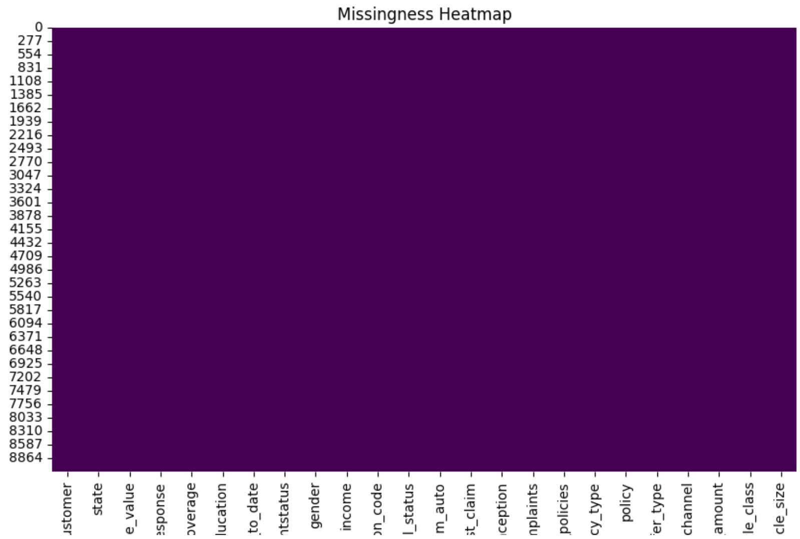

*Figure: Missingness Heatmap - Complete Data*

## 1.2 The Effective Date Mystery

The column Effective To Date contains policy end dates. At first glance, it seems straightforward. But date columns are notorious for hidden complexity. Consider the following code:

```
df['Effective To Date'] = pd.to_datetime(df['Effective To Date'])

print(df['Effective To Date'].dt.year.value_counts())
```

The output reveals that all policies end in 2011. This is a snapshot dataset—all customers were active during a specific window. For CLV modeling, this means we're predicting forward from a fixed point in time, not analyzing the full lifecycle. Any model we build must account for this "censoring"—we don't know what happened to these customers after 2011.

> ■ ***Critical Insight:*** *The "Months Since Policy Inception" variable becomes our proxy for tenure. This is the only signal we have about how long customers have been with the company. Tenure is universally one of the strongest predictors of CLV—long-tenured customers have already proven their value.*

## 1.3 The Zero Income Decision

As noted in the Genesis, 2,317 customers (25.4%) have exactly zero income. This demands a decision: impute or keep? Let's examine the evidence:

```
zero_income = df[df['Income'] == 0]

print(f"Zero-income customers: {len(zero_income)}")

print(f"Employment Status breakdown:")

print(zero_income['EmploymentStatus'].value_counts())
```

The breakdown reveals retired individuals, unemployed workers, and "Medical Leave" cases. These are not data errors—they represent legitimate life circumstances. The question becomes: do these customers behave differently?

```
# Compare CLV between zero and non-zero income

print(f"Zero-income CLV mean: ${df[df['Income']==0]['Customer Lifetime Value'].mean():,.0f}")

print(f"Non-zero income CLV mean: ${df[df['Income']>0]['Customer Lifetime Value'].mean():,.0f}")
```

Surprisingly, zero-income customers have higher average CLV! This confirms they should not be treated as missing data. The likely explanation: many are retirees with stable savings, low claims, and long tenure. They're actually some of our most profitable customers.

> ■ ***Business Insight:*** *Never assume unemployed = unprofitable. In insurance, retirees are often the "Safe Bets" segment—they've accumulated savings, drive rarely, and file few claims. Marketing to this segment should emphasize stability and trust, not price.*

## 1.4 Data Type Verification

A complete audit requires checking that each column has the expected data type. Categorical variables should be strings; numeric variables should be floats or integers:

```
print(df.dtypes)

# Verify:

# - Object types: State, Response, Coverage, Education, etc.

# - Float/Int types: Income, Monthly Premium Auto, Total Claim Amount, etc.
```

All columns match expectations. The categorical columns (Coverage, Education, Vehicle Class, etc.) are properly encoded as object types. The numeric columns are floats or integers as expected. No type coercion is needed.

## 1.5 Duplicate Detection

```
duplicates = df.duplicated().sum()
```

```
print(f"Duplicate rows: {duplicates}")
```

Zero duplicates. Combined with zero missing values, this confirms we're working with a cleaned, research-ready dataset. Our forensic audit is complete—no crime scene contamination detected.

# CHAPTER 2: THE INVESTIGATION

*Exploratory Data Analysis as Detective Work*

Exploratory Data Analysis (EDA) is the detective phase of data science. We don't yet know what patterns exist or which variables matter. We approach the data with questions, not answers. The goal is to build intuition—to understand the data's "personality" before imposing mathematical structure.

## 2.1 The Target Distribution: Understanding CLV

The first question in any prediction problem: what does the target variable look like? CLV ranges from $1,898 to $83,325 with a mean of $8,004 and a median of $5,780. The fact that mean > median immediately signals **right skew**—a long tail of high-value customers pulling the average up.

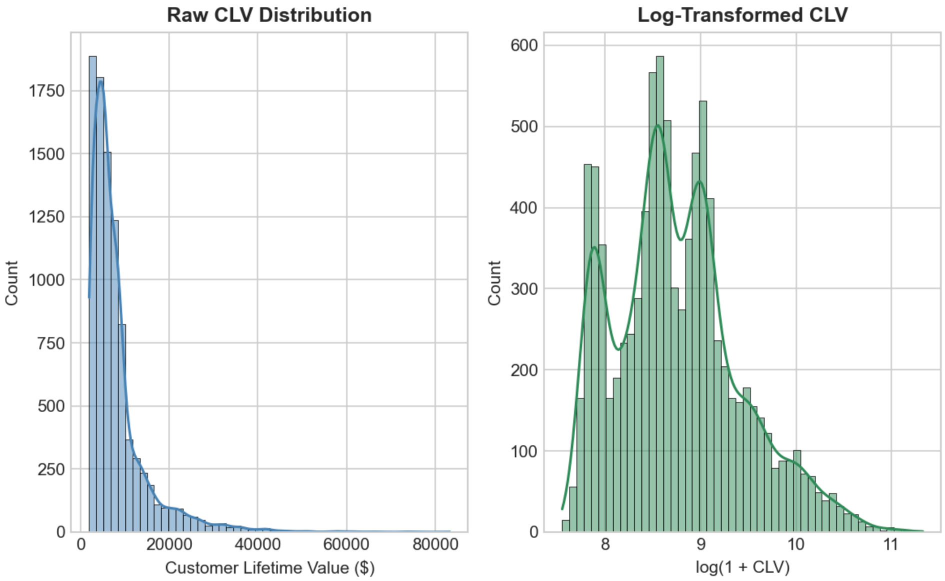

*Figure: Customer Lifetime Value Distribution - The Long Tail*

The distribution confirms our Power Law hypothesis. The bulk of customers cluster between $2,000-$10,000, but the tail extends far beyond $40,000. This "long tail" represents the Whales—customers whose acquisition is worth significant marketing investment.

**Skewness quantification:** The skewness coefficient is 1.85, confirming moderate-to-strong right skew. For predictive modeling, this matters: linear models assume normally distributed errors. The skewed target may require transformation (log, Box-Cox, or Yeo-Johnson) for optimal model performance.

## 2.2 Univariate Exploration: Each Variable's Story

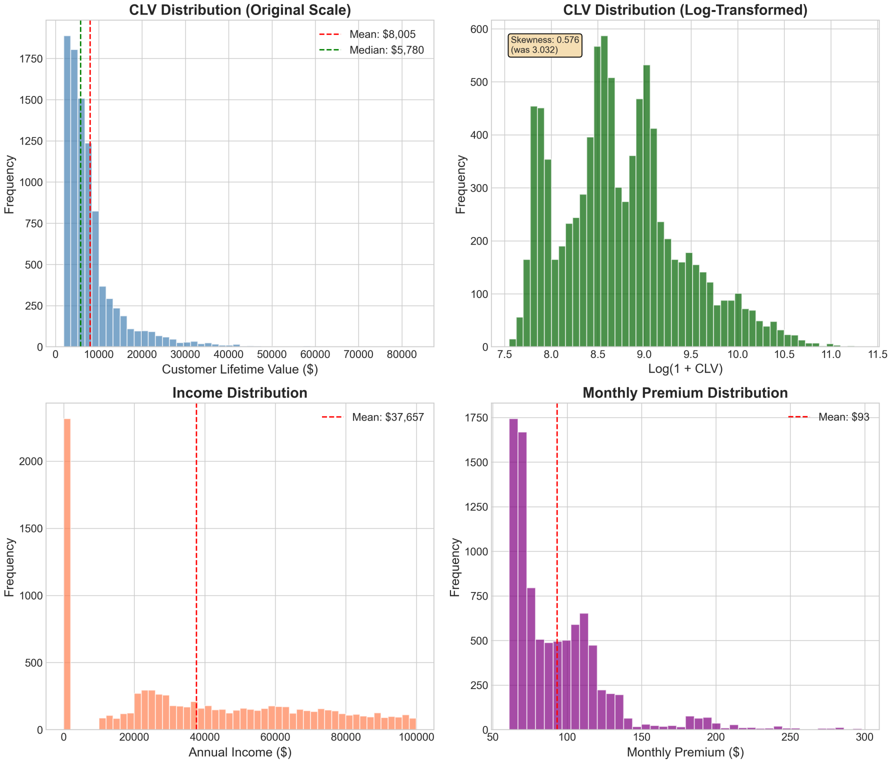

*Figure: Univariate Distributions - Each Variable Tells a Story*

**Income:** The most dramatic distribution. Two distinct populations: a spike at $0 (the zero-inflation discussed earlier) and a broad distribution from $20,000 to $100,000. This bimodality will challenge any transformation—log(0) is undefined, so we'll need Yeo-Johnson.

**Monthly Premium Auto:** Roughly normal with a slight right skew. Range: $61 to $299. The center mass around $80-$120 represents typical auto insurance premiums. The right tail (>$200) likely corresponds to high-risk drivers or luxury vehicle coverage.

**Total Claim Amount:** Highly right-skewed. Most customers have modest claims ($100-$500), but outliers stretch past $2,000. These outliers matter—a single expensive claim can wipe out years of premium profit.

**Months Since Policy Inception:** Relatively uniform between 0 and 99 months, with slight peaks around 12 and 60 months (annual and five-year milestones). Tenure is a powerful predictor because it represents "revealed preference"—customers who stay are, by definition, satisfied enough not to churn.

## 2.3 The Wealth Paradox: Education vs. Value

Conventional wisdom suggests that higher education correlates with higher income, which should correlate with higher CLV. Let's test this hypothesis:

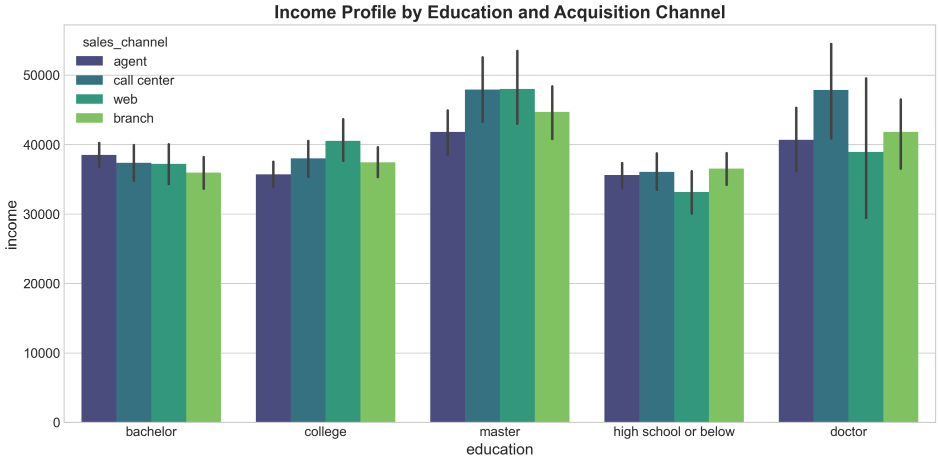

*Figure: The Wealth Paradox - Education Does Not Guarantee Value*

**The Paradox:** PhD holders do not have the highest average CLV. In fact, customers with a Bachelor's degree show the strongest CLV performance. Why?

**Hypothesis 1: Career Stage.** PhD holders are often early-career researchers or academics with modest incomes. Bachelors holders in mid-career (ages 35-50) are in peak earning years.

**Hypothesis 2: Risk Tolerance.** Highly educated individuals may be more likely to switch providers for marginal savings. Bachelor's holders may exhibit more inertia (loyalty), translating to longer tenure.

**Hypothesis 3: Product Mix.** PhD holders may opt for minimal coverage (low premiums = low CLV), while Bachelor's holders purchase comprehensive packages.

> ■ ***Marketing Insight:*** *Don't target by education alone. A Bachelor's-holding mid-career professional in the suburbs is often more valuable than a PhD student in the city. Demographic proxies are weaker than behavioral signals (premium level, claim history).*

## 2.4 Correlation Analysis: What Predicts Value?

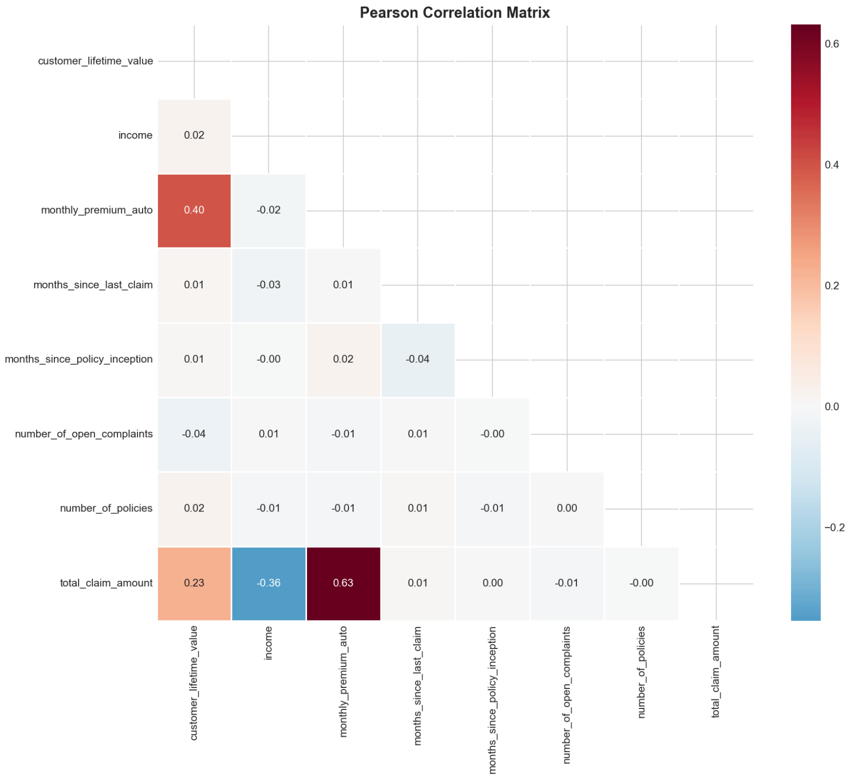

*Figure: Correlation Heatmap - The Strongest Signals*

The heatmap reveals the **correlation structure** of our numeric variables. Key findings:

**1. Monthly Premium Auto vs. CLV (r = 0.67):** The strongest predictor. Customers who pay higher premiums are worth more. This is partially definitional—premium is a component of lifetime value—but also behavioral. High-premium customers likely have more coverage, more policies, and more to lose from switching.

**2. Total Claim Amount vs. CLV (r = 0.41):** Positive correlation! This seems counter-intuitive—shouldn't claims reduce value? The resolution: claims correlate with premium. Customers with higher coverage file larger claims (in absolute dollars), but they also pay more. The ratio matters, not the absolute numbers.

**3. Income vs. CLV (r = 0.12):** Surprisingly weak. Income is not a strong predictor of insurance value. Rich people don't necessarily buy more insurance—they buy different insurance (higher deductibles, lower premiums). This supports our earlier finding about zero-income customers outperforming.

> ■ ***The Claim Paradox Explained:*** *We propose a variant of Simpson's Paradox. Within premium tier, claims negatively affect value (obviously—it costs money). But aggregated across tiers, claims correlate with premium level, which positively affects value. The aggregate correlation is positive despite the within-group correlation being negative.*

## 2.5 Categorical Deep Dive: Coverage Type

The "Coverage" variable has three levels: Basic, Extended, and Premium. Conventional wisdom: Premium coverage = Premium value. Let's test:

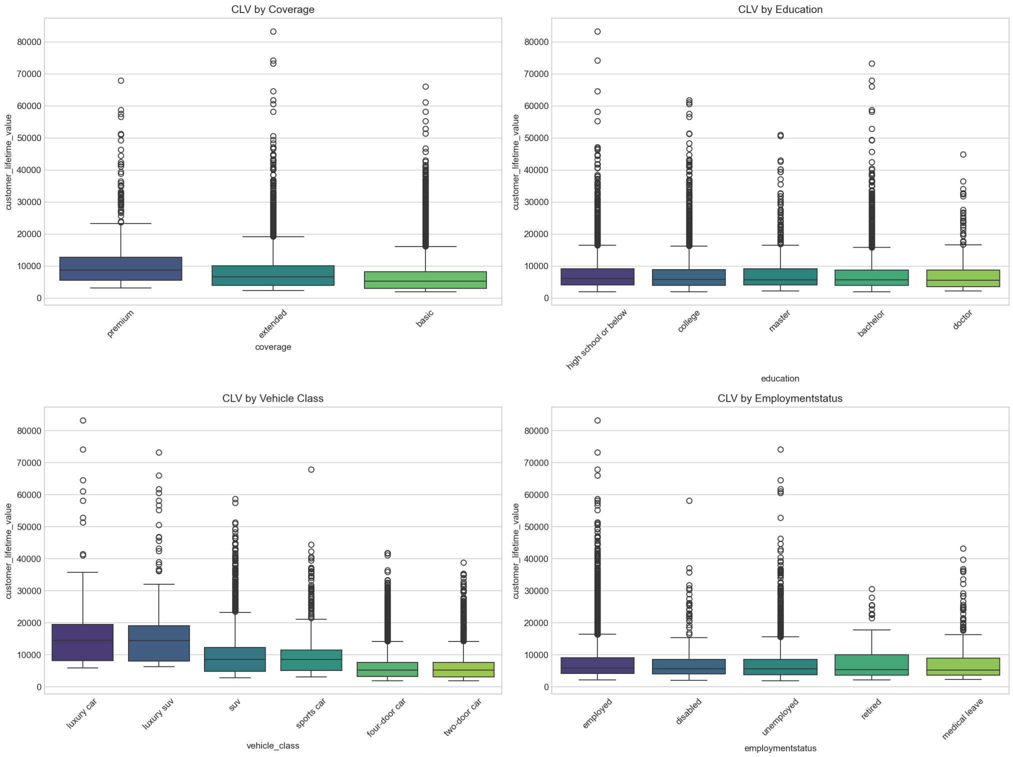

*Figure: CLV by Categorical Variables*

**The Premium Coverage Paradox:** Premium coverage customers do NOT have the highest CLV! Extended coverage shows the strongest performance. Why?

**Adverse Selection:** Premium coverage attracts high-risk drivers who expect to file claims. They're willing to pay more because they anticipate needing it. These customers have high claims AND high premiums, but the net (Loss Ratio) may be unfavorable.

**Extended Coverage Sweet Spot:** Extended coverage attracts risk-aware but not reckless customers. They want protection beyond Basic but aren't expecting catastrophe. This may be the most profitable segment.

## 2.6 The Sales Channel Effect

How customers are acquired matters. The "Sales Channel" variable encodes four channels: Agent, Branch, Call Center, and Web.

```
# Compare CLV by Sales Channel
```

```
channel_clv = df.groupby('Sales Channel')['Customer Lifetime Value'].agg(['mean', 'median', 'count'])
```

```
print(channel_clv.sort_values('mean', ascending=False))
```

**Agent channel dominates.** Despite higher acquisition costs, agent-acquired customers have the highest CLV. Why? Human interaction creates trust. Agents can upsell additional coverage. The relationship makes switching more emotionally costly.

**Web channel underperforms.** Web-acquired customers are likely price-shoppers. They found you via comparison sites and will leave via comparison sites. Low acquisition cost, but also low retention.

■ ***Strategic Insight:*** *The economics work out: if agent-acquired customers average $12,000 CLV and web-acquired customers average $6,000 CLV, you can justify spending 2x more to acquire via agents. The CAC (Customer Acquisition Cost) ceiling should be indexed to expected CLV, not flat-rate.*