Sample Business Report

NAME- SOUMIK MUKHOPADHYAY

ROLL- 1928127

Contents

1 Exploratory Data Analysis

1.1 Introduction of the business problem

The major objective of this data set is to help the company to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and up skill programs for low performing agents.

2: Structure of Data

2.1 Data Description

The data belongs to a leading life insurance company's agent. This dataset contains strong as well as weak attributes about the performance of the agents working in the company.

Variable	Description
CustID	Unique customer ID
AgentBonus	Bonus amount given to each agents in last month
Age	Age of customer
CustTenure	Tenure of customer in organization
Channel	Channel through which acquisition of customer is done
Occupation	Occupation of customer
EducationField	Field of education of customer
Gender	Gender of customer
ExistingProdType	Existing product type of customer
Designation	Designation of customer in their organization
NumberOfPolicy	Total number of existing policy of a customer
MaritalStatus	Marital status of customer
MonthlyIncome	Gross monthly income of customer
Complaint	Indicator of complaint registered in last one month by customer
ExistingPolicyTenure	Max tenure in all existing policies of customer
SumAssured	Max of sum assured in all existing policies of customer
Zone	Customer belongs to which zone in India. Like East, West, North and South
PaymentMethod	Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
LastMonthCalls	Total calls attempted by company to a customer for cross sell
CustCareScore	Customer satisfaction score given by customer in previous service call

2.2 Visual inspection of data (rows, columns, descriptive details)

The following dataset after inspection indicates :

- Name of the column attributes
- Datatype of the column attributes- int, float, object.
- The number of of rows contained in each column
- The total number of rows and columns in the dataset are 4520 * 20
- Dataset also shows the null values or missing values.

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   CustID                4520 non-null   int64
 1   AgentBonus            4520 non-null   int64
 2   Age                   4251 non-null   float64
 3   CustTenure            4294 non-null   float64
 4   Channel               4520 non-null   object
 5   Occupation            4520 non-null   object
 6   EducationField        4520 non-null   object
 7   Gender                4520 non-null   object
 8   ExistingProdType      4520 non-null   int64
 9   Designation           4520 non-null   object
 10  NumberOfPolicy        4475 non-null   float64
 11  MaritalStatus         4520 non-null   object
 12  MonthlyIncome         4284 non-null   float64
 13  Complaint             4520 non-null   int64
 14  ExistingPolicyTenure  4336 non-null   float64
 15  SumAssured            4366 non-null   float64
 16  Zone                  4520 non-null   object
 17  PaymentMethod         4520 non-null   object
 18  LastMonthCalls        4520 non-null   int64
 19  CustCareScore         4468 non-null   float64
dtypes: float64(7), int64(5), object(8)
memory usage: 706.4+ KB
```

	0	1	2	3	4	5	6	7	8	9
CustID	7000000	7000001	7000002	7000003	7000004	7000005	7000006	7000007	7000008	7000009
AgentBonus	4409	2214	4273	1791	2955	3252	3850	2073	2719	3247
Age	22.0	11.0	26.0	11.0	6.0	7.0	12.0	6.0	8.0	6.0
CustTenure	4.0	2.0	4.0	NaN	NaN	NaN	23.0	4.0	11.0	3.0
Channel	Agent	Third Party Partner	Agent	Third Party Partner	Agent	Third Party Partner	Agent	Agent	Agent	Online
Occupation	Salaried	Salaried	Free Lancer	Salaried	Small Business	Salaried	Salaried	Small Business	Salaried	Small Business
EducationField	Graduate	Graduate	Post Graduate	Graduate	UG	Graduate	Graduate	Under Graduate	Graduate	Under Graduate
Gender	Female	Male	Male	Fe male	Male	Male	Male	Female	Male	Male
ExistingProdType	3	4	4	3	3	3	4	3	4	2
Designation	Manager	Manager	Exe	Executive	Executive	Executive	VP	Executive	Manager	Exe
NumberOfPolicy	2.0	4.0	3.0	3.0	4.0	2.0	3.0	4.0	3.0	2.0
MaritalStatus	Single	Divorced	Unmarried	Divorced	Divorced	Single	Divorced	Unmarried	Divorced	Married
MonthlyIncome	20993.0	20130.0	17090.0	17909.0	18468.0	18068.0	34999.0	17279.0	20916.0	17089.0
Complaint	1	0	1	1	0	0	0	0	1	0
ExistingPolicyTenure	2.0	3.0	2.0	2.0	4.0	2.0	2.0	2.0	1.0	1.0
SumAssured	806761.0	294502.0	NaN	268635.0	366405.0	487836.0	392689.0	369079.0	405143.0	NaN
Zone	North	North	North	West	West	North	North	West	West	West
PaymentMethod	Half Yearly	Yearly	Yearly	Half Yearly	Half Yearly	Half Yearly	Yearly	Half Yearly	Yearly	Quarterly
LastMonthCalls	5	7	0	0	2	6	9	3	1	2
CustCareScore	2.0	3.0	3.0	5.0	5.0	5.0	2.0	3.0	4.0	4.0

2.3 Understanding of attributes (variable info, renaming if required)

	Column	Count	Data type	Remark
1	CustID	4520	int64	Numeric (Redundant Column and can be remove)
2	Agent Bonus	4520	int64	Numeric value and TARGET value
3	Age	4251	float64	Numeric

4	Cust Tenure	4294	float64	Numeric
5	Channel	4520	object	Categorical
6	Occupation	4520	object	Categorical
7	Educational Field	4520	object	Categorical
8	Gender	4520	object	Categorical
9	ExistingProdtype	4520	int64	Numeric
10	Designation	4520l	object	Categorical
11	Number of Policy	4475 l	float64	Numeric
12	Marital Status	4520	object	Categorical
13	Monthly Income	4284	float64	Numeric
14	Complaint	4520	int64	Numeric
15	Existing Policy Tenure	4336	float64	Numeric
16	Sum Assured	4366	float64	Numeric
17	Zone	4520	object	Categorical
18	Payment method	4520	object	Categorical
19	Last month Calls	4520	int64	Numeric
20	Cust Care Score	4468	float64	Numeric

Here we got the list of categorical values and we are focusing on changing spelling errors. We are also organizing the data in a logical form , for example , here we had UG and Undergraduate as different data blocks, since they mean the same thing , we have removed one.

<u>Unique values of various Categories</u>

Channel : 3
Online 468
Third Party Partner 858
Agent 3194
Name: Channel, dtype: int64

Occupation : 5
Freelancer 2
Large Business 153
Large Business 255
Small Business 1918
Salaried 2192
Name: Occupation, dtype: int64

EducationField : 7
MBA 74
UG 230
Post Graduate 252

Engineer 408
Diploma 496
Under Graduate 1190
Graduate 1870
Name: EducationField, dtype: int64

Gender : 3
Fe male 325
Female 1507
Male 2688
Name: Gender, dtype: int64

Designation : 6
Exe 127
VP 226
AVP 336
Senior Manager 676
Executive 1535
Manager 1620
Name: Designation, dtype: int64

MaritalStatus : 4
Unmarried 194
Divorced 804
Single 1254
Married 2268
Name: MaritalStatus, dtype: int64

Zone : 4
South 6
East 64
North 1884
West 2566
Name: Zone, dtype: int64

PaymentMethod : 4
Quarterly 76
Monthly 354
Yearly 1434
Half Yearly 2656
Name: PaymentMethod, dtype: int64

Post fixing of the data

Gender : 2
Female 1832
Male 2688

Name: Gender, dtype: int64

Occupation : 4
Free Lancer 2
Large Business 408
Small Business 1918
Salaried 2192
Name: Occupation, dtype: int64

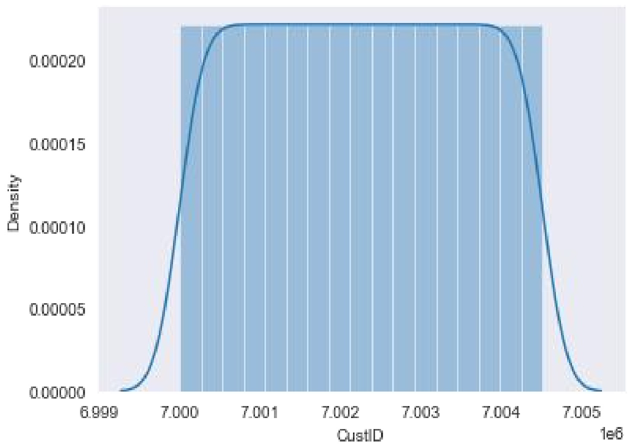
Designation : 5
VP 226
AVP 336
Senior Manager 676
Manager 1620
Executive 1662
Name: Designation, dtype: int64

EducationField : 7
MBA 74
Under Graduate 230
Post Graduate 252
Engineer 408
Diploma 496
Under Graduate 1190
Graduate 1870
Name: EducationField, dtype: int64

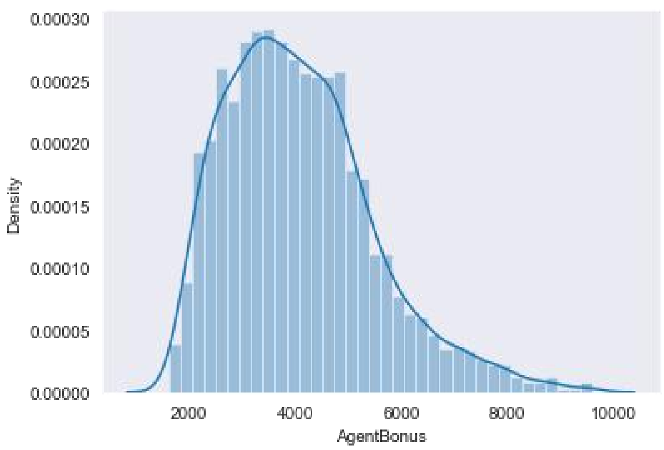
3. Predictive Power of Data

3.1 Uni variate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

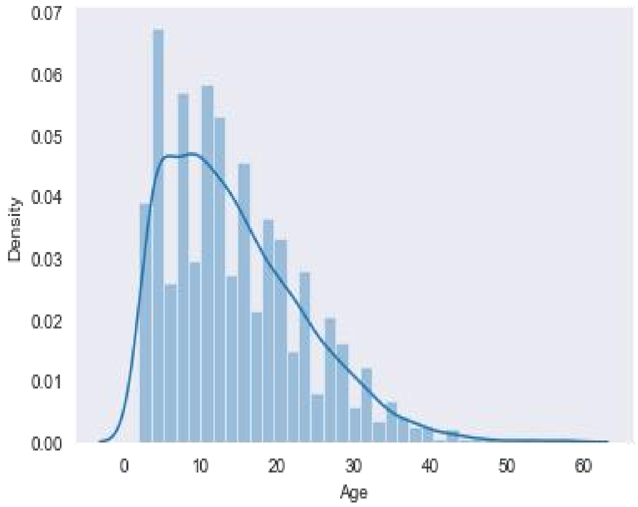
Analysing without any alteration in the dataset



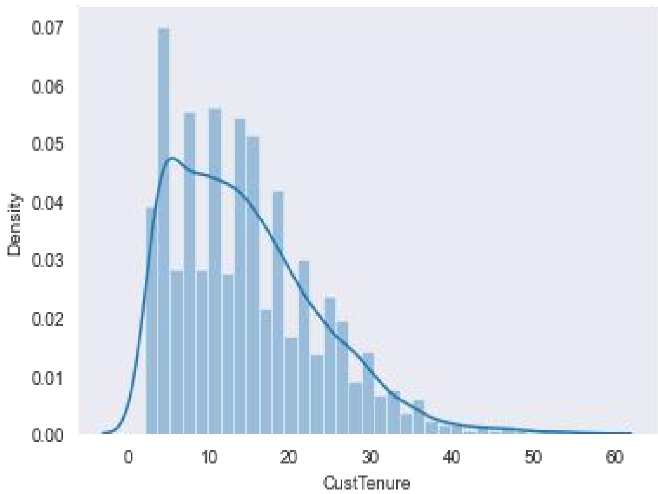
1: No Change



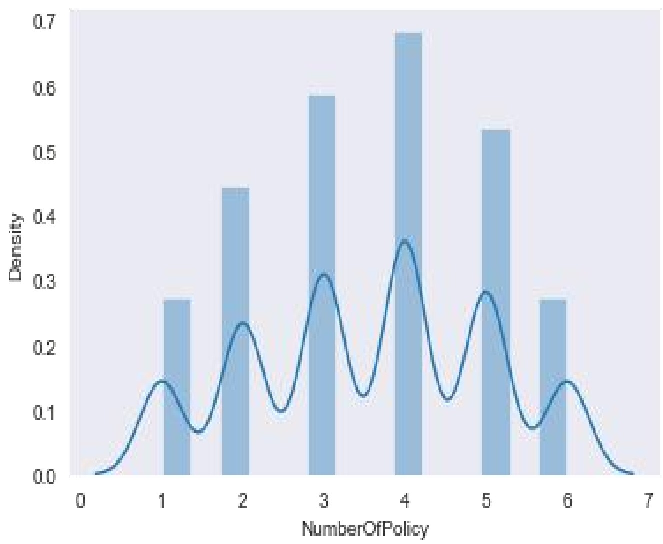
2: Slightly right skewed and continuous data



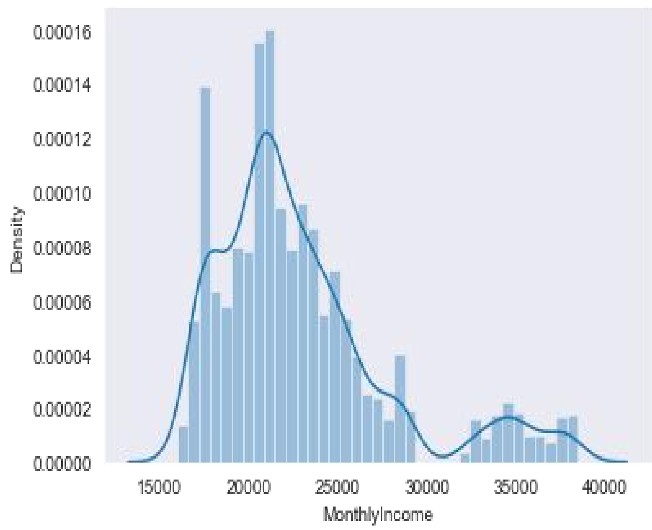
3:Slightly right skewed and continuous data



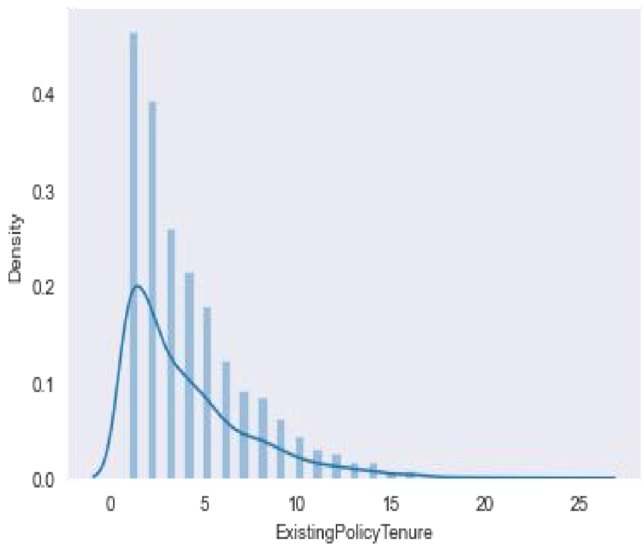
4:Slightly right skewed and continuous data



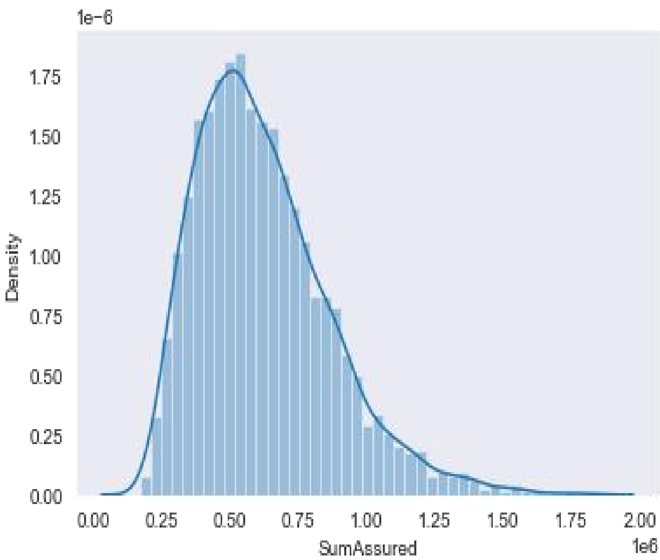
5: More Discrete Kind of data,4 is the most frequent observation



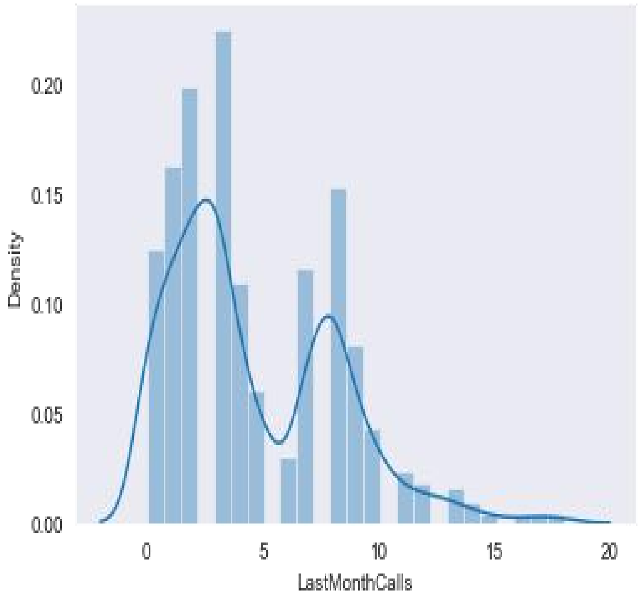
6: Discontinuous Kind of data



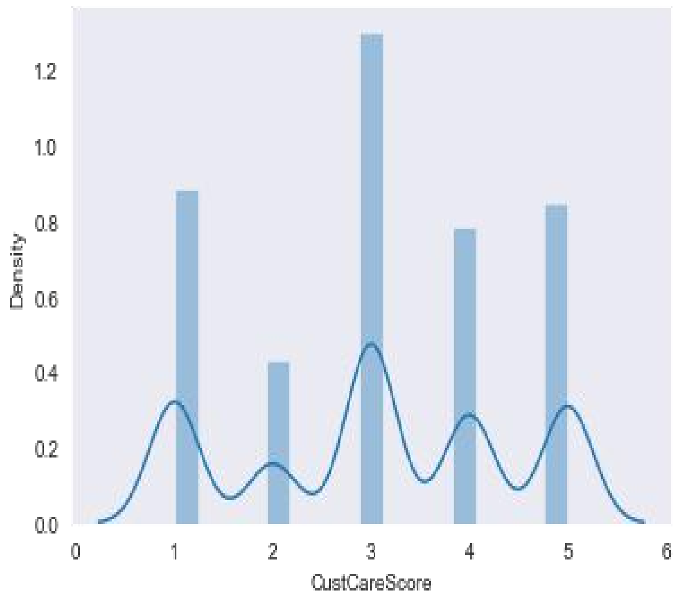
7: Discrete Kind of data,1 is the most frequent observation



8: Slightly right skewed and continuous data

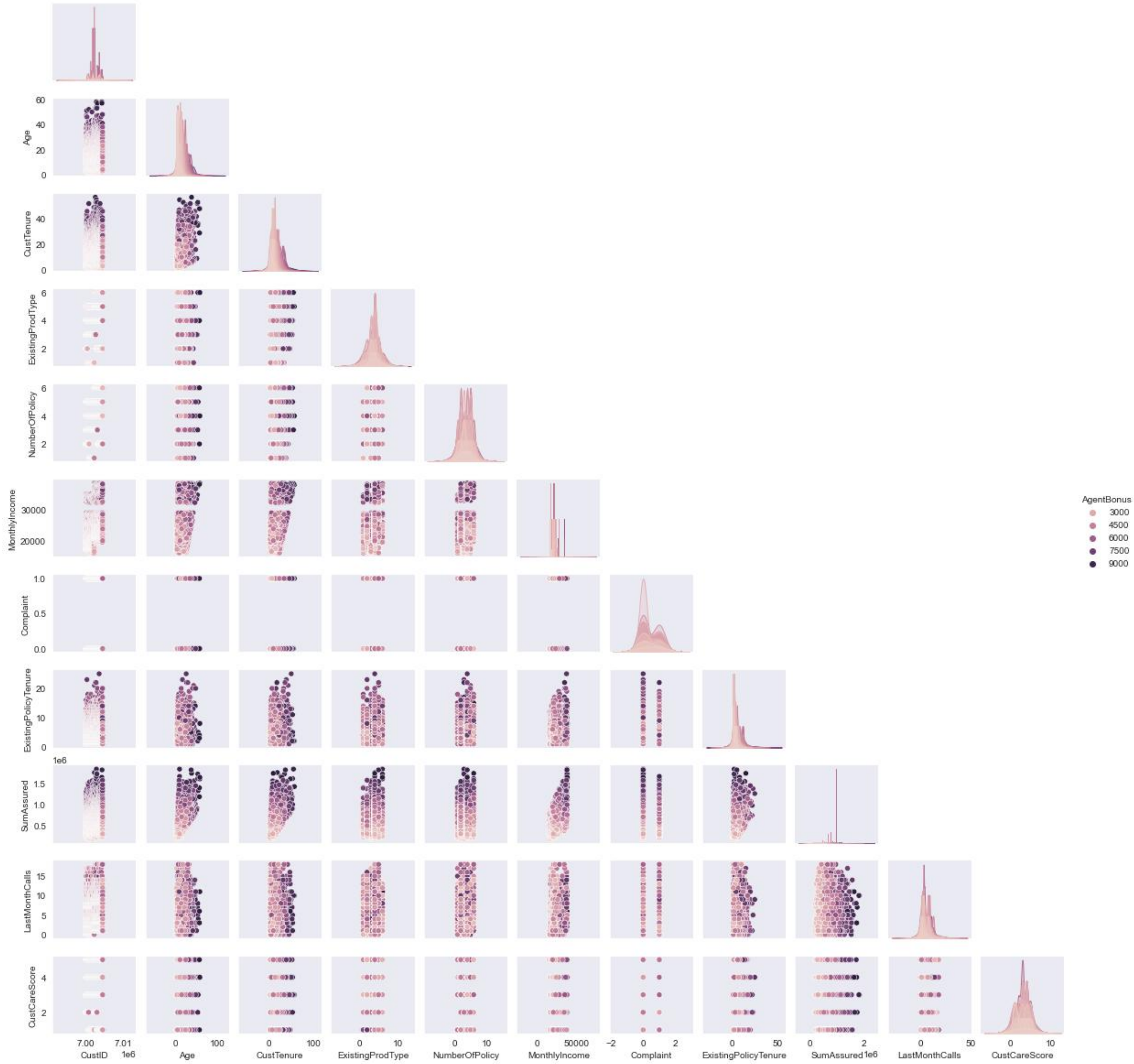


9: Discontinuous Kind of data

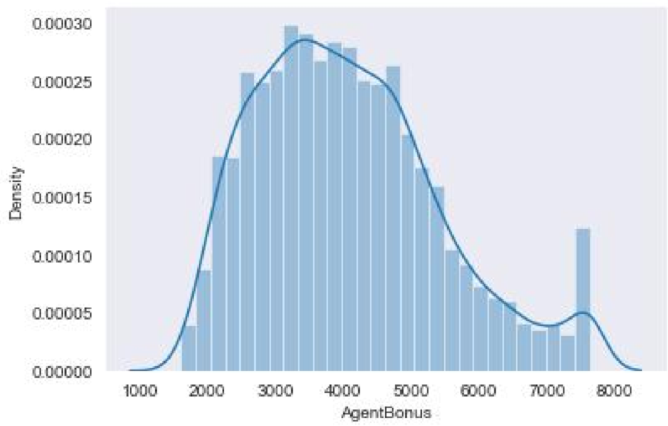


10: More Discrete Kind of data,3 is the most frequent

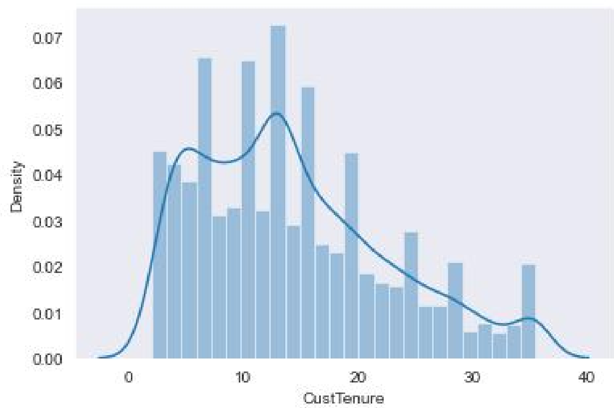
Most of the numerical data is discrete since the nature of the domain is such. So even if the data seems continuous but is limited to a range.



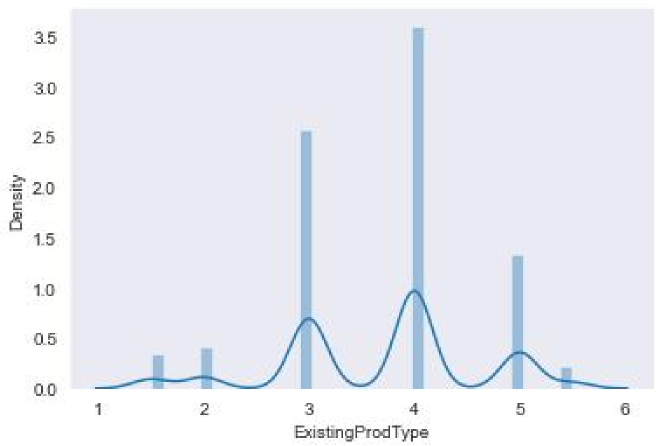
Analysing after alteration of the Dataset



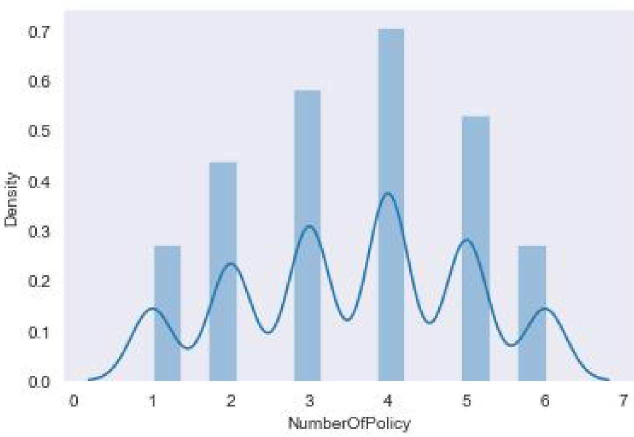
1: No Change



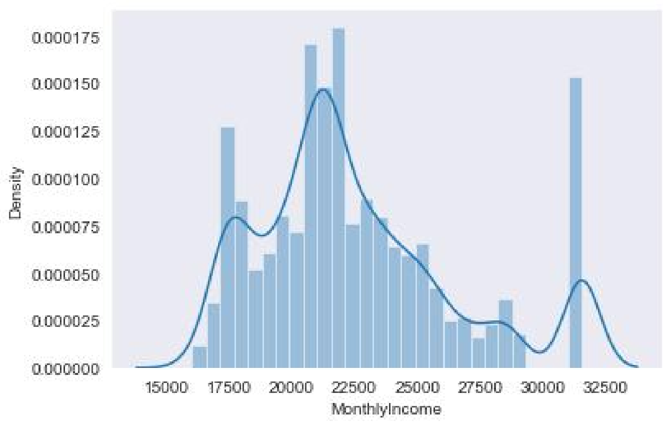
2: Slightly right skewed and continuous data



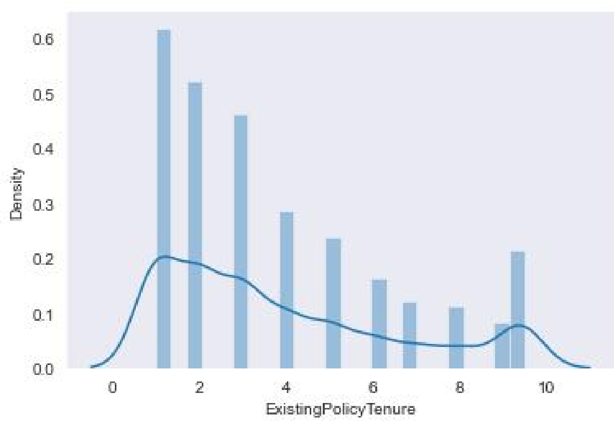
3: Discrete Kind of data,4 is the most frequent observation



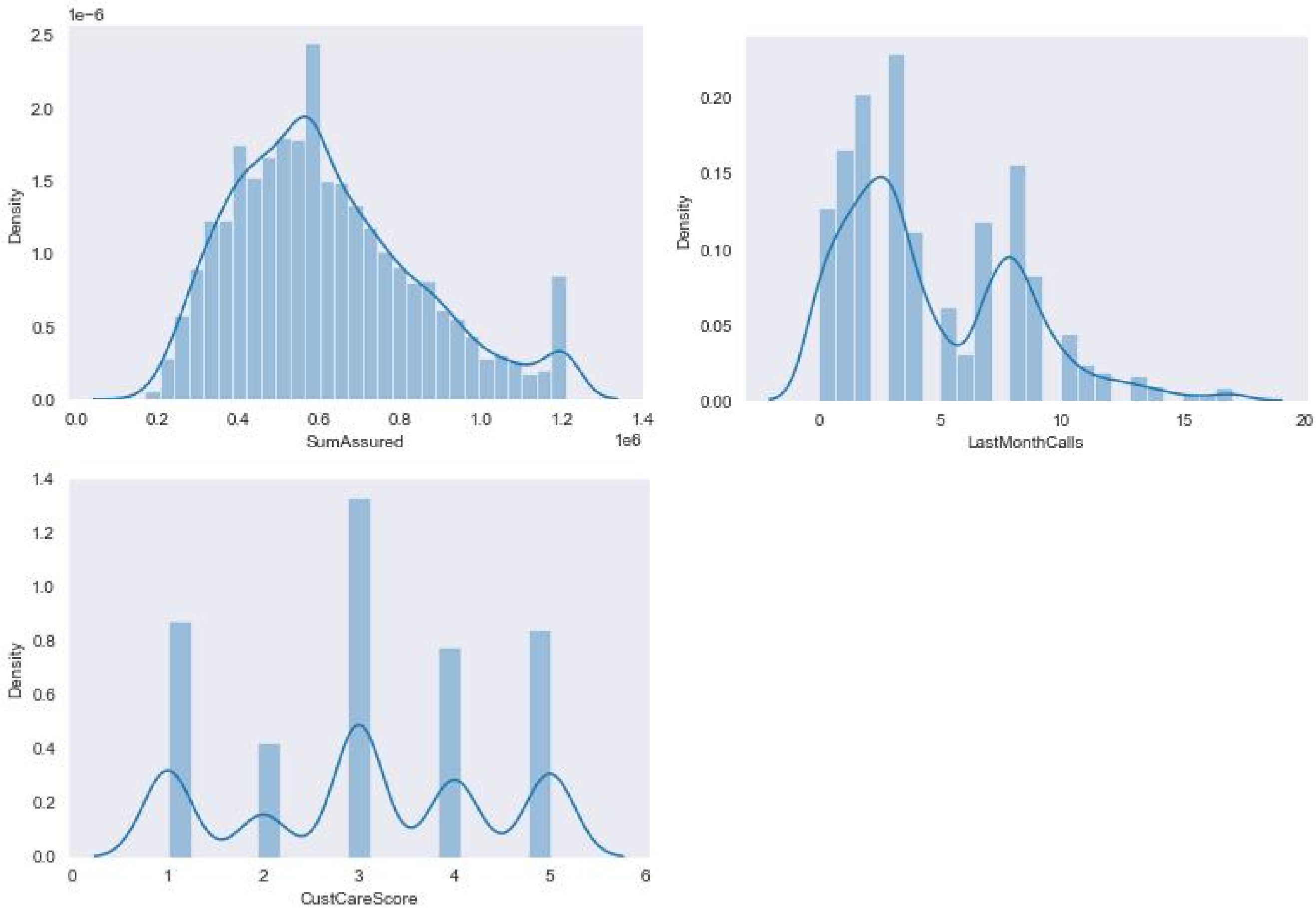
4: More Discrete Kind of data,4 is the most



5: Discontinuous Kind of data

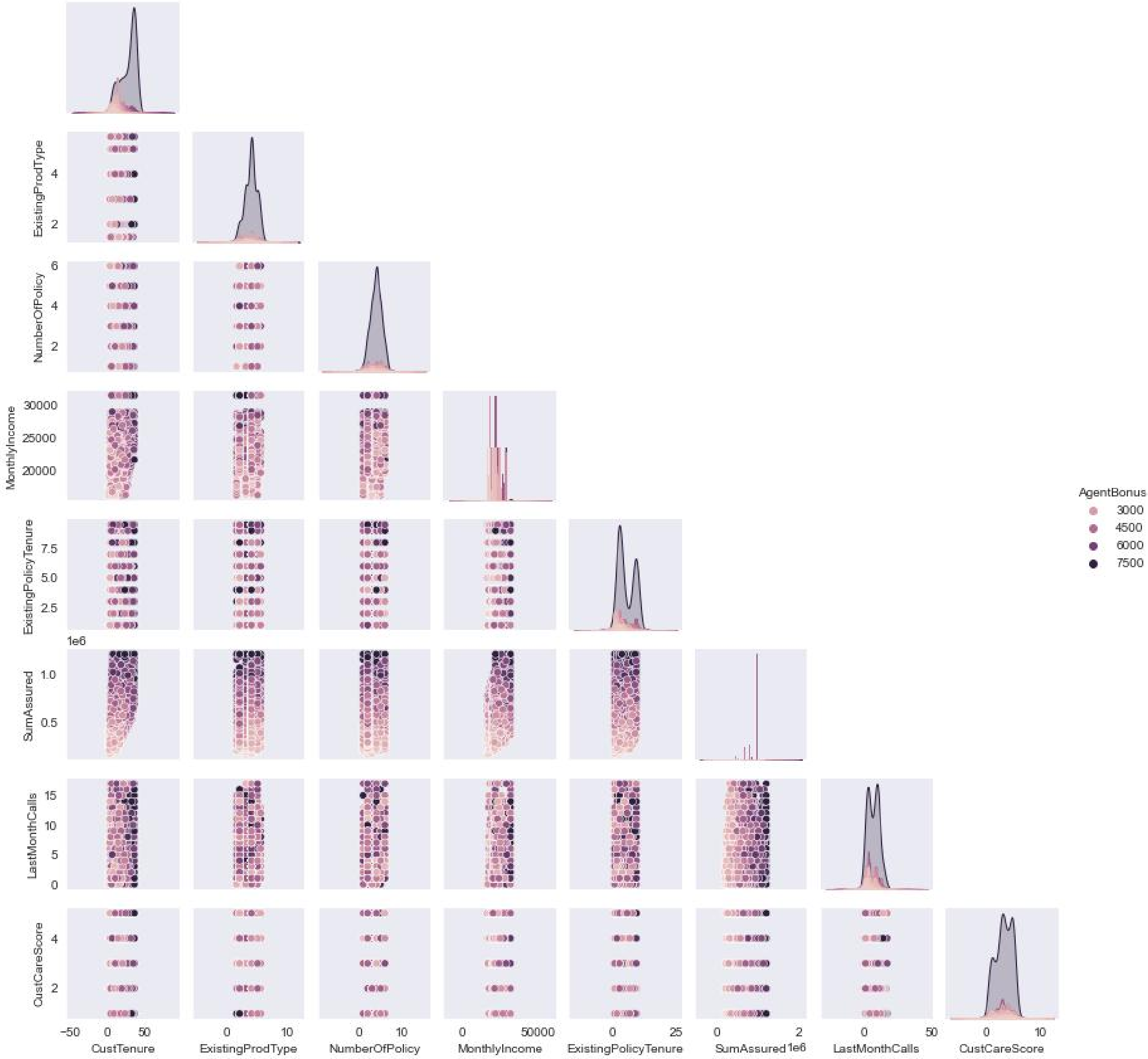


6: Discontinuous Kind of data



9: Continuous Kind of data

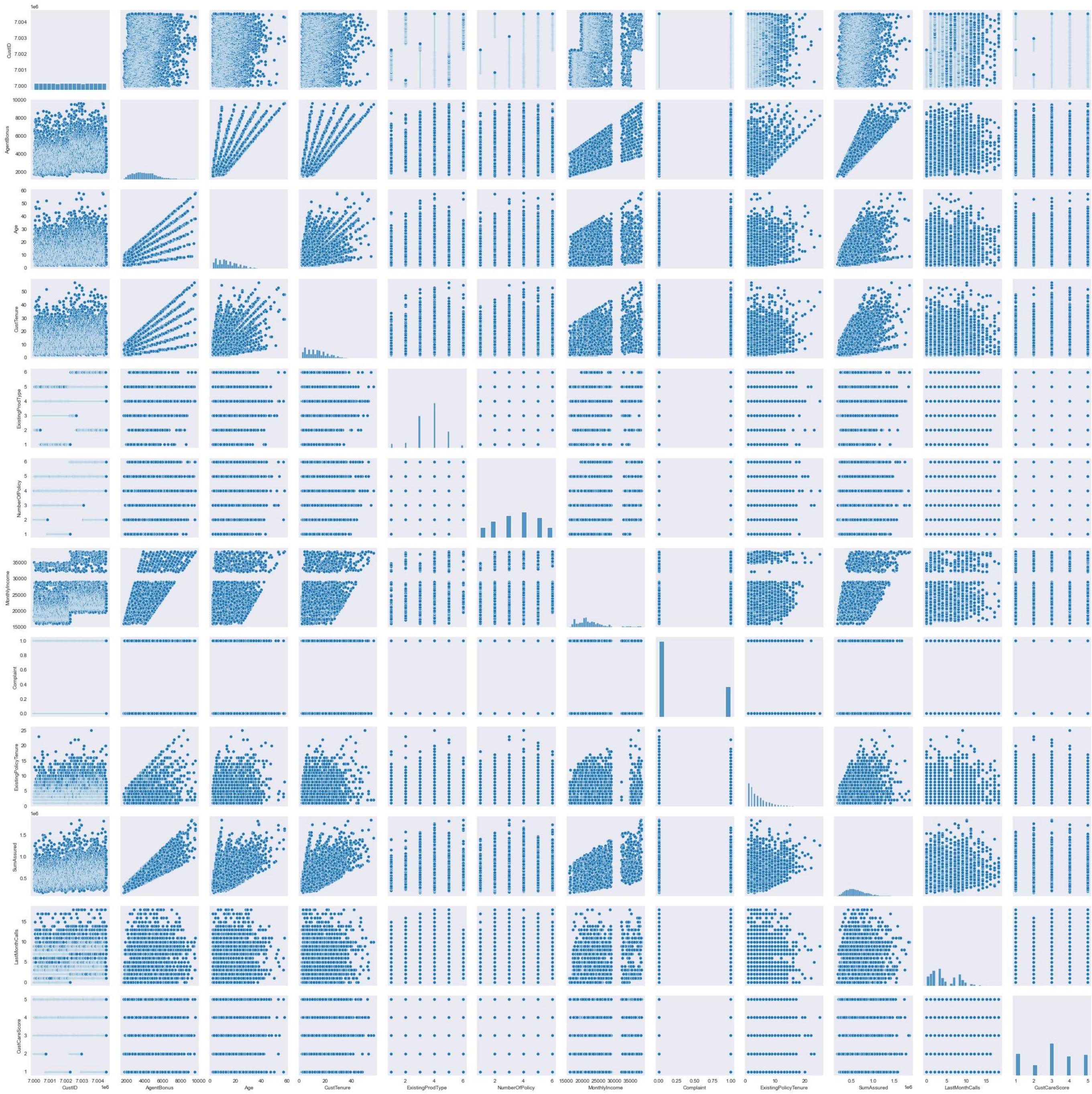
After Treatment there is no change in the UniVariate Analysis. Most of the numerical data is discrete since the nature of the domain is such. So even if the data seems continuous but is limited to a range.



Bivariate analysis (relationship between different variables , correlations)

Most of the variables don't seem to be related closely to each other which means there is low multi-collinearity in the data and each feature would have its importance in building the right model . Because of this we have not dropped any columns and would want to build the model to see the variable importance.

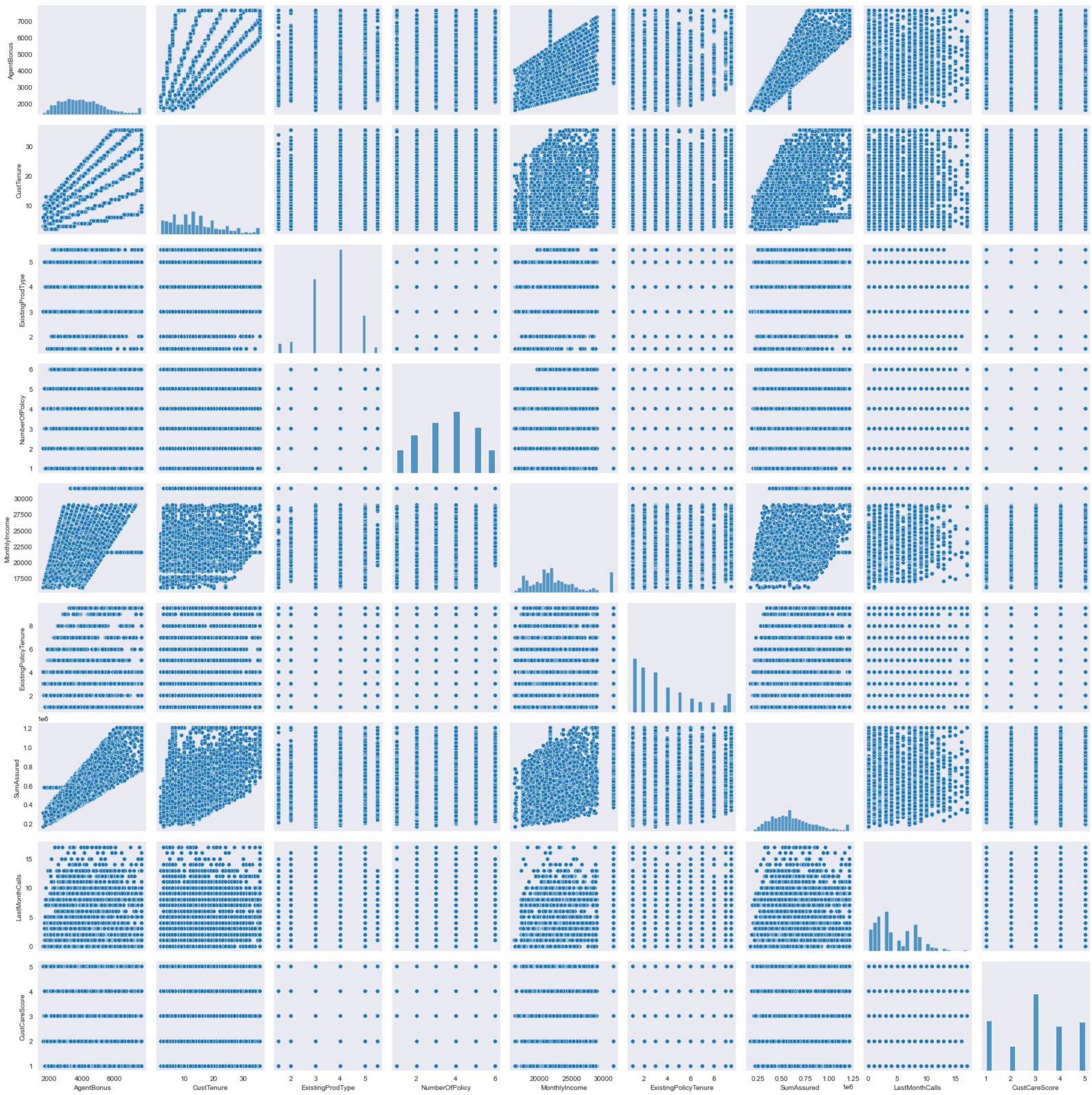
The pair plot also seems to suggest the same thing . But due to the huge number of columns, the pair plot was not providing very clear insight and hence resorted to bi variate plots with every combination possible.



Correlation matrix

	0	1	2	3	4	5	6	7	8	9
CustID	7000000	7000001	7000002	7000003	7000004	7000005	7000006	7000007	7000008	7000009
AgentBonus	4409	2214	4273	1791	2955	3252	3850	2073	2719	3247
Age	22.0	11.0	26.0	11.0	6.0	7.0	12.0	6.0	8.0	6.0
CustTenure	4.0	2.0	4.0	NaN	NaN	NaN	23.0	4.0	11.0	3.0
Channel	Agent	Third Party Partner	Agent	Third Party Partner	Agent	Third Party Partner	Agent	Agent	Agent	Online
Occupation	Salaried	Salaried	Free Lancer	Salaried	Small Business	Salaried	Salaried	Small Business	Salaried	Small Business
EducationField	Graduate	Graduate	Post Graduate	Graduate	UG	Graduate	Graduate	Under Graduate	Graduate	Under Graduate
Gender	Female	Male	Male	Fe male	Male	Male	Male	Female	Male	Male
ExistingProdType	3	4	4	3	3	3	4	3	4	2
Designation	Manager	Manager	Exe	Executive	Executive	Executive	VP	Executive	Manager	Exe
NumberOfPolicy	2.0	4.0	3.0	3.0	4.0	2.0	3.0	4.0	3.0	2.0
MaritalStatus	Single	Divorced	Unmarried	Divorced	Divorced	Single	Divorced	Unmarried	Divorced	Married
MonthlyIncome	20993.0	20130.0	17090.0	17909.0	18468.0	18068.0	34999.0	17279.0	20916.0	17089.0
Complaint	1	0	1	1	0	0	0	0	1	0
ExistingPolicyTenure	2.0	3.0	2.0	2.0	4.0	2.0	2.0	2.0	1.0	1.0
SumAssured	806761.0	294502.0	NaN	268635.0	366405.0	487836.0	392689.0	369079.0	405143.0	NaN
Zone	North	North	North	West	West	North	North	West	West	West
PaymentMethod	Half Yearly	Yearly	Yearly	Half Yearly	Half Yearly	Half Yearly	Yearly	Half Yearly	Yearly	Quarterly
LastMonthCalls	5	7	0	0	2	6	9	3	1	2
CustCareScore	2.0	3.0	3.0	5.0	5.0	5.0	2.0	3.0	4.0	4.0

Analysing after alteration of the Dataset



4. Quality of Data

4.1 Missing Value treatment (if applicable)

Before treating the values:

```
CustTenure              226
NumberOfPolicy           45
MonthlyIncome           236
ExistingPolicyTenure    184
SumAssured              154
CustCareScore            52
dtype: int64
```

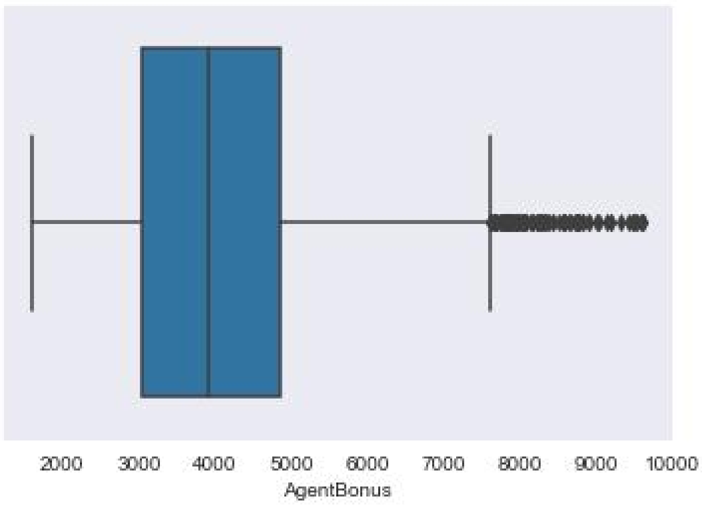
The missing values have been replaced with median and mode according to the pattern in the data set.For categorical data we have chosen MODE and for the numerical data we opted MEDIAN. The main reason for choosing mode or most frequent entry was it was making more sense considering to which the problem belongs.

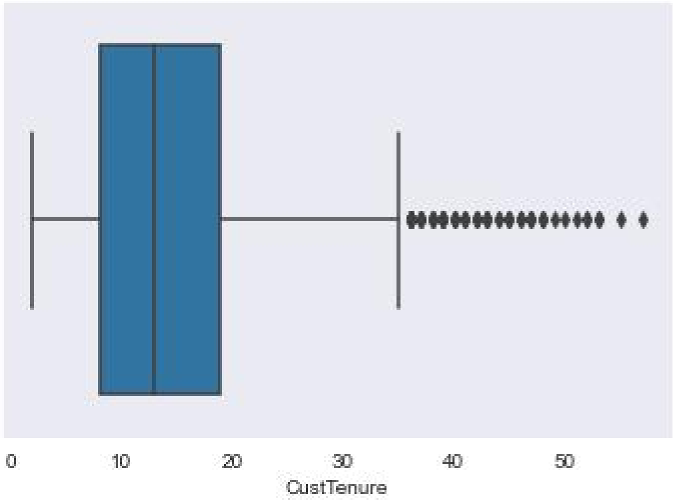
After treating the values:

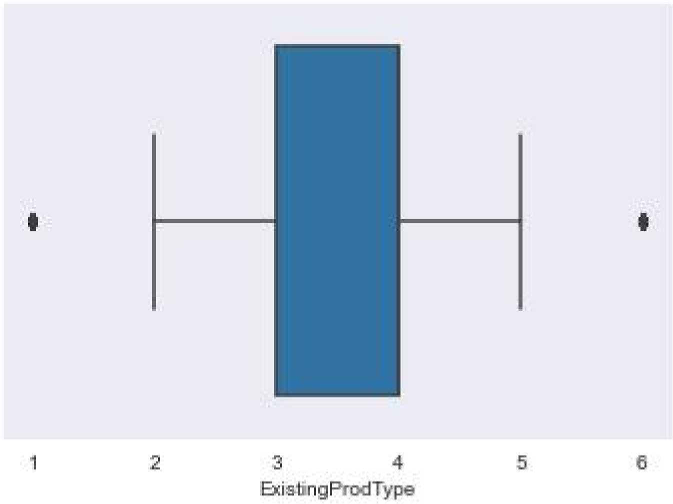
```
[ ]  df.isnull().sum()

     AgentBonus              0
     CustTenure              0
     Channel                 0
     ExistingProdType        0
     NumberOfPolicy          0
     MonthlyIncome           0
     Complaint               0
     ExistingPolicyTenure    0
     SumAssured              0
     Zone                    0
     PaymentMethod           0
     LastMonthCalls          0
     CustCareScore           0
     dtype: int64
```

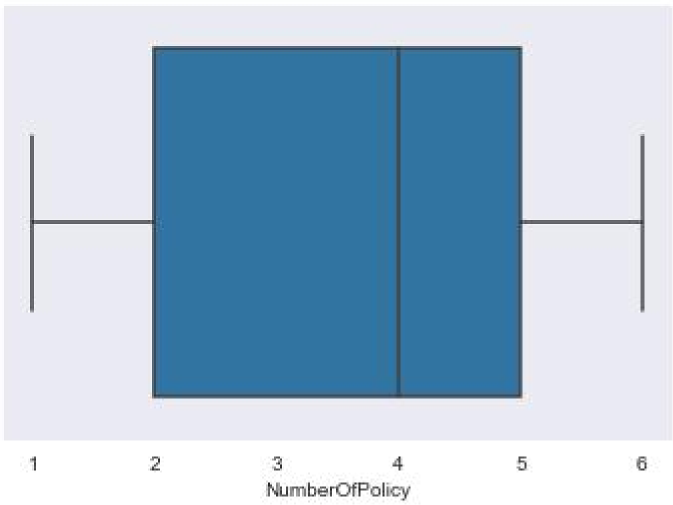
4.2 Outlier treatment (if required)

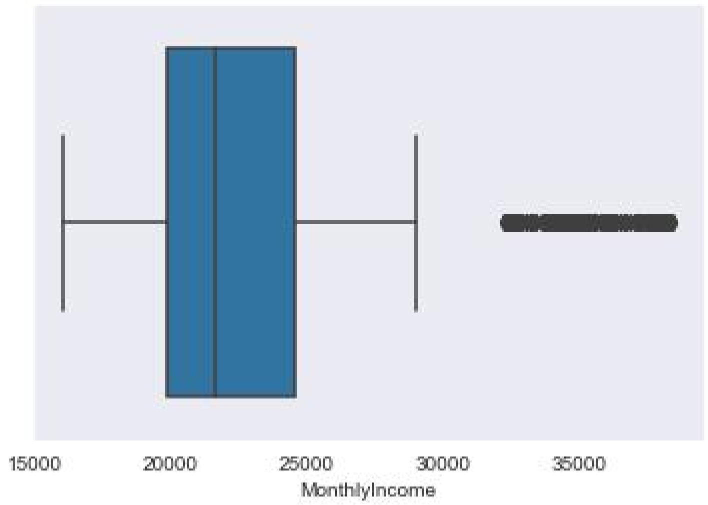
Before Outlier treatment

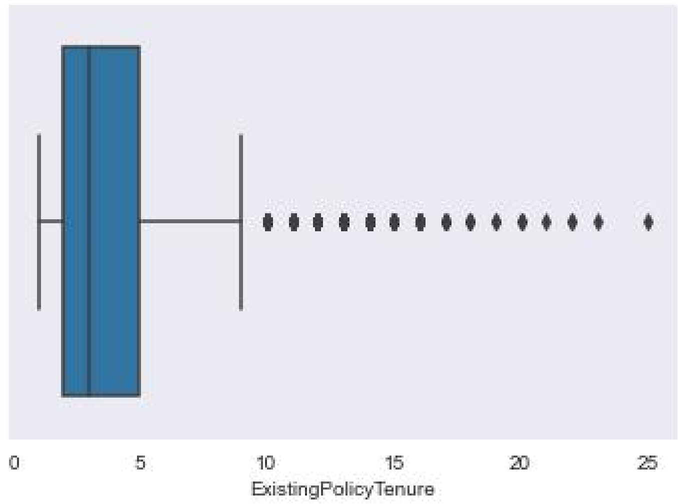


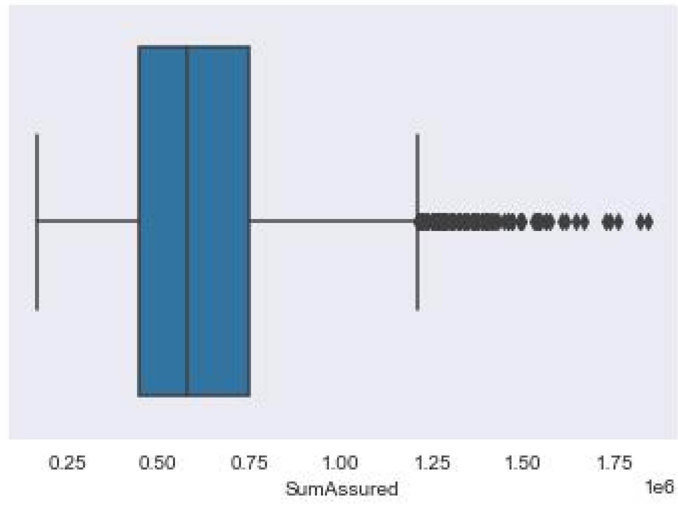

0
10
20
30
40
50
CustTenure

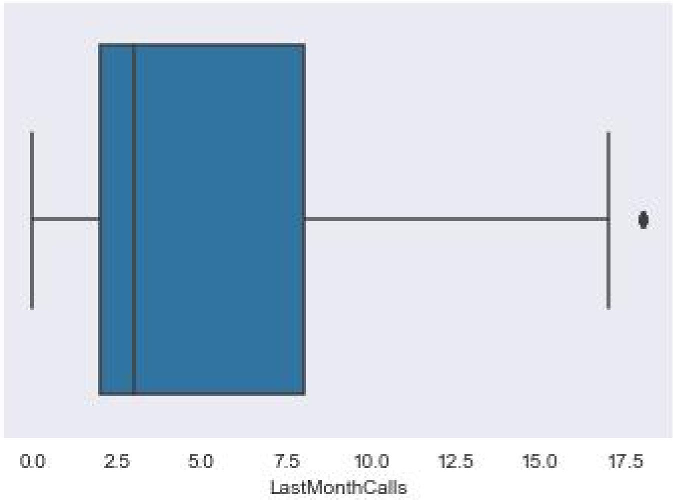

1
2
3
4
5
6
ExistingProdType

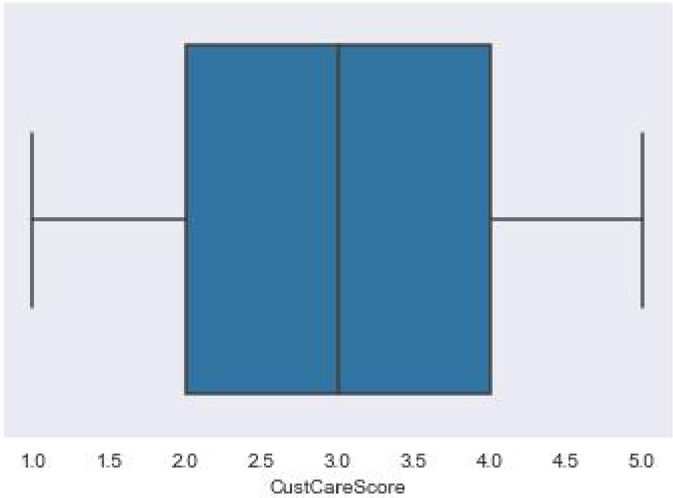

1
2
3
4
5
6
NumberOfPolicy


15000
20000
25000
30000
35000
MonthlyIncome

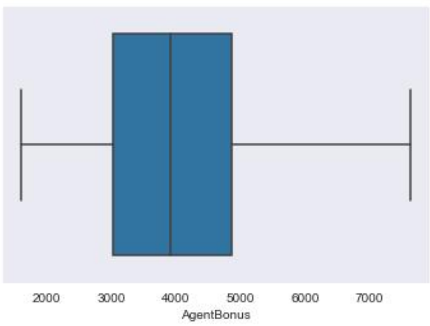

0
5
10
15
20
25
ExistingPolicyTenure

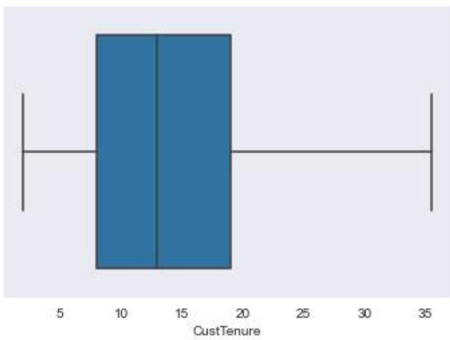

0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
SumAssured

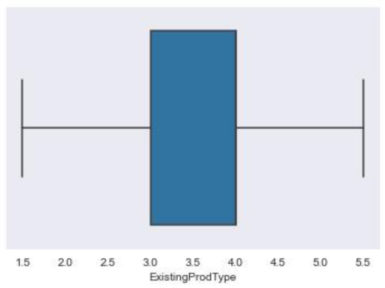


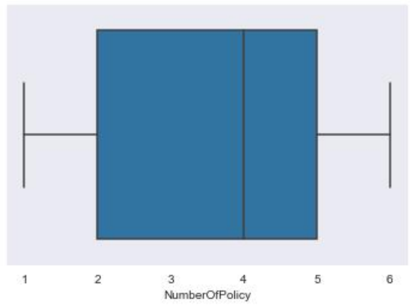


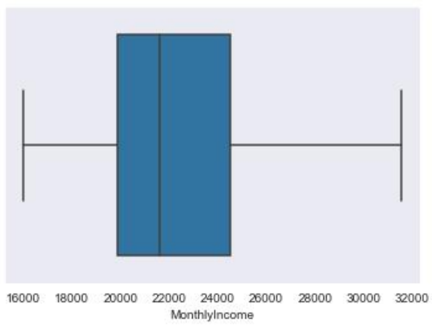
After Outlier treatment

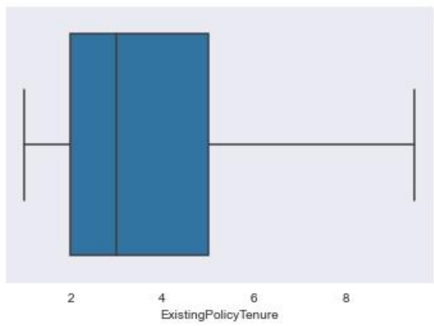


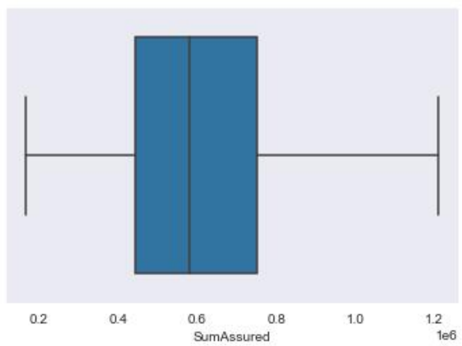


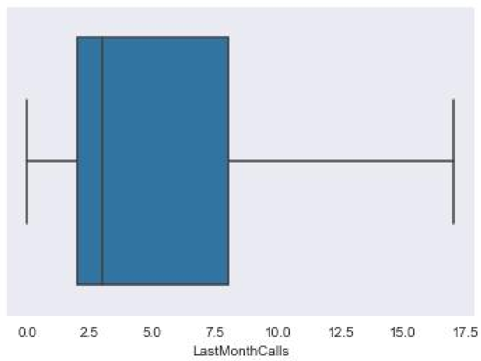


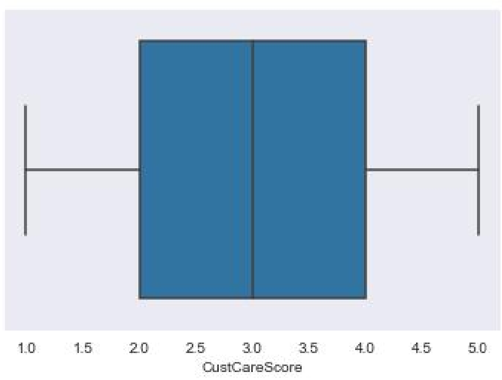












5. Feature Engineering

5.1 Removal of unwanted variables (if applicable)

CustID,Age,Occupation,EducationField,Gender,Designation,MaritalStatus are all redundant columns and have been removed. Chose not to remove any other columns and left to the model phase where the variable importance would be judged.

```
[ ] df.drop(['CustID','Age','Occupation','EducationField','Gender','Designation','MaritalStatus'],axis=1,inplace=True)
    df
```

5.2 Variable transformation (if applicable)

```
Occupation :  5
Free Lancer          2
Laarge Business    153
Large Business     255
Small Business    1918
Salaried          2192
Name: Occupation, dtype: int64


EducationField :  7
MBA                74
UG                230
Post Graduate     252
Engineer          408
Diploma           496
Under Graduate   1190
Graduate         1870
Name: EducationField, dtype: int64

Occupation :  5
Free Lancer          2
Laarge Business    153
Large Business     255
Small Business    1918
Salaried          2192
Name: Occupation, dtype: int64


Gender :  3
Fe male     325
Female     1507
Male       2688
Name: Gender, dtype: int64


Designation :  6
Exe               127
VP                226
AVP               336
Senior Manager    676
Executive        1535
Manager          1620
Name: Designation, dtype: int64
```

The highlighted data seems to be recorded incorrectly and required replacement and this was done to ensure the right categories are picked up by the model

In Gender Column “Female” is misspelled as “Fe male”. So we replace “Fe male” as “Female”

```
df['Gender'] = df['Gender'].replace(['Fe male'],'Female')
print('Gender',': ',df['Gender'].nunique())
print(df['Gender'].value_counts().sort_values())
print('\n')
```

In Occupation Column “Large” is misspelled as “Laarge”. So we replace “Laarge” as “Large”

```
df['Occupation'] = df['Occupation'].replace(['Laarge Business'],'Large Business')
print('Occupation',': ',df['Occupation'].nunique())
print(df['Occupation'].value_counts().sort_values())
print('\n')
```

In the Designation Column “Exe” and “Executive” refer the same.So we replace “Exe” as “Executive”

```
df['Designation'] = df['Designation'].replace(['Exe'],'Executive')
print('Designation',': ',df['Designation'].nunique())
print(df['Designation'].value_counts().sort_values())
print('\n')
```

In the Education Field Column “Under Graduate” and “UG” refer the same.So we replace “UG” as “Undergraduate”

```
df['EducationField'] = df['EducationField'].replace(['UG'],'Under Graduate ')
print('EducationField',': ',df['EducationField'].nunique())
print(df['EducationField'].value_counts().sort_values())
print('\n')
```

The variables has been encoded to numeric values from Categorical data for the following variables :-

```
df["Complaint"]= pd.Categorical(df['Complaint'])
df["Channel"]= pd.Categorical(df['Channel'])
df["Zone"]= pd.Categorical(df['Zone'])
df["PaymentMethod"]= pd.Categorical(df['PaymentMethod'])
```

5.3 Addition of new variables (if required)

No new variables were added at this stage .

6. Business Insights from EDA

6.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Total Agents = 4520

Low performance agent according to the AgentBonus 1130

High performance agent according to the AgentBonus 3390

Data is not balanced with more high performance than low performance but that would be the nature of each agent such that to perform better more and more on each sale so as to get more bonus and hence this should be the way the data is expected . Don't see any treatment on this would be needed .

As the problem statement is to predict the bonus of the agents so it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents.

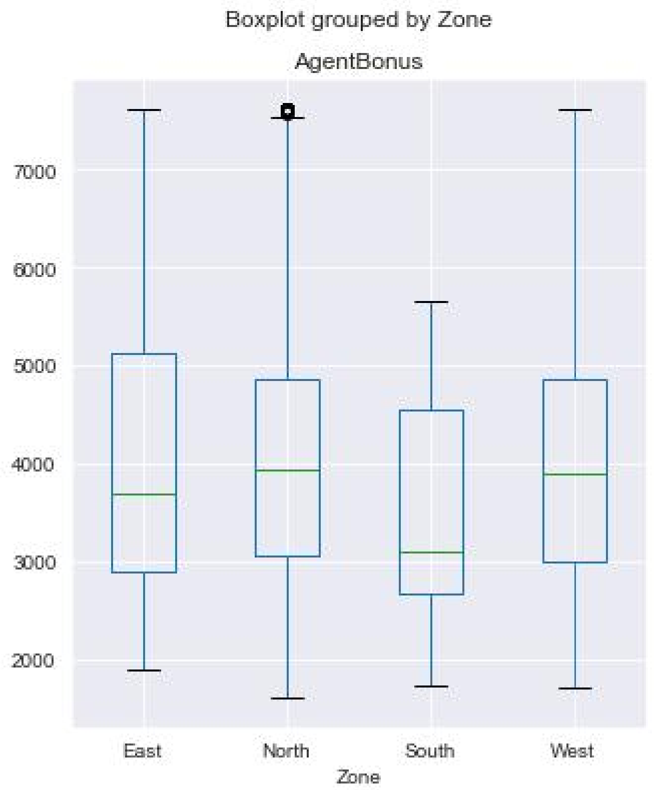
6.2 Any business insights using clustering (if applicable)

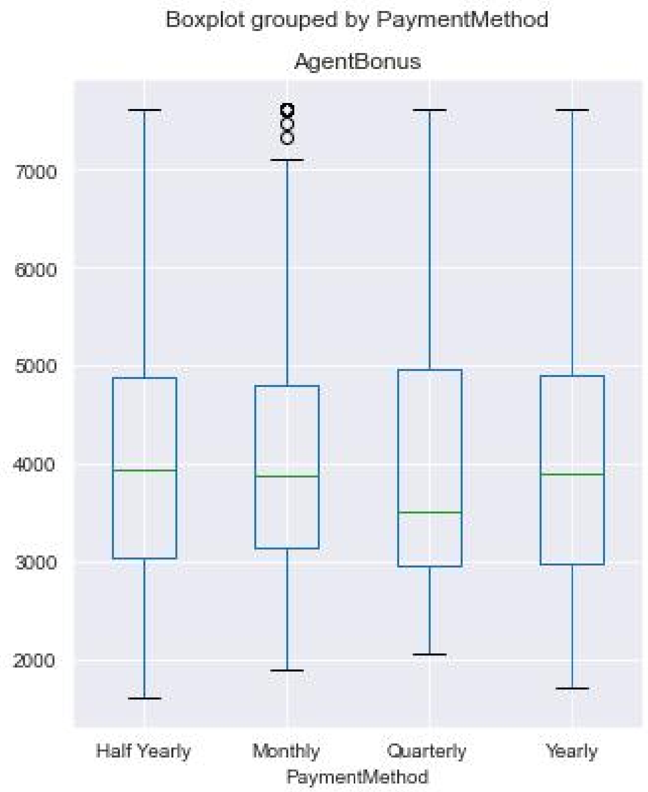
According to the observation, the North Zone will get the most Agent Bonus while the South Zone will get the least.

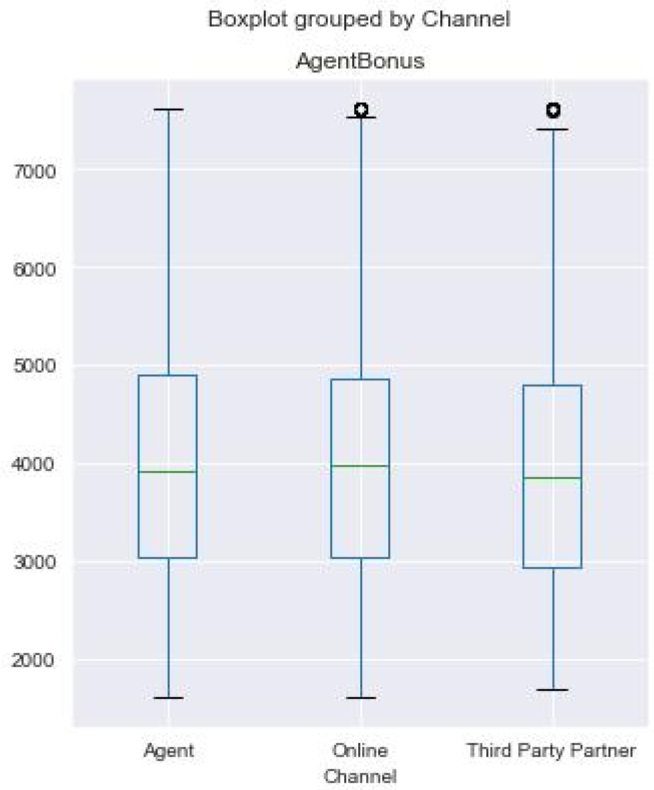
According to the observation, the Half-Yearly will get the most Agent Bonus while the Quarterly will get the least.

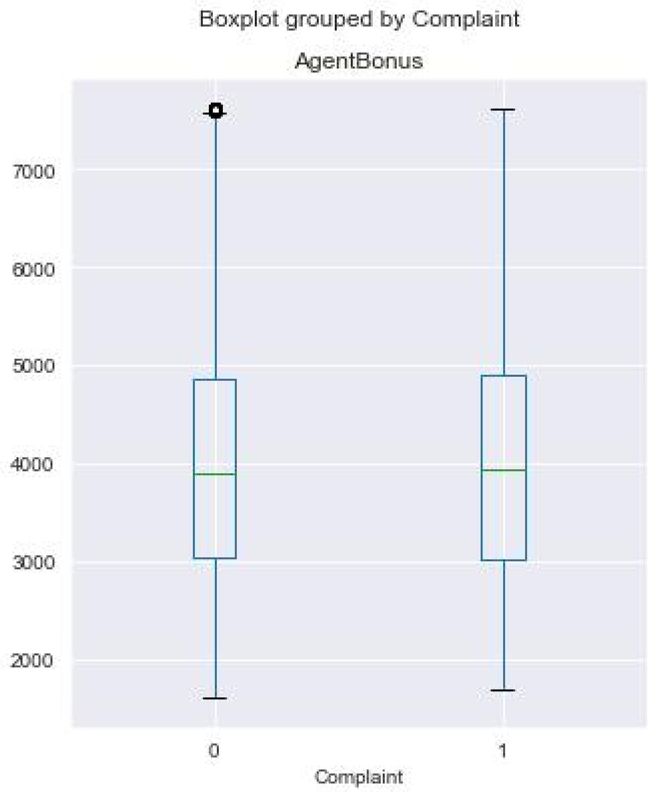
According to the observation, the Online Channel will get the most Agent Bonus while the Third-Party will get the least.

According to the observation, the Agent having one complaint will get the most Agent Bonus while the Agent having zero complaint will get the least.

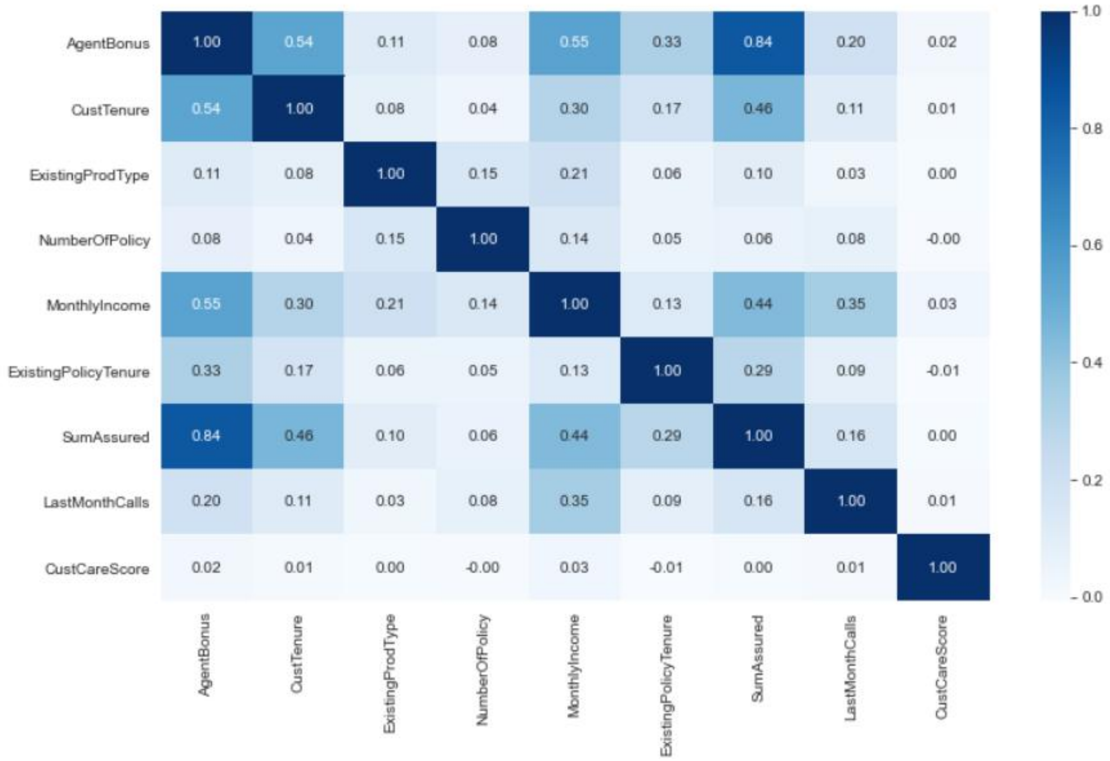








6.3 Any other business insights



- Agent Bonus having positively correlated with **SumAssured** and having least correlated with **Customer Care Score**.
- Quest Tenure having positively correlated with **AgentBonus** and having least correlated with **Customer Care Score.**
- Existing Product Type having positively correlated with **MonthlyIncome** and having least correlated with **Customer Care Score.**
- Number Of Policy having positively correlated with **ExistingProdType** and having least correlated with **Customer Care Score**.
- Monthly Income having positively correlated with **AgentBonus** and having least correlated with **Customer Care Score**.
- Existing PolicyTenure having positively correlated with **AgentBonus** and having least correlated with **Customer Care Score.**
- Sum Assured having positively correlated with **Sum Assured** and least correlated with **Customer Care Score**.
- Last Month Calls having positively correlated with Monthly Income and least correlated with **Customer Care Score**.
- Customer Care Score correlates best positively with **Monthly Income** and least correlates to **Existing Policies Tenure**.

From above observation least correlated is Customer Care score.