December 2021

This Business Report shall provide detailed explanation of how we approached each problem given in the assignment. It shall also provide relative resolution and explanation with regards to the problems

Life Insurance Sales - Capstone

Business Report

Thakur Arun Singh

Contents

[Problem 1: Introduction of the business problem 2](#_Toc91454630)

[Problem 1.a 2](#_Toc91454631)

[Executive Summary 2](#_Toc91454632)

[Project Approach 2](#_Toc91454633)

[Problem 1.b 2](#_Toc91454634)

[Problem 1.c 3](#_Toc91454635)

[Problem 2: Data Report 3](#_Toc91454636)

[Problem 2.a 3](#_Toc91454637)

[Problem 2.b 4](#_Toc91454638)

[Problem 2.c 5](#_Toc91454639)

[Problem 3: Exploratory data analysis 6](#_Toc91454640)

[Problem 3.A 6](#_Toc91454641)

[Problem 3.B 15](#_Toc91454642)

[Problem 3.c 21](#_Toc91454643)

[Problem 3.D 21](#_Toc91454644)

[Problem 3.e 23](#_Toc91454645)

[Problem 3.F 24](#_Toc91454646)

[Problem 3.G 25](#_Toc91454647)

[Problem 4: Business insights from EDA 25](#_Toc91454648)

[Problem 4.a 25](#_Toc91454649)

[Problem 4.b 28](#_Toc91454650)

[Problem 4.c 33](#_Toc91454651)

## Problem 1: Introduction of the business problem

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 up skill programs for low performing agents

### Problem 1.a

Defining problem statement

**Resolution:**

# Executive Summary

Academics and practitioners have studied over the years models for predicting firm’s various aspects, using statistical and machine-learning approaches. We are going to discuss one of the various aspects. An earlier sign that company employees are dissatisfied is the firm policies, work load, pay scale and bonuses. We are going to dive deeper and predict the bonus for its employees, so that we may designed an appropriate engagement activity for their high performing employees and up skill programs for low performing employees.

## Project Approach

The work that we have completed:

* Merged data from other sources like, demographic information, account details etc. for further deep analysis
* Data Quality and preparation activities were performed like missing value treatment, imputation, data type conversions for homogeneity in the data set
* Performed EDA on the data to understand the data and to determine any outlier and treatment for the same
* Also used ANNOVA to understand if the model performance can be improved

### Problem 1.b

Need of the study/project

**Resolution:**

All Life Insurance companies offer various incentive plans for their employees, to boost their sales and to have higher balance of insurance plans, **“The higher the Insurance amount the higher the bonus pay out”**. However, not all companies get successful in the above Mantra. Some companies fail to have a better plan for bonus payouts. For such companies we should study the data and need to understand the requirements to implement the above plan. Where in, the company can predict the bonus for its employees and design appropriate engagement activity for their high performing employees and also have training programs for non/low performing employees.

### Problem 1.c

Understanding business/social opportunity

**Resolution:**

Based on our analyses, once we thoroughly study the data we will be able to understand the business better and take decisions. Here, after all the analyses which were perform we are able to come to collusion whether the given approach will work or not.

## Problem 2: Data Report

### Problem 2.a

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

**Resolution:**

* Looking at the data we can see that the data collected with a wide verity of age range from 18 years to 58 years.
* There is good mix of gender where we have 40% Female 60% and 60% Male
* 50% of the data consist of married people
* From the entire data we have about 35% of the people who are at Manager Level.
* About 49% of the people who took Life Insurance policy are salaried employees.

**Describing the data:**

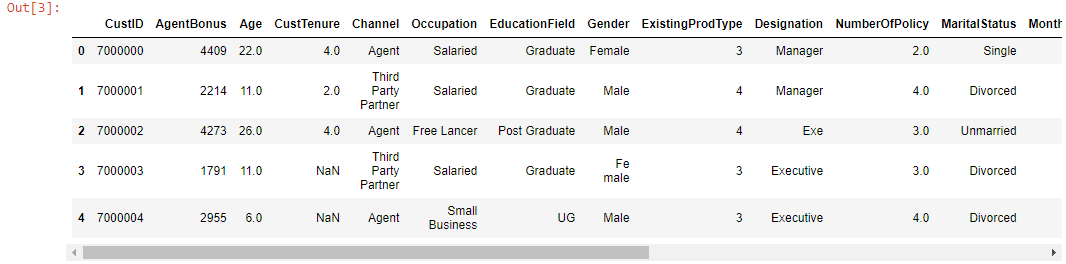
* First we import all the necessary libraries in Python, and then import the data file which is ‘LifeInsuranceSales’. Once we import the file we confirm whether the data has been uploaded correctly or not using ‘head’ function. Using this function we can view the data and all the columns and headers whether they are aligning correctly or not.
* Then using the ‘shape’ function we can understand how many row and columns are there in our data set.
* To check the data type of all the columns and also to check the null values, ‘info’ function. Has been used.
* To see the detail description of the data such as, Count, Mean, Median, Min, Max, Standard Deviations etc,
* Using the ‘isnull’ function, one can understand if there are any null values in the data set. And we do not have any null values in the existing data set.
* Using the ‘dups’ function we check for the duplicates and there were no duplicate values.
* We also identified the unique values in categorical data.

### Problem 2.b

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

**Resolution:**

To see if the data has been imported or not.



To know the shape of the data, we can see that we have 4520 Rows and 23 Colomns



We can see the date distribution

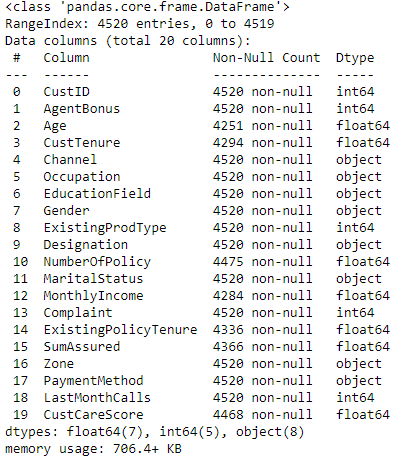
|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **unique** | **top** | **freq** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **AgentBonus** | 4520 | NaN | NaN | NaN | 4077.84 | 1403.32 | 1605 | 3027.75 | 3911.5 | 4867.25 | 9608 |
| **Age** | 4251 | NaN | NaN | NaN | 14.4947 | 9.03763 | 2 | 7 | 13 | 20 | 58 |
| **CustTenure** | 4294 | NaN | NaN | NaN | 14.469 | 8.96367 | 2 | 7 | 13 | 20 | 57 |
| **Channel** | 4520 | 3 | Agent | 3194 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Occupation** | 4520 | 5 | Salaried | 2192 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **EducationField** | 4520 | 7 | Graduate | 1870 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **Gender** | 4520 | 3 | Male | 2688 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **ExistingProdType** | 4520 | NaN | NaN | NaN | 3.68894 | 1.01577 | 1 | 3 | 4 | 4 | 6 |
| **Designation** | 4520 | 6 | Manager | 1620 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **NumberOfPolicy** | 4475 | NaN | NaN | NaN | 3.56536 | 1.45593 | 1 | 2 | 4 | 5 | 6 |
| **MaritalStatus** | 4520 | 4 | Married | 2268 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **MonthlyIncome** | 4284 | NaN | NaN | NaN | 22890.3 | 4885.6 | 16009 | 19683.5 | 21606 | 24725 | 38456 |
| **Complaint** | 4520 | NaN | NaN | NaN | 0.287168 | 0.452491 | 0 | 0 | 0 | 1 | 1 |
| **ExistingPolicyTenure** | 4336 | NaN | NaN | NaN | 4.13007 | 3.34639 | 1 | 2 | 3 | 6 | 25 |
| **SumAssured** | 4366 | NaN | NaN | NaN | 620000 | 246235 | 168536 | 439443 | 578976 | 758236 | 1.84E+06 |
| **Zone** | 4520 | 4 | West | 2566 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **PaymentMethod** | 4520 | 4 | Half Yearly | 2656 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **LastMonthCalls** | 4520 | NaN | NaN | NaN | 4.62699 | 3.62013 | 0 | 2 | 3 | 8 | 18 |
| **CustCareScore** | 4468 | NaN | NaN | NaN | 3.06759 | 1.38297 | 1 | 2 | 3 | 4 | 5 |

### Problem 2.c

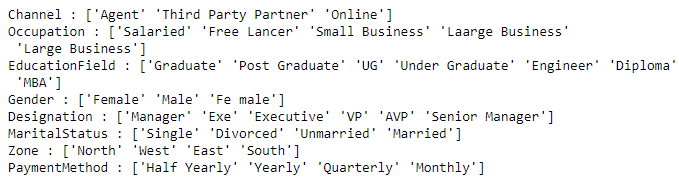
Understanding of attributes (variable info, renaming if required)

**Resolution:**

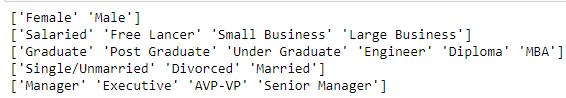
Below we can see the variable info about the data – data types of respective columns comprises of float, integer and object.



Converting some of the data types to Object data.



Replacing the duplicate words in Gender, Occupation, EducationField & Designation variable

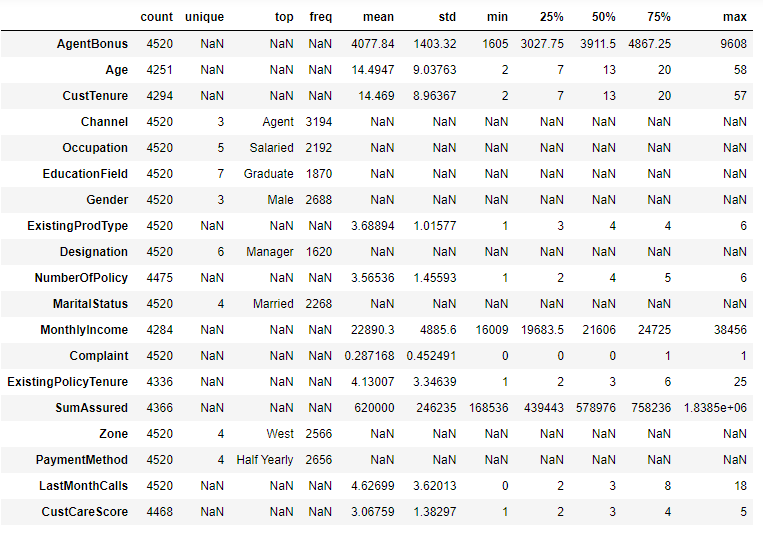


## Problem 3: Exploratory data analysis

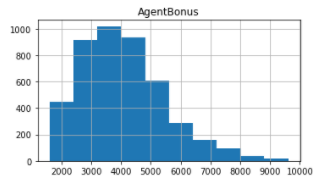
### Problem 3.A

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

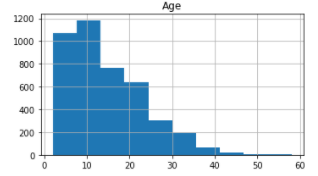
**Resolution:** First we can see the data distribution

****

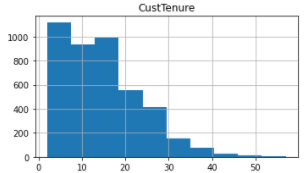
**Agent Bonus:** From the below graph we can see that the majority of the bonus falls between 3,000 to 4,000.



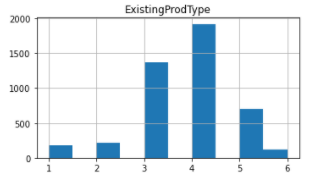
**Age:** Below the graph shows that the average age of the customer is about 30-35 years.

****

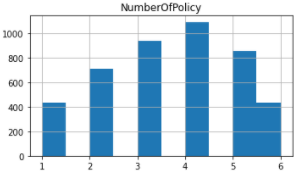
**Customer Tenure:** Below graph shows that there are over 1000 customer who are loyal customers for at least 10 years.



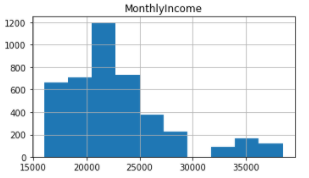
**Existing Product Type:** Below graph shows that there at about 2000 people who hold at least 4 products



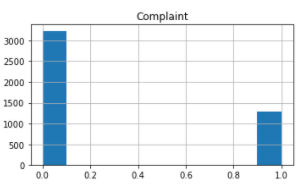
**Number of Policies:** Below graph shows that at least 1000 people hold 4 policies each.

****

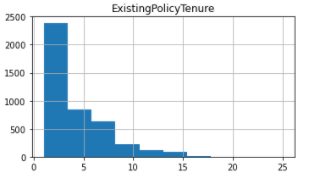
**Monthly Income:** About 1200 people have a monthly income between 20,000 – 25,000.



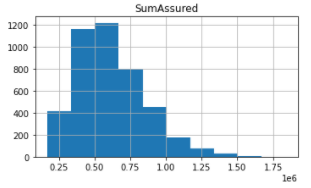
**Complaints:** There were about 3000+ people who had complained between 0-1 times.

****

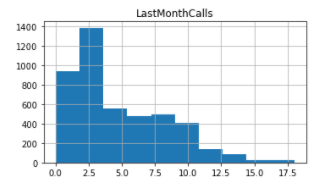
**Existing Policy Tenure:** There are about 2500 customer who took policies at least a year ago.



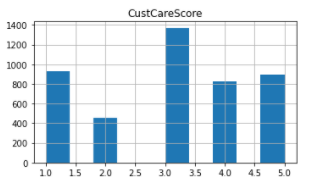
**Sum Assured:** The major count of the sum assured is between 400K and 600K



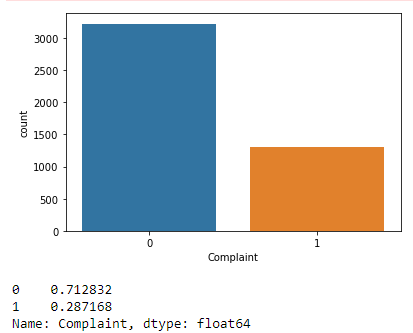
**Last Month Calls:** We have received at least 2 calls from nearly 1400 customers

****

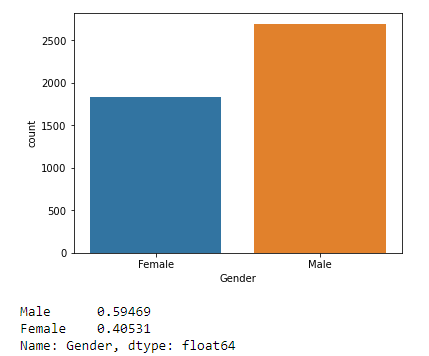
**Customer Care Score:** Marjory of the customer lies between the Customer care score of 3 – 3.5.



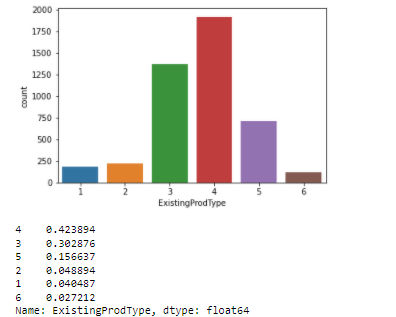
**Complaints:** Below graph shows the count of the customers who complained and not complained.



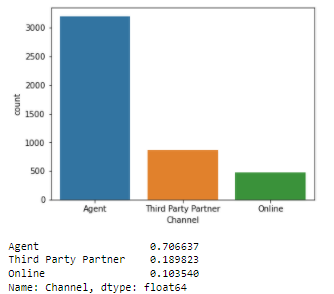
**Gender:** Below graph gives me the count of Gender



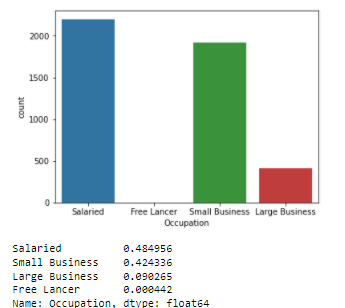
**Existing Policy Type:** Below Count Plot shows the existing policy types



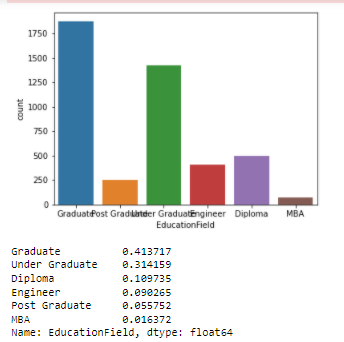
Blow graph shows the channel data



Below graph shows the occupation



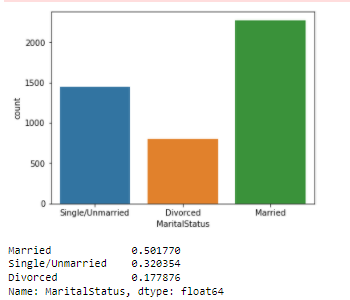
Below graph shows the education field



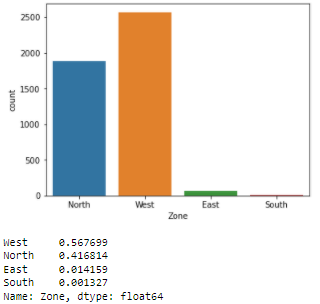
Below graph shows the designation



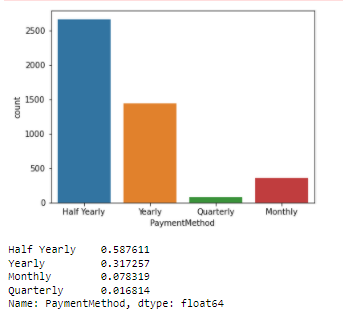
Below graph shows the marital status



Below graph shows the various Zones



Below graph shows various payment methods

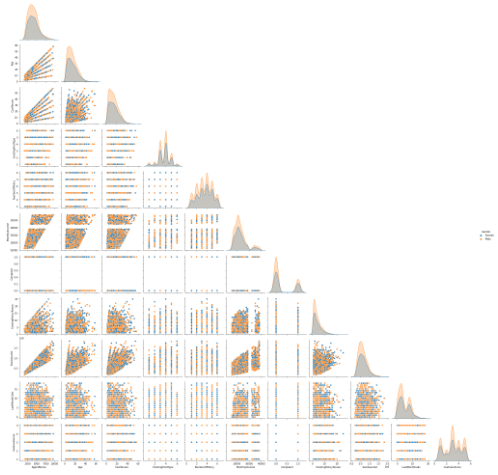


### Problem 3.B

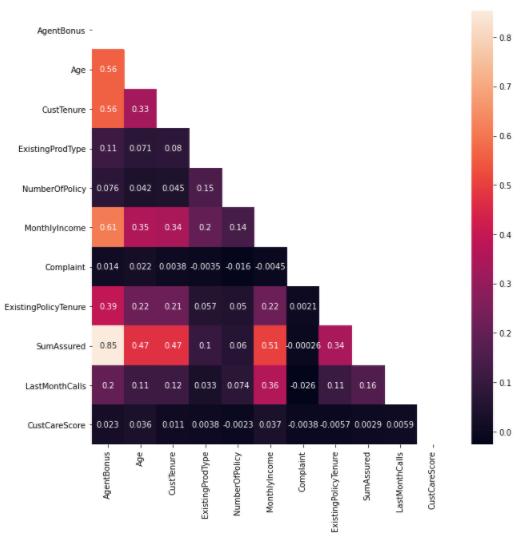
Bivariate analysis (relationship between different variables, correlations)

**Resolution:**

Below Pair plot shows the pairwise relationships in a dataset



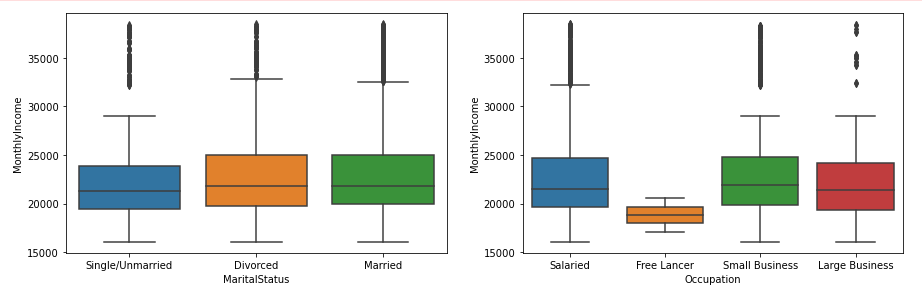
Below Correlation / Heat Map shows Strong Correlation

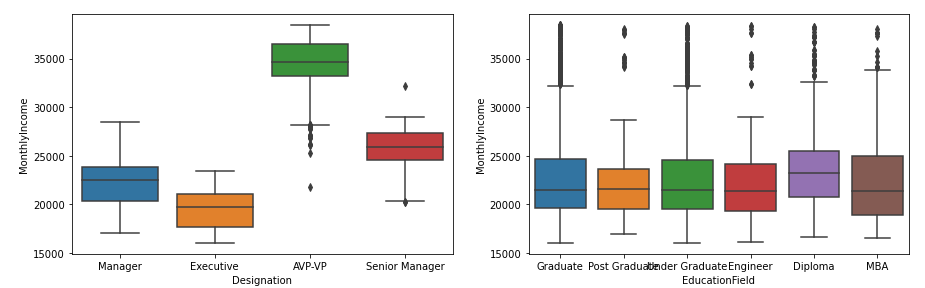


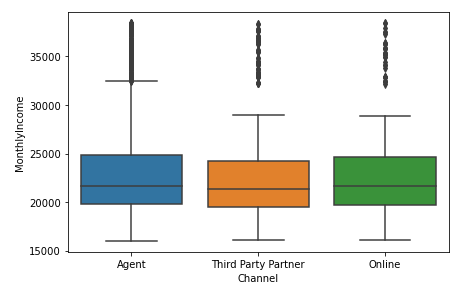
Top 5 strong correlations:

* SumAssured & AgentBonus
* MonthlyIncome & AgentBonus
* CustTenure & AgentBonus
* Age & AgentBonus
* MonthlyIncome & SumAssured

Below box plots shows relationship between MonthlyIncome & categorical variables

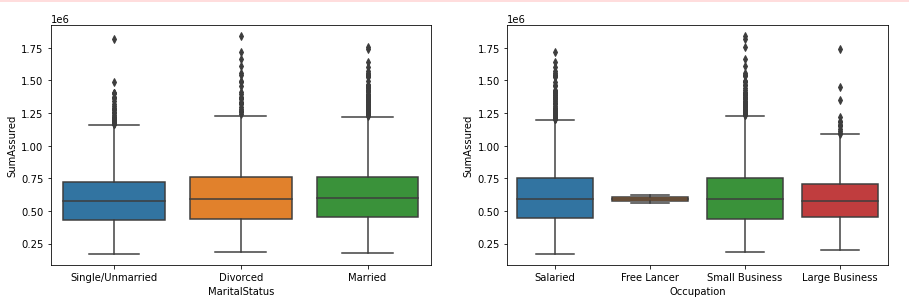


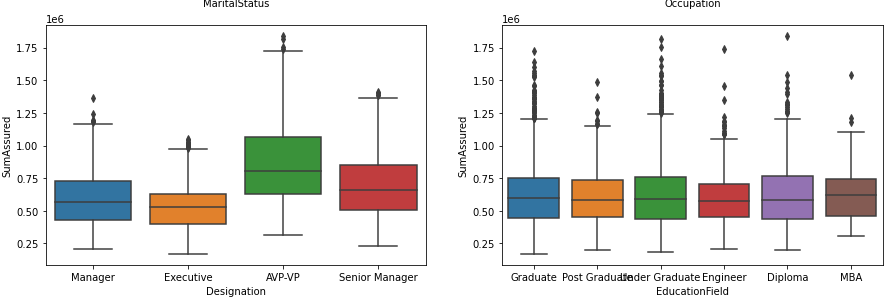


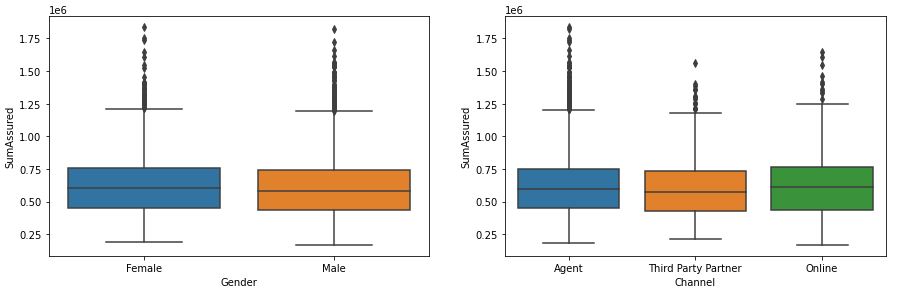


Customer Designation creates clear groups for MonthlyIncome of the customer so Missing Values in MonthlyIncome will be filled considering means of every group

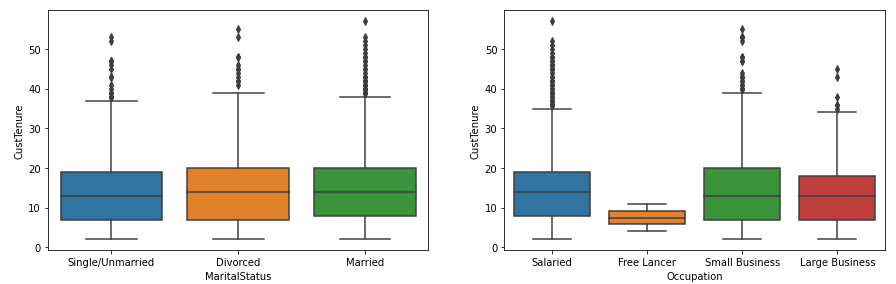
Below box plots shows relation between SumAssured & categorical variables

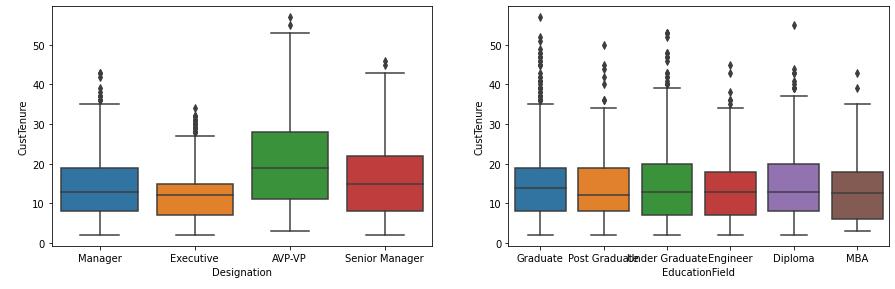


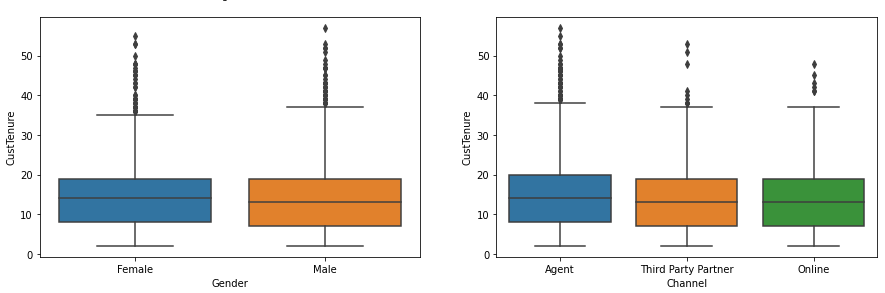




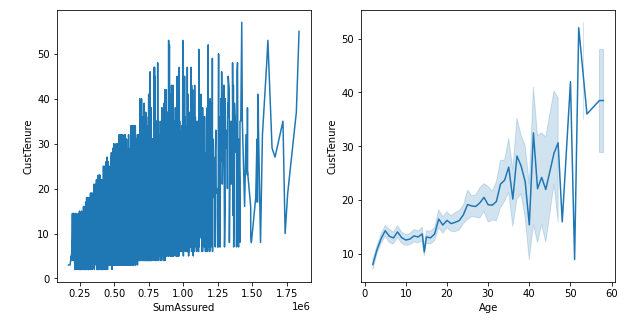
Below box plots shows the relationship between CustTenure & categorical variables



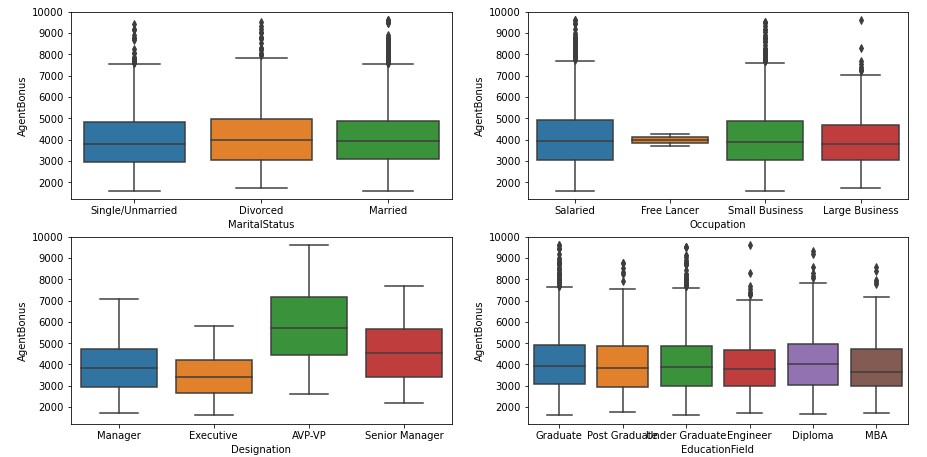


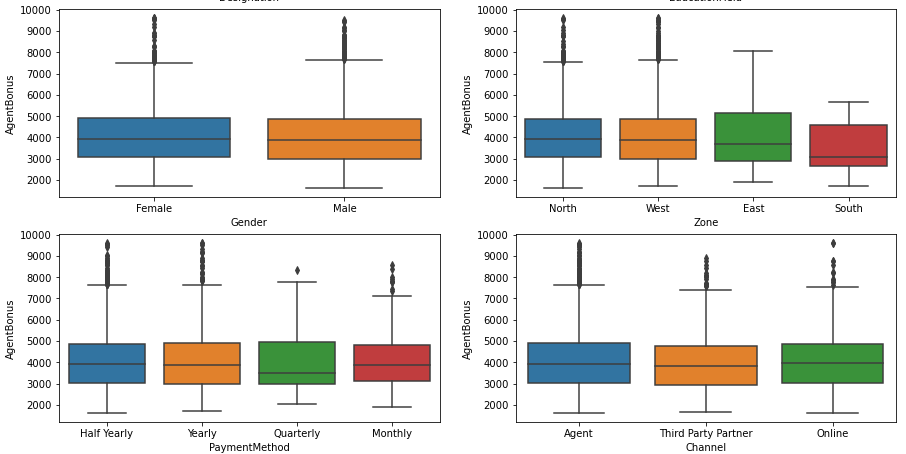


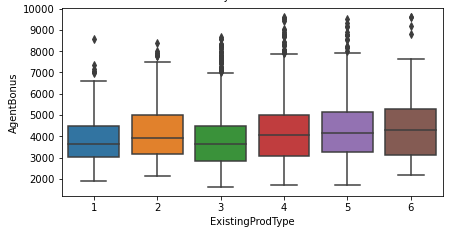
Below graphs shows the relationship between Customer tenure VS Sum Assured and age



Below Box plots shows relationship between AgentBonus & categorical variables







The dependent variable AgentBonus show some variation with Designation from above Boxplots, but doesn't seem to show any relationship otherwise with other categorical variable. We will test it further with ANOVA

### Problem 3.c

Removal of unwanted variables (if applicable)

**Resolution:**

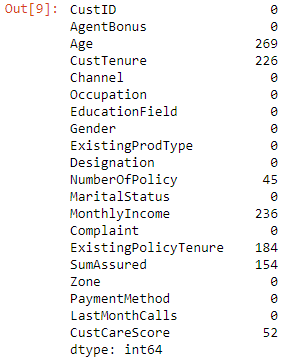
We have no removed any unwanted variables as there were not many, there was only customer ID and it was not making any difference so did not remove any.

### Problem 3.D

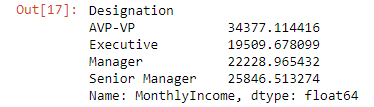
Missing Value treatment (if applicable)

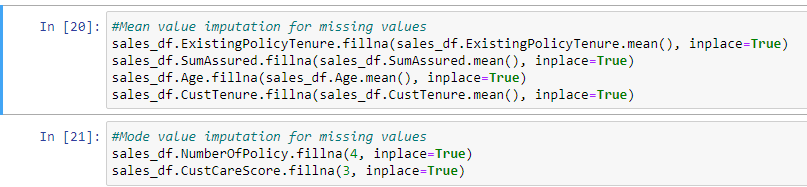
**Resolution:**

First we check for missing values : We can see that there are lot of missing values in multiple columns

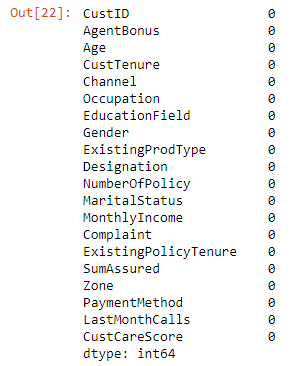
****

Then we are substituting Missing values for MonthlyIncome





Once the imputation is done we can check for data and there are no missing values

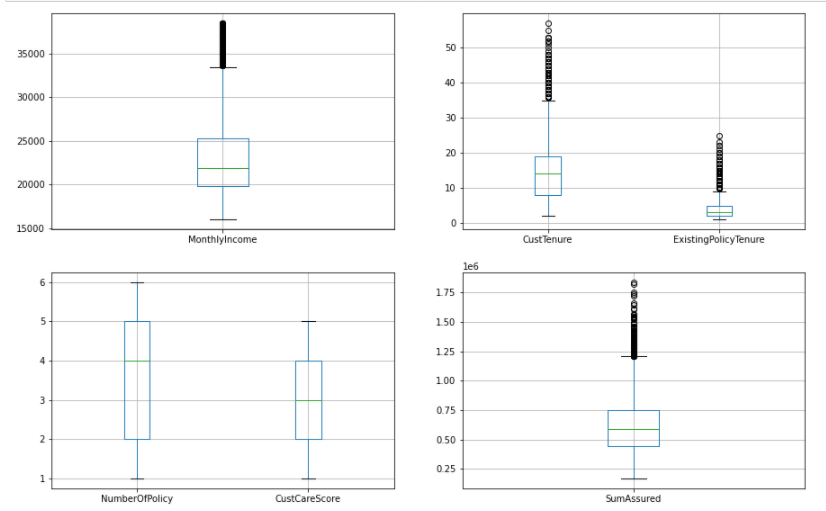


### Problem 3.e

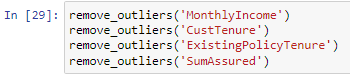
Outlier treatment (if required)

**Resolution:**

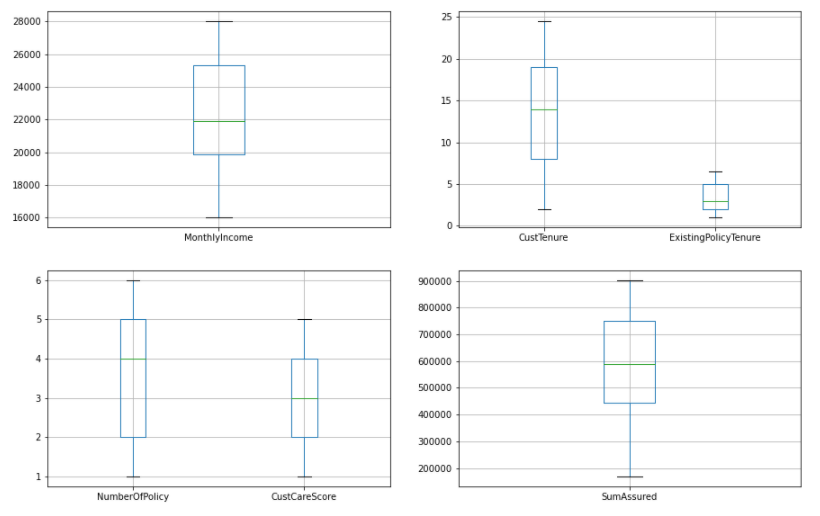
First we check for the outliers and box plots shows the outliers



Based on the above graph we remove the below outliers



After removing the outliers, we can see from the below box plots that there are no outliers



### Problem 3.F

Variable transformation (if applicable)

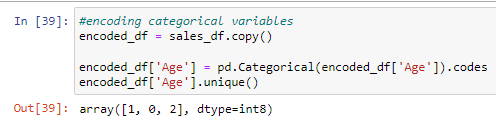
**Resolution:**

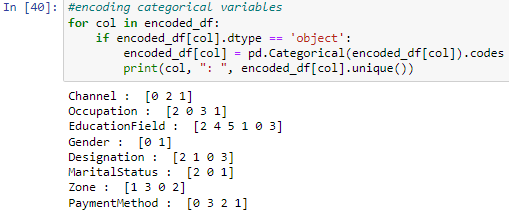
Yes, there were few variables which were transformation below are the transformed variables.

Categories created for age:



Encoding categorical variables



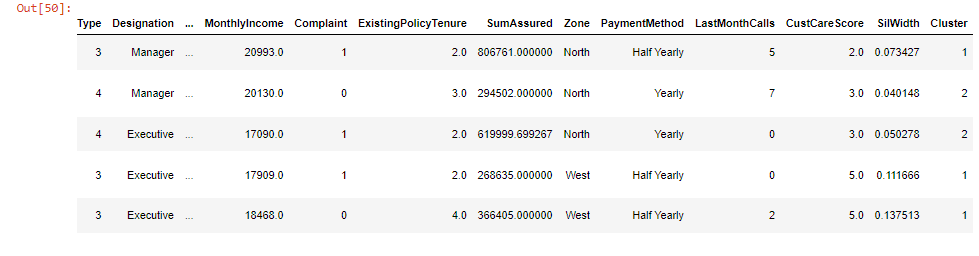


### Problem 3.G

Addition of new variables (if required)

**Resolution:**

There were 2 new variables added in the data set



We will consider 3 ultimate clusters as that is giving us very fewer negative silhouette widths than 4 clusters

Note: Positive silhouette width suggests that the observation belong to the correct cluster, negative would be opposite.

## Problem 4: Business insights from EDA

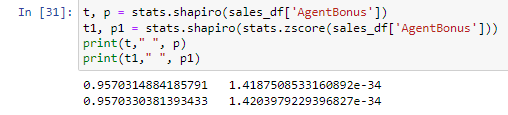
### Problem 4.a

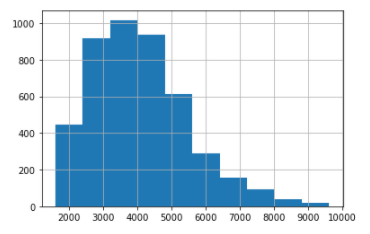
Is the data unbalanced? If so, what can be done? Please explain in the context of the business

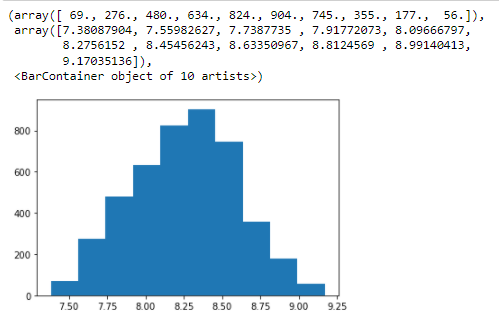
**Resolution:**

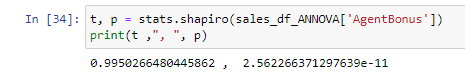
To check whether the data is unbalanced, we checked using Shapiro-wilk to test the normality of the continues variables

H0- Data is normal H1 - Data is not normal

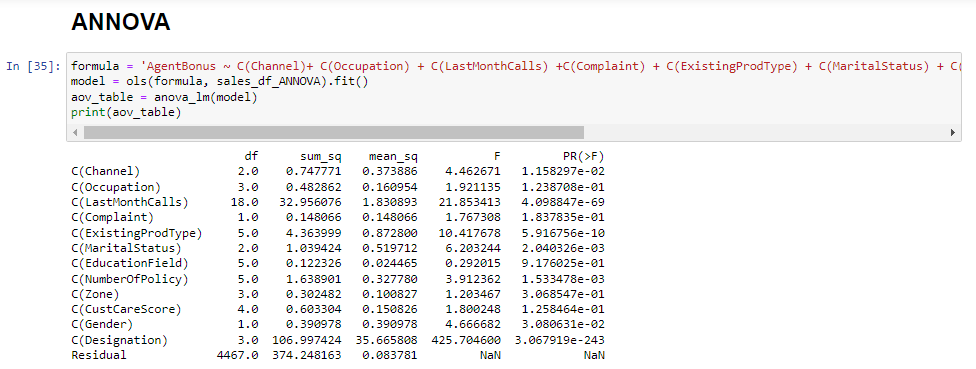


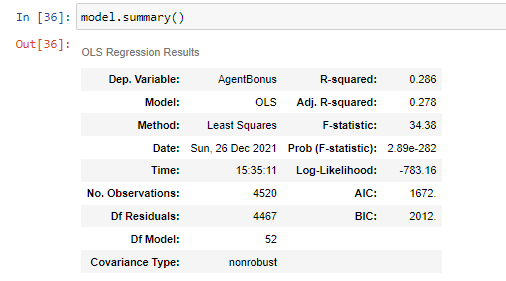






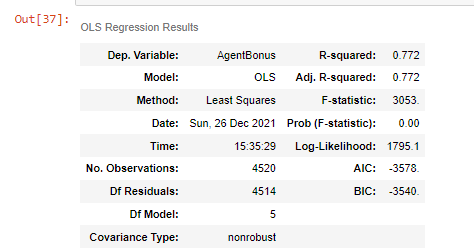
The dependent variable was tried to convert into a normal distribution, however the results were still unsuccessful. It will be assumed Normal for further ANNOVA test

****

****

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.



Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

[2] The condition number is large, 4.07e+06. This might indicate that there are

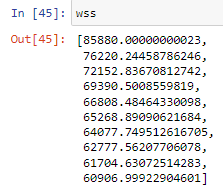
Strong multicollinearity or other numerical problems.

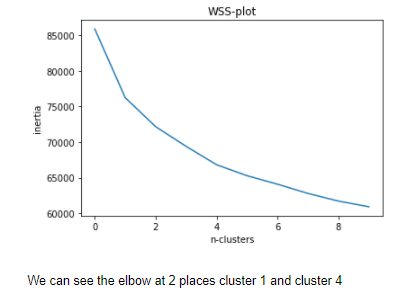
### Problem 4.b

Any business insights using clustering (if applicable)

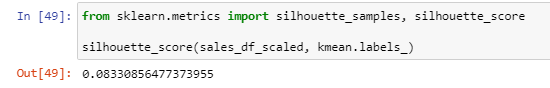
**Resolution:**

Below are the findings after using the clustering





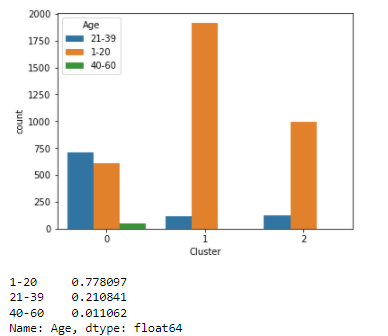
Below is the silhouette score

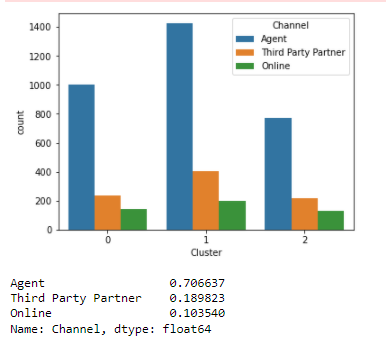


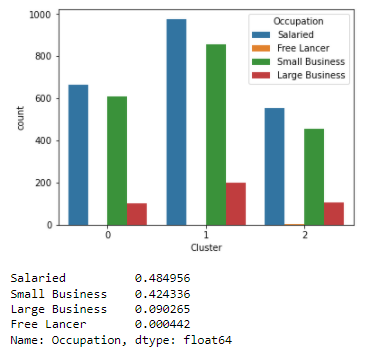
We will consider 3 ultimate clusters as that is giving us very fewer negative silhouette widths than 4 clusters

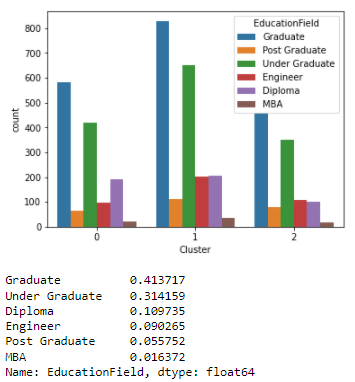
Note: Positive silhouette width suggests that the observation belong to the correct cluster, negative would be opposite.

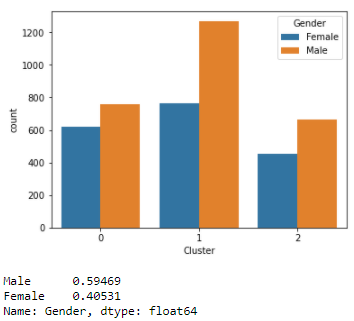
Below graphs shows the findings with variables VS Clusters

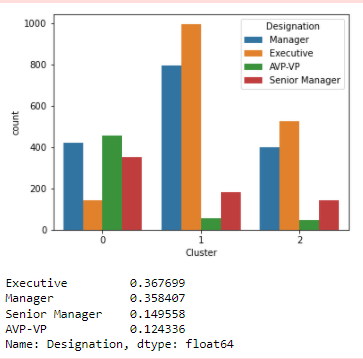


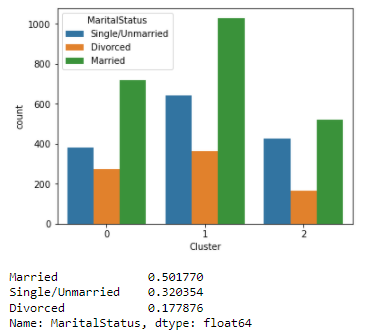


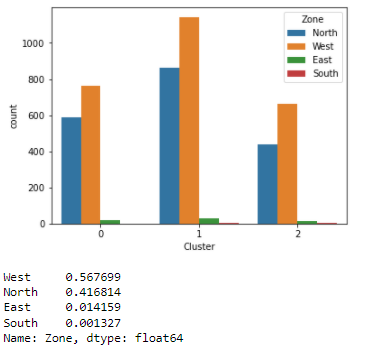


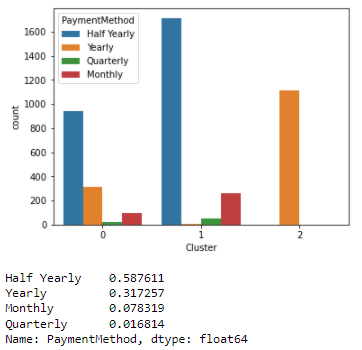










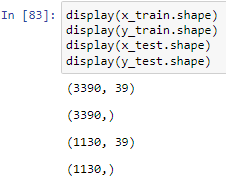


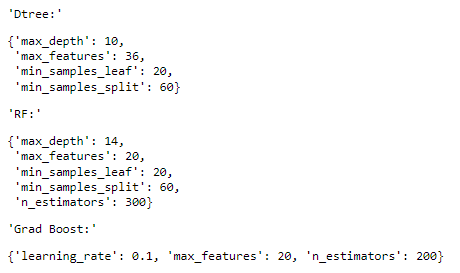
### Problem 4.c

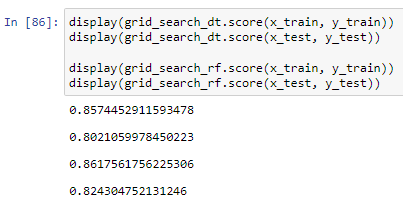
Any other business insights

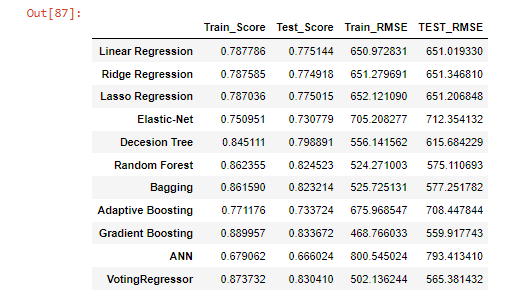
**Resolution:**

We have performed Train and Test as well for ‘sales\_df\_encoded’ data and below are the additional insights.









The End

Thakur Arun Singh

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*