11 APRIL 2021

CAPSTONE REPORT

INTRODUCTION

The dataset provided is the dataset of Insurance Sales. We need to analyse the data and predict the amount of bonus given to the agents based on the customer information provided in the dataset.

The dataset is a customer dataset which can help to generate new insights that paint a more complete picture of a customer. What are there buying habits? What is their risk profile? How appropriate are they to buy new insurance? Which in turn helps us to target specific audiences and spend time and resources accordingly. Also provide opportunities to cross sell and upsell.

Through this project , we tried to build various model and come up with the best fit model in order to provide a solution to company which will help them in categorising their agents in good or low performing agents which in turn will save company's valuable time in deciding for agents bonuses and upscale programs during company's appraisal cycle.

EXPLORATORY DATA ANALYSIS (EDA)

The data is sales data that is collected by an Insurance Company internally over a period of time describing how many insurance policies they have sold to the customers.

	0	1	2	3	4
CustID	7000000	7000001	7000002	7000003	7000004
AgentBonus	4409	2214	4273	1791	2955
Age	22	11	26	11	6
CustTenure	4	2	4	NaN	NaN
Channel	Agent	Third Party Partner	Agent	Third Party Partner	Agent
Occupation	Salaried	Salaried	Free Lancer	Salaried	Small Business
EducationField	Graduate	Graduate	Post Graduate	Graduate	UG
Gender	Female	Male	Male	Fe male	Male
ExistingProdType	3	4	4	3	3
Designation	Manager	Manager	Exe	Executive	Executive
NumberOfPolicy	2	4	3	3	4
MaritalStatus	Single	Divorced	Unmarried	Divorced	Divorced
MonthlyIncome	20993	20130	17090	17909	18468
Complaint	1	0	1	1	0
ExistingPolicyTenure	2	3	2	2	4
SumAssured	806761	294502	NaN	268635	366405
Zone	North	North	North	West	West

PaymentMethod	Half Yearly	Yearly	Yearly	Half Yearly	Half Yearly
LastMonthCalls	5	7	0	0	2
CustCareScore	2	3	3	5	5

Table 1

Although we need to calculate Agent Bonus but the dataset is a customer centric dataset.

There are a total of **4520** records in the dataset.

There are a total of **20** columns namely **-**

Variable	Discerption
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

Table 2

There were some data qualities issues that we worked on:

1. In Occupation 'Laarge Business' and 'Large Business' seems to be the same.
2. In EducationField we should consider 'UG' and 'Under Graduate' as the same.
3. In the Gender column variable 'Fe male' should be replaced with 'Female'.
4. In Designation 'Exe' can be replaced with 'Executive'.
5. In EducationField we can replace 'Engineer' with 'Graduate' and 'MBA' with 'Post Graduate'.

UNIVARIATE ANALYSIS

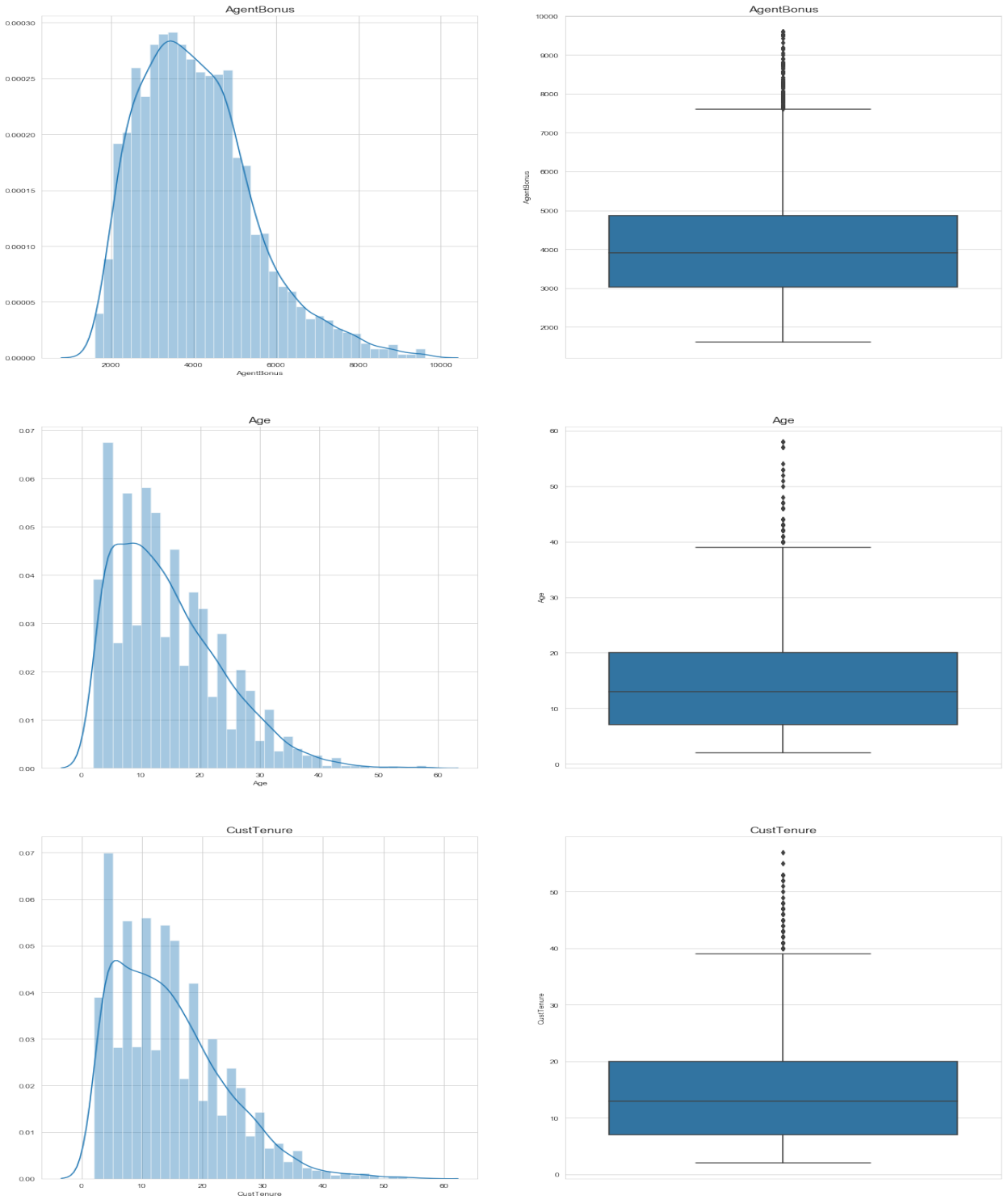

Figure 1

From the above analysis we can state -

- AgentBonus which is our target variable has skewness in the distribution, the distribution is right skewed and the box plot shows that there are outliers too.
- Age has right skewed distribution and has outliers as we can see the spread of Age is from 2 years to 58 years this is because there is an age limit to buy insurance. Further we have more customers of age 5 years this is due customers usually buy insurance for their children.
- The distribution of CustTenure is also right skewed and there are outliers. CustTenure is between 2 years and 58 years. Highest frequency is 4 years.

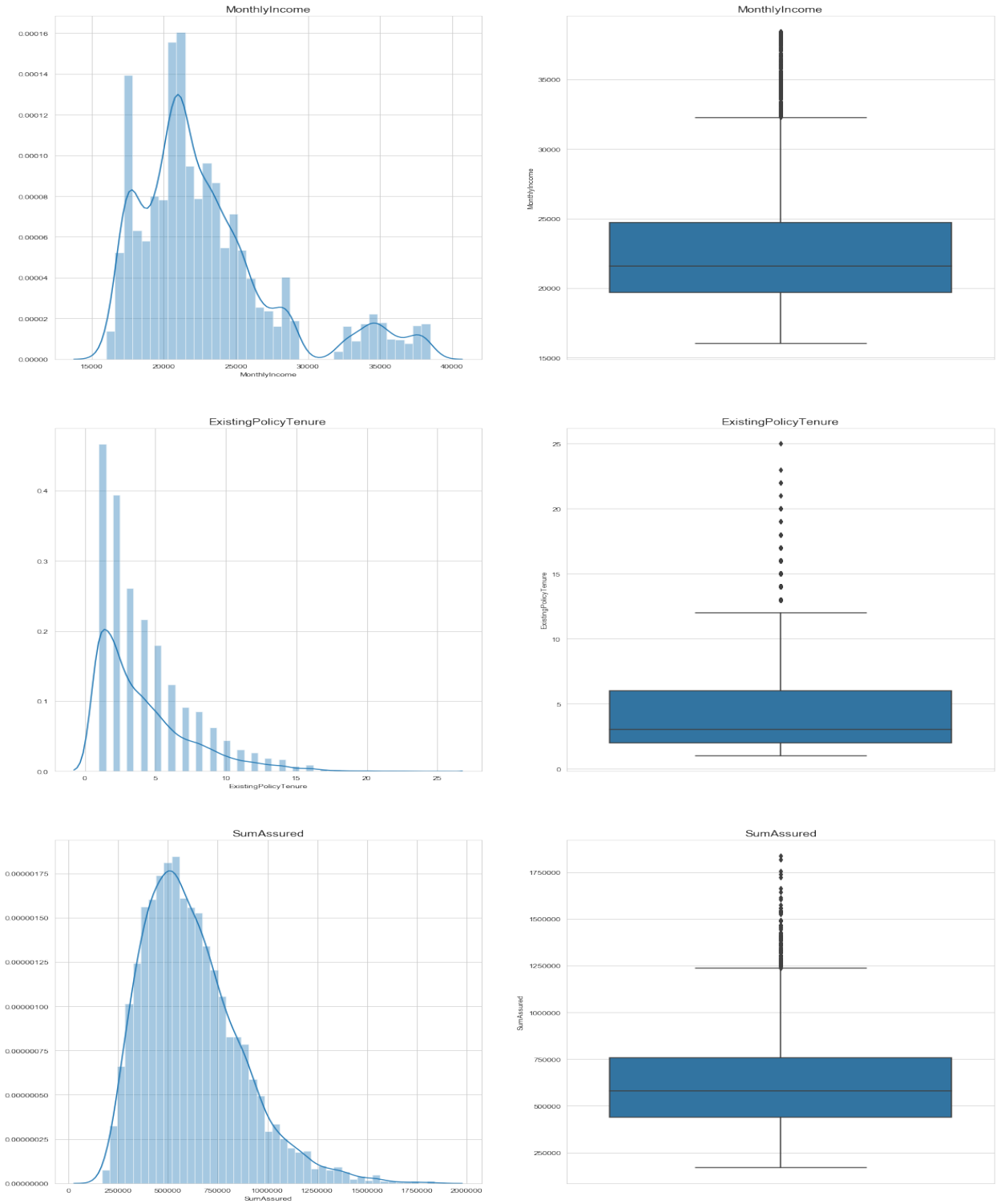

Figure 2

From the above analysis we can state the following:

1. The distribution of MonthlyIncome is right skewed, there are outliers as shown by boxplot, most of the customers have monthly income between 20000 to 25000.

2. ExistingPolicyTenure has a right skewed distribution and has outliers. Most frequent value is 1 year.
3. The distribution of SumAssured is also right skewed and there are outliers.

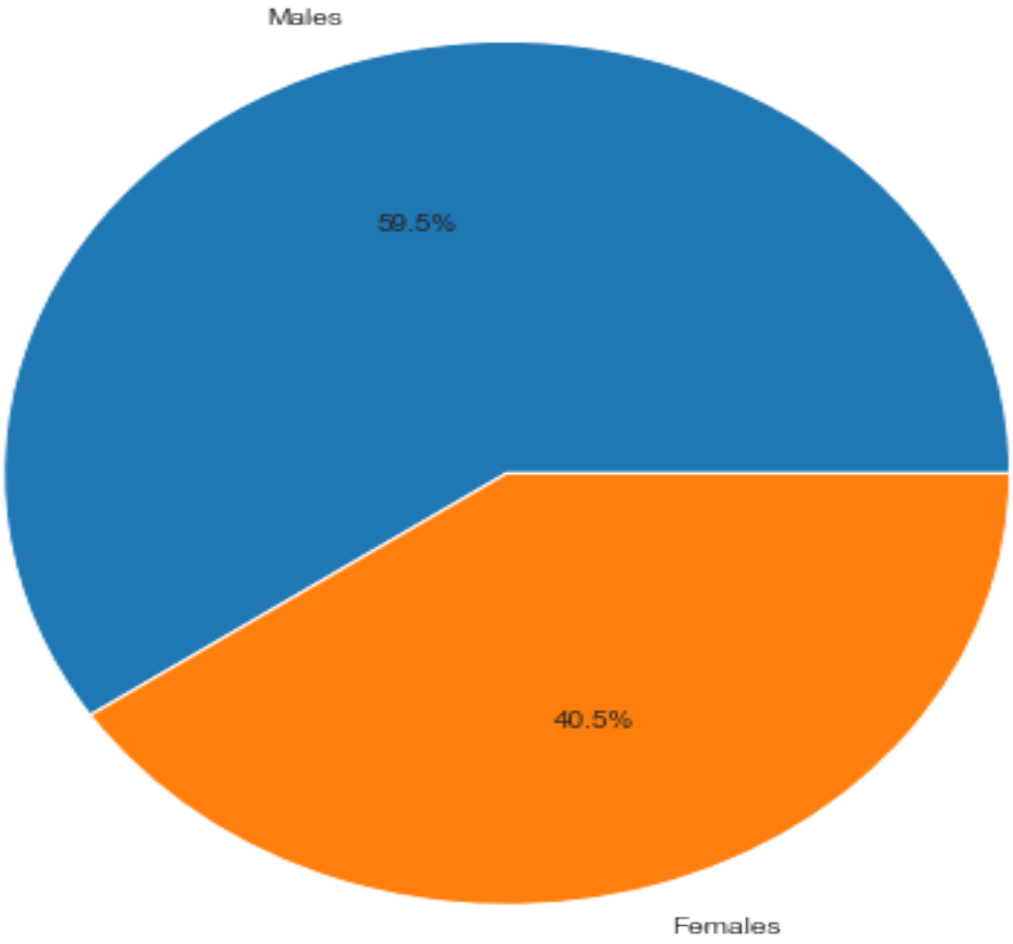

Figure 3

The customers are both Males and Females where 50.5% customers are Male and 40.5% are Females.

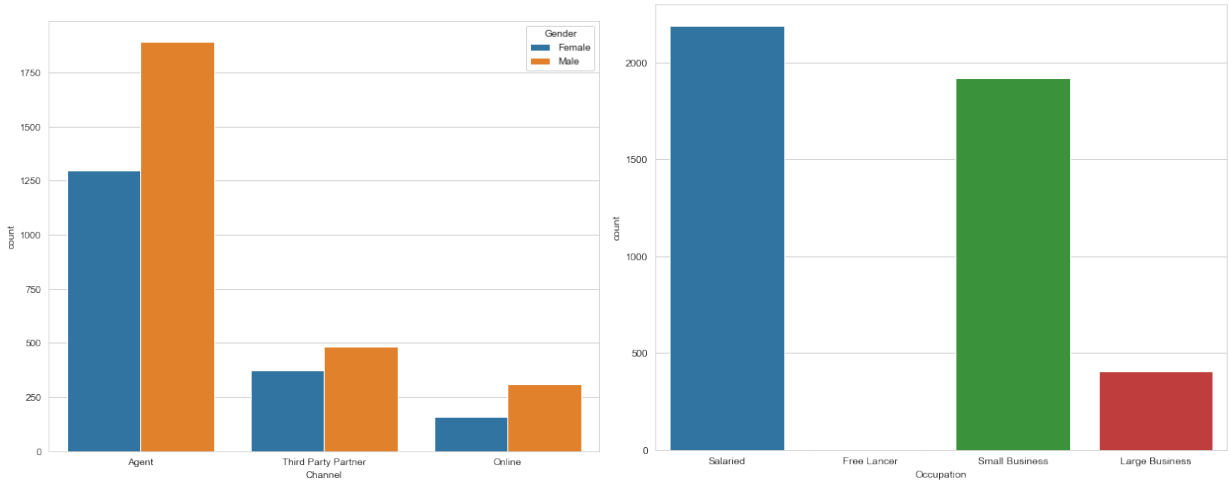

Figure 4

From the above figure we can say:

- The customer acquisition is done mostly through Agent channel, followed by Third Party Partner and Online.
- Male customers are more than Female customers in all the Channels.
- We can clearly see most of our customers are Salaried employees followed by customers that have Small Business, less than 500 customers have Large Business and to be specific Free Lancers are only two.

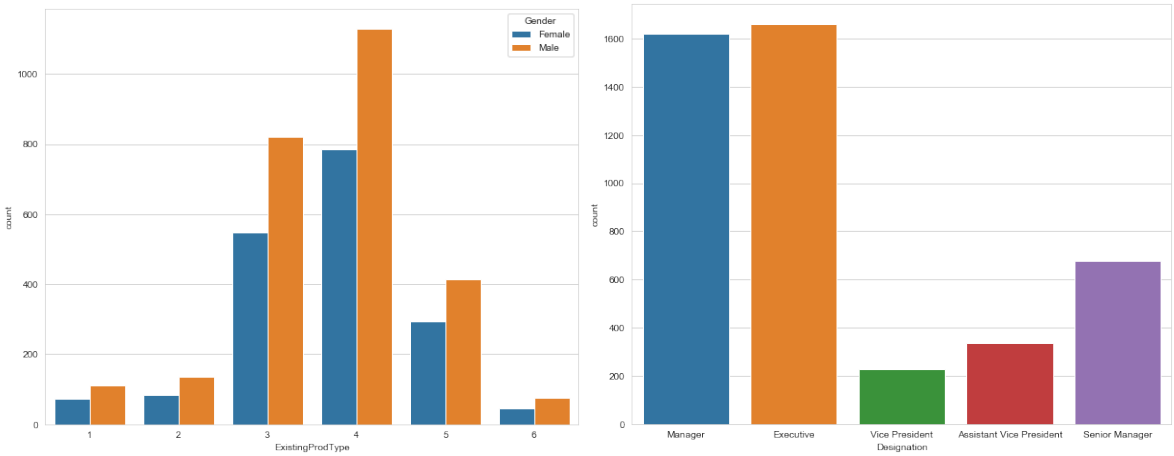

Figure 6

- As we can see from the figure above ExistingProdType ‘4’ is most popular followed by ‘3’, ‘5’, ‘2’, ‘1’ and ‘6’ respectively.
- Most of our customer are Executive followed by Manager, Senior Manager, Assistant Vice President and Vice President.

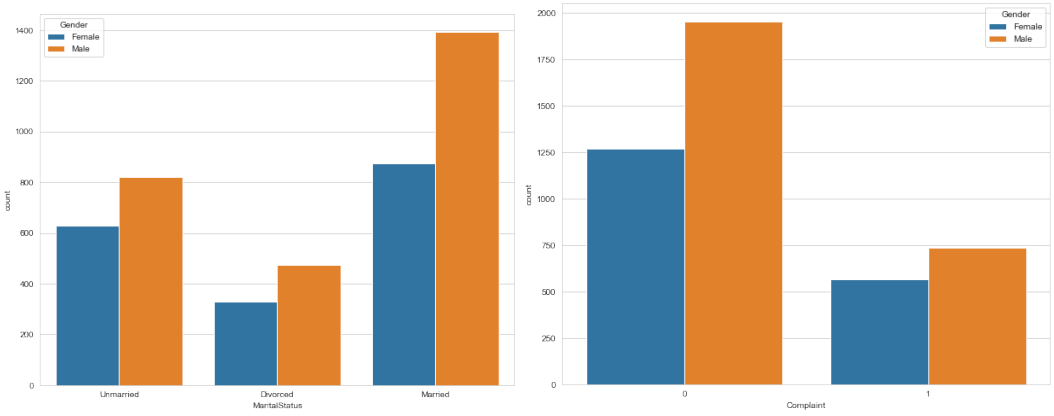

Figure 7

- As we can see most of the customers are Married followed by Unmarried and Divorced.
- Around 1300 customers have complained in the last one month.
- Around 550 Females customers and around 750 Male customers have complained in the last one month.

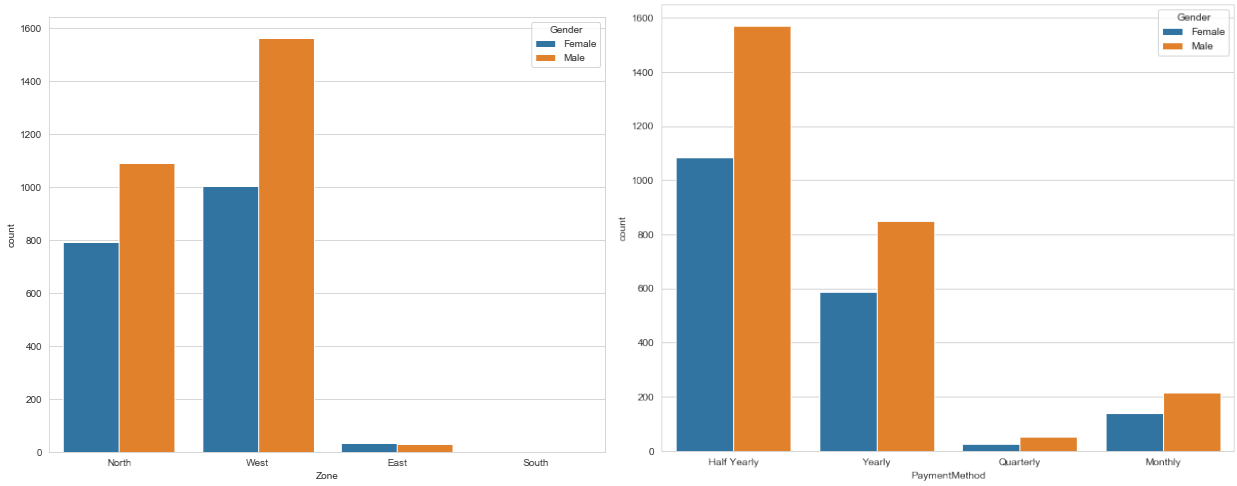

Figure 8

- We have more people that reside in the ‘West’ zone.
- Second highest is ‘North’ zone followed by ‘East’ and then ‘South’ which is the least.

- Most preferred Payment Method amongst the customers is ‘Half Yearly’.
- ‘Yearly’ is the second highest followed by ‘Monthly’ and then ‘Quarterly’.

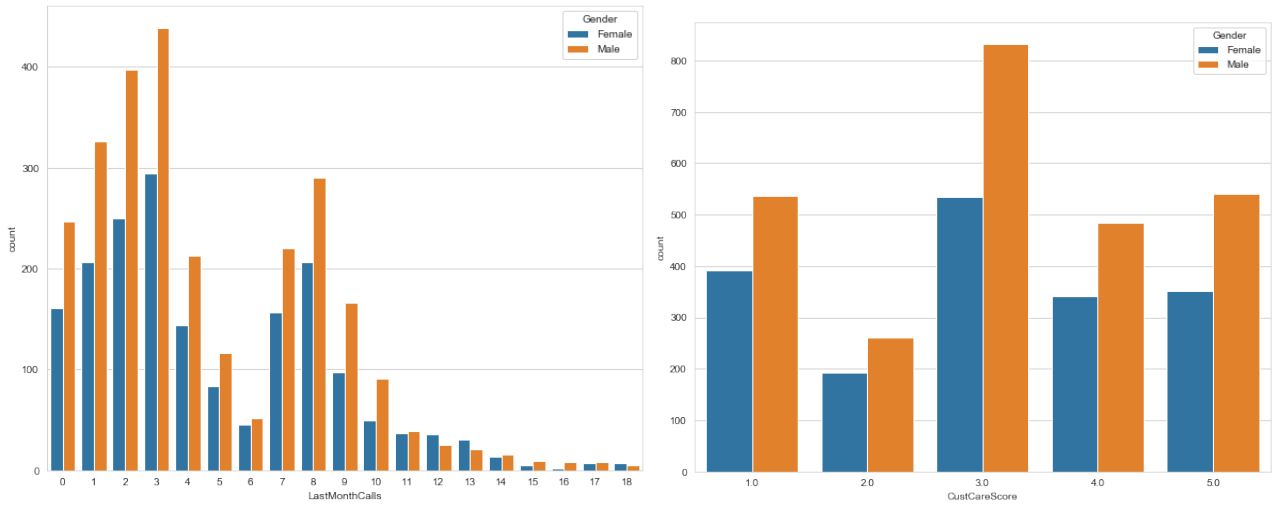

Figure 9

- Most number of calls generated for Cross Sell are generally 3 in the last one month.

-

- Most popular rating score given by customers is ‘3’ followed by ‘1’, ‘5’, ‘4’ and then ‘2’

BIVARIATE ANALYSIS-

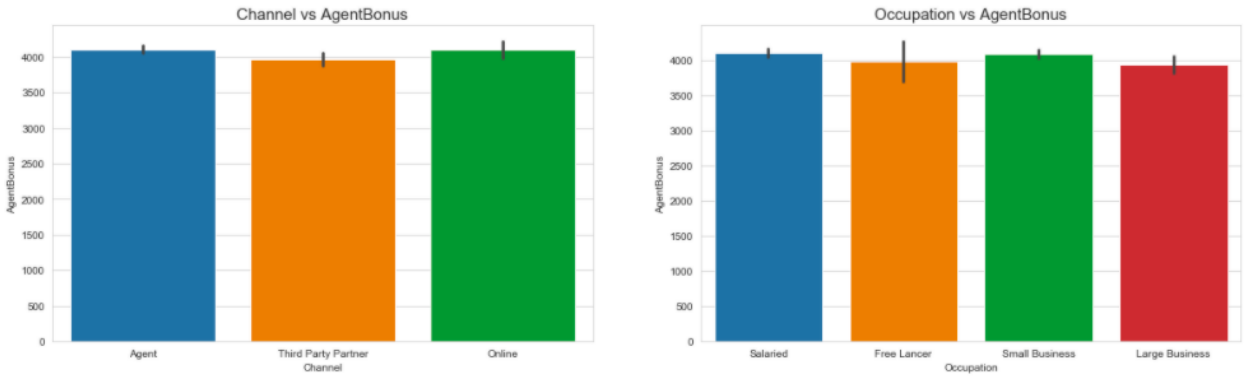

Figure 10

- Mean Agent Bonus is highest when the channel is ‘Agent’ followed by ‘Online’ and then ‘Third Party Partner’.
- In case of Occupation vs AgentBonus mean Agent Bonus is more ‘Salaried’ customers followed by ‘Small Business’, ‘FreeLancer’ and then ‘Large Business’.

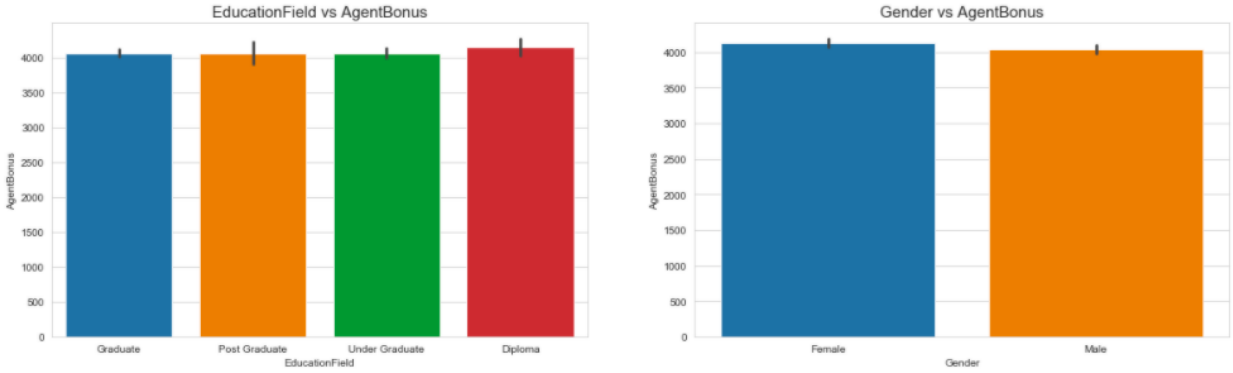

Figure 11

- AgentBonus is more when the customer is a ‘Diploma Holder’ and ‘Female’ customers have an edge over ‘Male’ customers.

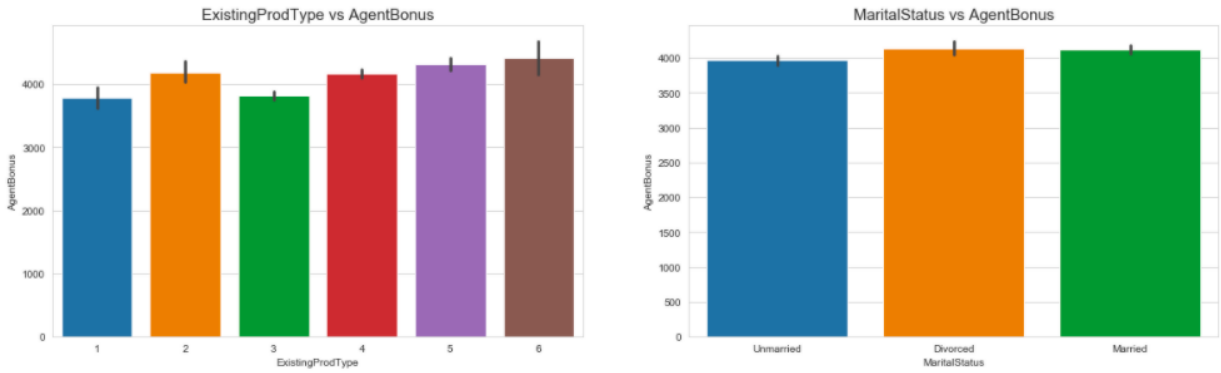

Figure 12

- Mean AgentBonus is highest for ExistingProdType ‘6’ and ‘5’ followed by ‘4’, ‘2’, ‘3’, ‘1’.
- The customers having marital status ‘Divorced’ and ‘Married’ can lead to more AgentBonus.

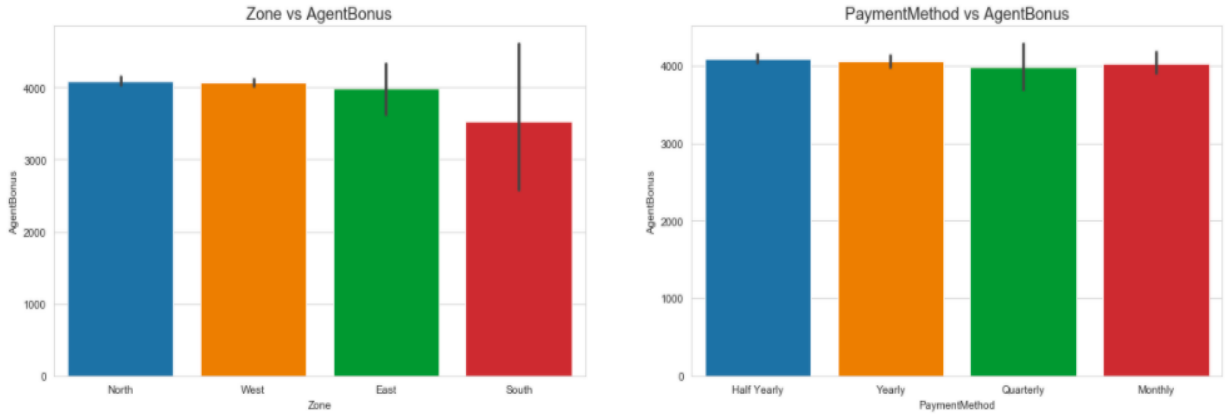

Figure 11

- The customers living in the ‘North’ zone and PaymentMethod ‘Half Yearly’ have edge over others.

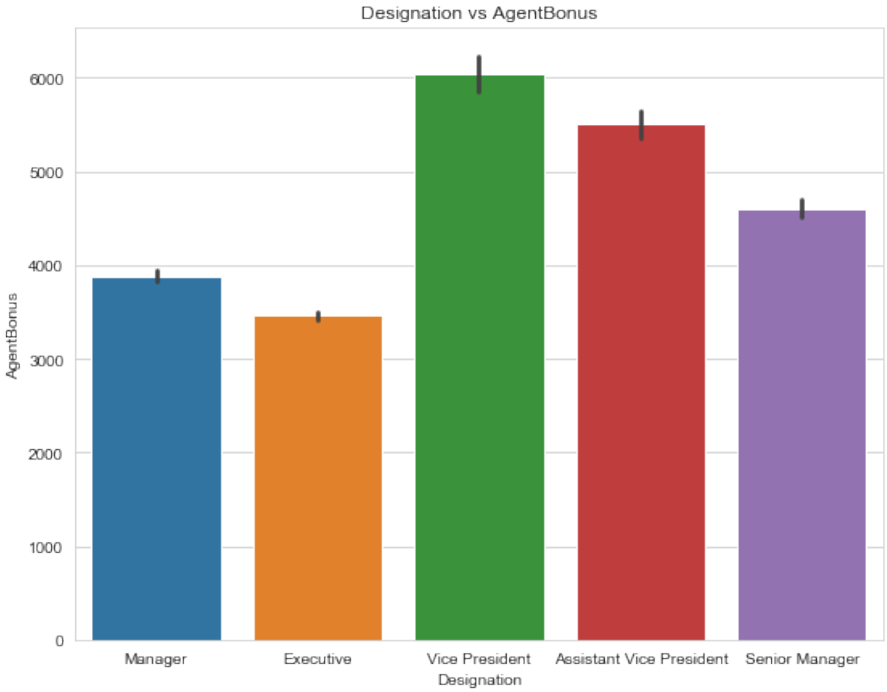

Figure 12

- The customer having Designation ‘Vice President’ can lead to more AgentBonus.
- Followed by ‘Assistant Vice President’, ‘Senior Manager’, ‘Manager’, and last is ‘Executive’,

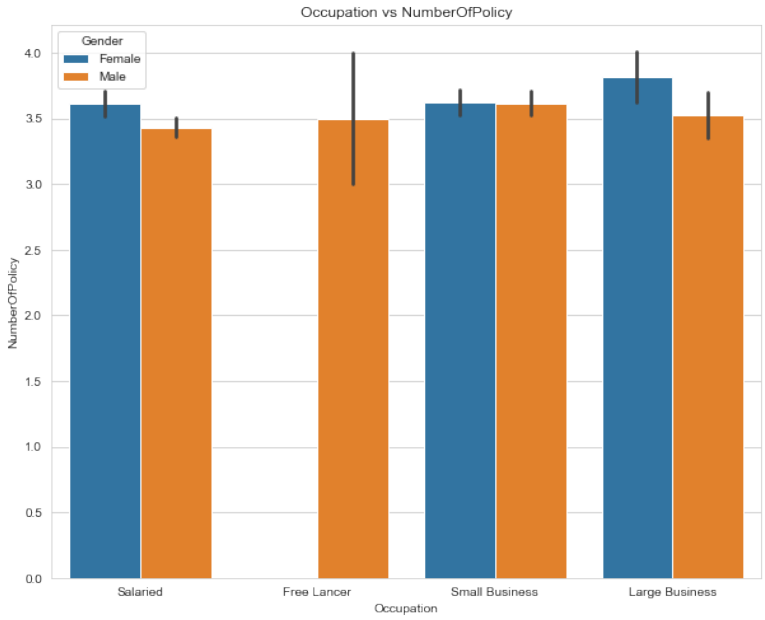

Figure 13

- Generally customers generally have existing 3 to 4 policies.

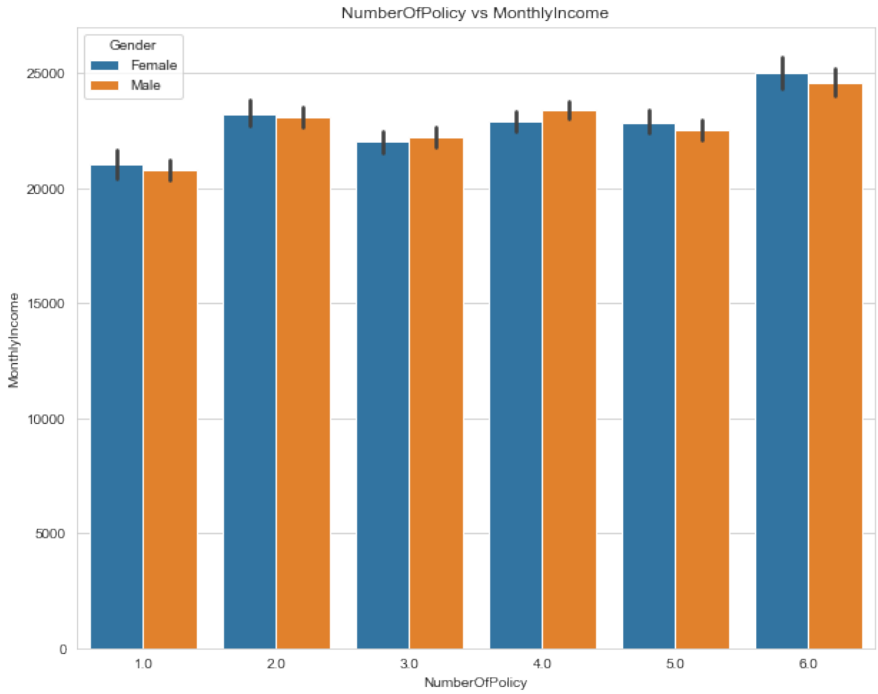

Figure 14

- We can see that customers having MonthlyIncome around 25000 have 6 existing policies.

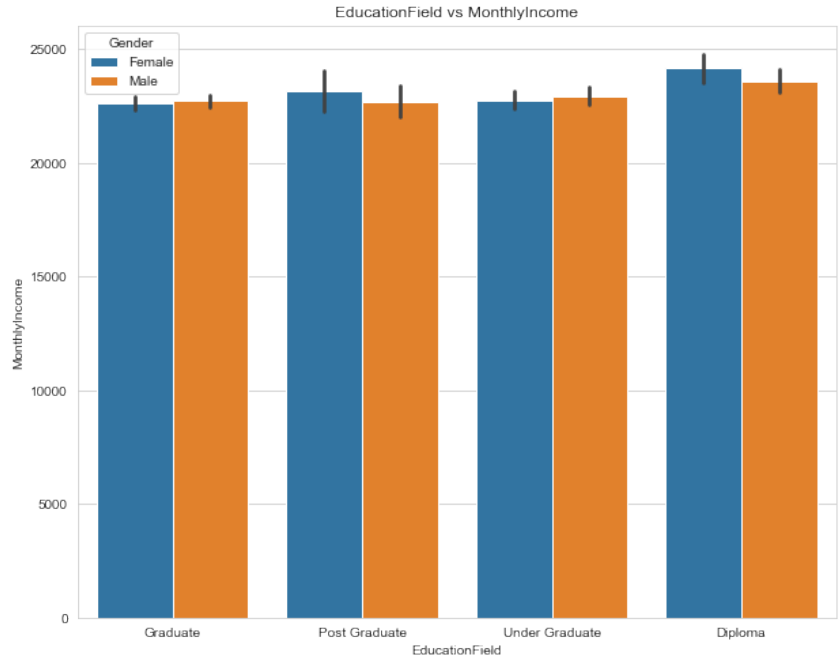

Figure 15

- Customers having ‘Diploma’ tend to earn more monthly as compared to others.

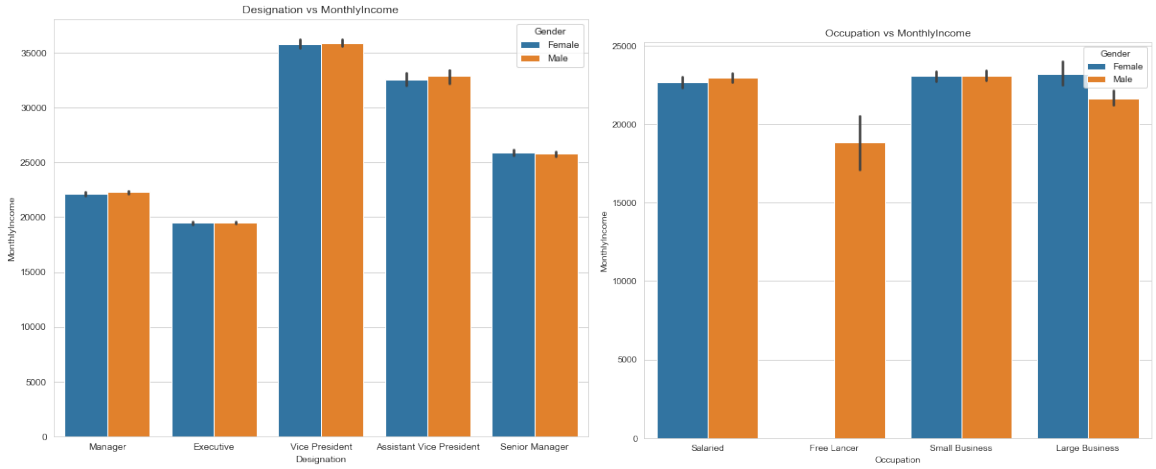

Figure 16

- Customers having Designation as 'Vice President' have the highest average MonthlyIncome followed by Assistant Vice President, Senior Manager, Manager and Executive.
- Average monthly income for 'Small Business' customers is highest followed by 'Salaried', 'Large Business' and 'Free Lancer' having lowest.

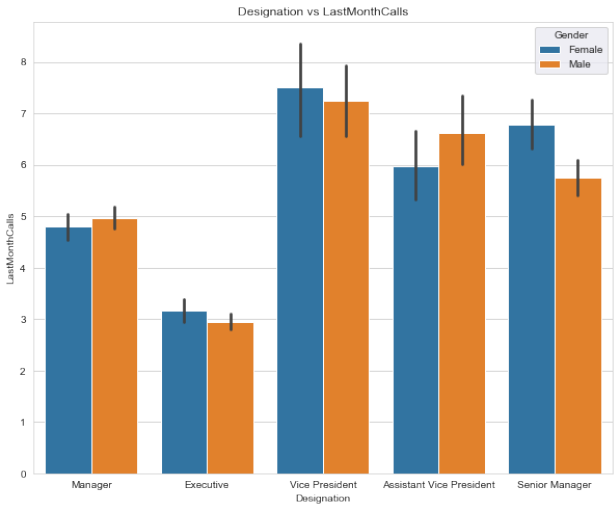

Figure 17

- Last Month Calls generated highest to the customers with Designation as 'Vice President' and lowest to 'Executives'.

MULTIVARIATE ANALYSIS -

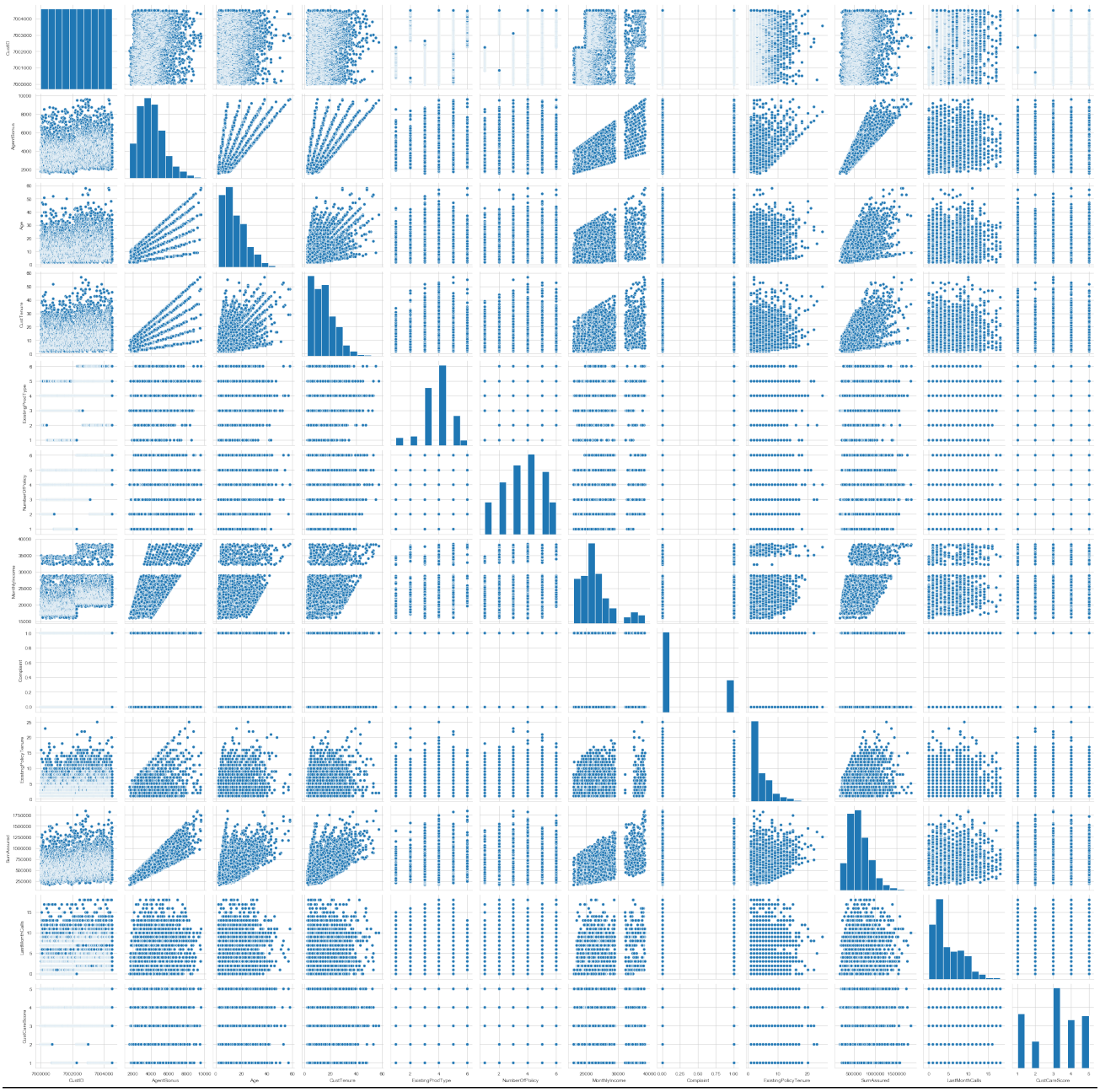

Figure 18

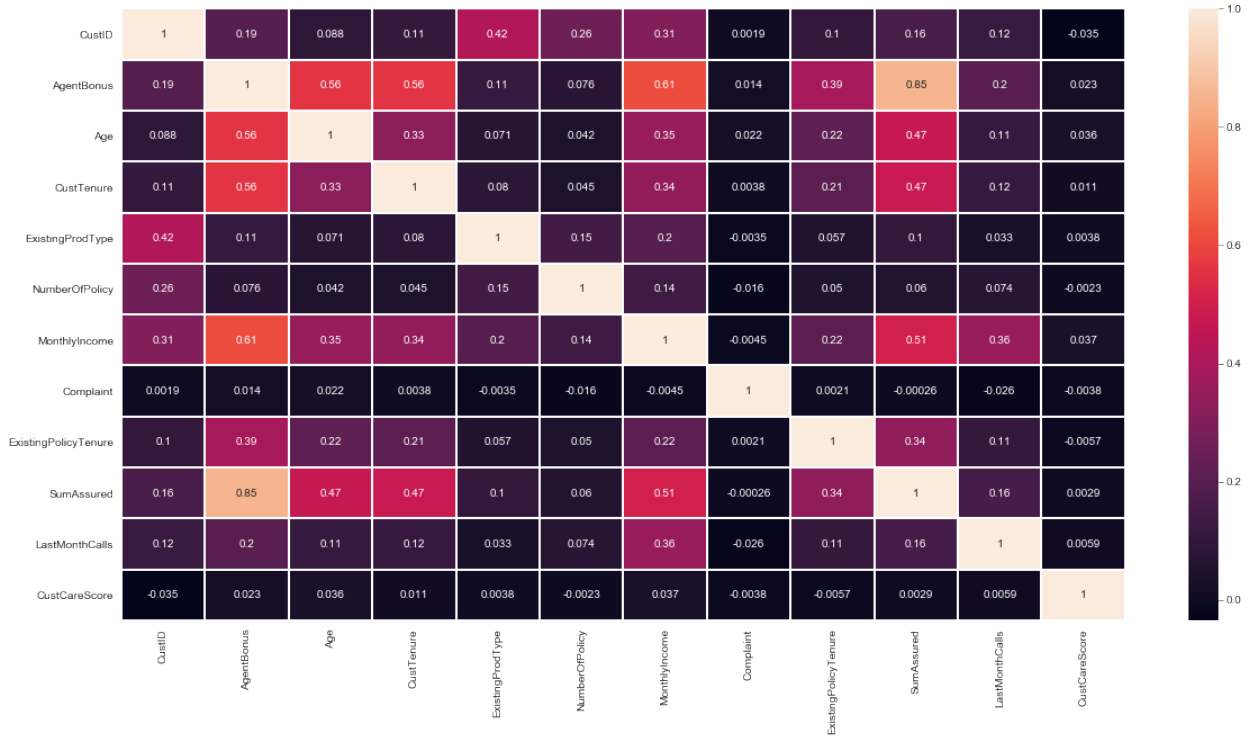

Figure 19

From Figure 18 and Figure 19 we can state:

- AgentBonus has a linear relationship with Age, CustTenure, MonthlyIncome, ExistingPolicyTenure and SumAssured.
- Age is linear related with CustTenure, MonthlyIncome, ExistingPolicyTenure and SumAssured.
- MonthlyIncome has some relation with SumAssured.
- SumAssured is highly correlated with AgentBonus as shown in heat map.
- AgentBonus has correlation with Age, CustTenure, MonthlyIncome and ExistingTenure.

DATA CLEANSING AND PREPROCESSING-

Customer ID is something that is not useful for model building. We can get rid of it.

Missing value and outliers treatment -

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

Table 3

As we can see from table above:

- Age has 269 missing values
- CustTenure has 226 missing values
- NumberOfPolicy has 45 missing values
- MonthlyIncome has 236 missing values
- ExistingPolicyTenure has 184 missing values
- SumAssured has 154 missing values
- CustCareScore has 54 missing values

Age	105
AgentBonus	100
Channel	0
Complaint	0
CustCareScore	0
CustTenure	97
Designation	0
EducationField	0
ExistingPolicyTenure	345
ExistingProdType	306
Gender	0
LastMonthCalls	12
MaritalStatus	0
MonthlyIncome	384
NumberOfPolicy	0
Occupation	0
PaymentMethod	0
SumAssured	110
Zone	0

Table 4

- Age has outliers the best value to impute missing value is median.
- CustTenure also has outliers we should impute missing values with median.
- NumberOfPolicy being a categorical variable we can use mode to impute missing values.
- MonthlyIncome being a continuous variable and having outliers median would be ideal to impute missing values.
- For SumAssured and ExistingPolicyTenure also we will use median.
- CustCareScore being categorical we will use mode to impute missing values.

In order to remove outliers we used a custom function which uses the upper and lower bound values of each variable to replace each outlier value. As we can see from the figure below all the outliers are removed.

Age	0
AgentBonus	0
Channel	0
Complaint	0
CustCareScore	0
CustTenure	0
Designation	0
EducationField	0
ExistingPolicyTenure	0
ExistingProdType	0
Gender	0
LastMonthCalls	0
MaritalStatus	0
MonthlyIncome	0
NumberOfPolicy	0
Occupation	0
PaymentMethod	0
SumAssured	0
Zone	0

Table 5

Variable Transformation-

We should convert all object data types to integer data types. For nominal variables we can encode the data through categorical codes.

For Ordinal variable 'Destination' we can either use label encoding or map custom values and order them accordingly.

For Designation variable we will rank the classes as shown below:

1. Executive (Lowest)
2. Manager
3. Senior Manager
4. Assistant Vice President
5. Vice President (Highest)

	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome	Complaint	ExistingPolicyTenure	SumAssured	Zone	PaymentMethod	LastMonthCalls	CustCareScore
0	4409.0	22.0	4.0	0	2	1	0	3.0	2	2.0	2	20993.0	1	2.0	806761.0	1	0	5.0	2.0
1	2214.0	11.0	2.0	2	2	1	1	4.0	2	4.0	0	20130.0	0	3.0	294502.0	1	3	7.0	3.0
2	4273.0	26.0	4.0	0	0	2	1	4.0	1	3.0	2	17090.0	1	2.0	578976.5	1	3	0.0	3.0
3	1791.0	11.0	13.0	2	2	1	0	3.0	1	3.0	0	17909.0	1	2.0	268635.0	3	0	0.0	5.0
4	2955.0	6.0	13.0	0	3	3	1	3.0	1	4.0	0	18468.0	0	4.0	366405.0	3	0	2.0	5.0

Table 6

Above table shows the encoded data and all variables are now numeric in nature.

MODEL BUILDING-

The dataset provided is the dataset of Insurance Sales. We need to analyse the data and predict the amount of bonus given to the agents based on the customer information provided in the dataset.

It is a regression problem. So for this model , we will be using different regression models like , linear regression , random forest regressor , SVM regressor and some of the regression and ensemble technique and tuning methods.

Further , we will see models with their interpretation. Before that we will split the data into x and y with regards to target variable.

	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome	Complaint
0	22.0	4.0	0	2	1	0	3.0	2	2.0	2	20993.0	1
1	11.0	2.0	2	2	1	1	4.0	2	4.0	0	20130.0	0
2	26.0	4.0	0	0	2	1	4.0	1	3.0	2	17090.0	1
3	11.0	13.0	2	2	1	0	3.0	1	3.0	0	17909.0	1
4	6.0	13.0	0	3	3	1	3.0	1	4.0	0	18468.0	0

Table 7

Checking the data in X after dropping the target variable.

	AgentBonus
0	4409.0
1	2214.0
2	4273.0
3	1791.0
4	2955.0

Table 8

After splitting the data in train and test , let's look at the data of X_train and X_test respectively.

	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome	Compla
2461	12.0	16.0	2	1	1	1	4.0	1	3.0	2	20742.0	
3681	31.0	15.0	0	3	3	1	4.0	2	5.0	1	23398.0	
1309	15.0	6.0	0	2	1	1	3.0	1	1.0	2	16232.0	
4254	5.0	16.0	1	1	1	1	4.0	2	2.0	1	23536.0	
1335	8.0	17.0	0	2	1	1	1.5	1	1.0	1	17269.0	
...	...	...	...	...	...	...	...	...	...	...	...	
2895	6.0	10.0	0	2	1	1	5.0	1	5.0	0	21658.0	
2763	14.0	5.0	0	3	3	1	4.0	1	4.0	2	20976.0	
905	12.0	12.0	0	3	3	1	3.0	2	1.0	1	19285.0	
3980	5.0	5.0	0	2	1	1	4.0	3	5.0	1	21606.0	
235	14.0	16.0	1	2	1	1	3.0	1	2.0	0	17097.0	

3164 rows × 18 columns

Table9

Although Scaling is not required in linear and trees based models and is optional. Here we can do scaling as it is optional in these models.

Lets check the data set after applying Z score scaling-

	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome
2461	-0.267690	0.204927	1.916737	-2.117232	-0.562649	0.841679	0.331647	-0.963113	-0.392079	1.246740	-0.448526
3681	1.988842	0.085165	-0.610161	1.030757	1.336254	0.841679	0.331647	-0.074172	0.984003	-0.195031	0.224800
1309	0.088604	-0.992686	-0.610161	-0.543237	-0.562649	0.841679	-0.738008	-0.963113	-1.768160	1.246740	-1.591863
4254	-1.099044	0.204927	0.653288	-2.117232	-0.562649	0.841679	0.331647	-0.074172	-1.080120	-0.195031	0.259785
1335	-0.742750	0.324688	-0.610161	-0.543237	-0.562649	0.841679	-2.342492	-0.963113	-1.768160	-0.195031	-1.328971
...	...	...	...	...	...	...	...	...	...	...	...
2895	-0.980279	-0.513641	-0.610161	-0.543237	-0.562649	0.841679	1.401303	-0.963113	0.984003	-1.636801	-0.216310
2763	-0.030160	-1.112448	-0.610161	1.030757	1.336254	0.841679	0.331647	-0.963113	0.295962	1.246740	-0.389205
905	-0.267690	-0.274119	-0.610161	1.030757	1.336254	0.841679	-0.738008	-0.074172	-1.768160	-0.195031	-0.817892
3980	-1.099044	-1.112448	-0.610161	-0.543237	-0.562649	0.841679	0.331647	0.814769	0.984003	-0.195031	-0.229492
235	-0.030160	0.204927	0.653288	-0.543237	-0.562649	0.841679	-0.738008	-0.963113	-1.080120	-1.636801	-1.372575

3164 rows × 18 columns

Table 10

1. Linear Regression Model -

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

Y = m(x) + C

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

Lets build the model for the given dataset -

```
regression_model = LinearRegression()
regression_model.fit(X_train, y_train)
```

— Lets see coefficients for each of the independent attributes -

The coefficient for Age is 0.13507381839546426
The coefficient for CustTenure is 0.13994384387200665
The coefficient for Channel is 0.0023134936612698664
The coefficient for Occupation is -0.004162840382476876
The coefficient for EducationField is 0.004565606461568633
The coefficient for Gender is 0.011964620419456578
The coefficient for ExistingProdType is -0.0054040127387138094
The coefficient for Designation is 0.093479798382732
The coefficient for NumberOfPolicy is 0.015671499908285647
The coefficient for MaritalStatus is 0.003963223397227654
The coefficient for MonthlyIncome is 0.115170024289234
The coefficient for Complaint is 0.012456292050586773
The coefficient for ExistingPolicyTenure is 0.08180788461972482
The coefficient for SumAssured is 0.6060471968821479
The coefficient for Zone is -0.000965376115857388
The coefficient for PaymentMethod is -0.005765434645315492
The coefficient for LastMonthCalls is -0.006201366114818894
The coefficient for CustCareScore is 0.011341976143869 56

The intercept for our model is 9.296261339145077e-17

The constant term in regression analysis is the value at which the regression line crosses the y-axis. The constant is also known as the y-intercept.

Coefficient of Determinant for train and test -

	Linear Regression Train	Linear Regression Test
Coefficient of Determinant	0.801	0.789

Table 11

Checking Intercept and coefficient Values -

```
Intercept               1.214306e-16
Age                     1.350738e-01
CustTenure              1.399438e-01
Channel                 2.313494e-03
Occupation             -4.162840e-03
EducationField          4.565606e-03
Gender                  1.196462e-02
ExistingProdType       -5.404013e-03
Designation             9.347980e-02
NumberOfPolicy          1.567150e-02
MaritalStatus           3.963223e-03
MonthlyIncome           1.151700e-01
Complaint               1.245629e-02
ExistingPolicyTenure    8.180788e-02
SumAssured              6.060472e-01
Zone                   -9.653761e-04
PaymentMethod          -5.765435e-03
LastMonthCalls         -6.201366e-03
CustCareScore           1.134198e-02
dtype: float64
```

Table 12

Lets check for summary table -

OLS Regression Results

Dep. Variable:	AgentBonus	R-squared:	0.801
Model:	OLS	Adj. R-squared:	0.800
Method:	Least Squares	F-statistic:	702.9
Date:	Wed, 17 Mar 2021	Prob (F-statistic):	0.00
Time:	17:52:32	Log-Likelihood:	-1936.1
No. Observations:	3164	AIC:	3910.
Df Residuals:	3145	BIC:	4025.
Df Model:	18		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.214e-16	0.008	1.53e-14	1.000	-0.016	0.016
Age	0.1351	0.009	14.784	0.000	0.117	0.153
CustTenure	0.1399	0.009	15.346	0.000	0.122	0.158
Channel	0.0023	0.008	0.290	0.772	-0.013	0.018
Occupation	-0.0042	0.009	-0.461	0.645	-0.022	0.014
EducationField	0.0046	0.009	0.506	0.613	-0.013	0.022
Gender	0.0120	0.008	1.495	0.135	-0.004	0.028
ExistingProdType	-0.0054	0.010	-0.542	0.588	-0.025	0.014
Designation	0.0935	0.014	6.479	0.000	0.065	0.122
NumberOfPolicy	0.0157	0.008	1.902	0.057	-0.000	0.032
MaritalStatus	0.0040	0.008	0.493	0.622	-0.012	0.020
MonthlyIncome	0.1152	0.015	7.799	0.000	0.086	0.144
Complaint	0.0125	0.008	1.561	0.119	-0.003	0.028
ExistingPolicyTenure	0.0818	0.008	9.765	0.000	0.065	0.098
SumAssured	0.6060	0.010	58.258	0.000	0.586	0.626
Zone	-0.0010	0.008	-0.121	0.904	-0.017	0.015
PaymentMethod	-0.0058	0.009	-0.612	0.541	-0.024	0.013
LastMonthCalls	-0.0062	0.009	-0.719	0.472	-0.023	0.011
CustCareScore	0.0113	0.008	1.415	0.157	-0.004	0.027

Omnibus:	123.874	Durbin-Watson:	1.993
Prob(Omnibus):	0.000	Jarque-Bera (JB):	139.107
Skew:	0.484	Prob(JB):	6.21e-31
Kurtosis:	3.342	Cond. No.	4.30

Table 13

Let's check for MSE -

	Linear Regression Train	Linear Regression Test
MSE	0.446	0.459

Table 14

Checking the linear graph w.r.t target variable -

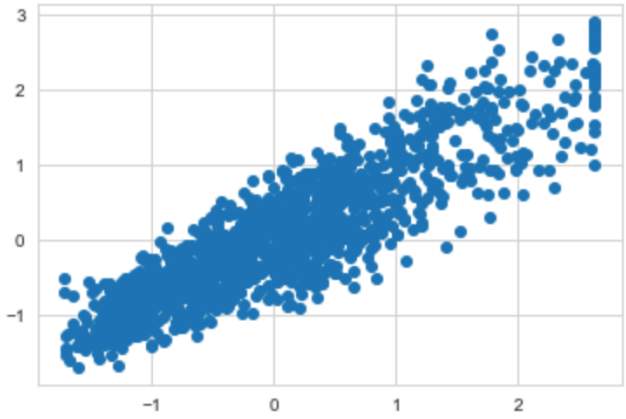

Figure 20

Interpretation of the Model -

(0.0) * Intercept + (0.14) * Age + (0.14) * CustTenure + (0.0) * Channel + (-0.0) * Occupation + (0.0) * EducationField + (0.01) * Gender + (-0.01) * ExistingProdType + (0.09) * Designation + (0.02) * NumberOfPolicy + (0.0) * MaritalStatus + (0.12) * MonthlyIncome + (0.01) * Complaint + (0.08) * ExistingPolicyTenure + (0.61) * SumAssured + (-0.0) * Zone + (-0.01) * PaymentMethod + (-0.01) * LastMonthCalls + (0.01) * CustCareScore

From the above linear equation we got we can say that -

1. Age , monthly income , CustTenure are some of the important factors that can be checked while deciding the Agent Bonuses.
2. There are some negative constants as well which indicates that payment method , zone these kind of variable should not be considered while deciding the agent bonus.

BEST MODEL -

In the model building exercise , we have built various regression models like linear , random forest regressor and SVM regressor.

We shall be choosing Linear regression model as our best fit model as the problem we are dealing with is a regression problem and linear regression is giving best score on train and test data set. Keeping in mind the concepts of under and overfitting , Linear regression model would be our best fit model for this problem.

Through this model it will help the business to determine the variables that needs to be taken into consideration for deciding the Agent Bonus and those as well who need training and up skill program.

MODEL TUNING/PERFORMANCE -

As part of model tuning, Ridge regression is a good option as it will reduce multicollinearity and it will not take those features into account which have less importance.

```
from sklearn.linear_model import Ridge

r_model = Ridge()
r_model.fit(X_train , y_train)
```

Lets check the R2 score - 0.7893899153598954

Conclusion -

The **r2 score** varies between 0 and 100%. It is closely related to the **MSE.** So if it is 100%, the two variables are perfectly correlated, i.e., with no variance at all. A low value would show a low level of correlation, meaning a regression model that is not valid, but not in all cases.

In part of model tuning exercise , we have seen that R2 score is low for Ridge Regression as compared to Ensemble technique Bagging Regressor.

So we can conclude here that Ridge Regression can be a better option for model tuning.

MODEL VALIDATION -

As part of model validation MSE and accuracy were taken into account. The smaller the means squared error, the closer you are to finding the line of best fit.

	Linear Regression Train	Linear Regression Test
Coefficient of Determinant	0.801	0.789

	Linear Regression Train	Linear Regression Test
MSE	0.446	0.459

Table 14

OTHER MODELS -

Random Forest Regressor -

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the `max_samples` parameter if `bootstrap=True` (default), otherwise the whole dataset is used to build each tree.

regressor = RandomForestRegressor(n_estimators=10 , random_state=0)
regressor.fit(X_train, y_train)

After training the model lets check the model score on train and test data -

	RF_Train	RF_Test
Score	0.971	0.789

Table 15

Model is performing good on train data and test data we can say.

Lets check the importance features , which can be helpful in deciding Agent Bonus -

	Imp
SumAssured	0.732008
Age	0.080762
CustTenure	0.060050
MonthlyIncome	0.037789
Designation	0.021688
ExistingPolicyTenure	0.013196
LastMonthCalls	0.012001
NumberOfPolicy	0.006980
CustCareScore	0.006520
Channel	0.004157
ExistingProdType	0.003744
EducationField	0.003568
MaritalStatus	0.003358
Occupation	0.003153
PaymentMethod	0.003100
Zone	0.002946
Gender	0.002822
Complaint	0.002157

Table 16

1. Sum Assured , Age , CustTenure , Monthly Income , Designation are the top 5 features which can be considered.
2. Zone , Gender , Occupation are some of the features which carry little less weightage.

SVM Regressor -

Support Vector Machines (SVM) are popularly and widely used for classification problems in machine learning.
But in regression cases SVM tries to find a line/hyperplane (in multidimensional space) that separates two classes. Then it classifies the new point depending on whether it lies on the positive or negative side of the hyperplane depending on the classes to predict.

SV_regressor = SVR(kernel='rbf')
SV_regressor.fit(X_train, y_train)

After training the model lets check the model score on train and test data -

	SV_Train	SV_Test
Score	0.79	0.895

Table 17

Model is performing good on train data and test data we can say.

Some Points on all the models used -

1. Linear regression is the most fit model for this problem as the prediction is better we can see though MSE.
2. SVM regressor has shown better results as compared to Random forest regressor.
3. Random regressor is a kind of ensemble technique which also tells the important features that can be considered and as per that we can tune the model for better results.

Let's compare the overall scores of our main models -

	Linear Regression Train	Linear Regression Test	Random Forest Train	Random Forest Test	SVM Train	SVM Test
Score	0.801	0.789	0.971	0.789	0.79	0.895

Table 18

All the regressor models (linear regression , Random forest regressor , SVM regressor) are performing good. But we can see Linear and SVM.

We can consider Linear regression as the most optimum model as explained above.

INTERPRETATIONS -

Through the various models built and our main model, following insights were drawn -

1. Variables like Age , complaints , Existing policy tenure , product type are some of the important factors which can be taken into consideration for agent bonus.

2. Gender , Zone , payment method are those variables in dataset which are of least importance while deciding agent bonus and agents up skill programs.

3. Designations , Education field , last month calls are secondary variable for deciding the agent bonuses.

4. Agent who have more complaints and less customer score should be given trainings and up skill programs.

RECOMMENDATIONS-

1. From the analysis it is recommended that company should target those agents which have good customer tenure and max sum assured in the existing polices.

2. Organisation should consider those agents as well who are associated with company from a very long time and are at a senior position.

3. Those agents having good existing policy tenure can be categorised under good performing agents.

4. Agents having more and frequent complaints can be categorised under low performing agents.

5. Company should target those agents who have less customer score and they can be considered for trainings/up skill programs.

6. Last month calls , Education field ,designation can also be considered for deciding bonuses and trainings within the company.

7. Gender , zone , payment methods are some of the variables which doesn't play major role in deciding bonuses and trainings within the organisation.