LIFE INSURANCE SALES

BUSINESS REPORT- PROJECT NOTES 1

Philjoy Dsilva

07-July-2024

PGPDSBA

Table of Contents

List of Tables

List of Figures

Life Insurance Sales

1) Introduction of the business problem

a) Defining 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

The objective of the above problem statement is to predict the monthly bonus of the insurance company in order to design engagement and upskilling programs for their agents.

b) Need of the study/project

The aim of this problem is to gain deeper insights into the performance of insurance company agents, ensuring fair compensation.

Predictive analytics will help the company pinpoint areas needing more focus; agents with lower policy sales may benefit from targeted training to enhance their performance, crucial as agents' portrayals heavily influence customer perceptions of policies.

High-performing agents, who sell more policies, deserve recognition and rewards to motivate sustained and improved future performance.

c) Understanding business/social opportunity

The company primarily sells life insurance financial products through agents to customers, with customer payments constituting its main income source. When customers' policies mature or they make valid claims, the company reimburses them according to the policy terms. In addition to managing customer funds, the company also engages in various ventures and investments.

Agents serve as the primary conduit for communicating the company's policies, goals, and benefits to customers. When customers are engaged by agents during policy discussions, they are more likely to be persuaded, thereby increasing sales and motivating the agents.

- This enhances the company's market share, positioning it ahead of competitors.
- Categorizing agents provides the company with valuable insights on where to focus efforts.
- Customer feedback helps in refining and updating policies/products to better meet their needs, enhancing customer retention.
- Ultimately, these efforts contribute to increased profitability for the company.

2) Data Report

a) Understanding how data was collected in terms of time, frequency and methodology

The dataset comprises details of life insurance policies along with the bonus amounts disbursed to agents in the previous month. This data is collected and updated on a monthly basis.

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

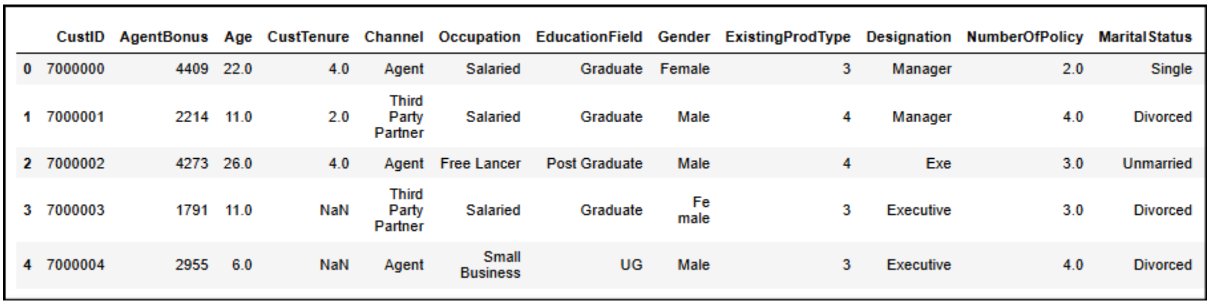
Data dictionary

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

Total rows in the dataset: 4520

Total columns in the dataset: 20

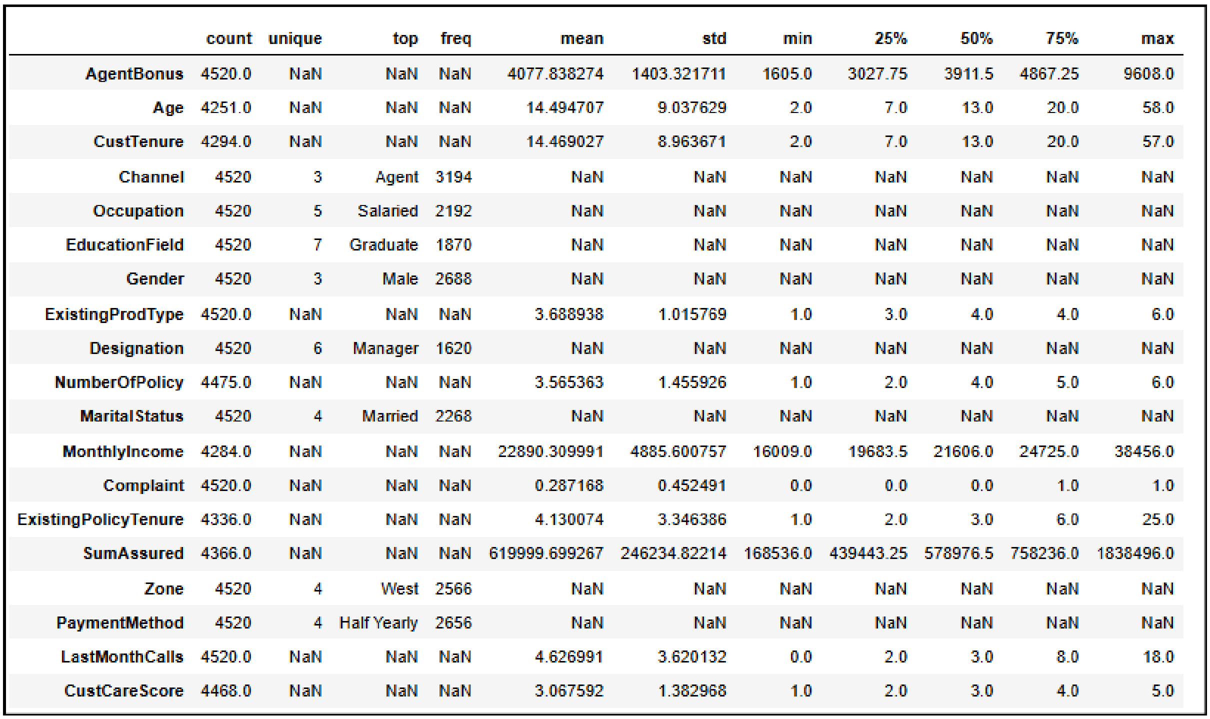
Head of the data



	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus
0	7000000	4409	22.0	4.0	Agent	Salaried	Graduate	Female	3	Manager	2.0	Single
1	7000001	2214	11.0	2.0	Third Party Partner	Salaried	Graduate	Male	4	Manager	4.0	Divorced
2	7000002	4273	26.0	4.0	Agent	Free Lancer	Post Graduate	Male	4	Exe	3.0	Unmarried
3	7000003	1791	11.0	NaN	Third Party Partner	Salaried	Graduate	Female	3	Executive	3.0	Divorced
4	7000004	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced

Table 1 - Head of the data

Descriptive stats of data

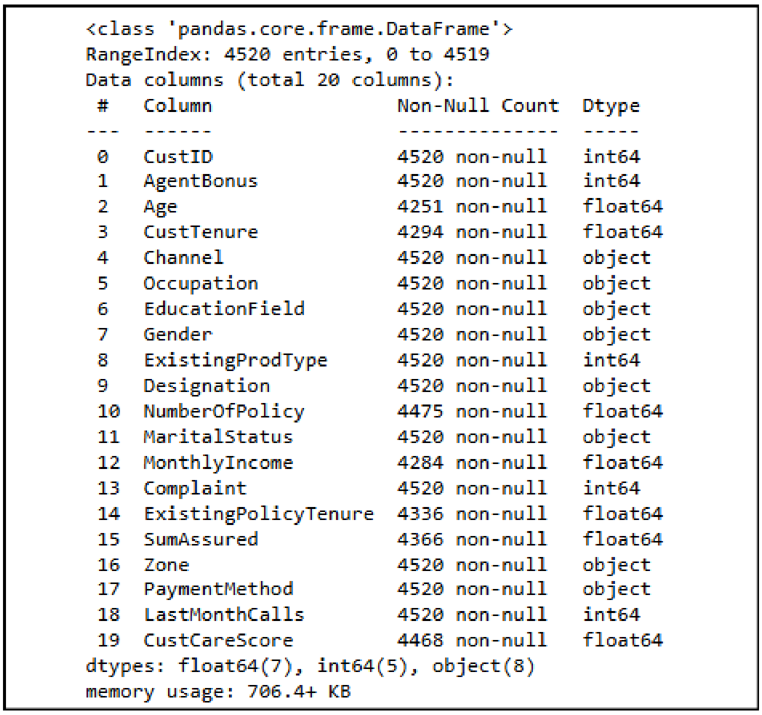


	count	unique	top	freq	mean	std	min	25%	50%	75%	max
AgentBonus	4520.0	NaN	NaN	NaN	4077.838274	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	NaN	NaN	NaN	14.494707	9.037629	2.0	7.0	13.0	20.0	58.0
CustTenure	4294.0	NaN	NaN	NaN	14.469027	8.963671	2.0	7.0	13.0	20.0	57.0
Channel	4520	3	Agent	3194	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	4520.0	NaN	NaN	NaN	3.688938	1.015769	1.0	3.0	4.0	4.0	6.0
Designation	4520	6	Manager	1620	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NumberOfPolicy	4475.0	NaN	NaN	NaN	3.565363	1.455926	1.0	2.0	4.0	5.0	6.0
MaritalStatus	4520	4	Married	2268	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthlyIncome	4284.0	NaN	NaN	NaN	22890.309991	4885.600757	16009.0	19683.5	21606.0	24725.0	38456.0
Complaint	4520.0	NaN	NaN	NaN	0.287168	0.452491	0.0	0.0	0.0	1.0	1.0
ExistingPolicyTenure	4336.0	NaN	NaN	NaN	4.130074	3.346386	1.0	2.0	3.0	6.0	25.0
SumAssured	4366.0	NaN	NaN	NaN	619999.699267	246234.82214	168536.0	439443.25	578976.5	758236.0	1838496.0
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	4520.0	NaN	NaN	NaN	4.626991	3.620132	0.0	2.0	3.0	8.0	18.0
CustCareScore	4468.0	NaN	NaN	NaN	3.067592	1.382968	1.0	2.0	3.0	4.0	5.0

Table 2 - Description of the data

The minimum age of the customer cannot be 2. Additionally, there are missing values in the data, which need to be addressed through data cleaning before performing exploratory data analysis (EDA).

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



```
<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
```

Table 3 - Info of data

We have 7 parameters having ‘float’ data type.

We have 5 parameters having ‘integer’ data type.

We have 8 parameters having ‘object’ data type

Given the range and distribution (minimum, maximum, and quartiles) of the data, three variables should be categorical instead of float/int (numerical). These variables are:

- CustCareScore
- ExistingProdType
- Complaint

3) Exploratory data analysis

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

Data cleaning and preprocessing are integral steps in the exploratory data analysis (EDA) process

The Analysis below is after data preprocessing.

70% insurance which is the majority is bought via agent

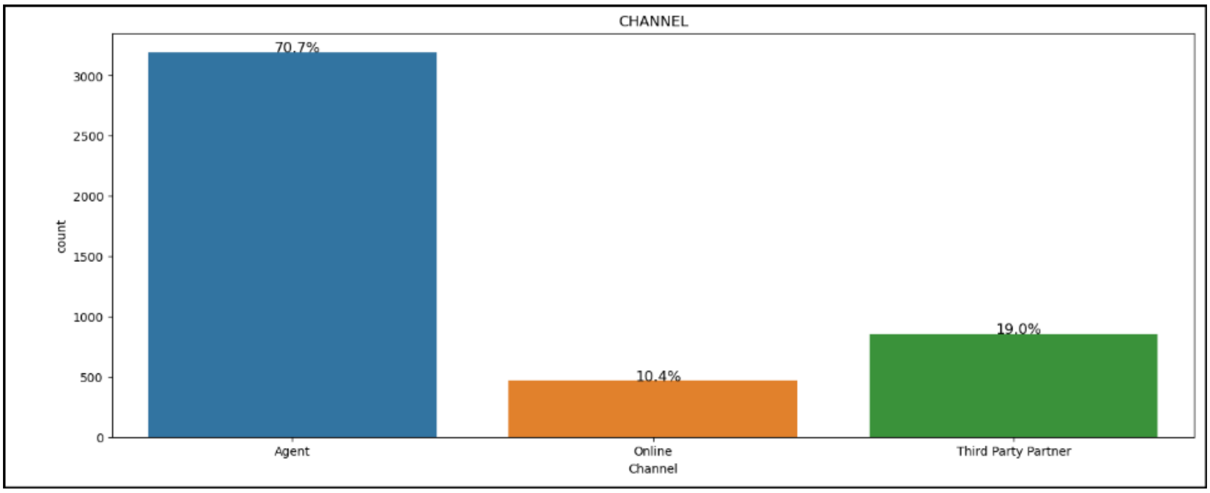


Figure 1 - Insurance channel

Salaried has the has the highest number of insurance

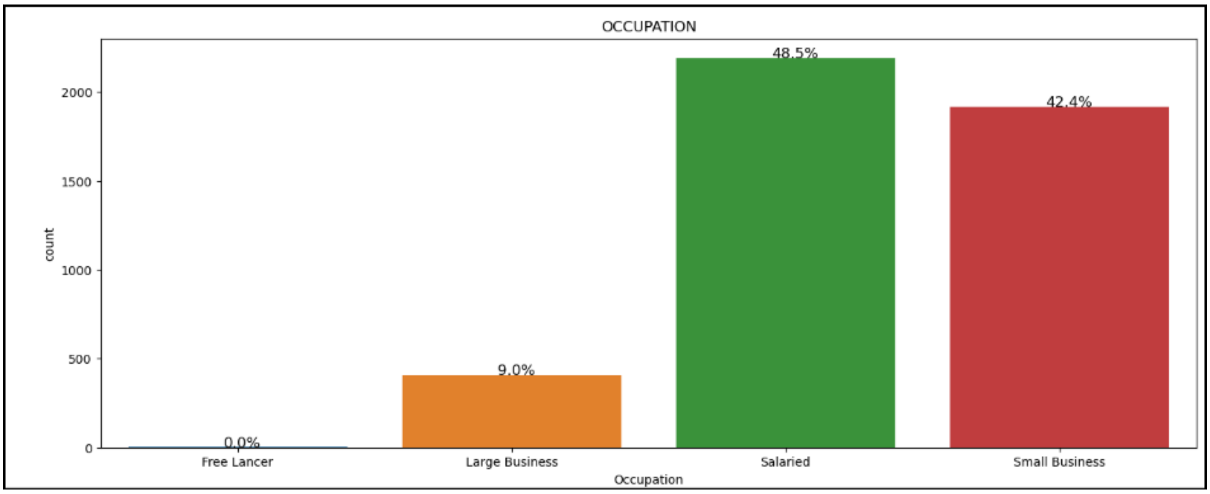


Figure 2 - Occupation

UG is the highest with a whopping 81%

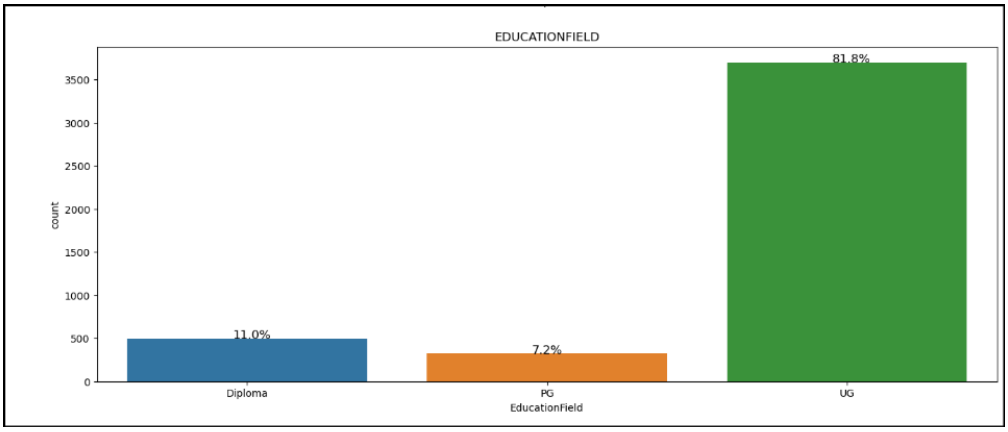


Figure 3 - Education Field

Males seem to have more insurance over females

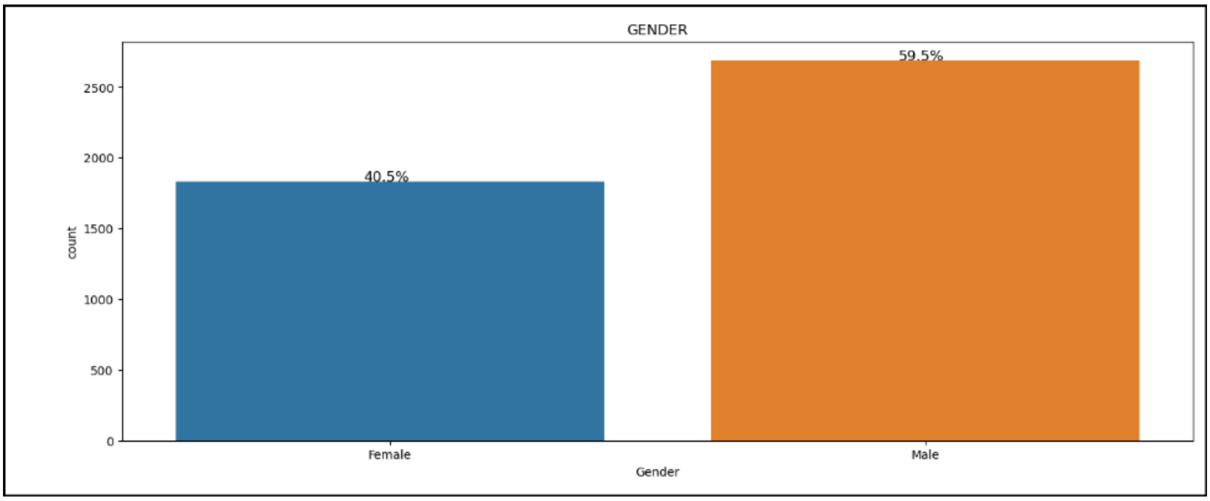


Figure 4 - Gender distribution

Prod type 4 is the most common type

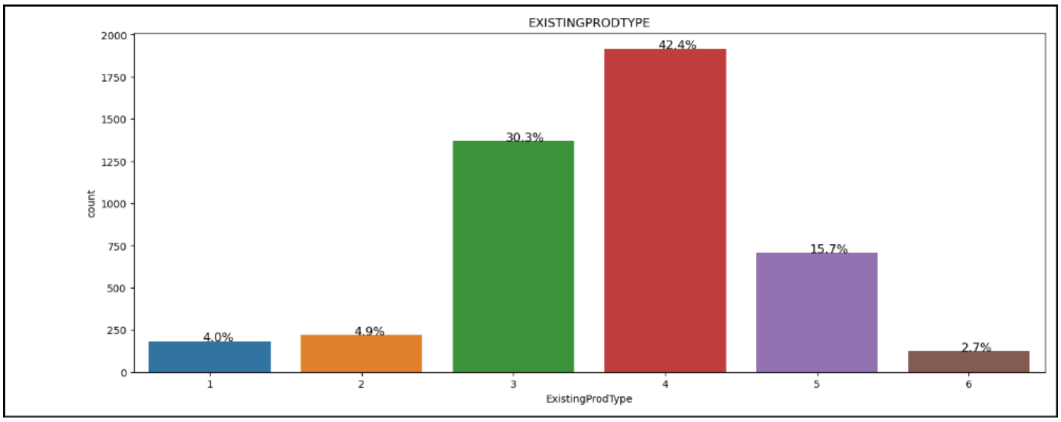


Figure 5 - Prod type

Executive and Managers have more insurance compared to others

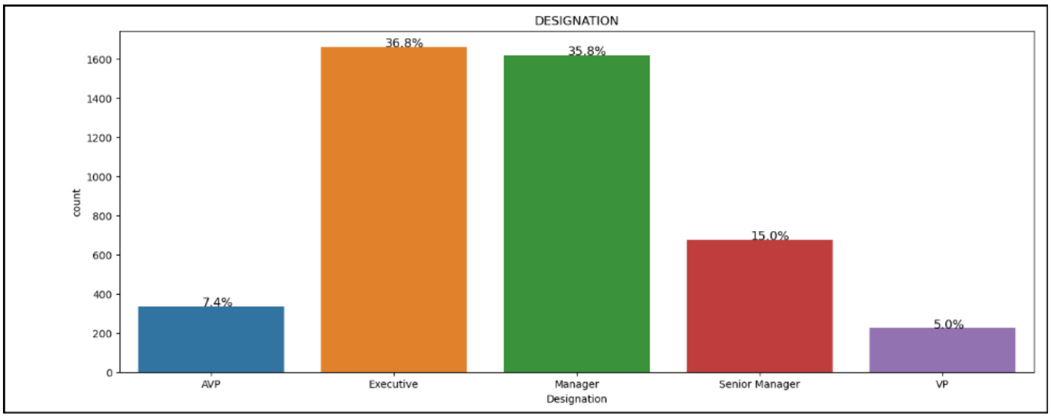


Figure 6 - Exec and Managers

Married men tend to go for insurance other than single

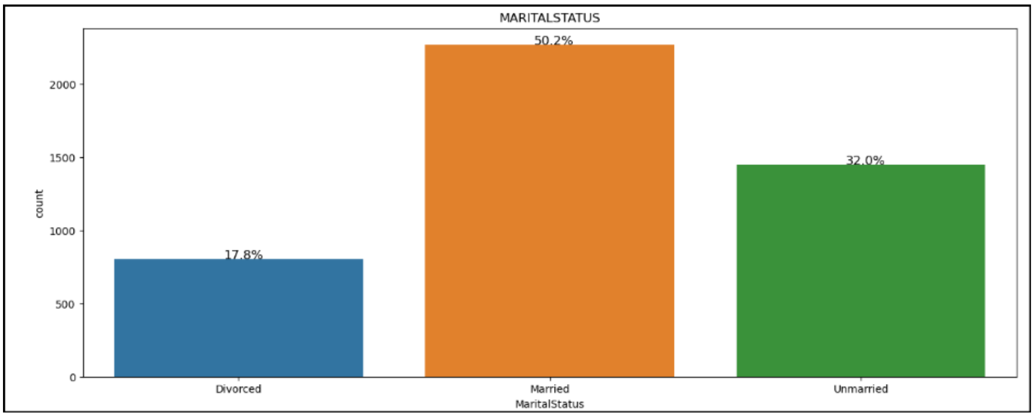


Figure 7 - Marital Status

West zone is the highest followed by north

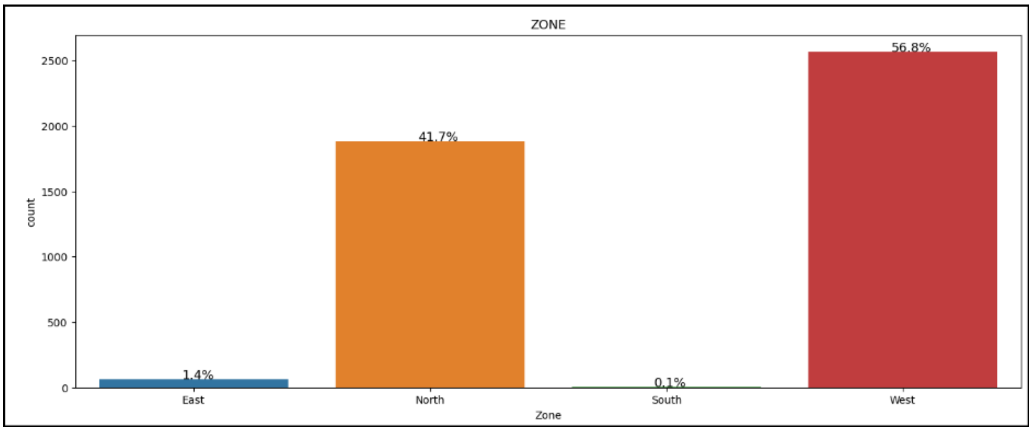


Figure 8 - Zonewise data

Half yearly is the most preferred mode of payment.

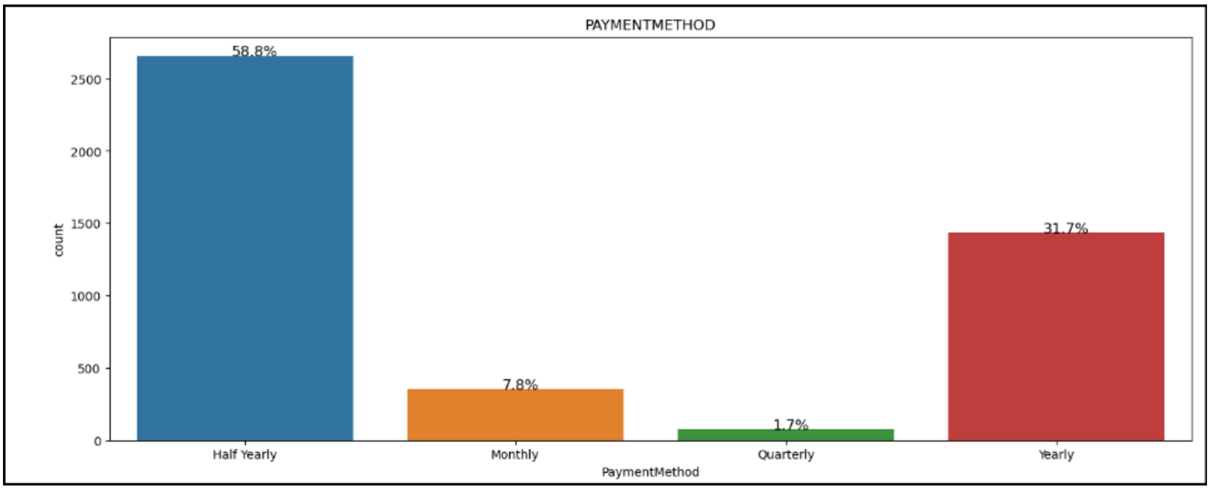


Figure 9 - Mode of Payment

Distribution of variables

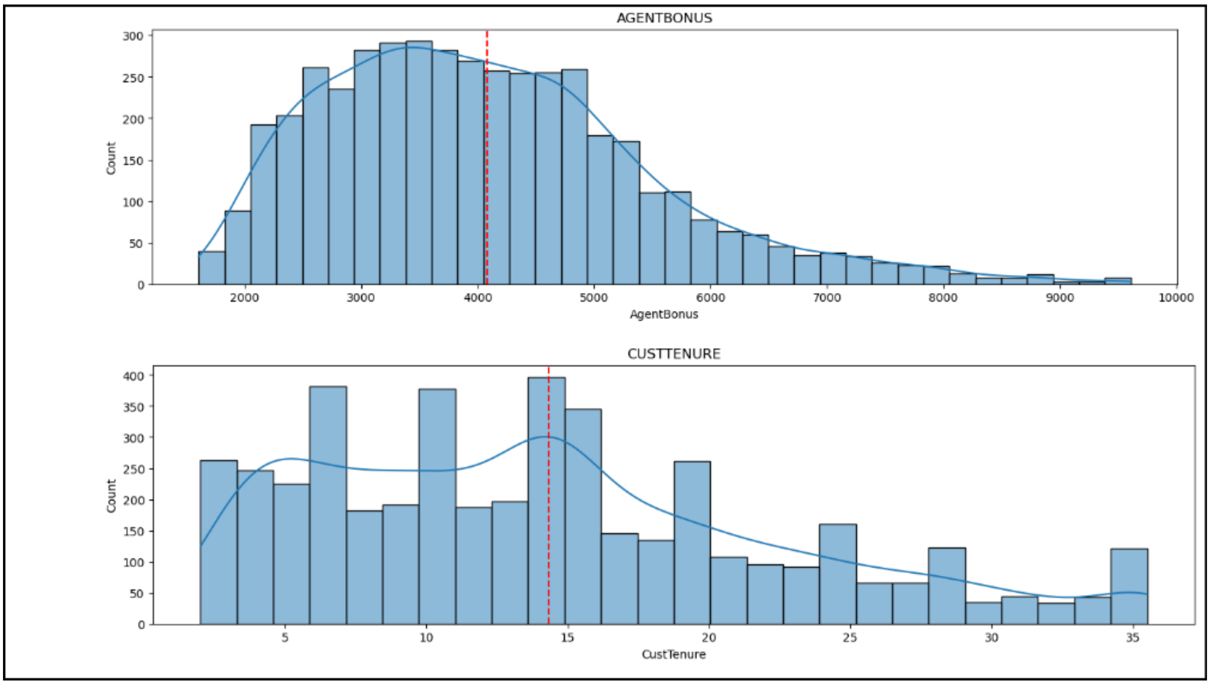


Figure 10 - Histogram 1

Agent bonus is postively skewed

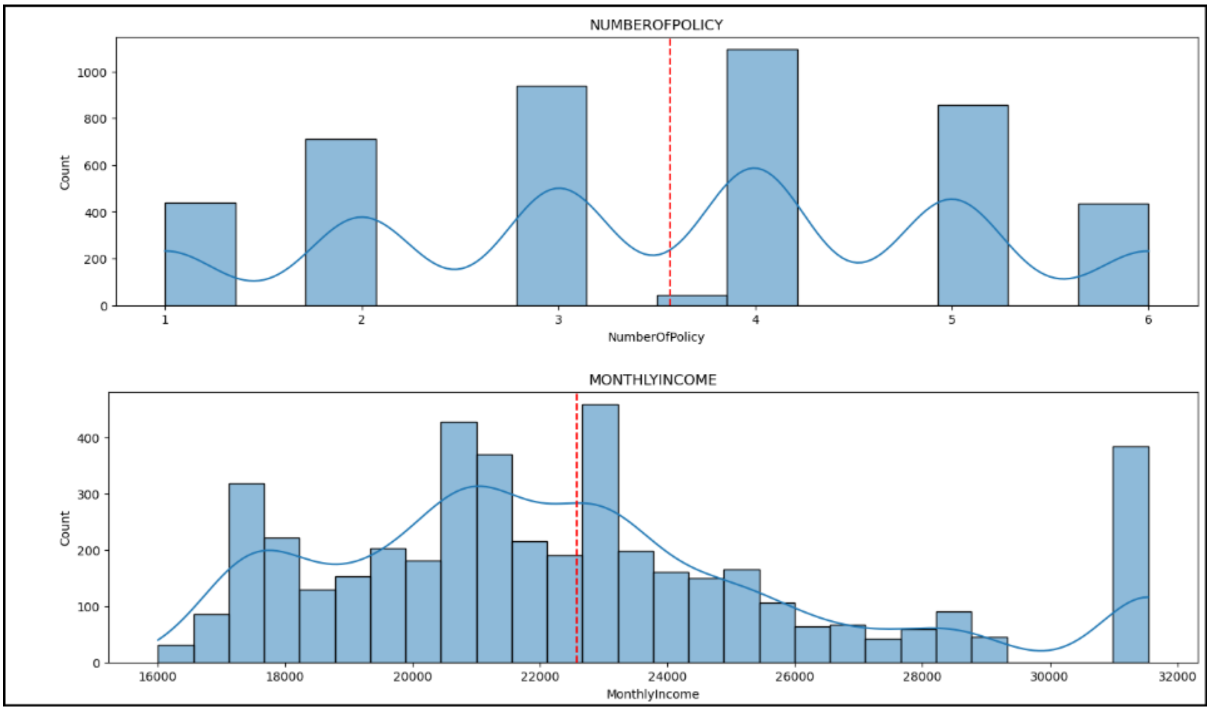


Figure 11 - Histogram 2

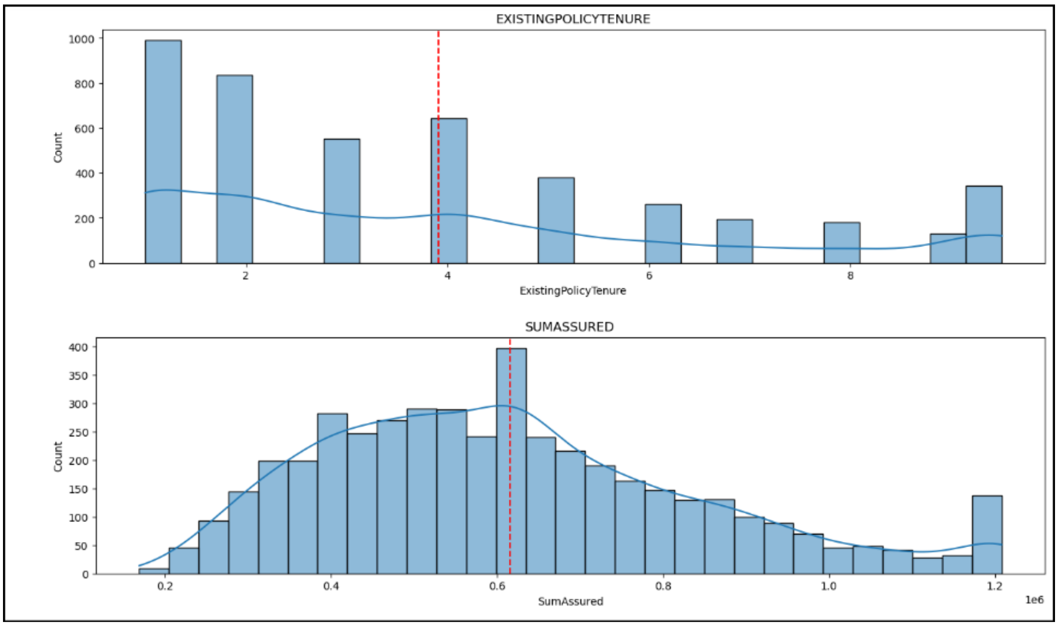


Figure 12 - Histogram 3

Sum assured in positively skewed.

b) Bivariate analysis (relationship between different variables , correlations)

The average Sum Assured is fairly consistent across all categories except for designation. The Sum Assured varies significantly with different designations.

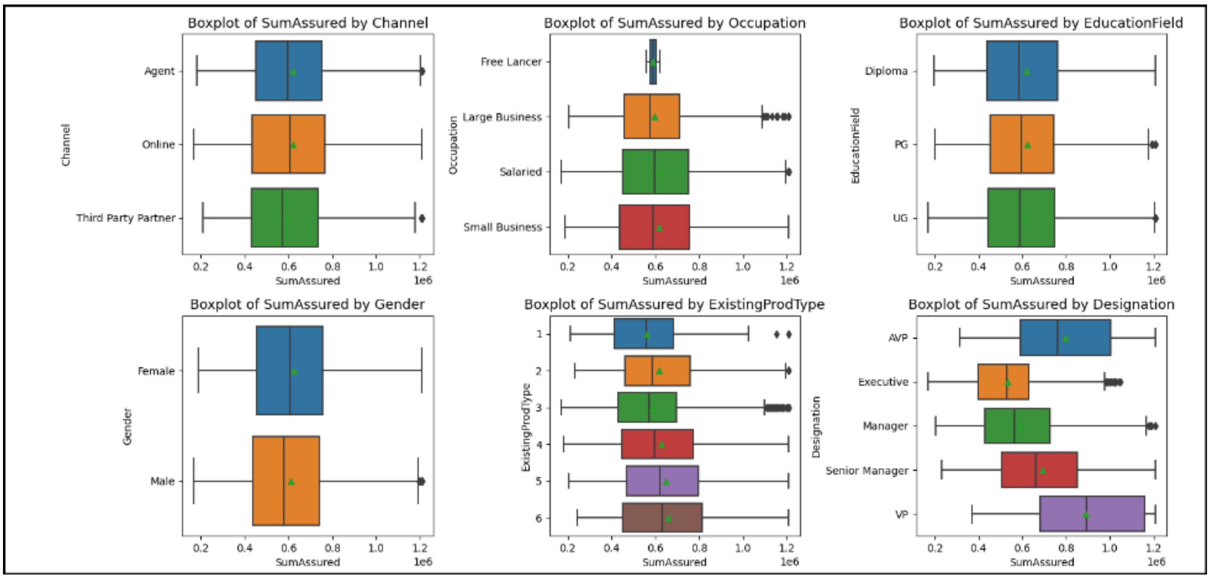


Figure 13 - Box plot bivariate

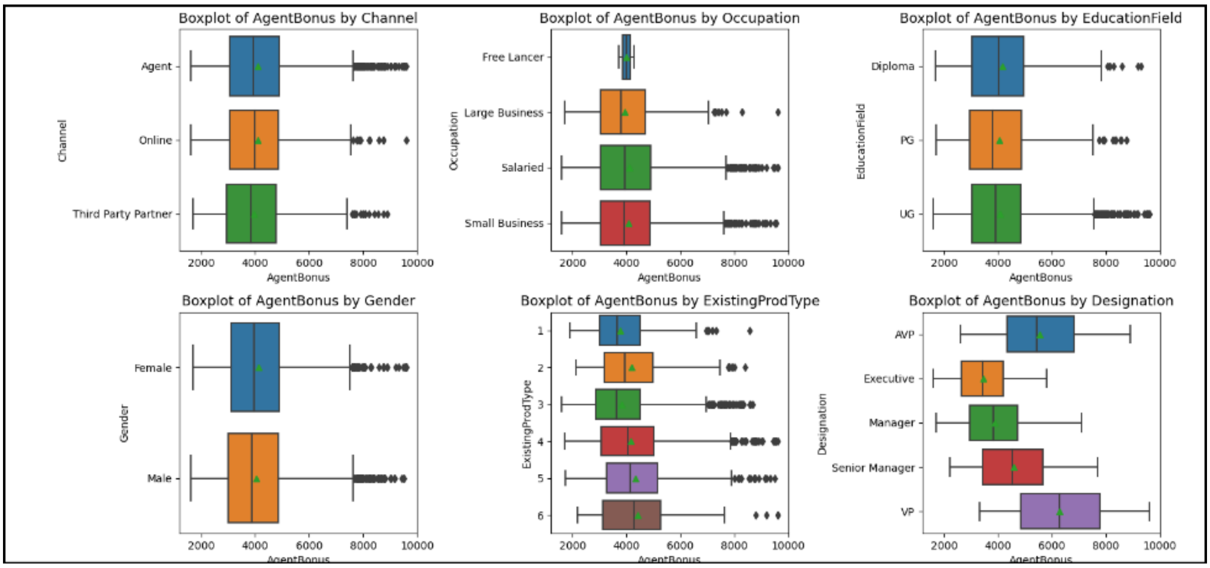


Figure 14 - Box blot 2 - Bivariate

We can see positive correlation amongst all variables

High correlation between agent bonus and sum assured

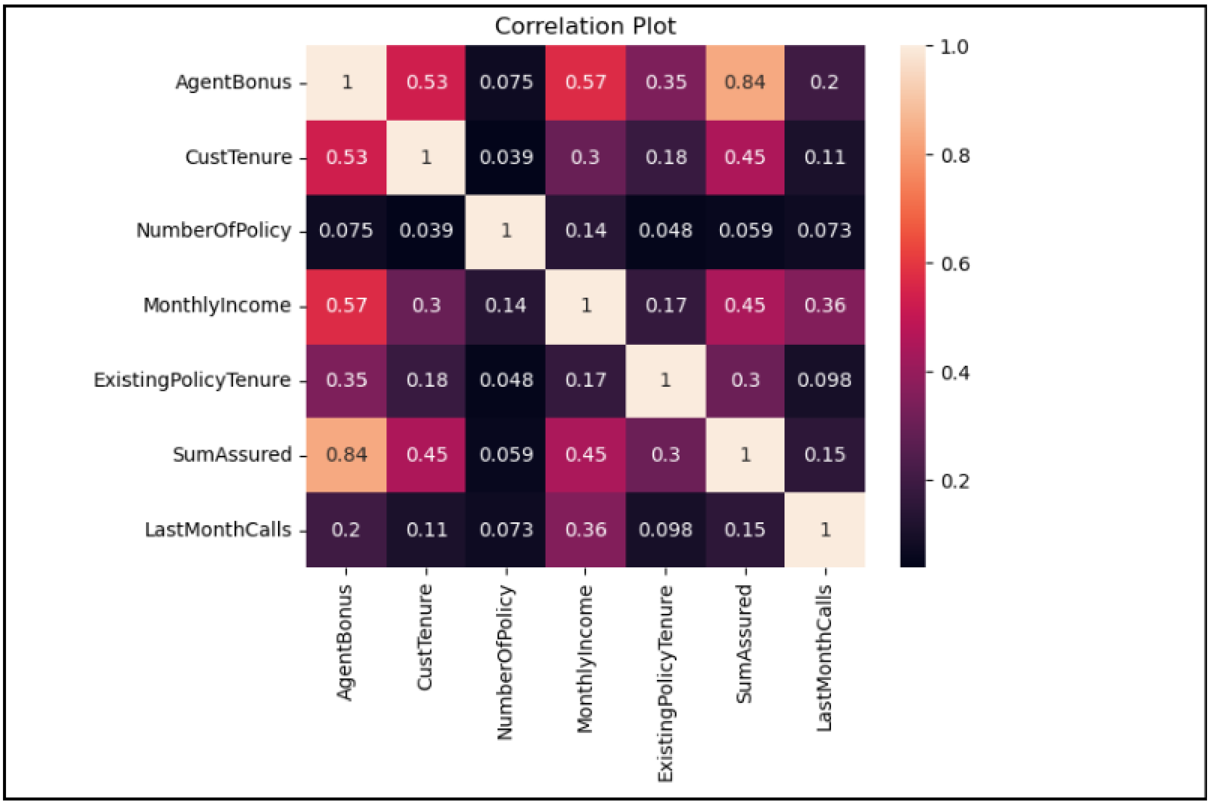


Figure 15 - Correlation Plot

As the sum assured increases bonus also increases

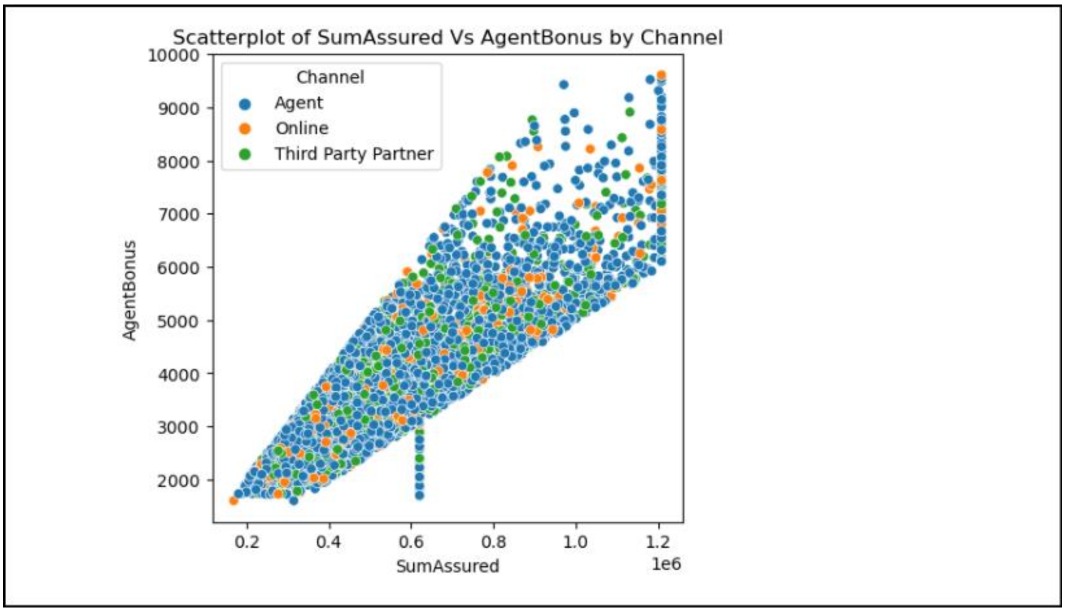


Figure 16 - Scatterplot of Sum assured vs Agent bonus

Same pattern is found for all multi variables below

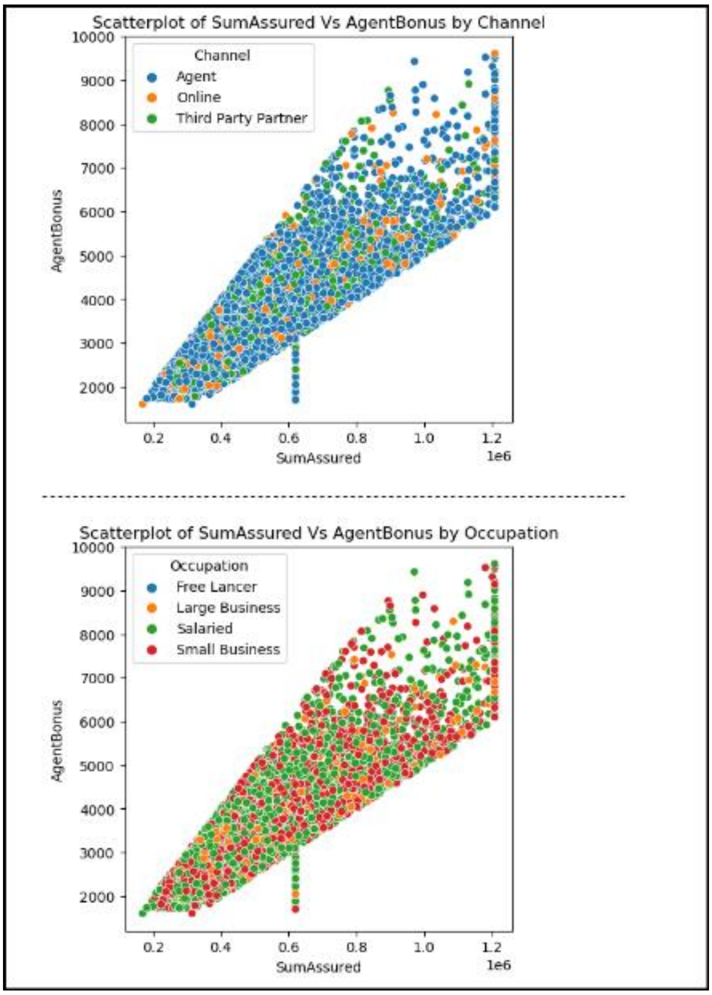


Figure 17 - Scatter plot for Multivariate analysis

Pairplot for all variables

A pair plot visualizes the relationships between all variables in a dataset. histograms for each variable, showing their distributions. From the plot, we can observe the relationships between each pair of variables.

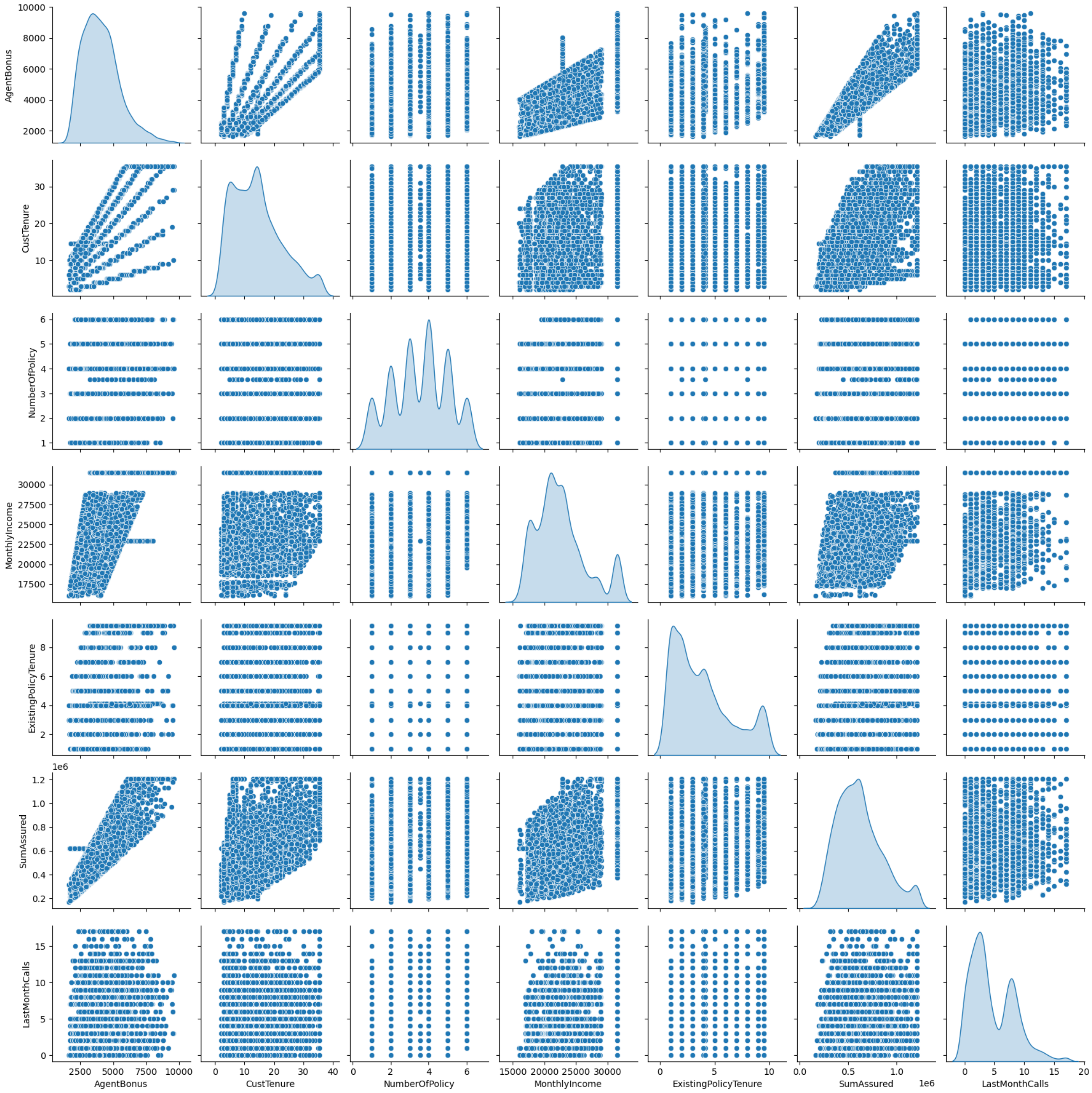
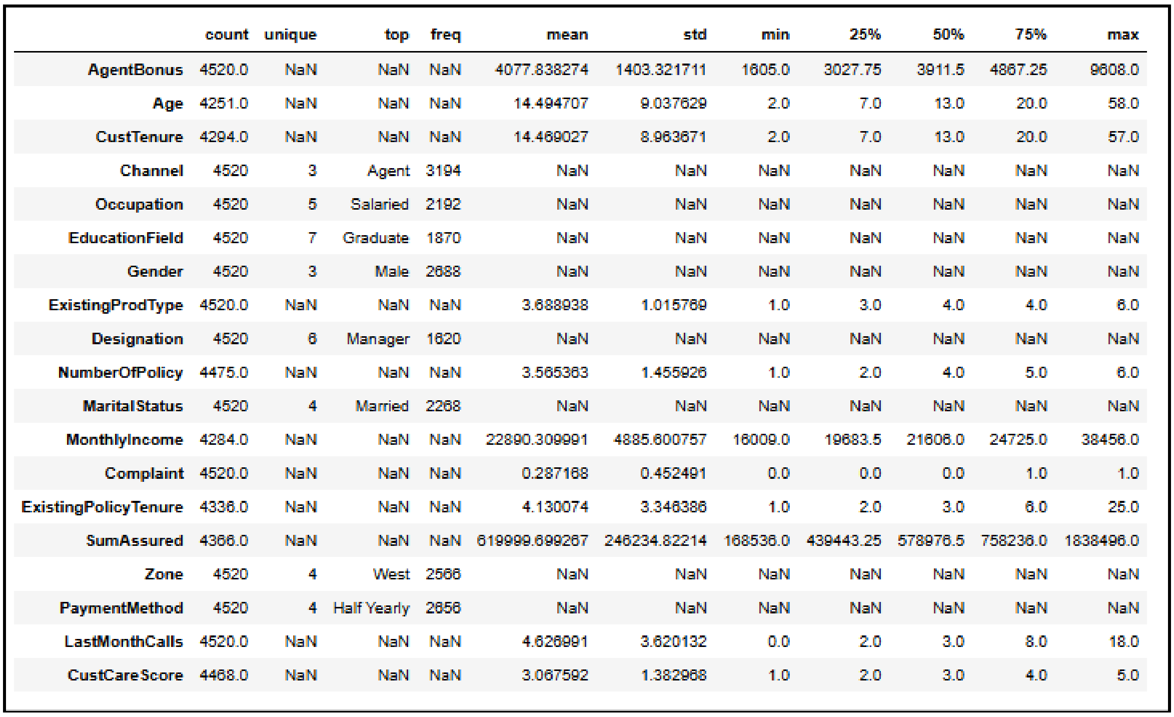


Figure 18 - Pair plot for all variables

c) Removal of unwanted variables (if applicable)

The CustID variable has been removed as it has no meaning.

DATAFRAME after dropping custID



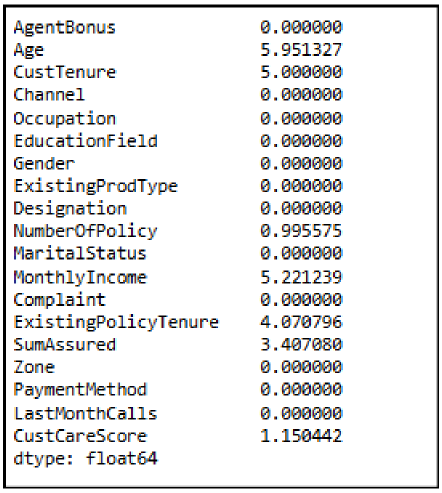
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
AgentBonus	4520.0	NaN	NaN	NaN	4077.838274	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	NaN	NaN	NaN	14.494707	9.037629	2.0	7.0	13.0	20.0	58.0
CustTenure	4294.0	NaN	NaN	NaN	14.469027	8.963671	2.0	7.0	13.0	20.0	57.0
Channel	4520	3	Agent	3194	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	4520.0	NaN	NaN	NaN	3.688938	1.015769	1.0	3.0	4.0	4.0	6.0
Designation	4520	6	Manager	1620	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NumberOfPolicy	4475.0	NaN	NaN	NaN	3.565363	1.455926	1.0	2.0	4.0	5.0	6.0
MaritalStatus	4520	4	Married	2268	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthlyIncome	4284.0	NaN	NaN	NaN	22890.309991	4885.600757	16009.0	19683.5	21606.0	24725.0	38456.0
Complaint	4520.0	NaN	NaN	NaN	0.287168	0.452491	0.0	0.0	0.0	1.0	1.0
ExistingPolicyTenure	4336.0	NaN	NaN	NaN	4.130074	3.346386	1.0	2.0	3.0	6.0	25.0
SumAssured	4366.0	NaN	NaN	NaN	619999.699267	246234.82214	168536.0	439443.25	578976.5	758236.0	1838496.0
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	4520.0	NaN	NaN	NaN	4.626991	3.620132	0.0	2.0	3.0	8.0	18.0
CustCareScore	4468.0	NaN	NaN	NaN	3.067592	1.382968	1.0	2.0	3.0	4.0	5.0

Table 4 - Data after removal of Cust ID

The mean age of customers, when compared to their designation, doesn't provide any meaningful insights. Based on descriptive statistics and boxplot analysis, the Age column contains some invalid data

The Age column has been dropped as it doesn't add significant value to the model or its inferences.

d) Missing Value treatment (if applicable)



```
AgentBonus              0.000000
Age                     5.951327
CustTenure              5.000000
Channel                 0.000000
Occupation              0.000000
EducationField          0.000000
Gender                  0.000000
ExistingProdType        0.000000
Designation             0.000000
NumberOfPolicy          0.995575
MaritalStatus           0.000000
MonthlyIncome           5.221239
Complaint               0.000000
ExistingPolicyTenure    4.070796
SumAssured              3.407080
Zone                    0.000000
PaymentMethod           0.000000
LastMonthCalls          0.000000
CustCareScore           1.150442
dtype: float64
```

Table 5 - Missing value

The are multiple missing values in the dataset

CustCareScore: 52 missing values.

The null values are imputed with the mode for this categorical variable.

The mode for CustCareScore is 3.0.

e) Outlier treatment (if required)

As we can see there are multiple outliers as per the below boxplot.

We will have to treat them.

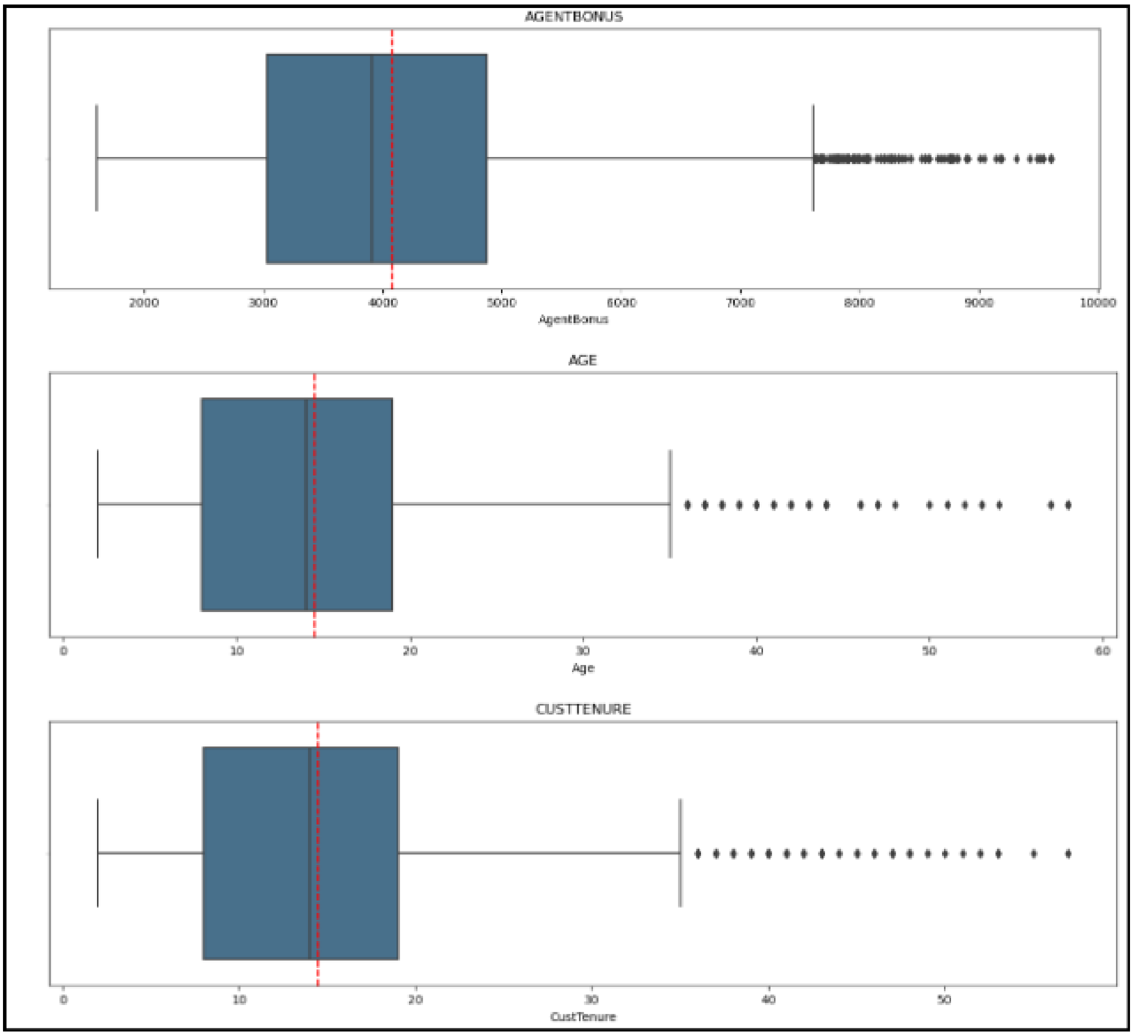


Figure 19 - Before outlier treatment 1

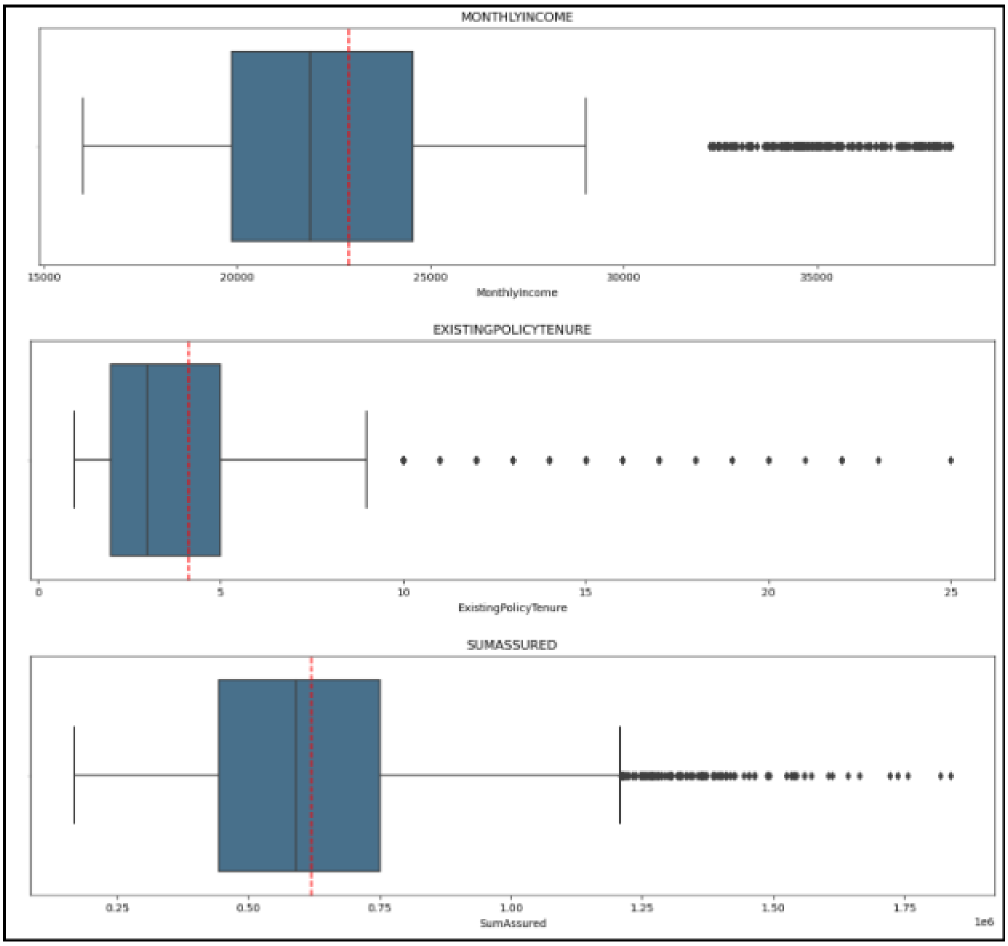


Figure 20 - Before outlier treatment 2

Post outlier treatment

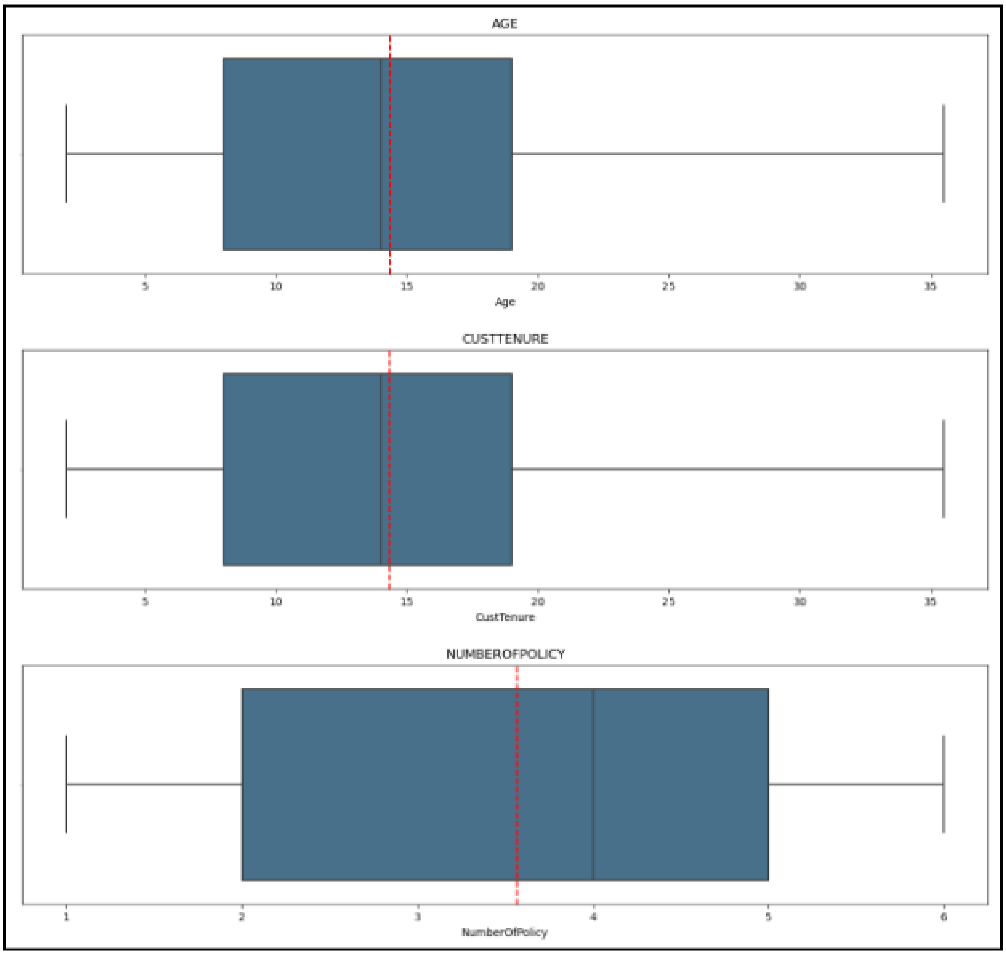


Figure 21 - Post outlier treatment 1

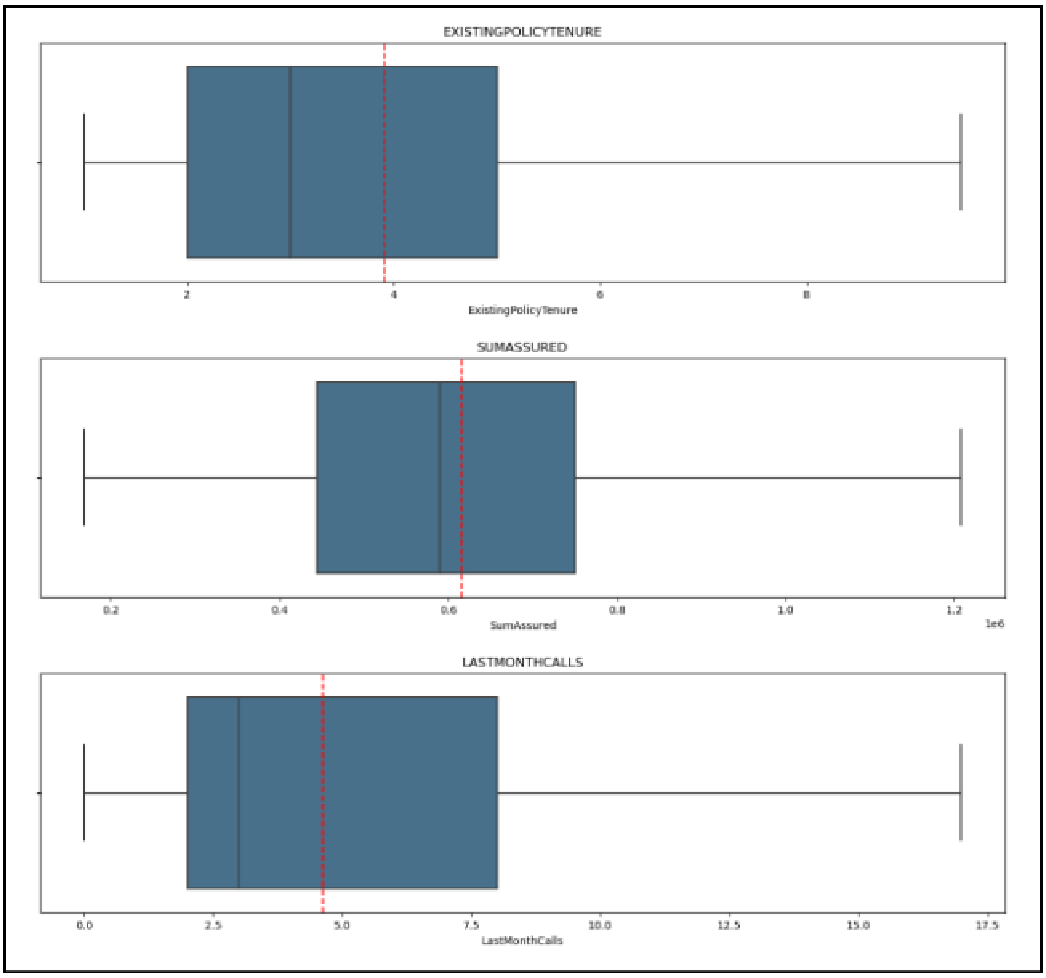
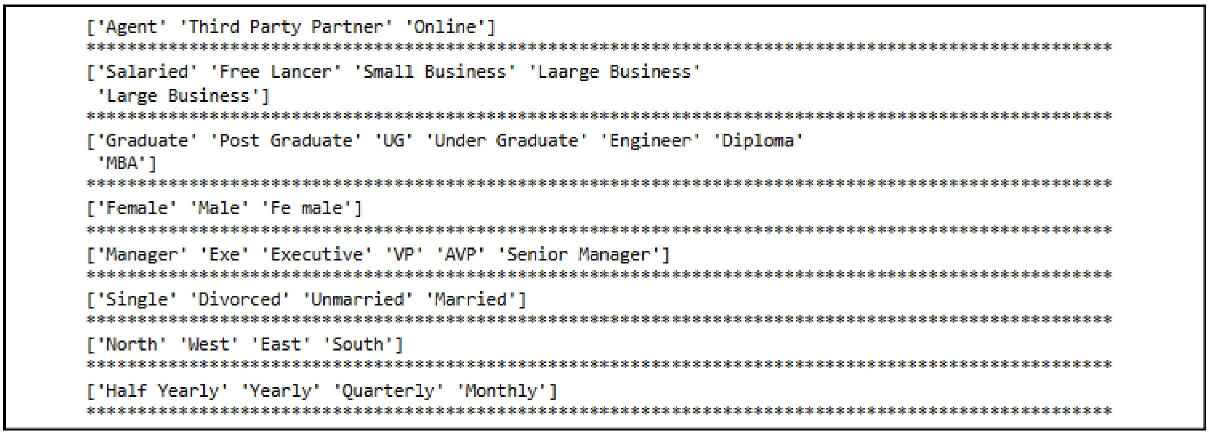


Figure 22 - Post outlier treatment 2

f) Variable transformation (if applicable)

Some categories in the object columns need to be cleaned to correct spelling mistakes and incorrect classifications.

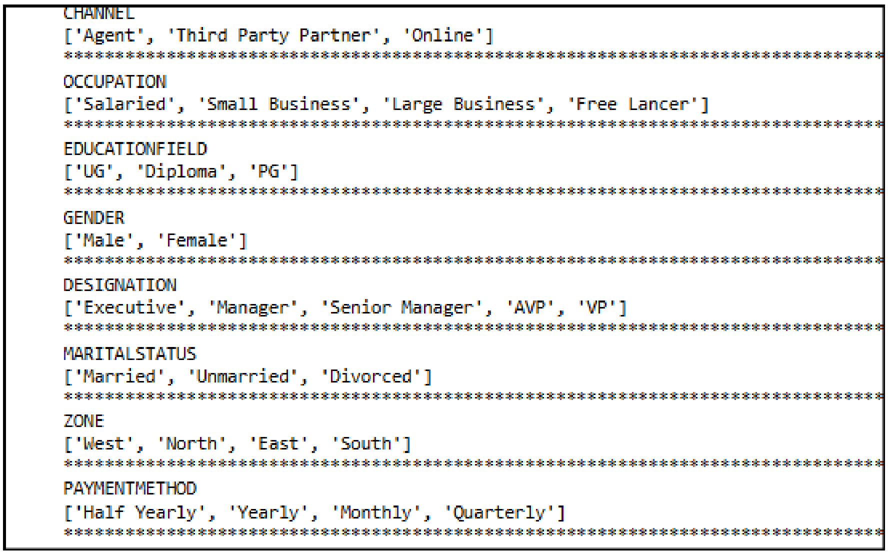


```
['Agent' 'Third Party Partner' 'Online']
*******************************************************************************************************
['Salaried' 'Free Lancer' 'Small Business' 'Laarge Business'
 'Large Business']
*******************************************************************************************************
['Graduate' 'Post Graduate' 'UG' 'Under Graduate' 'Engineer' 'Diploma'
 'MBA']
*******************************************************************************************************
['Female' 'Male' 'Fe male']
*******************************************************************************************************
['Manager' 'Exe' 'Executive' 'VP' 'AVP' 'Senior Manager']
*******************************************************************************************************
['Single' 'Divorced' 'Unmarried' 'Married']
*******************************************************************************************************
['North' 'West' 'East' 'South']
*******************************************************************************************************
['Half Yearly' 'Yearly' 'Quarterly' 'Monthly']
*******************************************************************************************************
```

Table 6 - Data before cleaning

- In the Occupation column, "Laarge Business" should be corrected to "Large Business".
- In the Gender column, "Fe male" should be corrected to "Female".
- In the Designation column, "Exe" and "Executive" should be combined into a single category, "Executive".
- In the Marital Status column, "Single" and "Unmarried" should be combined into the "Unmarried" category.
- In the Educationfield column:
- "Graduate," "Under Graduate," and "UG" should be combined into a single category, "UG".
- "Engineer" can be considered as "UG" since it is not a high-level education field category. Only 9% (refer to the table below*) are engineers. If engineers have completed a master's degree, they would have selected "Post Graduate" or "MBA".
- "Post Graduate" and "MBA" can be combined into a single category, "PG".

Cleaned data



```
CHANNEL
['Agent', 'Third Party Partner', 'Online']
********************************************************************************
OCCUPATION
['Salaried', 'Small Business', 'Large Business', 'Free Lancer']
********************************************************************************
EDUCATIONFIELD
['UG', 'Diploma', 'PG']
********************************************************************************
GENDER
['Male', 'Female']
********************************************************************************
DESIGNATION
['Executive', 'Manager', 'Senior Manager', 'AVP', 'VP']
********************************************************************************
MARITALSTATUS
['Married', 'Unmarried', 'Divorced']
********************************************************************************
ZONE
['West', 'North', 'East', 'South']
********************************************************************************
PAYMENTMETHOD
['Half Yearly', 'Yearly', 'Monthly', 'Quarterly']
********************************************************************************
```

Table 7 - Data after cleaning

g) Addition of new variables (if required)

Introducing new variables such as Premium is possible, but doing so may impact the model's performance and is therefore not recommended.

4) Business insights from EDA

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

The data is unbalanced. For example, the Zone category shows South has less representation, and the Occupation category shows Freelancers are underrepresented. To address this, more data is needed, or the existing data should be upsampled.

In a business context, this imbalance can lead to biased models that perform well for the majority class but poorly for the minority class. To address this:

Resampling Techniques: Use oversampling (like SMOTE) to increase the minority class samples or undersampling to reduce the majority class samples.

Class Weights: Assign higher weights to the minority class in the model to penalize misclassifications more heavily.

Data Augmentation: Create synthetic data for the minority class to balance the dataset.

Ensemble Methods: Use techniques like boosting or bagging that are robust to imbalanced data.

Balancing the data helps ensure that the model performs reliably across all classes, leading to more equitable business decisions and better resource allocation.

b) Any business insights using clustering (if applicable)

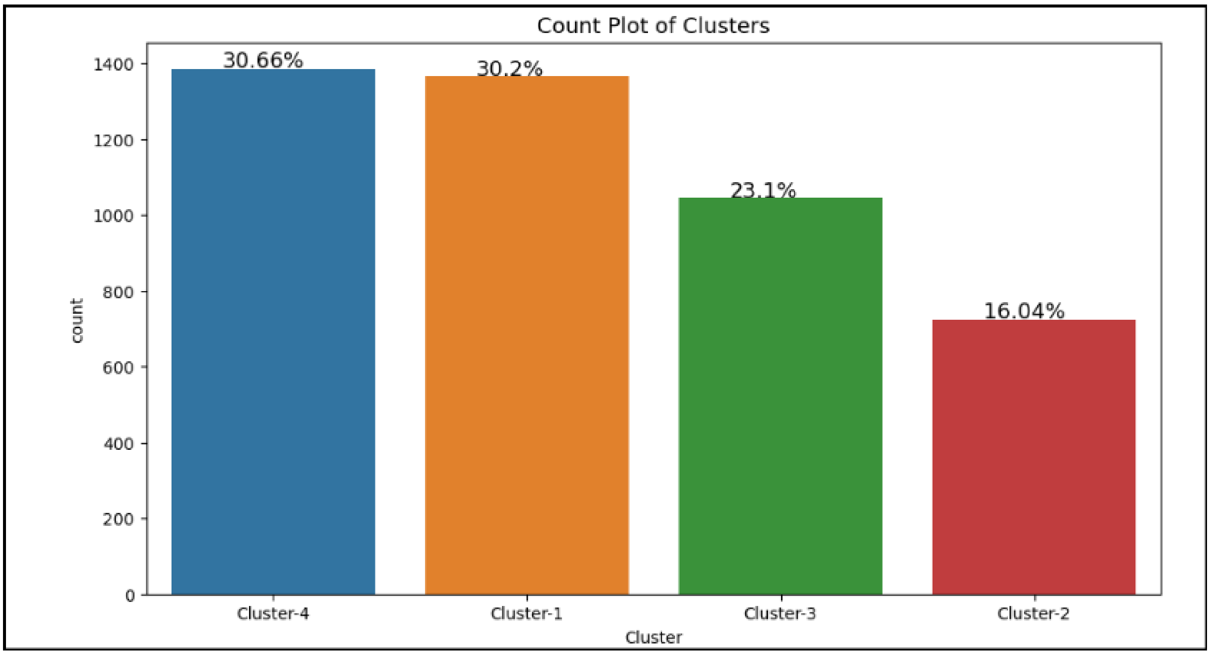
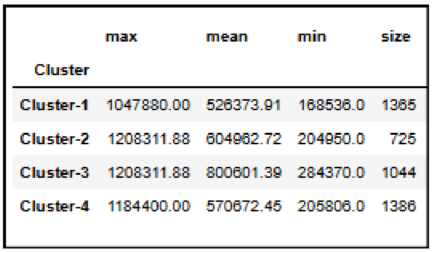


Figure 23 - Cluster

We have identified 4 clusters

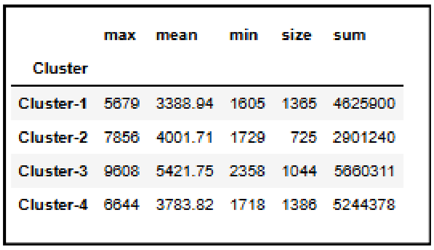
Cluster of sum assured



Cluster	max	mean	min	size
Cluster-1	1047880.00	526373.91	168536.0	1365
Cluster-2	1208311.88	604962.72	204950.0	725
Cluster-3	1208311.88	800601.39	284370.0	1044
Cluster-4	1184400.00	570672.45	205806.0	1386

Figure 24 - Cluster of Sum assured

Cluster of agent bonus



Cluster	max	mean	min	size	sum
Cluster-1	5679	3388.94	1605	1365	4625900
Cluster-2	7856	4001.71	1729	725	2901240
Cluster-3	9608	5421.75	2358	1044	5660311
Cluster-4	6644	3783.82	1718	1386	5244378

Figure 25 - Cluster of Agent bonus

Based on the 4 clusters, Cluster 4 has the highest number of customers followed by cluster 1. This cluster can be targeted for selling top-ups to increase their sum assured.

Cluster 2 should be focused on for top-ups and cross-selling other products to move them to the next cluster.

c) Any other business insights

- The AgentBonus ranges from 1,400 to 9,608.
- Customer work experience varies from 2 to 35 years.
- Most customers are salaried and have an undergraduate degree.
- The most common payment method is Half Yearly payments.
- The West region has the highest number of insurance policies.
- Most customers come through agents
- The most purchased product type is 4.