Assignment - Data Science
BUSINESS REPORT

NAME: MEGHA SINGH
ROLL NUMBER: 1928300
BRANCH: CSSE

Contents

1. Introduction to Problem statement

1.1 Problem statement

Objective of the problem statement is to predict the bonus awarded to its agents so that they can design appropriate engagement programs for their high performer agents and upskill programs for their low performers.

1.2 Need of the study

To analyze, interpret and deliver data in meaningful ways to help the business increase. Therefore it helps in productivity and enables effective decision-making.

1.3 Understanding business/social opportunity

Business owners, managers and co-workers will benefit by becoming highly skilled and more growth oriented.

2. Data Report

2.1 Data description

The dataset belongs to a leading life insurance company. It contains past information about its clients, as well as bonuses its agents have received.

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

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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 20 columns):
 #   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
```

Descriptive details: Central Tendency

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
CustID	4520.00	NaN	NaN	NaN	7002259.50	1304.96	7000000.00	7001129.75	7002259.50	7003389.25	7004519.00
AgentBonus	4520.00	NaN	NaN	NaN	4077.84	1403.32	1605.00	3027.75	3911.50	4867.25	9608.00
Age	4251.00	NaN	NaN	NaN	14.49	9.04	2.00	7.00	13.00	20.00	58.00
CustTenure	4294.00	NaN	NaN	NaN	14.47	8.96	2.00	7.00	13.00	20.00	57.00
Channel	4520	3	Agent	3194	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Occupation	4520	5	Salaried	2192	NaN	NaN	NaN	NaN	NaN	NaN	NaN
EducationField	4520	7	Graduate	1870	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Gender	4520	3	Male	2688	NaN	NaN	NaN	NaN	NaN	NaN	NaN
ExistingProdType	4520.00	NaN	NaN	NaN	3.69	1.02	1.00	3.00	4.00	4.00	6.00
Designation	4520	6	Manager	1620	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NumberOfPolicy	4475.00	NaN	NaN	NaN	3.57	1.46	1.00	2.00	4.00	5.00	6.00
MaritalStatus	4520	4	Married	2268	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthlyIncome	4284.00	NaN	NaN	NaN	22890.31	4885.60	16009.00	19683.50	21606.00	24725.00	38456.00
Complaint	4520.00	NaN	NaN	NaN	0.29	0.45	0.00	0.00	0.00	1.00	1.00
ExistingPolicyTenure	4336.00	NaN	NaN	NaN	4.13	3.35	1.00	2.00	3.00	6.00	25.00
SumAssured	4366.00	NaN	NaN	NaN	619999.70	246234.82	168536.00	439443.25	578976.50	758236.00	1838496.00
Zone	4520	4	West	2566	NaN	NaN	NaN	NaN	NaN	NaN	NaN
PaymentMethod	4520	4	Half Yearly	2656	NaN	NaN	NaN	NaN	NaN	NaN	NaN
LastMonthCalls	4520.00	NaN	NaN	NaN	4.63	3.62	0.00	2.00	3.00	8.00	18.00
CustCareScore	4468.00	NaN	NaN	NaN	3.07	1.38	1.00	2.00	3.00	4.00	5.00

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

Variable	Count	Dtype	Remarks
CustID	4520	int64	Numeric
AgentBonus	4520	int65	Numeric
Age	4251	float64	Numeric
CustTenure	4294	float65	Numeric
Channel	4520	object	Categorical
Occupation	4520	object	Categorical
EducationField	4520	object	Categorical
Gender	4520	object	Categorical
ExistingProdType	4520	int64	Numeric
Designation	4520	object	Categorical
NumberOfPolicy	4475	float64	Numeric
MaritalStatus	4520	object	Categorical
MonthlyIncome	4284	float64	Numeric
Complaint	4520	int64	Numeric
ExistingPolicyTenure	4336	float64	Numeric
SumAssured	4366	float64	Numeric
Zone	4520	object	Categorical
PaymentMethod	4520	object	Categorical
LastMonthCalls	4520	int64	Numeric
CustCareScore	4468	float64	Numeric

The name of the columns seems to be fine with no special characters or spaces between them and might delete the column if required while analyzing through the graph. There are no duplicate values in any column.

Unique values of various Categories

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


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
```

```
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
```

As it appears the highlighted data was recorded incorrectly and needed to be replaced, this was done to ensure the right categories would be picked up by the model.

Post fixing of the data

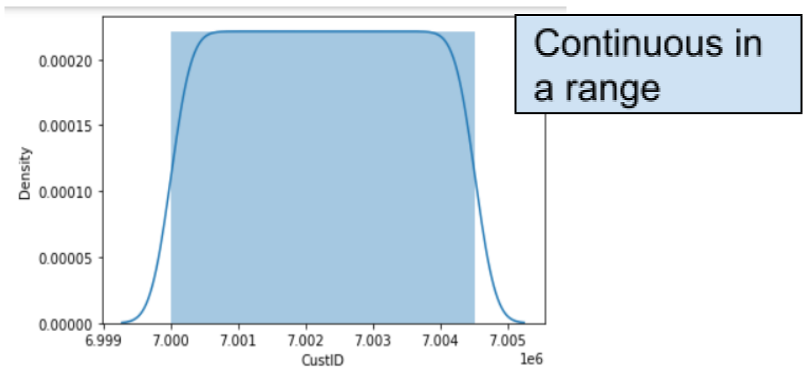
```
Channel :  3
Online                  468
Third Party Partner     858
Agent                  3194
Name: Channel, dtype: int64


Occupation :  4
Free Lancer            2
Large Business       408
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 :  2
Female    1832
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
```

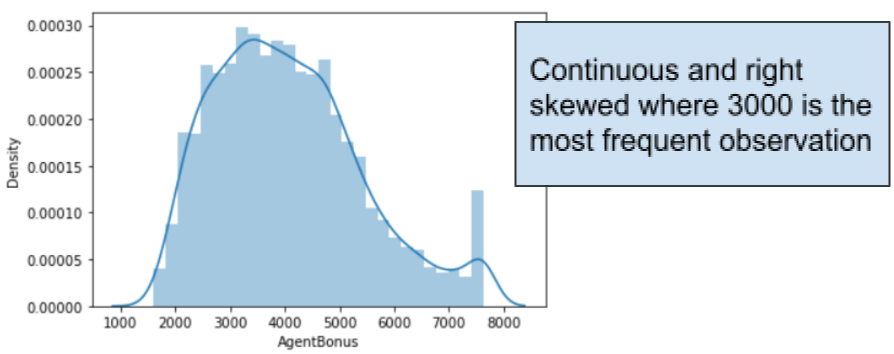
As a result of fixing the two highlighted columns, we can see the total number of females is 1832 and the total number of large businesses is 408.

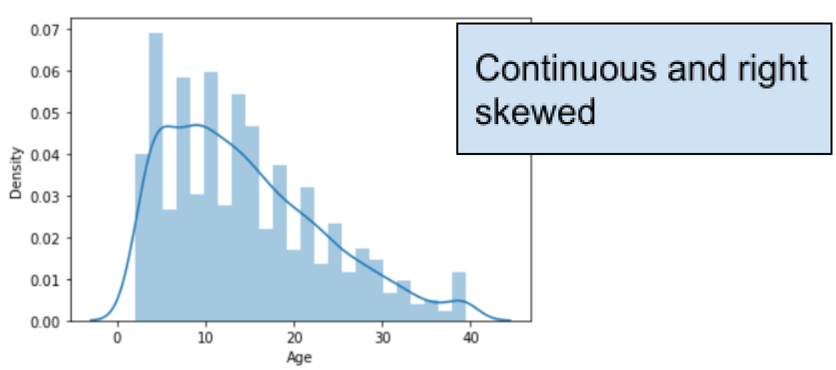
3. Exploratory Data Analysis

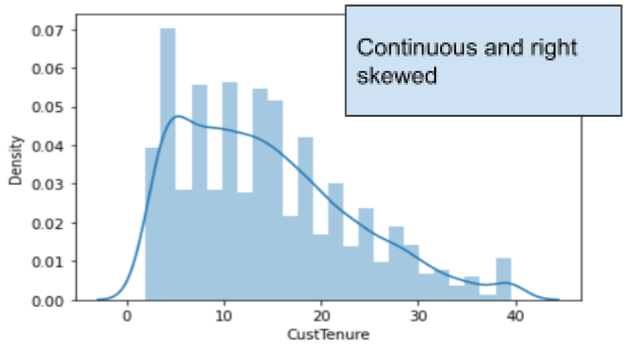
3.1 Univariate analysis

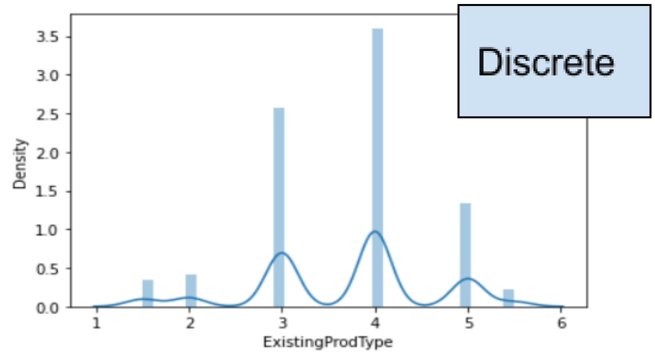
Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones).

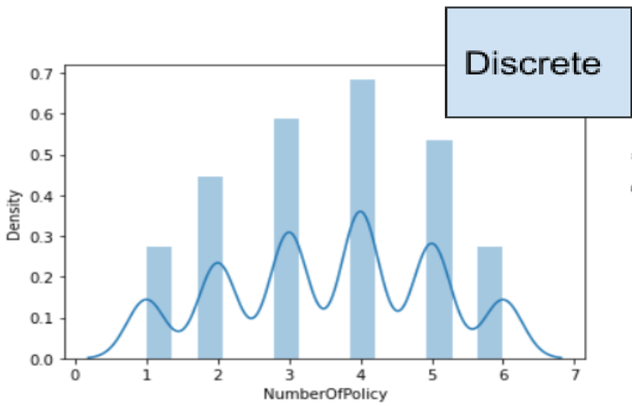


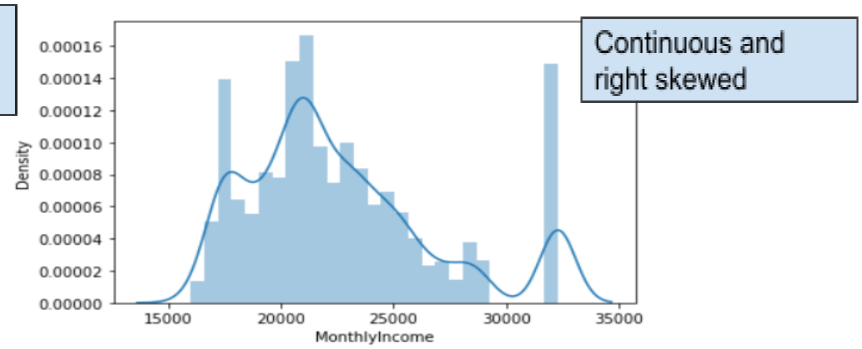


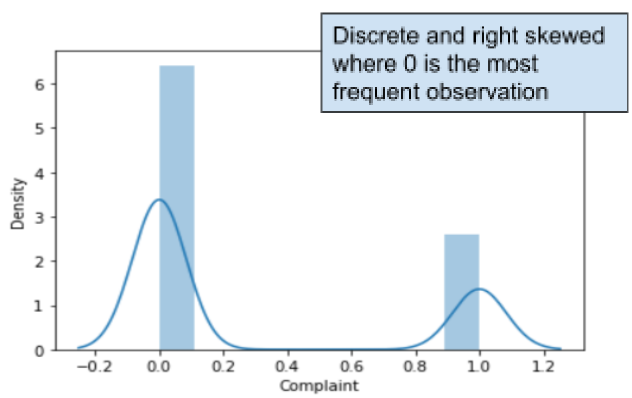


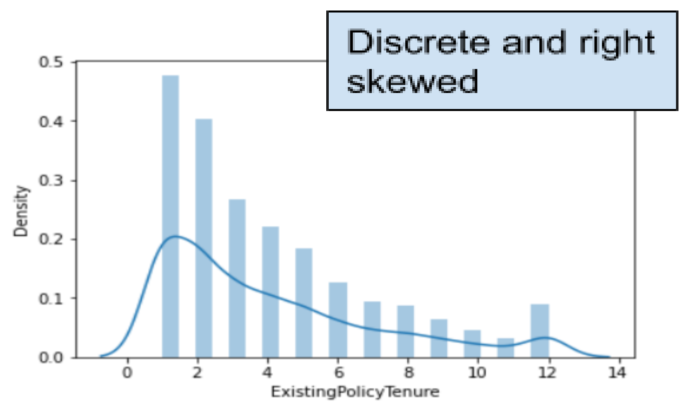

Continuous and right skewed
Density
CustTenure

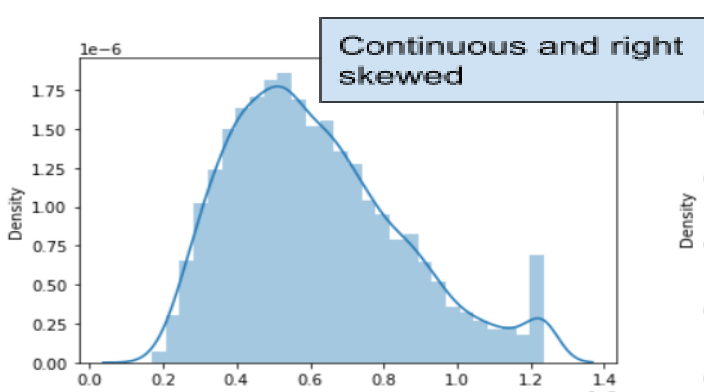

Discrete
Density
ExistingProdType

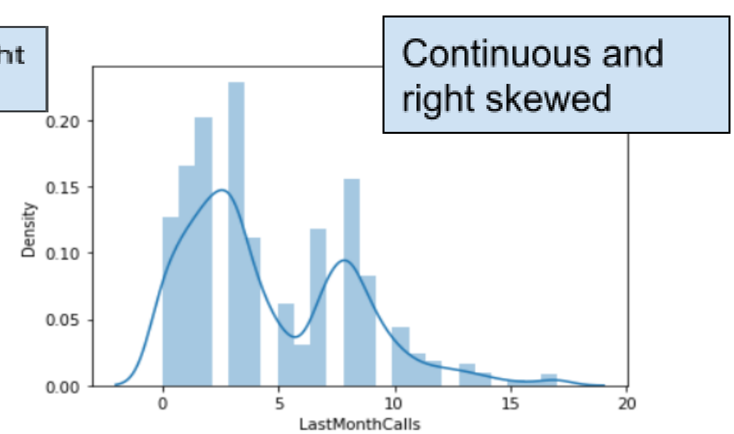

Discrete
Density
NumberOfPolicy


Continuous and right skewed
Density
MonthlyIncome


Discrete and right skewed where 0 is the most frequent observation
Density
Complaint

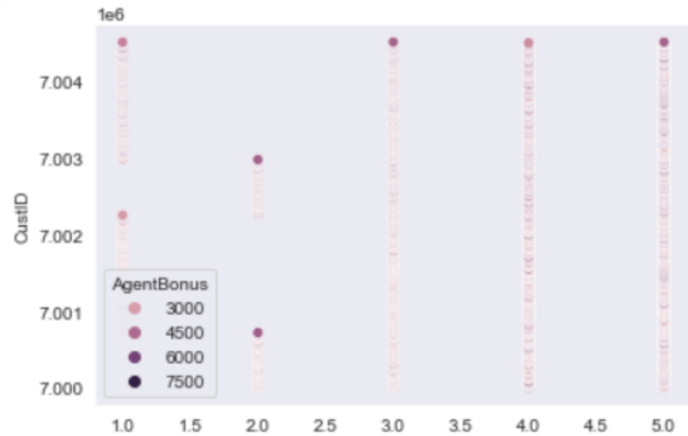

Discrete and right skewed
Density
ExistingPolicyTenure


Continuous and right skewed
Density
SumAssured

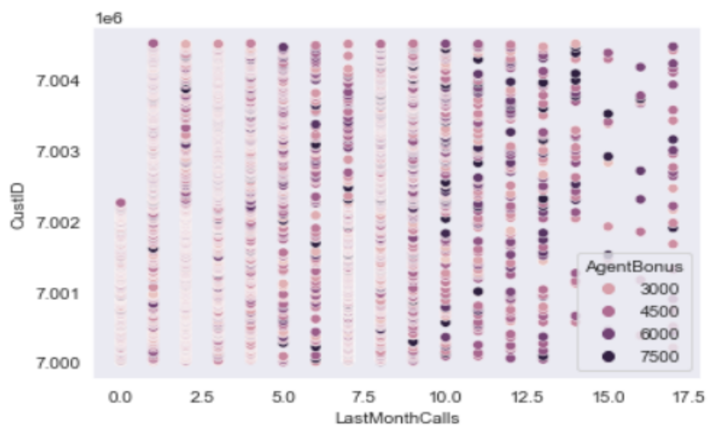

Continuous and right skewed
Density
LastMonthCalls

Since the nature of the domain is continuous, the numerical data is largely continuous. In addition, the bonus that agents receive will also depend upon other factors such as the sum assured by the customer, gross monthly income of the customer and customer tenure. And the highest bonus received by agents is 3000.

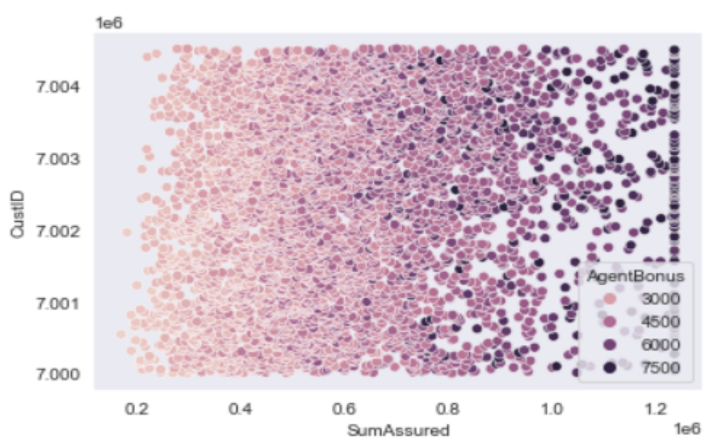
3.2 Bivariate Analysis



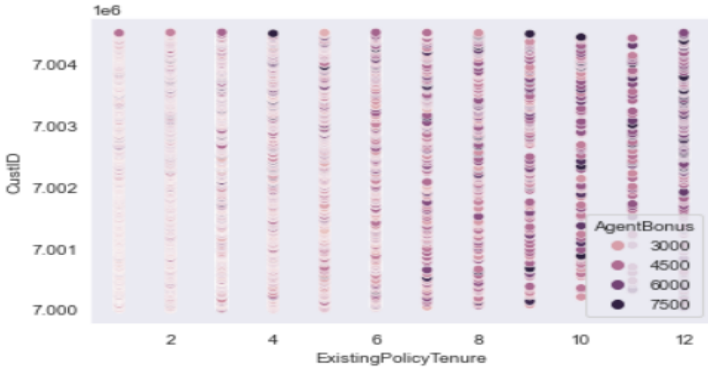
Average bonus is 3000 irrespective of CustID and CustCareScore



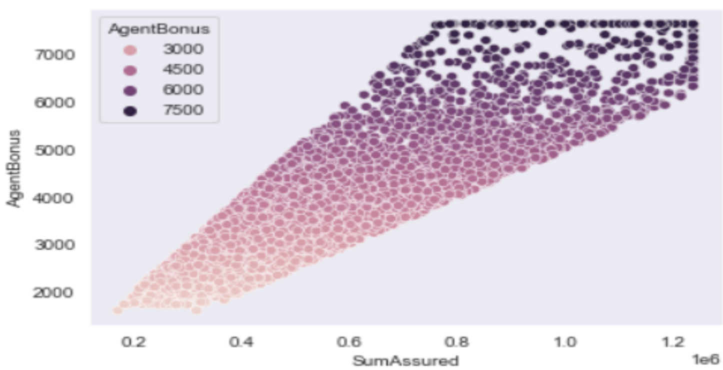
Weakly related



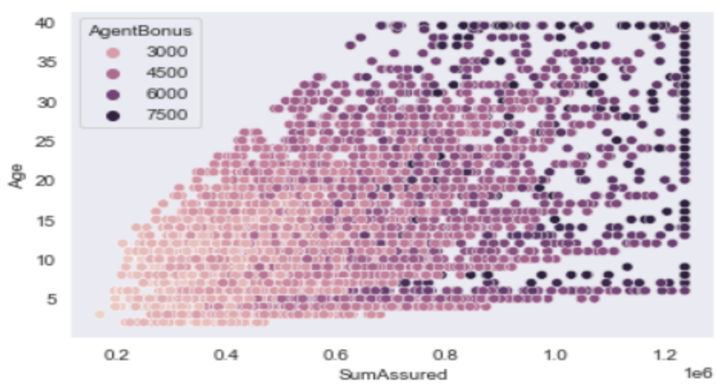
In tandem with both variables increasing continuously, the agent bonus is increasing as well.



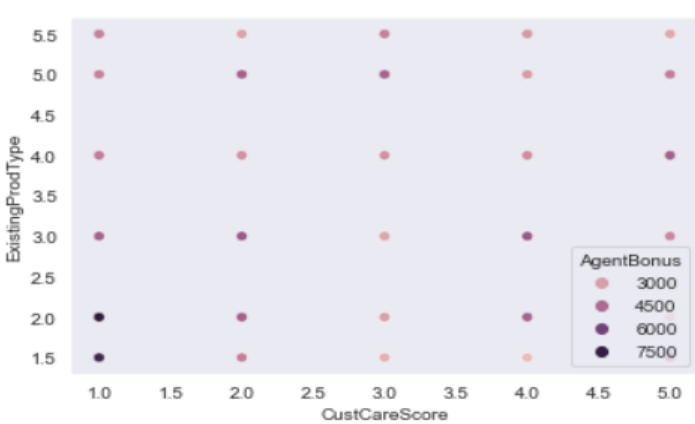
Increasing policy tenure means more customers, as a result there are increased bonuses for agents



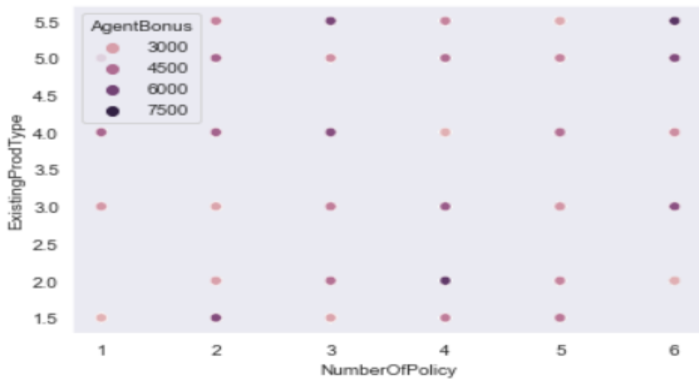
Positively related



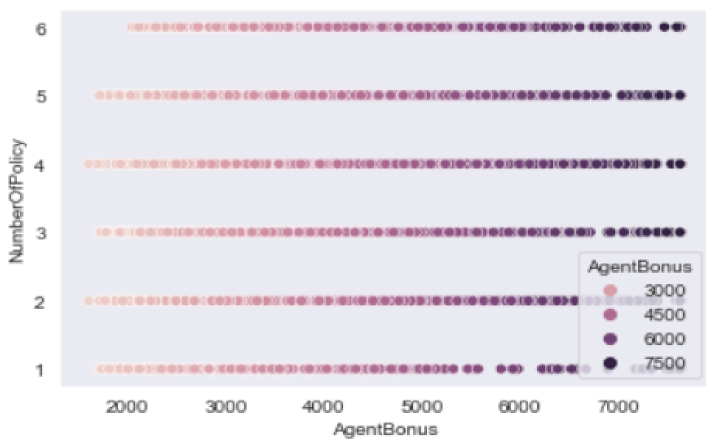
Positively related



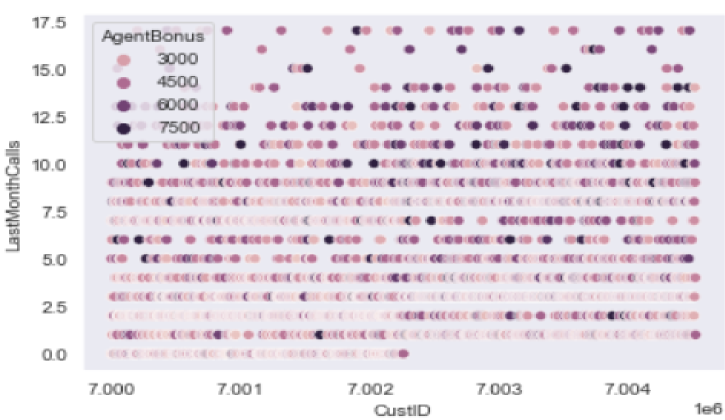
Not able to establish any relation between variables



Not able establish any relation between variables

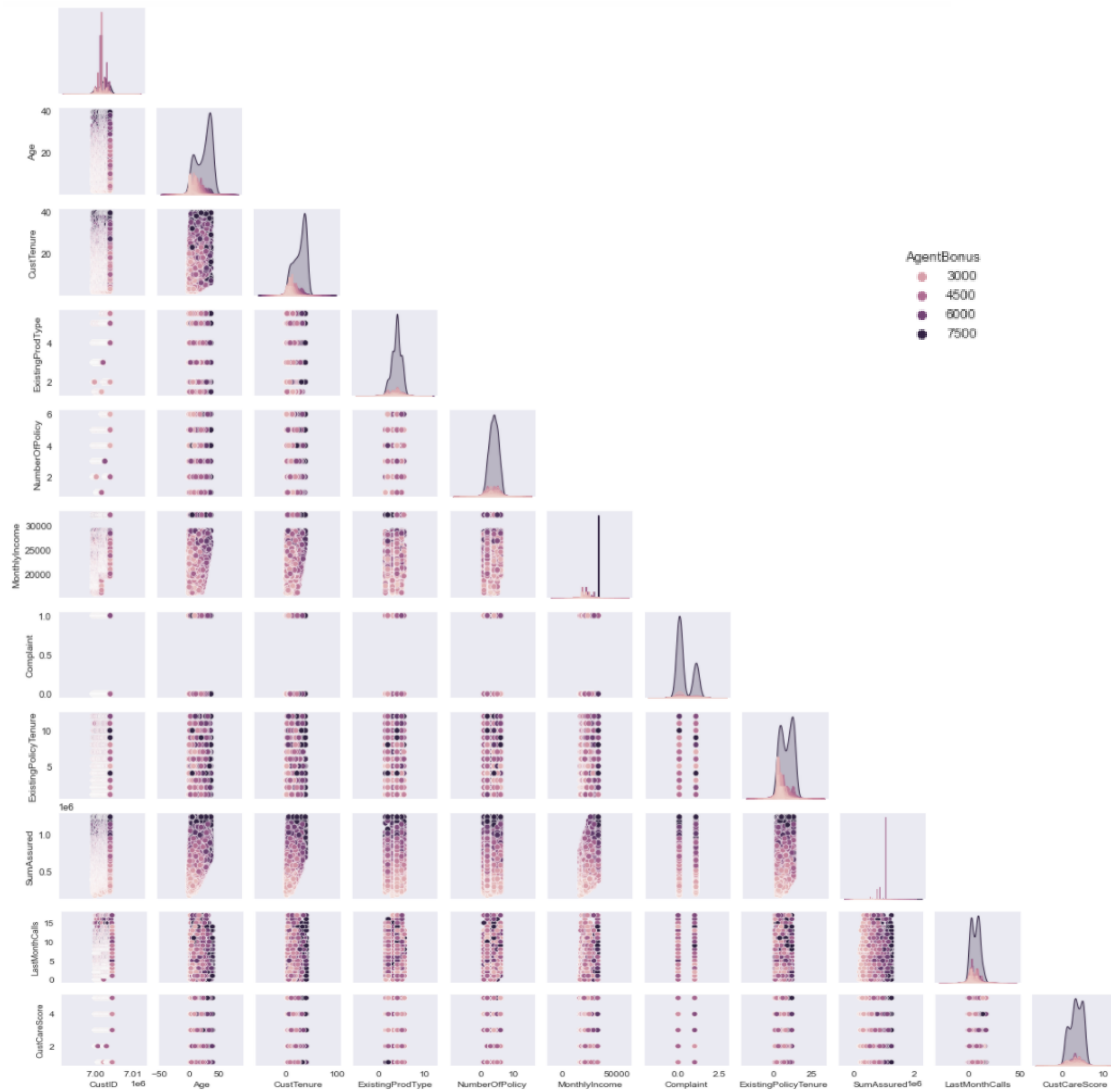


Positively related



Weakly related

Most of the variables are related to each other which means there is multi- collinearity in the data and each feature would have its importance in building the right model .It is not necessary to drop any columns since we would want to build the model to determine the variable importance. The pair plot also seems to suggest the same thing . Nevertheless, the large number of columns made the pair plot unable to provide very clear insight and so we resorted to bi-variate plots with every possible combination.



3.3 Removal of unwanted variables

Removal of columns is required because there is no relation between complaint and CustCareScore with other variables.

```
df.drop(['Complaint'],axis=1,inplace=True)
```

```
df.drop(['CustCareScore'],axis=1,inplace=True)
```

3.4 Missing Value Treatment

```
Age                     269
MonthlyIncome           236
CustTenure              226
ExistingPolicyTenure    184
SumAssured              154
CustCareScore            52
NumberOfPolicy           45
Gender                    0
ExistingProdType          0
Designation               0
AgentBonus                0
MaritalStatus             0
EducationField            0
Complaint                 0
Occupation                0
Channel                   0
Zone                      0
PaymentMethod             0
LastMonthCalls            0
CustID                    0
dtype: int64
```

The missing values have been treated with more frequent values than median for numeric data including categorical data. The primary reason for choosing mode or most frequent entry was that it made more sense on the basis of the sales domain. Furthermore, the numeric data has continuous patterns, so it has been treated as quantitative data as we have seen in the various plots.

```
from sklearn.impute import SimpleImputer
imputer = SimpleImputer(strategy='most_frequent',missing_values=np.nan)
```

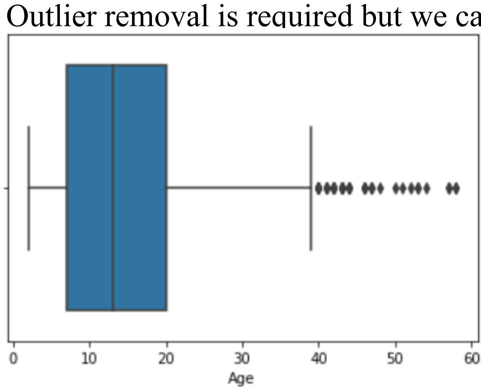
executed in 721ms, finished 00:36:21 2022-03-31

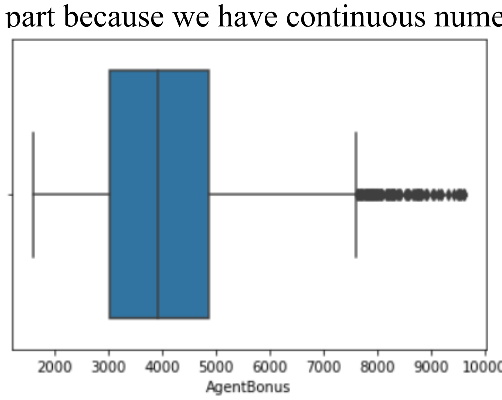
```
for i,col_val in enumerate(list(df.columns)):
    if df[col_val].isnull().sum()>0 :
        df[col_val]=imputer.fit_transform(df[col_val].values.reshape(-1,1))[:,0]
```

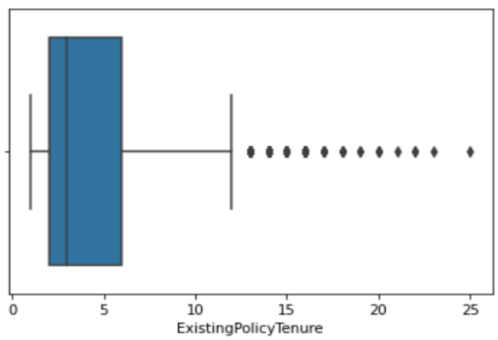
executed in 22ms, finished 00:36:37 2022-03-31

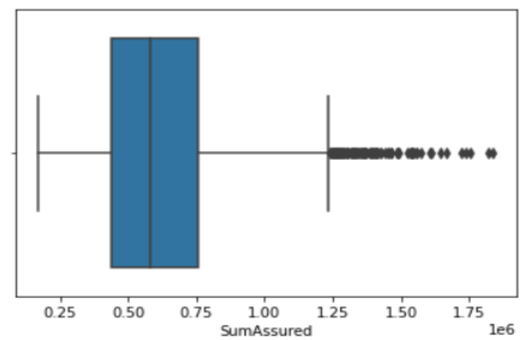
3.5 Outlier treatment

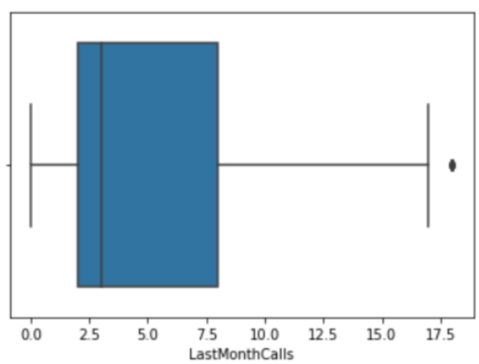
Outlier removal is required but we can skip this part because we have continuous numeric data.

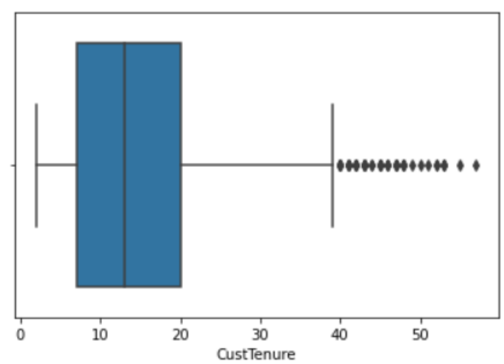


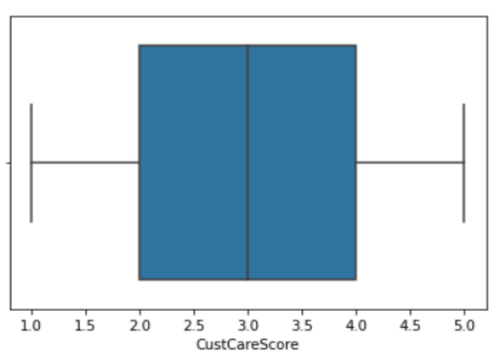


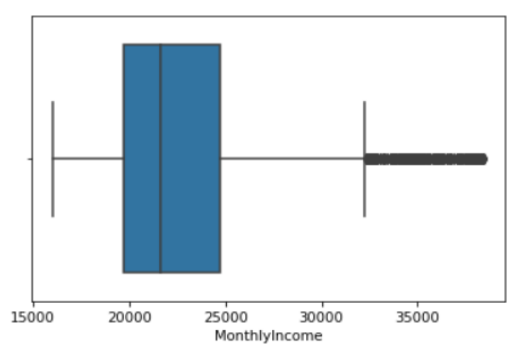

0 5 10 15 20 25
ExistingPolicyTenure


0.25 0.50 0.75 1.00 1.25 1.50 1.75
1e6
SumAssured


0.0 2.5 5.0 7.5 10.0 12.5 15.0 17.5
LastMonthCalls


0 10 20 30 40 50
CustTenure


1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
CustCareScore


15000 20000 25000 30000 35000
MonthlyIncome

3.6 Variable transformation

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


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

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
```

As a result of fixing the two highlighted columns, we can see the total number of females is 1832 and the total number of large businesses is 408.

```
df['Occupation']=df['Occupation'].replace(to_replace='Laarge Business',value='Large Business')
```
executed in 17ms, finished 02:45:55 2022-04-03

```
df['Gender']=df['Gender'].replace(to_replace='Fe male',value='Female')
```
executed in 8ms, finished 02:45:56 2022-04-03

The variables have been encoded to numeric values for the following variables.

```
df['Gender'] = df['Gender'].replace(to_replace='Female',value=1)
df['Gender'] = df['Gender'].replace(to_replace='Male',value=0)
```

3.7 Addition of new variables

No new variables were added at this stage , but before proceeding with the model one hot encoding would be required on a few categories which would increase the number of columns not essentially the number of variables.

4. Business Insights

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

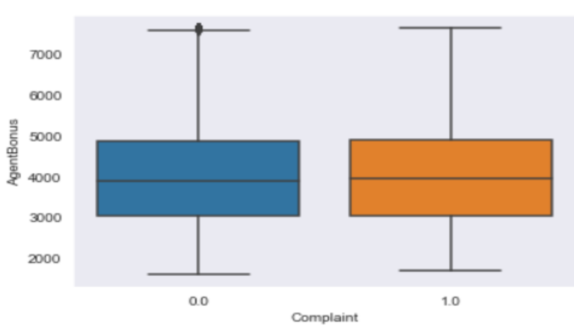
Yes, the data is unbalanced. Since the agents will get bonuses depending on their skills also customers belong to different zones and it's natural therefore there is no need to balance the dataset.

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

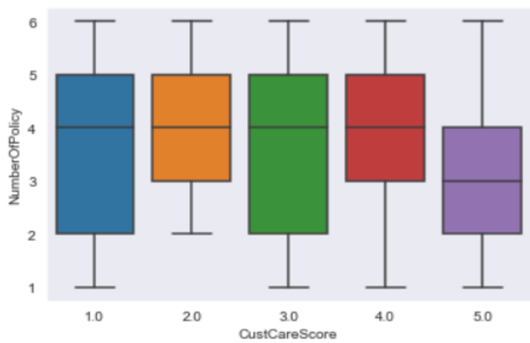
```
df['AgentBonus'].value_counts()
```
executed in 70ms, finished 15:39:08 2022-04-03

```
2581    8
3642    7
2952    7
4135    7
2906    6
5146    6
3788    6
4434    5
3379    5
2582    5
```

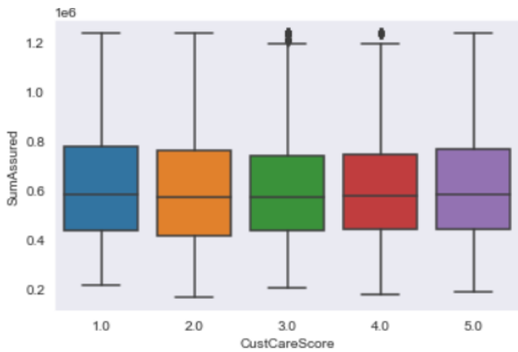
4.2 Any business insights using clustering (if applicable)



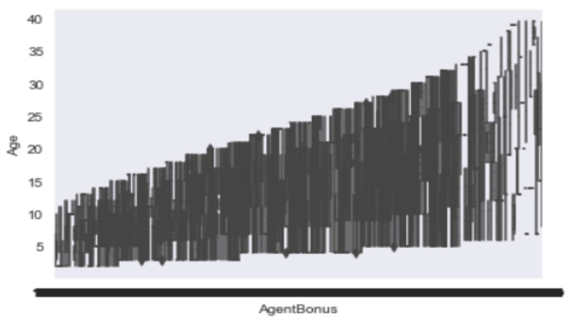
Average bonus is 4000



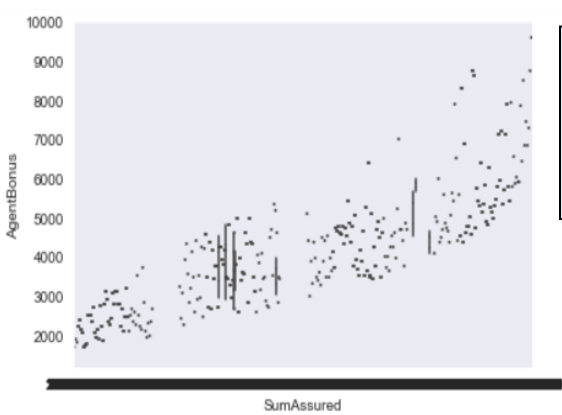
Average number of policy is 4 depending upon customer care score



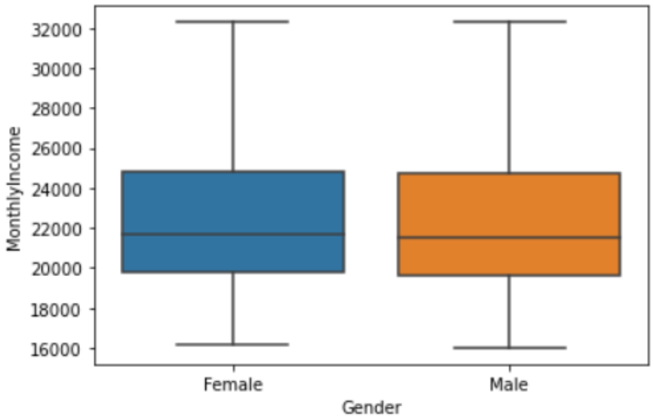
Average sum assured of the customer is 6



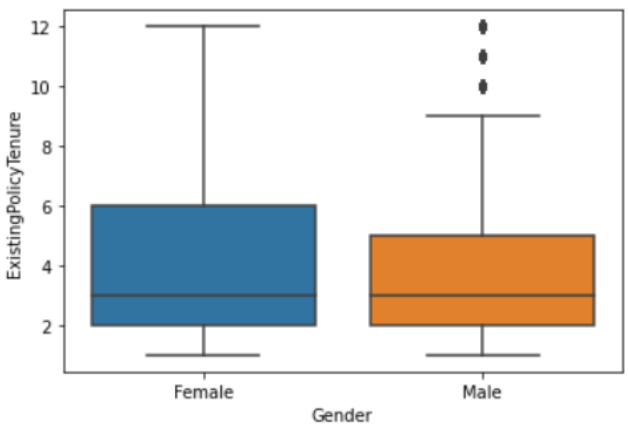
Based on the continuous data, we can only conclude that bonuses are increasing as customers age is increasing



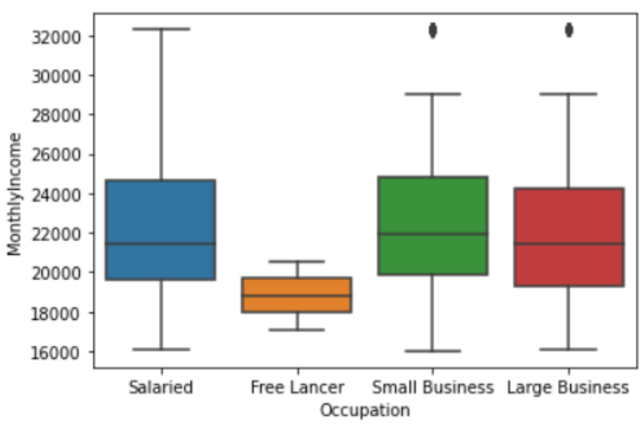
The continuous data indicates that as the sum assured of the customers increases, the bonus for agents is also increasing.



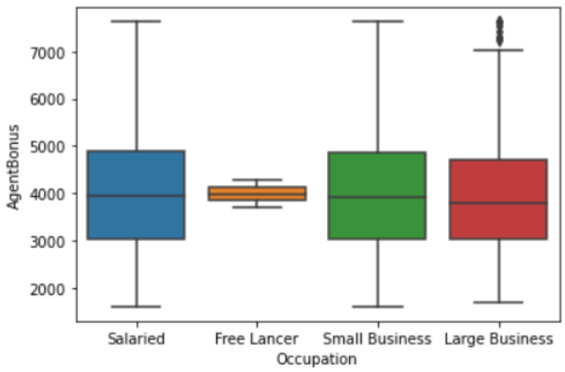
Average monthly income for male and female is 22000



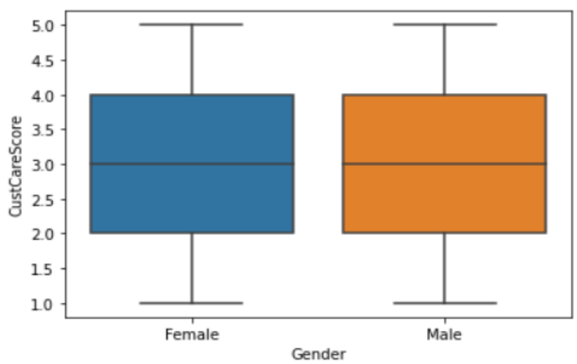
Existing policy tenure for female is the highest



Freelancer has the lowest MonthlyIncome

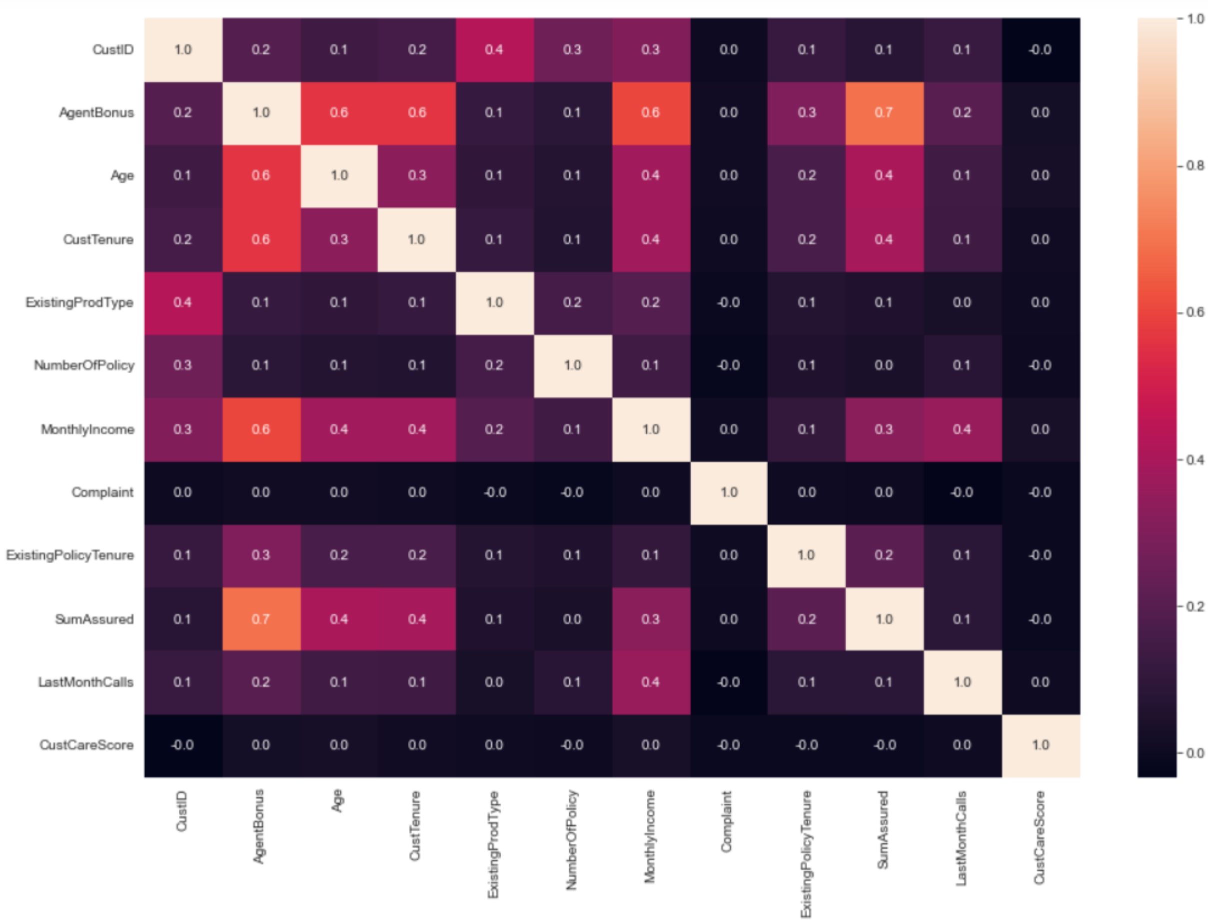


Freelancer gets the min agent bonus



Customer care score for both is 3 on an average

4.3 Any other business insights



- SumAssured is positively correlated with AgentBonus.
- MonthlyIncome is positively correlated with AgentBonus.
- CustTenure is positively correlated with AgentBonus.
- Complaint and CustCareScore have no correlation with any variables.
- AgentBonus is not much correlated with LastMonthCalls.

Recommendation:
If agents are getting low bonuses, they should target customers who have high monthly incomes, a high maximum sum assured on all their policies, and a long tenure with the organization. And the agents who are getting high bonuses should keep up their work.