Final Project

Insurance Customer Data

You NP-Complete Me

Meet the Team:
You NP-Complete Me

Adrian Aryaputra Hamzah
2206811474

Abbilhaidar Farras Zulfikar
2206026012

Fikri Dhiya Ramadhana
2206819533

Ravie Hasan Abud
2206031864

Outlines

- Dataset Overview
- Exploratory Data Analysis
- Data Preprocessing
- Feature Engineering
- Modeling
- Hyperparameter Tuning
- Clustering

Dataset Overview

Insurance Customer

You NP-Complete Me

Cuplikan Isi Dataset

Customer	State	Customer Lifetime V	Response	Coverage	Education	Effective To Date	EmploymentStatus	Gender	Income	Location Code	Marital Status	Monthly Premium A	Months Since Last C	Months Since Policy	Number of Open Cor	Number of Policies	Policy Type	Policy	Renew Offer Type	Sales Channel	Total Claim Amount	Vehicle Class	Vehicle Size
XA28126	California	29199.81072	0	Premium	Bachelor	2/12/11	Employed	M	67739	Rural	Divorced	149	4	76	0	2	Personal Auto	Personal L3	Offer1	Agent	270.373444	SUV	Medsize
EH27304	California	2955.495724	0		Doctor	1/21/11	Employed	F	56502	Suburban	Single	74	15	78	0	1	Personal Auto	Personal L1	Offer3	Branch	355.2	Four-Door Car	Medsize
QS83113	Oregon	5032.386371	0	Basic	Bachelor	1/10/11	Employed	M	30735	Suburban	Married	63	4	43	0	7	Personal Auto	Personal L3	Offer1	Agent		Two-Door Car	Small
YP32443	Oregon	4372.194536	0	Basic	Bachelor	1/23/11	Medical Leave	M	20354	Suburban	Married	116	23	0	0	1	Personal Auto	Personal L1	Offer3	Agent	556.8	SUV	Medsize
HP94242	Arizona	5288.173294	0		Bachelor	1/17/11	Employed	F	42621	Suburban	Married	66	8	3	0	3	Personal Auto	Personal L1	Offer2	Call Center	316.8	Four-Door Car	Medsize
OM50417	Arizona	29973.41592	0	Extended	Bachelor	2/16/11	Employed	F	48471	Rural	Married	83	5	20	1	2	Personal Auto	Personal L2	Offer1	Call Center	208.904212	Four-Door Car	Medsize
RP19541	Arizona	7582.113842	1	Basic	College	2/3/11	Employed	M	64801	Urban	Married	64	23	19	0	2	Personal Auto	Personal L3	Offer2	Web	268.471802	Four-Door Car	Medsize
IT37924	Oregon	9752.698201	0	Extended	College	1/17/11	Employed	F	37345	Rural	Married	82	27	12	0	2	Personal Auto	Personal L3	Offer4	Agent	31.652079	Four-Door Car	Medsize
VW65429	Arizona	6899.709949	0	Extended	High School or Belo	1/20/11	Employed	M	37308	Urban	Married	86	19	69	0	9	Personal Auto	Personal L2	Offer1	Agent	301.011106	Four-Door Car	Medsize
CT47738	Oregon	18205.67367	0	Extended	College	1/10/11	Unemployed	F		Suburban	Divorced	128	11	57	0	2	Corporate Auto	Corporate L3	Offer1	Branch	614.4	Sports Car	Medsize
FD79560	Arizona	10796.44832	0	Basic	High School or Belo	1/9/11	Employed	F	30761	Suburban	Divorced	69	16	16	2	2	Special Auto	Special L3	Offer1	Branch	331.2	Two-Door Car	Small
TM23788	California	6039.020697	0	Extended	High School or Belo	2/9/11	Unemployed	M		Suburban	Divorced	89	6	49	1	3	Personal Auto	Personal L3	Offer2	Agent		Four-Door Car	Medsize
MF30956	California	8945.689398	0	Basic	College	1/6/11	Employed	M	83761	Rural	Married	112	16	70	0	8	Corporate Auto	Corporate L2	Offer1	Web		SUV	Medsize
HS45195	California	6732.202033	0	Extended	High School or Belo	1/29/11	Unemployed	M		Suburban	Single	99	15	79	0	5	Personal Auto	Personal L3	Offer4	Web	712.8	Two-Door Car	Small
FA59186	California	4660.006868	0	Basic	High School or Belo	1/14/11	Unemployed	M		Suburban	Single	67	11	87	0	3	Personal Auto	Personal L2	Offer1	Web	482.4	Two-Door Car	Medsize
EJ73184	California	4743.694528	0	Basic	Bachelor	2/10/11	Employed	F	83413	Rural	Married	118	0	1	0	1	Personal Auto	Personal L3	Offer2	Web	134.801691	SUV	Medsize
ZK21724	Oregon	4016.541995	0	Basic	High School or Belo	2/10/11	Unemployed	F		Suburban	Single	111	4	85	0	1	Personal Auto	Personal L1	Offer3	Call Center		SUV	Large
HU22390	California	5389.499465	1	Extended	College	1/7/11	Employed	M	66429	Rural	Married	136	35	93	0	1	Corporate Auto	Corporate L3	Offer2	Agent		Sports Car	Medsize
SN41301	Oregon	6535.560574	0	Basic	Bachelor	2/10/11	Unemployed	M		Suburban	Single	65	14	62	0	2	Corporate Auto	Corporate L2	Offer3	Branch	468	Four-Door Car	Medsize
JL97265	Oregon	4611.2543	0	Basic	Master	1/7/11	Unemployed	M		Suburban	Married	61	4	15	0	7	Personal Auto	Personal L3	Offer2	Web	292.8	Four-Door Car	Medsize
KS12490	Oregon	10594.13249	0	Extended	High School or Belo	1/19/11	Employed	M	64669	Suburban	Married	131	22	32	0	6	Personal Auto	Personal L3	Offer4	Call Center	628.8	Sports Car	Small
ZS36394	California	5504.139033	1	Basic	Bachelor	2/8/11	Unemployed	F		Suburban	Married	73	18	45	0	5	Corporate Auto	Corporate L1	Offer1	Call Center		Four-Door Car	Medsize
TC99043	California	13736.1325	1	Premium	Master	2/13/11	Disabled	F	16181	Suburban	Divorced	181	22	79	0	8	Personal Auto	Personal L2	Offer1	Web		SUV	Medsize
HU94030	Arizona	10103.65572	0	Extended	High School or Belo	1/27/11	Employed	F	46667	Rural	Single	129	27	84	1	3	Corporate Auto	Corporate L2	Offer1	Branch	59.526095	SUV	Large
PK24825	Oregon	3293.245706	0	Extended	College	2/15/11	Unemployed	F		Suburban	Single	94	3	40	0	1	Corporate Auto	Corporate L1	Offer3	Call Center		Four-Door Car	Large
NM22175	Oregon	11490.33618	0	Premium	College	1/9/11	Unemployed	M		Suburban	Married	161	0	94	0	3	Personal Auto	Personal L3	Offer2	Branch	1052.933035	SUV	Large
QG51953	Oregon	4835.783179	0	Basic	College	1/29/11	Employed	M	36655	Suburban	Married	61	15	81	0	9	Personal Auto	Personal L3	Offer1	Web	295.223164	Four-Door Car	Small
CB62406	Arizona	6803.596768	0	Premium	High School or Belo	2/4/11	Employed	M	43758	Suburban	Married	171	16	29	0	1	Corporate Auto	Corporate L3	Offer1	Call Center	1129.929433	SUV	Medsize
UK68427	Washington	6369.262355	1	Extended	High School or Belo	1/20/11	Employed	M	34498	Suburban	Single	83	30	72	0	7	Personal Auto	Personal L3	Offer2	Branch	398.4	Two-Door Car	Small
XB43418	California	4815.851943	0	Extended	College	2/9/11	Employed	M	59438	Urban	Single	124	33	55	0	1	Special Auto	Special L2	Offer3	Branch	562.667789	SUV	Small
WY53775	Arizona	8382.630118	1	Basic	High School or Belo	2/4/11	Retired	M	19683	Suburban	Married	117	17	41	1	9	Personal Auto	Personal L1	Offer1	Web		Sports Car	Medsize
IO73222	Washington	5407.745153	0	Basic	Bachelor	1/12/11	Disabled	F	22555	Suburban	Married	70	1	62	0	3	Personal Auto	Personal L1	Offer1	Agent	336	Two-Door Car	Medsize
BR50492	Arizona	49423.79557	0	Extended	Bachelor	1/4/11	Employed	M	85058	Urban	Married	137	34	82	0	2	Personal Auto	Personal L1	Offer1	Call Center		SUV	Medsize
SO43919	California	14228.23432	0	Extended	Bachelor	1/16/11	Employed	M	31650	Suburban	Married	121	25	17	0	2	Personal Auto	Personal L3	Offer1	Branch	580.8	SUV	Medsize
TS31417	Washington	39769.55524	0	Basic	College	1/26/11	Medical Leave	M	16645	Suburban	Married	117	19	92	0	2	Personal Auto	Personal L3	Offer1	Web	696.9628	SUV	Medsize
IX18485	California	9594.248898	0	Extended	High School or Belo	2/1/11	Retired	F	27443	Suburban	Divorced	86	3	1	2	2	Personal Auto	Personal L2	Offer1	Agent	685.048914	Four-Door Car	Medsize
XS94482	Arizona	4151.667368	0	Basic	Master	2/17/11	Unemployed	F		Urban	Single	61	7	71	3	9	Personal Auto	Personal L2	Offer1	Web	208.892861	Two-Door Car	Medsize
FY77803	California	10129.50677	0	Basic	High School or Belo	1/11/11	Unemployed	F		Suburban	Married	68	34	31	0	2	Corporate Auto	Corporate L1	Offer4	Call Center	326.4	Four-Door Car	Medsize
ZU73588	California	5989.773931	1	Premium	College	1/19/11	Employed	M	66839	Suburban	Single	154	33	24	0	1	Personal Auto	Personal L3	Offer2	Branch	739.2	Sports Car	Medsize
CD95333	California	5638.930838	0	Basic	Bachelor	1/10/11	Employed	M	20920	Rural	Single	74	19	14	0	4	Personal Auto	Personal L1	Offer2	Branch	14.784843	Two-Door Car	Medsize
ZU96776	California	6189.122455	0	Extended	Bachelor	1/1/11	Unemployed	M		Suburban	Single	90	4	18	0	5	Personal Auto	Personal L1	Offer1	Call Center		Two-Door Car	Large
ZB82406	Oregon	6650.456653	0		Bachelor	2/27/11	Unemployed	F		Suburban	Single	99	16	65	2	3	Personal Auto	Personal L1	Offer2	Branch	712.8	Four-Door Car	Small
HN95240	Arizona	4984.095342	0	Basic	College	2/14/11	Unemployed	F		Suburban	Divorced	70	25	48	0	9	Corporate Auto	Corporate L1	Offer1	Agent	336	Four-Door Car	Small
XH97711	California	2430.506586	0	Basic	Bachelor	1/14/11	Employed	M	83140	Urban	Married	61	25	70	0	1	Corporate Auto	Corporate L3	Offer4	Agent	179.161843	Four-Door Car	Medsize
HE84643	California	2175.624856	0	Basic	High School or Belo	2/1/11	Unemployed	F		Suburban	Single	63	5	79	0	1	Personal Auto	Personal L3	Offer2	Branch	453.6	Four-Door Car	Medsize
YF81002	Oregon	3269.013926	0		College	2/11/11	Employed	M	26140	Suburban	Married	84	23	37	1	1	Personal Auto	Personal L3	Offer4	Branch	403.2	Two-Door Car	Small
XS17260	Oregon	12280.76661	1	Premium	College	1/1/11	Employed	M	88340	Suburban	Married	102	16	1	0	2	Personal Auto	Personal L3	Offer1	Branch		Two-Door Car	Medsize
EP34249	Oregon	13869.92705	1	Basic	College	1/21/11	Employed	M	28432	Suburban	Married	118	1	59	0	2	Corporate Auto	Corporate L1	Offer2	Branch	612.300581	SUV	Small
CQ19667	Oregon	4257.477861	0	Premium	Bachelor	1/6/11	Employed	F	49853	Suburban	Single	109	17	47	0	1	Personal Auto	Personal L3	Offer1	Branch		Four-Door Car	Medsize
JX78176	California	12045.67286	0	Extended	Bachelor	2/22/11	Employed	F	55805	Urban	Married	99	16	30	0	2	Personal Auto	Personal L3	Offer1	Branch	425.527151	Four-Door Car	Medsize
IS55638	California	15503.72309	0	Basic	College	2/18/11	Employed	F	29932	Suburban	Married	64	34	5	0	2	Personal Auto	Personal L2	Offer1	Branch		Two-Door Car	Medsize
XE17270	California	3520.696547	0	Extended	High School or Belo	2/2/11	Employed	M	31708	Suburban	Single	94	3	43	1	1	Personal Auto	Personal L1	Offer1	Agent	451.2	Two-Door Car	Medsize

Numerikal

- Customer Lifetime Value
- Income
- Number of Open Complaints
- Months Since Policy Inception
- Monthly Premium Auto
- Number of Policies
- Total Claim Amount

Kategorikal

Biner

- Gender

Ordinal

- Coverage
- Education
- Renew Offer Type
- Vehicle Class

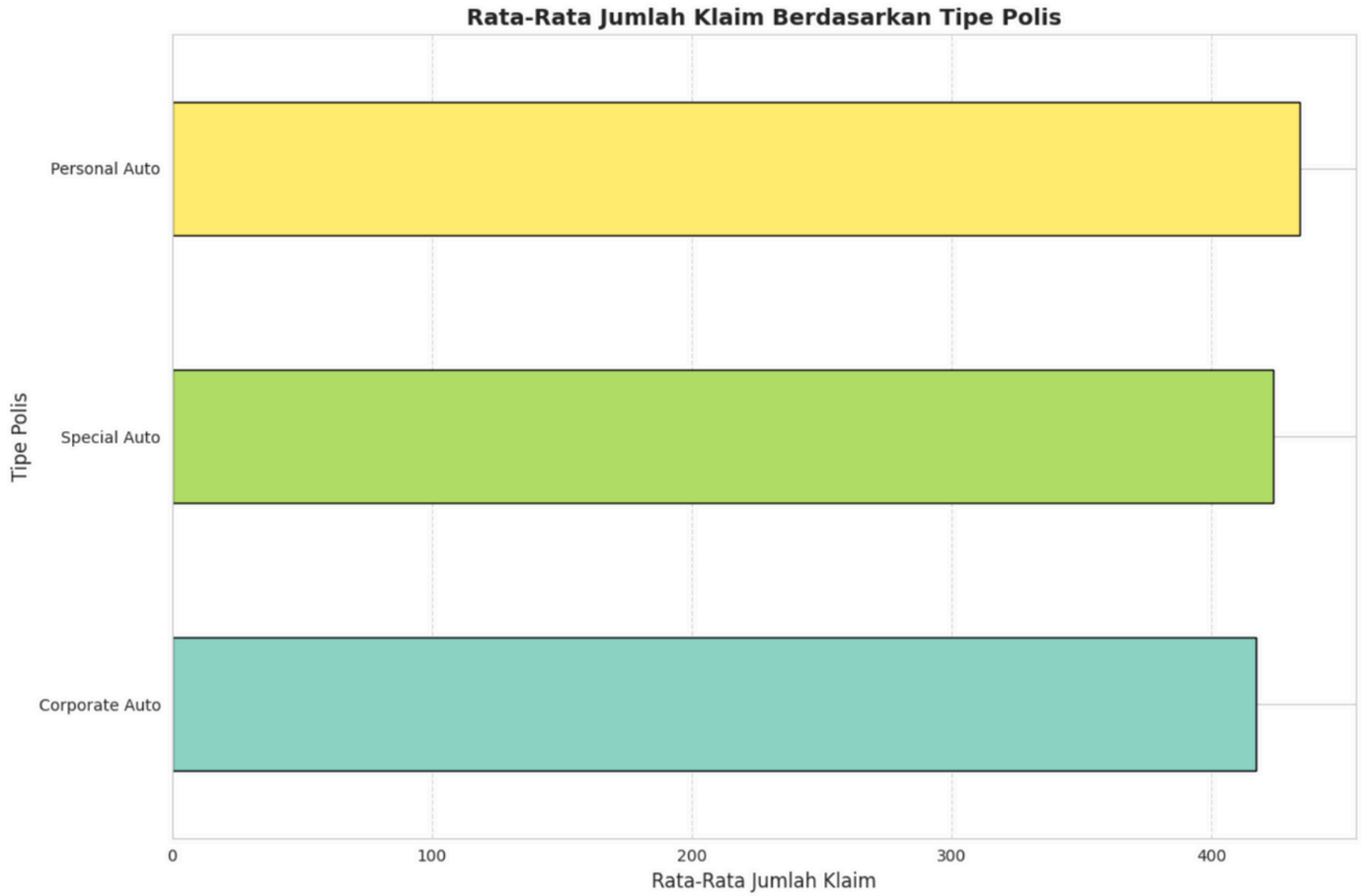
Nominal

- Policy
- Location Code
- Effective To Date
- Employment Status
- Marital Status
- Sales Channel
- Policy Type
- Vehicle Size
- State
- Response

Exploratory Data Analysis (EDA)

You NP-Complete Me

Insight 1: Tipe polis yang memiliki jumlah klaim rata-rata tertinggi

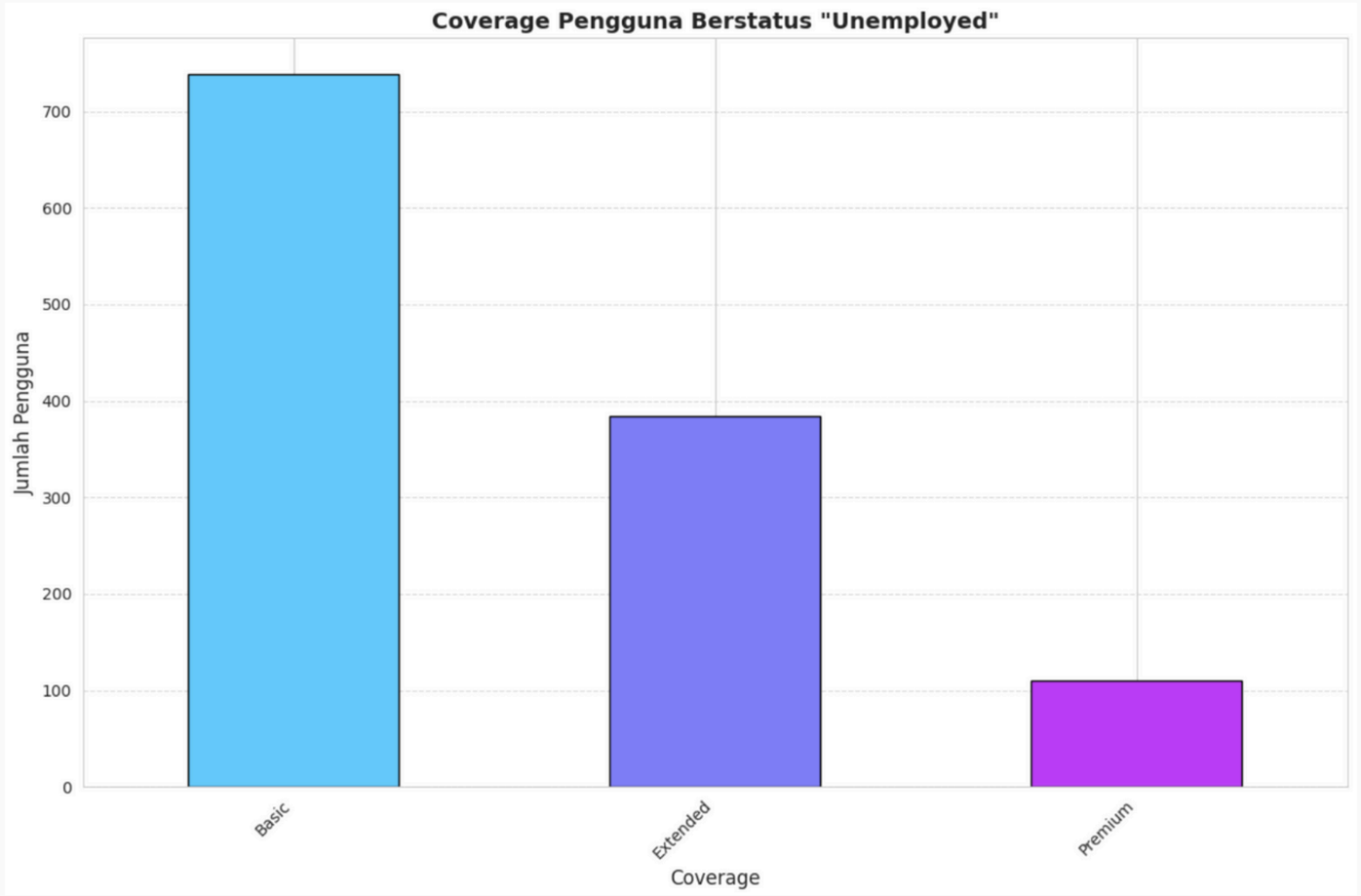


- Berdasarkan hasil visualisasi dengan *horizontal bar graph* di samping, dapat disimpulkan bahwa **tipe polis** yang memiliki jumlah klaim rata-rata tertinggi adalah **Personal Auto**.

- Alasan penggunaan *horizontal bar graph* adalah karena sifat dari tipe polis yang nominal (tidak ada ordering), sedangkan *vertical bar graph* secara implisit menyatakan bahwa ada logical order tertentu.

Figure: visualisasi rata-rata jumlah klaim untuk setiap tipe polis

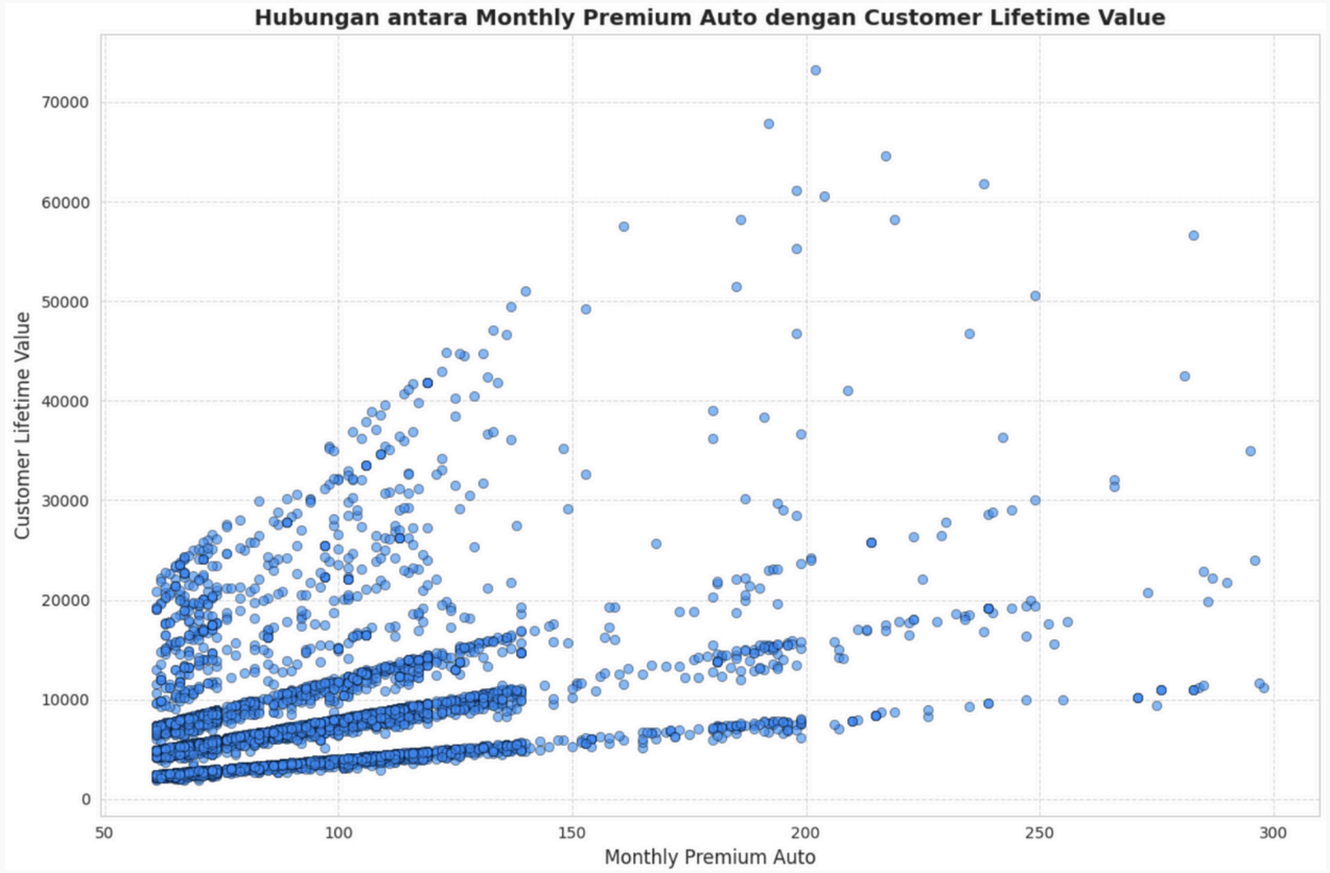
Insight 2: Coverage yang paling banyak dimiliki oleh pengguna yang berstatus pekerjaan Unemployed



Berdasarkan hasil visualisasi *bar graph* di samping, dapat disimpulkan bahwa **Coverage** yang paling banyak dimiliki pengguna yang berstatus pekerjaan **Unemployed** adalah **Basic**

Figure: visualisasi Coverage pada pengguna Unemployed

Insight 3a: Hubungan antara Monthly Premium Auto (MPA) dengan Customer Lifetime Value (CLV)



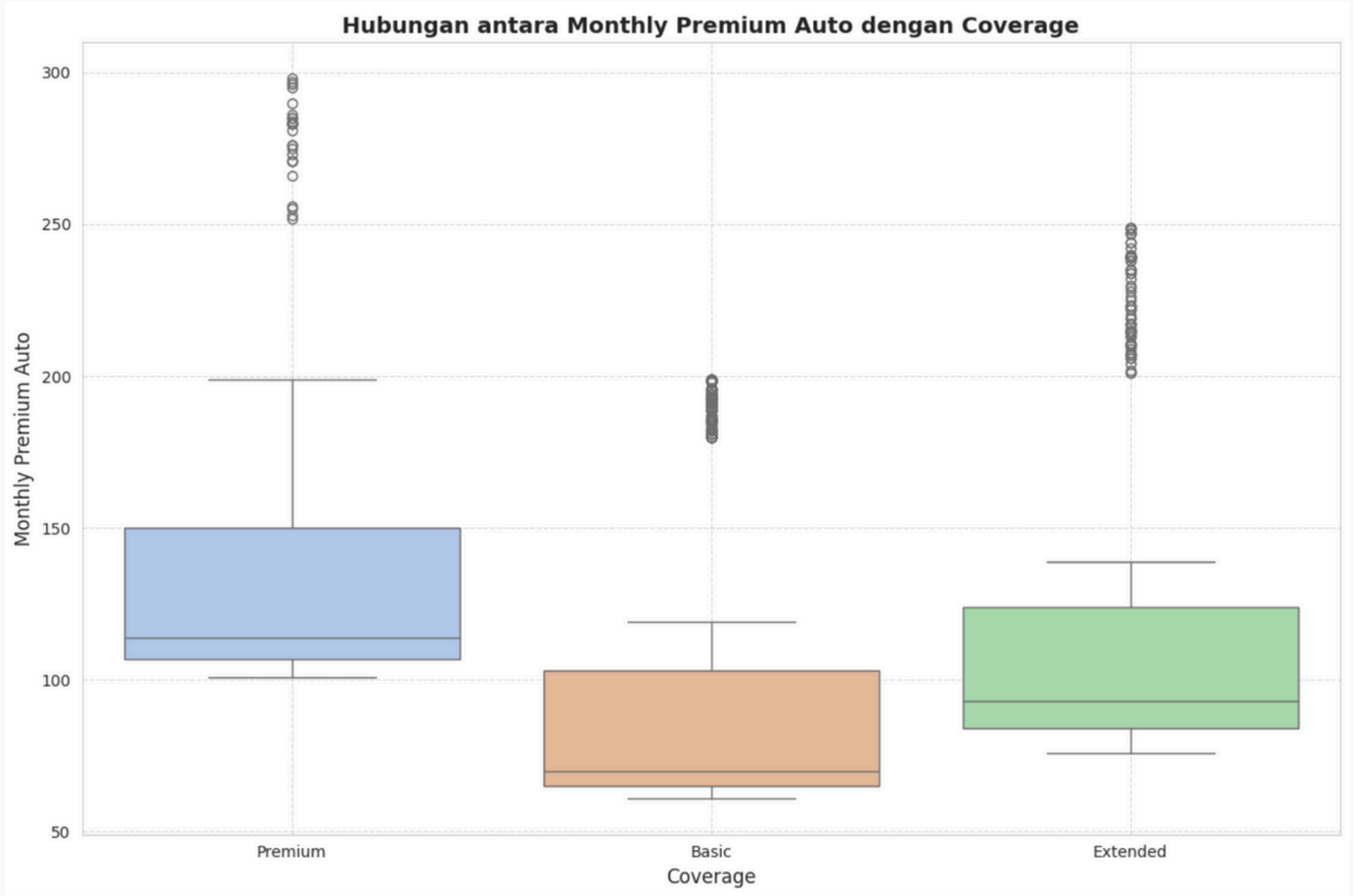
Korelasi antara Monthly Premium Auto dan Customer Lifetime Value sebesar 0.40

- Berdasarkan hasil visualisasi scatterplot di samping, terlihat bahwa terdapat korelasi sebesar 0.40 antara **Monthly Premium Auto** (MPA) dan **Customer Lifetime Value** (CLV).

- Ini menunjukkan adanya hubungan positif antar kedua variabel, semakin tinggi nilai MPA, cenderung semakin tinggi pula nilai CLV, meskipun tidak selalu.

Figure: Scatterplot MPA dan CLV

Insight 3b: Hubungan antara Monthly Premium Auto (MPA) dengan Coverage



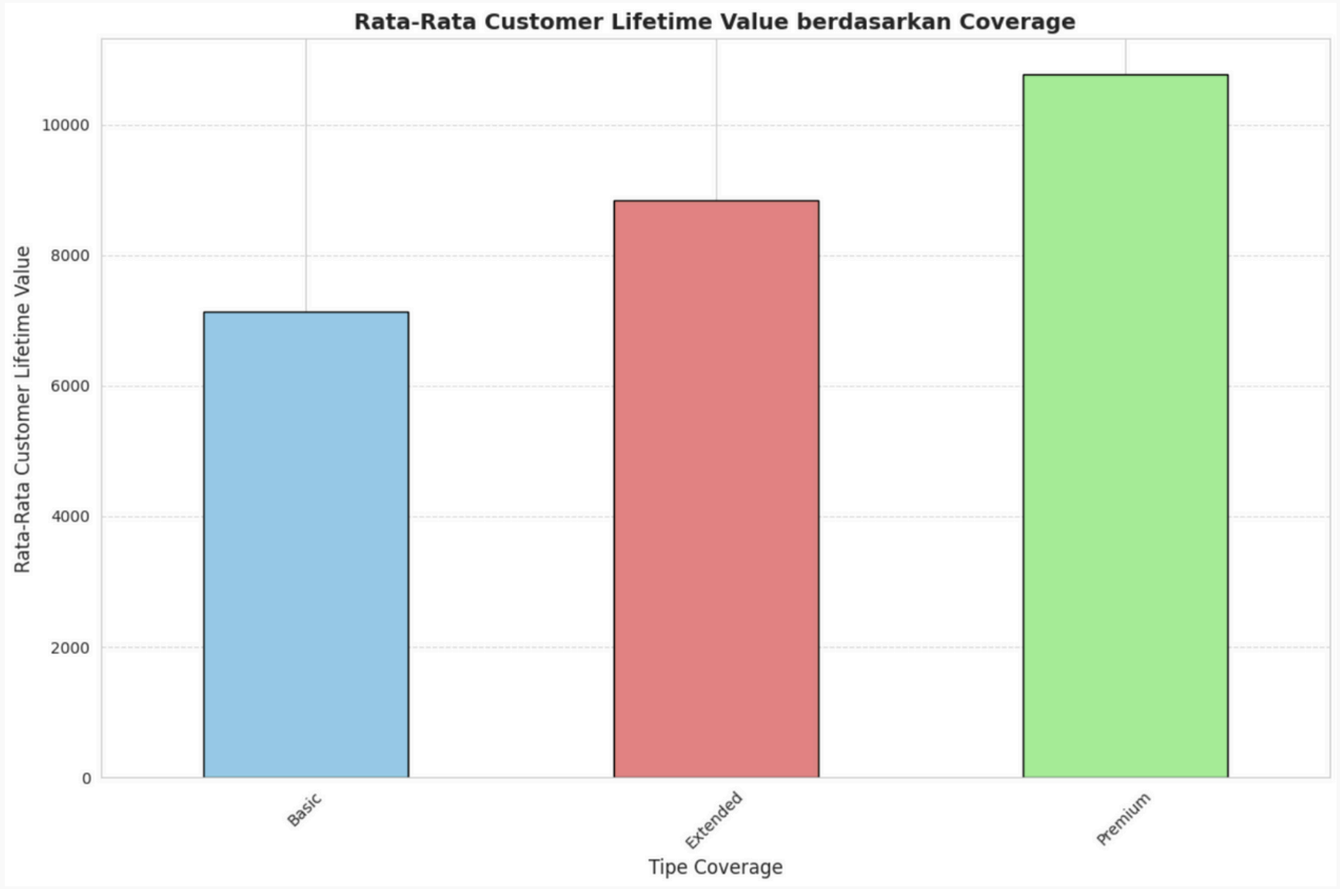
- Hasil dari plot di samping menunjukkan bahwa biaya bulanan asuransi otomotif (**Monthly Premium Auto**) berbeda berdasarkan jenis coverage yang dipilih. **Coverage Premium** memiliki median biaya bulanan tertinggi, diikuti oleh **Extended**, sementara **Basic** memiliki median terendah.

- Selain itu, **coverage Premium** menunjukkan variasi biaya yang lebih besar, dengan banyak outliers yang mencerminkan biaya bulanan jauh di atas mayoritas data. Outliers juga terlihat pada **Extended,** meskipun tidak sebanyak pada **Premium**.

- Secara keseluruhan, biaya bulanan tertinggi ditemukan pada **coverage Premium** sedangkan **Basic** memiliki biaya yang lebih rendah dan lebih stabil.

Figure: visualisasi Boxplot MPA untuk setiap tipe Coverage

Insight 4: Pelanggan dengan coverage 'Premium' cenderung memiliki Customer Lifetime Value (CLV) yang lebih tinggi dibandingkan pelanggan dengan coverage 'Basic'/'Extended'



- Berdasarkan tabel dan hasil visualisasi dengan bar graph, jelas bahwa pelanggan dengan jenis C**overage Premium** cenderung memiliki **CLV** yang lebih tinggi dibandingkan dengan pelanggan dengan jenis coverage lainnya.

- Alasan penggunaan bar graph vertikal adalah karena memang kolom **Coverage** sifatnya ordinal (ada logical order), dimulai dari **Basic**, **Extended**, setelah itu **Premium**.

Figure: visualisasi rata-rata CLV untuk setiap Coverage

Insight 5: Hubungan antara Total Claim Amount dan Renew Offer Type

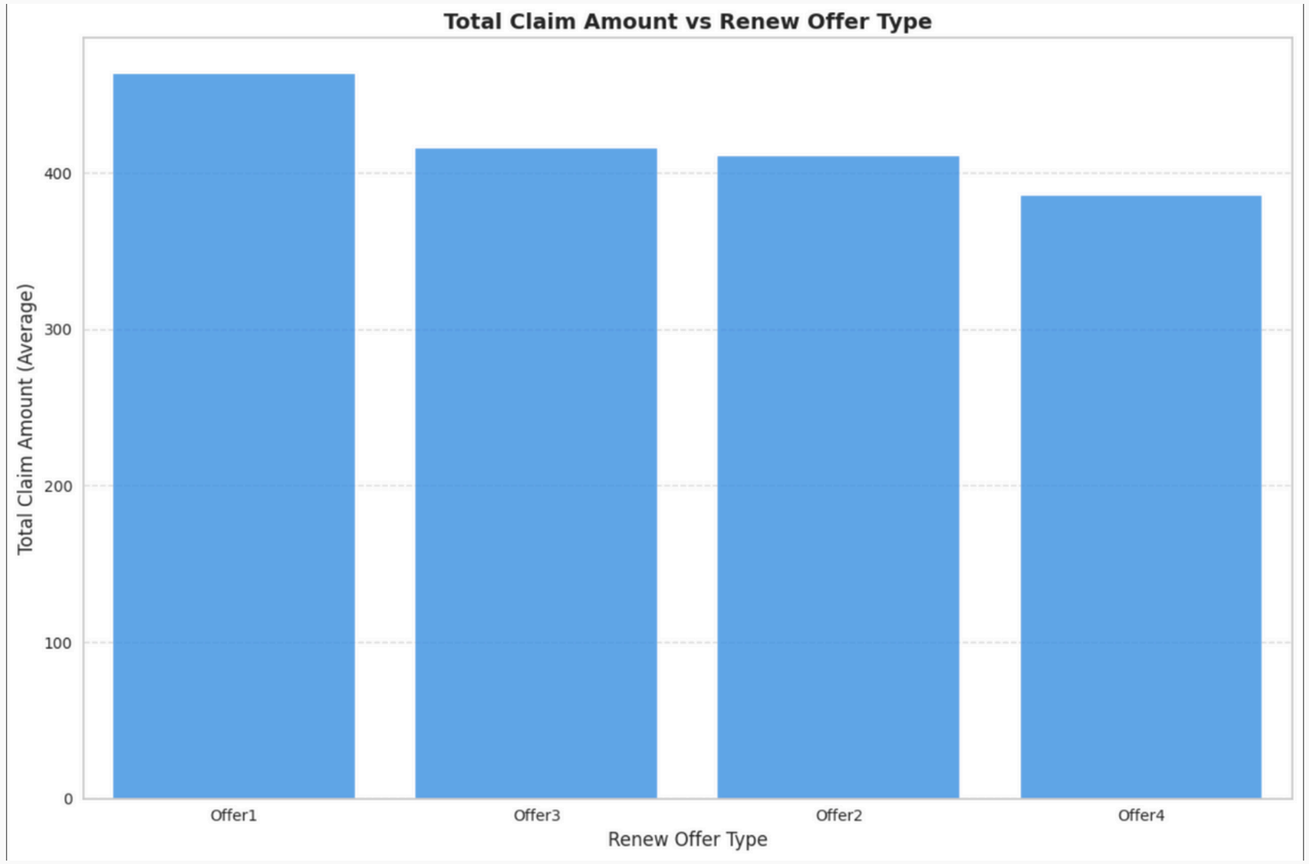
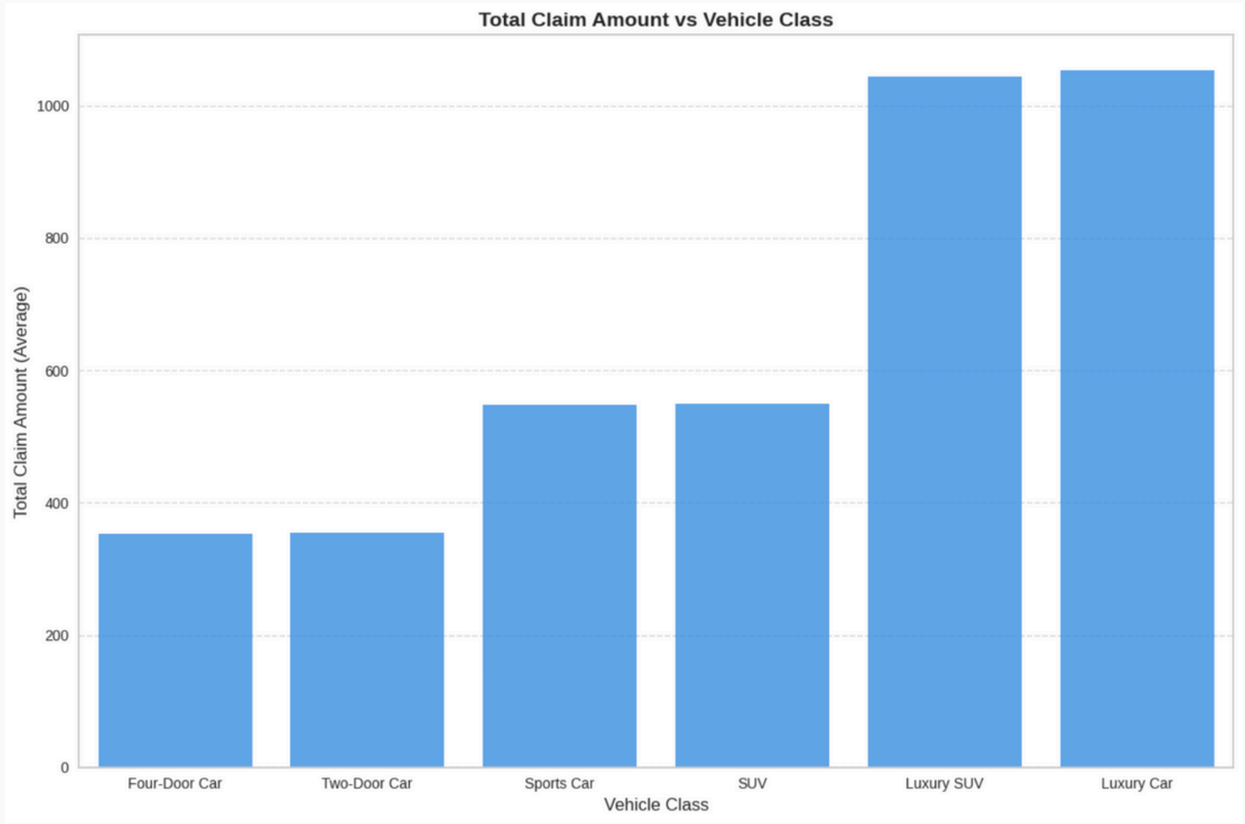


Figure: Total Claim Amount terhadap Renew offer Type

- Semakin besar jumlah **Total Claim Amount**, maka jenis penawaran perpanjangan akan semakin 'kecil'

- Pelanggan dengan **Total Claim Amount** yang rendah lebih memungkinkan untuk mendapatkan penawaran perpanjangan yang lebih menguntungkan (**Renew Offer Type**). Pada sisi lain, pelanggan dengan **Total Claim Amount** tinggi akan memerlukan penyesuaian dalam **Renew Offer Type.**

Insight 6: Hubungan antara Total Claim Amount dan Vehicle Classes



- Kendaraan yang mewah ataupun beresiko memiliki tendensi jumlah klaim yang lebih besar
- Hal ini disebabkan karena kendaraan mewah atau kendaranan dengan resiko besar cenderung untuk memiliki biaya perbaikan yang relatif lebih tinggi

Figure: Total Claim Amount terhadap Kelas Kendaraan

Data Preprocessing

You NP-Complete Me

Handling Missing Value

Terdapat 3 kolom yang memiliki missing value, yaitu **Coverage**, **Income** dan **Total Claim Amount**. Perlu diperhatikan bahwa **Coverage** bersifat kategorikal,. Sedangkan **Income** dan **Total Claim Amount** bersifat numerik, sehingga teknik untuk menghandle dapat berbeda.

	Column	# Missing	Persentase
0	Income	1404	25.620438
1	Total Claim Amount	917	16.733577
2	Coverage	666	12.153285

Handling Missing Value

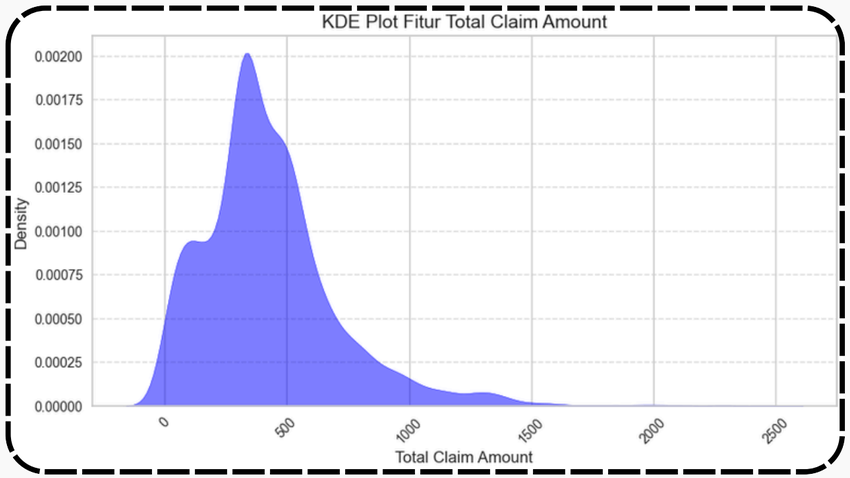
- Drop fitur **Income**
  sebab tidak relevan

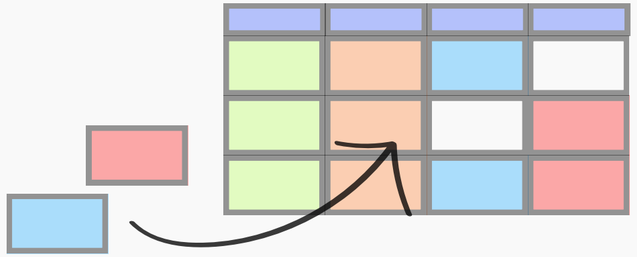
```
df = df.drop(columns=['Income'])
df_test_classif = df_test_classif.drop(columns=['Income'])
df_test_regress = df_test_regress.drop(columns=['Income'])
```

- Imputasi **Coverage** dengan mode
  sebab **Coverage** bertipe kategorikal

```
# Imputasi Fitur Coverage (kategorikal) dengan Mode
mode_imputer = SimpleImputer(strategy='most_frequent')
df['Coverage'] = mode_imputer.fit_transform(df[['Coverage']]).ravel()
mode_imputer = SimpleImputer(strategy='most_frequent')
df_test_classif['Coverage'] = mode_imputer.fit_transform(df_test_classif[['Coverage']]).ravel()
```

- Imputasi **Total Claim Amount** dengan median
  sebab data terdistribusi *skewed*





```
# Untuk Total Claim Amount (numerik)
median_imputer = SimpleImputer(strategy='median')
df['Total Claim Amount'] = median_imputer.fit_transform(df[['Total Claim Amount']]).ravel()
```

Handling Duplication

Periksa apakah ada baris duplikat pada dataset (setelah imputasi):

```
# Melihat banyaknya duplicated examples
print(f"Banyaknya baris duplikat: {df.duplicated().sum()}")
```

```
Banyaknya baris duplikat: 33
```

Kita akan drop row yang merupakan duplikat:

```
# Drop examples yang merupakan duplikat
df = df.drop_duplicates()
print(f"Banyaknya baris duplikat: {df.duplicated().sum()}")
```

```
Banyaknya baris duplikat: 0
```

Handling Outliers

Periksa outlier pada kolom numerik:

	Column	# Outlier	Persentase
0	Number of Open Complaints	1096	20.121168
1	Response	747	13.713971
2	Customer Lifetime Value	490	8.995777
3	Total Claim Amount	375	6.884524
4	Number of Policies	262	4.809987
5	Monthly Premium Auto	260	4.773270
6	Months Since Last Claim	0	0.000000
7	Months Since Policy Inception	0	0.000000

Karena sampai tahap ini kita belum menentukan model machine learning apa yang akan digunakan, maka kita akan tunda handling outliernya (menyesuaikan model).

Sebagai contoh: jika menggunakan tree-based model (robust terhadap outlier), maka kita dapat membiarkan outliernya.

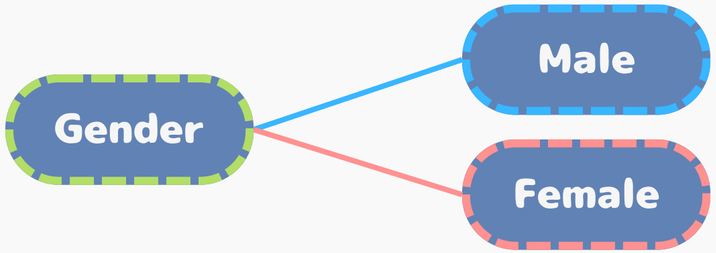
Feature Engineering

You NP-Complete Me

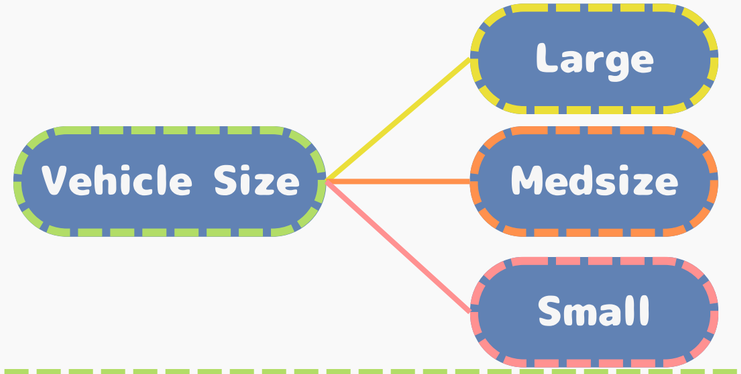
Encoding Categorical Columns

Secara umum, data kategorikal pada dataset dapat dibagi menjadi 3 jenis:

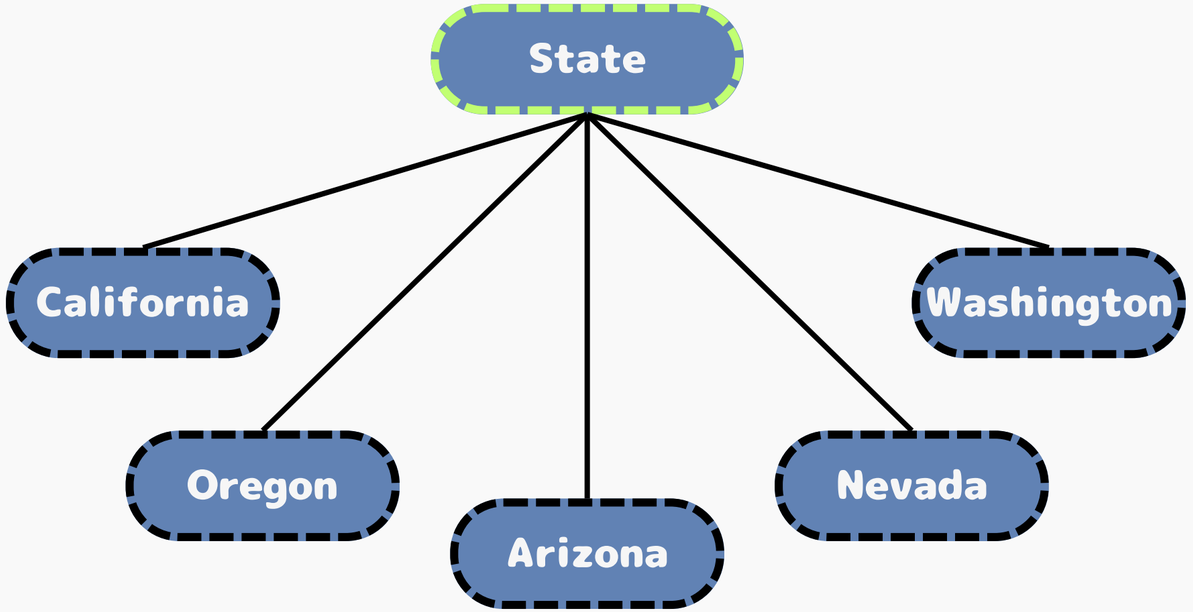
- **Biner** (hanya ada 2 kemungkinan nilai)



- **Ordinal** (memiliki urutan logikal)



- **Nominal** (tidak memiliki urutan logikal)



Encoding Categorical Columns

Teknik encoding yang dilakukan bergantung pada “jenis” data yang kategorikal

- **Biner:** Lakukan 0/1 label encoding

```
label_encoder = LabelEncoder()
insurance_copy['Gender'] = label_encoder.fit_transform(insurance['Gender'])
```

- **Ordinal:** Buat dictionary yang melakukan mapping pada setiap kemungkinan nilai kolom

```
ordinal_dict = {
    'Education': {
        'High School or Below': 0,
        'College': 1,
        'Bachelor': 2,
        'Master': 3,
        'Doctor': 4
    },
```

```
for col in ordinal_dict:
  insurance_copy[col] = insurance[col].map(ordinal_dict[col])
```

- **Nominal:** Lakukan One-Hot encoding

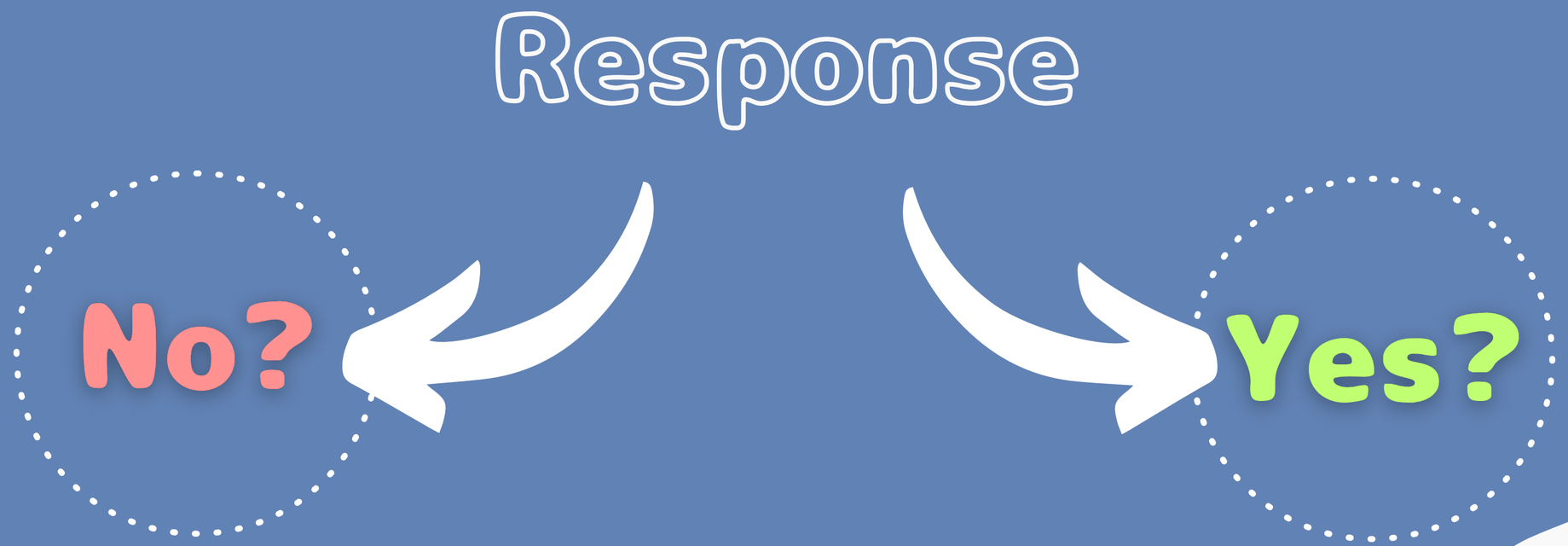
```
insurance_copy = pd.get_dummies(insurance_copy, columns=columns_to_encode)

bool_columns = insurance_copy.select_dtypes(include=['bool']).columns
for col in bool_columns:
    insurance_copy[col] = insurance_copy[col].astype(int)
```

Modeling

You NP-Complete Me

Classification



Imbalanced Dataset

Terdapat ketidakseimbangan jumlah data pada kolom target **Response**

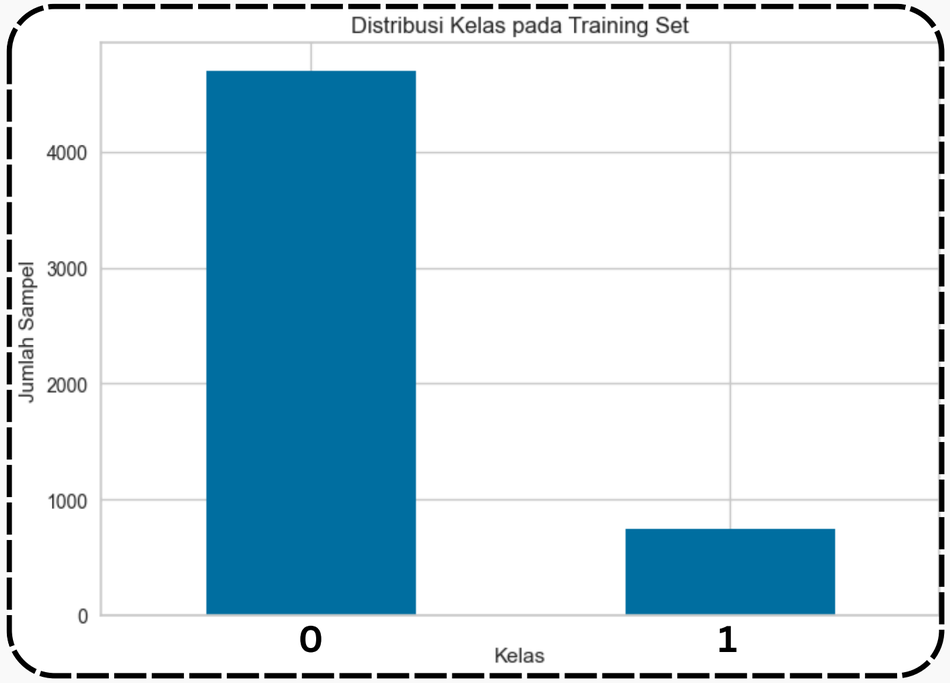
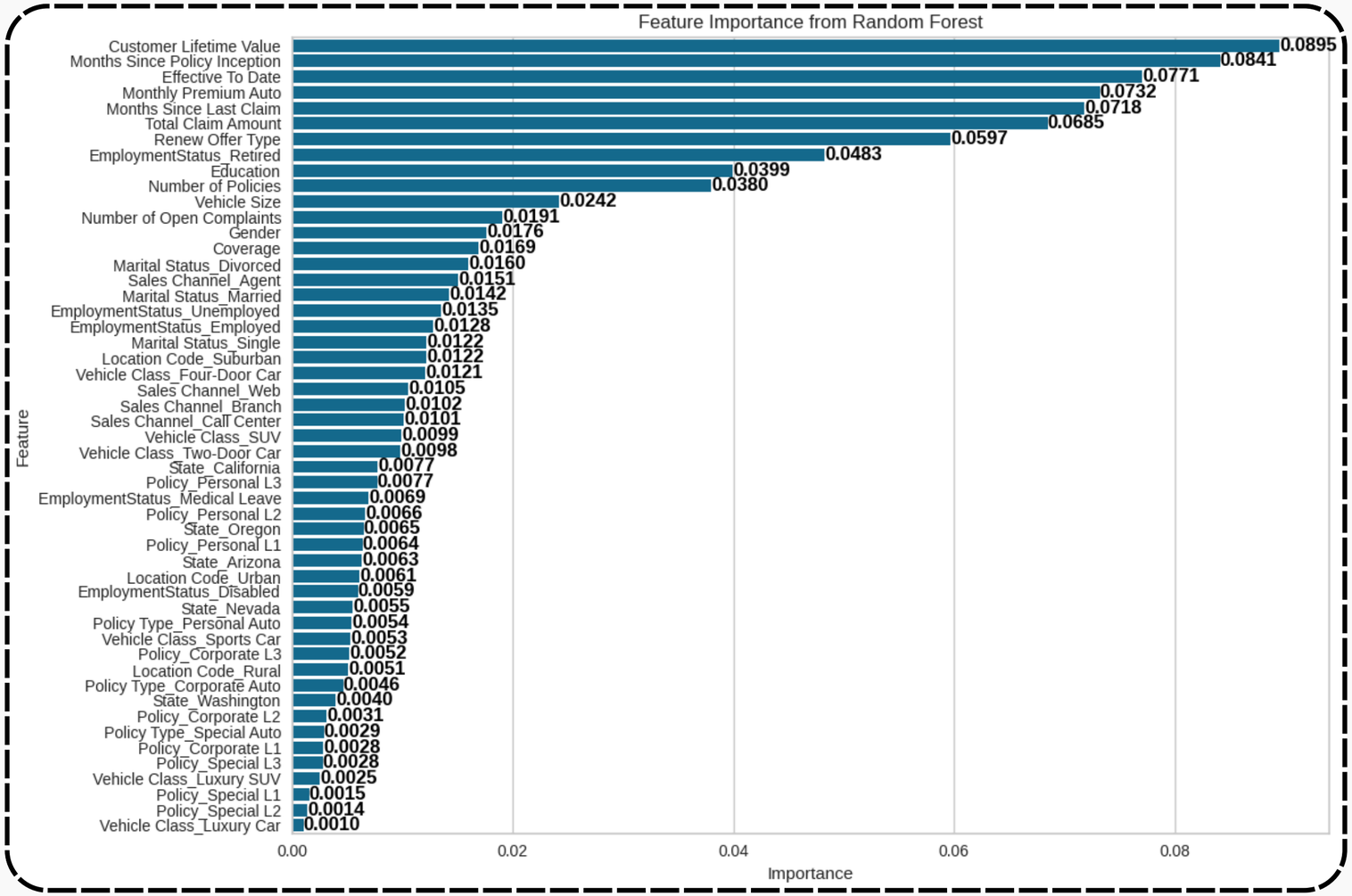


Figure: Distribusi Kelas pada kolom target **Response**

Namun pada preprocessing, **tidak dilakukan** Undersampling/Oversampling untuk meng-*handle* hal ini, sebab jumlah data/row pada dataset yang sudah kecil (~5000) sehingga perlakuan undersampling dapat mengurangi jumlah data secara signifikan. Selain itu, saat dilakukan resampling, performa model-model justru menurun karena alasan di atas.

Feature Engineering & Selection



Kami memutuskan untuk cukup mengambil sebagian *feature* dengan *importance score* tertinggi untuk menghindari *curse of dimensionality*.

Figure: *Feature importance* berdasarkan *machine learning model* (Random Forest)

Hyperparameter Tuning

You NP-Complete Me

Hyperparameter - Classification

Untuk klasifikasi, dilakukan ***hyperparameter tuning*** pada model berikut:

- **Random Forest Classifier**
- **Decision Tree Classifier**
- **Naive Bayes Classifier**
- **K-Neighbors Classifier**
- **MLP Classifier**
- **Logistic Regression**
- **XG Boost Classifier**

Dengan *hyperparameter tuning*, akan didapatkan parameter-parameter yang memberikan hasil terbaik untuk setiap model

Hyperparameter - Classification

```
'rf' : RandomForestClassifier(
    n_estimators=450,
    max_depth=50,
    min_samples_split=10,
    min_samples_leaf=2,
    max_features='log2',
    bootstrap=False,
    class_weight='balanced_subsample',
    random_state=42,
    n_jobs=-1
),
'knn' : KNeighborsClassifier(
    n_neighbors=5,
    weights='distance',
    metric='manhattan',
    n_jobs=-1
),
'xgb' : XGBClassifier(
    n_estimators = 500,
    max_depth = 30,
    learning_rate = 0.1,
    min_child_weight = 2,
    subsample = 0.8,
    colsample_bytree = 0.25,
    gamma = 2,
    reg_alpha = 1,
    reg_lambda = 1,
    random_state = 42,
    eval_metric = 'logloss'
),
```

```
'logistic' : LogisticRegression(
    C=10,
    class_weight=None,
    max_iter=100,
    solver='lbfgs',
    random_state=42
),
'mlp' : MLPClassifier(
    activation='tanh',
    alpha=0.0005717487921177058,
    batch_size=32,
    hidden_layer_sizes=(255, 70),
    learning_rate_init=0.0008730443503415081, max_iter=300,
    random_state=42
),
```

```
'dt' : DecisionTreeClassifier(
    max_depth=50,
    min_samples_split=2,
    min_samples_leaf=1,
    max_features=0.5,
    class_weight=None,
    criterion='entropy',
    random_state=42
),
'stacking' : StackingClassifier(
    estimators=[
        ('rf', RandomForestClassifier(
            n_estimators=450,
            max_depth=50,
            min_samples_split=10,
            min_samples_leaf=2,
            max_features='log2',
            bootstrap=False,
            class_weight='balanced_subsample',
            random_state=42,
            n_jobs=-1
        )),
        ('xgb', XGBClassifier(
            n_estimators = 450,
            max_depth = 50,
            learning_rate = 0.1,
            min_child_weight = 2,
            subsample = 0.8,
            colsample_bytree = 0.25,
            gamma = 2,
            reg_alpha = 1,
            reg_lambda = 1,
            random_state = 42,
            eval_metric = 'logloss'
        ))],
    final_estimator=LogisticRegression(),
    cv=5,
    n_jobs=-1
),
```

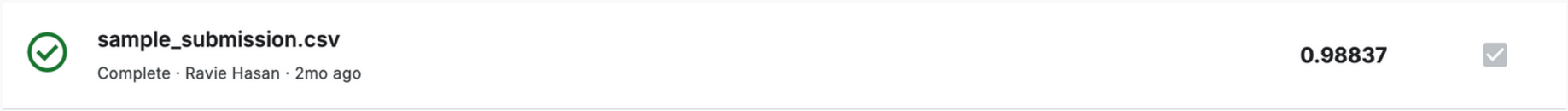
Model Evaluation with Statified 5-Fold CV

Rank	Model	F1	Precision	Recall
1	Random Forest	0.988773	0.988778	0.988801
2	Stacking (Random Forest + XGBoost + Logistic Regression)	0.987961	0.988001	0.988066
3	Multi Layer Perceptron	0.971025	0.972277	0.970441
4	Decision Tree	0.938003	0.945756	0.934641
5	K-Nearest Neighbors	0.924743	0.942550	0.918304
6	XGBoost	0.891261	0.907904	0.908388
7	Bernoulli Naive Bayes	0.839574	0.851305	0.873875
8	Logistic Regression	0.837006	0.856474	0.874792

--> Ketiga metrik menggunakan averaging="weighted"

Model dengan performa terbaik pada Stratified 5-Fold CV: **RandomForestClassifier**

Testing on Kaggle



Model dengan performa terbaik Kaggle (test set):

StackingClassifier & RandomForestClassifier

Regression

Customer Lifetime Value

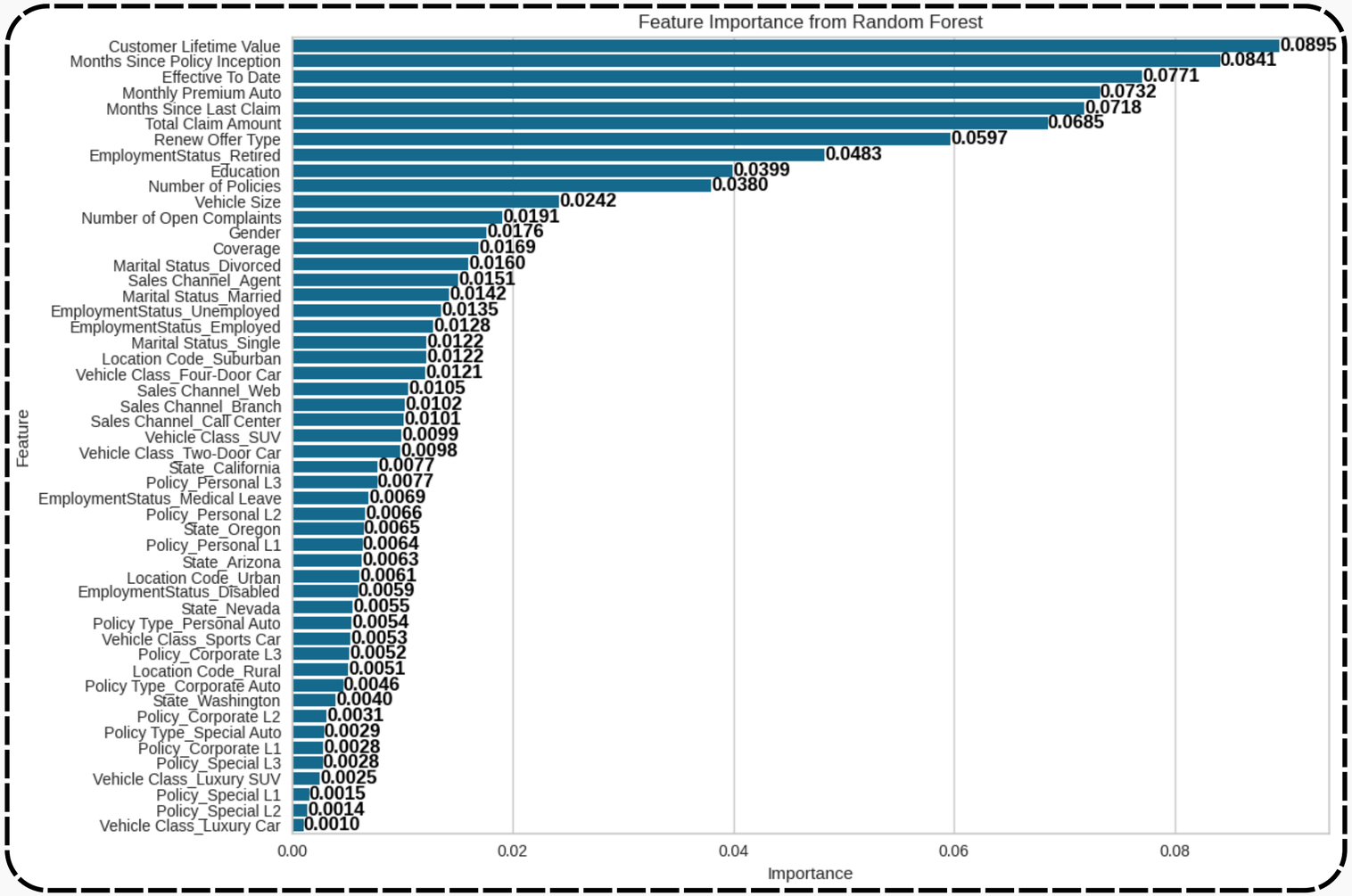
423,020

582,58

1231,422

???,??

Feature Engineering & Selection



Kami memutuskan untuk cukup mengambil sebagian *feature* dengan *importance score* tertinggi untuk menghindari *curse of dimensionality*.

Figure: *Feature importance* berdasarkan *machine learning model* (Random Forest)

Hyperparameter Tuning

You NP-Complete Me

Hyperparameter - Regression

Untuk regresi, dilakukan ***hyperparameter tuning*** pada model berikut:

- **Random Forest Regressor**
- **XG Boost Regressor**
- **MLP Regressor**
- **Linear Regression**
- **Ridge Regression**
- **Lasso Regression**

Dengan *hyperparameter tuning*, akan didapatkan parameter-parameter yang memberikan hasil terbaik untuk setiap model

Hyperparameter - Regression

```
'xgb' : xgb.XGBRegressor(
    learning_rate=0.006528323605514932,
    max_depth=24,
    n_estimators=600,
    min_child_weight=7,
    gamma=0.6277528998572619,
    subsample=0.6008332332699823,
    colsample_bytree=0.8811947381492168,
    reg_alpha=1.3903944538858608,
    reg_lambda=1.461494393149832,
    max_leaves=553,
    max_bin=414,
    scale_pos_weight=4.447507170690486,
    random_state=42,
    n_jobs=-1),
'mlp' : MLPRegressor(
    hidden_layer_sizes=(80, 256, 112),
    learning_rate_init=0.00024528595387536637,
    alpha=2.3749803745884497e-05,
    batch_size=128,
    activation='relu',
    solver='adam',
    max_iter=1000,
    early_stopping=True,
    random_state=42),
'rf' : RandomForestRegressor(
    n_estimators=350,
    max_depth=50,
    min_samples_split=10,
    min_samples_leaf=2,
    max_features=0.5,
    bootstrap=True,
    random_state=42,
    n_jobs=-1),
```

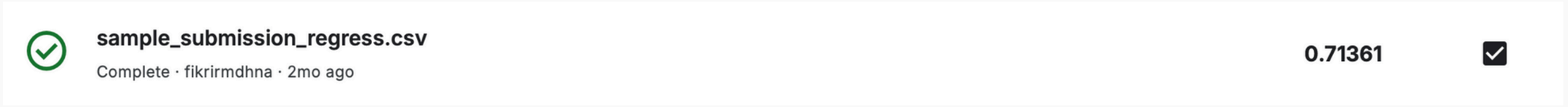
```
'rf2' : RandomForestRegressor(
    n_estimators=200,
    random_state=42),
'linear' : LinearRegression(),
'ridge' : Ridge(
    alpha=10,
    solver='sag',
    tol=0.1,
    random_state=42),
'lasso' : Lasso(
    alpha=100,
    max_iter=1000,
    tol=0.0001,
    random_state=42),
```

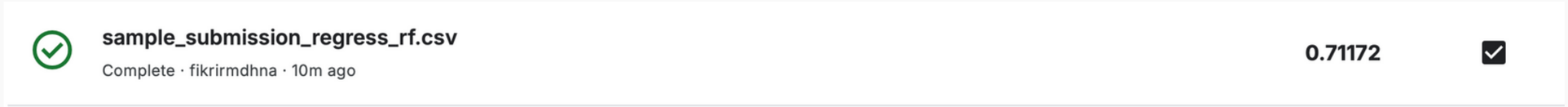
Model Evaluation with 5-Fold CV

Rank	Model	R-squared	MAE	RMSE
1	Random Forest	0.704046	1543.234522	3711.891165
2	XGBoost	0.698144	1541.378735	3747.922576
3	Random Forest (Different Hyperparameters)	0.697052	1495.279642	3756.928417
4	Multi Layer Perceptron	0.6715	1876.5218	3909.8918
5	Lasso Regression	0.159509	3859.837867	6255.040892
6	Linear Regression	0.154709	3895.441364	6273.025421
7	Ridge Regression	0.121072	4004.816108	6389.576384

Model dengan performa terbaik pada 5-Fold CV:
RandomForestRegressor

Testing on Kaggle





Model dengan performa terbaik Kaggle (test set):

RandomForestRegressor & XGBRegressor

Clustering

You NP-Complete Me

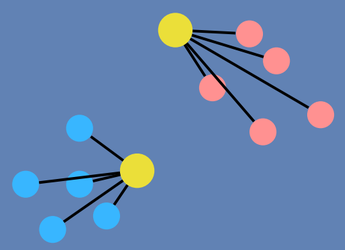
Clustering

Cleaned Data

Data Analysis

Select Feature

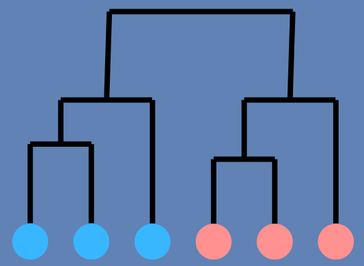
Partitional



K-Means

Clustering

Hierarchical



Agglomerative

Pemilihan Fitur

Customer Lifetime Value

Mengukur total kontribusi finansial pelanggan, penting untuk menilai nilai jangka panjang pelanggan dan mengoptimalkan strategi pemasaran

Total Claim Amount

Menunjukkan total klaim yang diajukan, penting untuk menilai risiko dan dampak finansial bagi perusahaan

Vehicle Class_Luxury Car

Kendaraan mewah memiliki risiko klaim lebih tinggi karena biaya perbaikan atau penggantian yang mahal sehingga penting untuk menilai risiko bagi perusahaan

Selain dari segi *domain knowledge*, pemilihan fitur ini didukung dengan percobaan yang juga dilakukan untuk seleksi fitur, seperti PCA, kombinasi fitur lain, dan pendekatan preprocess berbeda.

KMeans Clustering
Silhouette Method

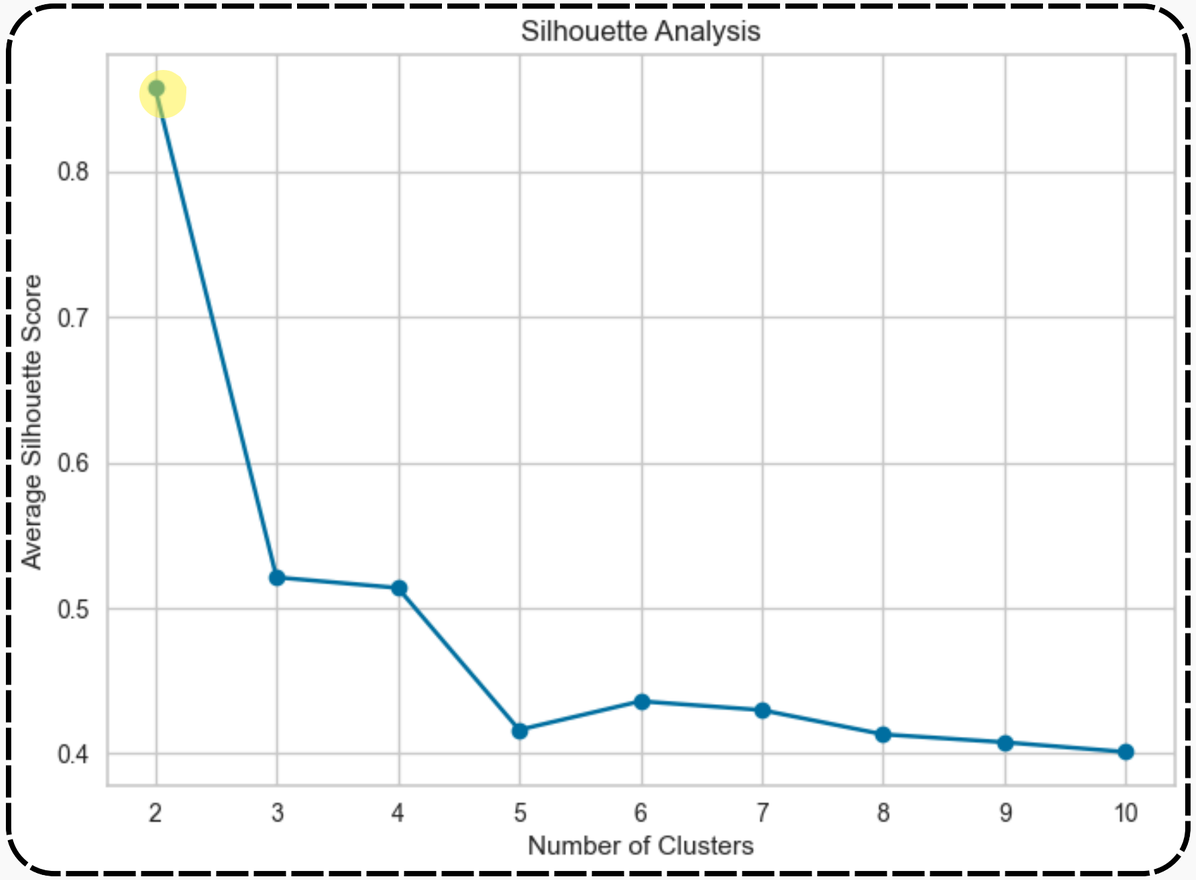


Figure: *Silhouette Coefficient* untuk setiap banyak *cluster*

Terlihat bahwa Silhouette Score di titik k = 2 merupakan yang paling tinggi dari titik k lainnya dengan score sebesar 0.8572.

KMeans Clustering
Elbow Method

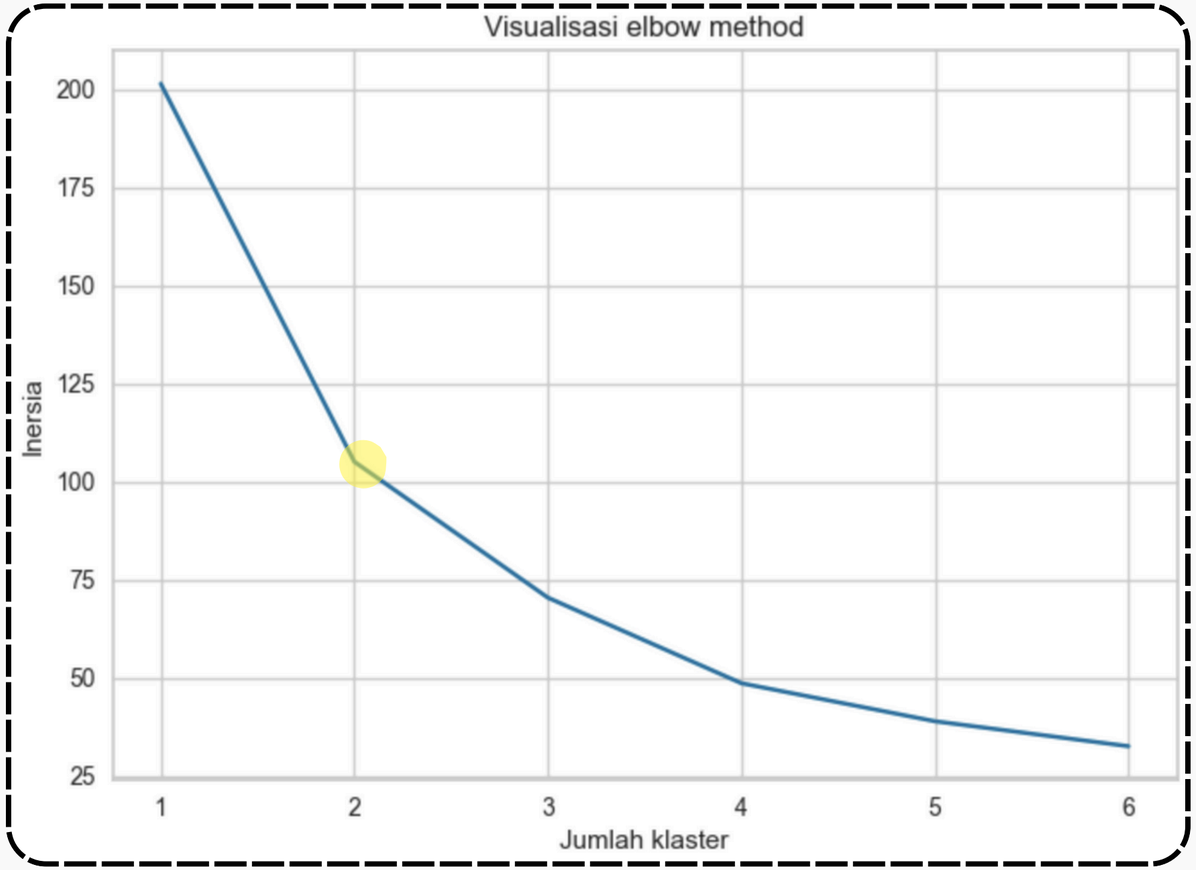
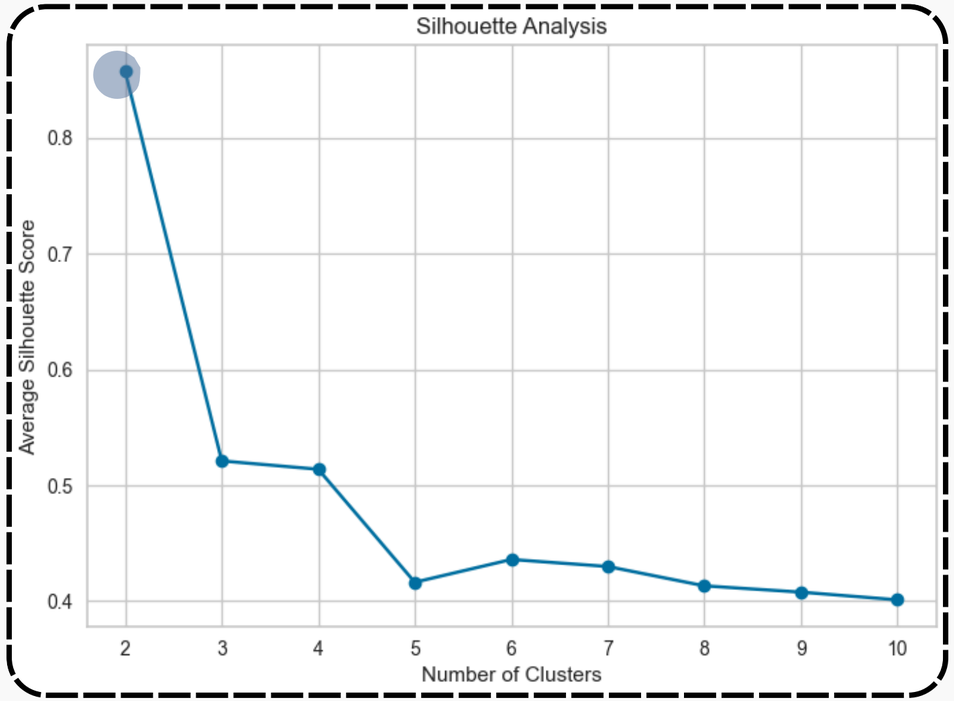


Figure: *Inersia Value* untuk setiap banyak *cluster*

- Terlihat adanya tikungan tajam ("elbow") di titik k = 2, hal ini menandakan penurunan inersia yang signifikan sebelum stabil pada nilai-nilai berikutnya

- Penurunan inersia setelah k = 2 semakin tidak drastis, berarti menambah jumlah cluster menjadi tidak memberikan keuntungan signifikan

Pemilihan Jumlah Cluster Optimal



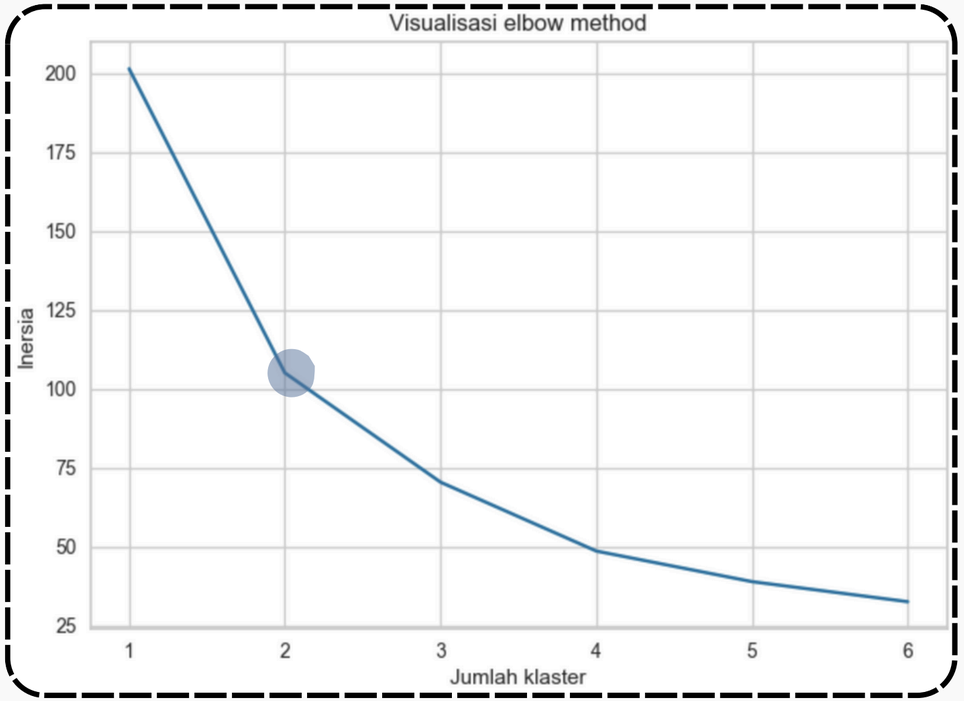


Figure: Perbandingan *Silhouette Method* dengan *Elbow Method*

Jadi, banyak **cluster optimal** yang dipilih adalah 2 karena:

- Tikungan paling tajam ("elbow") ada di titik k = 2
- Silhoutte coefficient tertinggi ada di titik k=2

Hierarchical Agglomerative Clustering
Ward Linkage

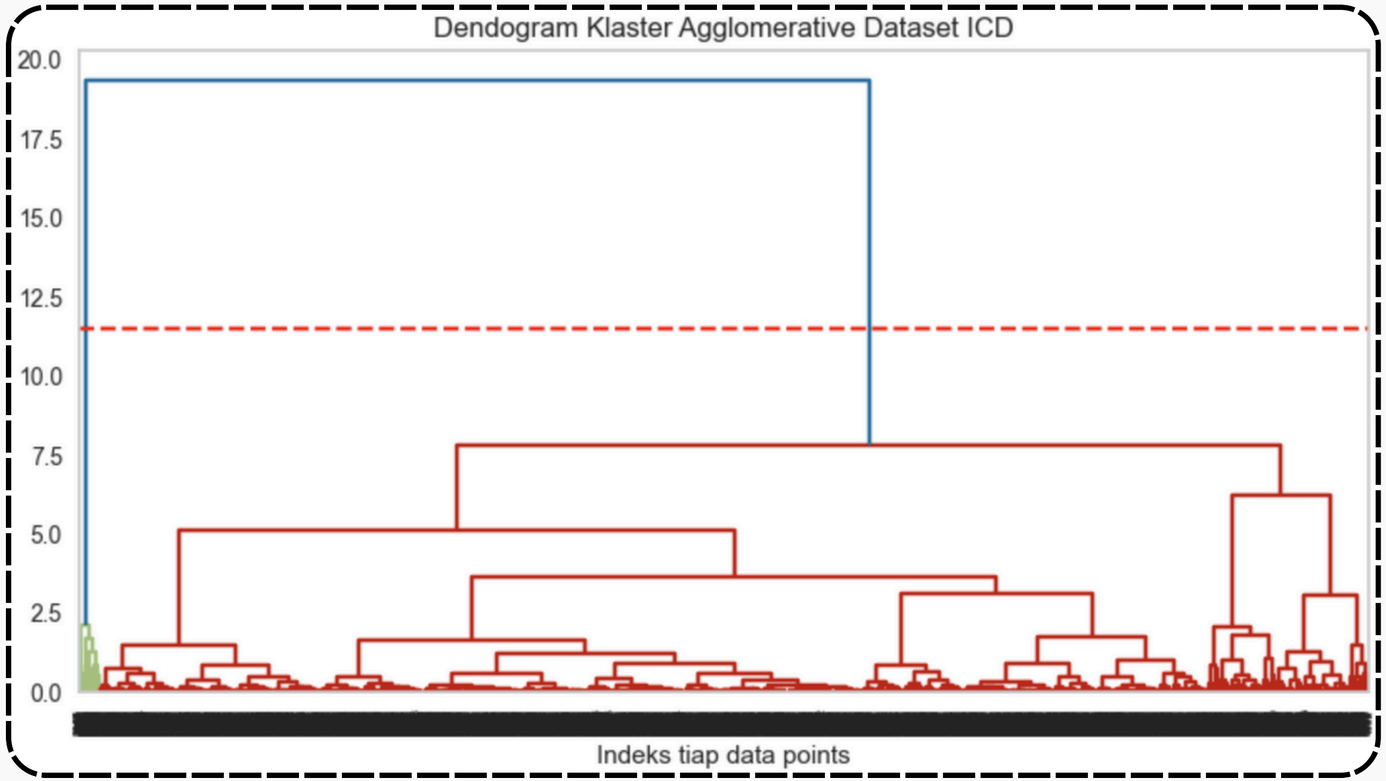


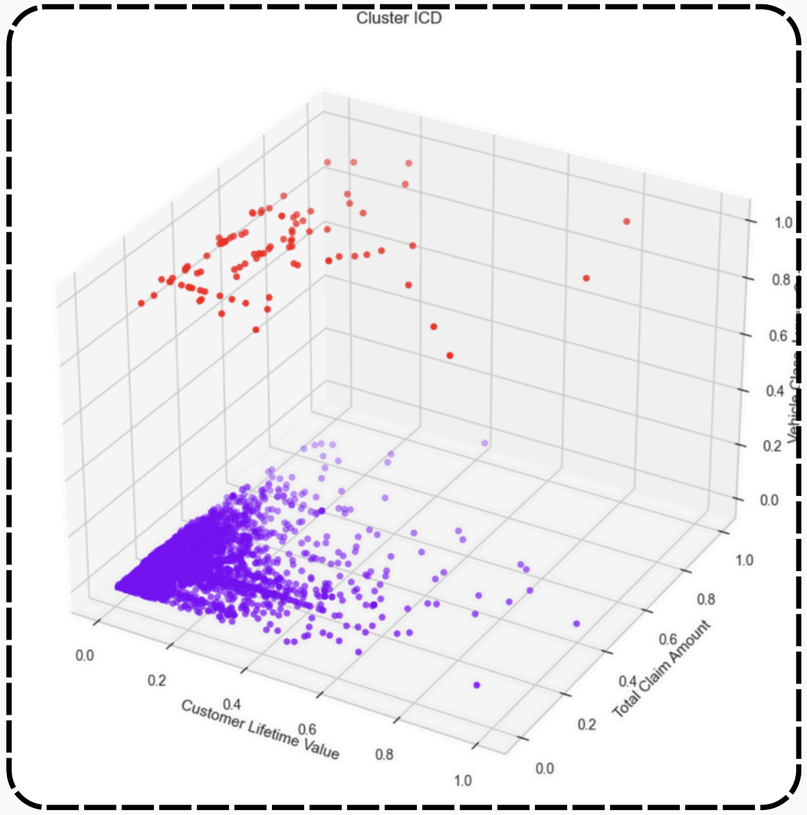
Figure: Visualisasi *Agglomerative Tree* dengan *method* Ward

- Dapat dilihat bahwa garis vertikal terpanjangnya ada pada yang berwarna biru dan memang berbeda jauh dengan panjang garis berwarna merah dan seterusnya

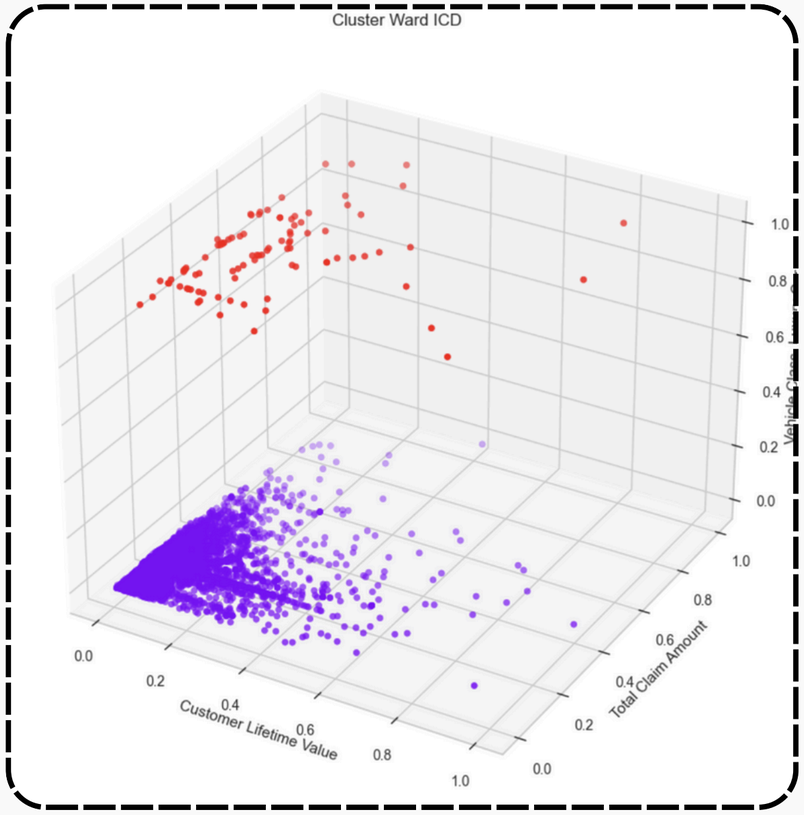
- Jadi, nilai k yang optimal adalah 2 karena dapat membagi data menjadi grup yang paling signifikan secara hierarchical

Hasil Clustering dengan k=2

K-Means



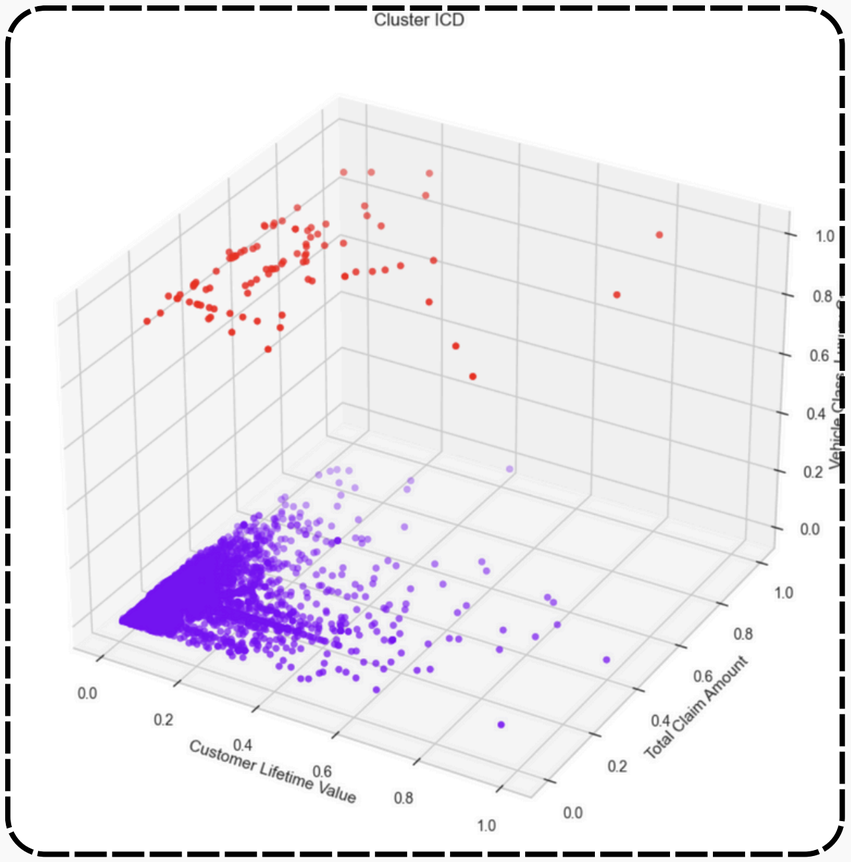
Agglomerative - Ward



- Dapat dilihat bahwa terbentuk 2 cluster (ungu dan merah) yang terbagi secara optimal
- Masing-masing cluster merepresentasikan 2 kelompok orang yang berbeda

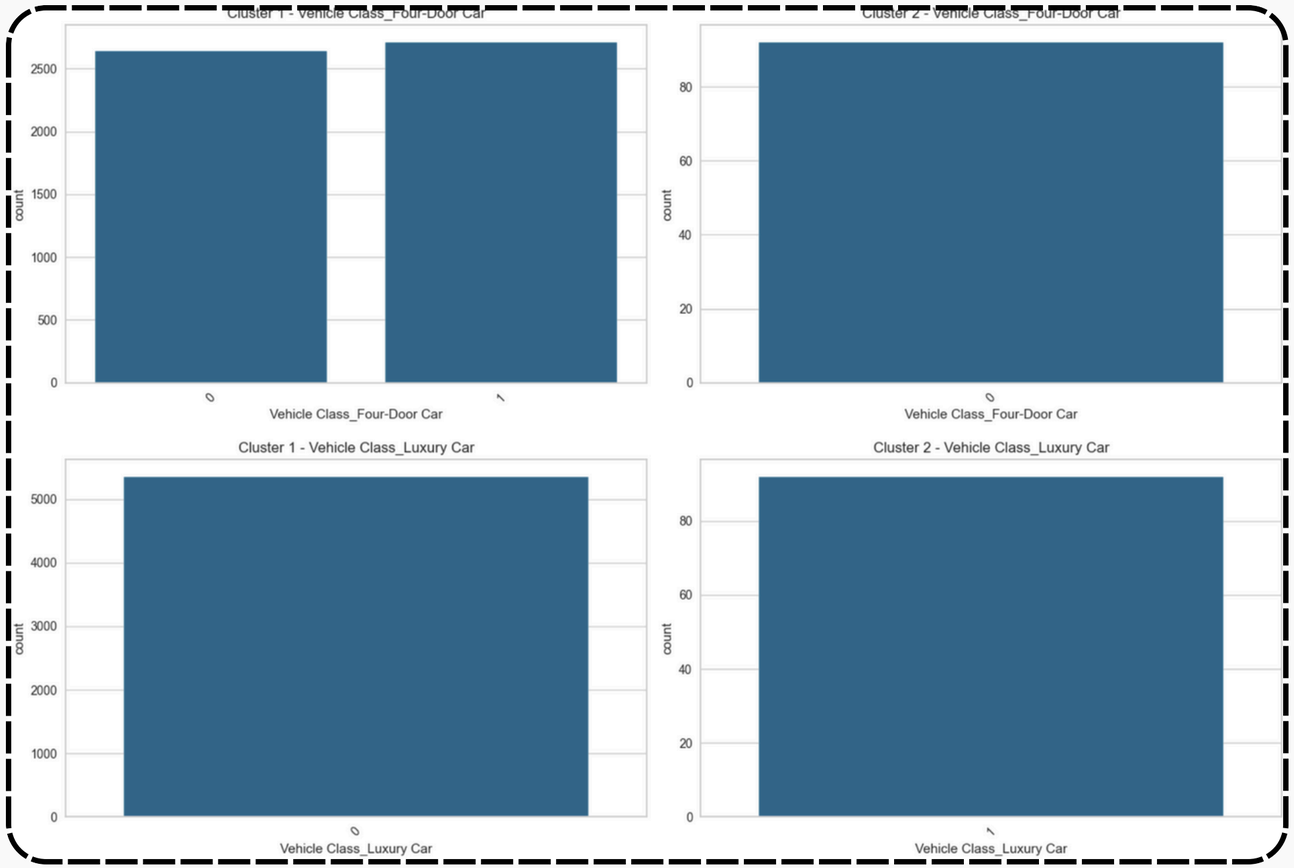
Figure: Perbandingan Clustering K-Means dengan Agglomerative

Analysis Clustering



- **Distribusi Data:**
  - Cluster Ungu: Jumlah titik data lebih banyak, tersebar di area dengan nilai **Customer Lifetime Value** dan **Total Claim Amount** yang lebih rendah.
  - Cluster Merah: Jumlah titik data lebih sedikit, terkonsentrasi di area dengan nilai **Customer Lifetime Value** dan **Total Claim Amount** yang lebih tinggi.
- **Interpretasi Cluster:**
  - Cluster Ungu: Memiliki **Customer Lifetime Value** dan **Total Claim Amount** yang rendah. Kemungkinan besar mereka merupakan pelanggan dengan cakupan asuransi dasar atau premi bulanan yang relatif rendah.
  - Cluster Merah: Memiliki **Customer Lifetime Value** yang tinggi dan mengajukan klaim dengan jumlah yang lebih besar. Kemungkinan besar merupakan pelanggan yang memiliki cakupan asuransi lebih luas atau menggunakan polis dengan premi yang lebih tinggi.

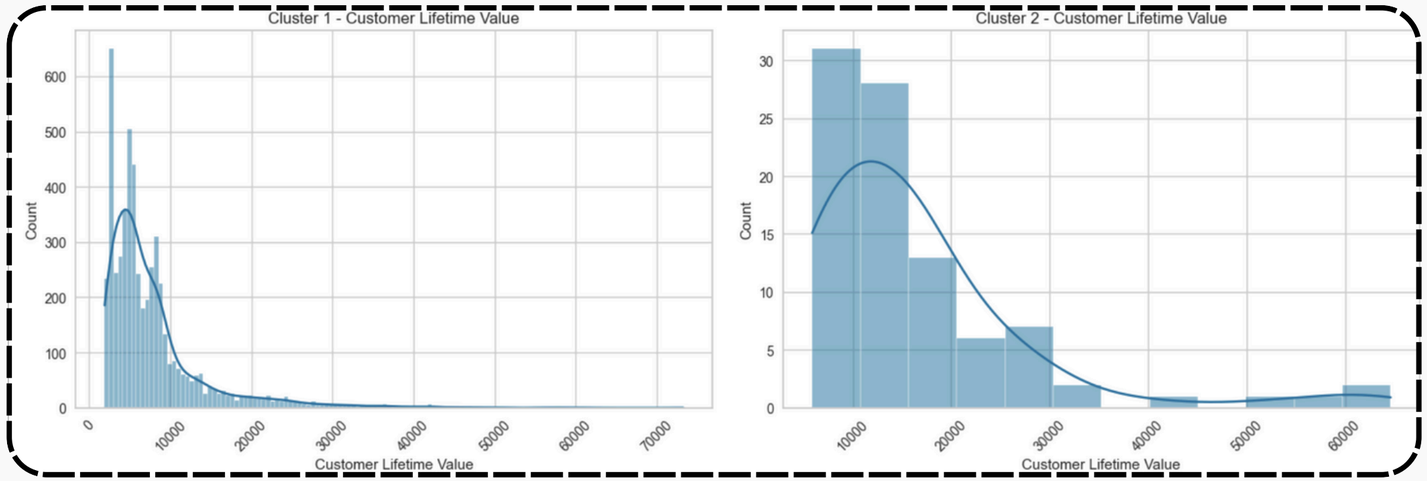
Analysis Clustering



Secara umum, hasil clustering ini menunjukkan bahwa pelanggan-pelanggan di **cluster 1** memiliki kendaraan yang lebih bervariasi karena **separuhnya menggunakan 4 *door car*** (secara umum bukan yang mewah) dan separuh lainnya menggunakan kendaraan lainnya (yang **bukan *luxury* car**).

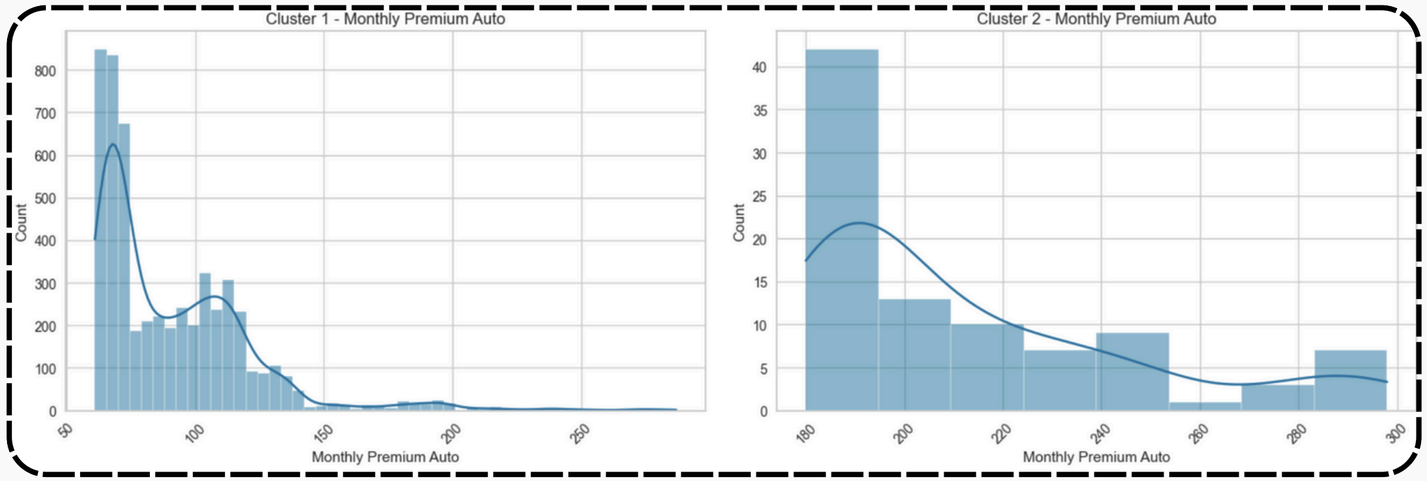
Di sisi lain, pelanggan-pelanggan di **cluster 2** cenderung menggunakan **kendaraan yang mewah/*luxury*** dan tidak ada yang menggunakan kendaraan 4 *door car* (secara umum bukan yang mewah).

Analysis Clustering



Customer Lifetime Value (CLV):

- **Di cluster 1**: mayoritas pelanggan memiliki CLV < 20.000 yang menunjukkan bahwa pelanggan di cluster ini cenderung memberikan nilai rendah bagi perusahaan
- **Di cluster 2**: pelanggan memiliki CLV yang lebih bervariasi dan tinggi yang menunjukkan bahwa pelanggan di cluster ini cenderung memberikan nilai tinggi bagi perusahaan



Monthly Premium Auto (MPA):

- **Di cluster 1**: mayoritas pelanggan memiliki MPA < 100 yang menunjukkan bahwa pelanggan di cluster ini lebih memilih paket dengan biaya bulanan yang lebih terjangkau
- **Di cluster 2**: pelanggan memiliki MPA yang lebih bervariasi dan tinggi yang menunjukkan bahwa pelanggan di cluster ini cenderung memilih paket dengan biaya bulanan yang lebih tinggi dan cakupan yang lebih luas

Any Question?

You NP-Complete Me

Thank you