Final Report of Capstone Project on Life Insurance Sales

Ujjwal Kumar

Post Graduation – DSBA (February, 2023)

Dated : 25th Feb, 2023

Table of Contents

1. Business Problem Summary

The given data set belongs to Life Insurance Sales data of a leading Life Insurance company which contains the attributes of life insurance sales and other related attributes of customer as well as sales details whereby the data consist of total 20 columns and 4520 rows and each rows specify the details of claims made by customers. There are total 20 columns and each column had the details of Life Insurance Sales such as customer ID, Age of Customer, Customer Tenure with the organization. The channel or the medium through which the customer has enrolled themselves or they have been acquired, Occupations, education level, gender of the customer, Existing product type which the customer has opted for, Designation of Customer, Number of policy, Marital Status of customer, Monthly income of customer, Complain indicator, Remaining tenure in existing policy, Sum assured, Geographical location like zone of Customer, Frequency of payment, Calls placed for next sales, Customer satisfaction score and Agent Bonus (Target variable).

As an analyst we have been assigned for the role of 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. As we have been provided with the 'Life insurance sales' data set of the leading insurance company, we need to deeply dive in the given data set and extract meaningful insights which in turn help the company to determine the bonus strategies for high performing and low performing agents.

2. Business Problem Definition

The purpose of this whole exercise is to predict the Agent Commission and group them according to their performances and to perform the above task we initially need to load the data and perform data cleansing such as Null value treatment, Anomalies treatment and Bad value treatment. Once we are done with it, we need to perform Outlier treatment if it exists. After that we need to perform Exploratory Data Analysis where we will be going through the Univariate, Bivariate and Multi-variate analysis from which we'll be taking the insights of the data. Later on we proceed to perform the process of Clustering of data into groups which is also called a process of segregating the performance of agents. Finally, we need to build the models and fine tune the parameters of the models to increase the performance of the models that are built. After creating all the possible models we need to evaluate the performance of all models such as Mean Square Error (MSE) and Root Mean Square Error (RMSE) from which we need to draw insights based on the model performance and select the best model and predict the Agent Bonus.

3. Need of the Study / Project

Insurance sector is highly data-driven industry. Every day a new company is formed and thus the competition is increasing exponentially. In order to stay ahead of the curve, around 86% companies are investing in insurance data analytics to optimize their mechanisms. It can be observed that the probability of the insurance companies achieving their long-term goals increases significantly by unleashing the power of the data that is collected over the years.

Here, in this problem statement 'Agent bonus' is the one of the key parameters which drives the enthusiasm in the agents to perform better. It not only rewards agents but also helps in retaining them for the longer period of time in the company. We need to predict the agents commission and to find the groups of Agents performance so that we can concentrate on low performing agents and upskill them to compete and to boost the high performing agents by appropriate engagement activities. So to do so we need to group the data using clustering and build the appropriate model, find the important feature and guide organization which is something we need to concentrate on.

4. Understanding Business / Social Opportunity

Revenues of Insurance companies depends mainly upon amount of premium received and amount spent in claim settlements. In order to maximize premium, the companies hire agents and offer them lucrative bonuses based on their performances. Social initiative for the product is covering as many lives possible under the ambit of life insurance policies. Following are the business opportunities that can be obtained by the data analysis in the insurance sectors.

• **Improving Employee Performance and Satisfaction**: By analysing the data about the employee performance, they can be rewarded with bonuses which will increase the employee satisfaction.

• **Improving Customer satisfaction:** By analysing the perspective customer data, the companies can predict the needs of the customers and thus increase the potential to make a sale when compared to a company following the conventional methods of selling. The existing customer data can be used to find the insights and thus improve customer satisfaction.

• **Lead Generation:** By analysing the data on the internet, the companies can deep dive into the customer behaviour and up-sell or cross-sell opportunities in the market.

• **Risk Analysis and Fraud detection:** By storing the previous fraudulent customer data and doing a predictive analysis on the new claim to calculate the risk of percentage, frauds can be prevented. This data can also be used to recognizes if any patterns or trends exists when a new insurance claim is made thus avoid risks and loses.

4.1 Data Collection:

The given data set is relating to a leading Insurance Organization with which we are expected to predict the agent commission of the Insurance Company. As per the given data dictionary below listed are the details of the columns

Serial No	Variables	Description
1	Cust ID	Unique customer ID
2	Agent Bonus	Bonus amount given to each agents in last month
3	Age	Age of Customer
4	Cust Tenure	Tenure of customer in organization
5	Channel	Channel through which acquisition of customer is done
6	Occupation	Occupation of customer
7	Education Field	Field of education of customer
8	Gender	Gender of Customer
9	Existing Prod Type	Existing product type of customer
10	Designation	Designation of customer in their organization
11	Number of Policy	Total number of existing policy of a customer
12	Marital Status	Marital Status of Customer
13	Monthly Income	Gross monthly income of customer
14	Complaint	Indicator of complaint registered in last one month by customer
15	Existing Policy Tenure	Max tenure in all existing policies of customer
16	Sum Assured	Max of sum assured in all existing policies of customer
17	Zone	Customer belongs to which zone in India like East, West, North and South
18	Payment Method	Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
19	Last Month Calls	Total calls attempted by company to a customer for cross sell
20	Cust Care Score	Customer satisfaction score given by customer in previous service call

4.2 Sample of the dataset:

Below mentioned is a sample of given dataset with all the columns and some rows

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	Monthly
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	Fe male	3	Executive	3.0	Divorced	
4	7000004	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced	

The given data set have 20 columns have 4250 rows as the given data set is a structured data. This data consists of both independent variables and dependent variables, and data set consists of both categorical and continuous data type which is (Integer and float).

4.3 Understanding of Attributes:

- Occupation is a Categorical variable with values small business, large business, Freelancer and salaried. There are eight object fields which indicate that there are 8 Categorical variables.
- The variable Customer ID (CustID) is a continuous integer variable. Similar to this there are a total of five integer variables.
- The variable Monthly Income (MonthlyIncome) indicates the discrete values which has the values of each customer. Similar to this there is a total of 7 float variables.

Variables	Data type
CustID	Int64
AgentBonus	Int64
Age	float64
CustTenure	float64
Channel	Object
Occupation	Object
EducationField	Object
Gender	Object
ExistingProdType	Int64
Designation	Object
NumberOfPolicy	float64
MaritalStatus	Object
MonthlyIncome	float64

Complaint	Int64
ExistingPoicyTenure	float64
SumAssured	float64
Zone	Object
PaymentMethod	Object
LastMonthCalls	Int64
CustCareScore	float64

We have total no of 4520 rows and 20 columns in the dataset. Out of 20, all 8 are object (categorical in nature) , 7 are of float data type whereas 5 are of integer data type. All object data type are of Categorical datatype fields present in the data.

4.4 Check for Duplicate values:

We need to check the duplicate values as well where we found that in the given data set we don't have any duplicate records present.

```
The total no of duplicate vaules = 0
```

CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	Monthlyl

4.5 Dropping the insignificant column: Since there is a column named "CustID" which we found is an insignificant and is also unique identifier of Customer data. It is further not going to add any valuable contribution in model building so we are going to drop it

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 39 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   AgentBonus                      4520 non-null   float64
 1   Age                             4520 non-null   float64
 2   CustTenure                      4520 non-null   float64
 3   ExistingProdType                4520 non-null   float64
 4   NumberOfPolicy                  4520 non-null   float64
 5   MonthlyIncome                   4520 non-null   float64
 6   Complaint                       4520 non-null   float64
 7   ExistingPolicyTenure            4520 non-null   float64
 8   SumAssured                      4520 non-null   float64
 9   LastMonthCalls                  4520 non-null   float64
 10  CustCareScore                   4520 non-null   float64
 11  Channel_Online                  4520 non-null   uint8
 12  Channel_Third Party Partner     4520 non-null   uint8
 13  Occupation_Laarge Business      4520 non-null   uint8
 14  Occupation_Large Business       4520 non-null   uint8
 15  Occupation_Salaried             4520 non-null   uint8
 16  Occupation_Small Business       4520 non-null   uint8
 17  EducationField_Engineer         4520 non-null   uint8
 18  EducationField_Graduate         4520 non-null   uint8
 19  EducationField_MBA              4520 non-null   uint8
 20  EducationField_Post Graduate    4520 non-null   uint8
 21  EducationField_UG               4520 non-null   uint8
 22  EducationField_Under Graduate   4520 non-null   uint8
 23  Gender_Female                   4520 non-null   uint8
 24  Gender_Male                     4520 non-null   uint8
 25  Designation_Exe                 4520 non-null   uint8
 26  Designation_Executive           4520 non-null   uint8
 27  Designation_Manager             4520 non-null   uint8
 28  Designation_Senior Manager      4520 non-null   uint8
 29  Designation_VP                  4520 non-null   uint8
 30  MaritalStatus_Married           4520 non-null   uint8
 31  MaritalStatus_Single            4520 non-null   uint8
 32  MaritalStatus_Unmarried         4520 non-null   uint8
 33  Zone_North                      4520 non-null   uint8
 34  Zone_South                      4520 non-null   uint8
 35  Zone_West                       4520 non-null   uint8
 36  PaymentMethod_Monthly           4520 non-null   uint8
 37  PaymentMethod_Quarterly         4520 non-null   uint8
 38  PaymentMethod_Yearly            4520 non-null   uint8
dtypes: float64(11), uint8(28)
memory usage: 512.2 KB
```

4.6 Data Description:

In Data Pre-processing we'll be performing some functions to understand the data which can be done through Data description. If we find any noises, any bad values surely that will be addressed (In this data set we don't have any bad value) and will be taken care of. We don't have any duplicate values either. If we find any variables insignificant, that variable will be dropped.

	count	mean	std	min	25%	50%	75%	max
AgentBonus	4520.0	4062.773894	1358.284526	1605.0	3027.75	3911.5	4867.25	7626.500
Age	4520.0	13.855863	8.800660	2.0	6.00	12.0	19.00	38.500
CustTenure	4520.0	13.865265	8.765148	2.0	6.00	12.0	19.00	38.500
ExistingProdType	4520.0	3.695575	0.936418	1.5	3.00	4.0	4.00	5.500
NumberOfPolicy	4520.0	3.569690	1.449302	1.0	2.00	4.0	5.00	6.000
MonthlyIncome	4520.0	22574.032557	3948.153973	16009.0	19858.00	21877.0	24531.75	31542.375
Complaint	4520.0	0.287168	0.452491	0.0	0.00	0.0	1.00	1.000
ExistingPolicyTenure	4520.0	3.876327	2.954770	1.0	1.00	3.0	5.00	11.000
SumAssured	4520.0	615902.262154	229255.422484	168536.0	444476.25	590012.5	750010.50	1208311.875
LastMonthCalls	4520.0	4.624336	3.610676	0.0	2.00	3.0	8.00	17.000
CustCareScore	4520.0	3.066814	1.375007	1.0	2.00	3.0	4.00	5.000
Channel_Online	4520.0	0.103540	0.304696	0.0	0.00	0.0	0.00	1.000
Channel_Third Party Partner	4520.0	0.189823	0.392204	0.0	0.00	0.0	0.00	1.000
Occupation_Laarge Business	4520.0	0.033850	0.180862	0.0	0.00	0.0	0.00	1.000
Occupation_Large Business	4520.0	0.056416	0.230749	0.0	0.00	0.0	0.00	1.000
Occupation_Salaried	4520.0	0.484956	0.499829	0.0	0.00	0.0	1.00	1.000
Occupation_Small Business	4520.0	0.424336	0.494297	0.0	0.00	0.0	1.00	1.000
EducationField_Engineer	4520.0	0.090265	0.286593	0.0	0.00	0.0	0.00	1.000
EducationField_Graduate	4520.0	0.413717	0.492553	0.0	0.00	0.0	1.00	1.000
EducationField_MBA	4520.0	0.016372	0.126914	0.0	0.00	0.0	0.00	1.000
EducationField_Post Graduate	4520.0	0.055752	0.229468	0.0	0.00	0.0	0.00	1.000
EducationField_UG	4520.0	0.050885	0.219787	0.0	0.00	0.0	0.00	1.000
EducationField_Under Graduate	4520.0	0.263274	0.440459	0.0	0.00	0.0	1.00	1.000
Gender_Female	4520.0	0.333407	0.471483	0.0	0.00	0.0	1.00	1.000
Gender_Male	4520.0	0.594690	0.491006	0.0	0.00	1.0	1.00	1.000
Designation_Exe	4520.0	0.028097	0.165269	0.0	0.00	0.0	0.00	1.000
Designation_Executive	4520.0	0.339602	0.473626	0.0	0.00	0.0	1.00	1.000
Designation_Manager	4520.0	0.358407	0.479586	0.0	0.00	0.0	1.00	1.000
Designation_Senior Manager	4520.0	0.149558	0.356677	0.0	0.00	0.0	0.00	1.000
Designation_VP	4520.0	0.050000	0.217969	0.0	0.00	0.0	0.00	1.000
MaritalStatus_Married	4520.0	0.501770	0.500052	0.0	0.00	1.0	1.00	1.000
MaritalStatus_Single	4520.0	0.277434	0.447782	0.0	0.00	0.0	1.00	1.000

4.7 Null value Check:

In the given dataset there are 1166 records with null value data, which are to be treated as with the null value records if not treated we will not be able to build a model that can predict the accurate results. Below are the list of variables which carries a missing values.

Before Missing values Treatment

```
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
dtype: int64
```

Out of 85 thousand data points we have 1000 data points which are missing values which is 1.3% of the total data. In our case other than the CustCareScore which is an object (categorical), rest all the columns with null values are numerical in nature with float data types, and they can be replaced with median values with respective columns and the Categorical variable CustCareScore can be imputed with Mode values of the column.

After Missing values Treatment

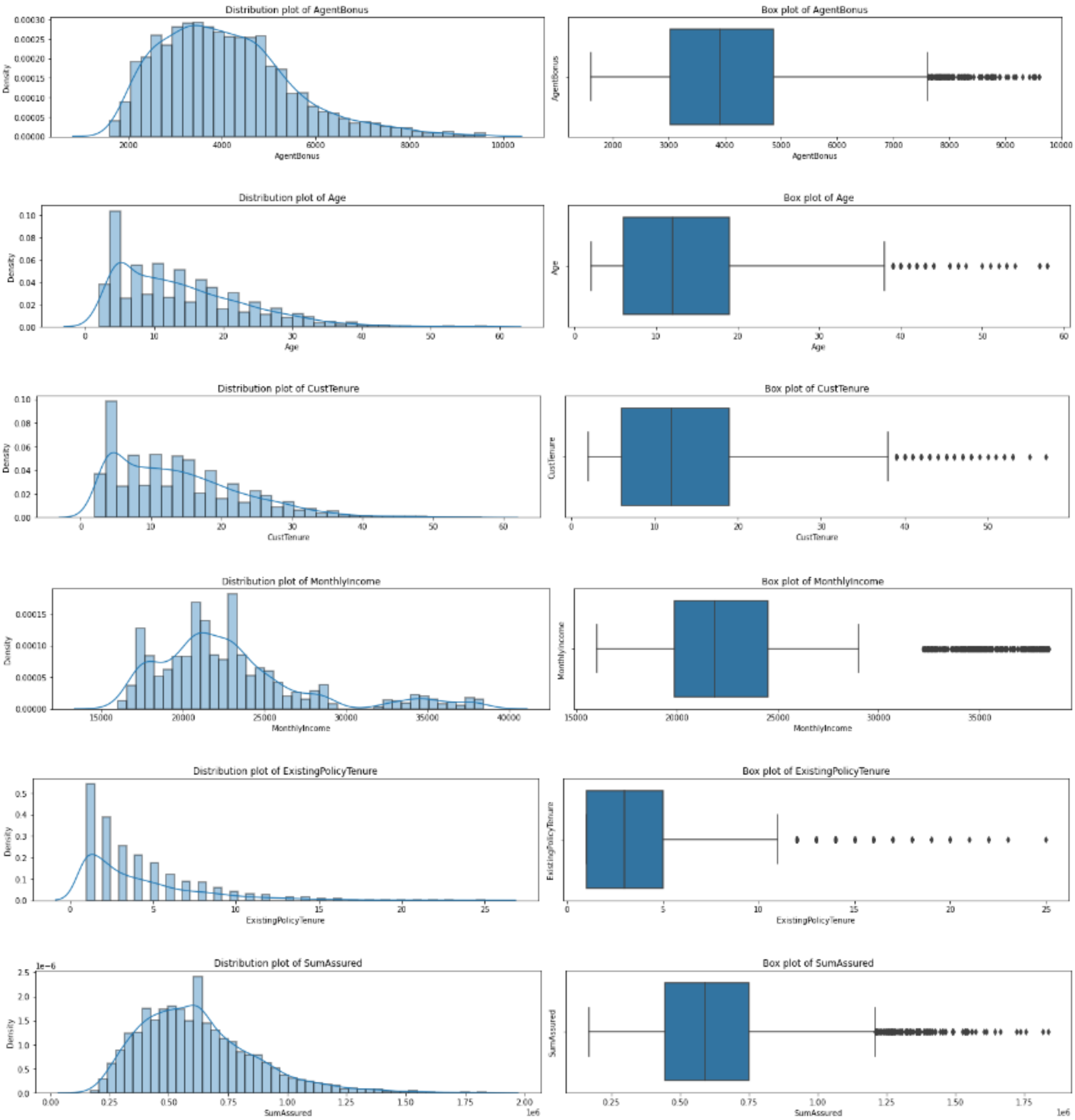
```
AgentBonus              0
Age                     0
CustTenure              0
Channel                 0
Occupation              0
EducationField          0
Gender                  0
ExistingProdType        0
Designation             0
NumberOfPolicy          0
MaritalStatus           0
MonthlyIncome           0
Complaint               0
ExistingPolicyTenure    0
SumAssured              0
Zone                    0
PaymentMethod           0
LastMonthCalls          0
CustCareScore           0
dtype: int64
```

We did the same by imputing the Fields like ‘Number of Policy’, ‘CustCareScore’, ‘Age’, ‘CustTenure’ are imputed with mode values and fields like ‘Sum Asssured’ and ‘Monthly Income’ are imputed with mean values.

4.8 Check for Outliers and its Treatment (Univariate Analysis)

Below are the specific variables who is suffering from Outliers

<u>Before Outliers Treatment</u>



Observations :

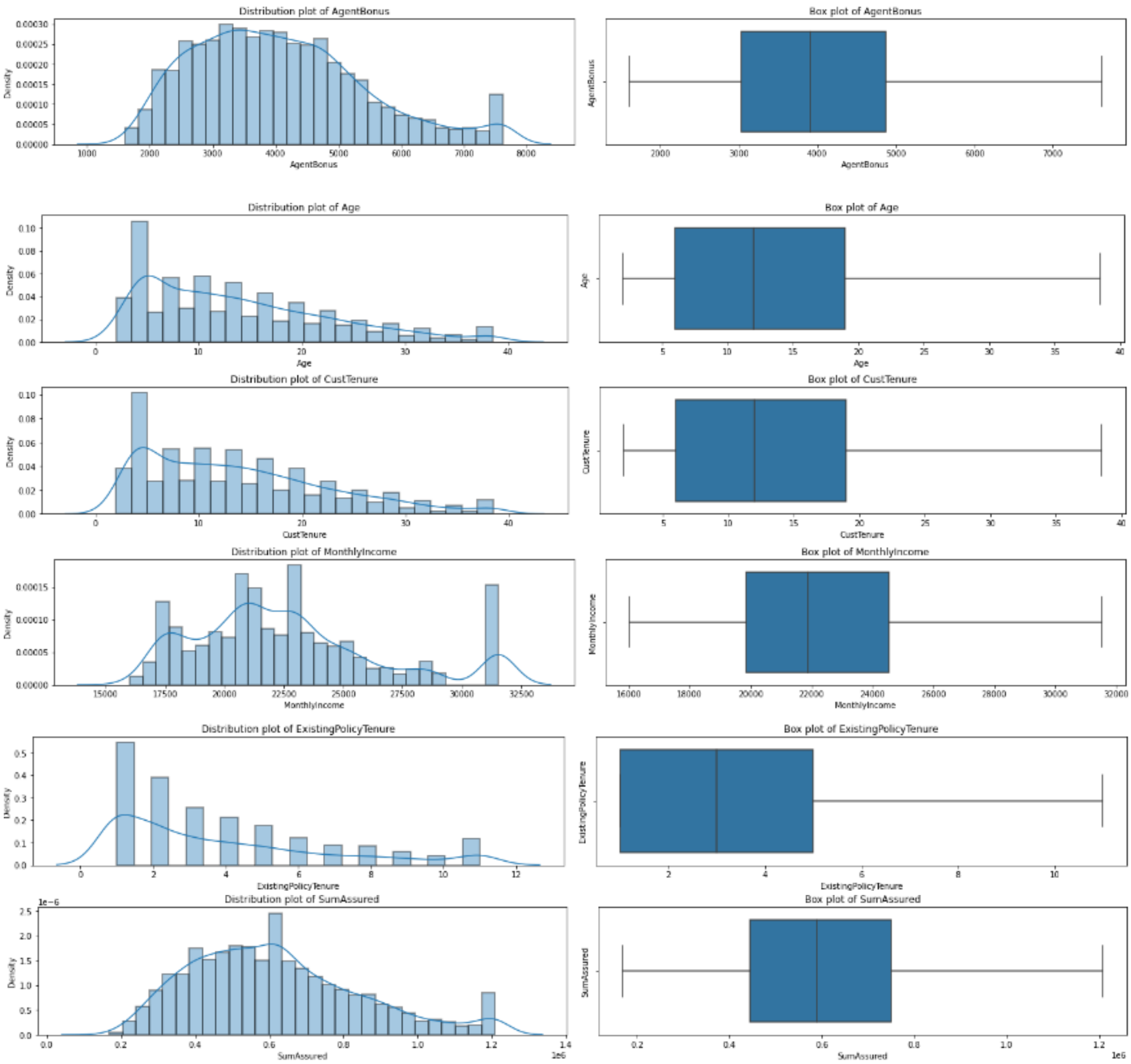
- We see that there are outliers present in many variables (AgentBonus, Age, CustTenure, MonthlyIncome, ExistingPolicyTenure, SummAssured) which needs to be treated through outlier treatment.

- Through density plot we observe that there is left skewness observed in few variables (Age, Cust Tenure) rest are following normal distribution.

- Also, we observe that most of policy holders are minor which means lower premium collected so lower revenue but at the same time claim rate is also lower

So, we need to treat the outliers and for that we have few ways to which one is dropping the Outlier and second is replacing the outlier with IQR (Inter Quartile Range) method where we treat the most extreme values be it upper or lower. In this case we are going to treat the Outlier with IQR method rather than dropping the values so that we're able to retain the existing count of the data.

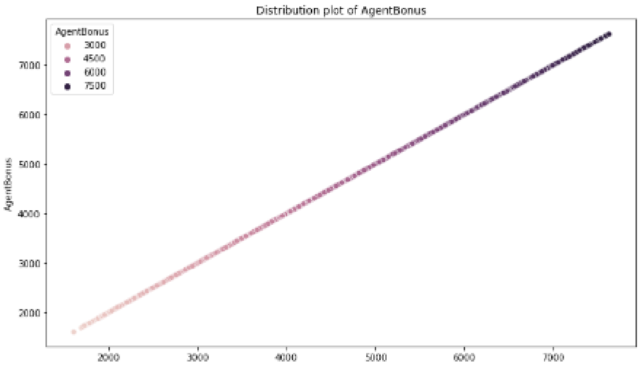
<u>After Outliers Treatment</u>



- Basically Outliers are those values which lie outside 1.5 *IQR, and in the above graphs only those variables are highlighted which is suffering from Outliers.

- From the box plot in univariate analysis, we observed that there are many outliers present in the given data which ultimately means that treating outliers becomes mandatory for the given data.
- So from the above graph it is clear that those variables who are dealing with Outliers has been treated successfully and is good to move on for further analysis.

4.9 Bivariate Analysis

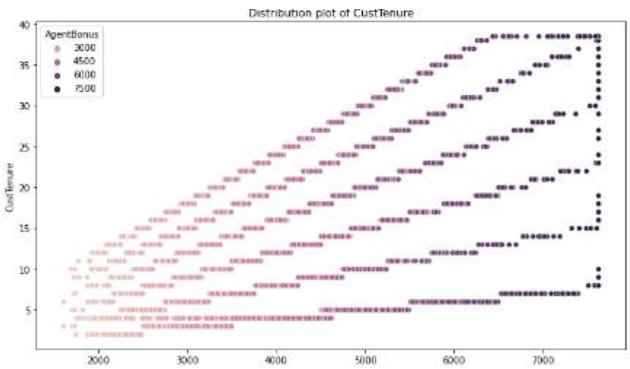


count	4520.000000
mean	4062.773894
std	1358.284526
min	1605.000000
25%	3027.750000
50%	3911.500000
75%	4867.250000
max	7626.500000

Distribution Plot of Agent bonus

count	4520.000000
mean	13.855863
std	8.800660
min	2.000000
25%	6.000000
50%	12.000000
75%	19.000000
max	38.500000

Distribution Plot of Age

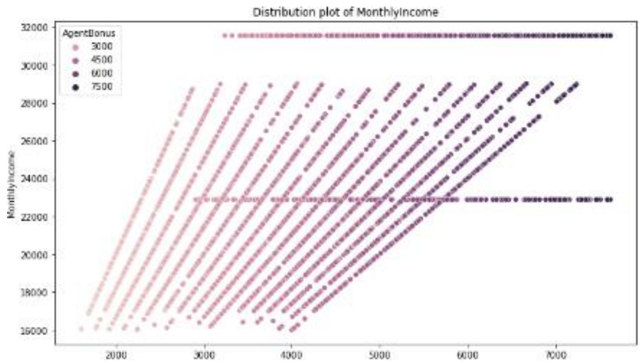


count	4520.000000
mean	13.865265
std	8.765148
min	2.000000
25%	6.000000
50%	12.000000
75%	19.000000
max	38.500000

Distribution Plot of Customer Tenure

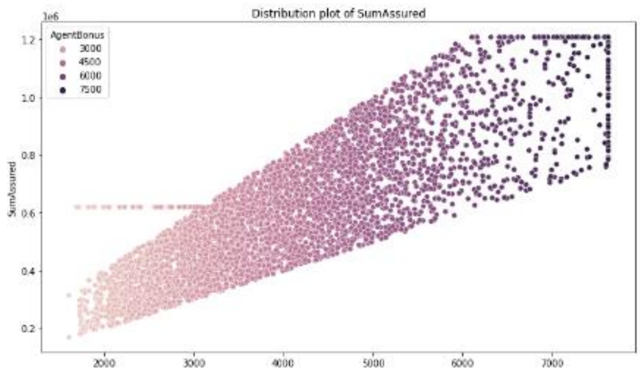
count	4520.000000
mean	3.695575
std	0.936418
min	1.500000
25%	3.000000
50%	4.000000
75%	4.000000
max	5.500000

Distribution Plot of Existing Production Type



count	4520.000000
mean	22890.309991
std	4756.317536
min	16009.000000
25%	19858.000000
50%	21877.000000
75%	24531.750000
max	38456.000000

Distribution Plot of Monthly Income



count	4.520000e+03
mean	6.199997e+05
std	2.420028e+05
min	1.685360e+05
25%	4.444762e+05
50%	5.900125e+05
75%	7.500105e+05
max	1.838496e+06

Distribution Plot of Sum Assured

Only those variables are listed above in this Bivariate Analysis which plays an significant role in Target level prediction.

4.10 Co-relation values & Heat map

scroll output; double click to hide

	CustID	AgentBonus	Age	CustTenure	Channel
CustID	1.000000	0.194469	0.137151	0.153822	-0.003346
AgentBonus	0.194469	1.000000	0.560300	0.559312	-0.035635
Age	0.137151	0.560300	1.000000	0.331396	-0.026437
[illegible]	[illegible]	0.559312	0.331396	1.000000	-0.036903
Channel	-0.003346	-0.035635	-0.026437	-0.036903	1.000000
Occupation	0.002776	0.024176	0.015726	0.027638	0.007956
EducationField	0.018621	0.008868	0.004500	0.006645	0.017410
Gender	0.085549	0.043575	0.022354	0.037631	-0.013682
ExistingProdType	0.426252	0.112871	0.098994	0.113379	0.005767
Designation	0.023020	0.153557	0.120095	0.109045	-0.000582
NumberOfPolicy	0.254446	0.079161	0.065757	0.061891	-0.018514
MaritalStatus	0.077783	-0.084234	-0.057079	-0.065455	0.015972
MonthlyIncome	0.321987	0.566234	0.376072	0.360205	-0.035724
Complaint	0.001921	0.013904	0.013925	0.005238	0.012836
ExistingPolicyTenure	0.117161	0.293780	0.160600	0.162035	-0.005560
SumAssured	0.163649	0.838631	0.477843	0.476221	-0.033335
Zone	-0.010998	-0.006068	0.012766	-0.001380	-0.002116
PaymentMethod	0.001127	-0.013060	0.006832	-0.007583	-0.011250
LastMonthCalls	0.121774	0.201466	0.140683	0.137942	0.005133
CustCareScore	-0.034245	0.021995	0.029872	0.013223	0.037541

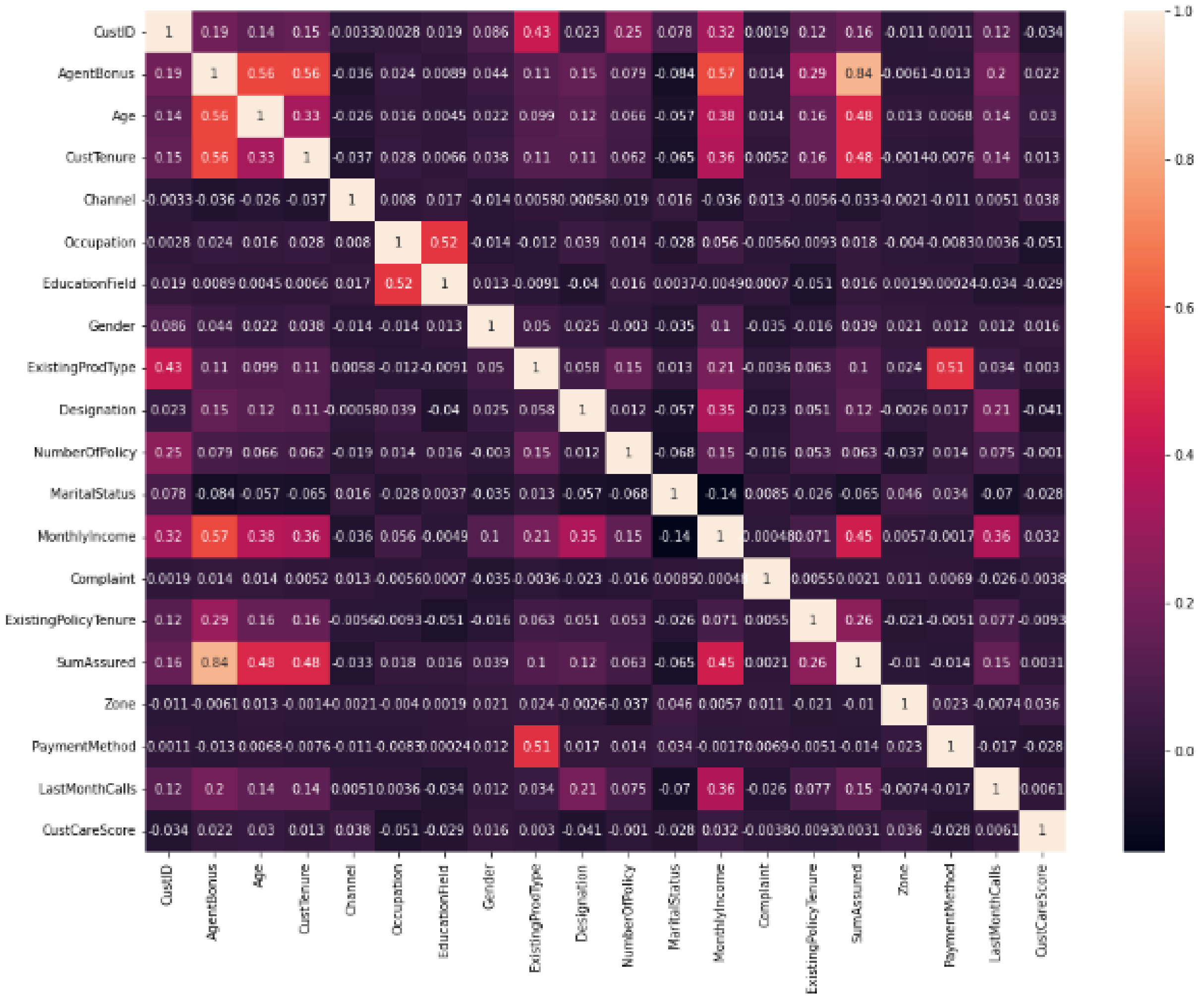
	Occupation	EducationField	Gender	ExistingProdType
CustID	0.002776	0.018621	0.085549	0.426252
AgentBonus	0.024176	0.008868	0.043575	0.112871
Age	0.015726	0.004500	0.022354	0.098994
CustTenure	0.027638	0.006645	0.037631	0.113379
Channel	0.007956	0.017410	-0.013682	0.005767
Occupation	1.000000	0.522222	-0.013999	-0.012056
EducationField	0.522222	1.000000	0.012886	-0.009148
Gender	-0.013999	0.012886	1.000000	0.049869
ExistingProdType	-0.012056	-0.009148	0.049869	1.000000
Designation	0.038848	-0.039797	0.025043	0.057538
NumberOfPolicy	0.013778	0.015699	-0.002965	0.153253
MaritalStatus	-0.027913	0.003700	-0.035331	0.013232
MonthlyIncome	0.056489	-0.004853	0.103181	0.207884
Complaint	-0.005633	0.000697	-0.034759	-0.003581
ExistingPolicyTenure	-0.009307	-0.050735	-0.015614	0.063208
SumAssured	0.018215	0.016025	0.038680	0.100265
Zone	-0.003981	0.001851	0.021027	0.023595
PaymentMethod	-0.008286	0.000241	0.012498	0.508807
LastMonthCalls	0.003567	-0.033820	0.011501	0.033810
CustCareScore	-0.050679	-0.028969	0.016196	0.002997

	Designation	NumberOfPolicy	MaritalStatus
CustID	0.023020	0.254446	0.077783
AgentBonus	0.153557	0.079161	-0.084234
Age	0.120095	0.065757	-0.057079
CustTenure	0.109045	0.061891	-0.065455
Channel	-0.000582	-0.018514	0.015972
Occupation	0.038848	0.013778	-0.027913
EducationField	-0.039797	0.015699	0.003700
Gender	0.025043	-0.002965	-0.035331
ExistingProdType	0.057538	0.153253	0.013232
Designation	1.000000	0.012356	-0.056558
NumberOfPolicy	0.012356	1.000000	-0.068103
MaritalStatus	-0.056558	-0.068103	1.000000
MonthlyIncome	0.347330	0.145314	-0.137657
Complaint	-0.023137	-0.016014	0.008482
ExistingPolicyTenure	0.051156	0.052628	-0.025910
SumAssured	0.120581	0.062680	-0.064630
Zone	-0.002594	-0.037167	0.046327
PaymentMethod	0.016862	0.014173	0.033759
LastMonthCalls	0.206669	0.075032	-0.069669
CustCareScore	-0.040520	-0.001005	-0.028218

	MonthlyIncome	Complaint	ExistingPolicyTenure
CustID	0.321987	0.001921	0.117161
AgentBonus	0.566234	0.013904	0.293780
Age	0.376072	0.013925	0.160600
CustTenure	0.360205	0.005238	0.162035
Channel	-0.035724	0.012836	-0.005560
Occupation	0.056489	-0.005633	-0.009307
EducationField	-0.004853	0.000697	-0.050735
Gender	0.103181	-0.034759	-0.015614
ExistingProdType	0.207884	-0.003581	0.063208
Designation	0.347330	-0.023137	0.051156
NumberOfPolicy	0.145314	-0.016014	0.052628
MaritalStatus	-0.137657	0.008482	-0.025910
MonthlyIncome	1.000000	-0.000484	0.071489
Complaint	-0.000484	1.000000	0.005549
ExistingPolicyTenure	0.071489	0.005549	1.000000
SumAssured	0.448113	0.002060	0.258702
Zone	0.005716	0.011059	-0.020735
PaymentMethod	-0.001677	0.006901	-0.005114
LastMonthCalls	0.357330	-0.026193	0.077118
CustCareScore	0.031993	-0.003814	-0.009349

	SumAssured	Zone	PaymentMethod	LastMonthCalls
CustID	0.163649	-0.010998	0.001127	0.121774
AgentBonus	0.838631	-0.006068	-0.013060	0.201466
Age	0.477843	0.012766	0.006832	0.140683
CustTenure	0.476221	-0.001380	-0.007583	0.137942
Channel	-0.033335	-0.002116	-0.011250	0.005133
Occupation	0.018215	-0.003981	-0.008286	0.003567
EducationField	0.016025	0.001851	0.000241	-0.033820
Gender	0.038680	0.021027	0.012498	0.011501
ExistingProdType	0.100265	0.023595	0.508807	0.033810
Designation	0.120581	-0.002594	0.016862	0.206669
NumberOfPolicy	0.062680	-0.037167	0.014173	0.075032
MaritalStatus	-0.064630	0.046327	0.033759	-0.069669
MonthlyIncome	0.448113	0.005716	-0.001677	0.357330
Complaint	0.002060	0.011059	0.006901	-0.026193
ExistingPolicyTenure	0.258702	-0.020735	-0.005114	0.077118
SumAssured	1.000000	-0.010429	-0.014352	0.152483
Zone	-0.010429	1.000000	0.022795	-0.007434
PaymentMethod	-0.014352	0.022795	1.000000	-0.017292
LastMonthCalls	0.152483	-0.007434	-0.017292	1.000000
CustCareScore	0.003136	0.036376	-0.028040	0.006126

	CustCareScore
CustID	-0.034245
AgentBonus	0.021995
Age	0.029872
CustTenure	0.013223
Channel	0.037541
Occupation	-0.050679
EducationField	-0.028969
Gender	0.016196
ExistingProdType	0.002997
Designation	-0.040520
NumberOfPolicy	-0.001005
MaritalStatus	-0.028218
MonthlyIncome	0.031993
Complaint	-0.003814
ExistingPolicyTenure	-0.009349
SumAssured	0.003136
Zone	0.036376
PaymentMethod	-0.028040
LastMonthCalls	0.006126
CustCareScore	1.000000



Plot of Heat Map

5. Business Insights from Exploratory Data Analysis

As we all know that Exploratory Data Analysis, or EDA is an exhaustive look at existing data from current and historical surveys conducted by a company. It allows us to prepare and analyse the proper model to interpret the correct results.

So far we've been through all the data cleansing process like removing missing values, outlier treatment and looking for Duplicate values but now we've to look ahead for the visualization part of the data.

5.1 Variable Transformation:

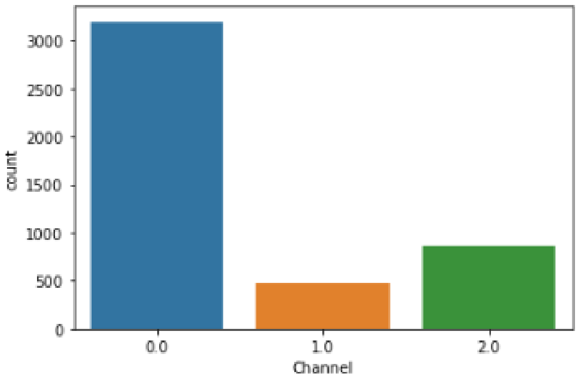
Variable transformation is basically a way to make the data work better in our model. Data variables can have two types of form: numeric variable and categorical variable, and their transformation should have different approaches.

And we've observed that there are many categorical variables present in existing dataset which need to be transformed. Since there are various methods of encoding data sets like, One-hot encoding, Binary encoding, Target encoding etc. but I have used here 'Label Encoding'. And the below variables are encoded into its numeric form.

```
# Encode labels in column 'species'.
sales['Channel']= label_encoder.fit_transform(sales['Channel'])
sales['Occupation']= label_encoder.fit_transform(sales['Occupation'])
sales['EducationField']= label_encoder.fit_transform(sales['EducationField'])
sales['Gender']= label_encoder.fit_transform(sales['Gender'])
sales['Designation']= label_encoder.fit_transform(sales['Designation'])
sales['MaritalStatus']= label_encoder.fit_transform(sales['MaritalStatus'])
sales['Zone']= label_encoder.fit_transform(sales['Zone'])
sales['PaymentMethod']= label_encoder.fit_transform(sales['PaymentMethod'])
```

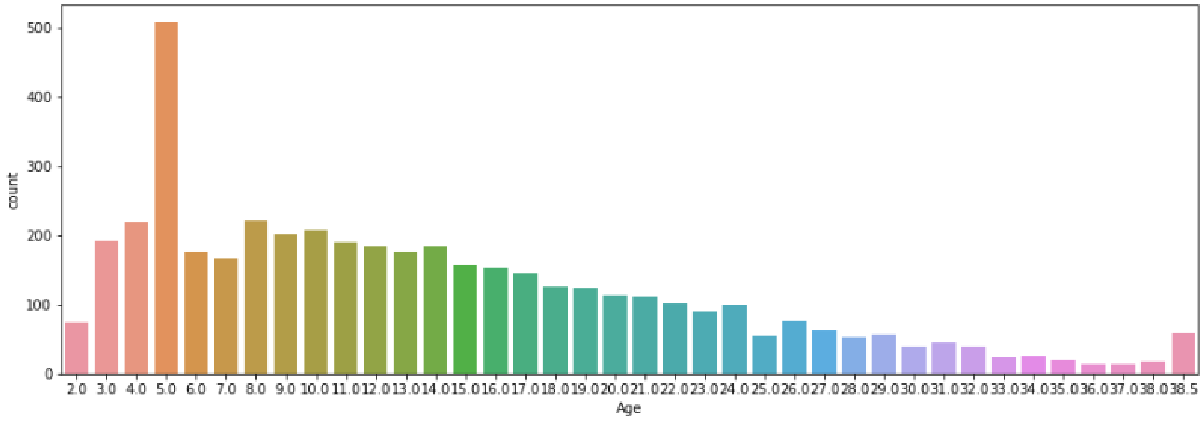
Variable encoded here are as follows:

- Gender
- Education Field
- Channel
- Occupation
- Designation
- Payment Method
- Marital Status
- Zone

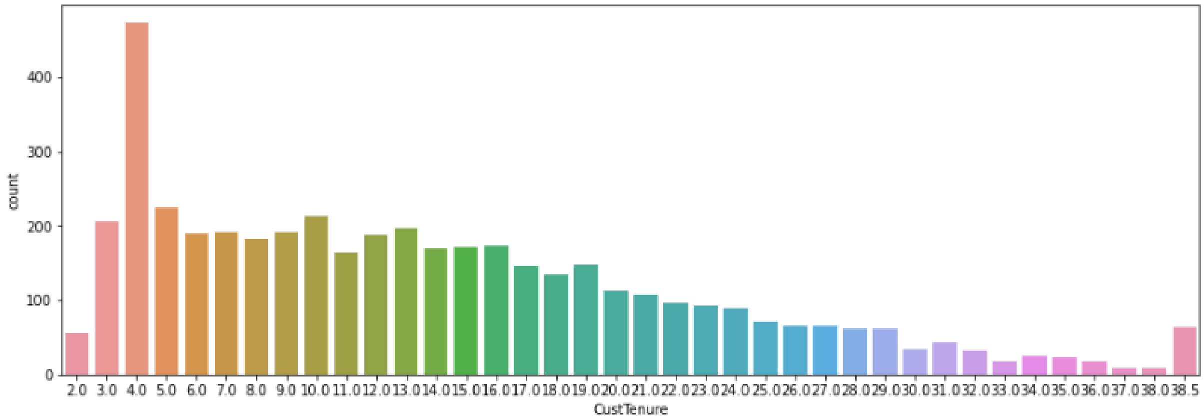



```
count    4520.000000
mean        0.483186
std         0.793412
min         0.000000
25%         0.000000
50%         0.000000
75%         1.000000
max         2.000000
Name: Channel, dtype: float64
```

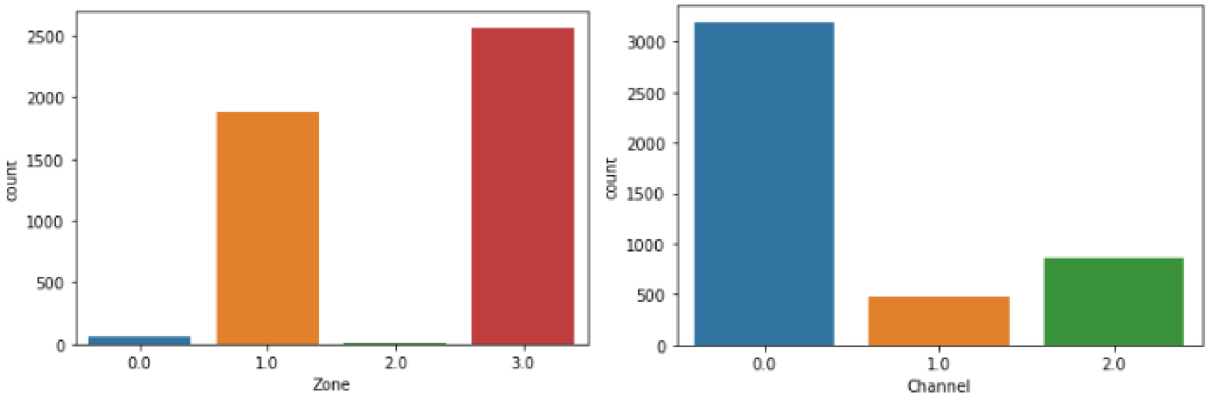
Here 0 refers to Agent, 1 refers to Online channel whereas 2 refers to Third Party Partner. As we can see that 'Channel' variable also plays an important factor in the data imbalance. The major channel of sourcing is 'Customer Agent' and rest are minimum.



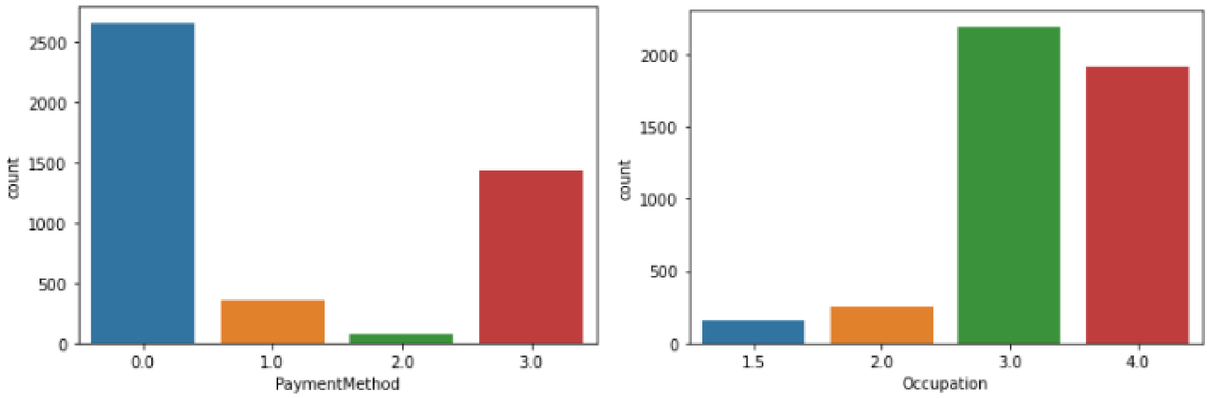
Plot of Customer Age and its counts



Plot of Customer Tenure and its counts



Plot of Customer Geographical location (Zone) and Channel



Plot of Payment method made and Occupation of Customers

Business Insights:

- The company should work on more web and mobile based applications for convincing or as penetration towards because online channel is the least along with third party based.
- We also have a variable of 'CustTenure' which clearly shows us that 22 years is max engagement with the organisation, which shows that products designed are serving mostly around 20-25 years segment.
- The company should produce some new products which could tie up the customers to itself through life time.
- When we talk about the geographical area or the zones which is mentioned in this dataset, then North and West are major contributors whereas the South and East shows very minimum hikes. It clearly points out that the company presence in these regions are limited therefore the Company should think of expanding their horizons to the respective zones.
- We have observed that the dataset comprises of age group less than 20 which means that less premium is expected out and so as the low mortality rate. This shows that company is operating with lower profit margins.

6. Model buildings and Interpretations

- Since we know that the given problems statement is continuous, however the variables involved are continuous in nature, so probably seems regression is better for this problem.
- But we also have seen that there are some categorical variables are also present in the data set. Since the regression model uses only numerical variables so we hav eto convert those categorical variables into the numerical form.
- We also see that some of the categorical variables have more than two categories, so we apply One-Hot encoding which means that it converts each categorical level within the category features into columns and makes it a binary feed.
  All in all it means that wherever the value holds 'true' it will give a value of '1' and wherever the value is not available for a particular observation, it will give a value of '0'.

	AgentBonus	Age	CustTenure	ExistingProdType	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	LastMonthCalls	...
0	4409.0	22.0	4.0	3.0	2.0	20993.0	1.0	2.0	806761.000000	5.0	...
1	2214.0	11.0	2.0	4.0	4.0	20130.0	0.0	3.0	294502.000000	7.0	...
2	4273.0	26.0	4.0	4.0	3.0	17090.0	1.0	2.0	619999.699267	0.0	...
3	1791.0	11.0	4.0	3.0	3.0	17909.0	1.0	2.0	268635.000000	0.0	...
4	2955.0	6.0	4.0	3.0	4.0	18468.0	0.0	4.0	366405.000000	2.0	...

MaritalStatus_Married	MaritalStatus_Single	MaritalStatus_Unmarried	Zone_North	Zone_South	Zone_West	PaymentMethod_Monthly	PaymentMethod_Quarterly
0	1	0	1	0	0	0	0
0	0	0	1	0	0	0	0
0	0	1	1	0	0	0	0
0	0	0	0	0	1	0	0
0	0	0	0	0	1	0	0

Sample of the dataset after Encoding

Well we are using the same data set so we don't need to have EDA every time because we have already treated the biasedness that it had like null values, Outliers, so we could directly proceed for our Model Exercise.

7. Model building

While building a model the first step that we need to do is to split the data into Train and Test data with their respective ratios. And here we have split the data into 75:25 ratio.

7.1 Data shape after Train and Test Split

```
Train data (3390, 38)
Test Data (1130, 38)
```

7.2 Insights of R Square and Root Mean Square Error (RMSE)

- The value of R Square on Train data is 0.809 and RMSE on Train data is 596
- The value of R Square on Test data is 0.781 and RMSE on Test data is 623

7.3 Using Linear Regression model

Now the first iteration towards building Linear Regression model is that we used all the independent variables that the dataset carries which is given below:-

```
#   Column
--- ------
0   AgentBonus
1   Age
2   CustTenure
3   ExistingProdType
4   NumberOfPolicy
5   MonthlyIncome
6   Complaint
7   ExistingPolicyTenure
8   SumAssured
9   LastMonthCalls
10  CustCareScore
11  Channel_Online
12  Channel_Third Party Partner
13  Occupation_Laarge Business
14  Occupation_Large Business
15  Occupation_Salaried
16  Occupation_Small Business
17  EducationField_Engineer
18  EducationField_Graduate
19  EducationField_MBA
20  EducationField_Post Graduate
21  EducationField_UG
22  EducationField_Under Graduate
23  Gender_Female
24  Gender_Male
25  Designation_Exe
26  Designation_Executive
27  Designation_Manager
28  Designation_Senior Manager
29  Designation_VP
30  MaritalStatus_Married
31  MaritalStatus_Single
32  MaritalStatus_Unmarried
33  Zone_North
34  Zone_South
35  Zone_West
36  PaymentMethod_Monthly
37  PaymentMethod_Quarterly
38  PaymentMethod_Yearly
```

List of Variables used in model building

8. Summary of Linear Model 1

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.799
Method:                 Least Squares   F-statistic:                     410.4
Date:                Thu, 23 Feb 2023   Prob (F-statistic):               0.00
Time:                        09:07:43   Log-Likelihood:                -26546.
No. Observations:                3390   AIC:                         5.316e+04
Df Residuals:                    3356   BIC:                         5.337e+04
Df Model:                          33
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
Intercept                    -308.8624    210.137     -1.470      0.142    -720.871     103.146
Age                            21.5843      1.411     15.294      0.000      18.817      24.351
CustTenure                     22.7989      1.408     16.189      0.000      20.038      25.560
ExistingProdType              -74.0961     23.404     -3.166      0.002    -119.983     -28.209
NumberOfPolicy                  0.0983      7.655      0.013      0.990     -14.911      15.108
MonthlyIncome                   0.0722      0.005     14.648      0.000       0.063       0.082
Complaint                      29.5719     23.490      1.259      0.208     -16.485      75.629
ExistingPolicyTenure           38.2599      3.748     10.208      0.000      30.911      45.608
SumAssured                      0.0035   6.01e-05     58.759      0.000       0.003       0.004
LastMonthCalls                  0.6478      3.147      0.206      0.837      -5.522       6.817
CustCareScore                   8.6291      7.749      1.114      0.266      -6.564      23.822
Channel_Online                 24.9877     35.035      0.713      0.476     -43.704      93.679
Channel_Third_Party_Partner    -3.2896     27.360     -0.120      0.904     -56.933      50.353
Occupation_Large_Business     -27.6162     74.344     -0.371      0.710    -173.380     118.148
Occupation_Salaried            -0.4077    147.813     -0.003      0.998    -290.220     289.405
Occupation_Small_Business      -0.4595    148.313     -0.003      0.998    -291.252     290.333
EducationField_Engineer       -17.6585    139.264     -0.127      0.899    -290.710     255.393
EducationField_MBA           -127.4832     90.602     -1.407      0.160    -305.125      50.159
EducationField_Post_Graduate   12.8045     49.439      0.259      0.796     -84.128     109.737
EducationField_Under_Graduate -33.5090     32.425     -1.033      0.301     -97.084      30.066
Gender_Male                    15.1673     21.646      0.701      0.484     -27.273      57.608
Designation_Executive         105.4200     46.669      2.259      0.024      13.917     196.923
Designation_Manager           -70.6496     40.914     -1.727      0.084    -150.868       9.569
Designation_Senior_Manager     -5.7249     43.342     -0.132      0.895     -90.704      79.255
Designation_VP                 47.1871     64.562      0.731      0.465     -79.398     173.772
MaritalStatus_Married         -52.9494     29.153     -1.816      0.069    -110.109       4.211
MaritalStatus_Single           11.4256     32.325      0.353      0.724     -51.953      74.804
MaritalStatus_Unmarried      -137.8457     60.636     -2.273      0.023    -256.734     -18.957
Zone_North                     49.1826     93.357      0.527      0.598    -133.860     232.225
Zone_South                    201.2722    289.706      0.695      0.487    -366.746     769.291
Zone_West                      42.9594     92.886      0.462      0.644    -139.158     225.077
PaymentMethod_Monthly         -49.9428     57.238     -0.873      0.383    -162.167      62.282
PaymentMethod_Quarterly        -9.2841     86.238     -0.108      0.914    -178.368     159.800
PaymentMethod_Yearly           44.1151     34.186      1.290      0.197     -22.913     111.143
==============================================================================
Omnibus:                      136.383   Durbin-Watson:                   1.998
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              156.022
Skew:                           0.479   Prob(JB):                     1.32e-34
Kurtosis:                       3.430   Cond. No.                     1.94e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.94e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Intercept	-308.862450
Age	21.584280
CustTenure	22.798862
ExistingProdType	-74.096130
NumberOfPolicy	0.098293
MonthlyIncome	0.072213
Complaint	29.571853
ExistingPolicyTenure	38.259870
SumAssured	0.003531
LastMonthCalls	0.647778
CustCareScore	8.629110
Channel_Online	24.987669
Channel_Third_Party_Partner	-3.289607
Occupation_Large_Business	-27.616164
Occupation_Salaried	-0.407712
Occupation_Small_Business	-0.459524
EducationField_Engineer	-17.658530
EducationField_MBA	-127.483172
EducationField_Post_Graduate	12.804460
EducationField_Under_Graduate	-33.509031
Gender_Male	15.167335
Designation_Executive	105.419991
Designation_Manager	-70.649609
Designation_Senior_Manager	-5.724853
Designation_VP	47.187149
MaritalStatus_Married	-52.949351
MaritalStatus_Single	11.425564
MaritalStatus_Unmarried	-137.845696
Zone_North	49.182622
Zone_South	201.272204
Zone_West	42.959429
PaymentMethod_Monthly	-49.942768
PaymentMethod_Quarterly	-9.284073
PaymentMethod_Yearly	44.115083

Summary of Linear Model 1 and its Parameters

Observations:

- We know that we have the RMSE (Root Mean Square Error) value which is 608.92 and here the variation in R Square and Adjusted R Square is not that significant.
- So, in the second iteration we are going to consider only those independent variables, whose P value is less than 0.05 and Hence we are going to drop all redundant variables or we can just ignore those variables to reduce the multicollinearity levels which is also the reason behind the change in values.

9. Summary of Linear Model 2

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.801
Model:                            OLS   Adj. R-squared:                  0.799
Method:                 Least Squares   F-statistic:                     541.4
Date:                Thu, 23 Feb 2023   Prob (F-statistic):               0.00
Time:                        09:07:46   Log-Likelihood:                -26550.
No. Observations:                3390   AIC:                         5.315e+04
Df Residuals:                    3364   BIC:                         5.331e+04
Df Model:                          25
Covariance Type:            nonrobust
==================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                      -134.5120     94.870     -1.418      0.156    -320.522      51.498
Age                              21.4809      1.409     15.243      0.000      18.718      24.244
CustTenure                       22.6278      1.406     16.095      0.000      19.871      25.384
ExistingProdType                -60.1828     21.281     -2.828      0.005    -101.907     -18.459
NumberOfPolicy                    0.6519      7.626      0.085      0.932     -14.300      15.604
MonthlyIncome                     0.0676      0.004     18.866      0.000       0.061       0.075
Complaint                        29.6797     23.455      1.265      0.206     -16.308      75.668
ExistingPolicyTenure             38.7012      3.738     10.352      0.000      31.371      46.031
SumAssured                        0.0035   5.96e-05     59.012      0.000       0.003       0.004
LastMonthCalls                    0.0369      3.121      0.012      0.991      -6.081       6.155
CustCareScore                     9.0957      7.719      1.178      0.239      -6.038      24.230
Channel_Online                   22.2588     34.481      0.646      0.519     -45.346      89.864
EducationField_Engineer         -21.6990     37.380     -0.580      0.562     -94.990      51.592
EducationField_MBA             -118.1478     89.715     -1.317      0.188    -294.050      57.754
EducationField_Post_Graduate     20.2667     47.923      0.423      0.672     -73.695     114.228
Gender_Male                      18.3735     21.570      0.852      0.394     -23.919      60.666
Designation_Manager            -147.1686     23.883     -6.162      0.000    -193.996    -100.342
Designation_Senior_Manager      -68.0573     33.280     -2.045      0.041    -133.308      -2.807
MaritalStatus_Married           -51.5621     29.122     -1.771      0.077    -108.661       5.537
MaritalStatus_Single             16.7452     32.164      0.521      0.603     -46.318      79.808
MaritalStatus_Unmarried        -150.3305     60.216     -2.497      0.013    -268.394     -32.267
Zone_South                      156.9224    274.897      0.571      0.568    -382.060     695.905
Zone_West                        -4.8346     21.371     -0.226      0.821     -46.737      37.068
PaymentMethod_Monthly           -26.6094     54.455     -0.489      0.625    -133.379      80.160
PaymentMethod_Quarterly           4.4188     85.496      0.052      0.959    -163.211     172.048
PaymentMethod_Yearly             30.6159     32.555      0.940      0.347     -33.215      94.446
==============================================================================
Omnibus:                      123.398   Durbin-Watson:                   1.995
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              140.124
Skew:                           0.451   Prob(JB):                     3.74e-31
Kurtosis:                       3.422   Cond. No.                     1.72e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.72e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

```
Intercept                      -134.512039
Age                              21.480889
CustTenure                       22.627846
ExistingProdType                -60.182839
NumberOfPolicy                    0.651886
MonthlyIncome                     0.067571
Complaint                        29.679669
ExistingPolicyTenure             38.701215
SumAssured                        0.003515
LastMonthCalls                    0.036937
CustCareScore                     9.095713
Channel_Online                   22.258761
EducationField_Engineer         -21.699018
EducationField_MBA             -118.147848
EducationField_Post_Graduate     20.266736
Gender_Male                      18.373475
Designation_Manager            -147.168605
Designation_Senior_Manager      -68.057342
MaritalStatus_Married           -51.562084
MaritalStatus_Single             16.745167
MaritalStatus_Unmarried        -150.330458
Zone_South                      156.922437
Zone_West                        -4.834612
PaymentMethod_Monthly           -26.609384
PaymentMethod_Quarterly           4.418765
PaymentMethod_Yearly             30.615870
dtype: float64
```

Summary of Linear Model 2 and its Parameters

10. Variance Inflation Factors

Moving ahead we are going to look for VIF which is (Variance Inflation Factor) but before that we have to understand what VIF actually is. Basically a variance inflation factor is a measure of the amount of multicollinearity in regression analysis. The Variance Inflation Factor is nothing but a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity.

Higher the value of Variance Inflation Factor is, higher the correlation between the variables. Below is the pictorial view of VIF

VIF values (Before Variables dropped)

```
Age   VIF =  1.41
CustTenure   VIF =  1.38
ExistingProdType   VIF =  4.75
NumberOfPolicy   VIF =  1.12
MonthlyIncome   VIF =  5.24
Complaint   VIF =  1.01
ExistingPolicyTenure   VIF =  1.12
SumAssured   VIF =  1.76
LastMonthCalls   VIF =  1.2
CustCareScore   VIF =  1.03
Channel_Online   VIF =  1.05
Channel_Third_Party_Partner   VIF =  1.04
Occupation_Laarge Business   VIF =  62.39
Occupation_Large_Business   VIF =  101.63
Occupation_Salaried   VIF =  432.81
Occupation_Small_Business   VIF =  440.93
EducationField_Engineer   VIF =  18.07
EducationField_Graduate   VIF =  17.29
EducationField_MBA   VIF =  2.0
EducationField_Post_Graduate   VIF =  4.44
EducationField_UG   VIF =  1.57
EducationField_Under_Graduate   VIF =  2.58
Gender_Female   VIF =  4.77
Gender_Male   VIF =  4.54
Designation_Exe   VIF =  2.3
Designation_Executive   VIF =  8.62
Designation_Manager   VIF =  6.08
Designation_Senior_Manager   VIF =  2.82
Designation_VP   VIF =  1.84
MaritalStatus_Married   VIF =  1.92
MaritalStatus_Single   VIF =  1.89
MaritalStatus_Unmarried   VIF =  1.37
Zone_North   VIF =  19.24
Zone_South   VIF =  1.12
Zone_West   VIF =  19.21
PaymentMethod_Monthly   VIF =  2.22
PaymentMethod_Quarterly   VIF =  1.12
PaymentMethod_Yearly   VIF =  2.4
```

VIF Values (After Variables dropped)

```
Age   VIF =  1.4
CustTenure   VIF =  1.37
ExistingProdType   VIF =  3.73
NumberOfPolicy   VIF =  1.11
MonthlyIncome   VIF =  1.98
Complaint   VIF =  1.01
ExistingPolicyTenure   VIF =  1.11
SumAssured   VIF =  1.74
LastMonthCalls   VIF =  1.18
CustCareScore   VIF =  1.02
Channel_Online   VIF =  1.02
Occupation_Laarge Business   VIF =  1.58
EducationField_Engineer   VIF =  1.68
EducationField_Graduate   VIF =  1.26
EducationField_MBA   VIF =  1.04
EducationField_Post_Graduate   VIF =  1.09
EducationField_UG   VIF =  1.26
Gender_Female   VIF =  4.74
Gender_Male   VIF =  4.51
Designation_Exe   VIF =  1.19
Designation_Manager   VIF =  1.22
Designation_Senior_Manager   VIF =  1.29
MaritalStatus_Married   VIF =  1.92
MaritalStatus_Single   VIF =  1.88
MaritalStatus_Unmarried   VIF =  1.36
Zone_South   VIF =  1.01
Zone_West   VIF =  1.02
PaymentMethod_Monthly   VIF =  1.98
PaymentMethod_Quarterly   VIF =  1.1
PaymentMethod_Yearly   VIF =  2.11
```

Chart of Variable Inflations Factors

We see in the above case that there are many variables which expresses multicollinerity, they also have VIF value which is more than 5, so we dropped those variables and highlighted again in the next column.

10.1 Comparing Linear Model Results

The RMSE of Linear Model 1 on Train data is 608 and RMSE of Linear Model 2 on Train Data is 609.

The RMSE of Linear Model 1 on Test data is 633 and RMSE of Linear Model 2 on Test Data is 632

- So, we have observed that there is no significant changes in R Square or Root Mean Saure Values in both the interations, so this may not be the ideal way to choose the best model.

- It is required for us to check for different models like Decision Tree, Random Forest or Artifical Neural Network with base parameters and then compare their results to choose the best model.

10.2 Data Scaling

First we need to understand what Data Scaling is and why do we need it. Scaling is reltaed too the numeric features in the data.

We need Scaling because when we observe the numeric features in the data, the scale of the numeric features differes, and some of the algorithm are sensitive to this. They start giving higher weightage to the features that have higher values comparatively to the features who have smaller values.

Talking about the data that we have had, has the following observations:-

- We've observed that the features like 'Sum', 'Sum Assured', are carrying higher weightage, so in order to make our decision based on them we have to normalize the data and bring them to common scale using the Data Scaling method.
- As discussed abouve, it is true that Scaling does not impact the coefficient of attributes or its intercept values, however it is useful in reducing multicollinearity.

11. Different Models and their Scores (Base Parameter)

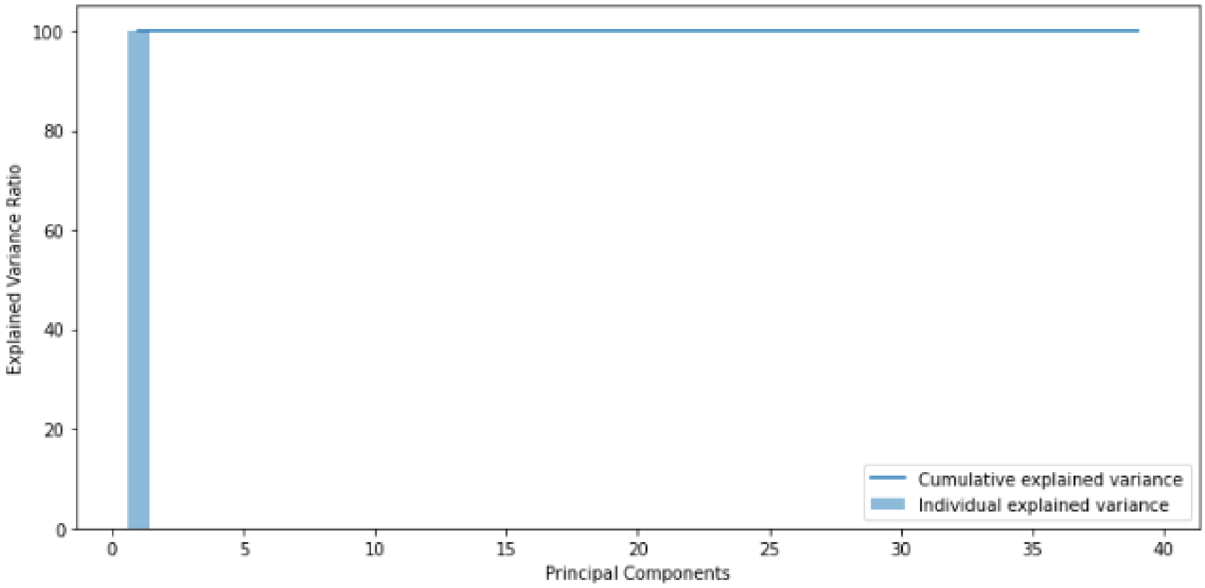
	Train RMSE	Test RMSE	Train Score	Test Score
Linear Regression	608.208934	590.392992	0.803620	0.798161
Decision Tree Regressor	0.000000	760.245991	1.000000	0.665318
Random Forest Regressor	190.483906	519.251378	0.980738	0.843773
ANN Regressor	497.488121	606.842724	0.868612	0.786756

Below are the observations and the analysis of Scores we have from different models on Train and Test data:

- So far we have observed that our Linear Model is performing better when compared to other models as the variation between the Train and test data is very minimum.
- If we compare the moidels, we find that most of the othe rmodels are in overfitting zone with respect to Linear Regression model.
- We have also observed that we are dealing with overfitting problems, and to overcome with that we can use hyperparameter tuning using Grid Search

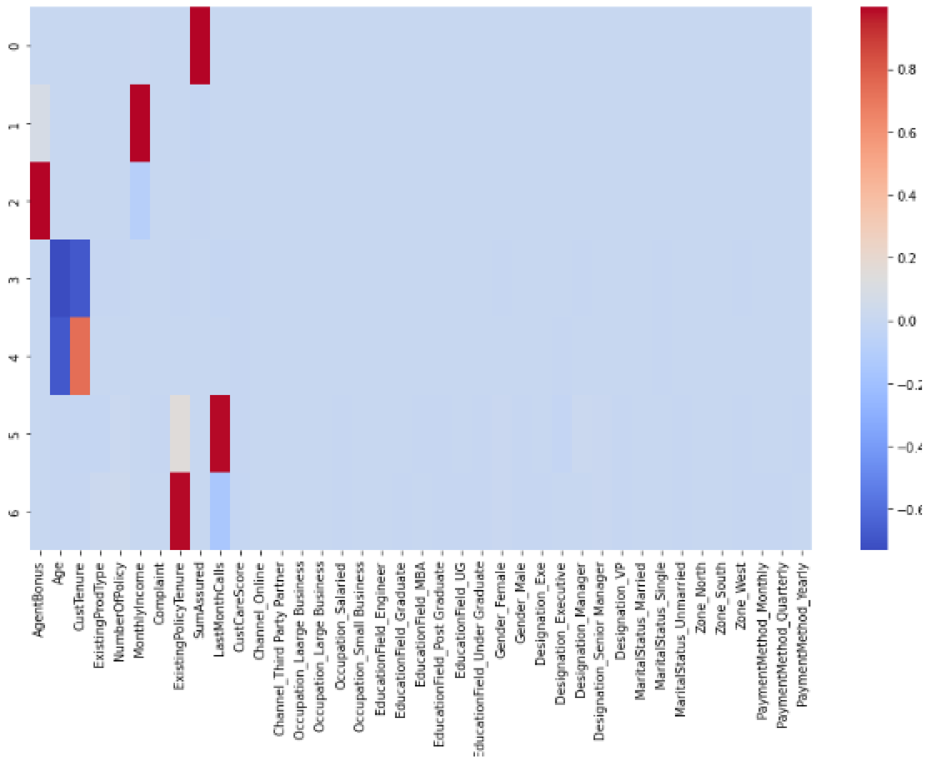
11.1 Checking if PCA can be applied:

```
Cumulative Variance Explained [ 99.97526512  99.99912394  99.99999975  99.99999985  99.99999995
  99.99999997  99.99999998  99.99999999  99.99999999  99.99999999
  99.99999999 100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         ]
```



Plot of Principal Components

11.2 Principle Components Vs Explained Variance Ratio: **(PCA Heat map)**



We see that not much can be observed about the components from the heatmap, hence dropping the need to perform PCA as all the variables holds almost the good deal of importance in the predictions.

Now we are going for Model Tuning but before that we have to understand what Model Tuning is and why is it used.

Model Tuning is basically the experimental process of finding the optimal values of hyper parameters to maximize the model performance. The purpose of model tuning a model is just to ensure that it performs at its best. This process involves adjusting various elements of the model to achieve optimal results. By fine tuning the model, we can actually maximise its performance without overfitting and reduce the variance error in our model.
Basically, there are different types of Model Tuning but here we are going to consider “Grid Search” technique to optimize best results.

Grid Search on Decision Tree

```
{'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 50}
```

Grid Search on Random Forest

```
{'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 300}
```

Grid Search on ANN

```
{'activation': 'relu', 'hidden_layer_sizes': 500, 'solver': 'adam'}
```

12. Different Models and their Scores (After Hyper Parameter Tuning)

	Train RMSE	Test RMSE	Train Score	Test Score
Linear Regression	608.208934	590.392992	0.803620	0.798161
Decision Tree Regressor	495.236882	573.495484	0.869798	0.809549
Random Forest Regressor	540.850048	583.163906	0.844709	0.803073
ANN Regressor	497.488121	606.842724	0.868612	0.786756

Observations drawn after the Hyper parameter Tuning are as follows:-

- We have observed that most of the variables in this data set has now moved out of overfitting zone.

- We have also observed that our model which is Linear Regression is still stable and the difference between Train and test data is minimal.
- And if the models accuracies are some what to be watched for, then we can clearly say that Random Forest Regressor model also does a good job as there is less than 5% difference between Train – Test data.
- Similarly, there is another and one more model which is ANN (Aerial Neural Network). This model also plays a pivotal role in target level prediction because after the Hyper Parameter Tuning we can see that the difference between the Train RMSE and Test RMSE score is minimum as compared to the base parameter.

12.1 Feature Importance:

We see the "Sum Assured" as our most important feature and on the second hand Geographical locations which is Zone South and East being the least.

Serial No	Feature	Feature Importance
1	Sum Assured	0.430643
2	Customer Tenure	0.147939
3	Age	0.135314
4	Monthly Income	0.122517

13. Interpretations and Recommendations

<u>Interpretations :</u>

- The Company basically wants to predict the ideal bonus for agents and the level of engagement of high and low performing agents respectively.
- The Customers who has high designations has a higher chances of buying more policies like if the Designation is Vice President of any company or a person with higher designation ranks then that person is going to buy more policies.
- Hence, for high performing agents we can create a healthy contest and for low performing agents what better can be done is, we can train them, or suggest them to purchase or get policies with high sum assured as it is very significant to our model.
- There is one more important feature which is Customer Tenure where the agents need to focus on the customer policy tenure which is ranging between 8 – 20 and this is where the majority of customer falls.
- Now keeping an eye on those customers who has higher monthly incomes because the higher monthly incomes the Customer has, the greater possibilities of Customer to buy higher valued policies.

- And through the models, we have the agents who are the higher performer. For them we have a few variables which is quite significant for them which is 'Sum Assured',

Recommendations :

- Firstly, we have to see for high performing agents, for them we can create a healthy contest with a threshold, like if they achieve the target like the desired 'Sum Assured' then they are eligible for certain perks and incentives like exotic family vacation packages, latest gadgets and many more.
- Secondly, we have to look for low performing agents as well, for them we can upgrade or introduce some feedbacks or up skill certain programs to train them generating higher 'Sum Assured' policies, reaching out to certain people to ultimately becoming the higher or top performers.
- As we also know that we have one column named as a 'Zone'. For high performing agent, they are performing good but form the business point of view we see that there are very less sales in South and East zone. So high performing agents can focus to acquire more customers from those regions.
- We can also add another predictor as our Customer geographical location or Regions not just the zones but also the people who usually lives in rural areas or remote areas are less likely to buy a policy whereas those living in highly developed location are likely to be belonging to the upper class should be targeted.
- Similarly, there can be another predictor like "AgentID" which can be introduced that will help us to make it easier to observe the high and low performing agents and their trends.
- The amount or the policy premium collected from the customers acts as a very good predictor in terms of analysing the agent bonus, which gives us the real insights towards the monetary business agent who is doing on regular basis.

** Thank you **