Business Report

Capstone Project

LI_BFSI_01+Life+Insurance+Sales

1st Submission

Date: 11-Dec-2022

Created by Amit Jain

Table of Contents

List of Figure

1. Introduction of the business Problem

Introduction: This report explains the business requirements and provide the detailed solution based on the data provided for each problem statement. given in the assignment.

1.1 Defining Problem Statement:

"The dataset belongs to a leading life insurance company. The company wants to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents are most important."

Dataset for Problem : **Sales.xlsx**

To understand the problem, Life insurance Company has given randomly collected sample of 4520 Customer records data in the sales.xlsx file, which have pattern of the Customer information about, their purchased Insurance Plans, their tenure with Insurance company, Sum insured and some more information about customer. Company has also given information about AgentBonus given to insurance company Agents who made customer purchase their plans.

Insurance company wants to analyze this sample data and Predict Bonus for their agents, based on past sample data , so that they can understand more about internal Agents, who bring sell to the Company. Company also want to know, if there are any low performing Agents, which requires any special training to increase growth or if they need any assistance. Company also want to build environment to encourage Agents, who are very good in selling plans, and motivate others by example of giving rewards to good performing Agents.

1.2 Need of the study/project:

It is very important for any company to know their Customers, at the same time its equally important to know their own employees, who serve end Customers. This is a very generic problem as well as requirement, specially in Insurance and Sales sectors to know their own employees, identifying good performing Agents/Sales person and low performing employees. So that they can know own capacity and can plan for the future growth. And based on this analysis, they can deploy their good performing Agents/Sales person in tough market and plan for good trainings to up scaling low performing employees.

1.3 Understanding business/social opportunity:

This kind of analysis for knowing own Agents/Sales person gives a good opportunity for any company to grow. They build a good healthy environment to encouraging good agents and upscaling and providing trainings assistance to not so good employees. It will also give a proper benchmark , to motivate other gents. By looking at the formula, management can understand, what parameters give more weightage for getting good compensation and good bonus amount. And it will remove partiality and bad judgement of the top level management.

2. Data Export :

2.1 Understanding how data was collected in terms of time, frequency and methodology:

Assume that the data follows a normal distribution. In reality, the normality assumption may not always hold if the sample size is small.

Also we have not given specific time frame, when this data elements were collected and how frequently this kind of information is being collected. Though we assume this all sample data belong to same time frame window for consigning it as a Randomly collected records, without any influence .

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

Import the data: Imported the data using Python notebooks and analyzed the effects of Education and Occupations over salary field.

This is how the data look like:

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

Data dictionary:

CustID =>Unique customer ID
AgentBonus =>Bonus amount given to each agents in last month
Age =>Age of customer
CustTenure =>Tenure of customer in organization
Channel =>Channel through which acquisition of customer is done
Occupation =>Occupation of customer
EducationField =>Field of education of customer
Gender =>Gender of customer
ExistingProdType =>Existing product type of customer
Designation =>Designation of customer in their organization
NumberOfPolicy =>Total number of existing policy of a customer
MaritalStatus =>Marital status of customer
MonthlyIncome =>Gross monthly income of customer

Complaint =>Indicator of complaint registered in last one month by customer
ExistingPolicyTenure =>Max tenure in all existing policies of customer
SumAssured =>Max of sum assured in all existing policies of customer
Zone =>Customer belongs to which zone in India. Like East, West, North and South
PaymentMethod =>Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
LastMonthCalls =>Total calls attempted by company to a customer for cross sell
CustCareScore =>Customer satisfaction score given by customer in previous service call

2.3 Understanding of attributes (variable info, renaming if required):

Data description:

	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

Insights:

1. Agent Bonus : Bonus given to Agents, as well as Target variable. Minimum bonus given as 1605 and Max is 9608
2. Age: Customers of all age group from 2 years to 58 years, there are also customers with 0 Age, needs to be corrected
3. Customer Tenure: Many customers associated with Company from their Birth
4. Existing Prod Type : There are only 6 insurance products

5. Number of Policy: Customer can have multiple policies, from insurance company , for their family members or Self , if it is blank, that needs to be corrected
6. Monthly income Ranging between about 16K to 38K , if it is blank, needs to be corrected
7. Complaint: 0 to 1 complaint in last one month
8. Gender showing as 3 ,: this needs to be corrected
9. Occupation : Different Occupations, there can be multiple business with same name, needs to be corrected
10. Education : can have similar names of degrees, needs to be corrected
11. Designation: can be duplicate
12. existing Policy tenure. if it is blank needs to be corrected
13. Sum Assured : if it is blank needs to be corrected
14. Customer care score. needs to be corrected if it is blank

Data info:

```
#   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
```

Insights:

1. Customer id can be removed from data, as it will not be required for Bonus prediction
2. Channel , occupation, Education, Gender, Designation , Marital Status and all other Object type fields should be converted to Numeric format in order to use it in prediction Model.

3. Exploratory data analysis

3.1 Addition of new variables (if required)

AgentBonus is the amount of Bonus given to Insurance agents based, what policy he/she has sold to customers, and what profit Company might have taken. Since one customer can buy multiple policies together, so it will not be fare to Compare AgentBonus amount of one agent with another.

May be AgentBonus for one agent is more (because hist Customer bought 5 policies together). but Average of per policy of that bonus amount can be lower than Single Policy of bigger SumAssured.

- So lets create additional Field, "AgentBonus_Per_Policy", which should be calculated by:

  AgentBonus_Per_Policy = AgentBonus/NumberOfPolicy

- Also lets create a new field called SumAssured_Per_Policy, which should be calculated by :

  SumAssured_Per_Policy = SumAssured/NumberOfPolicy

Now we have got this list of fields, which in new data set:

```
Index(['CustID', 'AgentBonus', 'Age', 'CustTenure', 'Channel', 'Occupation',
       'EducationField', 'Gender', 'ExistingProdType', 'Designation',
       'NumberOfPolicy', 'MaritalStatus', 'MonthlyIncome', 'Complaint',
       'ExistingPolicyTenure', 'SumAssured', 'Zone', 'PaymentMethod',
       'LastMonthCalls', 'CustCareScore', 'AgentBonus_Per_Policy',
       'SumAssured_Per_Policy'],
      dtype='object')
```

whereas Old data set fields are:

```
Index(['CustID', 'AgentBonus', 'Age', 'CustTenure', 'Channel', 'Occupation',
       'EducationField', 'Gender', 'ExistingProdType', 'Designation',
       'NumberOfPolicy', 'MaritalStatus', 'MonthlyIncome', 'Complaint',
       'ExistingPolicyTenure', 'SumAssured', 'Zone', 'PaymentMethod',
       'LastMonthCalls', 'CustCareScore'],
      dtype='object')
```

We have got 2 new fields , in our data set.

3.2 Missing Value treatment (if applicable)

We have taken NULL counts for all of our attributes and NULL value counts are as follows:

```
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
AgentBonus_Per_Policy     45
SumAssured_Per_Policy    199
dtype: int64


Total NULL values in data : 1410
Total data elements in Sample data are: 99440
```

1410 is the total Null counts, which includes NULL counts for our 2 newly created fields as well, which are :

```
AgentBonus_Per_Policy     45
SumAssured_Per_Policy    199
```

If we subtract this count, then actual total number of missing values count is 1166

NULL Treatment one by one for each NULL field:

- **Treating NULL Age:**

Age should be Greater than or Equals to Customer Tenure as well as Existing Policy Tenure. So let's give first preference to Customer Tenure and then existing Policy Tenure for correcting NULL "**Age**"

There were total 269 NULL records in Age field. We will consider Policy Started , as child gets Birth and Age should be replaced with "CustomerTenure " and then "ExistingPolicyTenure".

- **Treating NULL CustomerTenure:**

We will replace NULL "CustomerTenure " field with "ExistingPolicyTenure" and if "ExistingPolicyTenure" is also not available then we should replace "CustomerTenure " with "Age" field.

- **Treating "Existing Policy tenure":**

Similarly, if Cust "Existing Policy tenure" is NULL, then replace it with "Customer Tenure" and second preference will be to replace it with "Age"

- **Treating "NumberOfPolicy":**

Best way to Treat this field is to replace NULL with Mode.

- **Treating NULLs for "AgentBonus_Per_Policy":**

Since we have already computed NumberOfPolicy Field, then for treating missing values for "AgentBonus_Per_Policy", we will again use our formula for this field.

df2['AgentBonus_Per_Policy']=df2['AgentBonus']/df2['NumberOfPolicy']

and NULL has been treated.

- **Treating NULL for "MonthlyIncome" and "SumAssured":**

Best way to Treat this field is to use KNN Imputers for missing NULL values.

I have used KNNImputer from SKLearn library and used this method for imputing missing values for **"MonthlyIncome" and "SumAssured".**

- **Treating NULL for "CustCareScore":**

I have used Median for this value and treating missing values for "CustCareScore".

Now we are left with only SumAssured field , which have NULL values, but we will impute this, as we are using SumAssured_per_policy field instead, which we have already corrected and we are going to drop field SumAssured in our next Step.

3.3 Duplicate checks:

We checked duplicity in CustID field, and we did not find any Duplicate records.

Total Count for Duplicate records are : 0

3.4 Removal of unwanted variables (if applicable)

Since We have used AgentBonus and SumAssured for building our new field, so we don't need them any more for our analysis, Also We dont need field CustId for our analysis, as it will not add any benefits in our analysis.

So we will drop these 3 fields from our data:

AgentBonus , SumAssured , CustID

And we have added these 2 fields in our data :

AgentBonus_per_policy , SumAssured_per_policy

We will also check for co-relations, in further sections and if we find that any field don't have relation with Target field , then we will drop those fields as well.

	Complaint	LastMonthCalls	CustCareScore
Age	0.019496	0.123837	0.029853
CustTenure	0.00685	0.142982	0.013919
ExistingProdType	-0.003486	0.033191	0.00411
NumberOfPolicy	-0.016014	0.075138	-0.001005
MonthlyIncome	-0.004815	0.34393	0.035751
Complaint	1	-0.02632	-0.003814
ExistingPolicyTenure	-0.005082	0.126951	0.013532
LastMonthCalls	-0.02632	1	0.006386
CustCareScore	-0.003814	0.006386	1
AgentBonus_Per_Policy	0.025091	0.038717	-0.005319
SumAssured_Per_Policy	0.023838	0.03474	-0.013488

We have built co-relation matrix for all the fields and checked it's relation with Targeted field. This is co-relation with AgentBonus:

Complaint : 0.025091

LastMonthCalls: 0.038717

CustCareScore: -0.005319

we have checked that following fields have very minute impact on targeted fields and these can be dropped as well:

3.5 Outlier removal (if applicable):

We have analyzed data from the boxplot:

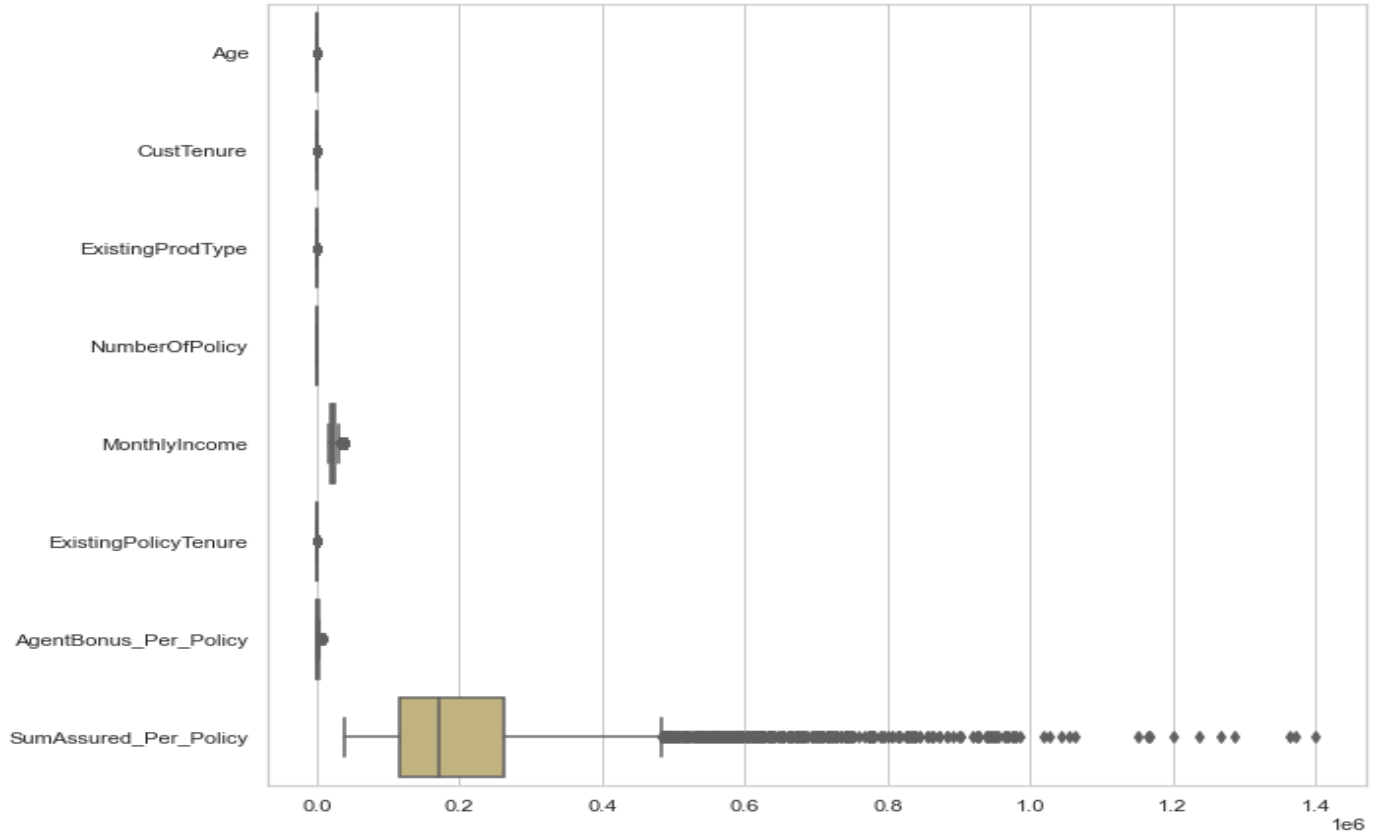

Figure 1 Boxplot

Action :

We do see outliers in about each of the field of sample data, but data scales are different,

Example: Age have limited discrete values, AgentBonus will be always lesser than SumAssured and it will be just a fraction of percentage of SumAssured . Customer Tenure, Existing Policy tenure, and other Age type fields are also discrete in nature.

So I would not like to treat outliers, as People have different income slaves and can take multiple or Single Policy as well.

And I really want to understand nature of other fields, how they are affecting our target field, AgentBonus , so we will not treat the outliers

3.6 Univariate Analysis

3.6.1 Taking counts for each Categorical field :

We have taken counts for all the Categorical fields in Sample data set and this is the result for it :

```
Field name is  CHANNEL : and Count for this categories are  3
Online                  468
Third Party Partner     858
Agent                  3194
Name: Channel, dtype: int64


Field name is  OCCUPATION : and Count for this categories are  5
Free Lancer              2
Laarge Business        153
Large Business         255
Small Business        1918
Salaried              2192
Name: Occupation, dtype: int64


Field name is  EDUCATIONFIELD : and Count for this categories are  7
MBA                     74
UG                     230
Post Graduate          252
Engineer               408
Diploma                496
Under Graduate        1190
Graduate              1870
Name: EducationField, dtype: int64


Field name is  GENDER : and Count for this categories are  3
Fe male     325
Female     1507
Male       2688
Name: Gender, dtype: int64


Field name is  DESIGNATION : and Count for this categories are  6
Exe                    127
VP                     226
AVP                    336
Senior Manager         676
Executive             1535
Manager               1620
Name: Designation, dtype: int64

Field name is  MARITALSTATUS : and Count for this categories are  4
Unmarried      194
Divorced       804
Single        1254
Married       2268
Name: MaritalStatus, dtype: int64
```

```
Field name is  ZONE : and Count for this categories are  4
South       6
East       64
North    1884
West     2566
Name: Zone, dtype: int64

Field name is  PAYMENTMETHOD : and Count for this categories are  4
Quarterly        76
Monthly         354
Yearly         1434
Half Yearly    2656
Name: PaymentMethod, dtype: int64
```

Insights:

1. Channel as "Agent" has maximum count of 3194, and least no of channel is Online. with 468
2. OCCUPATION : Maximum Policy Holder are Salaried Employees, where as free Lancer and Large business people are very less in Insurance .
3. OCCUPATION : "Laarge Business" and "Large Business" are same, it needs to be merged.
4. EDUCATIONFIELD: "UG" and "Under Graduate" are same category and needs to merged.
5: EDUCATIONFIELD: Highest Policy holders are Graduate and Least are MBA degree Holder.
6: GENDER : "Fe male" and "Female" are same, and needs to merge into one category
7. Maximum policy Holders are male with 2688 Count.
8. DESIGNATION : Exe and Executive are same and can be merge into one Category.
9. South Zone and followed by east zone have very less number of customers , need to work in that area
10. Customers like to pay half yearly and yearly payments method most for the payment of their premiums.

3.6.2 Merge similar categorical values:

From taking counts of each Categorical field, we have seen that , there are some values, which we can merge into Single value.

Example:

OCCUPATION : "Laarge Business" and "Large Business can be merged into Single field “Large business”
EDUCATIONFIELD: "UG" and "Under Graduate" can be merged into “Under graduate”
GENDER : "Fe male" and "Female" are same and should be merged into “Female”
DESIGNATION : Exe and Executive are same and should be merged into “Executive”

First of all take the counts for above listed 4 Categorical field and check before merger counts:

```
Name: EducationField, dtype: int64
Graduate          1870
Under Graduate    1190
Diploma            496
Engineer           408
Post Graduate      252
UG                 230
MBA                 74


Name: Gender, dtype: int64
Male       2688
Female     1507
Fe male     325


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

Name: Designation, dtype: int64
Manager           1620
Executive         1535
Senior Manager     676
AVP                336
VP                 226
Exe                127
```

Above highlighted in yellow field should be merged . we have performed merging of above listed fields

3.6.3 Taking counts for each Categorical fields after merging similar values:

After performing merging of similar values, this is how the counts look like :

```
Field name is  CHANNEL : and Count for this categories are  3
Online                  468
Third Party Partner     858
Agent                  3194
Name: Channel, dtype: int64


Field name is  OCCUPATION : and Count for this categories are  4
Free Lancer           2
Large Business      408
Small Business     1918
Salaried           2192
```

```
Field name is  EDUCATIONFIELD : and Count for this categories are  6
MBA                 74
Post Graduate      252
Engineer           408
Diploma            496
Under Graduate    1420
Graduate          1870

Field name is  GENDER : and Count for this categories are  2
Female    1832
Male      2688

Field name is  DESIGNATION : and Count for this categories are  5
VP                 226
AVP                336
Senior Manager     676
Manager           1620
Executive         1662

Field name is  MARITALSTATUS : and Count for this categories are  4
Unmarried     194
Divorced      804
Single       1254
Married      2268

Field name is  ZONE : and Count for this categories are  4
South       6
East       64
North    1884
West     2566

Field name is  PAYMENTMETHOD : and Count for this categories are  4
Quarterly        76
Monthly         354
Yearly         1434
Half Yearly    2656
```

Insights:

1. Maximum number of policies sold by Agents and very less policies sold by Online channel.
2. Maximum number of policy holders are Salaried with count of 2192 records and least number of Policy holders are free Lancer.
3. Maximum number of policy holders are Graduate in Education with count of 1870 records, and MBA holders are least with no of 74.
4. Male are Maximum number of policy holders as compared to female records.
5. Married people are Maximum number of policy holders whereas unmarried people don’t prefer taking Insurance policy.
6. There are maximum policies sold in West region of the country, whereas South reason have very less 6 policies sold.
7. Maximum number of policy holders prefer to pay Half yearly premiums .

3.6.4 Generate Histogram and Boxplot for Sample data:

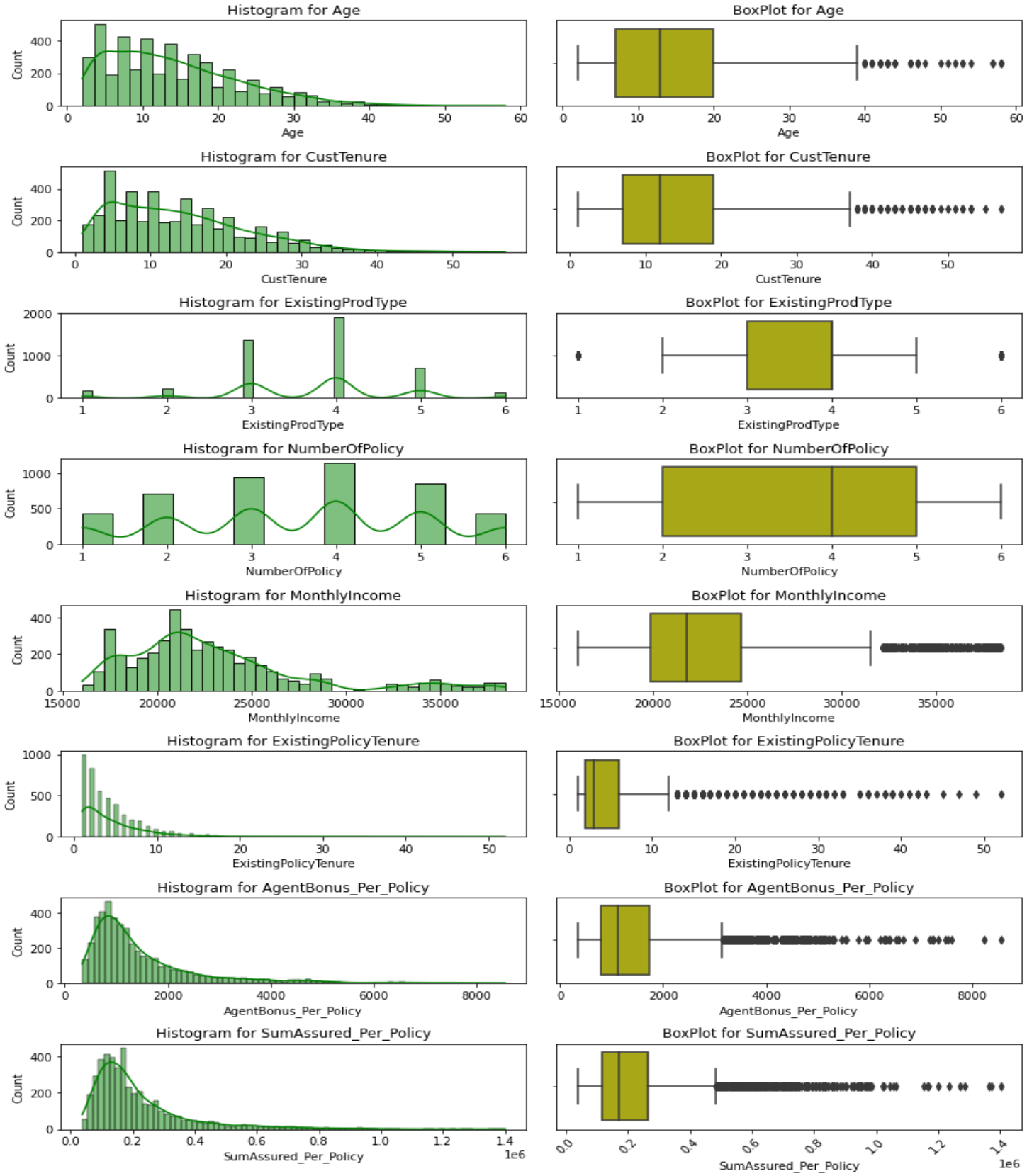

Figure 2 : Histogram and Boxplot

Insights:

1. Data is not 100% normally distributed .
    a. Age: we can have any range of Age for customer, maximum Age is 58 years in data, which is valid.
    b. CustomerTenure is about similar to Age field, Customertenure can also range from 1 year to Maximum of Age limit of the customer.
    c. Number of Policy is a discrete field, which have 6 possible values in it, maximum customers have taken 4 policies.
    d. Monthly Income field is a continuous field, and its ranging from 15K to about 40k , which is normal. Data is right Skewed.
    e. Existing Policy tenure shows maximum Customer are part of under 10 years and there are a very few customers from 10 to 58 years of existing Policies.
    f. AgentBonus_per_policy and SumAssured_per_policy is Continuous fields, and have similar distribution of data , this data is also right skewed.
2. About all the fields have some outliers.

3.6.5 Check data skewness :

```
Age                      0.960101
CustTenure               0.928995
ExistingProdType        -0.401100
NumberOfPolicy          -0.108161
MonthlyIncome            1.373508
ExistingPolicyTenure     3.440053
AgentBonus_Per_Policy    2.144651
SumAssured_Per_Policy    2.373659
dtype: float64
```

Insights:

1. data is not 100% Normally distributed
2. About all data have Outliers, and which is possible because , insurance Policies , can be different for many customers, depending on their Need , worth and sum insured.
3. All data is Slightly Right Skewed .

3.7 Bivariate Analysis

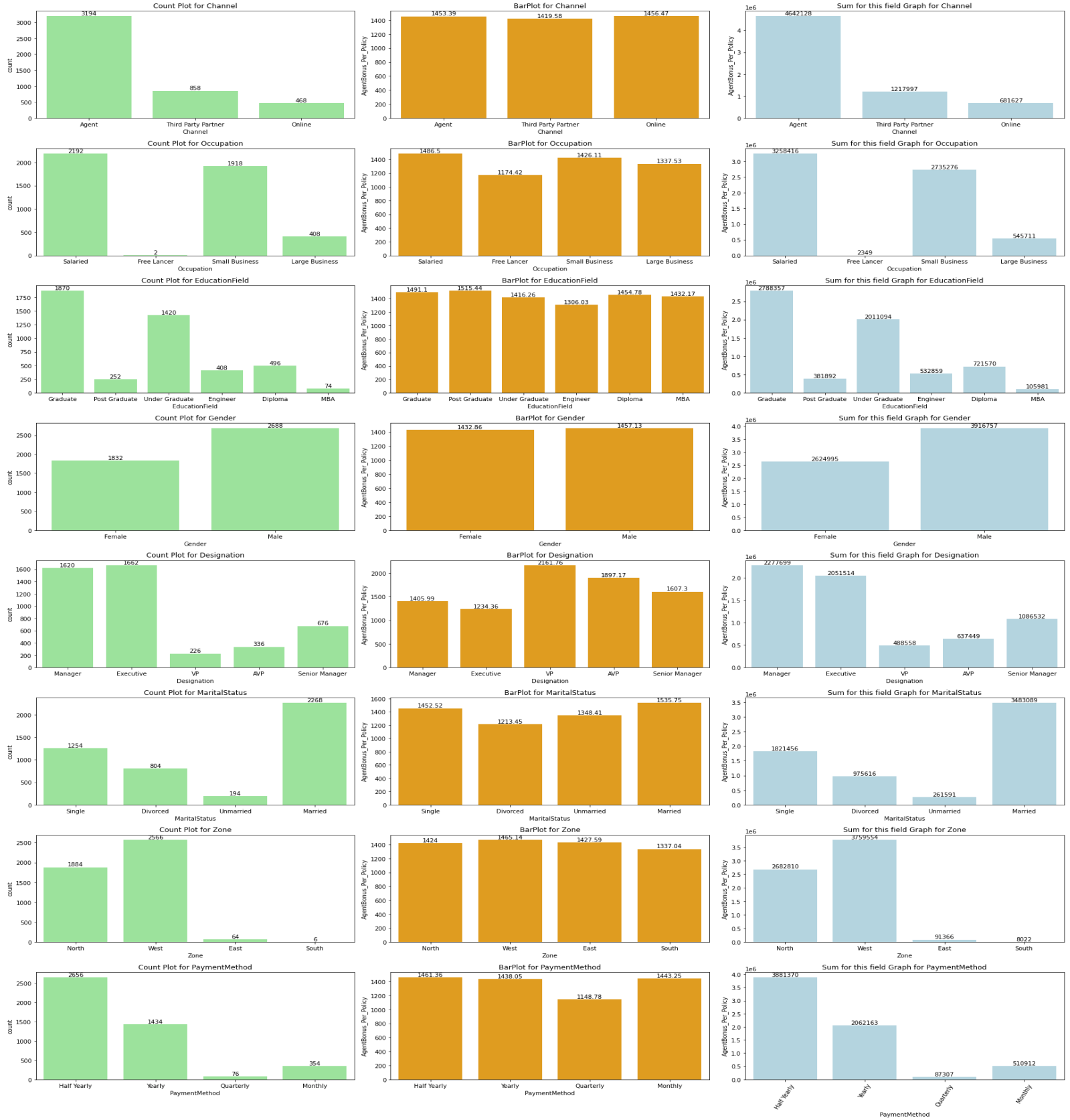

Figure 3 Mean, Count and Sum graph for categorical fields

I have performed, Bivariate analysis of all other the fields, with respect to Agent Bonus per policy as this is our target variable, and take Mean, Counts and Sum of Categorical fields.

```
Field name is  CHANNEL : and Distinct Count for this Column is  3
Field name is  CHANNEL : and it's distinct  categories are  ['Agent' 'Third Party Par
tner' 'Online']
Counts are each categories of this column are :
Online                 468
Third Party Partner     858
Agent                  3194
Name: Channel, dtype: int64

Average AgentBonus for each categories of this Column are:  Channel
Third Party Partner    1419.576690
Agent                  1453.390106
Online                 1456.468768
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  Channel
Online                 6.816274e+05
Third Party Partner    1.217997e+06
Agent                  4.642128e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  OCCUPATION : and Distinct Count for this Column is  4
Field name is  OCCUPATION : and it's distinct  categories are  ['Salaried' 'Free Lanc
er' 'Small Business' 'Large Business']
Counts are each categories of this column are :
Free Lancer          2
Large Business     408
Small Business    1918
Salaried          2192
Name: Occupation, dtype: int64

Average AgentBonus for each categories of this Column are:  Occupation
Free Lancer       1174.416667
Large Business    1337.526348
Small Business    1426.108568
Salaried          1486.503817
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  Occupation
Free Lancer       2.348833e+03
Large Business    5.457107e+05
Small Business    2.735276e+06
Salaried          3.258416e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  EDUCATIONFIELD : and Distinct Count for this Column is  6
Field name is  EDUCATIONFIELD : and it's distinct  categories are  ['Graduate' 'Post
Graduate' 'Under Graduate' 'Engineer' 'Diploma' 'MBA']
Counts are each categories of this column are :
MBA                 74
Post Graduate      252
```

```
Engineer           408
Diploma            496
Under Graduate    1420
Graduate          1870
Name: EducationField, dtype: int64

Average AgentBonus for each categories of this Column are:  EducationField
Engineer          1306.026511
Under Graduate    1416.263650
MBA               1432.172297
Diploma           1454.777688
Graduate          1491.099947
Post Graduate     1515.442857
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  EducationField
MBA               1.059808e+05
Post Graduate     3.818916e+05
Engineer          5.328588e+05
Diploma           7.215697e+05
Under Graduate    2.011094e+06
Graduate          2.788357e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  GENDER : and Distinct Count for this Column is  2
Field name is  GENDER : and it's distinct  categories are  ['Female' 'Male']
Counts are each categories of this column are :
Female    1832
Male      2688
Name: Gender, dtype: int64

Average AgentBonus for each categories of this Column are:  Gender
Female    1432.857442
Male      1457.126990
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  Gender
Female    2.624995e+06
Male      3.916757e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  DESIGNATION : and Distinct Count for this Column is  5
Field name is  DESIGNATION : and it's distinct  categories are  ['Manager' 'Executive
' 'VP' 'AVP' 'Senior Manager']
Counts are each categories of this column are :
VP                 226
AVP                336
Senior Manager     676
Manager           1620
Executive         1662
Name: Designation, dtype: int64

Average AgentBonus for each categories of this Column are:  Designation
Executive         1234.364811
Manager           1405.986770
Senior Manager    1607.296031
```

```
AVP               1897.170685
VP                2161.760324
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  Designation
VP                4.885578e+05
AVP               6.374494e+05
Senior Manager    1.086532e+06
Executive         2.051514e+06
Manager           2.277699e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  MARITALSTATUS : and Distinct Count for this Column is  4
Field name is  MARITALSTATUS : and it's distinct  categories are  ['Single' 'Divorced
' 'Unmarried' 'Married']
Counts are each categories of this column are :
Unmarried     194
Divorced      804
Single       1254
Married      2268
Name: MaritalStatus, dtype: int64

Average AgentBonus for each categories of this Column are:  MaritalStatus
Divorced     1213.452177
Unmarried    1348.409622
Single       1452.517145
Married      1535.753380
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  MaritalStatus
Unmarried    2.615915e+05
Divorced     9.756155e+05
Single       1.821456e+06
Married      3.483089e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  ZONE : and Distinct Count for this Column is  4
Field name is  ZONE : and it's distinct  categories are  ['North' 'West' 'East' 'Sout
h']
Counts are each categories of this column are :
South       6
East       64
North    1884
West     2566
Name: Zone, dtype: int64

Average AgentBonus for each categories of this Column are:  Zone
South    1337.038889
North    1423.996665
East     1427.593229
West     1465.141959
Name: AgentBonus_Per_Policy, dtype: float64
```

```
Total sum value of AgentBonus for each category for Column are :  Zone
South    8.022233e+03
East     9.136597e+04
North    2.682810e+06
West     3.759554e+06
Name: AgentBonus_Per_Policy, dtype: float64


Field name is  PAYMENTMETHOD : and Distinct Count for this Column is  4
Field name is  PAYMENTMETHOD : and it's distinct  categories are  ['Half Yearly' 'Yea
rly' 'Quarterly' 'Monthly']
Counts are each categories of this column are :
Quarterly        76
Monthly         354
Yearly         1434
Half Yearly    2656
Name: PaymentMethod, dtype: int64

Average AgentBonus for each categories of this Column are:  PaymentMethod
Quarterly      1148.781579
Yearly         1438.049361
Monthly        1443.254661
Half Yearly    1461.359130
Name: AgentBonus_Per_Policy, dtype: float64

Total sum value of AgentBonus for each category for Column are :  PaymentMethod
Quarterly      8.730740e+04
Monthly        5.109121e+05
Yearly         2.062163e+06
Half Yearly    3.881370e+06
Name: AgentBonus_Per_Policy, dtype: float64
```

Insights:

Based on the Above counts and unique values of Different Categorical fields of the data, we can come to this conclusion:

1. Field name is CHANNEL : and Distinct Count for this Column is 3
    - and it's distinct categories are ['Agent' 'Third Party Partner' 'Online']
    - There are maximum no of insurance taken from Agents with no of 3194 , and least no of insurance taken by online channel with count of 468.
    - On an average all three type of channels receiving about similar bonus, there is very slight difference in Bonus amount, if channel is different.
    - maximum total Bonus received by Agents Categories and Online policy giving agents get least amount of Bonus.
2. Field name is OCCUPATION : and Distinct Count for this Column is 4
    - it's distinct categories are ['Salaried' 'Free Lancer' 'Small Business' 'Large Business']
    - Salaried person are among highest Insurance holder with count of 2192, whereas Free lancer taken very less insurance of only 2 records in sample data
    - Again, on an average Bonus remain about same for all agents, who sold policies to any occupation of customer, though Bonus is highest for Salaries Customer.
    - Undouble Total bonus for Agents, who sold policy to Salaried Customers are maximum, because Salaried person are maximum in numbers as well.

3. Field name is EDUCATIONFIELD : and Distinct Count for this Column is 6
    - and it's distinct categories are ['Graduate' 'Post Graduate' 'Under Graduate' 'Engineer' 'Diploma' 'MBA']
    - there are maximum Customers are Graduate, whereas MBA holder are least.
4. Field name is GENDER : and Distinct Count for this Column is 2
    - and it's distinct categories are ['Female' 'Male']
    - Maximum Customers are Male
5. Field name is DESIGNATION : and Distinct Count for this Column is 5
    - and it's distinct categories are ['Manager' 'Executive' 'VP' 'AVP' 'Senior Manager']
    - Counts are each categories of this column are :
    - maximum customers are Executives and least customers are on VP post, this is obvious data.
6. Maximum Average bonus given to Agents, who sold policies to VP, it seems, VP takes most Sum insured, thats why Bonus is high for them.
7. Field name is MARITALSTATUS : and Distinct Count for this Column is 4
    - and it's distinct categories are ['Single' 'Divorced' 'Unmarried' 'Married']
    - Counts are each categories of this column are :
    - Married person baught maximum policies.
8. Field name is ZONE : and Distinct Count for this Column is 4
    - and it's distinct categories are ['North' 'West' 'East' 'South']
    - Counts are each categories of this column are :
    - maximum customers belong to West region of Country with the count of 2566, whereas there are only 6 customers from South region, this is interesting data to analyze, why we have very less customers from south. Also customers from east region are also very less with no of 64.
9. Field name is PAYMENTMETHOD : and Distinct Count for this Column is 4
    - and it's distinct categories are ['Half Yearly' 'Yearly' 'Quarterly' 'Monthly']
    - Counts are each categories of this column are :
    - Customers like to give EMI for policy maximumm is half yearly and then yearly.

3.8 Multivariate Analysis

Pair plot:

Figure 4 pair plot against AgentBonus per policy

insights:

1. there is clear separation of Bonus for Agents, as all variables increasing, Bonus also increasing.

Heat Map:

Figure 5 heat Map

Co-relation Matrix:

	Age	CustTenure	ExistingProdType	NumberOfPolicy	MonthlyIncome	ExistingPolicyTenure	AgentBonus_Per_Policy	SumAssu
Age	1.000000	0.359631	0.076609	0.053143	0.347942	0.263640	0.211933	
CustTenure	0.359631	1.000000	0.115561	0.066859	0.368522	0.347473	0.207415	
ExistingProdType	0.076609	0.115561	1.000000	0.149862	0.191194	0.009800	-0.085003	
NumberOfPolicy	0.053143	0.066859	0.149862	1.000000	0.128328	0.021839	-0.733911	
MonthlyIncome	0.347942	0.368522	0.191194	0.128328	1.000000	0.410641	0.163630	
ExistingPolicyTenure	0.263640	0.347473	0.009800	0.021839	0.410641	1.000000	0.188274	
AgentBonus_Per_Policy	0.211933	0.207415	-0.085003	-0.733911	0.163630	0.188274	1.000000	
SumAssured_Per_Policy	0.197145	0.193545	-0.072151	-0.683045	0.154269	0.185564	0.937796	

Insights:

1. AgentBonus_Per_Policy have very strong relationship with about all of the other fields
2. As Sum insured of the Customer increases, Bonus also increases.
3. Bonus increased with the experience of the Age of the Customer, Customer tenure, Monthly income of the Customer, as well as Existing policy tenure,
4. There is very minute positive relationship between Agent Bonus with Existing Prod type, Number of Policies, and Customer care Score.
5. CustCareScore, LastMonthCalls and Compaint dont have any relation with any other field, its almost 0 co-relation with every other field. So we have dropped them in initial steps

4. Business insights from EDA

4.1 Is the data unbalanced? If so, what can be done? Explain in Business context:

Since this requirement is for predicting continuous variable and its not a classification model, where we have separate classes for 1s and 0s, which are Targeted values, so we can't say that data is unbalanced.

Also, we have given data set with Agent Bonus for Total SumAssured for N number of policies purchases by any csutomers, so for building good prediction model, we have created 2 new fields , which are

AgentBonus_per_policy

SumAssured_per_Policy

And we have used NumberOfPolicy field for generating above fields.

Some of requirements of doing this are as follows:

- This is very important to know what is Bonus given for each Customers, irrespective of how many policies Customer purchased.
- We need to know, how is SumAssured affected AgenBonus, but if we not created new field SumAsuured_per_policy (This is actually a average of Totsl SumAssured / Number of Policy), Our model wont be able to understand, this relations.
- For an example , if Customer purchased 4 policy of total 1000 Rs, and AgentBonus is 400 Rs, which means Bonus is 1000/(400*4) = 6.25% per policy Bonus
  And for any other Customer purchased Single Policy of amount 1000 and his Agent get 250 Rs Bonus, which means, he received 1000/250 = 40% Bonus,
  so for making them all on same Scale, we need to create above listed both variable to know Average of AgentBonus per Policy as well as Average of SumAssured per policy.
- We have also checked for the co-relation between all the fields and targeted field, which is AgentBonus and we found that their co-relation coefficient was very low and we dropped those fields.
  Complaint : 0.025091
  LastMonthCalls: 0.038717
  CustCareScore: -0.005319

  we have checked that following fields have very minute impact on targeted fields and these can be dropped as well:

4.2 Any business insights using clustering (if applicable)

Though this prediction Model don't need any Clustering methodology, but we can build Clusters based on "**Number of Policy"** purchased by any Customers and build separate model for each data set.

And counts in each Clusters would be like this :

```
NumberOfPolicy Counts
1.0            438
2.0            711
3.0            939
4.0            1139
5.0            856
6.0            437
```

Clustering can be also performed using K-mean Clusters, and we can have N number of Clusters, needed. But for this model we don't need this.

Though based on Bivariate analysis, if we are supposed to build any Clusters based on our categorical fiels, we can build Clusters against these fields:

```
Based on Columns: Channel
Cluster names would be : ['Agent', 'Third Party Partner', 'Online']

Based on Columns: Occupation
Cluster names would be :
['Salaried', 'Free Lancer', 'Small Business', 'Large Business']


Based on Columns : EducationField
Cluster names would be :
['Graduate', 'Post Graduate', 'Under Graduate', 'Engineer', 'Diploma', 'MBA']

Based on Columns: Gender
Cluster names would be : ['Female', 'Male']

Based on Columns: Designation
Cluster names would be :  ['Manager', 'Executive', 'VP', 'AVP', 'Senior Manager']

Based on Columns: MaritalStatus
Cluster names would be : ['Single', 'Divorced', 'Unmarried', 'Married']

Based on Columns: Zone
Cluster names would be :  ['North', 'West', 'East', 'South']

Based on Columns: PaymentMethod
Cluster names would be :  ['Half Yearly', 'Yearly', 'Quarterly', 'Monthly']
```

4.3 Any other business insights

We performed following actions on our data :

1. Created 2 new fields AgentBonus_per_policy and SumAssured_per_policy, which is basically Average AgentBonus and Average SumAssured
2. We dropped fields CustID, AgentBonus and SumAssured for building our model.
3. Model building will be done in future submissions.
4. We have imputed NULL values for each of our missing field values.
5. I have analyzed data for Univariate analysis, checked individual column's distribution, their counts, in it's category, mean and Sum of values. Checked outliers, distribution and checked normality of data.
6. I have also done Co-relation check, data skewness and checked distribution of all the fields against target variable.

Based on above analysis , following are business insights:

1. We can see clear co-relation between SumAssured_per_policy and AgentBonus_per_policy field with co-relation of 0.94 , which means, as Sumassured increases, AgentBonus also increases.
2. AgentBonus also have good positive co-relation with field Age, CustomerTenure and existingPolicyTenure of with Positive 0.2 with each field mentioned.
3. AgentBonus have very minute relation with fields Complaint and LastMonthCalls, which makes sense.
4. There are very less policies sold in South region, and we have only 6 customers from South region, insurance company needs to give good campaign in that region, and there is good amount of chance to spread into different geolocation. And spread their legs in that.
5. Also, Company should run motivational trainings, good sales skill development program to their South region Agents.