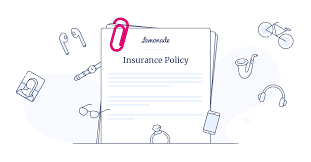
**CAPSTONE PROJECT**

**LIFE INSURANCE SALES**



**Table of contents:**

|  |  |  |
| --- | --- | --- |
| S.no | Content | Pg. no |
|  | [**Introduction of the business problem**](#Introduction) | 4 |
|  | [Need of the study/project](#Needofstudy) | 4 |
|  | [Visual inspection of data (rows, columns, descriptive details)](#Vis) | 5 |
|  | [Statistics of the data (Descriptive Statistics):](#stat) | 6 |
|  | [**Data Cleaning and Pre-processing**](#data_clean) | 8 |
|  | [Understanding of attributes](#understand) | 8 |
|  | [Variable transformation](#variable) | 9 |
|  | [Removal of unwanted variable](#Removal) | 9 |
|  | [Null values Treatment](#Null) | 9 |
|  | [Outlier Treatment](#outlier) | 10 |
|  | [**EDA and Business Implication**](#EDA) | 13 |
|  | [Insights](#insights1) | 16 |
|  | [Bi variate analysis](#bi_variate) | 17 |
|  | [Insights](#insights) | 23 |
|  | [**Model Building**](#Model_building) | 25 |
|  | [Why these models](#why_these) | 26 |
|  | [Clustering model](#cluster) | 29 |
|  | [**Model Validation**](#model_val) | 30 |
|  | [**Recommendation**](#recom) | 31 |

**List of Figures:**

|  |  |  |
| --- | --- | --- |
| **S.no** | **Figures** | **Pg. no** |
|  | Outliers in Insurance data before treatment | 11 |
|  | Outliers in Insurance data after treatment | 12 |
|  | Univariate Analysis | 13 |
|  | Bi-variate analysis | 17 |
|  | Age vs Agent Bonus (average | 18 |
|  | Agent Bonus vs Existing Policy Tenure vs Sum Assured | 19 |
|  | Customer Tenure vs Agent Bonus vs Gender | 19 |
|  | Gender vs Marital vs Agent Bonus | 20 |
|  | Gender vs Customer Tenure vs Agent Bonus | 20 |
|  | Gender vs Number of Policies vs Agent Bonus | 21 |
|  | Monthly Income vs Number of policies vs Sum Assured | 21 |
|  | Customer Care Score vs Last Month Calls vs Agent Bonus vs Complaints | 22 |
|  | Occupation vs Education field vs Agent Bonus | 22 |
|  | Models used | 25 |
|  | Clustering Model | 29 |

**List of tables:**

|  |  |  |
| --- | --- | --- |
|  | Basic information of Insurance data | 5 |
|  | Descriptive Statistics of Insurance data | 6 |
|  | Null values in the Insurance data before treatment | 9 |
|  | Null values in the Insurance data after treatment | 10 |
|  | MSE and R square of all the Models | 27 |
|  | MSE values after Cross-validation | 27 |
|  | Ensemble Learning for Random Forest | 28 |

**Introduc****tion**

The leading life insurance company wants to predict the bonus for its agents. This prediction is crucial in designing appropriate engagement activity for high-performing agents and upskilling programs for low-performing agents. By accurately estimating the bonus amounts, the company can allocate resources effectively and implement targeted strategies to enhance agent performance. These agents play a critical role in acquiring new clients, maintaining customer relationships, and driving sales. This, in turn, can lead to improved customer service, increased sales, and overall business growth.

The objective is to develop a machine learning model that considers historical agent performance data, customer satisfaction ratings, sales targets, training hours completed, and other relevant features to predict the bonus amount for each agent.

**Need of study:**

**Performance Incentives:** High-performing agents can be appropriately rewarded and allocated performance incentives, which help boost their morale, motivation, and job satisfaction.

**Resource Allocation:** By identifying high-performing agents, the company can invest in engagement activities, training, allocating its resources effectively, and rewards.

**Performance Evaluation:** The model's predictions can serve as a valuable tool for performance evaluation and be compared against actual bonus amounts to evaluate the effectiveness of the performance management system and identify areas for improvement.

**Business Growth:** Satisfied and motivated agents are more likely to deliver exceptional customer service, resulting in increased customer satisfaction, retention, and positive word-of-mouth referrals. Moreover, by improving the performance of low-performing agents through upskill programs, the company can maximize its sales potential and market competitiveness.

**Data-Driven Decision-Making:** The project promotes data-driven decision-making within the life insurance company. By leveraging historical agent performance data and other relevant features, the model provides insights into the factors influencing bonus amounts.

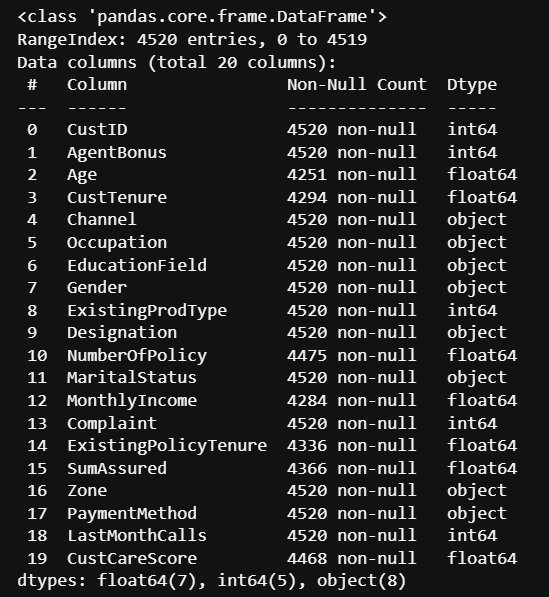
**Visual inspection of data (rows, columns, descriptive details)**

1. **Shape of Data:** The dataset contains 4520 rows and 20 columns.

2. **Data Types:** The dataset consists of various data types, including int64 and float64 for numerical variables, and object for categorical variables.

3. **Missing Values:** Some columns have missing values. For example, the 'Age' column has 4251 non-null values, 'Cust Tenure' has 4294 non-null values, 'Monthly Income' has 4284 non-null values, and so on.

Table 1: Basic information of Insurance data



4. **Column Descriptions:** The dataset contains several columns with different names and data types. For example, 'CustID' is an integer column representing customer IDs, 'Agent Bonus' is an integer column representing agent bonuses, 'Channel' is an object column representing the channel through which customers were acquired, and so on.

5. **Categorical Variables:** There are several categorical variables in the dataset, such as 'Channel', 'Occupation', 'Education Field', 'Gender', 'Designation', 'Marital Status', 'Zone', and 'Payment Method'. These variables may require encoding or conversion into numerical form for certain analysis or modeling techniques.

6. **Numerical Variables:** The dataset also includes numerical variables like 'Age', 'Cust Tenure', 'NumberOfPolicy', 'MonthlyIncome', 'Complaint', 'ExistingPolicyTenure', 'SumAssured', 'LastMonthCalls', and 'CustCareScore'. These variables represent different aspects of customers and their interactions with the life insurance company.

**Statistics of the data (Descriptive Statistics):**

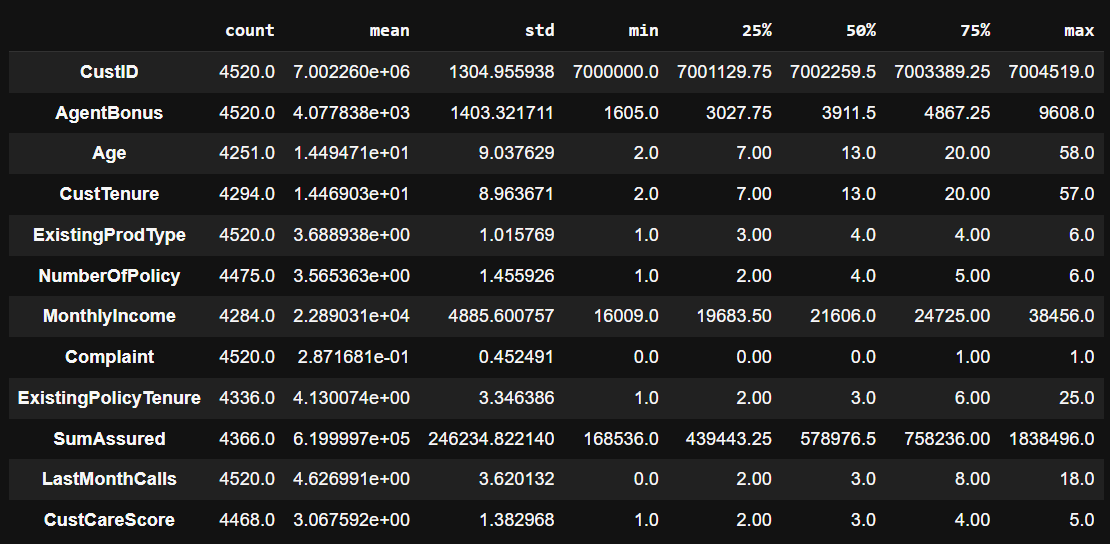
1. **Customer ID:**

* The CustID values range from 7000000 to 7004519, indicating a sequential identifier for customers.

1. **Agent Bonus:**

* The average Agent Bonus is around 4077.838, with a standard deviation of 1403.322.
* The minimum Agent Bonus is 1605, and the maximum is 9608.

Table 2: Descriptive Statistics of Insurance data



1. **Age:**

* The average age of customers is approximately 14.495 years, with a standard deviation of 9.038.
* The minimum age is 2, and the maximum age is 58.
* The age distribution appears to be positively skewed, with the median age at 13.

1. **Customer Tenure:**

* The average customer tenure is around 14.469 months, with a standard deviation of 8.964.
* The minimum tenure is 2 months, and the maximum is 57 months.
* The distribution seems to be positively skewed, with the median tenure at 13 months.

1. **Existing Product Type:**

* The Existing Product Type values range from 1 to 6, with an average of 3.689 and a standard deviation of 1.016.
* The majority of customers have an Existing Product Type of 3 or 4

1. **Number of Policy:**

* The average number of policies held by customers is approximately 4 (rounded), with a standard deviation of 1.456.
* The minimum number of policies is 1, and the maximum is 6.
* The distribution seems to be slightly positively skewed, with the median at 4 policies.

1. **Monthly Income:**

* The average monthly income of customers is around 22890.311, with a standard deviation of 4885.601.
* The minimum income is 16009, and the maximum is 38456.
* The income distribution appears to be positively skewed, with the median income at 21606.

1. **Complaint:**

* The Complaint variable takes binary values (0 or 1), indicating whether a customer has made a complaint or not.
* Approximately 28.72% of customers have made a complaint, as indicated by the mean.

1. **Existing Policy Tenure:**

* The average tenure of existing policies is approximately 4 years, with a standard deviation of 3.346.
* The minimum tenure is 1 year, and the maximum is 25 years.
* The distribution seems to be positively skewed, with the median tenure at 3 years.

1. **Sum Assured:**

* The average sum assured is approximately 619999.7, with a standard deviation of 246234.822.
* The minimum sum assured is 168536, and the maximum is 1838496.
* The distribution appears to be positively skewed, with the median sum assured at 578976.5.

1. **Last Month Calls:**

* The average number of calls made by customers in the last month is approximately 5, with a standard deviation of 3.620.
* The minimum number of calls is 0, and the maximum is 18.
* The distribution appears to be positively skewed, with the median at 3 calls.

1. **Customer Care Score:**

* The average customer care score is around 3, with a standard deviation of 1.383
* The minimum score is 1, and the maximum is 5.
* The majority of customer care scores fall between 2 and 4, as indicated by the 25th and 75th percentiles.

**Data** **Cleaning and Pre-processing**

**Understanding of attributes (variable info, renaming if required)**

**AgentBonus Unique values are :** [4409. 2214. 4273. ... 4377. 2541. 3953.]

**Age Unique values are :** [22. 11. 26. 6. 7. 12. 8. 20. 18. 10. 9. 5. 30. 14.

16. 13. 2. 4. 15. 27. 23. 37. 33. 19. 17. 25. 21. 24. 29. 39.5 31. 28. 3. 35. 32. 34. 39. 38. 36. ]

**CustTenure Unique values are :**[ 4. 2. 23. 11. 3. 7. 27. 22. 17. 5. 21. 18. 25. 14.

8. 15. 31. 9. 26. 16. 10. 6. 20. 24. 30. 12. 38. 29.

39. 28. 19. 13. 39.5 36. 32. 34. 35. 33. 37. ]

**Channel Unique values are :** ['Agent' 'Third Party Partner' 'Online']

**Occupation Unique values are** : ['Salaried' 'Free Lancer' 'Small Business' 'Laarge Business'

'Large Business']

**EducationField Unique values are :** ['Graduate' 'Post Graduate' 'UG' 'Under Graduate' 'Engineer' 'Diploma' 'MBA']

**Gender Unique values are :** ['Female' 'Male' 'Fe male']

**ExistingProdType Unique values are :**  [3. 4. 2. 5. 1.5 5.5]

**Designation Unique values are :** ['Manager' 'Exe' 'Executive' 'VP' 'AVP' 'Senior Manager']

**NumberOfPolicy Unique values are :**  [2. 4. 3. 5. 1. 6.]

**MaritalStatus Unique values are :** ['Single' 'Divorced' 'Unmarried' 'Married']

**MonthlyIncome Unique values are:** [20993. 20130. 17090. ... 21213. 28004. 21250.]

**Complaint Unique values are :** [1. 0.]

**ExistingPolicyTenure Unique values are :** [ 2. 3. 4. 1. 5. 6. 12. 7. 8. 9. 10. 11.]

**SumAssured Unique values are :** [806761. 294502. 268635. ... 667371. 943999. 700308.]

**Zone Unique values are :** ['North' 'West' 'East' 'South']

**PaymentMethod Unique values are :** ['Half Yearly' 'Yearly' 'Quarterly' 'Monthly']

**LastMonthCalls Unique values are :**

[ 5. 7. 0. 2. 6. 9. 3. 1. 8. 4. 13. 15. 12. 10. 11. 17. 16. 14.]

**CustCareScore Unique values are :** [2. 3. 5. 4. 1.]

**Variable transformation:**

Here the object type data types are converted to Categorical data type to make the further process convenient

The above highlighted are to be changed – [‘Fe male’ to ‘Female’, ‘UG’ to ‘Under Graduation’, ‘Exe’ to ‘Executive’, ‘Laarge Business’ to ‘Large Business’]

**Removal of unwanted variables**

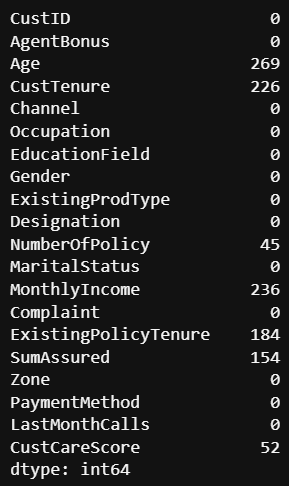
In the data set for further analysis, we can drop the customer id after checking for duplicates.

In the data, there are no duplicates. Customer Id can be removed.

**Null Value Treatment:**

There are few null values in the data.

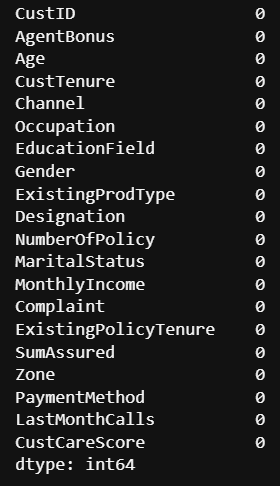
Table 3: Null values in the Insurance data before treatment



From the above table we could see that Age, CustTenure, NumberOfPolicy, MonthlyIncome, ExistingPolicyTenure, SumAssured, CustCareScore.

These null values are filled using using back fill, front fill methods.Table 4: Null values in the Insurance data after treatment

Table 4: Null values in the Insurance data before treatment



**Outlier Treatment:**

The lower range and upper range for a given column using the Interquartile Range (IQR) method are calculated. The IQR is a measure of statistical dispersion and is used to identify outliers.

The outlier removal is done by capping the values above the upper range and below the lower range for each numeric column. This helps in identifying any remaining outliers and understanding the distributional characteristics of the data.

Figure 1: Outliers in Insurance data before treatment

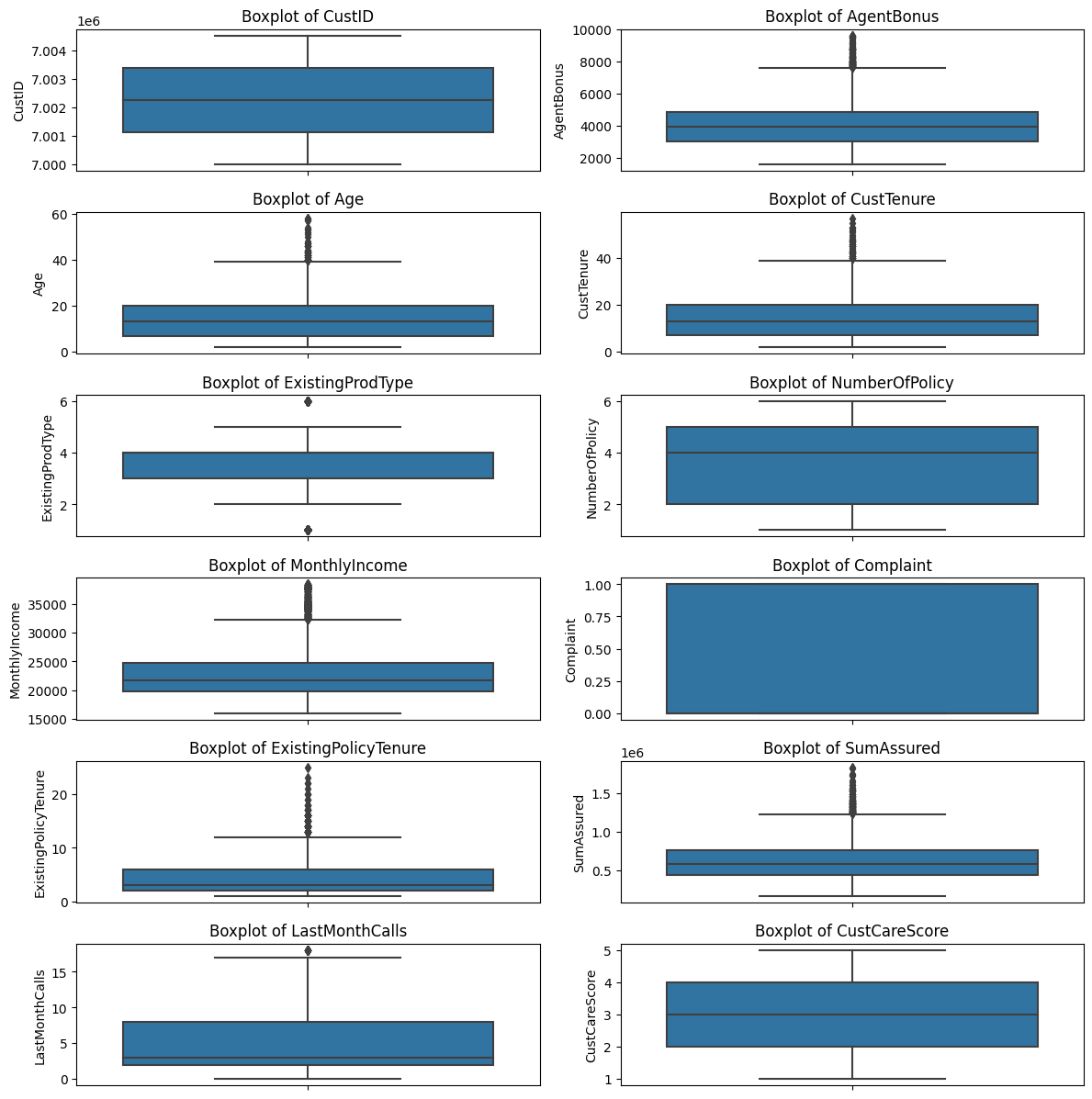
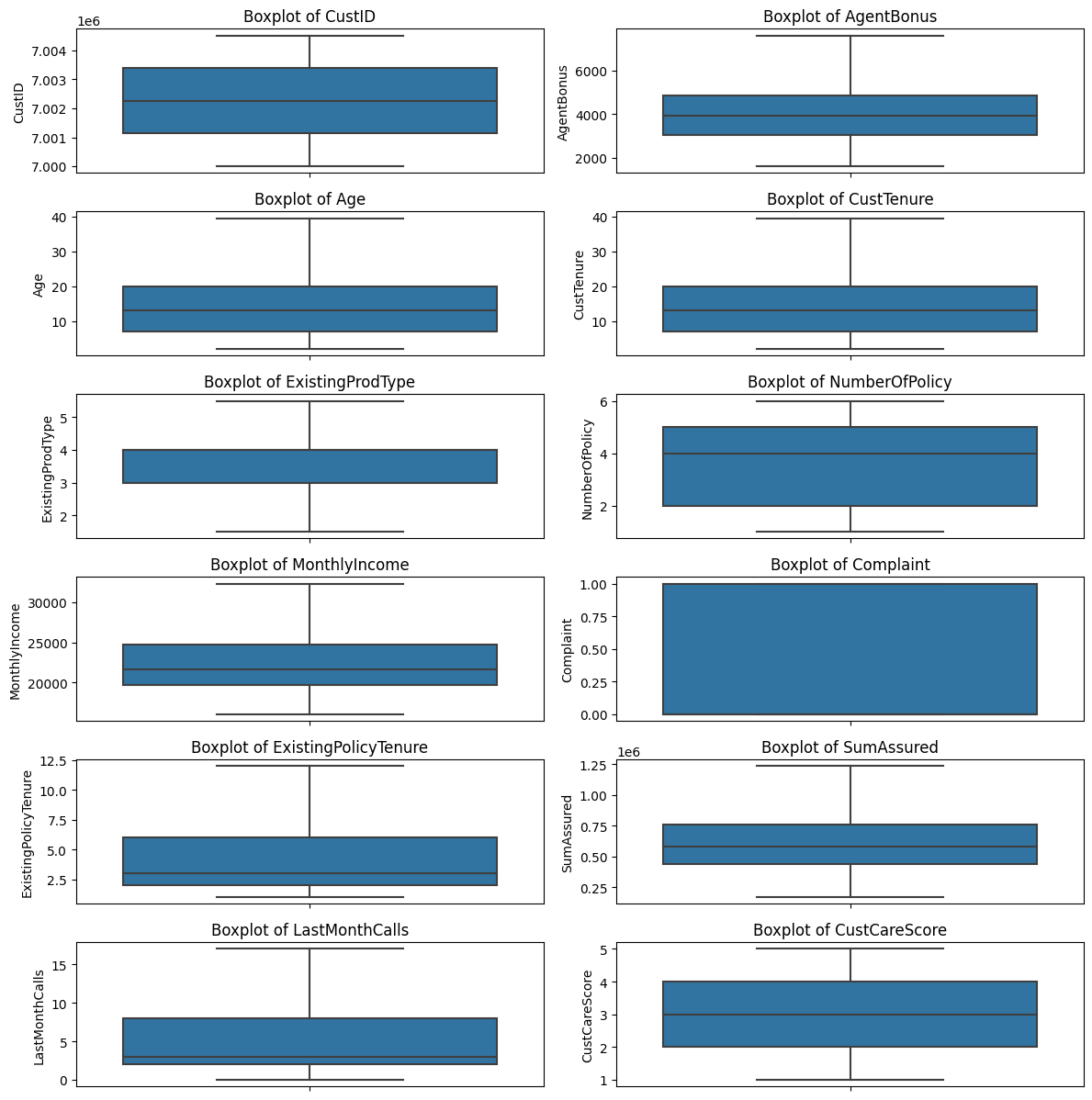
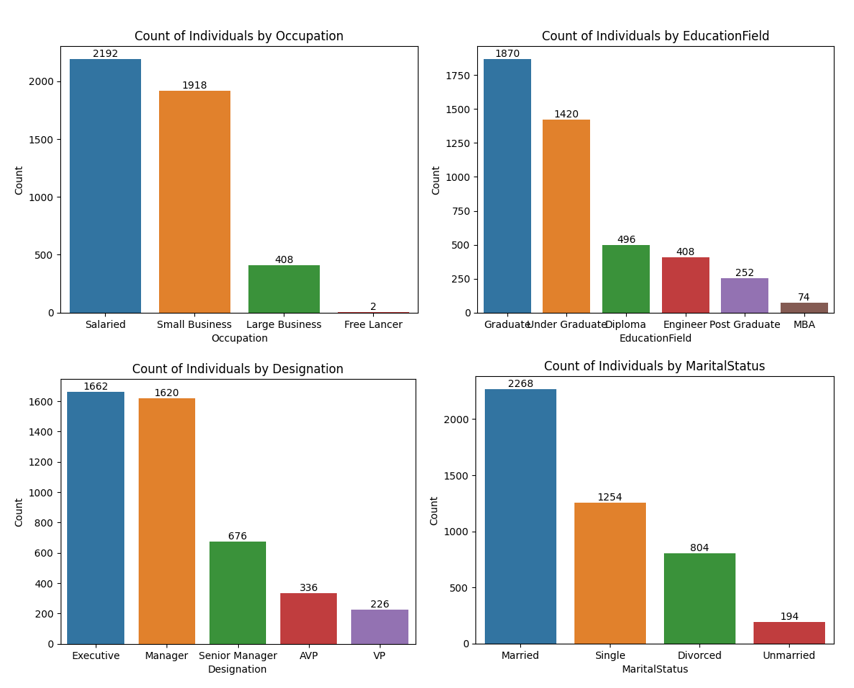


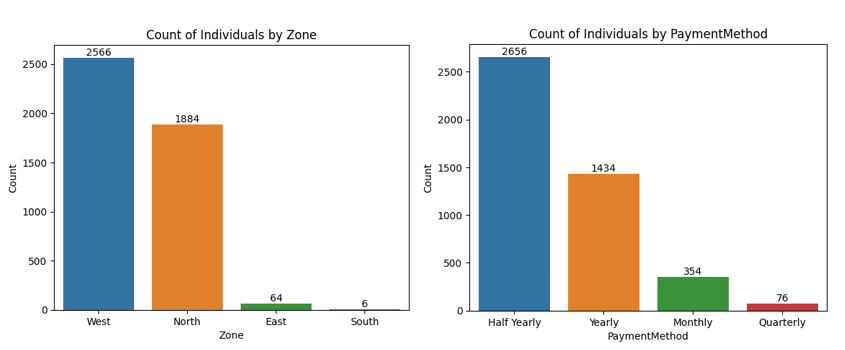
Figure 2: Outliers in Insurance data after treatment

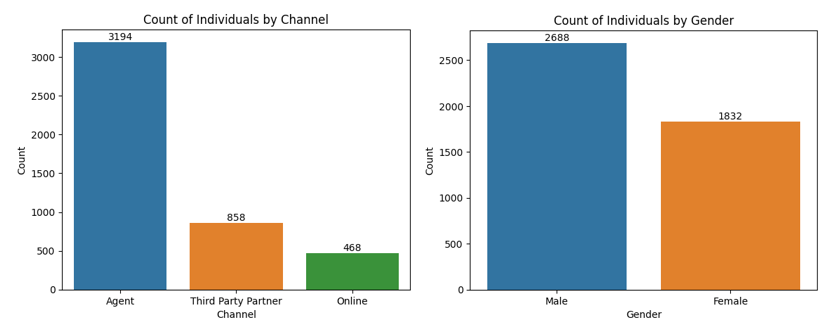


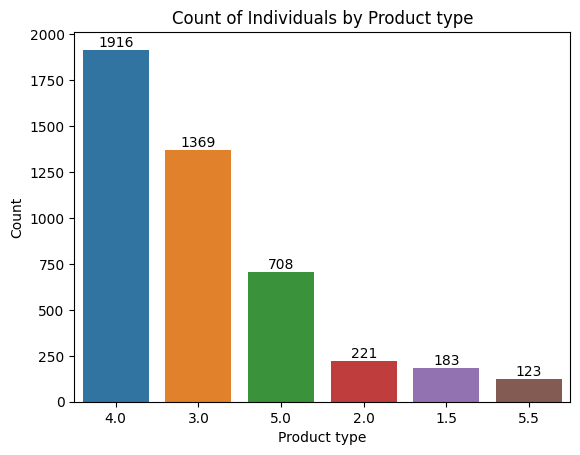
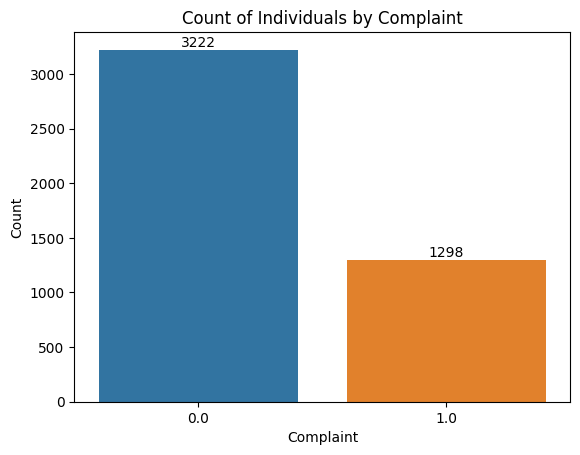
**EDA and Business Implication**

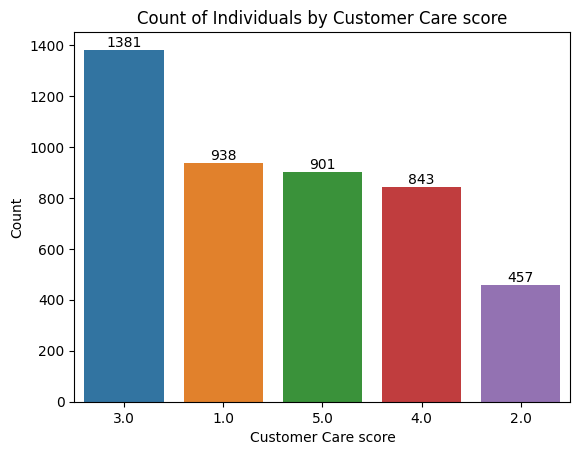
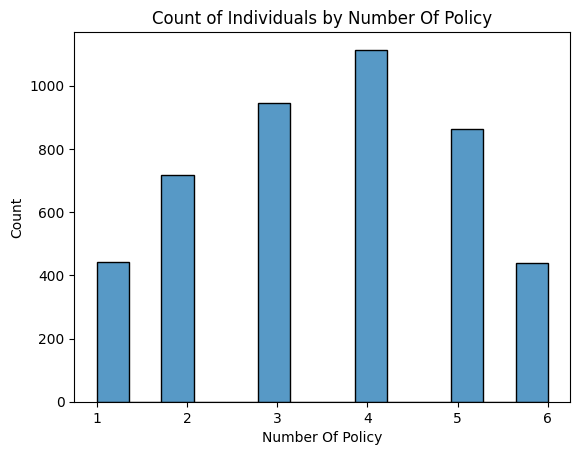
Figure 3: Univariate Analysis

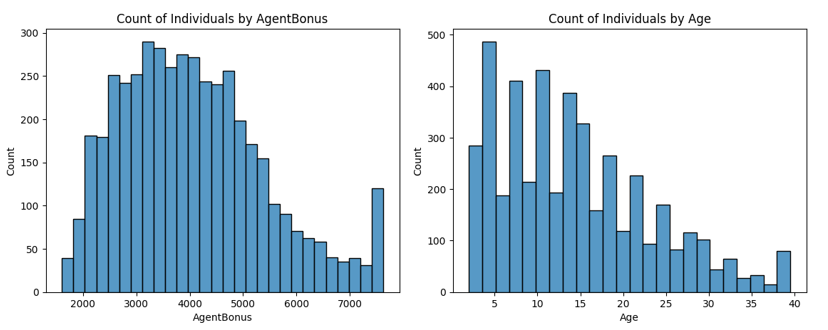
****

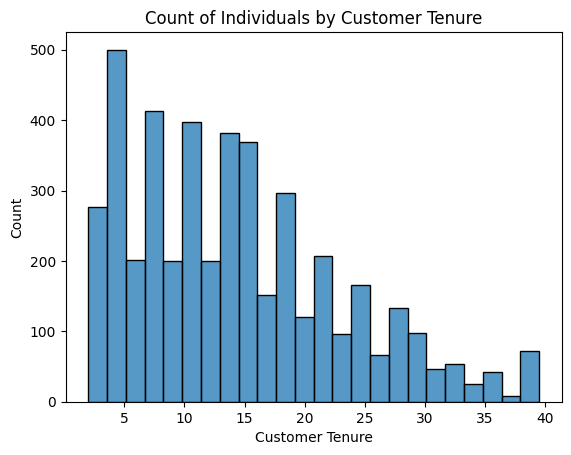
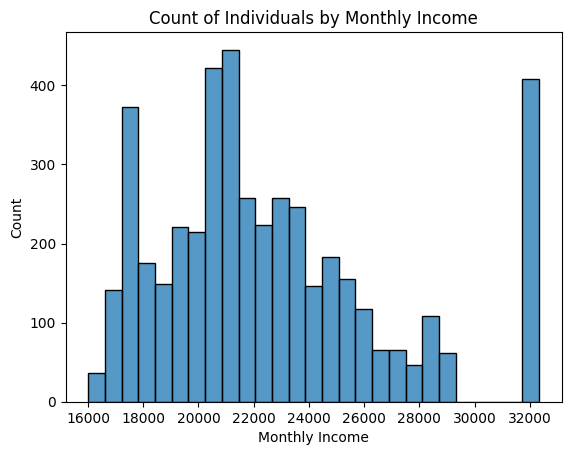
****

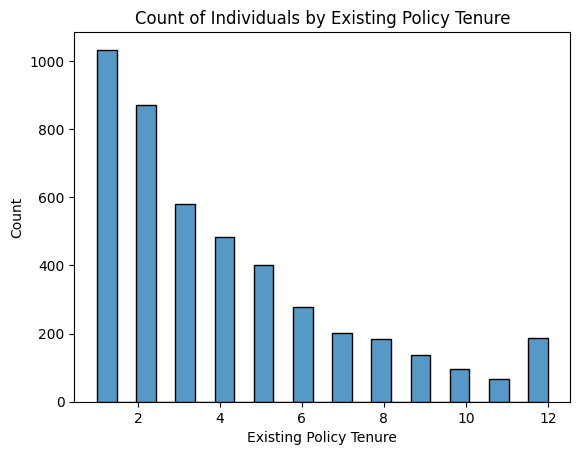
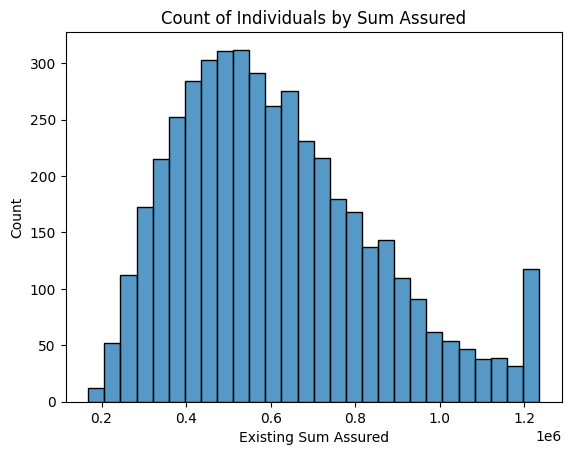


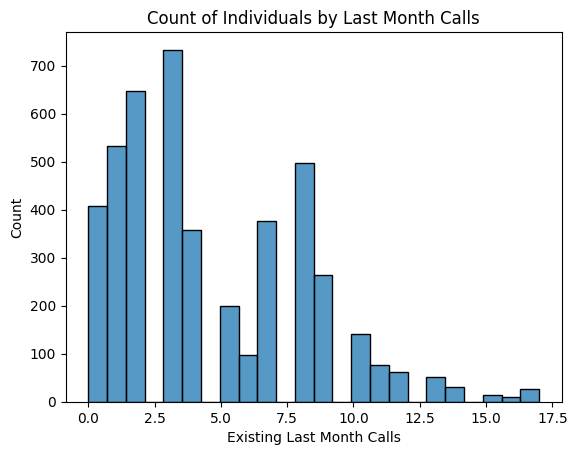
** **

**** 



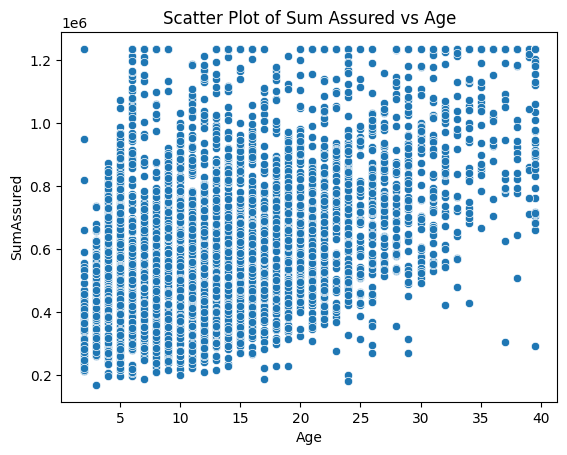
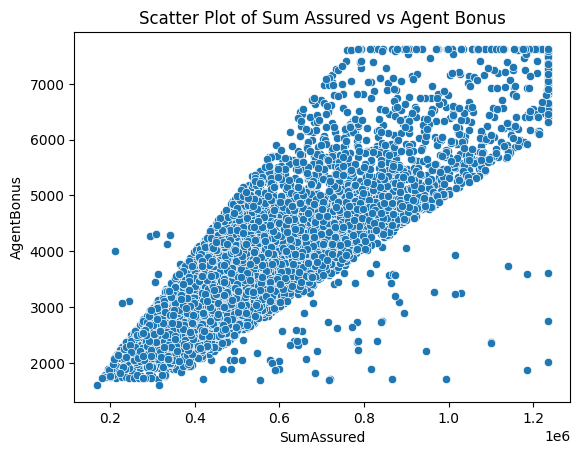


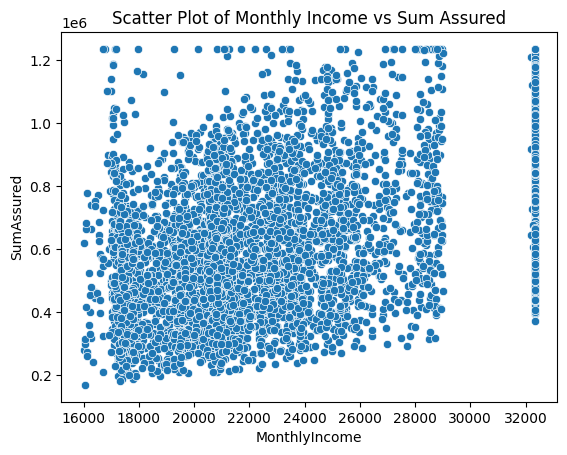
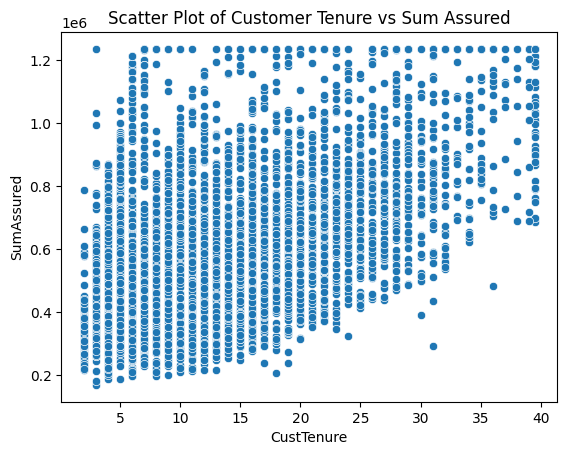
**Insights:**

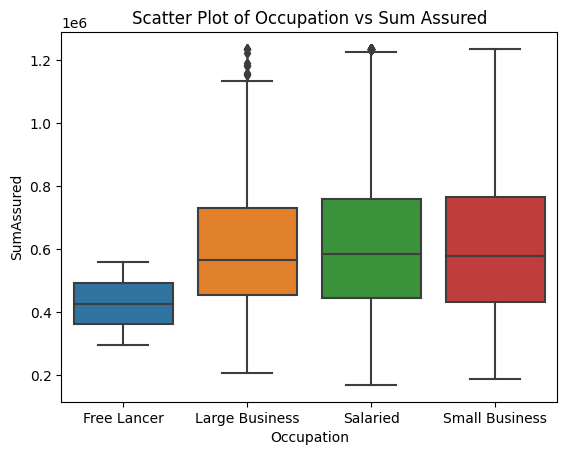
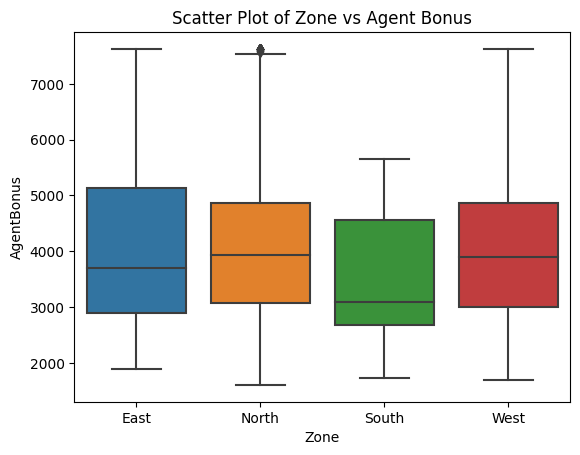
* High Aged customers are tend to pay high Bonuses
* Customers from high tenure are paying high agent bonus
* Less tenure policies are selected most
* Customers are more likely to opt for product number 4.
* As age increases, policies are decreasing
* The salaried and small business people are more interested in taking policies.
* In the Education Field, Graduates and Under Graduates are more applying for policies.
* Comparing Designations, Executives and Managers are making more policies.
* Married people are taking more policies.
* Comparing Zones, West and North people are taking more policies.
* Most of the people are making Half-yearly payments and the least is quarterly.
* Customers are reaching the organization mostly through Agent.
* Comparing Females and Males, Males are more likely to take policies.
* Agent Bonus is positively skewed.
* Customers are taking more policies for younger people.
* As age is increasing, the number of policies decreases.
* Fewer tenure policies are more liked by customers.
* Most of the customers are satisfied and have no complaints.

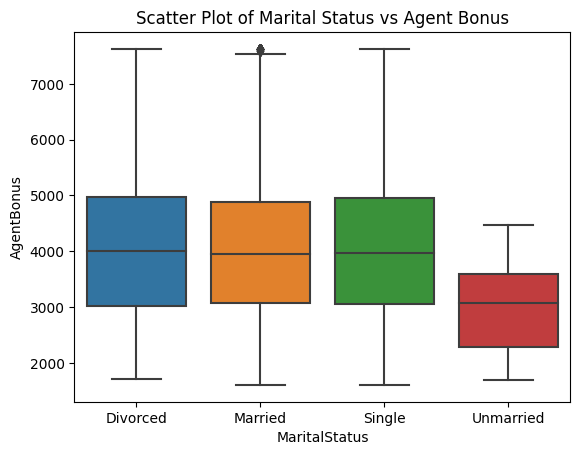
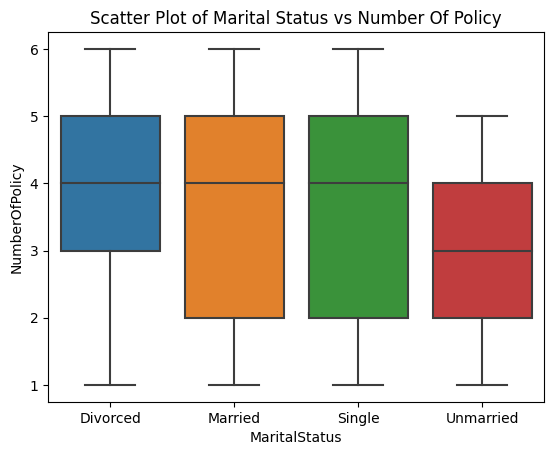
**Bi-Variate Analysis:**

Figure 5: Bi-variate analysis

****

** **

** **

****

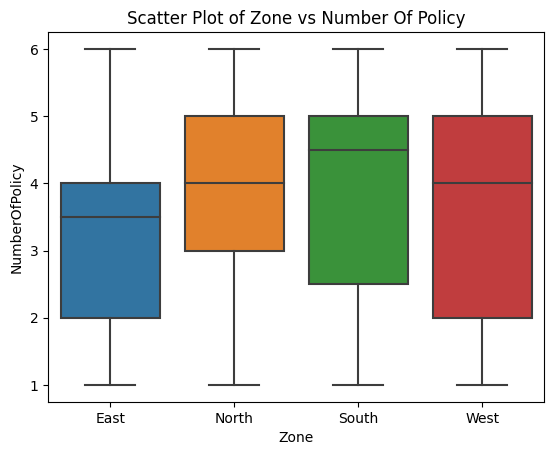
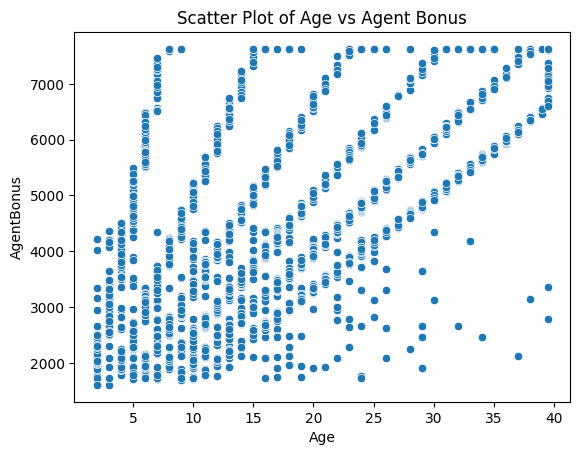
****

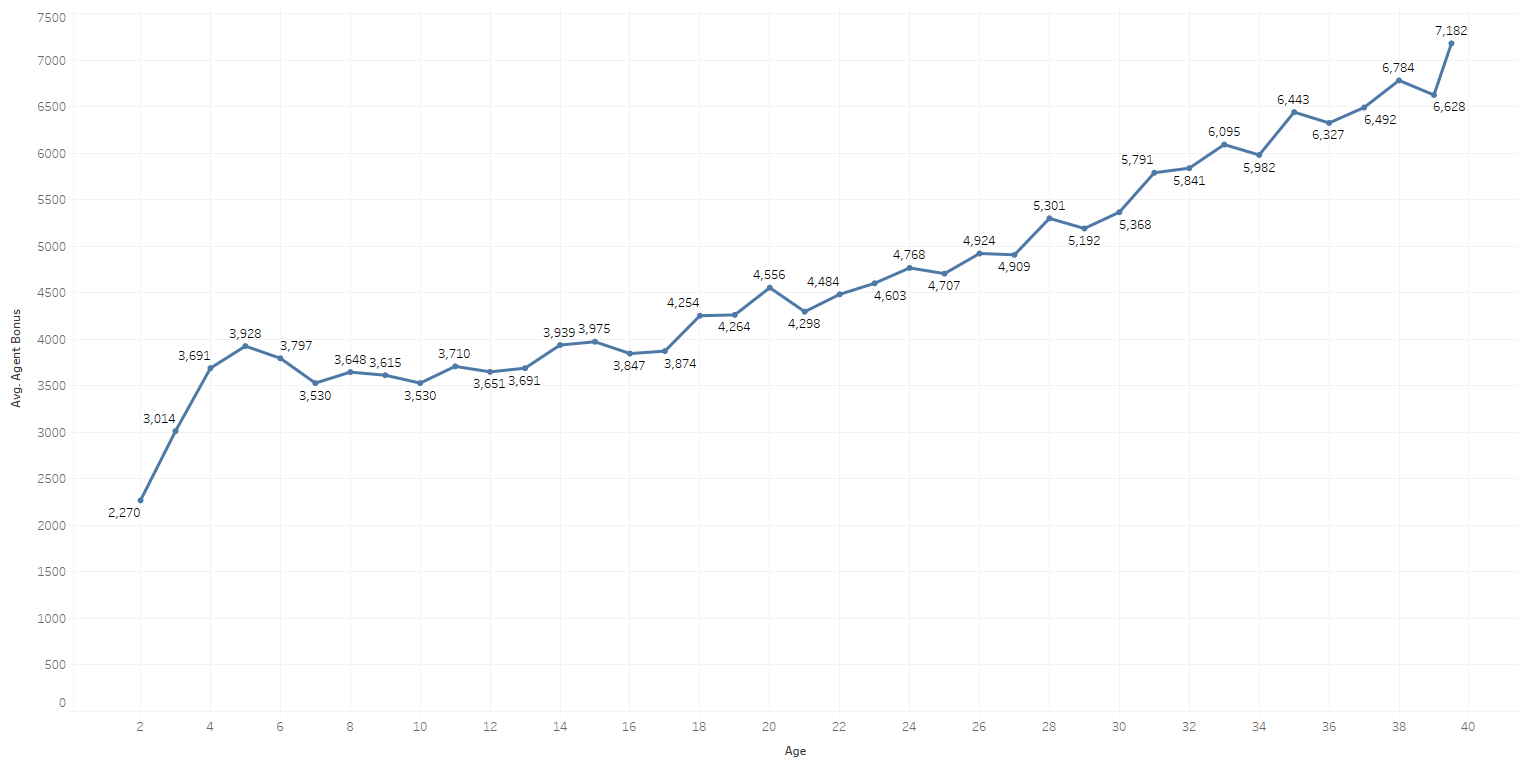
Figure 6: Age vs Agent Bonus(average)****

Figure 7: Agent Bonus vs Existing Policy Tenure vs Sum Assured

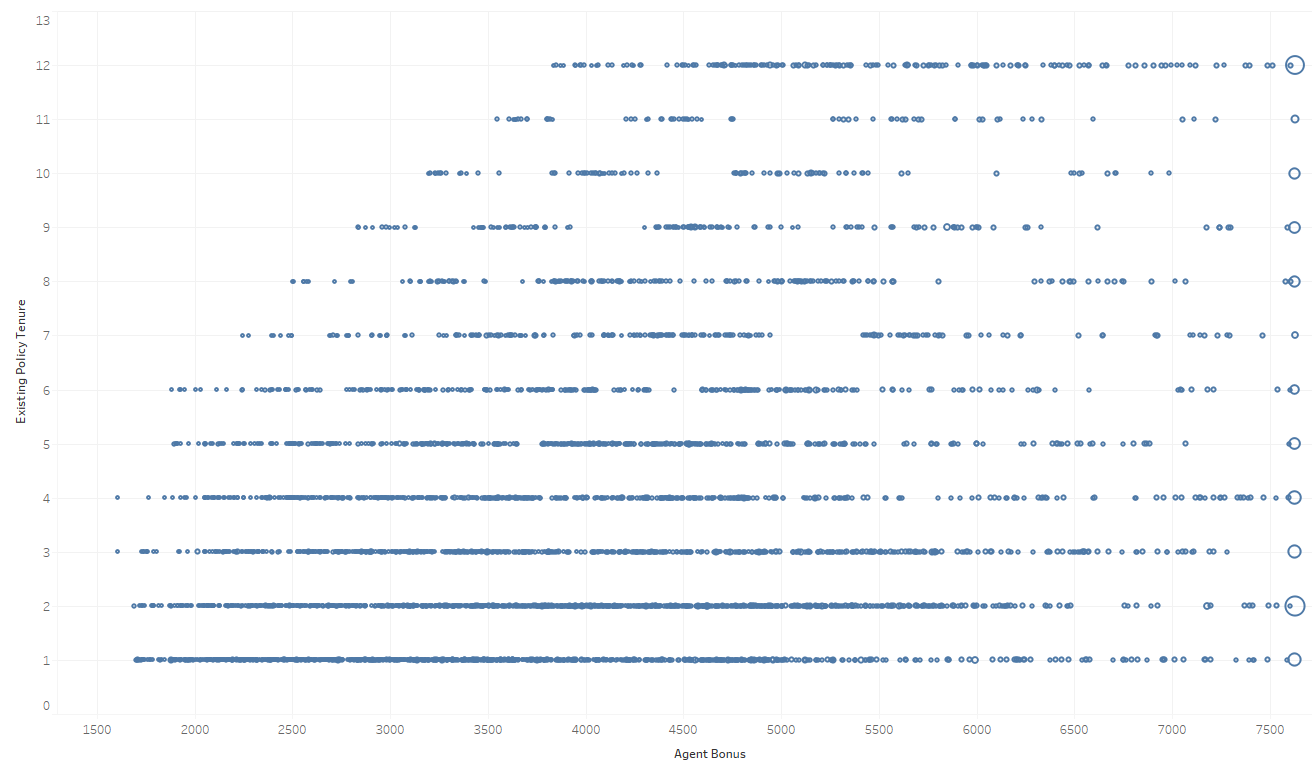
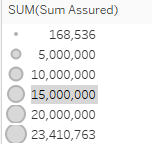
****

Figure 8: Customer Tenure vs Agent Bonus vs Gender

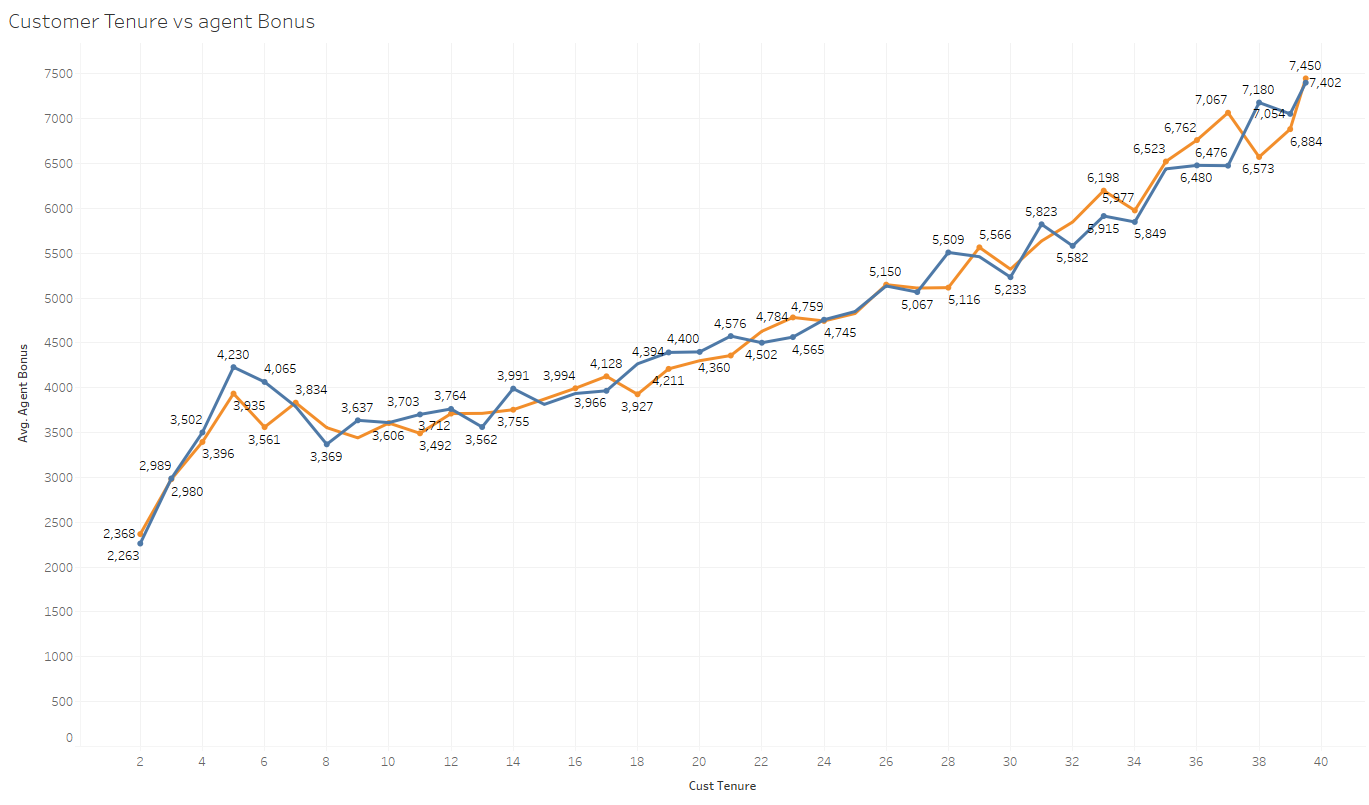
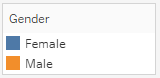
****

Figure 9: Gender vs Marital vs Agent Bonus

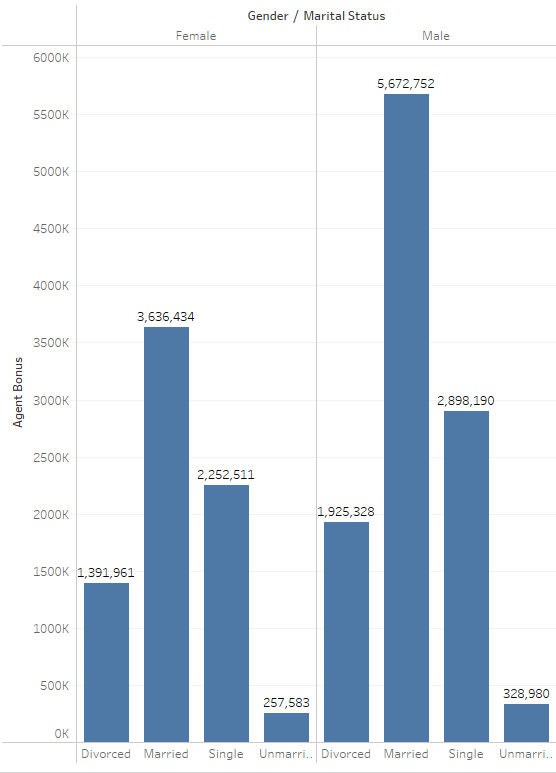
****

Figure 10: Gender vs Customer Tenure vs Agent Bonus

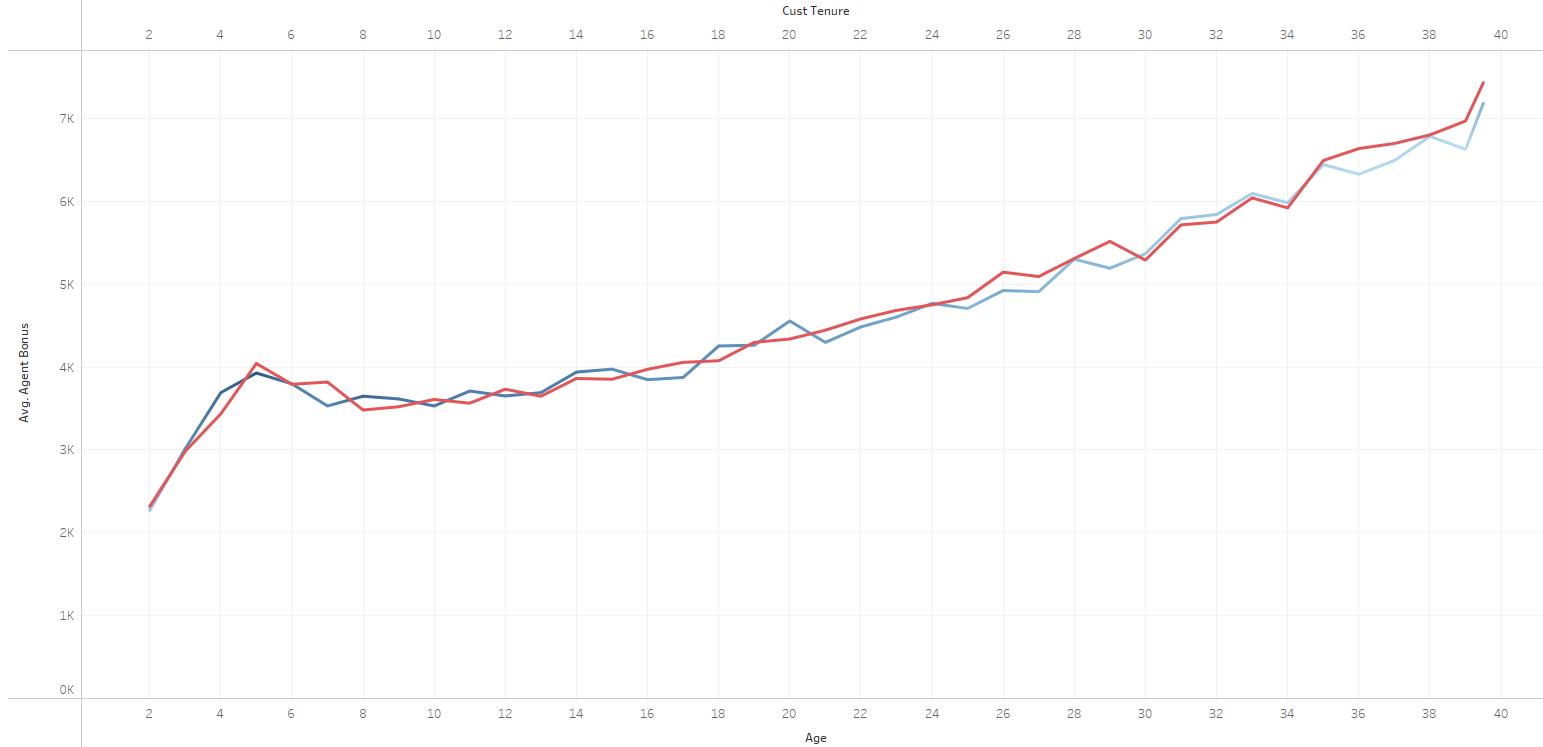
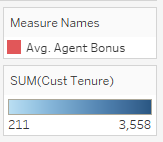
****

Figure 11: Gender vs Number of Policies vs Agent Bonus

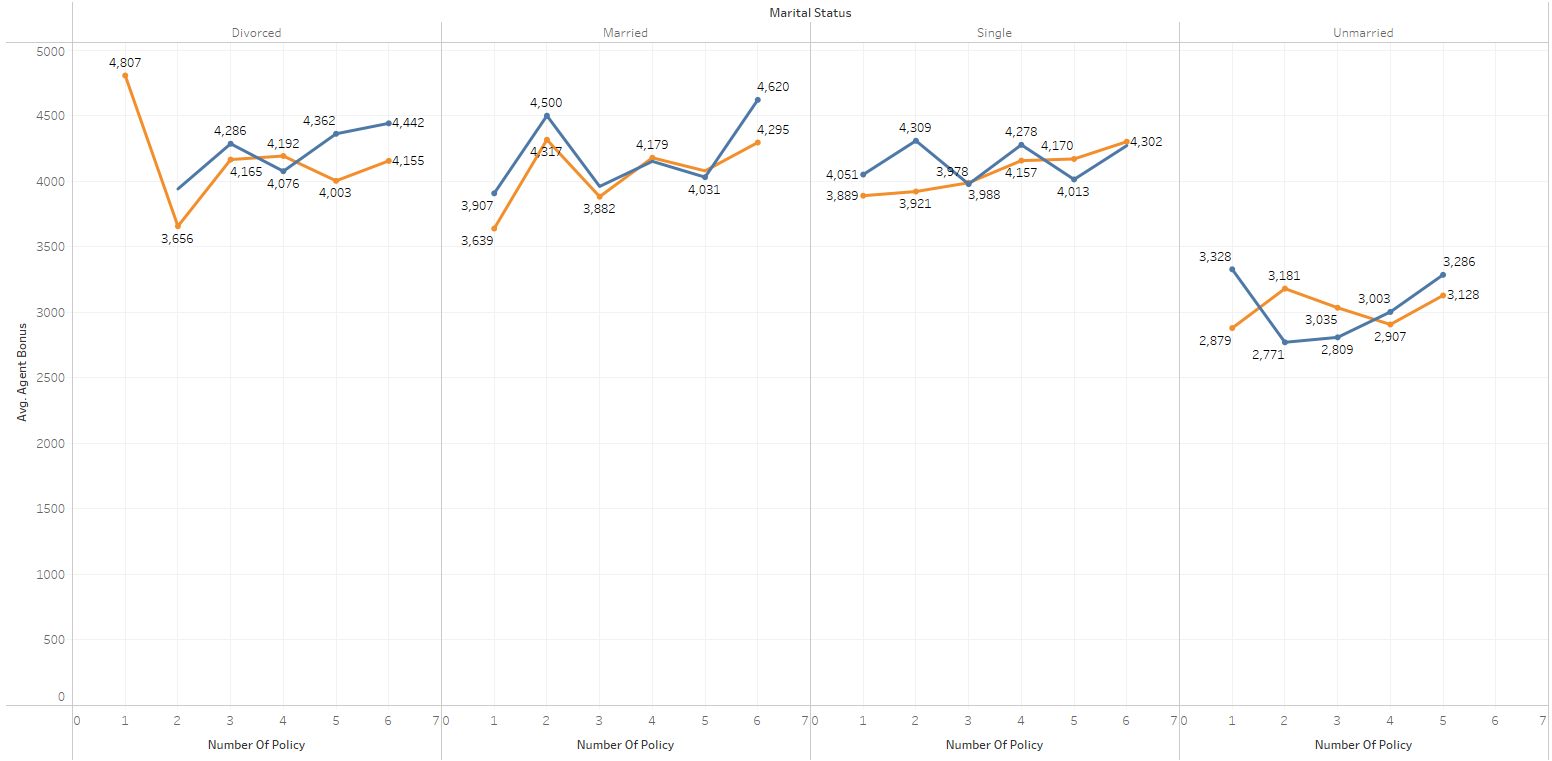
****

Figure 12: Monthly Income vs Number of policies vs Sum Assured

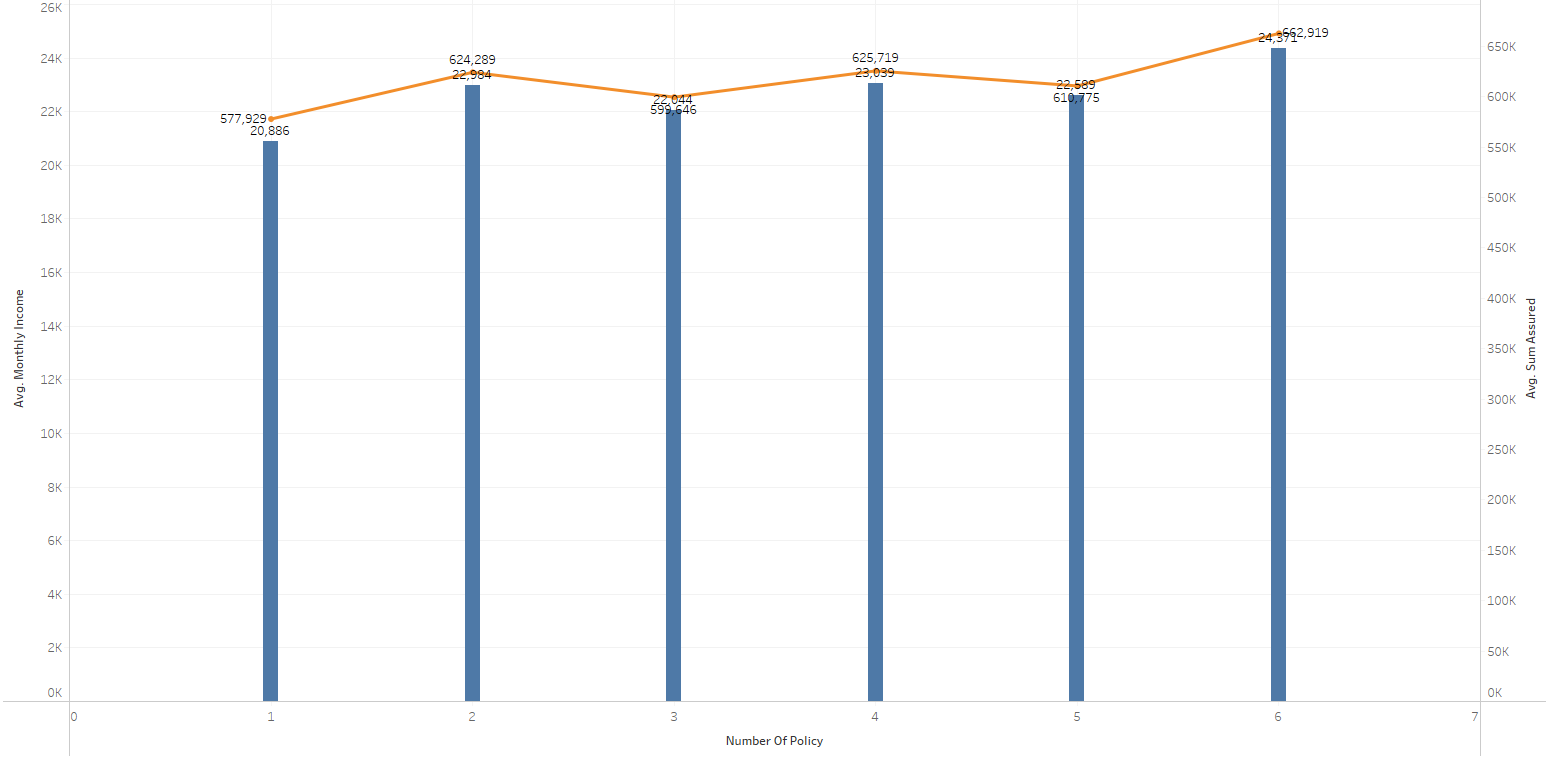
****

Figure 13: Customer Care Score vs Last Month Calls vs Agent Bonus vs Complaints

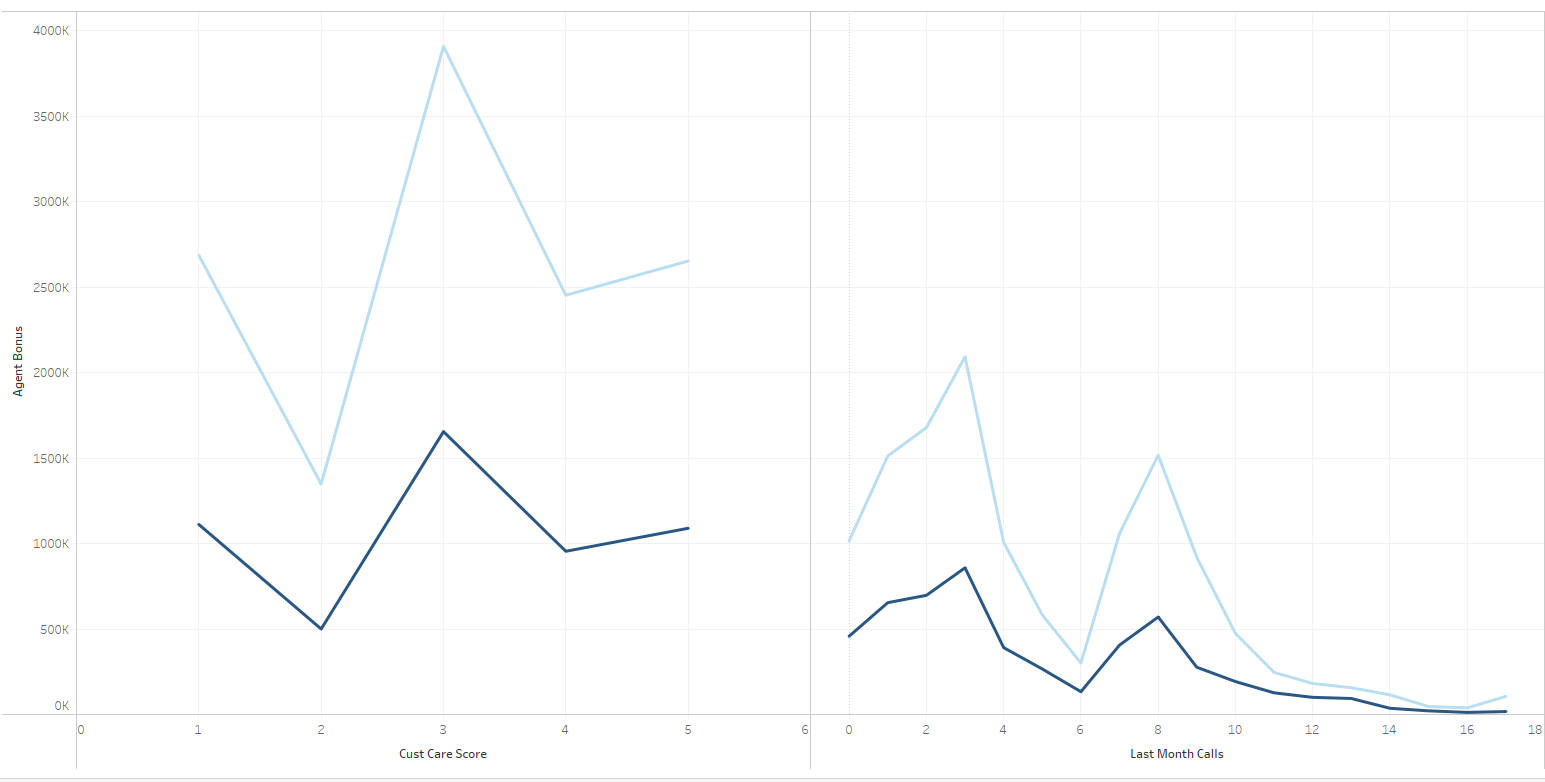
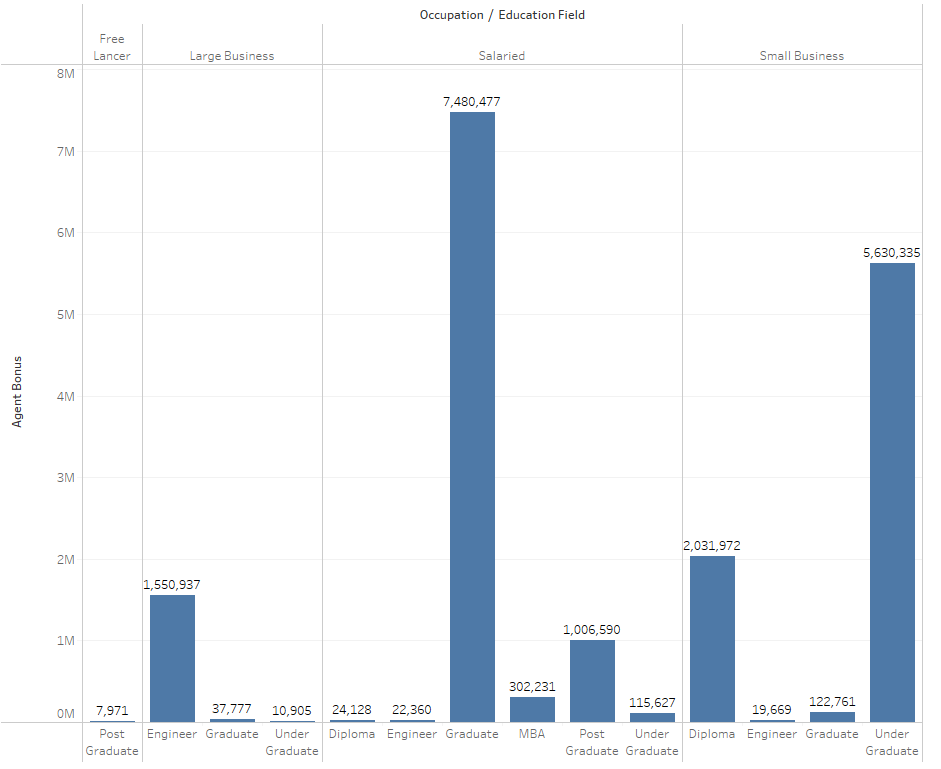
****

Figure 14: Occupation vs Education field vs Agent Bonus

****

**Insights from EDA:**

**From Age and Sum Assured plot:**

* Generally, there is an increasing trend in Sum Assured values as age increases from 2.0 to around 35.0.
* There seems to be some variability in Sum Assured values for age categories between 35.0 and 39.5, indicating potential differences in insurance coverage within this range.
* The age categories 36.0 and 37.0 show slightly lower Sum Assured values compared to the neighboring categories.
* It suggests that as individuals grow older, they may opt for higher insurance coverage or higher Sum Assured values.

**From Sum Assured and Agent Bonus plot:**

* There is some variation in Agent Bonus values for age categories between 35.0 and 39.5, indicating potential differences in bonus amounts within this range.
* The age categories 36.0 and 37.0 show relatively lower Agent Bonus values compared to the neighboring categories.
* There is a general increasing trend in Agent Bonus values as the Sum Assured increases.
* The highest Agent Bonus values are observed for higher Sum Assured categories, indicating a potential correlation between the insured amount and the bonus amount.
* It suggests that as the insured amount increases, there is a tendency for the bonus amount to increase as well.

**From Agent Bonus and Channel plot:**

* The average Agent Bonus for the "Agent" channel is 4,088.29.
* The average Agent Bonus for the "Online" channel is 4,085.31.
* The average Agent Bonus for the "Third Party Partner" channel is 3,955.50.
* The "Agent" and "Online" channels have similar average bonus amounts, while the "Third Party Partner" channel has a slightly lower average bonus amount.

**From Sum Assured and Occupation plot:**

* The average Sum Assured for the "Free Lancer" occupation is 426,484.50.
* The average Sum Assured for the "Large Business" occupation is 599,346.54.
* The average Sum Assured for the "Salaried" occupation is 618,792.27.
* The average Sum Assured for the "Small Business" occupation is 616,861.98.
* The "Salaried" occupation has the highest average sum assured, followed by "Small Business" and "Large Business" occupations. The "Free Lancer" occupation has the lowest average Sum Assured among the listed occupations.

**From Agent Bonus and Customer Tenure plot:**

* The variation in Agent Bonus across different tenure categories, indicating that customers with longer tenure tend to have higher average Agent Bonuses.
* **From the Age and Bonus plot**, as age increases customers are more likely to give higher bonuses to agents.
* **From Sum Assured vs Agent Bonus vs Customer Tenure**, Higher tenure customers are giving higher bonuses and a higher amount assured tends to higher bonus.
* **From Customer Tenure vs Agent Bonus vs Gender**, comparing the sum of agent bonus, Males are paying higher bonuses than females. The plot is the average of Agent Bonus which has not much difference between gender.
* **From Gender vs Marital vs Agent Bonus,** Married people are paying higher bonuses regardless of gender.
* **From Customer Care Score vs Last Month Calls vs Agent Bonus**, we could see the customer satisfaction, where customers with lower number of calls and lower complaints are giving higher bonuses.

**MODEL BUILDING**

Figure 15: Regression Models used

**Why these models?**

Each of these models offers unique advantages and addresses different aspects of the prediction problem.

* **Interpretability:** Linear Regression provides transparency in understanding the relationship between features and the target variable. This insight can help explain how various factors influence agent bonuses.
* **Non-linearity Handling:** Decision Tree Regressor and Random Forest Regressor can capture non-linear relationships in the data, accommodating complex interactions between features.
* **Ensemble Learning:** Random Forest Regressor, being an ensemble model, combines the predictions of multiple decision trees, mitigating overfitting and improving robustness.
* **Regularization:** Ridge and Lasso Regression models introduce regularization to prevent overfitting, making them suitable for datasets with many features.
* **High-Dimensional:** SVR can handle non-linear relationships and high-dimensional data and is suitable for non-linear relationships between the features and the target variable, especially when we have limited training data.
* **Cross Validation:** Cross Validation estimates the model's generalization performance and identify potential issues like overfitting or underfitting and ensure that the model's performance metrics are reliable and can be trusted for making predictions on new, unseen data.

Table 5: MSE and R square of all the Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.no** | **Regression Models** | **MSE** | | **R square** | |
|  |  | **Training Data** | **Testing Data** | **Training Data** | **Testing Data** |
| **1.** | Linear Regression | 0.20 | 0.22 | 80 | 76 |
| **2.** | Decision Tree Regressor | 0.13 | 0.21 | 86 | 77 |
| **3.** | Random Forest Regressor | 0.021 | 0.15 | 97 | 83 |
| **4.** | Ridge Regression | 0.20 | 0.22 | 80 | 76 |
| **5.** | Lasso Regression | 0.236 | 0.241 | 76 | 74 |
| **6.** | Support Vector Regression | 0.21 | 0.22 | 80 | 76 |

Note:

* For a model to be best, MSE(Mean Square Root) to be low and R square must be high.

From the above regression models, comparing MSE and R square values we could decide that Random Forest is the best model.

|  |  |  |
| --- | --- | --- |
| **S.no** | **Cross Validation Models** | **MSE** |
| **1.** | Linear Regression | 402557 |
| **2.** | Decision Tree Regressor | 572564.6 |
| **3.** | Random Forest Regressor | 285600.4 |

Table 6: MSE values after Cross-validation

Table 7: Ensemble Learning for Random Forest

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **S.no** | **Regression Models** | **MSE** | | **R square** | |
|  |  | **Training Data** | **Testing Data** | **Training Data** | **Testing Data** |
| **1.** | Bagging | 5 | 15 | 86 | 77 |
| **2.** | Boosting | 12 | 16 | 86 | 77 |

The Ensemble learning techniques are performed on Random Forest and the results are as above.

The Random Forest Regressor has an R-squared value of 97% for the training data and 83% for the testing data, indicating a high level of variance explained by the model.

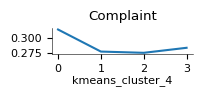
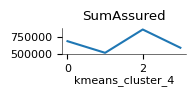
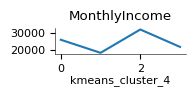
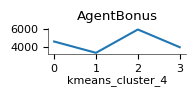
Based on the provided information, the Bagging model (specifically using Random Forest) has an MSE value of 5 for the training data and 15 for the testing data. This indicates that the model performs well in terms of minimizing prediction errors.

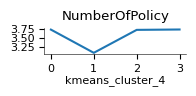
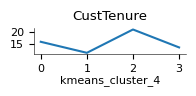
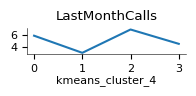
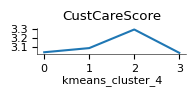
The Bagging model has an R-squared value of 86% for both the training and testing data, it explains a good variance in the target variable Agent Bonus.

Comparing the Bagging model with the other models, it shows competitive performance in terms of MSE and R-squared. Therefore, Bagging with Random Forest can be considered as one of the best models for this dataset.

**Clustering Model**

Figure 16: Clustering Models



**Model Validation**

In model validation, various evaluation metrics are used to assess the performance of the models, and it goes beyond just accuracy. While accuracy is a common metric, especially for classification tasks, other metrics are used based on the specific type of model and the nature of the data.

The target variable is continuous, we used regression tasks which mainly include MSE and R-square.

**Mean Squared Error (MSE):** It measures the average squared difference between the predicted and actual values in regression tasks. Lower MSE indicates better model performance.

**R-squared (R2):** It represents the proportion of variance in the target variable that is explained by the model. Higher R-squared values indicate better model fit.

**Root Mean Squared Error (RMSE):** The square root of MSE, giving a more interpretable measure of the model's error.

**Cross-Validation Scores:** Using techniques like k-fold cross-validation to estimate the model's generalization performance on unseen data.

**Rec****ommendations**

* Unmarried customers are to be focused and singles are to be offered better policies to increase sales.
* Insurance company must focus on East, and South zones for increasing policies.
* Online services and advertisement could help in increase of reach making the increase in sales through online and third parties.
* Customers with middle and higher aged are to be focused.
* According to the product type, other product types can be modified comparing product type 4 which is most opted.
* The Customer care score is better but must be increased, as could see the score of 1 at little raise.
* Agents should focus on providing high sum insurance to get high bonus

Clustering is also build for detailed explanation for customer satisfaction.

* Consider cluster 1, Last month’s calls, Customer Tenure, Number of policies, and Customer care score are the lowest.
* Consider cluster 2, Last month’s calls, Customer Tenure, Number of policies, and Customer care score are the highest.
* Lowest calls might be resulting in a lower customer care score this might also effect the number of policies taken by the customer.
* Customer care score in cluster 0 is the lowest. This effects the agent bonus. The customer care score has to be taken care of in all the other clusters.