Life Insurance Sales

Project Notes – 1

Ujjwal Kumar

Post-Graduation (DSBA)

January 21, 2023

Table of Contents:

Life Insurance Sales

Introduction:

So, Insurance is one of the vital products for both business and human life. It not only provides necessary financial support in case of uncertainties but also 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. Basically, Insurance is a financial safety net, helping us and our loved ones recover after something bad happens such as fire, theft, lawsuit or car accident etc. which is a legal contract between us and our insurance provider.

Defining the Business problem:

As an analyst we have been assigned for the role to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents. We are provided with the life insurance 'sales' data set of the leading insurance company.

We need to dive deeply 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.

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 '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.

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.

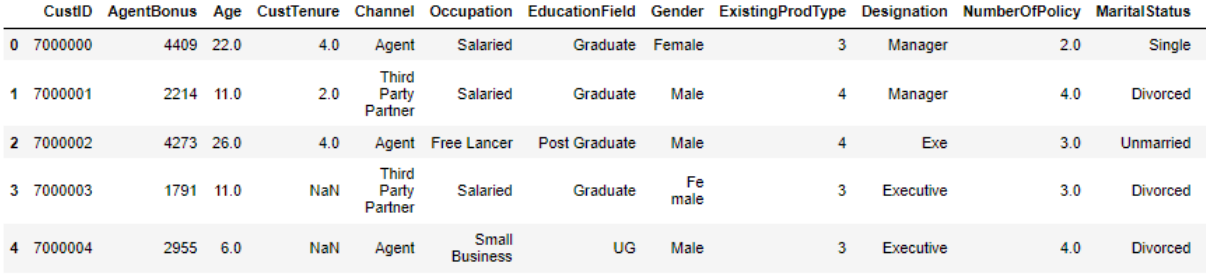
- **Improving Customer satisfaction:** By analysing the perspective customer data, the companies can predict the needs of the customers and thus increase the potential to make a sale when compared to a company following the conventional methods of selling. The existing customer data can be used to find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

Data reports

- Since the data we use, have 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 "Last Month Calls" 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)

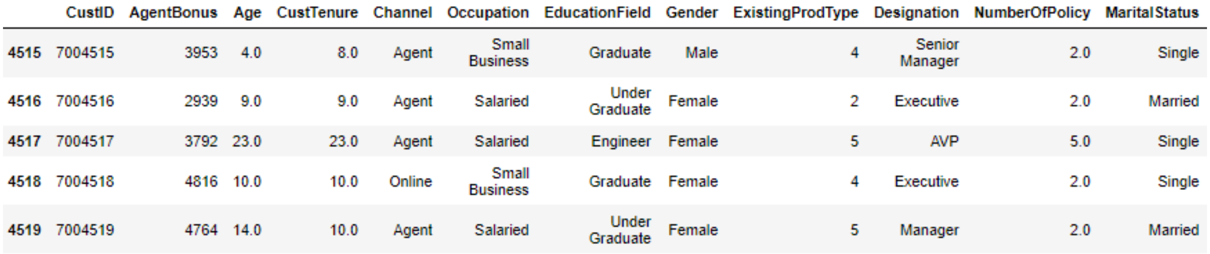
Data Visualization

Below are the first five rows of the given dataset



	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	Fe male	3	Executive	3.0	Divorced
4	7000004	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced

Last five rows of the dataset

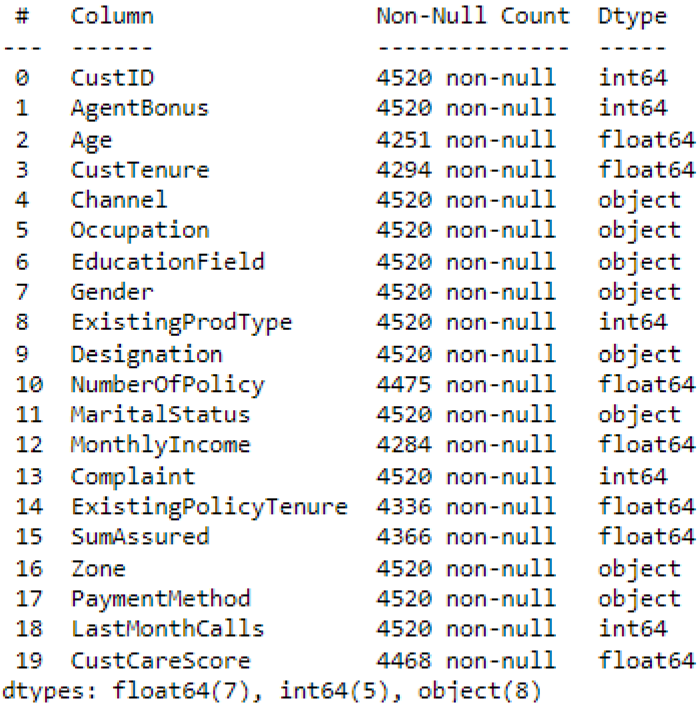


	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus
4515	7004515	3953	4.0	8.0	Agent	Small Business	Graduate	Male	4	Senior Manager	2.0	Single
4516	7004516	2939	9.0	9.0	Agent	Salaried	Under Graduate	Female	2	Executive	2.0	Married
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

Data variable information

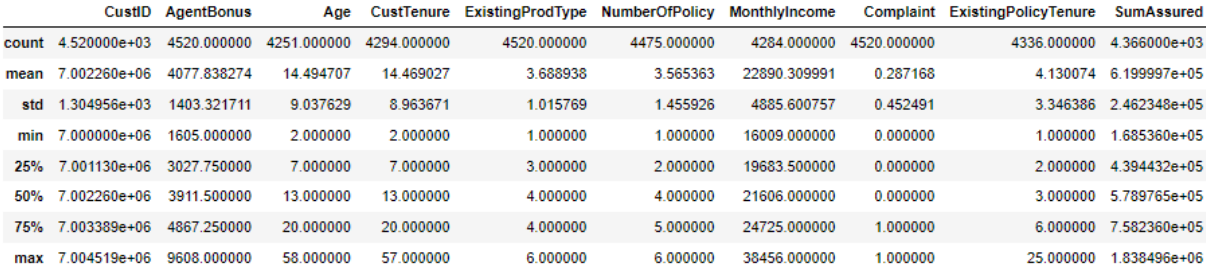
We’ve observed from the given dataset that there are total 19 variables out of which there are 5 integer type, 7 float and rest are categorical variables which needs to be encoded into numeric data before proceeding with model building.

Data Information



```
 #   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)
```

Below is the Summarised data provided



	CustID	AgentBonus	Age	CustTenure	ExistingProdType	NumberOfPolicy	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured
count	4.520000e+03	4520.000000	4251.000000	4294.000000	4520.000000	4475.000000	4284.000000	4520.000000	4336.000000	4.366000e+03
mean	7.002260e+06	4077.838274	14.494707	14.469027	3.688938	3.565363	22890.309991	0.287168	4.130074	6.199997e+05
std	1.304956e+03	1403.321711	9.037629	8.963671	1.015769	1.455926	4885.600757	0.452491	3.346386	2.462348e+05
min	7.000000e+06	1605.000000	2.000000	2.000000	1.000000	1.000000	16009.000000	0.000000	1.000000	1.685360e+05
25%	7.001130e+06	3027.750000	7.000000	7.000000	3.000000	2.000000	19683.500000	0.000000	2.000000	4.394432e+05
50%	7.002260e+06	3911.500000	13.000000	13.000000	4.000000	4.000000	21606.000000	0.000000	3.000000	5.789765e+05
75%	7.003389e+06	4867.250000	20.000000	20.000000	4.000000	5.000000	24725.000000	1.000000	6.000000	7.582360e+05
max	7.004519e+06	9608.000000	58.000000	57.000000	6.000000	6.000000	38456.000000	1.000000	25.000000	1.838496e+06

On summarising the data, we found the below observations:

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

- Rest variables depicted are making a bell curve therefore following normal distribution, which is further illustrated through density curve in univariate analysis.

EDA (Exploratory Data Analysis)

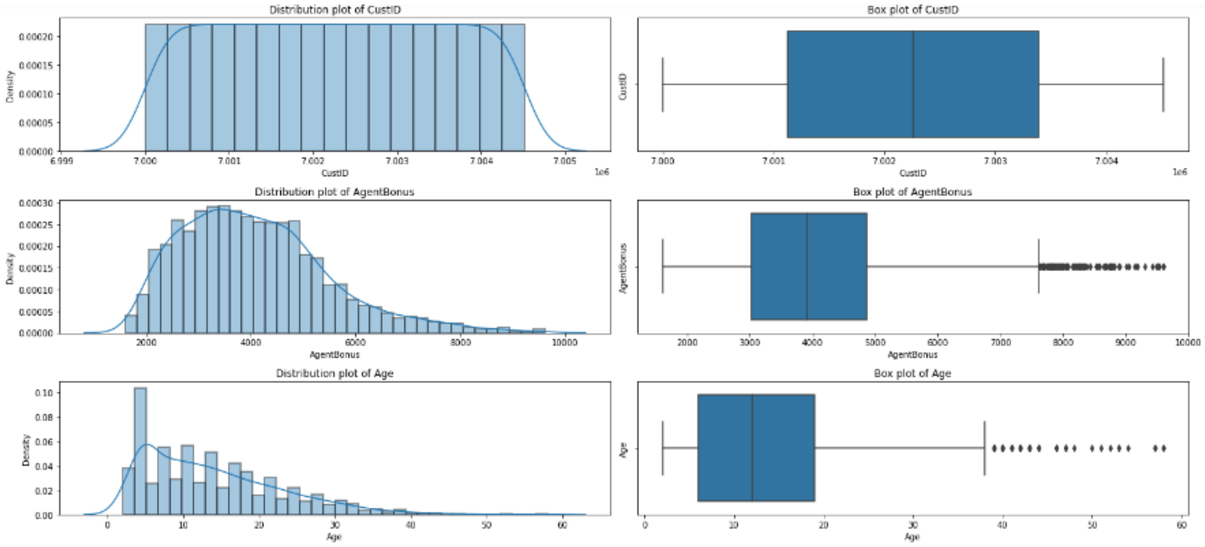
Firstly, we have to understand what EDA is? So, basically EDA is used by data scientists to analyse and investigate the data sets and summarize their main characteristics which means employing data visualization methods. It helps determine how best to manipulate data sources to get the answers we need.

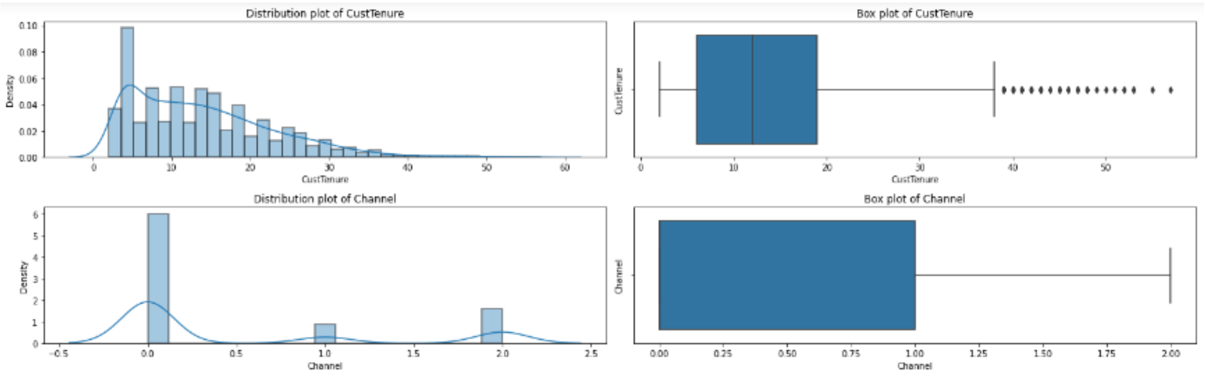
EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. So, various steps involved in EDA of the given 'life insurance sales' data set are covered below:

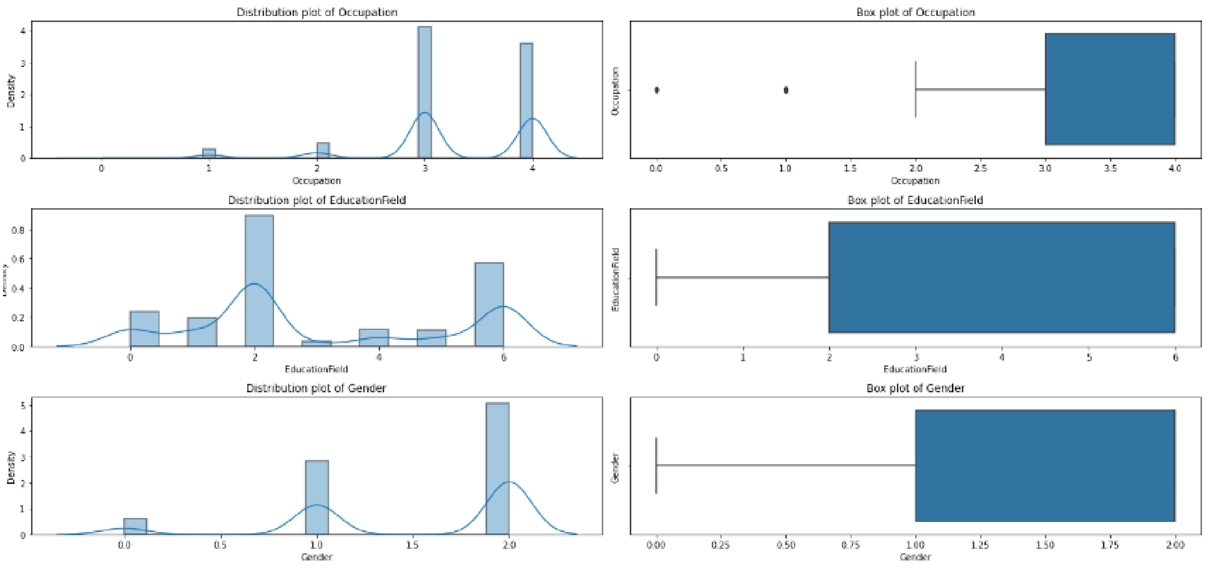
Univariate Analysis

Basically Univariate analysis are conducted for the purpose of making data easier to interpret and to understand how data is distributed within a sample of population being studied. So is done below:-

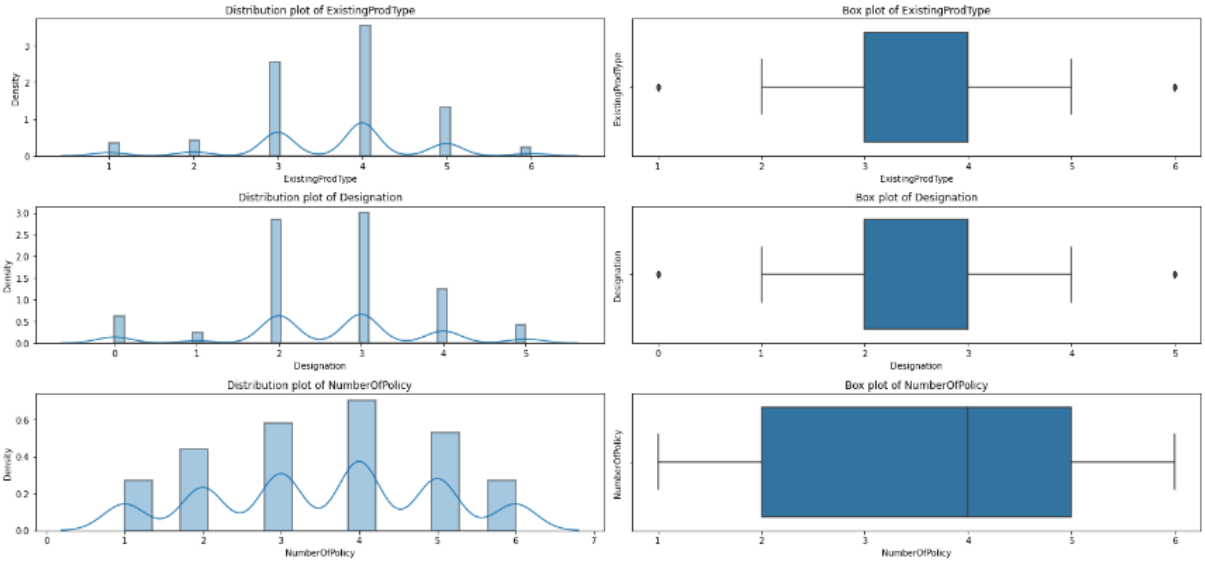
Graphs to show Outliers in the plots



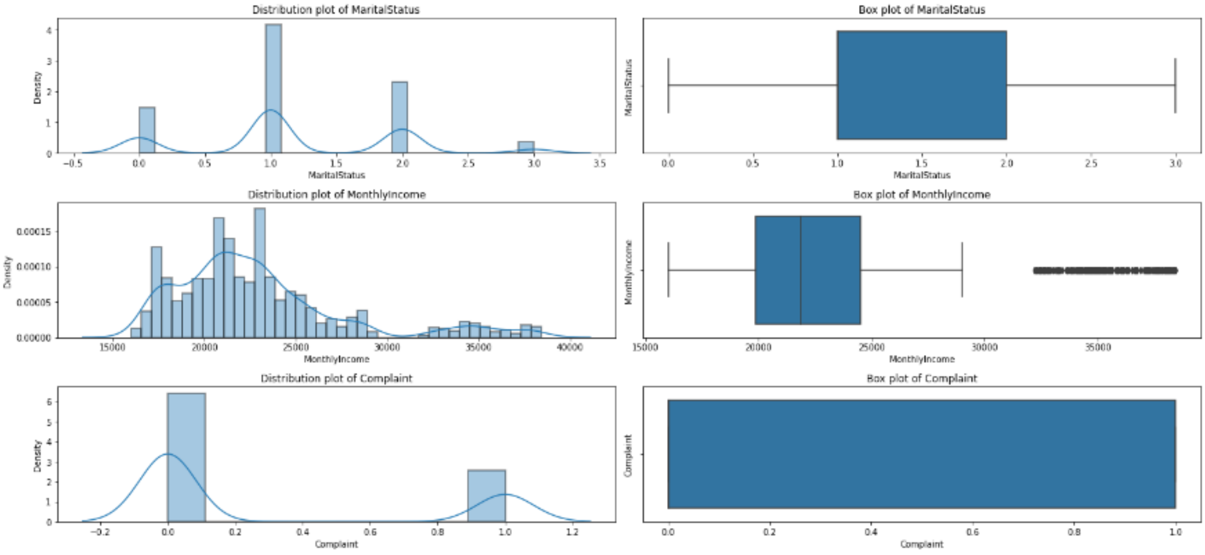




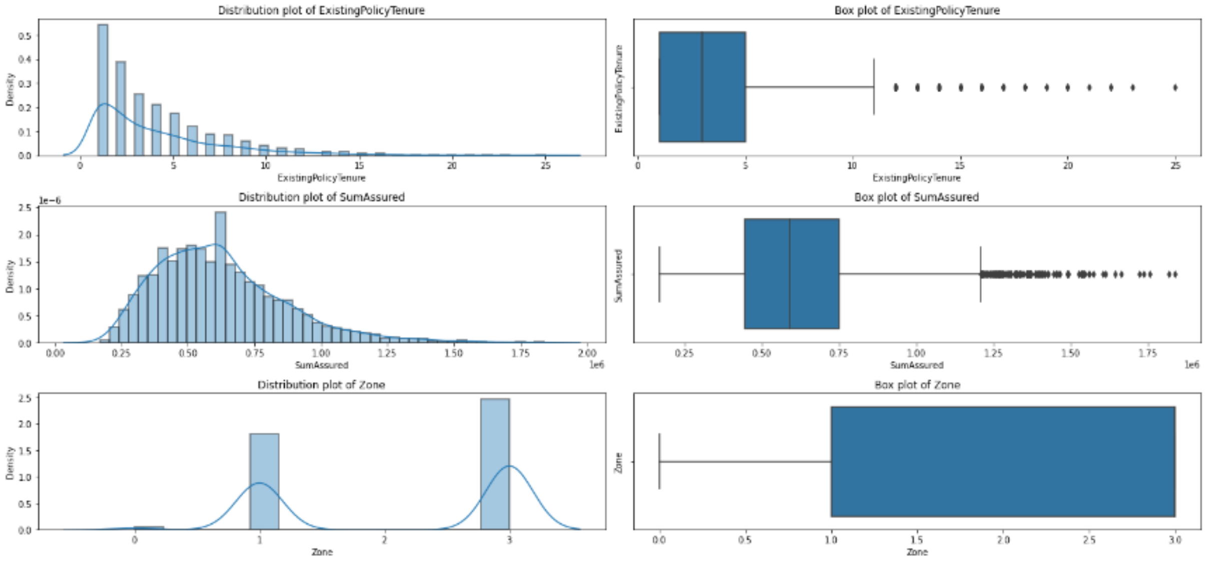
Distribution plot of Occupation
Density
0
1
2
3
4
Occupation
Box plot of Occupation
Occupation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Distribution plot of EducationField
0.0
0.2
0.4
0.6
0.8
EducationField
Box plot of EducationField
5
6
Distribution plot of Gender
Gender
Box plot of Gender
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00

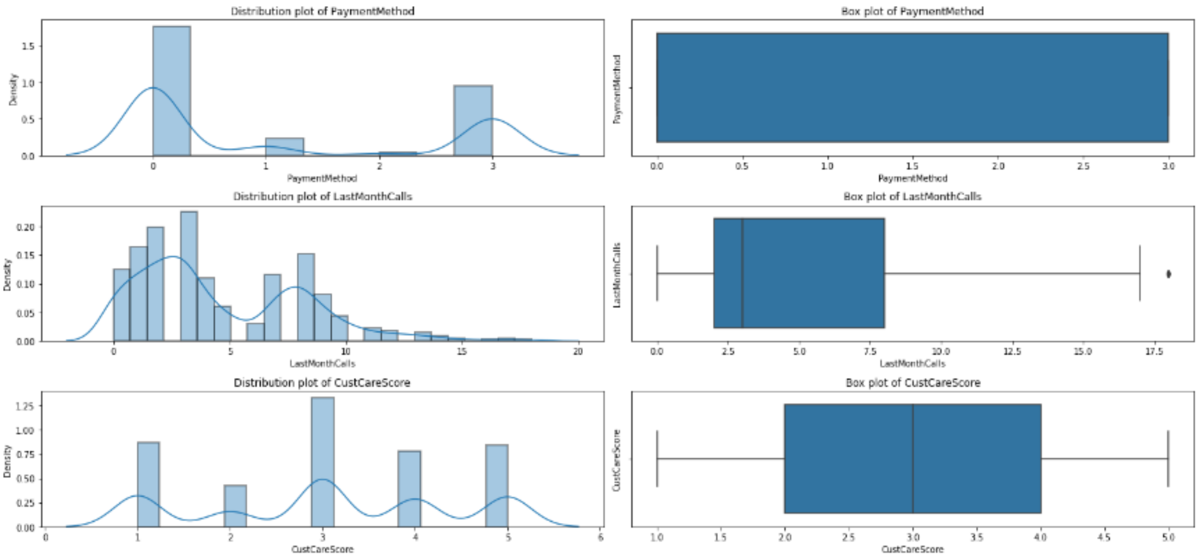


Distribution plot of ExistingProdType
Density
ExistingProdType
Box plot of ExistingProdType
Distribution plot of Designation
Designation
Box plot of Designation
Distribution plot of NumberOfPolicy
NumberOfPolicy
Box plot of NumberOfPolicy
0.0
0.2
0.4
0.6
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
6
7



Distribution plot of MaritalStatus
Density
MaritalStatus
Box plot of MaritalStatus
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
Distribution plot of MonthlyIncome
0.00000
0.00005
0.00010
0.00015
15000
20000
25000
30000
35000
40000
MonthlyIncome
Box plot of MonthlyIncome
Distribution plot of Complaint
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Complaint
Box plot of Complaint





Observations from the above plots:-

- We observe that the policy holders are minor which basically means that the lower premium collects lower revenue but at the same time claim rate is also lower.
- We also see that there are many outliers in many variables like (Age, CustTenure, ExistingPolicyTenure, SumAssured, AgentBonus, MonthlyIincome) which needs to be treated.
- Through density plot we observe that there is left skewness which is observed in a few variables like CustTenure, Age, and rest lies in normal distribution.

Heat Map

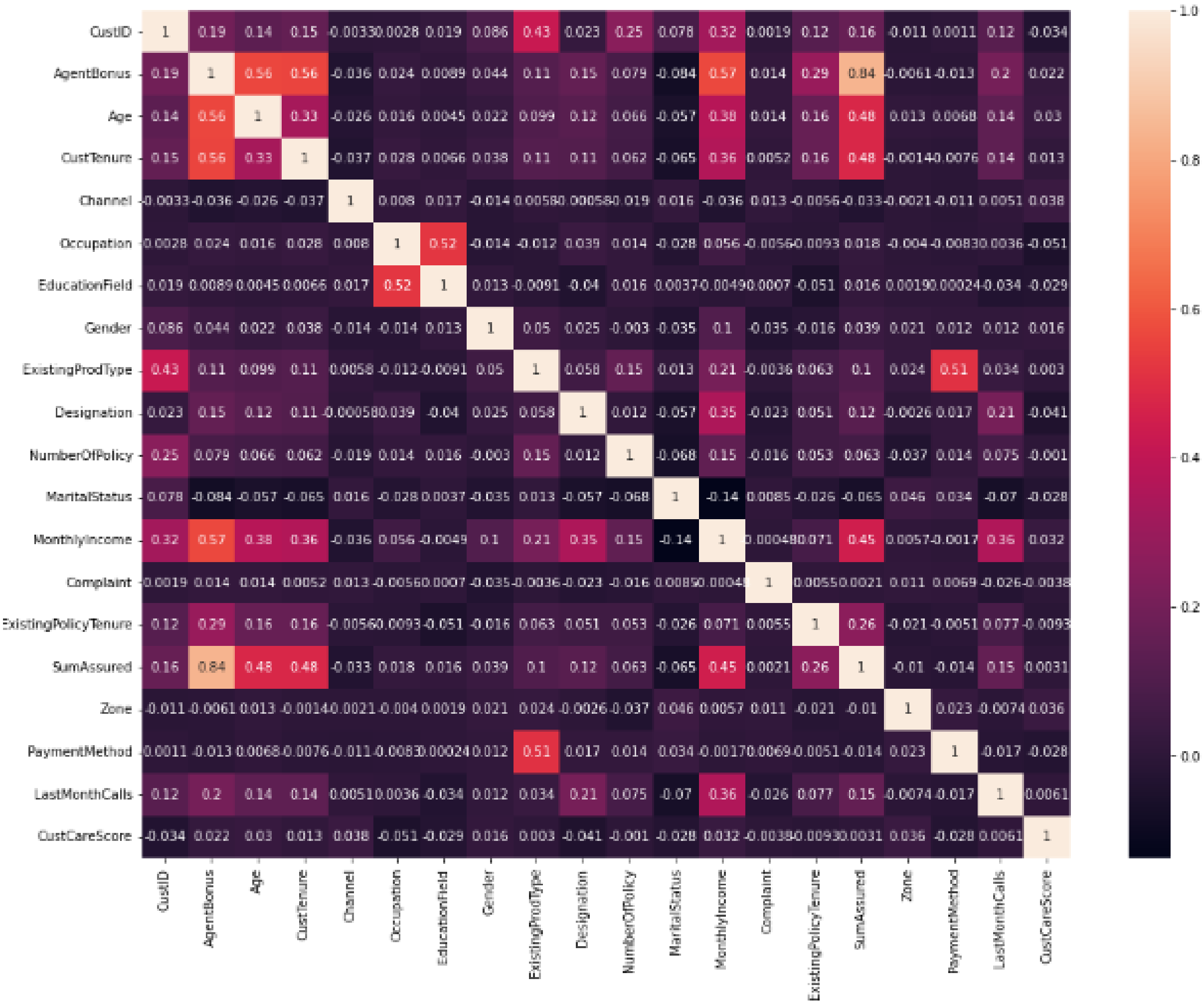
As we know that a heatmap contains all the values representing various shades of the same colour for each value to be plotted. In this heat map, we see that the darker shades of the chart represent higher values than the lighter shades. It provides us with an easy tool to understand the correlation between two entities.

Correlations

Correlation is a statistical measure that expresses the extent to which two variables are linear related which means that they change together at a constant rate.

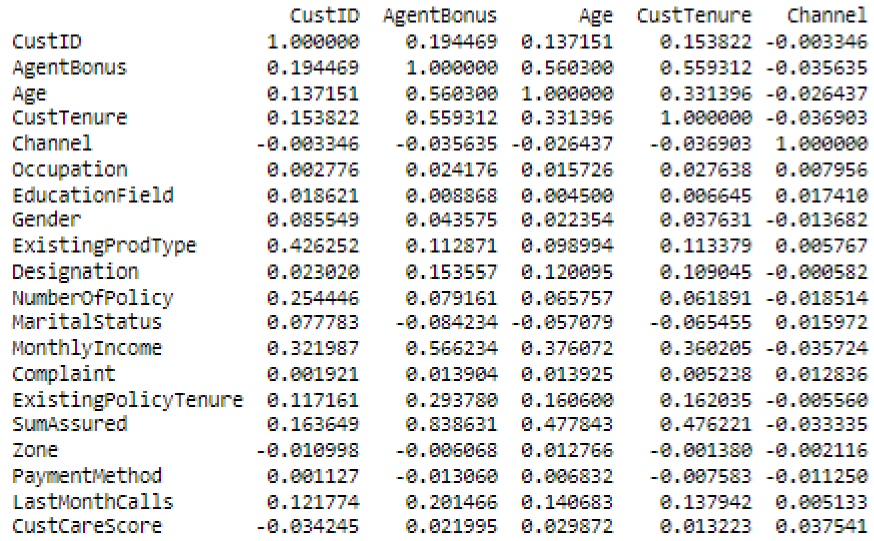
Whereas Correlation heatmap is a type of plot that visualize the strength of a relationships between numerical variables.

Heat Map

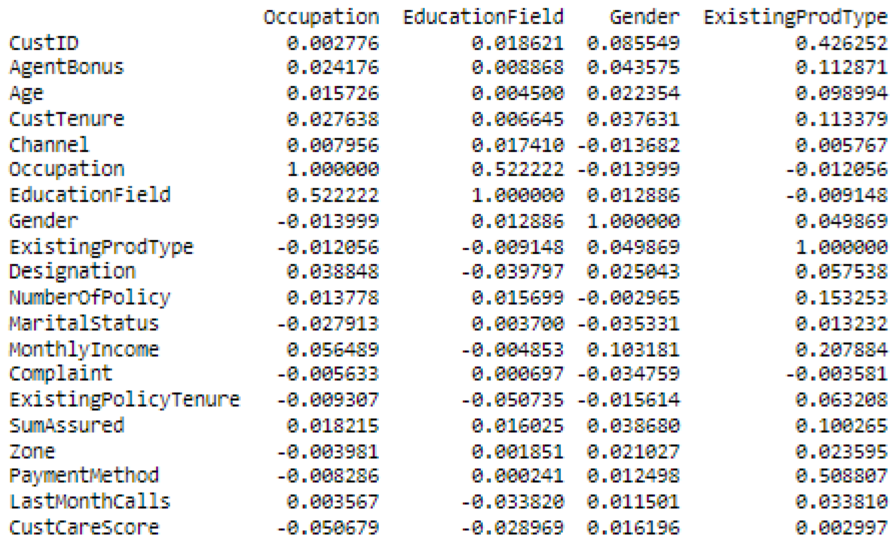


	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	Zone	PaymentMethod	LastMonthCalls	CustCareScore
CustID	1	0.19	0.14	0.15	-0.0033	0.0028	0.019	0.086	0.43	0.023	0.25	0.078	0.32	0.0019	0.12	0.16	-0.011	0.0011	0.12	-0.034
AgentBonus	0.19	1	0.56	0.56	-0.036	0.024	0.0089	0.044	0.11	0.15	0.079	-0.084	0.57	0.014	0.29	0.84	-0.0061	-0.013	0.2	0.022
Age	0.14	0.56	1	0.33	-0.026	0.016	0.0045	0.022	0.099	0.12	0.066	-0.057	0.38	0.014	0.16	0.48	0.013	0.0068	0.14	0.03
CustTenure	0.15	0.56	0.33	1	-0.037	0.028	0.0066	0.038	0.11	0.11	0.062	-0.065	0.36	0.0052	0.16	0.48	-0.0014	-0.0076	0.14	0.013
Channel	-0.0033	-0.036	-0.026	-0.037	1	0.008	0.017	-0.014	0.0058	-0.00058	-0.019	0.016	-0.036	0.013	-0.0056	-0.033	-0.0021	-0.011	0.0051	0.038
Occupation	0.0028	0.024	0.016	0.028	0.008	1	0.52	-0.014	-0.012	0.039	0.014	-0.028	0.056	-0.0056	0.0093	0.018	-0.004	-0.0083	0.0036	-0.051
EducationField	0.019	0.0089	0.0045	0.0066	0.017	0.52	1	0.013	-0.0091	-0.04	0.016	0.0037	-0.0049	0.0007	-0.051	0.016	0.0019	0.00024	-0.034	-0.029
Gender	0.086	0.044	0.022	0.038	-0.014	-0.014	0.013	1	0.05	0.025	-0.003	-0.035	0.1	-0.035	-0.016	0.039	0.021	0.012	0.012	0.016
ExistingProdType	0.43	0.11	0.099	0.11	0.0058	-0.012	-0.0091	0.05	1	0.058	0.15	0.013	0.21	-0.0036	0.063	0.1	0.024	0.51	0.034	0.003
Designation	0.023	0.15	0.12	0.11	-0.00058	0.039	-0.04	0.025	0.058	1	0.012	-0.057	0.35	-0.023	0.051	0.12	-0.0026	0.017	0.21	-0.041
NumberOfPolicy	0.25	0.079	0.066	0.062	-0.019	0.014	0.016	-0.003	0.15	0.012	1	-0.068	0.15	-0.016	0.053	0.063	-0.037	0.014	0.075	-0.001
MaritalStatus	0.078	-0.084	-0.057	-0.065	0.016	-0.028	0.0037	-0.035	0.013	-0.057	-0.068	1	-0.14	0.0085	-0.026	-0.065	0.046	0.034	-0.07	-0.028
MonthlyIncome	0.32	0.57	0.38	0.36	-0.036	0.056	-0.0049	0.1	0.21	0.35	0.15	-0.14	1	0.00048	0.071	0.45	0.0057	-0.0017	0.36	0.032
Complaint	0.0019	0.014	0.014	0.0052	0.013	-0.0056	0.0007	-0.035	-0.0036	-0.023	-0.016	0.0085	0.00048	1	0.0055	0.0021	0.011	0.0069	-0.026	-0.0038
ExistingPolicyTenure	0.12	0.29	0.16	0.16	-0.0056	0.0093	-0.051	-0.016	0.063	0.051	0.053	-0.026	0.071	0.0055	1	0.26	-0.021	-0.0051	0.077	-0.0093
SumAssured	0.16	0.84	0.48	0.48	-0.033	0.018	0.016	0.039	0.1	0.12	0.063	-0.065	0.45	0.0021	0.26	1	-0.01	-0.014	0.15	0.0031
Zone	-0.011	-0.0061	0.013	-0.0014	-0.0021	-0.004	0.0019	0.021	0.024	-0.0026	-0.037	0.046	0.0057	0.011	-0.021	-0.01	1	0.023	-0.0074	0.036
PaymentMethod	0.0011	-0.013	0.0068	-0.0076	-0.011	-0.0083	0.00024	0.012	0.51	0.017	0.014	0.034	-0.0017	0.0069	-0.0051	-0.014	0.023	1	-0.017	-0.028
LastMonthCalls	0.12	0.2	0.14	0.14	0.0051	0.0036	-0.034	0.012	0.034	0.21	0.075	-0.07	0.36	-0.026	0.077	0.15	-0.0074	-0.017	1	0.0061
CustCareScore	-0.034	0.022	0.03	0.013	0.038	-0.051	-0.029	0.016	0.003	-0.041	-0.001	-0.028	0.032	-0.0038	-0.0093	0.0031	0.036	-0.028	0.0061	1

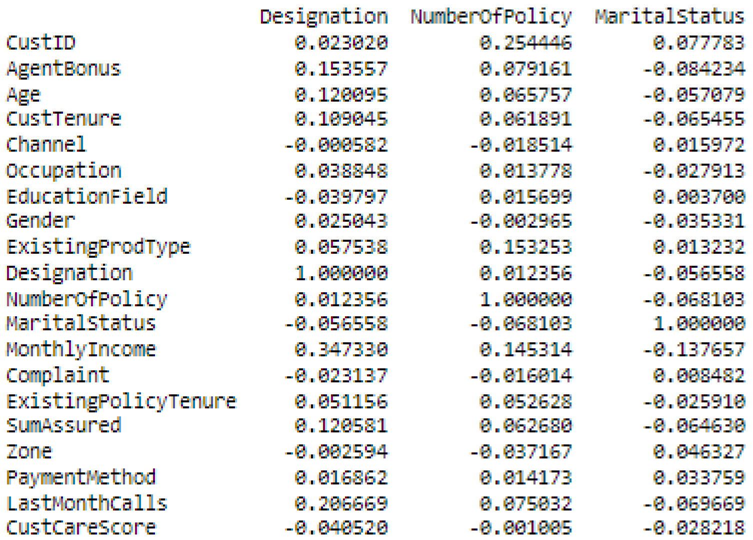
Correlations with all variables



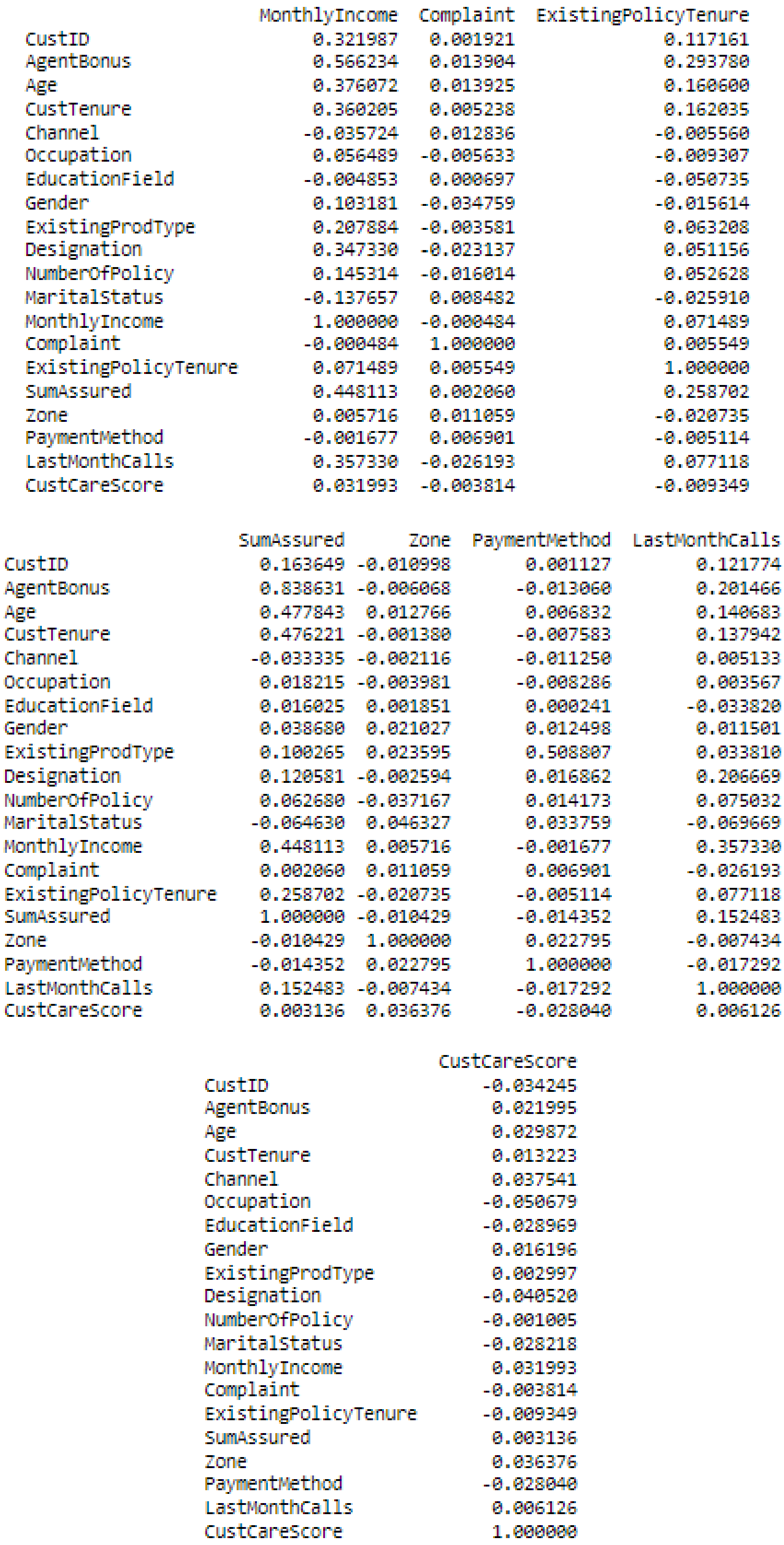
```
                      CustID  AgentBonus       Age  CustTenure   Channel
CustID              1.000000    0.194469  0.137151    0.153822 -0.003346
AgentBonus          0.194469    1.000000  0.560300    0.559312 -0.035635
Age                 0.137151    0.560300  1.000000    0.331396 -0.026437
CustTenure          0.153822    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
```



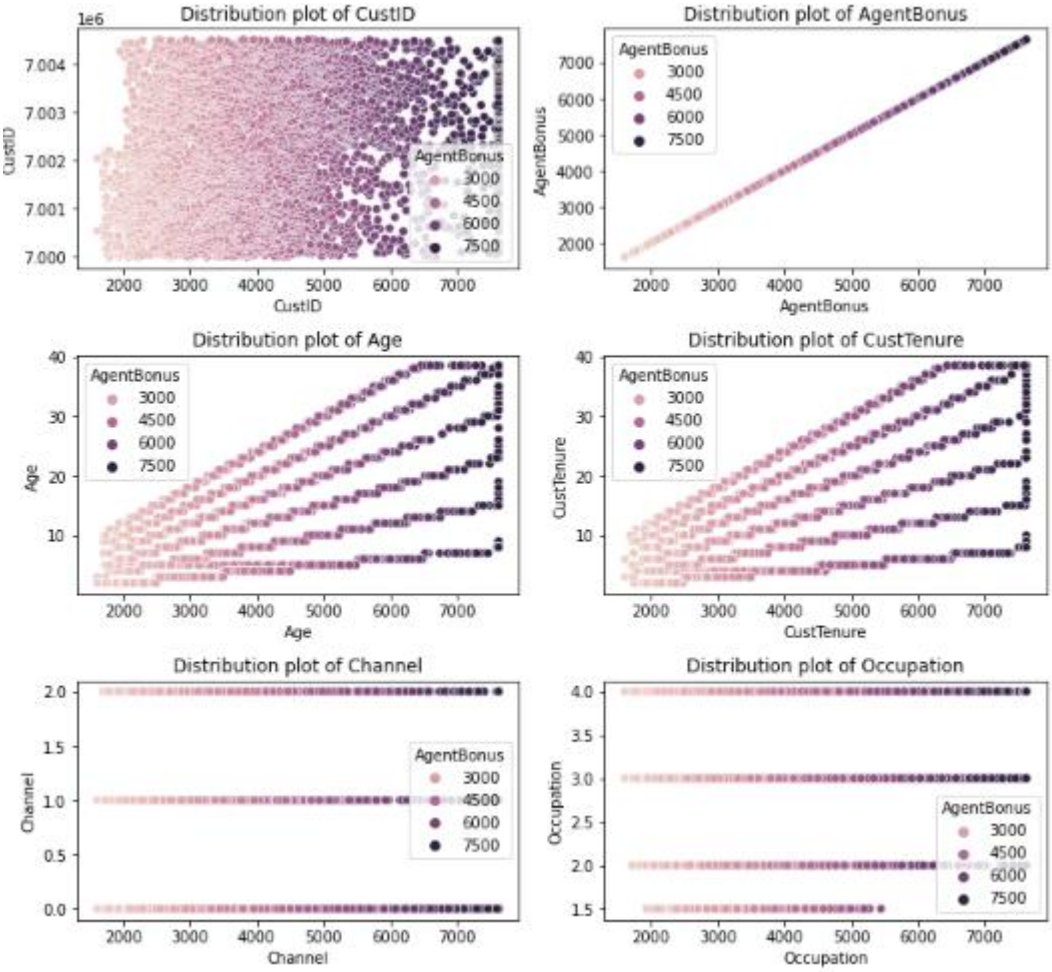
```
                     MonthlyIncome  Complaint  ExistingPolicyTenure
CustID                    0.321987   0.001921              0.117161
AgentBonus                0.566234   0.013904              0.293780
Age                       0.376072   0.013925              0.160600
CustTenure                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
```

So, from the above heat map we could observe that the highest correlations exists with Sum Assured, which simply means that the Sum Assured is the most dependent factor in determining the AgentBonus.

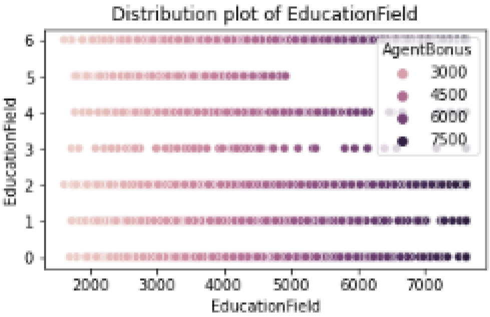
We also get some more variables which are co-related to Agent bonus and they are CUstTenure, Age, MonthlyIincome, ExistingPolicyTenure, SumAssured

Bivariate Analysis

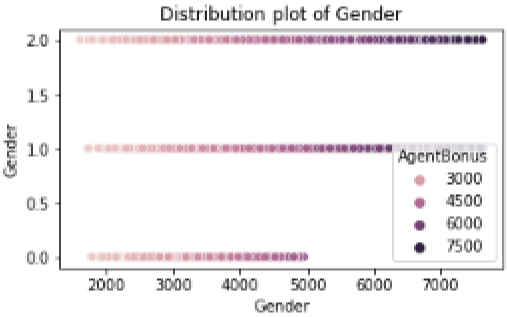
Bivariate analysis means the analysis of bivariate data. This is basically a single statistical analysis that is used to find out the relationship that exists between two value sets or two variables. The variables that are involved are X and Y. Whereas Univariate analysis is when only one variable is analysed.



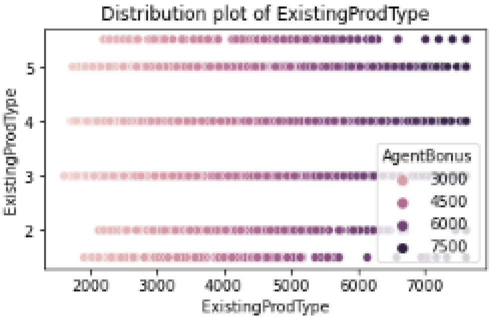
Distribution plot of EducationField



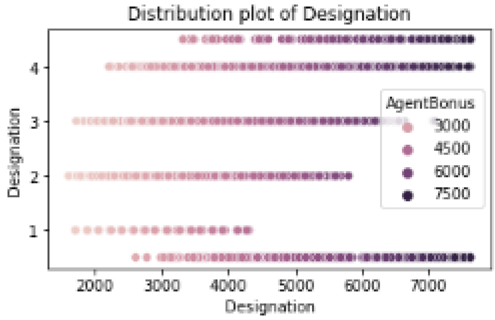
Distribution plot of Gender



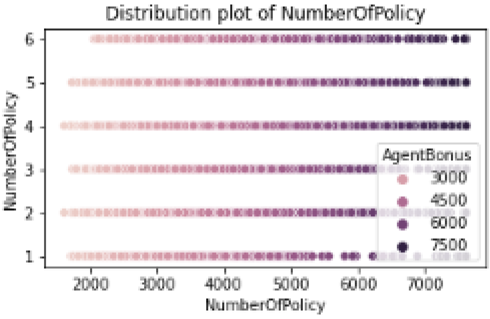
Distribution plot of ExistingProdType



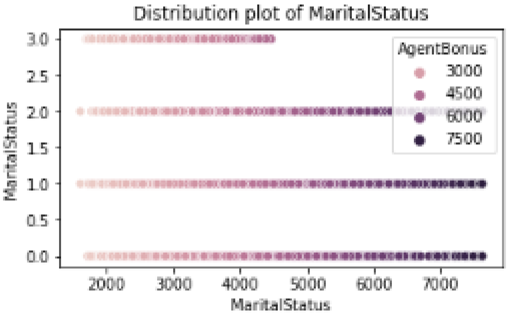
Distribution plot of Designation



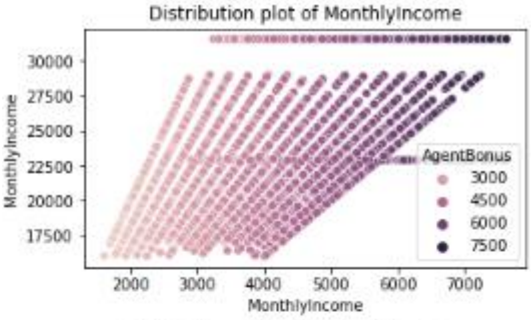
Distribution plot of NumberOfPolicy



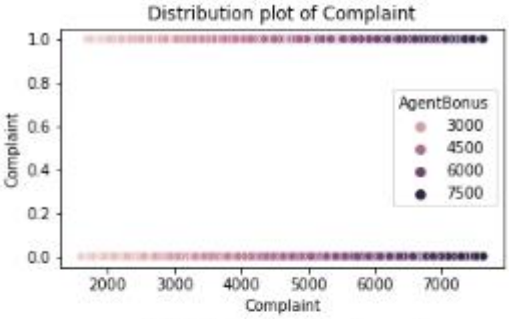
Distribution plot of MaritalStatus



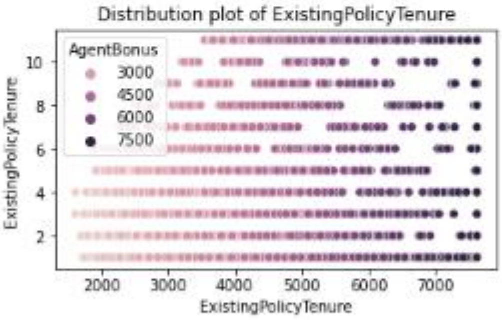
Distribution plot of MonthlyIncome



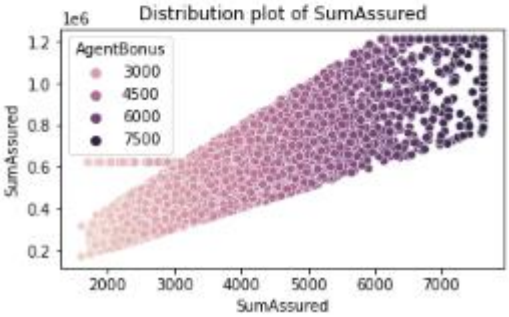
Distribution plot of Complaint

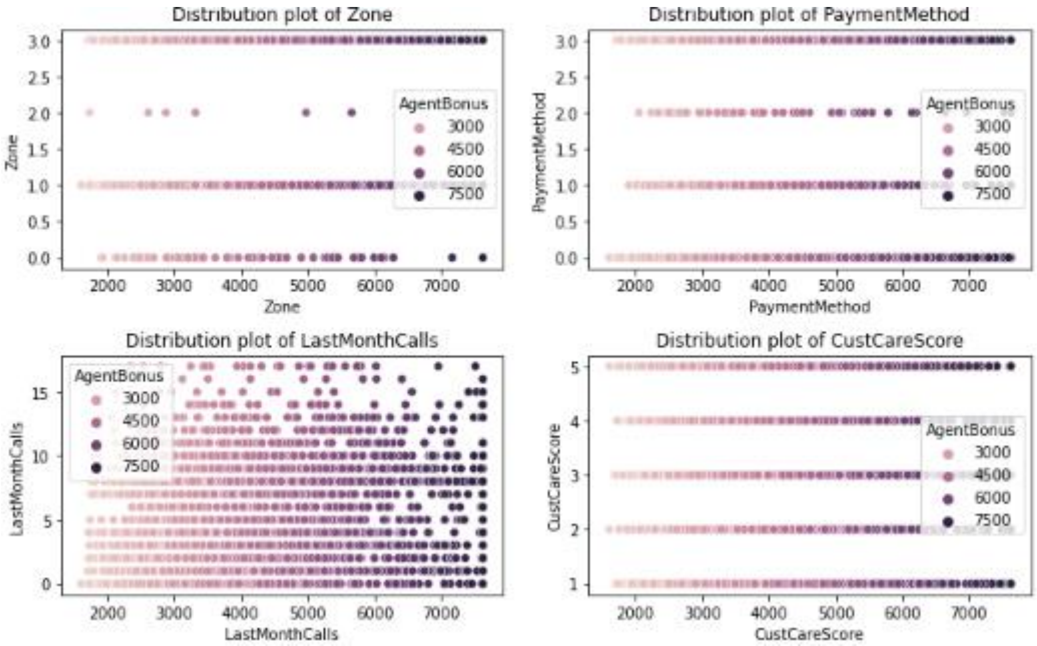


Distribution plot of ExistingPolicyTenure



Distribution plot of SumAssured





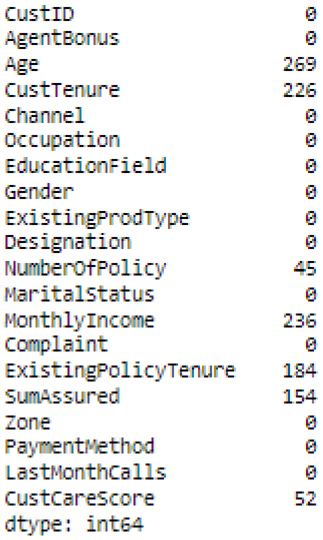
So, from the above graphs we can draw the following inferences and they are as follows:-

- Since the Customer monthly income plays an important role in any factor. So here the monthly income plays crucial role in determining the agent bonus because we know that the higher the Customer monthly income is, bigger the (sum assured), and greater the agent bonus will be.
- Here CustTenure also plays a factor in helping in determining that tells us how long the Customer sticks with a single 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 as the customer flows with the organisation.
- We do observed from the above graphs that the agent bonuses are low for the customers with lower age category on the other hand there is a higher chances of customer penetration in lower age group category.
- The Policy Tenure that exists is highly correlated as it says that, for how long does the Agents are going to be paid with the bonuses with the existing policy.
- We have also observed that the Agent bonus is highly correlated to Sum Assured which means that change in one variable would cause change to another variable and so the model results may fluctuate significantly. But Generally, it is recommended to avoid having correlated features in our dataset.

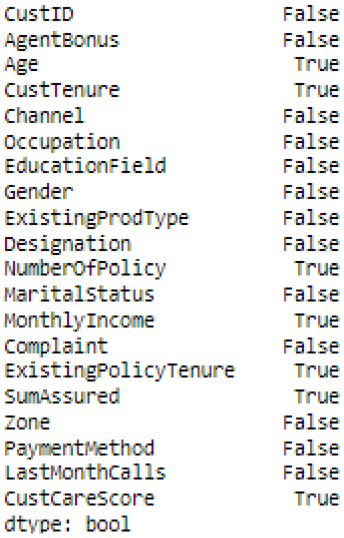
Now, certainly we are going ahead and look for the missing values present in the dataset and also use methods how to treat them.

As we can see that we have observed earlier that we have missing values in the dataset. Highlighting those values below :-

Missing values (Before Treatment)

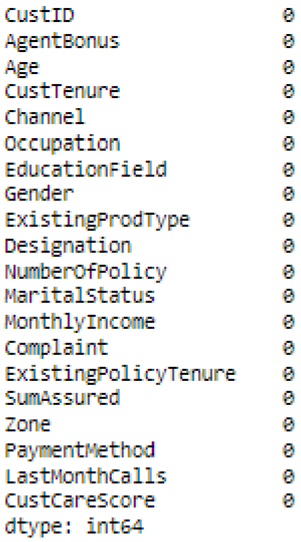


```
CustID                  0
AgentBonus              0
Age                   269
CustTenure            226
Channel                 0
Occupation              0
EducationField          0
Gender                  0
ExistingProdType        0
Designation             0
NumberOfPolicy         45
MaritalStatus           0
MonthlyIncome         236
Complaint               0
ExistingPolicyTenure  184
SumAssured            154
Zone                    0
PaymentMethod           0
LastMonthCalls          0
CustCareScore          52
dtype: int64
```

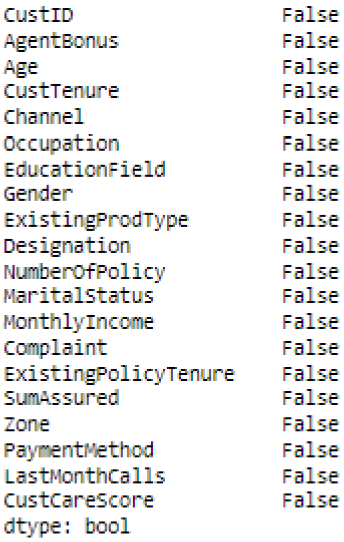


```
CustID                  False
AgentBonus              False
Age                      True
CustTenure               True
Channel                 False
Occupation              False
EducationField          False
Gender                  False
ExistingProdType        False
Designation             False
NumberOfPolicy           True
MaritalStatus           False
MonthlyIncome            True
Complaint               False
ExistingPolicyTenure     True
SumAssured               True
Zone                    False
PaymentMethod           False
LastMonthCalls          False
CustCareScore            True
dtype: bool
```

Missing values (After 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
dtype: int64
```



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

-

So, it is clear from the above graphs that the resultant dataset has no missing values because we have treated the missing values using distinctive strategies. We have imputed the two variables ‘SumAssured’ and ‘MonthlyIncome’ with mean values and rest of the variables like ‘Number of Policy’, CustCareScore, ExistingPolicyTenure, Age, CustTenure with median value.

Now, moving ahead we need to treat the Outliers that we saw, lies in the dataset, but before that we need to understand what “Outliers” actually is.

Outliers and its Treatment:

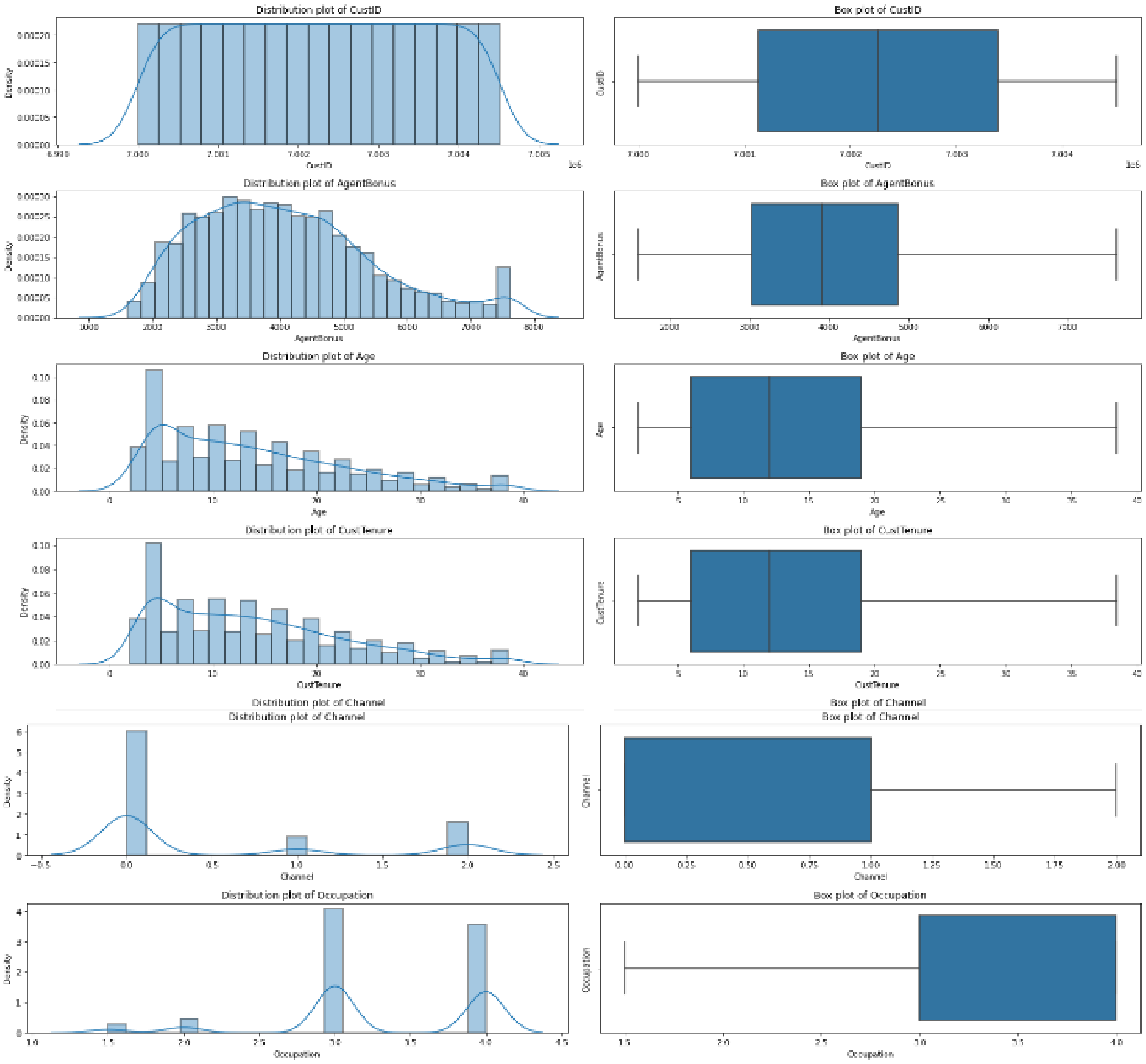
Outliers are nothing but an observation that lies an abnormal distance from other values in a random sample from a population. For example in the scores 25, 29, 3, 32, 33, 28, 27, 85, 40. In this scores 3 and 85 are clearly outliers.

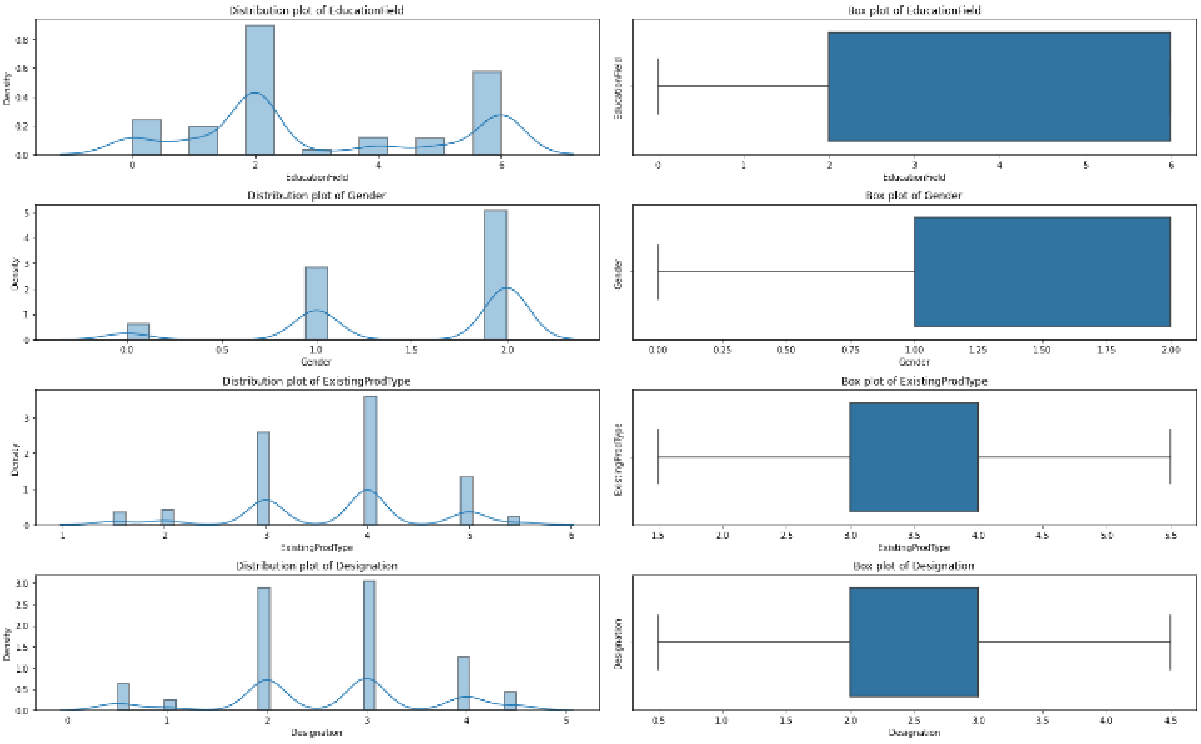
In another words Outliers are those values which lies outside 1.5*IQR (Inter Quartile Range)

From the box plots in univariate analysis, we've observed that are many outliers present in the given data, and that clearly means that treating outliers becomes compulsory for the given data.

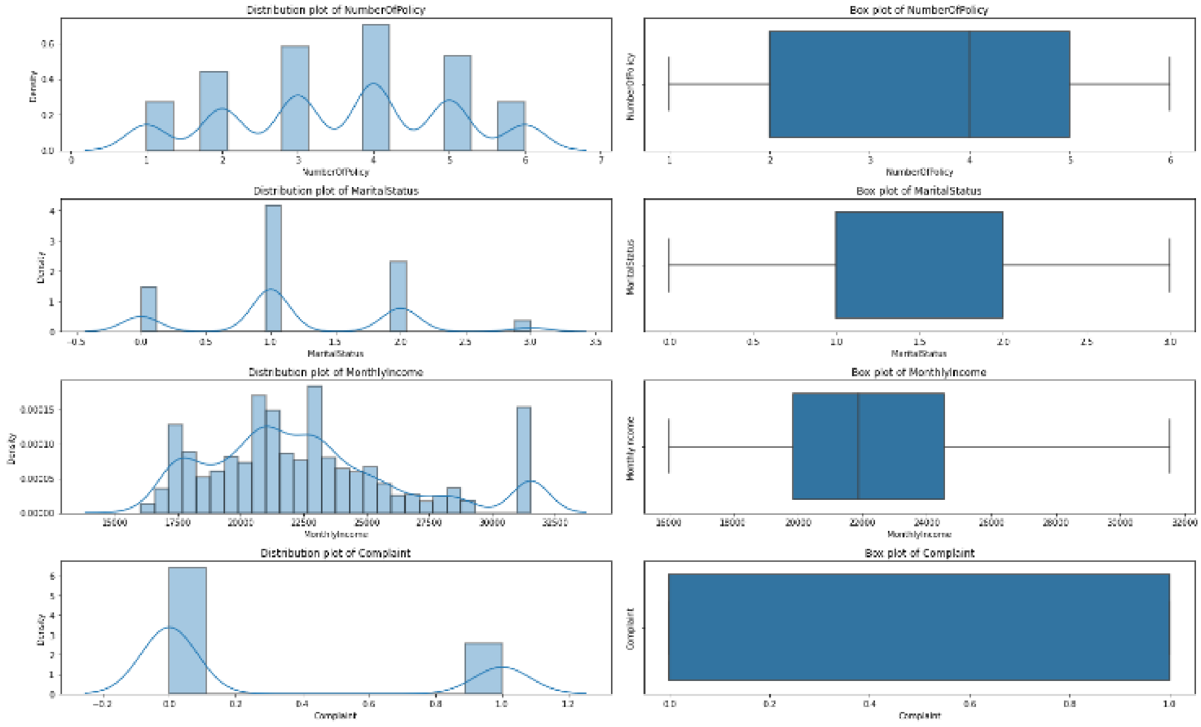
In Outliers treatment we could cap the values by replacing those observations outside the lower limit with the value of fifth percentile and those that lie above the upper limit, with the value of ninety-fifth percentile.

Graphs to show Treated Outliers

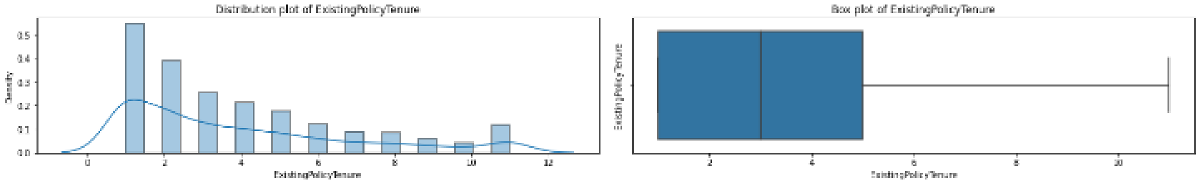




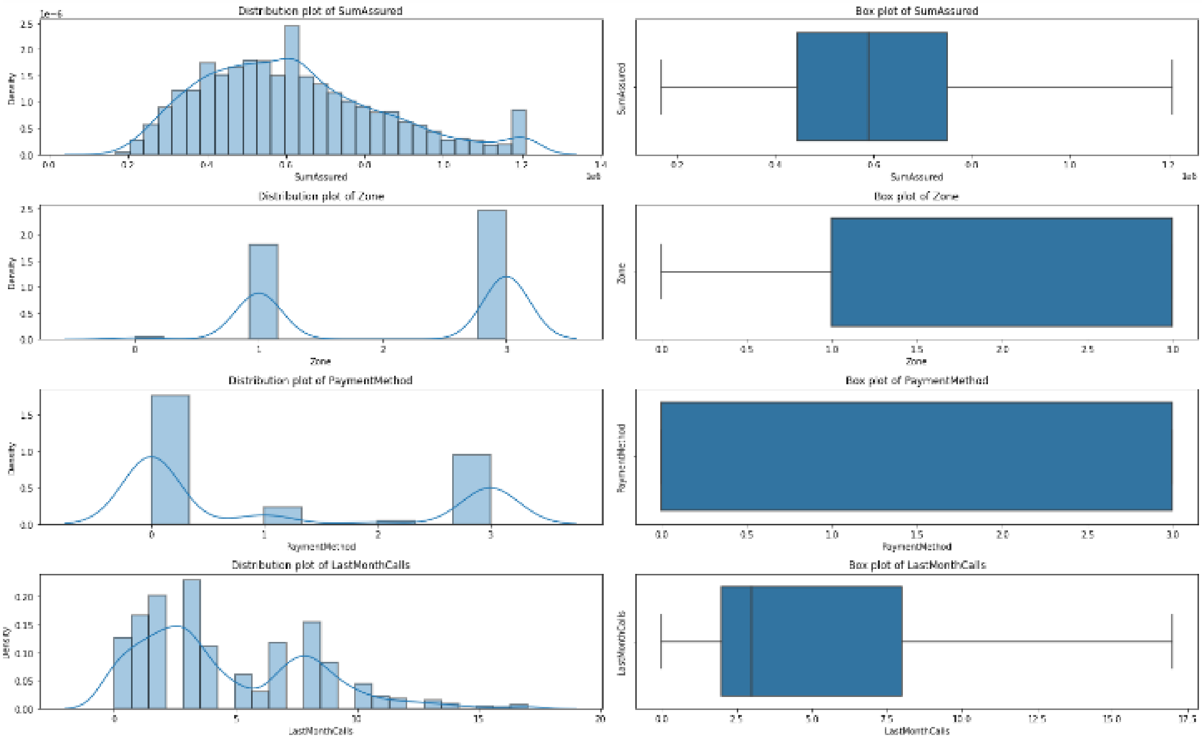
Distribution plot of EducationField
Density
0.0
0.2
0.4
0.6
0.8
0
2
4
6
EducationField
Box plot of EducationField
EducationField
0
1
2
3
4
5
6
EducationField
Distribution plot of Gender
Density
0
1
2
3
4
5
0.0
0.5
1.0
1.5
2.0
Gender
Box plot of Gender
Gender
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Gender
Distribution plot of ExistingProdType
Density
0
1
2
3
1
2
3
4
5
6
ExistingProdType
Box plot of ExistingProdType
ExistingProdType
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
ExistingProdType
Distribution plot of Designation
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
1
2
3
4
5
Designation
Box plot of Designation
Designation
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Designation

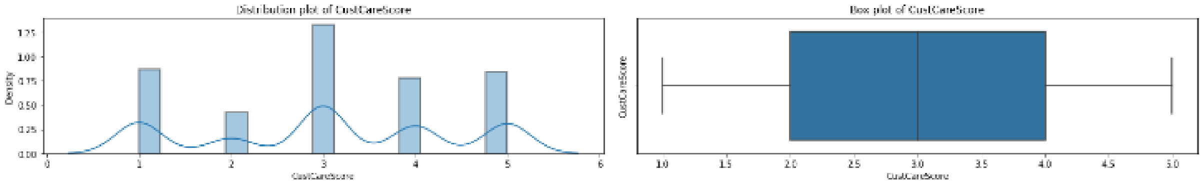


Distribution plot of NumberOfPolicy
Density
0.0
0.2
0.4
0.6
0
1
2
3
4
5
6
7
NumberOfPolicy
Box plot of NumberOfPolicy
NumberOfPolicy
1
2
3
4
5
6
NumberOfPolicy
Distribution plot of MaritalStatus
Density
0
1
2
3
4
-0.5
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
MaritalStatus
Box plot of MaritalStatus
MaritalStatus
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MaritalStatus
Distribution plot of MonthlyIncome
Density
0.00000
0.00005
0.00010
0.00015
15000
17500
20000
22500
25000
27500
30000
32500
MonthlyIncome
Box plot of MonthlyIncome
MonthlyIncome
16000
18000
20000
22000
24000
26000
28000
30000
32000
MonthlyIncome
Distribution plot of Complaint
Density
0
1
2
3
4
5
6
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Complaint
Box plot of Complaint
Complaint
0.0
0.2
0.4
0.6
0.8
1.0
Complaint



Distribution plot of ExistingPolicyTenure
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
2
4
6
8
10
12
ExistingPolicyTenure
Box plot of ExistingPolicyTenure
ExistingPolicyTenure
2
4
6
8
10
ExistingPolicyTenure





So, from the above graphs it is very clear that there are Outliers not any more in this dataset, as we have removed them from the data.

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 here 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 am using here Label encoding. And the below variables are encoded into its numeric form

Variables encoded are as follows:

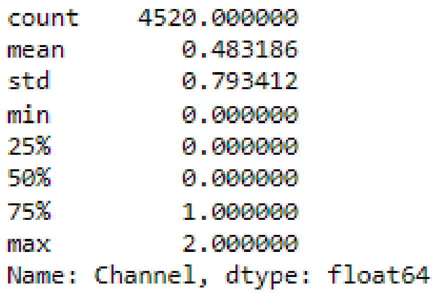
- Gender
- Education Field
- Channel
- Occupation
- Designation
- Payment Method

- Marital Status
- Zone

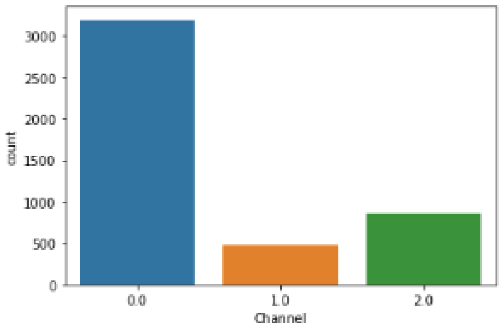
Business Insights from Exploratory Data Analysis

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 analyze the proper model to interpret the correct results.

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



```
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.

So, firstly we have to understand what unbalanced data is. Data imbalance usually used to reflect an unequal distributon of classes within a dataset.

And an effective way to handle imbalanced data is to move towards the most straightforward method for dealing with highly imbalanced dataset is called resampling which is categorised under two category.

a) **Under Sampling** b) **Over Sampling**

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

As we have observed that data set is highly imbalanced, and predicting using this data could lead to to potential danger to business decisions, so we need first to balance the imbalance data using the above mentioned techniques under 'Sampling' and then we'll be in a better positions to make recommendations.

Any business insights using Clustering (if applicable)

Clustering is basically the task of dividing the data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of siimilarity and dissimilarity between them. Clustering is generally done over unsuperivised data.

Commonly used clustering techniques includes:

- Centroid-based Clustering
- Density-based Clustering
- Distribution-based Clustering
- Hierarchial Clustering
- K means Clustering

And in the given dataset we have used K means clustering technique which has divided the dataset in three clusters

Any other 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 thorugh 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 Esat shows very minimum hikes. It clearly points out that the company presnce in these regions are limited therefore the Compnay 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.