Life Insurance sales – capstone

Project Notes – 1

SARATH KUMAR V

Table of Contents

1. **Problem Statement: Life Insurance Data** 3
2. **Need for this Study/Project** 3
3. **Why is this (agent bonus) important for the business/company?** 3
4. **Data Report/Dictionary** 4
5. **Performing Exploratory Data Analysis (EDA)** 4
6. **Checking for Unique Categorical Values** 7
7. **Univariate/Bivariate Analysis** 8
8. **Categorical Variable's Univariate Analysis** 13
9. **Bivariate Analysis (Pairplot/Heatmap)** 18

Table of Figures

Figure 1 – Univariate Analysis – Numerical Columns 8
Figure 2 - Univariate Analysis – Categorical Columns 13
Figure 3 - Categorical variable w.r.t Target Variable (Bivariate) 15
Figure 4 - Bivariate Analysis- Numerical Data Pairplot 18
Figure 5 - Bivariate Analysis- Numerical Data Correlation Heatmap 19

Problem Statement: Life Insurance Data

- 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.

Need for this Study/Project

- With this problem we want to better understand how the insurance company agents are performing, it's not to underpay or overpay, as the payment is regulated by IRDA.
- With the predictions it's better for the company to understand where they need to focus more as for agents selling less policies the company needs some booster training performs. As the policies are as good as the agents portray it to be to the potential customer.
- While the agents performing good i.e. selling more policies there needs to be a way to reward them, to make their contribution known so that they perform the same and even better in future.

Why is this (agent bonus) important for the business/company?

- A company is as good as their employers.
- For a Life Insurance Company, their agents are the best way to make the companies policies, aims, and perks known to the customer. Once the customer is intrigued by the policy delivery by the agent, its easier to convince the customer hence improving the sales and thereby motivating the agent as well.
- With this, the market share of the company will gain more ground dominating the potential opponents.
- Moreover, the agents can be classified into categories giving the company better insight where the need to put more effort.
- The customer feedback can help the company develop improved and updated policies/products. Meeting customer needs.
- Hereby, the easiest way to retain their agents.
- Overall, multiplying and adding to company's profit.

Data Report/Dictionary

The following data is provided by Great Learning cover the Life Insurance Sales made by the company, the data dictionary consists of:

Variable	Description
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
MaritialStatus	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

Performing Exploratory Data Analysis (EDA).

Head of the Data

	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome
0	4409	22.0	4.0	Agent	Salaried	Graduate	Female	3	Manager	2.0	Single	20993.0
1	2214	11.0	2.0	Third Party Partner	Salaried	Graduate	Male	4	Manager	4.0	Divorced	20130.0
2	4273	26.0	4.0	Agent	Free Lancer	Post Graduate	Male	4	Exe	3.0	Unmarried	17090.0
3	1791	11.0	NaN	Third Party Partner	Salaried	Graduate	Fe male	3	Executive	3.0	Divorced	17909.0
4	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced	18468.0

- I've removed CustID as it is irrelevant to agent bonus.
- Head gives us the idea of what the basic dataset looks like.
- Complete list of all variables is not presented.

Shape of the dataset

Total rows in the dataset: 4520
Total columns in the dataset: 19

Descriptive Statistics of the Columns

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
AgentBonus	4520	NaN	NaN	NaN	4077.84	1403.32	1605	3027.75	3911.5	4867.25	9608
Age	4251	NaN	NaN	NaN	14.4947	9.03763	2	7	13	20	58
CustTenure	4294	NaN	NaN	NaN	14.469	8.96367	2	7	13	20	57
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	NaN	NaN	NaN	3.68894	1.01577	1	3	4	4	6
Designation	4520	6	Manager	1620	NaN	NaN	NaN	NaN	NaN	NaN	NaN
NumberOfPolicy	4475	NaN	NaN	NaN	3.56536	1.45593	1	2	4	5	6
MaritalStatus	4520	4	Married	2268	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthlyIncome	4284	NaN	NaN	NaN	22890.3	4885.6	16009	19683.5	21606	24725	38456
Complaint	4520	NaN	NaN	NaN	0.287168	0.452491	0	0	0	1	1
ExistingPolicyTenure	4336	NaN	NaN	NaN	4.13007	3.34639	1	2	3	6	25
SumAssured	4366	NaN	NaN	NaN	620000	246235	168536	439443	578976	758236	1.8385e+06
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	NaN	NaN	NaN	4.62699	3.62013	0	2	3	8	18
CustCareScore	4468	NaN	NaN	NaN	3.06759	1.38297	1	2	3	4	5

- The table includes the complete description for all variable with categorical variables included.
- The description includes, variable count, unique values, top frequently occurring categories like Agent-3194, mean, standard deviation, minimum, 25%, 50%(median), 75%, and maximum values present in the respective variables.
- Hence the 'NaN' here is observed for Categorical Variables as a string object cannot have numeric values.
- This we will change by encoding the data in future if needed.
- We can also observe the missing values as the count is not constant for all the variables.
- The unique is only present for categorical variables which hold a specific category
- Example: Gender has male and female hence it should hold unique value of 2 but later we observed some subcategories needs to be renamed.

Info of the parameters

```
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   AgentBonus            4520 non-null   int64
 1   Age                   4251 non-null   float64
 2   CustTenure            4294 non-null   float64
 3   Channel               4520 non-null   object
 4   Occupation            4520 non-null   object
 5   EducationField        4520 non-null   object
 6   Gender                4520 non-null   object
 7   ExistingProdType      4520 non-null   int64
 8   Designation           4520 non-null   object
 9   NumberOfPolicy        4475 non-null   float64
 10  MaritalStatus         4520 non-null   object
 11  MonthlyIncome         4284 non-null   float64
 12  Complaint             4520 non-null   int64
 13  ExistingPolicyTenure  4336 non-null   float64
 14  SumAssured            4366 non-null   float64
 15  Zone                  4520 non-null   object
 16  PaymentMethod         4520 non-null   object
 17  LastMonthCalls        4520 non-null   int64
 18  CustCareScore         4468 non-null   float64
dtypes: float64(7), int64(4), object(8)
memory usage: 671.1+ KB
```

- We have 7 parameters having 'float' data type.
- We have 4 parameters having 'integer' data type.
- We have 8 parameters having 'object' data type.
- Age is shown as float, however we will later observe is its needed to change it to int or not, it won't make any difference in our observations.
- We can clearly observe some missing values.
- Further count of missing values is provided below.

- 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

- **Number of duplicate rows = 0**
- The **Missing values** can affect the prediction's hence need to be treated, hence the missing values are imputed with the **median values** in the respective column.

Checking for Unique Categorical Values.

CHANNEL has 3 Unique Values.
Online 468
Third Party Partner 858
Agent 3194
Name: Channel, dtype: int64

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

EDUCATIONFIELD has 7 Unique Values.
MBA 74
UG 230
Post Graduate 252
Engineer 408
Diploma 496
Under Graduate 1190
Graduate 1870
Name: EducationField, dtype: int64

GENDER has 3 Unique Values.
Fe male 325
Female 1507
Male 2688
Name: Gender, dtype: int64

DESIGNATION has 6 Unique Values.
Exe 127
VP 226
AVP 336
Senior Manager 676
Executive 1535
Manager 1620
Name: Designation, dtype: int64

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

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

PAYMENTMETHOD has 4 Unique Values.
Quarterly 76
Monthly 354
Yearly 1434
Half Yearly 2656

- Here it can be observed that subcategories highlighted with a different colour shows an error in naming convention hence have to be renamed.
- Example: 'Laarge' and 'Large' Business can be put in the same category, the same for 'UG' and 'Under Graduate', 'Graduate' and 'Post Graduate', 'Fe male' and 'Female', and 'Exe' and 'Executive'.

Univariate/Bivariate Analysis

AgentBonus

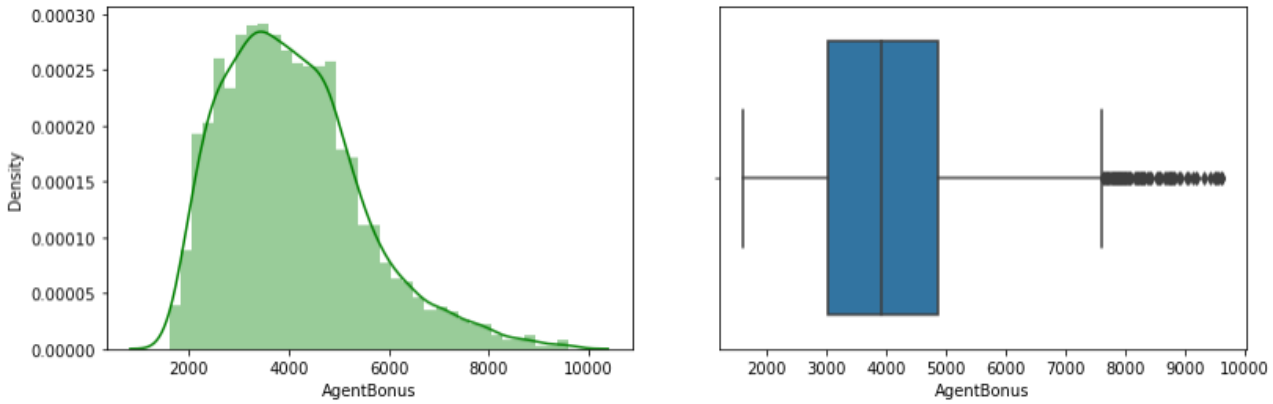

Figure 1(a) Distplot/Histplot - AgentBonus

- **The distribution of "AgentBonus" seems to be positively/right skewed.**
- **The data ranges from 1605 to 9600.**
- **The box plot holds many outliers.**

Age:

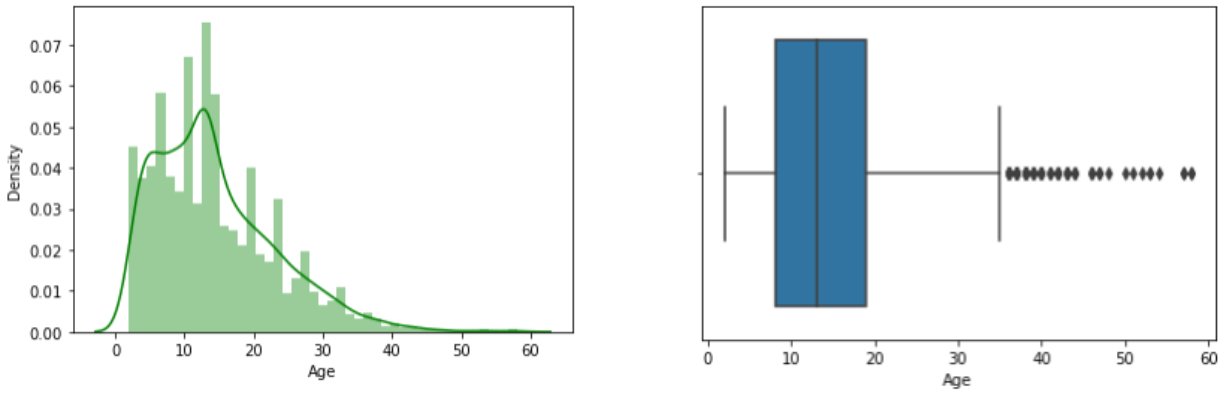

Figure 1(b) Distplot/Histplot - Age

- **The distribution of "Age" seems to be positively/right skewed.**
- **The data ranges from 2 to 58.**
- **The box plot holds many outliers.**

CustTenure:

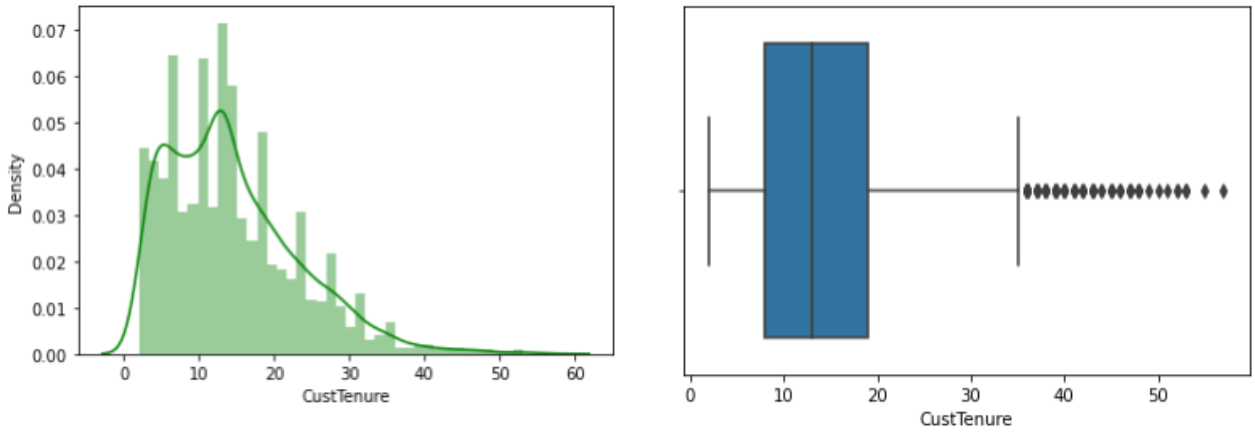

Figure 1(c) Distplot/Histplot - CustTenure

- **The distribution of "CustTenure" seems to be positively/right skewed.**
- **The data ranges from 2 to 57.**
- **The box plot holds many outliers.**

ExistingProdType:

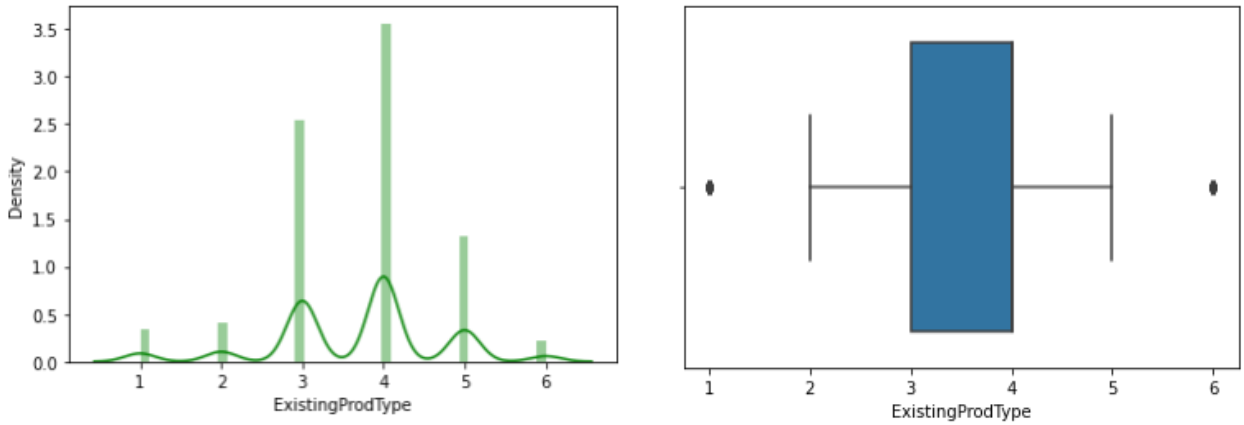

Figure 1(d) Distplot/Histplot - ExistingProdType

- **The distribution of "ExistingProdType" seems to be slightly left skewed.**
- **The data ranges from 1 to 6.**
- **The box plot holds outliers.**

NumberOfPolicy

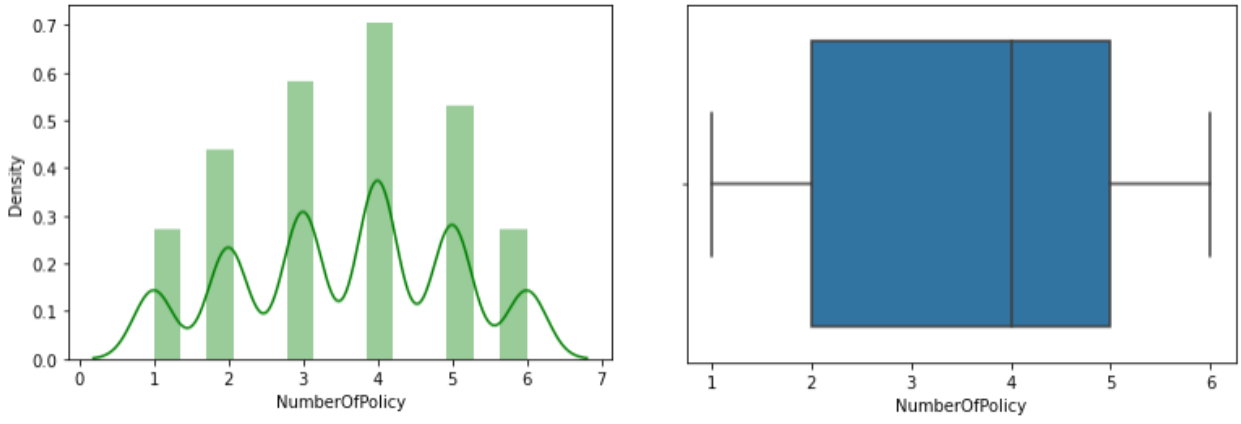

Figure 1(e) Distplot/Histplot - NumberofPolicy

- **The distribution of "NumberOfPolicy" seems to be slightly left skewed.**
- **The data ranges from 1 to 6.**
- **The box plot has no outliers.**

MonthlyIncome

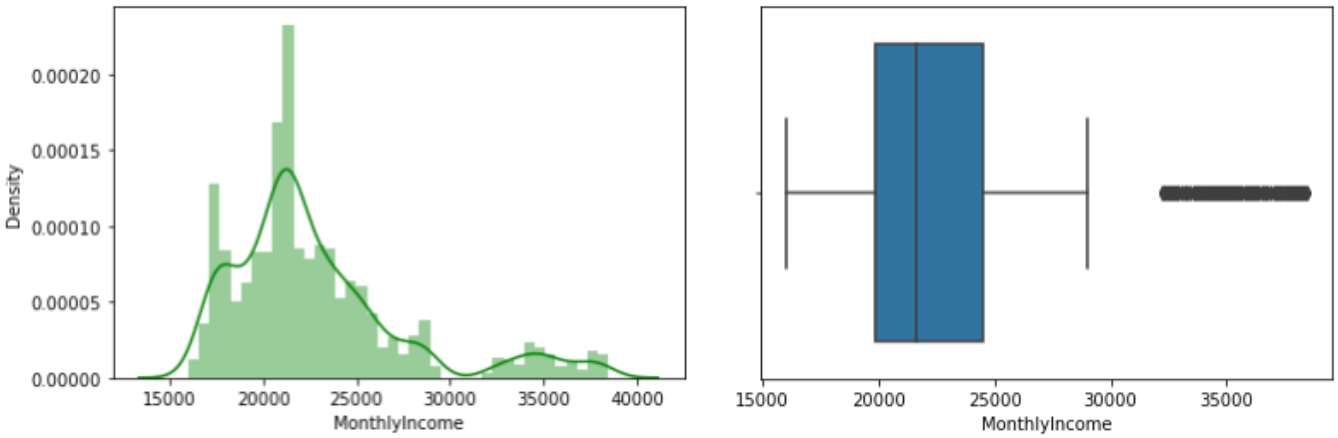

Figure 1(f) Distplot/Histplot - MonthlyIncome

- **The distribution of "MonthlyIncome" seems to be positively/right skewed.**
- **The data ranges from 16000 to 38500.**
- **The box plot holds many outliers.**

Complaint

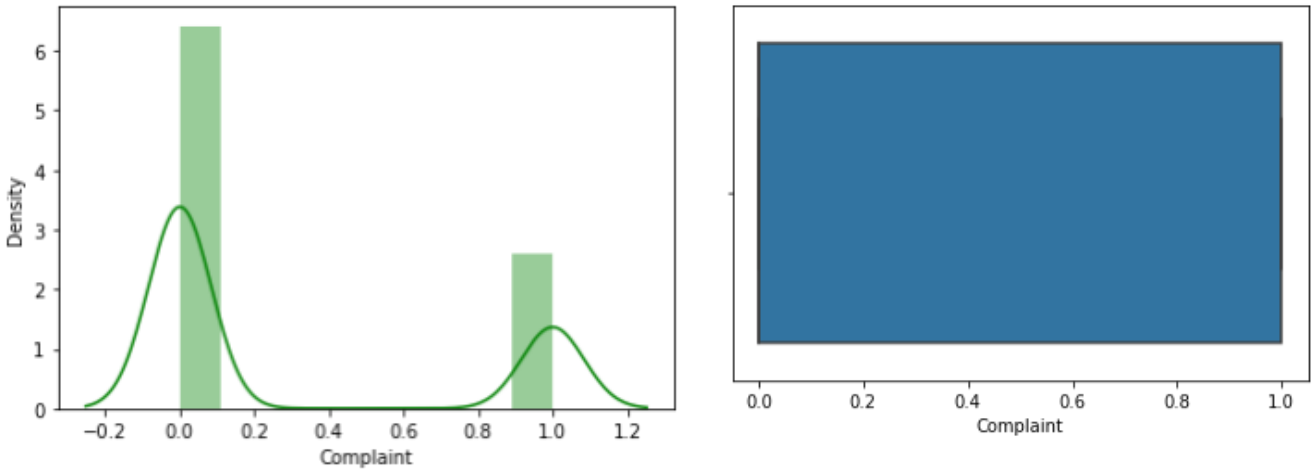

Figure 1(g) Distplot/Histplot - Complaint

- **The distribution of "Complaint" seems to be positively/right skewed.**
- **The data ranges from 0 to 1.**
- **The box plot holds no outliers.**

ExistingPolicyTenure

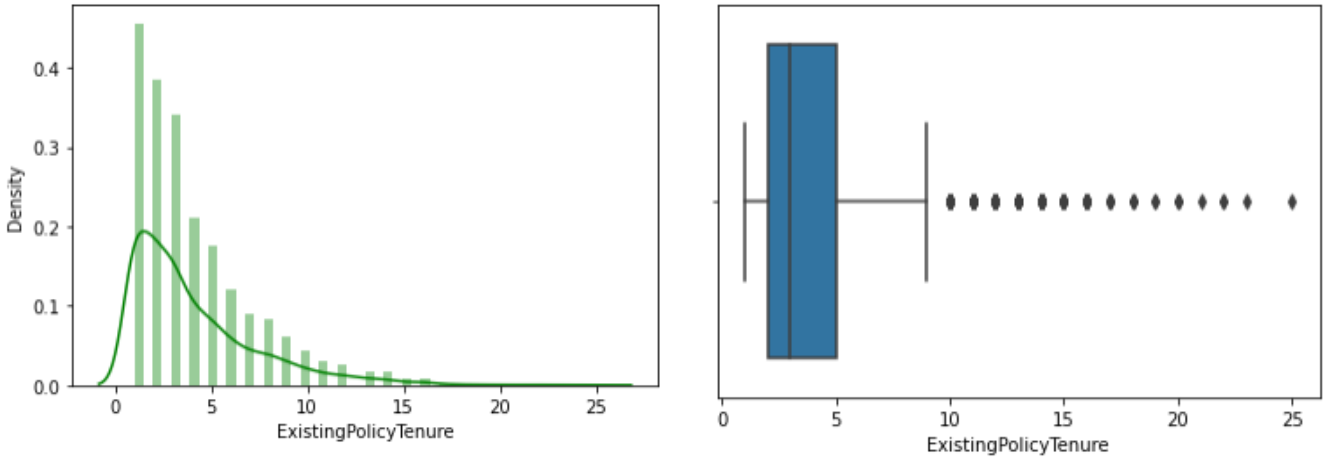

Figure 1(h) Distplot/Histplot - ExistingPolicyTenure

- **The distribution of "ExistingPolicyTenure" seems to be positively/right skewed.**
- **The data ranges from 1 to 25.**
- **The box plot holds many outliers.**

SumAssured

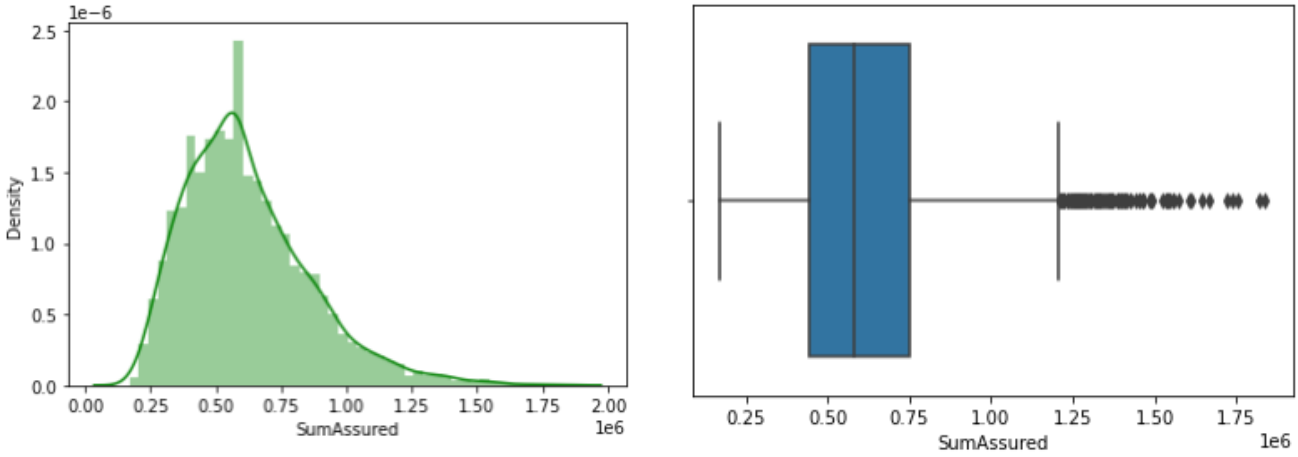

Figure 1(i) Distplot/Histplot - SumAssured

- **The distribution of "SumAssured" seems to be positively/right skewed.**
- **The data ranges from 1.68 * 10^5 to 1.83 * 10^5.**
- **The box plot holds many outliers.**

LastMonthCalls

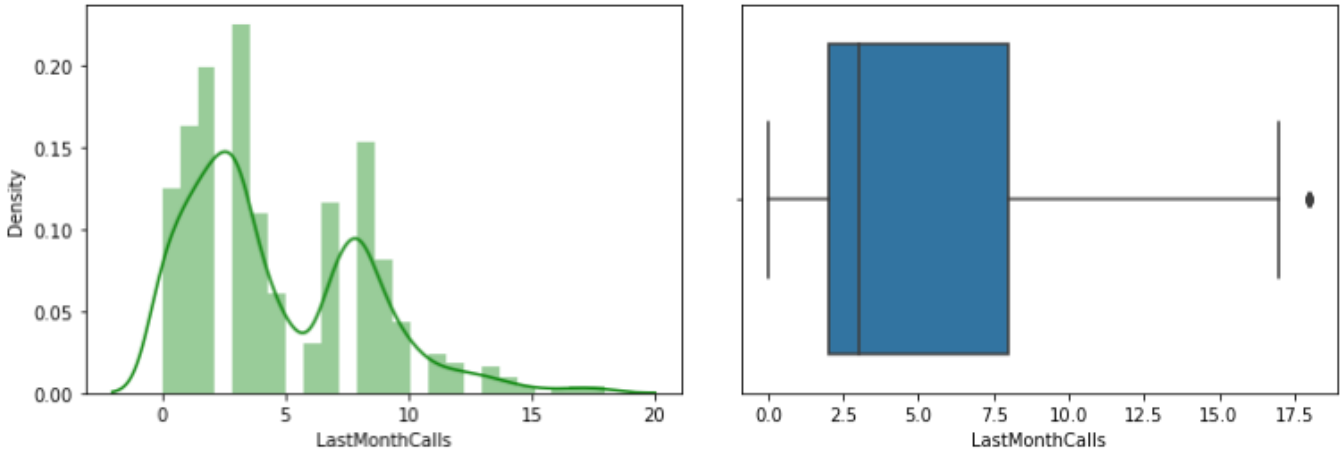

Figure 1(j) Distplot/Histplot - LastMonthCalls

- **The distribution of "LastMonthCalls" seems to be positively/right skewed.**
- **The data ranges from 0 to 18.**
- **The box plot holds outliers.**

CustCareScore

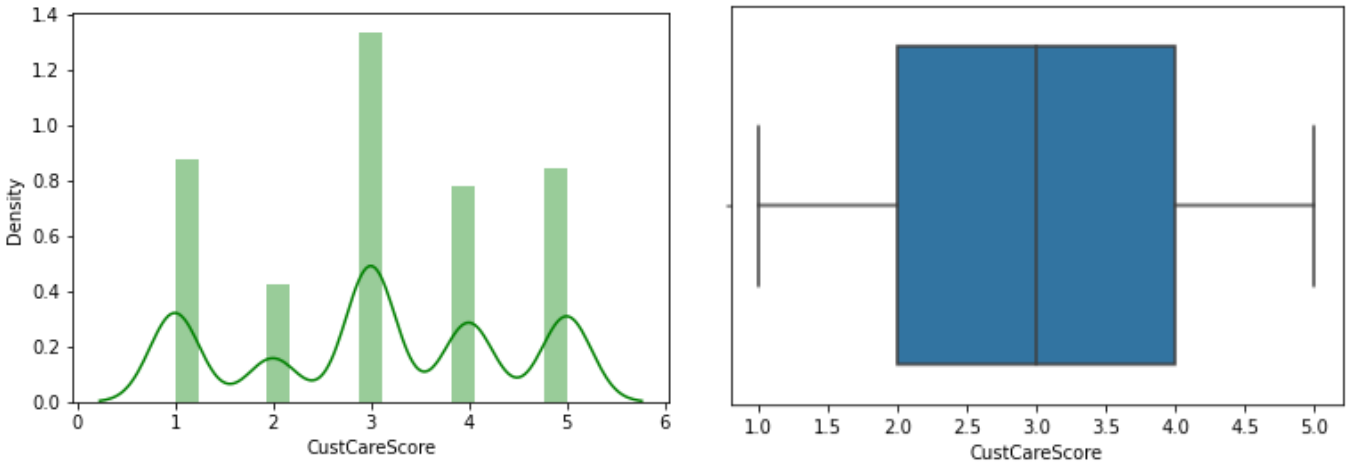

Figure 1(k) Distplot/Histplot – CustCareScore

- **The distribution of "CustCareScore" seems to be slightly left skewed.**
- **The data ranges from 1 to 5.**
- **The box plot holds no outliers**

Skewness

AgentBonus	0.822348
Age	0.998425
CustTenure	0.981002
ExistingProdType	-0.401100
NumberOfPolicy	-0.108161
MonthlyIncome	1.434315
Complaint	0.941129
ExistingPolicyTenure	1.601730
SumAssured	1.002018
LastMonthCalls	0.810417
CustCareScore	-0.138120

- We can observe skewness in the data with ExistingProdType, NumberofPoilicy and CustCareScore being negatively skewed.
- Rest all other parameters holds positive skewness the max being for ExistingPolicyTenure.

Categorical Variable's Univariate Analysis

Education Field

Post Graduate 0.47
Under Graduate 0.31
Diploma 0.11
Engineer 0.09
MBA 0.02

Most Customers approached are Post Graduates having 47% weightage.

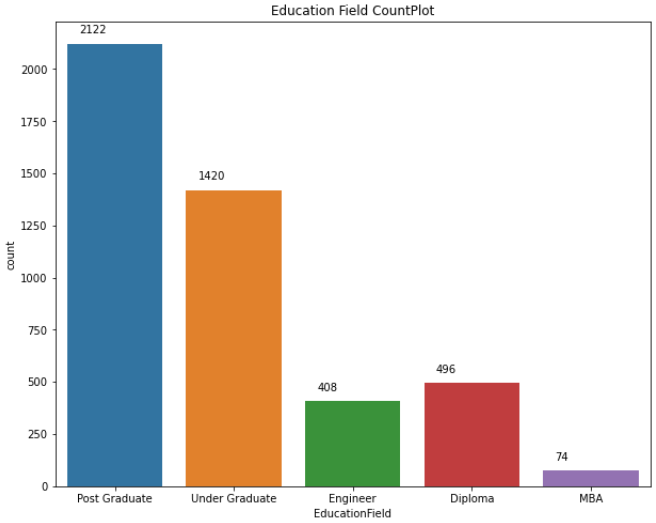

Figure 2(a) Count Plot - EducationField

Channel

Agent 0.71
Third Party Partner 0.19
Online 0.10

Acquisition of a customer is mostly done Via an Agent having 71% weightage.

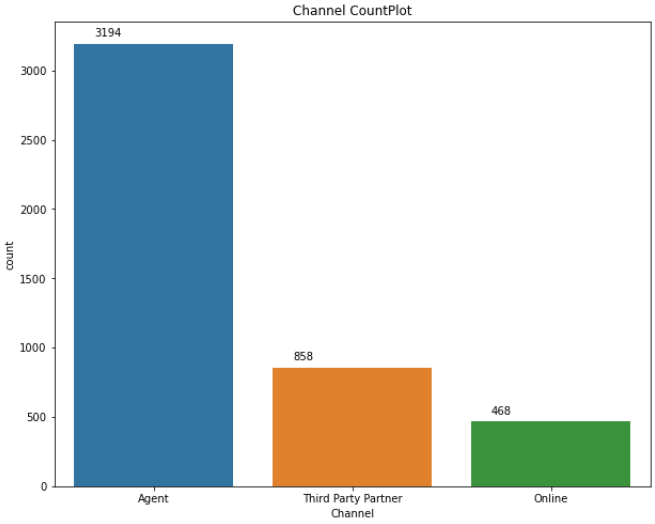

Figure 2(b) Count Plot - Channel

Occupation

Salaried 0.48
Small Business 0.42
Large Business 0.09
Free Lancer 0.00

Most customers have Salaried Occupations Around 48%.
Here freelancers have a minute weightage.

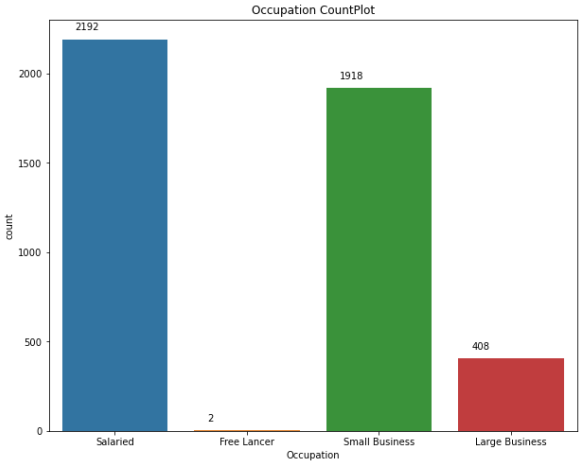

Figure 2(c) Count Plot - Occupation

Gender

Male 0.59
Female 0.41

Approximately 59% of customers Are males.

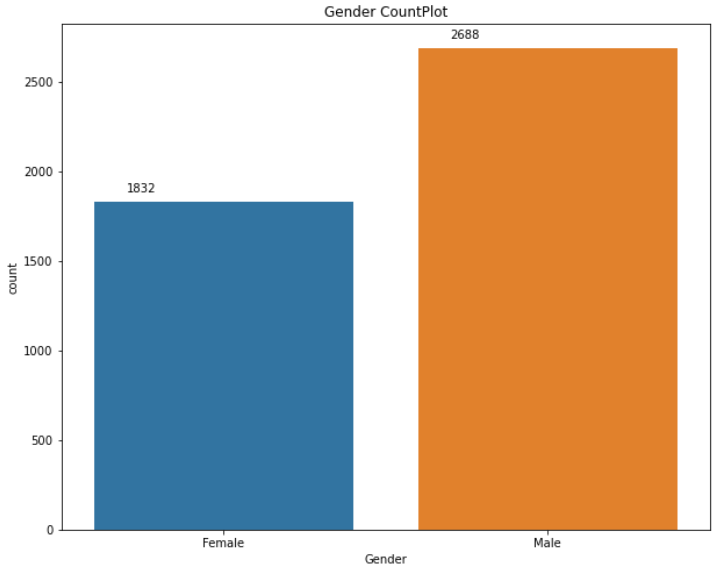

Figure 2(d) Count Plot - Gender

Designation

Executive 0.37
Manager 0.36
Senior Manager 0.15
AVP 0.07
VP 0.05

Most customers are either a Executive or Managers having Weightage of 37% and 36% Respectively.

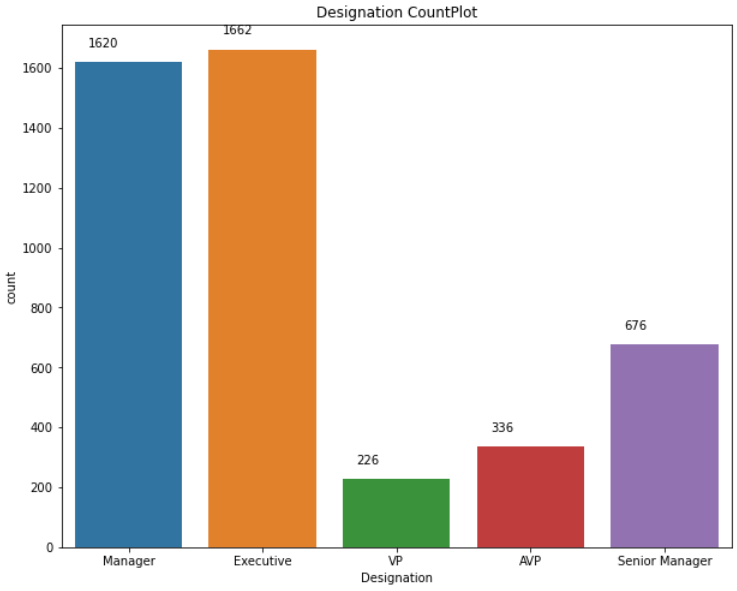

Figure 2(e) Count Plot - Designation

Marital Status

Married 0.50
Single 0.28
Divorced 0.18
Unmarried 0.04

Around 50% of the customers Are married.

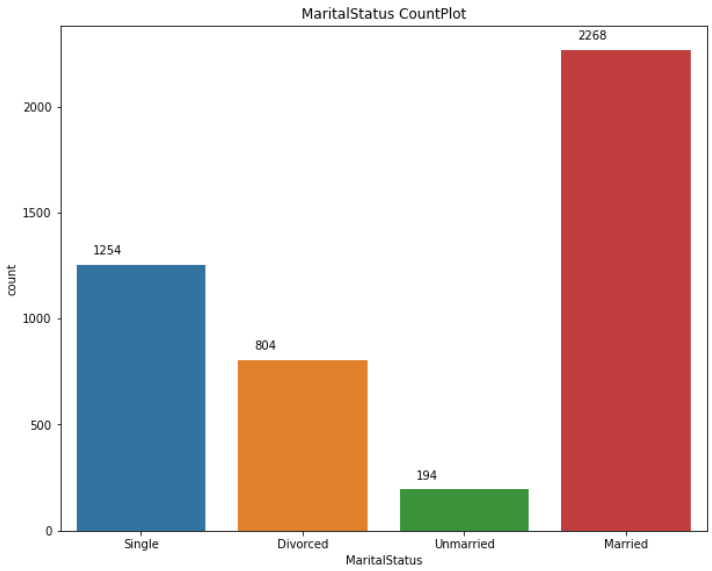

Figure 2(f) Count Plot -Marital Status

Zone

West 0.57
North 0.42
East 0.01
South 0.00

West Zone brings the most Customers with 57% weightage. Here freelancers have a minute weightage.

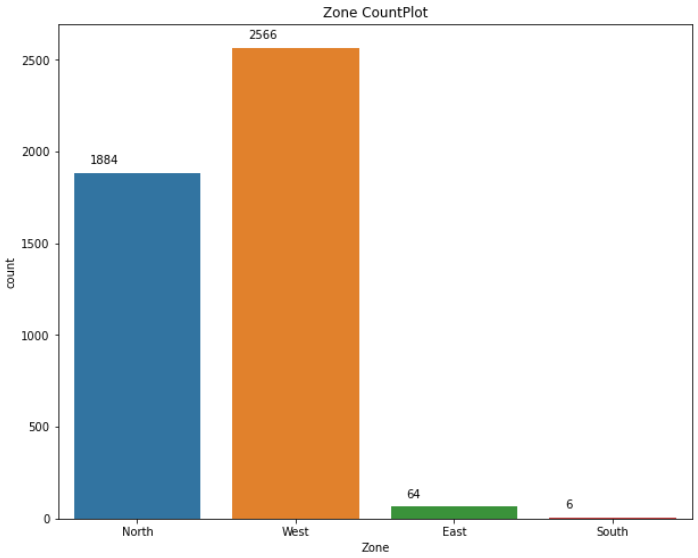

Figure 2(g) Count Plot - Zone

PaymentMethod

Half Yearly 0.59
Yearly 0.32
Monthly 0.08
Quarterly 0.02

Around 59% of Customers went For half-yearly payment plan

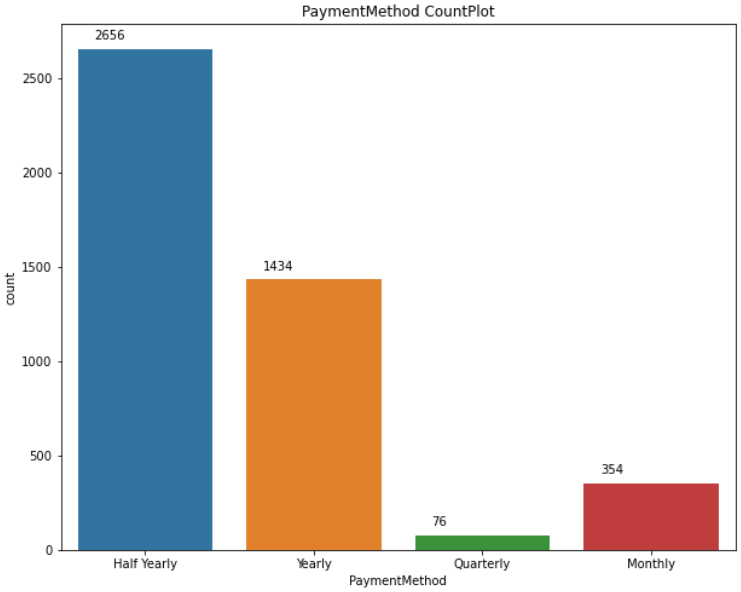

Figure 2(h) Count Plot - PaymentMethod

Categorical Variables Bivariate Analysis w.r.t Agent Bonus

- Agent Bonus has a lot of outlier values for every channel with almost similar mean values for all 3 channels.

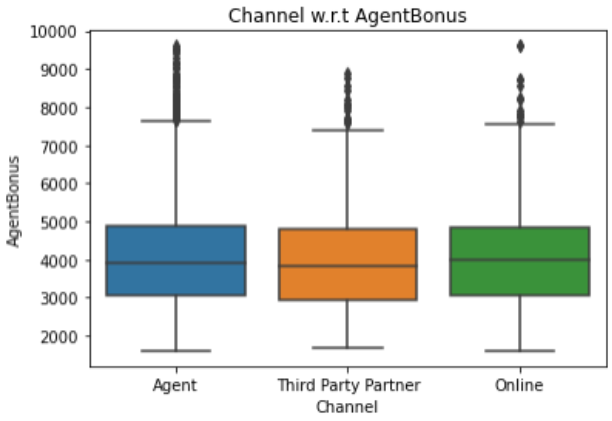

Figure 3(a) Boxplot – Channel w.r.t AgentBonus

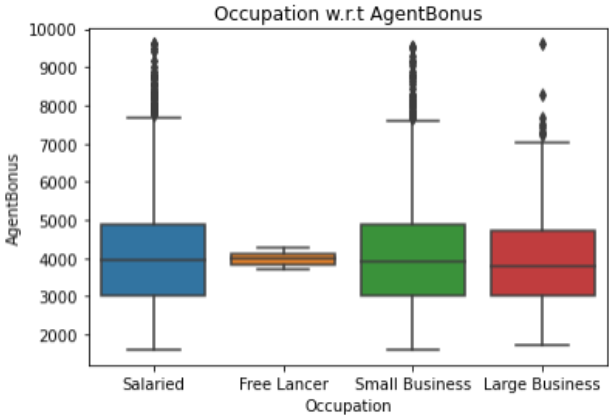

Figure 3(b) Boxplot – Occuapation w.r.t AgentBonus

Almost similar mean value for all Occupations.

NO outliers present for Free Lancer
Could be because we have only 2 data points for Free Lancer.

- Agent Bonus has a lot of outlier values
  for both Genders with almost similar mean values for b
  oth Male and Female.

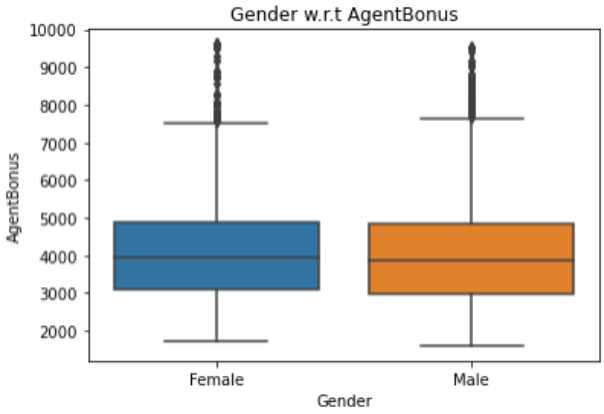

Figure 3(c) Boxplot – Genderl w.r.t AgentBonus

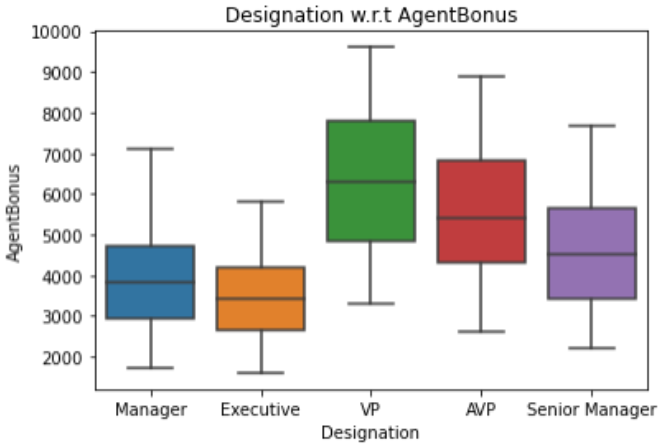

Figure 3(d) Boxplot – Desgnationl w.r.t AgentBonus

No outliers present.

VP Designation has the highest mean
As compared to other Designations.

- Agent Bonus has a lot of outlier values for all MaritalStatus except Unmarried customers.
- With almost similar mean values for all 3 customers except unmarried.

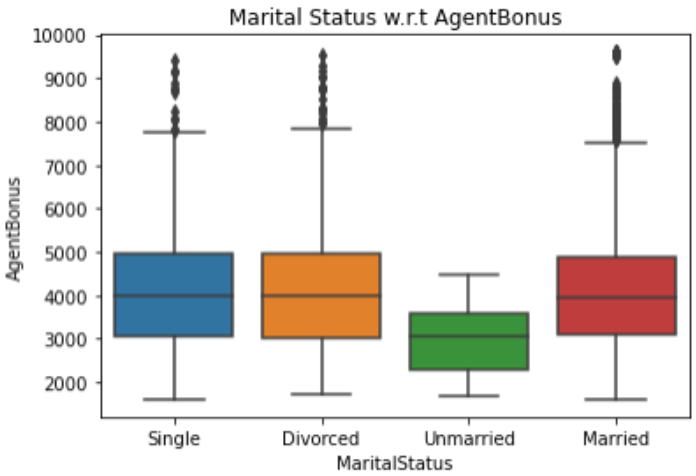

Figure 3(e) Boxplot – MaritalStatus w.r.t AgentBonus

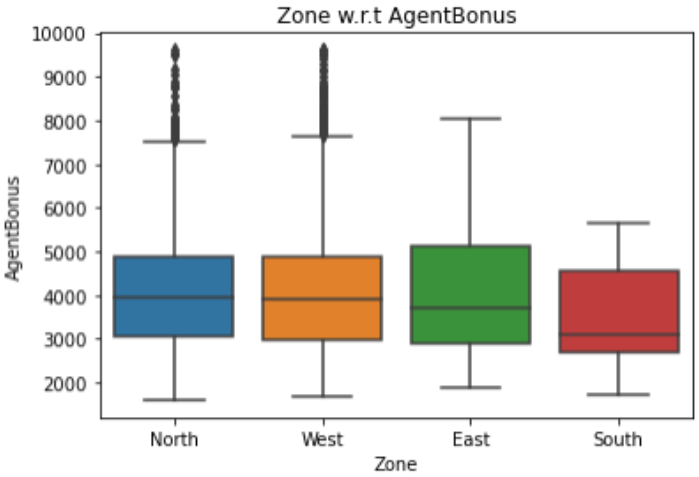

Figure 3(f) Boxplot – Zone w.r.t AgentBonus

Outliers present only for North and West Zones. Both having almost
Similar means.

No outliers present in East and South Zones possibly due to less Customer traffic from those Zones.

- Outliers present for all Payment methods chosen by the customer.
- Quarterly paying customers having the low est mean.

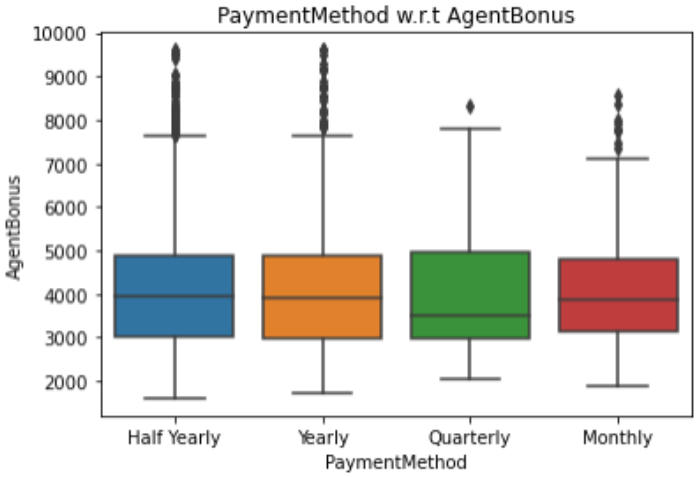

Figure 3(g) Boxplot – Channel w.r.t AgentBonus

Pairplot

A pair plot plots the relationships between all numeric variables in a dataset. The diagonal below is the histogram for each variable and shows the distribution. From the below plot, we can observe if there are relationships between every two pair of variables.

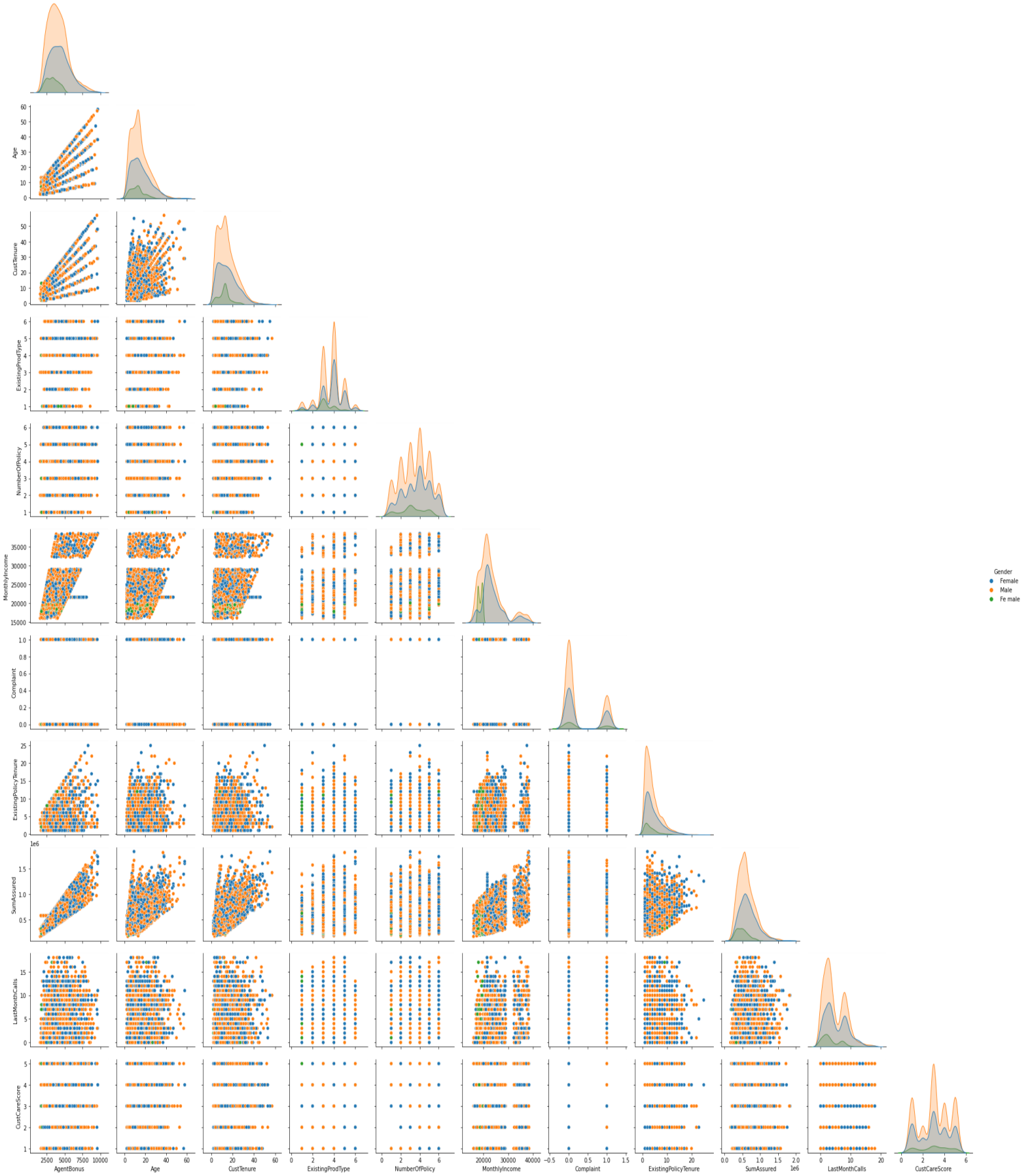

Figure 4 – Pairwise Distribution Plot

Correlation Heatmap.

The correlation coefficient shown in the table below shows the degree of correlation between the two variables represented in X axis and Y axis. It varies between -1 (maximum negative correlation) to +1 (maximum positive correlation).

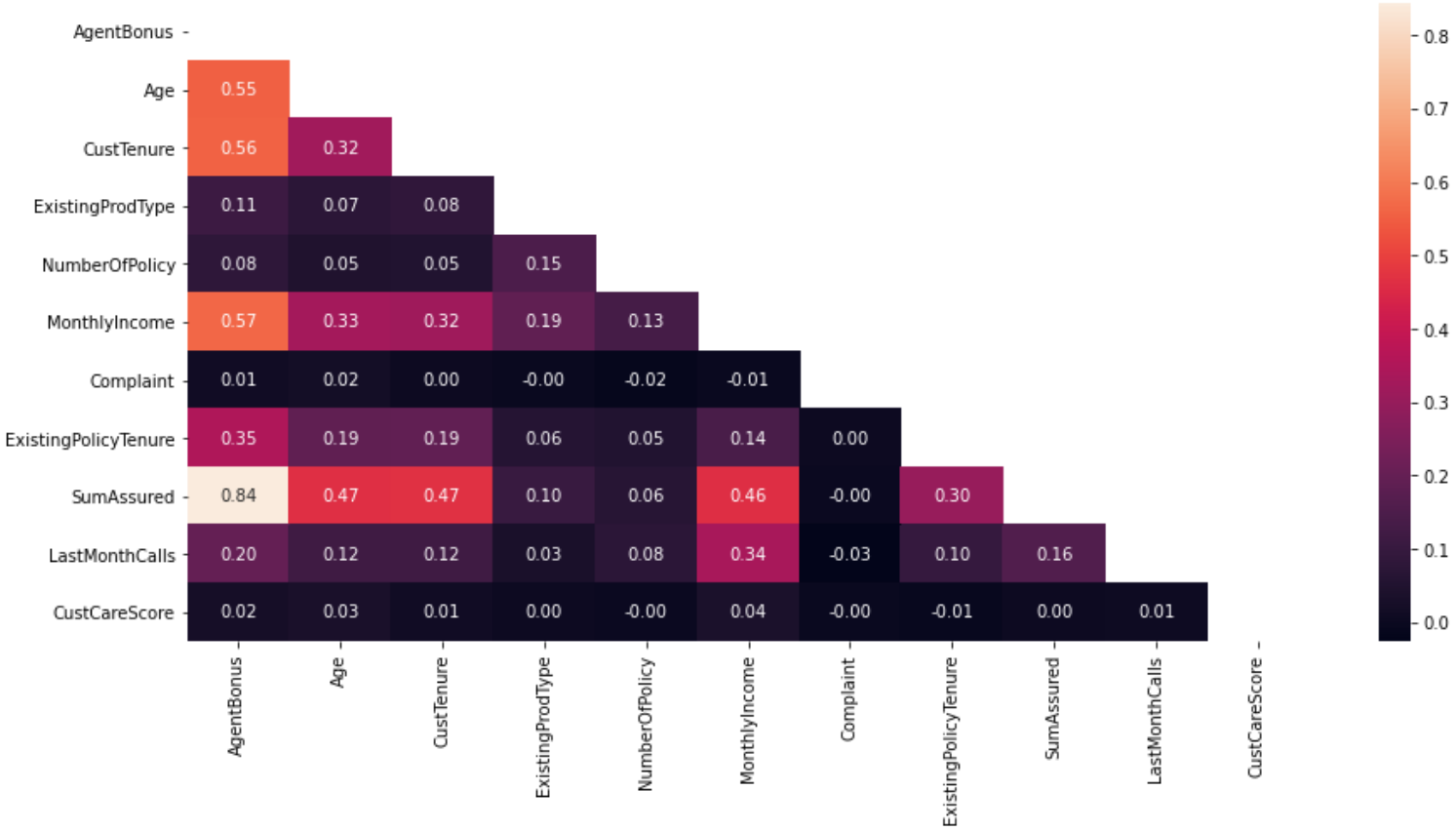

Figure 5 - Correlation Heatmap

- We can observe that there is almost no multicollinearity in the data.
- Complaint and CustCareScore have almost no correlation with any other parameter, hence dropping these columns will not make a difference.
- AgentBonus and SumAssured have high correlation with each other of 0.84.
- Here the lighter colors depict high correlation and darker colors depict low correlation.

1. Outlier Removal is performed but it does not seem as the correct approach as some variables like
2. SumAssured are allowed to have some outliers however our model will be affected if outliers are not removed.
3. We can add new variables like Premium but adding new variables can affect the model, hence not recommended.
4. With this we've completed the EDA in the coming exercises we'll build the model as this is a Classific ation problem, Regression Techniques for model building will be our go-to approach.
5. The data is highly unbalanced eg: Zone, South has less weightage similar for Occupation- Freelancer, more data is needed or upscale the data.

Thank You!