BUSINESS REPORT

Capstone Project Notes 1

RAMESH PANDEY
PGP-DSBA Online
May’ 23
Date: 02/07/2023

Table of Contents

PRESENTATION TITLE

Table of Contents

Table of Contents

PRESENTATION TITLE

List of Figures

List of Figures

List of Figures

Question 1 - Problem Understanding

a) Problem Statement

The dataset belongs to a leading life insurance company. The company wants to predict the
bonus for its agents so that it may design appropriate engagement activity for their high
performing agents and upskill programs for low performing agents.

b) Need of the study

The study is needed to optimize the allocation of resources and improve the performance of agents based on their predicted bonus. By identifying high-performing and low-performing agents, the company can incentivize and provide training to enhance their productivity.

c) Understanding business

By accurately predicting agent bonuses, the company can create a more effective and targeted engagement strategy. This can lead to improved agent performance, increased customer satisfaction, and ultimately, higher profitability for the company.

Question 2 - Data Report

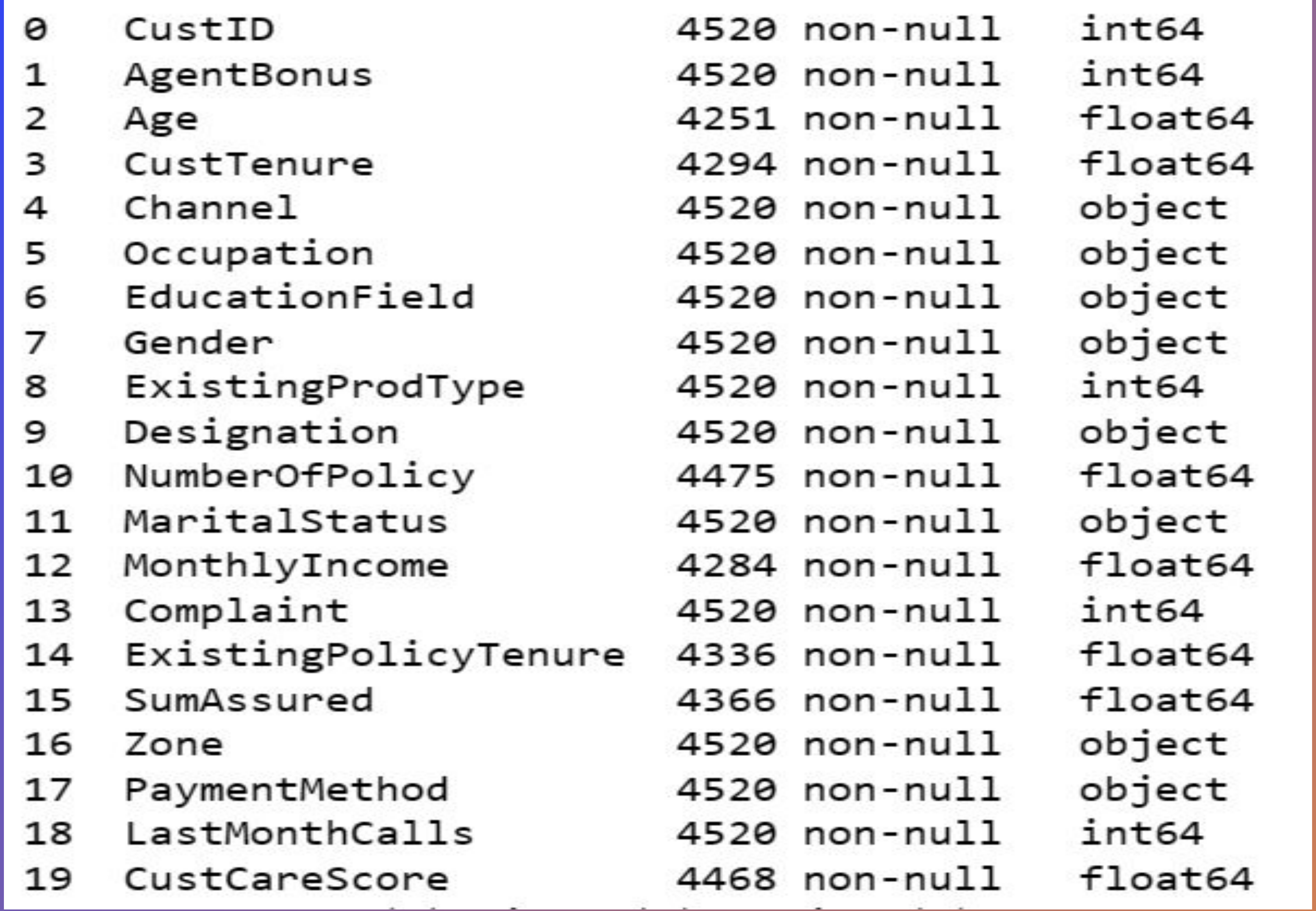
a) Understanding how data was collected

Information about the data collection process, time frame, and frequency is not provided. But looking at the amount of data, we can say that the data was collected within one year as the number of customers are very less and also looking at the feature LastMonthCalls, we can say that the data is of one month. It represents a snapshot of the insurance company's customer and agent information. The data may be collected from online or through the agents.

b) Visual inspection of data

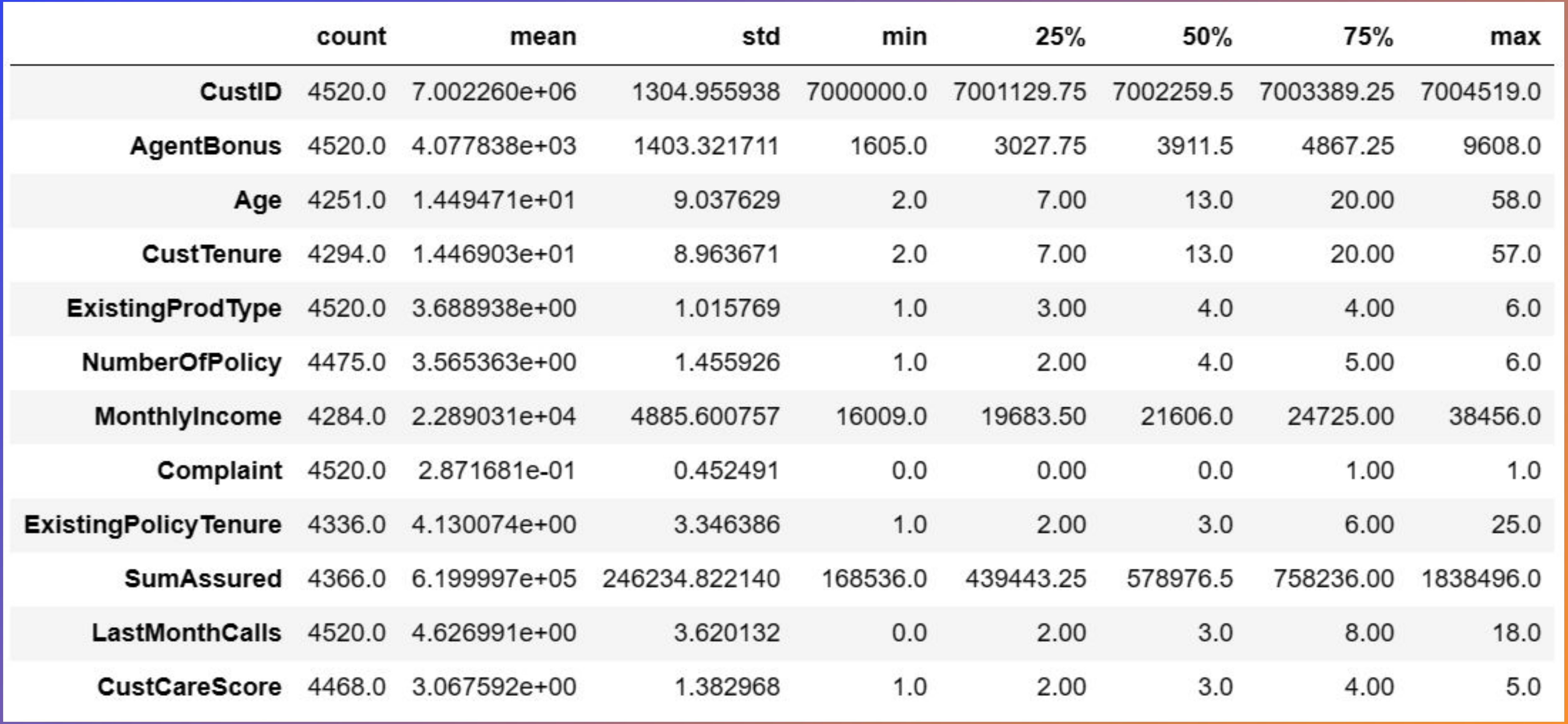
The shape of the data is (4520, 20). This tells that the number of rows are 4520 and the number of column or features are 20.
There are 5 int64, 7 float64 and 8 object type of columns in the data.
The mean bonus given to agents is 4077.8. The minimum is 1605 and the maximum is 9608, where as the median or 50 percentile is 3911.5 which is closer to the mean and differs by 150. Hence, we can say that it is normally distributed.

b) Visual inspection of data



```
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
```

b) Visual inspection of data



	count	mean	std	min	25%	50%	75%	max
CustID	4520.0	7.002260e+06	1304.955938	7000000.0	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

b) Visual inspection of data

The mean age of customer is around 15 years which means that the insurance is mainly taken by parents for their kids.
The median is 13 which is slightly deviated from the mean.
customer tenure is approx 14.5 years that means customers tends to stay for a longer period with the company.
But there are customers who have been associated for just 2 years and maximum is 57 years.

b) Visual inspection of data

existingprodtype shows the type of product which customers have opted. In the data it shows that the data type of
this column is int64 but it should be object data type as it shows the product classes. Hence, we need to change it.
Most of the customers opted for product type 4.

b) Visual inspection of data

Complaint column's data type is boolean but it shows int64, hence we need to change this as well.
CustCareScore data type is float 64 but it seems to be ordinal data type, hence we need to convert it.

c) Understanding of attributes

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

c) Understanding of attributes

Variable	Description
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

c) Understanding of attributes

Variable	Description
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

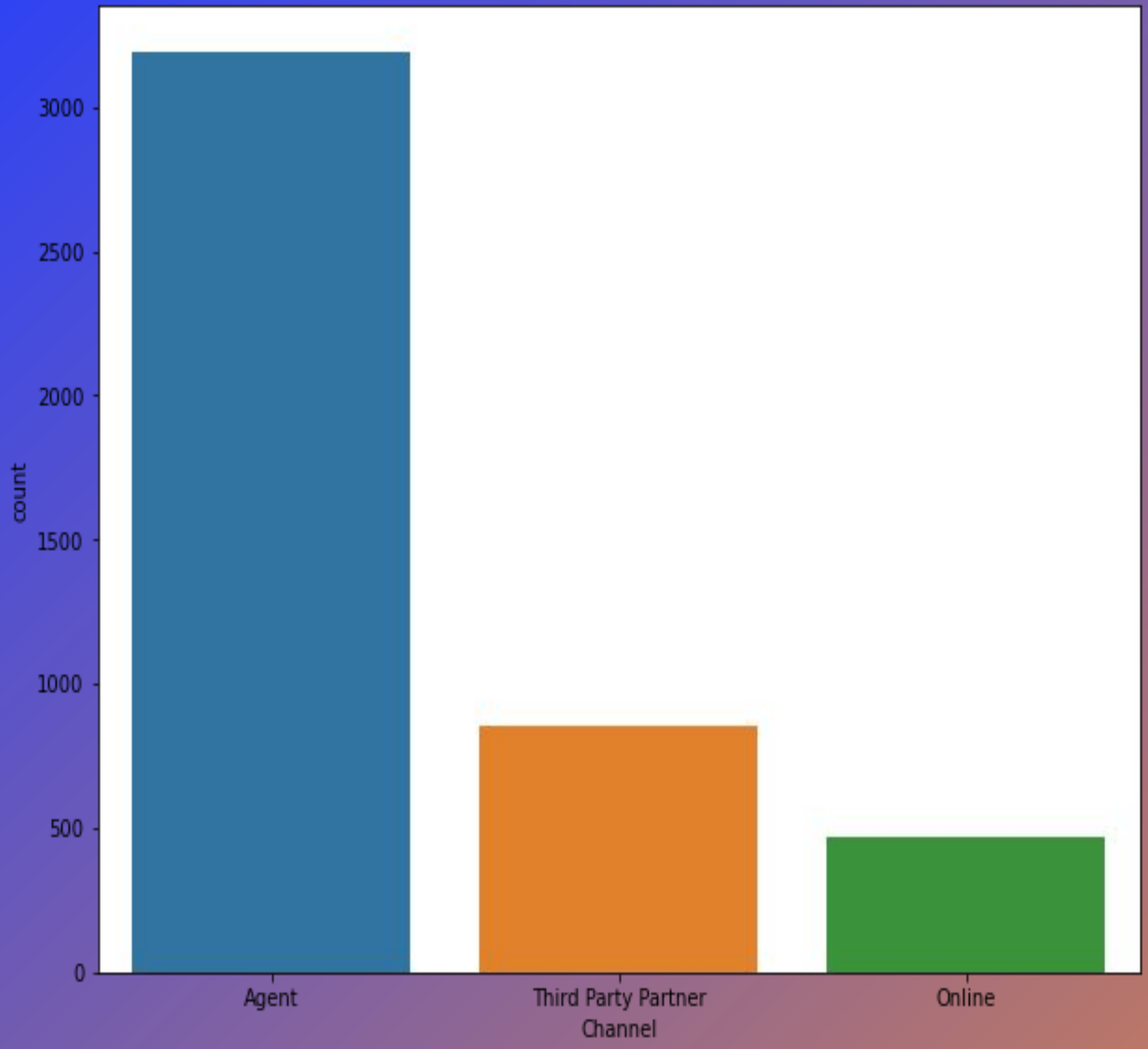
c) Understanding of attributes

Variable	Description
LastMonthCalls	Total calls attempted by company to a customer for cross sell
CustCareScore	Customer satisfaction score given by customer in previous service call

There is no need to rename any column name as all are variable and there is no space or special characters in them.

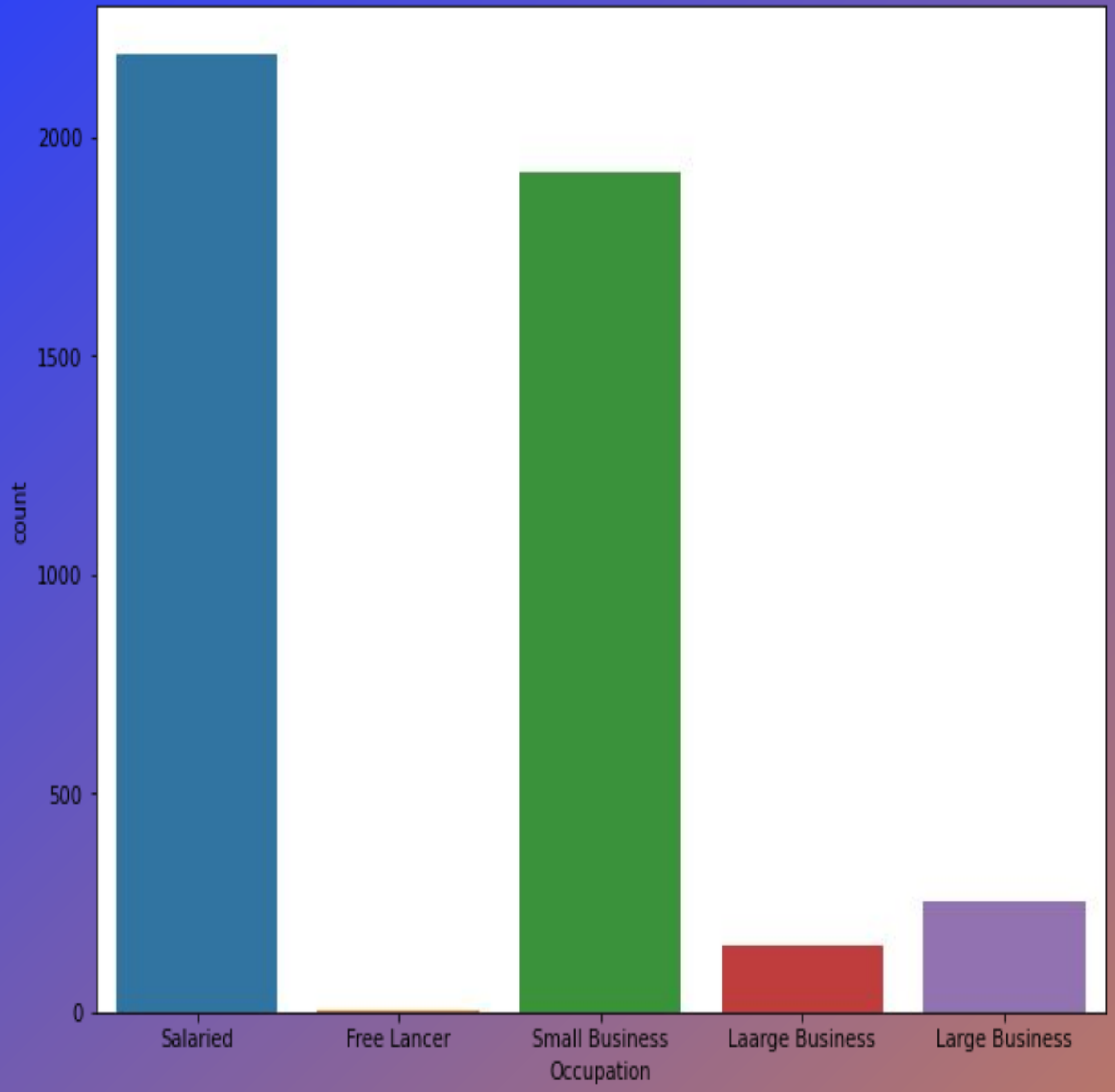
Question 3 - Exploratory Data Analysis

a) Univariate Analysis: Channel



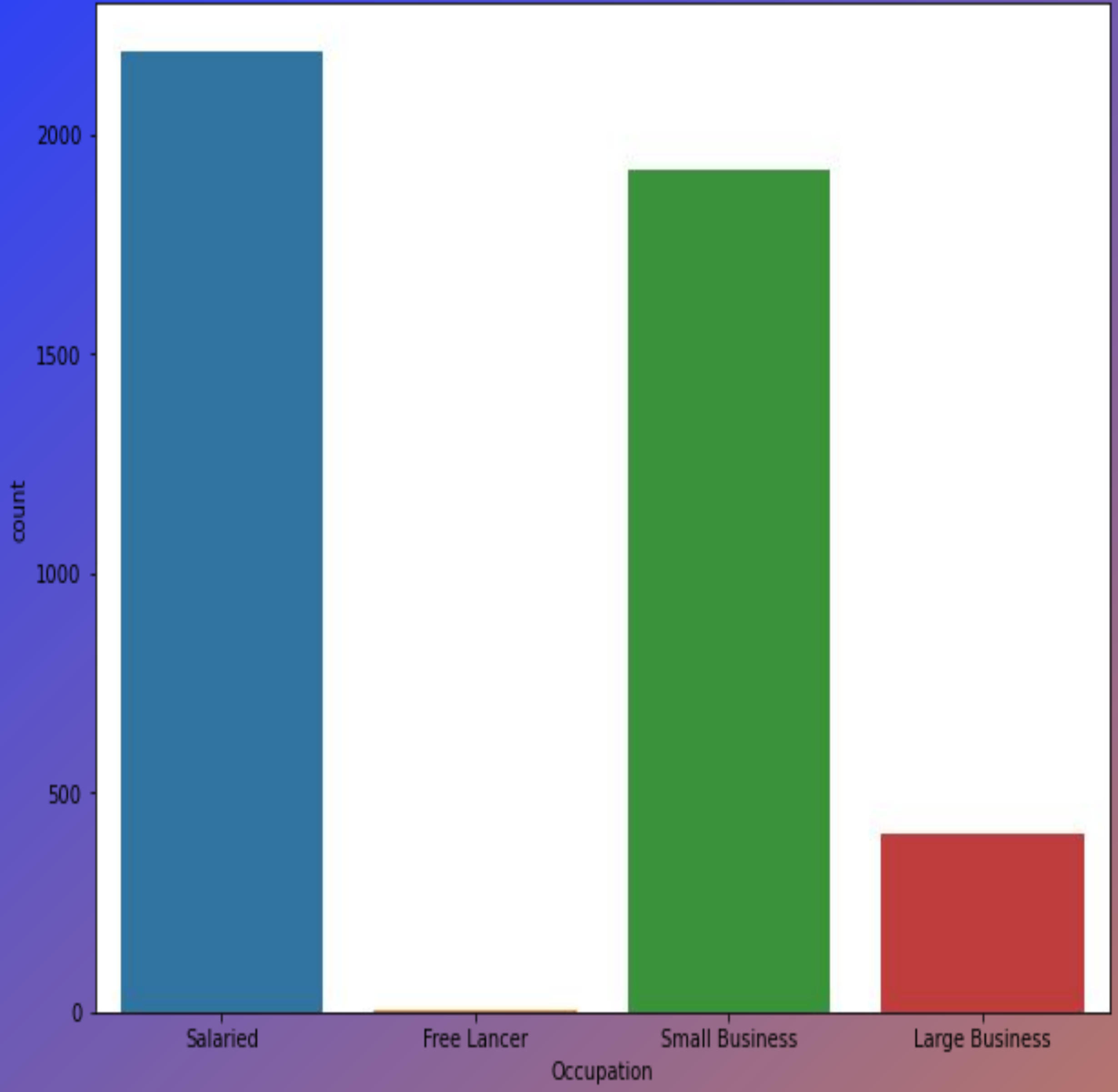
From the graph of Channel which tells about the channel through insurance was done. There are 3194 customers who took insurance through Agents, followed by Third Party Partner through which 858 customers took insurance and 468 customers took through online.

a) Univariate Analysis: Occupation



From the graph of Occupation I can see that the Laarge Business and Large Business class are the same, hence they need to me merged together as Large Business and I have replaced all Laarge Business by Large Business.

a) Univariate Analysis: Occupation



From the graph of Occupation I can tell that most of the insurance were purchased by Salaried person which are 2192 in number, followed by Small business owners which are 1918 in number.

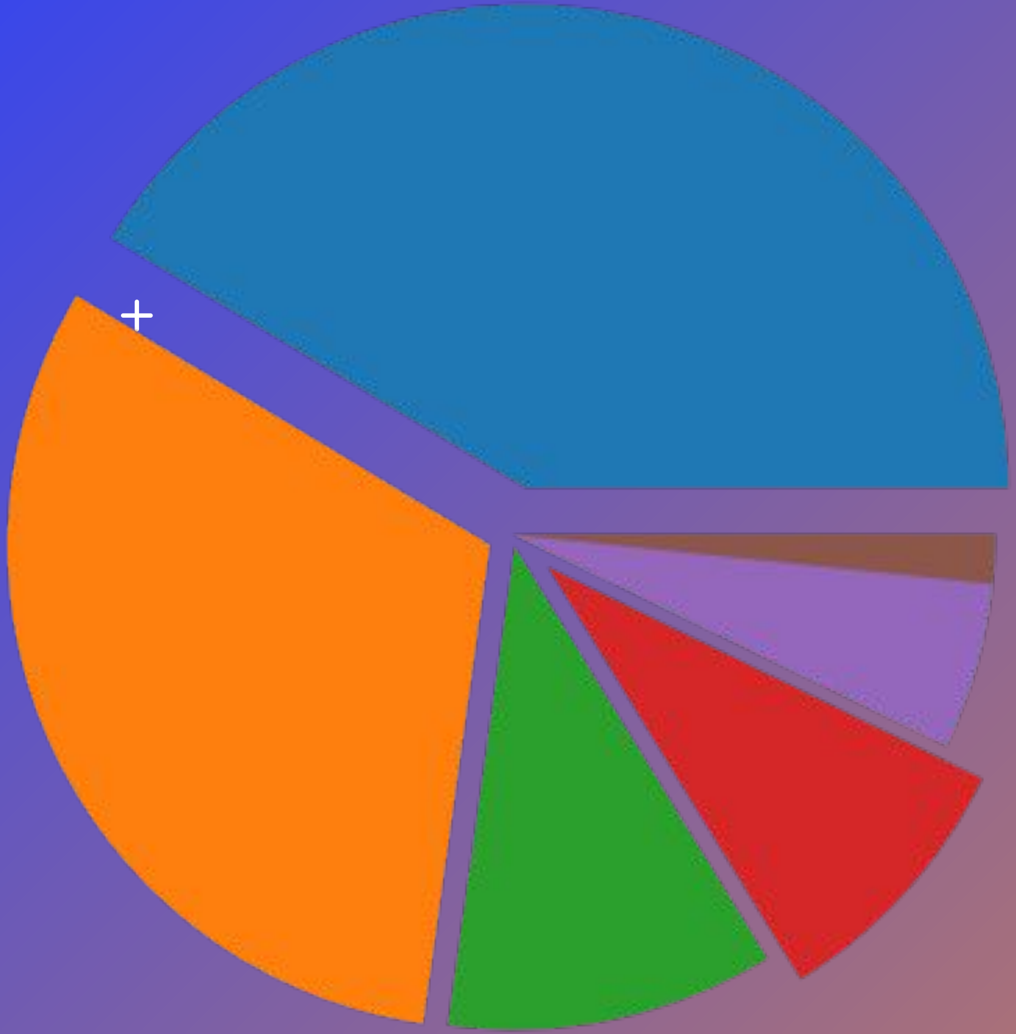
So, the company can target more salaried and small business owners to sell their insurance.

There are 408 Large business owner and only 2 free lancer.

a) Univariate Analysis: EducationField

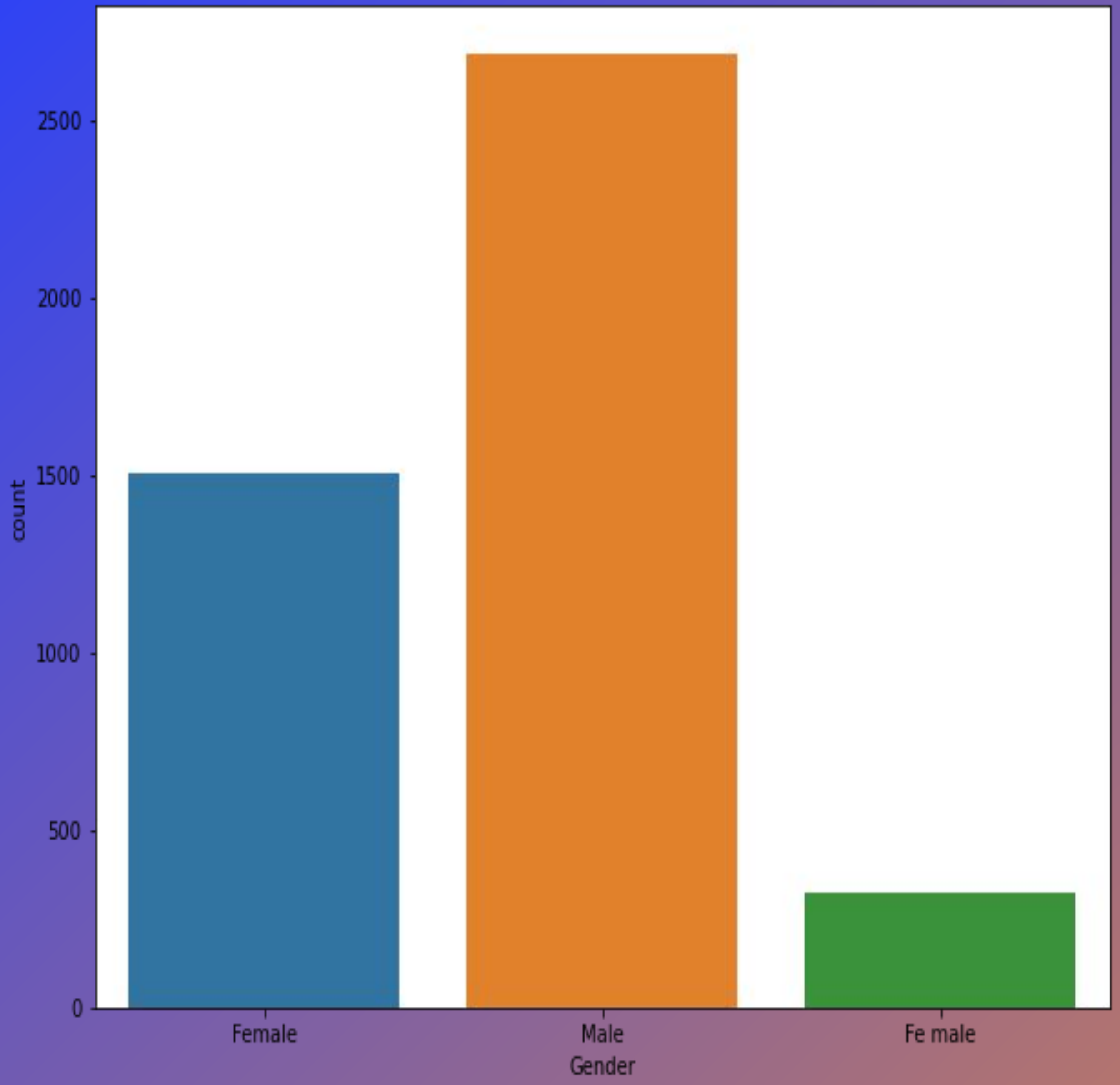
From the graph of EducationalField I can see that the UG and Under Graduate class are the same, hence they need to me merged together as Under Graduate and I have replaced all UG by Under Graduate.

a) Univariate Analysis: EducationField



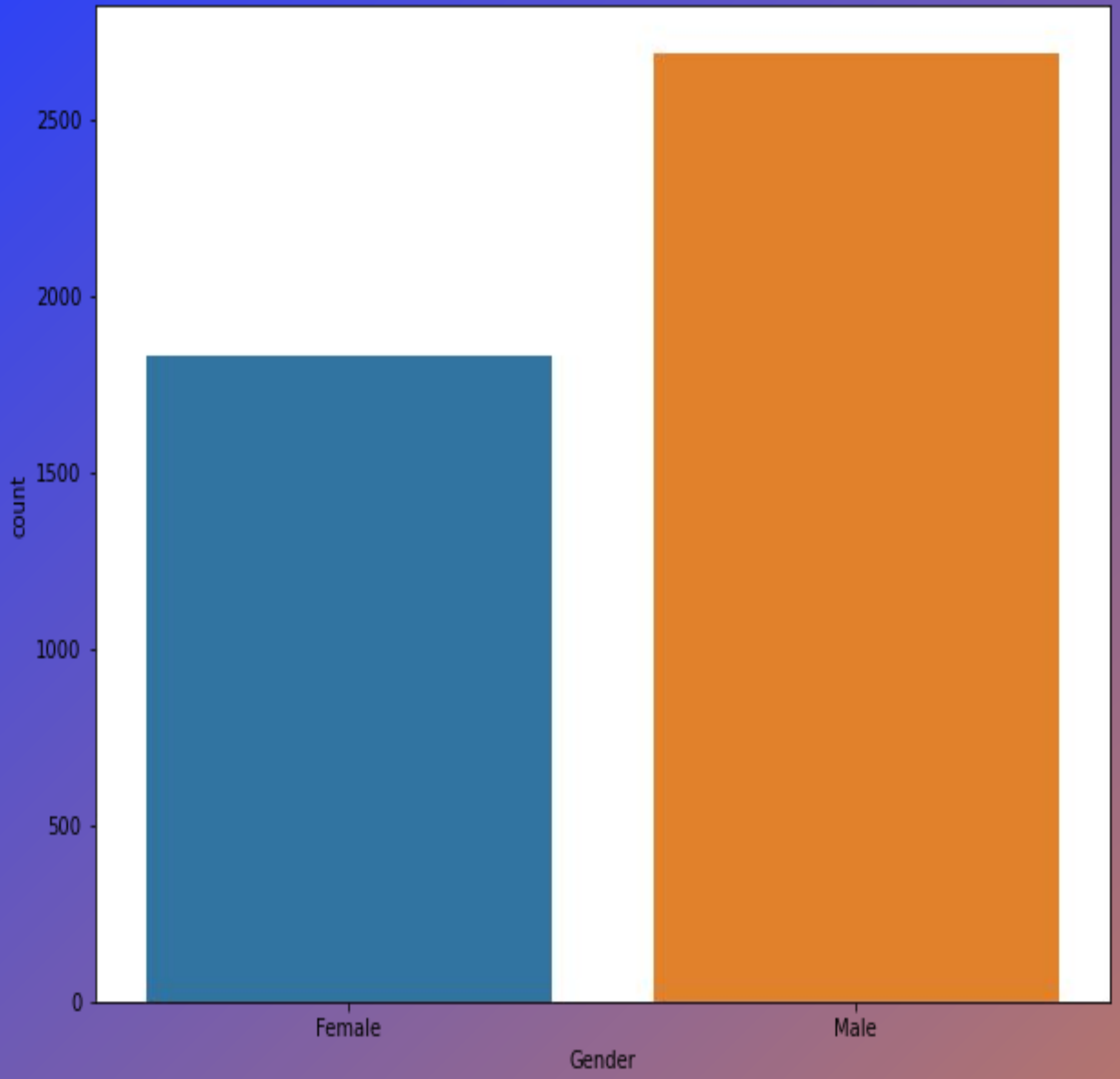
From the graph of EducationalField I can tell that most of the insurance were purchased by Graduate, 1870 in number followed by Under Graduate, 1420 in number. There are 496 Diploma holders and 408 Engineers. Post graduate are 252 in numbers and MBA are 74.

a) Univariate Analysis: Gender



From the graph of Gender I can see that in the class Fe male, one space has been given and it should be Female. Hence, I have replaced all Fe male with Female.

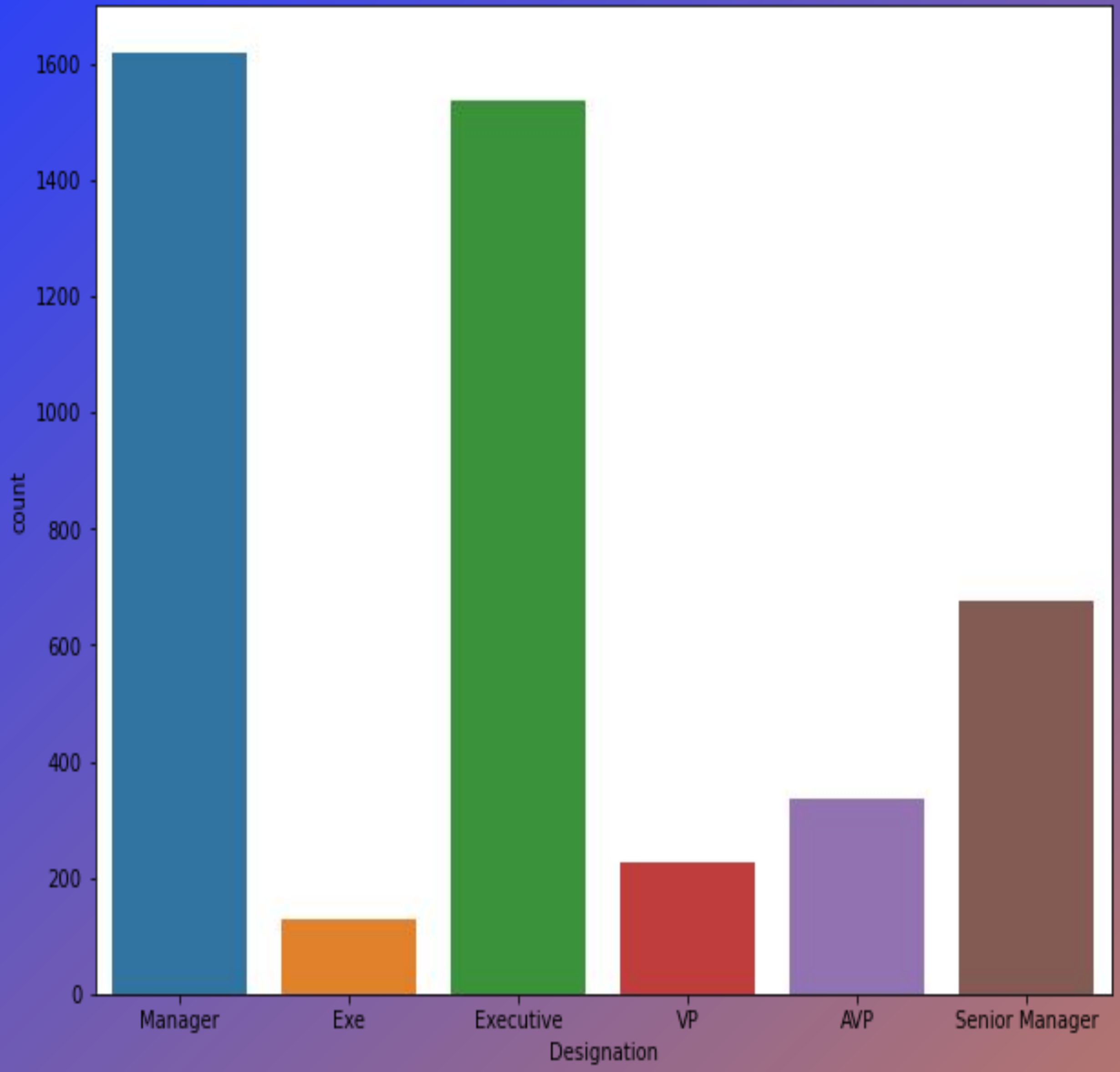
a) Univariate Analysis: Gender



From the graph of Gender I can tell that the data is slightly imbalanced as there are 2688 males and 1832 Females in the data.

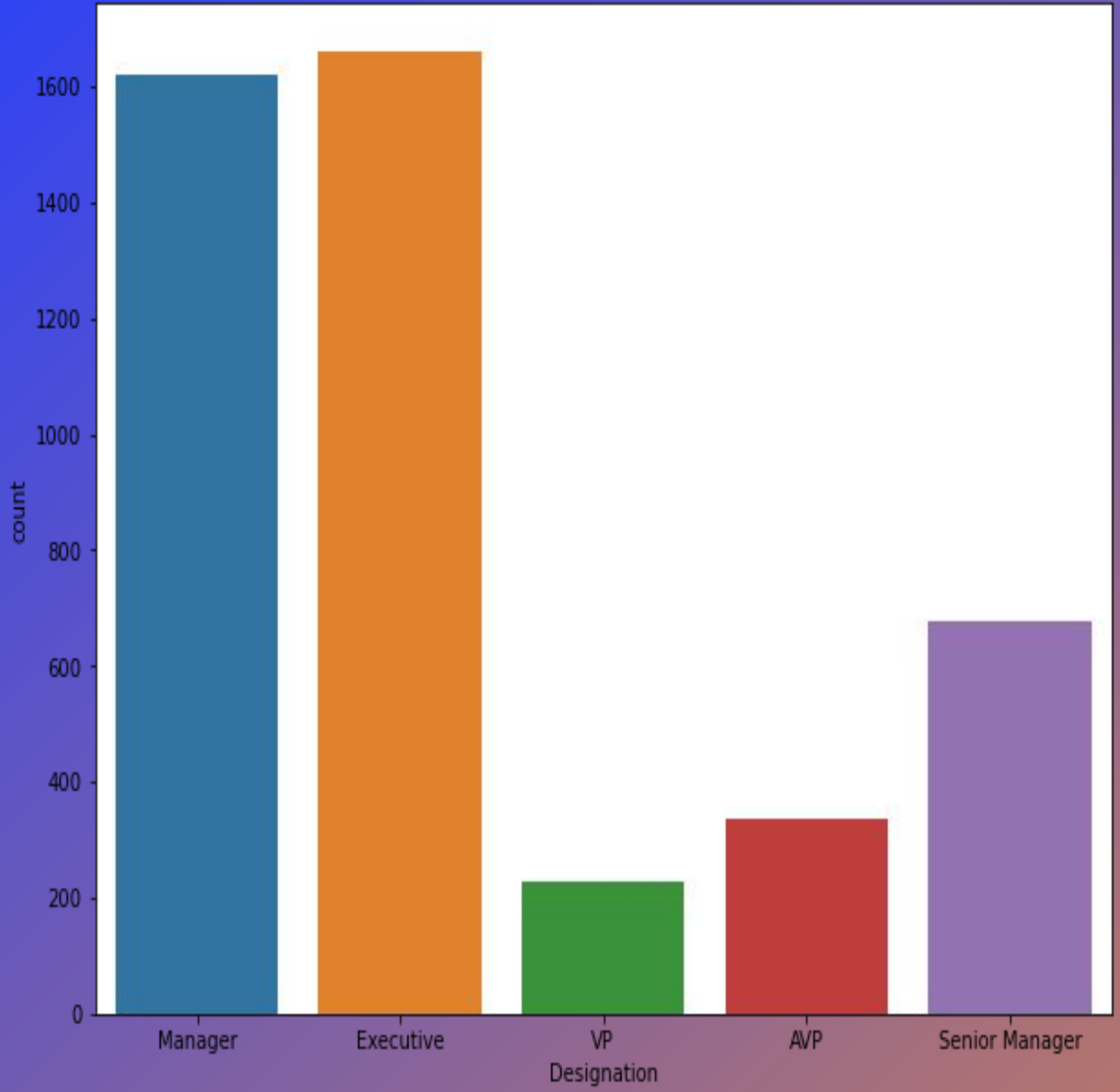
But we need not use SMOTE as we need not add new data having gender as Female.

a) Univariate Analysis: Designation



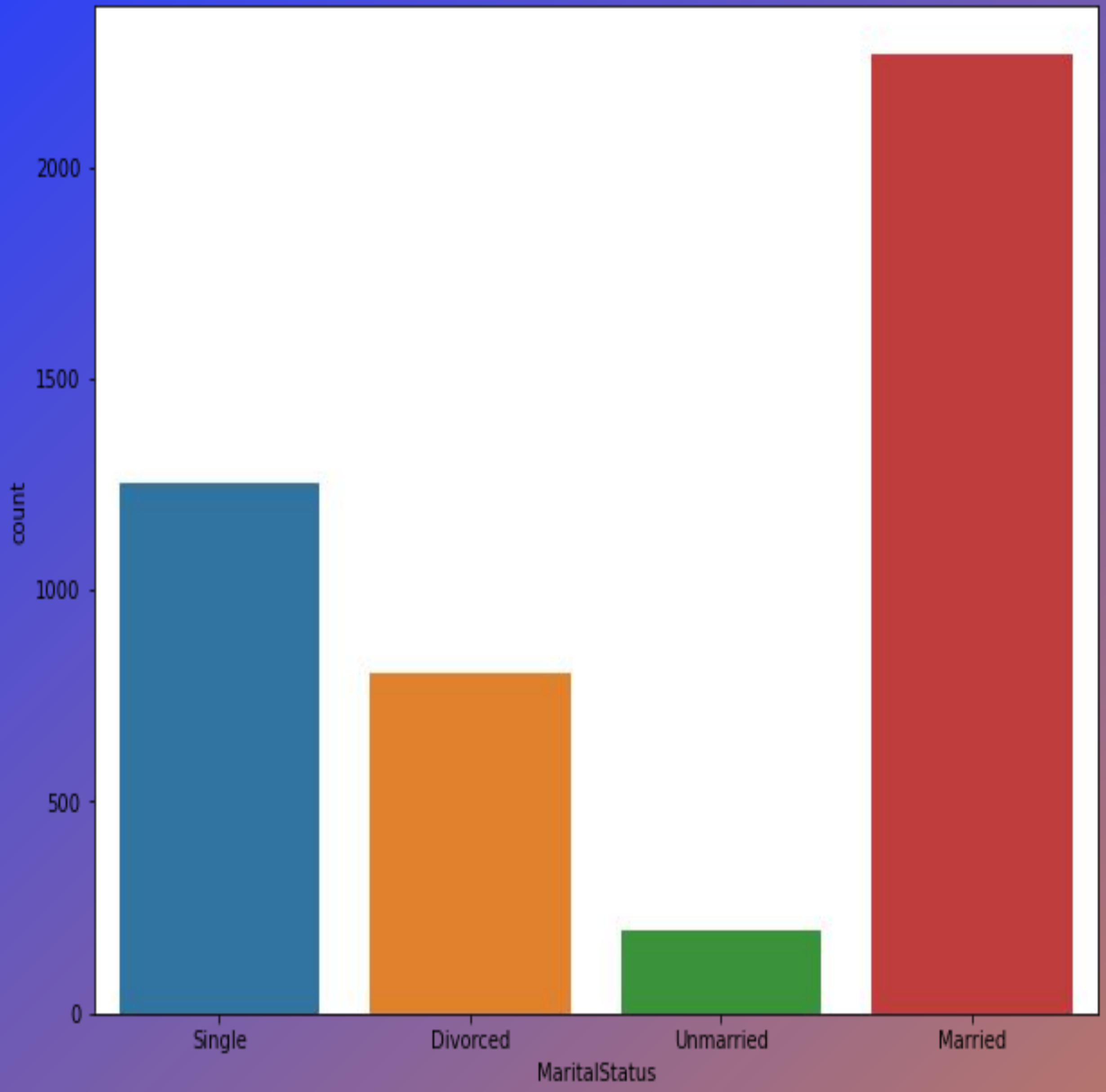
I can see that the two classes named as Exe and Executive are same as Exe is short of Executive, hence I have merged Exe with Executive by changing its name as Executive.

a) Univariate Analysis: Designation



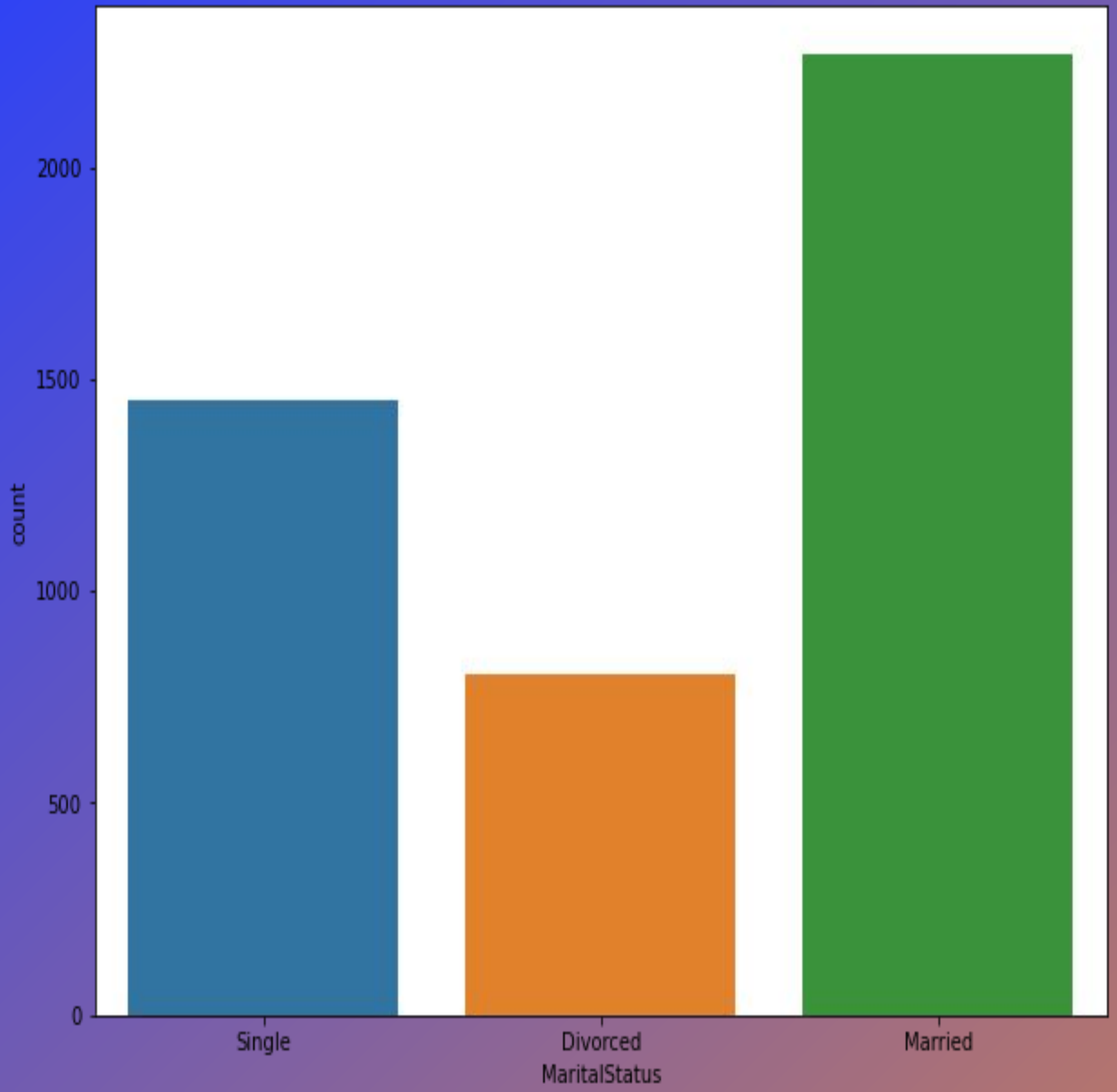
I can see that many Executives have purchased the insurance followed by Managers. Hence, we can target these working class for insurance. Moreover, we need to find new plans for VP and AVP, so that the number of insurance purchased by them increases.

a) Univariate Analysis: MaritalStatus



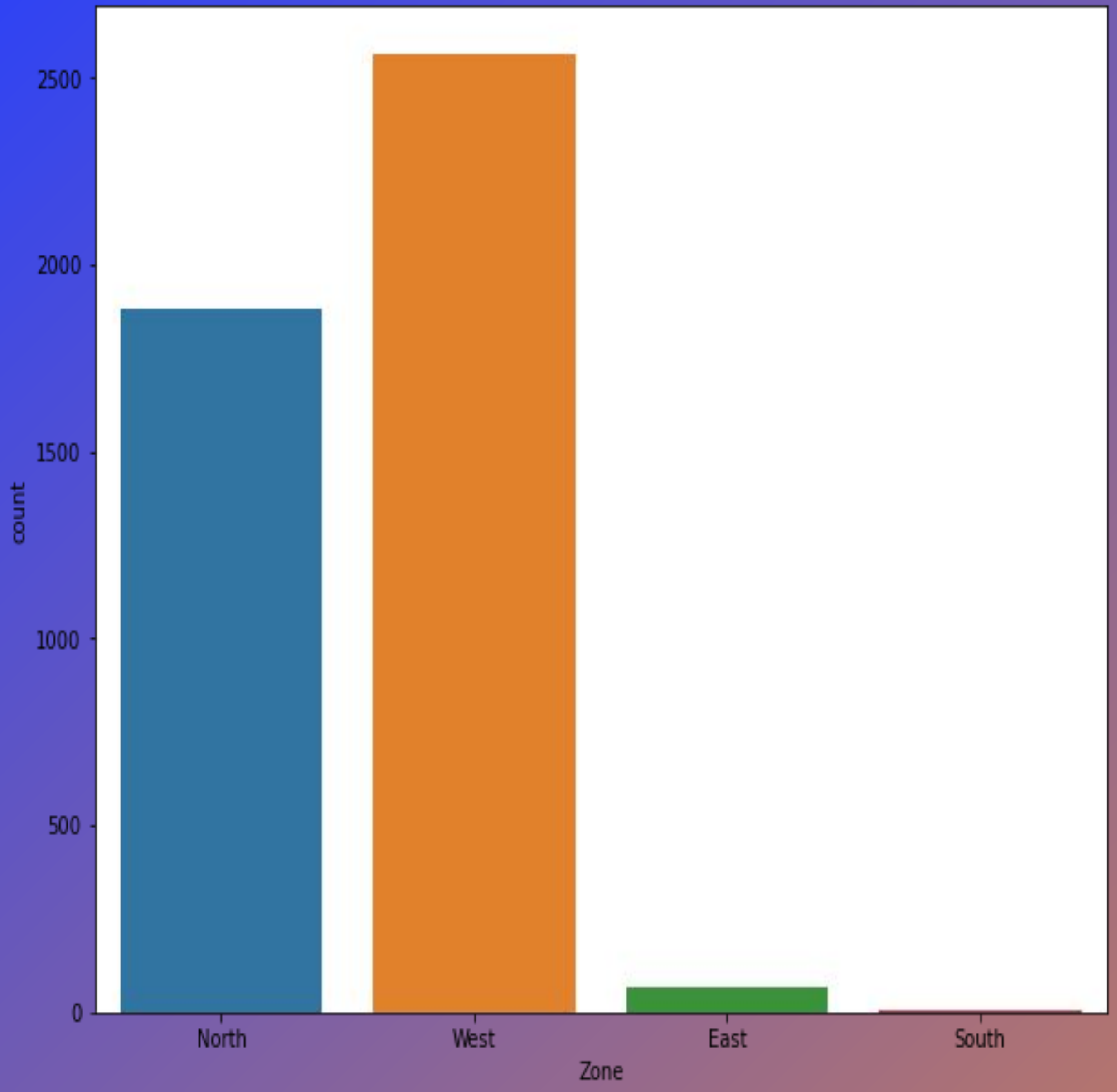
I can see that the two classes named as SIngle and Unmarried means the same, hence I have merged Unmarried with Single by changing its name as Single.

a) Univariate Analysis: MaritalStatus



Most of the insurance purchased are by the Married people followed by Singles and then Divorced people. The numbers are 2268, 1448 and 804 respectively. The company needs to make singles and divorced people aware about the benefits of insurance so that their numbers increases.

a) Univariate Analysis: Zone

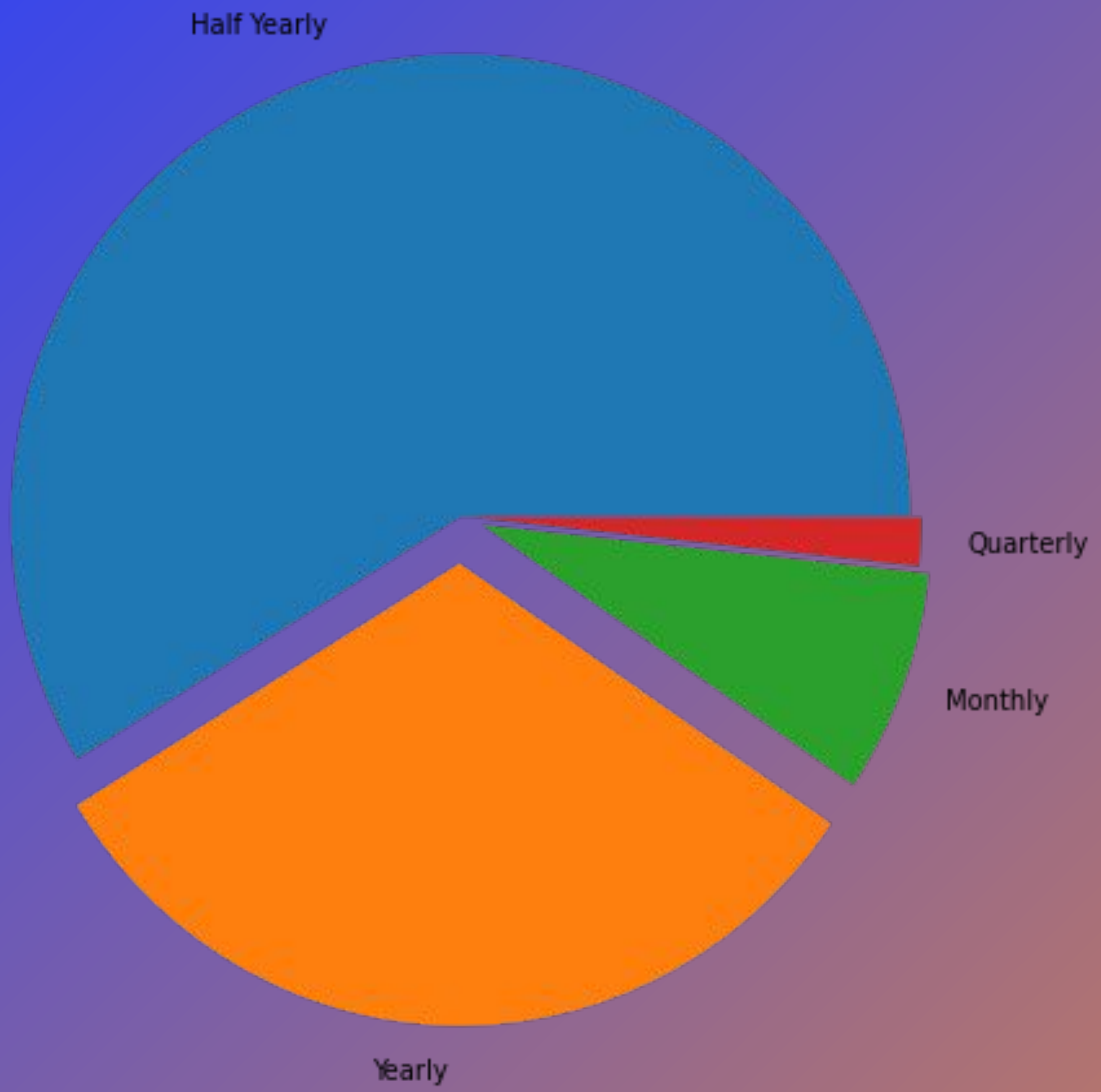


Most of the insurance purchased are from the West which are 2566 in numbers followed by North which are 1884 in numbers.

There are very few customers from East and South which are 64 and 6 in numbers respectively.

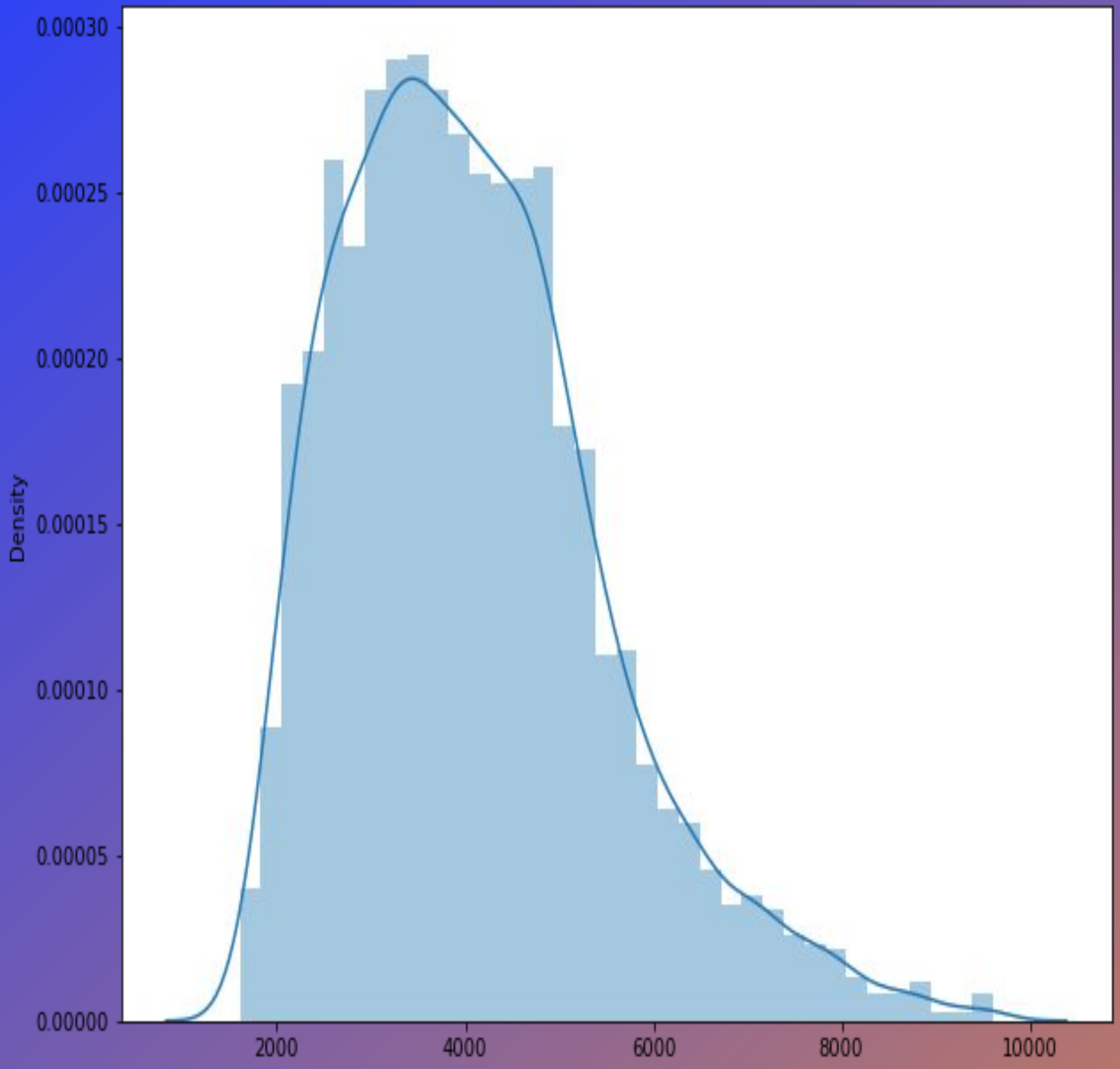
The company need to get more agents or third party in East and South zone so as to gain more customers.

a) Univariate Analysis: PaymentMethod



2656 customer prefer to pay the premium Half yearly where as 1434 customers prefer paying every Year. There are few customers who prefer to pay premium monthly and quarterly. They are 354 and 76 in numbers respectively.

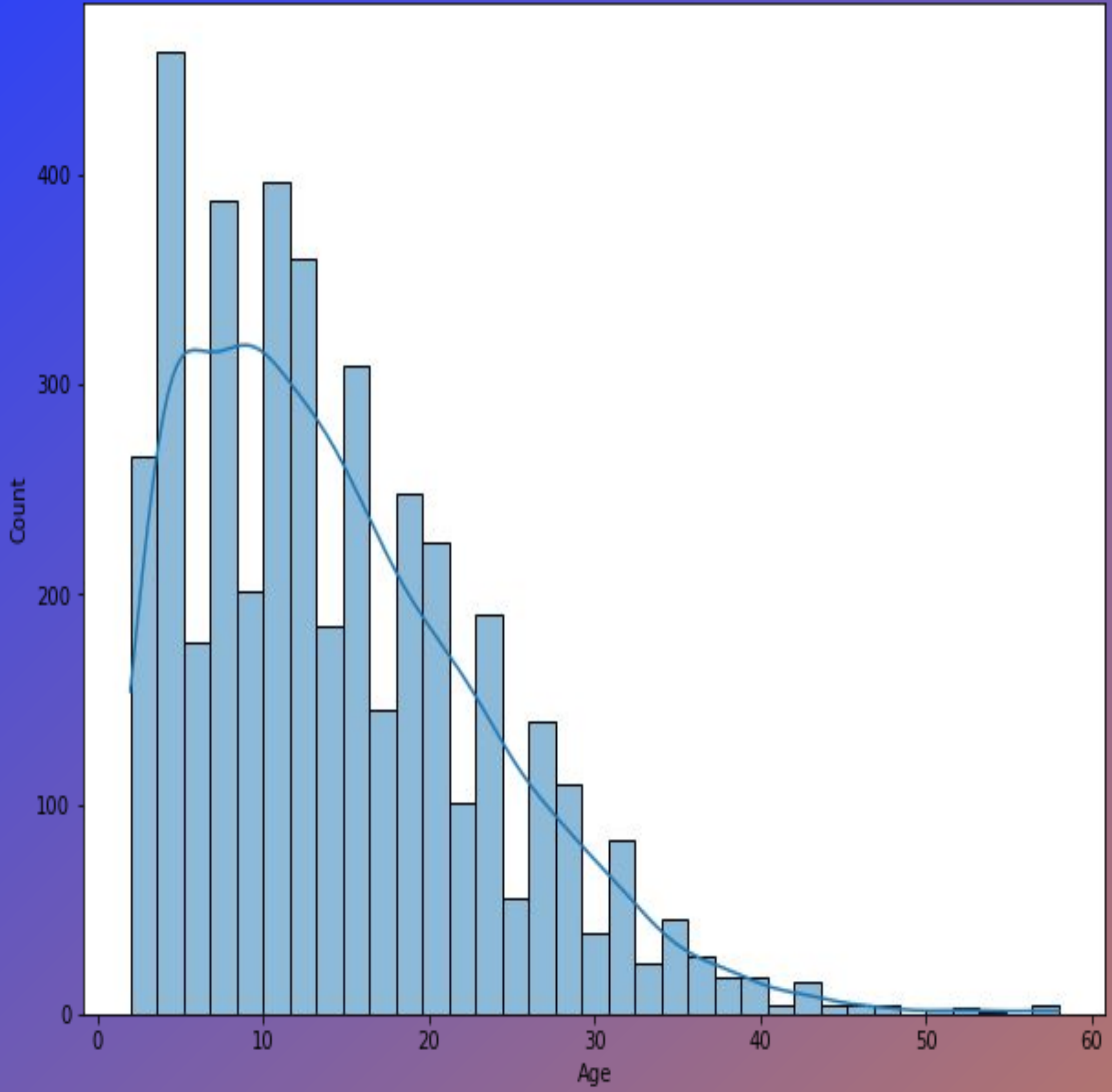
a) Univariate Analysis: AgentBonus



The agentbonus seems to be slightly normally distributed. This is my target column as I need to find or predict Bonus which should be given to the agents.

The mean bonus given to the agents is .

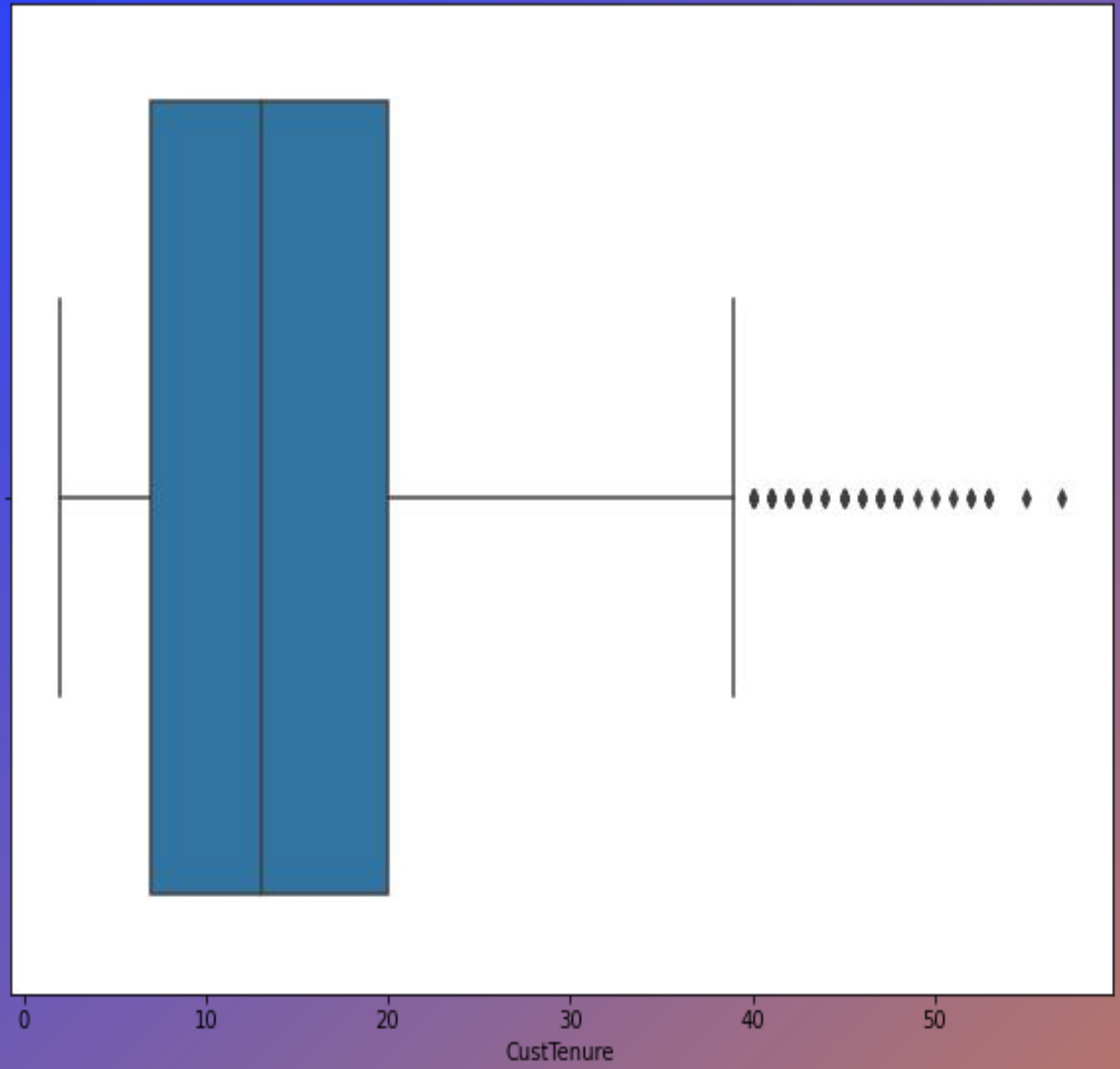
a) Univariate Analysis: Age



The age seems to be right skewed which means that most of the customers are young and must be purchased by their parents.

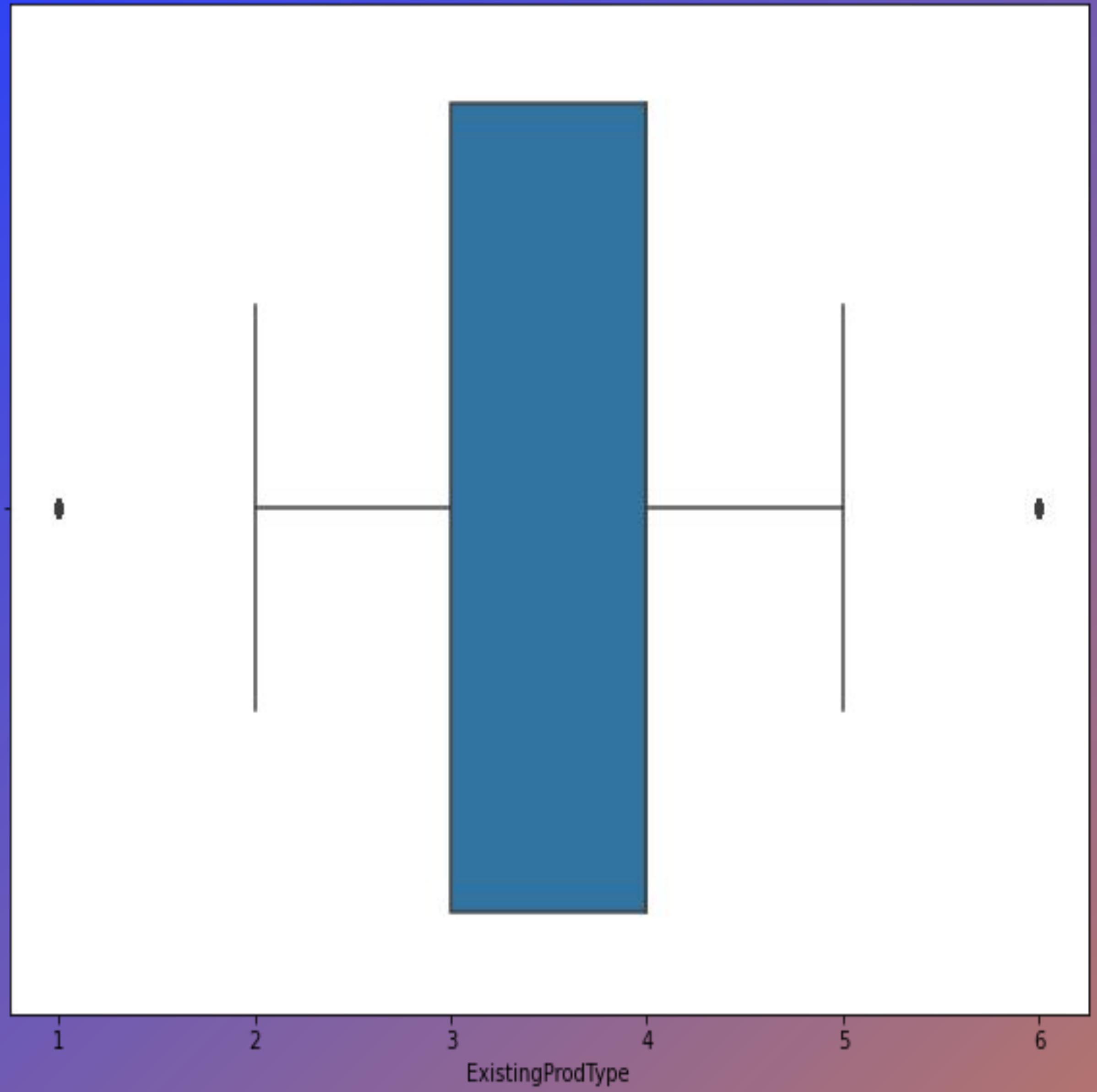
The average age of the customers for whom insurance were purchased is .

a) Univariate Analysis: CustTenure



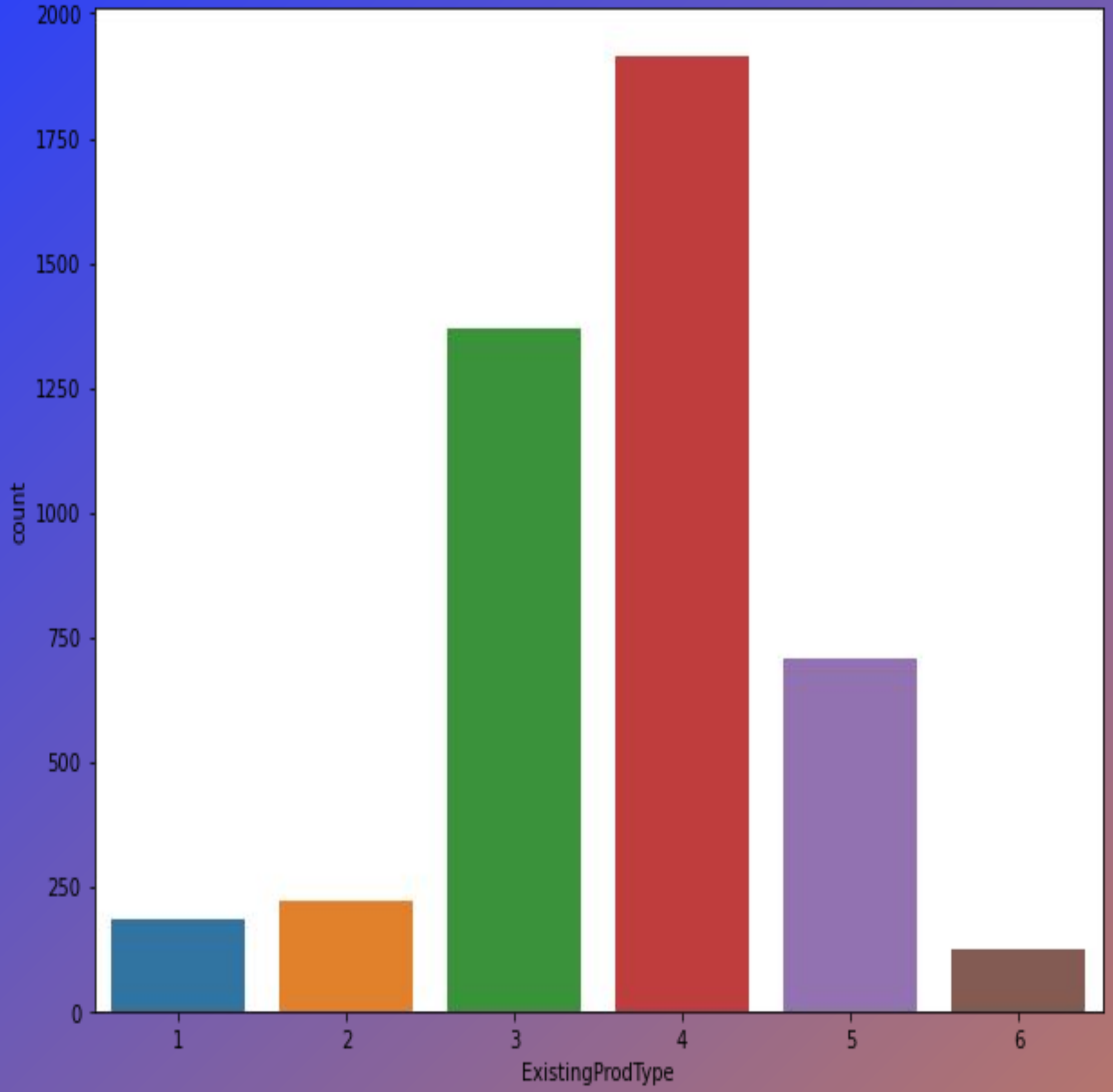
The average customer tenure is around 13 years.
I can see that there are outliers in this feature which needs to be treated.

a) Univariate Analysis: ExistingProdType



ExistingProdType feature is been given as integer type column but the data type should be categorical as can be seen from the boxplot of the data when it is assumed to be numerical type.
In order to better understand this feature, I have converted the data type from integer to categorical.

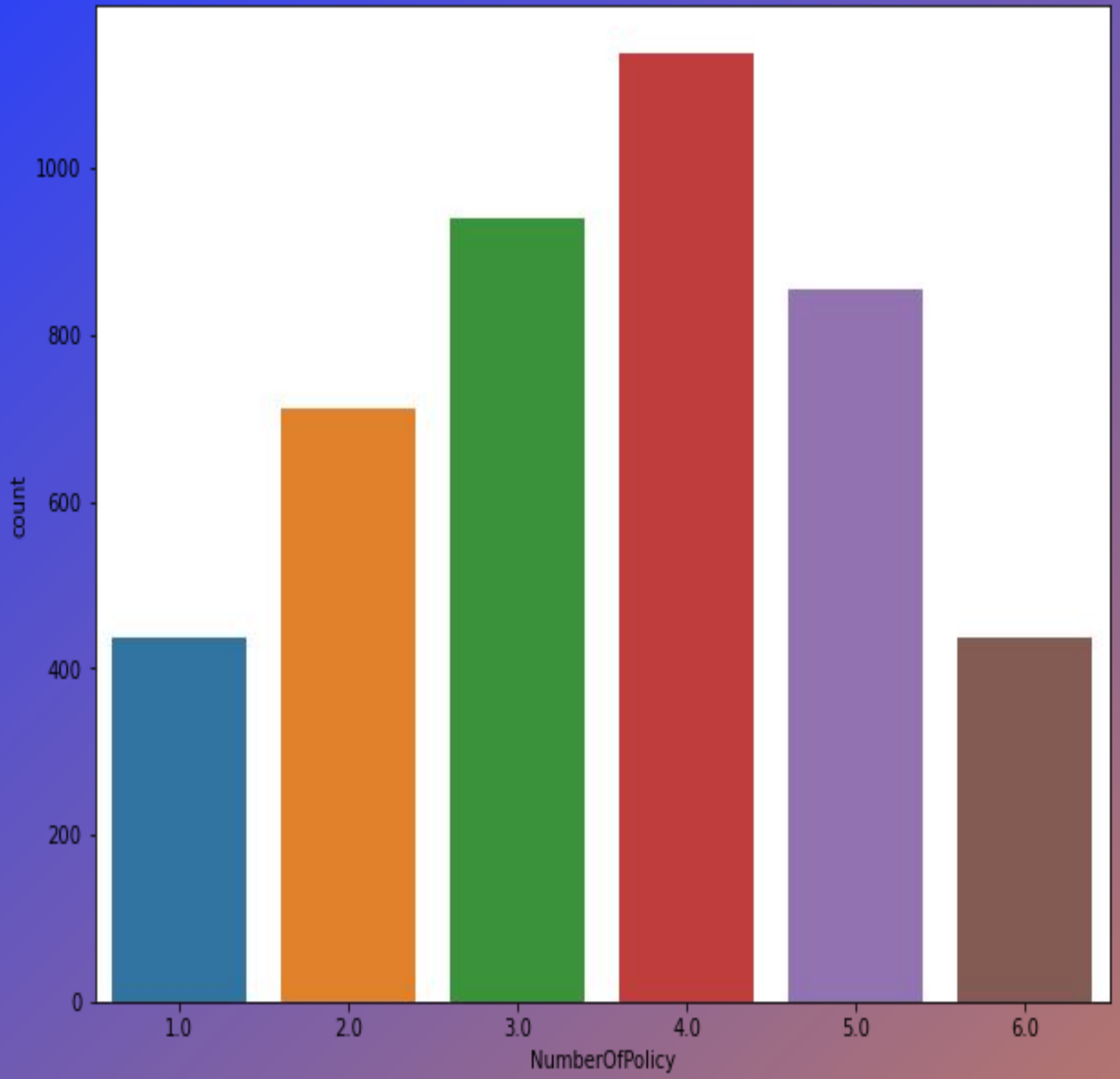
a) Univariate Analysis: ExistingProdType



I can see that most people who have purchased insurance have opted for the type 4 followed by type 3 and then type 5.

So, we can try to find why other products are not that successful and act accordingly, either by making some changes in the plan or completely scrapping it.

a) Univariate Analysis: NumberOfPolicy

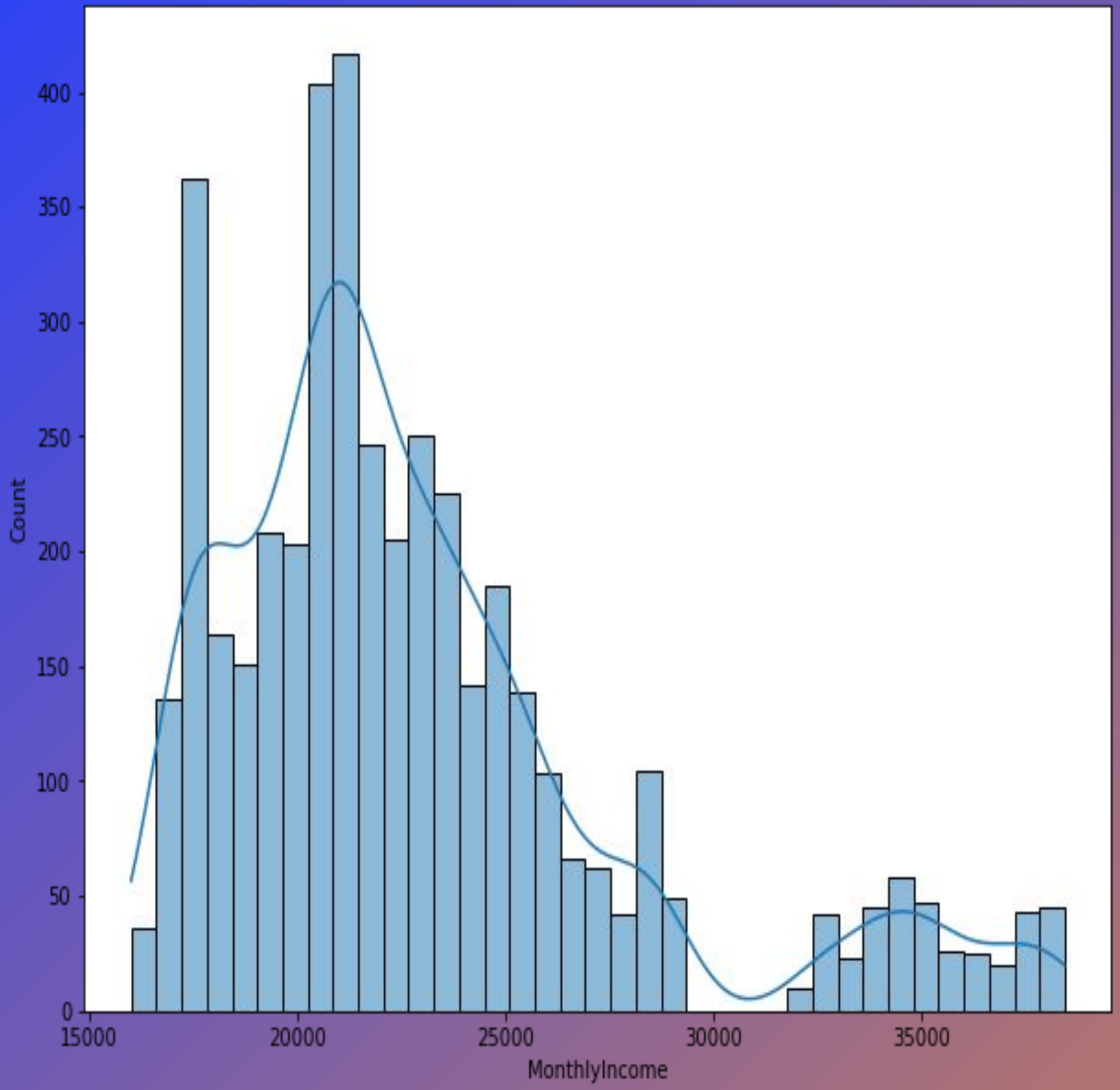


1094 customers have purchased 4 policies, followed by 939 customers purchasing 3 policies.

There are 856 customers who have purchased 5 policies.

Very less people have purchased 6 policies.

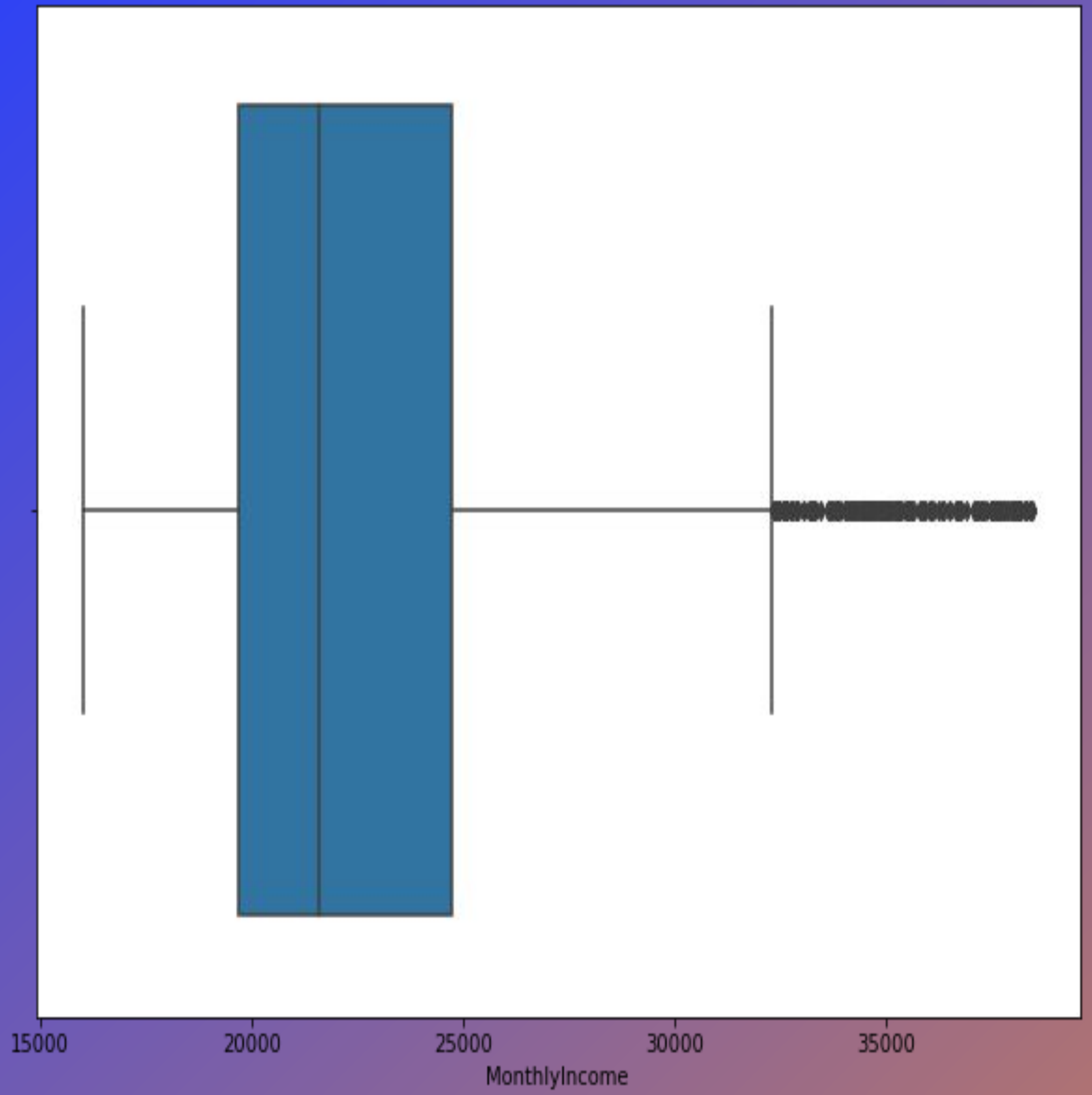
a) Univariate Analysis: MonthlyIncome



The Univariate Analysis of monthly income of the customers seems to be right skewed.

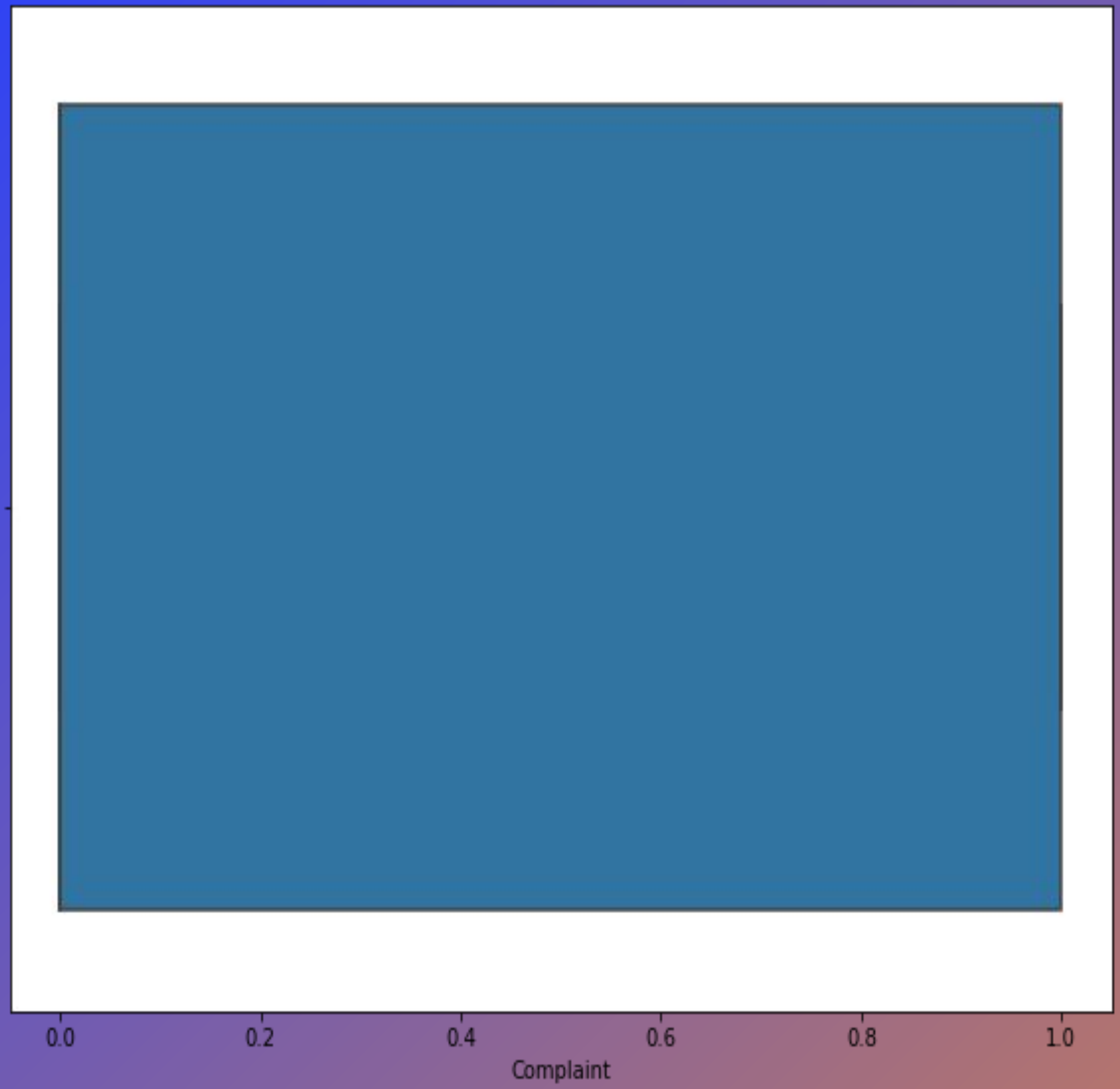
There are many outliers in the data which are genuine as there might be customers with high income.

a) Univariate Analysis: MonthlyIncome



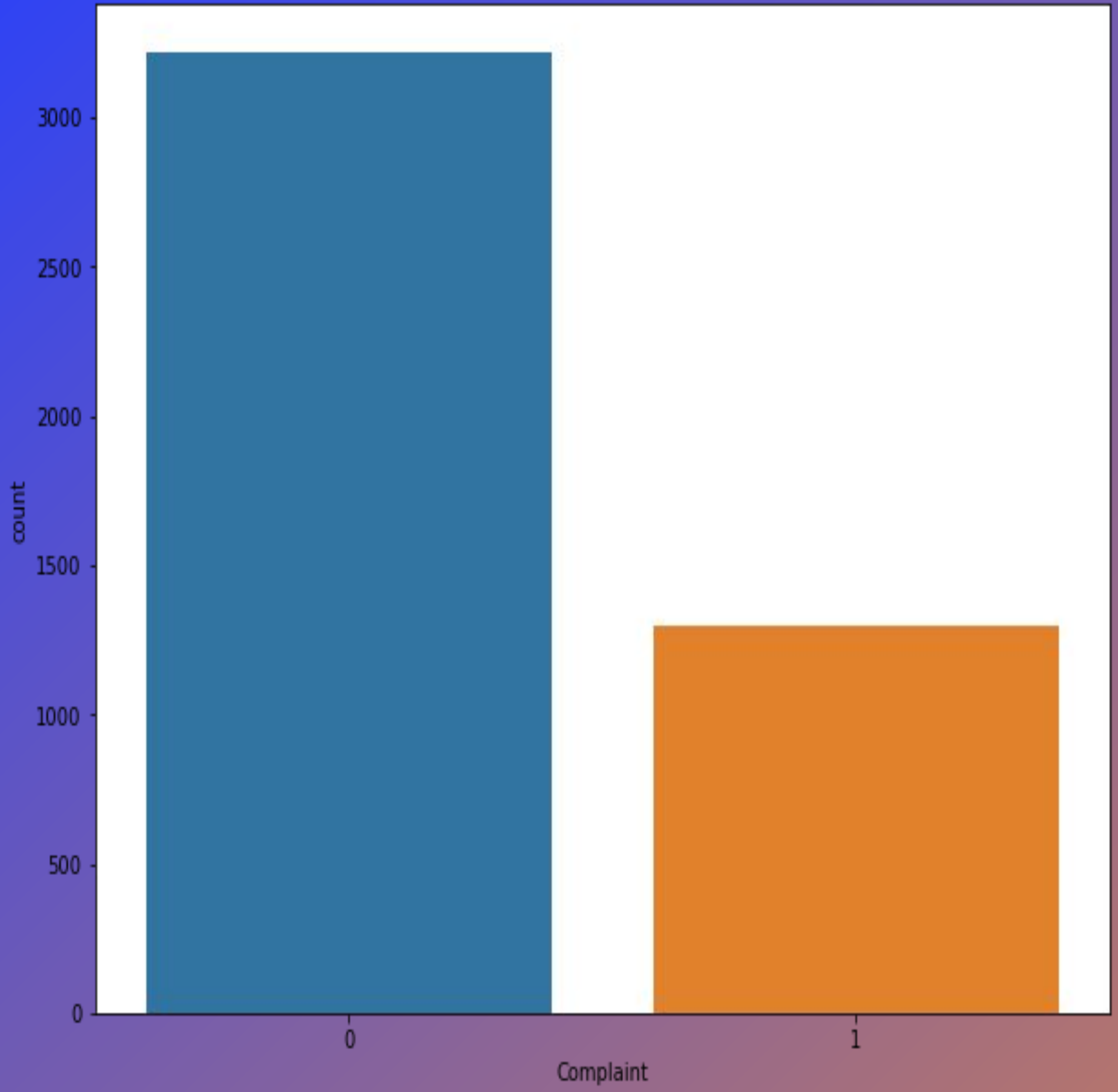
There are many outliers in the data which are genuine as there might be customers with high income.
The median of the data is which suggest that most of the customer’s monthly income is this.

a) Univariate Analysis: Complaint



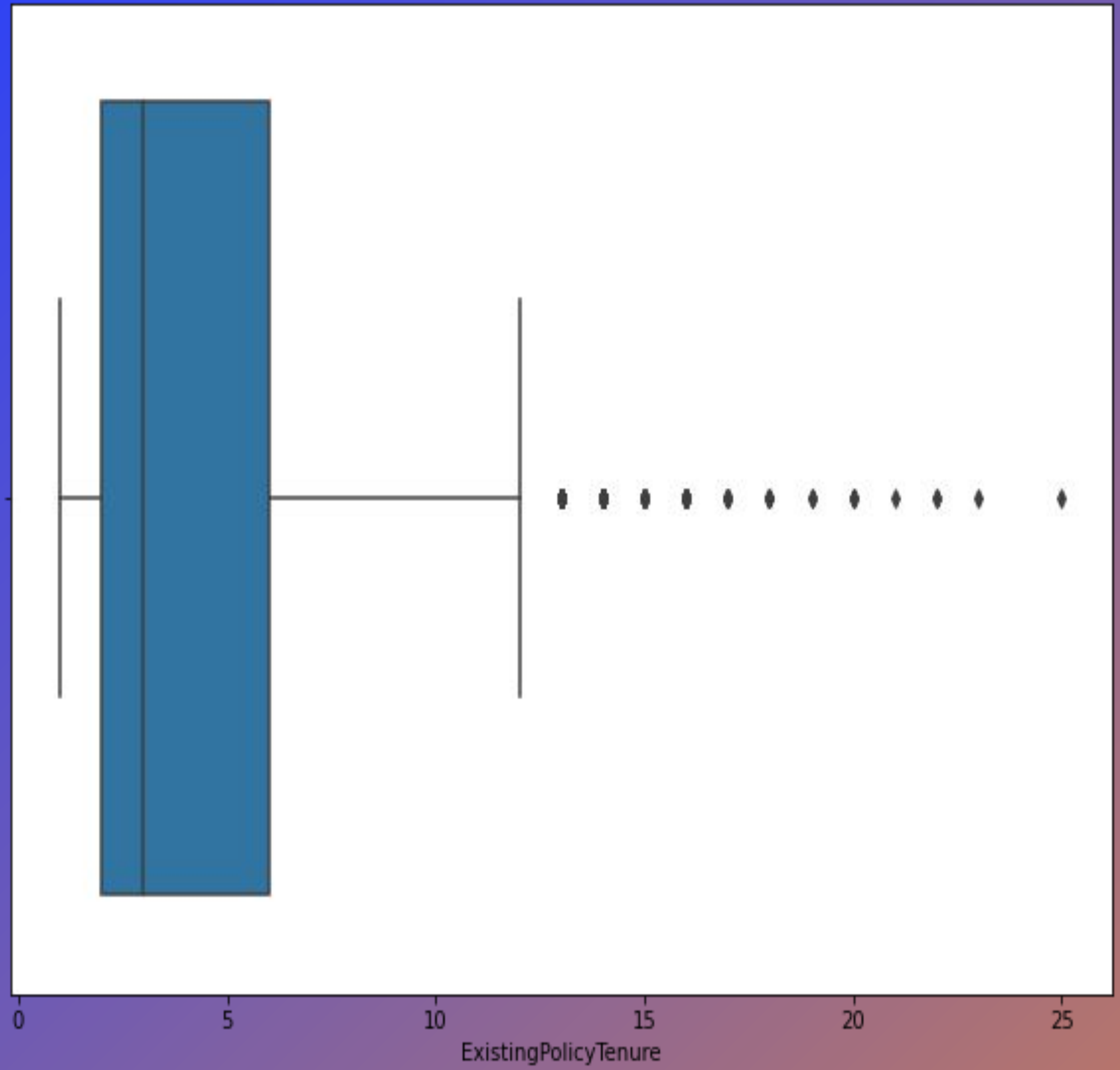
Complaint feature is been given as integer type column but the data type should be categorical as can be seen from the boxplot of the data when it is assumed to be numerical type.
In order to better understand this feature, I have converted the data type from integer to categorical.

a) Univariate Analysis: Complaint



I can see that most of the customers didn’t have any complaint against the agents or third party.
1298 customers have complained which is huge in number as compared to the total number of data. Hence, we need to check what is the reason for the complaints and work upon it.

a) Univariate Analysis: ExistingPolicyTenure

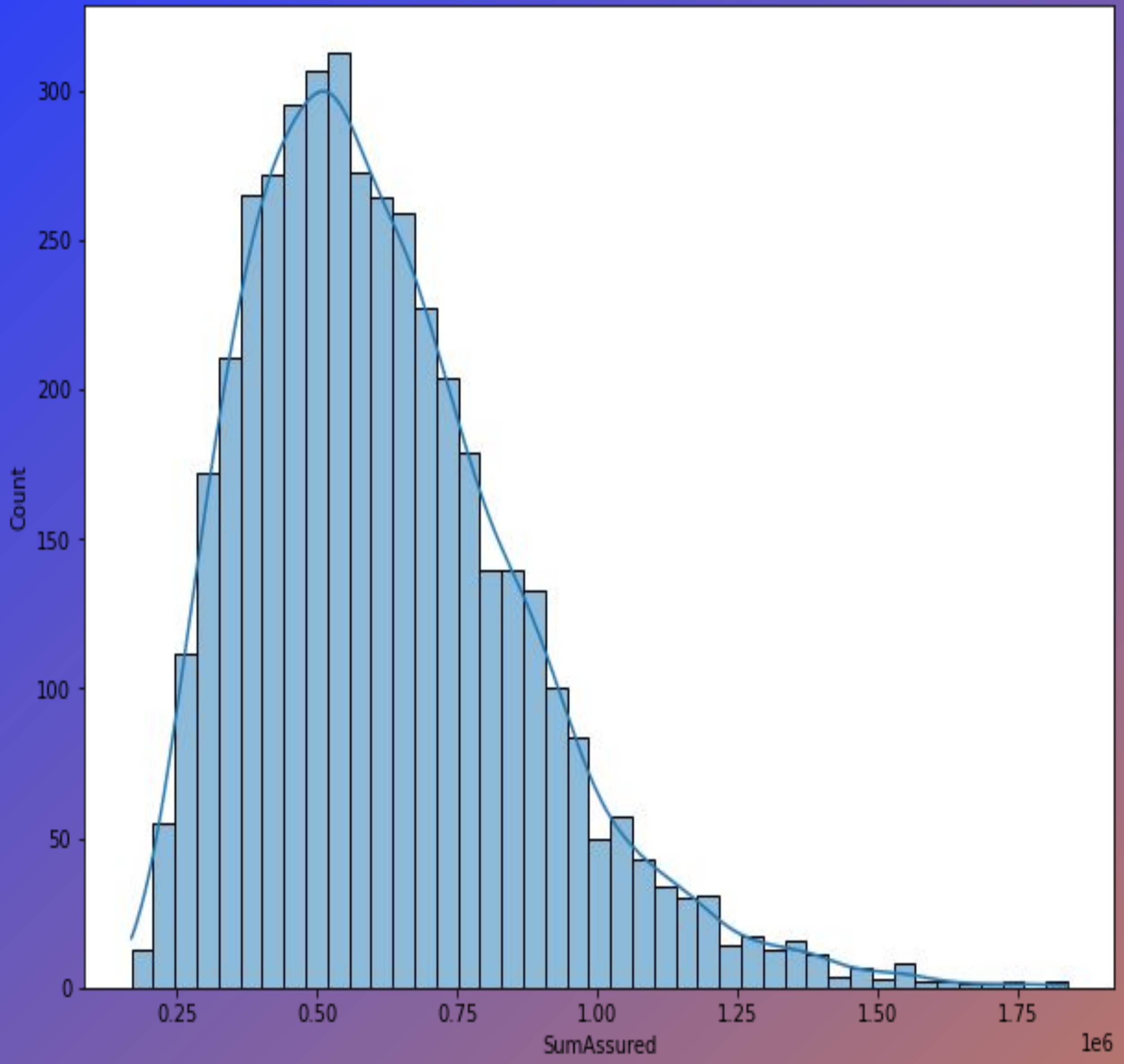


Average Policy tenure is approx 3 years.

From the graph I can see that there are many outliers.

Maximum customers are having 1 to 5 years of policy tenure.

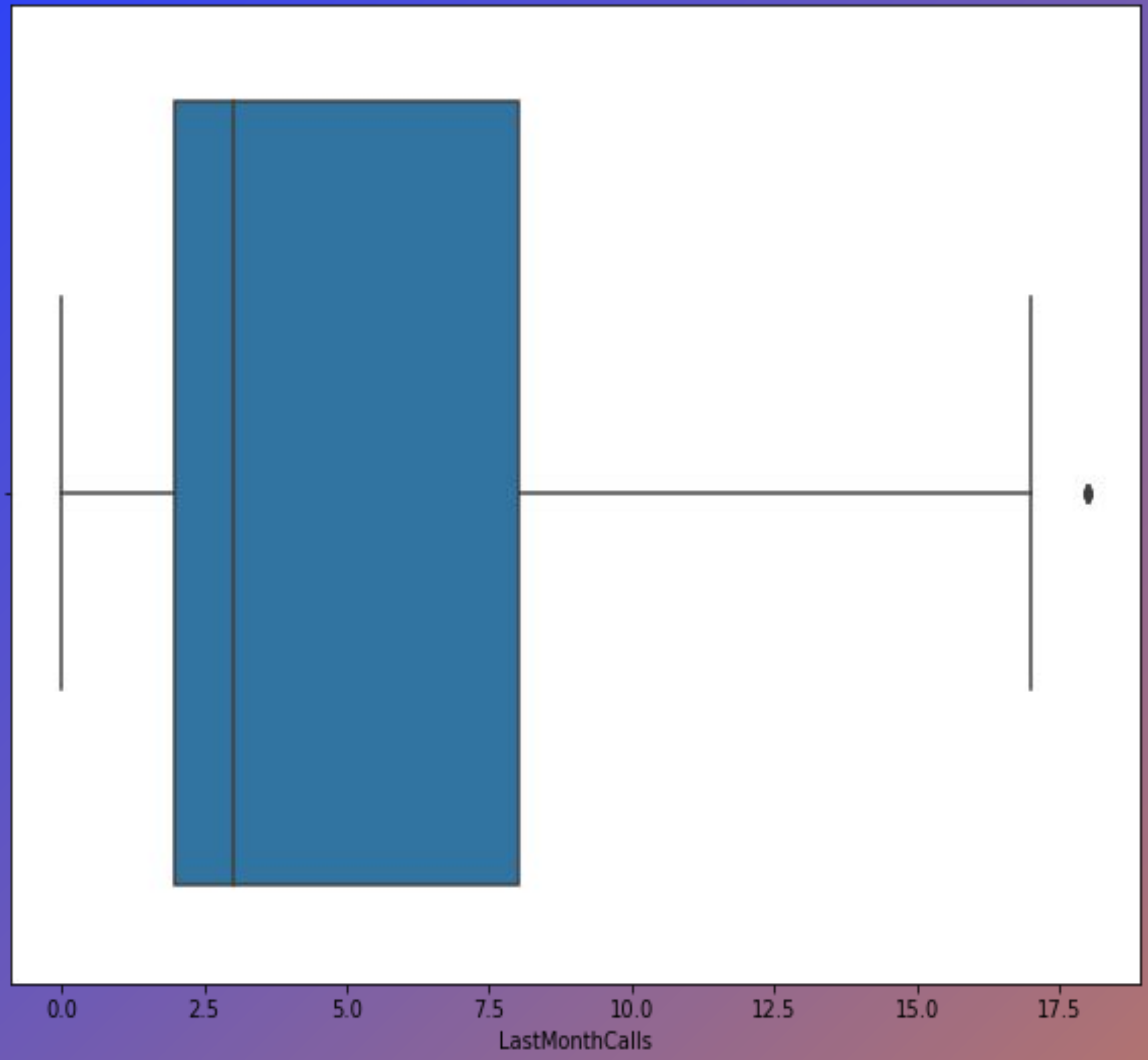
a) Univariate Analysis: SumAssured



Sum assured feature seem to be slightly right skewed. Hence, the outliers needs to be treated.

The average sum assured is around .

a) Univariate Analysis: LastMonthCalls

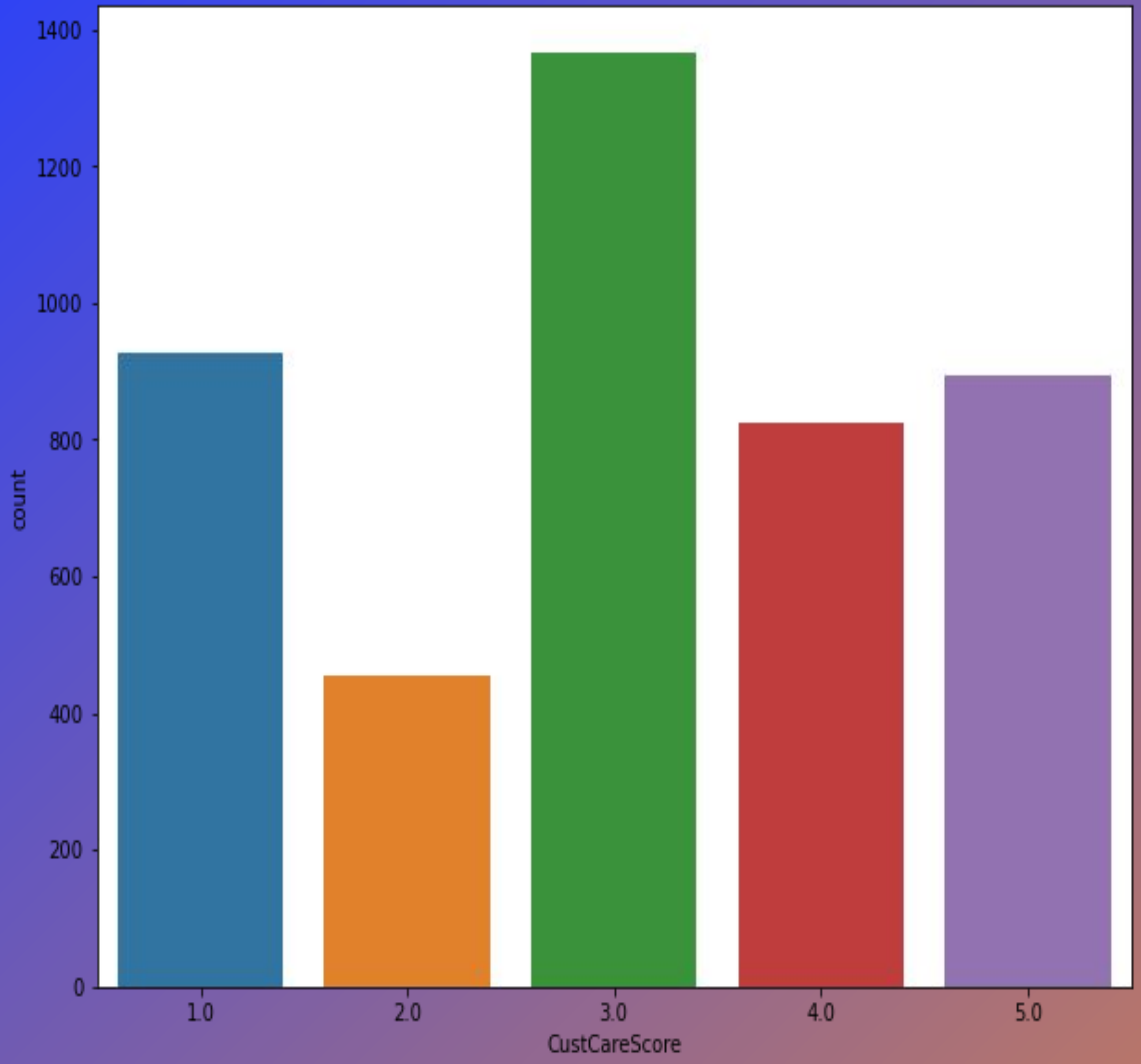


On an average there were 3 calls made in the last month.

There is very less outlier in the data.

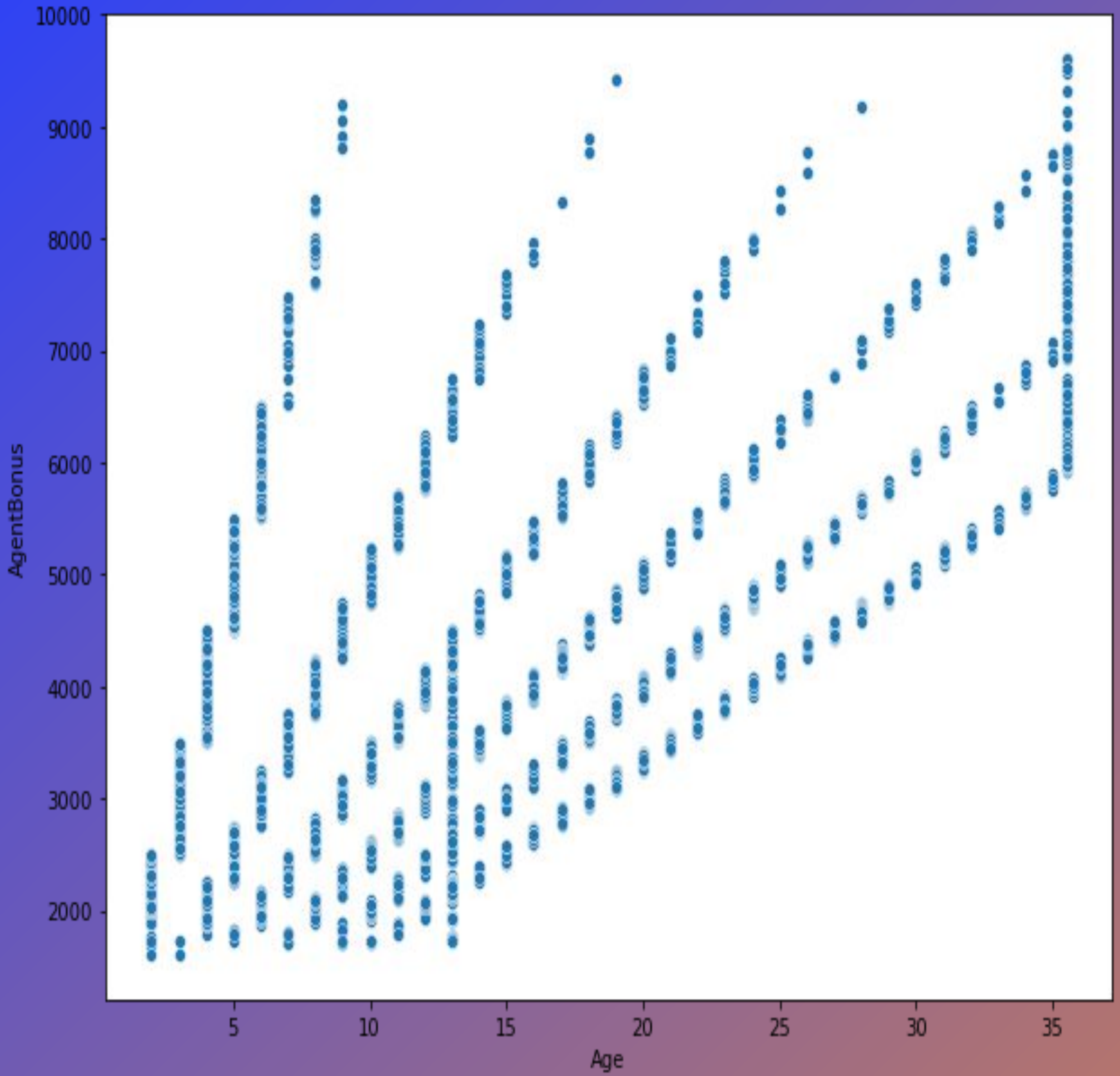
The data seem to be right skewed.

a) Univariate Analysis: CustCareScore



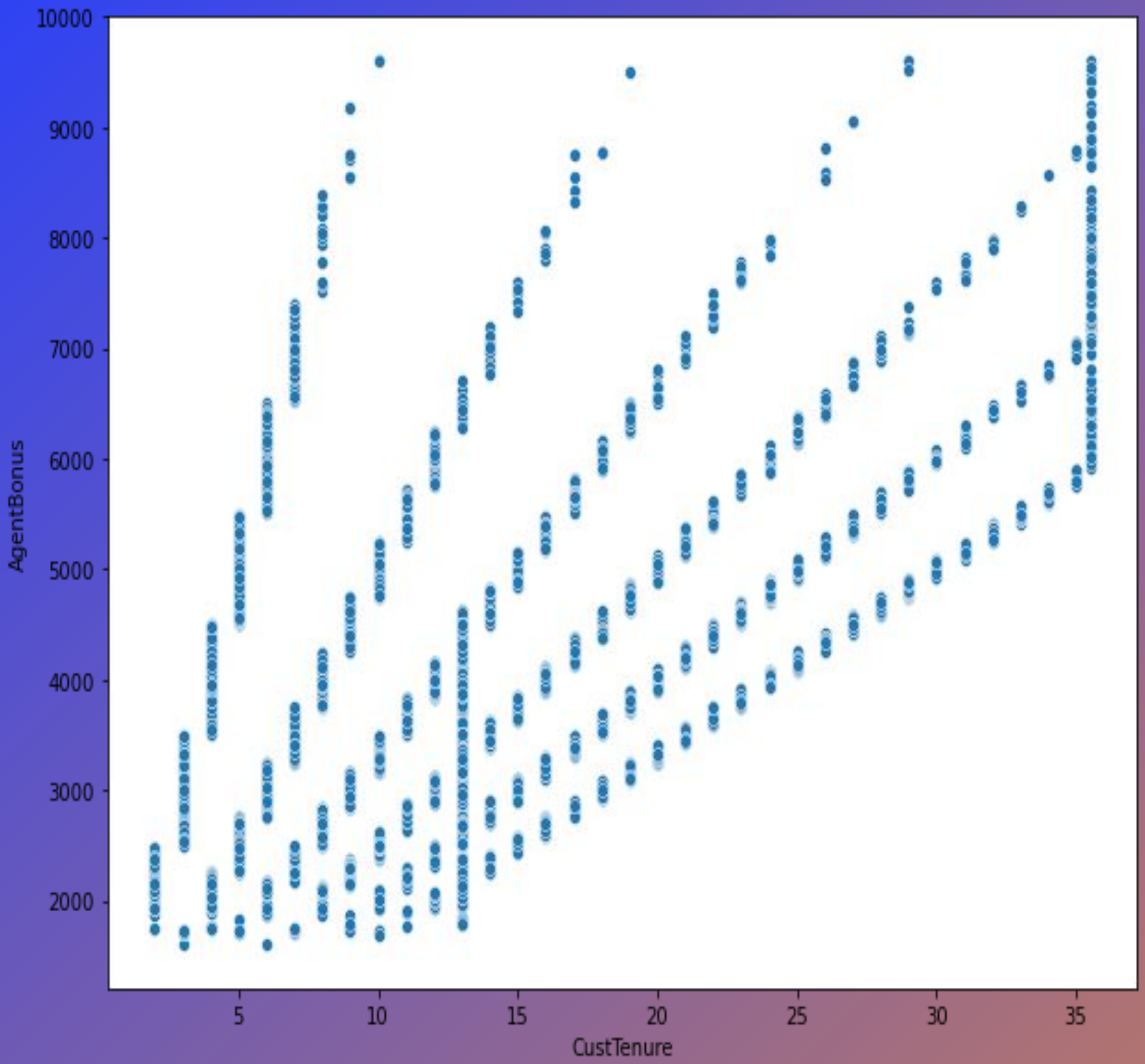
Customer care score is highest for 3 for 1367 agents, followed by 1 which is got by 928 agents or third party. The highest score of 5 is obtained by 893 agents followed by 4 which are 826 in numbers and the least is of rating 2 which is obtained by 454 agents. Company needs to work on or train the agents so that the score increases.

b) Bivariate Analysis: Age vs AgentBonus



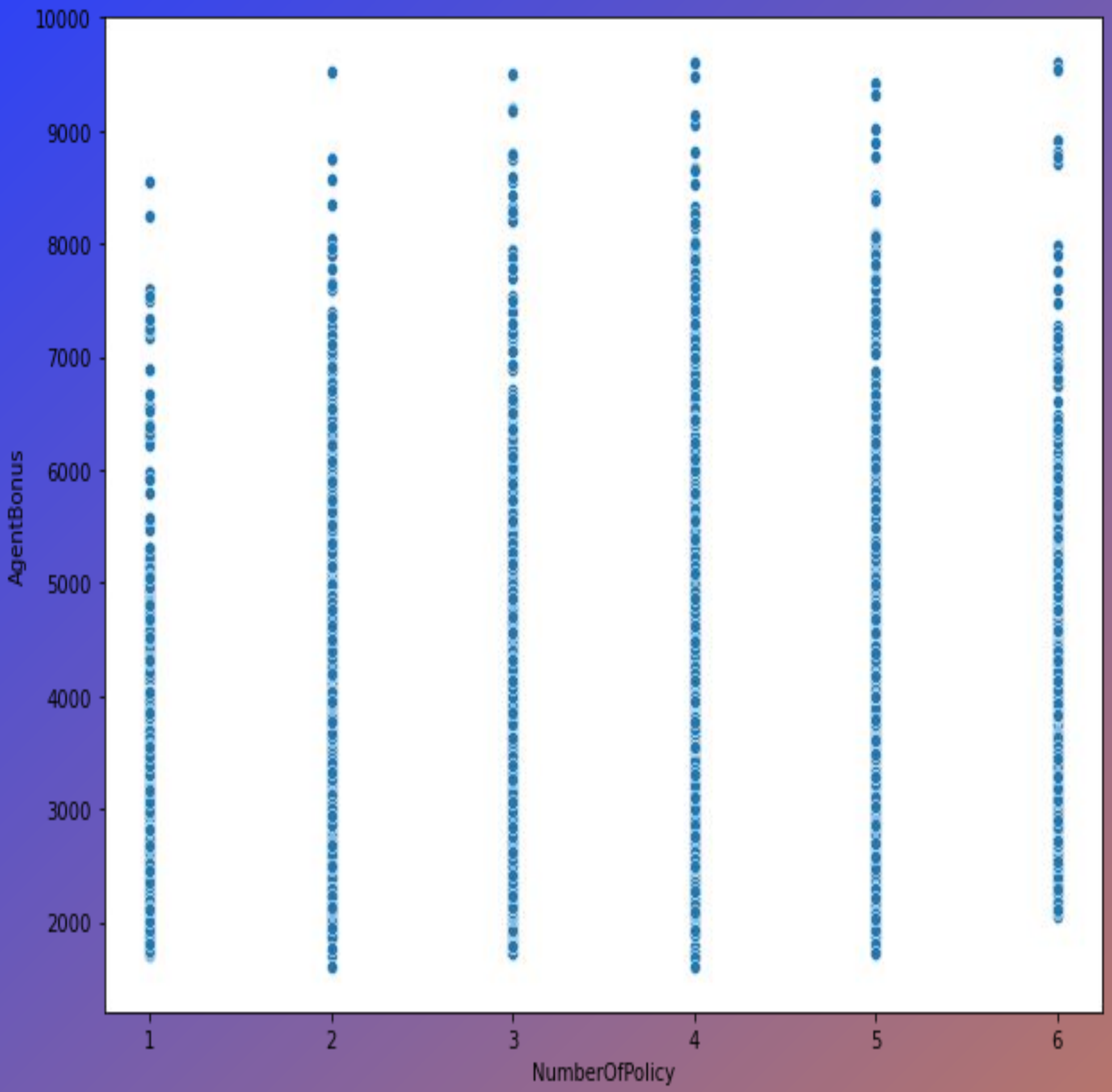
Agent Bonus increases with increase in Age of customers. Bonus increases faster when the customer’s age is lower.

b) Bivariate Analysis: CustTenure vs AgentBonus



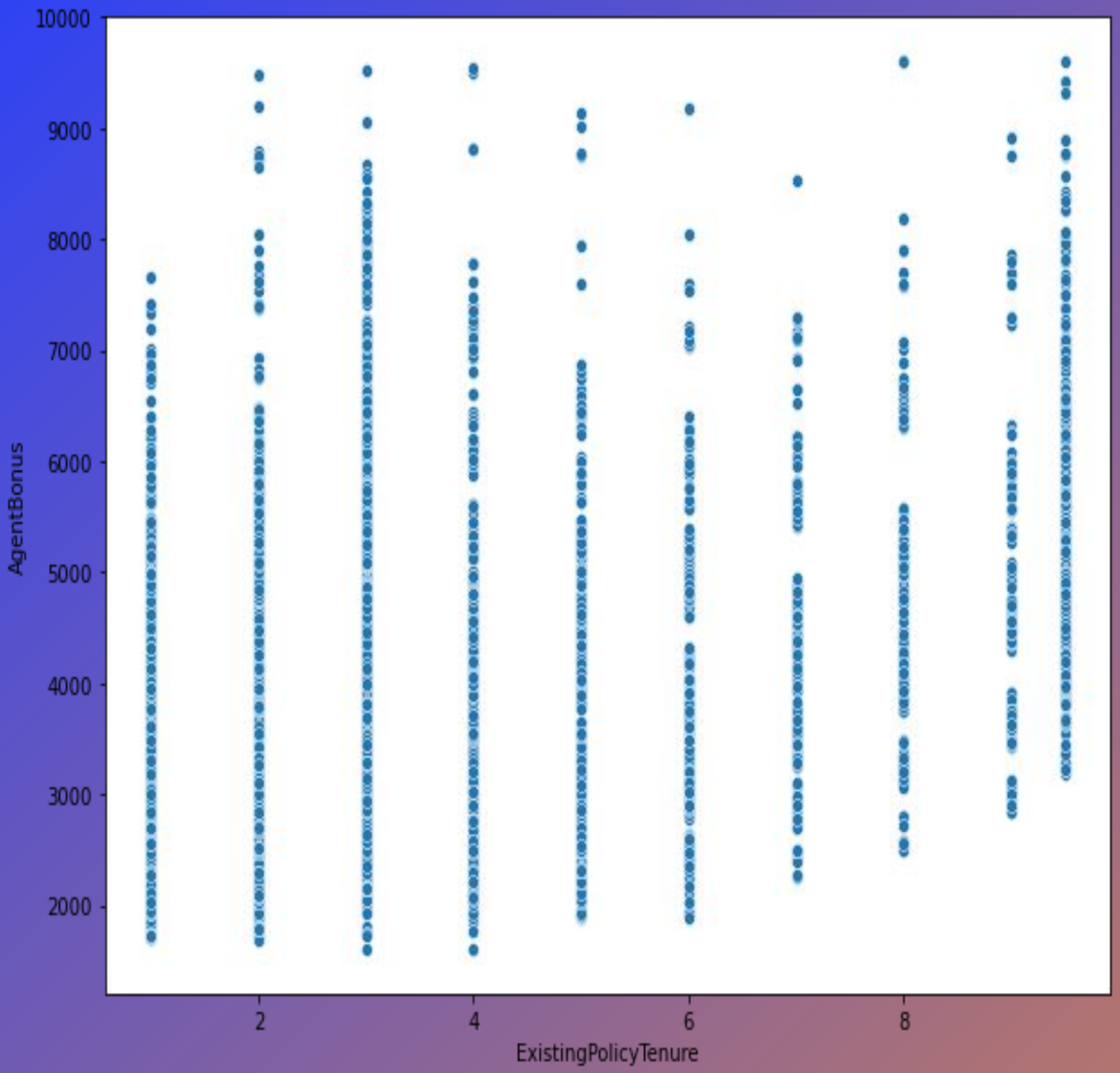
Agent Bonus increases with increase in customer tenure. Bonus increases faster when the customer’s age is lower.

b) Bivariate Analysis: NumberOfPolicy vs AgentBonus



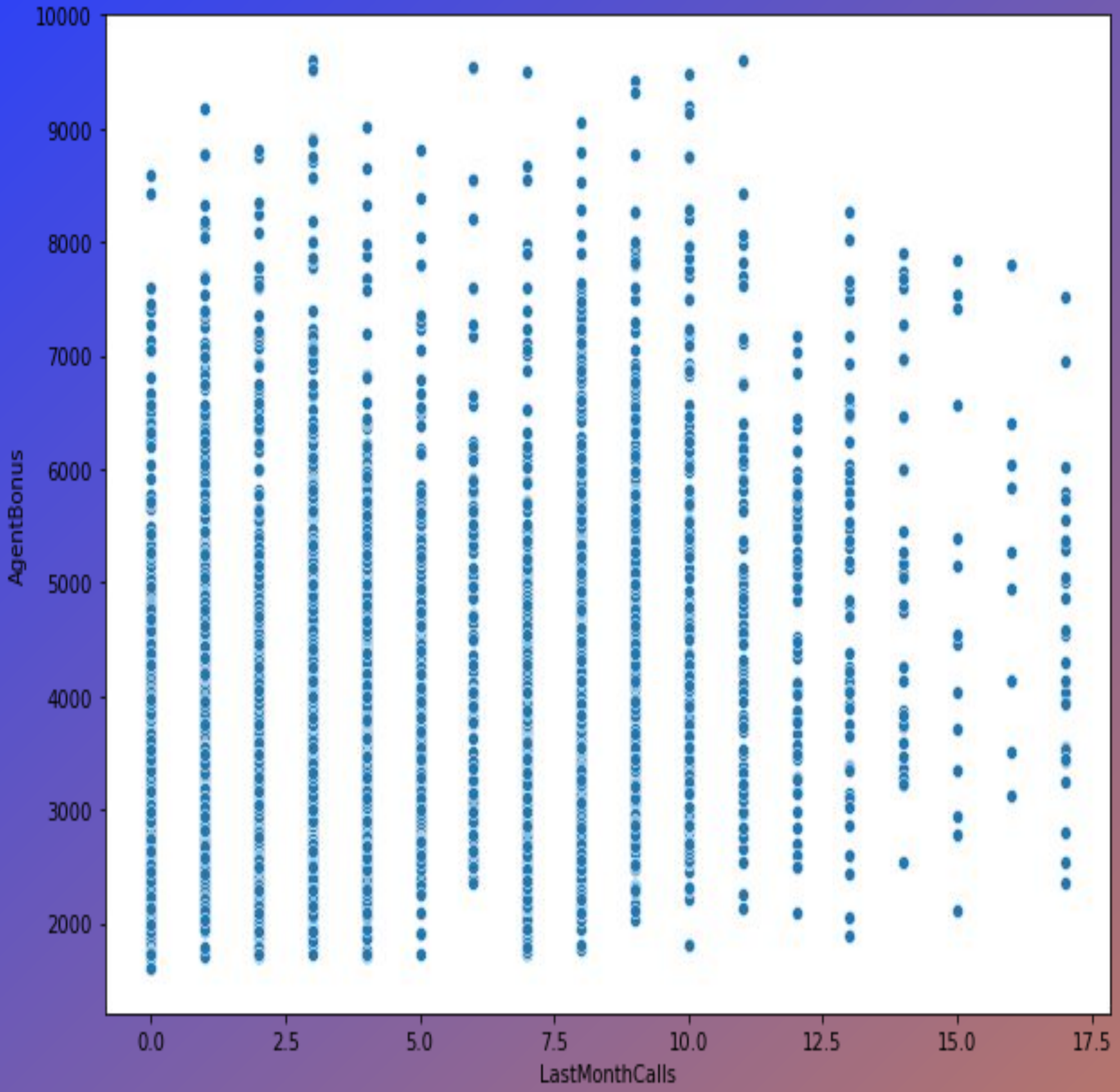
There doesn’t seem to be much of a solid relation here. With more number of policies the agent bonus increases.

b) Bivariate Analysis: ExistingPolicyTenure vs AgentBonus



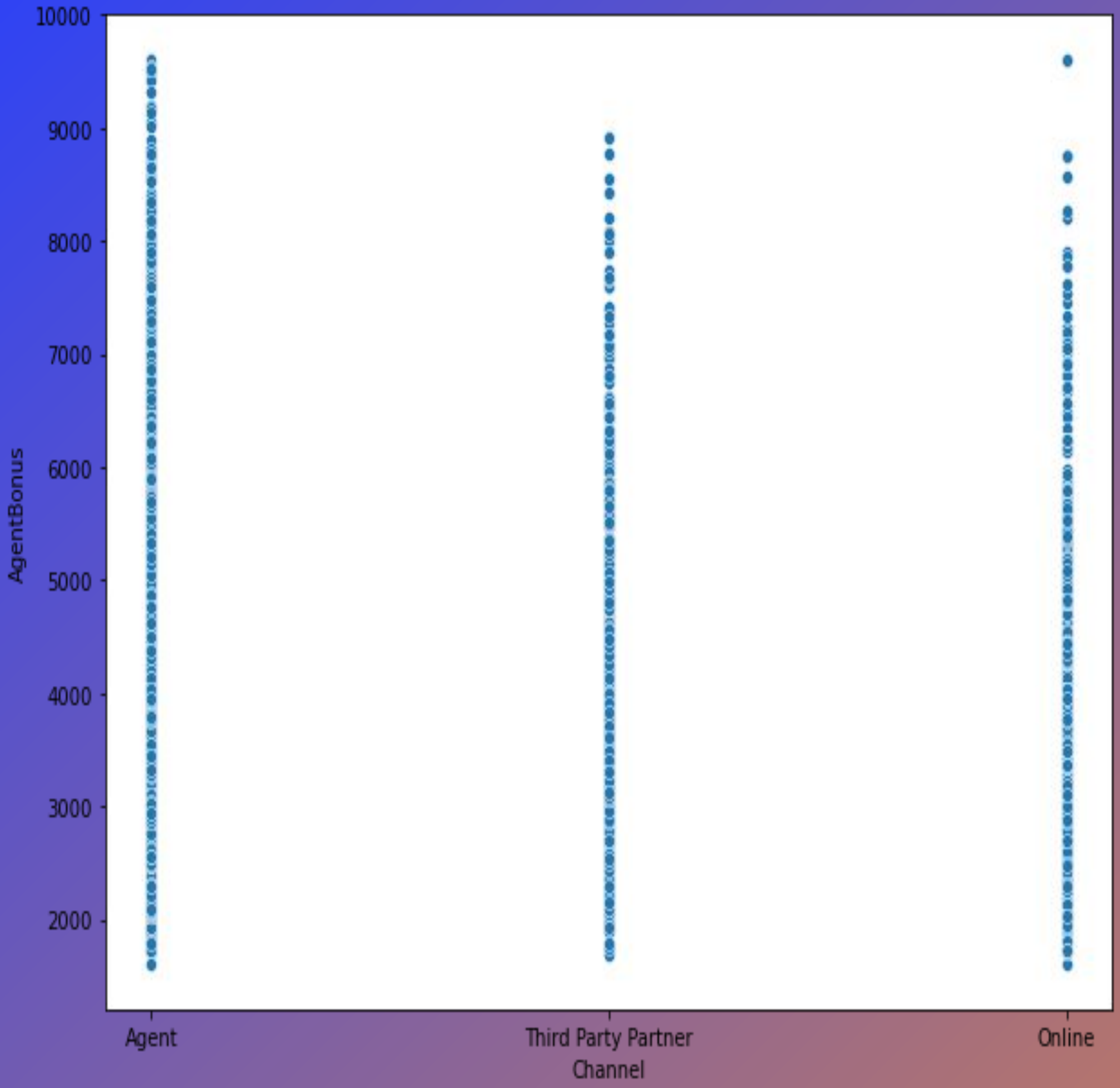
More the existing policy tenure more is the agent bonus but it is not definitive. But the average agent bonus increases with increase in existing policy tenure.

b) Bivariate Analysis: LastMonthCalls vs AgentBonus



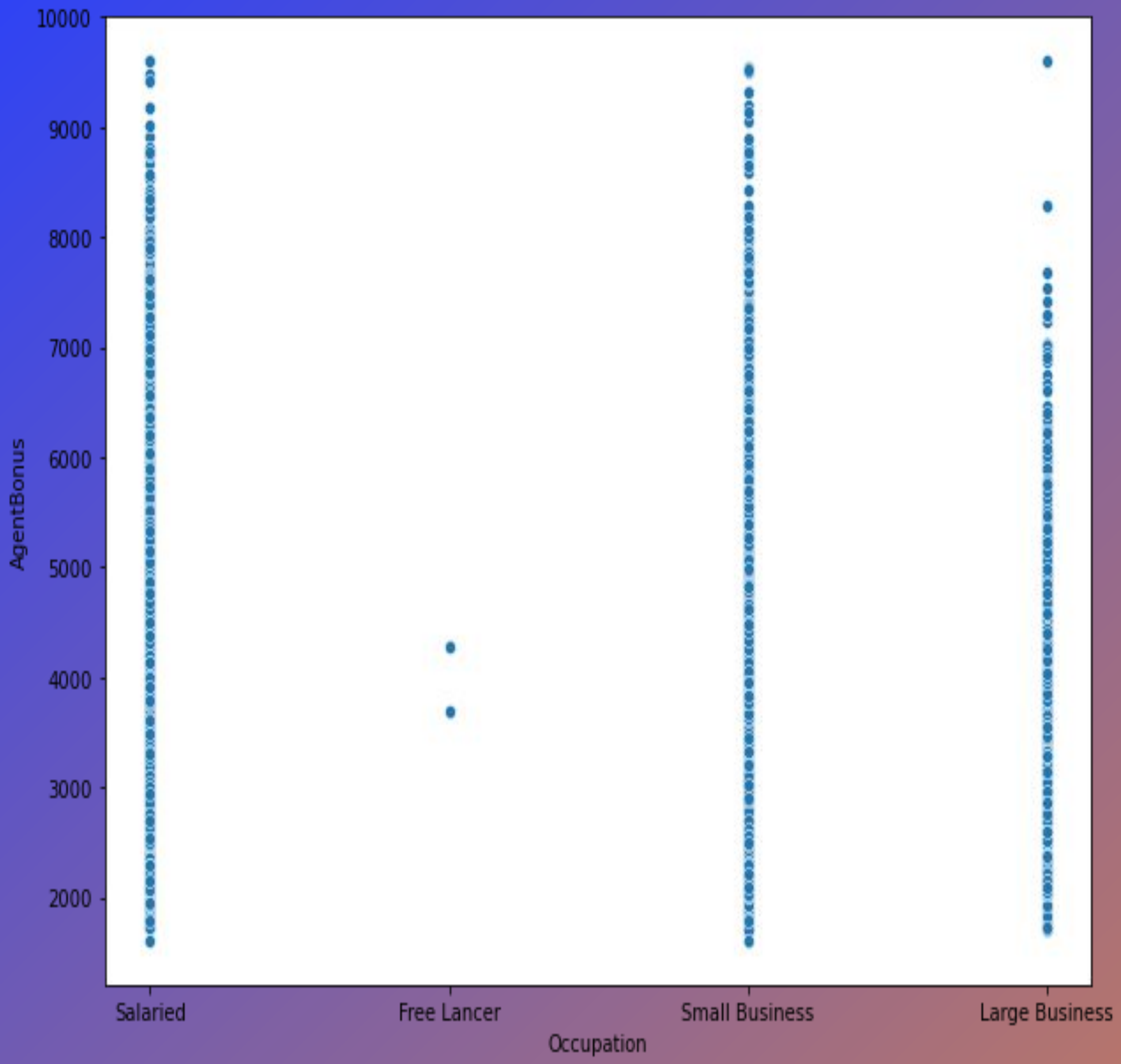
There doesn’t seem to be any pattern or relation in the number of calls and agent bonus.

b) Bivariate Analysis: Channel vs AgentBonus



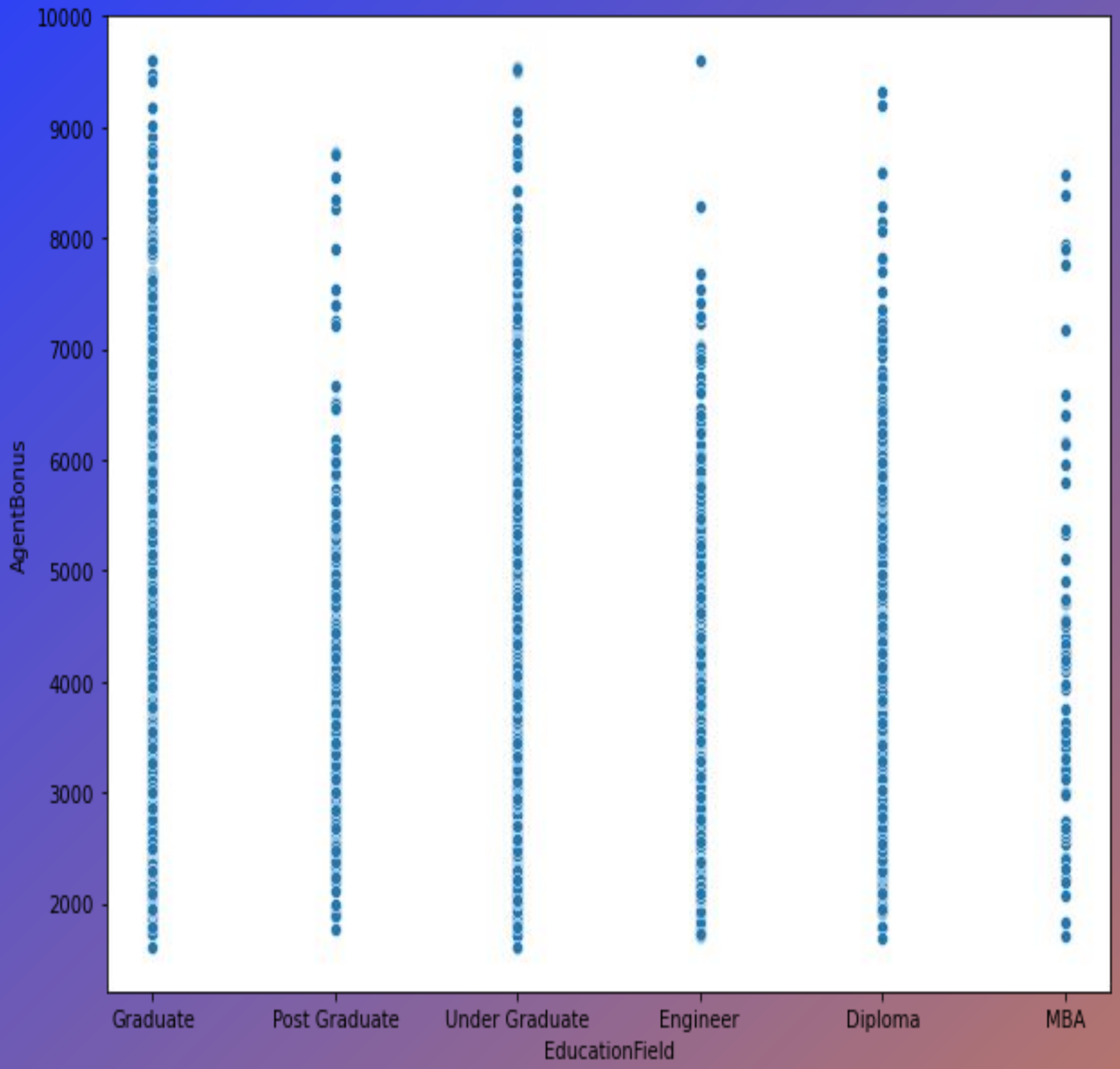
Higher bonus are obtained by agents. Third party partner highest bonus is around 9000 where as for agents it reaches 10000.

b) Bivariate Analysis: Occupation vs AgentBonus



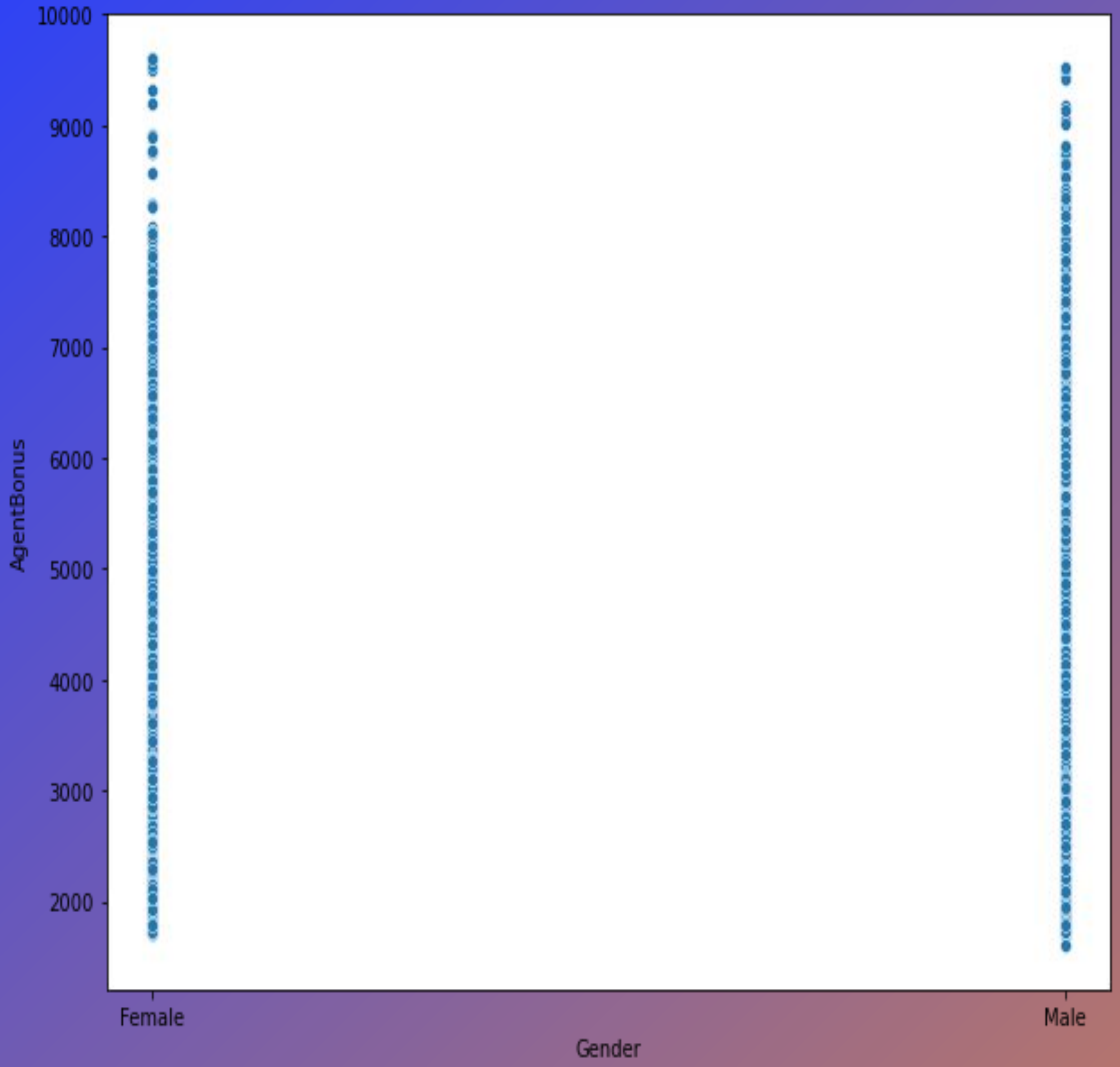
Agents who have salaried as customers tend to earn little higher profit.
Freelancer earns very less and we don't have much info about them.
Agents having small business owner as customer also get similar bonus as the agents.
Agents having large business as customer doesn't earn that high bonus.

b) Bivariate Analysis: EducationField vs AgentBonus



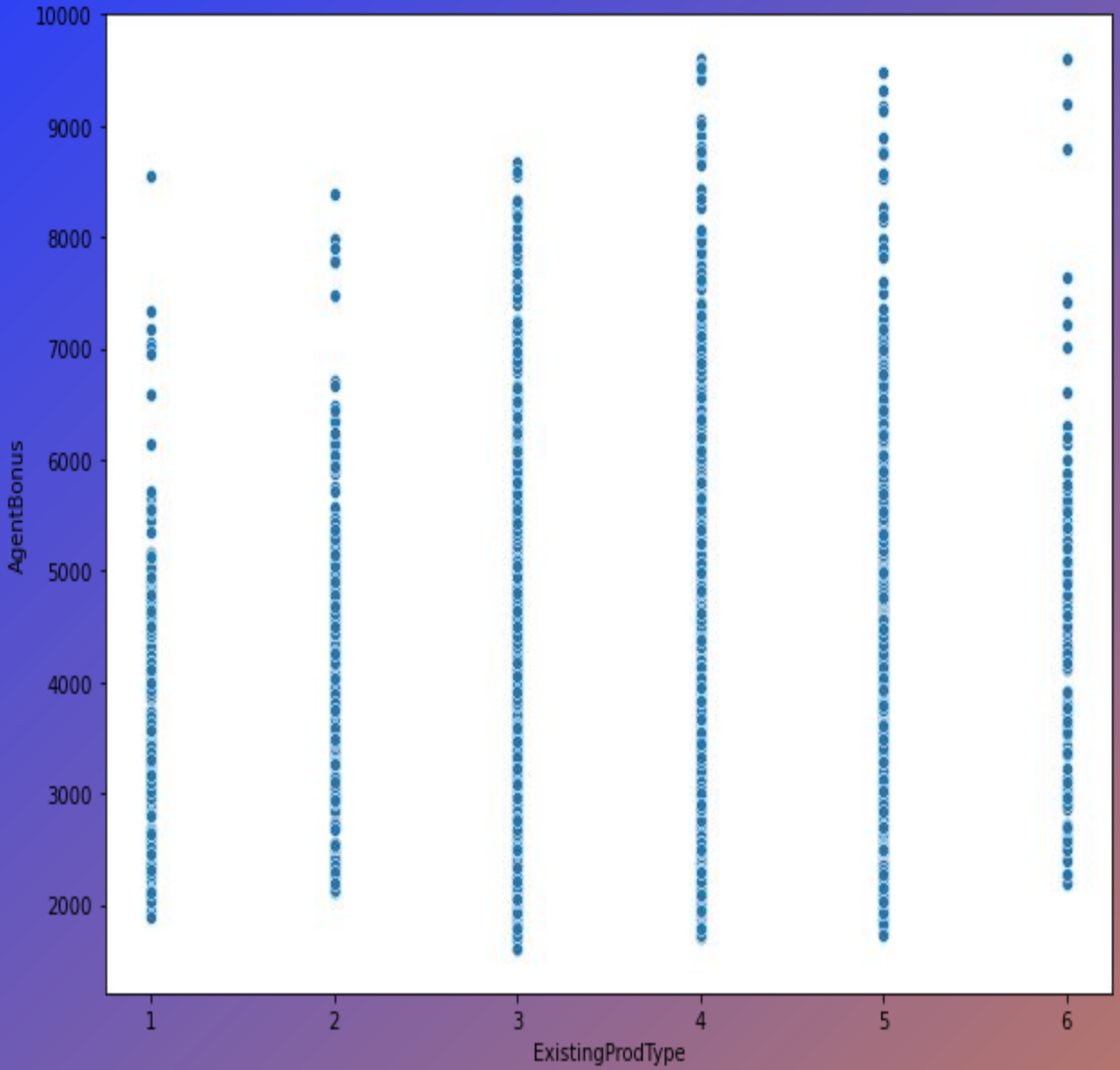
Agents having Graduate and Under graduate as customers get higher bonus as compared to MBA or Post graduate.

b) Bivariate Analysis: Gender vs AgentBonus



Agents having male or female as customers get similar bonus. More number of agents having male as customer tend to get high bonus.

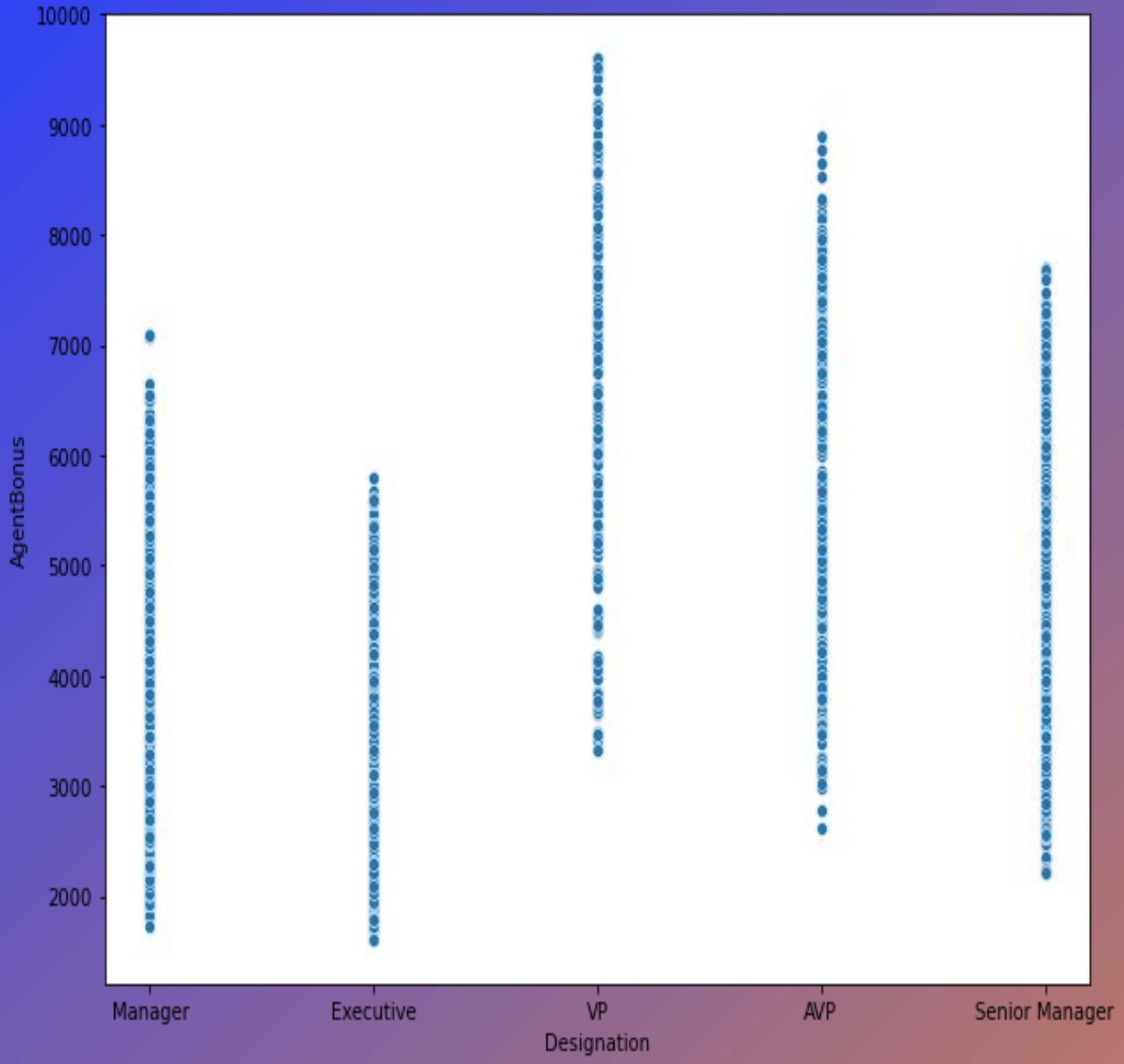
b) Bivariate Analysis: ExistingProdType vs AgentBonus



Agents having customers having product 4 gets more bonus as compared to others. Product 4 is followed by product 5.

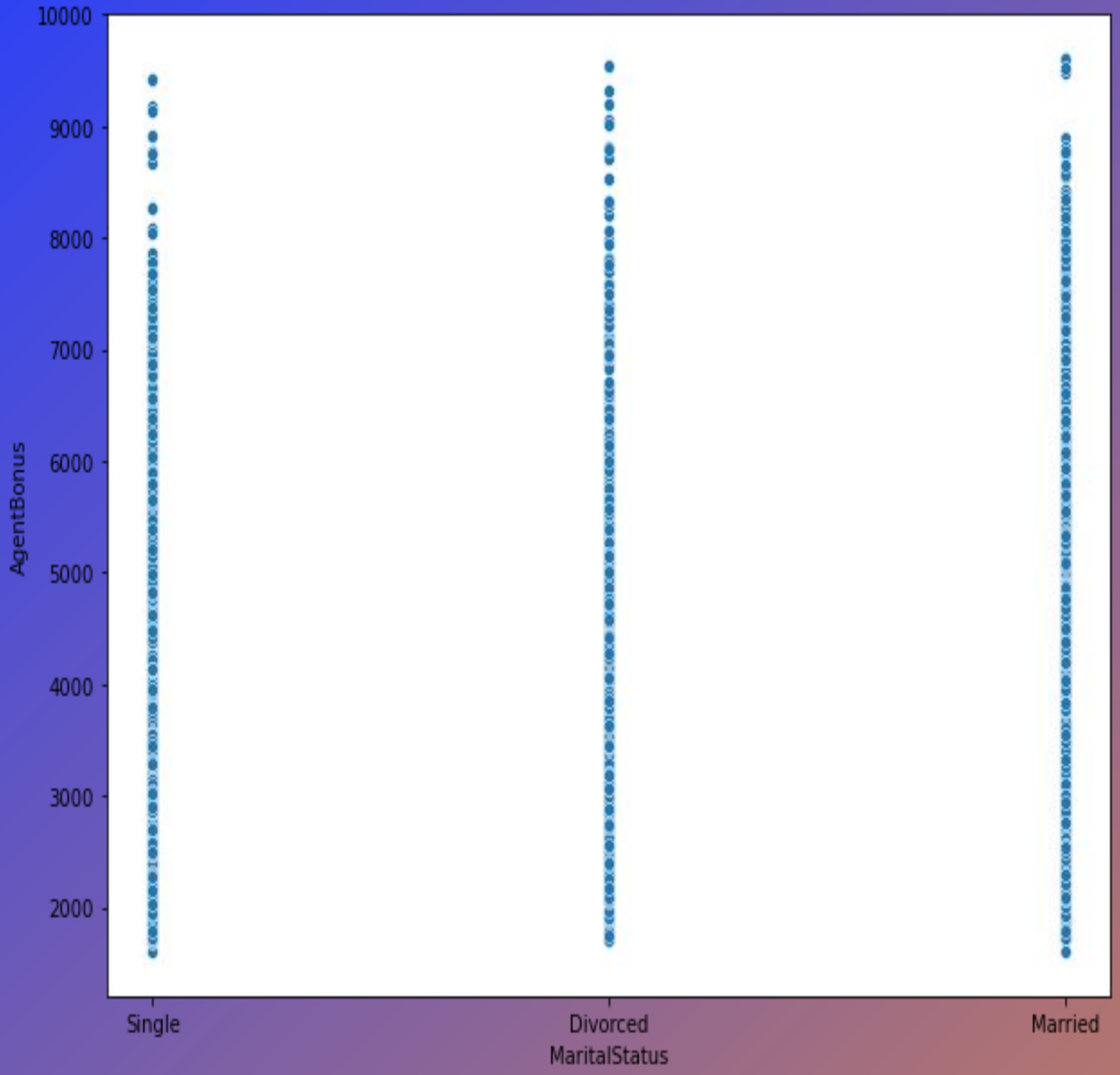
Least is for product 1.

b) Bivariate Analysis: Designation vs AgentBonus



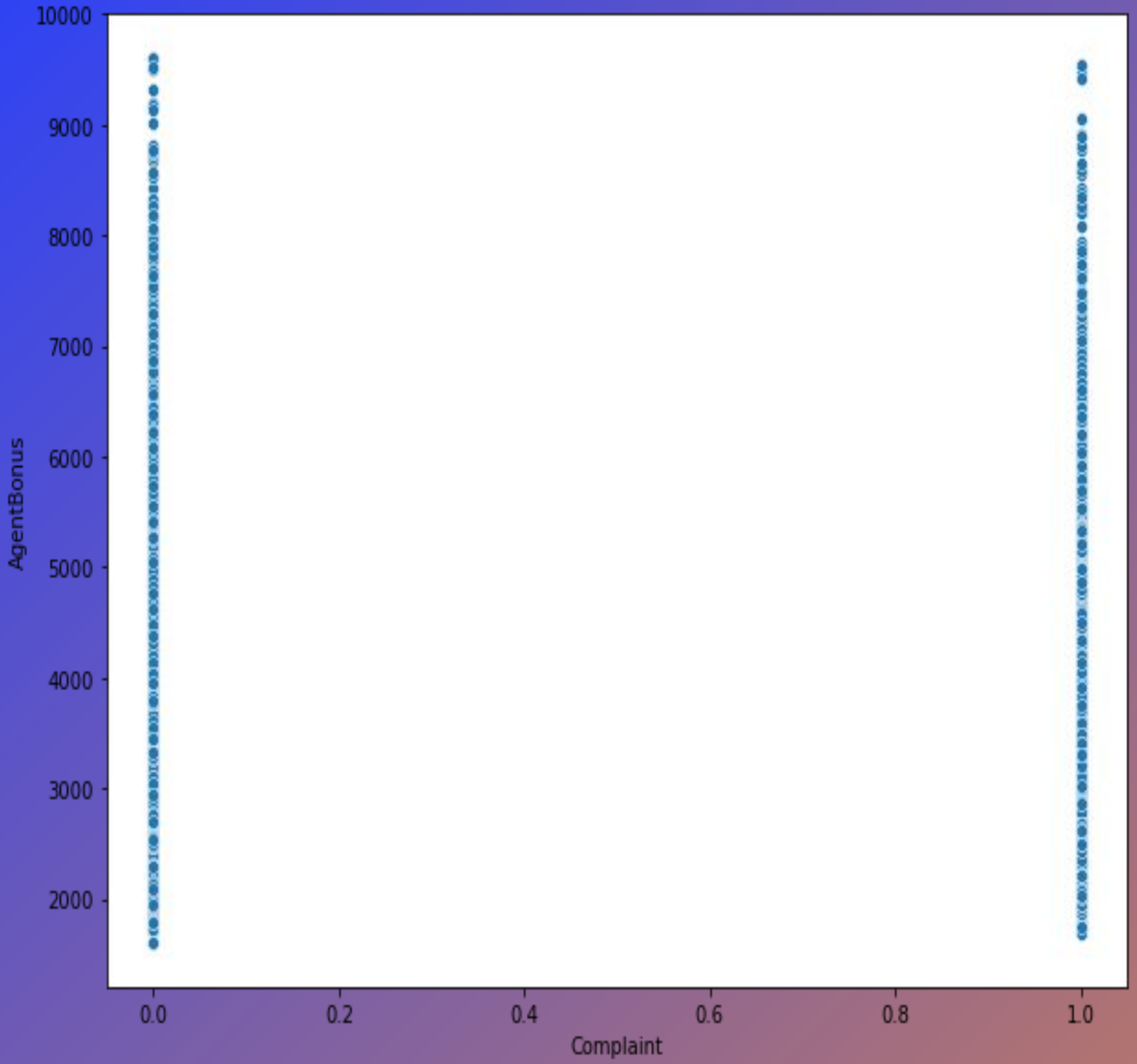
Agents whose customers are VP tends to get high bonus followed by AVP then senior manager. Least is got by agents having executive as their customers.

b) Bivariate Analysis: MaritalStatus vs AgentBonus



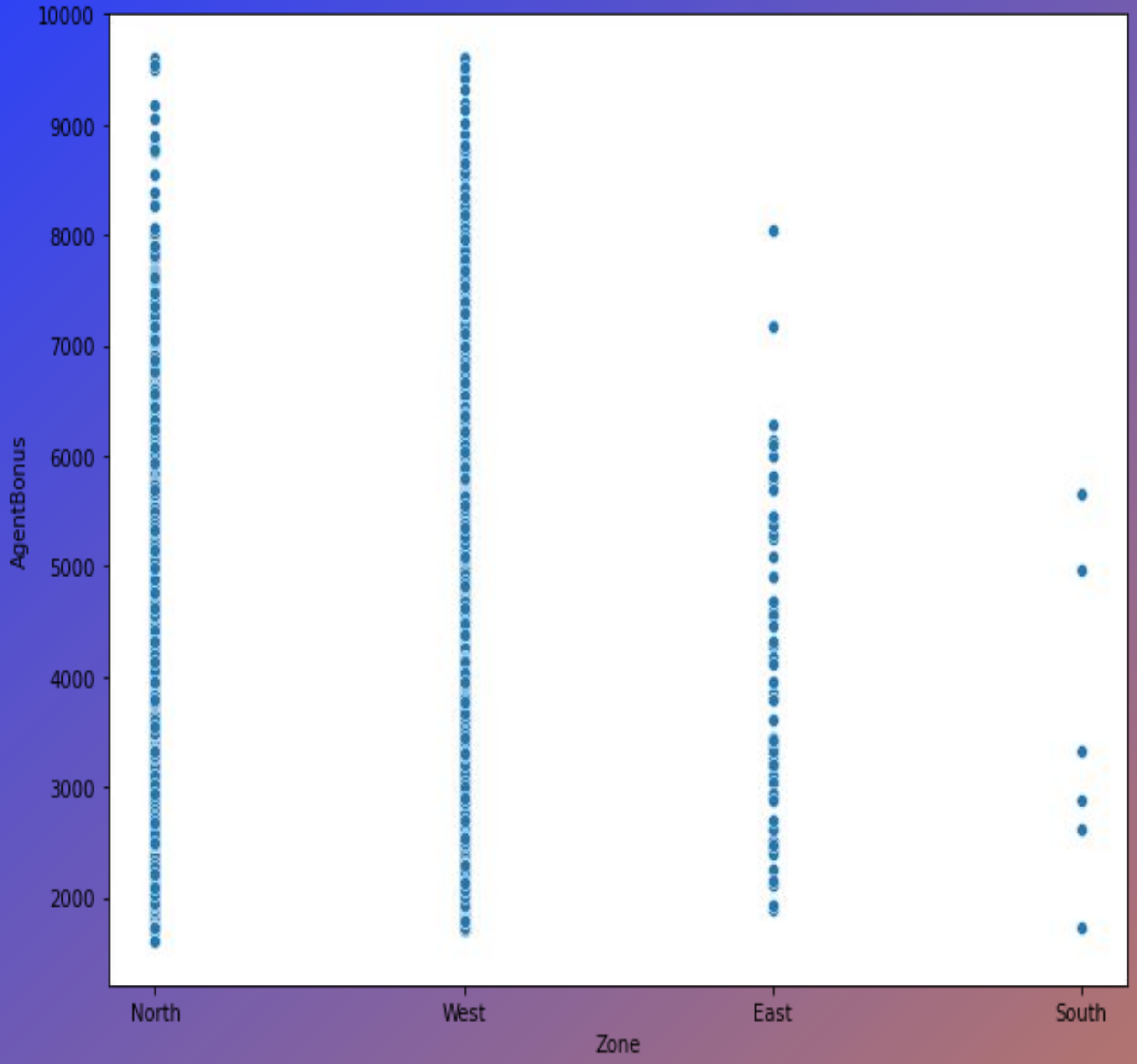
Marital status doesn’t seem to impact much on the agent bonus as bonus are similar for agents having single, married and divorced as customers.

b) Bivariate Analysis: Complaint vs AgentBonus



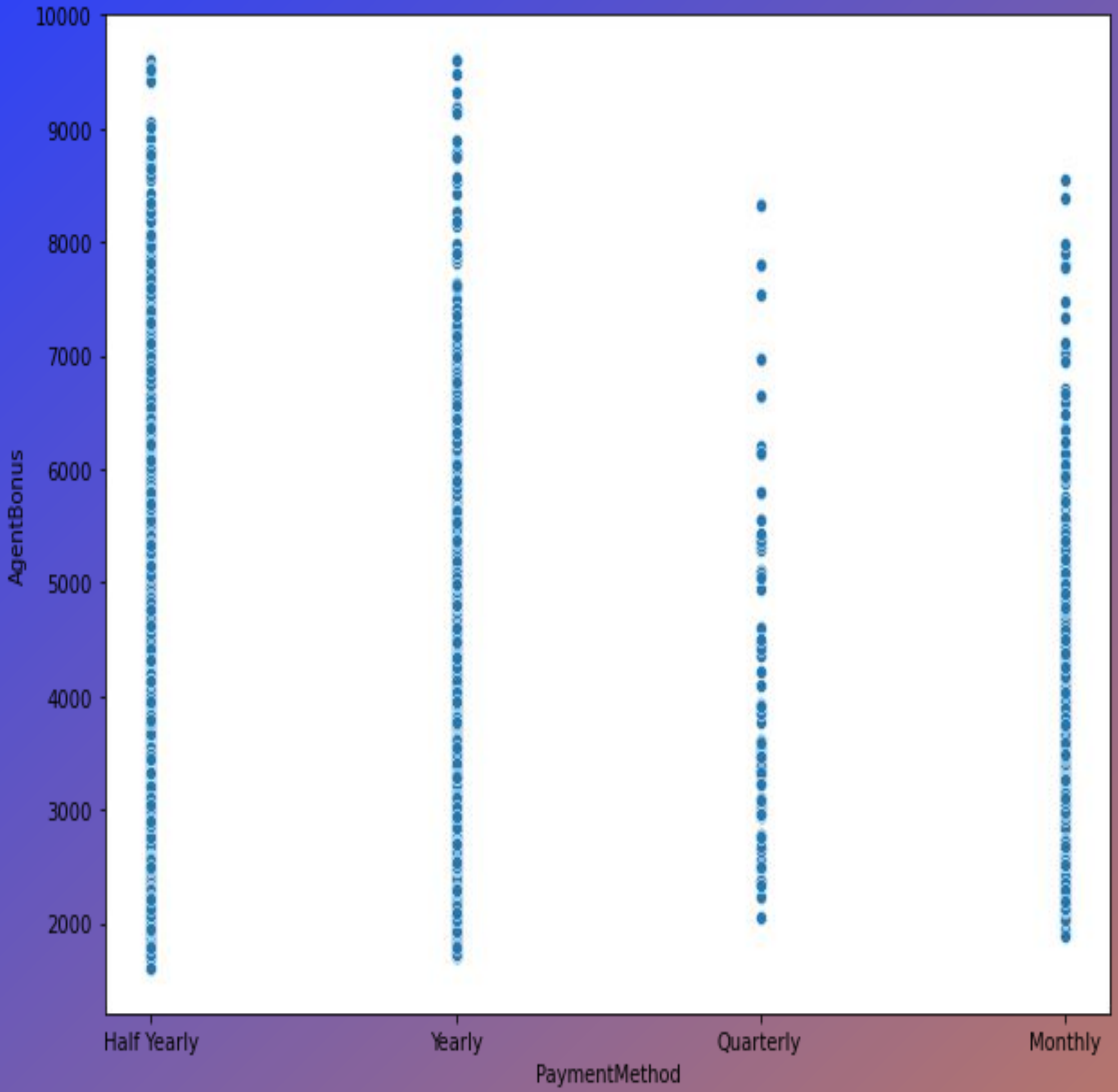
Complaint doesn’t seem to have much effect on agent bonus as bonus are similar for agents having complaint or not. Just that little more number of agents gets higher bonus.

b) Bivariate Analysis: Zone vs AgentBonus



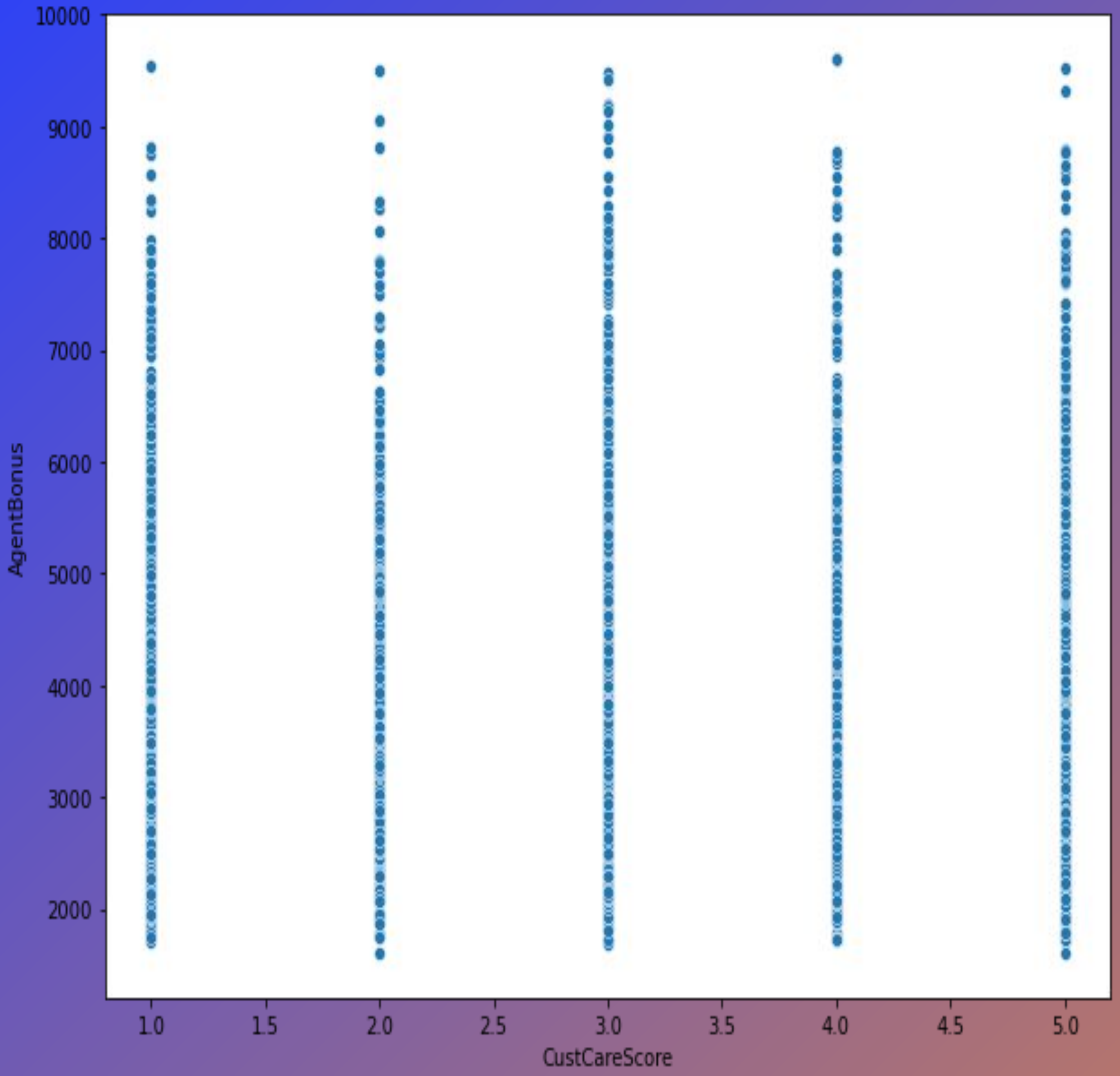
Agents having customers from the west gets higher bonus followed by customers from North.
Not much be predicted about East and South as we do not have much data about them.

b) Bivariate Analysis: PaymentMethod vs AgentBonus



Yearly and half yearly payments gets more bonus to the agents as compared to quarterly or monthly.

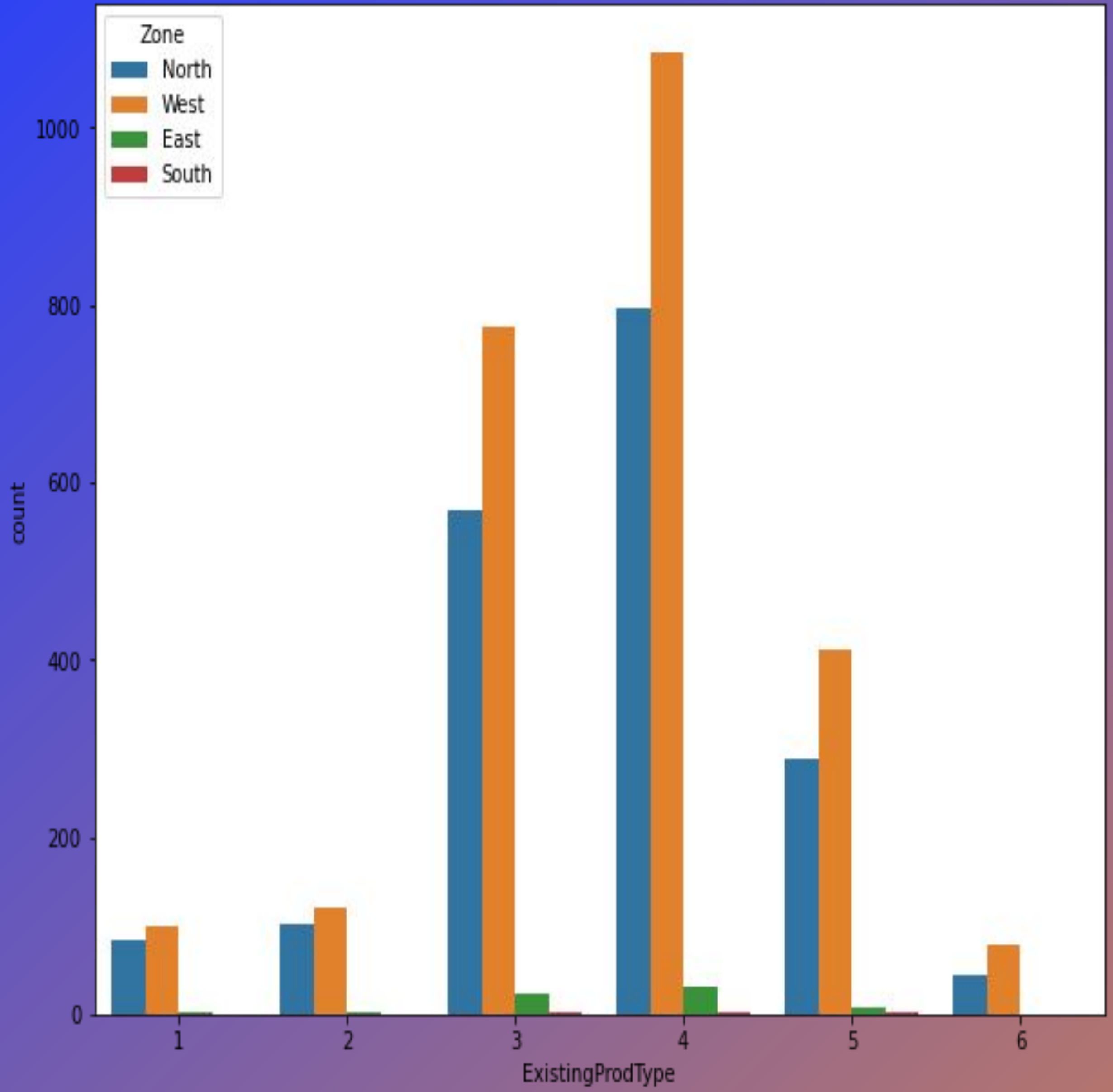
b) Bivariate Analysis: CustCareScore vs AgentBonus



Agents getting 3 score gets the higher bonus followed by 5 score.

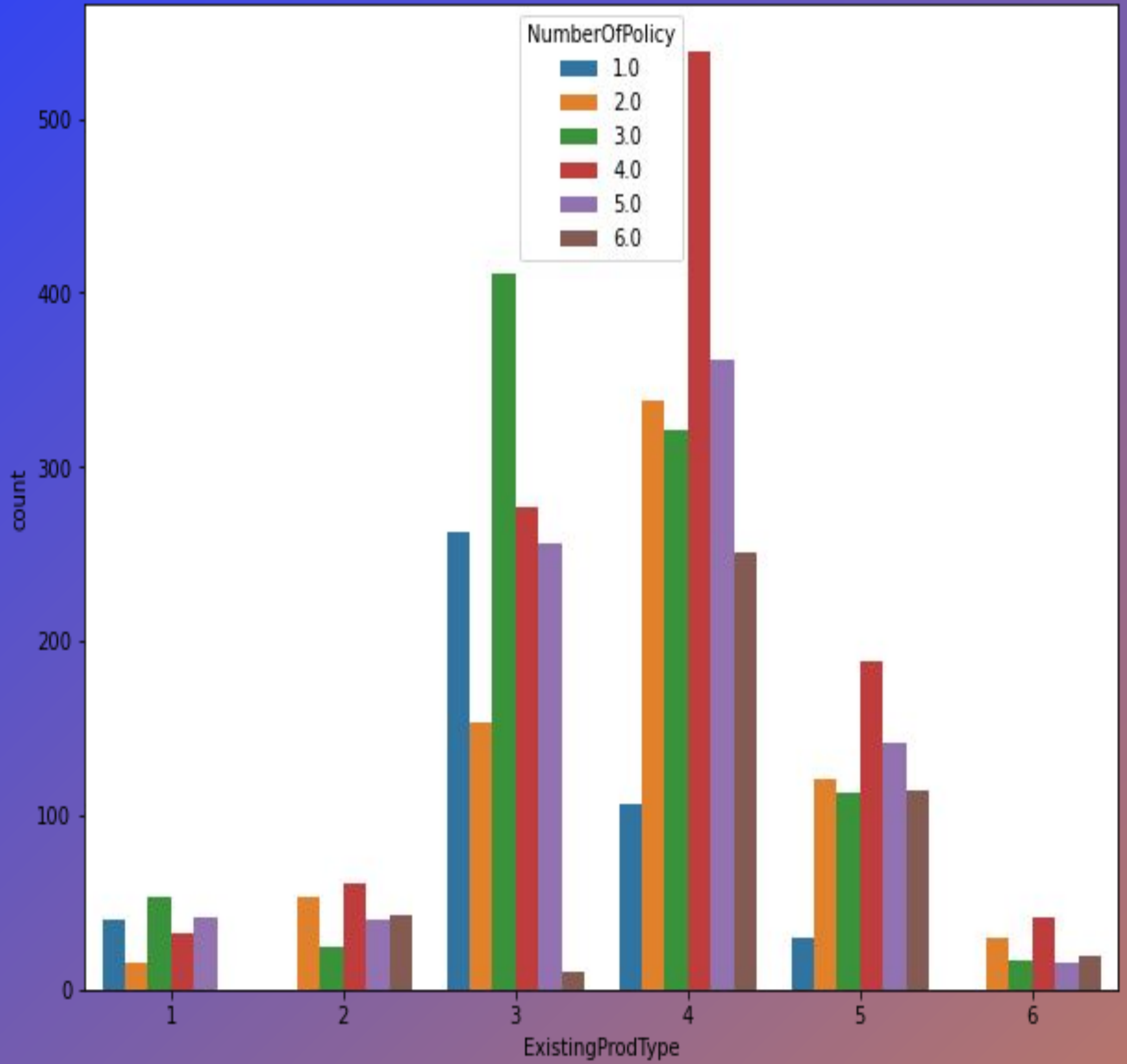
Least seem to be got by score of 2.

b) Bivariate Analysis: ExistingProductType vs Zone



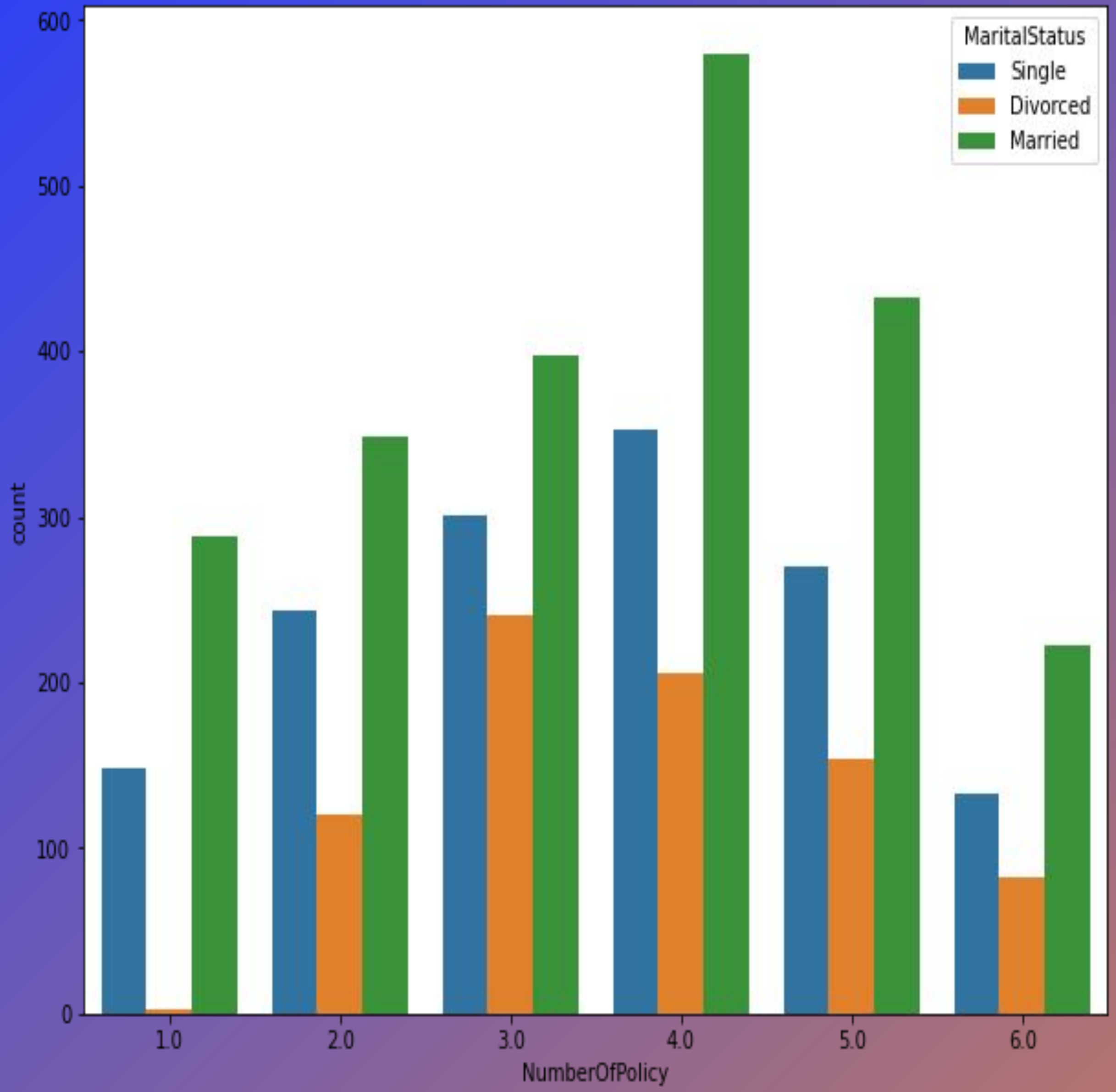
Product 4 is most popular in west and east zone followed by product 3 in west and east zone.

b) Bivariate Analysis: ExistingProductType vs NumberOfPolicy



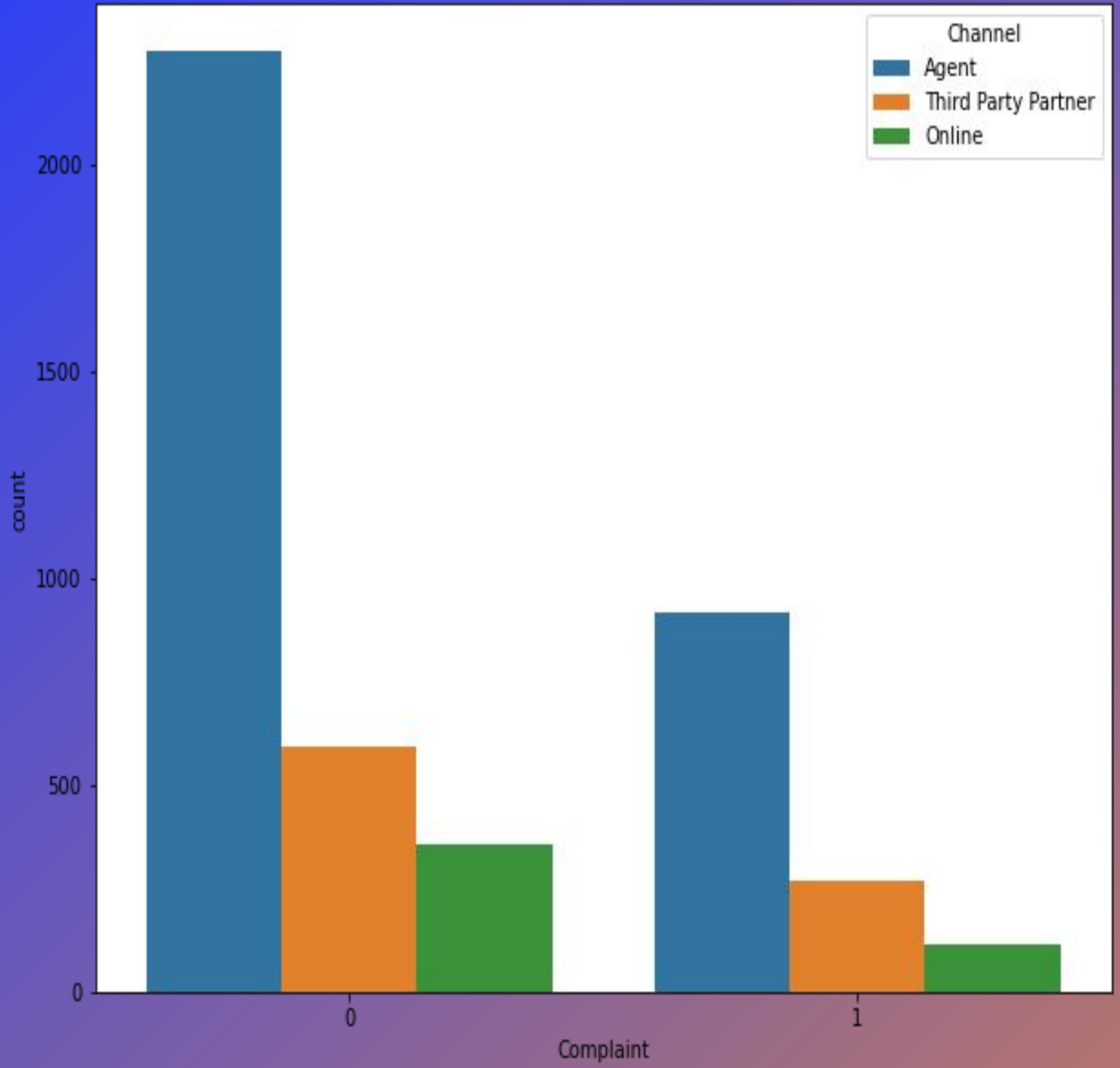
4th product type is having highest number of policies purchased as 4.

b) Bivariate Analysis: MaritalStatus vs NumberOfPolicy



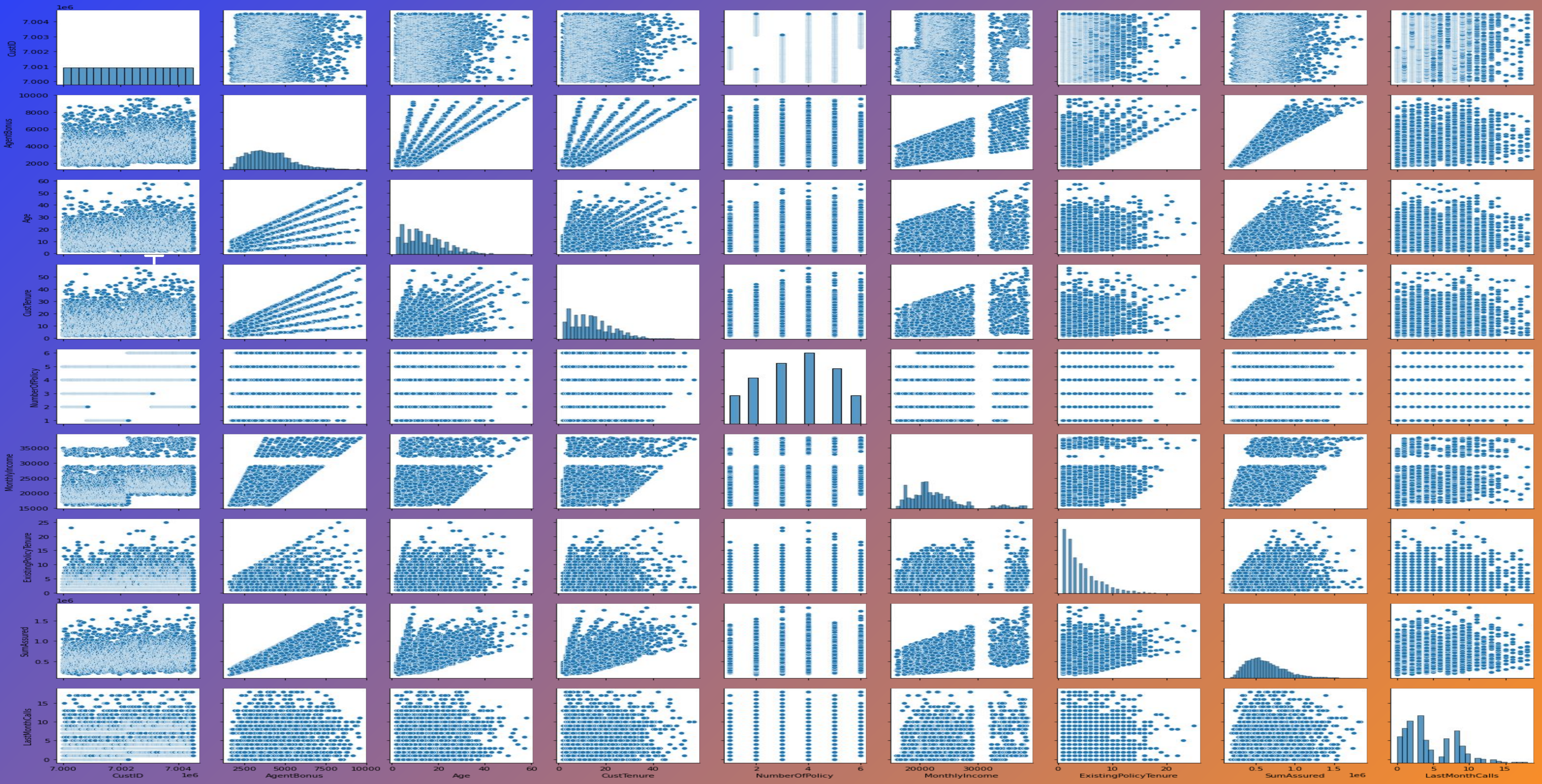
All type of number of policies are mostly purchased by married people followed by singles.
4 number of policies is purchased by married people.

b) Bivariate Analysis: Channel vs Complaint

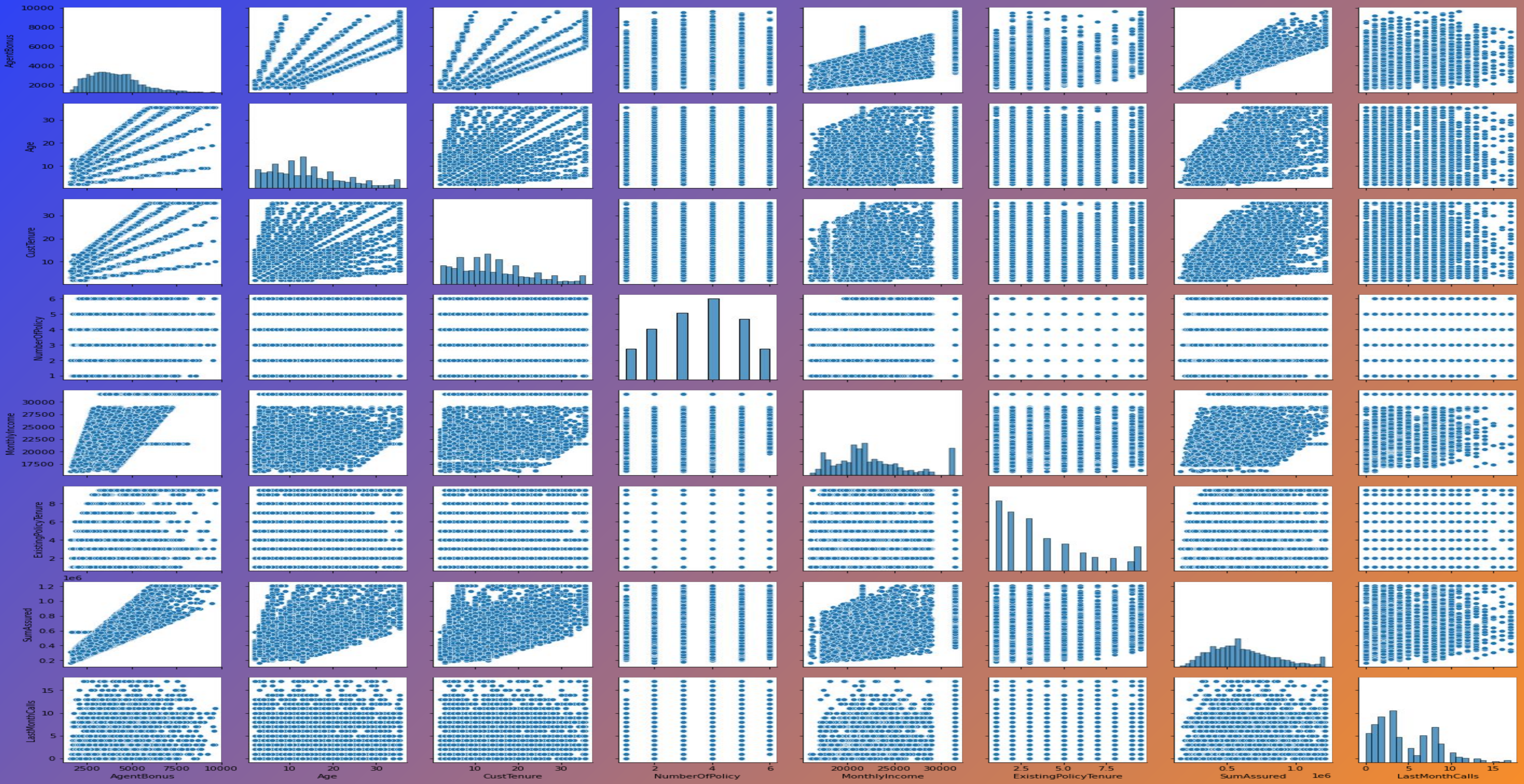


Highest complaint is received by the agents.

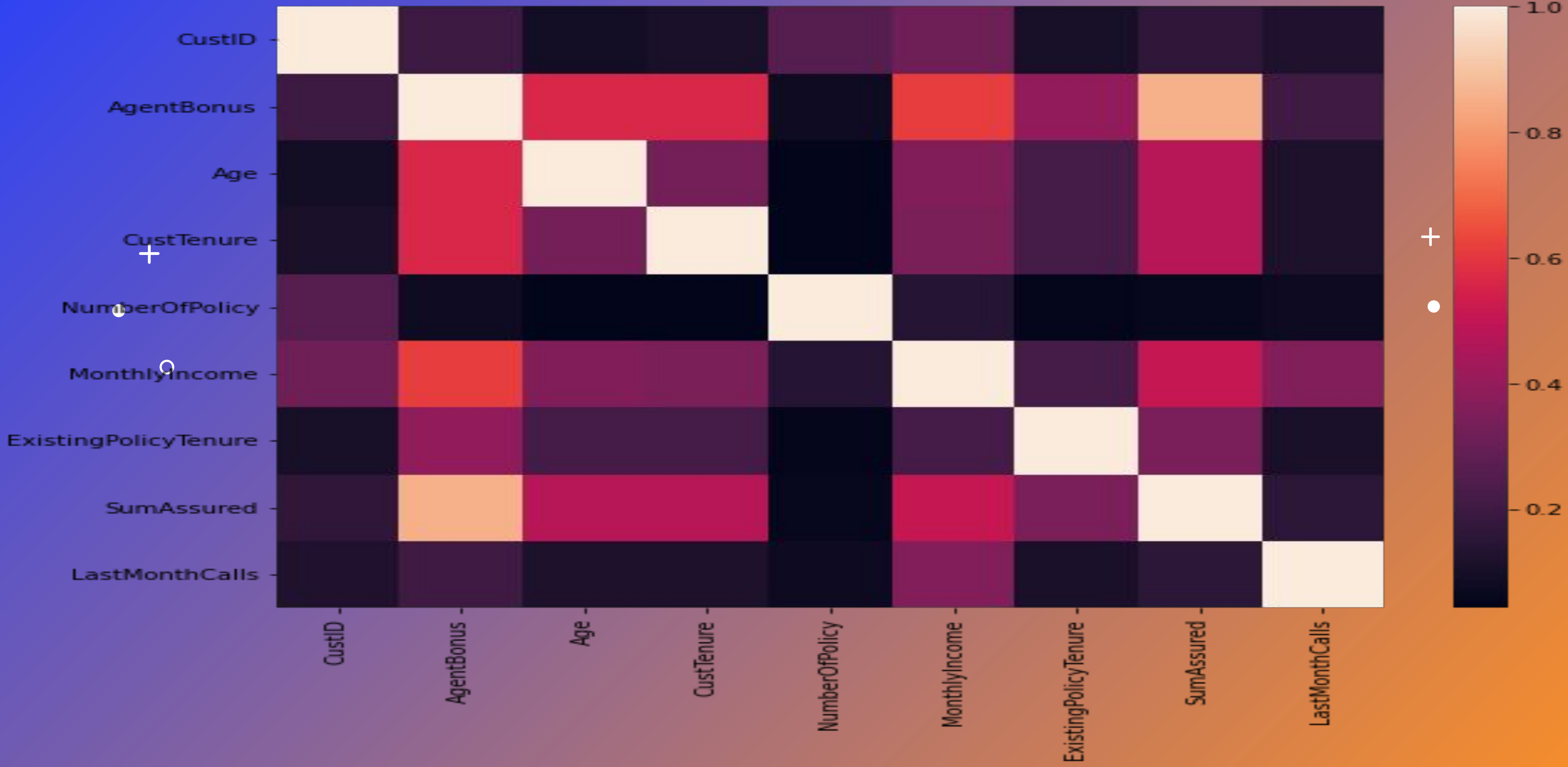
b) Multivariate Analysis: Pairplot of all columns before VIF



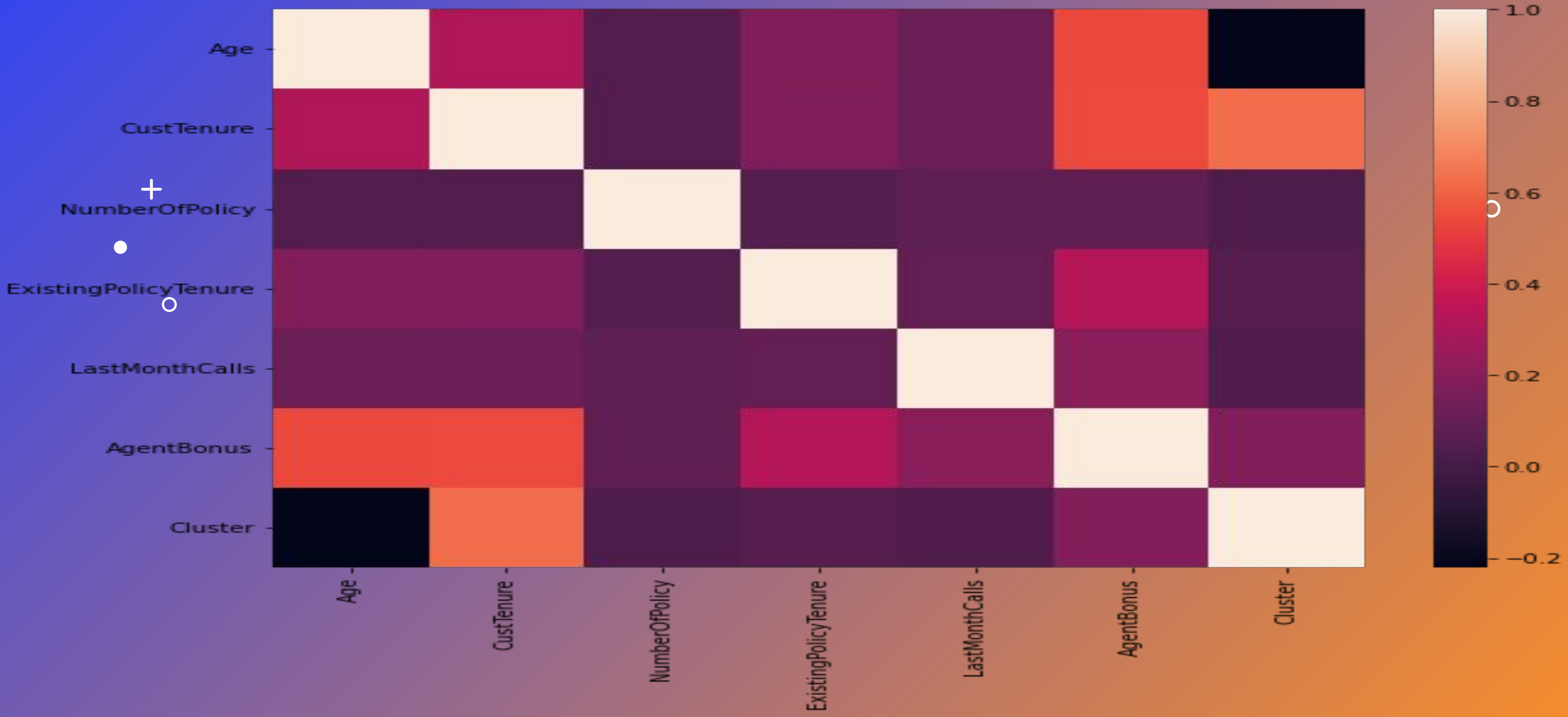
b) Multivariate Analysis: Pairplot of all columns after VIF



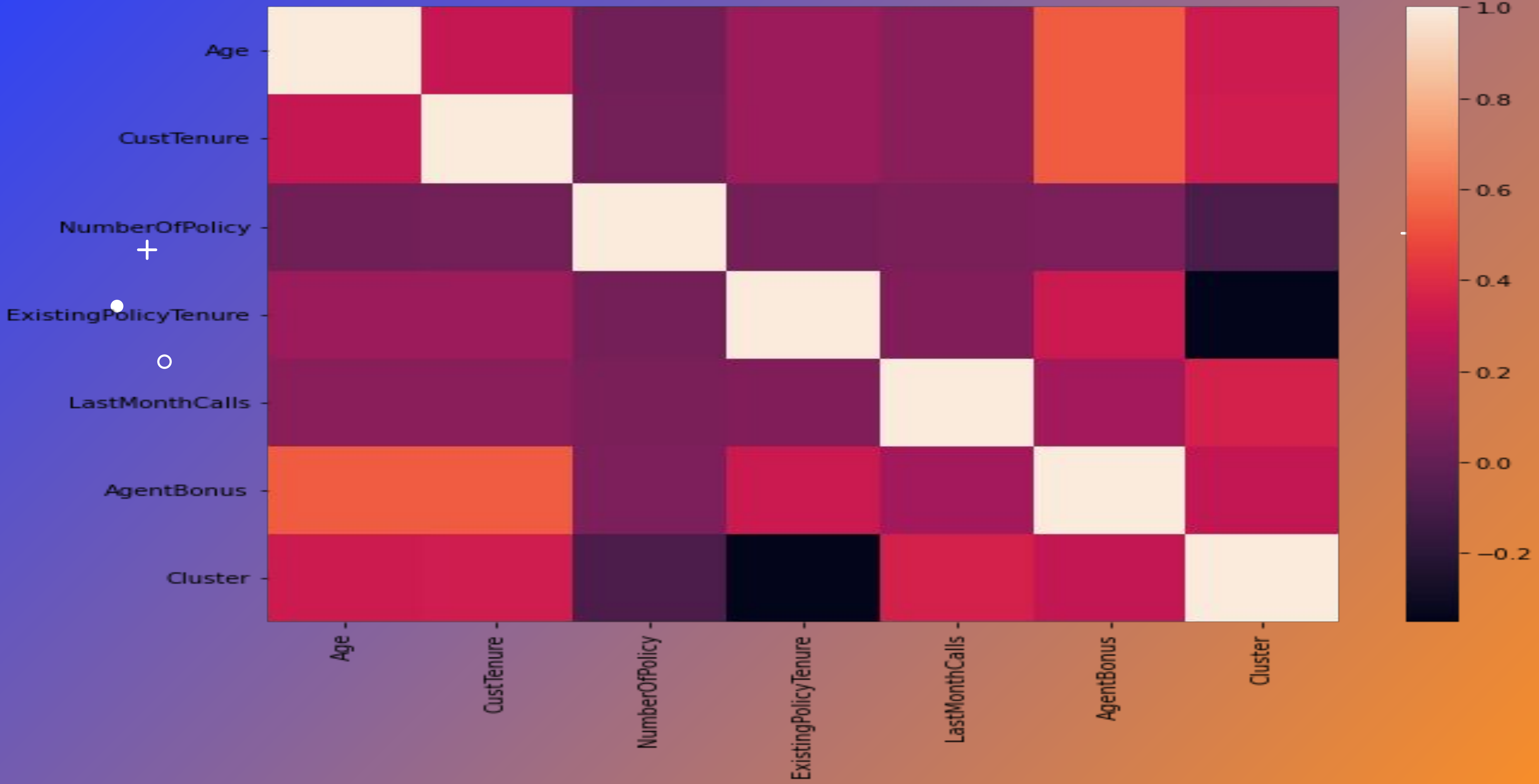
b) Multivariate Analysis: Heatmap of all columns before VIF



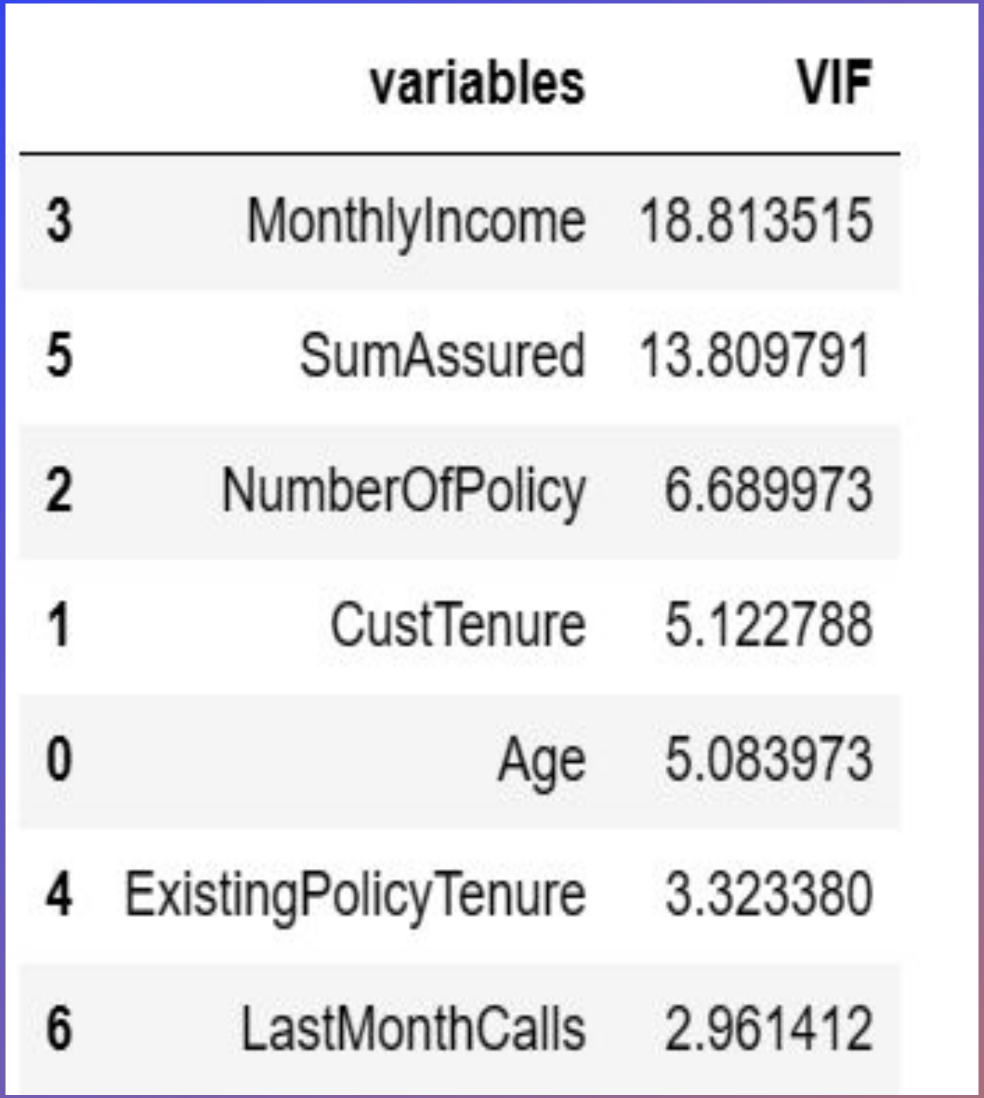
b) Multivariate Analysis: Heatmap of all columns after VIF before scaling



b) Multivariate Analysis: Heatmap of all columns after VIF



c) Removal of unwanted variables



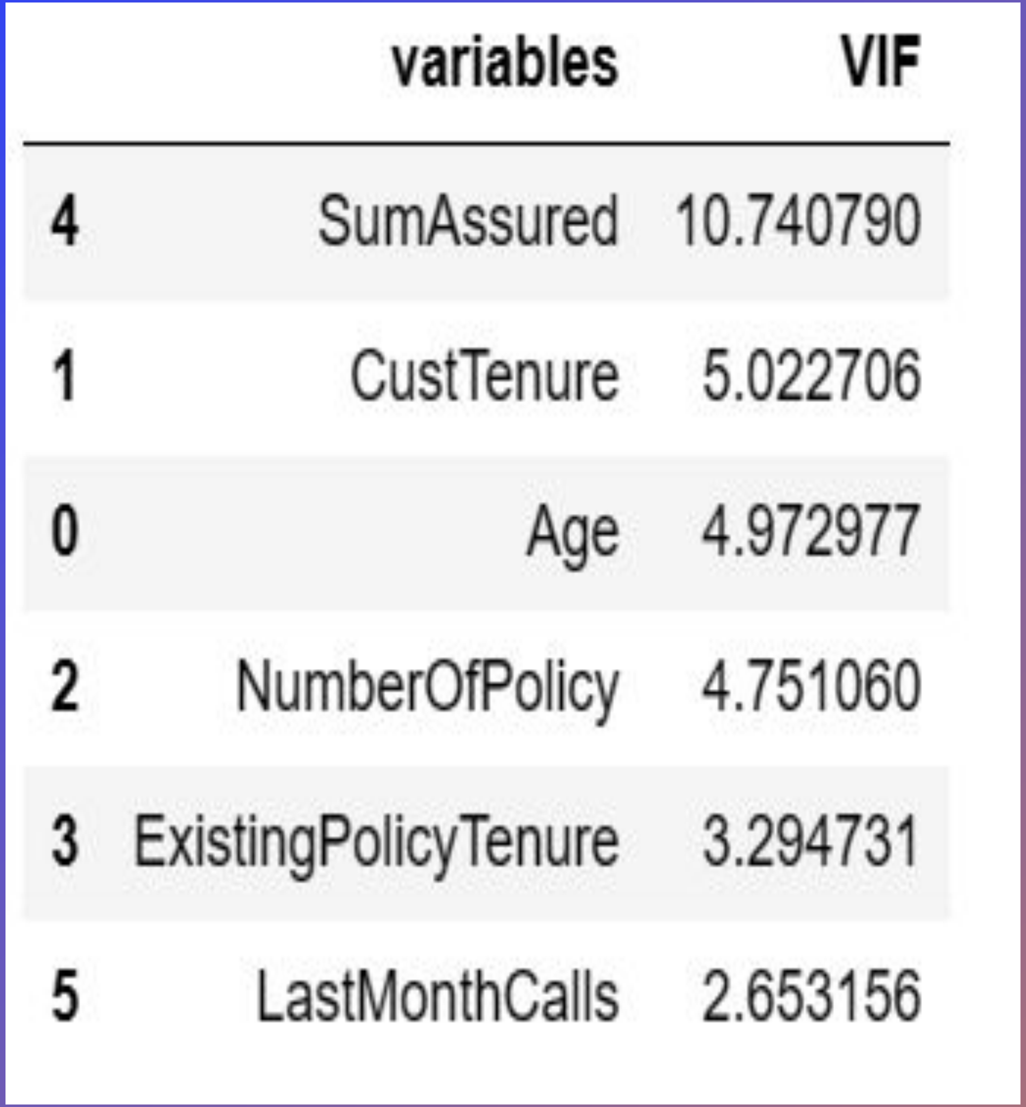
	variables	VIF
3	MonthlyIncome	18.813515
5	SumAssured	13.809791
2	NumberOfPolicy	6.689973
1	CustTenure	5.122788
0	Age	5.083973
4	ExistingPolicyTenure	3.323380
6	LastMonthCalls	2.961412

To remove unwanted variables, I have used VIF which is used to measure multicollinearity between the variables.

I have one by one removed variables having 5 or more as their VIF score.

In the first iteration of VIF, I found that the VIF score of MonthlyIncome is approx 19, hence I removed it.

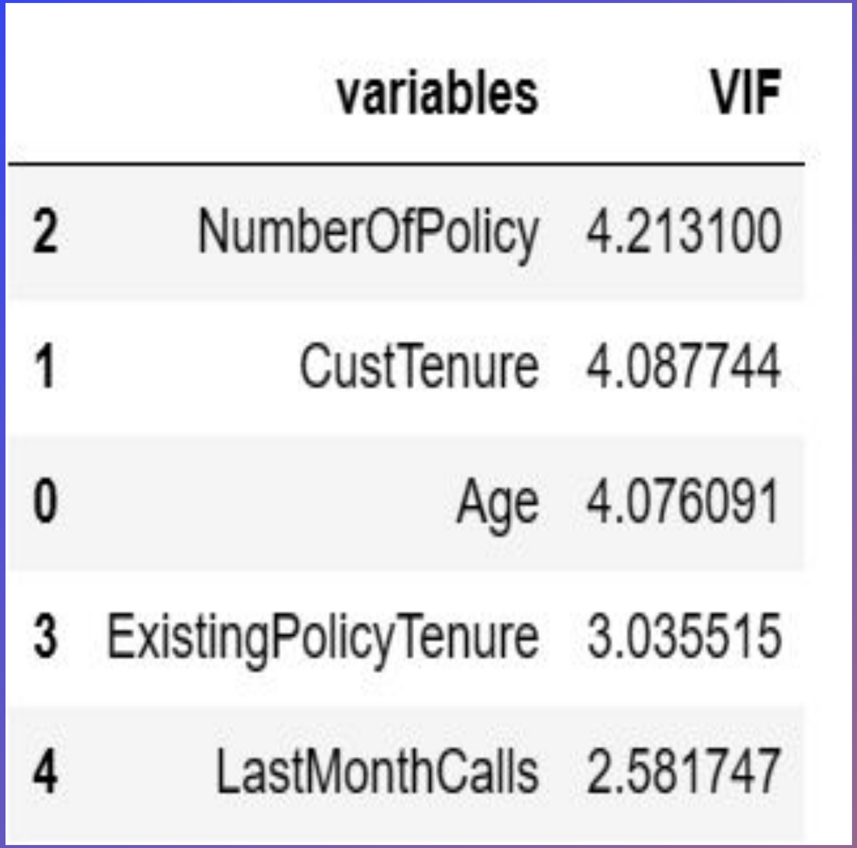
c) Removal of unwanted variables



	variables	VIF
4	SumAssured	10.740790
1	CustTenure	5.022706
0	Age	4.972977
2	NumberOfPolicy	4.751060
3	ExistingPolicyTenure	3.294731
5	LastMonthCalls	2.653156

After removing MonthlyIncome I have repeated the VIF treatment and found in the second iteration, that the SumAssured has VIF of around 11, which means it has high multicollinearity in the data, hence I have removed it and again did the VIF treatment.

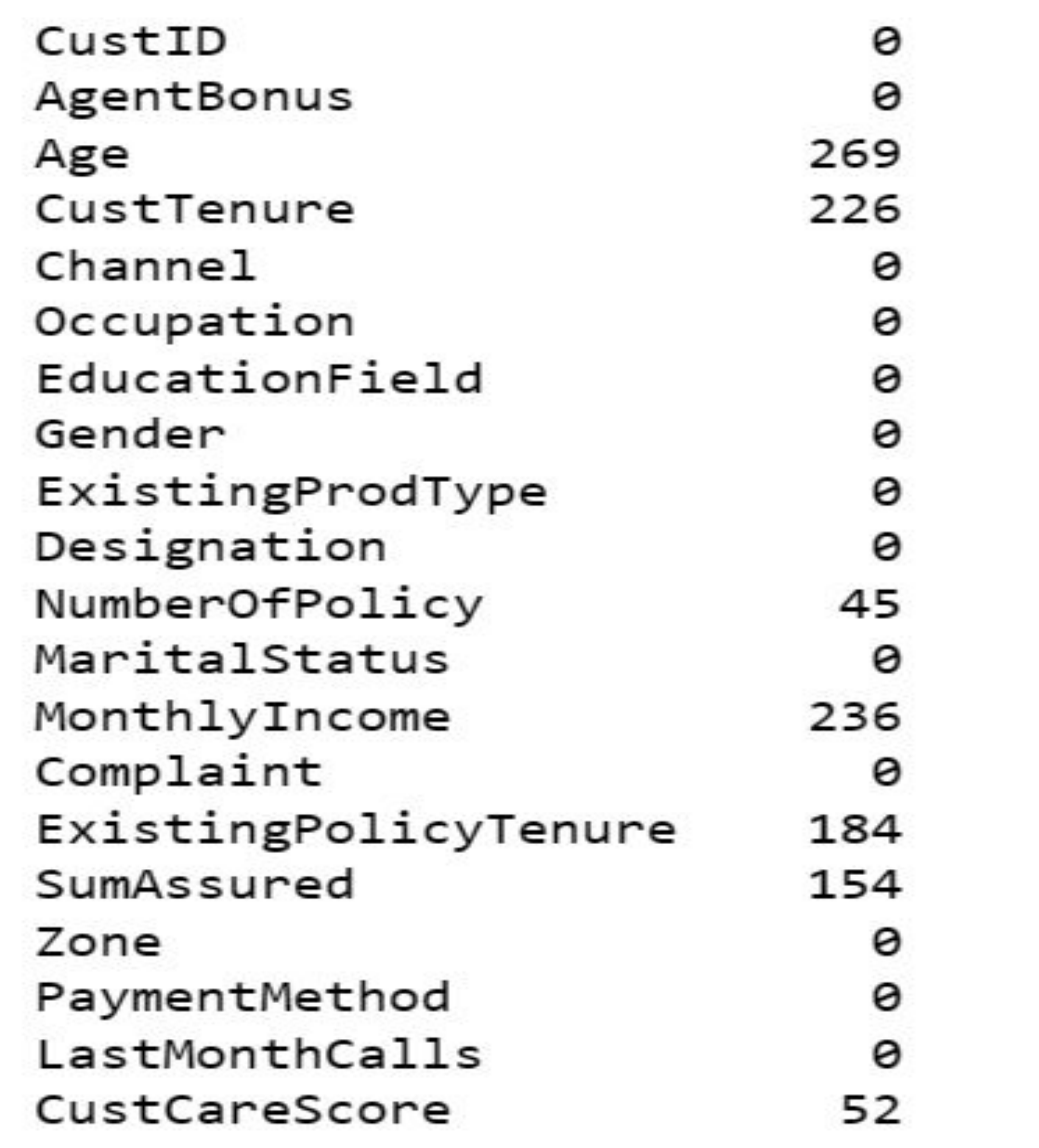
c) Removal of unwanted variables



	variables	VIF
2	NumberOfPolicy	4.213100
1	CustTenure	4.087744
0	Age	4.076091
3	ExistingPolicyTenure	3.035515
4	LastMonthCalls	2.581747

In the next iteration, I found that the VIF score of all the numerical column is less than 5, hence I have stopped the VIF treatment further as VIF score or 5 or more is usually taken to be standard value to remove the feature.

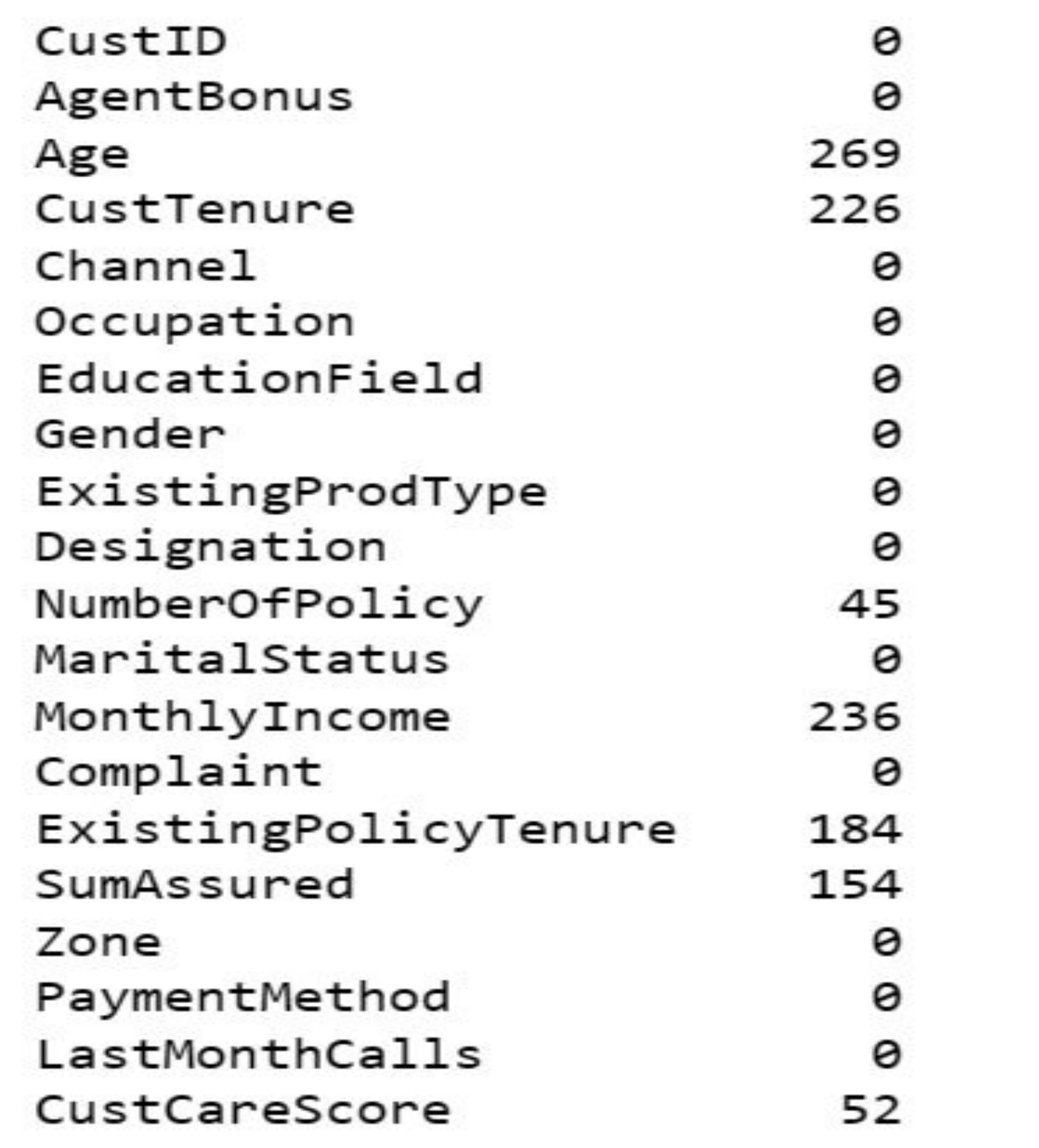
d) Missing Value treatment



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

In the data, I found that there are missing values which are mainly found in the Age, CustTenure, MonthlyIncome, NumberOfPolicy, ExistingPolicyTenure, SumAssured and CustCareScore columns. Hence these needs to be treated.

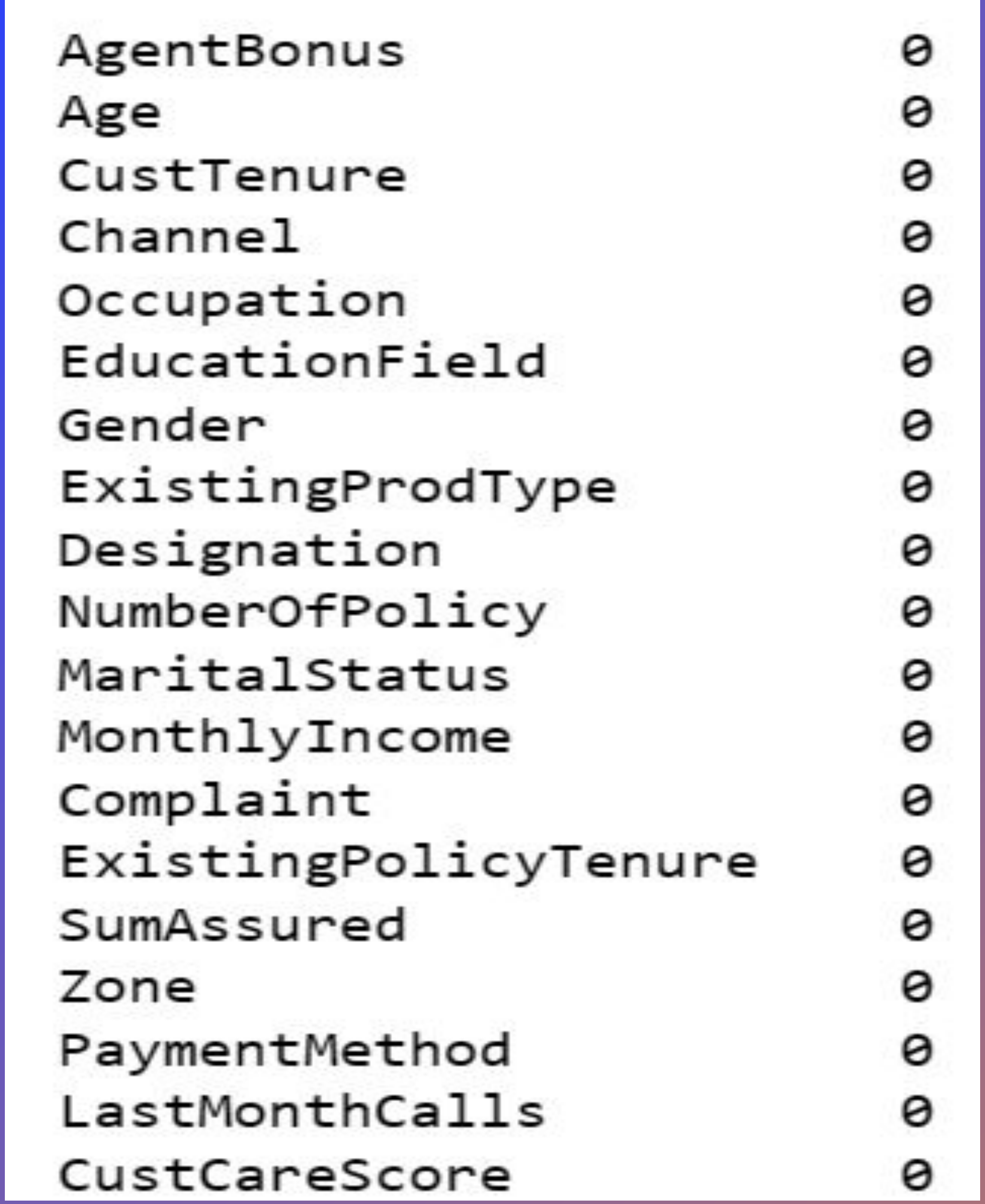
d) Missing Value treatment



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

For missing value treatment, we can use various method like KNN imputer or impute the missing value with mean, median or mode depending on the data type of the column. Here, I have used median imputation for numerical column as they had outliers in them and mode imputation for categorical columns.

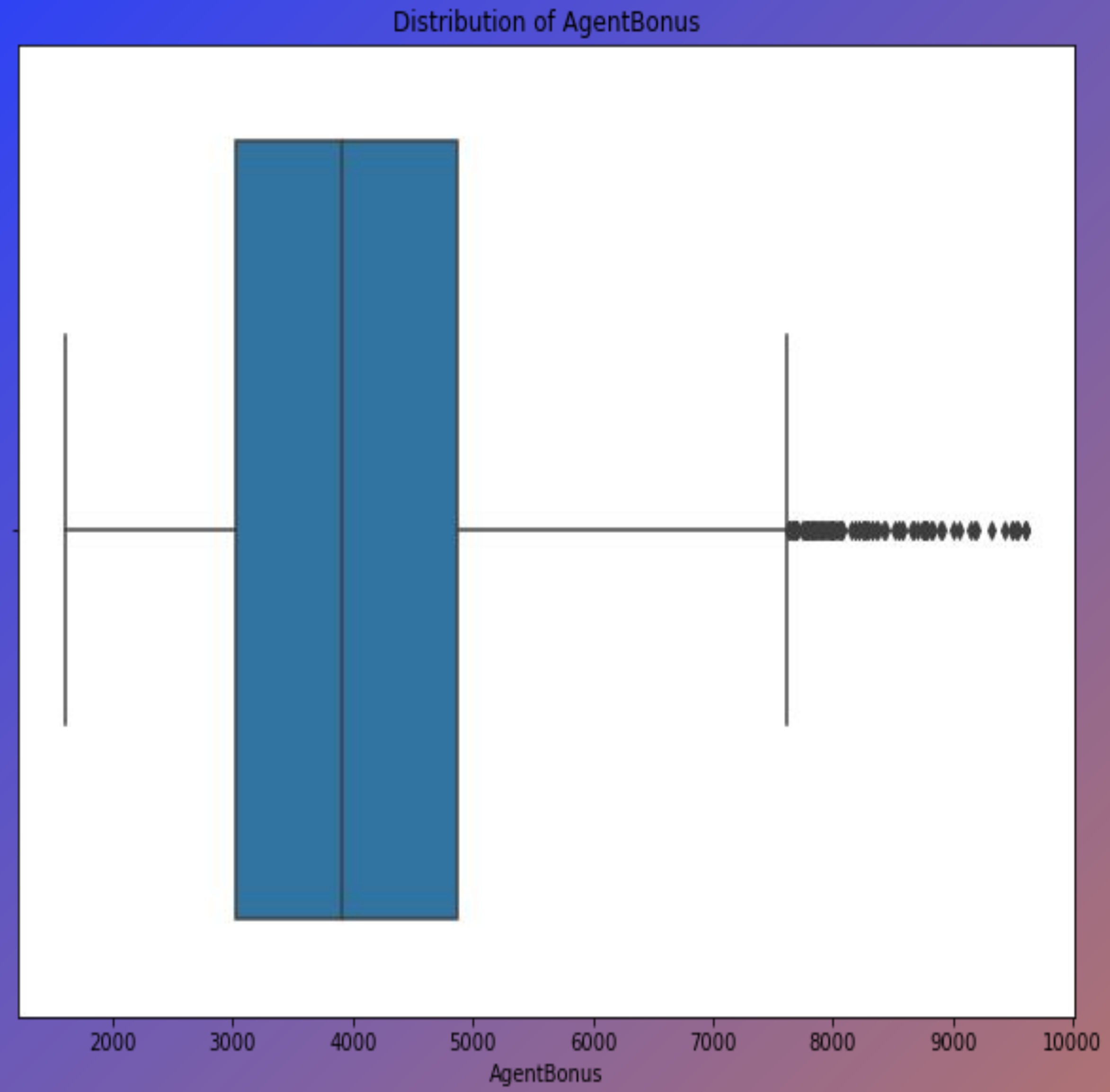
d) Missing Value treatment

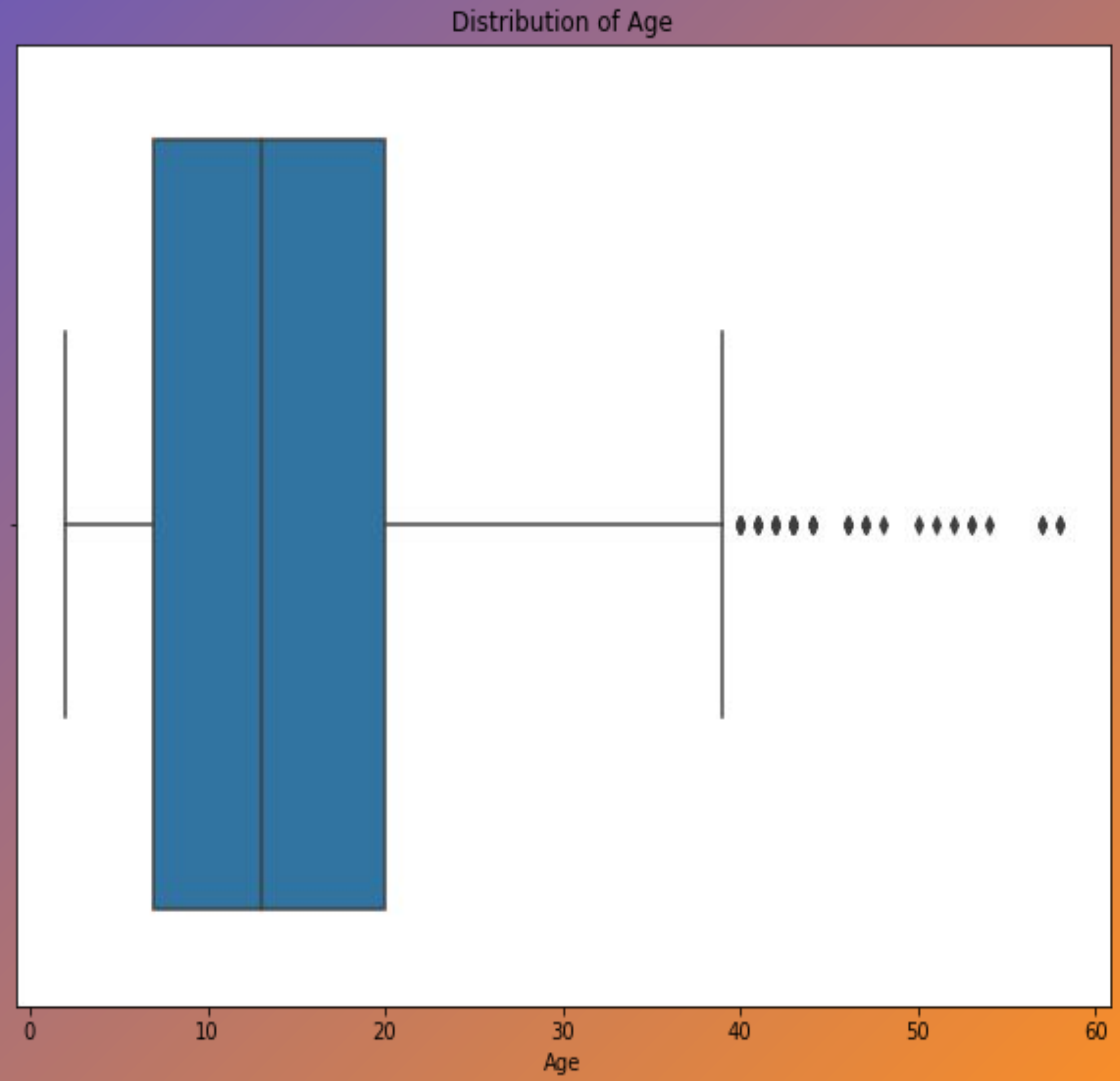


```
AgentBonus              0
Age                     0
CustTenure              0
Channel                 0
Occupation              0
EducationField          0
Gender                  0
ExistingProdType        0
Designation             0
NumberOfPolicy          0
MaritalStatus           0
MonthlyIncome           0
Complaint               0
ExistingPolicyTenure    0
SumAssured              0
Zone                    0
PaymentMethod           0
LastMonthCalls          0
CustCareScore           0
```

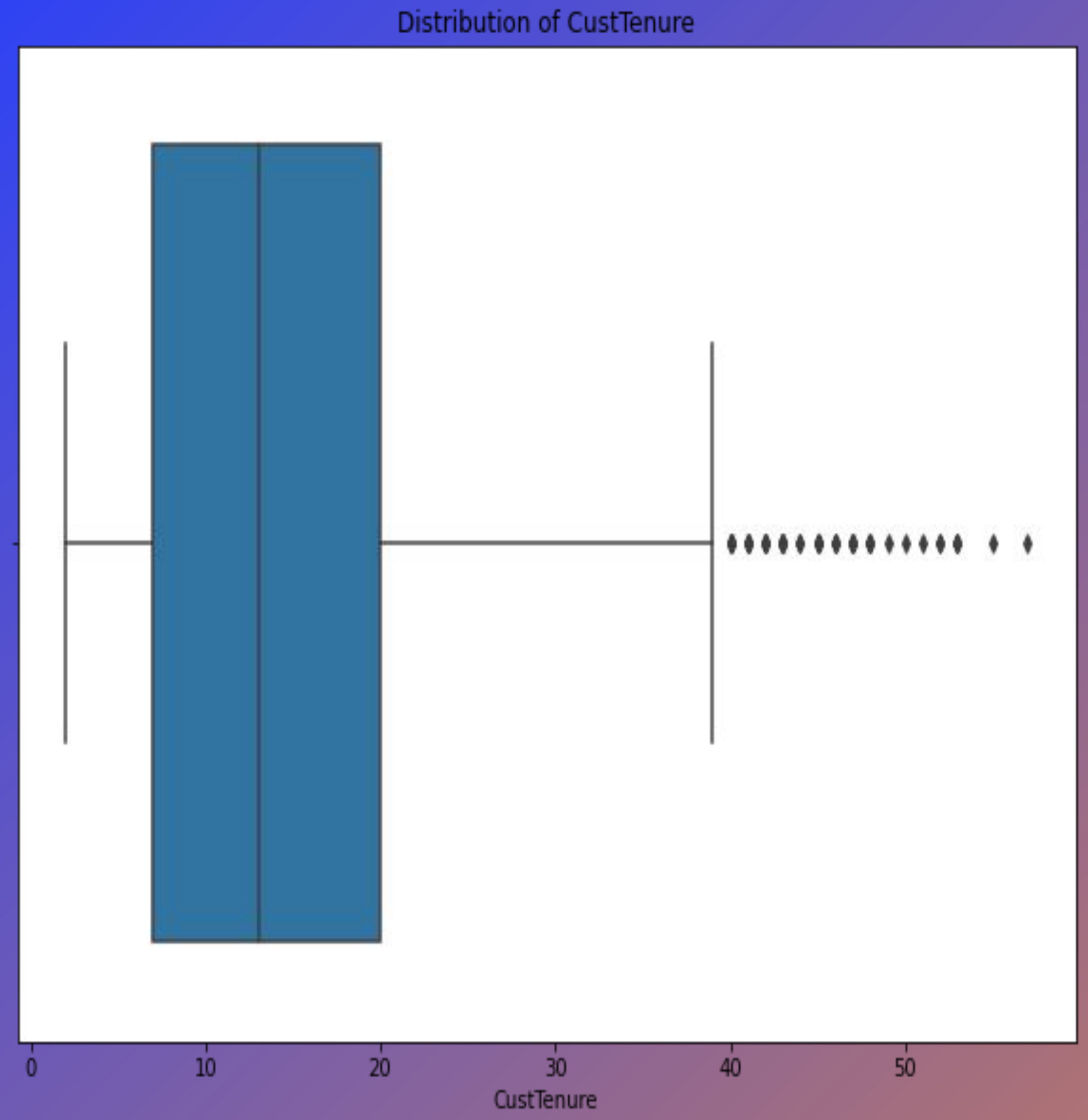
I can see that after imputation, there are no missing or null value present in the data.

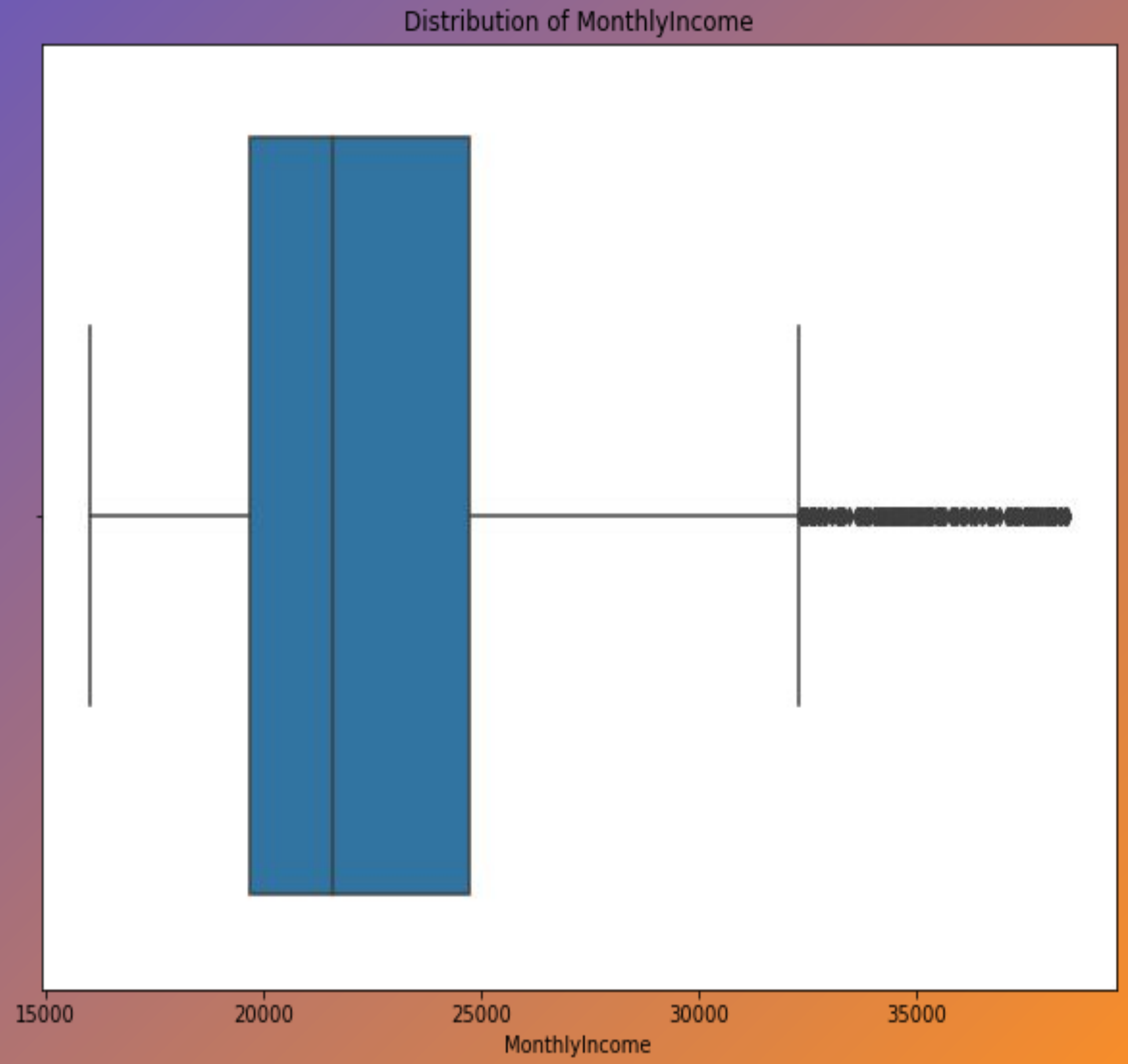
e) Outlier treatment



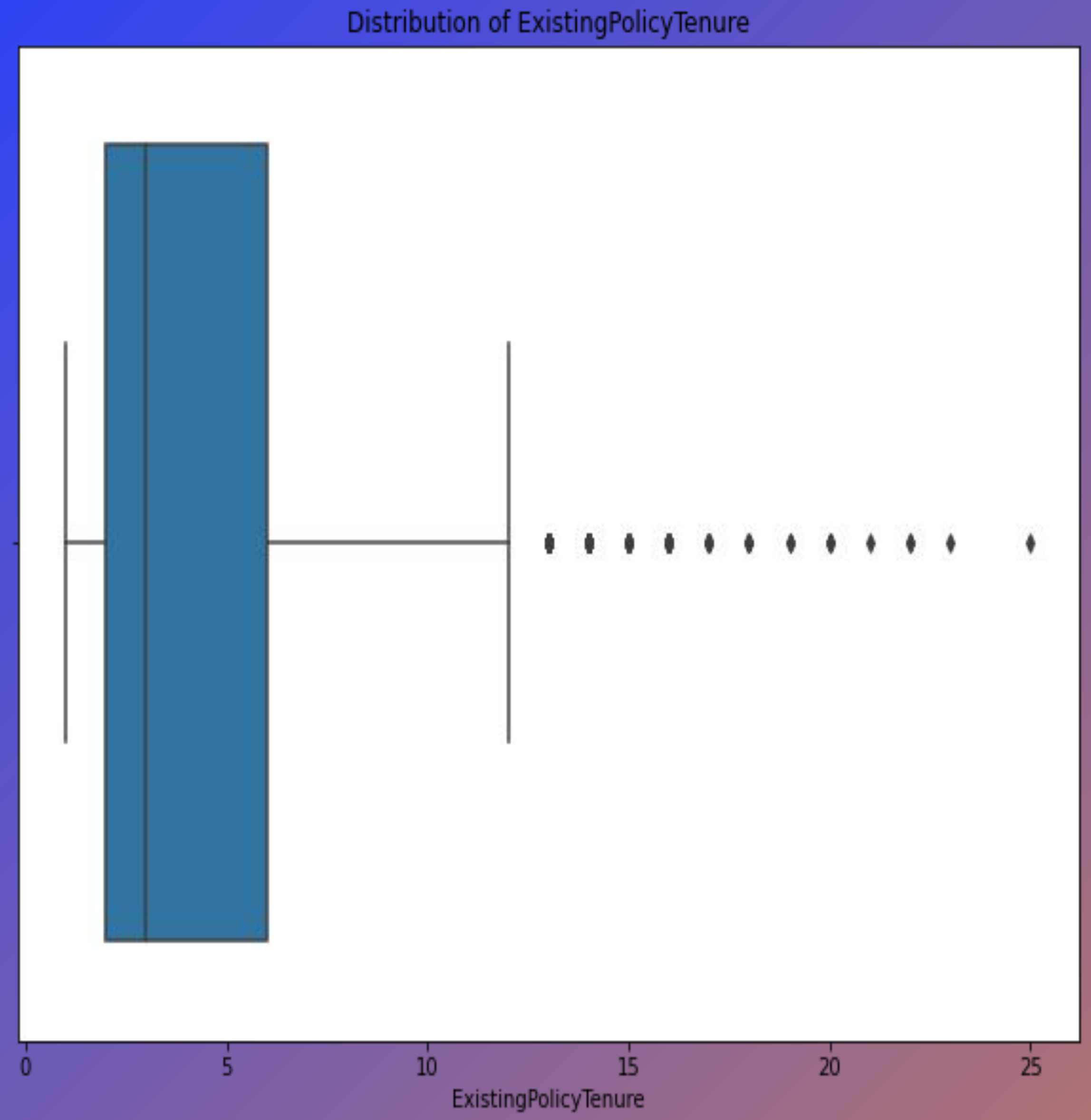


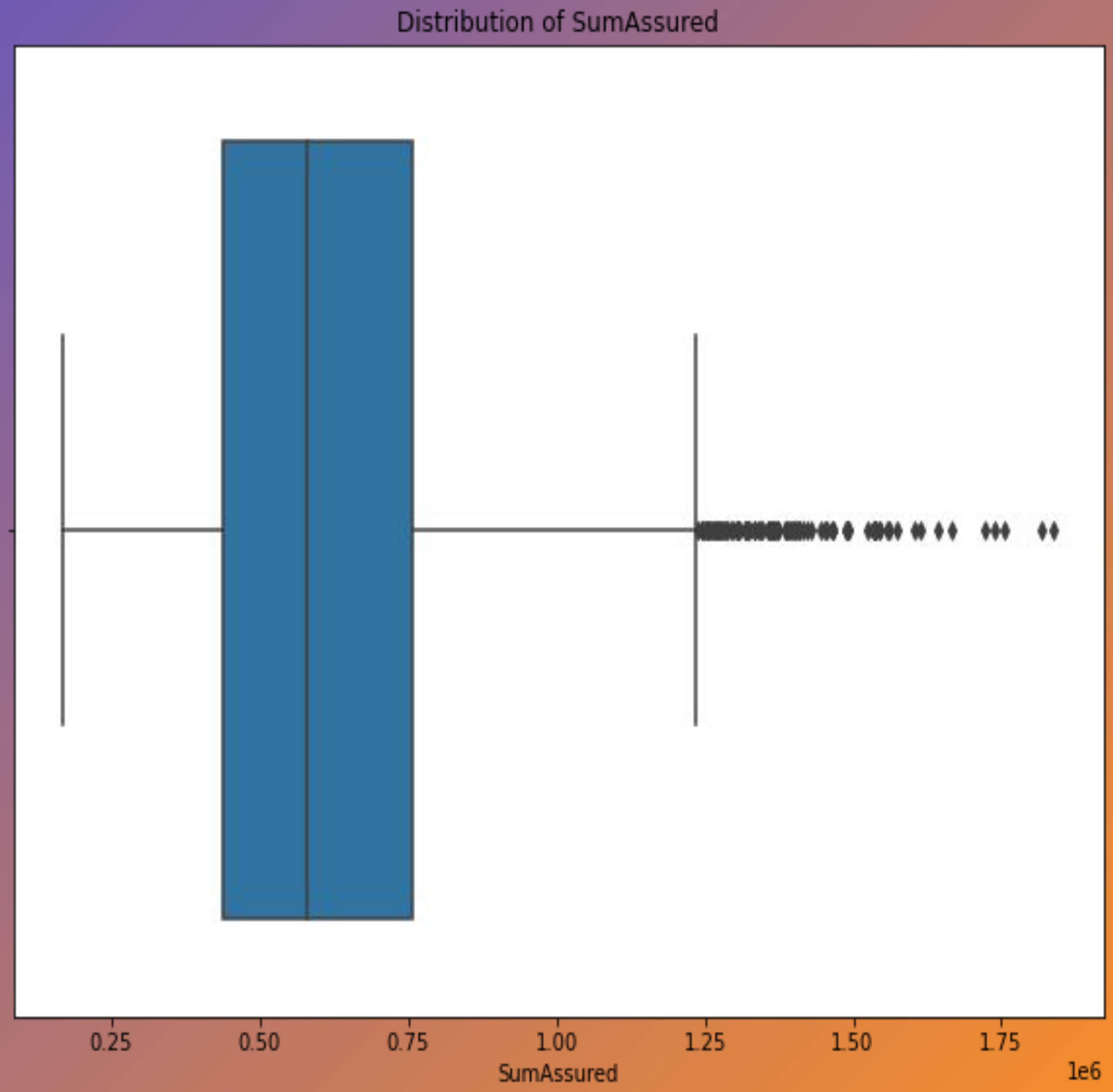
e) Outlier treatment



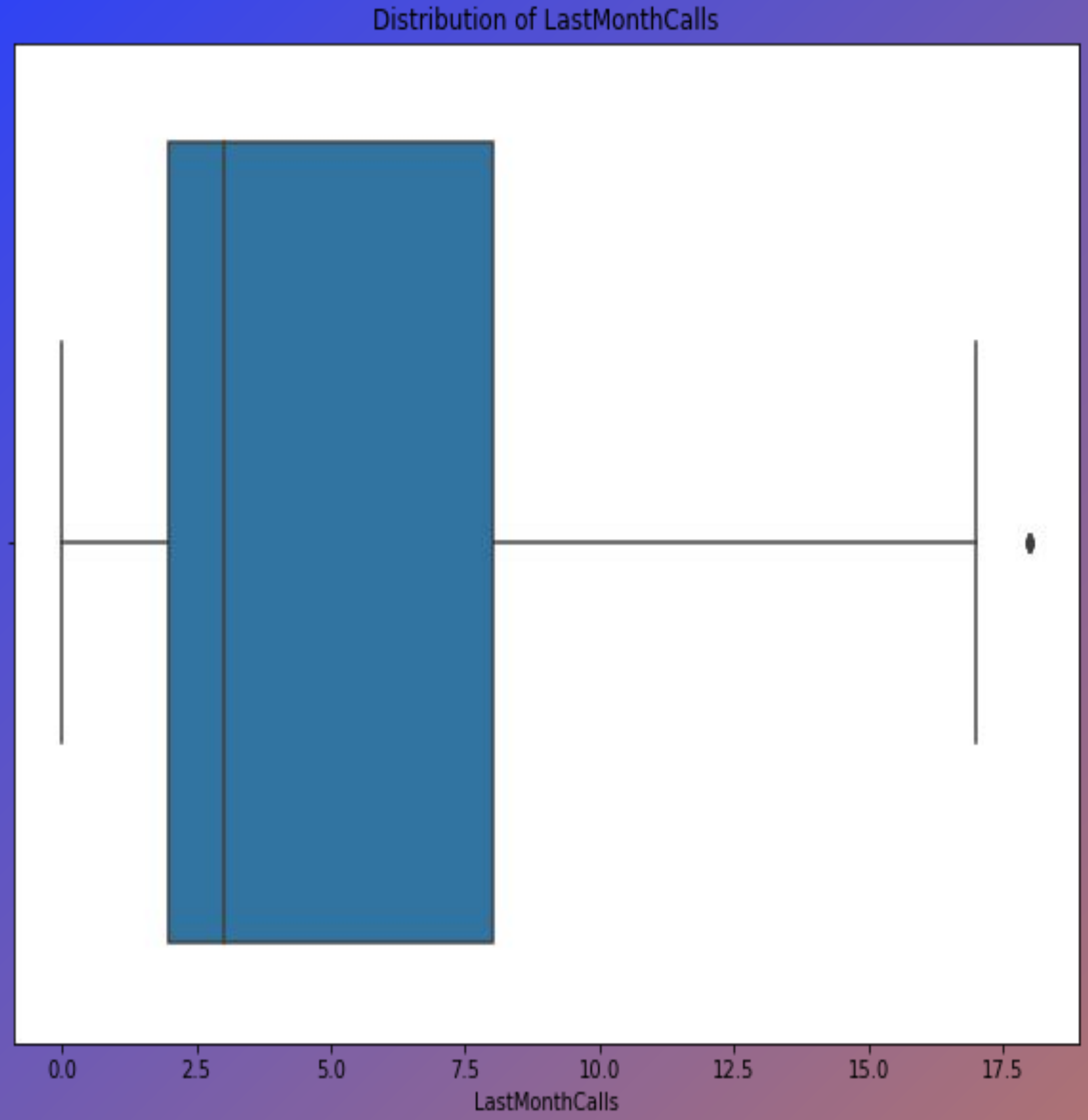


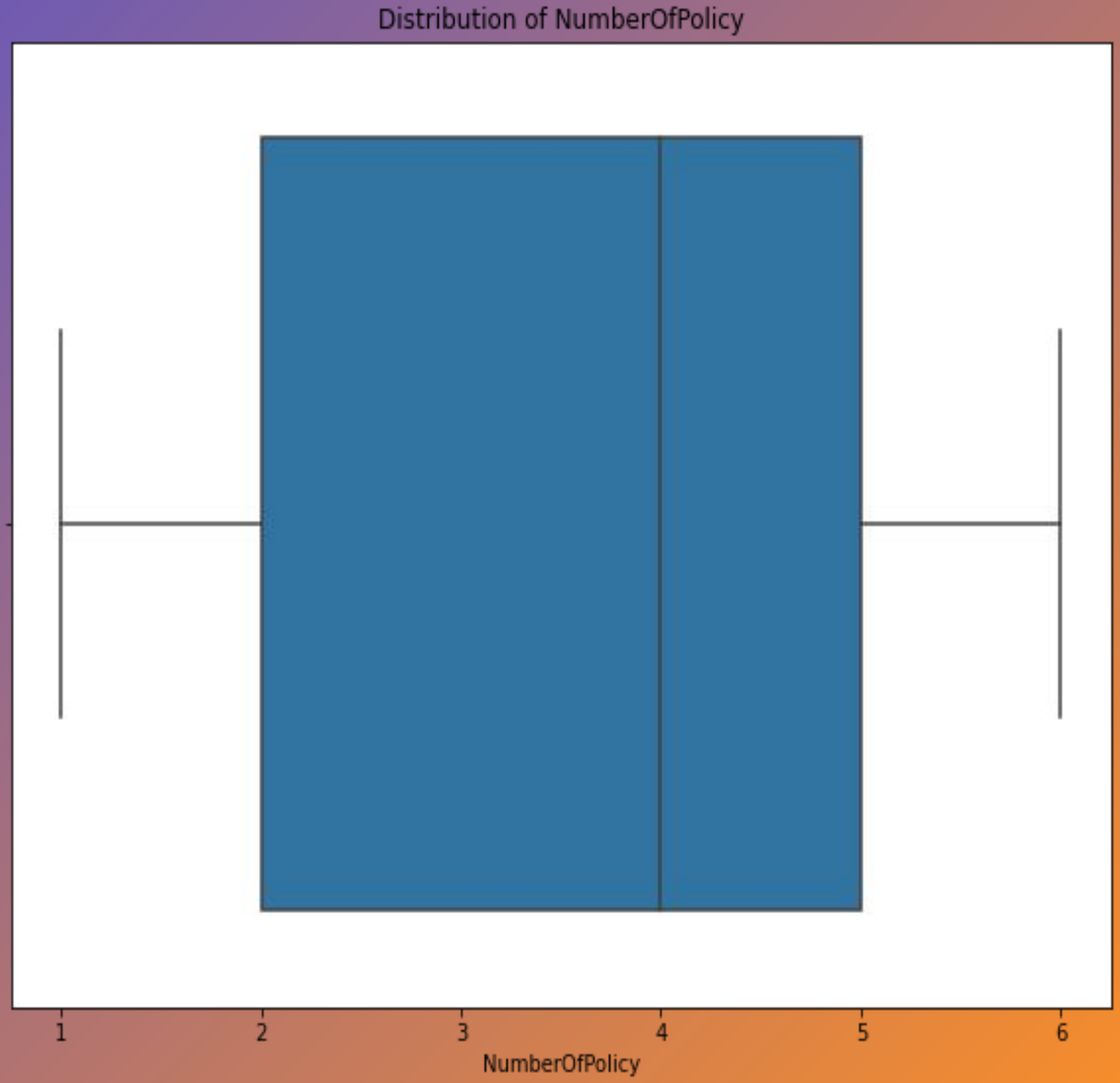
e) Outlier treatment





e) Outlier treatment





e) Outlier treatment

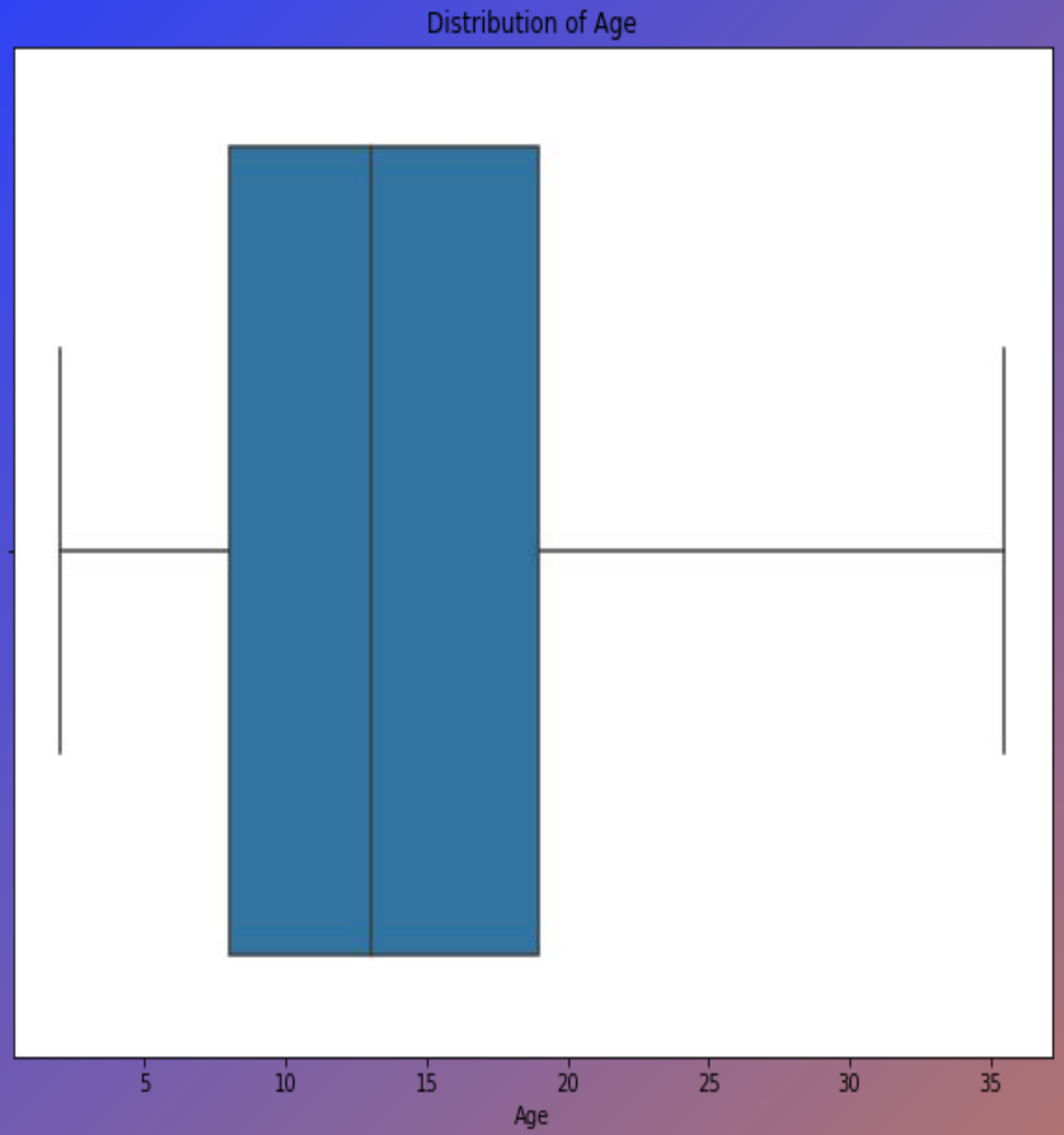
I can see from all the boxplots for the categorical columns that there are outliers in most of the features, hence they need to be treated.

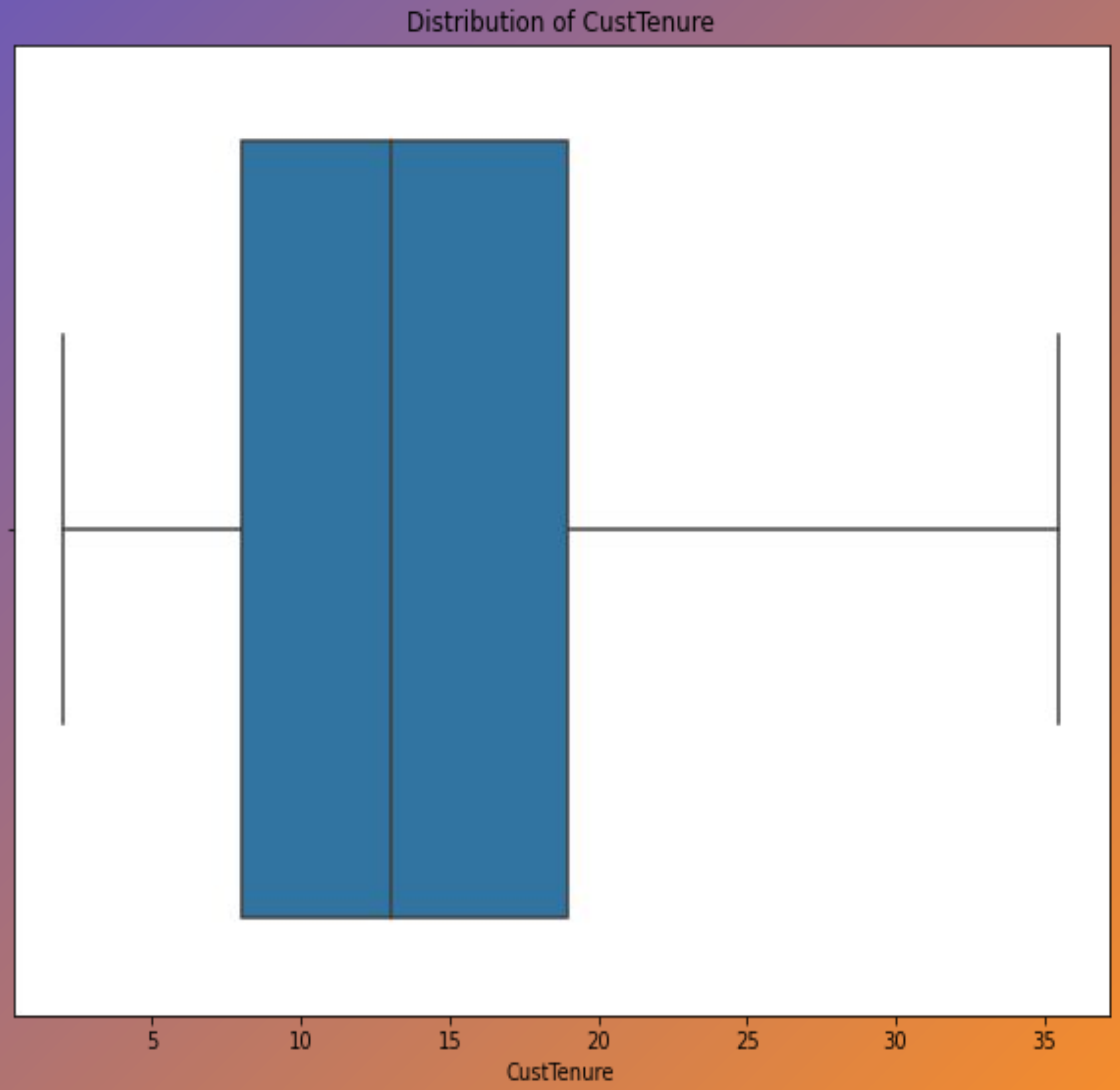
I have treated outliers by replacing lower side outliers to minimum value (which are mostly not present in the data) and upper side outliers to 1.5 times IQR which is equal to 75 percentile minus 25 percentile.

I haven’t treated outliers in the Target column which is AgentBonus.

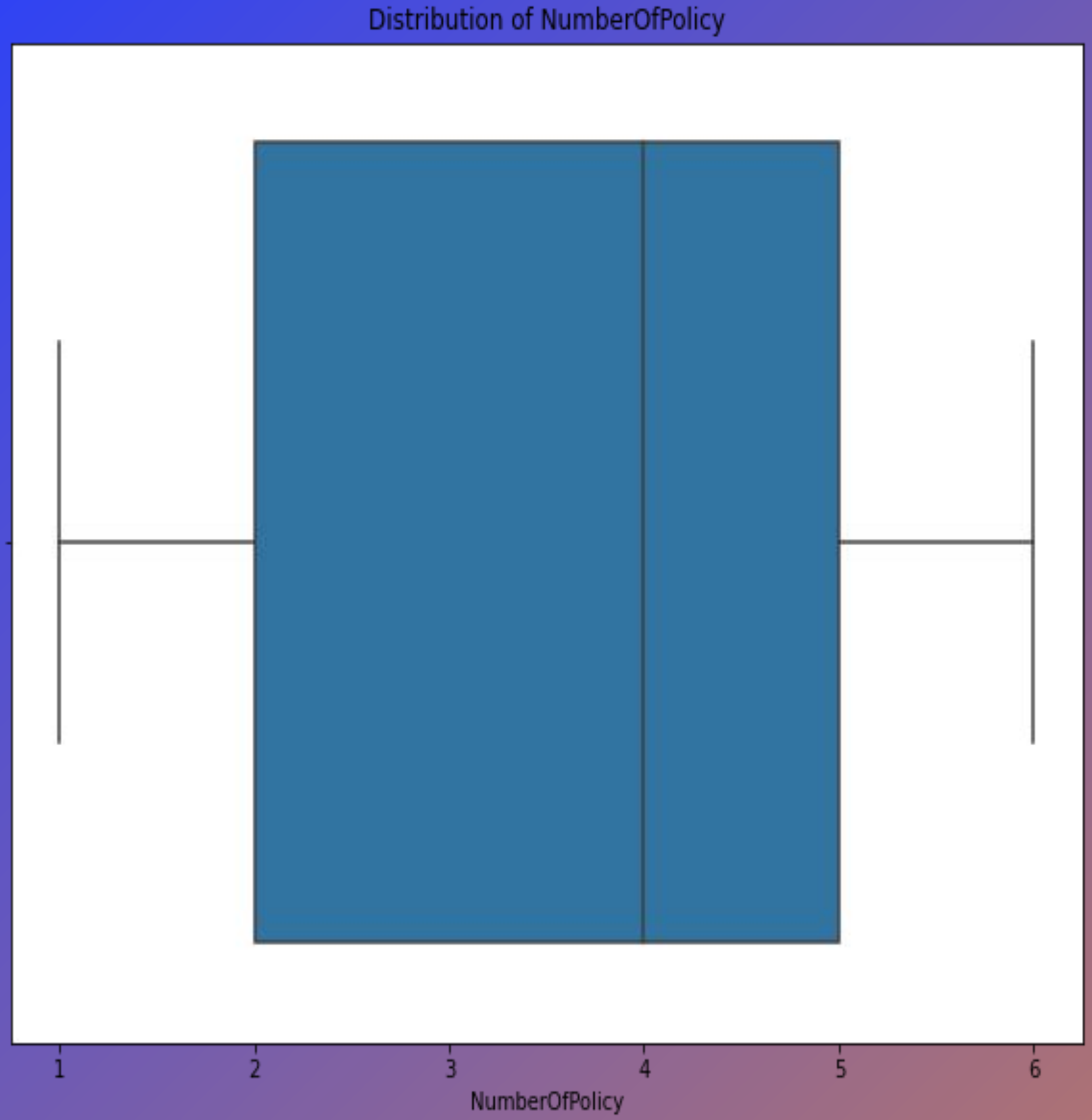
I can see after treatment, there are no outliers present which can be seen in the text few slides.

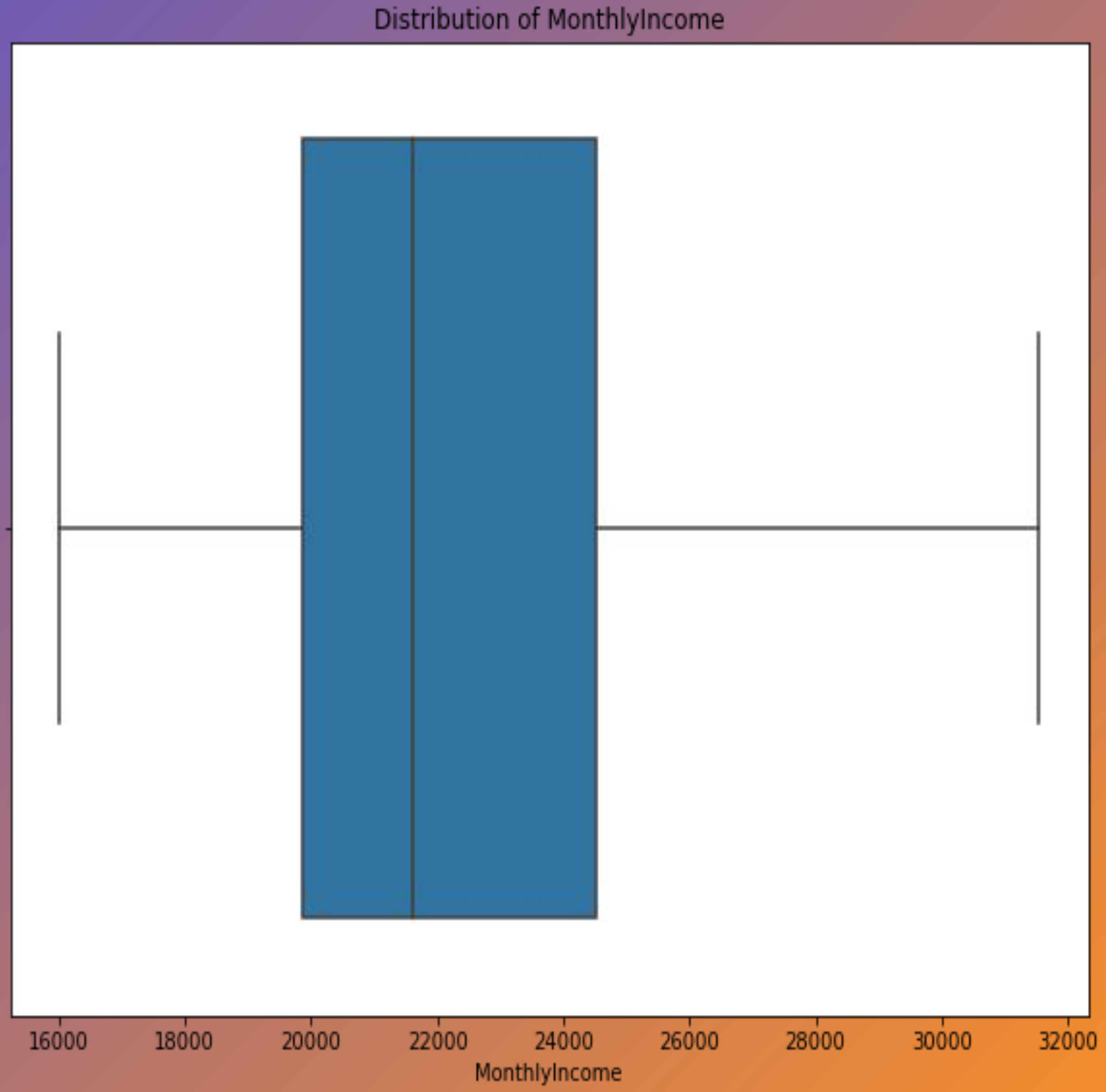
e) Outlier treatment



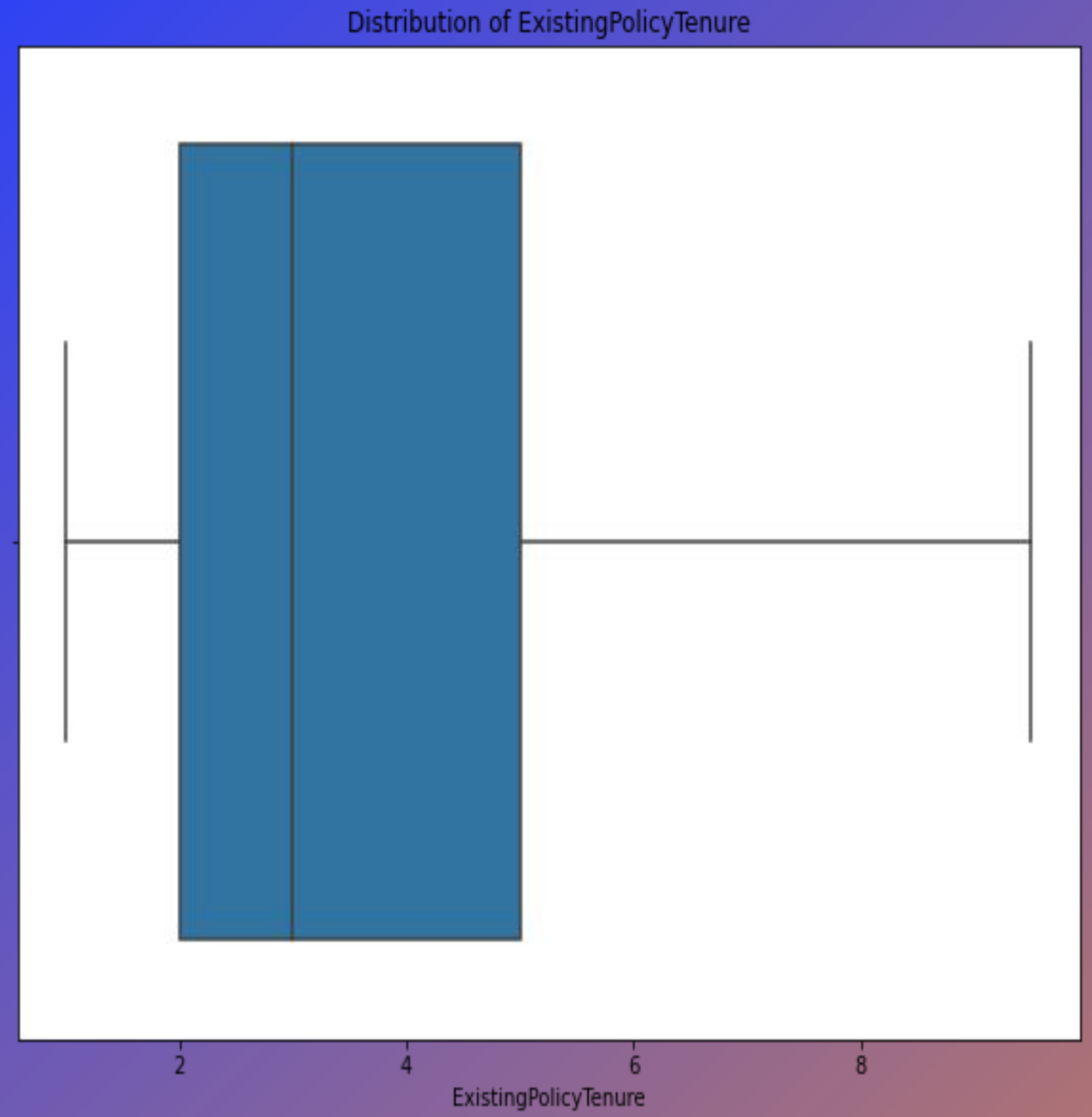


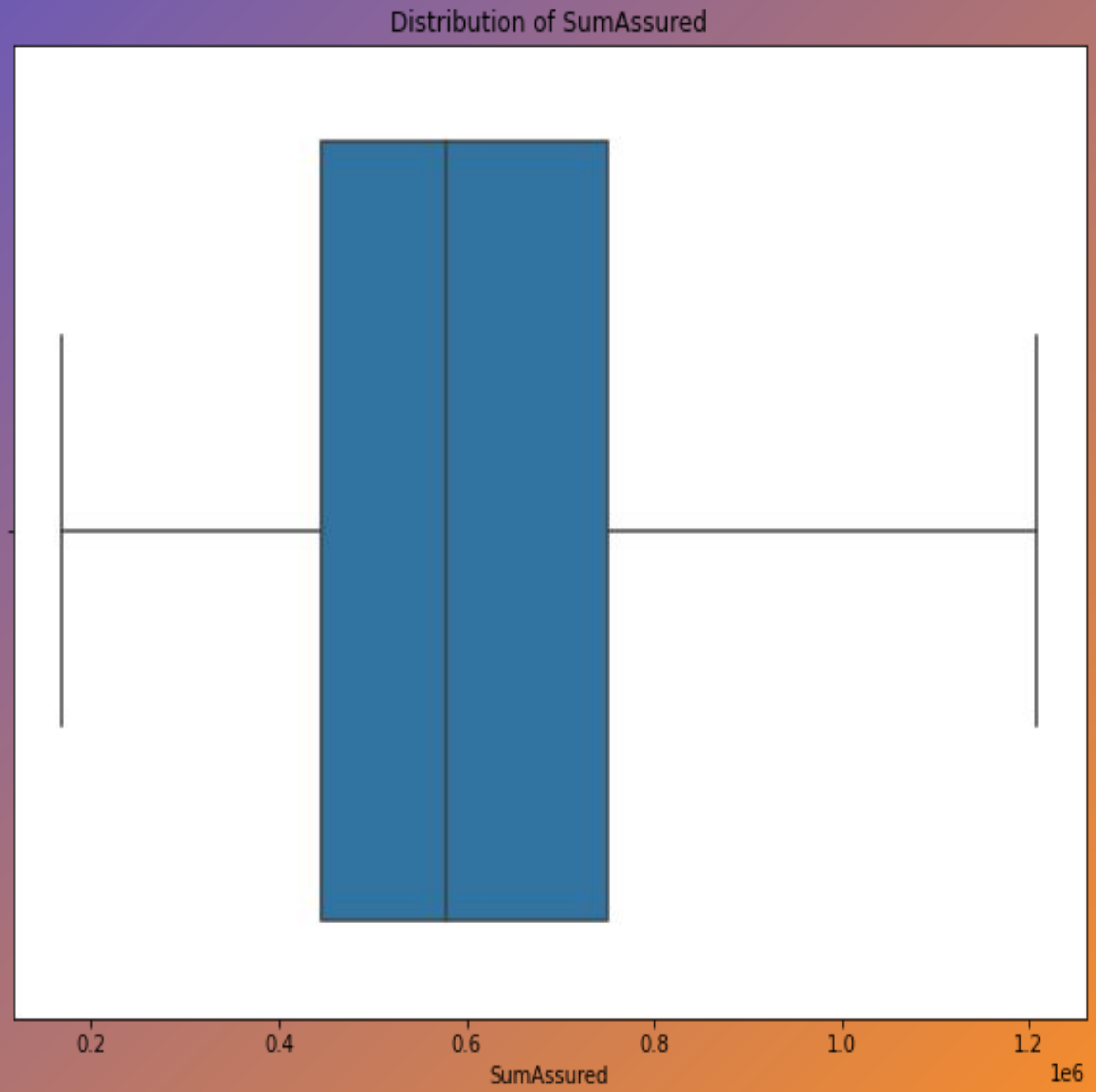
e) Outlier treatment



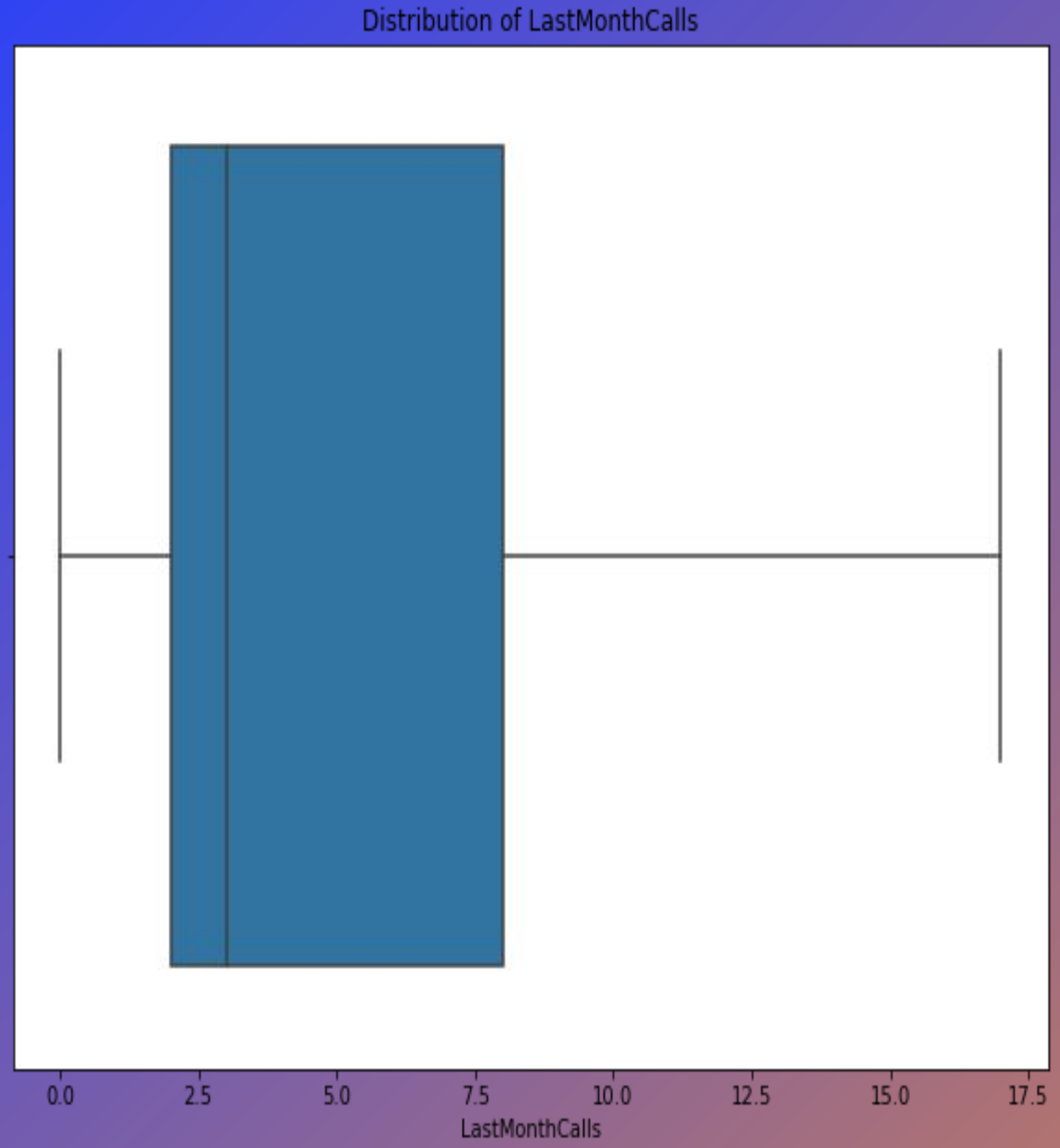


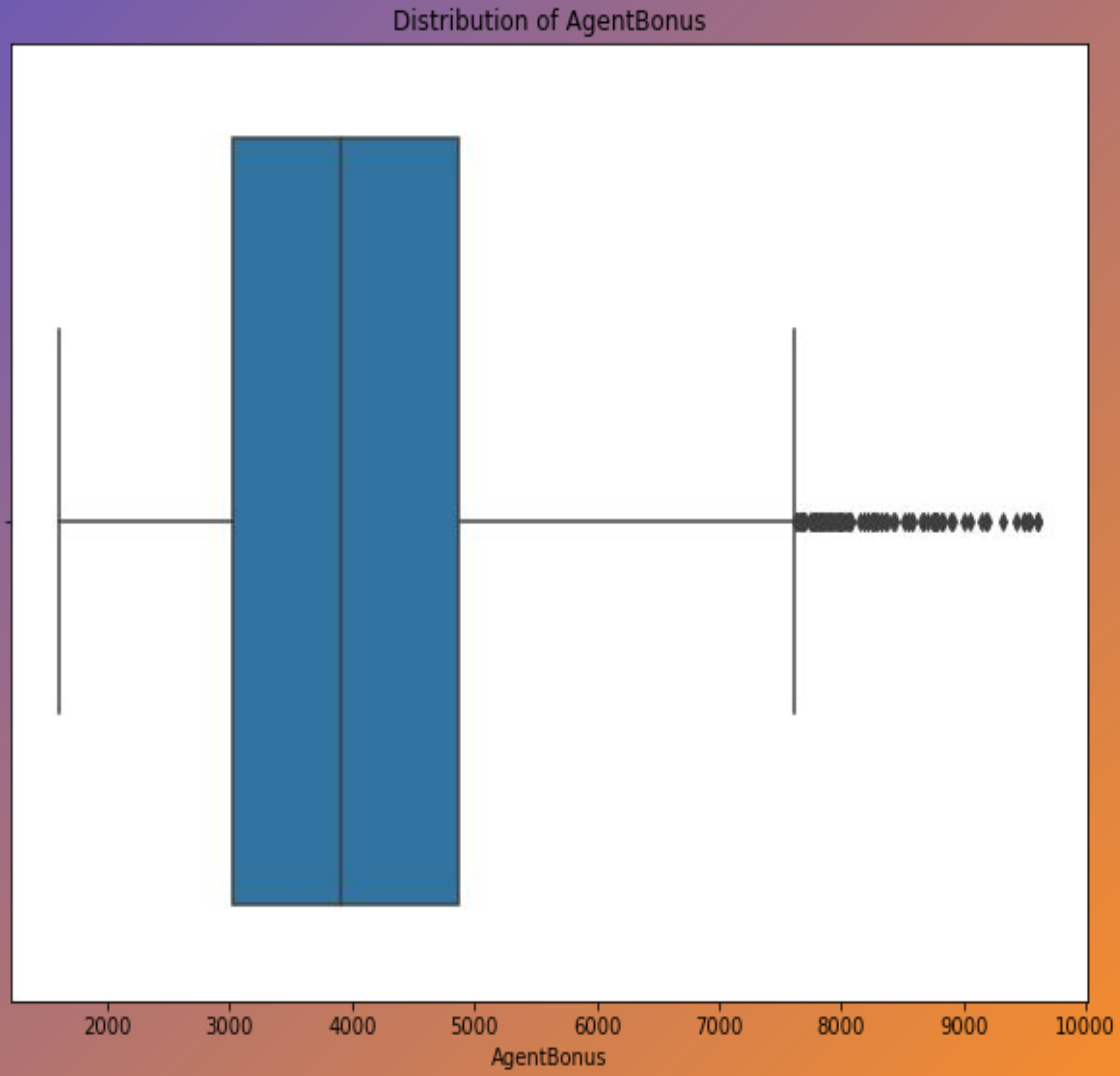
e) Outlier treatment





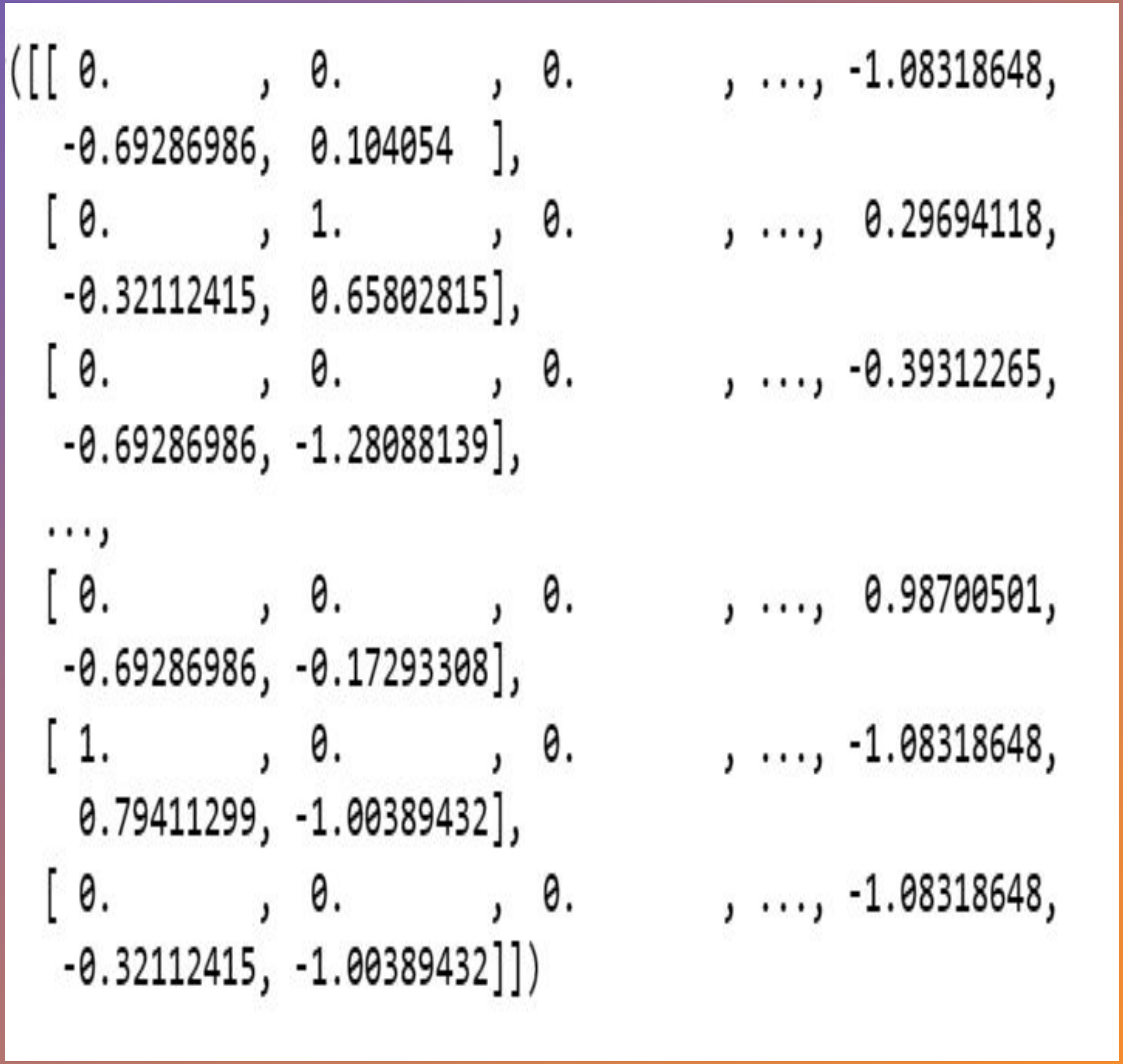
e) Outlier treatment





f) Variable transformation

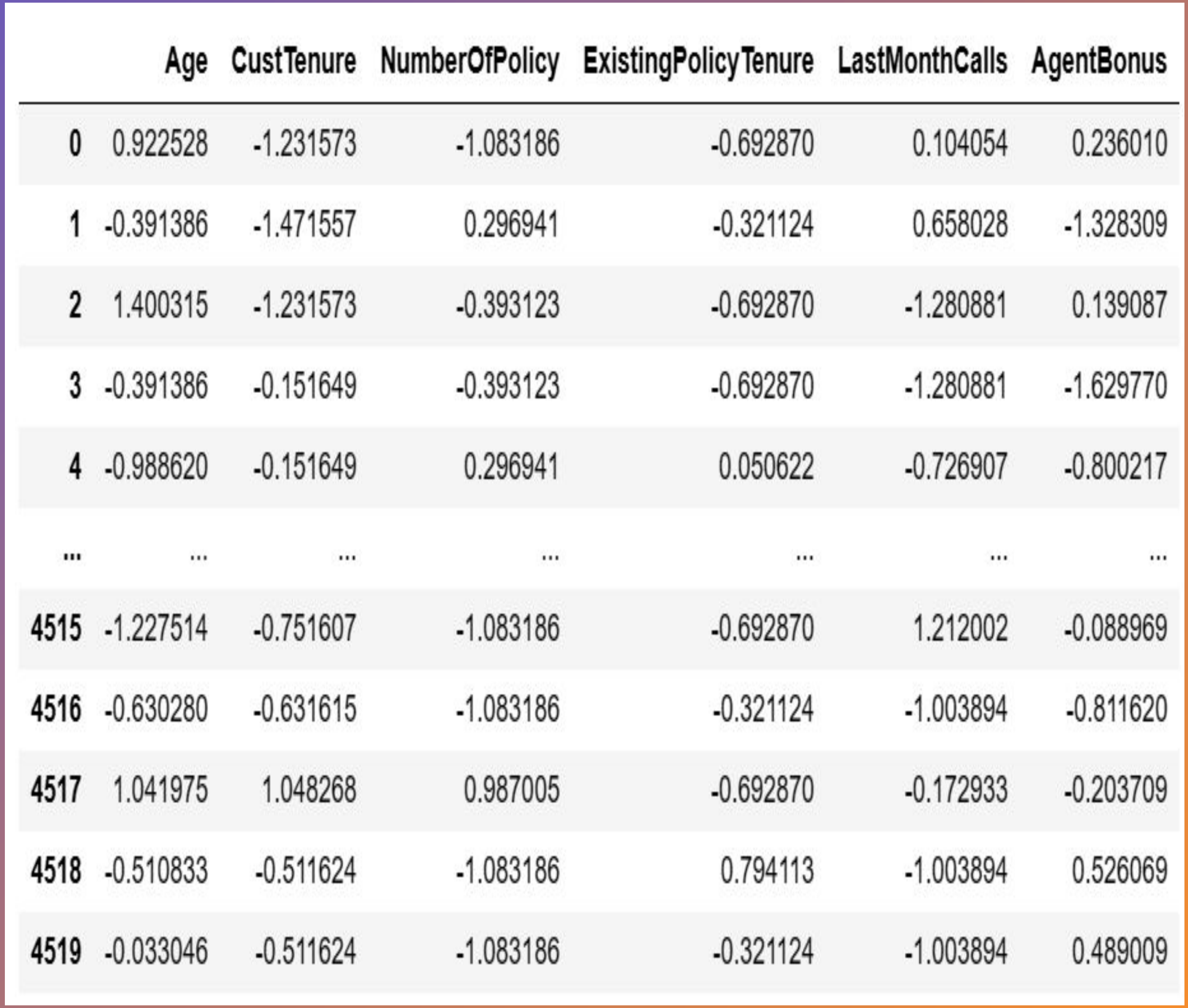
I have used OneHotEncoder to encode the categorical column as I have to use kmeans clustering for which categorical column must be encoded.
It has created new columns for the categories and added values of 0 or 1 whether that class is present or not.



```
([[ 0.        , 0.        , 0.        , ..., -1.08318648,
   -0.69286986, 0.104054  ],
  [ 0.        , 1.        , 0.        , ...,  0.29694118,
   -0.32112415, 0.65802815],
  [ 0.        , 0.        , 0.        , ..., -0.39312265,
   -0.69286986, -1.28088139],
  ...,
  [ 0.        , 0.        , 0.        , ...,  0.98700501,
   -0.69286986, -0.17293308],
  [ 1.        , 0.        , 0.        , ..., -1.08318648,
    0.79411299, -1.00389432],
  [ 0.        , 0.        , 0.        , ..., -1.08318648,
   -0.32112415, -1.00389432]])
```

f) Variable transformation

I have scaled the numerical data by using StandardScaler because if we don’t scale then kmeans might be biased towards the column which have higher mean as it is distance based algorithm.

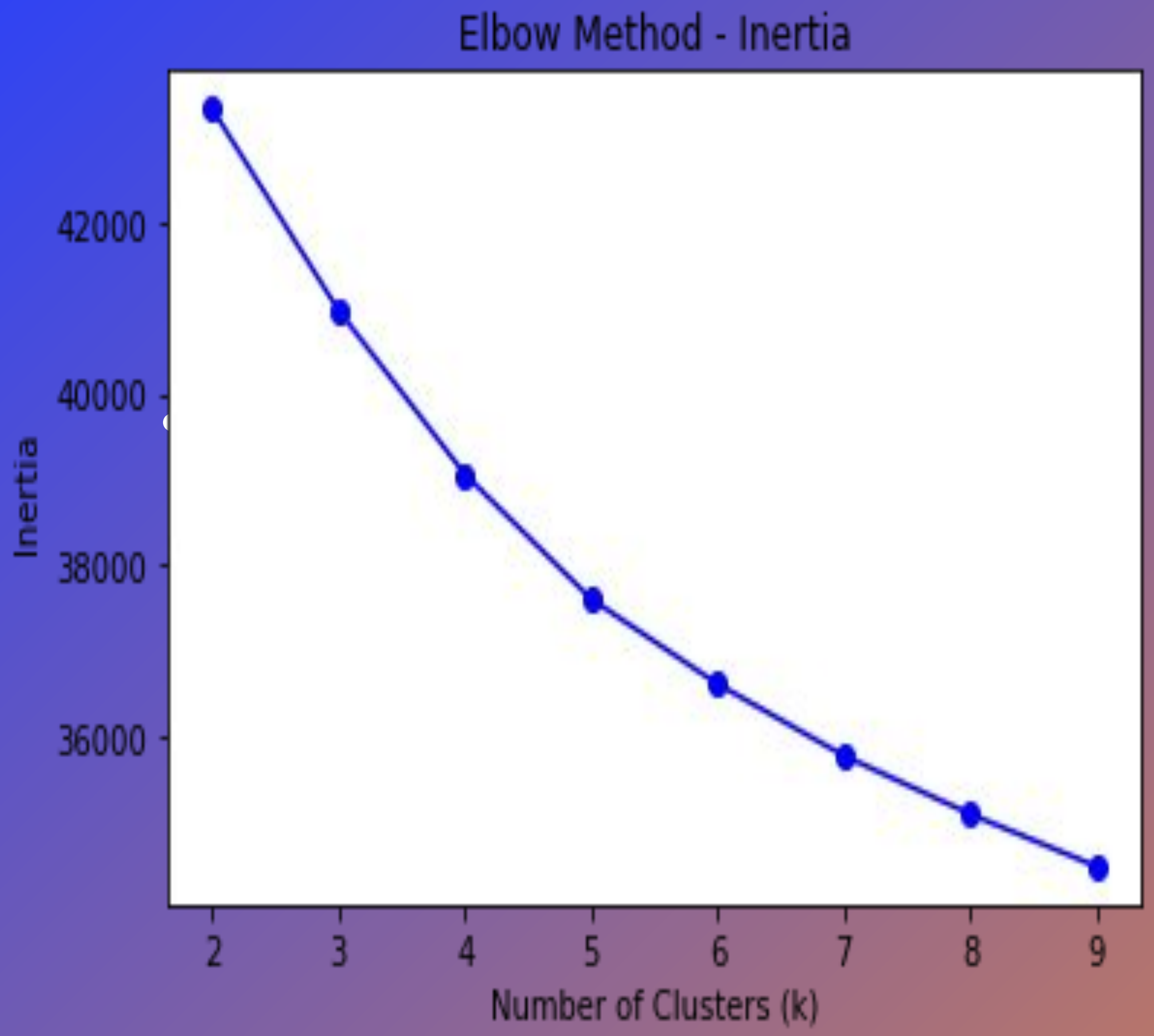


	Age	CustTenure	NumberOfPolicy	ExistingPolicyTenure	LastMonthCalls	AgentBonus
0	0.922528	-1.231573	-1.083186	-0.692870	0.104054	0.236010
1	-0.391386	-1.471557	0.296941	-0.321124	0.658028	-1.328309
2	1.400315	-1.231573	-0.393123	-0.692870	-1.280881	0.139087
3	-0.391386	-0.151649	-0.393123	-0.692870	-1.280881	-1.629770
4	-0.988620	-0.151649	0.296941	0.050622	-0.726907	-0.800217
...	...	...	...	...	...	...
4515	-1.227514	-0.751607	-1.083186	-0.692870	1.212002	-0.088969
4516	-0.630280	-0.631615	-1.083186	-0.321124	-1.003894	-0.811620
4517	1.041975	1.048268	0.987005	-0.692870	-0.172933	-0.203709
4518	-0.510833	-0.511624	-1.083186	0.794113	-1.003894	0.526069
4519	-0.033046	-0.511624	-1.083186	-0.321124	-1.003894	0.489009

g) Addition of new variables

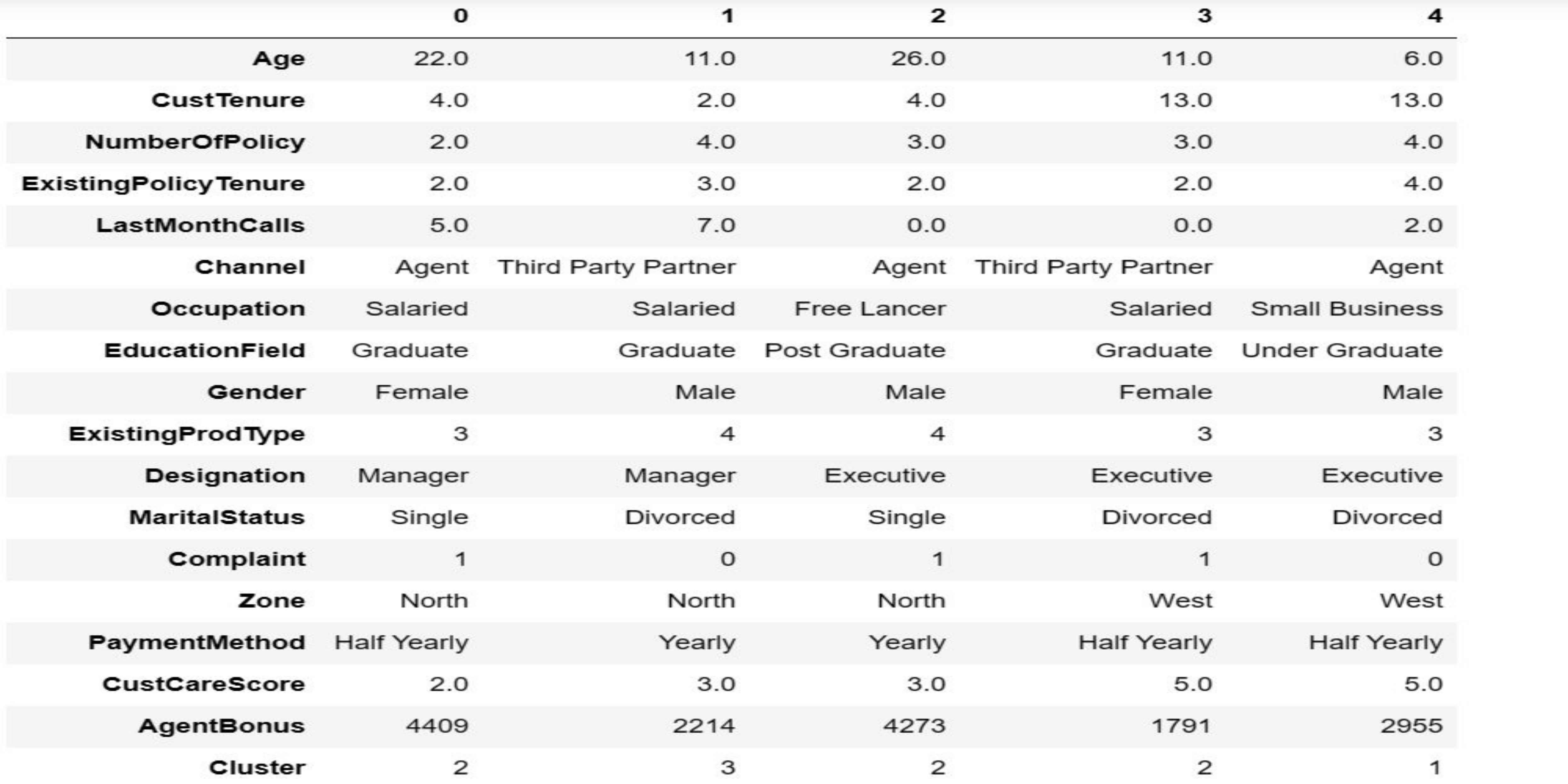
I have clustered the data using kmeans and added a new column showing which cluster the row belongs to.
Clustering will give idea on how the agents have performed. Which agents need to be given good bonus, which agent needs training and how and to which type of customers we have to focus on.

g) Addition of new variables



To find the number of clusters to be formed, I have used Elbow method to find it and the optimum number of clusters came out to be 5 as elbow is formed at this point and I have clustered the data into 5 clusters.
The clusters can be seen in the next slide.

g) Addition of new variables



	0	1	2	3	4
Age	22.0	11.0	26.0	11.0	6.0
CustTenure	4.0	2.0	4.0	13.0	13.0
NumberOfPolicy	2.0	4.0	3.0	3.0	4.0
ExistingPolicyTenure	2.0	3.0	2.0	2.0	4.0
LastMonthCalls	5.0	7.0	0.0	0.0	2.0
Channel	Agent	Third Party Partner	Agent	Third Party Partner	Agent
Occupation	Salaried	Salaried	Free Lancer	Salaried	Small Business
EducationField	Graduate	Graduate	Post Graduate	Graduate	Under Graduate
Gender	Female	Male	Male	Female	Male
ExistingProdType	3	4	4	3	3
Designation	Manager	Manager	Executive	Executive	Executive
MaritalStatus	Single	Divorced	Single	Divorced	Divorced
Complaint	1	0	1	1	0
Zone	North	North	North	West	West
PaymentMethod	Half Yearly	Yearly	Yearly	Half Yearly	Half Yearly
CustCareScore	2.0	3.0	3.0	5.0	5.0
AgentBonus	4409	2214	4273	1791	2955
Cluster	2	3	2	2	1

Question 4 - Business insights from EDA

a) Is the data unbalanced?

Yes the data seem to be imbalance.
If we see the column of Zone, the data consists most of the customers and agents from West and North and very less from South and East, so it will be difficult to predict about these zones.
Similarly data is imbalanced as per the gender as more data is about male and less about female.
Similarly for Occupation, we don’t have much data about freelancers, so we can’t comment whether freelancers don’t purchase the insurance or we are not targeting them.

a) Is the data unbalanced?

Yes the data seem to be imbalance.
If we see the column of Zone, the data consists most of the customers and agents from West and North and very less from South and East, so it will be difficult to predict about these zones.
Similarly data is imbalanced as per the gender as more data is about male and less about female.
Similarly for Occupation, we don’t have much data about freelancers, so we can’t comment whether freelancers don’t purchase the insurance or we are not targeting them.
…

a) Is the data unbalanced?

Hence, its difficult to predict to the bonuses on these features whether we should give more bonus to North and West zone agents.
Similarly more bonus should be given to agents having male customer or female customer.
Similarly, we can’t predict what bonus should be given to the agents having freelancers as the customers.

The company can take steps to ensure fairness and motivation among agents. This could include adjusting the bonus allocation methodology, setting clear performance metrics, or implementing reward systems that consider individual agent performance relative to their peers.

a) Is the data unbalanced?

For the data we can first check with the team whether we can get more data from different zones and different occupation where the data is less.

We can create artificial data using SMOTE so as to create balanced data, but creating data for all columns may lead to bias in the data.

Hence, we need to proceed with given data to do the analysis.

b) Any business insights using clustering

The clustering has divided data into 5 parts as

3	1141
0	1093
1	899
2	731
4	656

Cluster 4 has the highest average age which is approx 25 years with the highest average agent bonus which is 5977 and the highest average customer tenure which is approx 25 years.

b) Any business insights using clustering

Cluster 2 has the least average age which is approx 10 years with the least average agent bonus which is 3264 and the least average customer tenure which is approx 11.13 years.

Cluster 3 has the average age which is approx 12 years with the average agent bonus which is 3801 and the average customer tenure which is approx 12 years.

Cluster 0 has the average age which is approx 15 years with the average agent bonus which is 4678 and the average customer tenure which is approx 15 years.

b) Any business insights using clustering

Cluster 1 has the average age which is approx 11.14 years with the average agent bonus which is 3400 and the average customer tenure which is approx 11.15 years.

So the cluster have been divided based on age where 2 clusters 3 and 1 have similar age of customer and its tenure but average agent bonus have a difference of approx 400. This might be due to the Designation of customers. Cluster 3 has more VP and AVP and hence higher bonus for agents.

c) Any other business insights

- Since the mean age of customers is 15 years, so we can target married people to opt for company’s insurance.
- We can increases the benefits for customers who remain for longer period of time with the company.
- Since, product type 4 is more popular, we can try to sell it to different customers specially the married customers.
- We can focus on finding the issue for low customer score and depending on that conduct training for the agents.
- Since, the agents are preferred by customers as compared to third party, we can hire more agents and train them to get more clients.

c) Any other business insights

- We can target more salaried and small business owners to take up insurance and also focus on why freelancers are not taking insurance or whether we are not targeting them and depending on that focus on onboarding freelancers and large business owner.
- We can focus more on VP and AVP by giving some personalized insurance policy, as they might bring more revenue as they might be opting for higher premium.
- Focus more on South and East zone and find why there are less customers from here. Get more agents there and get more customers.
- Target married people having children.

THANK YOU

RAMESH PANDEY

ramesh.pandey.1015@gmail.com

9566793109