IBM Watson Marketing

Customer Value Analysis

IBM Watson Analytics

By Team 5

Aishwarya Panse
Farin Fukunaga
Jann Ang
Maggie Ding
Queenie Chao

IBM Watson

ABOUT OUR DATASET

Introduction

Purpose of this dataset

To predict customer behavior by examining all the relevant customer data of a car insurance company and create targeted customer retention strategies, optimizing renewal strategies, and maximizing customer lifetime value.

Origin of our dataset

IBM Official Website

Dataset released date

2018

Business Problems & Objectives

1 **New Customer Acquisition:**

- Which customers are the most valuable (in terms of CLV)?
- How can we identify potential valuable customers?

2 **Renewal and Churn Prevention:**

- How do we drive policy renewals?
- What renewal marketing efforts are effective?

TOOLS & TECHNIQUES

Our Methodology

All 9134 entries were examined

- Converted categorical variables such as Gender and Marital Status into dummy variables to facilitate inclusion in regression models

Model Selection & Validation

- Chose semi-log and logistic regression for their ability to model non-linear relationships and binary outcomes, respectively

Analytical Tools

- R Studio features (glm packages for model building and evaluation)
- Anaconda (data manipulation and visualization, employing packages like pandas, NumPy, seaborn and matplotlib)

Visualization Techniques

- Data visualizations (histograms, Q-Q plots, and scatter plots)
- Interactive visualizations (Python's Bokeh and Plotly)

```
df.isnull().sum()
```

```
Customer                         0
State                            0
Customer Lifetime Value          0
Response                         0
Coverage                         0
Education                        0
Effective To Date                0
EmploymentStatus                 0
Gender                           0
Income                           0
Location Code                    0
Marital Status                   0
Monthly Premium Auto             0
Months Since Last Claim          0
Months Since Policy Inception    0
Number of Open Complaints        0
Number of Policies               0
Policy Type                      0
Policy                           0
Renew Offer Type                 0
Sales Channel                    0
Total Claim Amount               0
Vehicle Class                    0
Vehicle Size                     0
dtype: int64
```

IBM Watson

Data Summary

Our Dependent Variables

Variable	Type	Description
CLV (Customer Lifetime Value)	Numerical	The total revenue the car insurance company can expect to bring in from the customer if he/she remains a client
Respond	Categorical	Whether the customer has responded to the marketing calls

Predicted Variables: Factors that can affect insurance premium and reflect insure behavior pattern

Demographics & Basic Information

Car Plate Number, Gender, Education Level, Employment Status, Income, Marital Status, Living State, Type of Living Area

Insure Behavior

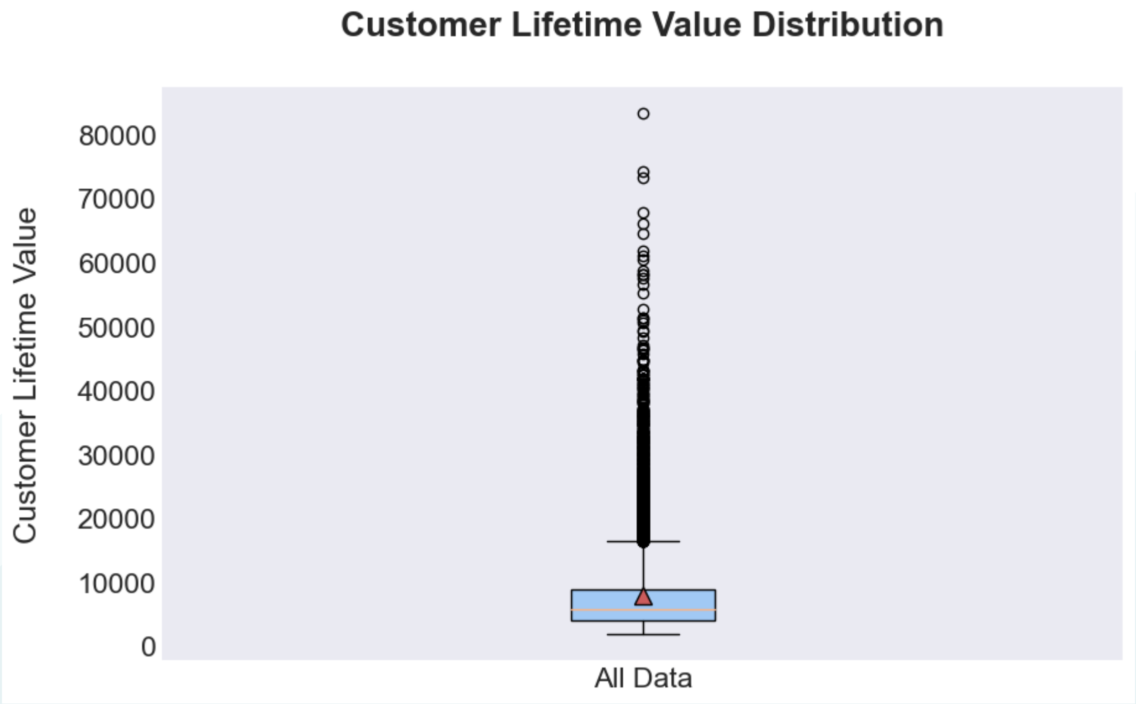
Vehicle Class & Size, Insurance Coverage Level, Monthly Premium, Sales Channel, Renew Offer, Number of Open Complaints

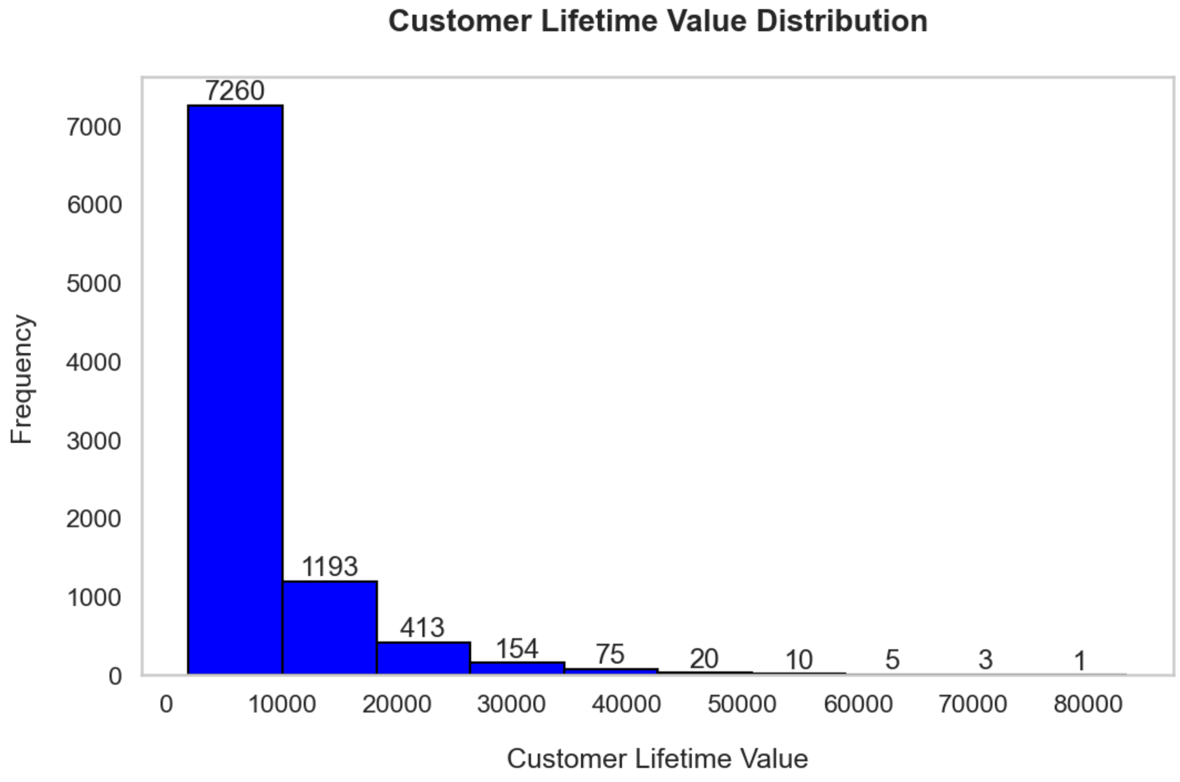
Policy Record

How long since last claim, Number of Policies, Policy & Policy Type (Purpose of vehicle use), Total Claim Amount, Duration Since Policy Inception

Data Visualization (CLV)

The sample has the median CLV of around $7,000 and the mean CLV of around $10,000. Over 80% customers have CLV under $10,000.





Predicting CLV

USING SEMI-LOG REGRESSION MODEL

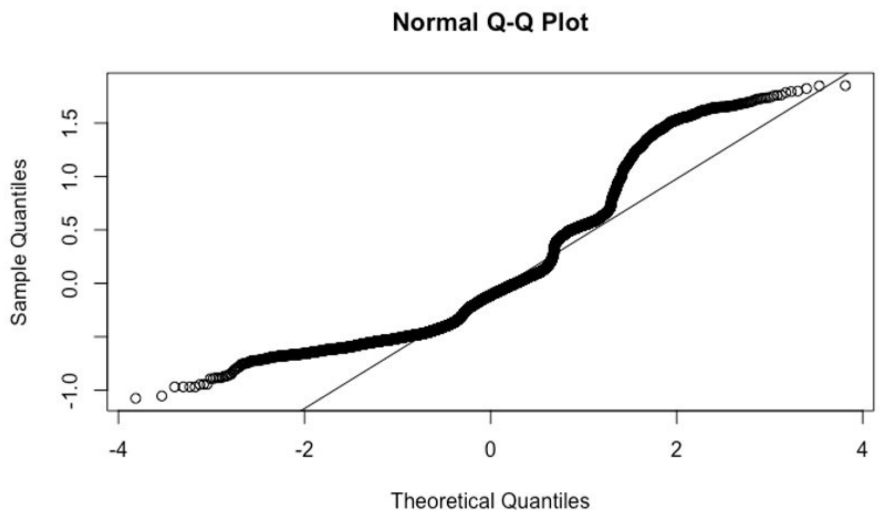
- Y-Variable: Log(Customer Lifetime Value)
- X-Variables are as follows:

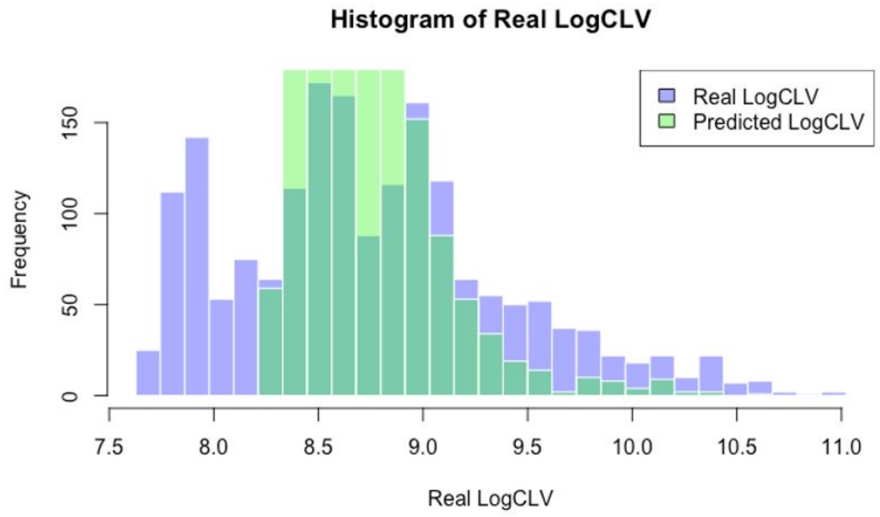
Variable Names	Variable Type	Note
Coverage	Categorical	Basic, Extended, Premium
Employment Status	Categorical	Employed, Unemployed, Retired, Disabled, Medical Leave
Monthly Premium Auto	Numerical	
Number of Open Complaints	Numerical	
Number of Policies	Numerical	Personal, Corporate, Special
Policy	Numerical	
Renew Offer	Categorical	Offer1, Offer2, Offer3, Offer4

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.0757 -0.4592 -0.1114  0.2659  1.8509

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               7.9352335  0.0254911 311.295  < 2e-16 ***
Coverage_Ext              0.0431497  0.0154534   2.792  0.00525 **
Coverage_Pre              0.0239962  0.0259835   0.924  0.35577
Employ_Unemp             -0.1195313  0.0158453  -7.544 5.12e-14 ***
Employ_Dis               -0.0668636  0.0329290  -2.031  0.04234 *
Employ_Med               -0.0753617  0.0317513  -2.373  0.01765 *
Employ_Ret               -0.1183461  0.0388741  -3.044  0.00234 **
Monthly.Premium.Auto      0.0081876  0.0002166  37.801  < 2e-16 ***
Number.of.Open.Complaints -0.0286805 0.0072762  -3.942 8.17e-05 ***
Number.of.Policies        0.0522081  0.0027538  18.959  < 2e-16 ***
Policy_Cor               -0.0241376  0.0164113  -1.471  0.14139
Policy_Spe                0.0912785  0.0318373   2.867  0.00416 **
Renew_02                 -0.1299291  0.0160120  -8.114 5.68e-16 ***
Renew_03                 -0.0679791  0.0198515  -3.424  0.00062 ***
Renew_04                 -0.1471387  0.0224736  -6.547 6.26e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.5647 on 7306 degrees of freedom
Multiple R-squared:  0.2574,    Adjusted R-squared:  0.256
F-statistic: 180.9 on 14 and 7306 DF,  p-value: < 2.2e-16
```

IBM Watson

The Results





Diagnosis

The Central Limit Value (CLV) in the middle is accurately predicted. However, the model tends to underestimate in extreme values.

Key Findings

1. Employment status is also a significant contributor to CLV. When a customer is employed, CLV tends to increase.
2. The number of policies has a positive impact on CLV since the coefficient is positive. Therefore, a greater focus should be placed on customers with a higher number of policies.
3. The Renew Offer plays a pivotal role in influencing CLV. The coefficients for Renew_O2 to Renew_O4 are all negative. Adjusting Renew Offer 1 positively contributes to increasing CLV.

IBM Watson

Predicting Response Rate

MODEL (TRAINING)

```
Call:
glm(formula = response_binary ~ EmploymentStatus + Renew.Offer.Type +
    Sales.Channel + Education, family = binomial(link = "logit"),
    data = data.train)

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -1.20168    0.16581  -7.247 4.26e-13 ***
EmploymentStatusEmployed      -0.43865    0.15712  -2.792  0.00524 **
EmploymentStatusMedical Leave  0.02990    0.20624   0.145  0.88473
EmploymentStatusRetired        2.66542    0.22405  11.897  < 2e-16 ***
EmploymentStatusUnemployed    -0.91732    0.17222  -5.326 1.00e-07 ***
Renew.Offer.TypeOffer2         0.67095    0.07867   8.528  < 2e-16 ***
Renew.Offer.TypeOffer3        -2.10124    0.21631  -9.714  < 2e-16 ***
Renew.Offer.TypeOffer4       -16.78815  222.92553  -0.075  0.93997
Sales.ChannelBranch           -0.66737    0.09390  -7.107 1.19e-12 ***
Sales.ChannelCall Center      -0.48808    0.10430  -4.679 2.88e-06 ***
Sales.ChannelWeb              -0.55519    0.11843  -4.688 2.76e-06 ***
EducationCollege               0.16489    0.09492   1.737  0.08237 .
EducationDoctor                0.48699    0.18139   2.685  0.00726 **
EducationHigh School or Below -0.04270    0.09850  -0.434  0.66465
EducationMaster                0.23962    0.14446   1.659  0.09716 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6016.2  on 7320  degrees of freedom
Residual deviance: 4850.9  on 7306  degrees of freedom
AIC: 4880.9

Number of Fisher Scoring iterations: 17
```

- **Using logistic regression model on the Training list**

- **Y dependent variable:**
  - Response (Binary)

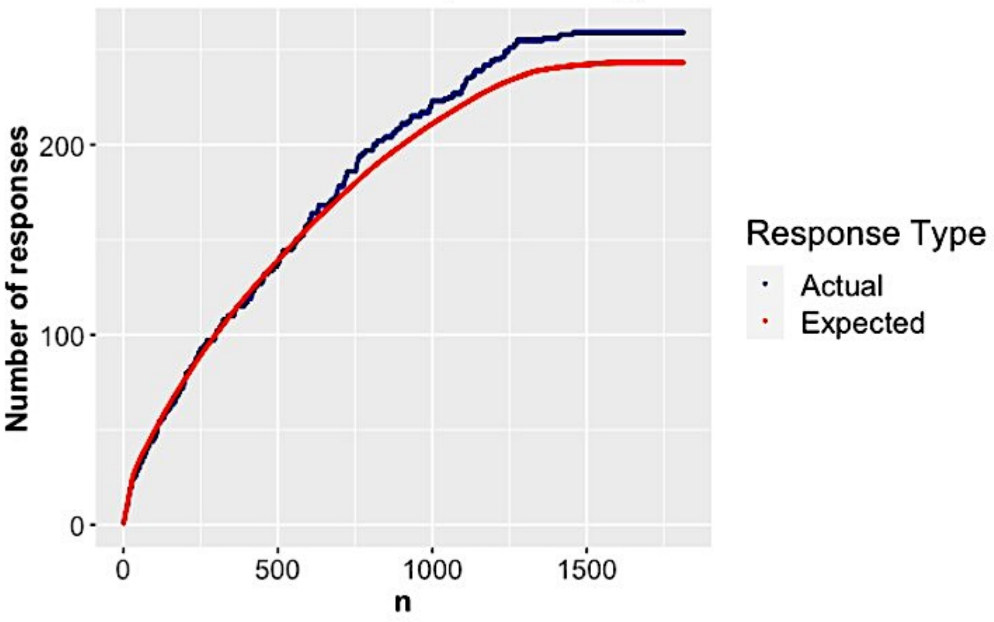
- **4 X independent variables:**
  - Employment Status – Employed, Unemployed, Retired, etc.
  - 4 Renew Offer Type – Offer 1, Offer 2, Offer 3, Offer 4
  - Sales Channel – Agent, Branch, Call Center, or Web
  - Education – High School, Bachelor, Masters, etc.

- **Coefficients Interpretation**
  - E.g. EmploymentStatusRetired has a positive coefficient, suggesting that retired customers are more likely to renew

Quality of Fit

Predict	Actual 0	1	Row Total
0	1546	234	1780
1	8	25	33
Column Total	1554	259	1813

Expected and Actual Positive Responses vs. Number of Prospects Targeted



Confusion Matrix

- The model stuffers from false negatives where the model predicts a non-renewal but the customer actually renewed (234 cases).

Model Accuracy

- Predicted many true negatives correctly, which is good for identifying low-potential customers.

Lift Curve

- The model predicted the response rate well for the first ~1000 customers but becomes less accurate as it moves to customers with a lower propensity to respond.

Key Findings

- **Demographics:** Older people respond more often. Highly educated people respond more often.
- **Sales-Related:** Agent is the best channel. Offers 1 and 2 are the best.
- **Response rates do not differ by:** Location, Vehicle Type, Number of Claims, Type of Policy (Personal vs. Corporate vs Special), CLV

IBM Watson

Recommendation

Outcome from CLV

1. **Enhance focus on Employment Status**
   - Create tailored communication for employed individuals emphasizing the security a policy provides against potential income loss.

2. **Leverage Policy Count**
   - Include bundle offers or discounts for customers who hold or add multiple policies

3. **Streamline Policy Upgrades**
   - For customers who are likely to increase their CLV, make it easier to upgrade or add policies

Outcome from Response Rate

1. **Target Elderly Retired & Highly Educated Individuals Via Their Personal Agents**
   - Capitalizes on the personal touch that can be very effective with older demographics
   - Higher education levels often correlate with a greater understanding of the benefits and complexities of insurance policies

2. **Reevaluate Sales Channels (Branch, Call Center & Web) and Offers (3 and 4)**
   - Gather sentiment analysis on why these offers are not performing well could lead to more effective offer structuring
   - Conduct A/B Testing with variations to measure

IBM Watson

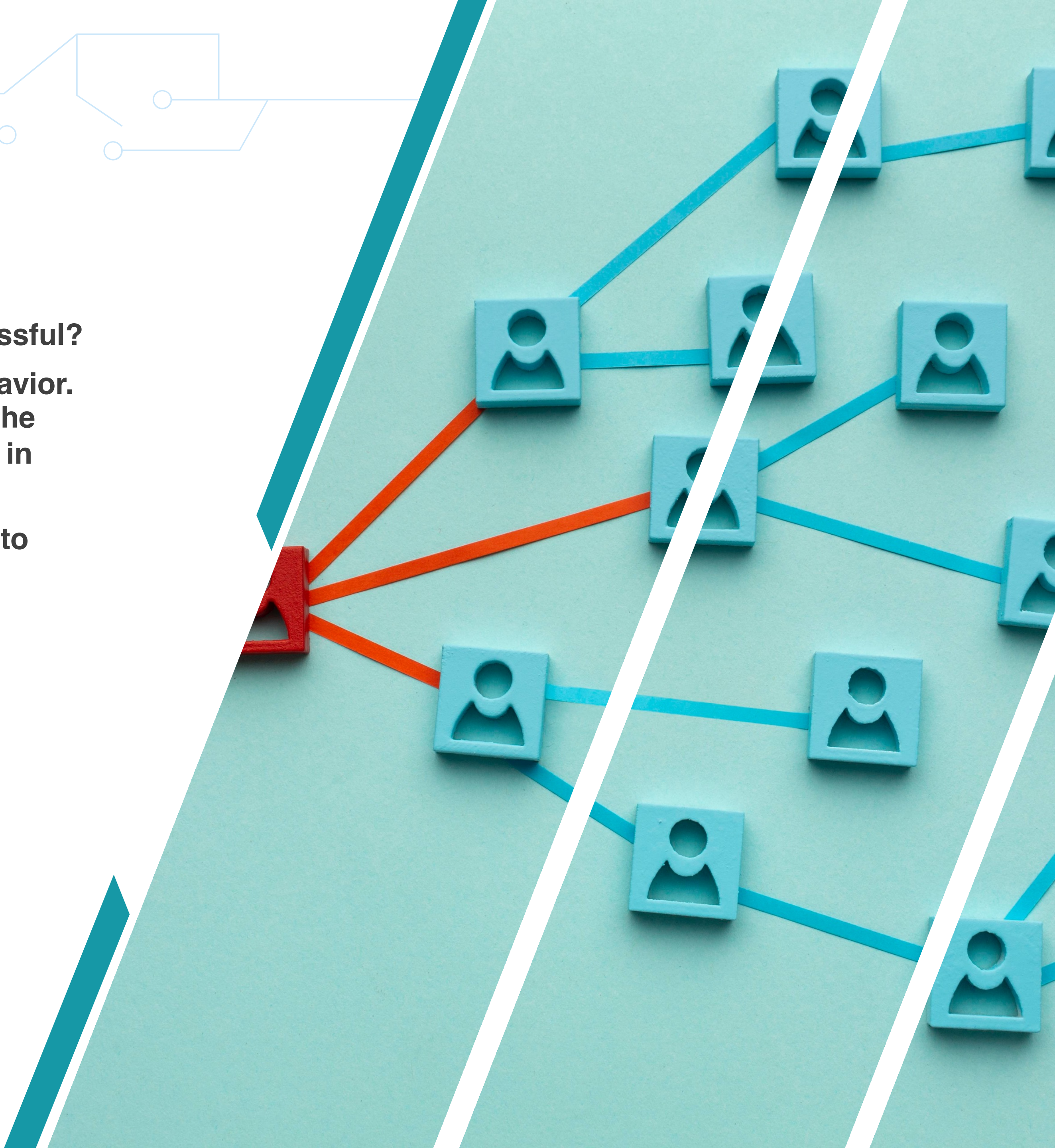
Future Analysis

1 Longitudinal Tracking

- Is there any evidence that the campaigns were successful?
- Compare initial predictions with actual customer behavior. This includes whether non-responsive customers in the dataset eventually renewed and if there was a real lift in response rates from campaigns.
- Make changes to advertising campaigns in response to patterns and outcomes seen over the long run.

2 Clustering & Micro-Segmentation

- How should we aggregate and target specific customer groups?
- Use clustering algorithms to identify distinct groups within the customer base based on a variety of factors beyond CLV, like behavior patterns, policy preferences, and demographic details.
- To maximize the efficacy of outreach, create cluster-specific marketing tactics.



IBM Watson

THANK YOU

FOR LISTENING

By Team 5

Appendix

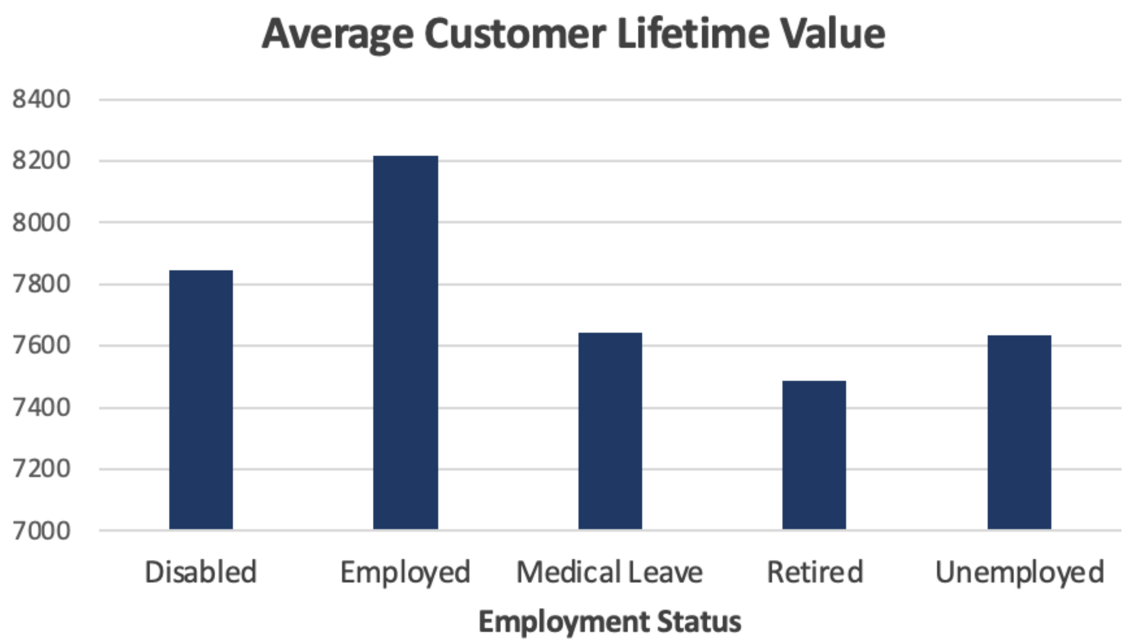
```
RangeIndex: 9134 entries, 0 to 9133
Data columns (total 24 columns):
 #   Column                         Non-Null Count  Dtype
---  ------                         --------------  -----
 0   Customer                       9134 non-null   object
 1   State                          9134 non-null   object
 2   Customer Lifetime Value        9134 non-null   float64
 3   Response                       9134 non-null   int64
 4   Coverage                       9134 non-null   object
 5   Education                      9134 non-null   object
 6   Effective To Date              9134 non-null   object
 7   EmploymentStatus               9134 non-null   object
 8   Gender                         9134 non-null   object
 9   Income                         9134 non-null   int64
 10  Location Code                  9134 non-null   object
 11  Marital Status                 9134 non-null   object
 12  Monthly Premium Auto           9134 non-null   int64
 13  Months Since Last Claim        9134 non-null   int64
 14  Months Since Policy Inception  9134 non-null   int64
 15  Number of Open Complaints      9134 non-null   int64
 16  Number of Policies             9134 non-null   int64
 17  Policy Type                    9134 non-null   object
 18  Policy                         9134 non-null   object
 19  Renew Offer Type               9134 non-null   object
 20  Sales Channel                  9134 non-null   object
 21  Total Claim Amount             9134 non-null   float64
 22  Vehicle Class                  9134 non-null   object
 23  Vehicle Size                   9134 non-null   object
dtypes: float64(2), int64(7), object(15)
memory usage: 1.7+ MB
```

In [120]: df.describe()

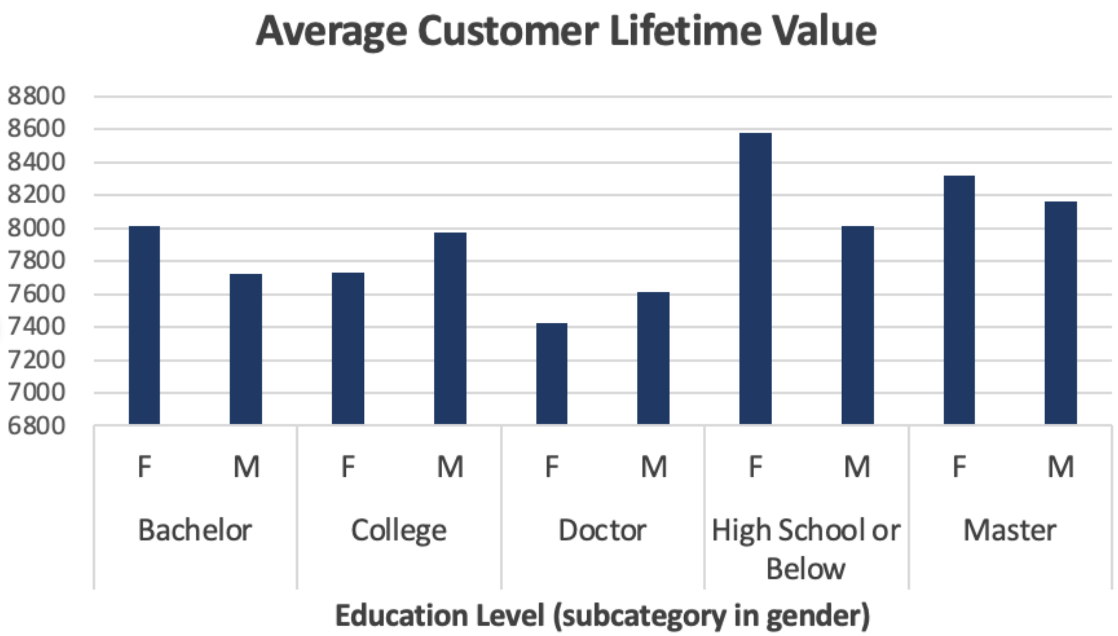
Out[120]:

	Customer Lifetime Value	Response	Income	Monthly Premium Auto	Months Since Last Claim	Months Since Policy Inception	Number of Open Complaints	Number of Policies	Total Claim Amount
count	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000	9134.000000
mean	8004.940475	0.143201	37657.380009	93.219291	15.097000	48.064594	0.384388	2.966170	434.088794
std	6870.967608	0.350297	30379.904734	34.407967	10.073257	27.905991	0.910384	2.390182	290.500092
min	1898.007675	0.000000	0.000000	61.000000	0.000000	0.000000	0.000000	1.000000	0.099007
25%	3994.251794	0.000000	0.000000	68.000000	6.000000	24.000000	0.000000	1.000000	272.258244
50%	5780.182197	0.000000	33889.500000	83.000000	14.000000	48.000000	0.000000	2.000000	383.945434
75%	8962.167041	0.000000	62320.000000	109.000000	23.000000	71.000000	0.000000	4.000000	547.514839
max	83325.381190	1.000000	99981.000000	298.000000	35.000000	99.000000	5.000000	9.000000	2893.239678

Appendix

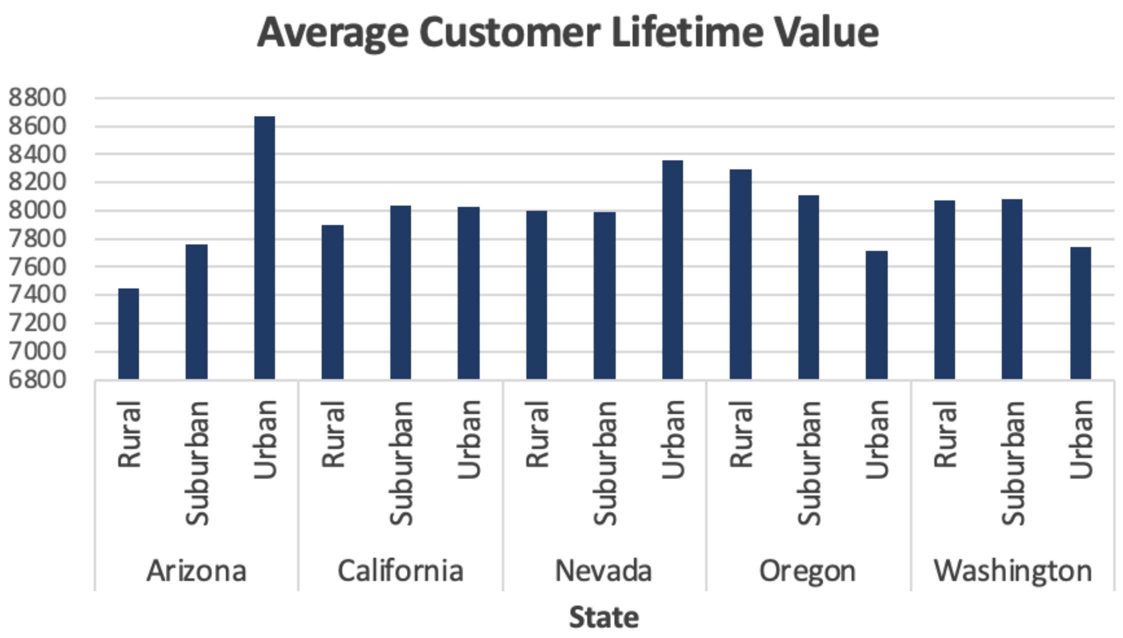


The clients who are employed are with higher consuming power and more auto use, leading to much higher CLV on average

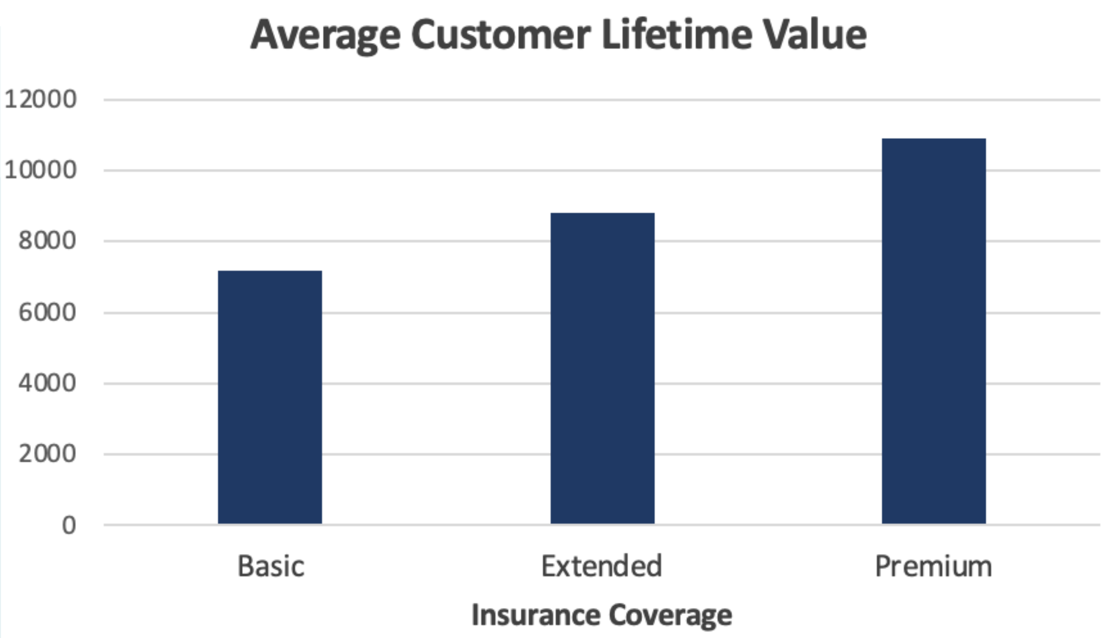


Female with education level of high school or below and education level of master have much higher CLV

Appendix

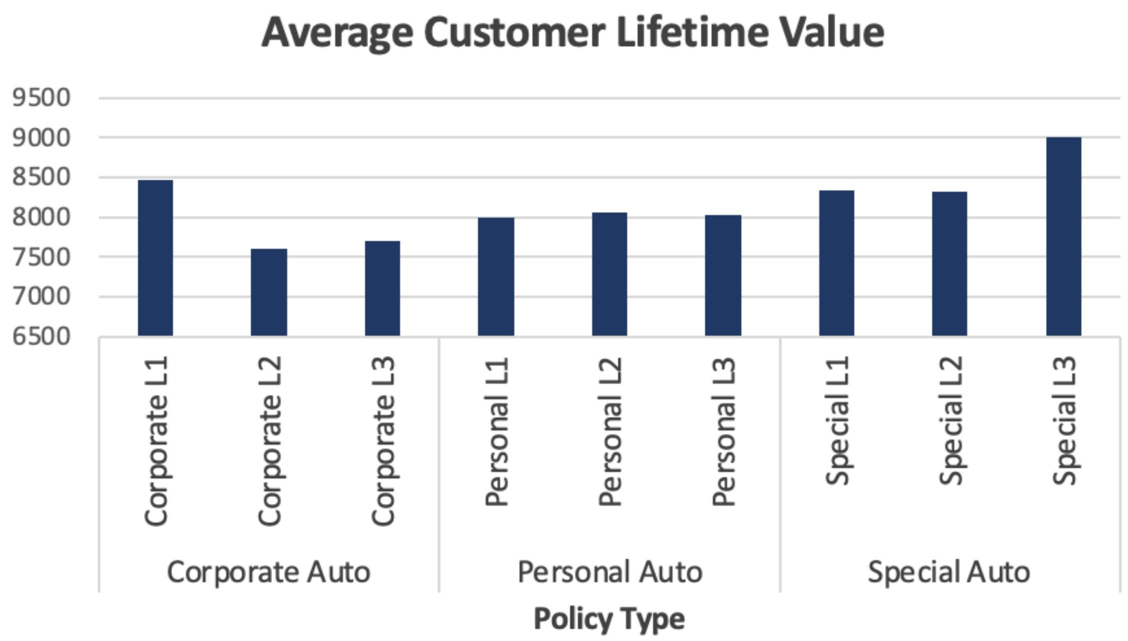


The clients living in Arizona and Nevada Urban have higher risk of accidents, leading to much higher CLV

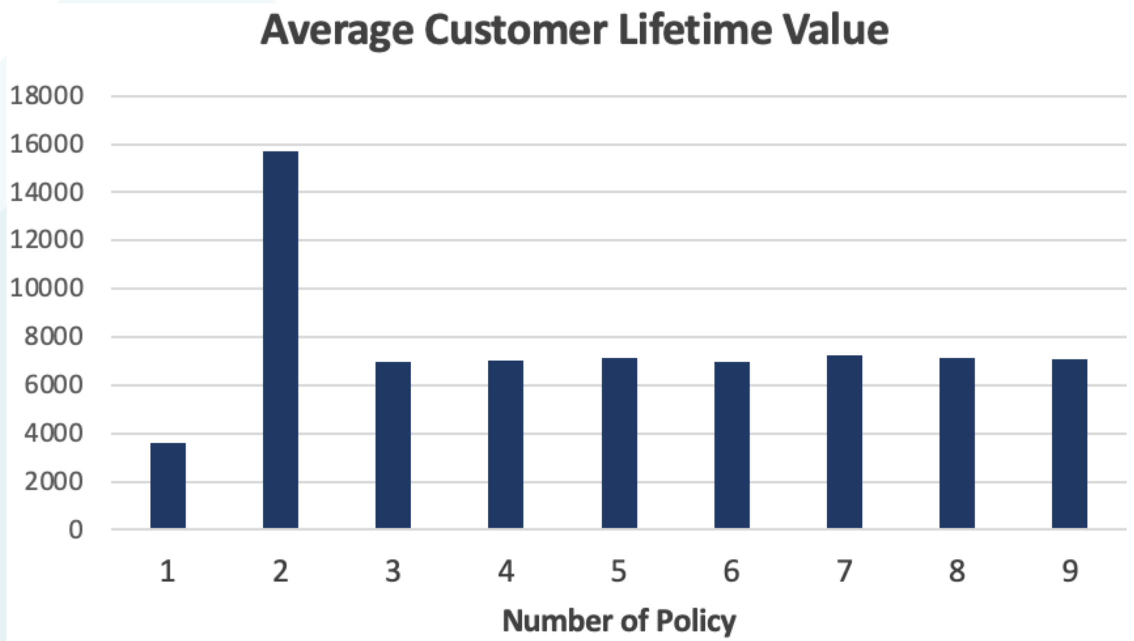


The clients purchasing premium insurance coverage have much higher CLV

Appendix

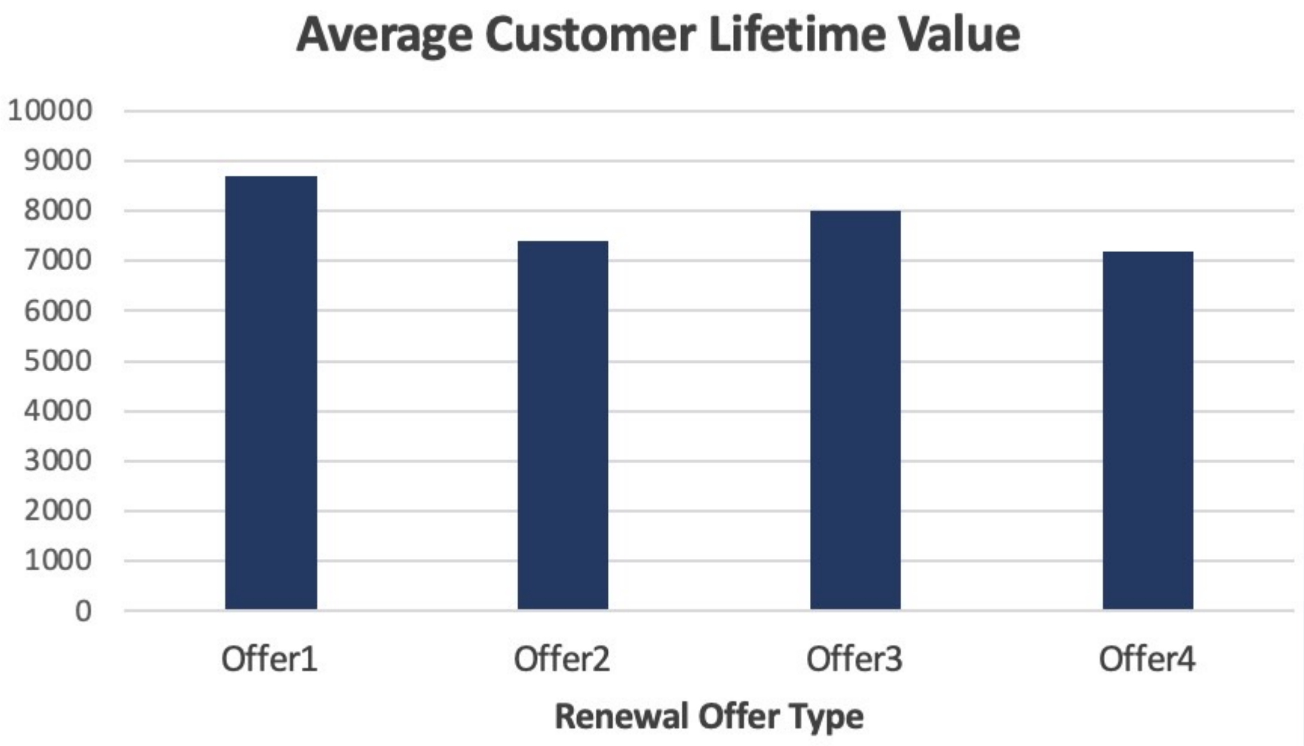


The clients using their vehicles for business or special purpose have much higher CLV



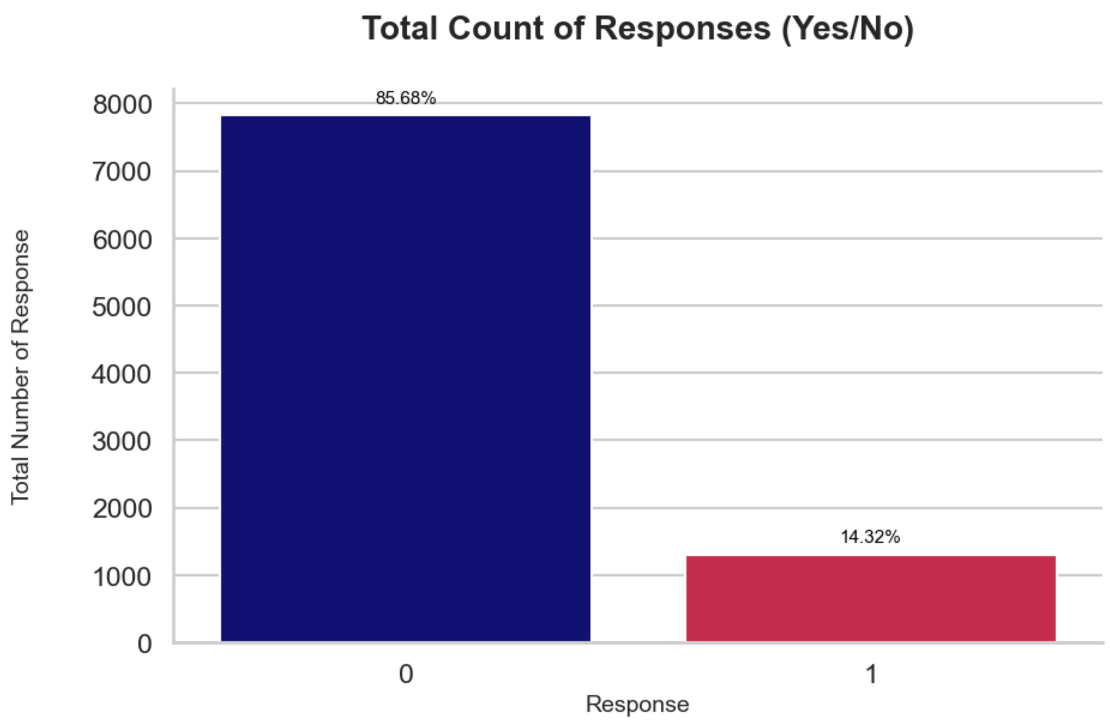
The clients having two policies have much higher CLV on average

Appendix

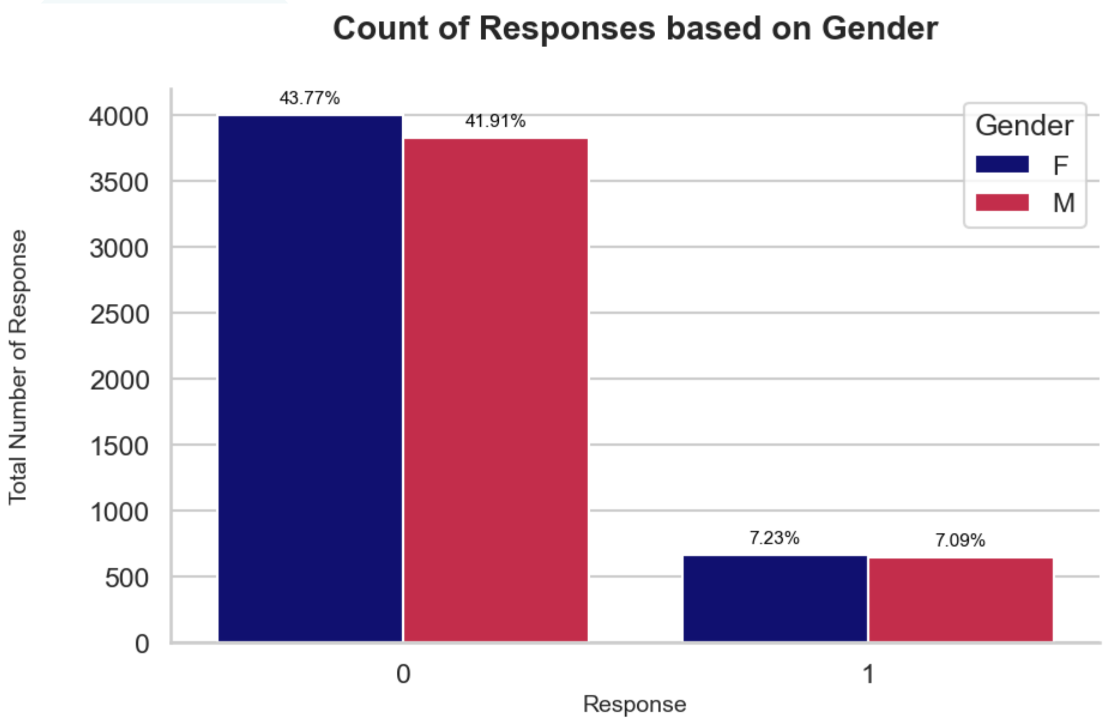


The clients of Offer 1 and 3 have higher CLV

Appendix

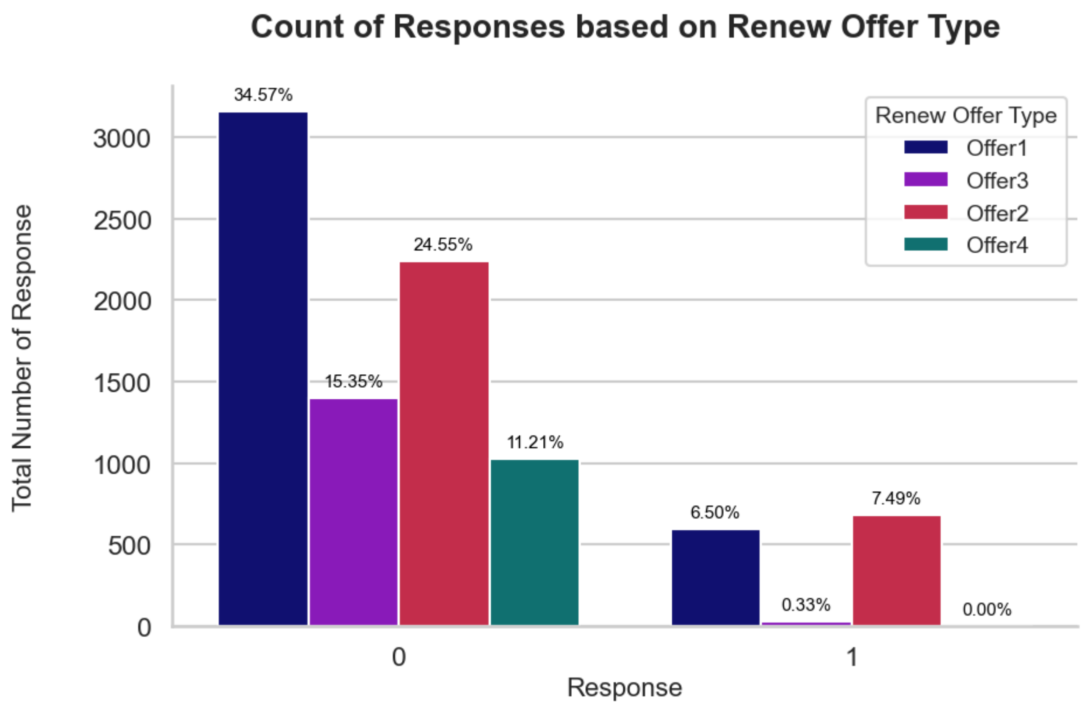


It is worth noting that approximately 14% of customers have replied to marketing calls, while the remaining 86% have not.

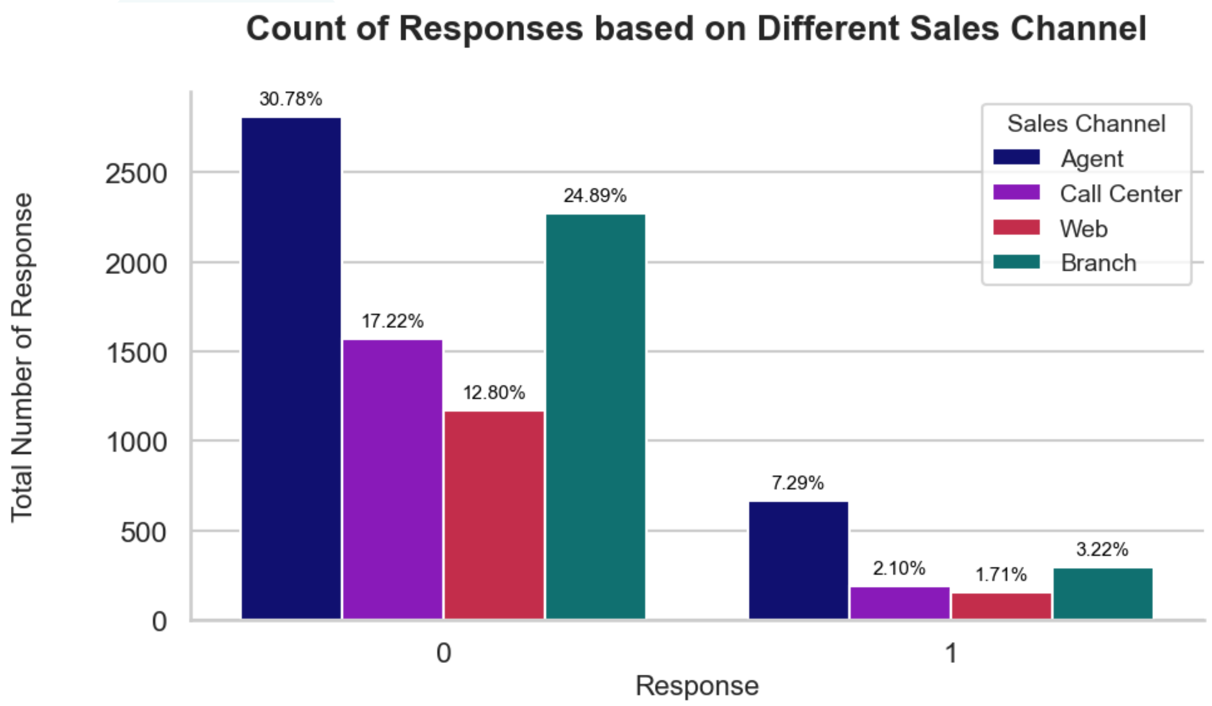


A marketing call will get nearly the same number of responses from males and females.

Appendix



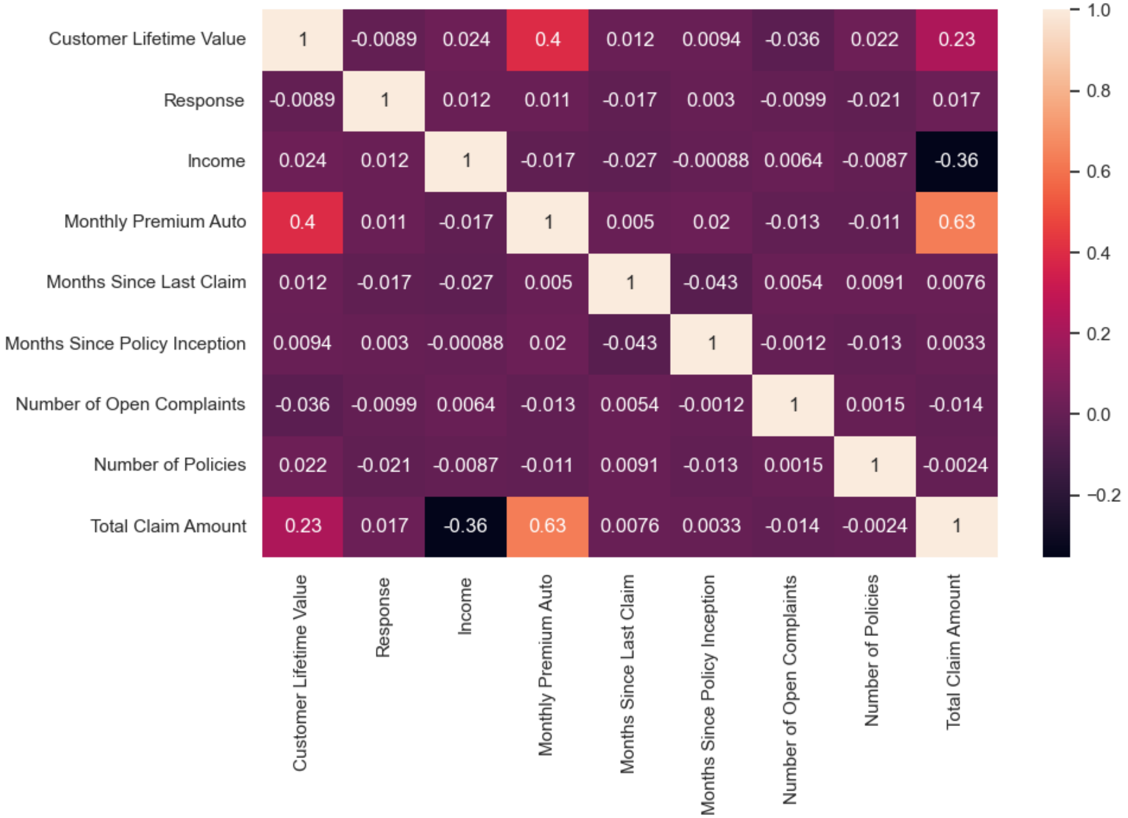
Customers have answered marketing calls for offers 1 and 2, but for offers 3 and 4, nearly no one has answered.



Response rate through sales agent garnered the highest response rate of 7.29%.

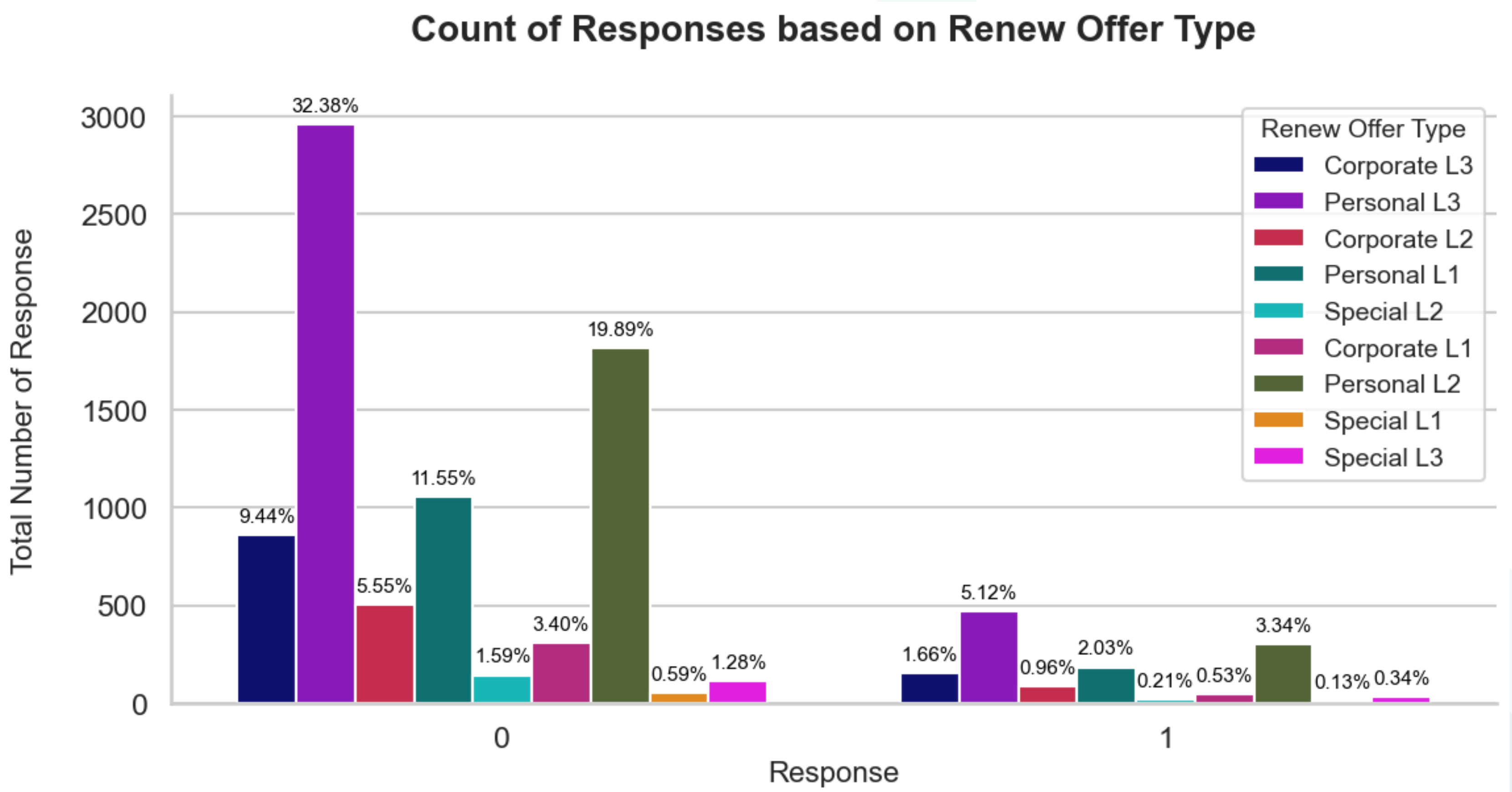
Appendix

Heatmap of a Correlation Matrix Using Continuous Variable Only



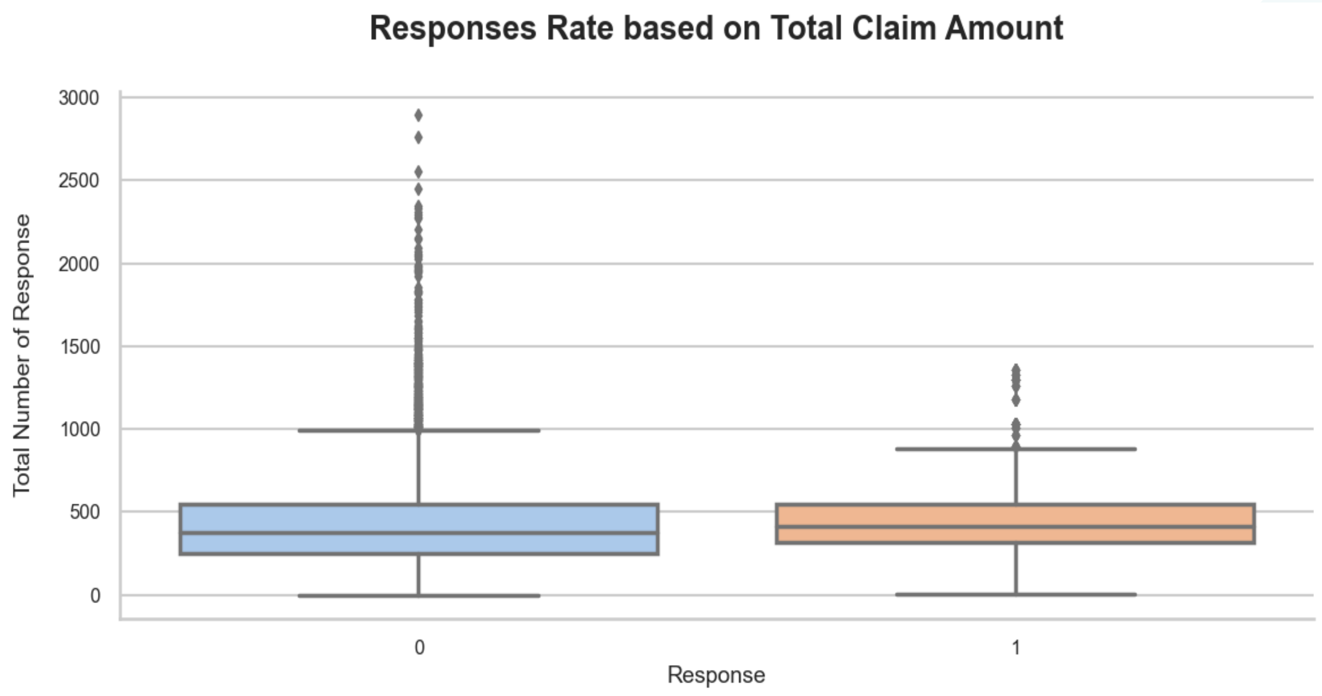
- **Monthly Premium Auto** and **Total Claim Amount** have a strong positive correlation (0.63), suggesting that as the monthly auto premium increases, the total claim amount tends to increase as well.
- **Income** and **Total Claim Amount** have a moderate negative correlation (-0.36), suggesting that higher income levels are associated with lower total claim amounts.
- **Customer Lifetime Value** and **Monthly Premium Auto** also show a positive correlation (0.4), implying that customers with higher lifetime values tend to pay higher monthly premiums.

Appendix



Count of Responses based on Renew Offer Type
Total Number of Response
3000
2500
2000
1500
1000
500
0
Response
0
1
32.38%
9.44%
5.55%
11.55%
1.59%
3.40%
19.89%
0.59%
1.28%
1.66%
5.12%
0.96%
2.03%
0.21%
0.53%
3.34%
0.13%
0.34%
Renew Offer Type
Corporate L3
Personal L3
Corporate L2
Personal L1
Special L2
Corporate L1
Personal L2
Special L1
Special L3

Appendix



Response '0':

- Has a higher median Total Claim Amount compared to Response '1'.
- Displays a wider interquartile range, indicating more variability in the Total Claim Amount.
- Has a longer upper whisker and more outliers, suggesting that there are more claims with higher amounts in this category.

Response '1':

- Has a lower median, indicating that the central tendency of claims is less than that of Response '0'.
- The interquartile range is narrower, suggesting less variability in the Total Claim Amount.
- There are fewer outliers, indicating fewer extreme claim amounts.