Life-Insurance Sale
Capstone Project

JUN2023
Final Report

Kandan Arumugam
PGP-DSBA – JUN2022-B

Table of Contents

Table of Figures

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?

- 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, it's 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 agent 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	4520.0	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

- 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-4077, 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

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 19 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   AgentBonus            4520 non-null   int64
 1   Age                   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
Name: PaymentMethod, dtype: int64
```

- 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

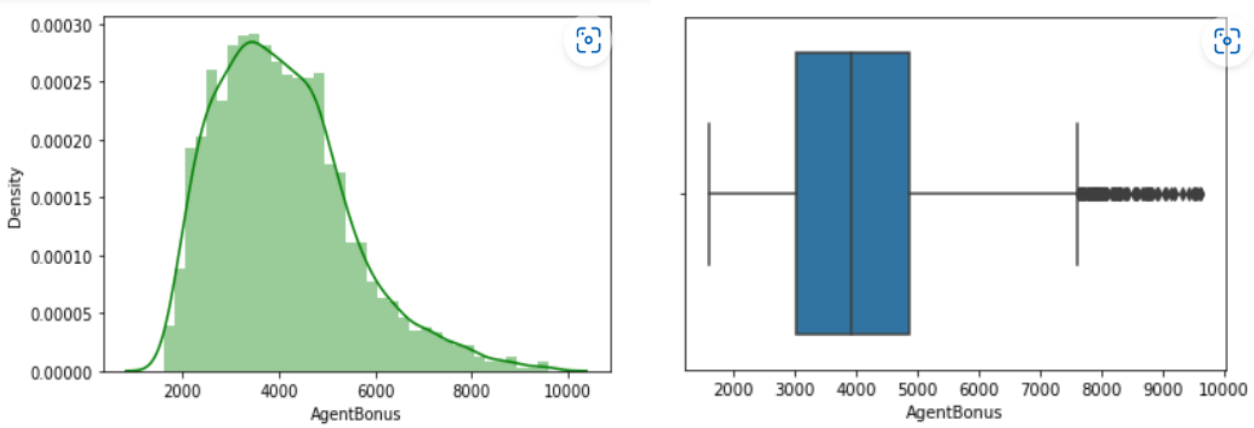

Figure 1(a) Distplot/Boxplot – AgentBonus

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

Age:

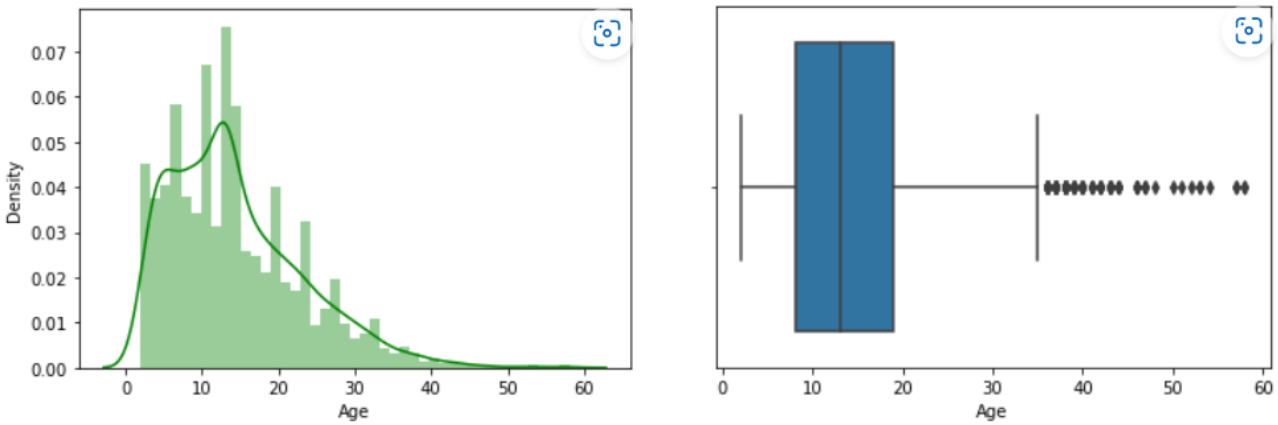

Figure 1(b) Distplot/Boxplot - Age

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

CustTenure:

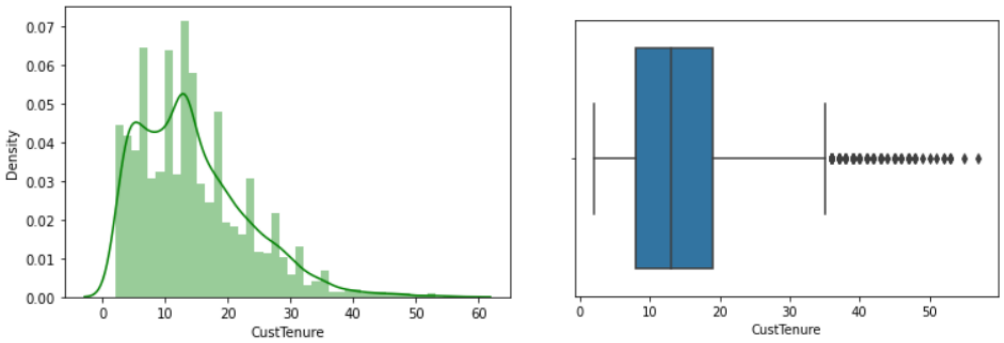

Figure 1(c) Distplot/Boxplot - CustTenure

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

ExistingProdType:

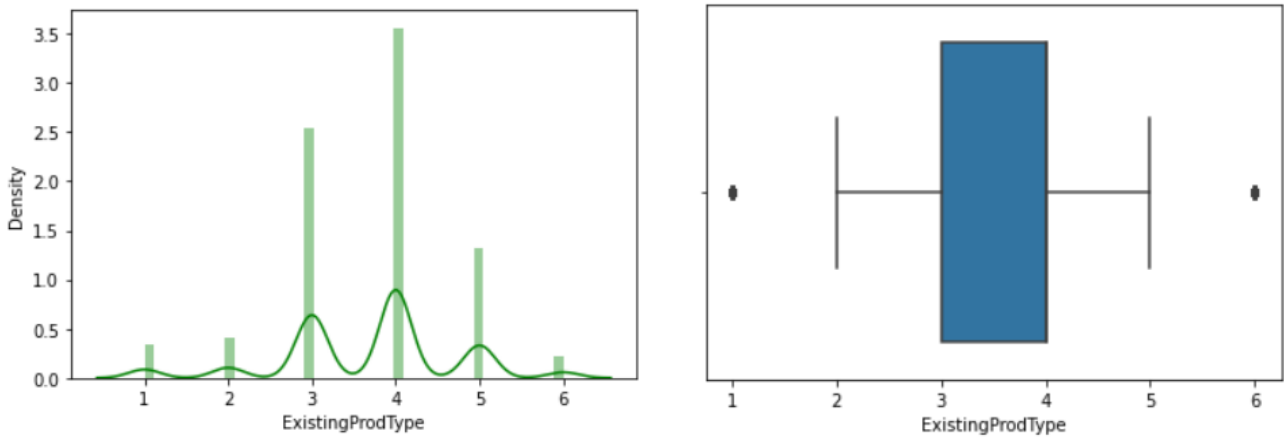

Figure 1(d) Distplot/Boxplot - ExistingProdType

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

NumberOfPolicy:

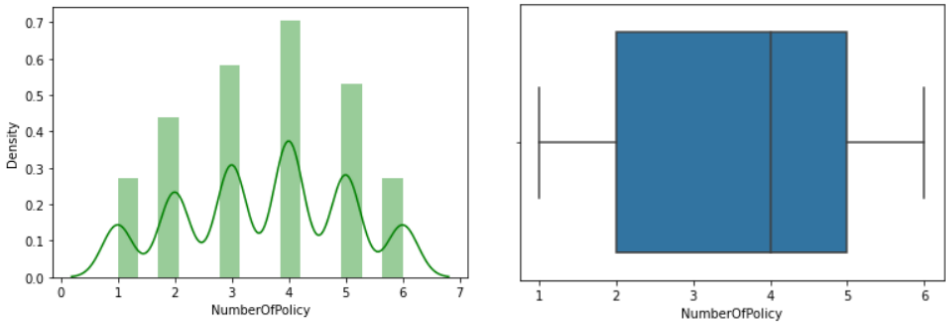

Figure 1(e) Distplot/Boxplot – NumberofPolicy

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

MonthlyIncome:

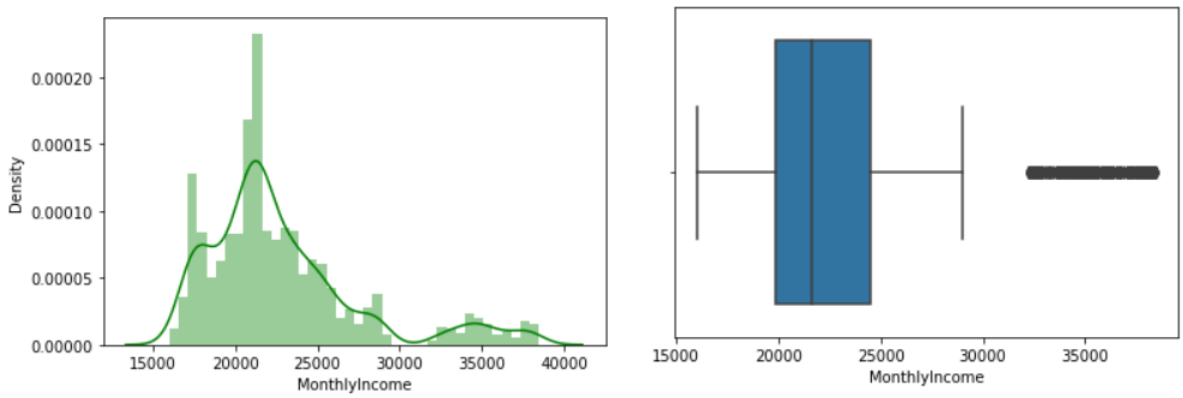

Figure 1(f) Distplot/Boxplot - MonthlyIncome

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

Complaint:

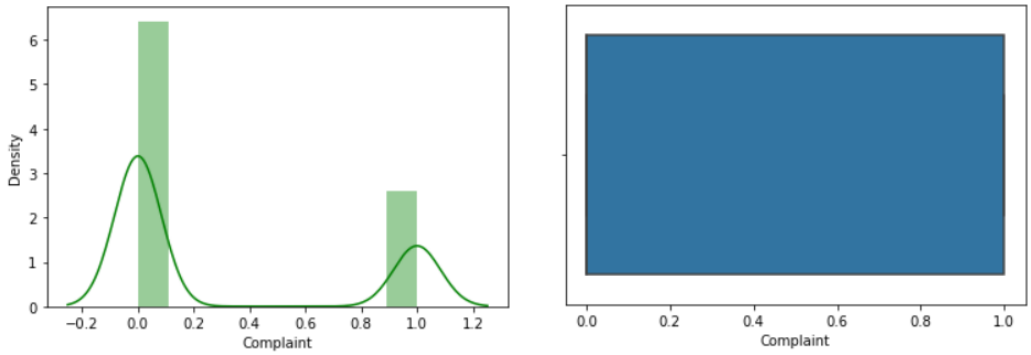

Figure 1(g) Distplot/Boxplot - Complaint

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

ExistingPolicyTenure

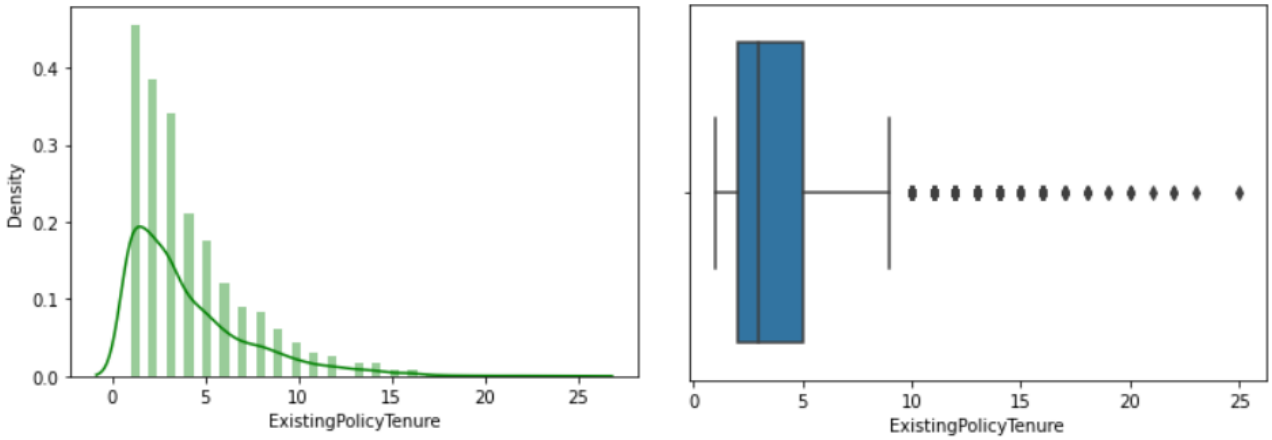

Figure 1(h) Distplot/Boxplot - ExistingPolicyTenure

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

SumAssured:

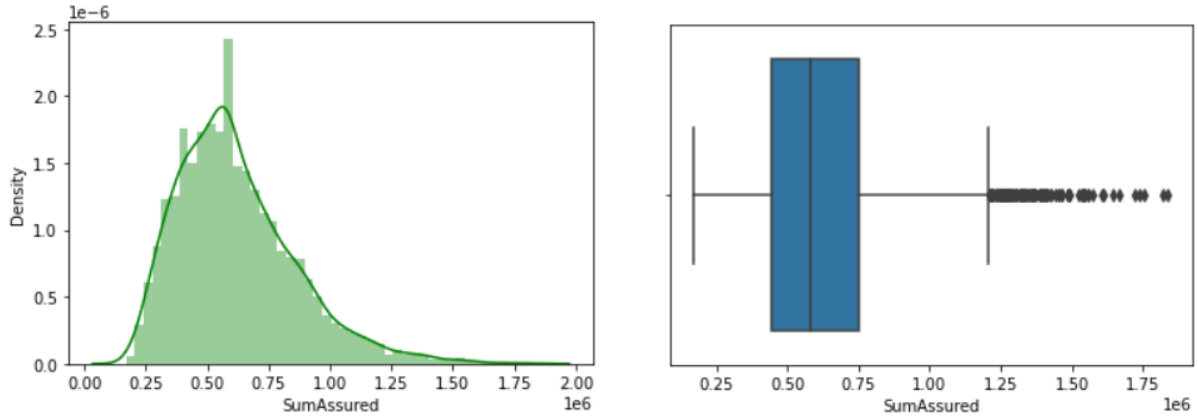

Figure 1(i) Distplot/Boxplot - SumAssured

- **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:

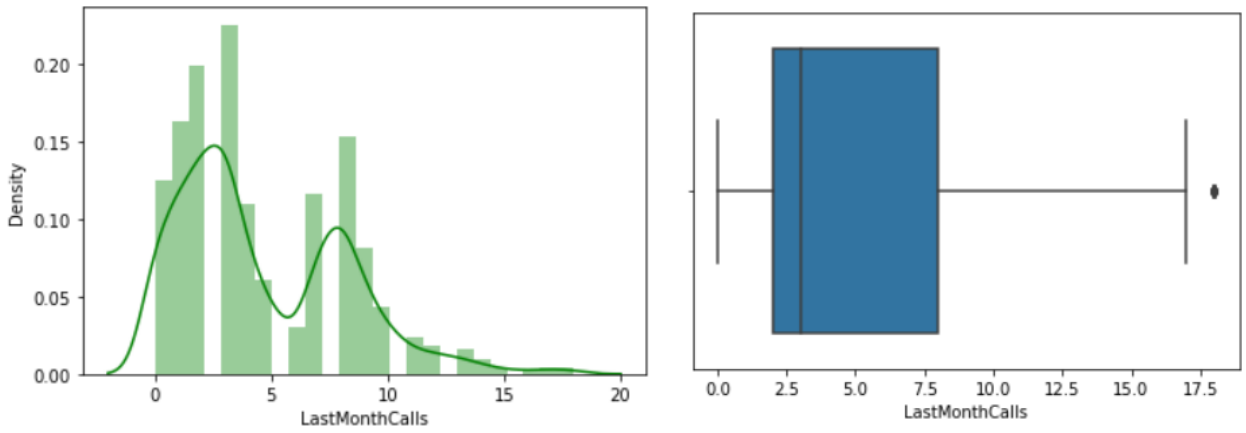

Figure 1(j) Distplot/Boxplot - LastMonthCalls 12

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

CustCareScore:

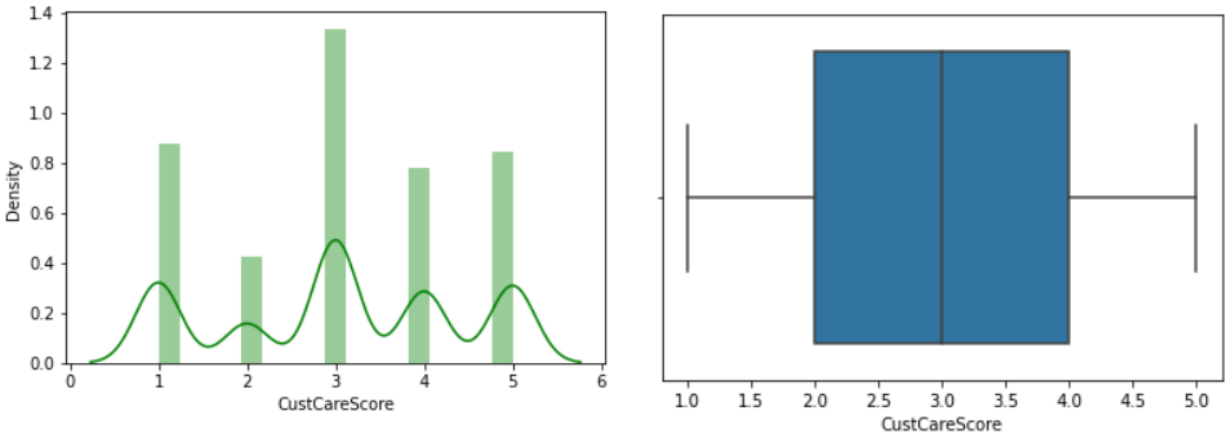

Figure 1(k) Distplot/Boxplot – CustCareScore

- **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
dtype: float64
```

- 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

Name: EducationField, dtype: float64

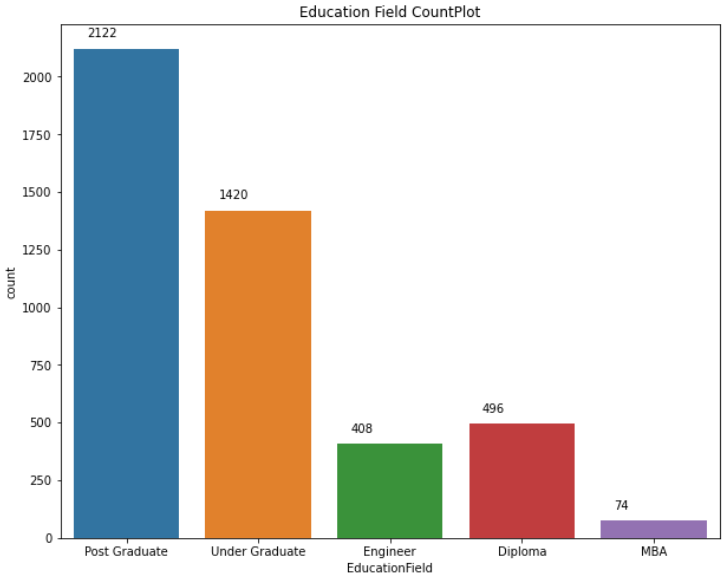

Figure 2(a) Count Plot - EducationField

- **Most Customers approached are Post Graduates having 47% weightage.**

Channel:

Agent	0.71
Third Party Partner	0.19
Online	0.10

Name: Channel, dtype: float64

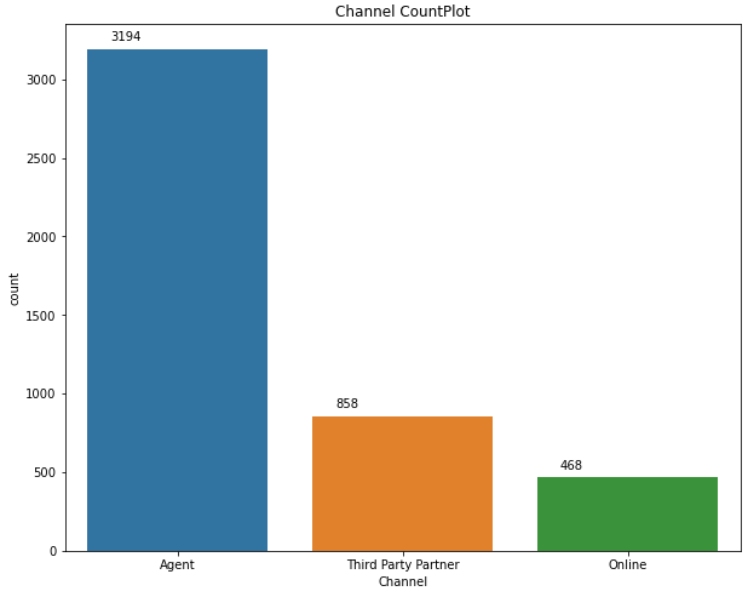

Figure 2(b) Count Plot - Channel

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

Occupation:

```
Salaried          0.48
Small Business    0.42
Large Business    0.09
Free Lancer       0.00
Name: Occupation, dtype: float64
```

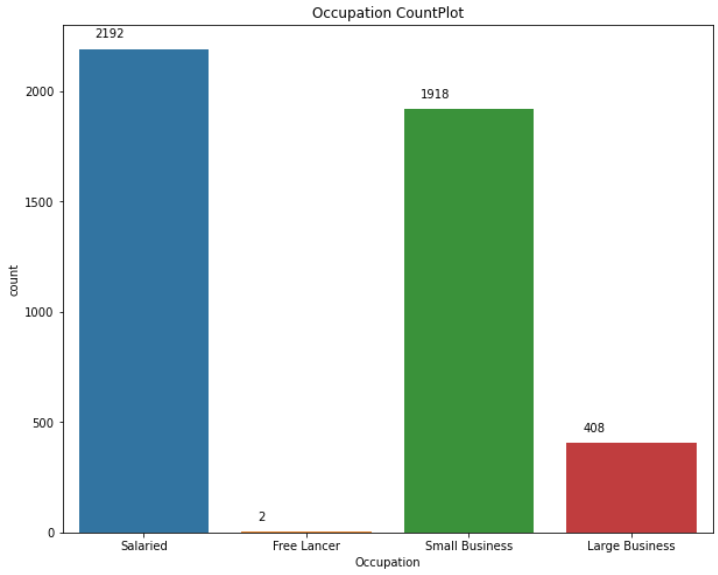

Figure 2(c) Count Plot - Occupation

- **Most customers have Salaried Occupations around 48%.**
- **Here freelancers have a low weightage.**

Gender:

Male 0.59
Female 0.41
Name: Gender, dtype: float64

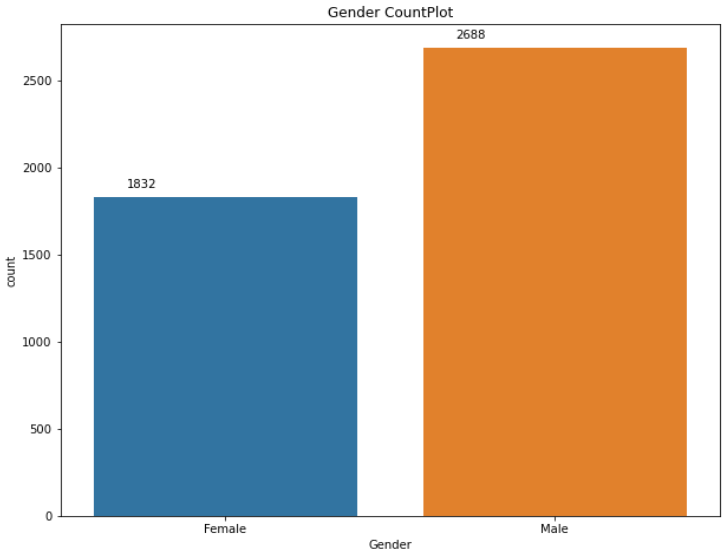

Figure 2(d) Count Plot - Gender

- **Approximately 59% of customers are males.**

Designation:

```
Executive          0.37
Manager            0.36
Senior Manager     0.15
AVP                0.07
VP                 0.05
Name: Designation, dtype: float64
```

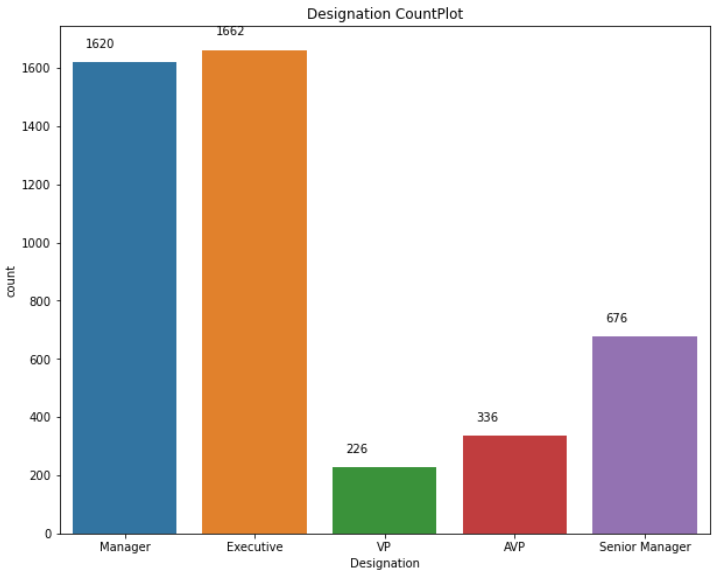

Figure 2(e) Count Plot – Designation

- **Most customers are either an executive or managers having weightage of 37% and 36% respectively.**

Marital Status:

```
Married      0.50
Single       0.28
Divorced     0.18
Unmarried    0.04
Name: MaritalStatus, dtype: float64
```

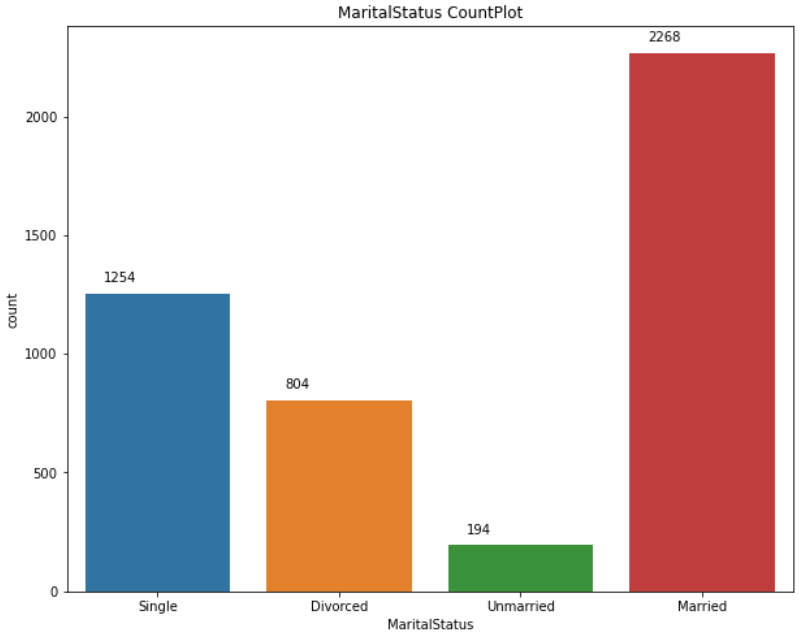

Figure 2(f) Count Plot -Marital Status

- **Around 50% of the customers are married.**

Zone:

```
West      0.57
North     0.42
East      0.01
South     0.00
```

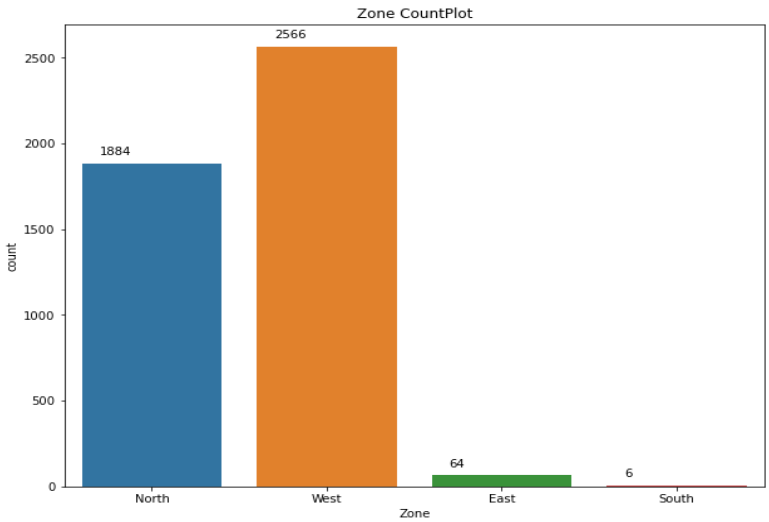

Figure 2(g) Count Plot – Zone

- **West Zone brings the most Customers with 57% weightage.**

PaymentMethod:

```
Half Yearly    0.59
Yearly         0.32
Monthly        0.08
Quarterly      0.02
```

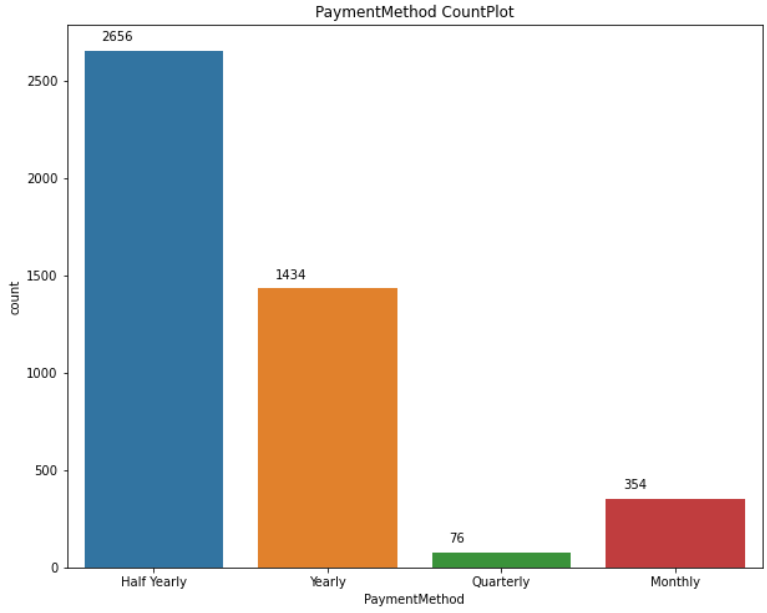

Figure 2(h) Count Plot – PaymentMethod

- **Around 59% of Customers went for half-yearly payment plan**

Categorical Variables Bivariate Analysis w.r.t Agent Bonus:

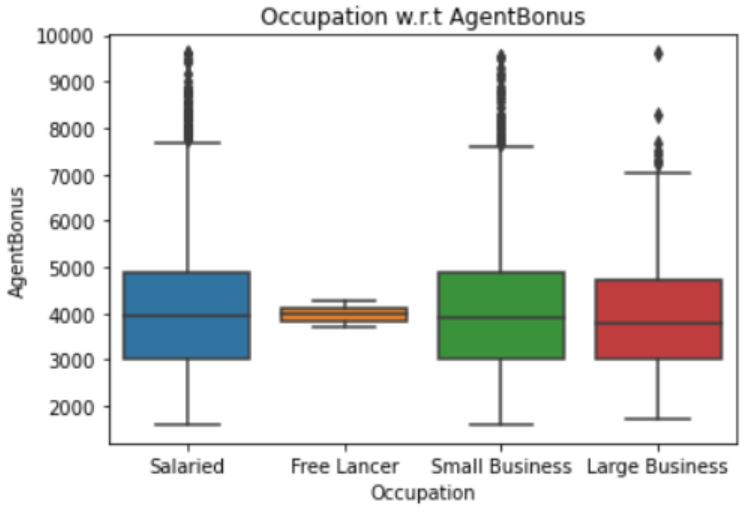

Figure 3(a) Boxplot – Occuapation w.r.t AgentBonus

- **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.**

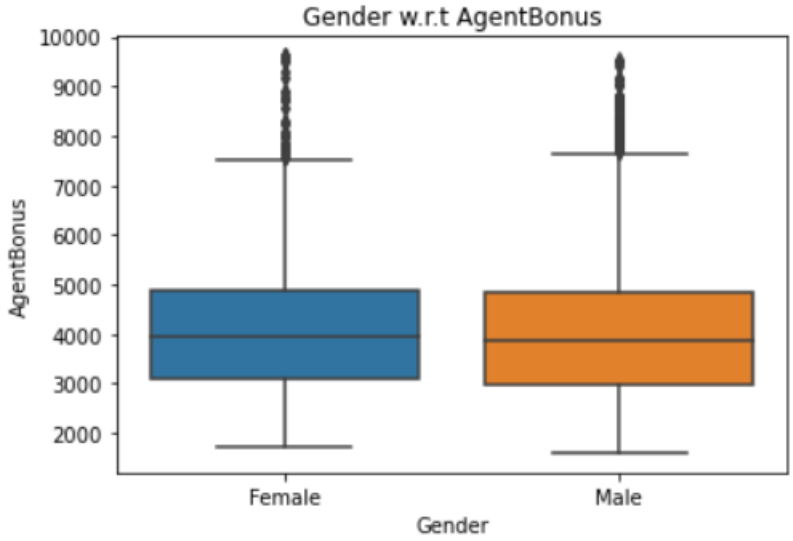

Figure 3(b) Boxplot – Gender w.r.t AgentBonus

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

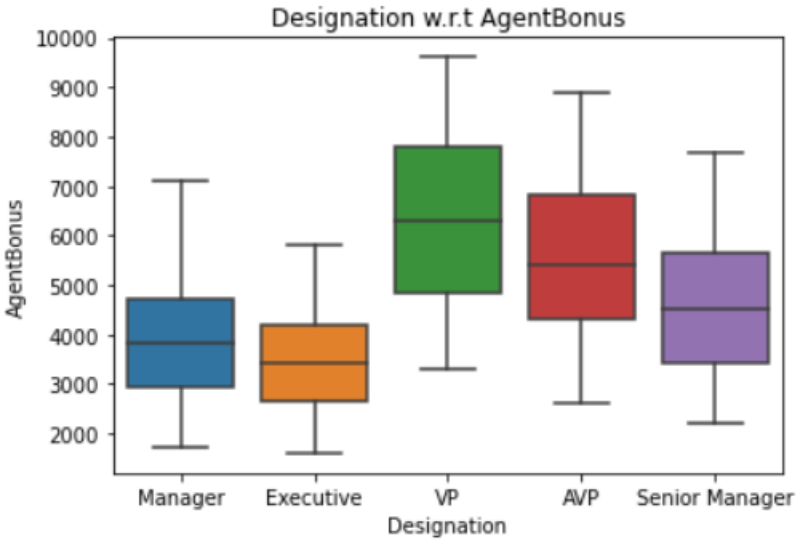

Figure 3(c) Boxplot – Designation w.r.t AgentBonus

- **No outliers present.**
- **VP Designation has the highest mean as compared to other Designations.**

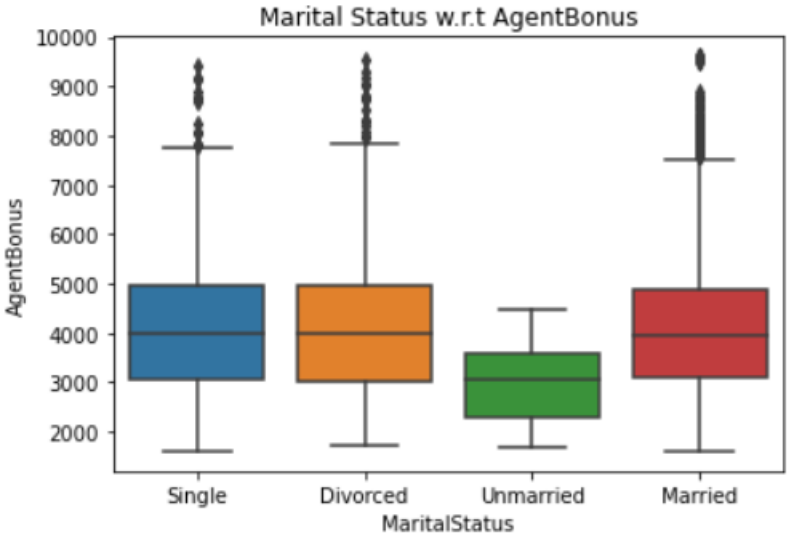

Figure 3(d) Boxplot – MaritalStatus w.r.t AgentBonus

- **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.**

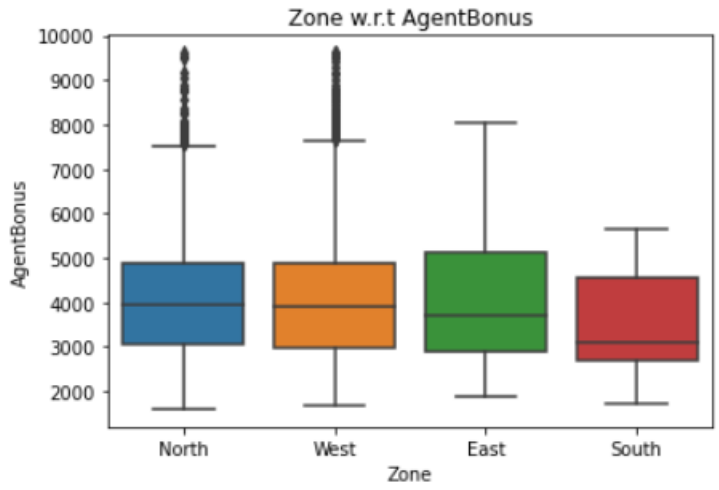

Figure 3(e) Boxplot – Zone w.r.t AgentBonus

- **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.**

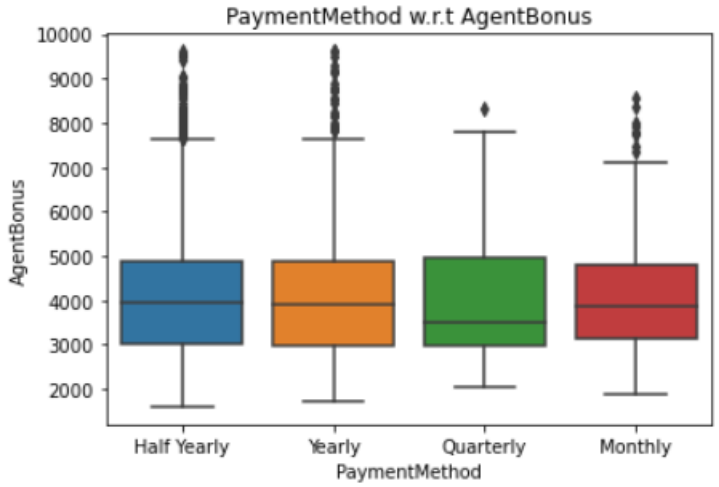

Figure 3(f) Boxplot – Payment w.r.t AgentBonus

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

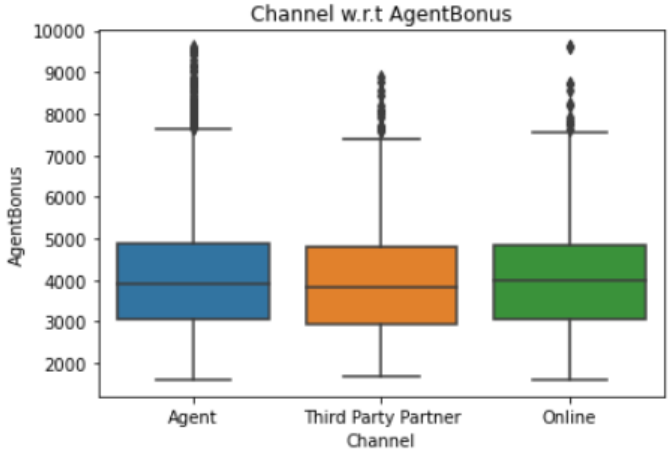

Figure 3(g) Boxplot – Channel w.r.t AgentBonus

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

Pair plot:

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 - Bivariate Analysis- Numerical Data Pairplot

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).

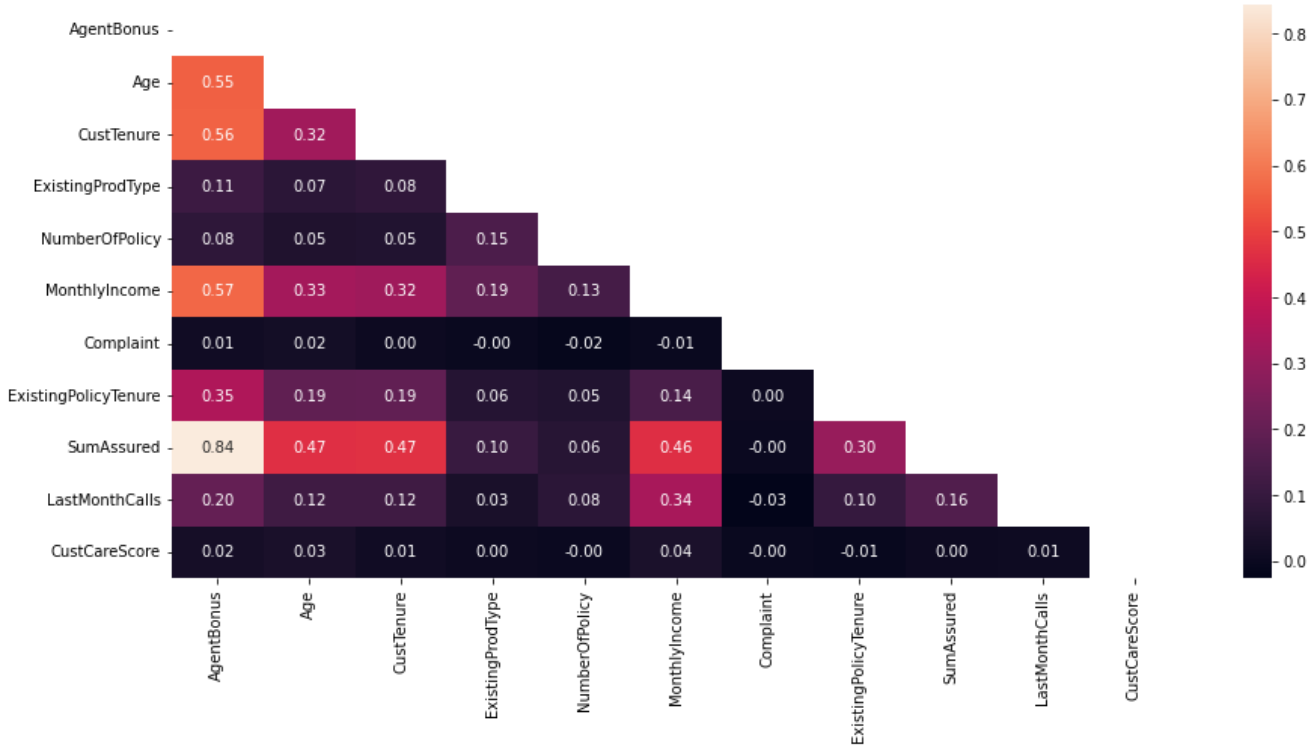

Figure 5 - Correlation Heatmap

Inferences:

- Here the lighter colours depict high correlation and darker colours depict low correlation.
- We can observe that there is almost no multicollinearity in the data.
- Multicollinearity refers to more variables affect our dependant variables, here from the graph above only SumAssured makes the cut as a variable affecting the AgentBonus
- Complaint and CustCareScore have almost no correlation with any other parameter, hence dropping these columns will not make a difference as they hold no weightage in predictions for our dependant variable, i.e., AgentBonus where these columns ultimately are ignored in the prediction, hence are removed.
- AgentBonus and SumAssured have high correlation with each other of 0.84.

Business insights from EDA:

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

We might have to convert some categorical variables by encoding them into numeric values for out model Building. Stay Tuned to find more.

Model Building and Interpretation:

- Regression uses numerical variables,
- But we have a lot of categorical variables we wish to use in our models further,
- And since most of the categorical variables have categories more than 2, therefore applying **one-hot encoding.**
- One-Hot encoding takes every level of the category and turns it into a variable with two level (yes/no).

The data looks like this after one-hot encoding:

ExistingPolicyTenure	SumAssured	LastMonthCalls	CustCareScore	...	MaritalStatus_Married	MaritalStatus_Single	MaritalStatus_Unmarried	Zone_North	Zone_So
1.0	570217.0	4.0	2.0	...	1	0	0	0	
2.0	398096.0	11.0	5.0	...	1	0	0	0	
4.0	1034360.0	8.0	3.0	...	1	0	0	1	
9.0	568846.0	2.0	3.0	...	1	0	0	0	
1.0	704583.0	2.0	3.0	...	0	1	0	0	

- Building our Linear Regression Model with the unprocessed data above.
- Note, this data holds no outliers as they were removed in EDA – PN1
- **Split X and Y into training and test set in 67:33 ratio:**

- **The coefficient for Age is 21.20745043230412**
- **The coefficient for CustTenure is 24.65855753517184**
- **The coefficient for ExistingProdType is 58.108212779272584**
- **The coefficient for NumberOfPolicy is 1.4104600051924319**
- **The coefficient for MonthlyIncome is 0.02492390722157958**
- **The coefficient for Complaint is 57.48038648121199**
- **The coefficient for ExistingPolicyTenure is 41.38698386079497**
- **The coefficient for SumAssured is 0.003569644886851935**
- **The coefficient for LastMonthCalls is -1.024472304939965**
- **The coefficient for CustCareScore is 5.536304592187767**
- **The coefficient for Channel_Online is 16.21576458259404**
- **The coefficient for Channel_Third Party Partner is -1.1867932896097877**
- **The coefficient for Occupation_Large Business is -125.42887885225629**
- **The coefficient for Occupation_Salaried is -121.44999892648654**
- **The coefficient for Occupation_Small Business is -219.05873148371518**
- **The coefficient for EducationField_Engineer is -109.84759428280452**
- **The coefficient for EducationField_MBA is -202.82156088533264**
- **The coefficient for EducationField_Post Graduate is -89.48036653686641**
- **The coefficient for EducationField_Under Graduate is -10.665753874612415**
- **The coefficient for Gender_Male is 7.862346290989826**
- **The coefficient for Designation_Executive is -578.9440452255853**
- **The coefficient for Designation_Manager is -526.3556284301319**
- **The coefficient for Designation_Senior Manager is -350.81090341439216**
- **The coefficient for Designation_VP is -28.06000542072714**
- **The coefficient for MaritalStatus_Married is -38.628525308598526**
- **The coefficient for MaritalStatus_Single is 22.22527865431414**
- **The coefficient for MaritalStatus_Unmarried is -142.17530103601595**
- **The coefficient for Zone_North is -71.28209216101781**
- **The coefficient for Zone_South is -0.19030454466694863**
- **The coefficient for Zone_West is -70.34142852177685**
- **The coefficient for PaymentMethod_Monthly is 208.22905347605132**
- **The coefficient for PaymentMethod_Quarterly is 89.30444398717106**
- **The coefficient for PaymentMethod_Yearly is -62.91527712851256**
- **The intercept for our model is 1010.4018101785405**

R-Squared	RMSE

Training	0.7996844080706142	607.412867349152
Testing	0.8033418997423953	603.1679916043022

Checking the same using statsmodel, to get more insights on p-value, r-squared and adjusted r-squared value.

Before we move to statsmodel,

- We need to rename some columns created after encoding as they have some spaces which will not be accepted my statsmodel.

COLUMN NAMES:

```
Index(['Age', 'CustTenure', 'ExistingProdType', 'NumberOfPolicy',
       'MonthlyIncome', 'Complaint', 'ExistingPolicyTenure', 'SumAssured',
       'LastMonthCalls', 'CustCareScore', 'Channel_Online',
       'Channel_Third Party Partner', 'Occupation_Large Business',
       'Occupation_Salaried', 'Occupation_Small Business',
       'EducationField_Engineer', 'EducationField_MBA',
       'EducationField_Post Graduate', 'EducationField_Under Graduate',
       'Gender_Male', 'Designation_Executive', 'Designation_Manager',
       'Designation_Senior Manager', 'Designation_VP', 'MaritalStatus_Married',
       'MaritalStatus_Single', 'MaritalStatus_Unmarried', 'Zone_North',
       'Zone_South', 'Zone_West', 'PaymentMethod_Monthly',
       'PaymentMethod_Quarterly', 'PaymentMethod_Yearly', 'AgentBonus'],
      dtype='object')
```

RENAMED COLUMNS (SPACES REMOVED):

```
Index(['Age', 'CustTenure', 'ExistingProdType', 'NumberOfPolicy',
       'MonthlyIncome', 'Complaint', 'ExistingPolicyTenure', 'SumAssured',
       'LastMonthCalls', 'CustCareScore', 'Channel_Online',
       'Channel_Third_Party_Partner', 'Occupation_Large_Business',
       'Occupation_Salaried', 'Occupation_Small_Business',
       'EducationField_Engineer', 'EducationField_MBA',
       'EducationField_Post_Graduate', 'EducationField_Under_Graduate',
       'Gender_Male', 'Designation_Executive', 'Designation_Manager',
       'Designation_Senior_Manager', 'Designation_VP', 'MaritalStatus_Married',
       'MaritalStatus_Single', 'MaritalStatus_Unmarried', 'Zone_North',
       'Zone_South', 'Zone_West', 'PaymentMethod_Monthly',
       'PaymentMethod_Quarterly', 'PaymentMethod_Yearly', 'AgentBonus'],
      dtype='object')
```

Building a Multiple Linear Regression Model, with 'AgentBonus' as the independent variable and all other variables as dependent variables - LINEAR MODEL 1

(LM1):

```
Intercept                       1092.348510
Age                               21.645436
CustTenure                        22.620905
ExistingProdType                  46.508784
NumberOfPolicy                     6.254332
MonthlyIncome                      0.031885
Complaint                         33.050381
ExistingPolicyTenure              40.229015
SumAssured                         0.003548
LastMonthCalls                    -2.308710
CustCareScore                      7.559057
Channel_Online                    22.691901
Channel_Third_Party_Partner        3.495278
Occupation_Large_Business       -616.860010
Occupation_Salaried             -474.972964
Occupation_Small_Business       -581.637241
EducationField_Engineer           26.675848
EducationField_MBA              -177.273687
EducationField_Post_Graduate     -92.609498
EducationField_Under_Graduate      2.331225
Gender_Male                       25.187256
Designation_Executive           -493.361225
Designation_Manager             -481.419266
Designation_Senior_Manager      -277.421219
Designation_VP                    -2.956791
MaritalStatus_Married            -48.203783
MaritalStatus_Single              29.658244
MaritalStatus_Unmarried         -188.879075
Zone_North                        62.354153
Zone_South                       193.510577
Zone_West                         49.998087
PaymentMethod_Monthly            141.951935
PaymentMethod_Quarterly          112.028794
PaymentMethod_Yearly             -79.920805
dtype: float64
```

OLS Regression Results

Dep. Variable:	AgentBonus	R-squared:	0.807
Model:	OLS	Adj. R-squared:	0.805
Method:	Least Squares	F-statistic:	424.7
Date:	Mon, 01 May 2023	Prob (F-statistic):	0.00
Time:	23:58:06	Log-Likelihood:	-26499.
No. Observations:	3390	AIC:	5.307e+04
Df Residuals:	3356	BIC:	5.327e+04
Df Model:	33		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1092.3485	467.264	2.338	0.019	176.198	2008.499
Age	21.6454	1.420	15.245	0.000	18.862	24.429
CustTenure	22.6209	1.428	15.840	0.000	19.821	25.421
ExistingProdType	46.5088	23.229	2.002	0.045	0.964	92.054
NumberOfPolicy	6.2543	7.560	0.827	0.408	-8.569	21.078
MonthlyIncome	0.0319	0.005	5.954	0.000	0.021	0.042
Complaint	33.0504	23.172	1.426	0.154	-12.381	78.482
ExistingPolicyTenure	40.2290	4.066	9.894	0.000	32.257	48.201
SumAssured	0.0035	5.88e-05	60.294	0.000	0.003	0.004
LastMonthCalls	-2.3087	3.109	-0.743	0.458	-8.405	3.787
CustCareScore	7.5591	7.644	0.989	0.323	-7.429	22.547
Channel_Online	22.6919	34.552	0.657	0.511	-45.054	90.438
Channel_Third_Party_Partner	3.4953	26.973	0.130	0.897	-49.389	56.380
Occupation_Large_Business	-616.8600	453.438	-1.360	0.174	-1505.902	272.182
Occupation_Salaried	-474.9730	428.923	-1.107	0.268	-1315.949	366.003
Occupation_Small_Business	-581.6372	436.329	-1.333	0.183	-1437.134	273.860
EducationField_Engineer	26.6758	155.095	0.172	0.863	-277.414	330.766
EducationField_MBA	-177.2737	123.966	-1.430	0.153	-420.330	65.783
EducationField_Post_Graduate	-92.6095	87.381	-1.060	0.289	-263.934	78.715
EducationField_Under_Graduate	2.3312	36.703	0.064	0.949	-69.631	74.293
Gender_Male	25.1873	21.339	1.180	0.238	-16.652	67.027
Designation_Executive	-493.3612	59.744	-8.258	0.000	-610.500	-376.222
Designation_Manager	-481.4193	50.448	-9.543	0.000	-580.330	-382.508
Designation_Senior_Manager	-277.4212	48.283	-5.746	0.000	-372.088	-182.755
Designation_VP	-2.9568	63.911	-0.046	0.963	-128.266	122.352
MaritalStatus_Married	-48.2038	28.749	-1.677	0.094	-104.572	8.164
MaritalStatus_Single	29.6582	31.785	0.933	0.351	-32.662	91.978
MaritalStatus_Unmarried	-188.8791	59.461	-3.177	0.002	-305.462	-72.296
Zone_North	62.3542	91.992	0.678	0.498	-118.011	242.720
Zone_South	193.5106	285.551	0.678	0.498	-366.362	753.383
Zone_West	49.9981	91.518	0.546	0.585	-129.439	229.435
PaymentMethod_Monthly	141.9519	56.403	2.517	0.012	31.363	252.541
PaymentMethod_Quarterly	112.0288	85.052	1.317	0.188	-54.730	278.787
PaymentMethod_Yearly	-79.9208	33.879	-2.359	0.018	-146.346	-13.496

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 5.53e+07. This might indicate that there are
strong multicollinearity or other numerical problems.

- **RMSE – value - 600.5900784990948**
- The variation in R-squared and Adjusted R-squared is not too significant.

VIF Calculation:

Age VIF	= 1.33
CustTenure VIF	= 1.32
ExistingProdType VIF	= 4.36
NumberOfPolicy VIF	= 1.12
MonthlyIncome VIF	= 4.17
Complaint VIF	= 1.01
ExistingPolicyTenure VIF	= 1.11
SumAssured VIF	= 1.73
LastMonthCalls VIF	= 1.2
CustCareScore VIF	= 1.03
Channel_Online VIF	= 1.05
Channel_Third_Party_Partner VIF	= 1.04
Occupation_Large_Business VIF	= 153.84
Occupation_Salaried VIF	= 427.21
Occupation_Small_Business VIF	= 434.53
EducationField_Engineer VIF	= 18.0
EducationField_MBA VIF	= 2.0
EducationField_Post_Graduate VIF	= 17.68
EducationField_Under_Graduate VIF	= 2.73
Gender_Male VIF	= 1.03
Designation_Executive VIF	= 7.73
Designation_Manager VIF	= 5.43
Designation_Senior_Manager VIF	= 2.73
Designation_VP VIF	= 1.84
MaritalStatus_Married VIF	= 1.92
MaritalStatus_Single VIF	= 1.88
MaritalStatus_Unmarried VIF	= 1.34
Zone_North VIF	= 19.18
Zone_South VIF	= 1.12
Zone_West VIF	= 19.15
PaymentMethod_Monthly VIF	= 2.13
PaymentMethod_Quarterly VIF	= 1.11
PaymentMethod_Yearly VIF	= 2.31

- Wherever VIF score > 5, multicollinearity is present.

- **Multicollinearity** is detected for **Occupation_Large_Business**, **Occupation_Salaried, Occupation_Small_Business, EducationField_Engineer, EducationField_Post_Graduate, Zone_North & Zone_West**

- **Designation_Executive, Designation_Manager** (can be omitted), **Zone_North, Zone_West**.

We still find we have multi collinearity in the dataset, to drop these values to a further lower level we can drop columns after performing stats model.

- *From stats model we can understand the features that do not contribute to the Model*
- *We can remove those features after that the VIF Values will be reduced.*
- ***Ideal value of VIF is less than 5%.***

Calculating VIF again after dropping variables having VIF>5:

Age VIF = 1.32
CustTenure VIF = 1.31
ExistingProdType VIF = 3.53
NumberOfPolicy VIF = 1.11
MonthlyIncome VIF = 1.7
Complaint VIF = 1.01
ExistingPolicyTenure VIF = 1.11
SumAssured VIF = 1.71
LastMonthCalls VIF = 1.17
CustCareScore VIF = 1.02
Channel_Online VIF = 1.02
EducationField_Engineer VIF = 1.11
EducationField_MBA VIF = 1.03
EducationField_Post_Graduate VIF = 1.13
Gender_Male VIF = 1.02
Designation_Manager VIF = 1.18
Designation_Senior_Manager VIF = 1.25
MaritalStatus_Married VIF = 1.92
MaritalStatus_Single VIF = 1.87
MaritalStatus_Unmarried VIF = 1.33
Zone_South VIF = 1.01
Zone_West VIF = 1.02
PaymentMethod_Monthly VIF = 1.92
PaymentMethod_Quarterly VIF = 1.09
PaymentMethod_Yearly VIF = 2.06

Stats-model Implementation:

- **Running statsmodel again after dropping the necessary variables above - LINEAR MODEL 2 (LM2):**

```
Intercept                   643.616060
Age                          21.878587
CustTenure                   22.719343
MonthlyIncome                 0.037191
ExistingPolicyTenure         40.175194
SumAssured                    0.003551
Designation_Executive      -427.448405
Designation_Manager        -436.759878
Designation_Senior_Manager -258.644927
MaritalStatus_Married       -67.607811
MaritalStatus_Unmarried    -226.243427
dtype: float64
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.806
Model:                            OLS   Adj. R-squared:                  0.805
Method:                 Least Squares   F-statistic:                     1399.
Date:                Thu, 18 May 2023   Prob (F-statistic):               0.00
Time:                        14:24:39   Log-Likelihood:                -26511.
No. Observations:                3390   AIC:                         5.304e+04
Df Residuals:                    3379   BIC:                         5.311e+04
Df Model:                          10
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                      643.6161    129.776      4.959      0.000     389.168     898.064
Age                             21.8786      1.416     15.451      0.000      19.102      24.655
CustTenure                      22.7193      1.424     15.955      0.000      19.927      25.511
MonthlyIncome                    0.0372      0.004      8.473      0.000       0.029       0.046
ExistingPolicyTenure            40.1752      4.037      9.951      0.000      32.259      48.091
SumAssured                       0.0036   5.85e-05     60.654      0.000       0.003       0.004
Designation_Executive         -427.4484     52.722     -8.108      0.000    -530.818    -324.079
Designation_Manager           -436.7599     45.193     -9.664      0.000    -525.367    -348.152
Designation_Senior_Manager    -258.6449     43.277     -5.977      0.000    -343.496    -173.794
MaritalStatus_Married          -67.6078     21.235     -3.184      0.001    -109.243     -25.973
MaritalStatus_Unmarried       -226.2434     55.495     -4.077      0.000    -335.050    -117.437
==============================================================================
Omnibus:                      128.393   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              143.854
Skew:                           0.475   Prob(JB):                     5.79e-32
Kurtosis:                       3.341   Cond. No.                     9.23e+06
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.23e+06. This might indicate that there are
strong multicollinearity or other numerical problems.
```

- **As it can be observed above the P-value for multiple variables are greater than our alpha i.e 0.05, depicting multicollinearity present therefore we will drop the variables and perform the stats model again.**
- To ideally bring down the values to lower levels we can drop one of the variables that is highly correlated.
- Dropping variables would bring down the multi collinearity level down

	RMSE(LM1)	RMSE(LM2)
Training	600.5900784990952	602.6246250878111
Testing	621.5274260070425	620.4861930401262

- Since for model 2 our RMSE value has increased, it is not an optimal way to choose the new model. Not a significant change in R-squared either.
- Removing variables until all the insignificant variables are removed.

OLS Regression Results

Dep. Variable:	AgentBonus	R-squared:	0.806
Model:	OLS	Adj. R-squared:	0.805
Method:	Least Squares	F-statistic:	1399.
Date:	Thu, 18 May 2023	Prob (F-statistic):	0.00
Time:	14:24:39	Log-Likelihood:	-26511.
No. Observations:	3390	AIC:	5.304e+04
Df Residuals:	3379	BIC:	5.311e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	643.6161	129.776	4.959	0.000	389.168	898.064
Age	21.8786	1.416	15.451	0.000	19.102	24.655
CustTenure	22.7193	1.424	15.955	0.000	19.927	25.511
MonthlyIncome	0.0372	0.004	8.473	0.000	0.029	0.046
ExistingPolicyTenure	40.1752	4.037	9.951	0.000	32.259	48.091
SumAssured	0.0036	5.85e-05	60.654	0.000	0.003	0.004
Designation_Executive	-427.4484	52.722	-8.108	0.000	-530.818	-324.079
Designation_Manager	-436.7599	45.193	-9.664	0.000	-525.367	-348.152
Designation_Senior_Manager	-258.6449	43.277	-5.977	0.000	-343.496	-173.794
MaritalStatus_Married	-67.6078	21.235	-3.184	0.001	-109.243	-25.973
MaritalStatus_Unmarried	-226.2434	55.495	-4.077	0.000	-335.050	-117.437

Omnibus:	128.393	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	143.854
Skew:	0.475	Prob(JB):	5.79e-32
Kurtosis:	3.341	Cond. No.	9.23e+06

The overall P value is less than alpha, so rejecting H0 and accepting Ha that at least 1 regression co-efficient is not 0. Here all regression co-efficient are not 0

We can see all variables are having p-value < 0.05 and the r-squared value hasn't changes much either

	RMSE (LM2)	RMSE (LM1)
Training	602.6246250878111	600.5900784990952
Testing	620.4861930401804	621.5274260080358

Since for model 2 our RMSE value has increased, it is not an optimal way to choose the new model.

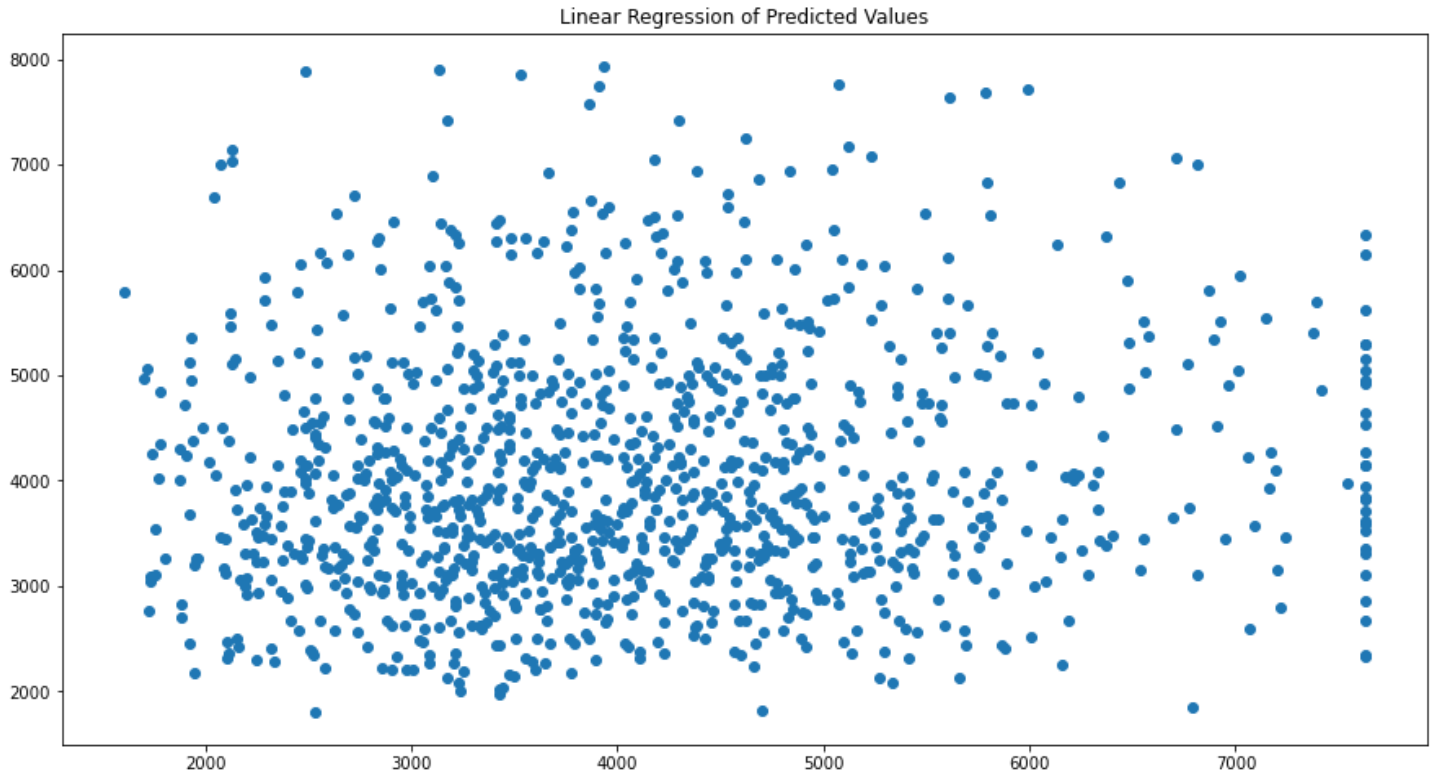

Figure 6 – Linear Regression Scatterplot

The variables are following a linear trend with a little homoscedasticity.

Comparing Linear Regression Model with Other models like Random Forest, Artificial Neural Network and Decision Trees – With base parameter values are no hyperparameter tuning the parameters.

We are scaling the data for ANN. Without scaling it will give very poor results. Computations becomes easier.

Scaling is done as some variables with greater weight will affect the predictions more, hence scaling is done to bring all variables in a common range e.g., 0 to 1. Due to which the predictions can be unbiased and not biased to one specific variable with higher weights. For e.g., age and sum assured.

SCALING:

- Scaling can be useful to reduce or check the multi collinearity in the data, so if scaling is not applied, I find the VIF – variance inflation factor values very high. Which indicates presence of multi collinearity

- *These values are calculated after building the model of linear regression. To understand the multi collinearity in the model*

- *The scaling had no impact in model score or coefficients of attributes nor the intercept.*

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	612.550689	585.514819	0.800806	0.801482
Decision Tree Regressor	0.000000	725.006753	1.000000	0.695626
Random Forest Regressor	189.614010	519.044211	0.980913	0.843997
ANN Regressor	225.889011	701.144120	0.972912	0.715332

- **Here Linear Regression is the best performing model with almost same Training and Testing Accuracies.**
- **On the other hand, we can observe that the other three models namely, Decision Tree, Random Forest, and ANN are Overfitting the model, i.e., the model is performing better while training but poorly while testing.**
- **To fix this we will use Hyperparameter Tuning, this will be done by performing grid search.**
- **Checking if PCA can be applied here**

```
Cumulative Variance Explained [ 99.97511098  99.99912638  99.99999976  99.99999986  99.99999995
  99.99999997  99.99999998  99.99999999  99.99999999  99.99999999
  99.99999999 100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.         100.
 100.         100.         100.         100.        ]
```

- **Since cumulative variance is almost 99%, hence there is no need to perform PCA.**

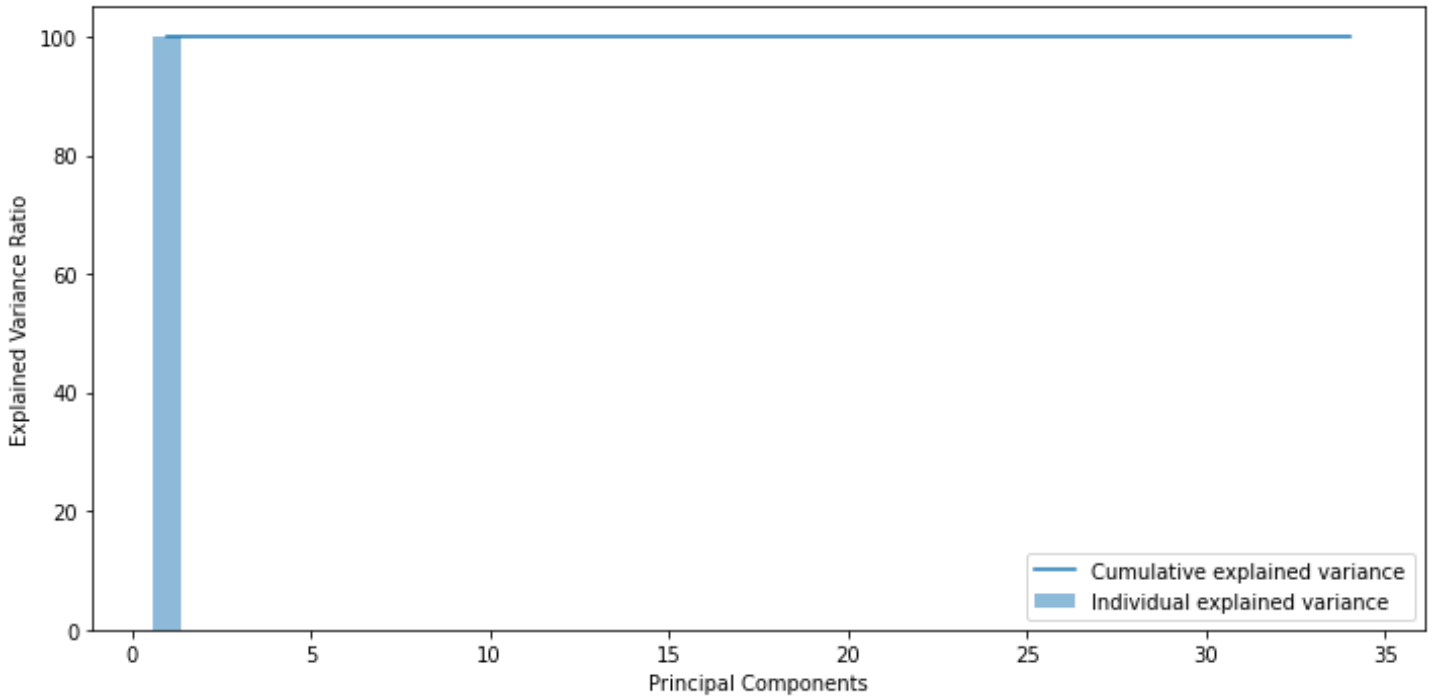

Figure 7 – Principal Components vs Variance Ratio

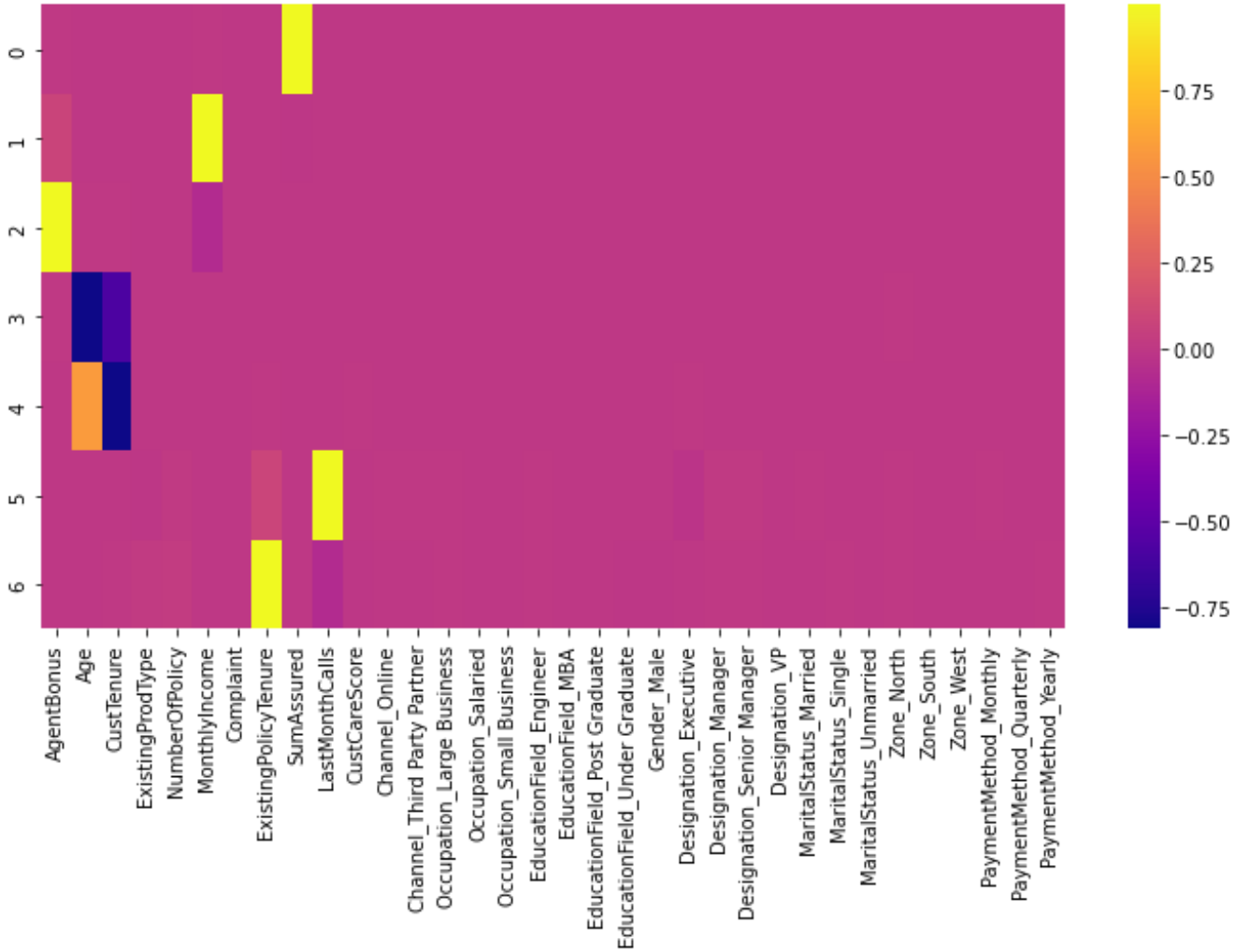

Figure 8 – PCA Heatmap

- **Not much can be observed about the components from the heatmap, therefore dropping the need to perform PCA as almost all these variables hold a good deal of significance in the predictions.**

MODEL TUNING:

- **We will perform grid search for hyperparameter tuning and check if that makes a difference in our accuracies.**

Grid Search on Decision Tree

Best parameters - `{'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 40}`

Grid Search on Random Forest:

```
GridSearchCV(cv=3, estimator=RandomForestRegressor(random_state=123),
             param_grid={'max_depth': [7, 10], 'max_features': [4, 6],
                         'min_samples_leaf': [3, 15, 30],
                         'min_samples_split': [30, 50, 100],
                         'n_estimators': [300, 500]})
```

Best parameters - `{'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 500}`

Using Grid Search for ANN:

```
GridSearchCV(cv=3, estimator=MLPRegressor(max_iter=10000, random_state=123),
param_grid={'activation': ['tanh', 'relu'],
'hidden_layer_sizes': [500, (100, 100)],
'solver': ['sgd', 'adam']})
```

Best parameters - `{'activation': 'tanh', 'hidden_layer_sizes': 500, 'solver': 'adam'}`

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	612.550689	585.514819	0.800806	0.801482
Decision Tree Regressor	495.463438	569.694730	0.869679	0.812065
Random Forest Regressor	527.410585	572.885614	0.852331	0.809954
ANN Regressor	28.117642	670.444991	0.999580	0.739715

- ***After Hyperparameter tuning it can be observed the problem of overfitting is removed for most of the models however some overfitting can be observed in ANN.***
- ***Apart from this, we can observe Linear Regression is still the most stable having not much variation between training and testing sets.***
- ***If you're looking for more stable Model, definitely go for Linear Regression model, else Decision Tree and Random Forest can be chosen for higher accuracy and are good models as the there's only 5% fluctuations between training and testing model.***
- ***Random forest is the better choice between the Regressors as random forest is the more advanced version of decision trees where we can further tweak the parameters according to the needs.***

Feature Importance from the model can be observed here:

	Imp
SumAssured	0.422989
CustTenure	0.158176
Age	0.148770
MonthlyIncome	0.111202
ExistingPolicyTenure	0.037939
Designation_Executive	0.032872
Designation_VP	0.028287
Designation_Manager	0.012450
LastMonthCalls	0.010957
Designation_Senior Manager	0.007725
ExistingProdType	0.004555
NumberOfPolicy	0.003833
MaritalStatus_Unmarried	0.003550
CustCareScore	0.002795
MaritalStatus_Single	0.001091
Gender_Male	0.001069
Zone_North	0.001028
Complaint	0.001018
MaritalStatus_Married	0.000984
Zone_West	0.000983
Channel_Third Party Partner	0.000940
EducationField_Post Graduate	0.000831
Occupation_Salaried	0.000819
Occupation_Small Business	0.000818
PaymentMethod_Yearly	0.000793
Channel_Online	0.000674
EducationField_Under Graduate	0.000668
EducationField_Engineer	0.000646
PaymentMethod_Monthly	0.000556
Occupation_Large Business	0.000493
EducationField_MBA	0.000276
PaymentMethod_Quarterly	0.000211
Zone_South	0.000003

- **Sum Assured is the most important feature here, Zone_South being the least important.**

MODEL SELECTION:

- From the previous results, it is evident that Linear Regression is a better model.
- Why Linear Regression?
  - Post removal of variables causing multicollinearity, Linear Regression provided a good R-squared value and similarly a high adjusted R squared value. Hence a good percentage of variance can be successfully explained by our model.
  - A very important factor being the train and test set accuracy scores are ~80% and consistent.
  - Unlike other models where overfitting and inconsistency in the performance metrics can be observed. Linear Regression model does not show these inconsistencies in the observation.

(Here by overfitting, we mean, the model is performing very good for training set and giving poor results for the testing set)

- The LR model makes it easier to understand the model, multicollinearity in the data. Also, unlike other model its computational time is quick therefore we can run it multiple times whereas ANN and Random Forests needs capable machines as they are very time-consuming models. Might have to wait for hours and in our case, they still don't perform better than LR.

Note: 100 % accuracy cannot be achieved in real life data as there is always some unexplainable factors and noise that's always present in our data.

MODEL EVALUATION:

The Equation:

```
(1092.35) * Intercept + (21.65) * Age + (22.62) * CustTenure + (46.51) * ExistingProdType
+ (6.25) * NumberOfPolicy + (0.03) * MonthlyIncome + (33.05) * Complaint + (40.23) * Exis
tingPolicyTenure + (0.0) * SumAssured + (-2.31) * LastMonthCalls + (7.56) * CustCareScore
+ (22.69) * Channel_Online + (3.5) * Channel_Third_Party_Partner + (-616.86) * Occupation
_Large_Business + (-474.97) * Occupation_Salaried + (-581.64) * Occupation_Small_Business
+ (26.68) * EducationField_Engineer + (-177.27) * EducationField_MBA + (-92.61) * Educati
onField_Post_Graduate + (2.33) * EducationField_Under_Graduate + (25.19) * Gender_Male +
(-493.36) * Designation_Executive + (-481.42) * Designation_Manager + (-277.42) * Designa
tion_Senior_Manager + (-2.96) * Designation_VP + (-48.2) * MaritalStatus_Married + (29.66
) * MaritalStatus_Single + (-188.88) * MaritalStatus_Unmarried + (62.35) * Zone_North + (
193.51) * Zone_South + (50.0) * Zone_West + (141.95) * PaymentMethod_Monthly + (112.03) *
PaymentMethod_Quarterly + (-79.92) * PaymentMethod_Yearly +
```

- **From the equation the variables with a low or no coefficient value depicts that the variable is very important to the independent variable's prediction. As the coefficients value increase it shows the variable has become comparatively less significant.**

The variable significance can be explained using the * method, where * depicts highly significant, ** less significant, and * and **** least significant**

Variables	Significance
SumAssured, MonthlyIncome	*
LastMonthCalls, CustCareScore, Channel_Third_Party_Partner, EducationField_Under_Graduate, Designation_VP, NumberOfPolicy	**
Age, CustTenure, Channel_Online, EducationField_Engineer, Gender_Male, MaritalStatus_Single, Complaint, ExistingPolicyTenure, MaritalStatus_Married, Zone_West, Zone_North, PaymentMethod_Yearly, EducationField_Post_Graduate	***
Occupation_Large_Business, Occupation_Salaried, Occupation_Small_Business, EducationField_MBA, Designation_Executive, Designation_Manager, Designation_Senior_Manager, MaritalStatus_Unmarried, Zone_South, Paymentmethod_Monthly, PaymentMethod_Quaterly	****

- **R-Squared Obtained from final Linear Regression Model:0.807 (LM1) & 0.806 (LM2)**
- **Adjusted R-Squared Obtained from final Linear Regression Model:0.805**
- **Decision Trees, Random Forest, and ANN (Before Hyperparameter Tuning):**
  - It can be observed that all the 3 models have overfitting problems where we have ideal accuracies of ~100% for our training set. However, the models are performing poorly on our testing set having accuracies ~70% – 84%. There is a major accuracy difference between the training and testing set which is not acceptable for predictions.
  - If the accuracy difference is greater than 6-10% it is advised to not accept the model as the predictions can be unreliable.
- **Decision Trees, Random Forest, and ANN (After Hyperparameter Tuning):**
  - After Hyperparameter Tuning Decision Trees and Random Forest models showed no overfitting errors.
  - The training accuracies were ~85% and testing accuracies were ~80%.
  - ANN still showed no improvement in results and was still overfitting.

- Although the Decision Trees and Random Forest were performing good, **I went with Linear Regression** as it gave more stable results and Variable importance could be calculated more easily from the Linear Regression Equation and stats-model performed to predict the results.

Interpretation and Business Recommendations:

- *Company wants to predict the ideal bonus and what is the engagement for high and low performing agents respectively.*
- *From the model, the high performing agent we will find variable significance, for e.g., Sum Assured is highly significant here.*
- *If the Designation is VP the person buys more policy or high value policies.*
- *Therefore, for high and low performing agents, we will train them, suggesting them to purchase or get policies with high sum assured as it is very significant to our model.*
- *Another important feature is Customer tenure where the agents need to focus on the customers who've a tenure ranging between 8-20 this where the majority of the customer are.*
- *Focusing on customers with greater monthly incomes as greater the monthly income, greater is the possibility of the customer buying a higher valued policy.*

Recommendations:

- *For High Performing Agents we can create a healthy contest with a threshold.*
- *Where, if they achieve the desired sum assured, they are eligible for certain incentives like latest gadgets, exotic family vacation packages and some extra perks as well.*
- *For low performing agents, we can introduce certain feedback upskill programs to train them into closing higher sum assured policies, reaching certain people to ultimately becoming top/high performers.*
- *Apart from this, we need more data/predictors like Premium Amount, this will help us to solve the business problem even better as well have more variables to test upon thereby having more accurate results in real time problems like this.*
- *I also feel another predictor can be added as customers geographical location or Region and not just the zones as people living in rural areas are less likely to buy a policy whereas those living in a highly developed location are likely to be belonging to the upper class and should be targeted.*
- *Similarly, another predictor can be '**AgentID** 'can be introduced which will make it easier to observe the high and low performing agent trend.*