LIFE INSURANCE CASE STUDY

(CAPSTONE Project Notes -1)

JULY 31, 2022

ALIND KHANNA
GREAT LAKES

Table of Contents

DATA DICTIONARY

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

LIFE INSURANCE CASE STUDY

Introduction

Insurance is one of the vital products for both business and human life. It provides necessary financial support in case of uncertainties also it safeguards against unpredictable events.

It gives necessary cover and peace of mind against any catastrophic events which are not even in control of human being.

Defining of the business problem

As an analyst we have been assigned the role to predict the bonus for the agents in order to reward them for their accomplishments and also upskill the lower performing agents. We have been provided with life insurance 'Sales' data set of the leading insurance company.

Data Set: Sales

No of Records: 4520

We need to dive deeply in the given data set and extract meaningful insights, this would in turn help the company to determine the bonus strategies for high performing agents.

Need of Study/Project

Bonus is 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.

Company, generally divides bonus based on agent's performance in various bonus categories.

This helps in boosting company's market share.

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 for their performances. Social initiative for the product is covering as many lives

possible under the ambit of life insurance policies, so as the person doesn't face any difficulty addressing any financial obligations arising due to catastrophic life events.

Data report

- Data has been provided us academically so we consider it as a primary source data
- On analysing data, we come to know data has been collected on a monthly basis, which is clearly depicted by the variable 'LastMonthCalls' in given data set. This variable tells us about how many calls were made on monthly basis to the given customer.
- Methodology used is depicted by variable 'Channel' which shows by which way the customer was acquired (Agent/Third Party/online)

Visual data inspection

- Data set consists of 4520 rows and 20 columns.

First 5 rows of given data set

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

MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	Zone	PaymentMethod	LastMonthCalls	CustCareScore
20993.0	1	2.0	806761.0	North	Half Yearly	5	2.0
20130.0	0	3.0	294502.0	North	Yearly	7	3.0
17090.0	1	2.0	NaN	North	Yearly	0	3.0
17909.0	1	2.0	268635.0	West	Half Yearly	0	5.0
18468.0	0	4.0	366405.0	West	Half Yearly	2	5.0

Variable Information:

We observe there are 19 variables in given data set out of which 5 are integer type, 7 float and rest are categorical variables which needs to be encoded into numeric data before proceeding with model building.

```
 #   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
```

Statical Analysis of data provided

	count	mean	std	min	25%	50%	75%	max
CustID	4520.0	7.002260e+06	1304.955938	7000000.0	7001129.75	7002259.5	7003389.25	7004519.0
AgentBonus	4520.0	4.077838e+03	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	1.449471e+01	9.037629	2.0	7.00	13.0	20.00	58.0
CustTenure	4294.0	1.446903e+01	8.963671	2.0	7.00	13.0	20.00	57.0
ExistingProdType	4520.0	3.688938e+00	1.015769	1.0	3.00	4.0	4.00	6.0
NumberOfPolicy	4475.0	3.565363e+00	1.455926	1.0	2.00	4.0	5.00	6.0
MonthlyIncome	4284.0	2.289031e+04	4885.600757	16009.0	19683.50	21606.0	24725.00	38456.0
Complaint	4520.0	2.871681e-01	0.452491	0.0	0.00	0.0	1.00	1.0
ExistingPolicyTenure	4336.0	4.130074e+00	3.346386	1.0	2.00	3.0	6.00	25.0
SumAssured	4366.0	6.199997e+05	246234.822140	168536.0	439443.25	578976.5	758236.00	1838496.0
LastMonthCalls	4520.0	4.626991e+00	3.620132	0.0	2.00	3.0	8.00	18.0
CustCareScore	4468.0	3.067592e+00	1.382968	1.0	2.00	3.0	4.00	5.0

On analysing data, we found below observations:

- Variable ‘Age’ shows minimum policy holder age as 2 and 50% data falls into the range on 13 yrs. age, which shows minor dominate, which shows left skewness in data.
- Variable ‘CustTenure’ also shows data is left skewed which means most of the persons (50%) who bought policy have sticked with present organisation for 13 years

- Most of the rest variables depicted are making bell curve therefore following normal distribution, which is further illustrated through density curve in univariate analysis.

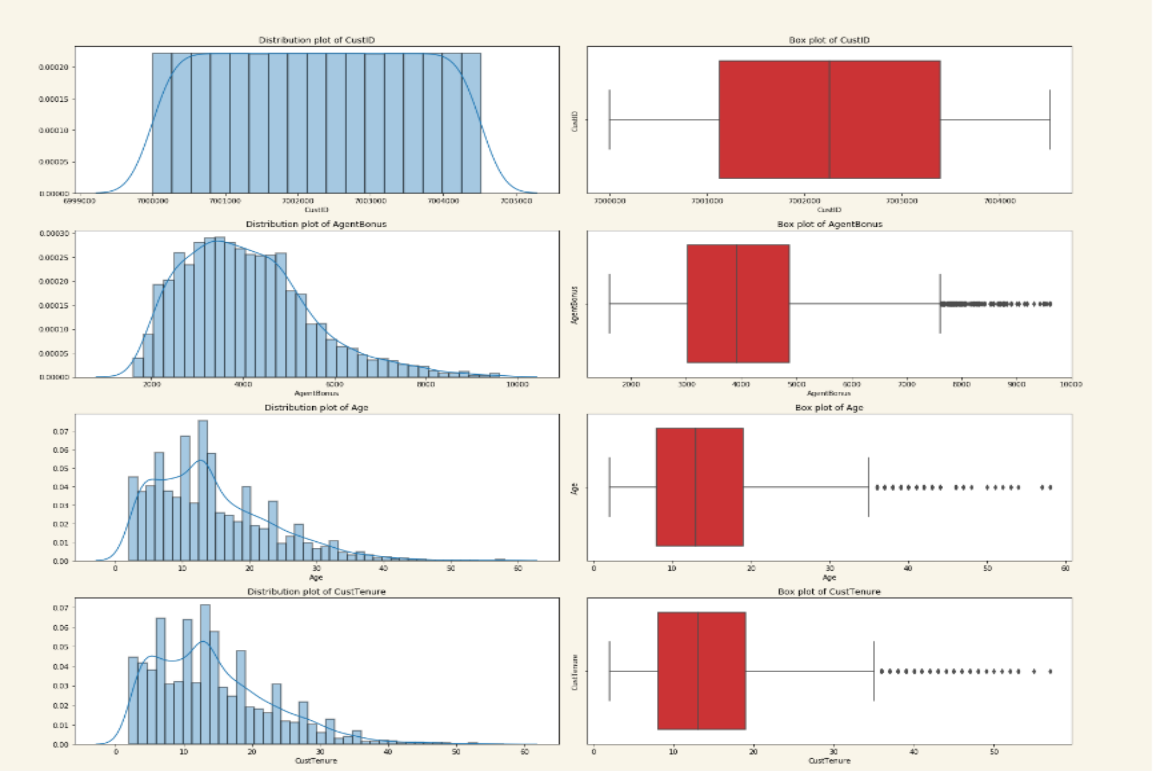
Exploratory Data Analysis

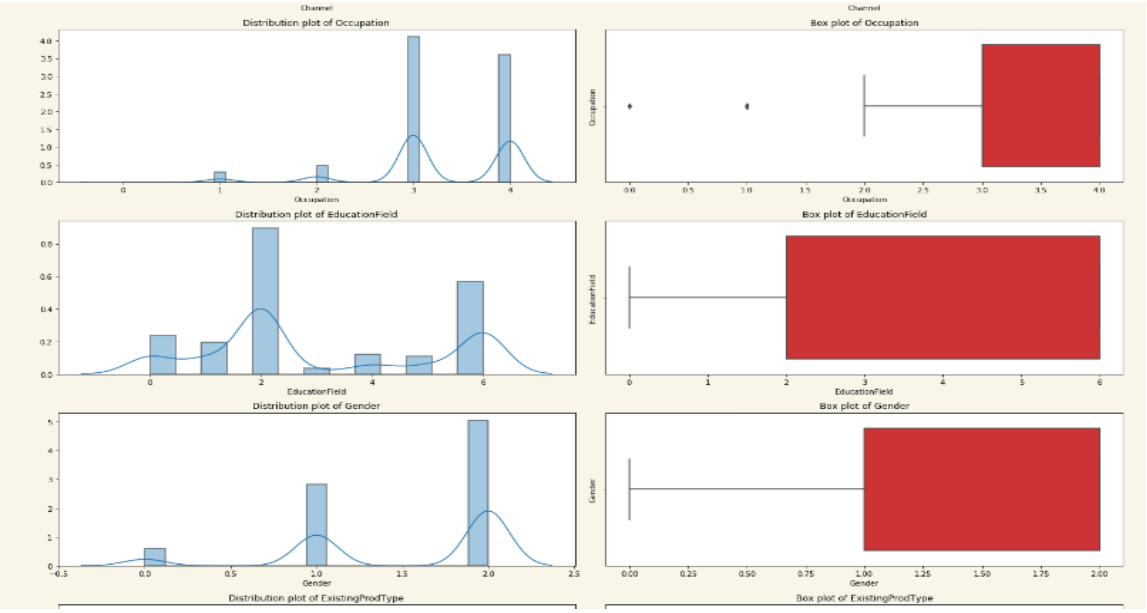
Exploratory data analysis (EDA) is the pre primary component of any model building exercise. This consumes up to 70 percent time of any data science project. This involves steps like visualising data, data cleaning, data filtering, data mining, outlier treatment etc.

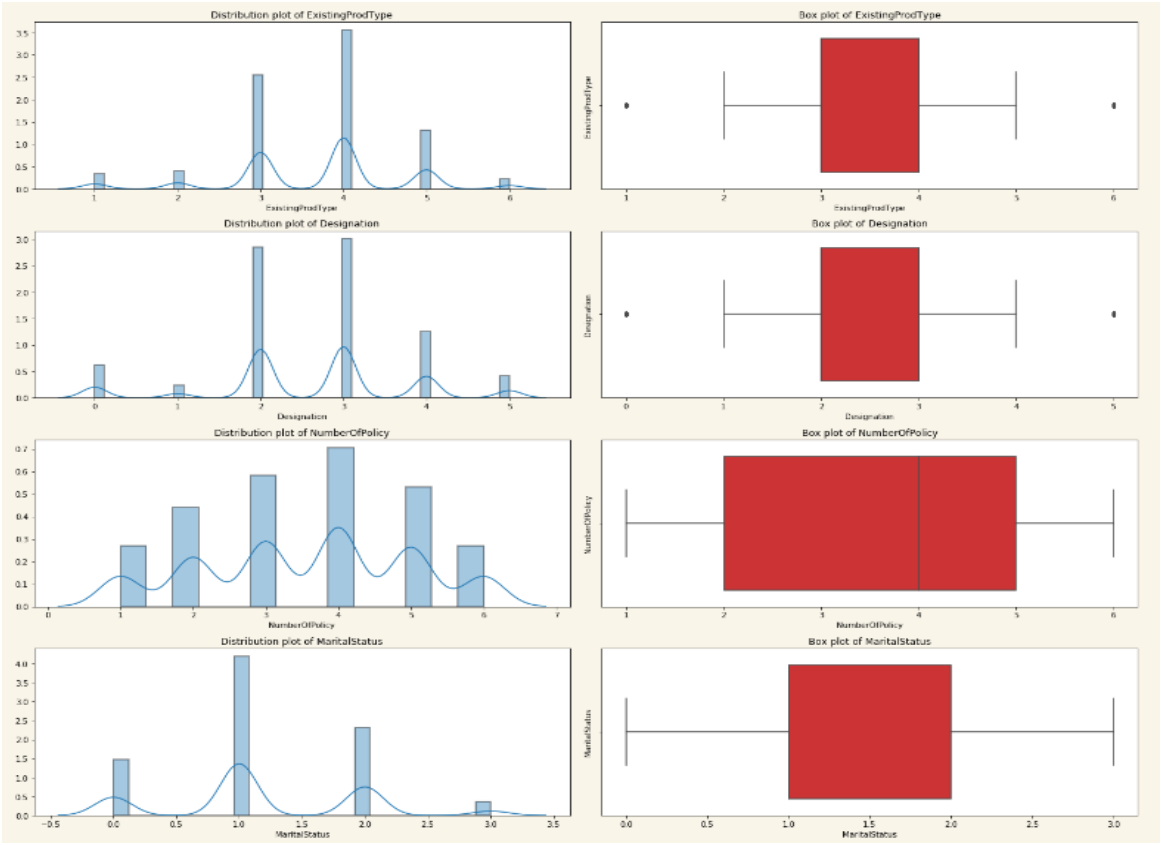
Various steps involved in EDA of the given 'life insurance sales' data set are covered below:

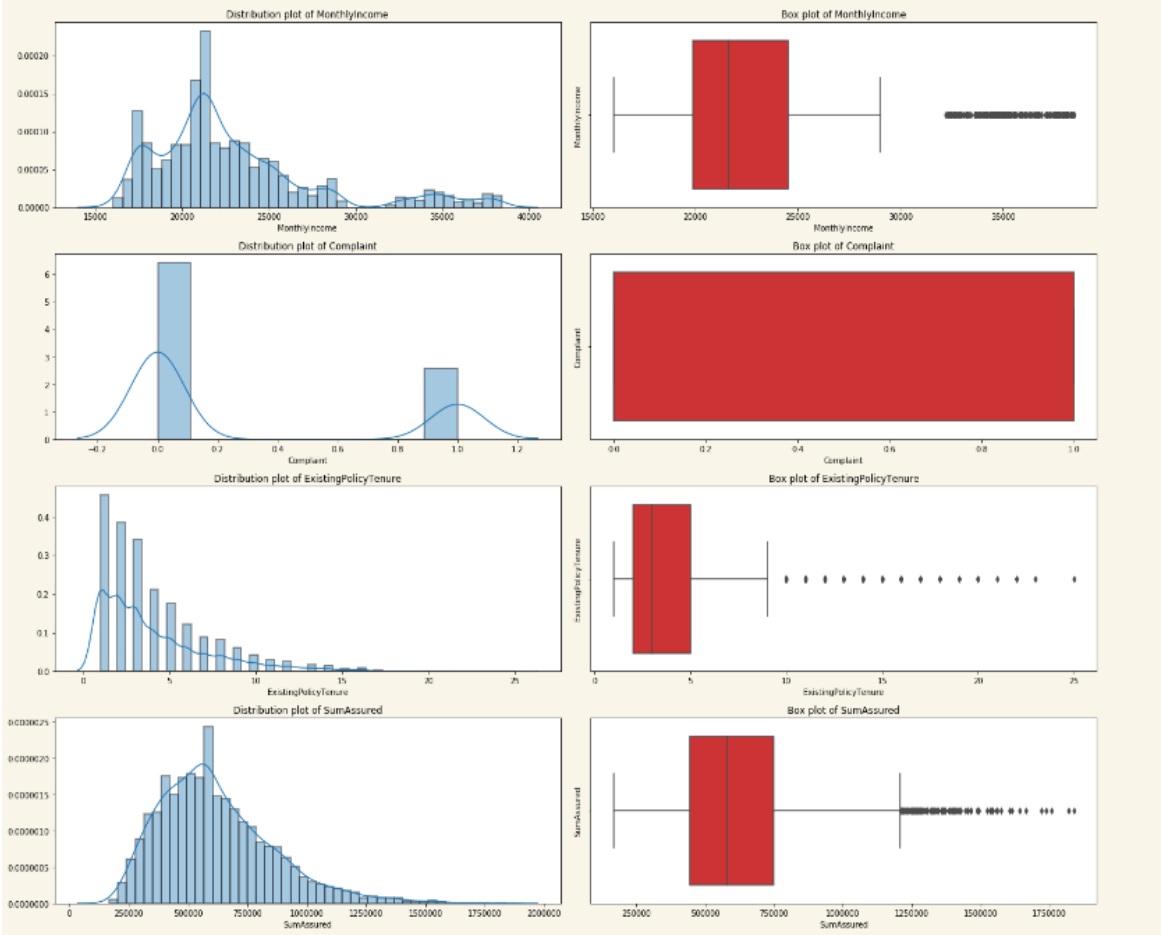
Univariate Analysis:

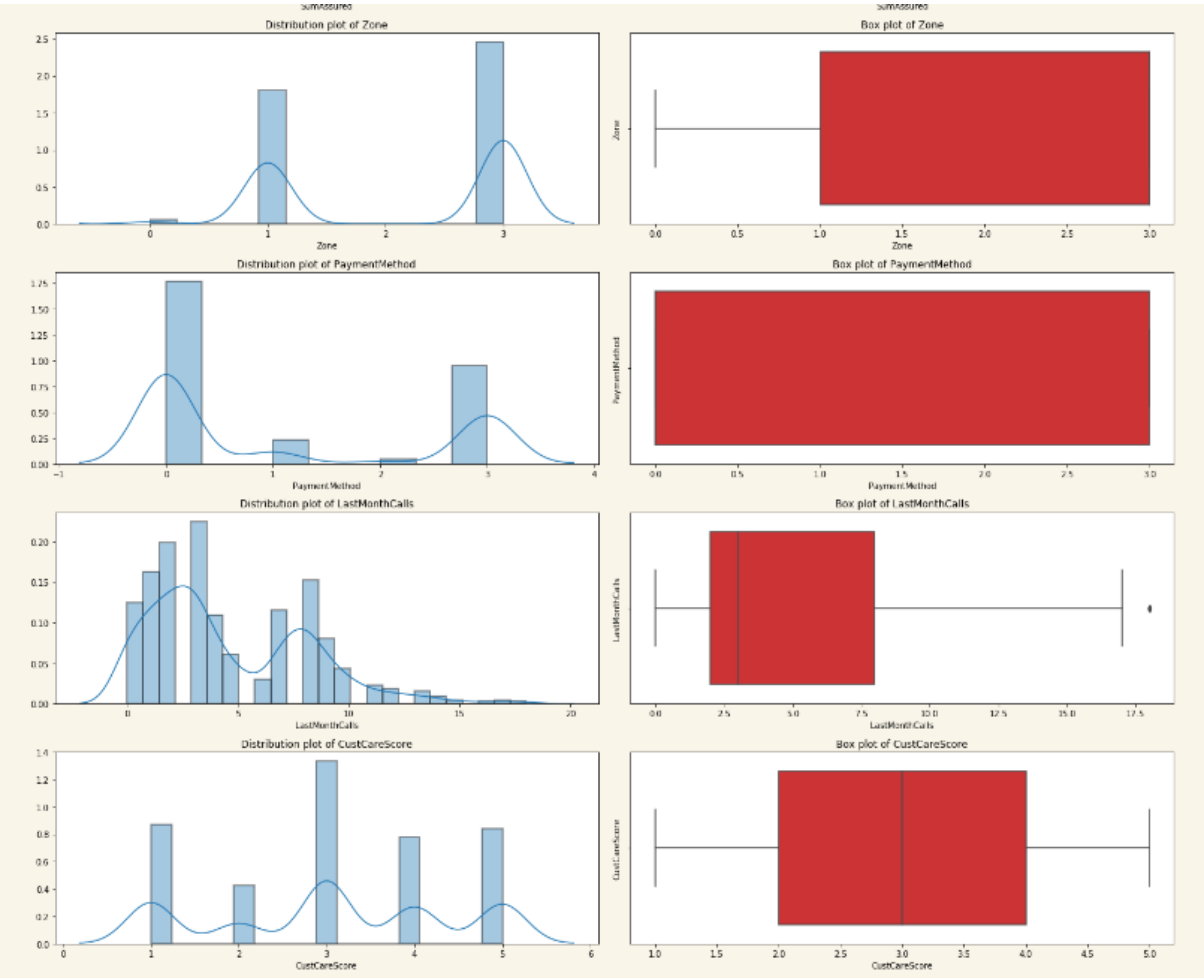
Box plots and Density plots before outlier treatment




Distribution plot of Occupation
Box plot of Occupation
Occupation
Distribution plot of EducationField
Box plot of EducationField
EducationField
Distribution plot of Gender
Box plot of Gender
Gender
Distribution plot of ExistingProdType
Box plot of ExistingProdType


Distribution plot of ExistingProdType
Box plot of ExistingProdType
Distribution plot of Designation
Box plot of Designation
Distribution plot of NumberOfPolicy
Box plot of NumberOfPolicy
Distribution plot of MaritalStatus
Box plot of MaritalStatus

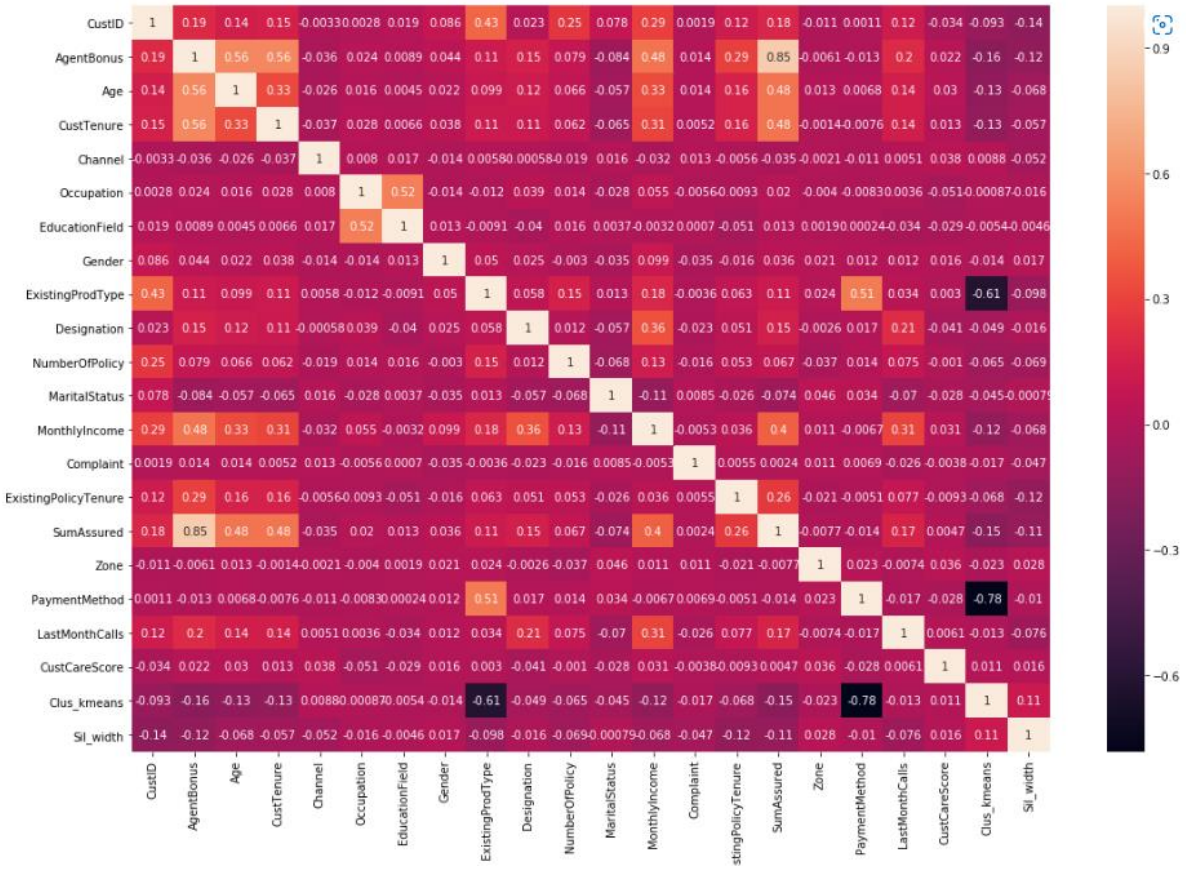

Distribution plot of MonthlyIncome
Box plot of MonthlyIncome
Distribution plot of Complaint
Box plot of Complaint
Distribution plot of ExistingPolicyTenure
Box plot of ExistingPolicyTenure
Distribution plot of SumAssured
Box plot of SumAssured



Observations from above density and box plots

- We see that there are outlier present in many variables (AgentBonus, Age, CustTenure, MonthlyIncome, ExistingPolicyTenure, SumAssured) 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

Heat map for the correlated variables



From the heat map we get below variables are highly co-related to Agent Bonus

- SumAssured
- ExistingPolicyTenure
- MonthlyIncome
- CustTenure
- Age

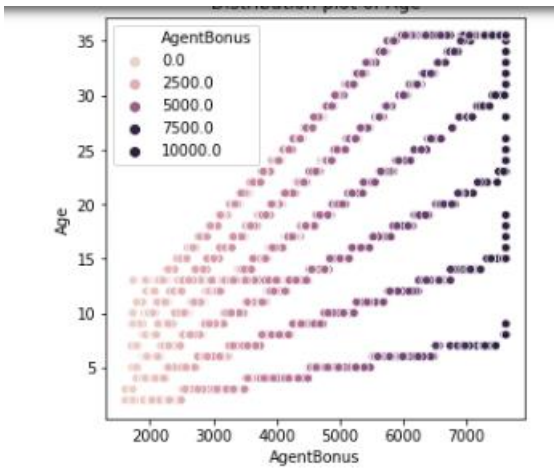
From the above heatmap we could easily observe that the highest correlation exists with Sum Assured, which means Sum Assured is the most dependent factor in determining the AgentBonus

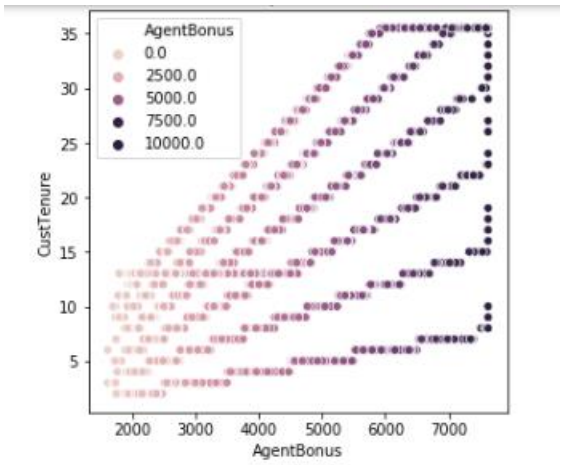
Bivariate Analysis:

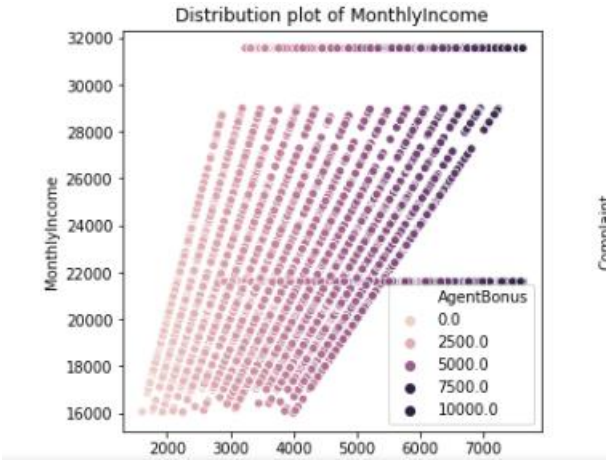
We have introduced the bivariate analysis for the highly correlated variables

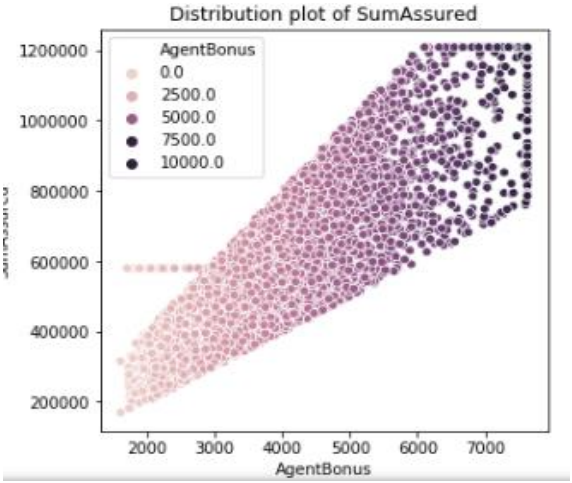
From the below graph we can draw below inferences

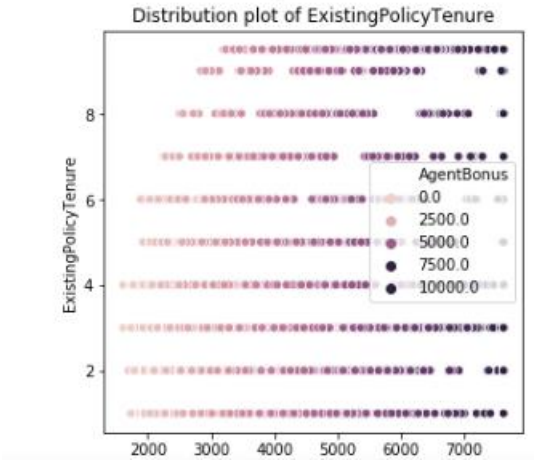
- Agent bonus is highly correlated to Sum Assured, which means if sum assured is high then agent bonus is also high.
- We also observe that agent bonus is low for the customers with lower age group, however customer penetration is higher in lower age group.
- Cust Tenure is also a deterministic factor determining which tells how long the customer sticks with the organisation and since premium is a recursive event which customer needs to pay each year, so the agent is paid with the bonus for the policy as long the customer is with the organisation.
- Customer monthly income also plays important role determining the agent bonus as higher the monthly income bigger is the ticket size (sum assured) and more is the agent bonus.
- Existing Policy Tenure is also slightly high correlated since, it determines how long the agent will be paid with the recursive bonus for the existing policy.






Distribution plot of MonthlyIncome
MonthlyIncome
AgentBonus
0.0
2500.0
5000.0
7500.0
10000.0


Distribution plot of SumAssured
AgentBonus
0.0
2500.0
5000.0
7500.0
10000.0
AgentBonus


Distribution plot of ExistingPolicyTenure
ExistingPolicyTenure
AgentBonus
0.0
2500.0
5000.0
7500.0
10000.0

Missing Value Treatment:

We observe that there are many variables with missing values

Before missing value treatment

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

After missing value treatment

CustID	0
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

Resultant dataset has no missing values.

We have used below imputing strategies while dealing with different variables

Mean

- SumAssured
- MonthlyIncome

Median

- Age
- CustTenure
- ExistingPolicyTenure
- CustCareScore
- NumberOfPolicy

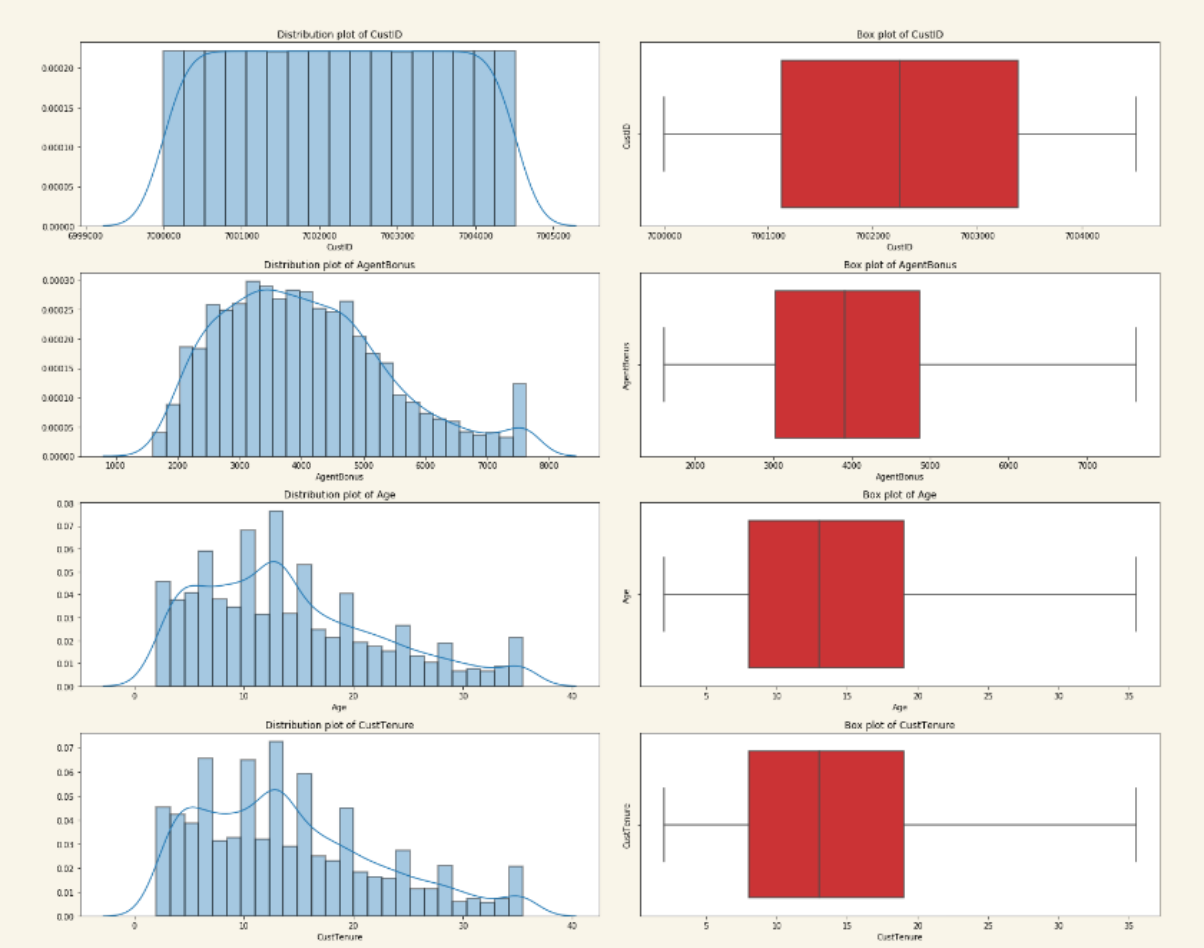
Outlier Treatment:

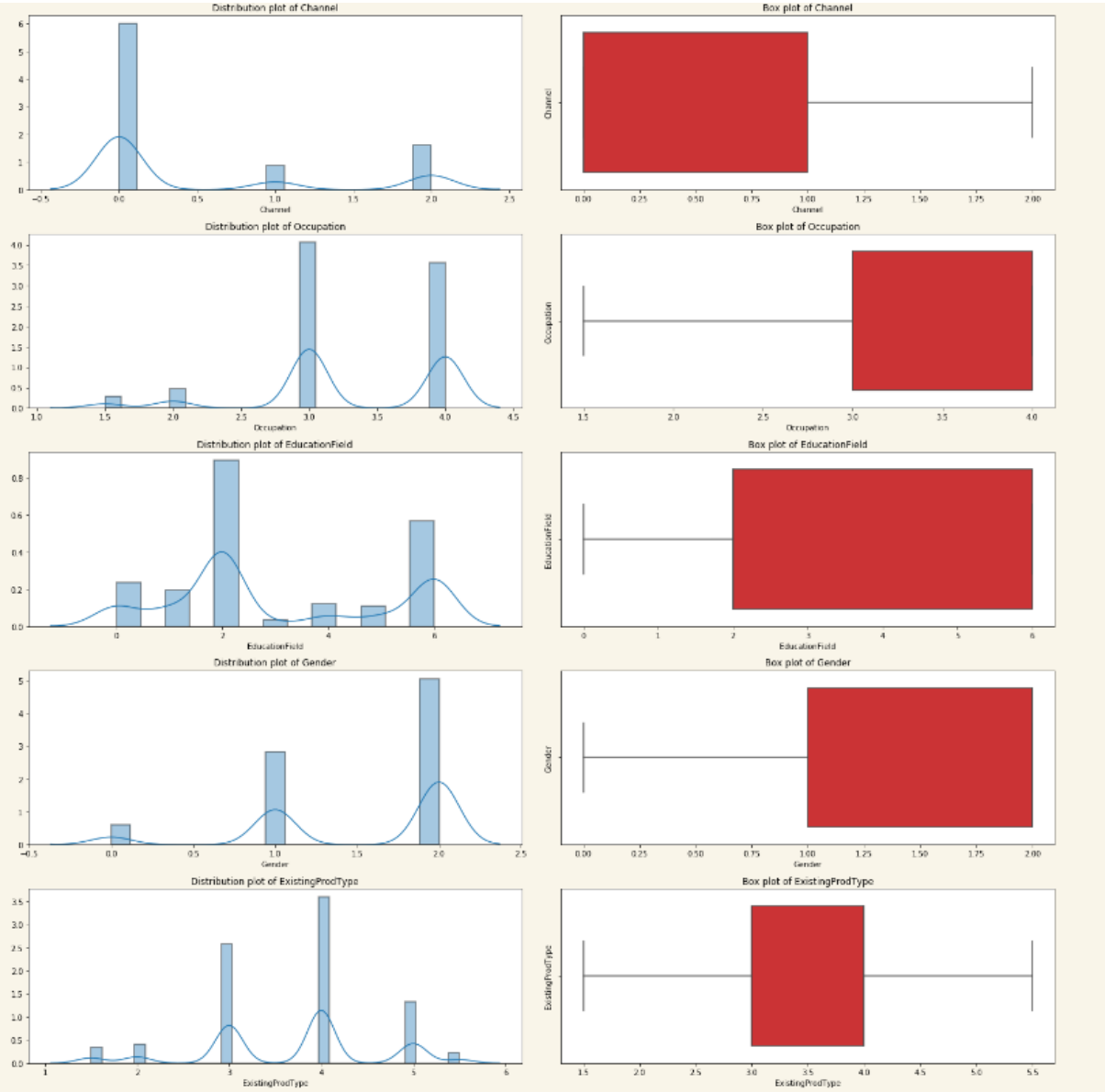
Outliers are those values which lie outside 1.5 * *IQR*.

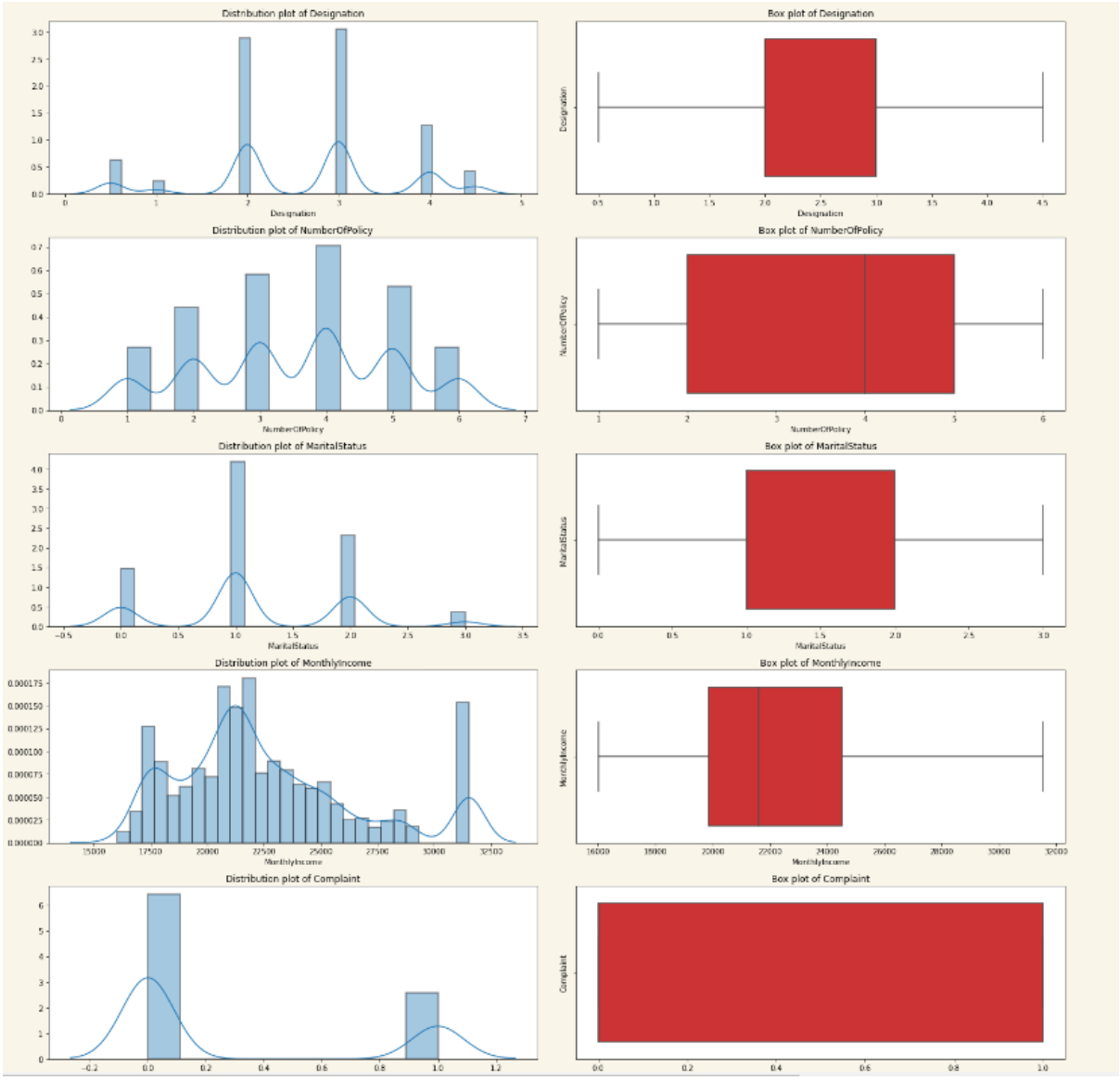
From the box plot in univariate analysis, we observe there are many outliers present in the given data, which means outlier treatment become mandatory for the given data.

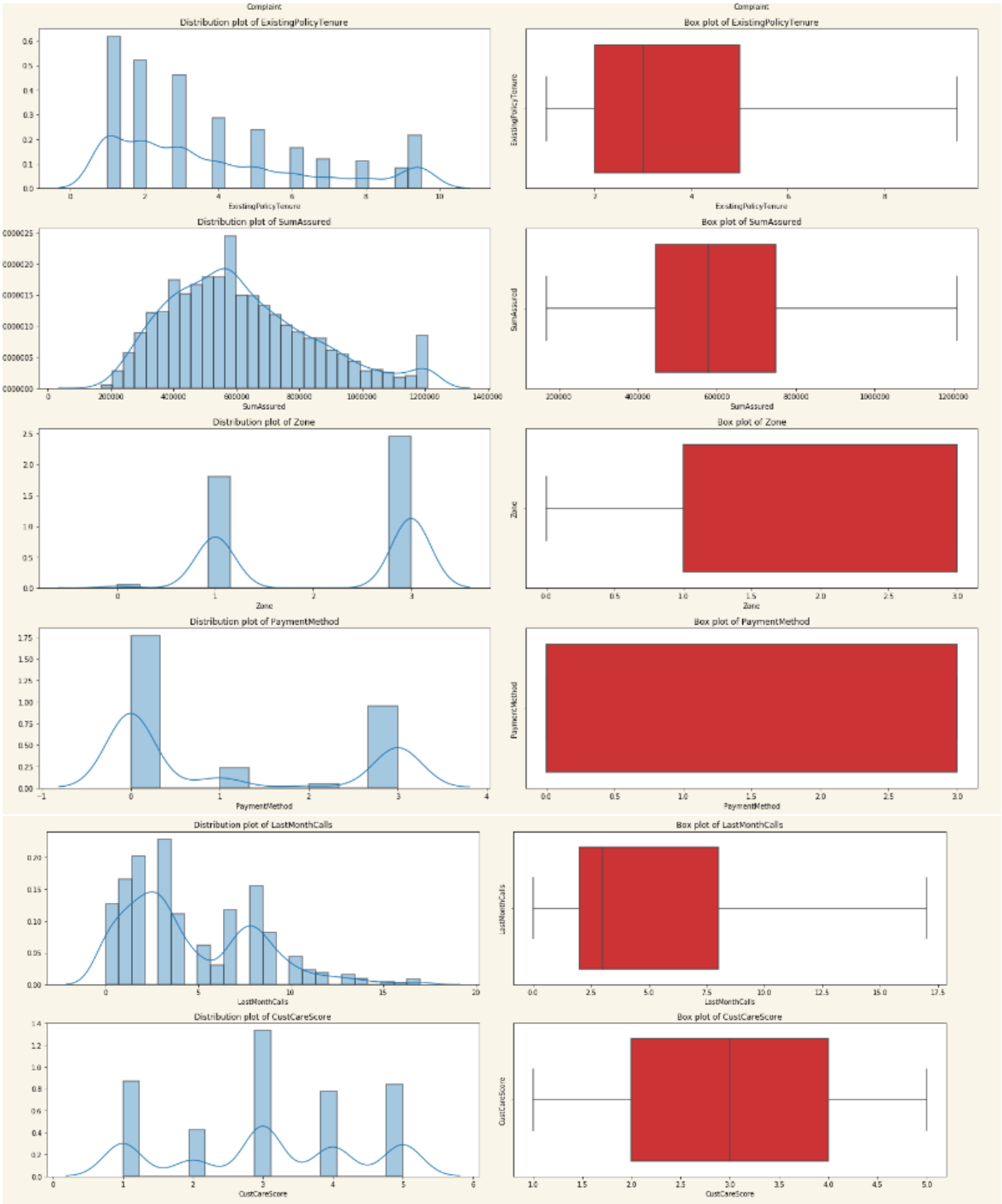
In outlier treatment we could cap the values by replacing those observations outside the lower limit with the value of 5th percentile and those that lie above the upper limit, with the value of 95th percentile.

Box Plots after outlier treatment




Distribution plot of Channel
Box plot of Channel
Channel
Distribution plot of Occupation
Box plot of Occupation
Occupation
Distribution plot of EducationField
Box plot of EducationField
EducationField
Distribution plot of Gender
Box plot of Gender
Gender
Distribution plot of ExistingProdType
Box plot of ExistingProdType
ExistingProdType


Distribution plot of Designation
Box plot of Designation
Designation
Distribution plot of NumberOfPolicy
Box plot of NumberOfPolicy
NumberOfPolicy
Distribution plot of MaritalStatus
Box plot of MaritalStatus
MaritalStatus
Distribution plot of MonthlyIncome
Box plot of MonthlyIncome
MonthlyIncome
Distribution plot of Complaint
Box plot of Complaint
Complaint



From the above graph we observe that all the outliers are removed from the data.

Variable Transformation

We observe that there are many categorical variables present in existing dataset which need to be transformed, there are various methods of encoding data set I have selected Label encoding for the purpose

Variables which are encoded:

- Channel
- Occupation
- EducationField
- Gender
- Designation
- MaritalStatus
- Zone
- PaymentMethod

Above variables are encoded into numeric form depending upon there levels.

Addition Of Variables:

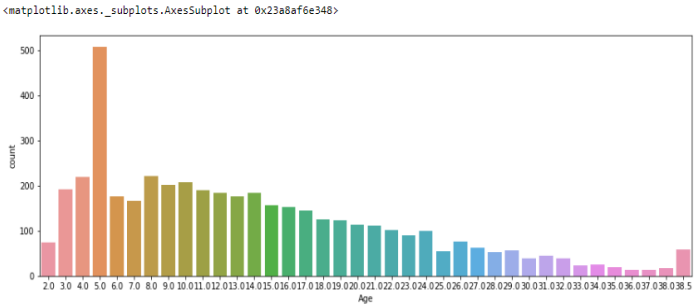
Since the given data set, is the primary source data set and we could not get more data, if possible, we could have asked for ‘premium’ collected from the customer as it is also a vital component in determining the agent bonus.

Basic Insights From Exploratory Data Analysis

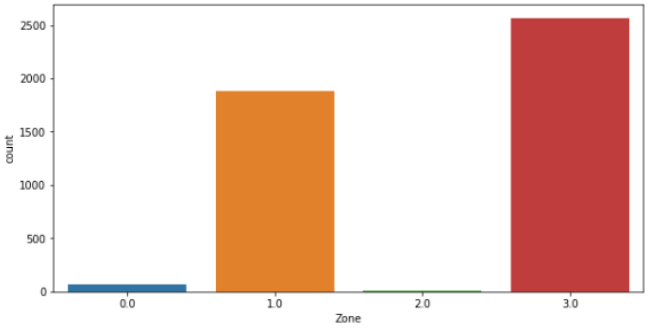
Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Dataset shared with us is highly unbalanced dataset. There are few observations while looking at data:

- Age, which is one of the foremost deterministic factors while determining agent bonus is highly unbalanced, we see that most of the data set corresponds to minor below 18.

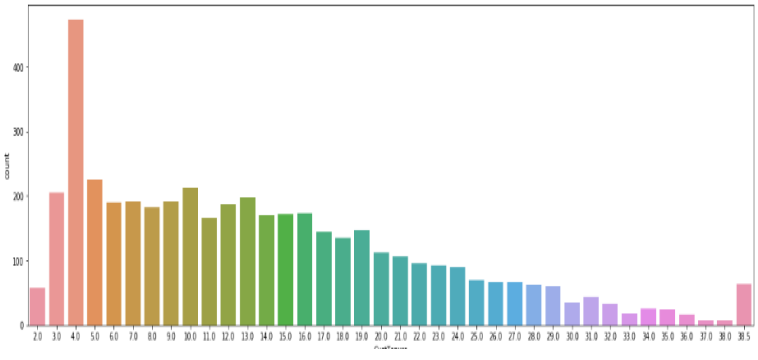


- Zone, we see major penetration of data is in West followed by North but almost no contribution from east and south which also cause data unbalancing.

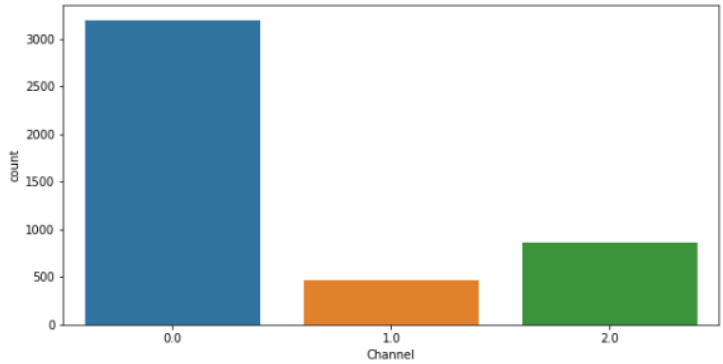


0-East 1-North 2-South 3-West

- Cust Tenure, is also major contributor to unbalancing data set as the data is skewed to left also we see customer is sticking to the organization is mainly below 20 years.

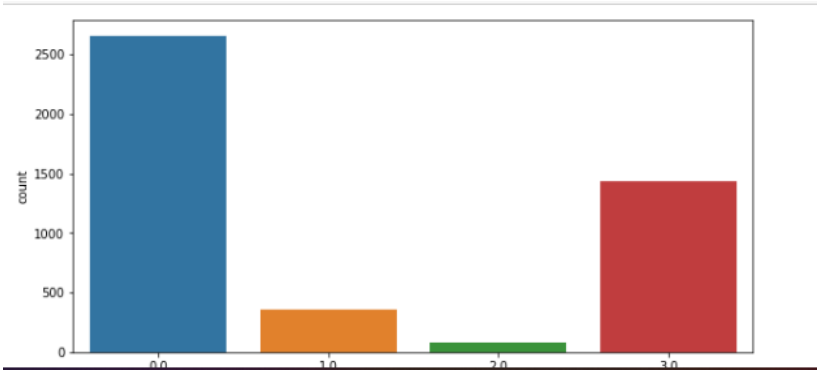


- Channel, also contributes to unbalancing. As the major channel of sourcing is ‘Customer Agent’ and rest are minimal.



0-Agent 1-Online 2-Third Party Partner

- We see ‘payment method’ also contribute to unbalancing as major chunk is for half yearly and yearly mode rest are minimal.



0-Half Yearly1-Monthly 2-Quarterly 3-Yearly

In order to treat the unbalancing issue with the data we can use below methods to treat the data to certain extent.

- Under sampling and Over sampling

Under sampling consists of removing samples from majority class whereas Oversampling is adding more samples from minority class.

As we saw, data set is highly unbalanced, predicting using this data could lead to potential danger to business decisions, we need to firstly balance data using above mentioned techniques only when we could make better recommendations,

Any business insights using clustering (if applicable)

Clustering is done generally over unsupervised data; we divide data in different clusters based on pattern formation.

Commonly used clustering techniques include

- Hierarchal Clustering
- K means Clustering

In the given data set we have used K means clustering technique which has divided dataset in 3 clusters.

Row Labels	Count of Clus_kmeans
0	1089
1	1362
2	2069
(blank)	
Grand Total	**4520**

We find cluster 2 is larger cluster compared to 0 and 1.

Any other business insights

- We find that most of data set comprises of age group less than 20 which means less premium is expected out and so as the low mortality rate. This shows company is operating with lower profit margins
- CustTenure also shows that 22 years is the maximum engagement with the organisation, this shows products designed are serving mostly at max 20-25 years segment, company should tailor some new products which could engage customers throughout life time.
- Company should work on more web and mobile based applications as penetration towards as online channel is the least along with third party based.
- North and West are major contributors whereas the south and east shows minimal engagements, this shows company presence in these regions are limited therefore company should think of expanding their horizons to the respective zones.