CAPSTONE PROJECT

Notes

Life Insurance Sales

Done by:
Kewal Kumar Singh

Table of Content:

Data Dictionary

Data	Variable	Description
Sales	CustID	Unique customer ID
Sales	AgentBonus	Bonus amount given to each agents in last month
Sales	Age	Age of customer
Sales	CustTenure	Tenure of customer in organization
Sales	Channel	Channel through which acquisition of customer is done
Sales	Occupation	Occupation of customer
Sales	EducationField	Field of education of customer
Sales	Gender	Gender of customer
Sales	ExistingProdType	Existing product type of customer
Sales	Designation	Designation of customer in their organization
Sales	NumberOfPolicy	Total number of existing policy of a customer
Sales	MaritalStatus	Marital status of customer
Sales	MonthlyIncome	Gross monthly income of customer
Sales	Complaint	Indicator of complaint registered in last one month by customer
Sales	ExistingPolicyTenure	Max tenure in all existing policies of customer
Sales	SumAssured	Max of sum assured in all existing policies of customer
Sales	Zone	Customer belongs to which zone in India. Like East, West, North and South
Sales	PaymentMethod	Frequency of payment selected by customer like Monthly, quarterly, half yearly and yearly
Sales	LastMonthCalls	Total calls attempted by company to a customer for cross sell
Sales	CustCareScore	Customer satisfaction score given by customer in previous service call

Introduction of the Business Problem

a) Defining the Problem Statement

In the highly competitive life insurance sector, agent performance plays a crucial role in driving both sales and customer satisfaction. A leading life insurance company is seeking to predict the **bonus amount** for its agents based on their performance data. The main objective is to develop a predictive model that can estimate bonuses accurately. This will enable the company to proactively identify top-performing agents and design tailored incentive programs, while also recognizing underperformers and offering them targeted support and training.

b) Need of the Study/Project

The life insurance industry thrives on a motivated and efficient agent workforce. However, maintaining consistent performance across the board remains a challenge. Currently, the company relies on retrospective assessments to award bonuses, which often delays corrective measures or rewards. By predicting bonus payouts in advance:

- The company can **retain high-performing agents** by recognizing their efforts earlier.

- **Upskilling initiatives** can be planned proactively for agents who may need support.

- It enables **data-driven resource allocation**, helping management make informed decisions about agent training, marketing support, and performance monitoring.

Ultimately, the project aims to bring in more **transparency, objectivity, and foresight** into agent engagement and performance management strategies.

c) Understanding the Business/Social Opportunity

This project offers both **business and social impact opportunities**:

- **Business Opportunity:** Predictive modeling will improve operational efficiency and enhance the company's ability to nurture talent, which is critical for sales-driven industries like insurance. This will directly contribute to better revenue generation, reduced attrition, and higher customer satisfaction.

- **Social Opportunity:** For the agents, especially those from underrepresented or rural backgrounds, early identification of performance gaps through this model can lead to timely interventions such as mentorship or skill development. This fosters a more inclusive and supportive work environment, where individuals are given fair chances to grow.

Dataset Report

a) Data Collection Overview

The dataset, provided for academic analysis, is treated as a **primary data source**, reflecting real-life operational data from a life insurance company. It captures **monthly customer and policy activity**, as indicated by the variable LastMonthCalls, which represents the number of contact attempts made to customers for cross-selling purposes in the past month.

The **data collection methodology** is further illustrated by the variable Channel, showing the mode of customer acquisition, such as through agents, third-party channels, or online platforms, offering insight into sales strategy and lead sourcing.

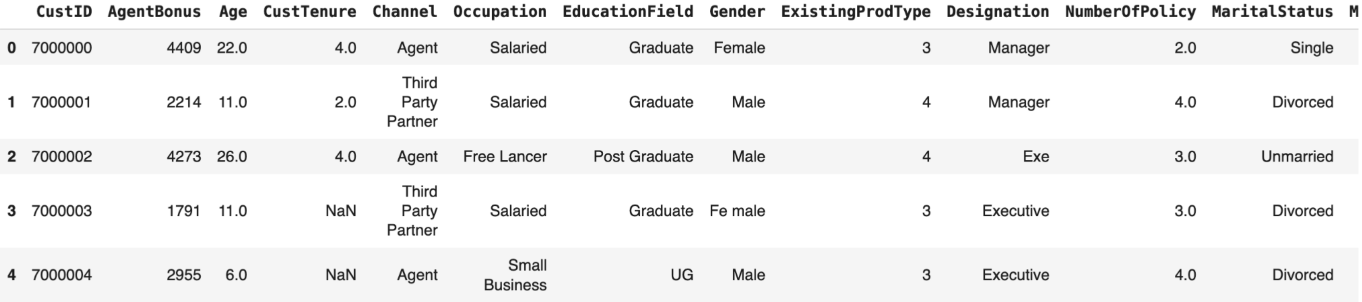
Additionally, variables such as CustTenure, ExistingPolicyTenure, and NumberOfPolicy suggest that **customer history and relationship depth** were tracked, possibly through a centralized CRM system. Complaint, CustCareScore, and PaymentMethod reflect **service interactions**, while SumAssured, MonthlyIncome, and EducationField highlight financial and demographic profiling.

Together, these attributes indicate that the data was gathered systematically through both **customer interactions and policy records**, likely updated regularly within an internal sales and service database. The focus appears to be on capturing the full lifecycle of customer engagement, from acquisition to servicing and feedback.

b) Visual Inspection of Data

The dataset comprises **4,520 rows and 20 columns**, each representing individual customer records with various demographic, financial, and behavioral attributes.

First 5 rows of the dataset:



	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	M
0	7000000	4409	22.0	4.0	Agent	Salaried	Graduate	Female	3	Manager	2.0	Single	
1	7000001	2214	11.0	2.0	Third Party Partner	Salaried	Graduate	Male	4	Manager	4.0	Divorced	
2	7000002	4273	26.0	4.0	Agent	Free Lancer	Post Graduate	Male	4	Exe	3.0	Unmarried	
3	7000003	1791	11.0	NaN	Third Party Partner	Salaried	Graduate	Fe male	3	Executive	3.0	Divorced	
4	7000004	2955	6.0	NaN	Agent	Small Business	UG	Male	3	Executive	4.0	Divorced	

Fig 1. First 5 rows of the dataset

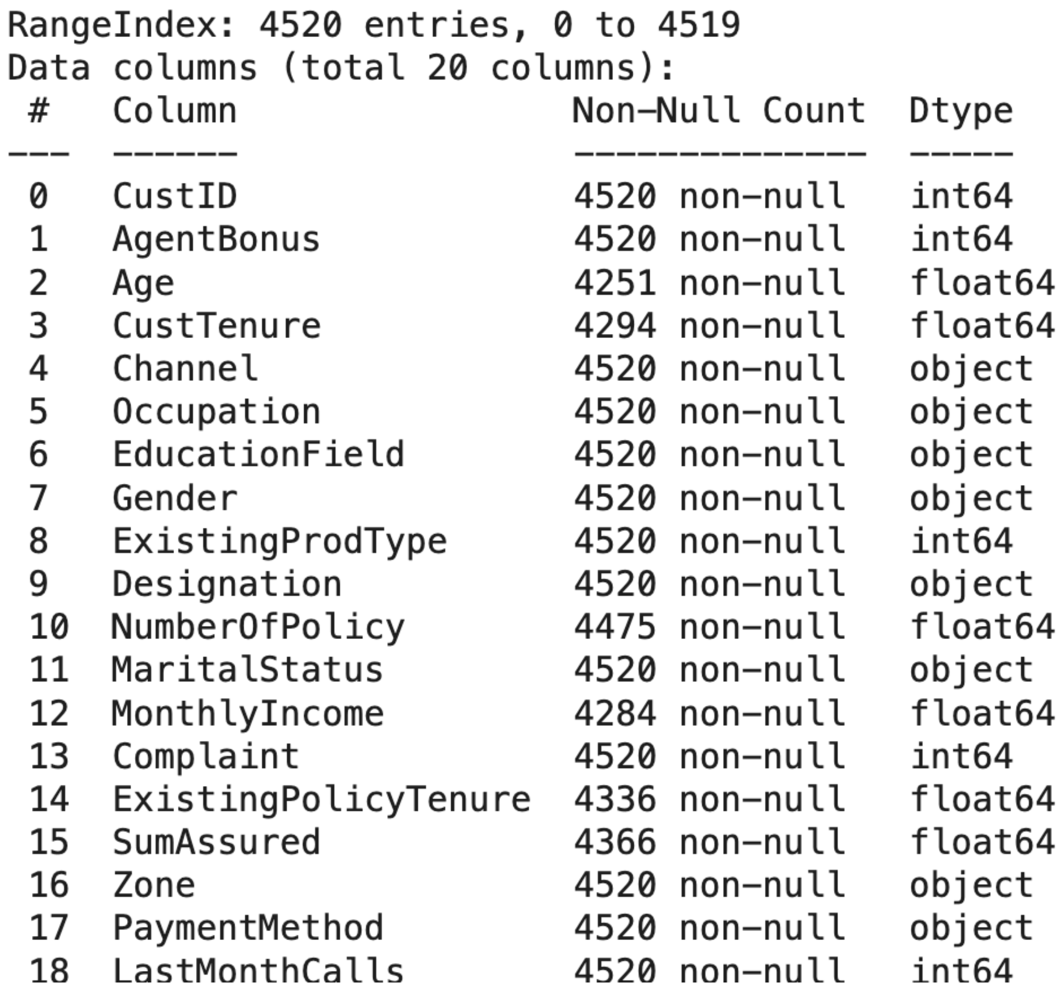
c) Variable Information:

Out of the 19 features (excluding the target variable), we observe the following:

- **5 variables** are of integer type

- **7 variables** are of float type

- The **remaining variables** are categorical in nature and will require encoding before they can be used in machine learning models.



```
RangeIndex: 4520 entries, 0 to 4519
Data columns (total 20 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   CustID                4520 non-null   int64  
 1   AgentBonus            4520 non-null   int64  
 2   Age                   4251 non-null   float64
 3   CustTenure            4294 non-null   float64
 4   Channel               4520 non-null   object 
 5   Occupation            4520 non-null   object 
 6   EducationField        4520 non-null   object 
 7   Gender                4520 non-null   object 
 8   ExistingProdType      4520 non-null   int64  
 9   Designation           4520 non-null   object 
 10  NumberOfPolicy        4475 non-null   float64
 11  MaritalStatus         4520 non-null   object 
 12  MonthlyIncome         4284 non-null   float64
 13  Complaint             4520 non-null   int64  
 14  ExistingPolicyTenure  4336 non-null   float64
 15  SumAssured            4366 non-null   float64
 16  Zone                  4520 non-null   object 
 17  PaymentMethod         4520 non-null   object 
 18  LastMonthCalls        4520 non-null   int64  
```

Fig 2: Variable Information

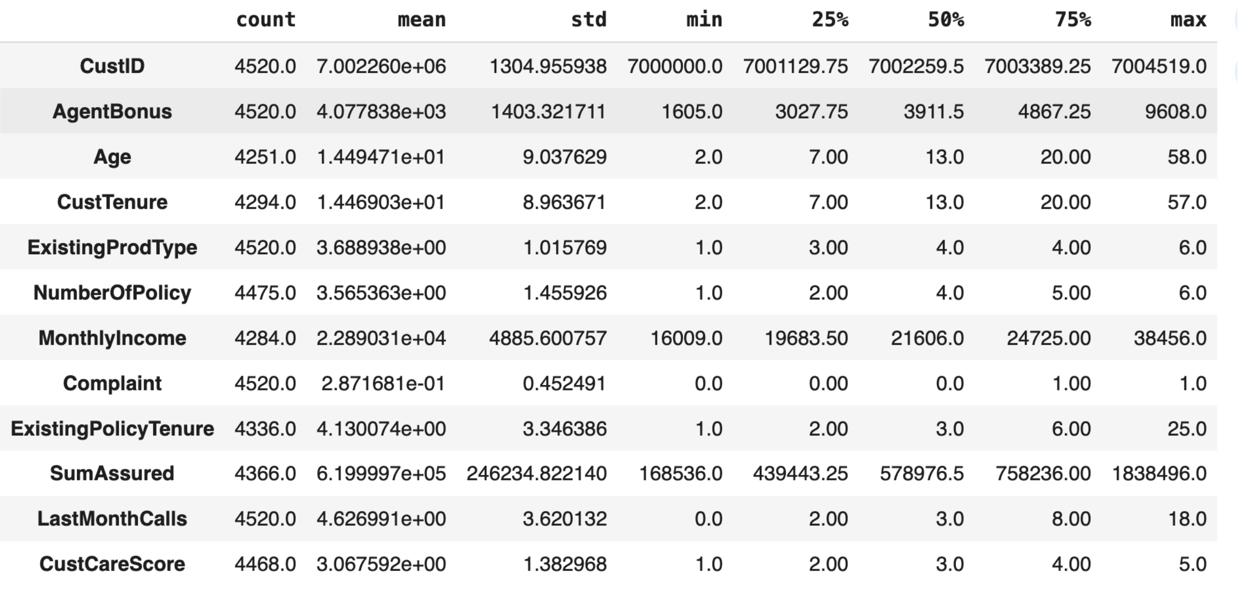
Statistical Observations:

A preliminary analysis of the numerical variables reveals the following insights:

- The variable Age has a minimum value of 2 years, and 50% of the values fall around 13 years, suggesting a concentration of younger policyholders. This distribution is left-skewed, indicating that minors form a large portion of the dataset.

- The variable CustTenure, which denotes the number of years a customer has been associated with the company, also exhibits a left-skewed distribution, with over half of the customers having a tenure of 13 years or less.

- Most of the remaining numerical variables follow a roughly normal distribution, as observed in their bell-shaped density curves, which will be further visualized during univariate analysis.



	count	mean	std	min	25%	50%	75%	max
CustID	4520.0	7.002260e+06	1304.955938	7000000.0	7001129.75	7002259.5	7003389.25	7004519.0
AgentBonus	4520.0	4.077838e+03	1403.321711	1605.0	3027.75	3911.5	4867.25	9608.0
Age	4251.0	1.449471e+01	9.037629	2.0	7.00	13.0	20.00	58.0
CustTenure	4294.0	1.446903e+01	8.963671	2.0	7.00	13.0	20.00	57.0
ExistingProdType	4520.0	3.688938e+00	1.015769	1.0	3.00	4.0	4.00	6.0
NumberOfPolicy	4475.0	3.565363e+00	1.455926	1.0	2.00	4.0	5.00	6.0
MonthlyIncome	4284.0	2.289031e+04	4885.600757	16009.0	19683.50	21606.0	24725.00	38456.0
Complaint	4520.0	2.871681e-01	0.452491	0.0	0.00	0.0	1.00	1.0
ExistingPolicyTenure	4336.0	4.130074e+00	3.346386	1.0	2.00	3.0	6.00	25.0
SumAssured	4366.0	6.199997e+05	246234.822140	168536.0	439443.25	578976.5	758236.00	1838496.0
LastMonthCalls	4520.0	4.626991e+00	3.620132	0.0	2.00	3.0	8.00	18.0
CustCareScore	4468.0	3.067592e+00	1.382968	1.0	2.00	3.0	4.00	5.0

Fig 3. Statistical summary

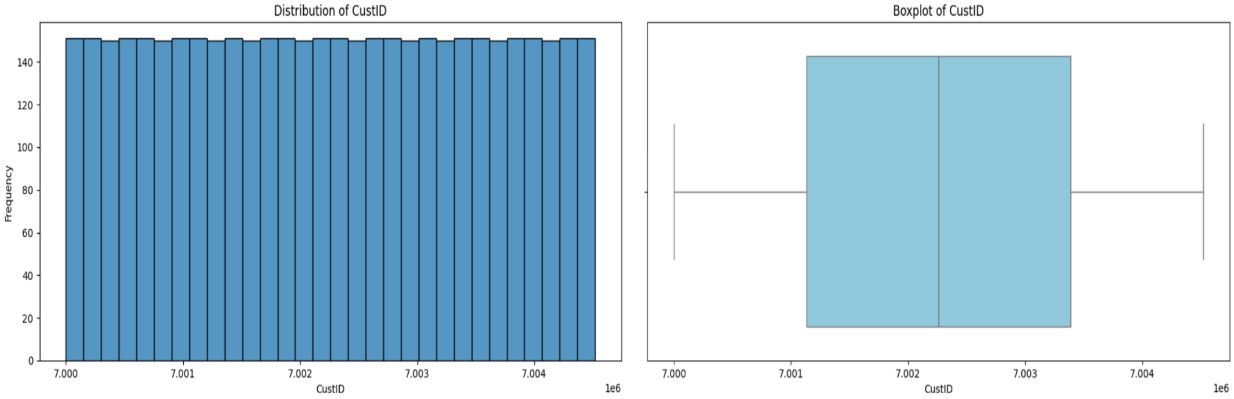
Exploratory Data Analysis (EDA)

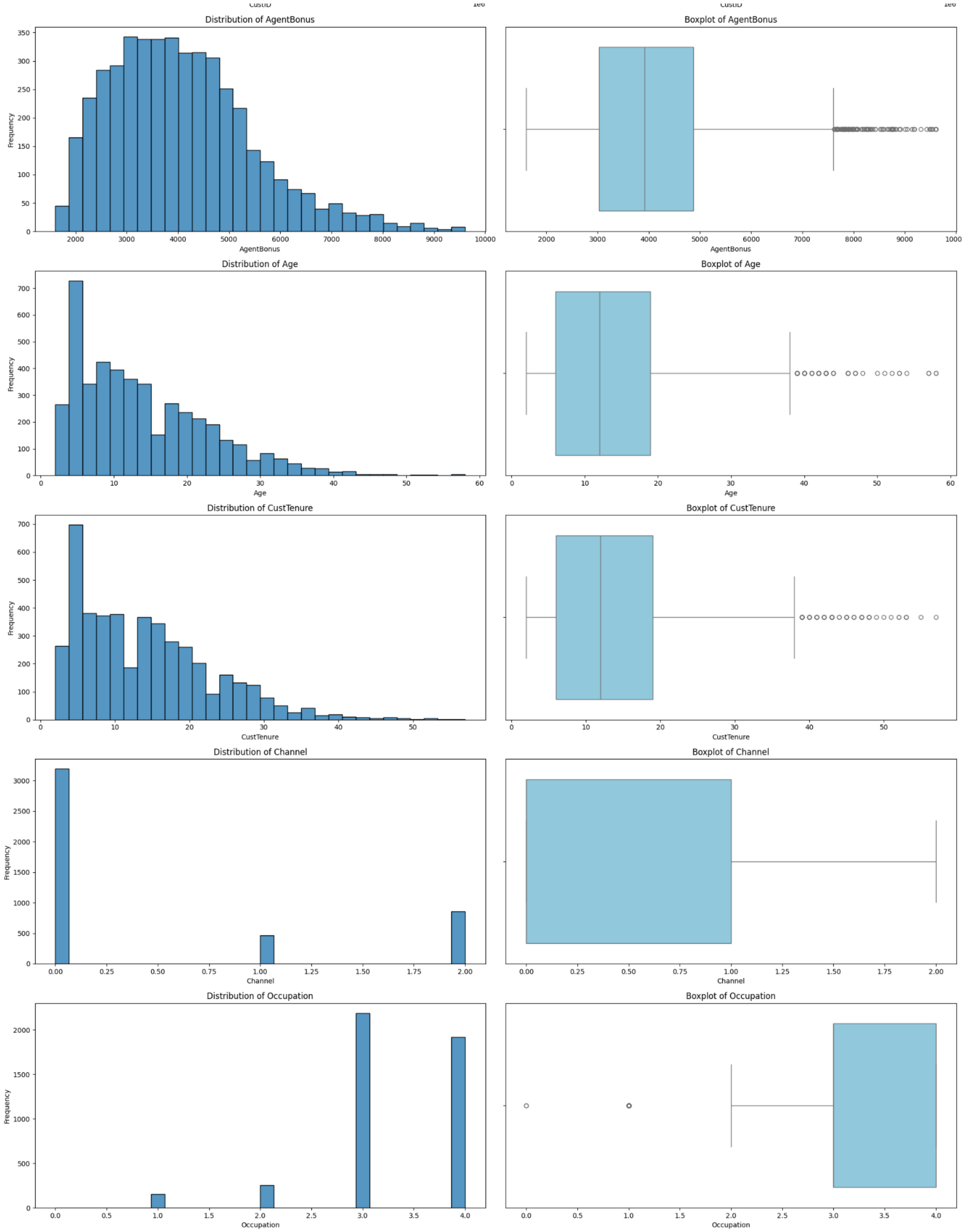
Exploratory Data Analysis (EDA) is a crucial first step in any data science project and often consumes up to 70% of the total project time. It lays the foundation for model building by helping us understand the structure, distribution, and relationships within the data. EDA not only reveals patterns and trends but also uncovers data quality issues that must be addressed before modeling.

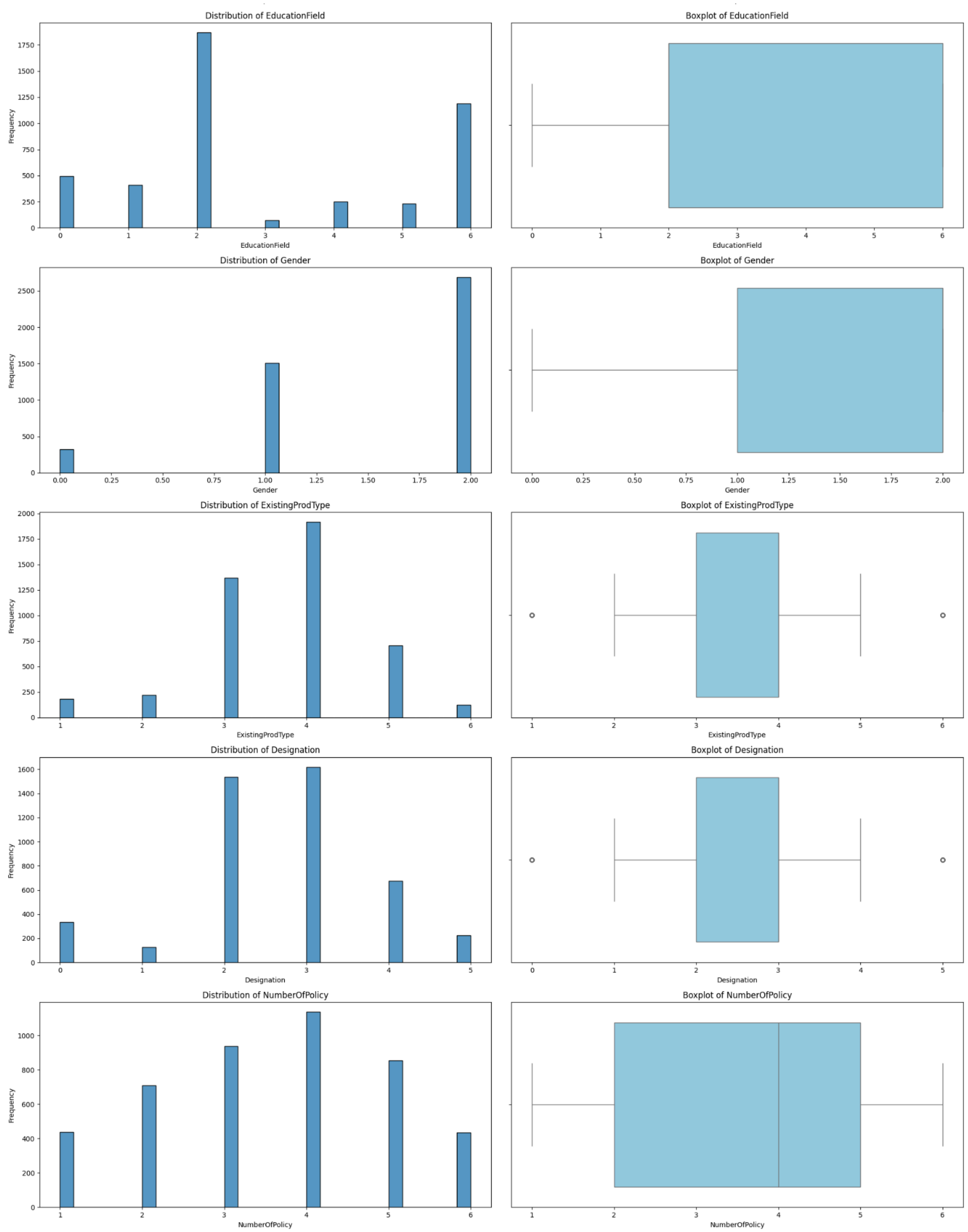
Key activities during EDA include:

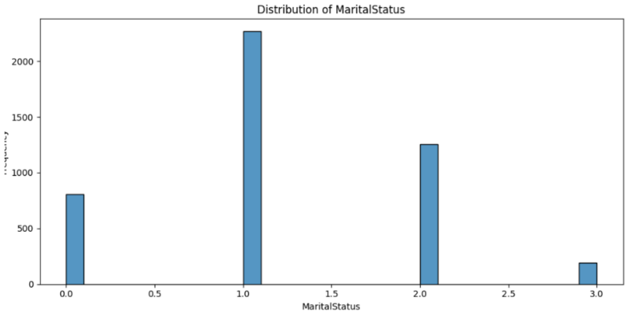
- Data visualization to understand distributions and relationships
- Data cleaning to handle missing or inconsistent values
- Filtering and feature exploration to focus on relevant information
- Outlier detection and treatment
- Mining hidden insights that may influence the target variable etc.

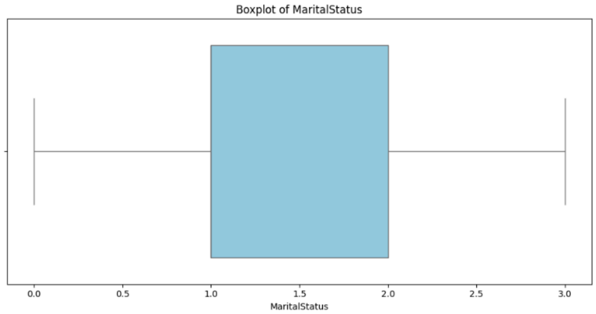
Univariate Analysis:

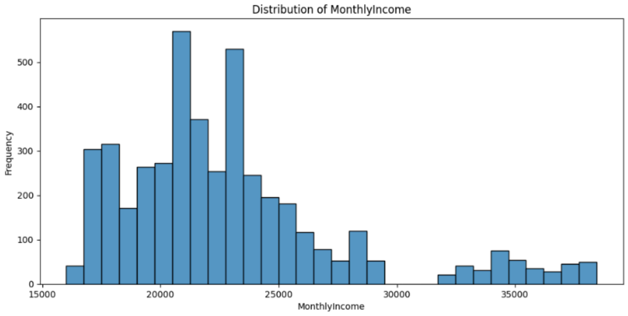


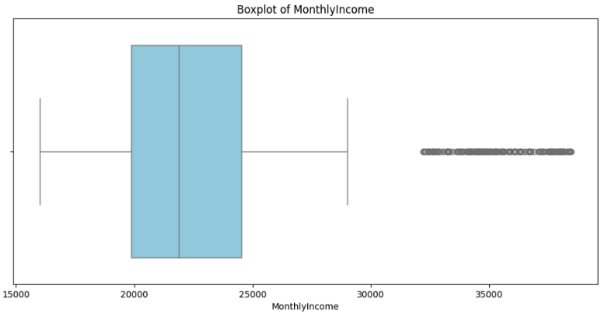

CustID
1e6
CustID
1e6
Distribution of AgentBonus
Frequency
0
50
100
150
200
250
300
350
2000
3000
4000
5000
6000
7000
8000
9000
10000
AgentBonus
Boxplot of AgentBonus
2000
3000
4000
5000
6000
7000
8000
9000
10000
AgentBonus
Distribution of Age
Frequency
0
100
200
300
400
500
600
700
0
10
20
30
40
50
60
Age
Boxplot of Age
0
10
20
30
40
50
60
Age
Distribution of CustTenure
Frequency
0
100
200
300
400
500
600
700
0
10
20
30
40
50
CustTenure
Boxplot of CustTenure
0
10
20
30
40
50
CustTenure
Distribution of Channel
Frequency
0
500
1000
1500
2000
2500
3000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Channel
Boxplot of Channel
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Channel
Distribution of Occupation
Frequency
0
500
1000
1500
2000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Occupation
Boxplot of Occupation
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Occupation

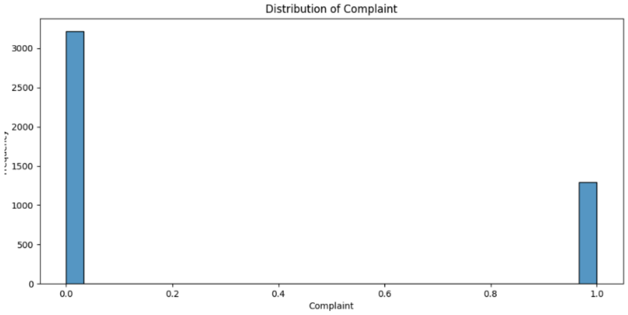

Distribution of EducationField
Frequency
0
250
500
750
1000
1250
1500
1750
0
1
2
3
4
5
6
EducationField
Boxplot of EducationField
0
1
2
3
4
5
6
EducationField
Distribution of Gender
Frequency
0
500
1000
1500
2000
2500
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Gender
Boxplot of Gender
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Gender
Distribution of ExistingProdType
Frequency
0
250
500
750
1000
1250
1500
1750
2000
1
2
3
4
5
6
ExistingProdType
Boxplot of ExistingProdType
1
2
3
4
5
6
ExistingProdType
Distribution of Designation
Frequency
0
200
400
600
800
1000
1200
1400
1600
0
1
2
3
4
5
Designation
Boxplot of Designation
0
1
2
3
4
5
Designation
Distribution of NumberOfPolicy
Frequency
0
200
400
600
800
1000
1
2
3
4
5
6
NumberOfPolicy
Boxplot of NumberOfPolicy
1
2
3
4
5
6
NumberOfPolicy

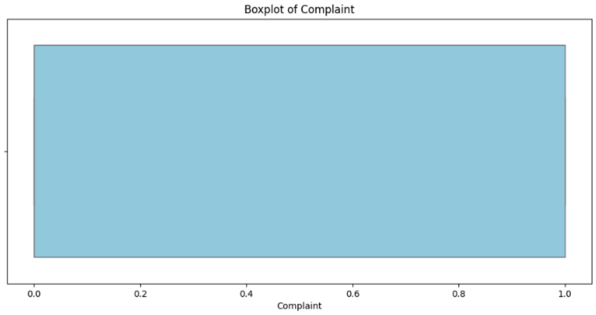

Distribution of MaritalStatus
Frequency
0
500
1000
1500
2000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MaritalStatus

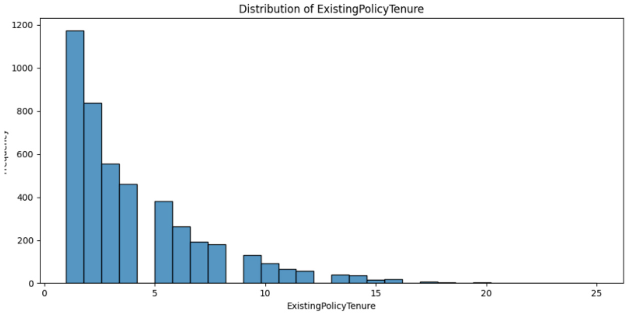

Boxplot of MaritalStatus
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MaritalStatus

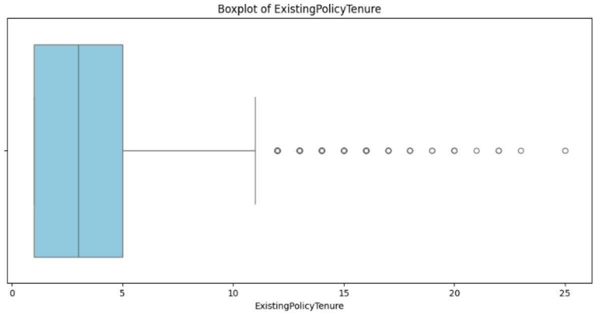

Distribution of MonthlyIncome
Frequency
0
100
200
300
400
500
15000
20000
25000
30000
35000
MonthlyIncome

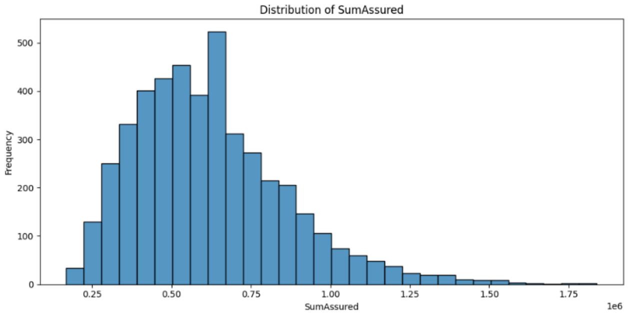

Boxplot of MonthlyIncome
15000
20000
25000
30000
35000
MonthlyIncome

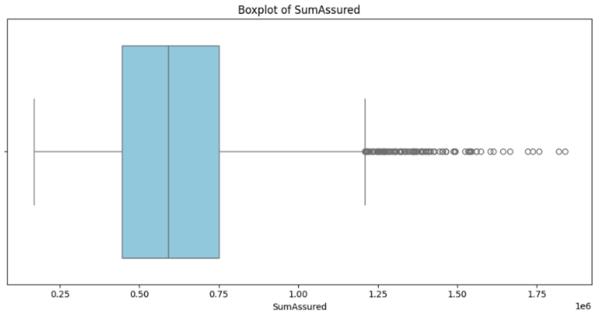

Distribution of Complaint
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
Complaint


Boxplot of Complaint
0.0
0.2
0.4
0.6
0.8
1.0
Complaint


Distribution of ExistingPolicyTenure
Frequency
0
200
400
600
800
1000
1200
0
5
10
15
20
25
ExistingPolicyTenure


Boxplot of ExistingPolicyTenure
0
5
10
15
20
25
ExistingPolicyTenure


Distribution of SumAssured
Frequency
0
100
200
300
400
500
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
SumAssured


Boxplot of SumAssured
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e6
SumAssured

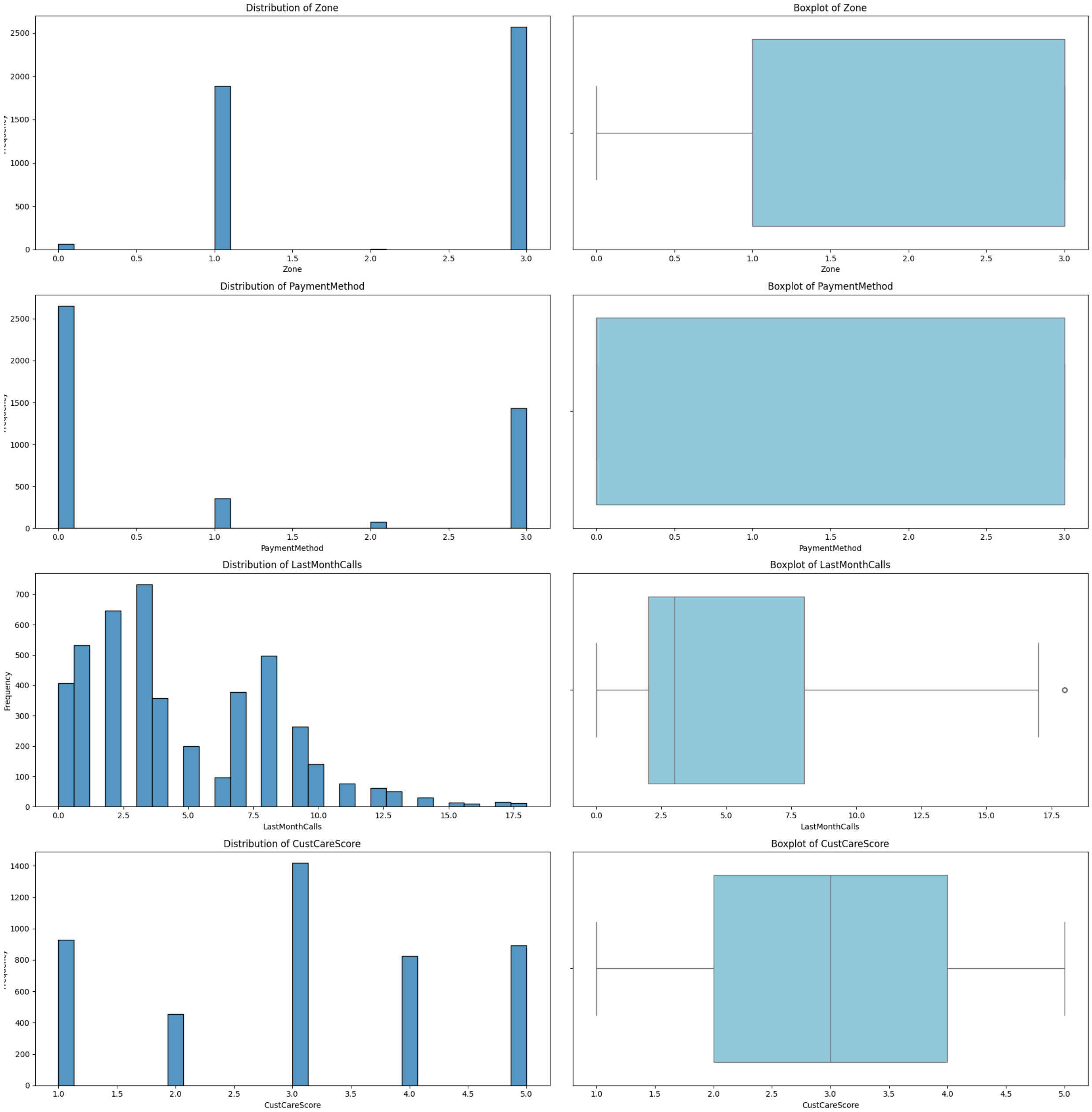


Fig 4. Univariate analysis

Observations from Univariate Analysis

- Most of the numerical variables exhibit positive skewness (right-skewed distribution), indicating that a large portion of values are concentrated toward the lower end, with a long tail on the higher side.

- Variables such as ExistingPolicyTenure and LastMonthCalls are discrete in nature, which explains the visible gaps between certain frequencies in their distributions.

- AgentBonus: The distribution shows clear right skewness, with the majority of values falling between 2000 to 6000 units.

- Age: The variable is positively skewed, with the majority of policyholders falling in the 5 to 20 years age range. The distribution also shows irregular fluctuations.

- CustTenure: Displays a right-skewed distribution similar to Age, with the bulk of customers having a tenure between 5 to 20 years.

- MonthlyIncome: This variable is also positively skewed and presents a bimodal pattern, with major concentration around 15,000–30,000 and 31,000–38,000 units.

- ExistingPolicyTenure: The data is right-skewed, with most values concentrated between 1 to 3 years, suggesting many customers have relatively new policies.

- SumAssured: Strong positive skewness is observed, with most customers having a sum assured in the range of 100,000 to 1,000,000 units.

- LastMonthCalls: The distribution appears bimodal and block-like, with clear groupings around 0–5 calls and 6–10 calls, and gaps between certain values due to its discrete nature.

Bivariate Analysis:

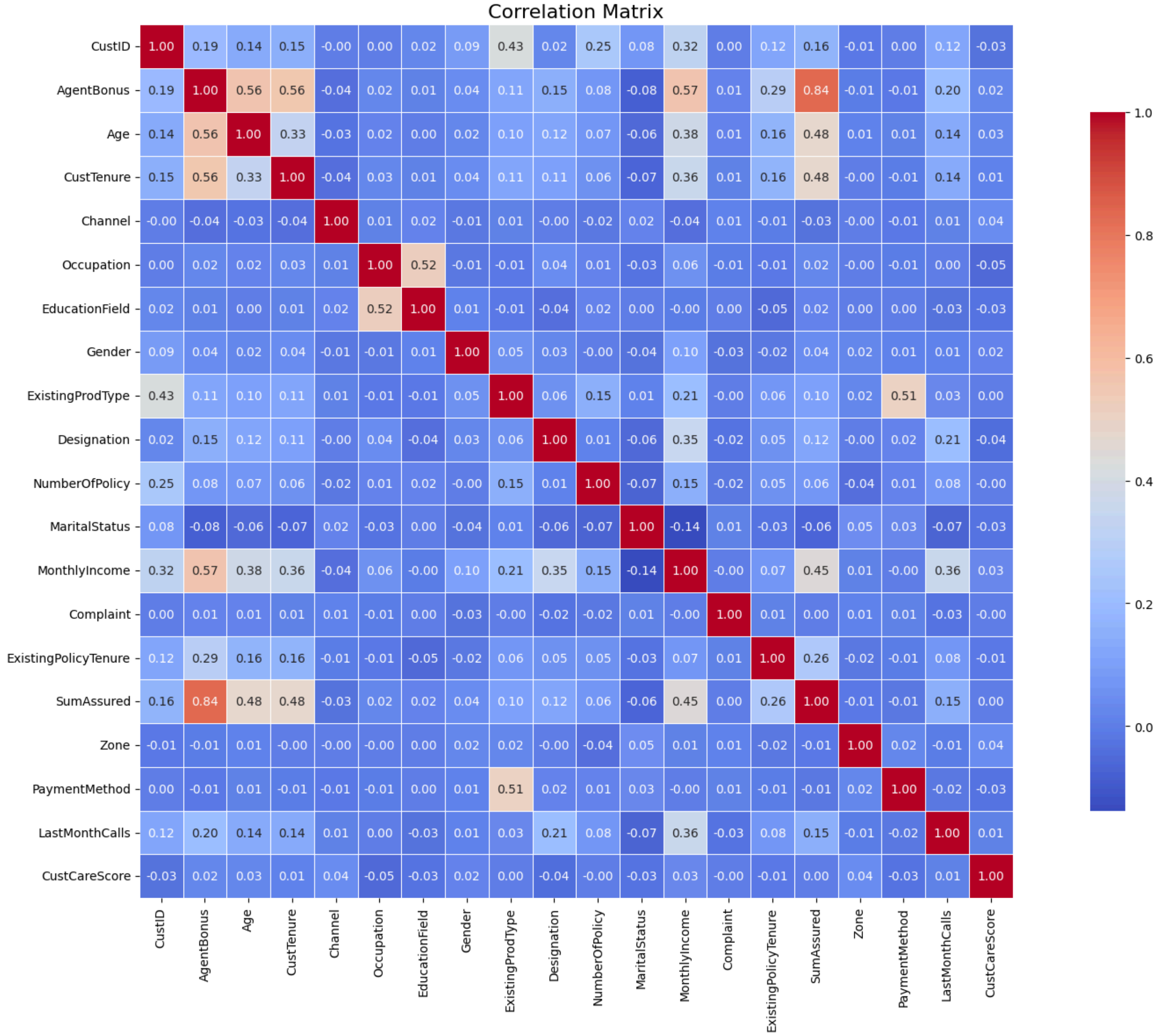


Fig 5. Bivariate Analysis (Heatmap)

Correlation Analysis (Heatmap Insights)

From the correlation heatmap, we can identify several variables that exhibit a strong positive correlation with the target variable AgentBonus. These include:

- SumAssured
- ExistingPolicyTenure
- MonthlyIncome

- CustTenure
- Age

Among these, the variable with the strongest correlation to AgentBonus is SumAssured. This suggests that the bonus awarded to agents is most strongly influenced by the total sum assured across policies, making it a critical factor in bonus determination.

This insight provides valuable direction for the company, as it highlights which customer or policy features most impact agent performance incentives.

Bivariate Analysis – Key Insights

Bivariate analysis was conducted between AgentBonus and the top correlated variables identified in the heatmap. The following observations were drawn from the resulting visualizations:

- SumAssured shows the strongest correlation with AgentBonus. As the sum assured increases, the agent's bonus also rises significantly. This indicates that higher-value policies lead to better incentives for agents.

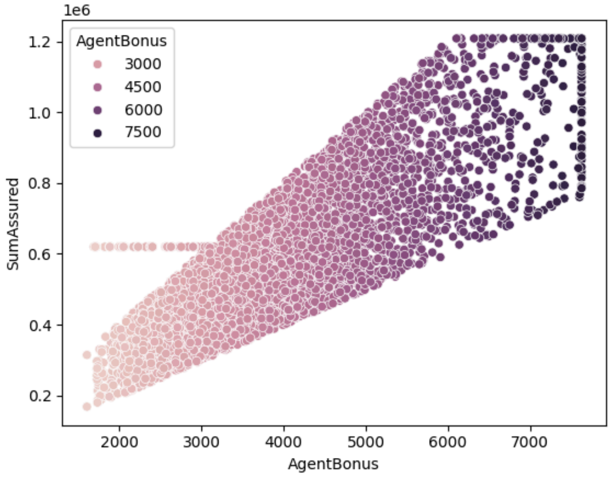


Fig 6. Sumassured vs Agentbonus

- Age of the customer exhibits an inverse relationship with agent bonus. While a large number of policies are sold to younger customers, these typically come with lower sums assured, resulting in lower agent bonuses.

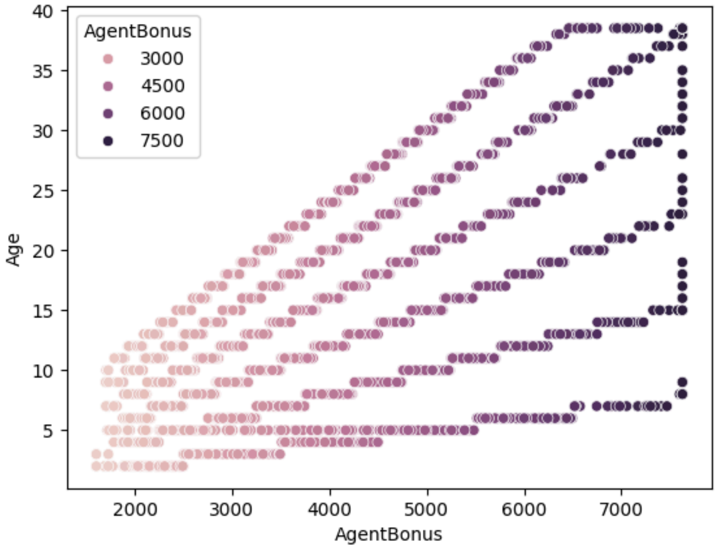


Fig 7. Age vs Agentbonus

- CustTenure (customer tenure) also shows a meaningful relationship with AgentBonus. Since life insurance premiums are typically recurring, longer customer retention results in continued revenue, incentivizing agents with sustained bonuses over time.

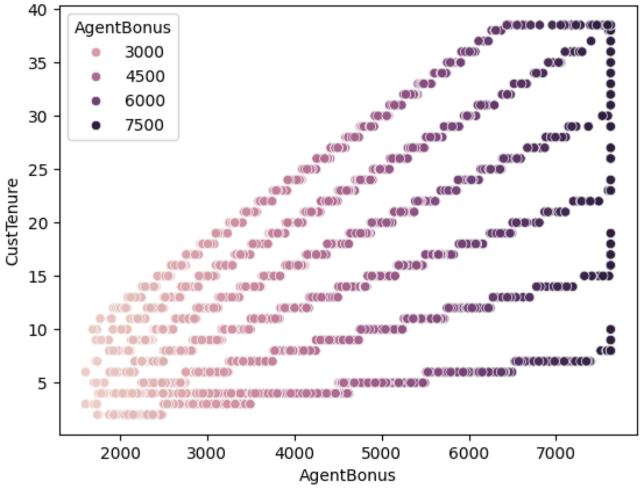


Fig 8. Custtenure vs Agentbonus

- MonthlyIncome is another influential variable. Customers with higher income levels tend to opt for higher coverage amounts (sum assured), which directly contributes to larger agent bonuses.

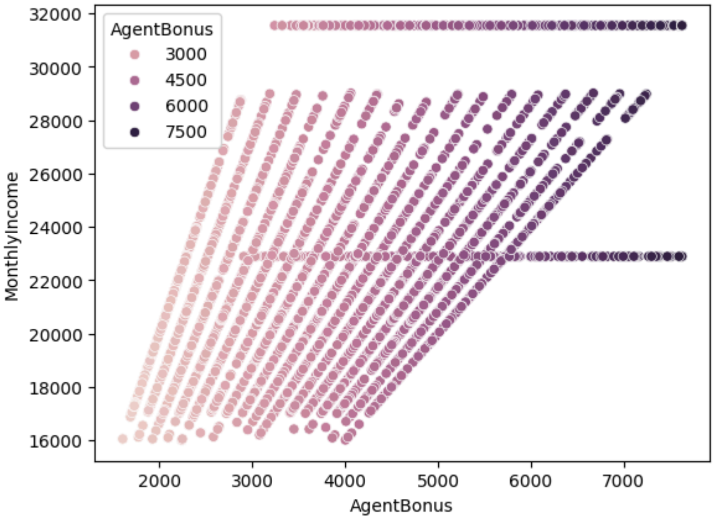


Fig 9.Monthlyincome vs Agentbonus

- ExistingPolicyTenure displays a moderate positive correlation, implying that the longer a customer's existing policy has been active, the longer an agent may continue to receive bonus payments tied to that policy.

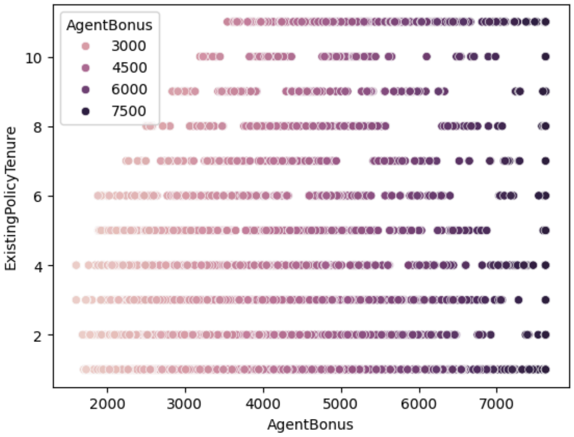
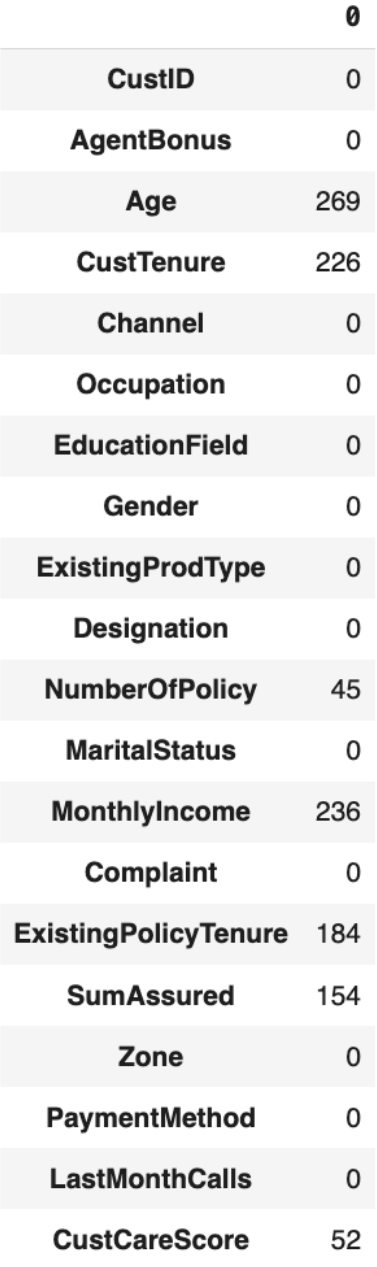


Fig 10. Existingpolicy vs Agentbonus

Missing Value Treatment:

We observe that there are many variables with missing values

Before missing value treatment



	0
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

Fig 11.Before missing value treatment

The final dataset used for modeling is free from missing values. The following strategies were used:

- **Mean Imputation** was applied to:
    - SumAssured
    - MonthlyIncome

- **Median Imputation** was applied to:

    - Age
    - CustTenure
    - ExistingPolicyTenure
    - CustCareScore
    - NumberOfPolicy

These methods were chosen to preserve the overall distribution of the data while minimizing distortion due to outliers or skewed variables.

Outlier Treatment

Outliers are data points that fall significantly outside the expected range, typically defined as values lying beyond 1.5 times the Interquartile Range (IQR) from the lower or upper quartiles.

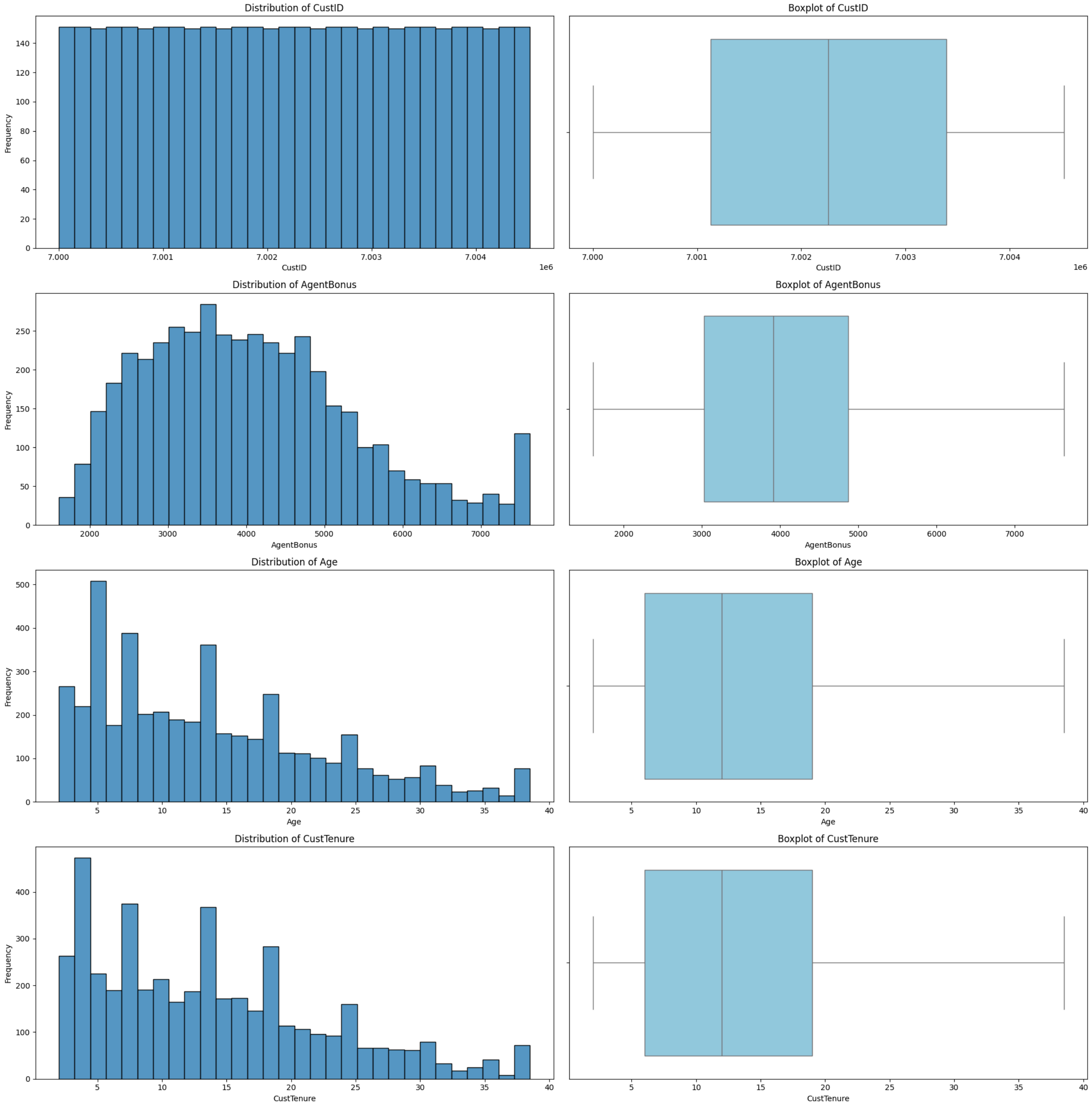
Based on the box plots analyzed during univariate analysis, several variables in the dataset were found to contain notable outliers. These extreme values can distort statistical models and reduce overall predictive accuracy, making outlier treatment a critical preprocessing step.

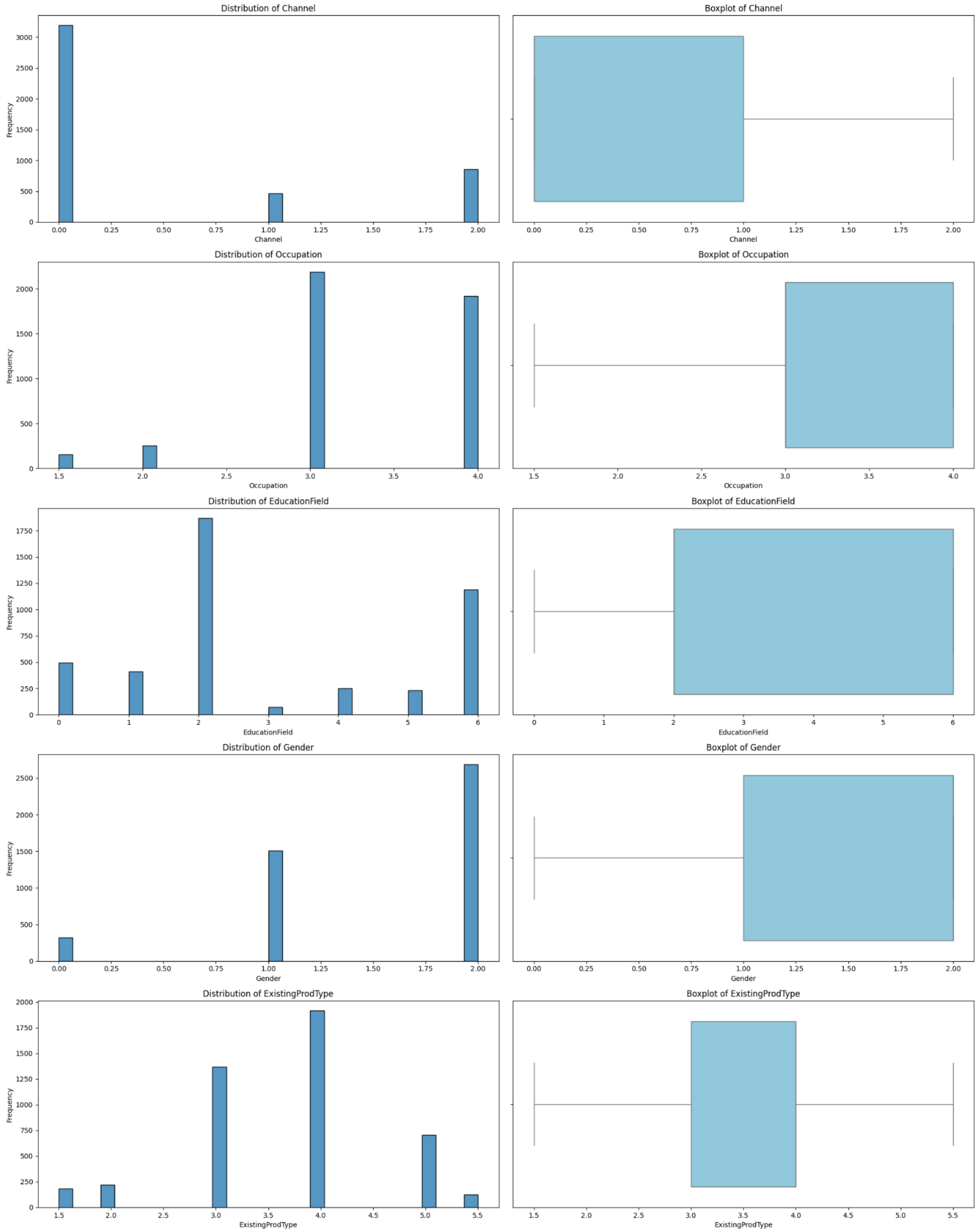
To address this, we used a capping (winsorization) technique:

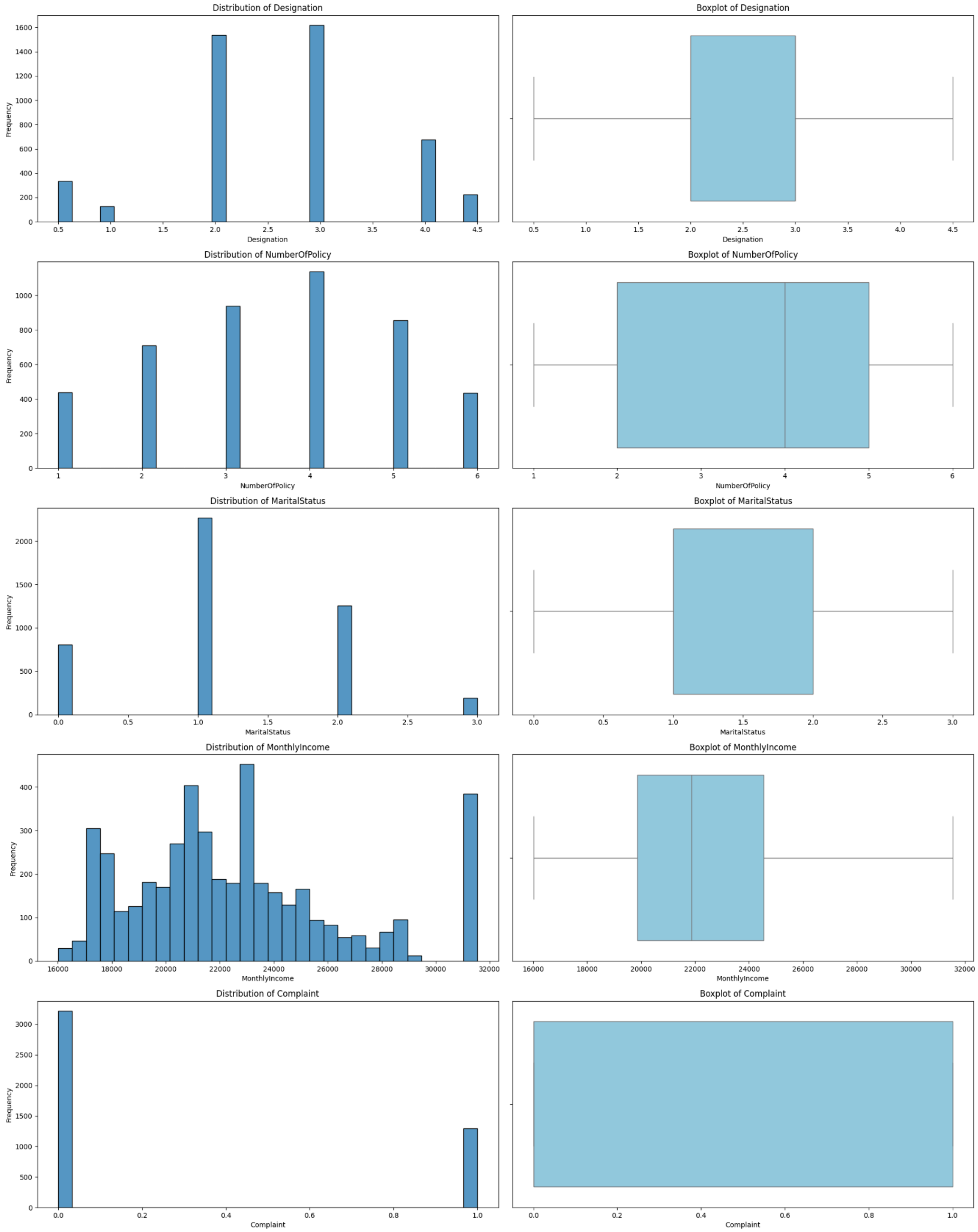
- Values below the lower bound were capped at the 5th percentile.

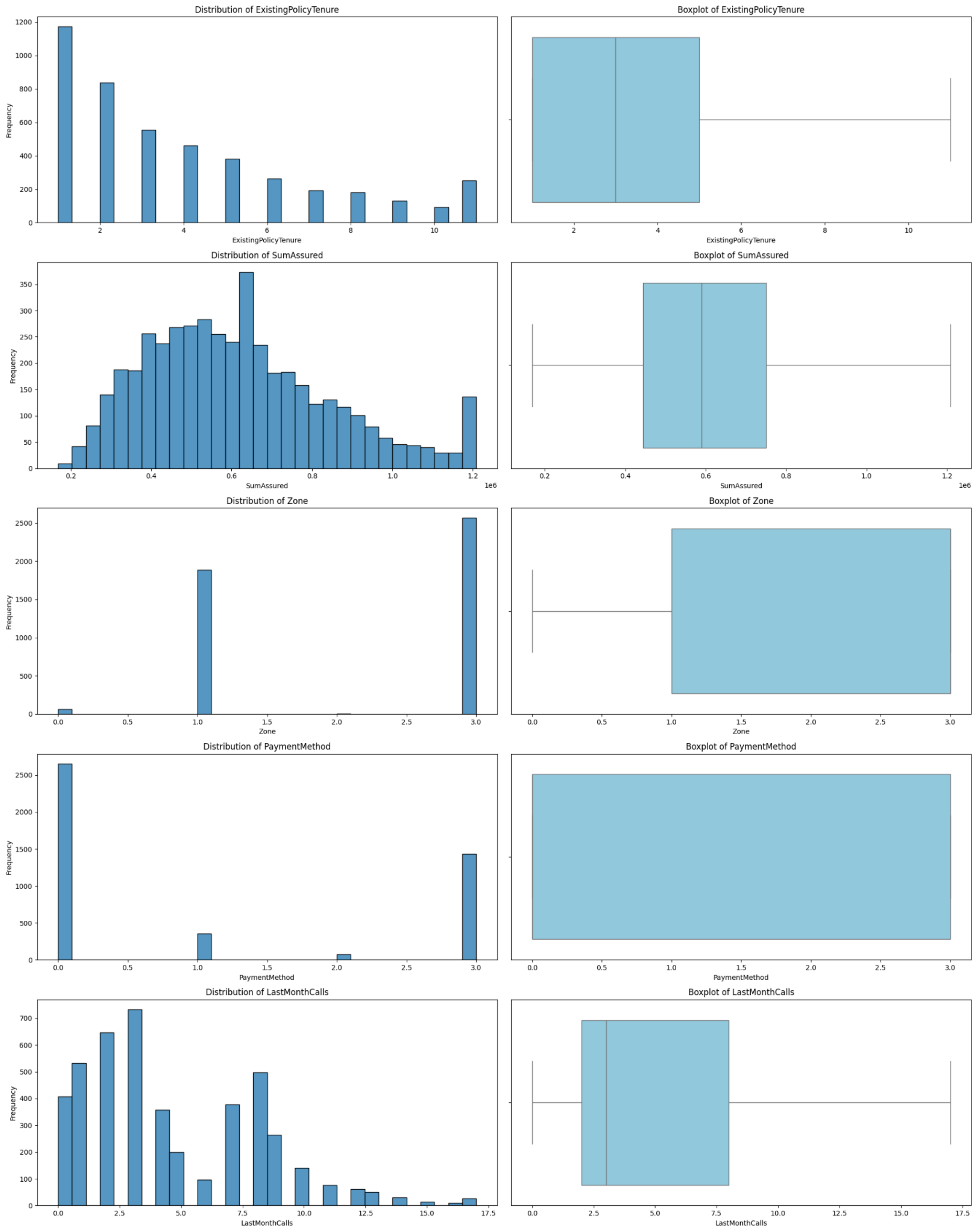
- Values above the upper bound were capped at the 95th percentile.

This method helps in minimizing the influence of extreme values while preserving the general structure and distribution of the data.




Distribution of Channel
Frequency
0
500
1000
1500
2000
2500
3000
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Channel
Boxplot of Channel
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Channel
Distribution of Occupation
Frequency
0
500
1000
1500
2000
1.5
2.0
2.5
3.0
3.5
4.0
Occupation
Boxplot of Occupation
1.5
2.0
2.5
3.0
3.5
4.0
Occupation
Distribution of EducationField
Frequency
0
250
500
750
1000
1250
1500
1750
0
1
2
3
4
5
6
EducationField
Boxplot of EducationField
0
1
2
3
4
5
6
EducationField
Distribution of Gender
Frequency
0
500
1000
1500
2000
2500
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Gender
Boxplot of Gender
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Gender
Distribution of ExistingProdType
Frequency
0
250
500
750
1000
1250
1500
1750
2000
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
ExistingProdType
Boxplot of ExistingProdType
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
ExistingProdType


Distribution of Designation
Frequency
0
200
400
600
800
1000
1200
1400
1600
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Designation
Boxplot of Designation
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Designation
Distribution of NumberOfPolicy
Frequency
0
200
400
600
800
1000
1
2
3
4
5
6
NumberOfPolicy
Boxplot of NumberOfPolicy
1
2
3
4
5
6
NumberOfPolicy
Distribution of MaritalStatus
Frequency
0
500
1000
1500
2000
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MaritalStatus
Boxplot of MaritalStatus
0.0
0.5
1.0
1.5
2.0
2.5
3.0
MaritalStatus
Distribution of MonthlyIncome
Frequency
0
100
200
300
400
16000
18000
20000
22000
24000
26000
28000
30000
32000
MonthlyIncome
Boxplot of MonthlyIncome
16000
18000
20000
22000
24000
26000
28000
30000
32000
MonthlyIncome
Distribution of Complaint
Frequency
0
500
1000
1500
2000
2500
3000
0.0
0.2
0.4
0.6
0.8
1.0
Complaint
Boxplot of Complaint
0.0
0.2
0.4
0.6
0.8
1.0
Complaint


Distribution of ExistingPolicyTenure
Frequency
0
200
400
600
800
1000
1200
2
4
6
8
10
ExistingPolicyTenure
Boxplot of ExistingPolicyTenure
2
4
6
8
10
ExistingPolicyTenure
Distribution of SumAssured
Frequency
0
50
100
150
200
250
300
350
0.2
0.4
0.6
0.8
1.0
1.2
1e6
SumAssured
Boxplot of SumAssured
0.2
0.4
0.6
0.8
1.0
1.2
1e6
SumAssured
Distribution of Zone
Frequency
0
500
1000
1500
2000
2500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Zone
Boxplot of Zone
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Zone
Distribution of PaymentMethod
Frequency
0
500
1000
1500
2000
2500
0.0
0.5
1.0
1.5
2.0
2.5
3.0
PaymentMethod
Boxplot of PaymentMethod
0.0
0.5
1.0
1.5
2.0
2.5
3.0
PaymentMethod
Distribution of LastMonthCalls
Frequency
0
100
200
300
400
500
600
700
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
LastMonthCalls
Boxplot of LastMonthCalls
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
LastMonthCalls

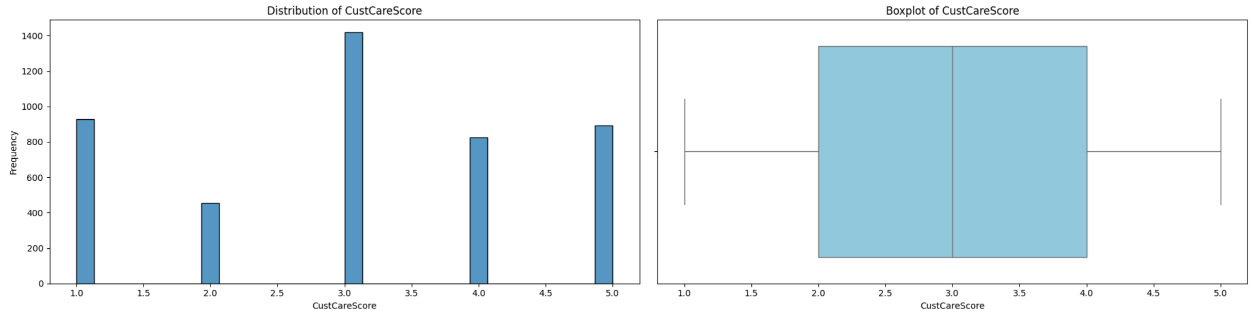


Fig 12. After Outlier Treatment

From the above graph we observe that all the outliers are removed from the data.

Variable Transformation

Encoding Categorical Variables

The dataset contains several categorical variables that need to be converted into a numerical format before modeling. For this purpose, Label Encoding was applied, which assigns a unique integer to each category level.

The following variables were encoded:

- Channel
- Occupation
- EducationField
- Gender
- Designation
- MaritalStatus
- Zone
- PaymentMethod

Label encoding was chosen for its simplicity and efficiency, especially since the model being considered can handle ordinal relationships where applicable.

Addition of New Variables

As this dataset is academic in nature and serves as a primary data source, we were limited to the available features. However, one important variable that could have significantly enhanced the model is the ‘Premium’ collected from the customer. Since agent bonuses are often directly tied to premiums paid, this variable would have added more predictive power in modeling the bonus structure accurately.

Basic Insights from EDA

Is the Data Imbalanced?

Yes, the dataset is significantly imbalanced, both in terms of numerical and categorical attributes. Key observations include:

- Age: A large proportion of policyholders are below 18 years, indicating a strong skew toward minors. Since minors generally have lower insurance premiums, this affects both revenue and agent bonus potential.

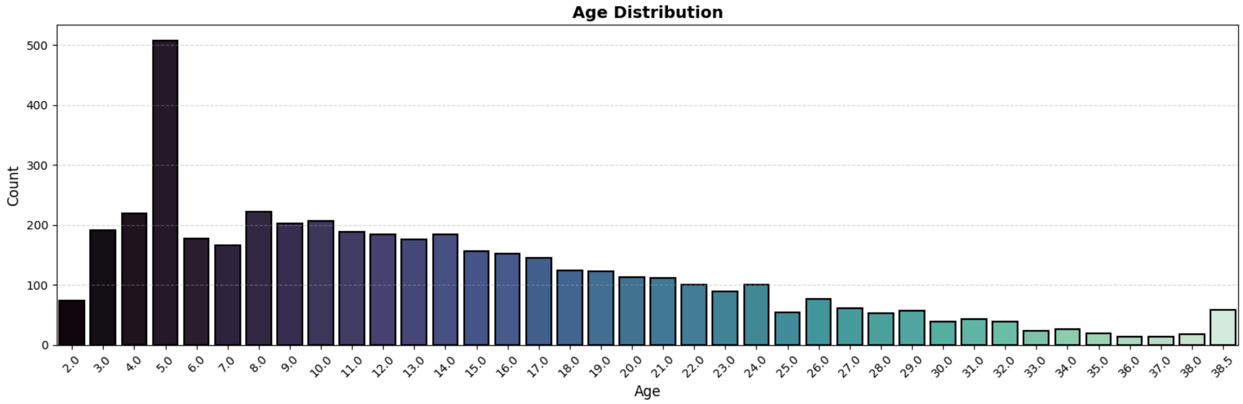


Fig 13. Age Distribution

- Zone: There is disproportionate representation across regions, West has the highest penetration, followed by North, while East and South are severely underrepresented.
 (Zone encoding: 0 - East, 1 - North, 2 - South, 3 - West)

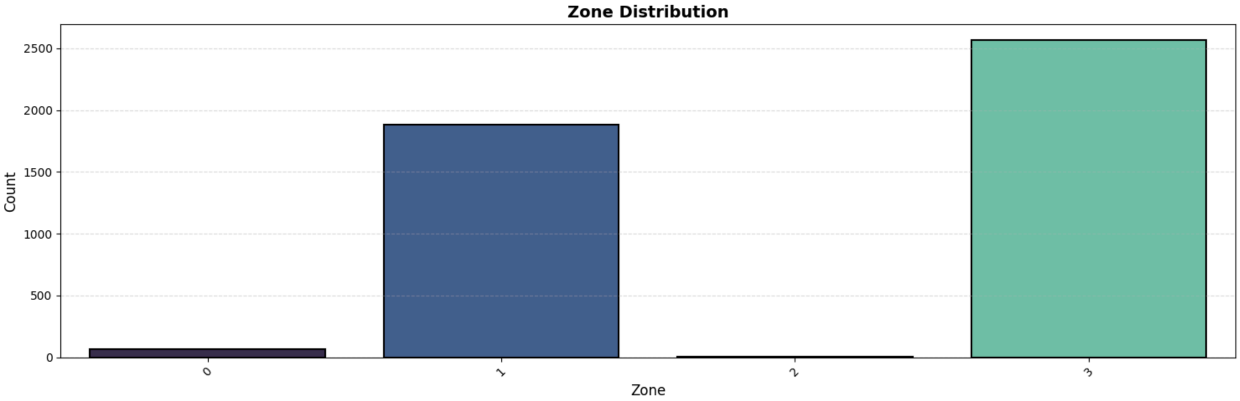


Fig 14. Zone Distribution

- CustTenure: The distribution shows most customers having tenures below 20 years, which reflects in shorter engagement and bonus periods for agents.

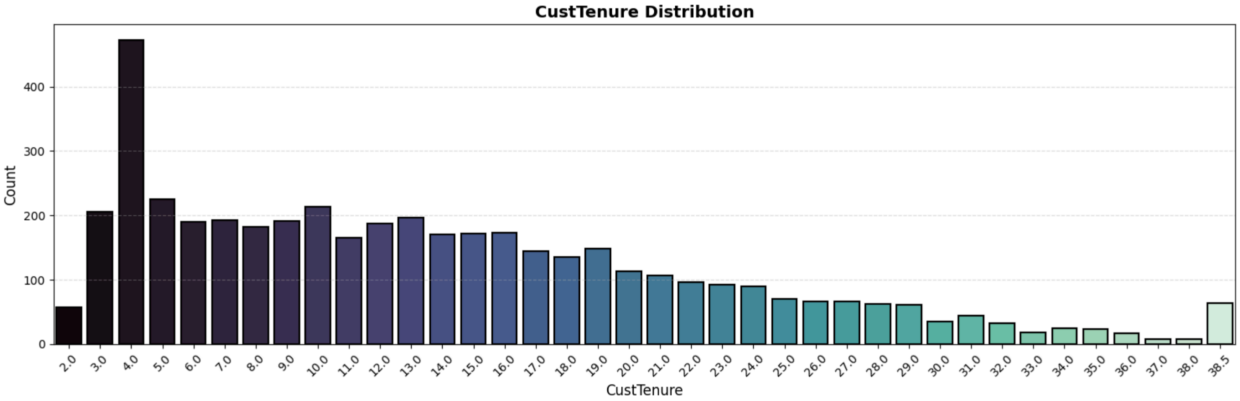


Fig 15. CustTenure Distribution

- Channel: Majority of customers are acquired through the Agent channel, while Online and Third Party Partner channels have minimal presence.
  (Channel encoding: 0 - Agent, 1 - Online, 2 - Third Party Partner)

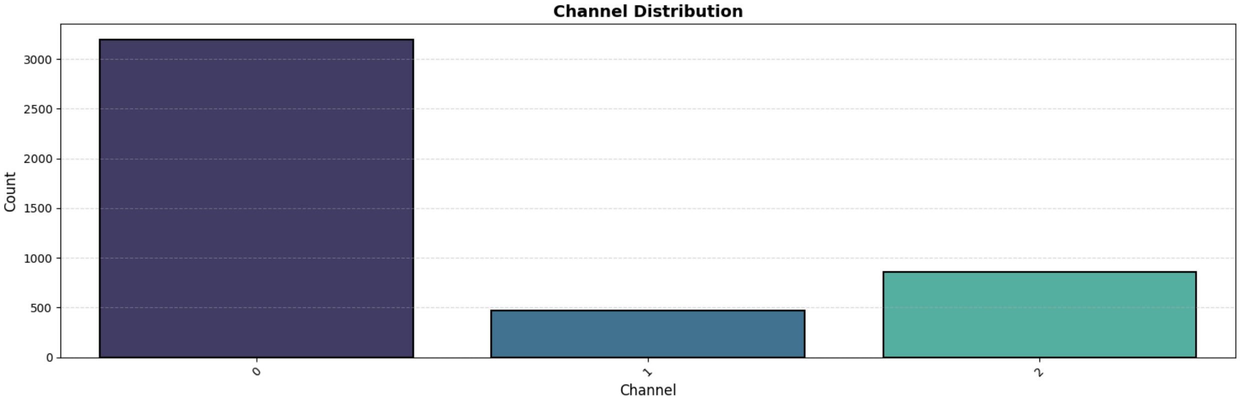


Fig 16. Channel Distribution

- PaymentMethod: Most customers opt for Half-Yearly and Yearly payment modes, while Monthly and Quarterly modes are rarely used.
  (Payment encoding: 0 - Half-Yearly, 1 - Monthly, 2 - Quarterly, 3 - Yearly)

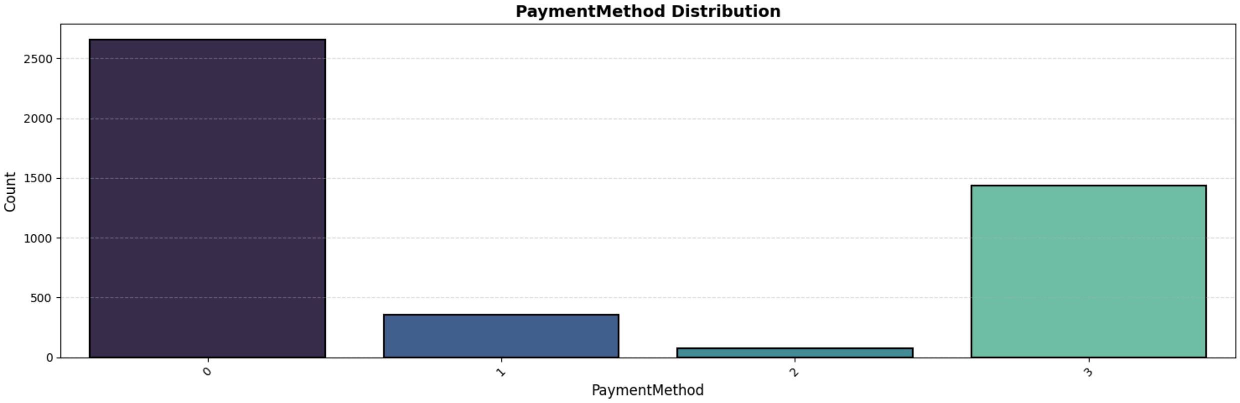


Fig 17. PaymentMethod Distribution

Addressing Data Imbalance

To address this imbalance and ensure fair and accurate modeling, the following techniques can be considered:

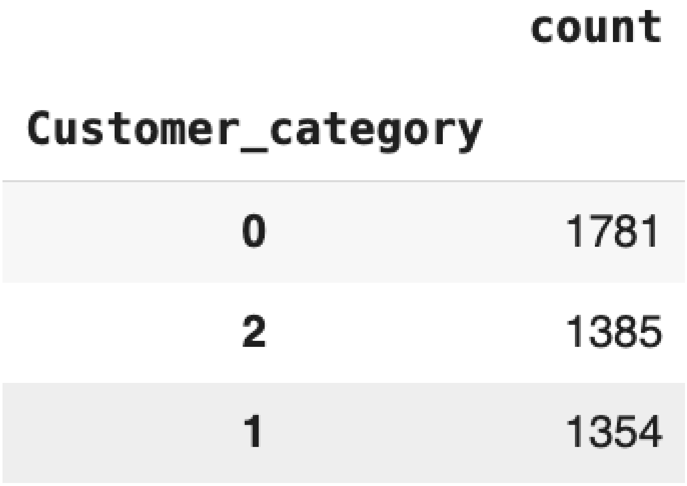
- Oversampling: Add more synthetic or duplicated samples from underrepresented groups (e.g., South/East zone, online channel).

- Undersampling: Reduce the number of observations from overrepresented groups to avoid model bias.

These methods are particularly important when building predictive models, as training on an imbalanced dataset can lead to biased decisions and poor generalization, especially in a business context like agent bonus prediction.

Business Insights from Clustering

To identify hidden patterns, K-Means Clustering was applied to the dataset, resulting in three distinct clusters:



	count
Customer_category	
0	1781
2	1385
1	1354

Fig 18. K-Means Clustering

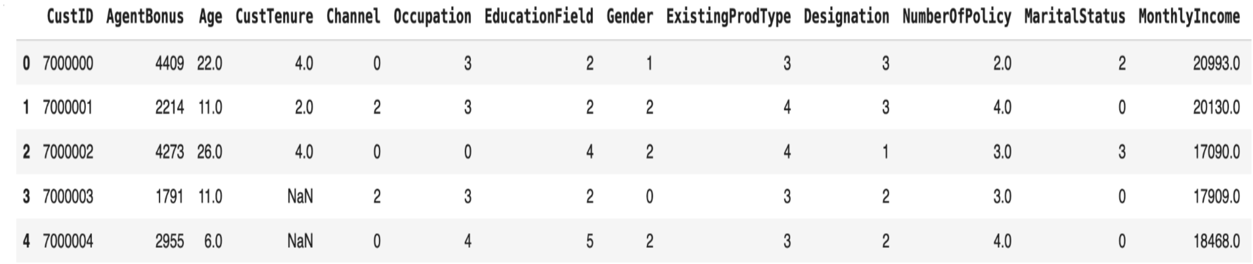
- Cluster 0 emerged as the largest segment, suggesting a common customer profile or behavior pattern within the majority.
- These clusters can help the company:
    - Design tailored engagement programs for different customer types.
    - Identify high-value agents or customers based on cluster behavior.
    - Align marketing and retention strategies to specific cluster profiles.

Other Key Business Insights

- The dominance of customers below age 20 implies low premiums and low mortality risk. While this leads to lower claims, it also results in limited profitability and agent bonuses.
- The maximum customer tenure is 22 years, indicating a product life cycle limitation. The company should consider introducing longer-term or lifetime products to improve customer retention and revenue.
- Digital adoption is low, the company should enhance its presence through online and mobile platforms, and also strengthen third-party partnerships to expand outreach.
- Regional imbalance suggests operational or marketing focus is lacking in the South and East zones. Expanding into these markets could unlock new business potential and customer segments.

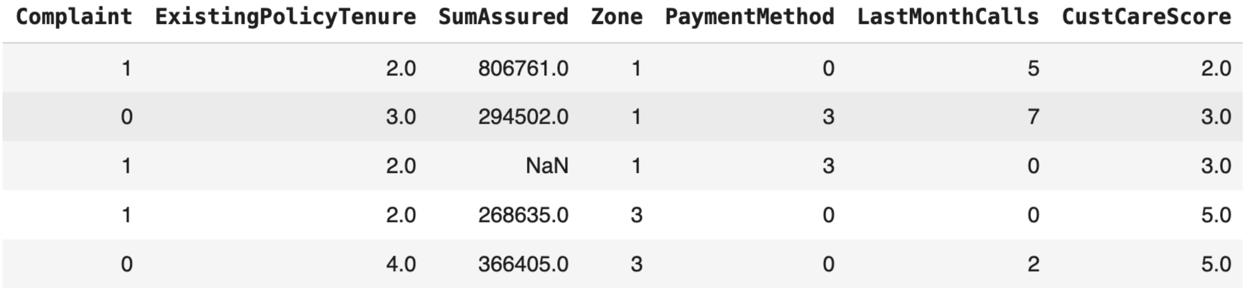
Model Building and Interpretation

The main task here is that, given the dataset of sales, we need to predict the AgentBonus using supervised learning algorithms. In the dataset we have almost around 20 features, in which few are categorical and few are numerical. There is no problem in numerical features, but the categorical features need to be converted/encoded. We can use one hot encoding or Label encoding. I have used the Label encoding and the columns of the first 5 rows of the dataset after encoding looks like this:



	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	MonthlyIncome
0	7000000	4409	22.0	4.0	0	3	2	1	3	3	2.0	2	20993.0
1	7000001	2214	11.0	2.0	2	3	2	2	4	3	4.0	0	20130.0
2	7000002	4273	26.0	4.0	0	0	4	2	4	1	3.0	3	17090.0
3	7000003	1791	11.0	NaN	2	3	2	0	3	2	3.0	0	17909.0
4	7000004	2955	6.0	NaN	0	4	5	2	3	2	4.0	0	18468.0

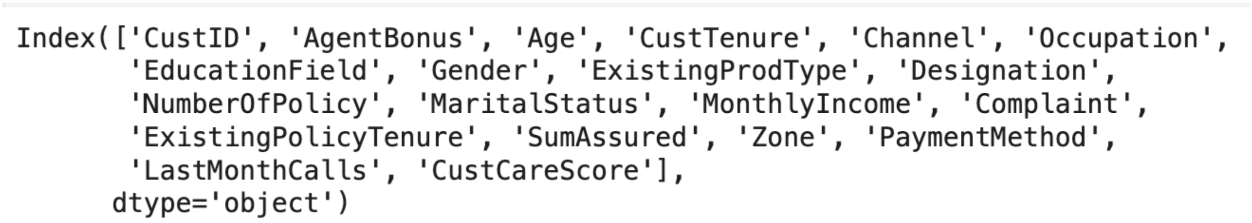
Fig 19. Label encoding First 5 rows



Complaint	ExistingPolicyTenure	SumAssured	Zone	PaymentMethod	LastMonthCalls	CustCareScore
1	2.0	806761.0	1	0	5	2.0
0	3.0	294502.0	1	3	7	3.0
1	2.0	NaN	1	3	0	3.0
1	2.0	268635.0	3	0	0	5.0
0	4.0	366405.0	3	0	2	5.0

Fig 20. Label encoding First 5 rows

Also there were many missing values, and outliers. As seen in notes1 we have already removed the missing values by replacing the numerical values by their means and the categorical values by their mode. Also we did box plots to see the outliers and removed them by clipping them to their 95th percentiles.



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

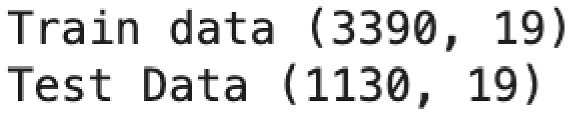
Fig 21. Columns after preprocessing

The above are the columns in the dataset after all the preprocessing.

For training the KNN regression model we need to scale the data, therefore I have done the normalization of the data so that they are in the same range.

MODEL BUILDING:

To evaluate the performance of our model, I have split the data set into a training and validation(test) set. I have split the dataset in the ratio of 75/25. Below we can see the number of records in the training and validation respectively.



```
Train data (3390, 19)
Test Data (1130, 19)
```

Evaluation metrics:

In the context of our model evaluation, **R² (R-squared)** represents the proportion of variance in the agent bonus that can be explained by the input features such as customer age, income, tenure, and sum assured. A higher R² value (closer to 1) indicates that the model fits the data well and effectively captures the relationship between variables. **RMSE (Root Mean Squared Error)**, on the other hand, measures the average difference between the predicted and actual bonus values, lower RMSE indicates better prediction accuracy. Together, these metrics help assess how reliably our model can forecast agent bonuses, enabling informed business decisions.

Linear Regression:

Linear Regression is one of the simplest and most widely used predictive modeling techniques. It helps us understand the relationship between one target variable (in this case, *Agent Bonus*) and several input features like *Age*, *Monthly Income*, *Sum Assured*, etc. The model tries to fit a straight line that best predicts the target variable based on the input features. In business terms, it helps us estimate how changes in factors like customer income or policy amount can impact the agent's bonus. The model is easy to interpret and useful for decision-making.

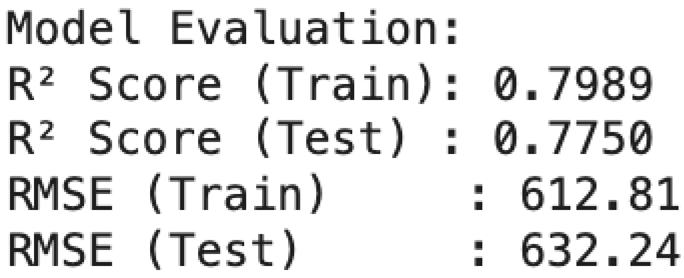
The mathematical form of the linear regression equation is:

$$\text{Agent Bonus} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Where:

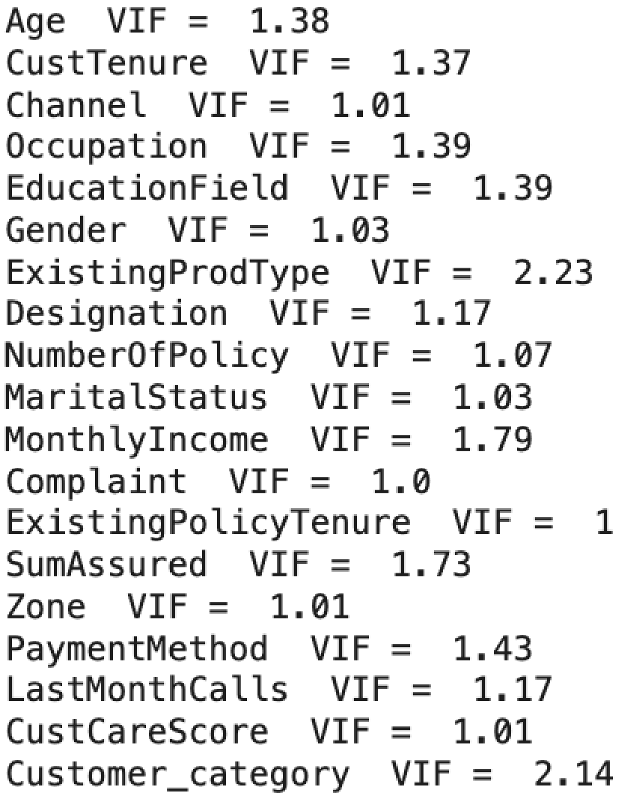
- β_0 is the intercept,
- β1,β2,...,βn are the coefficients for each input variable X_1,X_2,...,X_n
- ε is the error term (difference between actual and predicted values).

The evaluation score on the dataset is as below. We can see that we have achieved almost a 0.78 r2 score.



```
Model Evaluation:
R² Score (Train): 0.7989
R² Score (Test) : 0.7750
RMSE (Train)      : 612.81
RMSE (Test)       : 632.24
```

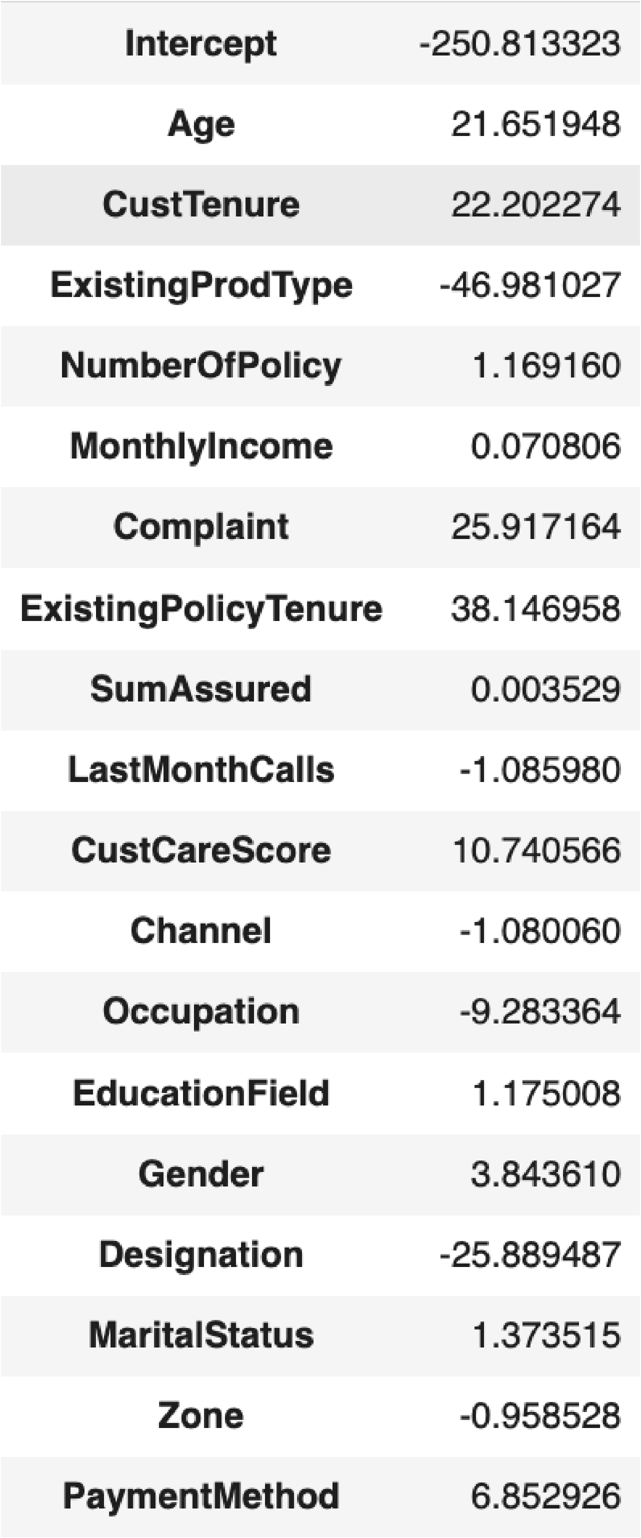
I have found the VIF for each feature too and the exact value of the coefficient of that feature. We see that all the features are under the control of high variance and hence we don’t drop any feature further.



```
Age  VIF =  1.38
CustTenure  VIF =  1.37
Channel  VIF =  1.01
Occupation  VIF =  1.39
EducationField  VIF =  1.39
Gender  VIF =  1.03
ExistingProdType  VIF =  2.23
Designation  VIF =  1.17
NumberOfPolicy  VIF =  1.07
MaritalStatus  VIF =  1.03
MonthlyIncome  VIF =  1.79
Complaint  VIF =  1.0
ExistingPolicyTenure  VIF =  1.1
SumAssured  VIF =  1.73
Zone  VIF =  1.01
PaymentMethod  VIF =  1.43
LastMonthCalls  VIF =  1.17
CustCareScore  VIF =  1.01
Customer_category  VIF =  2.14
```

Fig 22. Vif

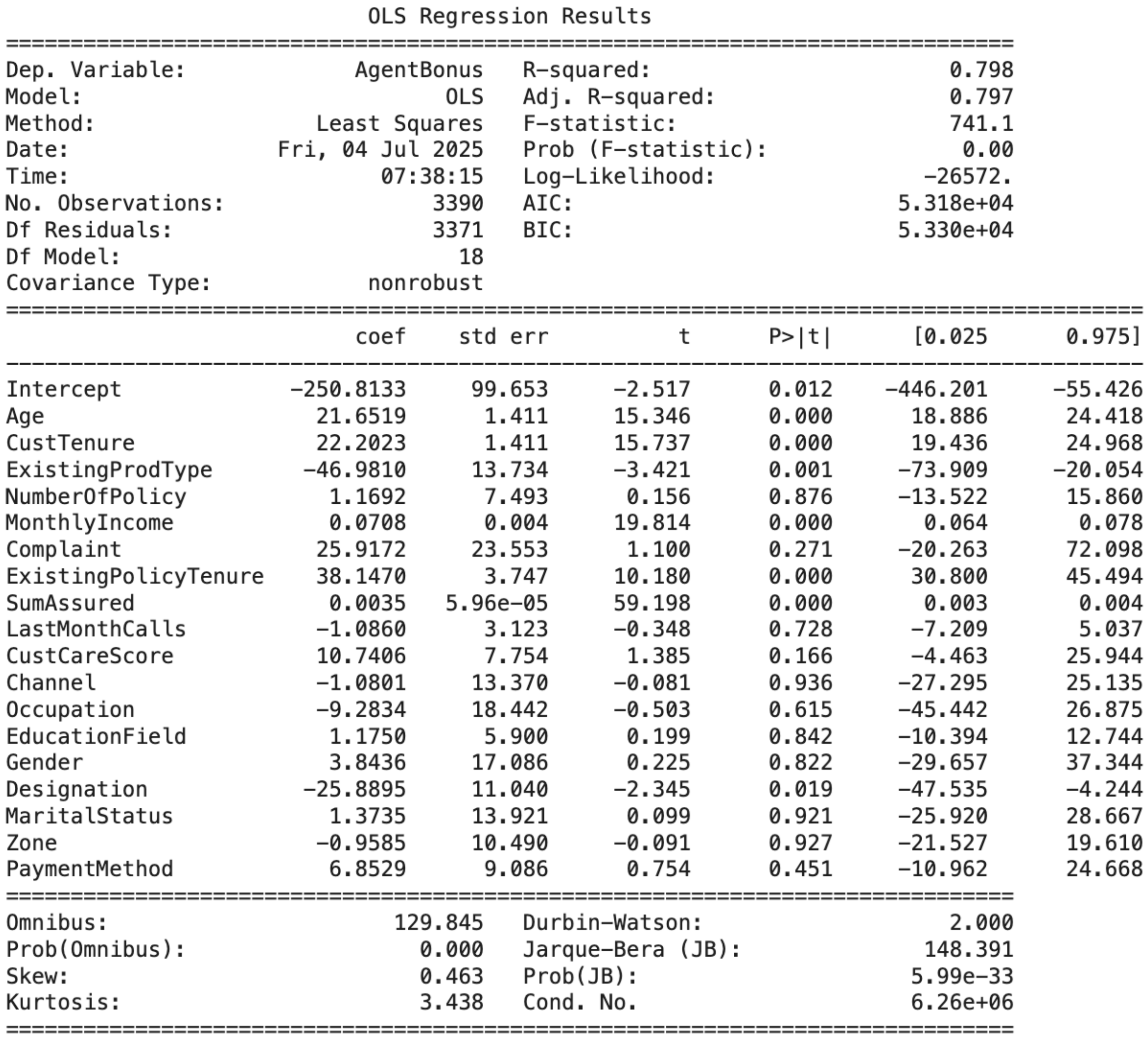
The values of the coefficient of each feature can be seen in the image below.



Intercept	-250.813323
Age	21.651948
CustTenure	22.202274
ExistingProdType	-46.981027
NumberOfPolicy	1.169160
MonthlyIncome	0.070806
Complaint	25.917164
ExistingPolicyTenure	38.146958
SumAssured	0.003529
LastMonthCalls	-1.085980
CustCareScore	10.740566
Channel	-1.080060
Occupation	-9.283364
EducationField	1.175008
Gender	3.843610
Designation	-25.889487
MaritalStatus	1.373515
Zone	-0.958528
PaymentMethod	6.852926

Fig 23. Value of the coefficient of each feature

Summary of OLS Model:



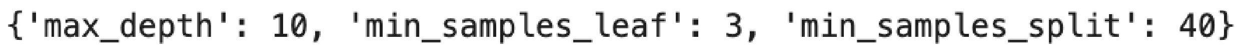
```
                            OLS Regression Results
==============================================================================
Dep. Variable:             AgentBonus   R-squared:                       0.798
Model:                            OLS   Adj. R-squared:                  0.797
Method:                 Least Squares   F-statistic:                     741.1
Date:                Fri, 04 Jul 2025   Prob (F-statistic):               0.00
Time:                        07:38:15   Log-Likelihood:                -26572.
No. Observations:                3390   AIC:                         5.318e+04
Df Residuals:                    3371   BIC:                         5.330e+04
Df Model:                          18
Covariance Type:            nonrobust
========================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept             -250.8133     99.653     -2.517      0.012    -446.201     -55.426
Age                     21.6519      1.411     15.346      0.000      18.886      24.418
CustTenure              22.2023      1.411     15.737      0.000      19.436      24.968
ExistingProdType       -46.9810     13.734     -3.421      0.001     -73.909     -20.054
NumberOfPolicy           1.1692      7.493      0.156      0.876     -13.522      15.860
MonthlyIncome            0.0708      0.004     19.814      0.000       0.064       0.078
Complaint               25.9172     23.553      1.100      0.271     -20.263      72.098
ExistingPolicyTenure    38.1470      3.747     10.180      0.000      30.800      45.494
SumAssured               0.0035   5.96e-05     59.198      0.000       0.003       0.004
LastMonthCalls          -1.0860      3.123     -0.348      0.728      -7.209       5.037
CustCareScore           10.7406      7.754      1.385      0.166      -4.463      25.944
Channel                 -1.0801     13.370     -0.081      0.936     -27.295      25.135
Occupation              -9.2834     18.442     -0.503      0.615     -45.442      26.875
EducationField           1.1750      5.900      0.199      0.842     -10.394      12.744
Gender                   3.8436     17.086      0.225      0.822     -29.657      37.344
Designation            -25.8895     11.040     -2.345      0.019     -47.535      -4.244
MaritalStatus            1.3735     13.921      0.099      0.921     -25.920      28.667
Zone                    -0.9585     10.490     -0.091      0.927     -21.527      19.610
PaymentMethod            6.8529      9.086      0.754      0.451     -10.962      24.668
==============================================================================
Omnibus:                      129.845   Durbin-Watson:                   2.000
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              148.391
Skew:                           0.463   Prob(JB):                     5.99e-33
Kurtosis:                       3.438   Cond. No.                     6.26e+06
==============================================================================
```

Fig 24. OLS Model Summary

Decision Tree Regression:

Decision Tree Regression is a machine learning technique that splits the data into smaller and smaller groups based on feature values, forming a tree-like structure. At each decision point (or node), the model asks a yes/no question like *Is the customer's income greater than ₹30,000?* and based on the answer, it moves down the appropriate branch. This process continues until the model reaches a final prediction value at a leaf node.

There are many parameters in the decision tree like max_depth, min_samples_leaf and min_samples_split and we have to find the value of these parameters. I have used grid search method to find these values and the result is as below:

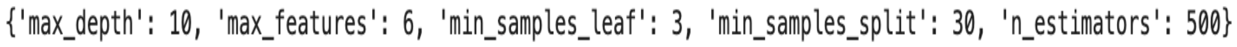


```
{'max_depth': 10, 'min_samples_leaf': 3, 'min_samples_split': 40}
```

Random Forest Regression:

Random Forest Regression is an advanced machine learning technique that builds on decision trees. Instead of relying on a single decision tree, it creates a "forest" of many decision trees, each trained on different random parts of the data. When a prediction is needed, like estimating an agent's bonus, the model takes the average of predictions from all the trees, which leads to more accurate and stable results.

There are many parameters in the decision tree like max_depth, max_features, min_samples_leaf, min_samples_split and n_estimators and we have to find the value of these parameters. I have used grid search method to find these values and the result is as below:

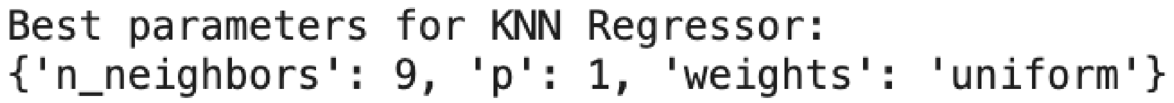


```
{'max_depth': 10, 'max_features': 6, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 500}
```

KNN Regression

KNN Regression is a simple, yet effective machine learning algorithm that predicts the target value (like *Agent Bonus*) based on the values of the 'K' most similar data points in the dataset. It doesn’t make any assumptions about the underlying relationship between features, it simply looks for the “nearest neighbors” in terms of input features like *Age*, *Monthly Income*, *Sum Assured*, etc., and takes the average of their bonus values to make a prediction.

There are many parameters in the decision tree like n_neighbors, weights and p and we have to find the value of these parameters. I have used grid search method to find these values and the result is as below:



```
Best parameters for KNN Regressor:
{'n_neighbors': 9, 'p': 1, 'weights': 'uniform'}
```

Model performance comparisons

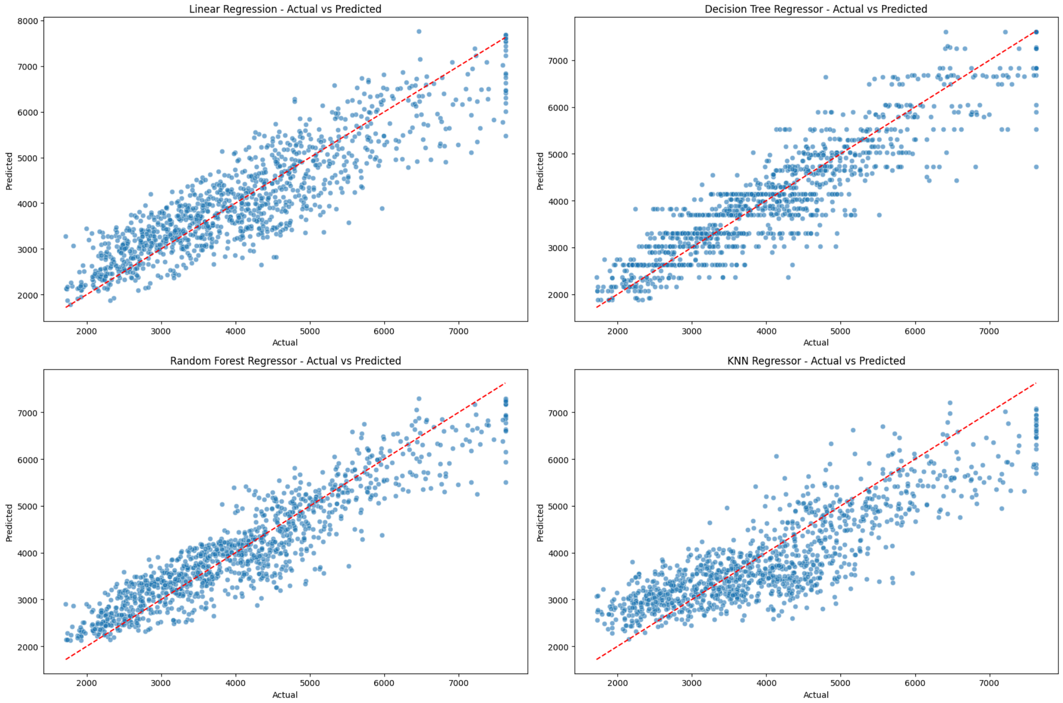
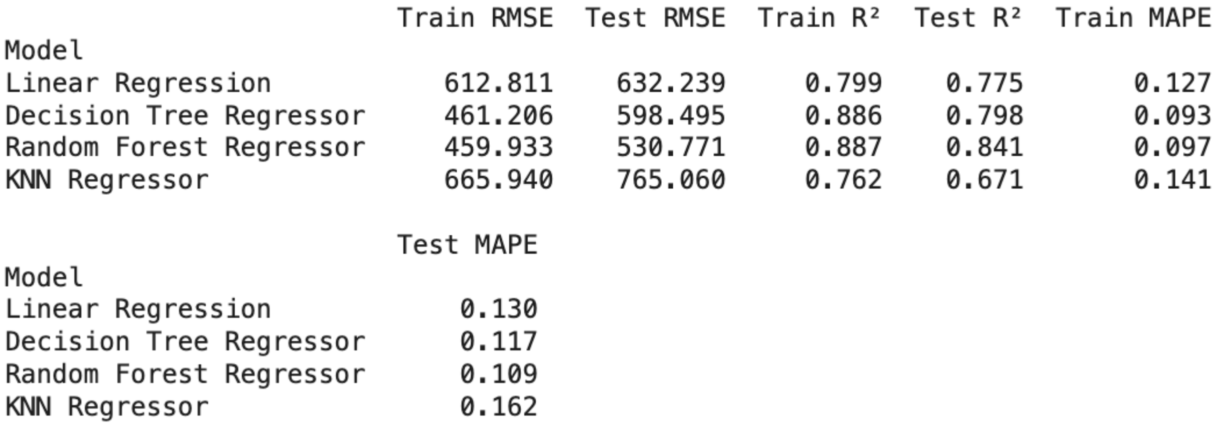


Fig 25. Model performance graph



```
                         Train RMSE  Test RMSE  Train R²  Test R²  Train MAPE
Model
Linear Regression           612.811    632.239     0.799    0.775       0.127
Decision Tree Regressor     461.206    598.495     0.886    0.798       0.093
Random Forest Regressor     459.933    530.771     0.887    0.841       0.097
KNN Regressor               665.940    765.060     0.762    0.671       0.141

                         Test MAPE
Model
Linear Regression            0.130
Decision Tree Regressor      0.117
Random Forest Regressor      0.109
KNN Regressor                0.162
```

Fig 26. Model performance Scores

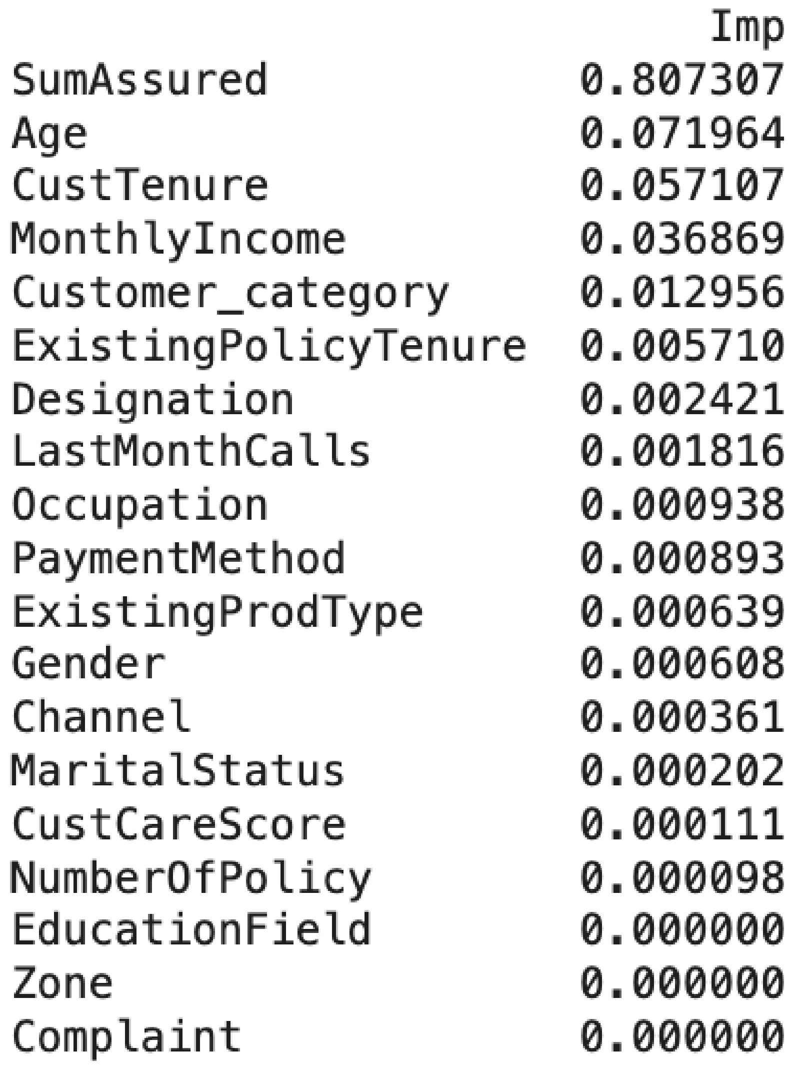
We can see that KNN is overfitting.

We can also see that the test score of random forest is the best compared to other models.

Feature Importance:

Since random forest performed the most better among the other models we would proceed with this model for testing. Before that I saw the importance of each feature for both decision tree and random forest. It is there below:

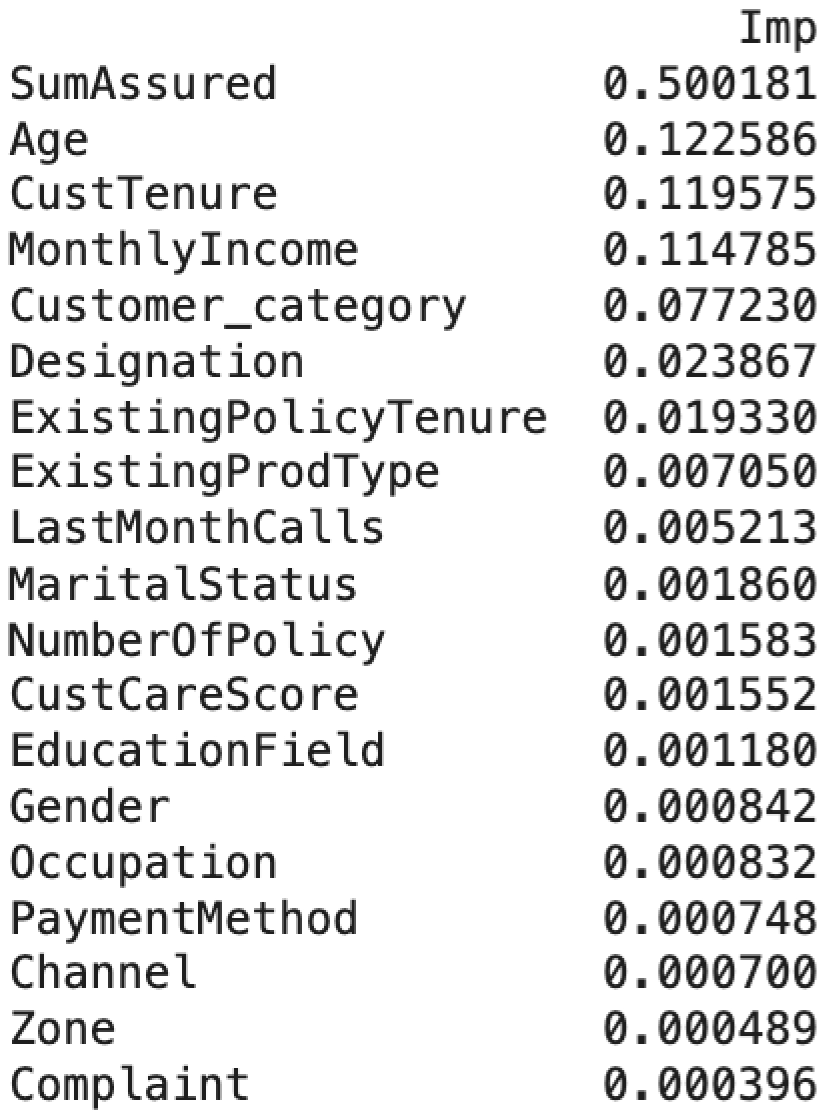
Decision Tree:



```
                             Imp
SumAssured              0.807307
Age                     0.071964
CustTenure              0.057107
MonthlyIncome           0.036869
Customer_category       0.012956
ExistingPolicyTenure    0.005710
Designation             0.002421
LastMonthCalls          0.001816
Occupation              0.000938
PaymentMethod           0.000893
ExistingProdType        0.000639
Gender                  0.000608
Channel                 0.000361
MaritalStatus           0.000202
CustCareScore           0.000111
NumberOfPolicy          0.000098
EducationField          0.000000
Zone                    0.000000
Complaint               0.000000
```

Fig 27. Feature Importance Decision Tree:

Random Forest:



```
                            Imp
SumAssured             0.500181
Age                    0.122586
CustTenure             0.119575
MonthlyIncome          0.114785
Customer_category      0.077230
Designation            0.023867
ExistingPolicyTenure   0.019330
ExistingProdType       0.007050
LastMonthCalls         0.005213
MaritalStatus          0.001860
NumberOfPolicy         0.001583
CustCareScore          0.001552
EducationField         0.001180
Gender                 0.000842
Occupation             0.000832
PaymentMethod          0.000748
Channel                0.000700
Zone                   0.000489
Complaint              0.000396
```

Fig 28. Feature Importance Random Forest:

Interpretation and Business Recommendations

The objective of this analysis was to help the company predict the ideal bonus for agents and to understand what differentiates high-performing agents from low-performing ones. Based on the models and variable importance, Sum Assured emerged as the most influential factor in determining agent bonuses. This suggests that agents closing high-value policies are likely to receive higher bonuses.

For example, customers with a VP designation tend to buy more or higher-value policies, indicating that targeting such profiles can increase performance. Similarly, Customer Tenure plays a significant role, customers who have been associated with the company for 8-20 years form the largest and most valuable group, making them a key focus area for agents. Another important driver is Monthly Income, as customers with higher incomes are more likely to opt for policies with greater sum assured.

Recommendations:

- **For high-performing agents**, the company can launch performance-based contests with clear targets. Incentives could include high-end gadgets, vacation packages, or bonus perks to maintain engagement and motivation.

- **For low-performing agents**, personalized upskilling programs and feedback mechanisms should be implemented. These should focus on improving their ability to target and convert high-value prospects, particularly those in the optimal income and tenure segments.

- To improve predictive power and insights, the company should consider adding more variables such as:

    - **Premium amount collected** – a direct financial indicator that aligns closely with bonuses.

    - **Geographic region** – detailed location data could help better understand customer segmentation beyond general zones.

    - **Agent ID** – to track and monitor individual agent performance more accurately and design targeted interventions.

By acting on these insights, the company can not only predict agent bonuses more accurately but also create more focused engagement strategies that drive sales performance and customer satisfaction.