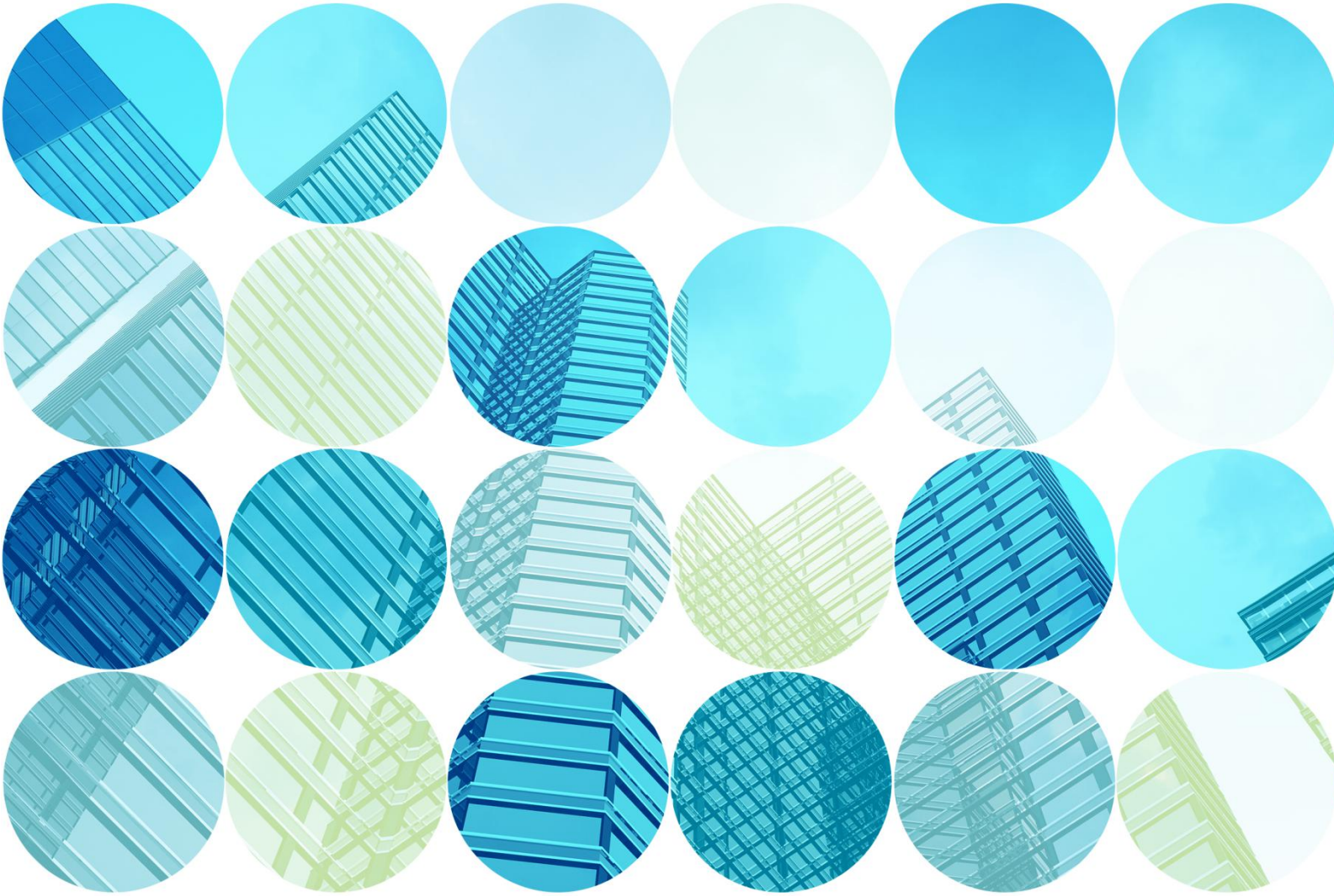
LIFE INSURANCE COMPANY BUSINESS REPORT



Khussal Pradhan
1928035

Contents

1. Introduction of the Business Problem

1.1 Defining problem statement

- The given life insurance company wants to predict the bonus and their working efficiency of its agents so that the higher performing agents get rewarded for their work and plan certain upskill programs for the lower performing agents.

1.2 Need of the study/project

- This study would give the the company key and valuable information about their employees according to the attributes provided in the dataset. Then the company can work on how to properly compensate, plan on their further skill improvement of their agents.

1.3 Understanding business/social opportunity

- A happy employee is the key for the profitability of the company. If the employee is being recognized for its work and also compensated properly or even if the employee feels the company is trying its best to provide proper opportunities he/she would stick with the company and also being a insurance company if the end users get to know how the company is treating its agents and how good is the work culture they would definitely get insurance policy from the company.

2. Data Report

2.1 Data Description

- The data belongs to all the agents of insurance company and various factors are given such as performance determining factors along with personal and useful details of each agent.

Variable	Discerption
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

The data belongs to the agents of the insurance company

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

```
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)
The number of rows (observations) is 4520
The number of columns (variables) is 20
```

	count	mean	std	min	25%	50%	75%	max
CustID	4520.0	7.002260e+06	1304.955938	700000.0	7001129.75	7002259.5	7003389.25	7004519.0
AgentBonus	4520.0	4.077838e+03	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	1.449471e+01	9.037629	2.0	7.00	13.0	20.00	58.0
CustTenure	4294.0	1.446903e+01	8.963671	2.0	7.00	13.0	20.00	57.0
ExistingProdType	4520.0	3.688938e+00	1.015769	1.0	3.00	4.0	4.00	6.0
NumberOfPolicy	4475.0	3.565363e+00	1.455926	1.0	2.00	4.0	5.00	6.0
MonthlyIncome	4284.0	2.289031e+04	4885.600757	16009.0	19683.50	21606.0	24725.00	38456.0
Complaint	4520.0	2.871681e-01	0.452491	0.0	0.00	0.0	1.00	1.0
ExistingPolicyTenure	4336.0	4.130074e+00	3.346386	1.0	2.00	3.0	6.00	25.0
SumAssured	4366.0	6.199997e+05	246234.822140	168536.0	439443.25	578976.5	758236.00	1838496.0
LastMonthCalls	4520.0	4.626991e+00	3.620132	0.0	2.00	3.0	8.00	18.0
CustCareScore	4468.0	3.067592e+00	1.382968	1.0	2.00	3.0	4.00	5.0

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

	Column	Count	Dtype	Remarks
0	CustID	4520	int64	Redundant
1	AgentBonus	4520	int64	Numeric
2	Age	4251	float64	Numeric
3	CustTenure	4294	float64	Numeric
4	Channel	4520	object	Categorical
5	Occupation	4520	object	Categorical
6	EducationField	4520	object	Redundant
7	Gender	4520	object	Redundant
8	ExistingProdType	4520	int64	Numerical
9	Designation	4520	object	Redundant
10	NumberOfPolicy	4475	float64	Numeric
11	MaritalStatus	4520	object	Redundant
12	MonthlyIncome	4284	float64	Numeric
13	Complaint	4520	int64	Numeric
14	ExistingPolicyTenure	4336	float64	Numerical
15	SumAssured	4366	float64	Numerical

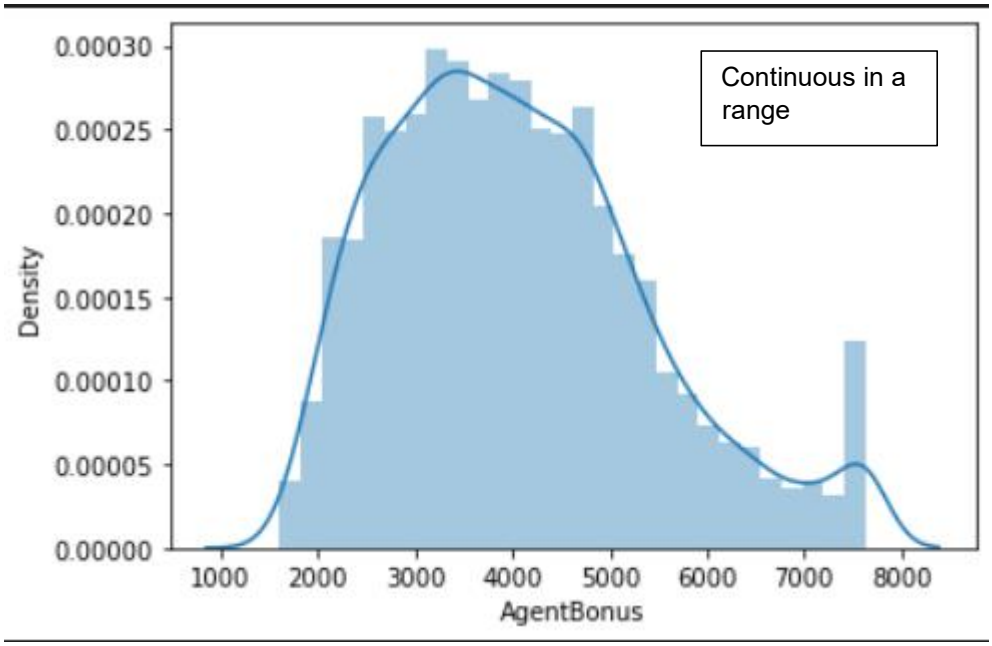
16	Zone	4520	object	Categorical
17	PaymentMethod	4520	object	Categorical
18	LastMonthCalls	4520	int64	Numeric
19	CustCareScore	4468	float64	Numeric

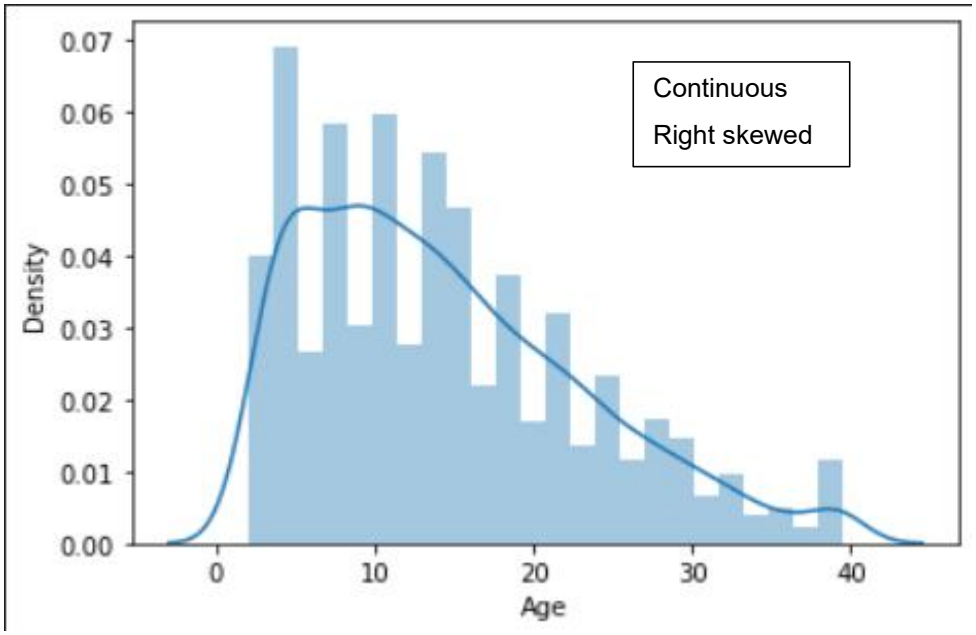
Unique values of various Categories

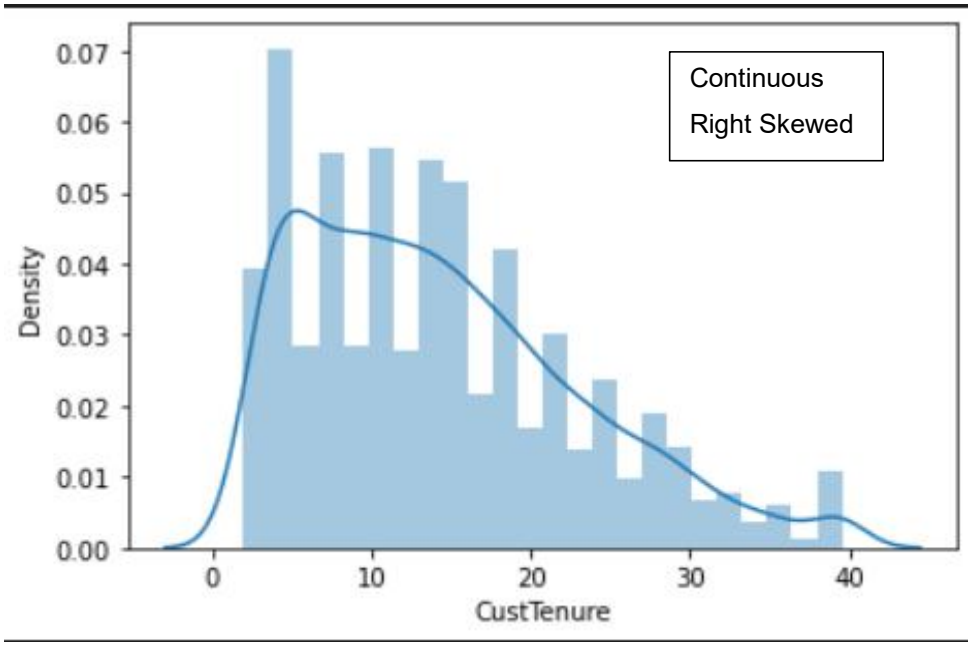
```
Channel :  3
Online                   468
Third Party Partner      858
Agent                   3194
Name: Channel, dtype: int64

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

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
```

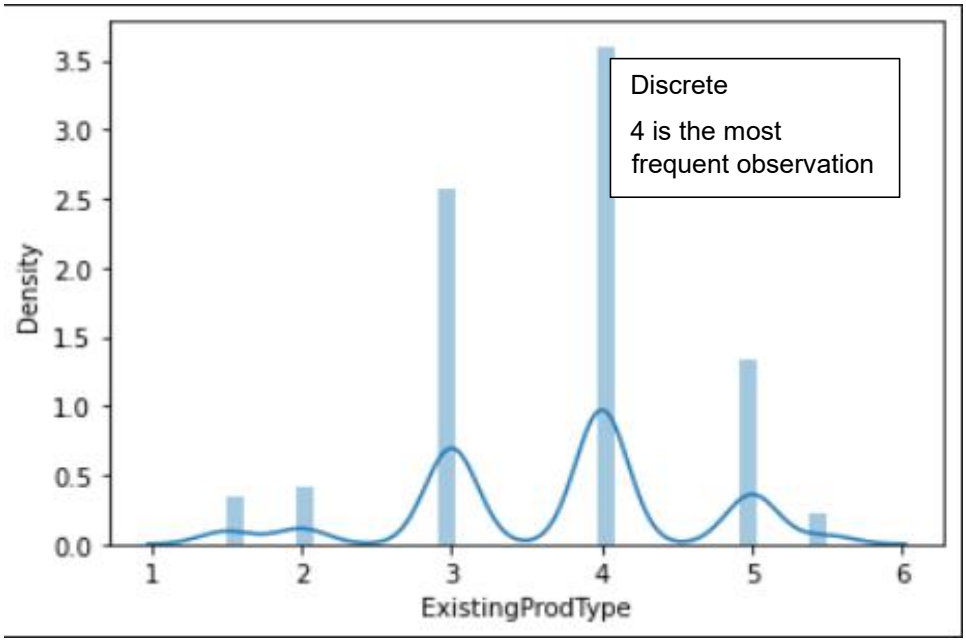
3. Exploratory data analysis

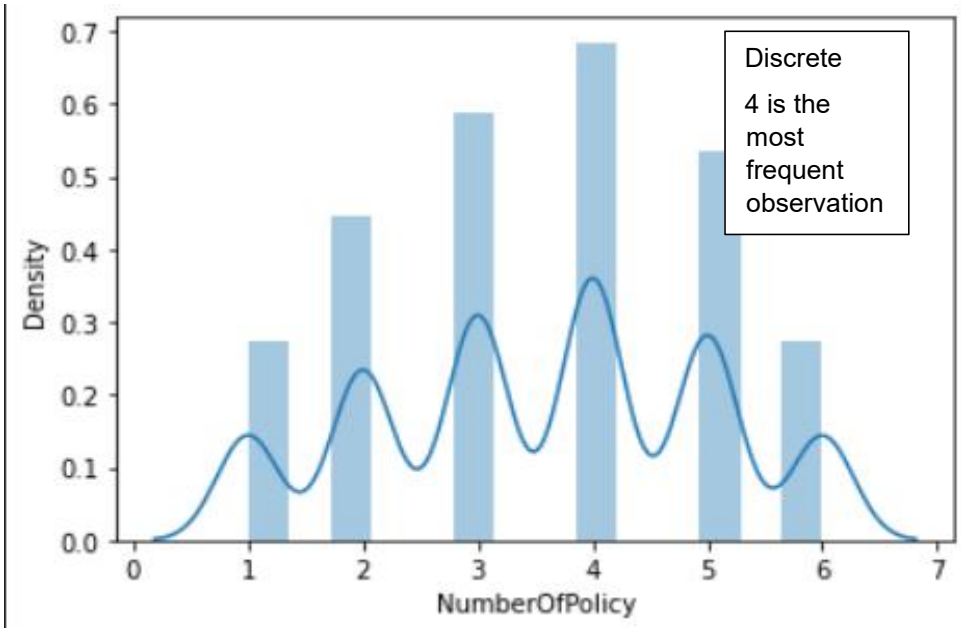
3.1 Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

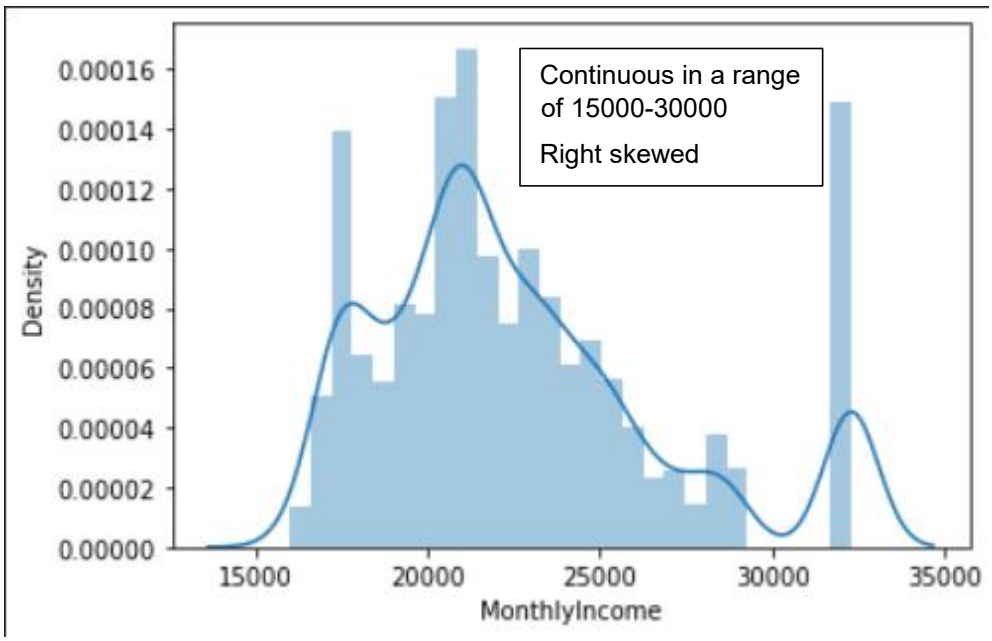


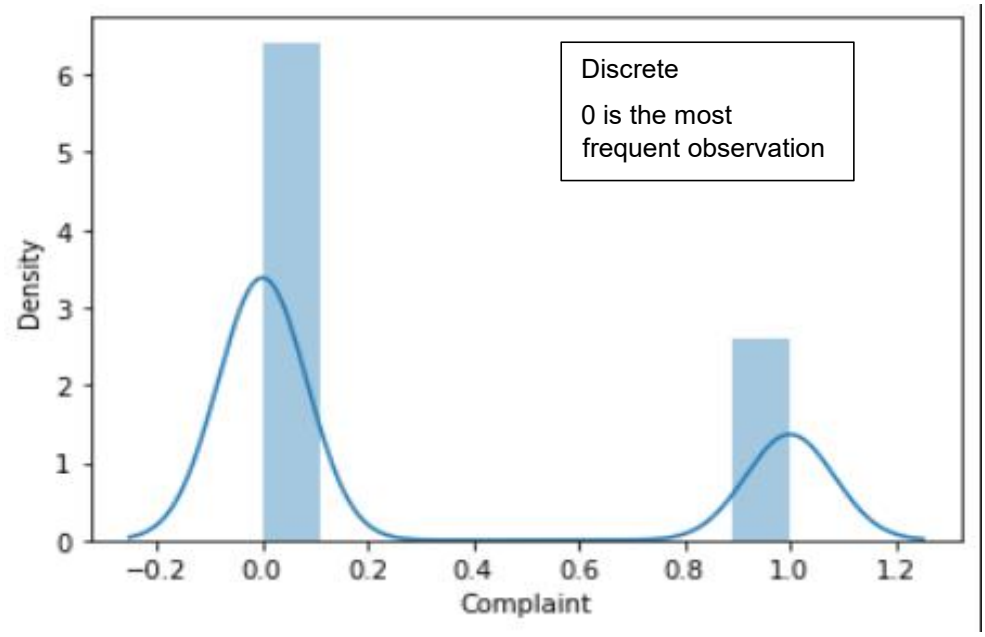


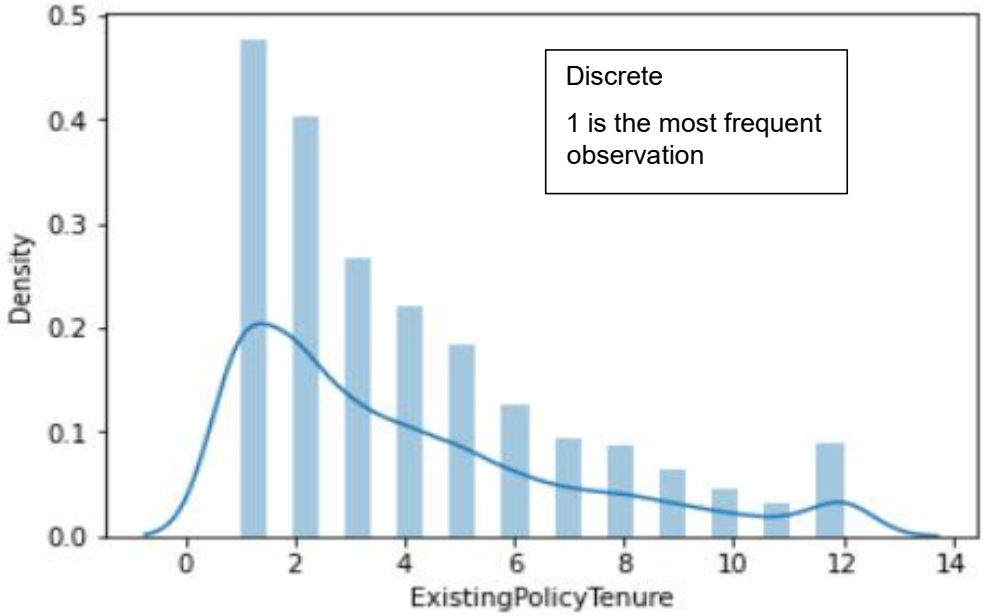

Continuous
Right Skewed
Density
CustTenure
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0
10
20
30
40

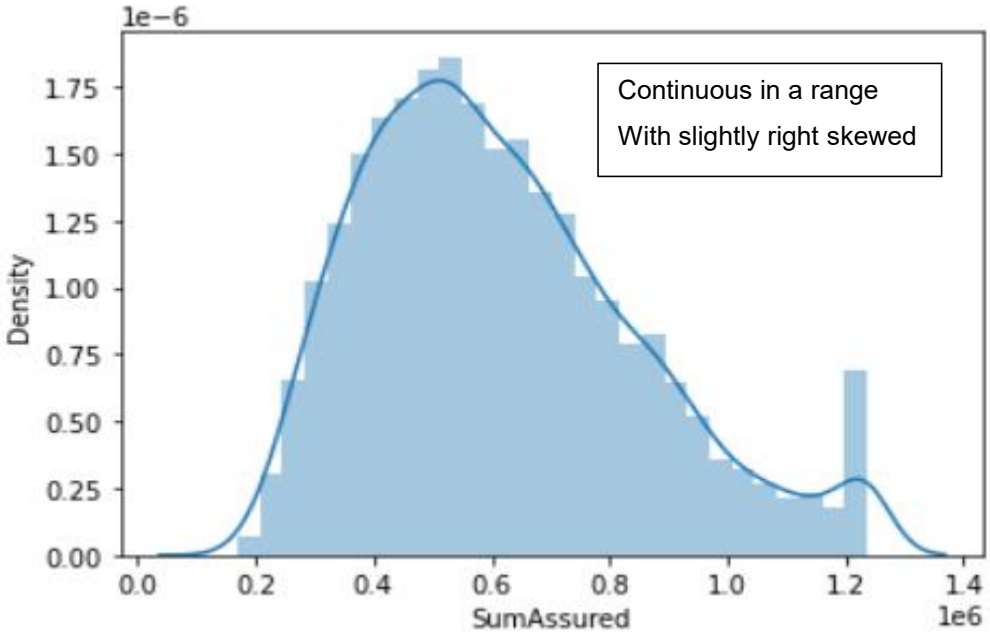

Discrete
4 is the most frequent observation
Density
ExistingProdType
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
1
2
3
4
5
6

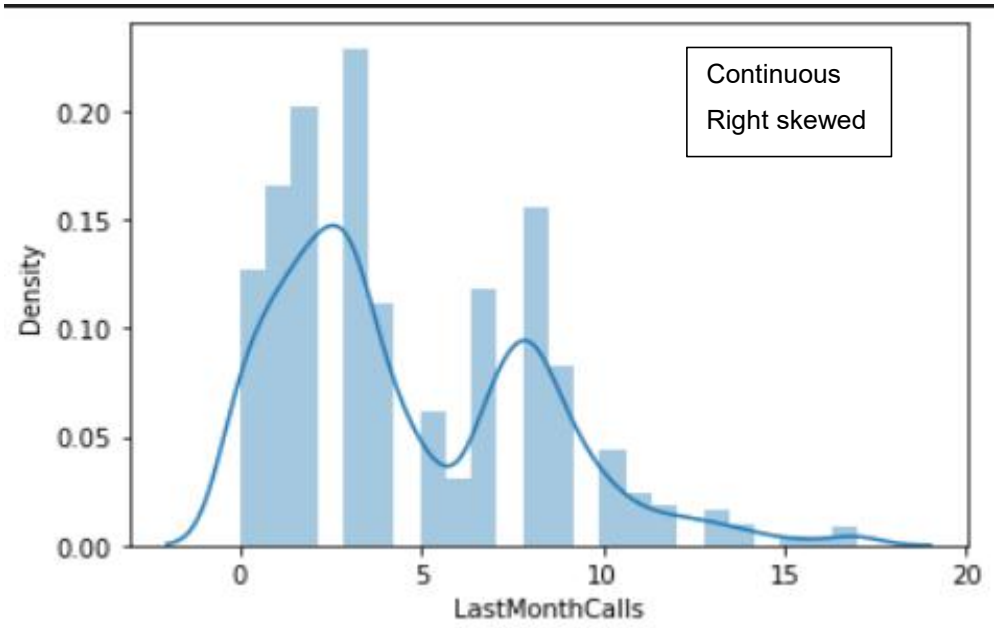

Discrete
4 is the most frequent observation
Density
NumberOfPolicy
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
1
2
3
4
5
6
7

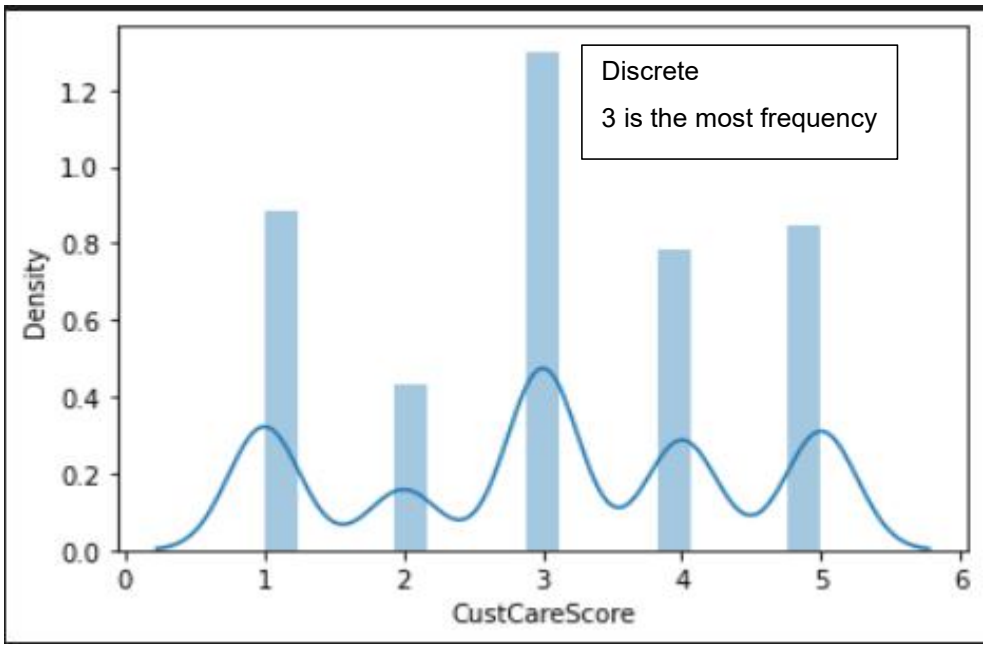

Continuous in a range of 15000-30000
Right skewed
Density
0.00016
0.00014
0.00012
0.00010
0.00008
0.00006
0.00004
0.00002
0.00000
15000
20000
25000
30000
35000
MonthlyIncome


Discrete
0 is the most frequent observation
Density
6
5
4
3
2
1
0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Complaint


Discrete
1 is the most frequent observation
Density
ExistingPolicyTenure

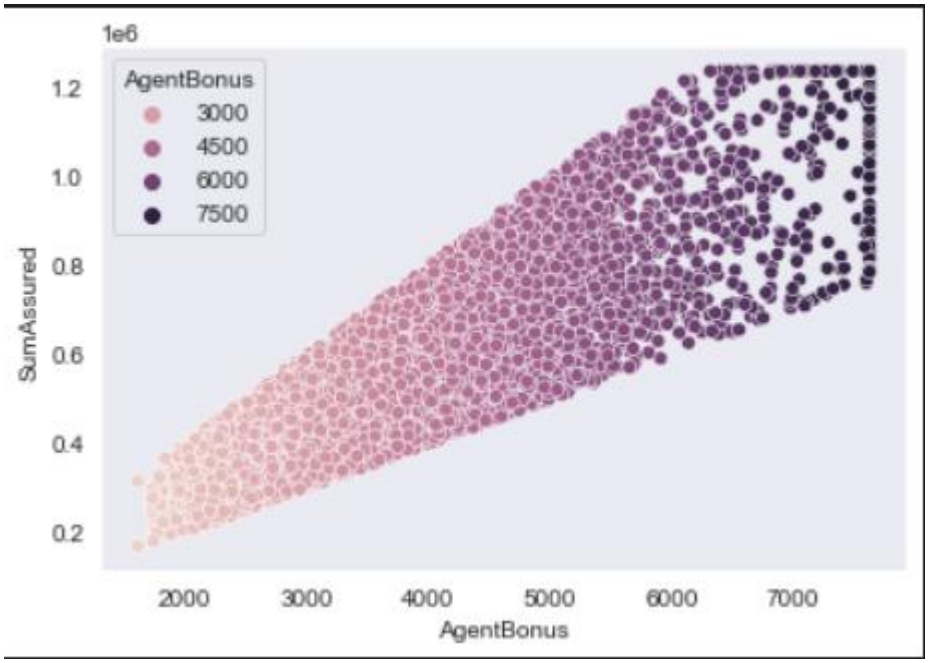

Continuous in a range
With slightly right skewed
1e−6
Density
SumAssured
1e6





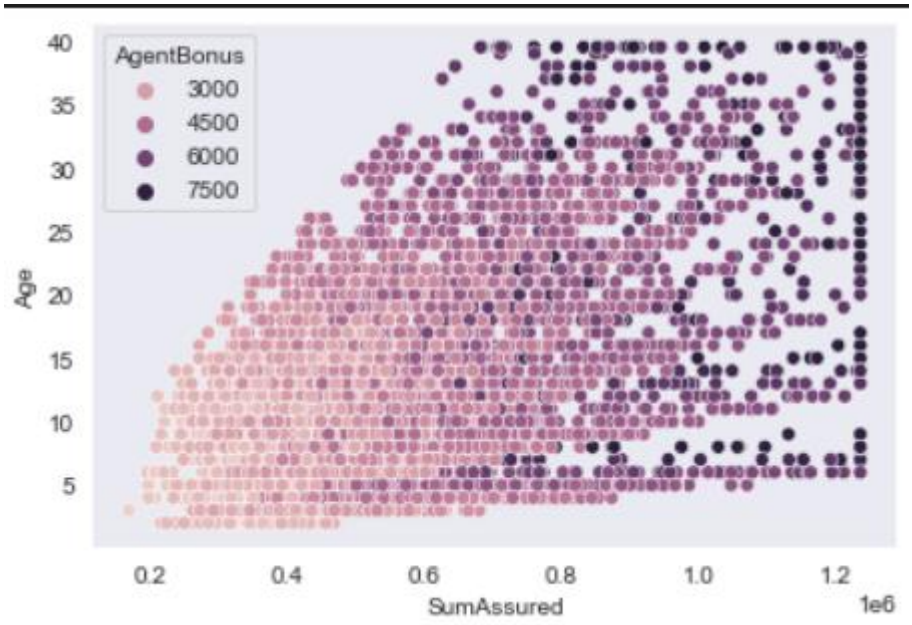
- Most of the numerical data is more or less equally divided between continuous and discrete type since the nature of the domain is such. There is no fixed assumption of integers for any column, every data can vary in any range, no predefined values are set. Only outlier is complaint in this case, where only two values are set, if their is a complaint it is set to 1 or if their is no complaint then it is set to 0.

3.2 Bivariate analysis (relationship between different variables , correlations)



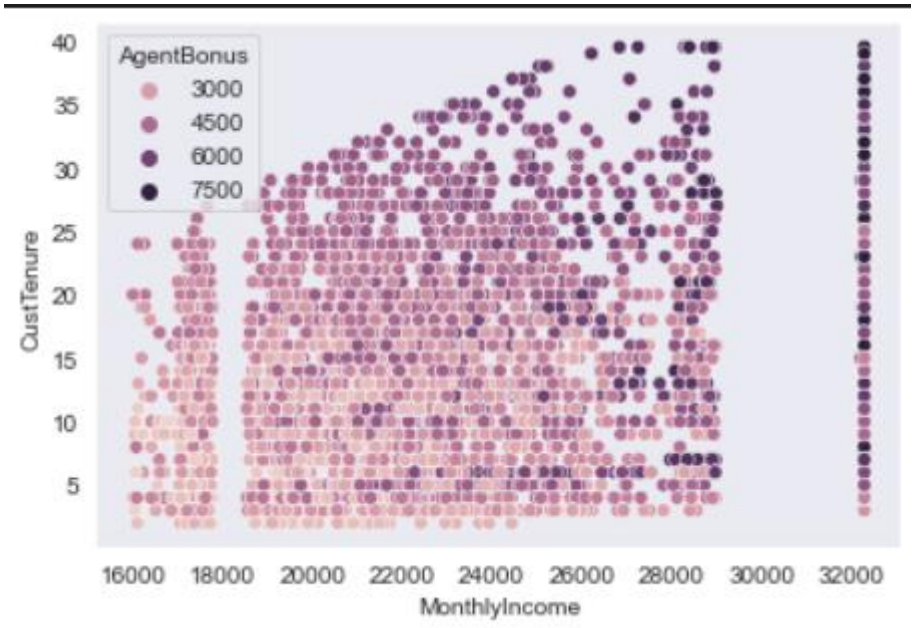
Higher the SumAssured higher the AgentBonus.

Positively related and almost a linear relation can be established.



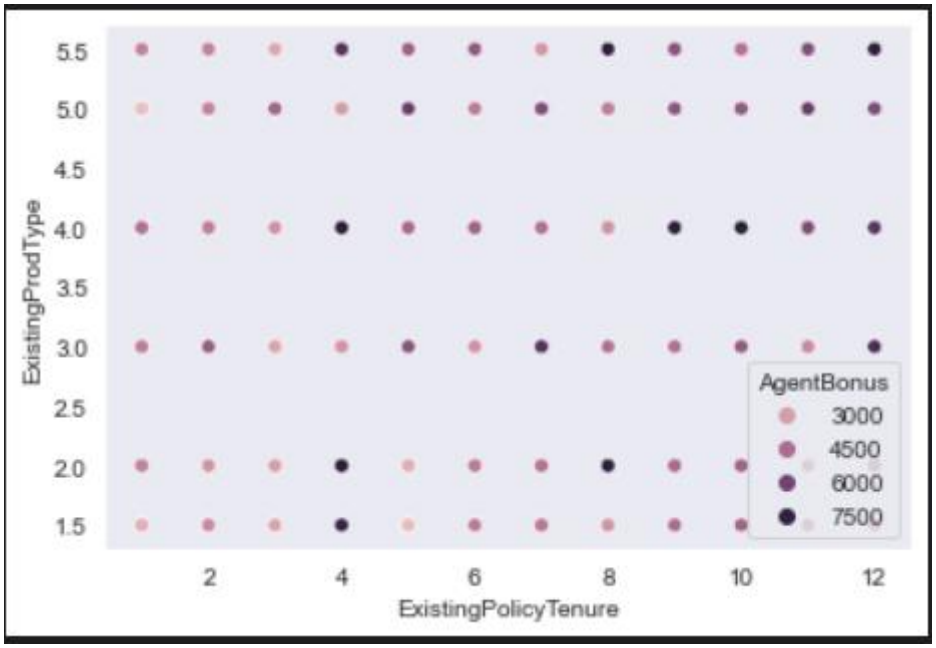
Average age lies between 10-30 and is more or less independent with agent bonus.

More the sum assured, more the bonus which is given to the agent.

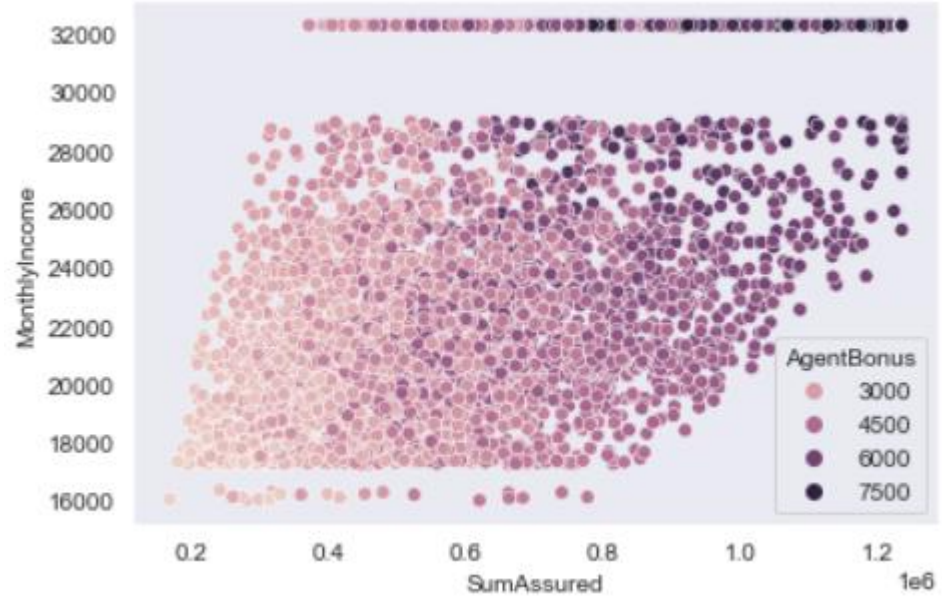


As the CustTenure increases a slight rise in agent bonus can be seen.

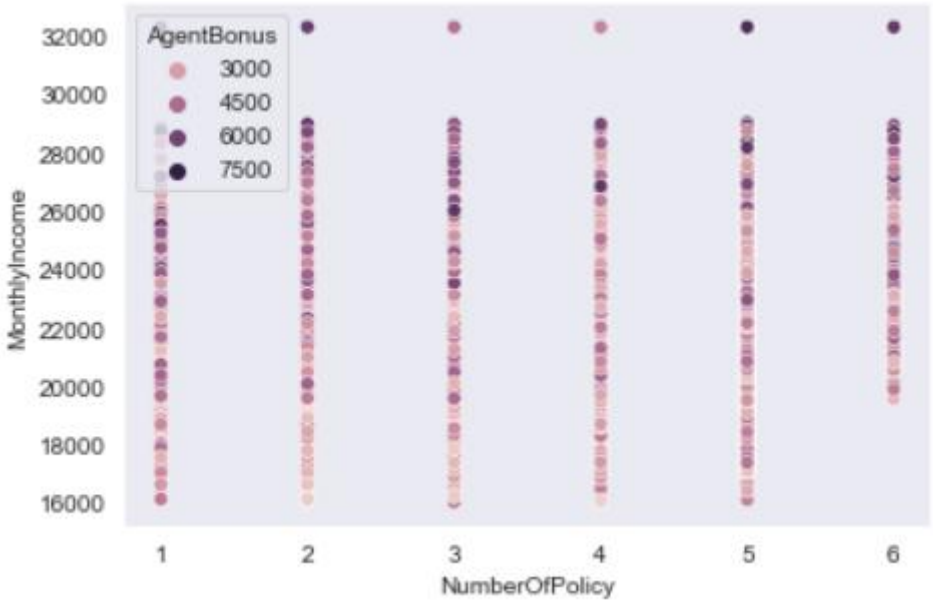
Greater the monthly income of the client, a higher agent bonus can be seen.



Not able to establish any relation.



As we can see, a higher SumAssured is generally taken by people with a higher MonthlyIncome as a result a higher AgentBonus is also observed.



Not able to establish any relation

- 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.

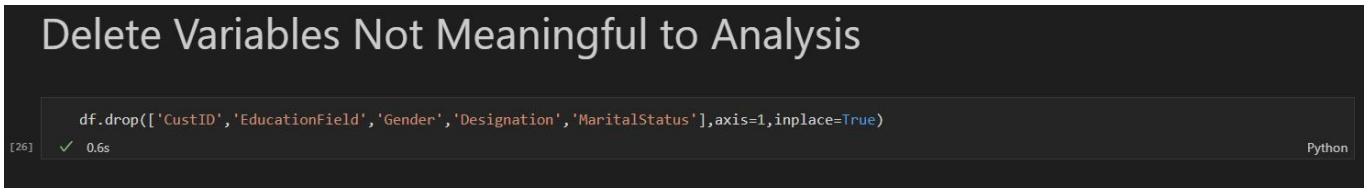
- The pair plot also seems to suggest the same thing . But due to the huge number of columns pair plot was not providing very clear insight and hence resorted to bi variate plots with every combination possible.


AgentBonus
3000
4500
6000
7500
Age
CustTenure
ExistingProdType
NumberOfPolicy
MonthlyIncome
Complaint
ExistingPolicyTenure
SumAssured
LastMonthCalls
CustCareScore
1e6


SumAssured
0.2
0.4
0.6
0.8
1.0
1.2
AgentBonus
Age
CustTenure
ExistingProdType
NumberOfPolicy
MonthlyIncome
Complaint
ExistingPolicyTenure
LastMonthCalls
CustCareScore

3.3 Removal of unwanted variables (if applicable)

- CustID, EducationField, Gender, Designation and MaritalStatus are 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.

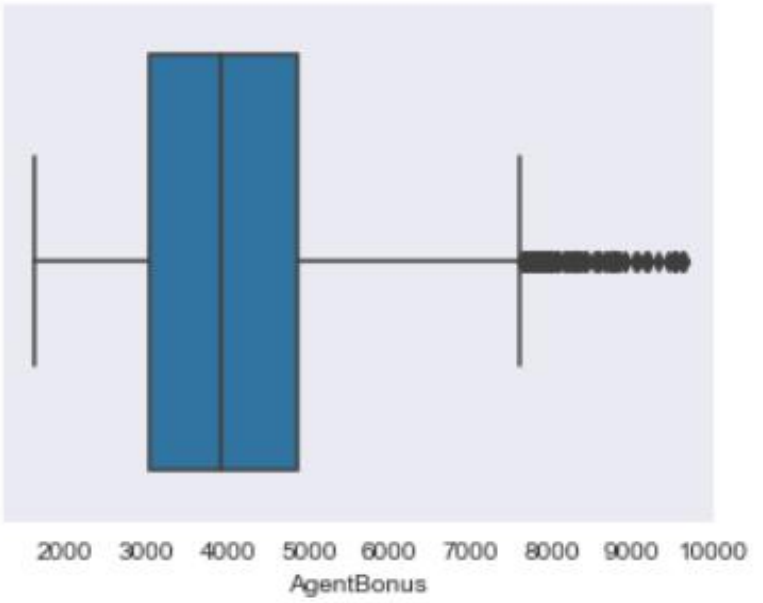


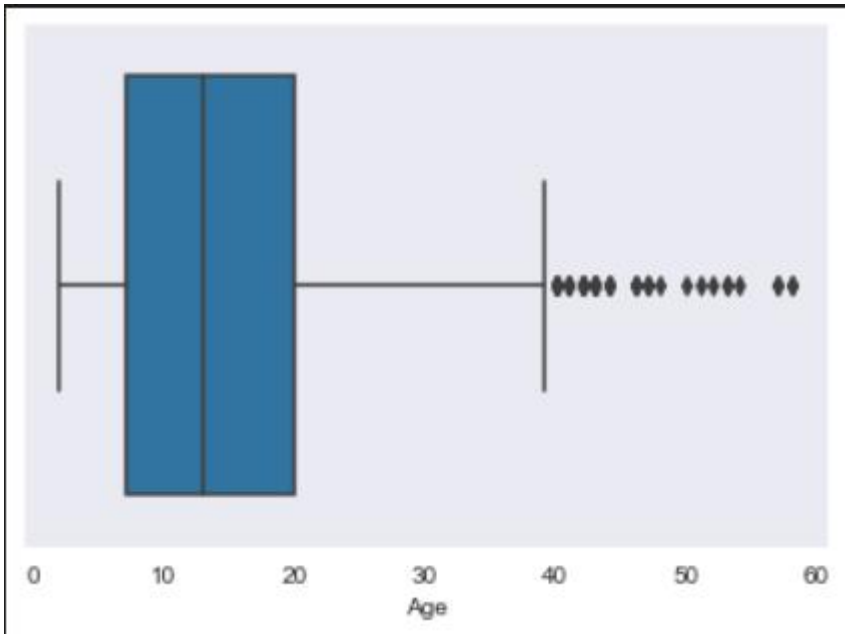
3.4 Missing Value treatment (if applicable)

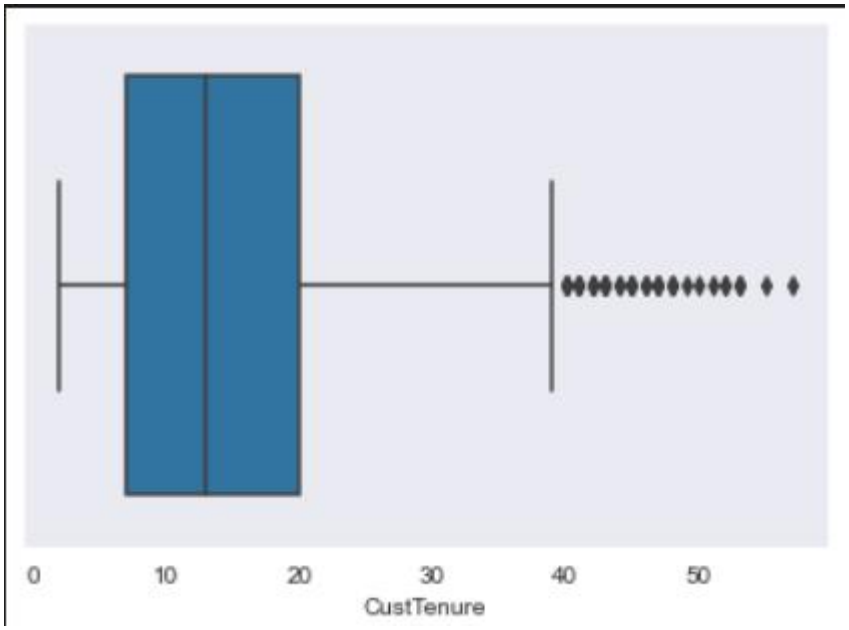
```
Age                       269
MonthlyIncome             236
CustTenure                226
ExistingPolicyTenure      184
SumAssured                154
CustCareScore              52
NumberOfPolicy             45
AgentBonus                  0
Channel                     0
Occupation                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

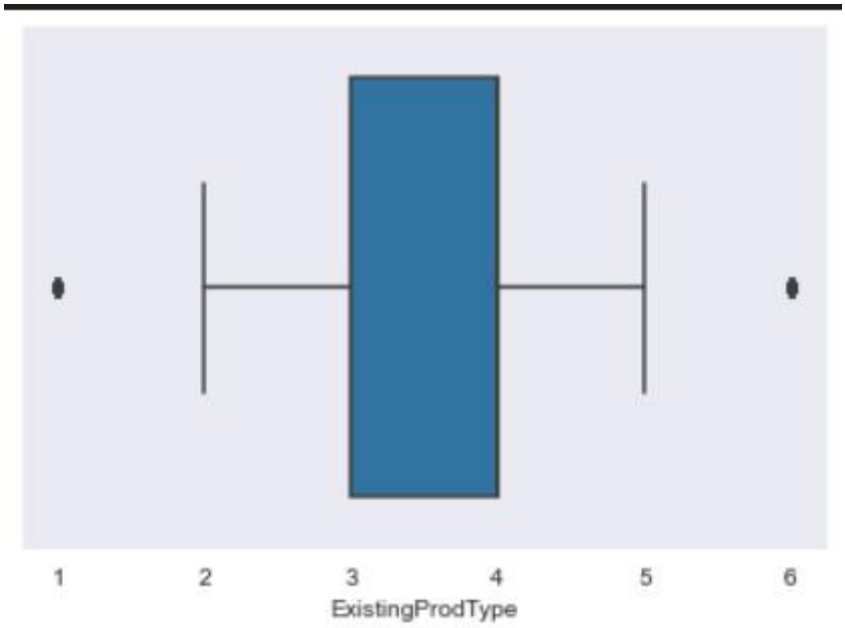
- The missing values have been treated with most frequent values than median for numeric data including categorical data. The main reason of choosing mode or most frequent entry was it was making more sense considering the domain of our problem statement. More so as we have been in the various plots as well the numeric data has discrete pattern due to which we treated them as categorical data.

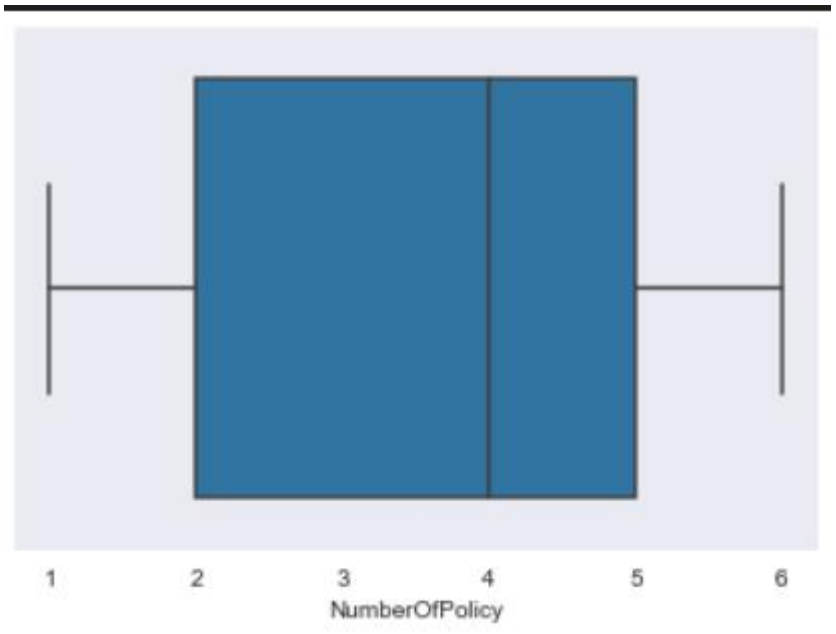
3.5 Outlier treatment (if required)

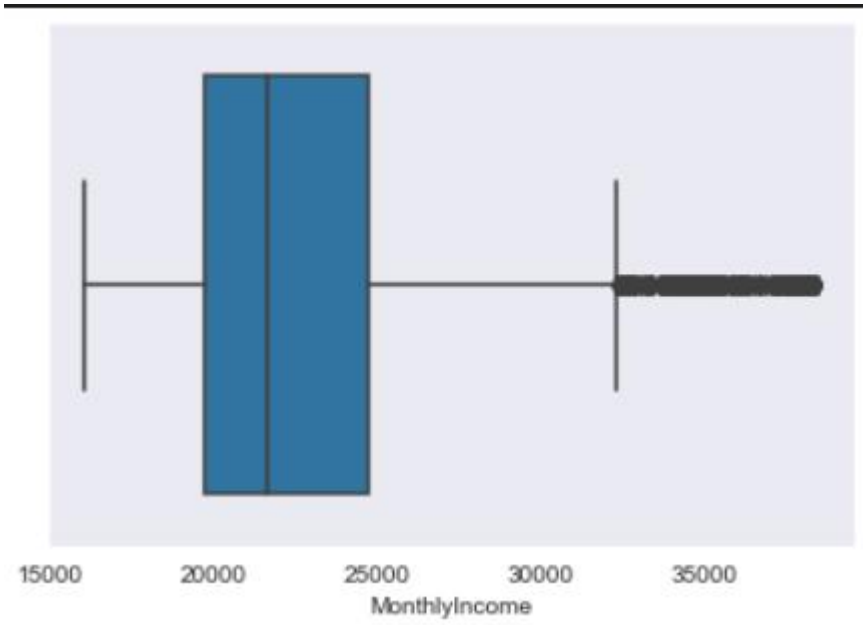


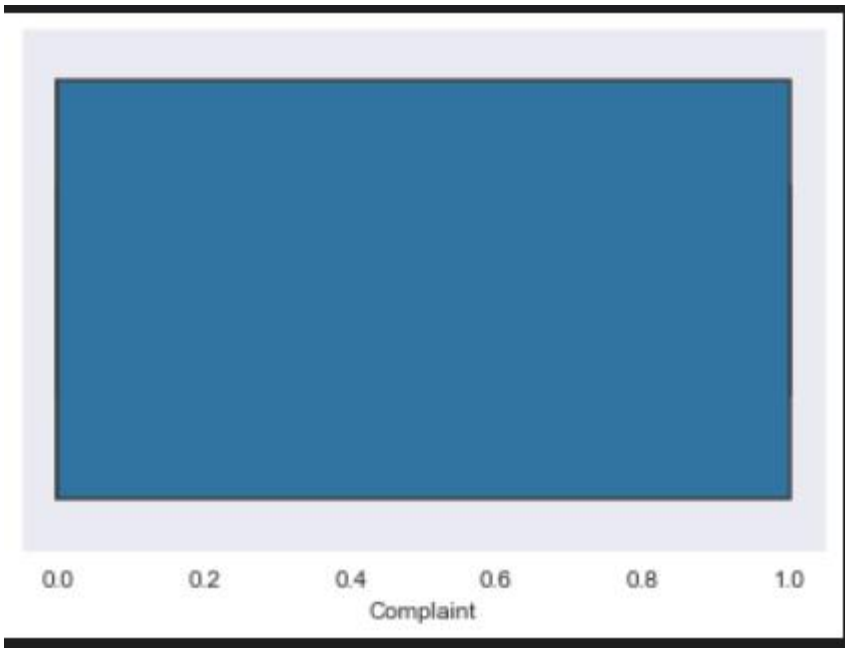


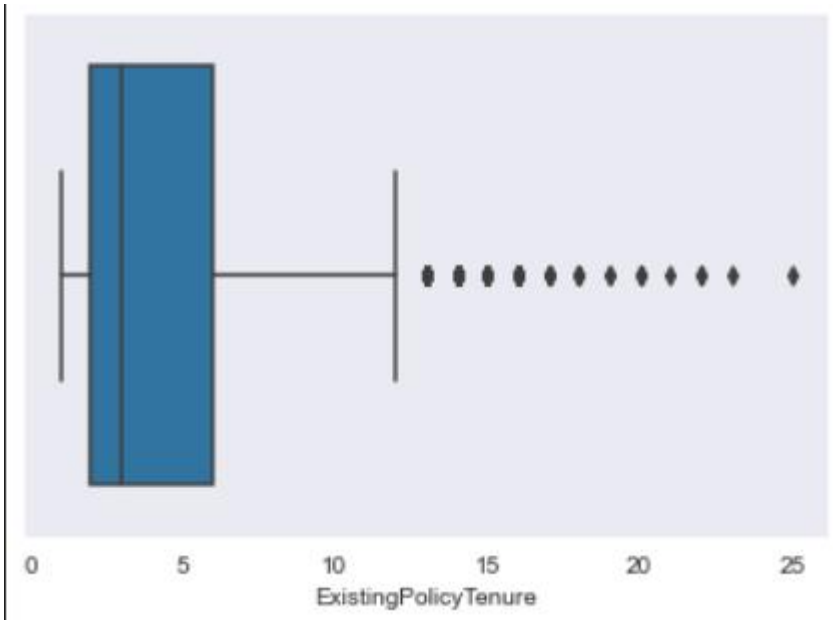

0
10
20
30
40
50
CustTenure

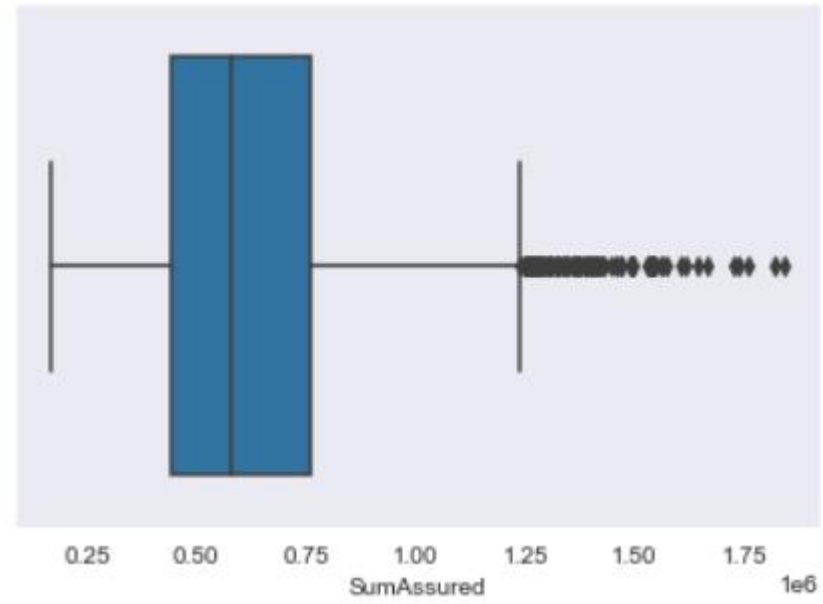

1
2
3
4
5
6
ExistingProdType

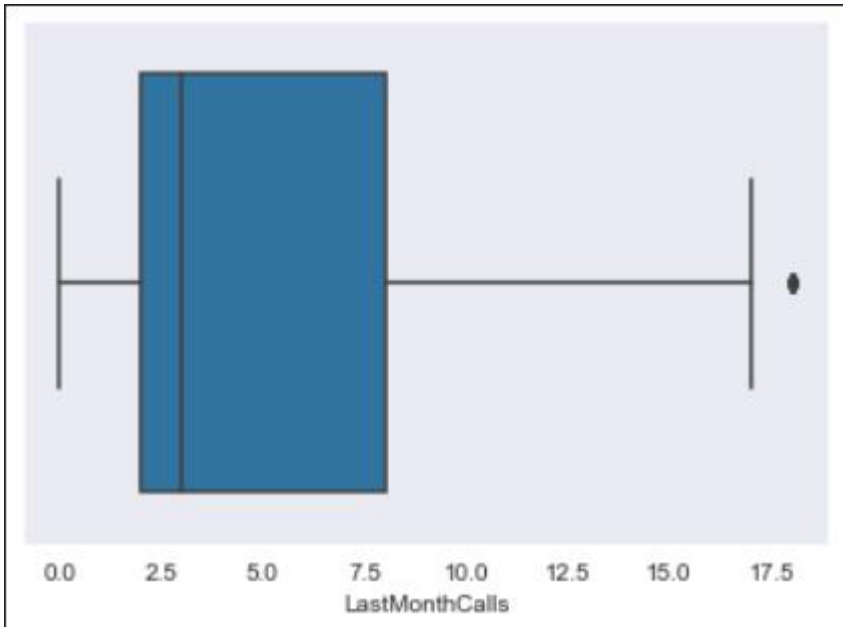

1
2
3
4
5
6
NumberOfPolicy

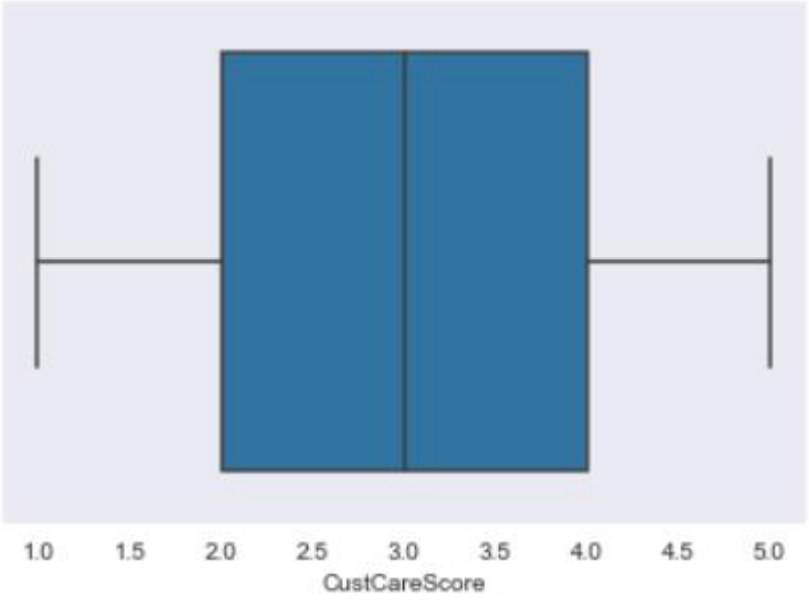

15000
20000
25000
30000
35000
MonthlyIncome


0.0
0.2
0.4
0.6
0.8
1.0
Complaint


0
5
10
15
20
25
ExistingPolicyTenure


0.25
0.50
0.75
1.00
1.25
1.50
1.75
SumAssured
1e6


0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
LastMonthCalls



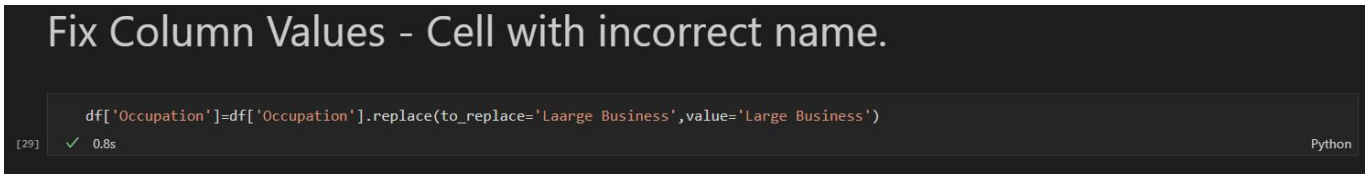
- Not in a favour of doing any outlier treatment as most of the numeric data here is discrete and hence the outliers might be able add value to the model. More so the numeric data which is continuous has minimal outliers . Like the LastMonthCalls has one observation which stands out and most of the others are in the right range.

3.6 Variable transformation (if applicable)

```
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

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
```

- The highlighted value has an incorrect spelling, so the spelling is corrected and merged with the correct one.



3.7 Addition of new variables (if required)

Not required in our problem statement.

4. Business Insights from EDA

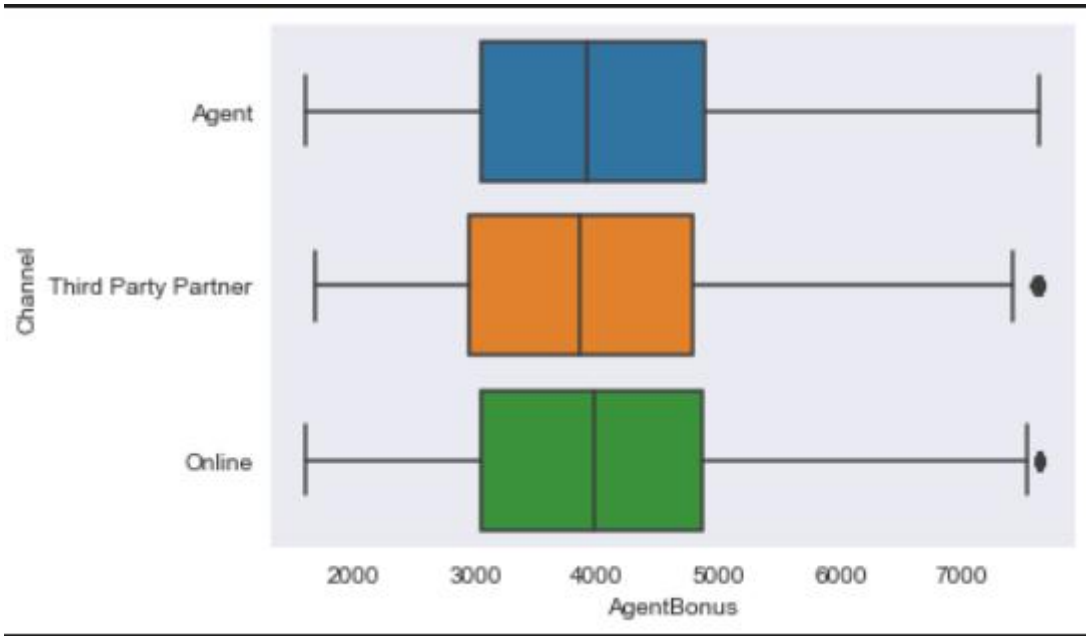
4.1 Is the data unbalanced? If so, what can be done? Please explain in the context of the business

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

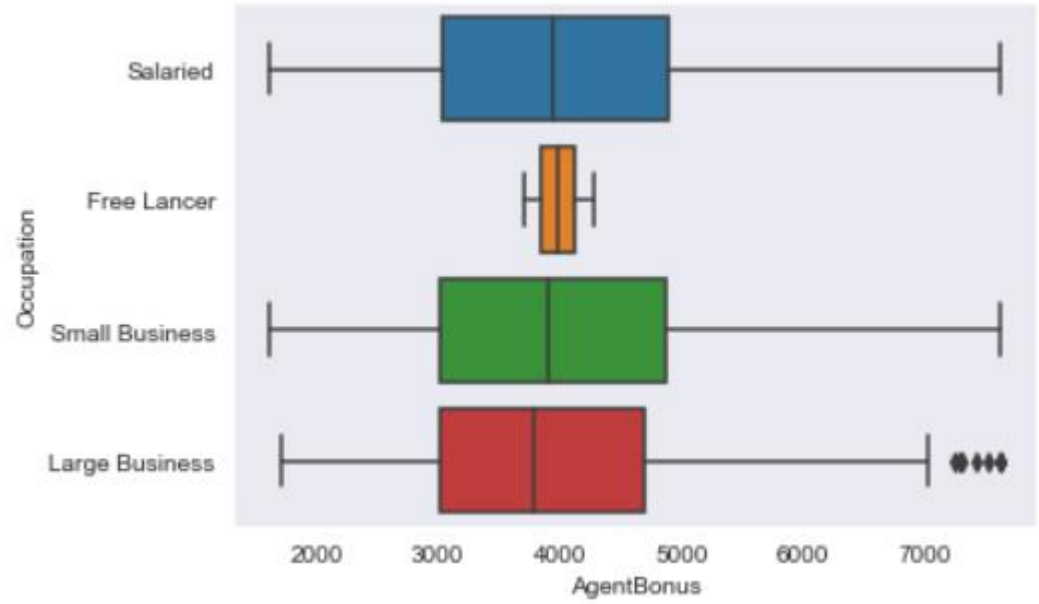
- Data is not balanced with more agents getting insurance policy to the customer rather than getting it through online portals or third party partner but it is the nature of this dataset. People generally prefer their insurance to be handled by agents on ground rather than any online platform because in case of emergency getting a response from those platform may get delayed or sometimes they don't even respond but whereas the customer can make a call to his/her agent through which they have done their insurance and can easily get the claim. Ultimately it is the hard earned money of the customer, he/she will do its best to keep it in someone's hand so that if anything goes wrong they can hold their agent liable.

- Don't see any treatment on this would be needed.
- More so the problem statement is also to predict the AgentBonus.

4.2 Any business insights using clustering (if applicable)

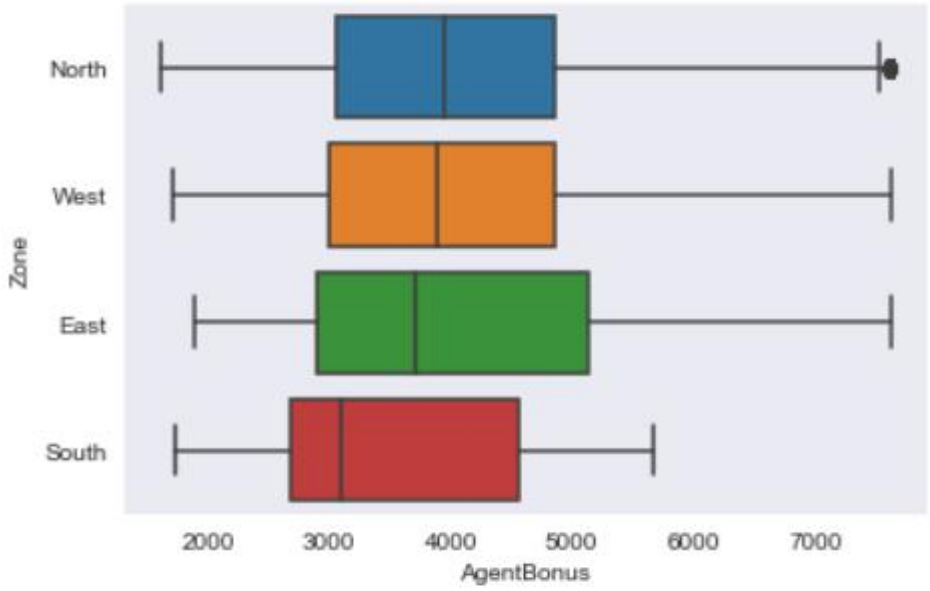


Bonus is nearly equally distributed between various agent types with more or less the same median for all the 3 channels.



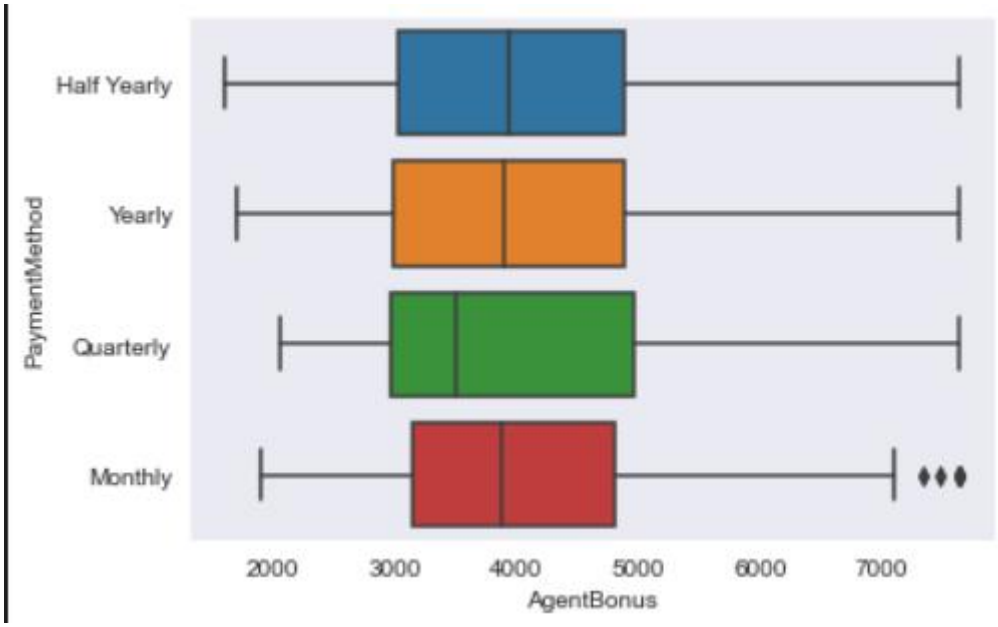
Securing a salaried client or a large business gives more or less the same bonus.

Free lancer client has the least scope for bonus.



East zone has the highest inter quartile range and south zone has the least median for the bonus.

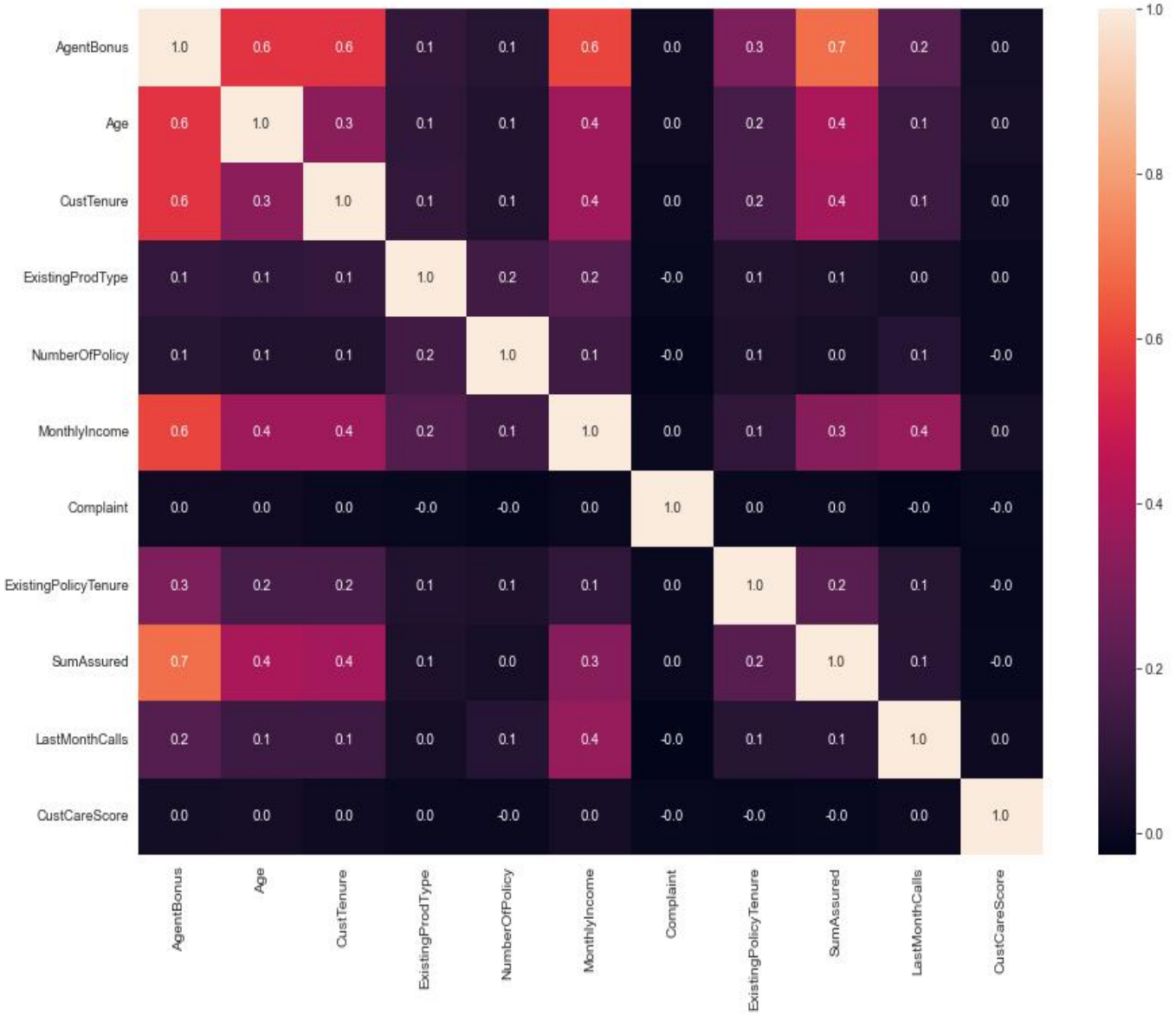
The median for AgentBonus is more or less the same for North, West and East zone.



Quarterly payment method has the least median of bonus. Monthly has the smallest scope for bonus among these 4 with few outliers

Half yearly an yearly have almost the same value against AgentBonus.

4.3 Any other business insights



- AgentBonus has no correlation with NumberOfPolicy but seems to be positively and nearly linearly correlated with SumAssured which means higher the SumAssured higher the bonus the agent gets.
- Age has a positive correlation with both SumAssured and MonthlyIncome. A higher MonthlyIncome is observed with gradual increase in Age and also the SumAssured tends to go up.
- CustTenure has a positive correlation with AgentBonus. Longer the tenure of the customer with the insurance company higher the bonus.
- MonthlyIncome is also postively correlated with the AgentBonus, higher income of the client leads to getting higher SumAssured insurance policy which leads to higher AgentBonus.
- Age has a positive correlation with AgentBonus. Generally people in a higher age group tend to get a larger SumAssured which has a positive impact on the bonus.