# 1. Problem Understanding

## a) 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

## b) Need of the study/project

In any sales, designing an appropriate engagement programme helps in improving the overall business. One way to ensure a successful engagement programme is to create incentive programme.

## c) Understanding business/social opportunity

with the Sudden rise in the need of life insurance due to covid, As a insurance company it is a right time to capture the market.in order to utilize this opportunity, helping in increasing the customer base

# 2. Data Report

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

The data seems to be collected from the customer database containing basic customer information and the interaction recency the agents had with them with some agent info

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

Graphical user interface, application

Description automatically generated

Table

Description automatically generated with medium confidence

## c) Understanding of attributes (variable info, renaming if required)

we found attributes in the data that needed the replacement some of them are listed below

data = data.replace('UG', 'Under Graduate')

data = data.replace('Fe male', 'Female')

data = data.replace('Exe', 'Executive')

data = data.replace('Laarge Business', 'Large Business')

# 3. Exploratory Data Analysis

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

below are the univariate analysis. All the chart shows that some of the categorical data are skewed

Chart, bar chart

Description automatically generated Chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart, bar chart, box and whisker chart

Description automatically generated

## b) Bivariate analysis (relationship between different variables , correlations)

Chart, waterfall chart

Description automatically generated Graphical user interface

Description automatically generated with medium confidence

Chart, bar chart

Description automatically generated Chart, treemap chart

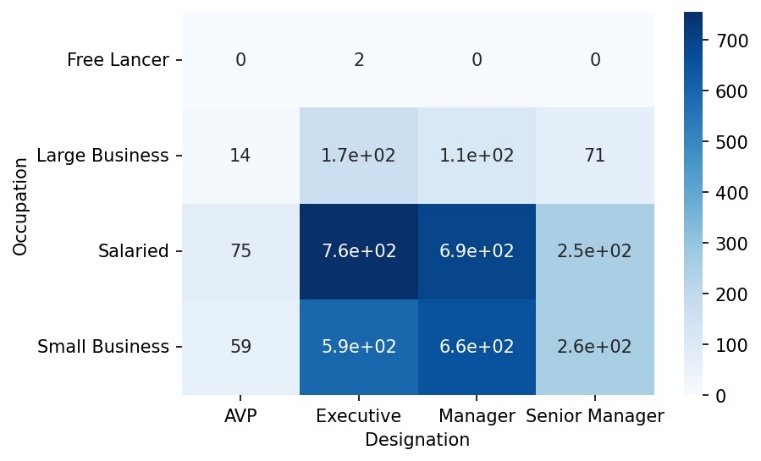
Description automatically generated

A picture containing table

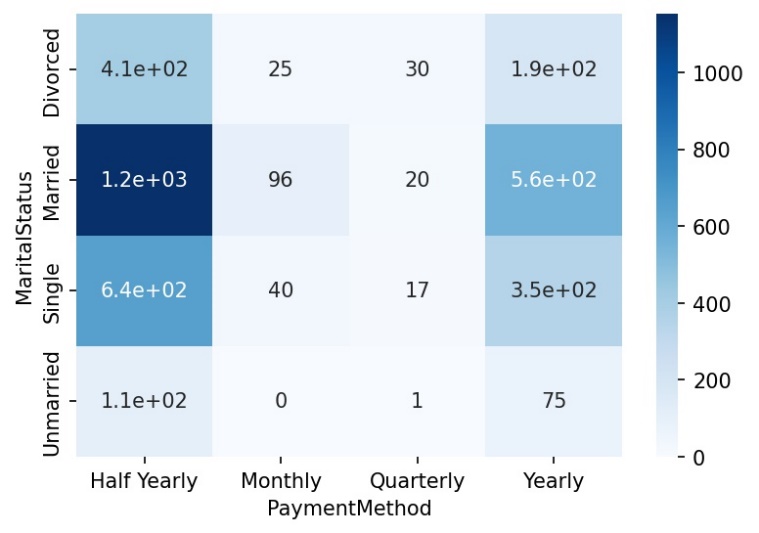
Description automatically generated Chart

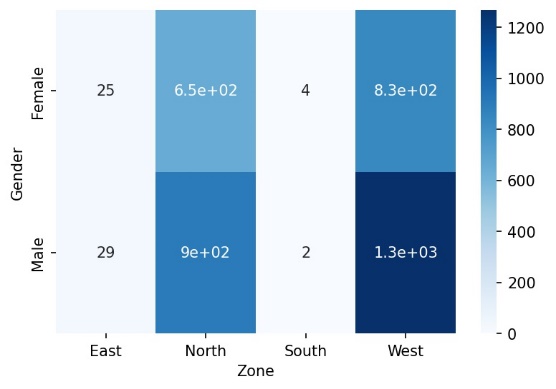
Description automatically generated

Chart

Description automatically generated with medium confidence 

Chart, bar chart

Description automatically generated 

 Chart, bar chart

Description automatically generated

Chart, treemap chart

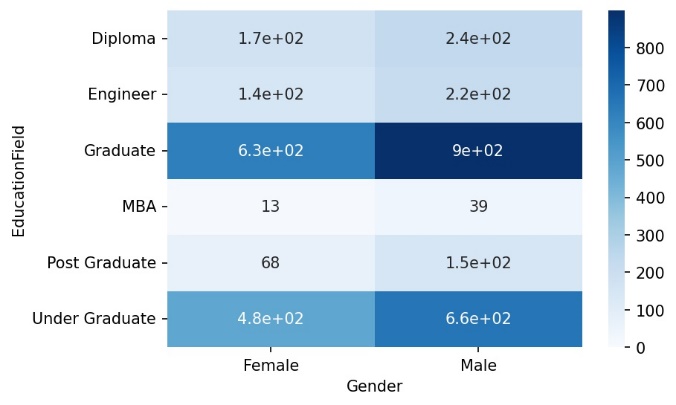
Description automatically generated 

Chart

Description automatically generated A picture containing table

Description automatically generated

Table

Description automatically generated 

Table

Description automatically generated Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated Chart

Description automatically generated

Chart, waterfall chart

Description automatically generated Chart

Description automatically generated

Chart, treemap chart

Description automatically generated Chart, treemap chart

Description automatically generated

Chart

Description automatically generated Chart, bar chart, waterfall chart

Description automatically generated

Chart

Description automatically generated

## c) Removal of unwanted variables (if applicable),Missing Value treatment (if applicable)

We removed CustID because it is a key field from database

There are couple of variables had missing value we used knn-imputation to replace missing value

|  |  |
| --- | --- |
| Pre correction of missing value | Post correction of missing value |
|  | Table  Description automatically generated |

## d) Outlier treatment (if required)

for outlier correction we plan to use iqr as our benchmark to correct it

There is also a need to modify the age column, which is currently underway of changes since some of the age are underage for the qualification or designation of customer

## e) Variable transformation (if applicable)

all the object datatype are converted to categorical datatype

feature: Channel

Categories (3, object): ['Agent', 'Online', 'Third Party Partner']

[0 2 1]

feature: Occupation

Categories (4, object): ['Free Lancer', 'Large Business', 'Salaried', 'Small Business']

[2 0 3 1]

feature: EducationField

Categories (6, object): ['Diploma', 'Engineer', 'Graduate', 'MBA', 'Post Graduate', 'Under Graduate']

[2 4 5 1 0 3]

feature: Gender

Categories (2, object): ['Female', 'Male']

[0 1]

feature: Designation

Categories (4, object): ['AVP', 'Executive', 'Manager', 'Senior Manager']

[2 1 0 3]

feature: MaritalStatus

Categories (4, object): ['Divorced', 'Married', 'Single', 'Unmarried']

[2 0 3 1]

feature: Zone

Categories (4, object): ['East', 'North', 'South', 'West']

[1 3 0 2]

feature: PaymentMethod

Categories (4, object): ['Half Yearly', 'Monthly', 'Quarterly', 'Yearly']

[0 3 2 1]

# 4. Business insights from EDA

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

The data is unbalanced in the following categories

* Designation
* Marital Status
* Occupation
* Payment Method
* Zone

Some them can merge couple of categories while for others we exclude certain categories

## b) Any other business insights

we see most of the business is from north and west zone, we can ask the company to create a program to focus on providing upskill programs for agents from east and south zone