FINAL CAPSTONE PROJECT

BUSINESS REPORT - SUBMISSION 1

Life Insurance Sales

PGP- DSBA (Feb Batch: 2022-2023)

BY SASHWAT RAJ BAKSHI

1. Introduction of the business problem

Problem Statement: Life Insurance Data

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

Introduction to problem statement

The data given is a company's internally acquired data, which is provided tous by a leading Life Insurance organization located in India.

We have been provided with the data with the objective of establishing a model for the organization which helps us in predicting the Bonus given tothe agents when they sell policies and onboard customers (with or withouttheir families) with one/multiple Life Insurance Policies.

Idea behind making a model is to understand the most important predictors in deciding the agent bonuses. With the help of this knowledge, we will be able to motivate higher performing agents by better incentivisation and also help their underperforming peers with strategies that may involve up-skilling.It may also involve the process of auditing their sales meets and calls to better their performance from a selling perspective.

Need of the Study

Understanding what type of bonuses should be allotted with appropriatesales is an important step towards building a high performance - high rewarding sales team.

A high performance - high reward environment also builds a strong internal team, which has a higher chance of retention, mainly due to the growth andrecognition an Agent or Employee receives for his/ her performance. Properrecognition of agent's contribution leads to agent feeling more satisfied andmotivated to work harder and better.

Better incentivisation leads to creating performance oriented employees andfuture leaders within the organization.

Good Performance also leads to internal appraisals and for a sales team, agood internal growth story is a benchmark which low-medium sales performers can strive for growth and thus seek improvement.

As for the low performing counterparts, up-skilling plans, call/visit shadowingplans (Where a low performing employee is grouped with the higher performing employee, with the desired objective of mirroring things like a solid sales pitch, asking the right questions to the customer, building an understanding of customer profile, ideal conversion atmospheres, even prospecting and pursuing the more likely customer, or evensimply - effective time management).

At the end of the day Sales is about providing an effective solution to customer’s problem in the best possible way.

Business Opportunities:

1. Better Bonuses, Higher Performance, more sales which in turn wouldmean more business to the organization.
2. A better Brand value can be expected considering higher volumes of sales being good and Ideal outcome would be the company's name becoming a household name in India for Life Insurance.
3. Higher Profits can be expected with increase in sales.
4. Increase in market share of the Life insurance company.

Social opportunities:

1. Increase in awareness of the benefits of Life Insurances.
2. Ideal outcome from a social perspective, would mean that more and more lives getting insured leading to financial security in case of an unfortunate mishappening
3. Life Insurance policies is the only security families have, when there is asingle bread winner in the household

Data Dictionary

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

Fig 1: Data Dictionary

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

Fig 2: Head of the data

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	Mc
4515	7004515	3953	4.0	8.0	Agent	Small Business	Graduate	Male	4	Senior Manager	2.0	Single	
4516	7004516	2939	9.0	9.0	Agent	Salaried	Under Graduate	Female	2	Executive	2.0	Married	
4517	7004517	3792	23.0	23.0	Agent	Salaried	Engineer	Female	5	AVP	5.0	Single	
4518	7004518	4816	10.0	10.0	Online	Small Business	Graduate	Female	4	Executive	2.0	Single	
4519	7004519	4764	14.0	10.0	Agent	Salaried	Under Graduate	Female	5	Manager	2.0	Married	

Fig 3: Tail of the data

2. Data Report

Non-Visual understanding of the data

First, we understand how the data was collected in terms of time, frequencyand methodology.

Methodology: In the initial inspection, the data seems to be collected after the Customer/ Customer's Family has purchased the Life insurance Plan or after the acquisition of the customer has been done.
Source: The source of the data is internal, and is given by the organization(Primary Data Source).
Time and Frequency: Data is collected at a monthly frequency after customer is on boarded with at least one insurance policy.

Approach for Exploratory Data Analysis:

1. We load the data dictionary and understand it thoroughly.
2. We then check if the data is loaded correctly using head and tail function.
3. We check the shape using the shape function.
4. Using the info function, we gather information around the Data Impurities(specifically Null Values) and the data type of each column.
5. We check the null values column wise and their total.
6. Check for duplicate values using duplicated function.
7. Describe function is used to understand the statistical side of our data inpertinence to distribution and count of the data.

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

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
CustID	4520.0	NaN	NaN	NaN	7002259.5	1304.955938	7000000.0	7001129.75	7002259.5	7003389.25	7004519.0
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

Fig 4: Basic data description

Insights from Visual Inspection:

1. The data has a total of 20 Features (Columns) and 4520 Observations inthe form of Customer Data points (Rows)
2. We have 12 Numeric Data Features and 8 Object Data Type Features in ourdata. We will have to convert data of object type to numeric considering this being a predictive problem statement. However, we will explore the data more in depth before doing so.
3. 1.29% of our data is with null values which we will explore in the dataanalysis segment.
4. There are no duplicate values in the data.

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
```

Fig 5: General Information about the data

```
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
dtype: int64
```

Fig 6: Null values in the dataset

There are a total of **1166** Null Values in our data as shown above:

1. Age has 269 Null Values
2. CustTenure has 226 NullValues
3. NumberOfPolicy has 45 NullValues
4. MonthlyIncome has 236 NullValues
5. ExistingPolicyTenure has 184Null Values
6. SumAssured has 154 NullValues
7. CustCareScore has 52 NullValues

Numerical Features total – 12

'CustID', 'AgentBonus', 'Age', 'CustTenure', 'ExistingProdType', 'NumberOfPolicy', 'MonthlyIncome', 'Complaint', 'ExistingPolicyTenure', 'SumAssured', 'LastMonthCalls', and 'CustCareScore' are Numerical.

Categorical Features total – 7

Channel', 'Occupation', 'EducationField', 'Gender', 'Designation','MaritalStatus', 'Zone'

and ‘PaymentMethod' are Categorical.

Renaming:

1. We observed that there is a data entry error in the Gender column where'Female' is mis-spelled as 'Fe male'.
2. We also observe that in the case of Designation, 'Exe' seems to be equivalent to 'Executive' - given that the Monthly Income also falls in the same bucket. We will go ahead and change 'Exe' to 'Executive'.
3. In column EducationField 'UG' and 'Undergraduate' are the same meaning.We will go ahead and change UG to Undergraduate.

3. EXPLORATORY DATA ANALYSIS

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

Presence of Outliers:

These are the following variables with outliers and we check them visuallywith the Boxplot visualization:

1. Monthly Income
2. Existing Policy Tenure
3. Existing Product Type
4. Sum Assured
5. Age
6. Agent Bonus
7. Customer Tenure
8. Last Month Calls

Skewness:

In statistics, skewness is a measure of the asymmetry of the probability distribution of a random variable about its mean. In other words, skewness tells you the amount and direction of skew (departure from horizontal symmetry). The skewness value can be positive or negative, or even undefined. If skewness is 0, the data are perfectly symmetrical, although it isquite unlikely for real-world data. As a general rule of thumb:

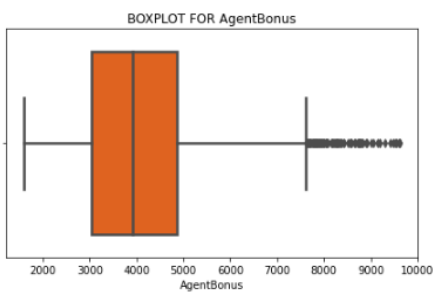
	NULL COUNT	ZERO COUNT	NEGATIVE COUNT	MIN	MAX	RANGE	VARIANCE	IQR	SKEWNESS	KURTOSIS
CustID	0.0	0.0	0.0	7000000.0	7004519.0	4519.0	1.702910e+06	2260.10	0.00	-1.20
AgentBonus	0.0	0.0	0.0	1605.0	9608.0	8003.0	1.969312e+06	1840.10	0.82	0.71
Age	269.0	0.0	0.0	2.0	58.0	56.0	8.168000e+01	14.00	0.94	0.84
CustTenure	226.0	0.0	0.0	2.0	57.0	55.0	8.035000e+01	13.00	0.93	0.85
ExistingProdType	0.0	0.0	0.0	1.0	6.0	5.0	1.030000e+00	1.00	-0.40	0.60
NumberOfPolicy	45.0	0.0	0.0	1.0	6.0	5.0	2.120000e+00	3.00	-0.10	-0.89
MonthlyIncome	236.0	0.0	0.0	16009.0	38456.0	22447.0	2.386909e+07	5433.75	1.36	1.58
Complaint	0.0	3222.0	0.0	0.0	1.0	1.0	2.000000e-01	1.00	0.94	-1.11
ExistingPolicyTenure	184.0	0.0	0.0	1.0	25.0	24.0	1.120000e+01	4.00	1.54	2.75
SumAssured	154.0	0.0	0.0	168536.0	1838496.0	1669960.0	6.063159e+10	339618.60	0.97	1.25
LastMonthCalls	0.0	408.0	0.0	0.0	18.0	18.0	1.311000e+01	6.00	0.81	0.17
CustCareScore	52.0	0.0	0.0	1.0	5.0	4.0	1.910000e+00	2.00	-0.14	-1.13

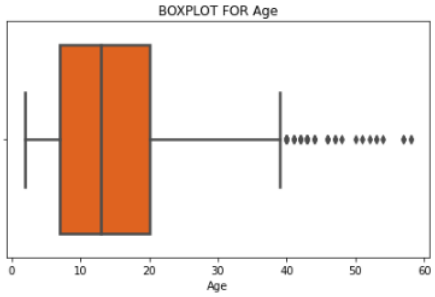
Fig 7: Advance data description

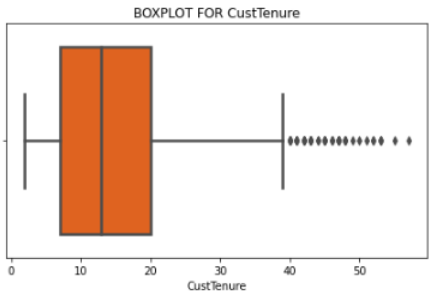
Convention:

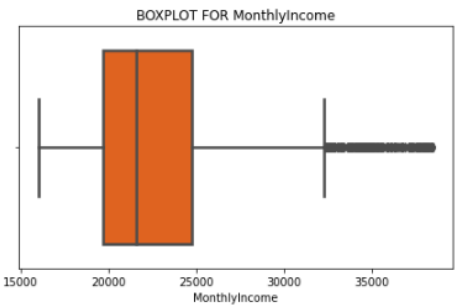
Positive Skewness - Right Skewed
Negative Skewness - Left Skewed
If skewness is less than -1 or greater than 1, the distribution is highly skewed.
If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

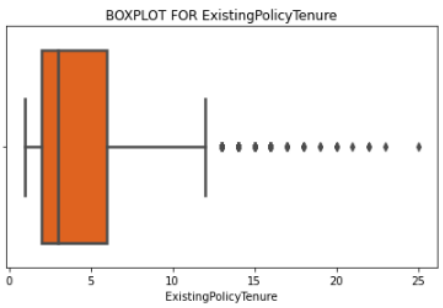
Boxplot:

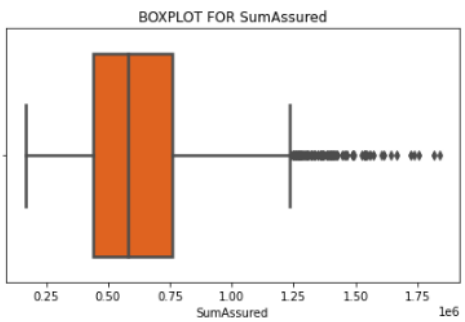












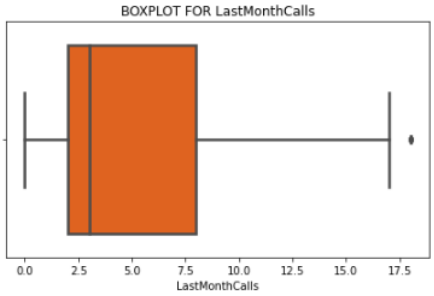


Fig 8: Boxplot of all the numeric variables

Distribution Plot

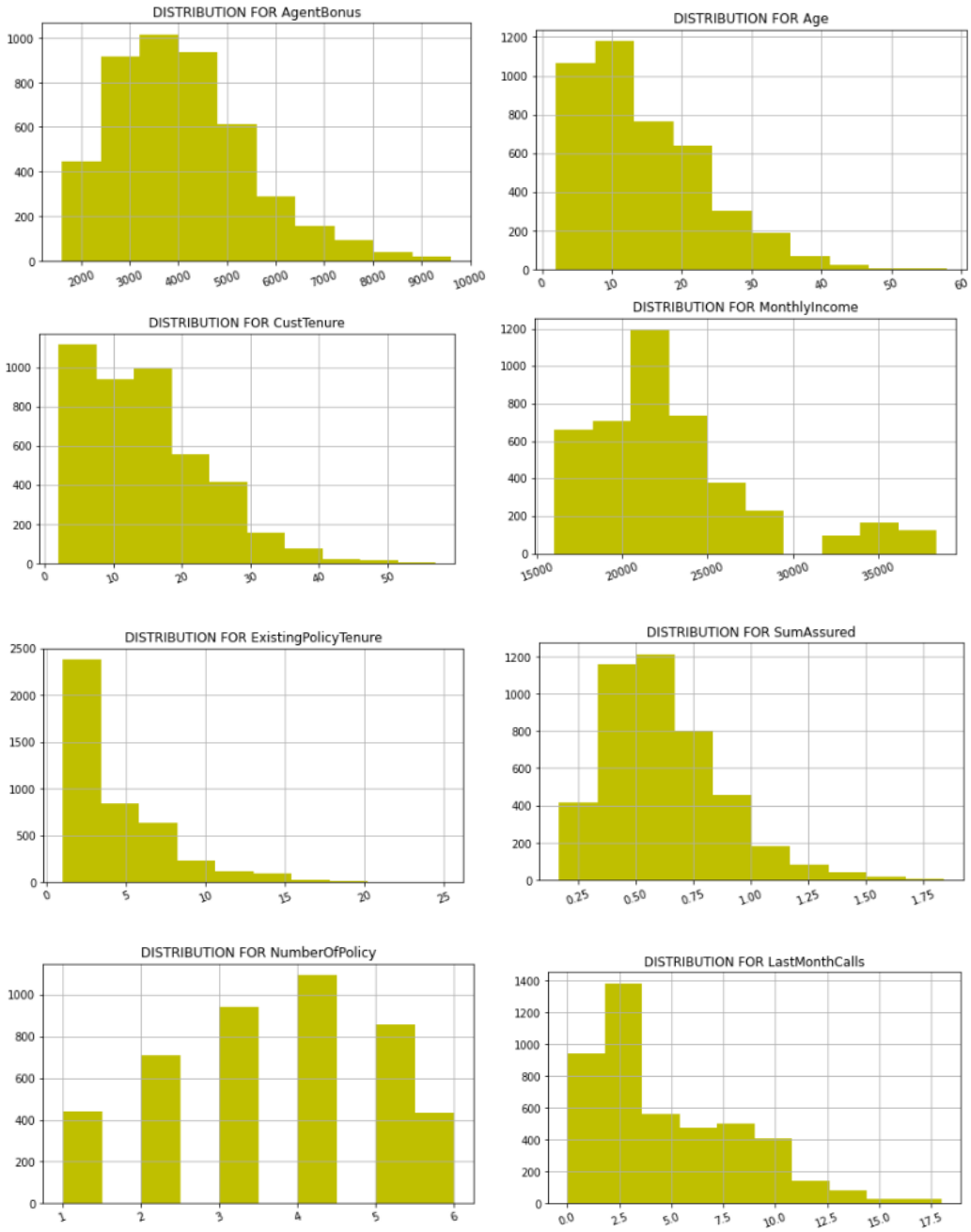


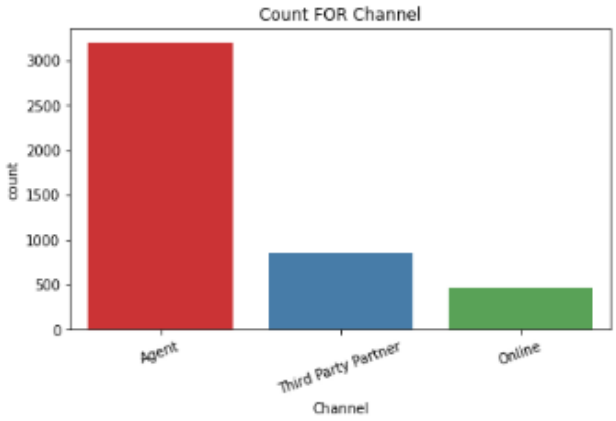
Fig 9: Distribution plot for all the numeric variables

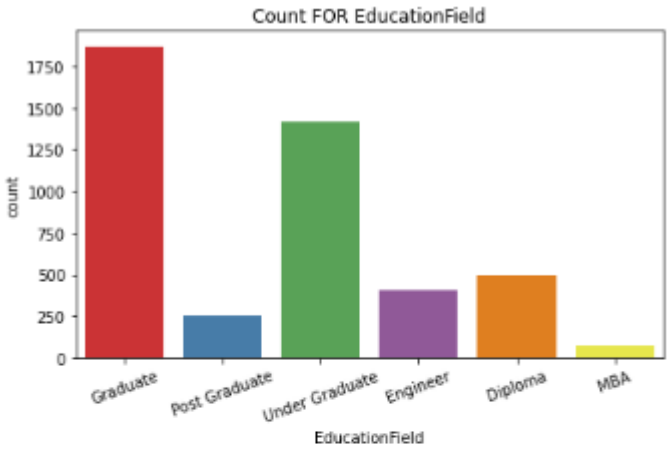
Inferences:

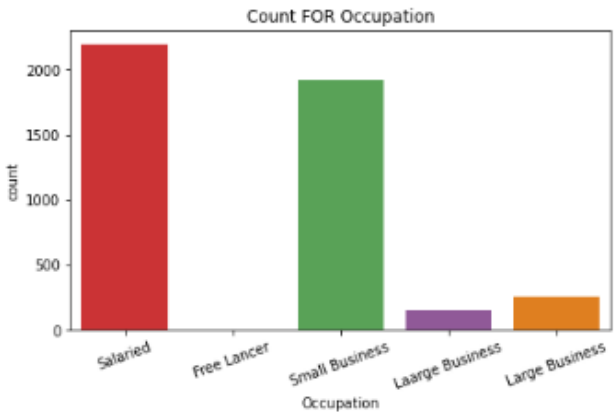
We can make the following observations with the help of the above table and boxplot/distributiongraphs below (Figure 2 and Figure 3):

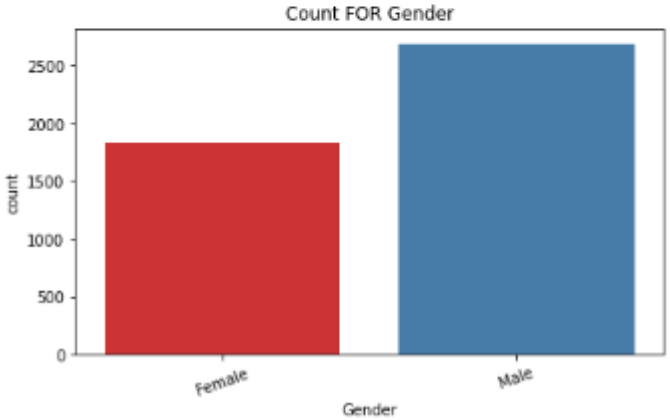
1. The distribution of the variable 'Agent Bonus' is moderately right skewed.
2. The distribution of the variable 'Age' is moderately right skewed.
3. The distribution of the variable 'Customer Tenure' is moderately right skewed.
4. The distribution of the variable 'Existing Prod Type' is approximately symmetric.
5. The distribution of the variable 'Number Of Policy' is approximately symmetric.
6. The distribution of the variable 'Monthly Income' is highly right skewed.
7. The distribution of the variable 'Complaint' is moderately right skewed.
8. The distribution of the variable 'Existing Policy Tenure' is highly right skewed.
9. The distribution of the variable 'Sum Assured' is moderately right skewed.
10. The distribution of the variable 'Last Month Calls' is moderately right skewed.
11. The distribution of the variable 'Customer Care Score' is approximately symmetric.
12. There are no negative values in our data.

Univariate Analysis for categorical variables:









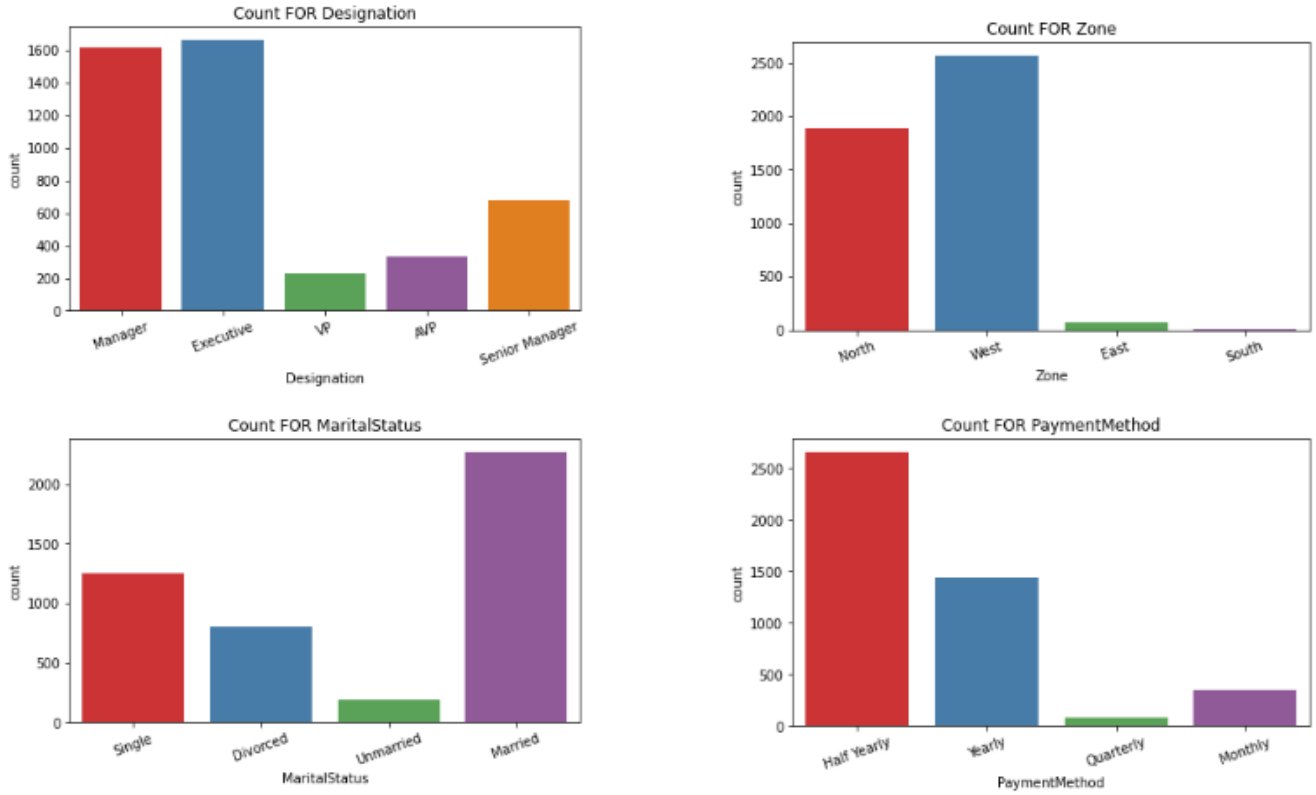


Fig 10: Categorical Variables count plot

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


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


EDUCATIONFIELD : 6
MBA                  74
Post Graduate       252
Engineer            408
Diploma             496
Under Graduate     1420
Graduate           1870
Name: EducationField, dtype: int64


GENDER : 2
Female    1832
Male      2688
Name: Gender, dtype: int64
```

```
DESIGNATION : 5
VP                226
AVP               336
Senior Manager    676
Manager          1620
Executive        1662
Name: Designation, dtype: int64


MARITALSTATUS : 4
Unmarried     194
Divorced      804
Single       1254
Married      2268
Name: MaritalStatus, dtype: int64


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


PAYMENTMETHOD : 4
Quarterly        76
Monthly         354
Yearly         1434
Half Yearly    2656
Name: PaymentMethod, dtype: int64
```

Fig 11: Count for categorical variables

Inferences for Visual Analysis of Categorical Variables:

1. Sales made by Agents dominate the other two channels (i.e. Third Party Partner and Online Channels).

2. Major Customers are from families where the main source of income is Salaried and Customers who operate a small business.
3. Major chunk of the Customers have an educational qualification of Graduate followed by Under Graduate.
4. The observations for the gender Male is about 2500 and Female is about 1800.
5. Most of the customers are Manger, Executive and Senior Manger Level.
6. Most of the people are working in Managerial or Executive level positions, followed by people working as Senior Managers.
7. More than half of the customers who have purchased life insurance plans are married or divorced.
8. There is a higher company presence in Northern and Western India, where as sales numbers in the Eastern and Southern India is very poor .
9. Most of the people opt for Half Yearly and Yearly plans.

Bivariate Analysis

Correlation is a statistical measure that expresses the extent to which two variables are linearly related (meaning they change together at a constant rate).

Convention:

Correlation coefficients whose magnitude are between 0.7 and 1 indicatevariables which can be considered highly correlated.
Correlation coefficients whose magnitude are between 0.5 and 0.7 indicatevariables which can be considered moderately correlated.
Correlation coefficients whose magnitude are between 0.3 and 0.5 indicatevariables which have a low correlation.
Correlation coefficients whose magnitude are less than 0.3 have little if any(linear) correlation.

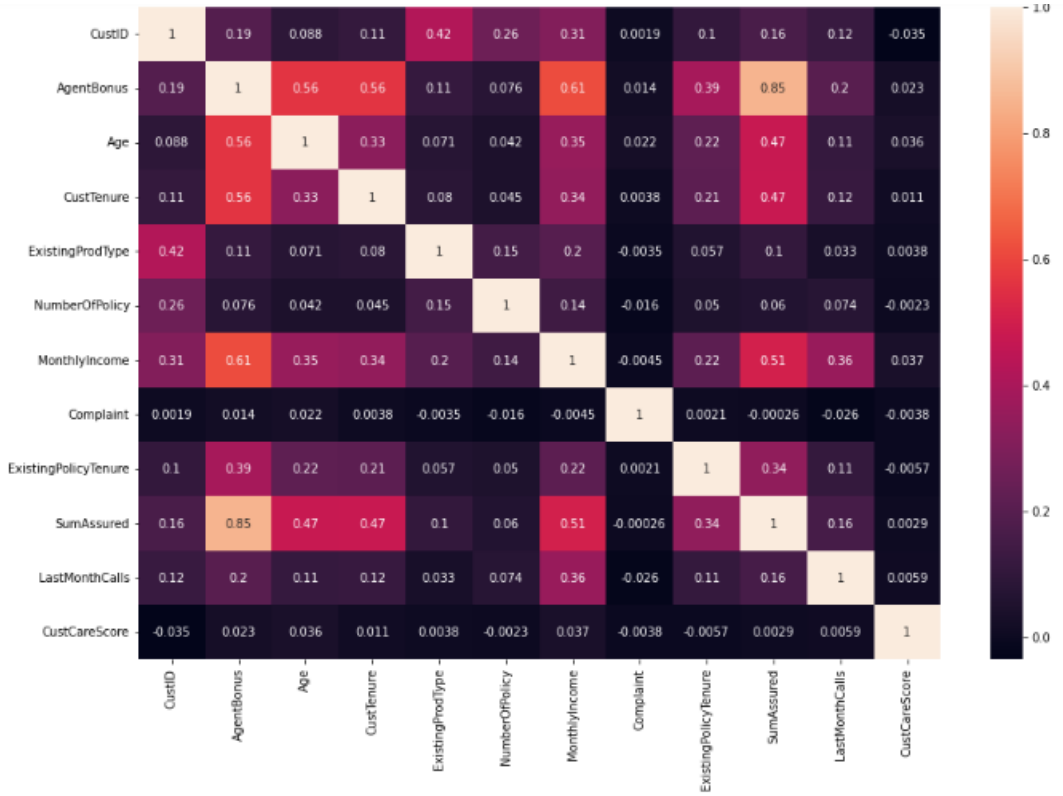


Fig 12: Heatmap

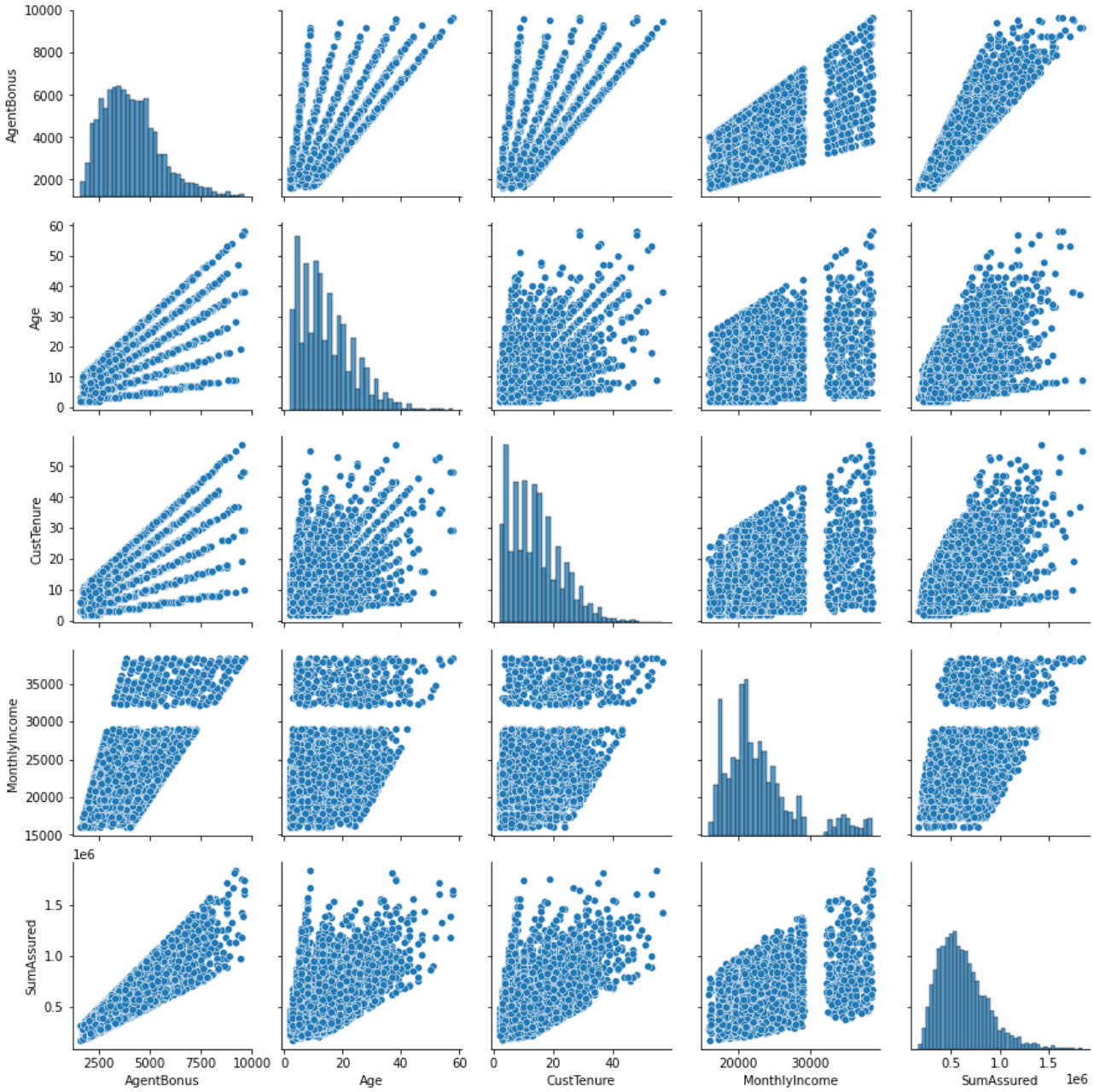


Fig 13: Pairplot

We can clearly draw the following inferences from the above heatmap and the following table (printed correlation higher than 0.7 that the Sum assuredis highly correlated to Agent Bonus (Target Variable).

Multicollinearity is a statistical concept where several independent variablesin a model are correlated. Two variables are considered to be perfectly collinear if their correlation coefficient is +/- 1.0. Multicollinearity amongindependent variables will result in less reliable statistical inferences.
In our data, we do not have any highly correlated independent variables.

We can also see how two continuous variables are associated with eachother with the help of the following pair plot below.

We can clearly see that Agent Bonus (our target variable) is positively correlated with Age, Customer Tenure, Monthly Income and Sum Assured.

Removal of unwanted variables

We see that the Customer ID cannot be a significant predictor since it is a unique identifier. Unique observations in general are not significant when itcomes to predictive modeling. We go ahead and remove it using the drop function.

Missing Value treatment

Approach:

If there are several or large numbers of data points that act as outliers. Outliers data points will have a significant impact on the mean and hence, in such cases, it is not recommended to use the mean for replacing the missing values. Using mean values for replacing missing values may not create a great model and hence gets ruled out. For symmetric data distribution, one can use the mean value for imputing missing values. Since we don't have a large (> 15% of missing values) in our data, we will go ahead and impute missing values in the following manner:

For Age, MonthlyIncome, SumAssured, CustomerTenure have Outliers, we will impute using Median.

For ExistingPolicyTenure, CustCarescore and Number of policy do not have Outliers; we will impute using the Mode since they are taking whole values.

Outlier treatment

Approach:

For Outlier Treatment We usually have 3 ways we can approach them:

1) Retain the outliers if data occurrence seems to be genuine.
2) Drop or remove the outliers.
3) Capping these data points to the upper and lower limit of the distribution,to ensure they lie within range of data points.

After carefully examining the features with outliers present, we reach a conclusion of not removing/treating these outliers as they seem to be genuine observations. All of these data points existing as outliers hold high importance for analysis.

	% OUTLIERS
MonthlyIncome	8.50
ExistingPolicyTenure	7.63
ExistingProdType	6.77
SumAssured	2.43
Age	2.32
AgentBonus	2.21
CustTenure	2.15
LastMonthCalls	0.27
NumberOfPolicy	0.00
Complaint	0.00
CustCareScore	0.00

Fig 14: Features with outliers (in %)

Monthly Income: Income is subjective and different for different individuals and is a genuine observation. We should not treat these outliers with high Monthly Income.

Existing Policy Tenure: Tenure can be dependent on multiple factors. Though most individuals lie between 1-5, there are customers on boarding for policies with higher tenures as well.

Existing Product Type: The type of products are categorized into 6 Categories. This is a data point which exists in numeric format however is categorical and need not be treated for outliers.

Sum Assured: Sum Assured may be different for different policies within the system and again, should not be dropped as they hold high importance in the analysis.

Age: In this case there are people with above 35 Age onboard with the insurance company which could be a part of a family package or even would have on boarded at a younger age and have been in the system.

Agent Bonus: Higher sales performers may receive higher bonus and since this is the target variable, this information of higher performing agents getting higher bonuses, is of utmost importance to us.

Customer Tenure: Again, similar to age, this can be a data point which indicates customers tenure with the company in the system, and completely normal.

Last Month Calls: Total calls attempted by company to a customer for cross sell can be high in case of a follow-up attempt made by an agent. This can be analyzed further. If done by high performing agent then deduction can be made that more number of follow ups usually lead to more cross selling.

Variable transformation

Categorical Variable Transformations:

Linear Regression models can only handle numeric variables. Therefore, if a column has categorical variable, models will only work when it is encoded tonumeric variable.

There are two types of categorical data:

Ordinal: Order based Payment Method
Nominal: Without any order or ranks like city Channel, Marital Status, Zone

Designation - Order can be on the basis from customers with higher designation to lower designation conventionally (i.e. VP > AVP > Senior Manger > Manager > Executive). All other features (Channel, Occupation, Education Field, Gender, Marital Status, Zone and Payment Method) will be considered as nominal for further steps.

In the case of Educational field we can do so too, however we see separate classes of information.

For example Post Graduation and MBA.

MBA is a Post Graduation however not all customers may have done an MBA. Hence, ranking them according to order may become difficult.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   AgentBonus            4520 non-null   int64
 1   Age                   4520 non-null   float64
 2   CustTenure            4520 non-null   float64
 3   Channel               4520 non-null   int8
 4   Occupation            4520 non-null   int8
 5   EducationField        4520 non-null   int8
 6   Gender                4520 non-null   int8
 7   ExistingProdType      4520 non-null   int64
 8   NumberOfPolicy        4520 non-null   float64
 9   MaritalStatus         4520 non-null   int8
 10  MonthlyIncome         4520 non-null   float64
 11  Complaint             4520 non-null   int64
 12  ExistingPolicyTenure  4520 non-null   float64
 13  SumAssured            4520 non-null   float64
 14  Zone                  4520 non-null   int8
 15  PaymentMethod         4520 non-null   int8
 16  LastMonthCalls        4520 non-null   int64
 17  CustCareScore         4520 non-null   float64
 18  DesignationOrdinal    4520 non-null   int64
dtypes: float64(7), int64(5), int8(7)
memory usage: 454.8 KB
```

Fig 15: Variable after transformation

All features have been successfully converted to numeric data types.

The scales of each x-variable need to be specified, because linear regression is heavily dependent on scale. If we were to train on a dataset where length is measured in feet and another measured in miles, the performance of both will be the same but the coefficients will differ.

We will go ahead and scale the entire data using Z score.

Z-score is a variation of scaling that represents the number of standard deviations away from the mean. You would use z-score to ensure your feature distributions have mean = 0 and std = 1. It's useful when there are a few outliers, but not so extreme that you need clipping.

Addition of new variables

At this moment we do not see the need to add a New Variable given theexisting features provided to us in the data set.

4. Business insights from EDA

Before we jump onto the insights, we need to check if the data and predictorvariables are balanced?

In this case since our target variable is continuous and our predictor variables are a mix of continuous and categorical, we will use Linear Regression for our model building exercise.

Linear regression being a model which predicts continuous variables, doesnot require data balancing (which is usually required by models which are based on classification).

Hence, we do not have a need of processing the data through a databalancing methodology.

Data Balancing:

In this case, since our target variable is continuous and our predictor variables are a mix of continuous and categorical.
We will use Linear Regression for our model building exercise.
Linear regression being a model which predicts continuous variables does not require data balancing (which is usually required by models which are based on classification).

Hence, we do not have a need of processing the data through a data balancing methodology.

Here are data imbalance observations for categorical variables. We need to keep these in mind when these variables come up as important predictors as there will be some features (eg: in Zone- there is very less data in Eastern and Southern Zone).

1. Sales made by Agents dominate the other two channels (i.e. Third Party Partner and Online Channels).
2. Majority Customers are from families where the main source of income is Salaried and people who operate a small business
3. Majority of the Customers have an educational qualificaiton of Graduate followed by Under Graduate
4. The obsevations for the gender Male is about 2500 and Female is about 1800
5. Most of the customers are Manger, Exececctive and Senior Manger Level
6. Most of the people are working in Managerial or Executive level positions, followed by people working as Senior Managers
7. More than half of the customers who have purchased life insurance plans are Married or Divorced
8. There is a higher company presence in Northern and Western India, where as sales numbers in the Eastern and Southern India is very poor.
9. Most of the people opted for Half Yearly and Yearly plan.

Business Insights Using Clustering:

We have performed Hierarchical Clustering.

Parameter used for Linkage is Average Linkage.

In average linkage, we define the distance between two clusters to be the average distance between data points in the first cluster and data points in the second cluster. On the basis of this definition of distance between clusters, at each stage of the process we combine the two clusters that have the smallest average linkage distance.

fcluster:

It is a methodology used to flatten the clusters.

We have used number of Clusters as 3 and the criterion as maxclust.

Maxclust: The maxclust criterion is a distance threshold; it rejects far points from being in the same cluster.

Insights:

Before we move onto insights from clustering, we have used the mean values to understand how each numerical continuous feature is in the Cluster.

For the categorical (Nominal and Ordinal), since we have encoded them, it is better if we look at the mode as it will give us the variable that has occured the most in each cluster.

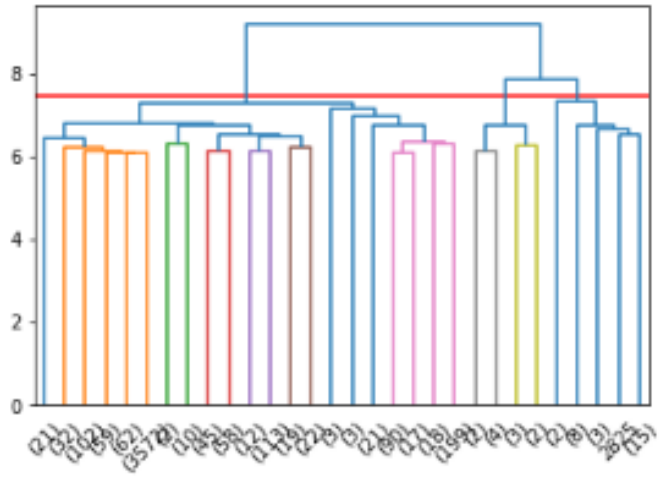


Fig 16: Dendogram for clustering

Number of Clusters - 3:

Using Number of Clusters as 3

Since, we can see from the above dendogram that major chunk of data can be bucketed into 6 such categories where we can segment the customer bucket.

We can also see that the distance between each of these 3 i.e. the volume of customers in each bucket is similar and 3 look optimal visually as well.

The number of clusters is the number of vertical lines in the dendogram cut by a horizontal line that can transverse the maximum distance vertically without intersecting a cluster.

In this case, we cut the model where we get 3 clusters.

Clusters=3	1	2	3
AgentBonus	4040.94	7059.82	8646.52
Age	14.26	19.73	34.76
CustTenure	14.21	28.27	38.41
NumberOfPolicy	3.56	3.18	4.26
MonthlyIncome	22711.33	31654.36	36764.24
ExistingProdType	3.68	2.82	4.72
ExistingPolicyTenure	4.06	19.45	8.57
Complaint	0.29	0.36	0.31
SumAssured	613056.77	997513.27	1331521.10
LastMonthCalls	4.61	5.18	6.79
CustCareScore	3.07	3.00	3.31

Fig 17: Clustering Result

Insights for Cluster 1 (Number of Customers in Cluster 1 - 4480):

Target Variable: Average Agent Bonus given for this cluster of customers is almost half of what is to the other 2 clusters (Average Agent Bonus of 4040.94)

Average Age of customer is 14

Average Tenure of Customer is the least amongst all 3 (around 14)

Existing Product Type (rounded off since original values are integers) - 4

Number of Policy (rounded off since original values are integers) - 4

Average Monthly Income of customers is lowest amongst the 3 clusters (22711.33)

Average number of Complaints is lowest of all (0 and 1)

Average Existing Policy Tenure is 4 which is on the lower side (Range 1-21)

Average Sum Assured is the least as compared to all 3 clusters (613056.77)

Average of Last Month Calls is the least to these customers in cluster 1 - 5 (Range of calls is 0 - 16)

Average Customer Care Score is 3 out of 5 which is average (Range is 1-5)

Insights for Cluster 2 (Number of Customers in Cluster 2 - 11):

Target Variable: Average Agent Bonus given for this cluster of customers is on the higher side (Average Agent Bonus of 7059.82)

Average Age of customer is 20

Average Tenure of Customer is the in the middle amongst all 3 (around 28)

Existing Product Type (rounded off since original values are integers) - 3

Number of Policy (rounded off since original values are integers) - 3

Average Monthly Income of customers is in between cluster 1 and cluster 3 (31654.36)

Average number of Complaints is highest of all 3 clusters (0 and 1)

Average Existing Policy Tenure is 19 and is the highest in cluster 2 (Range 1-21)

Average Sum Assured is in between cluster 1 and cluster 2 (997513.27)

Average of Last Month Calls is the least to these customers in cluster 2 - 5 (Range of calls is 0 - 16)

Average Customer Care Score is 3 out of 5 which is average (Range is 1-5)

Insights for Cluster 3 (Number of Customers in Cluster 3 - 29):

Target Variable: Average Agent Bonus given for this cluster of customers is on the highest amongst the 3 clusters (Average Agent Bonus of 8646.52)

Average Age of customer is 35

Average Tenure of Customer is the highest amongst all 3 (around 38)

Existing Product Type (rounded off since original values are integers) - 5

Number of Policy (rounded off since original values are integers) - 4

Average Monthly Income of customers is highest amongst the 3 clusters (22711.33)

Average number of Complaints is between cluster 1 and cluster 2 (0 and 1)

Average Existing Policy Tenure is 9 (Range 1-21)

Average Sum Assured is the highest as compared to all 3 clusters (1331521.10)

Average of Last Month Calls is the least to these customers in cluster 3 - 7 (Range of calls is 0 - 16)

Average Customer Care Score is 3 out of 5 which is average (Range is 1-5)

Business Implications

1. Agent Bonus is highest for cluster where average tenure of customer is highest indicating that reselling policy is strong for these customers. Thisalso is confirmed by the average number of calls made by agents as it is highest across all 3 customers.
2. Agent Bonus is highest for customer policies where average age of customer is higher. Long term customers are bringing in a good policy/reselling.
3. Sum Assured is also higher to customers who have a higher average age indicating trend. Higher age customers can be pursued more rigorously to up sell policies.
4. Monthly income is highest for customers where the sum assured is higherindicating right kind of follow-up done on cream customers which is good time management. Lower performing sales agents can learn this to maximize company realization.

Insights from EDA

1. Sales made by Agents dominate the other two channels (i.e. Third PartyPartner and Online Channels).
2. Majority Customers are from families where the main source of income isSalaried and people who operate a small business.
3. Majority of the Customers have an educational qualification of Graduatefollowed by Under Graduate.
4. The observations for the gender Male is about 2500 and Female is about1800.
5. More than half of the customers who have purchased life insurance plansare Married or Divorced indicating family members purchasing plans in higher volumes than independent people.
6. There is a higher company presence in Northern and Western India, whereas sales numbers in the Eastern and Southern India is very poor. Surveys can be done and improvements can be thought of for low performing regions.

- - - END - - -