

# CLV

Customer Lifetime Value

# Prediction using Machine Learning Algorithm

Athaya Zahrani Irmansyah

JCDS 0408 On Site Bandung

Capstone Project Module 3



# Table of Contents



1	<i>Customer Lifetime Value (CLV)</i>
2	<i>Business Problem and Goals</i>
3	<i>Analytic Approach and Evaluation Metrics</i>
4	<i>Data Preprocessing for Machine Learning (ML) Algorithm</i>
5	<i>Conclusion and Recommendation</i>

# *Customer Lifetime Value*



- Metrik yang sangat penting dalam **manajemen bisnis modern**
- **Pembuat keputusan strategis** dalam hal pemasaran, retensi, dan pengembangan produk
- **Jumlah total uang yang dihabiskan oleh pelanggan** untuk perusahaan selama hubungan bisnis antara pelanggan dan perusahaan tersebut berjalan
- Mempertahankan *customer* baru bisa sampai **25 kali lebih mahal** daripada mempertahankan *customer* lama

# Kenapa CLV itu penting?

- **Meningkatkan pendapatan** perusahaan
- **Membantu penganggaran**
- **Menganalisis kepuasan** pelanggan



# **Business Problem**

- Perusahaan asuransi mobil di Amerika Serikat **menghadapi masalah dalam meningkatkan pendapatan.**
- **Strategi pemasaran yang tidak tepat**, di mana perusahaan mengalokasikan anggaran yang sama untuk semua tipe pelanggan.
- Perusahaan **menghabiskan lebih banyak anggaran untuk pelanggan bernilai rendah** dan berisiko kehilangan pelanggan bernilai tinggi.



# Goals

**Membuat ‘alat’ untuk memprediksi CLV** dengan melihat dari data demografis dan data asuransi mobil customer, sehingga pengolahan data CLV **tidak lagi secara manual** dan dapat **mempercepat proses pengambilan keputusan** strategi marketing.

Machine Learning  
Algorithm



## MODEL REGRESI



# Data Preprocessing



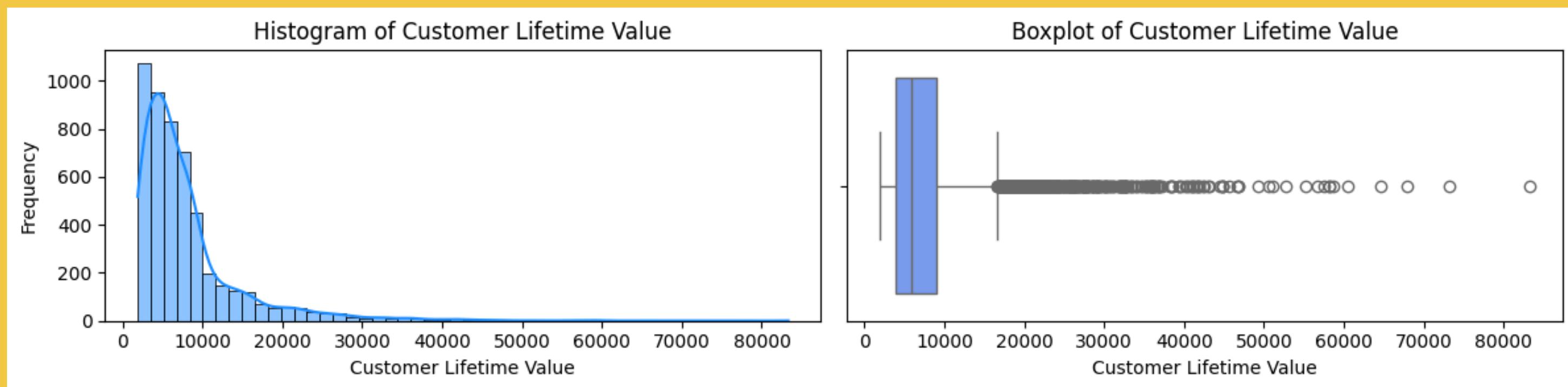
# Data Understanding



Attribute	Data Type	Description
Vehicle Class	Object	Vehicle type classification
Coverage	Object	Types of vehicle insurance coverage
Renew Type Offer	Object	Offer to renew policies that have been/will expire
EmploymentStatus	Object	Customer's employment status
Marital Status	Object	Customer's marital status
Education	Object	Customer's educational level
Number of Policies	Float	Number of policies owned by the customer
Monthly Premium Auto	Float	Monthly premium paid by the insured
Total Claim Amount	Float	Cumulative number of claims since the beginning of the policy
Income	Float	Customer's income (in dollar)
Customer Lifetime Value	Float	Customer Lifetime Value (Target)

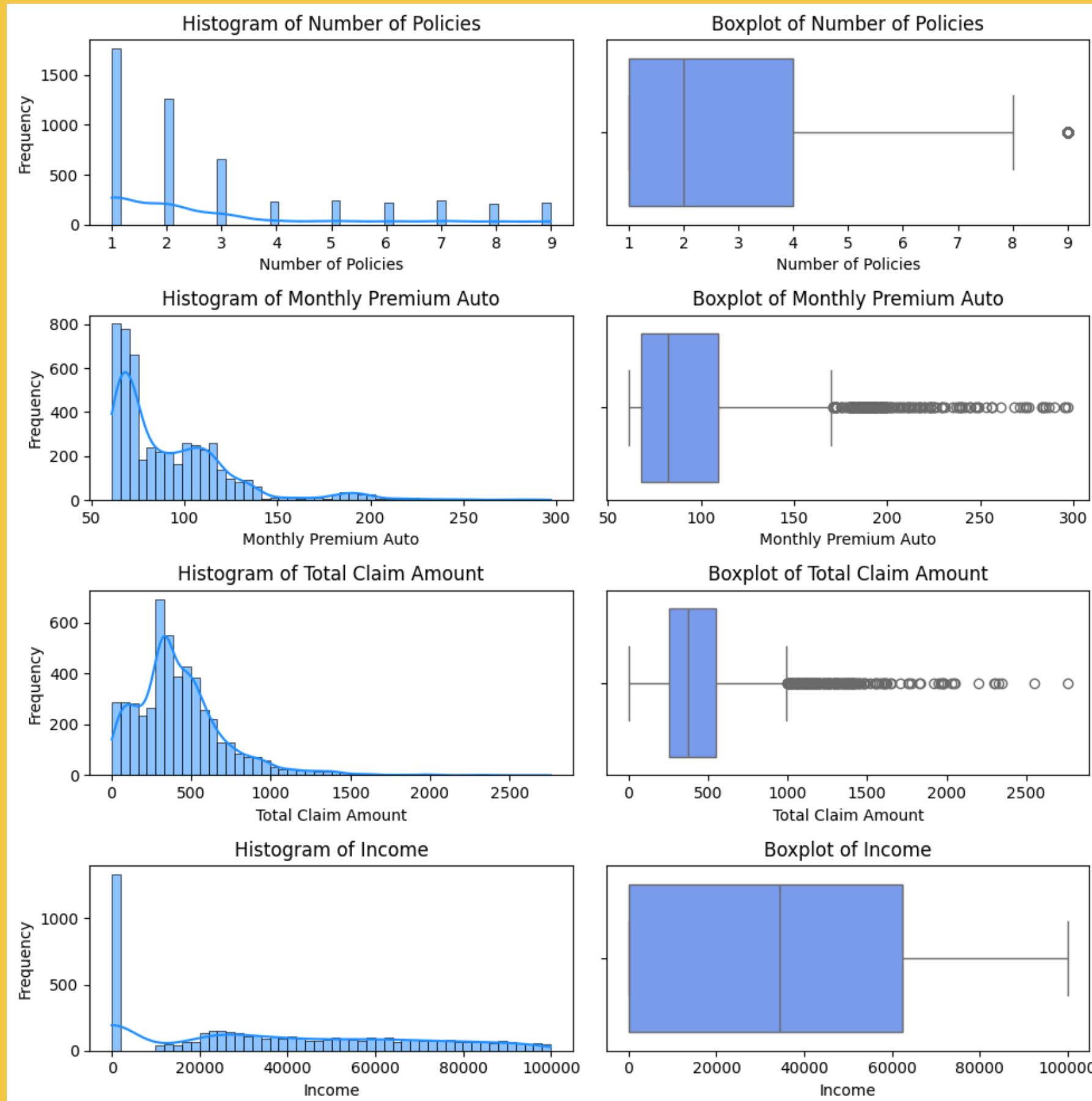
# Data Findings

## Customer Lifetime Value



- Banyak customer dengan ***low CLV***
- Terdapat customer dengan **CLV yang ekstrim** (di atas mediannya 16000)

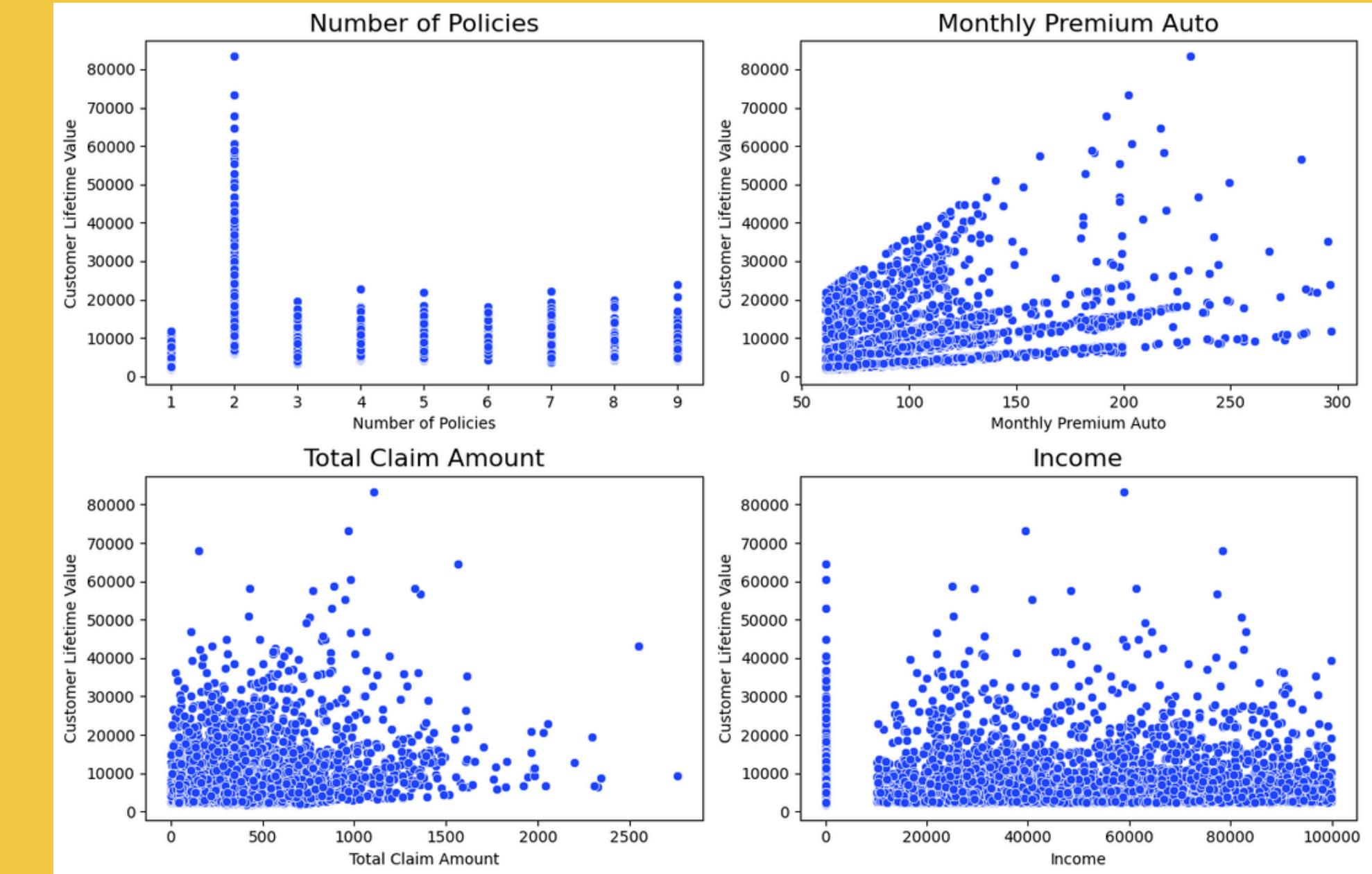
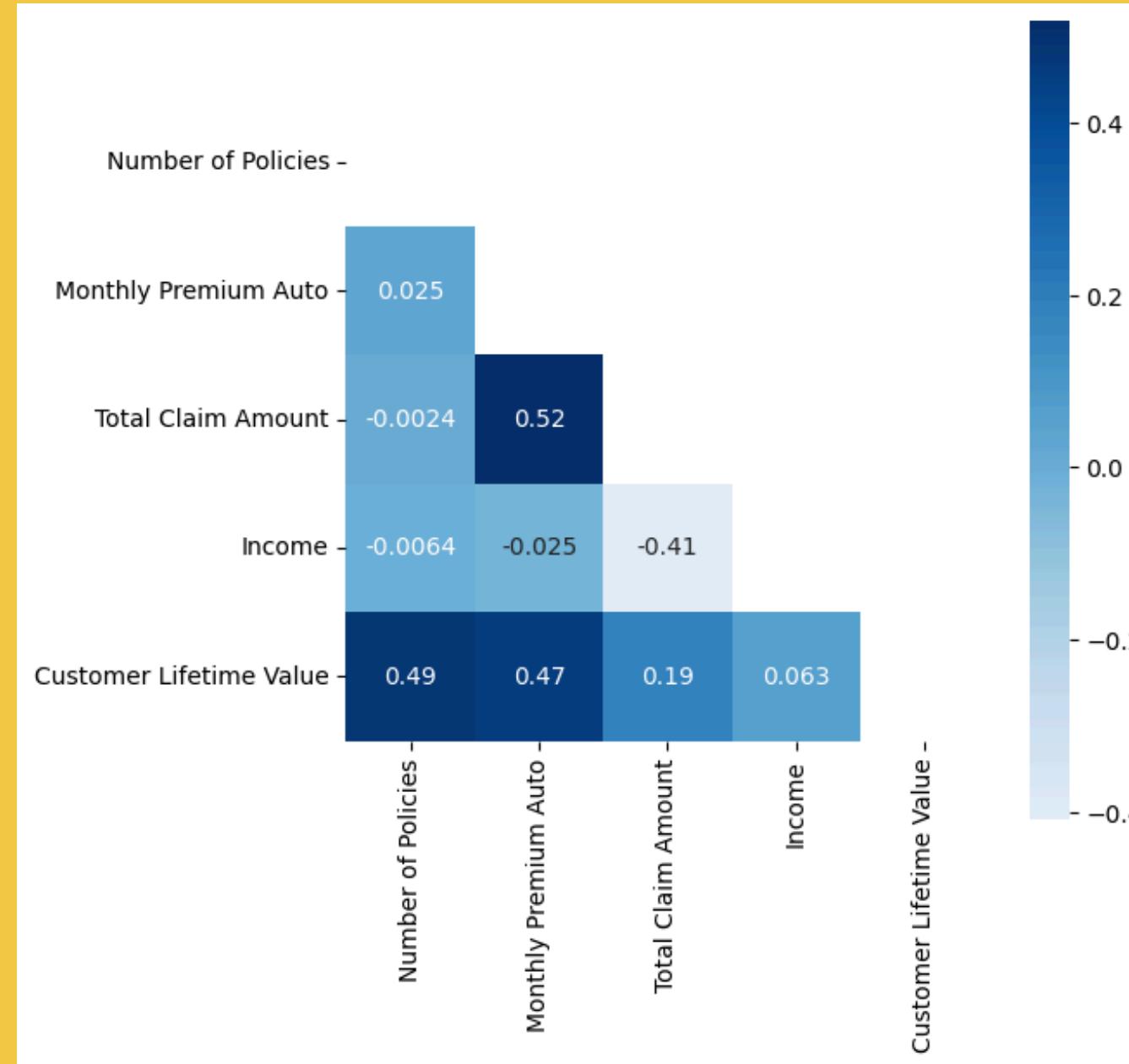
## Kolom Features Numerikal



Terdapat **outliers** pada kolom  
**Number of Policies, Monthly  
Premium Auto, dan Total  
Claim Amount**

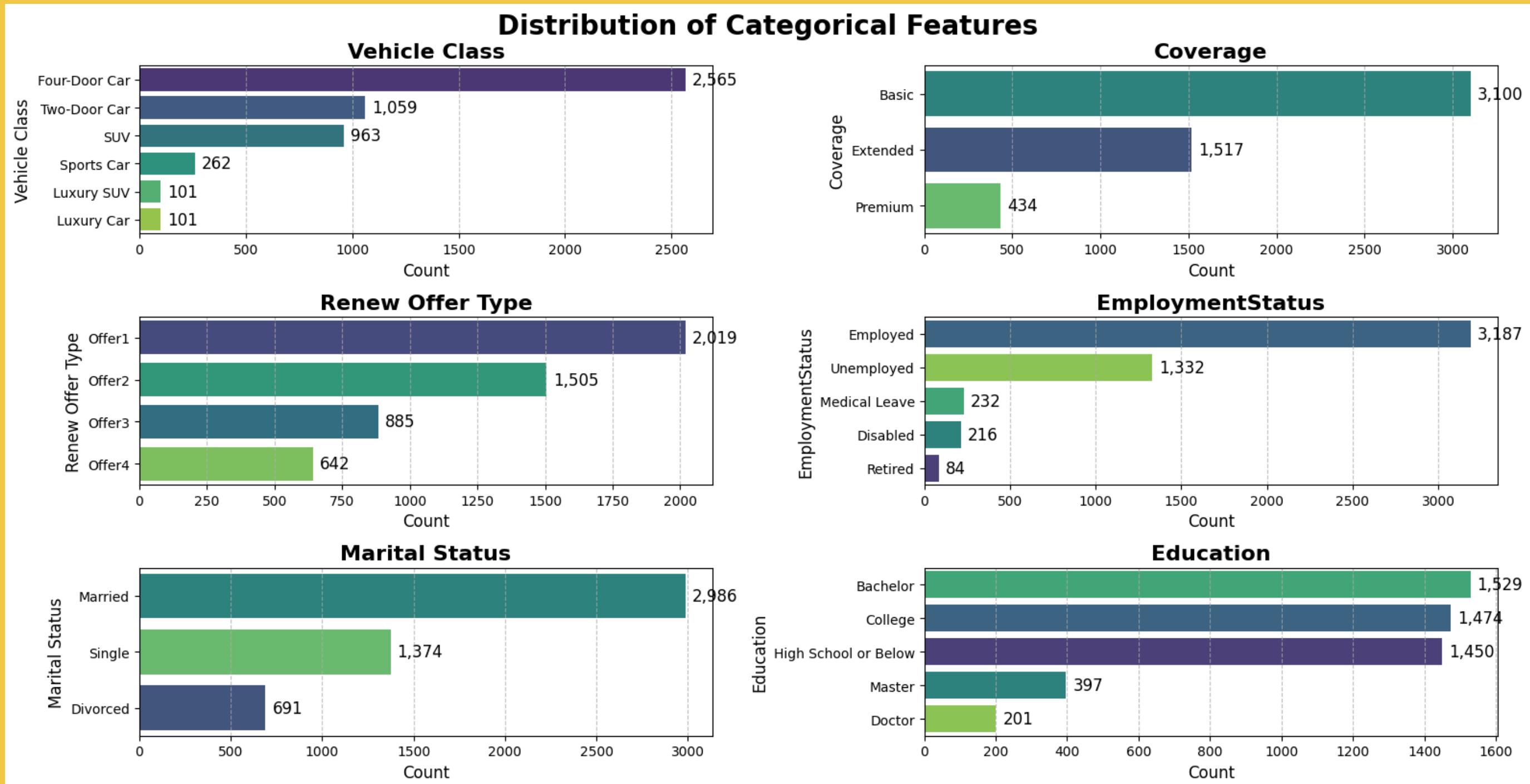


# Korelasi



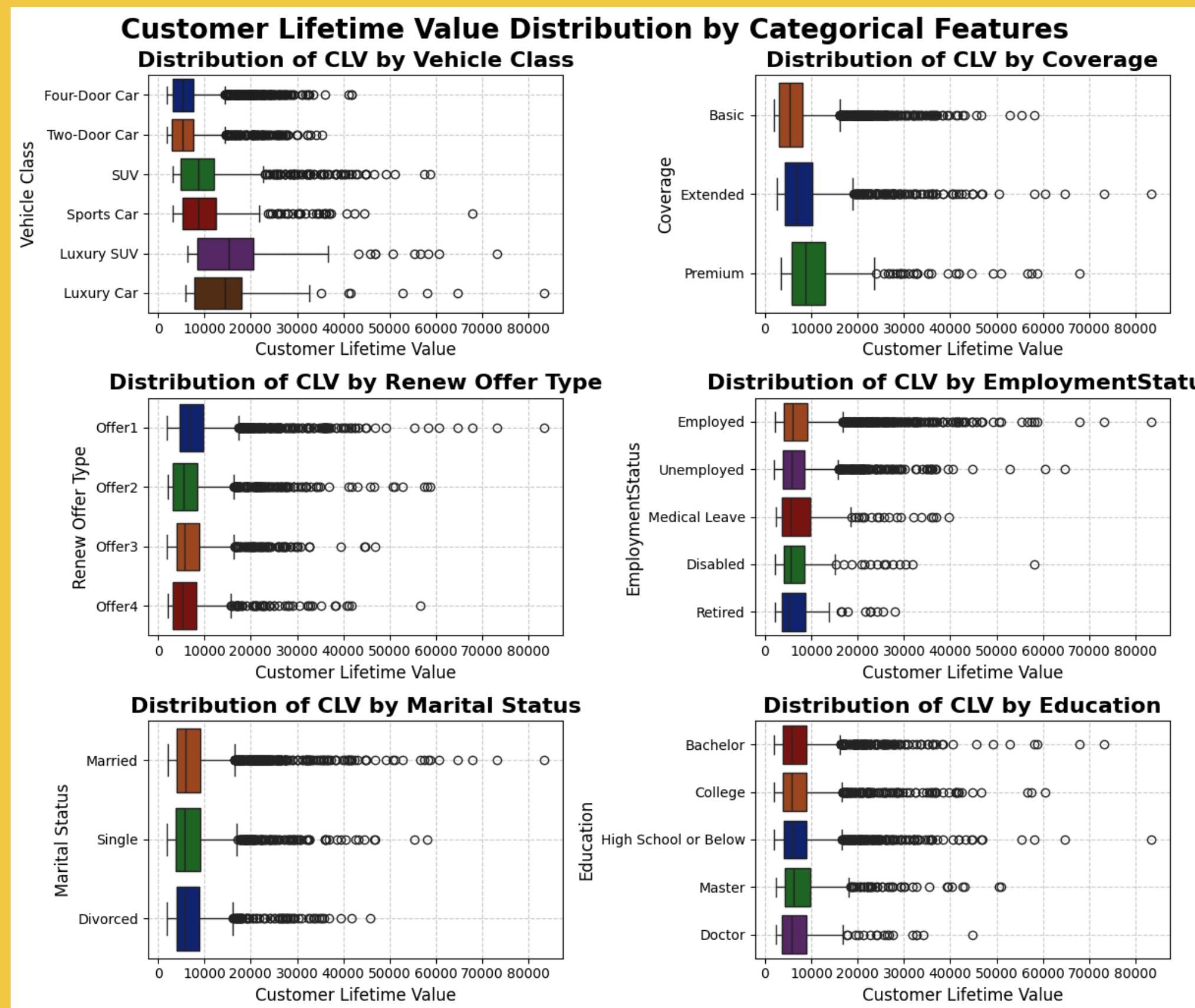
- Kenaikan CLV **sejalan** dengan kenaikan Number of Policies, Monthly Premium Auto, dan Total Claim Amount

# Kolom Features Kategorikal



Four-Door Car, Coverage Basic, Offer 1, bekerja, sudah menikah, sarjana menjadi customer terbanyak

# Kolom Features Kategorikal

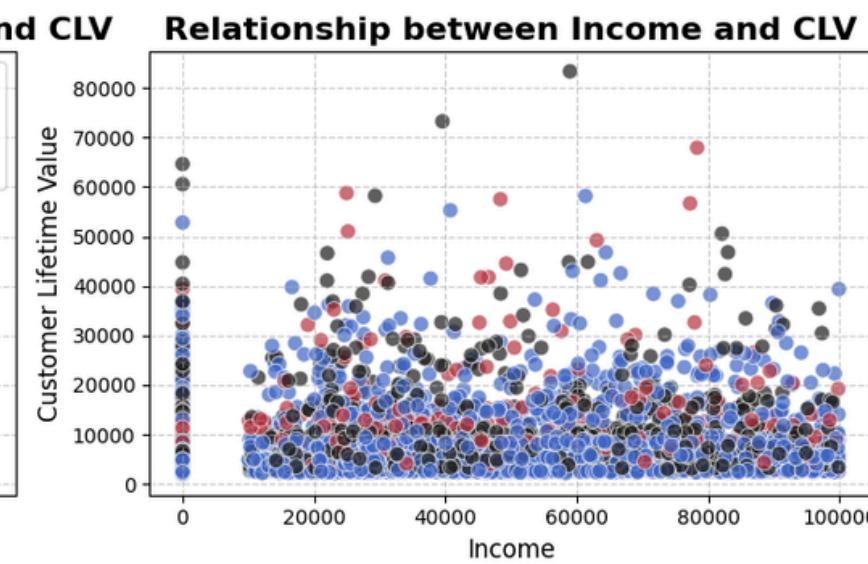
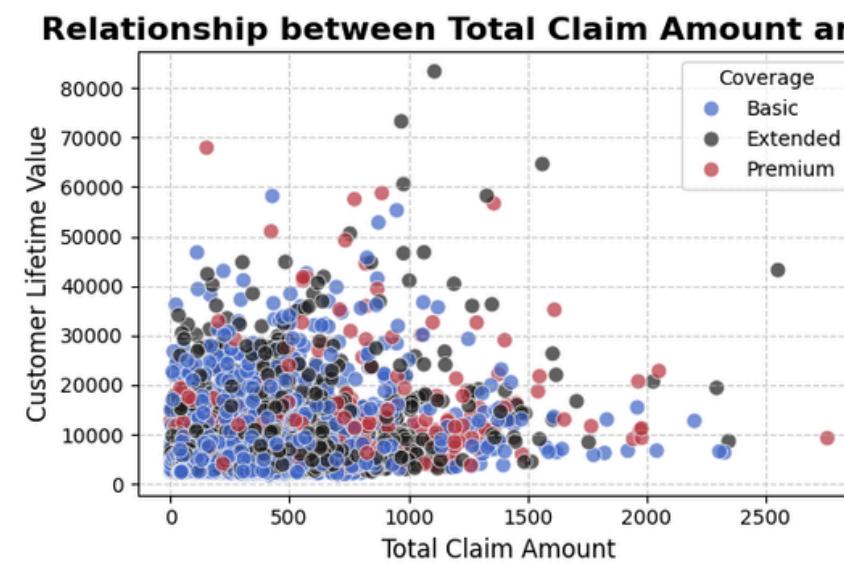
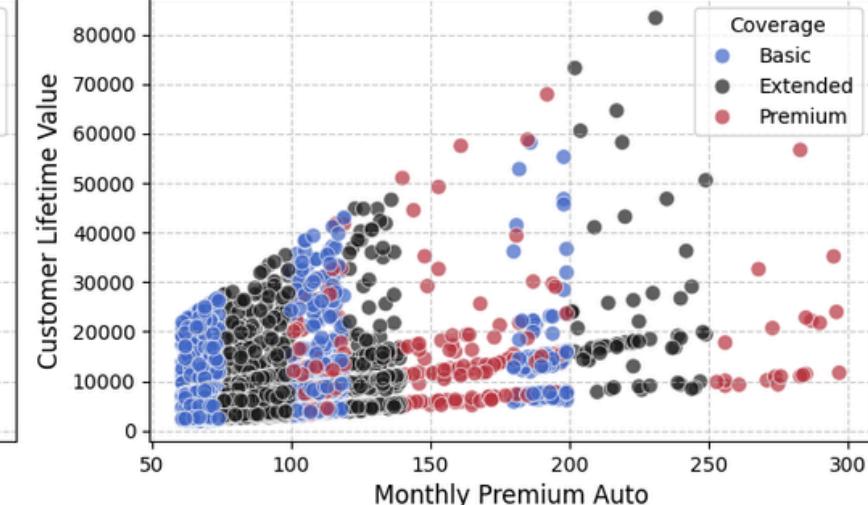
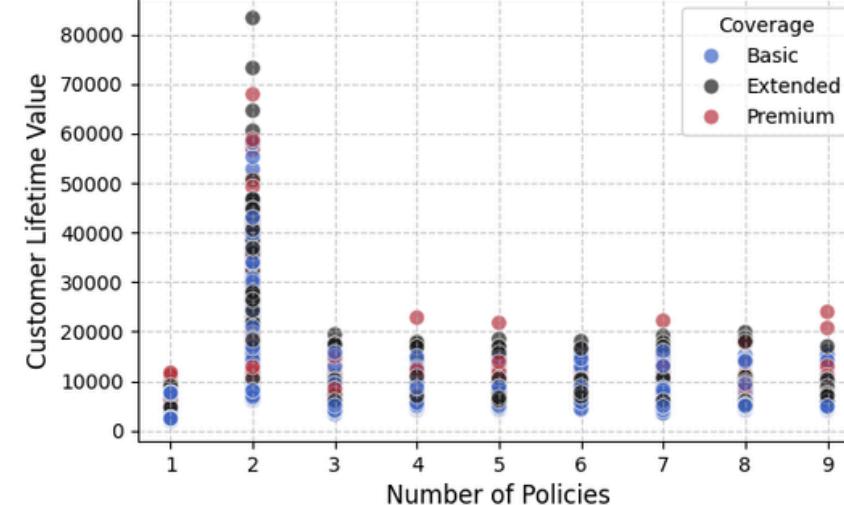


Kendaraan tipe luxury (*Luxury SUV/Luxury Car*), *Coverage Premium*, *Offer 1*, sudah bekerja, menikah, dan sarjana menjadi customer dengan *high CLV*



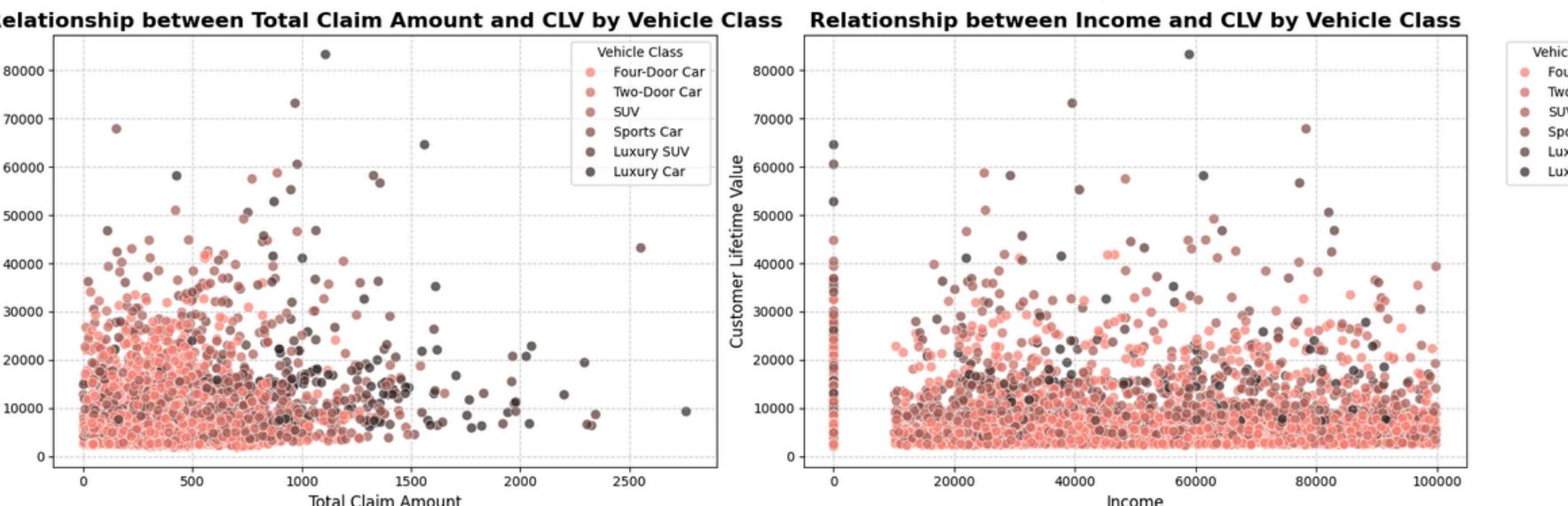
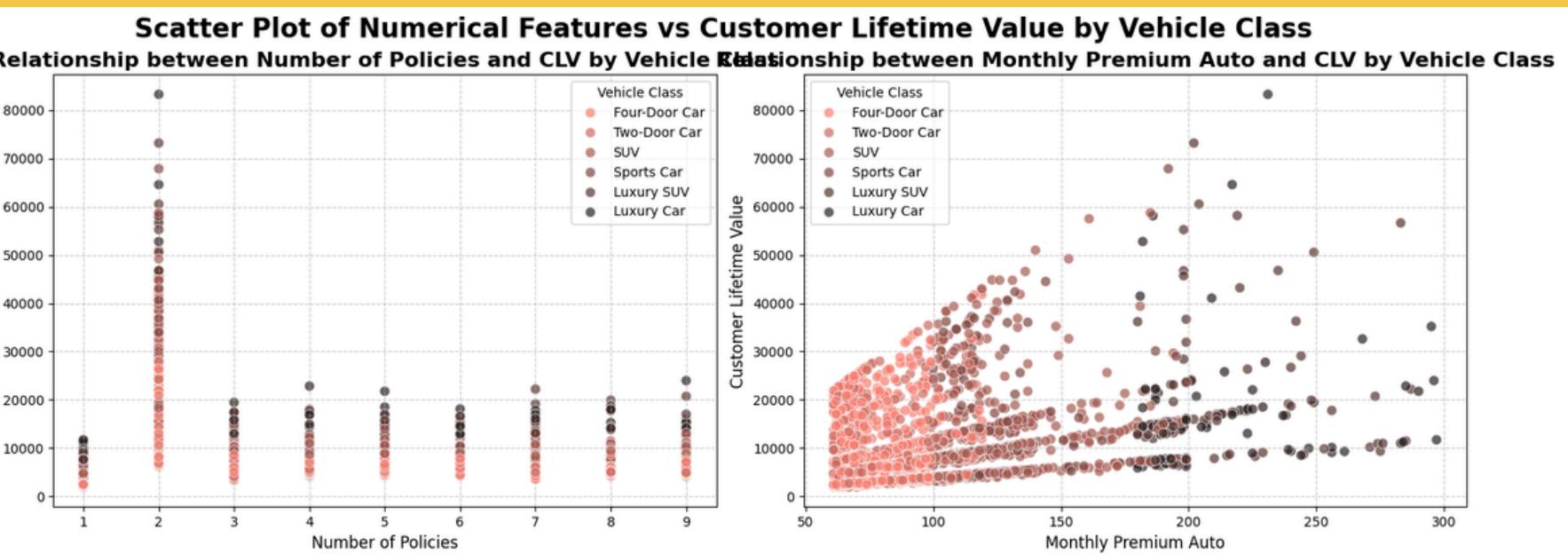
# Scatter Plot of Numerical Features vs Customer Lifetime Value by Coverage

Relationship between Number of Policies and CLV

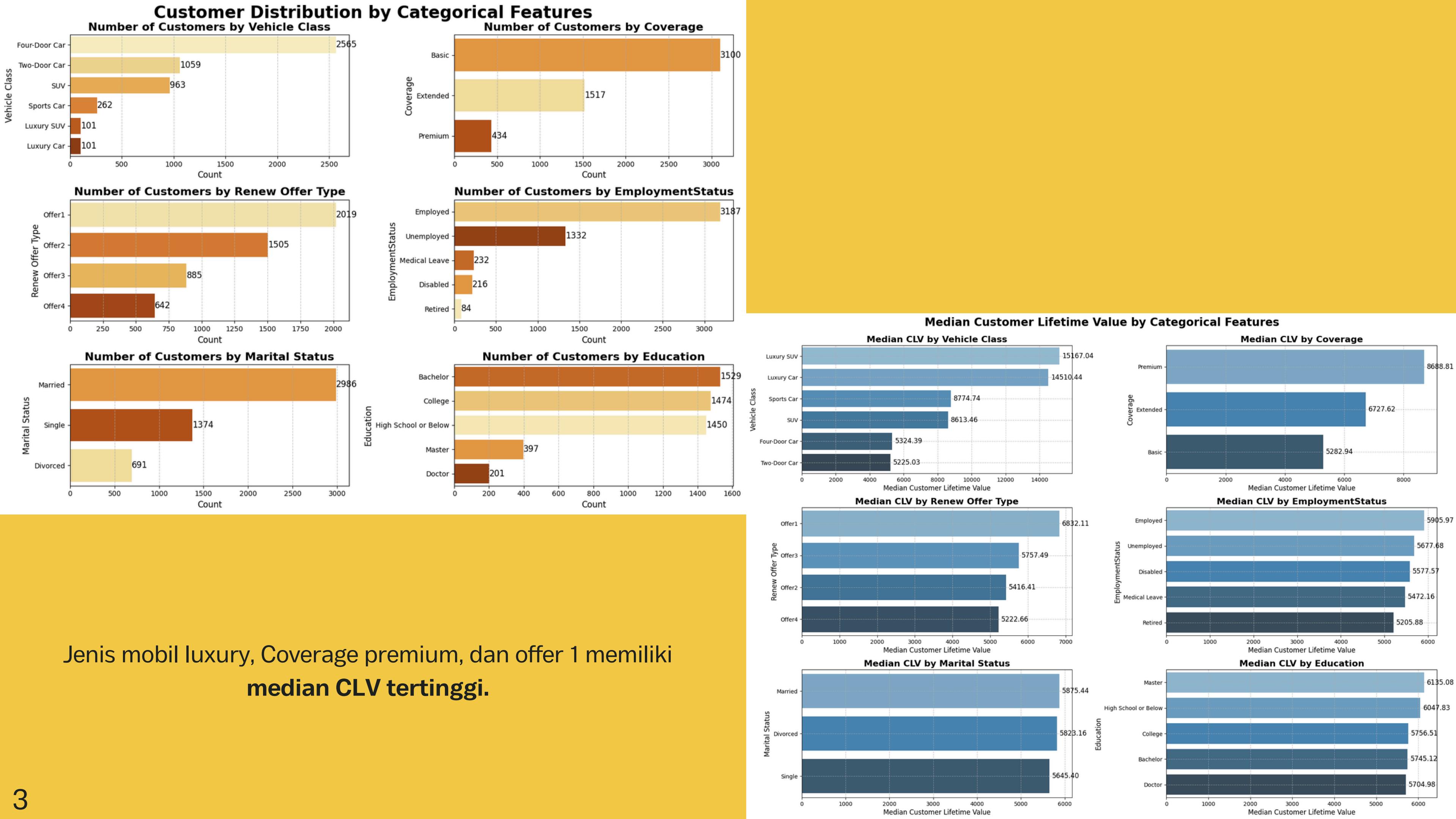


Semakin mahal jenis kendaraan, maka semakin besar angka CLV-nya (persebaran \*dot\* pada scatter plot semakin naik ke atas dengan warna yang lebih gelap).

Hal ini masuk akal karena customer tersebut lebih membutuhkan asuransi yang berjangka.

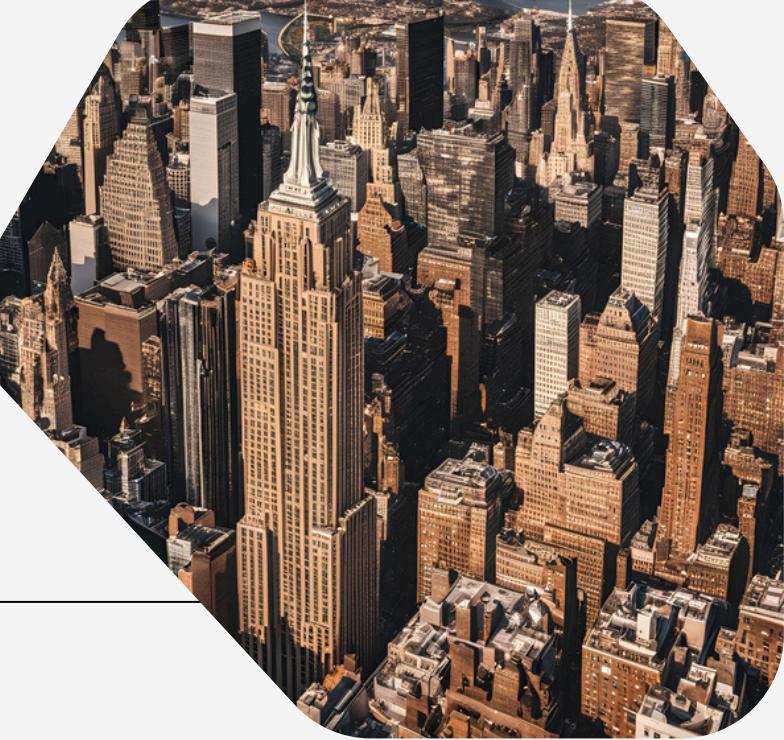


# Customer Distribution by Categorical Features



# Machine Learning Modeling





## Anomaly Handling

- Missing Value
- Outlier
- Duplicated
- Encoding (categorical feature)

## Evaluation Metrics

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

Where:

$\hat{y}_i$  = Predicted value for the  $i^{th}$  data point  
 $y_i$  = Actual value for the  $i^{th}$  data point  
 $n$  = number of observations

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i|$$

Where:

$\hat{y}_i$  = Predicted value for the  $i^{th}$  data point  
 $y_i$  = Actual value for the  $i^{th}$  data point  
 $n$  = number of observations

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{\hat{y}_i - y_i}{y_i} \right| \times 100$$

Where:

$\hat{y}_i$  = Predicted value for the  $i^{th}$  data point  
 $y_i$  = Actual value for the  $i^{th}$  data point  
 $n$  = number of observations

## Benchmark Model

	model	mean_RMSE	std_RMSE	mean_MAE	std_MAE	mean_MAPE	std_MAPE
0	Linear Regression	-2724.183	31.875	-2020.132	20.352	-0.372	0.011
1	KNN	-2760.071	43.115	-1868.387	50.082	-0.346	0.019
2	Decision Tree	-1240.130	67.248	-441.033	28.363	-0.056	0.004
3	Random Forest	-930.591	47.070	-361.231	16.846	-0.044	0.002
4	AdaBoost	-1376.433	36.232	-984.035	21.691	-0.147	0.003
5	XGBoost	-995.160	48.032	-446.744	6.670	-0.068	0.001
6	Gradient Boost	-900.118	52.492	-385.160	15.697	-0.051	0.002
7	Lasso	-2724.181	31.876	-2020.130	20.353	-0.372	0.011
8	Ridge	-2724.182	31.875	-2020.131	20.352	-0.372	0.011

-RMSE, Gradient Boost adalah model terbaik (-900.118)

MAE dan MAPE, RandomForest memiliki nilai paling rendah (-361.231 dan -0.044).

## Predict to Test Set

	RMSE	MAE	MAPE
rf	876.655	369.078	0.047
gbr	777.481	363.703	0.049

Gradient Boost menjadi model terpilih

## *Hyperparameter Tuning*

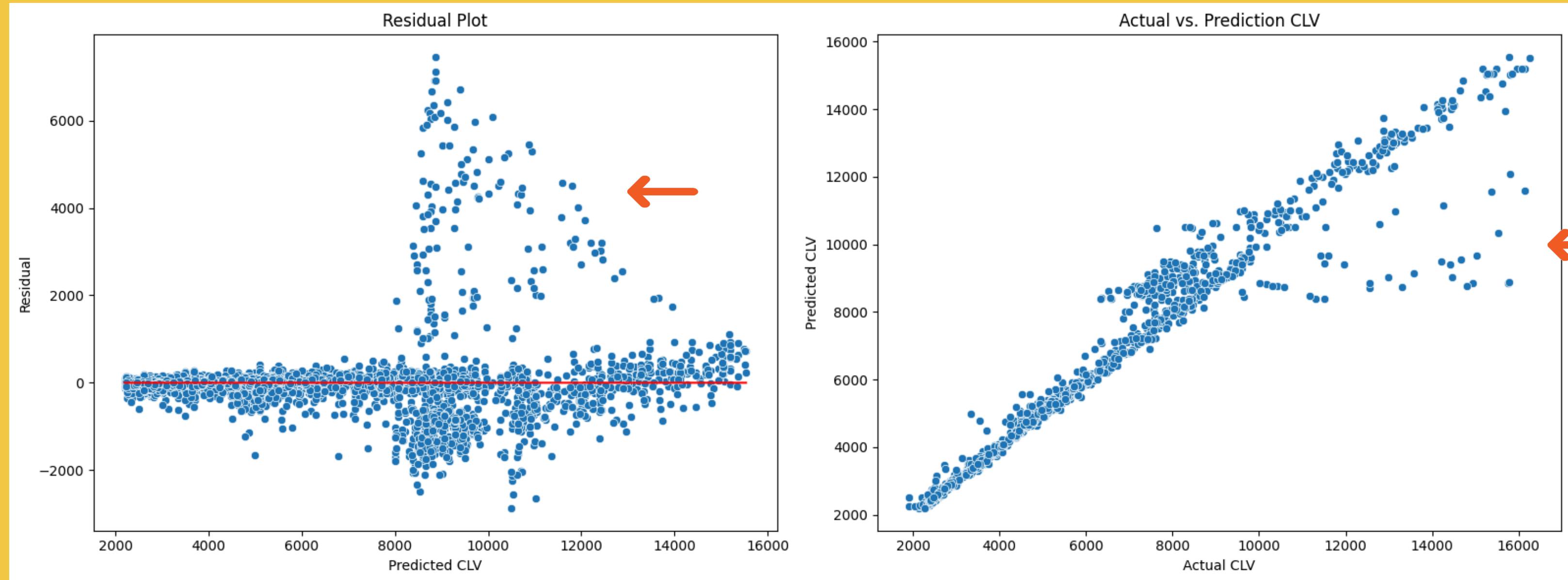
<b>Condition</b>	<b>RMSE</b>	<b>MAE</b>	<b>MAPE</b>
Before Tuning	-900.118	-385.160	-0.051
After Tuning (RandomizedSearch)	-890.543	-364.399	-0.044
After Tuning (GridSearch)	-889.465	-364.399	-0.044

Tuning menggunakan **GridSearch** menghasilkan **performa model yang lebih baik**.

## *Predict to Test Set with Tuned Model*

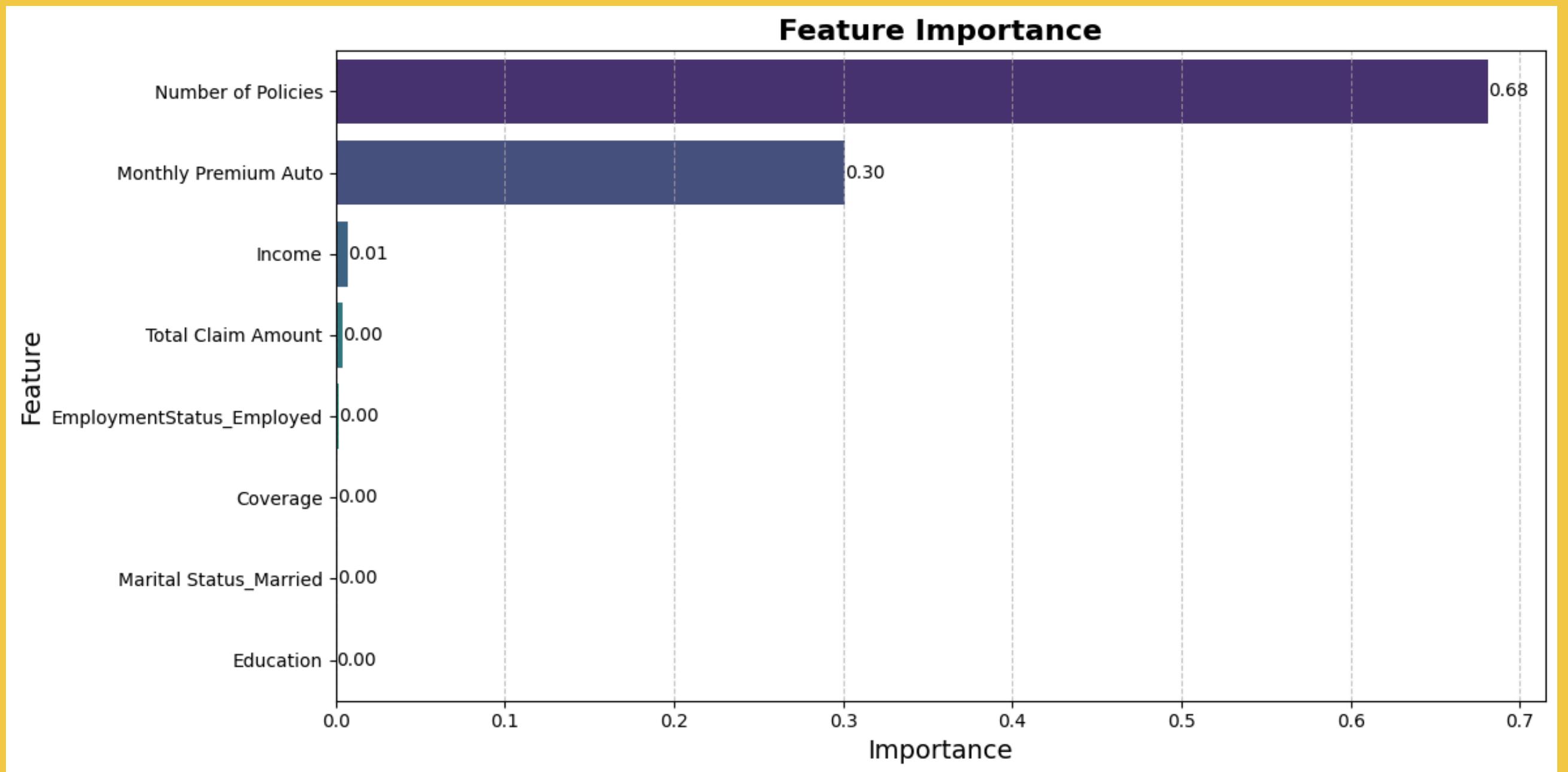
- RMSE, MAE & MAPE sebelum tuning: 777.481, 363.703, dan 0.049
- RMSE, MAE & MAPE setelah tuning: 778.689, 343.016, dan 0.045

# *Evaluation by Residual Plot*



- Prediksi model ini cukup akurat untuk memprediksi  $CLV < 8000$  karena terlihat dengan banyaknya nilai error yang mendekati nilai 0.
- Diatas 8000 menunjukkan bahwa variance dari residual tidak seragam.
- Pada plot Actual vs Predicted CLV, hasil prediksi juga menunjukkan hasil cukup akurat karena plotnya membentuk satu garis lurus dengan terdapat beberapa outlier mulai dari rentang  $\pm 9000$ .

# Feature Importance



Number of Policies menjadi fitur paling penting

# **Kesimpulan dan Saran**



# Kesimpulan

- Dalam pemodelan yang telah dilakukan, fitur `Number of Policies` dan `Monthly Premium Auto` menjadi faktor yang **paling signifikan**.
- Berdasarkan nilai MAE yang diperoleh setelah melakukan hyperparameter tuning, yaitu sebesar 343.016 (perkiraan nilai CLV bisa memiliki kesalahan sekitar  $\pm 343.016$  dari nilai CLV yang sebenarnya).
- Berdasarkan nilai MAPE, perkiraan nilai CLV dapat memiliki deviasi sebesar  $\pm 4.5\%$  dari nilai aktual CLV.



# Limitasi

- Maksimal biaya premi asuransi mobil (` Monthly Premium Auto `) = \$224.42
- Maksimal total klaim (` Total Claim Amount `) = \$1777
- Mampu memprediksi ` Customer Lifetime Value ` dengan baik pada rentang maksimal 16624.75. Diatas nilai tersebut, hasil akan bias.



# Rekomendasi

## Modeling

1. Identifikasi prediksi yang memiliki \*error\* tinggi.
2. Tambahkan fitur-fitur yang relevan seperti data pelanggan untuk memprediksi CLV
3. Model yang sudah dibangun dapat digunakan sebagai dasar untuk pengembangan model selanjutnya.

## Bisnis

1. Membuat penawaran yang dipersonalisasi kepada pelanggan. Ini akan membantu lebih baik memenuhi kebutuhan pelanggan berdasarkan jumlah polis dan biaya premi yang mereka bayarkan.
2. Melakukan *upselling* dan *cross-selling* kepada pelanggan. Strategi ini dapat meningkatkan jumlah polis dan premi yang dibayarkan, yang berpotensi meningkatkan CLV pelanggan.



# Thank You



Athaya Zahrani Irmansyah