CAPSTONE PROJECT NOTES 1
LIFE INSURANCE DATA

By Apoorva p

June 9th 2024

LI_BFSI_01

CONENTS

LIST OF FIGURES

Introduction of the business problem

Defining problem statement

The dataset belongs to a leading life insurance company. The company wants to predict the bonus for its agents so that it may design appropriate engagement activity for their high performing agents and upskill programs for low performing agents.

Need of the study/project

The need of study is as given below:-

- **Predict Agent Bonus:**

Objective: Predicting the bonus for agents using different features related to customer and agent.
Advantage: Predictions that are on point will find your top performers who deserve higher bonuses and more acknowledgment, while they can alert from poor performers that might require more guidance to ensure their performance does not drop lower.

- **Enhance Agent Performance:**

Objective: Structure specific actions of engagement in performance for high performance agents to remain in the same performance buckets or move up the performance buckets.
Advantage: Keeping top agents motivated, engaged, and productive in order to provide customers with the highest level of service and generate dollars for the company.

- **Implement Upskill Programs:**

Objective: Create upskill programs for low-performing agents from the data produced by the model predictions.
Advantage: Targeted training and development can enable these agents to perform better, which in turn, will help improve the overall sales and customer satisfaction.

- **Resource Allocation:**

Objective: To allocate resources for training and engagement activities in a smarter way by segregating performance predictions of the agents.
Advantage: Maximize return on human capital investment by concentrating efforts on areas of most significance.

Understanding business/social opportunity

1. **Increased Sales and Revenue:** Top-performing agents tend to be rainmakers for sales and revenue. This way the company can harness maximum sales potential by retaining their interest and motivation to make the sale.
2. **Satisfied Clients:** Agents are more professional, perform better from providing customer service due to higher satisfaction and retention.
3. **Cost Efficiency:** This allows the company to identify and resolve specific training needs for low-performing agents to see an enhanced ROI through increased efficiency and decreased costs from sub-optimal performance and high turnover.
4. **Strategic Planning:** Functions/Predicting agent performance, bonuses- Strategic planning of future predictions, how much budget can be planned and spent out of the given amount for bonuses and training.

Data Report

Understanding how data was collected in terms of time, frequency and methodology

Let look into a hypothetical scenario of how data collection might have taken place:
Customer Registration:
When a customer registers, their CustID, Age, Gender, Occupation, EducationField, Channel and MaritalStatus are captured.
Policy Purchase:
In the case of policy purchase by a customer, information such as ExistingProdType, NumberOfPolicy, SumAssured, PaymentMethod, ExistingPolicyTenure and Zone is collected for every transaction.
Monthly Updates:
At the end of each month AgentBonus data is updated monthly MonthlyIncome data latest LastMonthCalls information Complaints will be used to update CustCareScores
Event-Based Records:
These events are immediately recorded if a customer makes a complaint or makes a service call and then aggregated into larger units for analysis on a monthly basis.

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

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	Mon
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	

The first 5 rows of the table

	CustID	AgentBonus	Age	CustTenure	Channel	Occupation	EducationField	Gender	ExistingProdType	Designation	NumberOfPolicy	MaritalStatus	M
4515	7004515	3953	4.0	8.0	Agent	Small Business	Graduate	Male	4	Senior Manager	2.0	Single	
4516	7004516	2939	9.0	9.0	Agent	Salaried	Under Graduate	Female	2	Executive	2.0	Married	
4517	7004517	3792	23.0	23.0	Agent	Salaried	Engineer	Female	5	AVP	5.0	Single	
4518	7004518	4816	10.0	10.0	Online	Small Business	Graduate	Female	4	Executive	2.0	Single	
4519	7004519	4764	14.0	10.0	Agent	Salaried	Under Graduate	Female	5	Manager	2.0	Married	

The last 5 rows of the table

```
<class 'pandas.core.frame.DataFrame'>
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
 19  CustCareScore         4468 non-null   float64
dtypes: float64(7), int64(5), object(8)
memory usage: 706.4+ KB
```

The data contains 12 numerical columns (7 float data types and 5 integer data types)and 8 object data types.
The data contains 4520 records(rows) and 20 columns.

	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

The descriptive analysis of all the numerical columns are as shown in the figure above.

```
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
dtype: int64
```

The above picture give the null values in each column.
There are no duplicate values in the data.

Understanding of attributes (variable info, renaming if required)

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

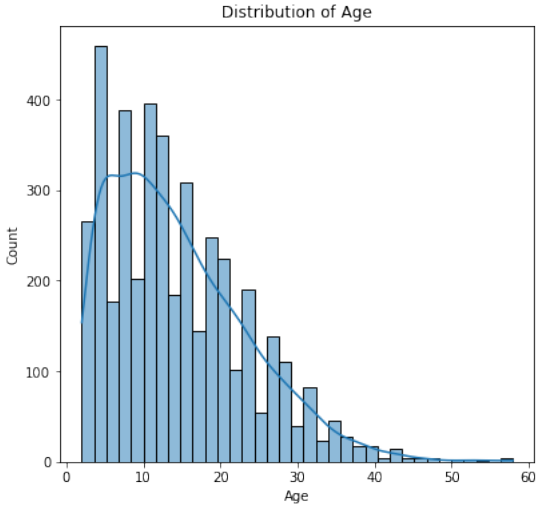
CustCareScore	Customer satisfaction score given by customer in previous service call

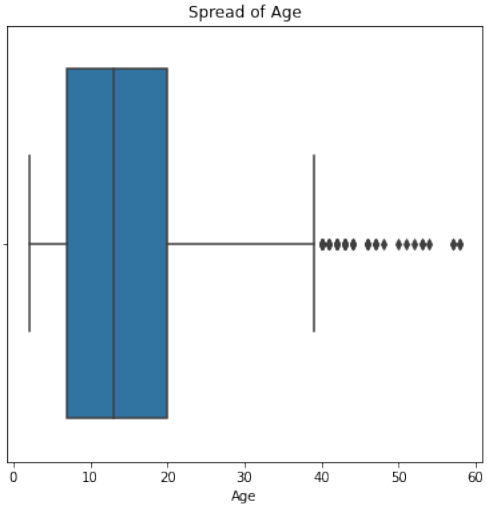
The data description of the columns of the data.
As all names follow the nomenclature no need to change variable names.

Exploratory data analysis

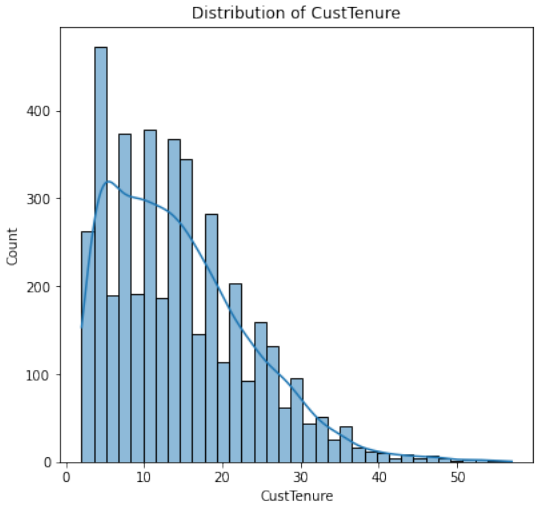
Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

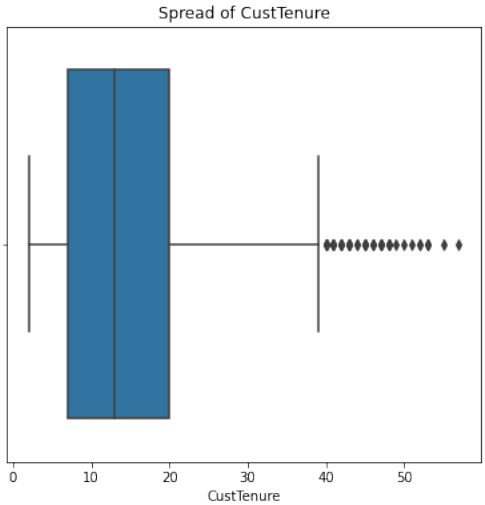
For continuous variables:



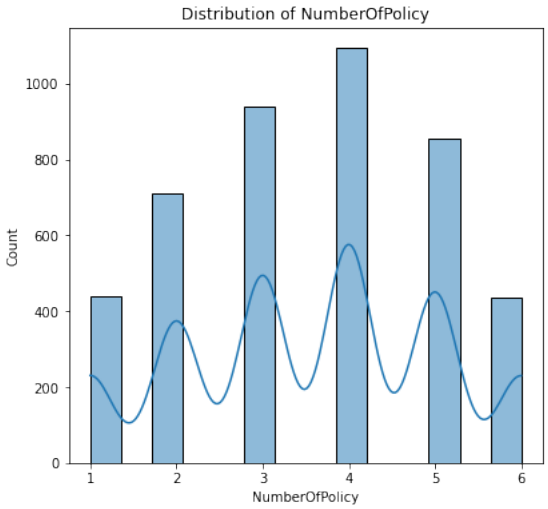


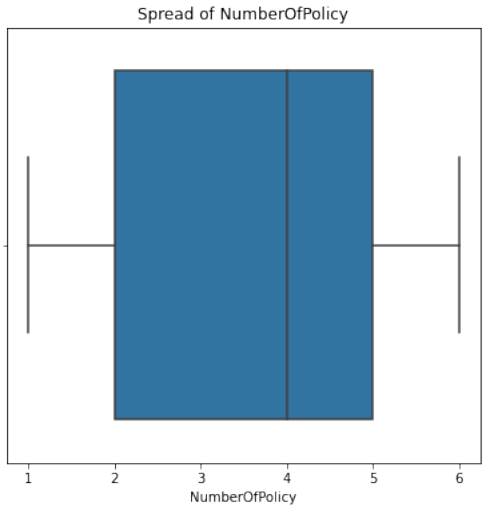
The graph show the customers are mostly children and as age goes above 40 the customers decline. The majority are of the age between 7 and 20.



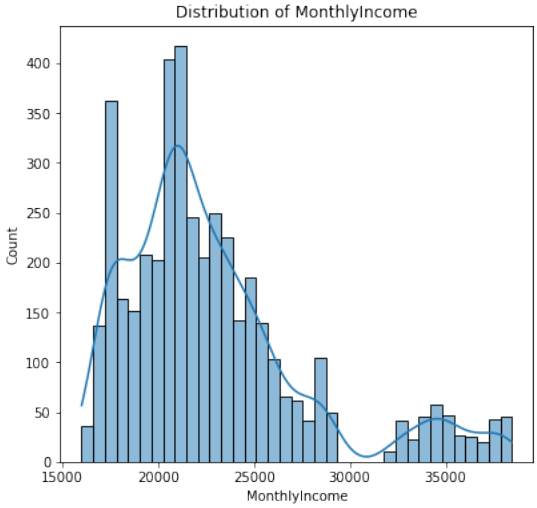


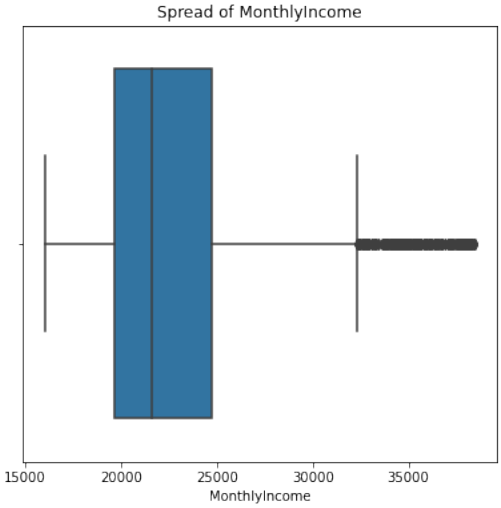
The tenure of customer is highest between 0 to 10 years and then sees a gradual decline.



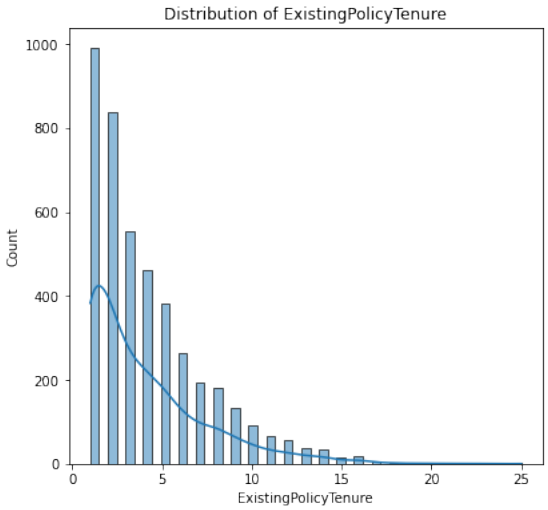


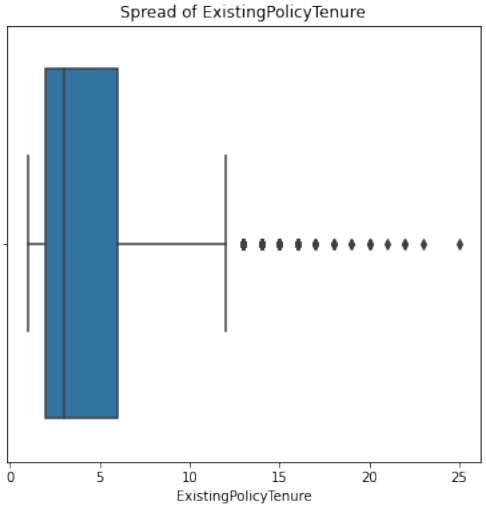
Maximum customers have 4 policies.



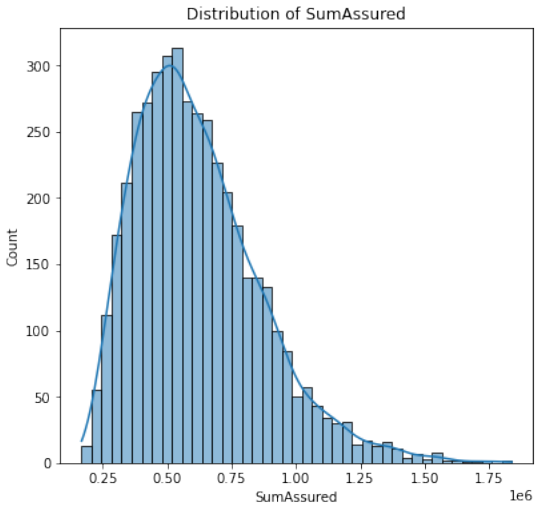


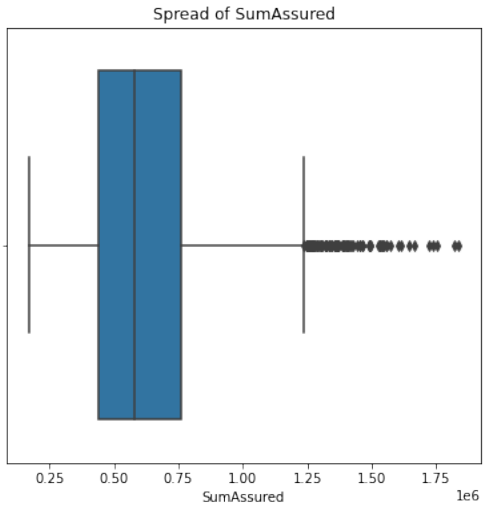
Majority have a monthly income between 15000 and 25000.



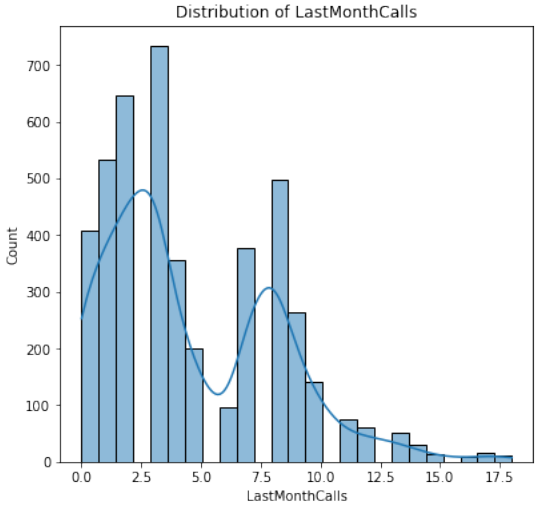


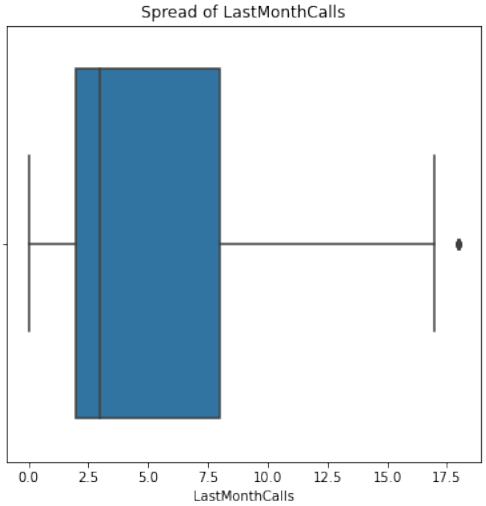
Maximum tenure is 0 to 5 years.



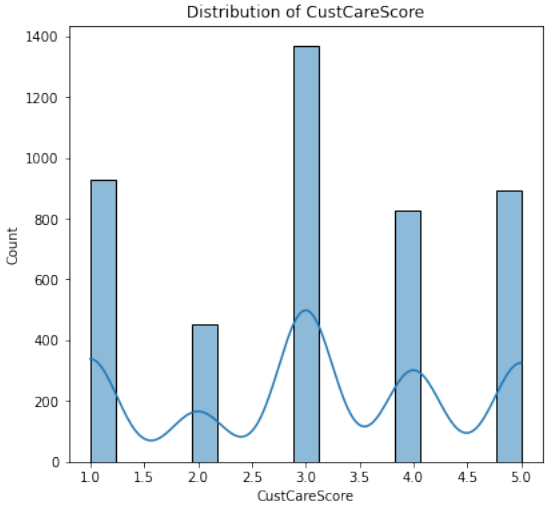


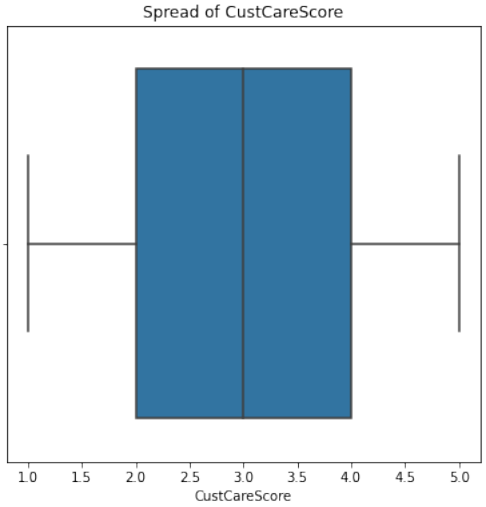
There is a gradual rise till 0.50-0.75 after which there's a gradual decline.



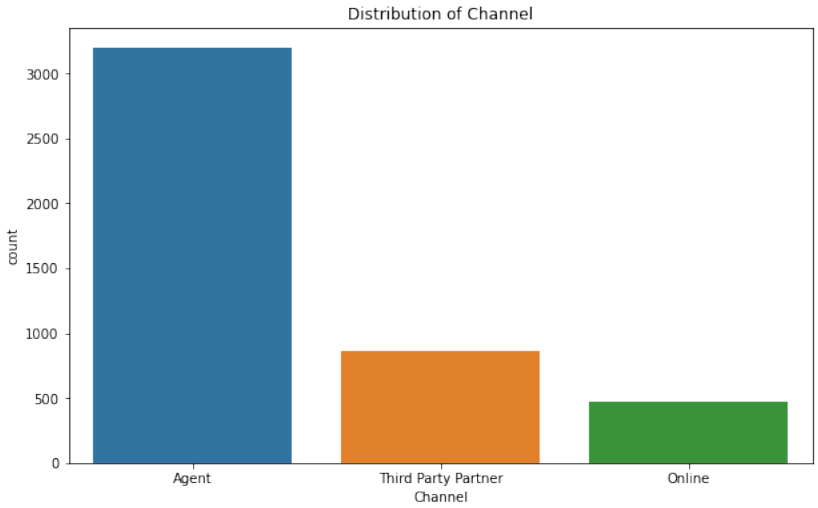


Theres a high after 2.5 and a drop which with a smaller high and low declines.

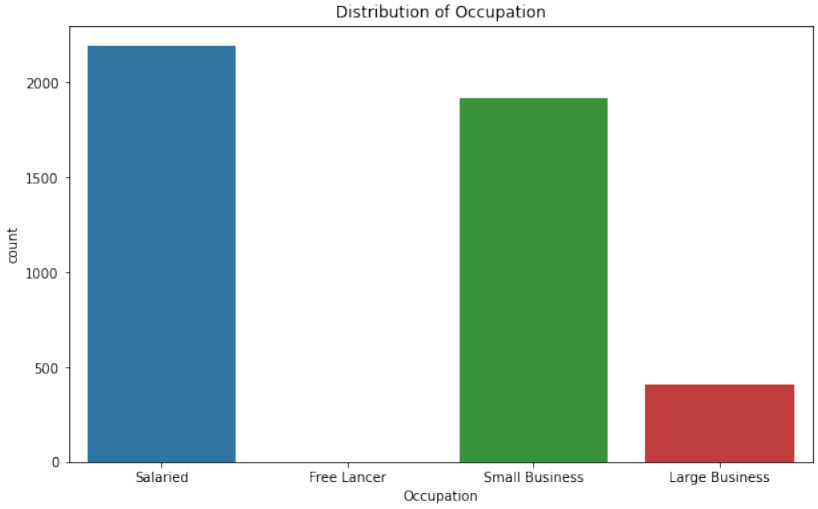




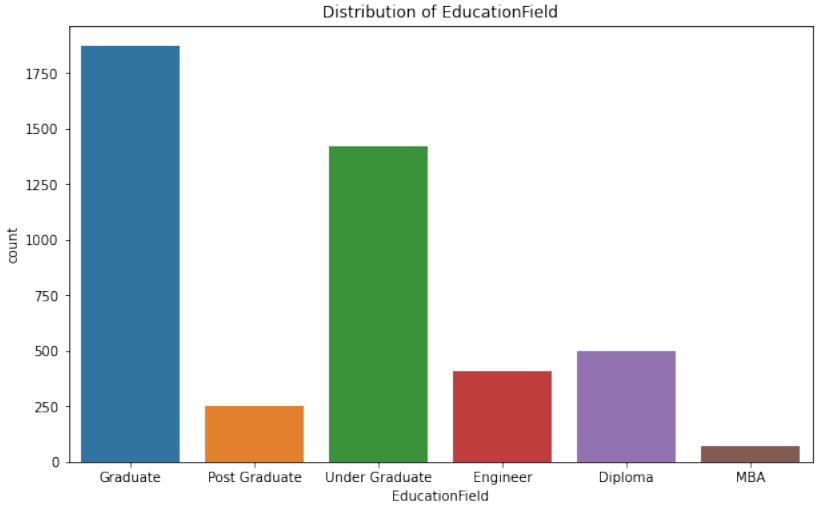
Maximum score given is 3 the score of 1 needs to be worked on.
For categorical variables:



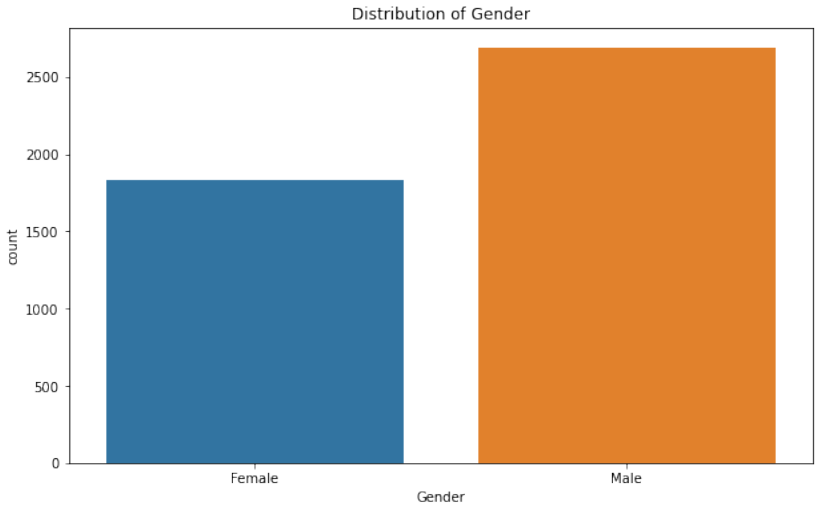
Maximum distribution is from agents.



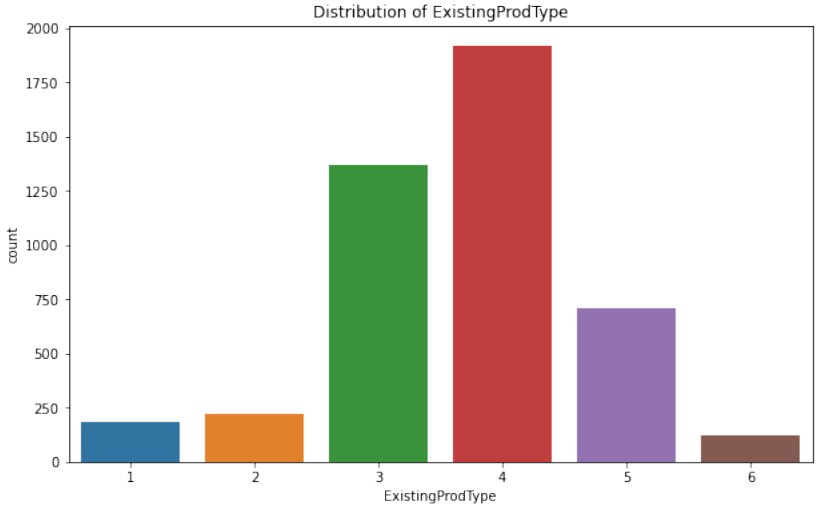
Maximum customers have salaried job followed by small businesses.



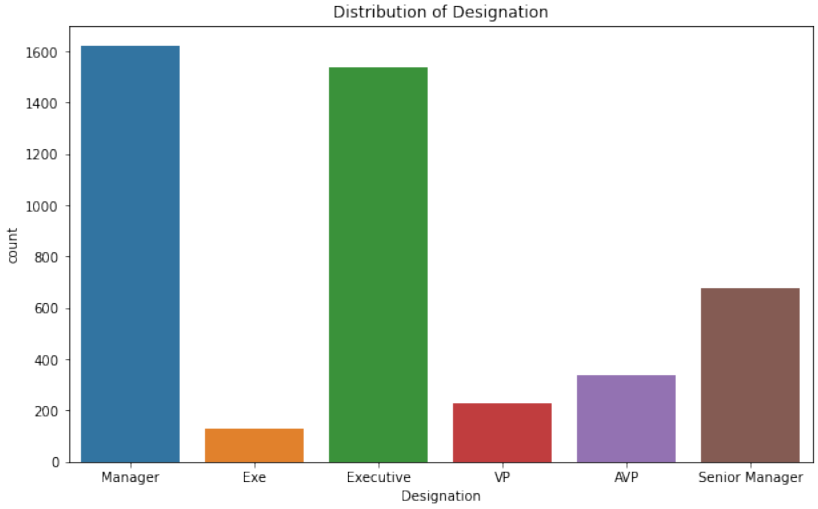
Majority of the customers are graduates. Followed by people in the education field.



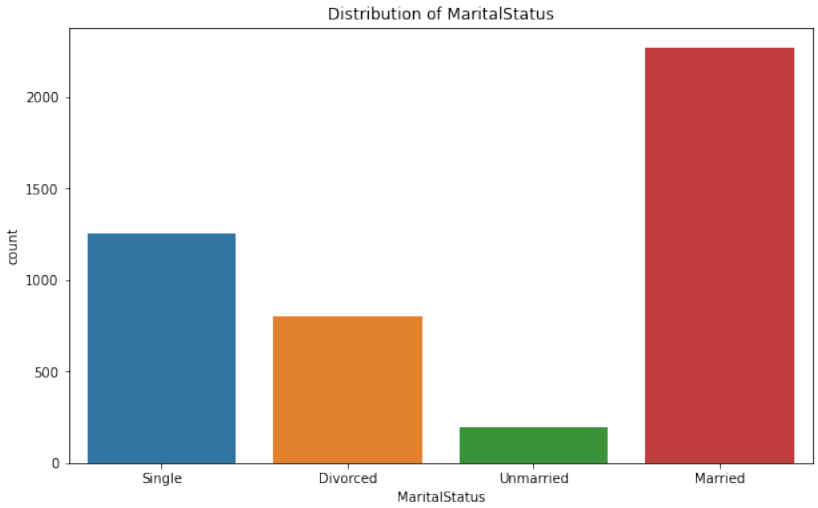
Majority customers are males.



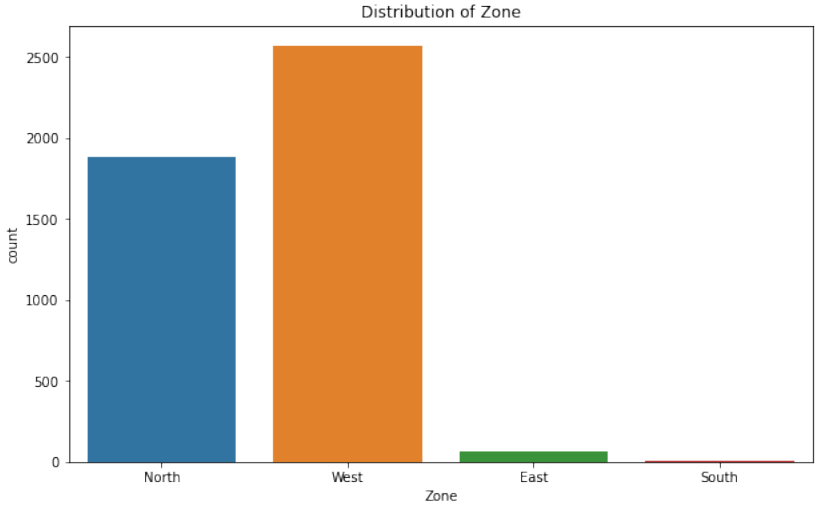
Maximum have opted for product type 4 followed by 3 and then 5. Lowest is 6



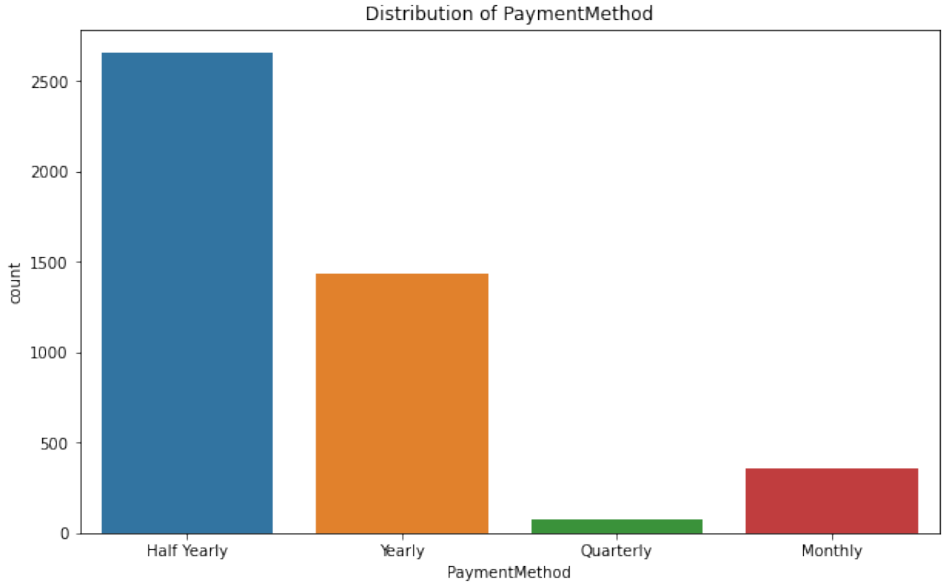
Maximum customers and managers followed by executives and least are EXe



Maximum customers and married

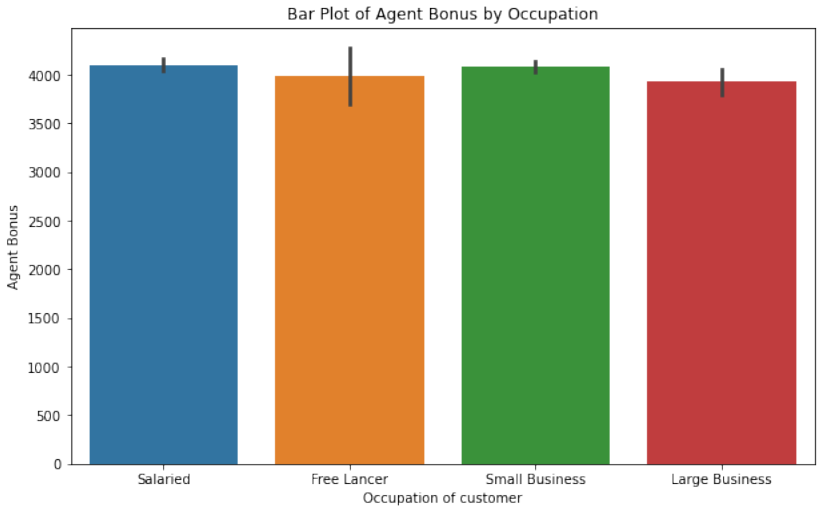


The highest customers are from west

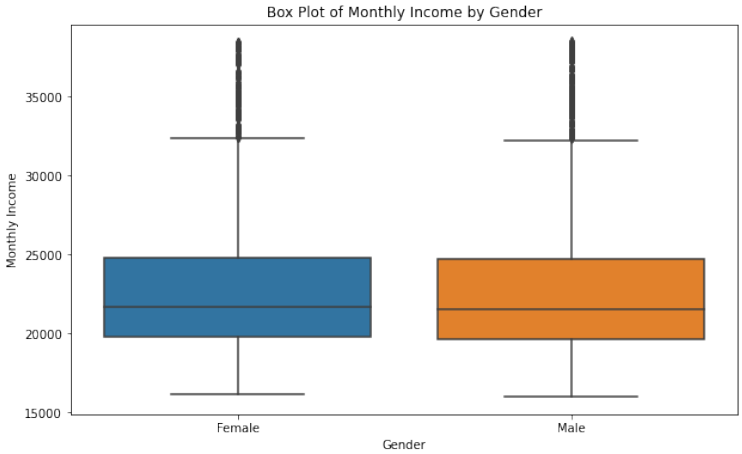


Maximum payment are done half yearly.

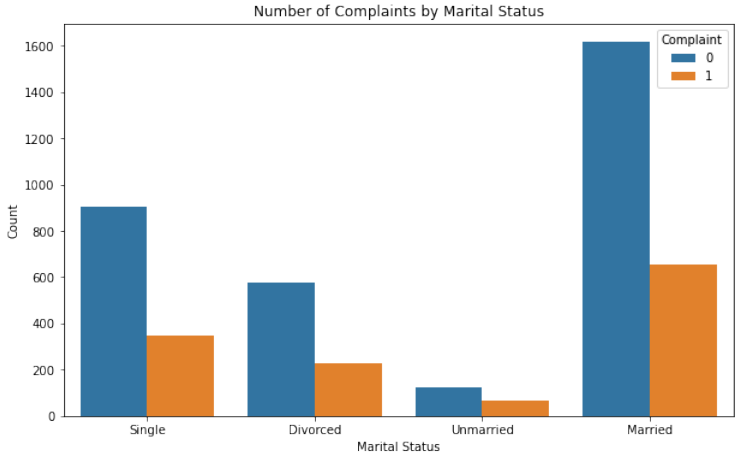
Bivariate analysis (relationship between different variables , correlations)



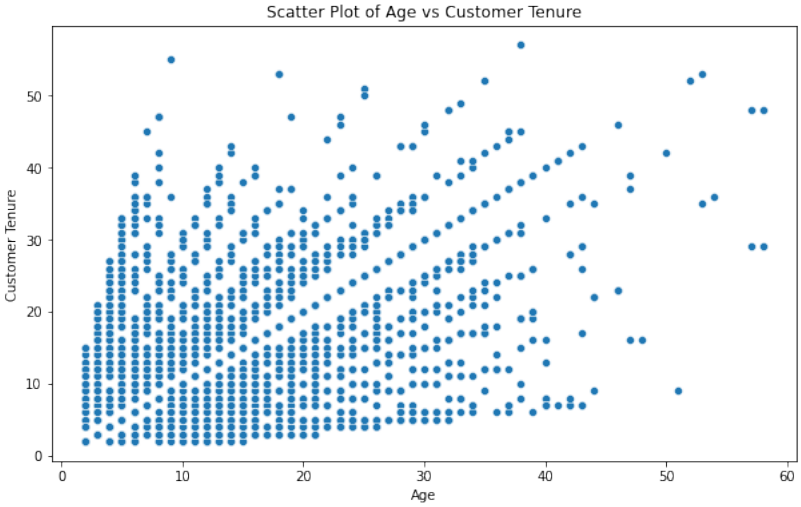
Small business has highest value and large business have the least value. Free lancers have larger error bar suggesting highest variability or uncertainty. And salaried and small business have small error bar



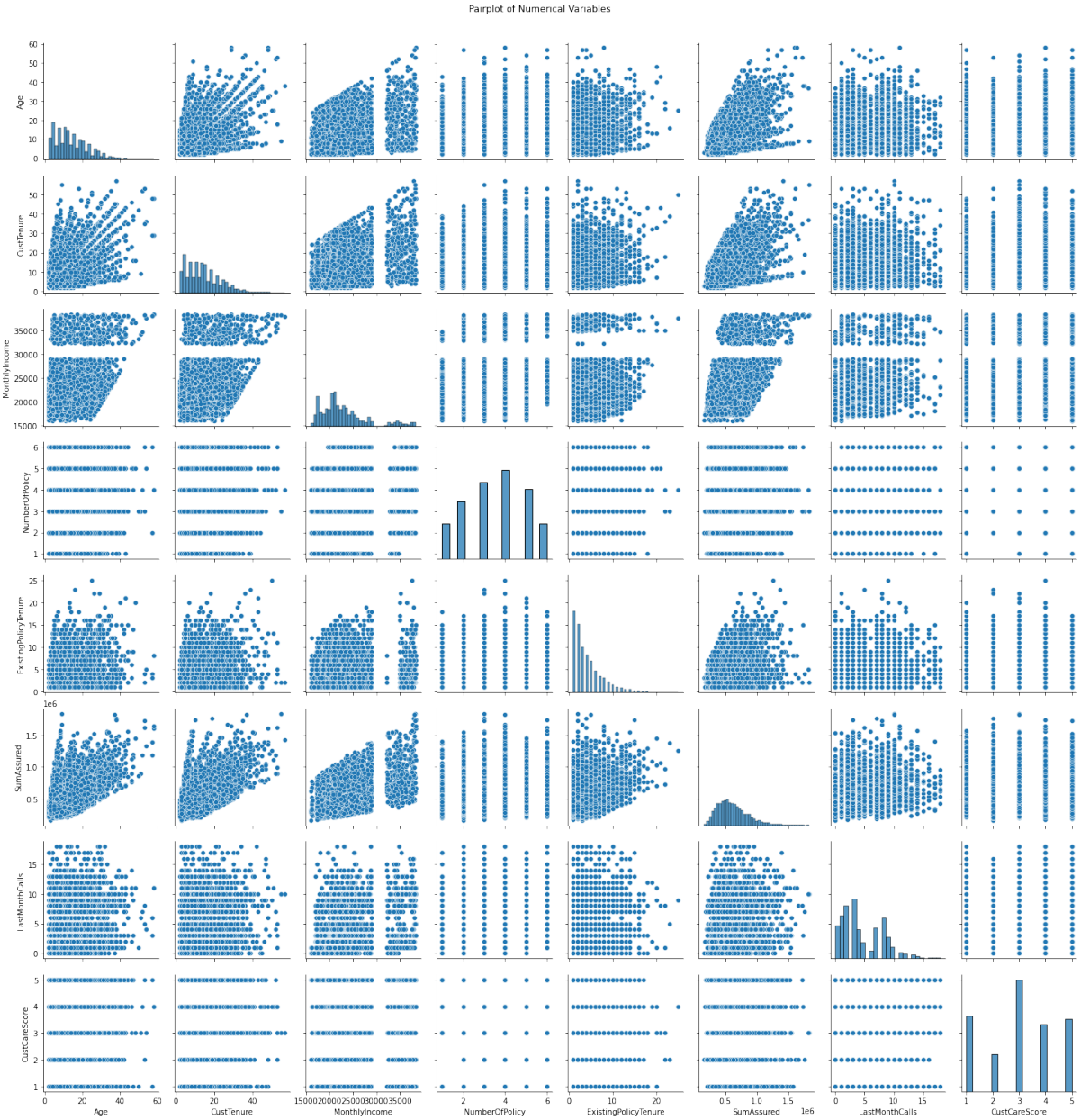
The boxplot show similar medians, interquartile ranges, and overall data ranges, indicating that the distributions of two categories are comparable. There are no significant outliners or skewness in either category.

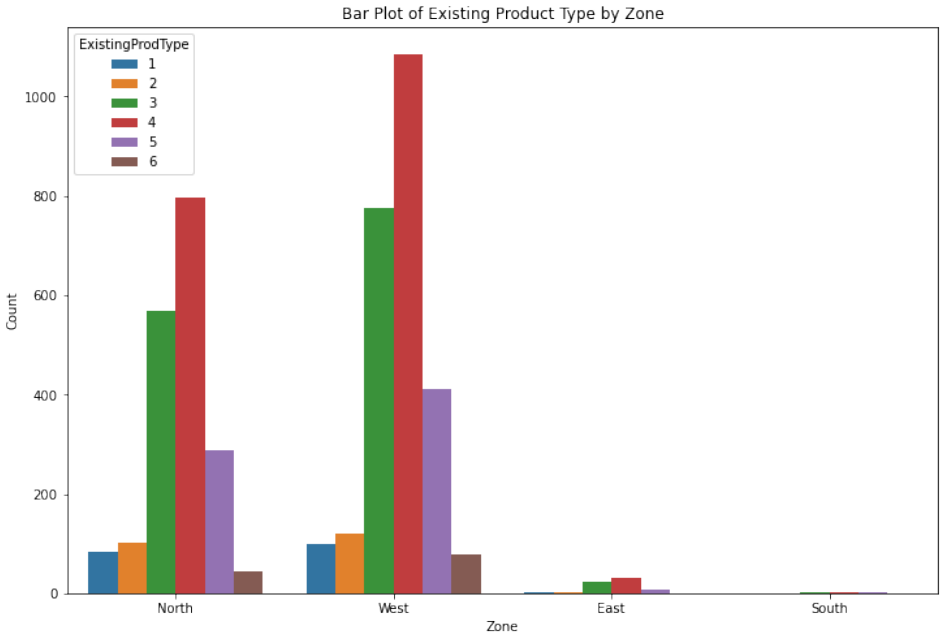


The shows that complaints(0) are consistently higher than 1 across all groups and married have highest complaints.

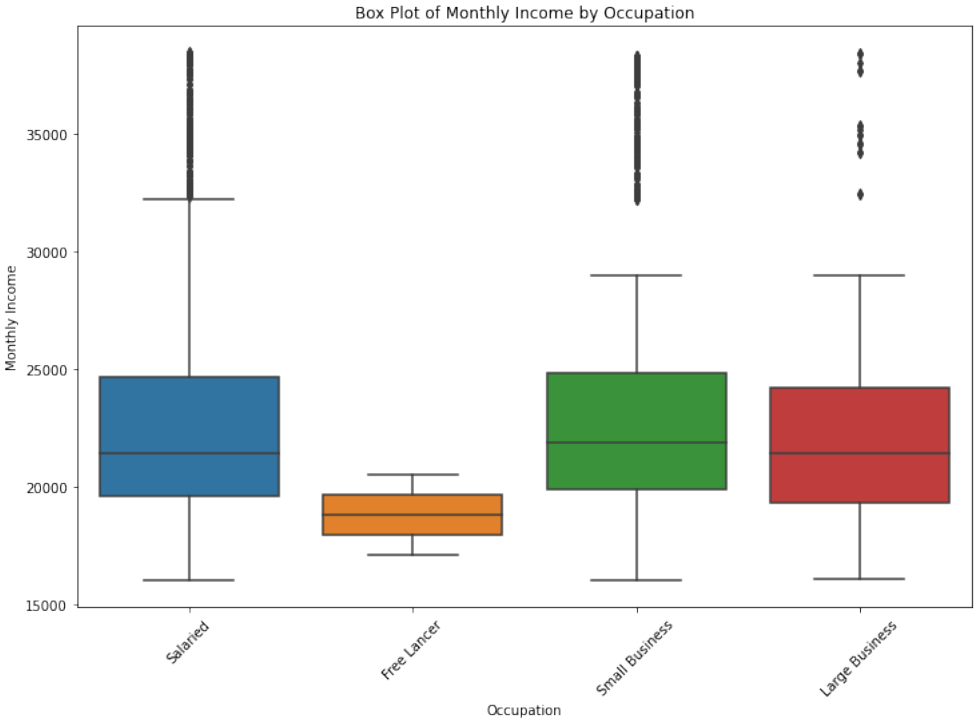


The graph shows customer tenure gradual increase.

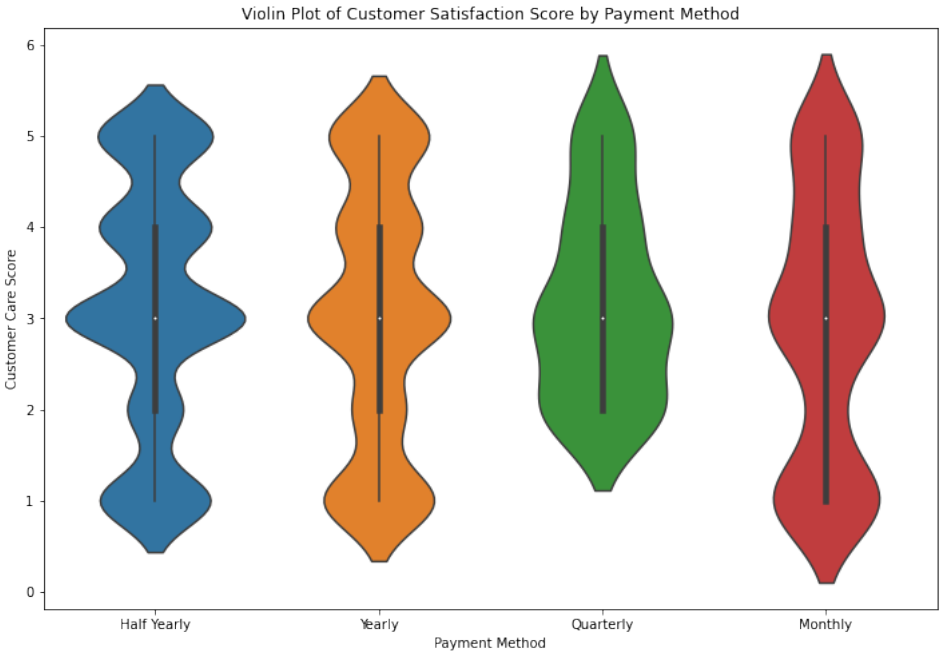




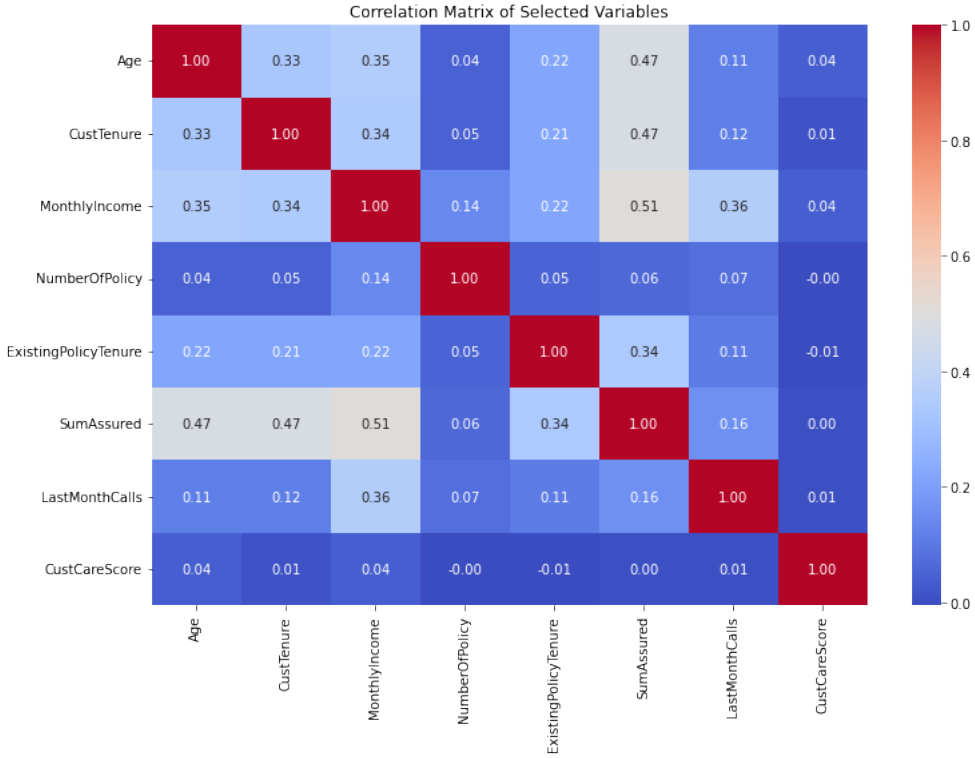
The plot shows maximum prod type is 4 and in west region. South has least .



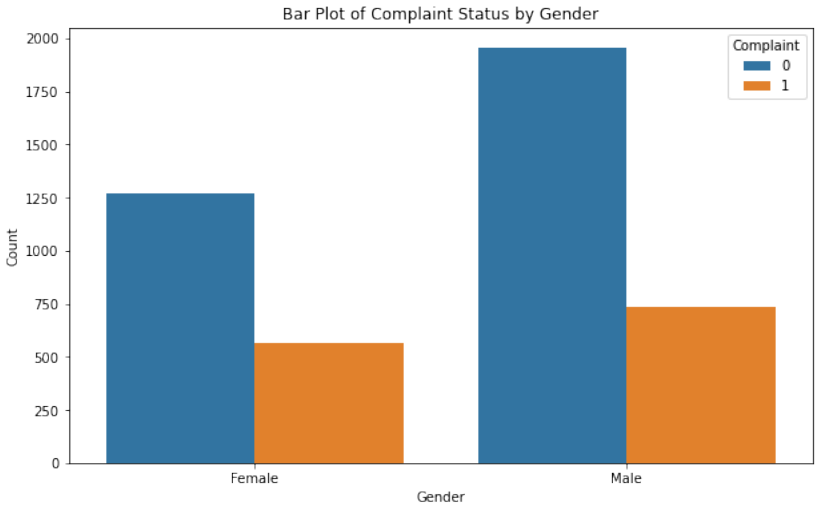
Although the large business has higher median small business and salaried have higher income.



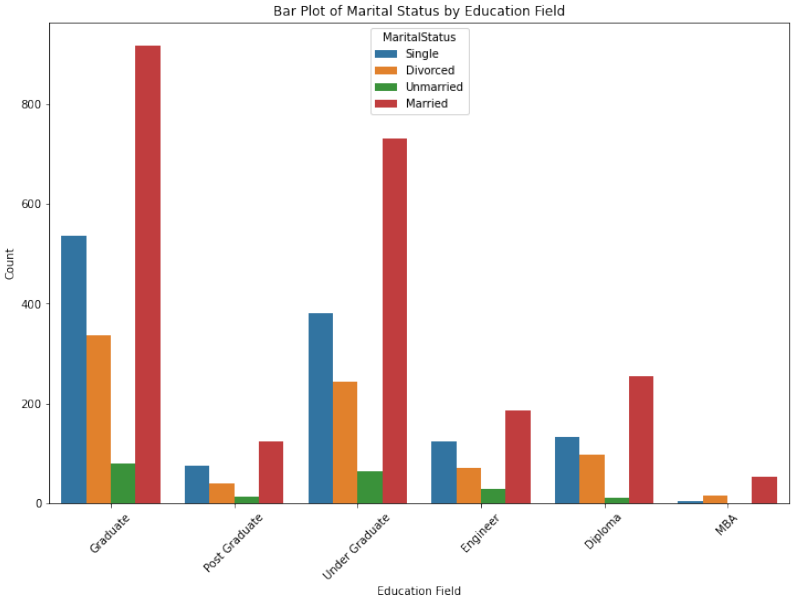
The plot shows maximum of the scores given are 3 but a good majority get 1 in monthly payment.



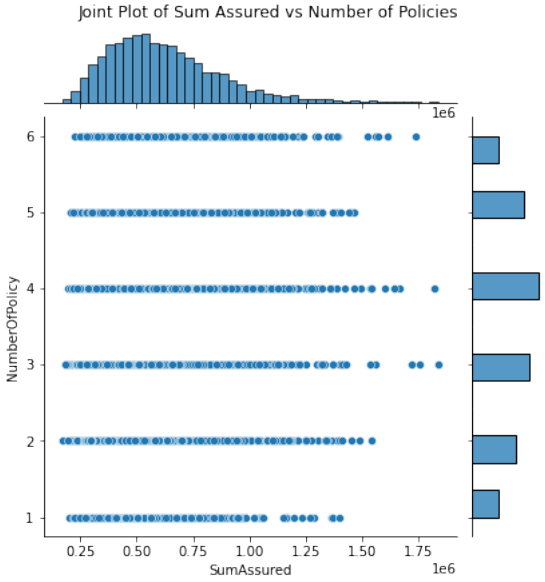
The heat map shows relation between sumassured and age , custtenure , monthlyincome. The other column don’t show much corelation.



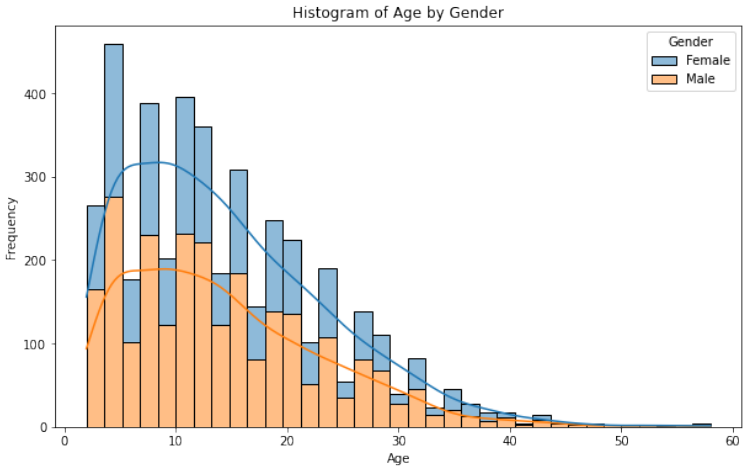
Majority males have 0 complaint



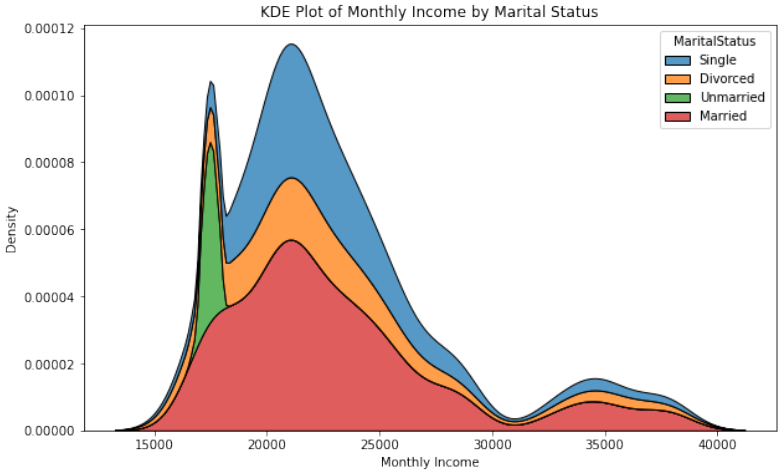
Maximum married people are salaried personals.



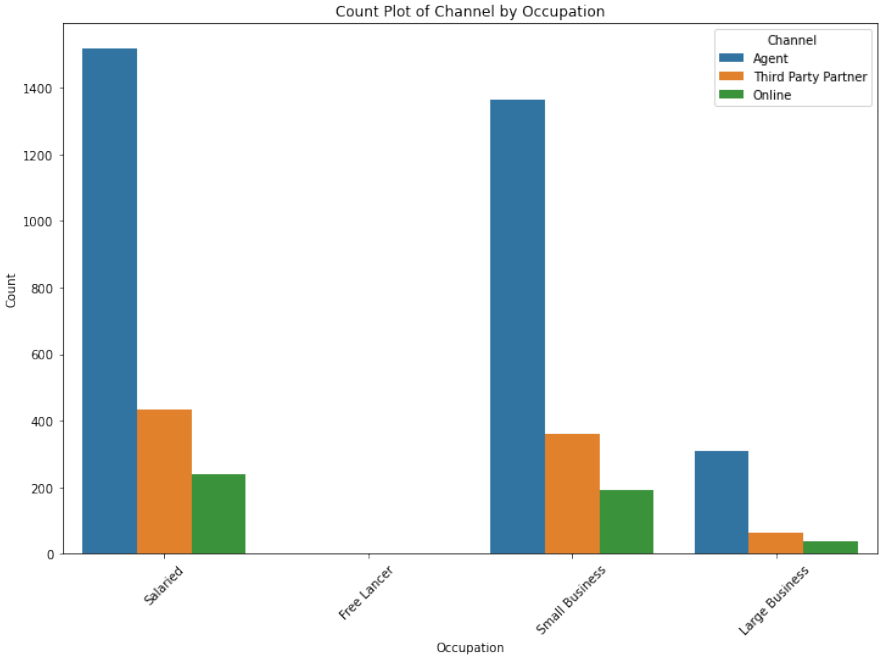
Maximum of 4 polies are taken and 0.50 is highest Sum assured.



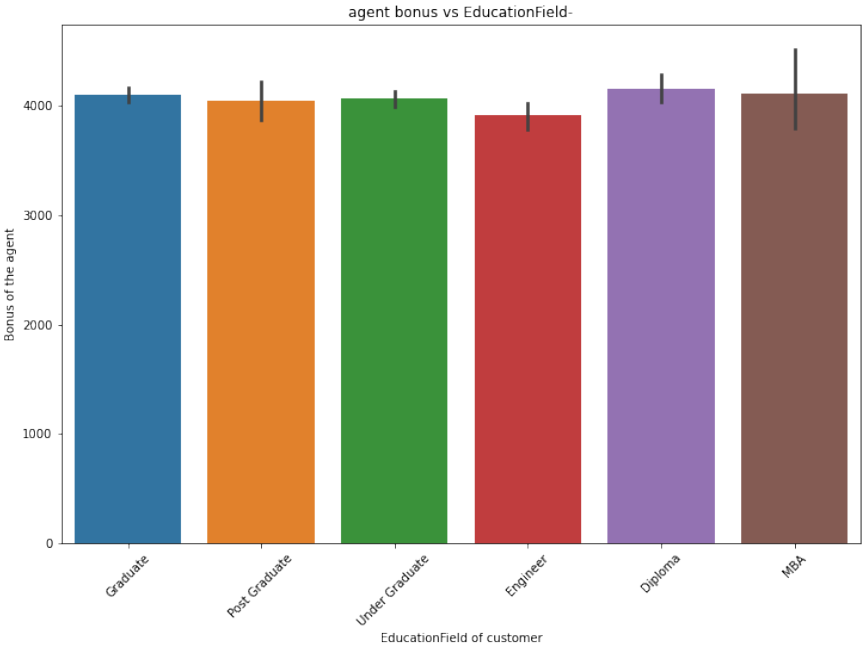
Maximum female are below age 10 and male are highest from 2 to 15.



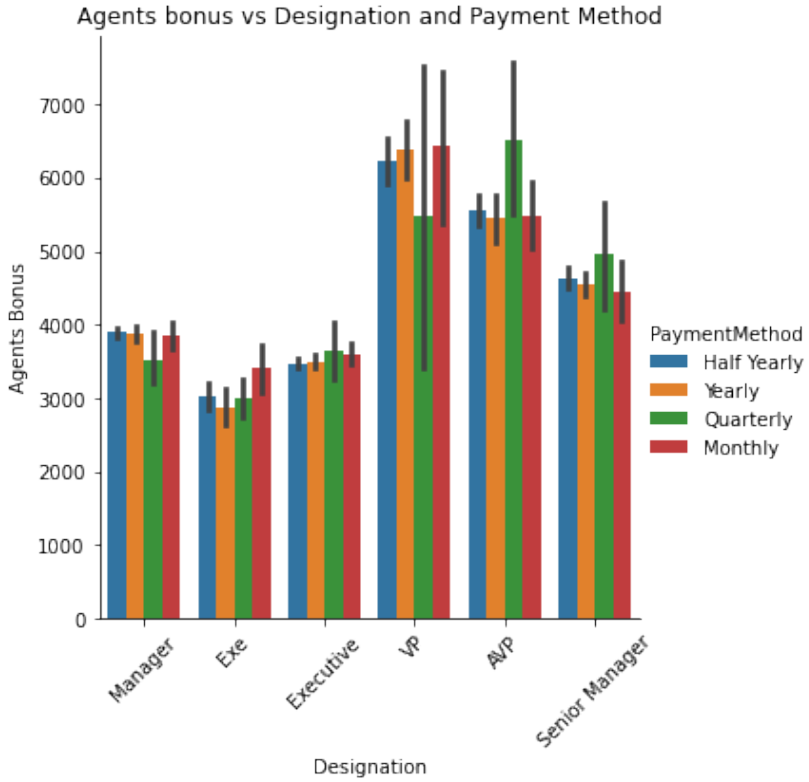
Unmarried have income range of 15000 to 20000. Maximum income ranges from 17500 to 25000.



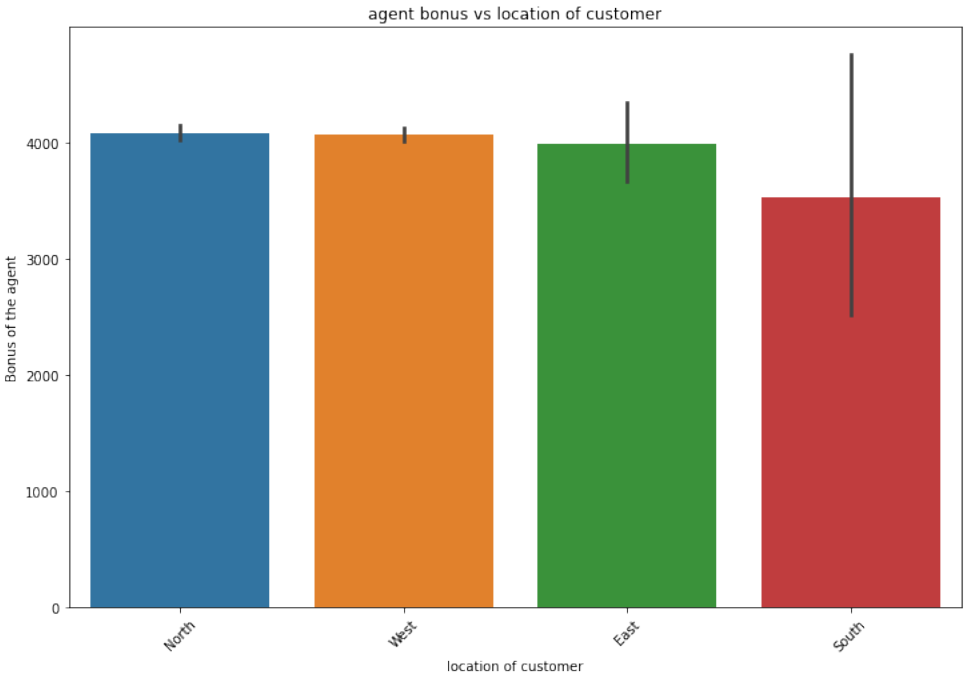
Maximum sales is from agents and to salaried customers. Least is to free lancers.



The graph shows the agent bonus is overall equivalent across all education fields but has a high in Under graduates in education field and lowest in UG.



The above two graph show that AVP customer has highest agents bonus in quarterly payments followed by VP customers in yearly or monthly payments.



Maximum bonus for agents is In north and west and least in south.

Removal of unwanted variables (if applicable)

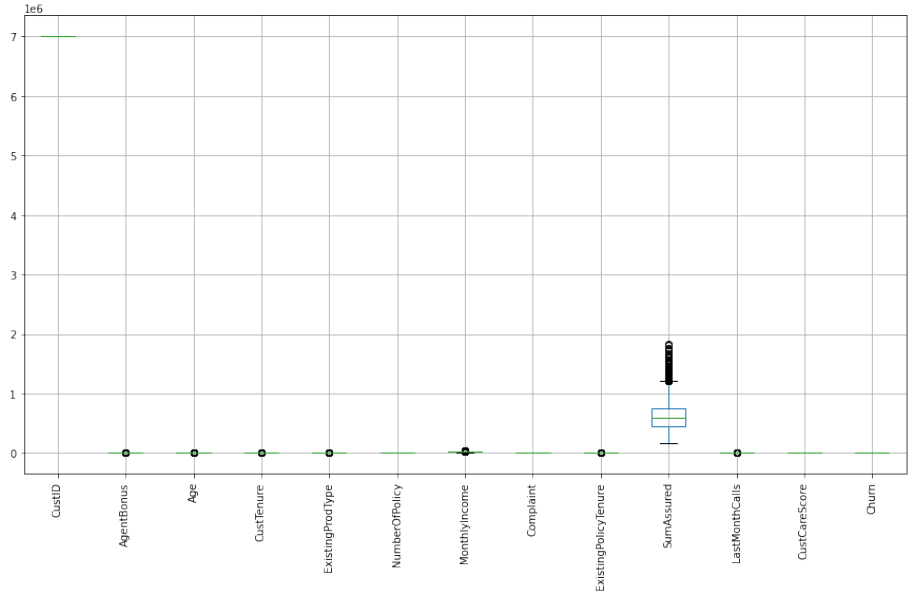
The data consisted of error

a. Gender column had fe male which was corrected to female
b. The education field column had UG and Under graduate which was combined to Under Graduate
c. The Occupation column had Laarge business which was rectified as large business.

Missing Value treatment (if applicable)

The missing vales were treated using mean through simple Imputer.

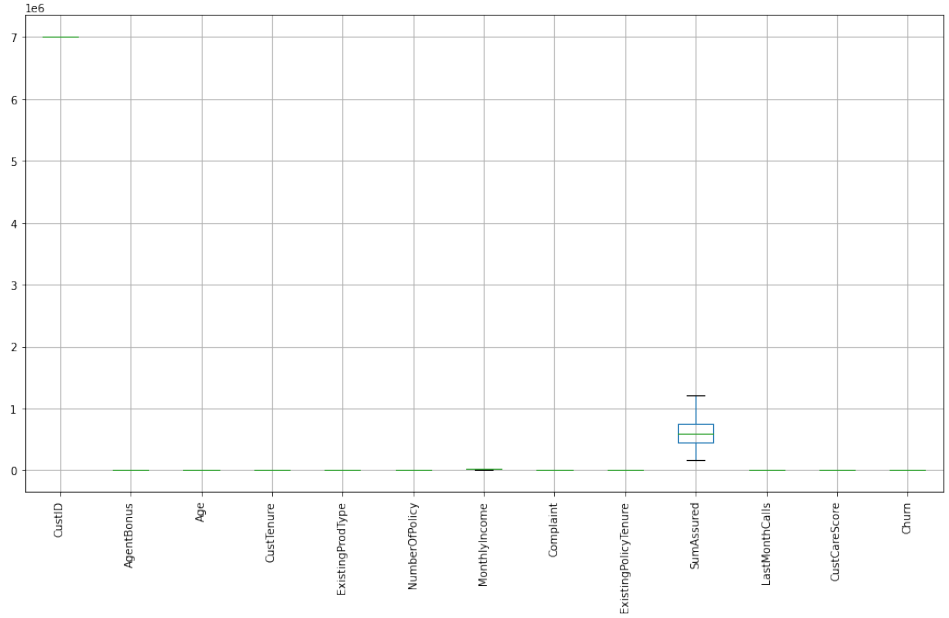
Outlier treatment (if required)



Number of outliers in each numeric column:

	Column	No. of outliers
0	CustID	0
1	AgentBonus	100
2	Age	105
3	CustTenure	97
4	ExistingProdType	306
5	NumberOfPolicy	0
6	MonthlyIncome	384
7	Complaint	0
8	ExistingPolicyTenure	345
9	SumAssured	110
10	LastMonthCalls	12
11	CustCareScore	0
12	Churn	0

This is before outliner treatment.
The outliner are treated using IQR method



Variable transformation (if applicable)

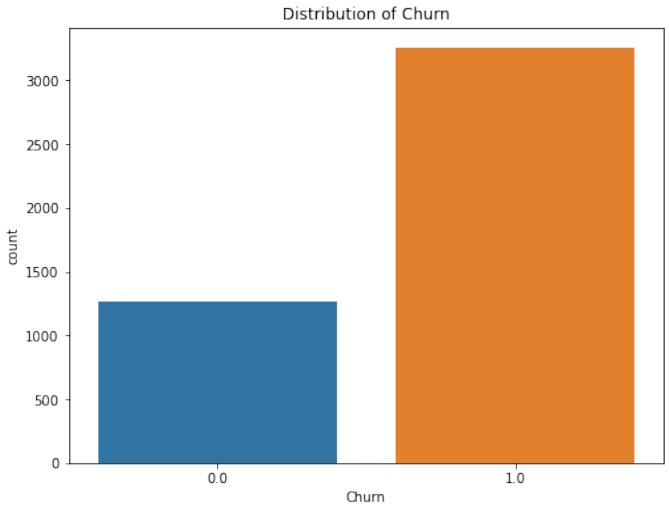
No variables were transformed.

Addition of new variables (if required)

An additional column called churn was added which was calculated using custcarescore and complaints which can be used a target variable for further analysis.

Business insights from EDA

Is the data unbalanced? If so, what can be done? Please explain in the context of the business




```
Class distribution of Churn:
1.0    3253
0.0    1267
```

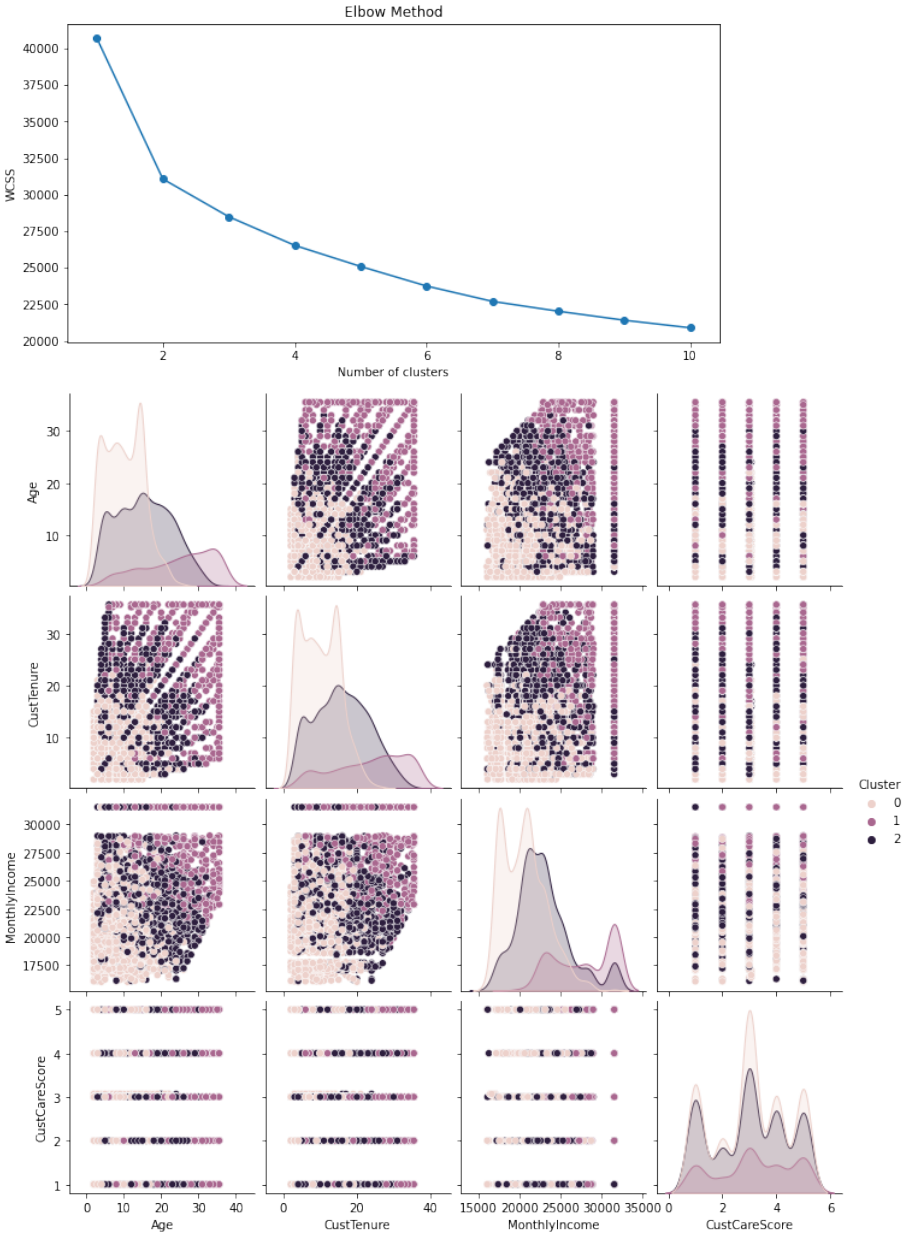
Yes the data is imbalanced which can cause model bias, i.e, A model trained on this data might be more biased towards predicting the majority class as it has more examples or it can cause poor minority class prediction I.e, when predicting minority class there can be errors that could lead to missed opportunity in identifying customers who are likely to stay.
The best ways to deal with this is:

a. Resampling techniques which include oversampling (increase the number of instances in minority class), undersampling (decrease the number of majority class) or smote (generate synthetic examples for minority class)
b. Algorithmic approach which is some algorithms allows you to assign weights to classes, giving more importance to minority class.

In the context of customer churn:

a. Better Minority Class Detection: Addressing imbalance can help in better identifying customers who are likely to stay (non-churn), allowing targeted retention strategies.
b. Improved Customer Satisfaction: By accurately predicting churn, businesses can proactively engage with at-risk customers to improve satisfaction and loyalty.
c. Resource Allocation: Helps in allocating resources efficiently towards customers who need attention the most.
d. Balancing the dataset will improve the model's ability to generalize and provide more reliable predictions, ultimately aiding in making informed business decisions.

Any business insights using clustering (if applicable)



1. Cluster 0:
   a. Characteristics: Younger customers with high tenure and moderate income.
   b. Business Strategy: These could be loyal, younger customers. Target them with loyalty programs and upsell higher-end products.
2. Cluster 1:
   a. Characteristics: Older customers with lower tenure and higher income.
   b. Business Strategy: Focus on retaining these high-income customers through premium services and personalized attention.
3. Cluster 2:
   a. Characteristics: Middle-aged customers with varying tenure and lower customer care scores.
   b. Business Strategy: Improve customer service and satisfaction for this segment. Consider offering better support and engagement programs.

Any other business insights

1. Channel Effectiveness:
   a. Insight: agents are the most effective of different acquisition channels in attracting and retaining customers.
   b. Action: Analyse the distribution of customer clusters across acquisition channels. Allocate resources to channels that bring in high-value customers.

2. Occupation Preferences:
   a. Insight: most of the customers salaried professionals
   b. Action: Customize marketing messages and product offerings to align with the needs and interests of different occupations. For example, large business might prefer premium services, while salaried and freelancers might seek affordability.

3. Education Impact:
   a. Insight: majority customers are graduates or higher degree.
   b. Action: Develop educational content or services tailored to the specific fields of customers.

4. Gender-Based Insights:
   a. Insight: have majority male customers
   b. Action: Create targeted marketing campaigns and product features that resonate with each gender. Consider offering gender-specific promotions or discounts.

5. Income Segmentation:
   a. Insight: Segment customers based on their income levels to understand their purchasing power and preferences. The majority fall under the segment of 15000 to 25000
   b. Action: Offer customized pricing plans or product bundles that cater to different income segments. Provide special incentives or rewards for high-income customers.

6. Marital Status Influence:
   a. Insight: Explore how marital status impacts customer behaviour and needs. The majority customers are married
   b. Action: Develop family-oriented products or services for married customers. Offer exclusive deals or benefits for couples or families.

7. Zone-Based Strategies:
   a. Insight: Identify regional trends and preferences to tailor marketing and operational strategies. West has highest followed by north and south is least.
   b. Action: Customize promotions, products, and services based on the characteristics of each zone. Consider localizing offerings to better resonate with customers in specific regions.

8. Payment Method Preferences:
   a. Insight: Determine which payment methods are preferred by different customer segments. Majority of payments are done half yearly but different segments have different preferences.
   b. Action: Optimize the sales by offering a variety of payment options. Provide incentives or discounts for using preferred payment methods.