

I. Dataset Preprocessing __ Harriet Onoriode Otomiewor (23103939)

2023-08-19

Read data

```
setwd('C:/Users/tinhl/OneDrive/Documents')
data <- read.csv(file = 'WA_Fn-UseC_-Marketing-Customer-Value-Analysis.csv')
head(data)
```

```
## Customer State Customer_Lifetime_Value Response Coverage Education
## 1 BU79786 Washington 2763.519 No Basic Bachelor
## 2 QZ44356 Arizona 6979.536 No Extended Bachelor
## 3 AI49188 Nevada 12887.432 No Premium Bachelor
## 4 WW63253 California 7645.862 No Basic Bachelor
## 5 HB64268 Washington 2813.693 No Basic Bachelor
## 6 OC83172 Oregon 8256.298 Yes Basic Bachelor
## Effective_To_Date EmploymentStatus Gender Income Location_Code Marital_Status
## 1 2/24/2011 Employed F 56274 Suburban Married
## 2 1/31/2011 Unemployed F 0 Suburban Single
## 3 2/19/2011 Employed F 48767 Suburban Married
## 4 1/20/2011 Unemployed M 0 Suburban Married
## 5 2/3/2011 Employed M 43836 Rural Single
## 6 1/25/2011 Employed F 62902 Rural Married
## Monthly_Premium_Auto Months_Since_Last_Claim Months_Since_Policy_Inception
## 1 69 32 5
## 2 94 13 42
## 3 108 18 38
## 4 106 18 65
## 5 73 12 44
## 6 69 14 94
## Number_of_Open_Complaints Number_of_Policies Policy_Type Policy
## 1 0 1 Corporate Auto Corporate L3
## 2 0 8 Personal Auto Personal L3
## 3 0 2 Personal Auto Personal L3
## 4 0 7 Corporate Auto Corporate L2
## 5 0 1 Personal Auto Personal L1
## 6 0 2 Personal Auto Personal L3
## Renew_Offer_Type Sales_Channel Total_Claim_Amount Vehicle_Class Vehicle_Size
## 1 Offer1 Agent 384.8111 Two-Door Car Medsize
## 2 Offer3 Agent 1131.4649 Four-Door Car Medsize
## 3 Offer1 Agent 566.4722 Two-Door Car Medsize
## 4 Offer1 Call Center 529.8813 SUV Medsize
## 5 Offer1 Agent 138.1309 Four-Door Car Medsize
## 6 Offer2 Web 159.3830 Two-Door Car Medsize
```

```
summary(data)
```

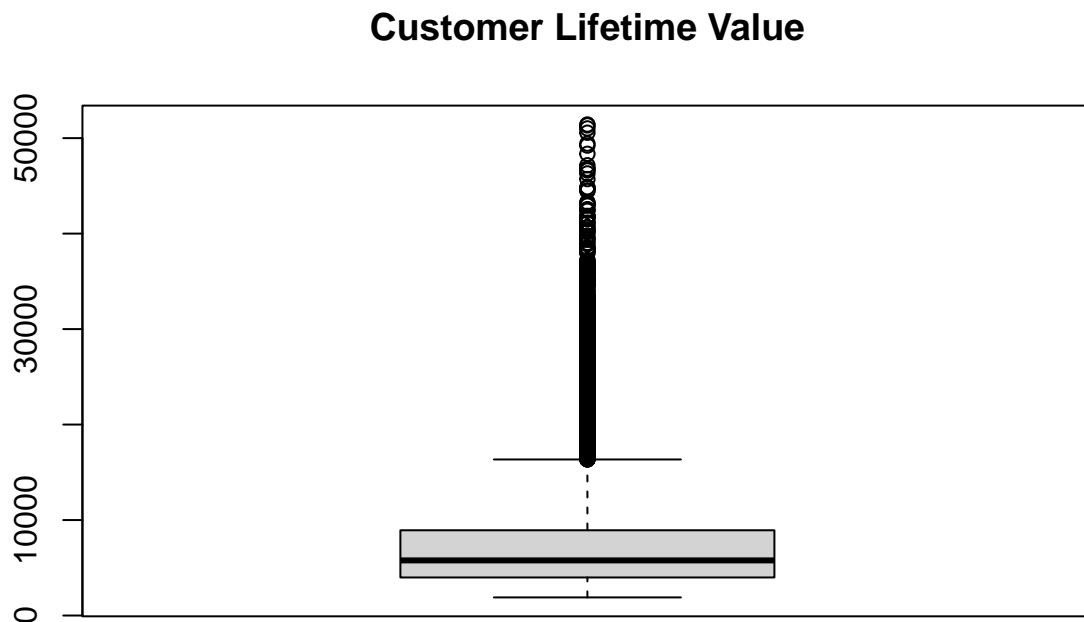
```
##      Customer          State      Customer_Lifetime_Value
## Length:9134      Length:9134      Min.   : 1898
## Class :character  Class :character  1st Qu.: 3994
## Mode  :character  Mode  :character  Median : 5780
##                                     Mean  : 8005
##                                     3rd Qu.: 8962
##                                     Max.   :83325
##      Response          Coverage      Education      Effective_To_Date
## Length:9134      Length:9134      Length:9134      Length:9134
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##      EmploymentStatus      Gender          Income      Location_Code
## Length:9134      Length:9134      Min.   :    0      Length:9134
## Class :character  Class :character  1st Qu.:    0      Class :character
## Mode  :character  Mode  :character  Median :33890      Mode  :character
##                                     Mean  :37657
##                                     3rd Qu.:62320
##                                     Max.   :99981
##      Marital_Status      Monthly_Premium_Auto      Months_Since_Last_Claim
## Length:9134      Min.   : 61.00      Min.   : 0.0
## Class :character  1st Qu.: 68.00      1st Qu.: 6.0
## Mode  :character  Median : 83.00      Median :14.0
##                                     Mean  : 93.22      Mean  :15.1
##                                     3rd Qu.:109.00      3rd Qu.:23.0
##                                     Max.   :298.00      Max.   :35.0
##      Months_Since_Policy_Inception      Number_of_Open_Complaints      Number_of_Policies
## Min.   : 0.00      Min.   :0.0000      Min.   :1.000
## 1st Qu.:24.00      1st Qu.:0.0000      1st Qu.:1.000
## Median :48.00      Median :0.0000      Median :2.000
## Mean   :48.06      Mean   :0.3844      Mean   :2.966
## 3rd Qu.:71.00      3rd Qu.:0.0000      3rd Qu.:4.000
## Max.   :99.00      Max.   :5.0000      Max.   :9.000
##      Policy_Type          Policy      Renew_Offer_Type      Sales_Channel
## Length:9134      Length:9134      Length:9134      Length:9134
## Class :character  Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character  Mode  :character
##
##
##
##      Total_Claim_Amount      Vehicle_Class      Vehicle_Size
## Min.   : 0.099      Length:9134      Length:9134
## 1st Qu.: 272.258      Class :character  Class :character
## Median : 383.945      Mode  :character  Mode  :character
## Mean   : 434.089
## 3rd Qu.: 547.515
## Max.   :2893.240
```


##	2191	KI58952	California	51337.91	No	Premium	
##	6570	LW64678	California	51016.07	No	Premium	
##	7836	QT84069	Oregon	50568.26	No	Extended	
##	6978	BR50492	Arizona	49423.80	No	Extended	
##	1559	RP30093	Oregon	49221.43	No	Premium	
##	1813	LU42720	Nevada	48356.96	No	Extended	
##	3458	MJ77630	Oregon	47155.63	No	Extended	
##	1859	CP92616	Nevada	46805.22	No	Extended	
##	3211	KB44286	Oregon	46770.95	No	Basic	
##		Education	Effective_To_Date	EmploymentStatus	Gender	Income	
##	6555	College	1/9/2011	Employed	F	84650	
##	2191	College	2/24/2011	Employed	F	72794	
##	6570	Master	2/19/2011	Employed	F	25167	
##	7836	Master	2/28/2011	Employed	M	82081	
##	6978	Bachelor	1/4/2011	Employed	M	85058	
##	1559	Bachelor	1/23/2011	Employed	F	63035	
##	1813	College	2/20/2011	Employed	M	52499	
##	3458	High School or Below	2/10/2011	Employed	M	39891	
##	1859	High School or Below	2/25/2011	Employed	M	83006	
##	3211	High School or Below	2/1/2011	Employed	F	64403	
##		Location_Code	Marital_Status	Monthly_Premium_Auto	Months_Since_Last_Claim		
##	6555	Urban	Married	185			13
##	2191	Rural	Single	164			3
##	6570	Urban	Married	140			3
##	7836	Urban	Married	249			1
##	6978	Urban	Married	137			34
##	1559	Suburban	Married	153			20
##	1813	Suburban	Divorced	138			0
##	3458	Urban	Married	133			12
##	1859	Urban	Married	235			8
##	3211	Rural	Single	198			11
##		Months_Since_Policy_Inception	Number_of_Open_Complaints	Number_of_Policies			
##	6555		39	3			2
##	2191		47	1			2
##	6570		76	0			2
##	7836		62	0			2
##	6978		82	0			2
##	1559		97	0			2
##	1813		61	0			2
##	3458		31	0			2
##	1859		61	1			2
##	3211		86	0			2
##		Policy_Type	Policy	Renew_Offer_Type	Sales_Channel		
##	6555	Personal Auto	Personal L2	Offer1	Agent		
##	2191	Personal Auto	Personal L2	Offer1	Web		
##	6570	Personal Auto	Personal L3	Offer2	Agent		
##	7836	Personal Auto	Personal L1	Offer2	Branch		
##	6978	Personal Auto	Personal L1	Offer1	Call Center		
##	1559	Personal Auto	Personal L3	Offer1	Agent		
##	1813	Personal Auto	Personal L2	Offer3	Call Center		
##	3458	Personal Auto	Personal L3	Offer1	Agent		
##	1859	Personal Auto	Personal L1	Offer3	Agent		
##	3211	Corporate Auto	Corporate L1	Offer1	Agent		
##		Total_Claim_Amount	Vehicle_Class	Vehicle_Size			

```
## 6555      660.47427   Luxury Car   Medsize
## 2191       50.45446      SUV      Large
## 6570      422.49429      SUV      Small
## 7836      753.76010   Luxury SUV   Small
## 6978      595.36978      SUV      Medsize
## 1559      734.40000      SUV      Medsize
## 1813     1171.53759      SUV      Medsize
## 3458      630.88866      SUV      Medsize
## 1859     1065.04989   Luxury SUV   Small
## 3211      111.17302   Luxury SUV   Small
```

Box plot for “Customer Lifetime Value”

```
boxplot(data$Customer_Lifetime_Value, main="Customer Lifetime Value",
        sub=paste("Outlier rows: ", boxplot.stats(data$Customer_Lifetime_Value)$out))
```



```
## Outlier rows: 6555 2191 6570 7836 6978 1559 1813 3458 1859 3211
```

Identify character features

```
character_cols <- sapply(data, is.character)
head(data[, character_cols])
```

```
##   Customer      State Response Coverage Education Effective_To_Date
## 1 BU79786 Washington      No    Basic  Bachelor      2/24/2011
## 2 QZ44356   Arizona      No Extended  Bachelor      1/31/2011
## 3 AI49188    Nevada      No  Premium  Bachelor      2/19/2011
## 4 WW63253 California      No    Basic  Bachelor      1/20/2011
```

```
## 5 HB64268 Washington No Basic Bachelor 2/3/2011
## 6 OC83172 Oregon Yes Basic Bachelor 1/25/2011
## EmploymentStatus Gender Location_Code Marital_Status Policy_Type
## 1 Employed F Suburban Married Corporate Auto
## 2 Unemployed F Suburban Single Personal Auto
## 3 Employed F Suburban Married Personal Auto
## 4 Unemployed M Suburban Married Corporate Auto
## 5 Employed M Rural Single Personal Auto
## 6 Employed F Rural Married Personal Auto
## Policy Renew_Offer_Type Sales_Channel Vehicle_Class Vehicle_Size
## 1 Corporate L3 Offer1 Agent Two-Door Car Medsize
## 2 Personal L3 Offer3 Agent Four-Door Car Medsize
## 3 Personal L3 Offer1 Agent Two-Door Car Medsize
## 4 Corporate L2 Offer1 Call Center SUV Medsize
## 5 Personal L1 Offer1 Agent Four-Door Car Medsize
## 6 Personal L3 Offer2 Web Two-Door Car Medsize
```

Check unique value

Print the unique values State

```
unique(data$State)
```

```
## [1] "Washington" "Arizona" "Nevada" "California" "Oregon"
```

Print the unique values Response

```
unique(data$Response)
```

```
## [1] "No" "Yes"
```

Print the unique values Coverage

```
unique(data$Coverage)
```

```
## [1] "Basic" "Extended" "Premium"
```

Print the unique values Education

```
unique(data$Education)
```

```
## [1] "Bachelor" "College" "Master"
## [4] "High School or Below" "Doctor"
```

Print the unique values EmploymentStatus

```
unique(data$EmploymentStatus)
```

```
## [1] "Employed" "Unemployed" "Medical Leave" "Disabled"
## [5] "Retired"
```

Print the unique values Gender

```
unique(data$Gender)
```

```
## [1] "F" "M"
```

Print the unique values Location_Code

```
unique(data$Location_Code)
```

```
## [1] "Suburban" "Rural" "Urban"
```

Print the unique values Marital_Status

```
unique(data$Marital_Status)
```

```
## [1] "Married" "Single" "Divorced"
```

Print the unique values Policy_Type

```
unique(data$Policy_Type)
```

```
## [1] "Corporate Auto" "Personal Auto" "Special Auto"
```

Print the unique values Policy

```
unique(data$Policy)
```

```
## [1] "Corporate L3" "Personal L3" "Corporate L2" "Personal L1" "Special L2"  
## [6] "Corporate L1" "Personal L2" "Special L1" "Special L3"
```

Print the unique values Renew_Offer_Type

```
unique(data$Renew_Offer_Type)
```

```
## [1] "Offer1" "Offer3" "Offer2" "Offer4"
```

Print the unique values Sales_Channel

```
unique(data$Sales_Channel)
```

```
## [1] "Agent" "Call Center" "Web" "Branch"
```

Print the unique values Vehicle_Class

```
unique(data$Vehicle_Class)
```

```
## [1] "Two-Door Car" "Four-Door Car" "SUV"          "Luxury SUV"  
## [5] "Sports Car"    "Luxury Car"
```

Print the unique values Vehicle_Size

```
unique(data$Vehicle_Size)
```

```
## [1] "Medsize" "Small"   "Large"
```

Convert Character to Numeric

Response: No- 0, Yes- 1

```
data$Response <- ifelse(data$Response == "No", 0, 1)
```

State

```
data$State <- as.numeric(factor(data$State))
```

EmploymentStatus

```
data$EmploymentStatus <- as.numeric(factor(data$EmploymentStatus))
```

Gender

```
data$Gender <- as.numeric(factor(data$Gender))
```

Marital_Status

```
data$Marital_Status <- as.numeric(factor(data$Marital_Status))
```

Sales_Channel

```
data$Sales_Channel <- as.numeric(factor(data$Sales_Channel))
```

Vehicle_Class

```
data$Vehicle_Class <- as.numeric(factor(data$Vehicle_Class))
```

Coverage ("Premium", "Extended", "Basic")

```
data$Coverage <- as.numeric(factor(data$Coverage, levels = c("Premium", "Extended", "Basic")))
```

Education ("Doctor", "Master", "Bachelor", "College", "High School or Below")


```
data$Education <- as.numeric(factor(data$Education, levels = c("Doctor", "Master", "Bachelor", "College", "I
```

```
Location_Code ("Urban", "Suburban", "Rural")
```

```
data$Location_Code <- as.numeric(factor(data$Location_Code, levels = c("Urban", "Suburban", "Rural")))
```

```
Policy_Type ("Special Auto", "Corporate Auto", "Personal Auto")
```

```
data$Policy_Type <- as.numeric(factor(data$Policy_Type, levels = c("Special Auto", "Corporate Auto", "P
```

```
Policy ("Special L1", "Special L2", "Special L3", "Corporate L1", "Corporate L2", "Corporate L3", "Personal  
L1", "Personal L2", "Personal L3")
```

```
data$Policy <- as.numeric(factor(data$Policy, levels = c("Special L1", "Special L2", "Special L3", "Cor
```

```
Renew_Offer_Type ("Offer1", "Offer2", "Offer3", "Offer4")
```

```
data$Renew_Offer_Type <- as.numeric(factor(data$Renew_Offer_Type, levels = c("Offer1", "Offer2", "Offer
```

```
Vehicle_Size ("Small", "Medsize", "Large")
```

```
data$Vehicle_Size <- as.numeric(factor(data$Vehicle_Size, levels = c("Small", "Medsize", "Large")))
```

```
summary(data)
```

```
##      Customer      State      Customer_Lifetime_Value      Response
## Length:9118      Min.    :1.000      Min.    : 1898      Min.    :0.0000
## Class :character  1st Qu.:2.000      1st Qu.: 3985      1st Qu.:0.0000
## Mode  :character  Median :2.000      Median : 5774      Median :0.0000
##                      Mean   :2.741      Mean   : 7908      Mean   :0.1435
##                      3rd Qu.:4.000      3rd Qu.: 8930      3rd Qu.:0.0000
##                      Max.    :5.000      Max.    :51426      Max.    :1.0000
##      Coverage      Education      Effective_To_Date      EmploymentStatus
## Min.    :1.00      Min.    :1.000      Length:9118      Min.    :1.000
## 1st Qu.:2.00      1st Qu.:3.000      Class :character  1st Qu.:2.000
## Median :3.00      Median :4.000      Mode  :character  Median :2.000
## Mean    :2.52      Mean    :3.711                      Mean    :2.825
## 3rd Qu.:3.00      3rd Qu.:5.000                      3rd Qu.:5.000
## Max.    :3.00      Max.    :5.000                      Max.    :5.000
##      Gender      Income      Location_Code      Marital_Status
## Min.    :1.00      Min.    : 0      Min.    :1.000      Min.    :1.00
## 1st Qu.:1.00      1st Qu.: 0      1st Qu.:2.000      1st Qu.:2.00
## Median :1.00      Median :33899      Median :2.000      Median :2.00
## Mean    :1.49      Mean    :37669      Mean    :2.021      Mean    :2.12
## 3rd Qu.:2.00      3rd Qu.:62358      3rd Qu.:2.000      3rd Qu.:3.00
## Max.    :2.00      Max.    :99981      Max.    :3.000      Max.    :3.00
## Monthly_Premium_Auto      Months_Since_Last_Claim      Months_Since_Policy_Inception
## Min.    : 61.00      Min.    : 0.00      Min.    : 0.00
## 1st Qu.: 68.00      1st Qu.: 6.00      1st Qu.:24.00
## Median : 83.00      Median :14.00      Median :48.00
```

```
## Mean : 93.02      Mean :15.09      Mean :48.06
## 3rd Qu.:109.00    3rd Qu.:23.00    3rd Qu.:71.00
## Max. :298.00      Max. :35.00      Max. :99.00
## Number_of_Open_Complaints Number_of_Policies Policy_Type Policy
## Min. :0.0000      Min. :1.000      Min. :1.000      Min. :1.000
## 1st Qu.:0.0000    1st Qu.:1.000    1st Qu.:2.000    1st Qu.:6.000
## Median :0.0000     Median :2.000     Median :3.000     Median :8.000
## Mean :0.3847       Mean :2.968       Mean :2.701       Mean :7.424
## 3rd Qu.:0.0000     3rd Qu.:4.000     3rd Qu.:3.000     3rd Qu.:9.000
## Max. :5.0000       Max. :9.000       Max. :3.000       Max. :9.000
## Renew_Offer_Type Sales_Channel Total_Claim_Amount Vehicle_Class
## Min. :1.000      Min. :1.000      Min. : 0.099      Min. :1.000
## 1st Qu.:1.000    1st Qu.:1.000    1st Qu.: 271.983    1st Qu.:1.000
## Median :2.000     Median :2.000     Median : 383.296     Median :1.000
## Mean :1.971       Mean :2.103       Mean : 432.906       Mean :3.036
## 3rd Qu.:3.000     3rd Qu.:3.000     3rd Qu.: 547.200     3rd Qu.:5.000
## Max. :4.000       Max. :4.000       Max. :2893.240       Max. :6.000
## Vehicle_Size
## Min. :1.00
## 1st Qu.:2.00
## Median :2.00
## Mean :1.91
## 3rd Qu.:2.00
## Max. :3.00
```

Scale Data

```
columns_to_scale <- c(
  "Customer_Lifetime_Value", "Income", "Monthly_Premium_Auto",
  "Months_Since_Last_Claim", "Months_Since_Policy_Inception",
  "Number_of_Open_Complaints", "Number_of_Policies", "Total_Claim_Amount"
)

scale_data <- scale(data[columns_to_scale])
head(scale_data)
```

```
## Customer_Lifetime_Value Income Monthly_Premium_Auto
## 1 -0.79535661 0.6123357 -0.7047628
## 2 -0.14356772 -1.2398258 0.0288120
## 3 0.76978278 0.3652558 0.4396139
## 4 -0.04055488 -1.2398258 0.3809279
## 5 -0.78759991 0.2029605 -0.5873909
## 6 0.05381747 0.8304849 -0.7047628
## Months_Since_Last_Claim Months_Since_Policy_Inception
## 1 1.6791513 -1.5425994
## 2 -0.2079686 -0.2171227
## 3 0.2886419 -0.3604175
## 4 0.2886419 0.6068223
## 5 -0.3072907 -0.1454753
## 6 -0.1086465 1.6457094
## Number_of_Open_Complaints Number_of_Policies Total_Claim_Amount
## 1 -0.4223209 -0.8227086 -0.1666108
```

```
## 2          -0.4223209          2.1037920          2.4199345
## 3          -0.4223209         -0.4046371          0.4626965
## 4          -0.4223209          1.6857205          0.3359389
## 5          -0.4223209         -0.8227086         -1.0211564
## 6          -0.4223209         -0.4046371         -0.9475350
```

```
data <- data[, !(names(data) %in% columns_to_scale)]
```

```
head(data)
```

```
## Customer State Response Coverage Education Effective_To_Date EmploymentStatus
## 1 BU79786      5         0         3         3         2/24/2011             2
## 2 QZ44356      1         0         2         3         1/31/2011             5
## 3 AI49188      3         0         1         3         2/19/2011             2
## 4 WW63253      2         0         3         3         1/20/2011             5
## 5 HB64268      5         0         3         3         2/3/2011              2
## 6 OC83172      4         1         3         3         1/25/2011             2
## Gender Location_Code Marital_Status Policy_Type Policy Renew_Offer_Type
## 1      1              2              2          2         6              1
## 2      1              2              3          3         9              3
## 3      1              2              2          3         9              1
## 4      2              2              2          2         5              1
## 5      2              3              3          3         7              1
## 6      1              3              2          3         9              2
## Sales_Channel Vehicle_Class Vehicle_Size
## 1              1              6              2
## 2              1              1              2
## 3              1              6              2
## 4              3              5              2
## 5              1              1              2
## 6              4              6              2
```

Combine Data & Scale Data

```
df <- cbind(data, scale_data)
head(df)
```

```
## Customer State Response Coverage Education Effective_To_Date EmploymentStatus
## 1 BU79786      5         0         3         3         2/24/2011             2
## 2 QZ44356      1         0         2         3         1/31/2011             5
## 3 AI49188      3         0         1         3         2/19/2011             2
## 4 WW63253      2         0         3         3         1/20/2011             5
## 5 HB64268      5         0         3         3         2/3/2011             2
## 6 OC83172      4         1         3         3         1/25/2011             2
## Gender Location_Code Marital_Status Policy_Type Policy Renew_Offer_Type
## 1      1              2              2          2         6              1
## 2      1              2              3          3         9              3
## 3      1              2              2          3         9              1
## 4      2              2              2          2         5              1
## 5      2              3              3          3         7              1
## 6      1              3              2          3         9              2
```

```
## Sales_Channel Vehicle_Class Vehicle_Size Customer_Lifetime_Value Income
## 1 1 6 2 -0.79535661 0.6123357
## 2 1 1 2 -0.14356772 -1.2398258
## 3 1 6 2 0.76978278 0.3652558
## 4 3 5 2 -0.04055488 -1.2398258
## 5 1 1 2 -0.78759991 0.2029605
## 6 4 6 2 0.05381747 0.8304849
## Monthly_Premium_Auto Months_Since_Last_Claim Months_Since_Policy_Inception
## 1 -0.7047628 1.6791513 -1.5425994
## 2 0.0288120 -0.2079686 -0.2171227
## 3 0.4396139 0.2886419 -0.3604175
## 4 0.3809279 0.2886419 0.6068223
## 5 -0.5873909 -0.3072907 -0.1454753
## 6 -0.7047628 -0.1086465 1.6457094
## Number_of_Open_Complaints Number_of_Policies Total_Claim_Amount
## 1 -0.4223209 -0.8227086 -0.1666108
## 2 -0.4223209 2.1037920 2.4199345
## 3 -0.4223209 -0.4046371 0.4626965
## 4 -0.4223209 1.6857205 0.3359389
## 5 -0.4223209 -0.8227086 -1.0211564
## 6 -0.4223209 -0.4046371 -0.9475350
```

```
dim(df)
```

```
## [1] 9118 24
```

```
summary(df)
```

```
## Customer State Response Coverage
## Length:9118 Min. :1.000 Min. :0.0000 Min. :1.00
## Class :character 1st Qu.:2.000 1st Qu.:0.0000 1st Qu.:2.00
## Mode :character Median :2.000 Median :0.0000 Median :3.00
## Mean :2.741 Mean :0.1435 Mean :2.52
## 3rd Qu.:4.000 3rd Qu.:0.0000 3rd Qu.:3.00
## Max. :5.000 Max. :1.0000 Max. :3.00
## Education Effective_To_Date EmploymentStatus Gender
## Min. :1.000 Length:9118 Min. :1.000 Min. :1.00
## 1st Qu.:3.000 Class :character 1st Qu.:2.000 1st Qu.:1.00
## Median :4.000 Mode :character Median :2.000 Median :1.00
## Mean :3.711 Mean :2.825 Mean :1.49
## 3rd Qu.:5.000 3rd Qu.:5.000 3rd Qu.:2.00
## Max. :5.000 Max. :5.000 Max. :2.00
## Location_Code Marital_Status Policy_Type Policy
## Min. :1.000 Min. :1.00 Min. :1.000 Min. :1.000
## 1st Qu.:2.000 1st Qu.:2.00 1st Qu.:2.000 1st Qu.:6.000
## Median :2.000 Median :2.00 Median :3.000 Median :8.000
## Mean :2.021 Mean :2.12 Mean :2.701 Mean :7.424
## 3rd Qu.:2.000 3rd Qu.:3.00 3rd Qu.:3.000 3rd Qu.:9.000
## Max. :3.000 Max. :3.00 Max. :3.000 Max. :9.000
## Renew_Offer_Type Sales_Channel Vehicle_Class Vehicle_Size
## Min. :1.000 Min. :1.000 Min. :1.000 Min. :1.00
## 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:1.000 1st Qu.:2.00
## Median :2.000 Median :2.000 Median :1.000 Median :2.00
```

```
## Mean :1.971 Mean :2.103 Mean :3.036 Mean :1.91
## 3rd Qu.:3.000 3rd Qu.:3.000 3rd Qu.:5.000 3rd Qu.:2.00
## Max. :4.000 Max. :4.000 Max. :6.000 Max. :3.00
## Customer_Lifetime_Value Income Monthly_Premium_Auto
## Min. :-0.9292 Min. :-1.2398 Min. :-0.9395
## 1st Qu.: -0.6065 1st Qu.: -1.2398 1st Qu.: -0.7341
## Median : -0.3300 Median : -0.1241 Median : -0.2940
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000
## 3rd Qu.: 0.1580 3rd Qu.: 0.8126 3rd Qu.: 0.4690
## Max. : 6.7278 Max. : 2.0509 Max. : 6.0148
## Months_Since_Last_Claim Months_Since_Policy_Inception
## Min. :-1.4992 Min. :-1.721718
## 1st Qu.: -0.9032 1st Qu.: -0.861949
## Median : -0.1086 Median : -0.002181
## Mean : 0.0000 Mean : 0.000000
## 3rd Qu.: 0.7853 3rd Qu.: 0.821764
## Max. : 1.9771 Max. : 1.824828
## Number_of_Open_Complaints Number_of_Policies Total_Claim_Amount
## Min. :-0.4223 Min. :-0.8227 Min. :-1.4993
## 1st Qu.: -0.4223 1st Qu.: -0.8227 1st Qu.: -0.5575
## Median : -0.4223 Median : -0.4046 Median : -0.1719
## Mean : 0.0000 Mean : 0.0000 Mean : 0.0000
## 3rd Qu.: -0.4223 3rd Qu.: 0.4315 3rd Qu.: 0.3959
## Max. : 5.0662 Max. : 2.5219 Max. : 8.5230
```

Identify numeric features

```
numeric_cols <- sapply(df, is.numeric)
df <- df[, numeric_cols]
dim(df)
```

```
## [1] 9118 22
```

```
head(df)
```

```
## State Response Coverage Education EmploymentStatus Gender Location_Code
## 1 5 0 3 3 2 1 2
## 2 1 0 2 3 5 1 2
## 3 3 0 1 3 2 1 2
## 4 2 0 3 3 5 2 2
## 5 5 0 3 3 2 2 3
## 6 4 1 3 3 2 1 3
## Marital_Status Policy_Type Policy Renew_Offer_Type Sales_Channel
## 1 2 2 6 1 1
## 2 3 3 9 3 1
## 3 2 3 9 1 1
## 4 2 2 5 1 3
## 5 3 3 7 1 1
## 6 2 3 9 2 4
## Vehicle_Class Vehicle_Size Customer_Lifetime_Value Income
## 1 6 2 -0.79535661 0.6123357
```

```
## 2      1      2      -0.14356772 -1.2398258
## 3      6      2      0.76978278  0.3652558
## 4      5      2     -0.04055488 -1.2398258
## 5      1      2     -0.78759991  0.2029605
## 6      6      2      0.05381747  0.8304849
##   Monthly_Premium_Auto Months_Since_Last_Claim Months_Since_Policy_Inception
## 1      -0.7047628      1.6791513      -1.5425994
## 2      0.0288120     -0.2079686     -0.2171227
## 3      0.4396139      0.2886419     -0.3604175
## 4      0.3809279      0.2886419      0.6068223
## 5     -0.5873909     -0.3072907     -0.1454753
## 6     -0.7047628     -0.1086465      1.6457094
##   Number_of_Open_Complaints Number_of_Policies Total_Claim_Amount
## 1      -0.4223209     -0.8227086     -0.1666108
## 2      -0.4223209      2.1037920      2.4199345
## 3      -0.4223209     -0.4046371      0.4626965
## 4      -0.4223209      1.6857205      0.3359389
## 5      -0.4223209     -0.8227086     -1.0211564
## 6      -0.4223209     -0.4046371     -0.9475350
```

Check Missing Value

```
library(Amelia)
```

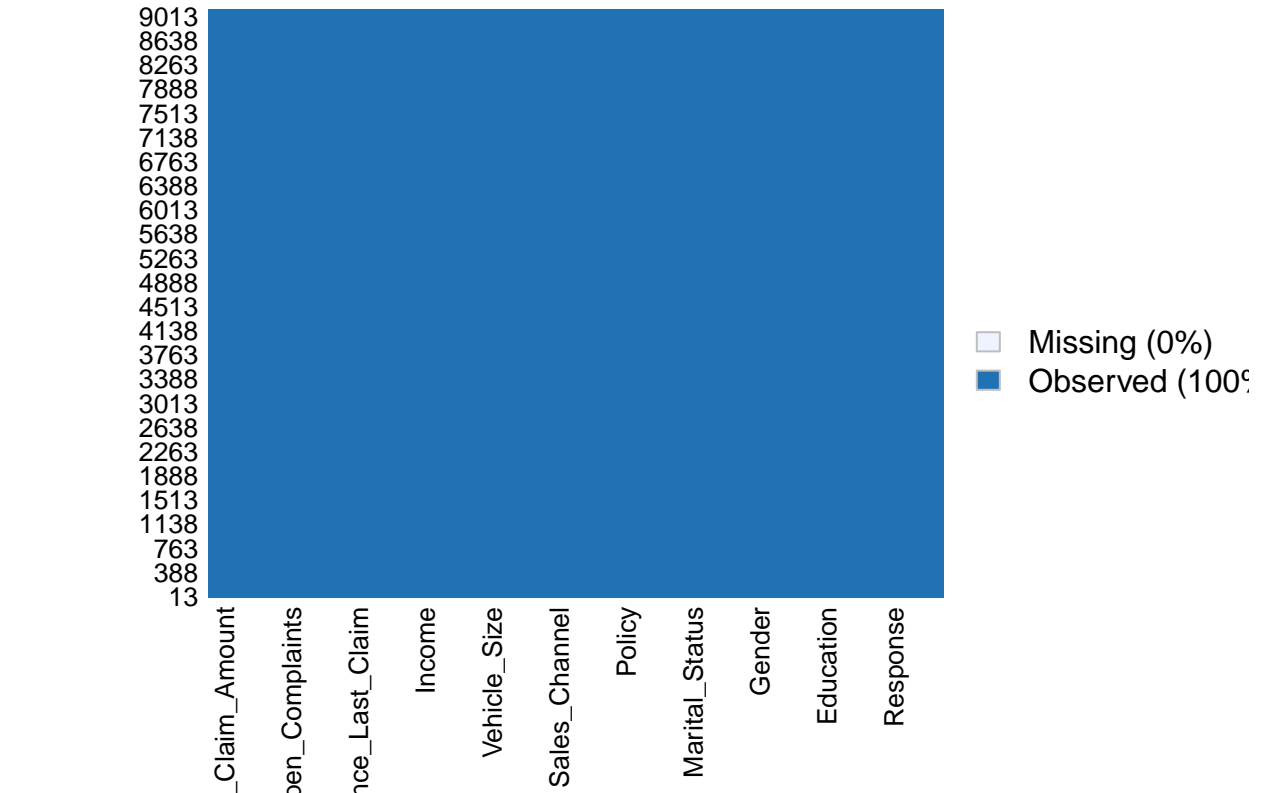
```
## Warning: package 'Amelia' was built under R version 4.2.3
```

```
## Loading required package: Rcpp
```

```
## ##
## ## Amelia II: Multiple Imputation
## ## (Version 1.8.1, built: 2022-11-18)
## ## Copyright (C) 2005-2023 James Honaker, Gary King and Matthew Blackwell
## ## Refer to http://gking.harvard.edu/amelia/ for more information
## ##
```

```
missmap(df, main="Missing values")
```

Missing values



Save file

```
write.csv(df, file = "data_processed.csv", row.names = FALSE)
```