CAPSTONE PROJECT

Insurance Cost Estimation

By:- Raghvendra Singh
-raghavsingh0027@gmail.com

INDEX

Contents

LIST OF FIGURES

DATA DICTIONARY

Name	Description
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
adventure_sports	Customer is involved with adventure sports like climbing, diving etc.
Occupation	Occupation of the customer
visited_doctor_last_1_year	Number of times customer has visited doctor in last one year
cholesterol_level	Cholesterol level of the customers while applying for insurance
daily_avg_steps	Average daily steps walked by customers
age	Age of the customer
heart_decs_history	Any past heart diseases
other_major_decs_history	Any past major diseases apart from heart like any operation
Gender	Gender of the customer
avg_glucose_level	Average glucose level of the customer while applying the insurance
bmi	BMI of the customer while applying the insurance
smoking_status	Smoking status of the customer
Year_last_admitted	When customer have been admitted in the hospital last time
Location	Location of the hospital
weight	Weight of the customer
covered_by_any_other_company	Customer is covered from any other insurance company
Alcohol	Alcohol consumption status of the customer
exercise	Regular exercise status of the customer
weight_change_in_last_one_year	How much variation has been seen in the weight of the customer in last year
fat_percentage	Fat percentage of the customer while applying the insurance
insurance_cost	Total Insurance cost

1. INTRODUCTION TO BUSINESS PROBLEM

DEFINING THE PROBLEM STATEMENT:

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

Goal & Objective:

The objective of this exercise is to **build a model, using data that provide the optimum insurance cost for an individual.** You have to use the health and habit related parameters for **the estimated cost of insurance**

NEED OF THE STUDY & PROJECT:

- The life insurance industry play an important role in securing the individuals from any health emergencies that might disrupt the financial wellbeing of families.
- Health insurance is important for financial security, financial safety and well being of the individual. In case of health emergencies a health coverage can ensure that the treatment costs are taken care of without adding financial burden on family.

Significance of a good health insurance plan:

1. **Financial Protection**: Health insurance helps cover the high costs of medical care, protecting you from unexpected expenses.
2. **Access to Medical Services**: Insurance often provides access to a wider network of healthcare providers and specialists.
3. **Chronic Disease Management**: Insurance can offer resources and support for managing chronic illnesses, improving quality of life.
4. **Prescription Drug Coverage**: Health plans often include coverage for medications, reducing out-of-pocket costs for prescriptions.
5. **Emergency Services**: In emergencies, health insurance covers the costs of ambulance services and emergency room visits.
6. **Maternity Coverage**: Health insurance often provides coverage for prenatal and postnatal care, as well as childbirth.
7. **Hospitalization Coverage**: Insurance helps cover the costs of hospital stays, surgeries, and other inpatient services.
8. **Peace of Mind**: Having health insurance can reduce stress and anxiety about potential medical costs.

9. **Family Coverage Options**: Many policies allow you to add family members, ensuring everyone has access to care.

10. **Cashless Transactions**: With certain plans, you can receive treatment without upfront payment, simplifying the process during emergencies.

Need of the project:

- The study will help us understand the cost which diffident individuals need to be charged to ensure that all the desirable features of a good health insurance policy is provided.
- Every individual is different by their way of living, physical activity, behaviour, education level, occupation, marital status and many other features that characterizes needs of each individual. Therefore, every person might not be able to bear the cost equally.
- Based on the need and several features we need to understand the costing and ensure that best cost of insurance is offered to customers while ensuring fulfilment of business goals & objectives.

UNDERSTANDING SOCIAL REPOSNIBILITY & BUSINESS OPPORTUNITY

- An insurance company plays a significant role in providing not merely financial security but social responsibility as well. By securing the health of and individuals an insurance company also protects people from poverty, financial losses, securing dependents, freedom for any debt. This ensures that families and overall social structure is protected from financial crisis as any major health crisis can lead to economic crisis as well.
- Insurance company should not just provide financial coverage against the health treatment needs but must proactively engage its customers for overall wellbeing, diet plan, fitness plans and community activities to ensure social wellbeing, emotional health as well as reducing the possibilities of health issues.
- By providing such engaging group/community social programs free of cost company can do better business by reducing number of claims while fulfilling its social responsibility.

Business Opportunity:

- Offering a customizable health insurance plan to individuals
- Catering to different individuals based on their needs and preferences.
- Providing corporate health insurance plans to ensure organizations are protected from loss of manpower due to health issues.

2. EDA & Business Implication

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

The dataset is collected & provided by the organization for analysis based on internal parameters. Little information available regarding the time, frequency and methodology.

Visual & Non-Visual inspection of data:

Shape:

```
(25000, 24)
```

- The data set has 24 columns and 25000 rows

Head:

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level
0	5000	3	1	1	Salried	2	125 to 150
1	5001	0	0	0	Student	4	150 to 175
2	5002	1	0	0	Business	4	200 to 225
3	5003	7	4	0	Business	2	175 to 200
4	5004	3	1	0	Student	2	150 to 175

5 rows × 24 columns

Figure 1 Head Dataset

- Applicant ID is not required column but is only for identification purpose of individual.
- Years of insurance in numerical and continuous column.
- Regular checkup last year is also numerical and continuous.
- Adventure sports is a binary column
- Occupation column is object columns.

Lets check all the column heads:

daily_avg_steps	age	heart_decs_history	...	smoking_status	Year_last_admitted	
4866	28	1	...	Unknown	NaN	
6411	50	0	...	formerly smoked	NaN	
4509	68	0	...	formerly smoked	NaN	
6214	51	0	...	Unknown	NaN	
4938	44	0	...	never smoked	2004.0	E

Figure 2 Head 2

- Daily avg, age are continuous variable.
- Heart_decs _history is a binary column
- Year last is a date time column

Location	weight	covered_by_any_other_company	Alcohol	exercise
Chennai	67	N	Rare	Moderate
Jaipur	58	N	Rare	Moderate
Jaipur	73	N	Daily	Extreme
Chennai	71	Y	Rare	No
Bangalore	74	N	No	Extreme

Figure 3 Head 3

weight_change_in_last_one_year	fat_percentage	insurance_cost
1	25	20978
3	27	6170
0	32	28382
3	37	27148
0	34	29616

Figure 4 Head 4

- Insurance cost is the target variable as per the problem. We need to predict the insurance cost for different individuals. It is a continuous variable and therefore a Regression problem.

Data Info:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    25000 non-null  int64
 1   years_of_insurance_with_us      25000 non-null  int64
 2   regular_checkup_lasy_year       25000 non-null  int64
 3   adventure_sports                25000 non-null  int64
 4   Occupation                      25000 non-null  object
 5   visited_doctor_last_1_year      25000 non-null  int64
 6   cholesterol_level               25000 non-null  object
 7   daily_avg_steps                 25000 non-null  int64
 8   age                             25000 non-null  int64
 9   heart_decs_history              25000 non-null  int64
 10  other_major_decs_history        25000 non-null  int64
 11  Gender                          25000 non-null  object
 12  avg_glucose_level               25000 non-null  int64
 13  bmi                             25000 non-null  float64
 14  smoking_status                  25000 non-null  object
 15  Year_last_admitted              13119 non-null  datetime64[ns]
 16  Location                        25000 non-null  object
 17  weight                          25000 non-null  int64
 18  covered_by_any_other_company    25000 non-null  object
 19  Alcohol                         25000 non-null  object
 20  exercise                        25000 non-null  object
 21  weight_change_in_last_one_year  25000 non-null  int64
 22  fat_percentage                  25000 non-null  int64
 23  insurance_cost                  25000 non-null  int64
dtypes: datetime64[ns](1), float64(1), int64(14), object(8)
memory usage: 4.6+ MB
```

Figure 5 Data Info

- Data set has 1 Datetime variable, 15 numeric and 8 categorical variables.
- Year last admistted has lot of missing values

Data Description:

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year
count	25000.000000	25000.000000	25000.000000
mean	17499.500000	4.089040	0.773680
min	5000.000000	0.000000	0.000000
25%	11249.750000	2.000000	0.000000
50%	17499.500000	4.000000	0.000000
75%	23749.250000	6.000000	1.000000
max	29999.000000	8.000000	5.000000
std	7217.022701	2.606612	1.199449

	adventure_sports	visited_doctor_last_1_year	daily_avg_steps
count	25000.000000	25000.000000	25000.000000
mean	0.081720	3.104200	5215.889320
min	0.000000	0.000000	2034.000000
25%	0.000000	2.000000	4543.000000
50%	0.000000	3.000000	5089.000000
75%	0.000000	4.000000	5730.000000
max	1.000000	12.000000	11255.000000
std	0.273943	1.141663	1053.179748

	age	heart_decs_history	other_major_decs_history
count	25000.000000	25000.000000	25000.000000
mean	44.918320	0.054640	0.098160
min	16.000000	0.000000	0.000000
25%	31.000000	0.000000	0.000000
50%	45.000000	0.000000	0.000000
75%	59.000000	0.000000	0.000000
max	74.000000	1.000000	1.000000
std	16.107492	0.227281	0.297537

Figure 6 Data Desctiption 1

- Average years of insurance with us: 4.08 yrs
- Avg. Regular checkup last year: 0.77 times
- Avg. Visit doctors last 1 year : 3 times
- Avg. Daily steps: 5215 steps

	avg_glucose_level	bmi	Year_last_admitted
count	25000.000000	25000.000000	13119
mean	167.530000	31.393328	2003-11-23 00:14:35.920420608
min	57.000000	12.300000	1990-01-01 00:00:00
25%	113.000000	26.300000	1997-01-01 00:00:00
50%	168.000000	30.800000	2004-01-01 00:00:00
75%	222.000000	35.300000	2010-01-01 00:00:00
max	277.000000	100.600000	2018-01-01 00:00:00
std	62.729712	7.718998	NaN

	weight	weight_change_in_last_one_year	fat_percentage
count	25000.000000	25000.000000	25000.000000
mean	71.610480	2.517960	28.812280
min	52.000000	0.000000	11.000000
25%	64.000000	1.000000	21.000000
50%	72.000000	3.000000	31.000000
75%	78.000000	4.000000	36.000000
max	96.000000	6.000000	42.000000
std	9.325183	1.690335	8.632382

	insurance_cost
count	25000.000000
mean	27147.407680
min	2468.000000
25%	16042.000000
50%	27148.000000
75%	37020.000000
max	67870.000000
std	14323.691832

Figure 7 Data Description 2

Let see the average values for the variables:

- Glucose level avg: 167
- Bmi avg: 31.39
- Weight avg: 71.61 kg
- Weight change last 1 year avg: 2.51 kg
- Fat percentage avg: 28.8
- Insurance cost avg: 27147

Understanding of attributes (variable info, renaming if required)

Unique Values of Categorical Variables:

Occupation : ['Student', 'Business', 'Salried']

cholesterol_level : ['150 to 175', '125 to 150', '200 to 225', '175 to 200', '225 to 250']

Gender : ['Male', 'Female']

smoking_status : ['never smoked', 'Unknown', 'formerly smoked', 'smokes']

Location : ['Bangalore', 'Jaipur', 'Bhubaneswar', 'Mangalore', 'Delhi', 'Ahmedabad', 'Guwahati', 'Chennai', 'Kanpur', 'Nagpur', 'Mumbai', 'Lucknow', 'Pune', 'Kolkata', 'Surat']

covered_by_any_other_company : ['N', 'Y']

Alcohol : ['Rare', 'No', 'Daily']

exercise : ['Moderate', 'Extreme', 'No']

- We can see that there is little anomaly in categorical values. We need not do the renaming for any value or variables.

Unique Values of Numerical Values:

Yeas of insurance with us:

array([3, 0, 1, 7, 8, 4, 6, 5, 2]

age:

array([28, 50, 68, 51, 44, 39, 40, 46, 45, 38, 35, 49, 30, 71, 54, 20, 62,

23, 33, 34, 65, 25, 55, 60, 29, 63, 18, 26, 24, 47, 19, 57, 67, 32,

58, 69, 53, 70, 41, 43, 48, 31, 72, 52, 42, 59, 17, 61, 36, 66, 56,

21, 74, 64, 22, 27, 16, 37, 73]

fat_percentage:

array([25, 27, 32, 37, 34, 13, 16, 12, 38, 11, 21, 19, 22, 31, 29, 18, 24,

33, 36, 39, 20, 23, 40, 42, 17, 35, 41, 30, 15, 26, 28, 14]

- We do not need to change the values in numerical fiels as well after checking the unique values.

Understanding Target Variable:

Insurance cost is the target variable. Lets check some key facts about the variable.

- Avg. Insurance cost is 27147.40
- Standard Deviation is 14323.69
- Minimum Insurance cost is 2468
- Maximum Insurance cost is 7870

Exploratory Data Analysis

Univariate analysis:

- Boxplot of Target Variable: Insuarance_cost

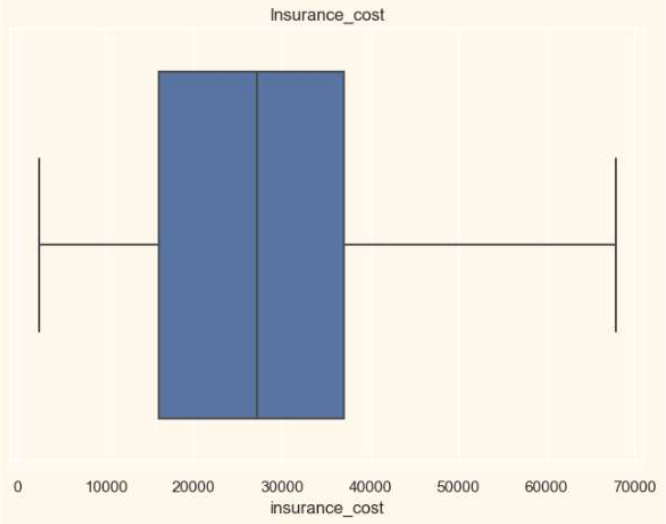

Figure 8 Boxplot: Insurance cost

- Insurance cost do not have any outliers. 20k to 40 k is range for which there are many individuals.

Lets see the distribution of data:

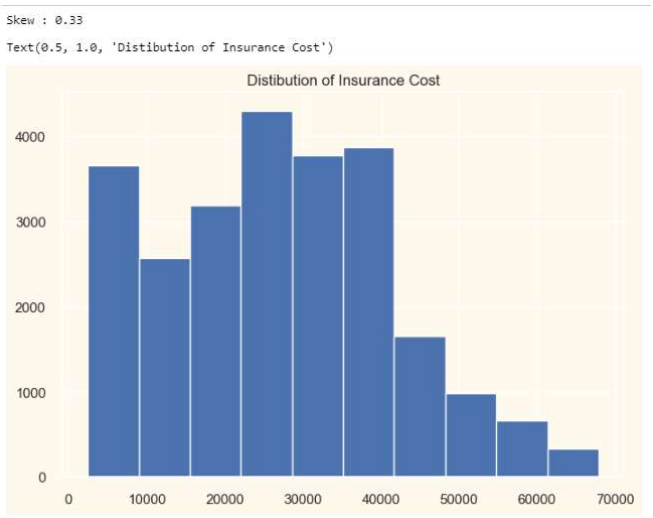

Figure 9 Histplot: Insurance cost

- The Insurance cost which is most frequently occurring is 25k. 25k to 40 k is the most occurring cost range.
- There is also many individuals with 10k insurance cost.

```
Years_of_insurance_with_us
Skew : -0.08
```

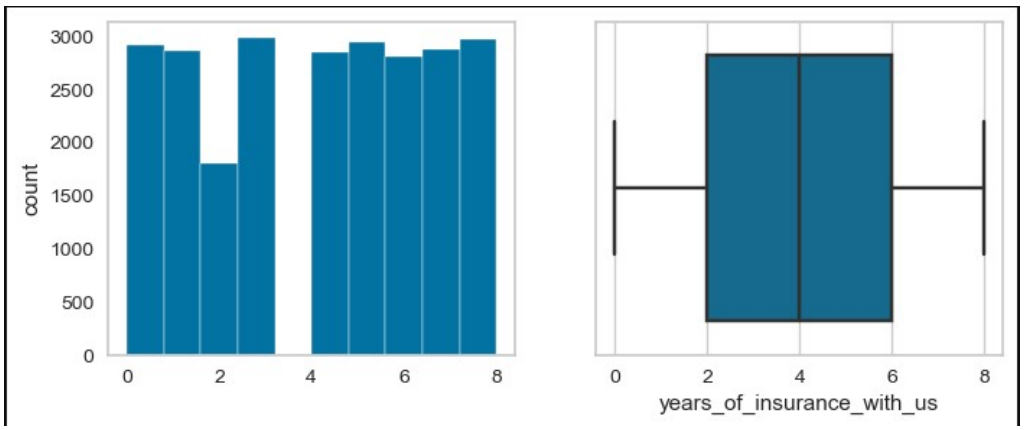

Figure 10 Histplot & Boxplot: Years of insurance with us

- Skewness is -.08 which shows symmetrically distributed data. No outliers found in boxplot.

```
regular_checkup_lasy_year
Skew : 1.61
```

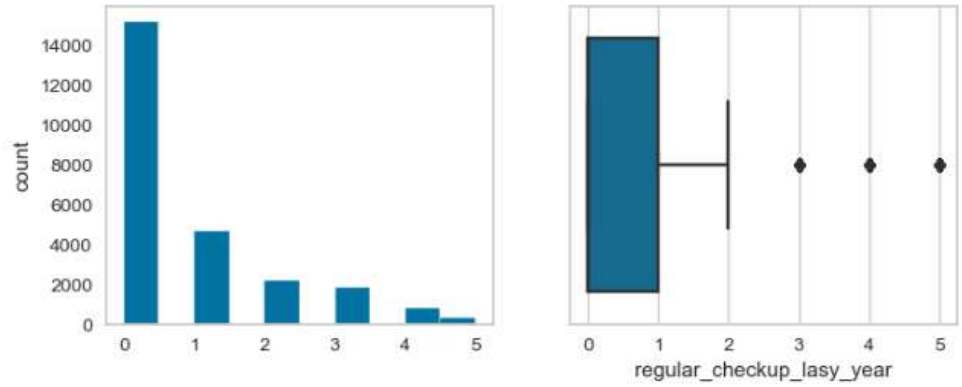

Figure 11 Histplot & Boxplot: Regular check-up last year

- Skewness is 1.61 which shows asymmetrically distributed data. Outliers found in boxplot.

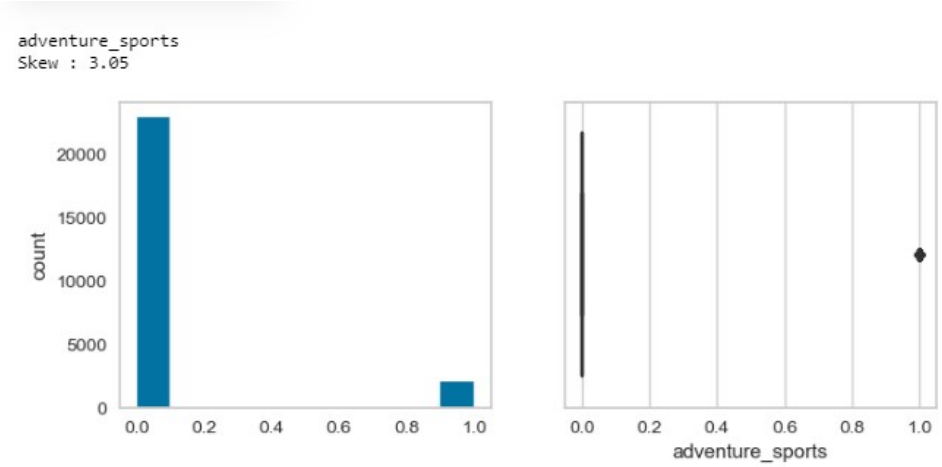

Figure 12 Histplot & Boxplot: adventure sports

- Skewness is 3.08 which shows asymmetrically distributed data. Binary data type outlier should not be treated for this.

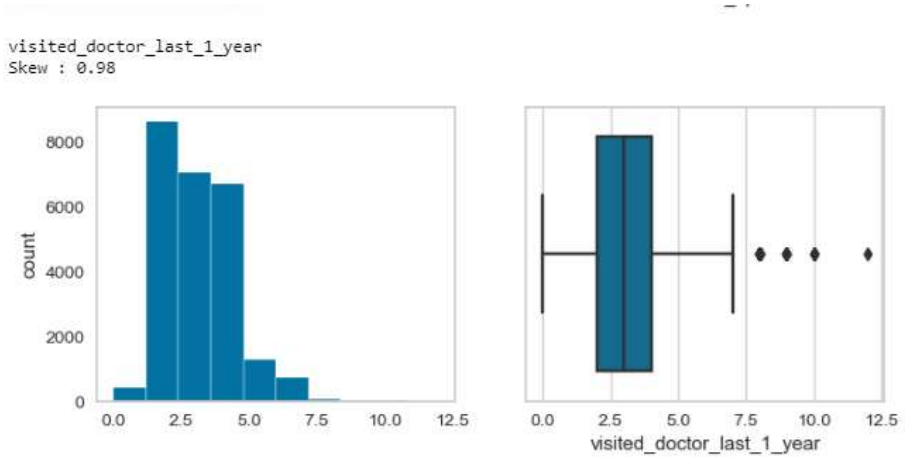

Figure 13 Histplot & Boxplot: visited doctor last 1 year

- Skewness is .98 which shows morderately asymmetrically distributed data. Outliers found in boxplot.

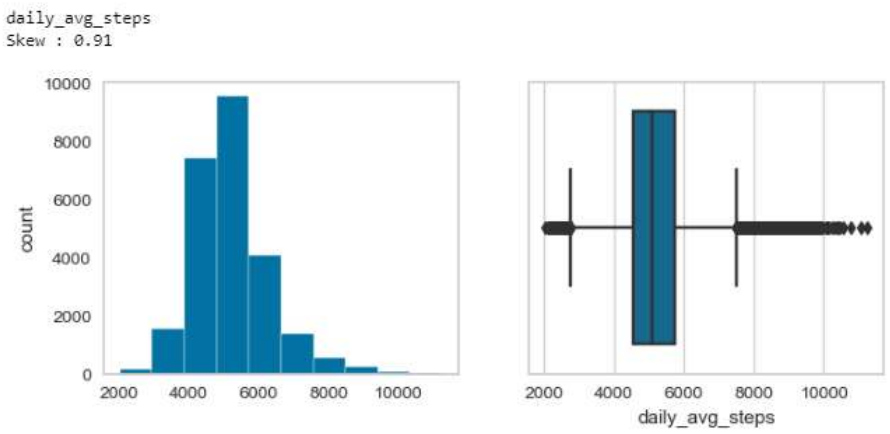

Figure 14 Histplot & Boxplot: Daily avg steps

- Skewness is .91 which shows morderately asymmetrically distributed data. Outliers found in boxplot.

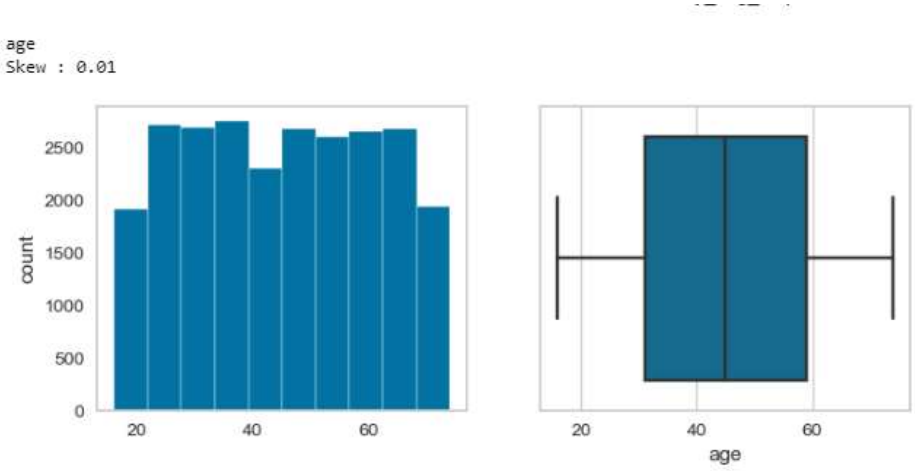

Figure 15 Histplot & Boxplot: age

- Skewness is .01 which shows symmetrically distributed data. No Outliers found in boxplot.

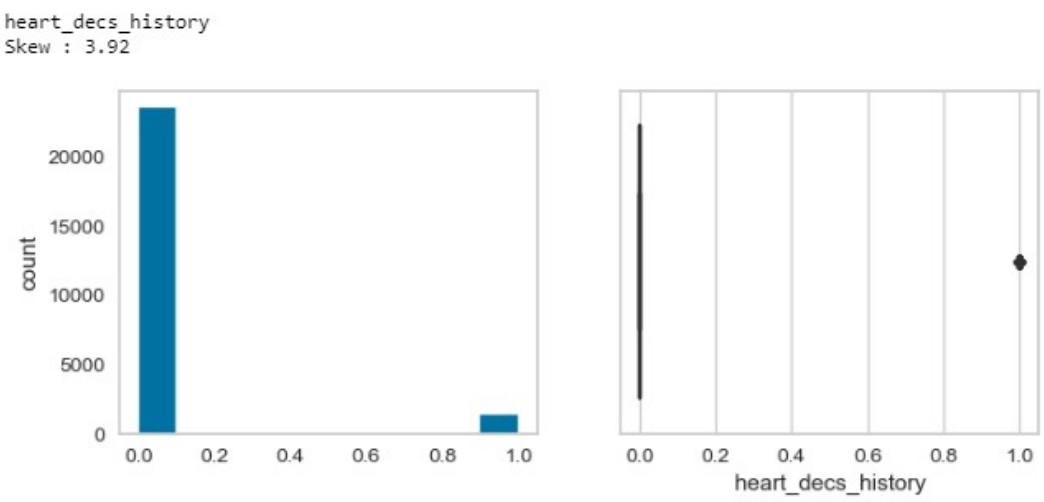

Figure 16 Histplot & Boxplot: heart decs history

- Skewness is 3.92. Outliers found in boxplot. However, binary type data so outliers should not be treated.

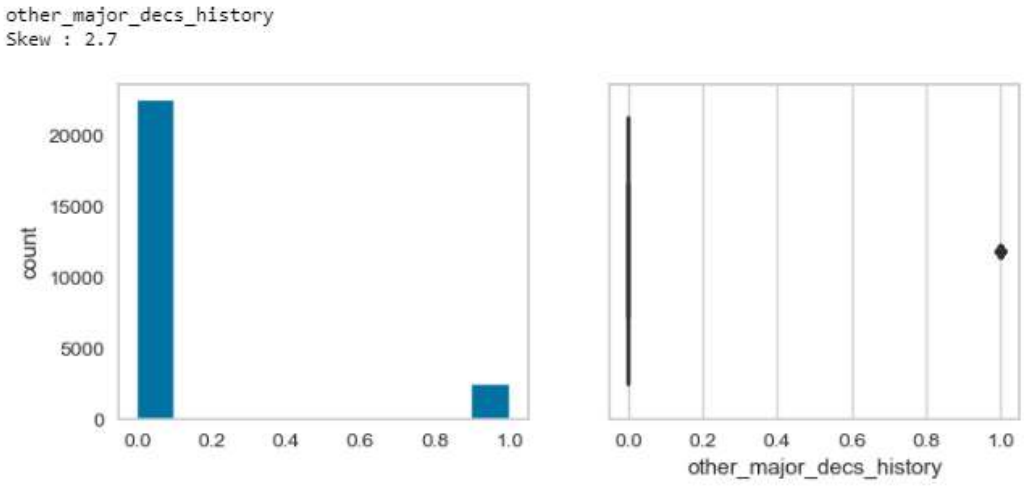

Figure 17 Histplot & Boxplot: other major decs history

- Skewness is 2.7. Outliers found in boxplot. However, binary type data so outliers should not be treated.

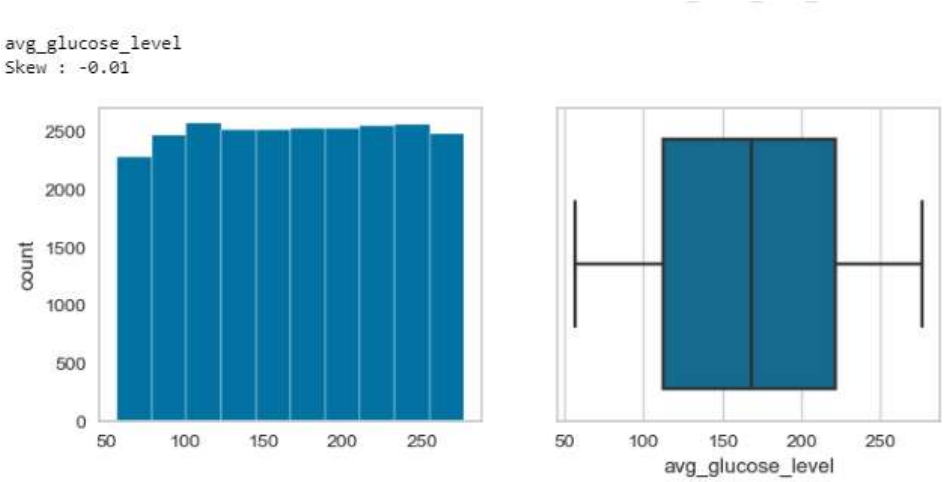

Figure 18 Histplot & Boxplot: avg glucose level

- Skewness is .3.92. Outliers found in boxplot. However, binary type data so outliers should not be treated.

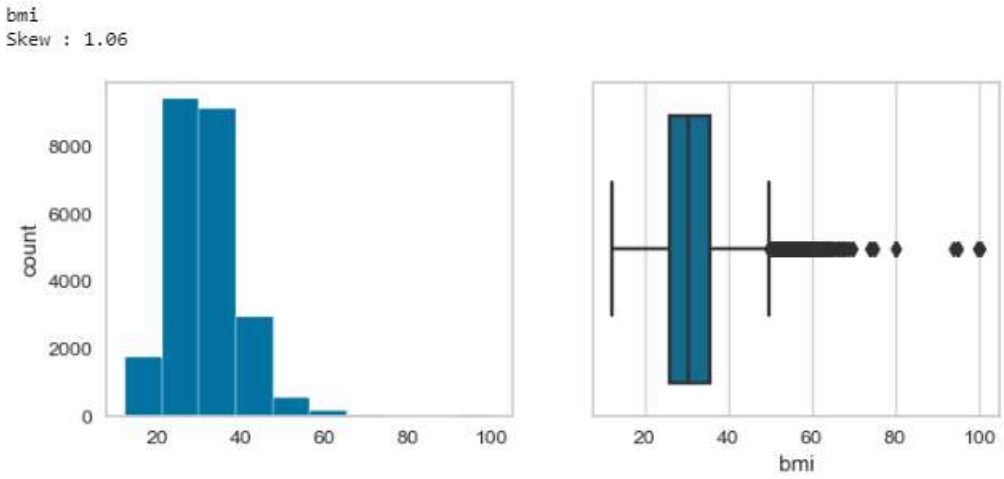

Figure 19 Histplot & Boxplot: bmi

- Skewness is 1.06 which shows morderately asymmetrically distributed data. Outliers found in boxplot.

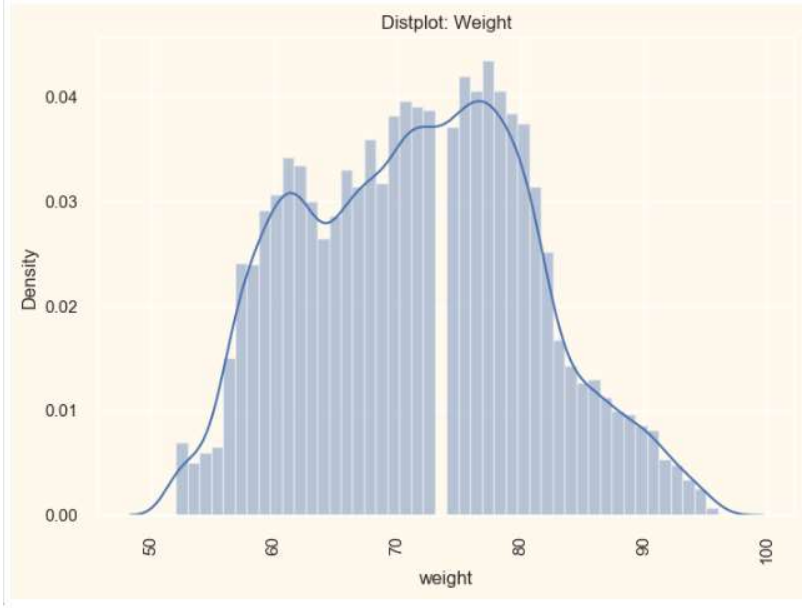

Figure 20 Distplot: weight

- The density plot of weight shows that a particular age range is completely missing between 70-80 age which needs to be checked.
- Most of the individuals range between 60-80 weight range

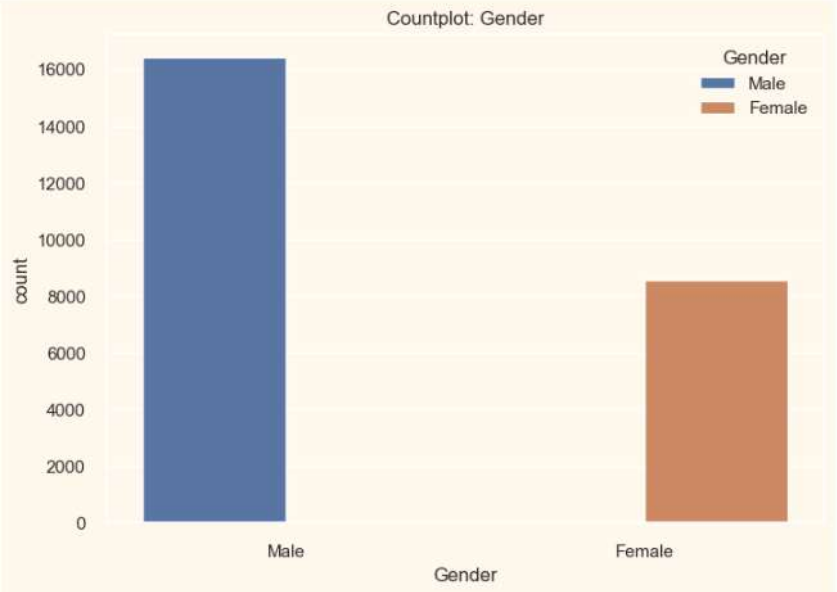

Figure 21 Countplot: Gender

- 65.6% of individuals are male and 34.4% are females.

```
Gender (%age share)
Male      0.65688
Female    0.34312
```

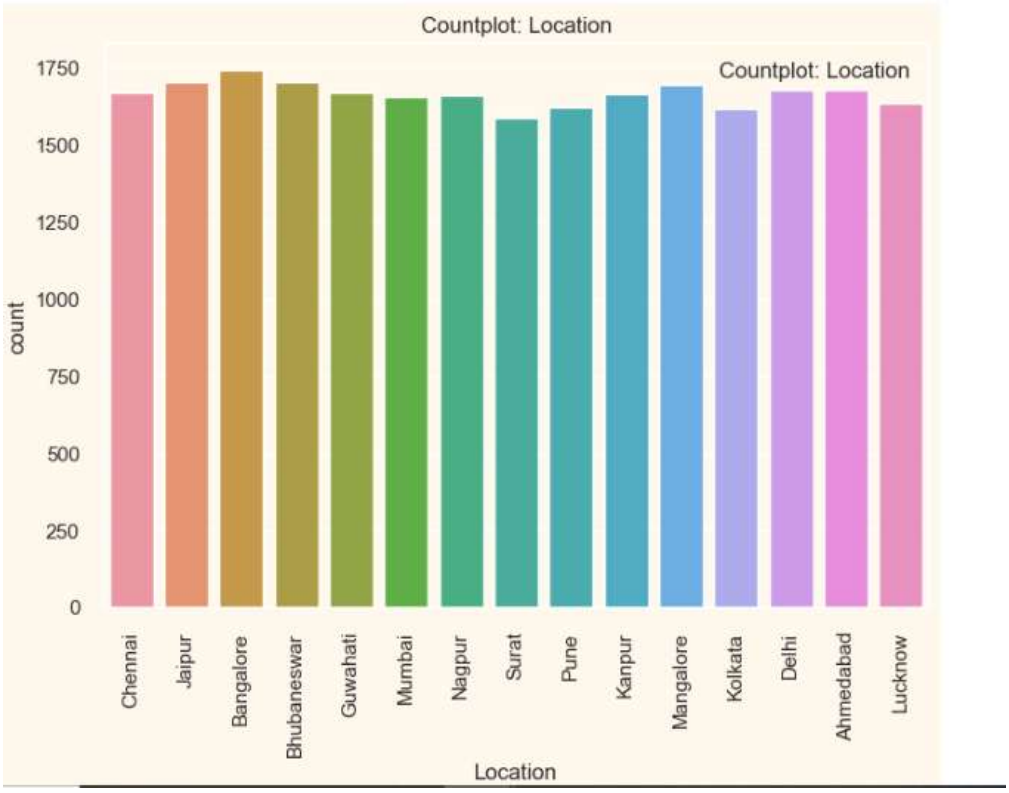

Figure 22 Countpot: Location

Location	(%age share)
Bangalore	0.06968
Jaipur	0.06824
Bhubaneswar	0.06816
Mangalore	0.06788
Delhi	0.06720
Ahmedabad	0.06708
Guwahati	0.06688
Chennai	0.06676
Kanpur	0.06656
Nagpur	0.06652
Mumbai	0.06632
Lucknow	0.06548
Pune	0.06488
Kolkata	0.06480
Surat	0.06356

- Bangalore contributed to maximum number of individuals. Surat has the lowest share of individuals.

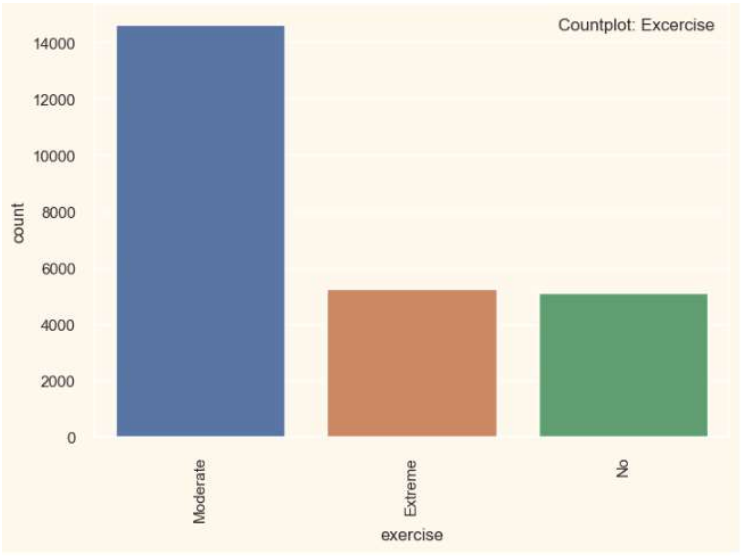

Figure 23 Countplot: excercise

- We see that most of the individuals are not fitness conscious and have morderate level of fitness activity.

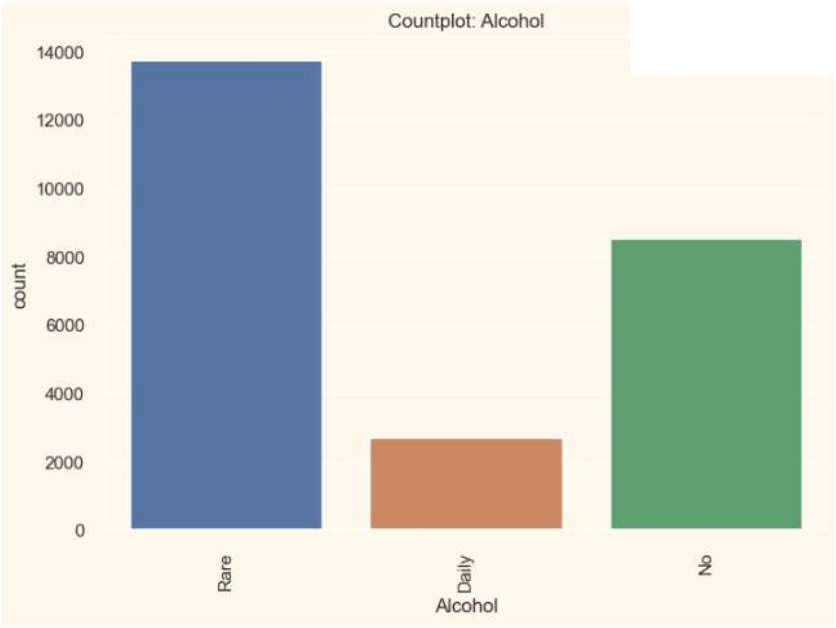

Figure 24 Countplot: alcohol

- Most of the individuals consume alcohol rarely or not at all.

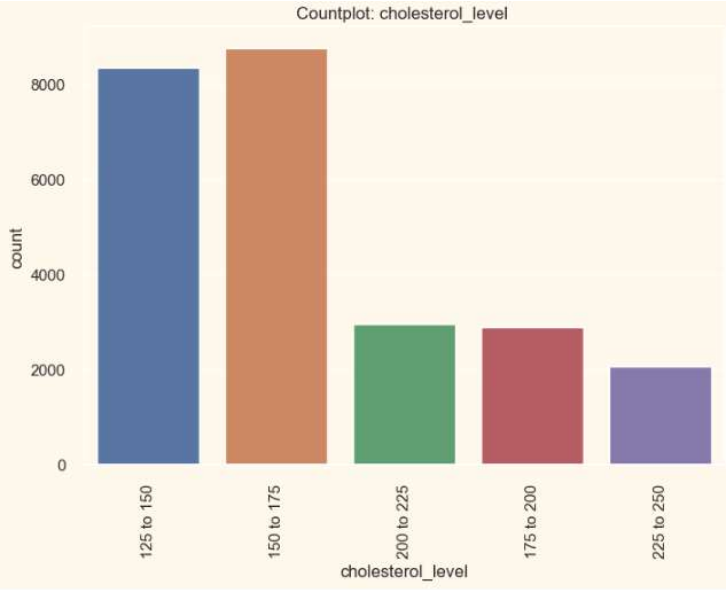

Figure 25 Countplot: cholestrol_level

- People with 125 to 175 of cholesterol level are very high.

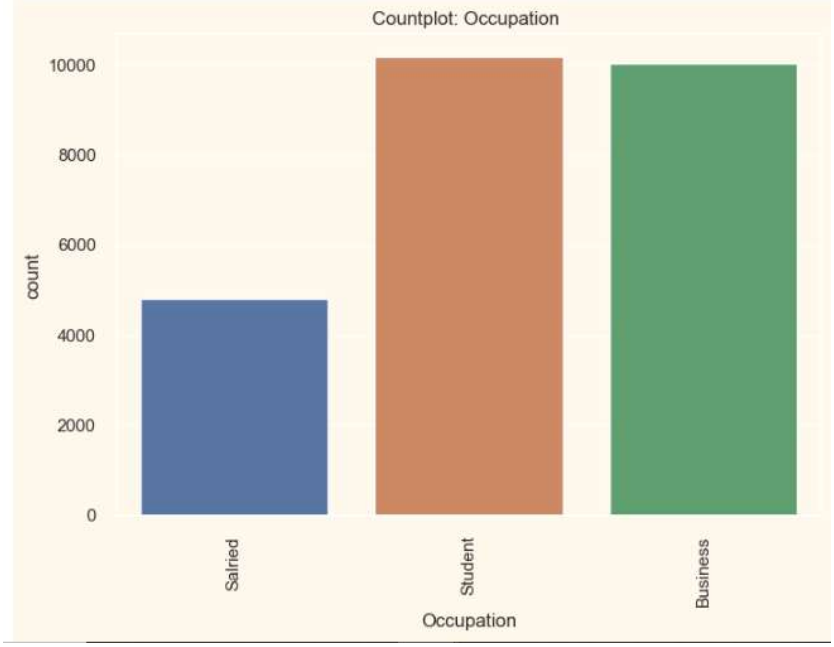

Figure 26 Countplot: Occupation

- Most of the individuals are either student or Business occupation. Share of salaried people is low.

Bivariate analysis (relationship between different variables , correlations):

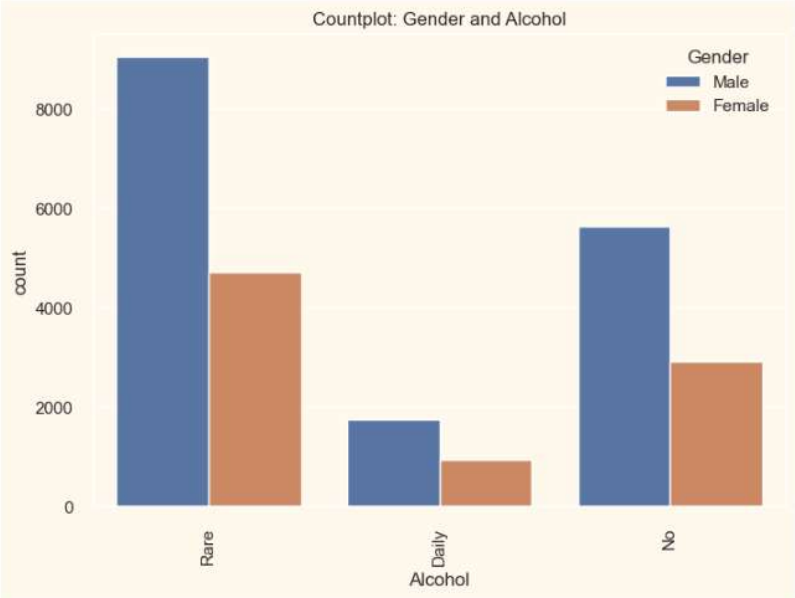

Figure 27 Countplot: Gender & alcohol

- If we see the Gender and Alcohol plot, we see that female are usually half in alcohol consumption. We can see both genders who are dailiy alcohol consumers as well.

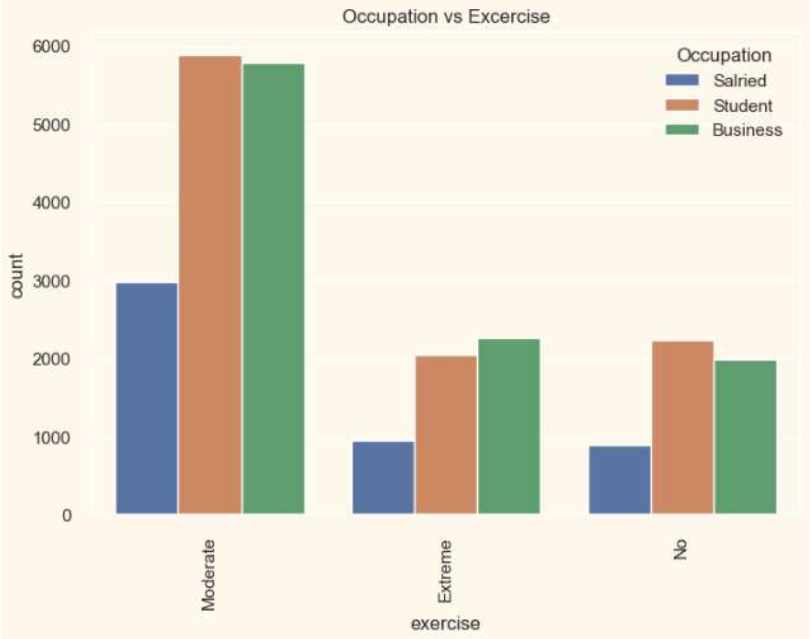

Figure 28 Countplot: Occupation vs Exercise

- Occupation and Execrcise plot has been created to see how much importance is given on fitness levels. We can see that Salaried are usually less in involved in fitness regime compared to students and business persons.

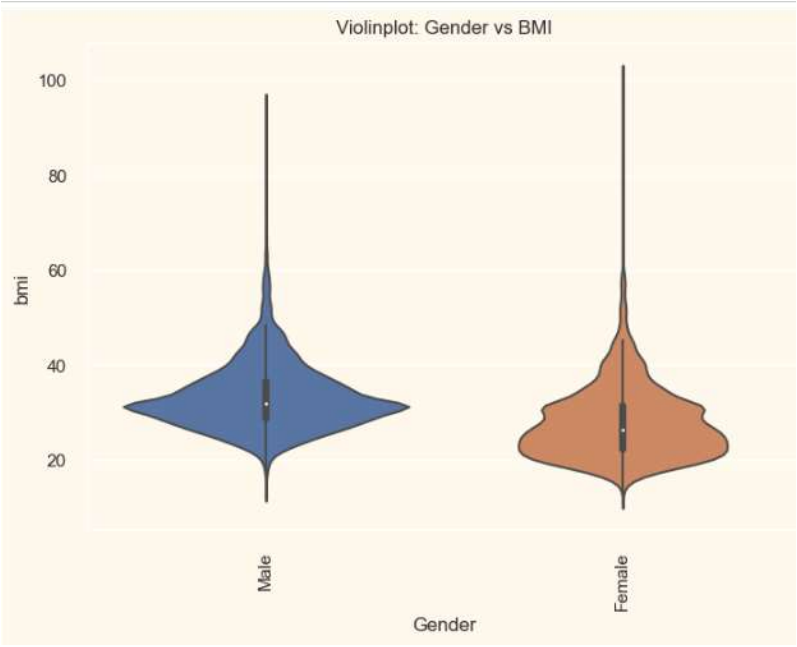

Figure 29 Violinplot: Gender vs bmi

- Gender vs Bmi shows comparision between bmi levels of genders. As per Fig 29 we can se that most women have lower bmi of 20 compared to men who have bmi of 30.

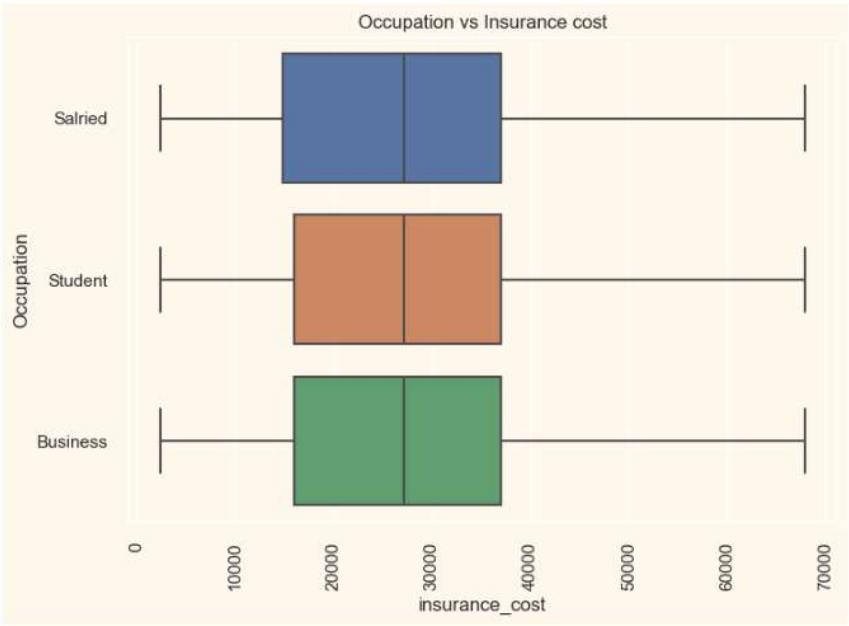

Figure 30 Boxplot: Occupation vs insurance cost

- If we compare the Insurance cost by occupation we can see no major distinction between the three occupations.

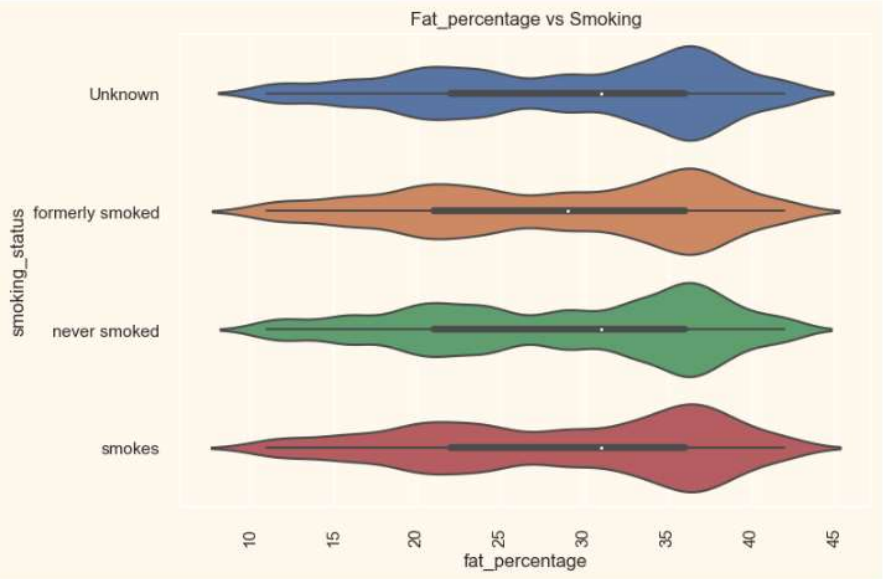

Figure 31 Violinplot: Fat percentage vs insurance cost

- How does smoking effect the fat percentage? Based on Fig 31 we can see no major difference in fat percentage among smokers and non-smokers. However, people who have smoked formerly have a lower mean avg fat percentage suggesting their fitness choices. This is a important insight as fitness conscious people are likely to pay more cost while claiming little amount.

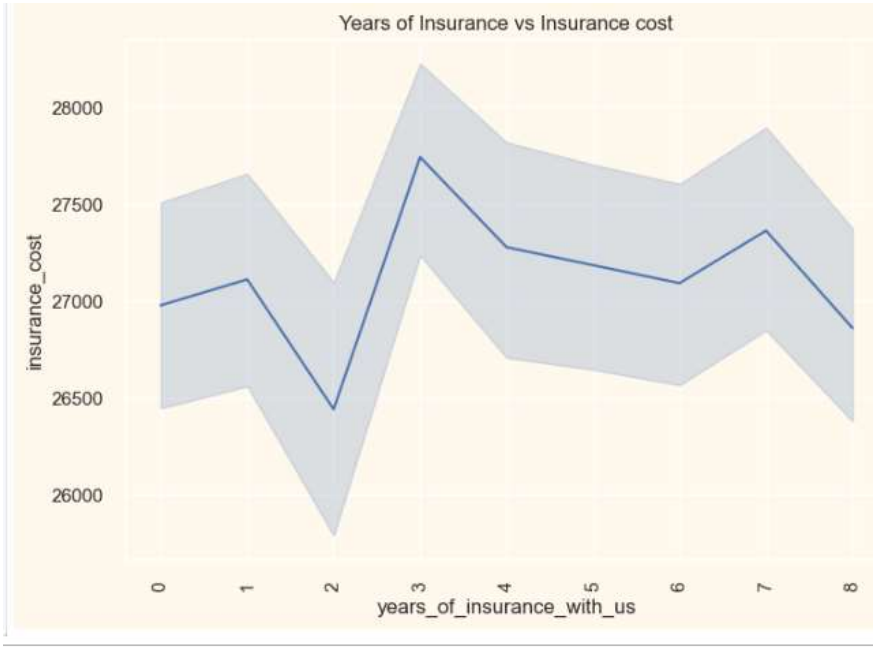

Figure 32 Lineplot: Insurance cost vs years of insurance with us

- Most of the individuals who are with us for 3 years pay highest insurance cost. The people who are with us for 2 years have lowest insurance cost.

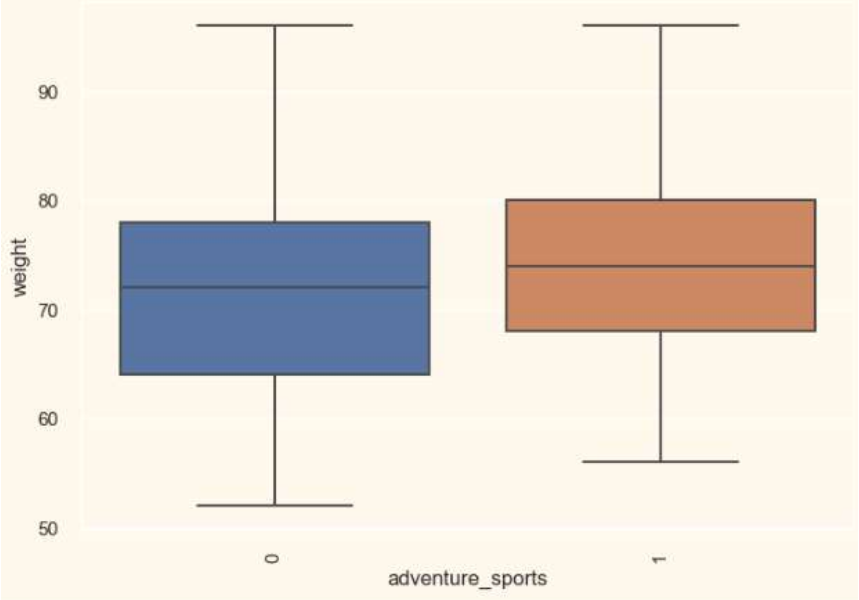

Figure 33 Weight vs adventure sports

- Individuals who are involved in adventure sports have slighltly higher avg weight.

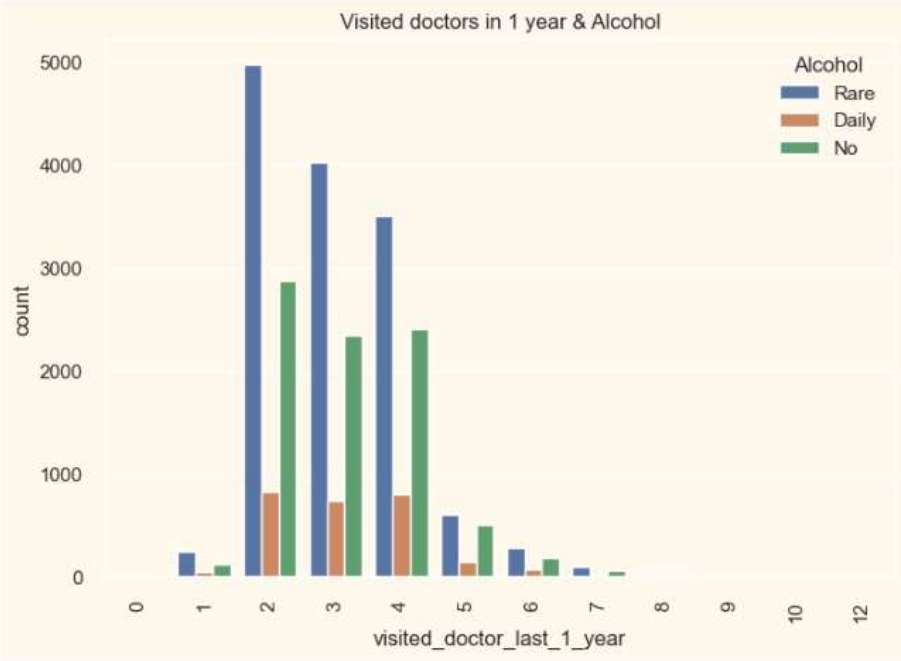

Figure 34 Visited doctor in 1 year vs Alcohol

- How does alcohol effect doctor visits in last 1 year? We see that people who consume alcohol usually have 2 to 4 visits to doctors. People who rarely consume alcohol have to visit 2 times atleast to doctors.

Figure 35 Visited doctor in 1 year vs Age

- Which age group visited doctor most in last 1 year? Individuals in age group of 50-55 visited docter 10 times which is highest among all age groups.

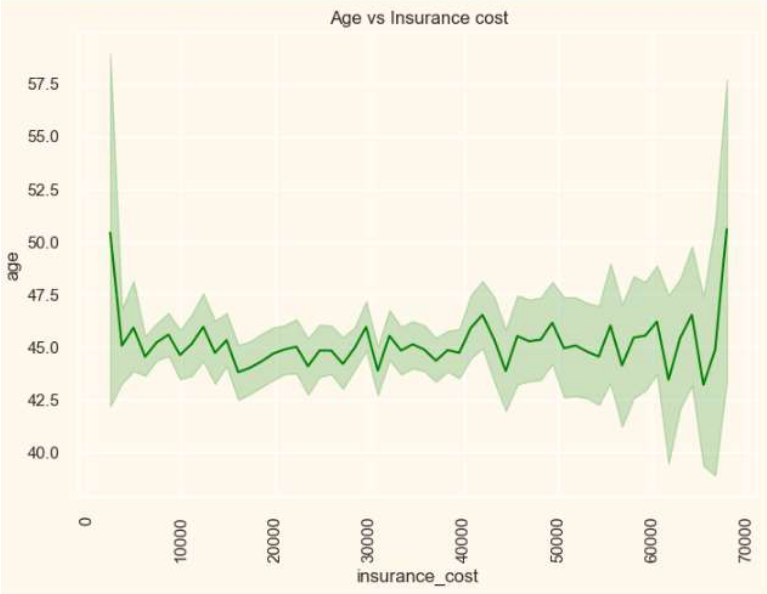

Figure 36 Age vs insurance cost

- How does age effect the insurance cost? : People in age group of 50 to 53 either pay minimum insurance cost lest than 10k or upto 70k. This might be due to financial security and income level. The age group is either affluent or depends on pension to bear insurance costs.

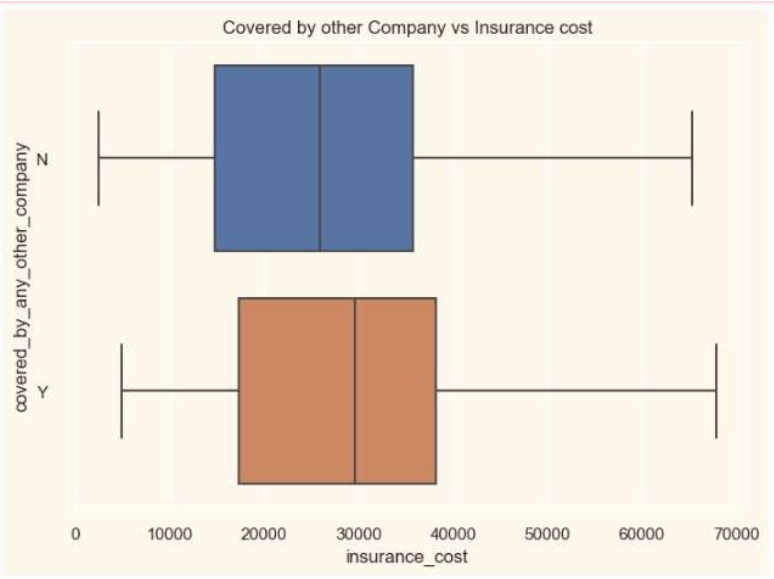

Figure 37 Covered by any other company vs Insurance cost

- Individuals who are covered by other company as well usually pay more insurance cost.

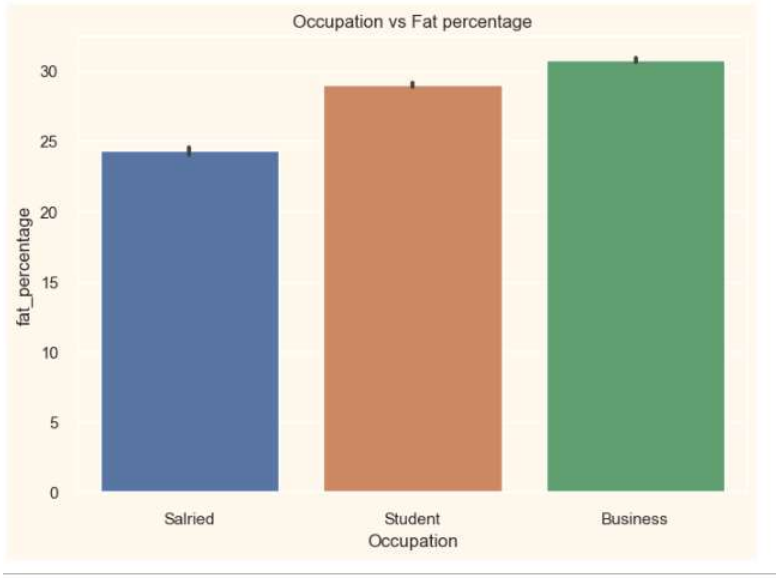

Figure 38 Occupation vs fat percentage

- Business class people have higher fat percentage and Salaried people have lowest fat percentage.

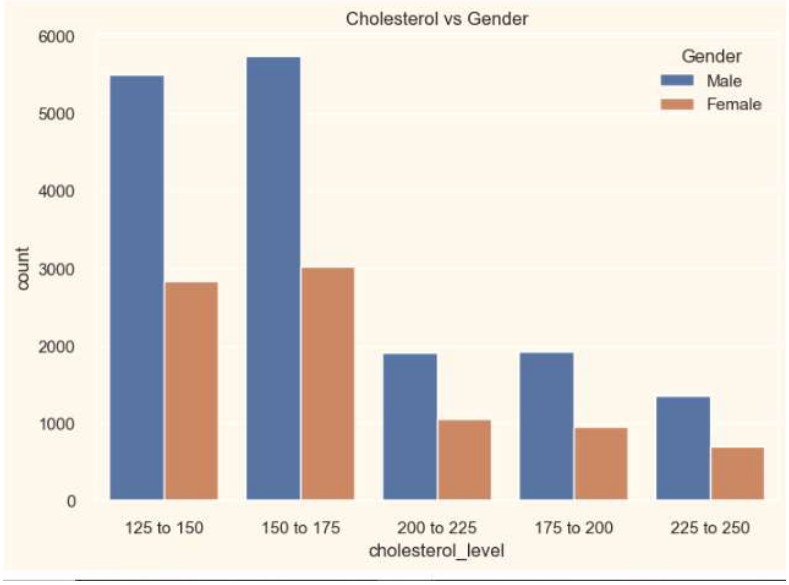

Figure 39 Cholestrol vs Gender

- Most of the men and women have cholesterol level between 150 to 175.

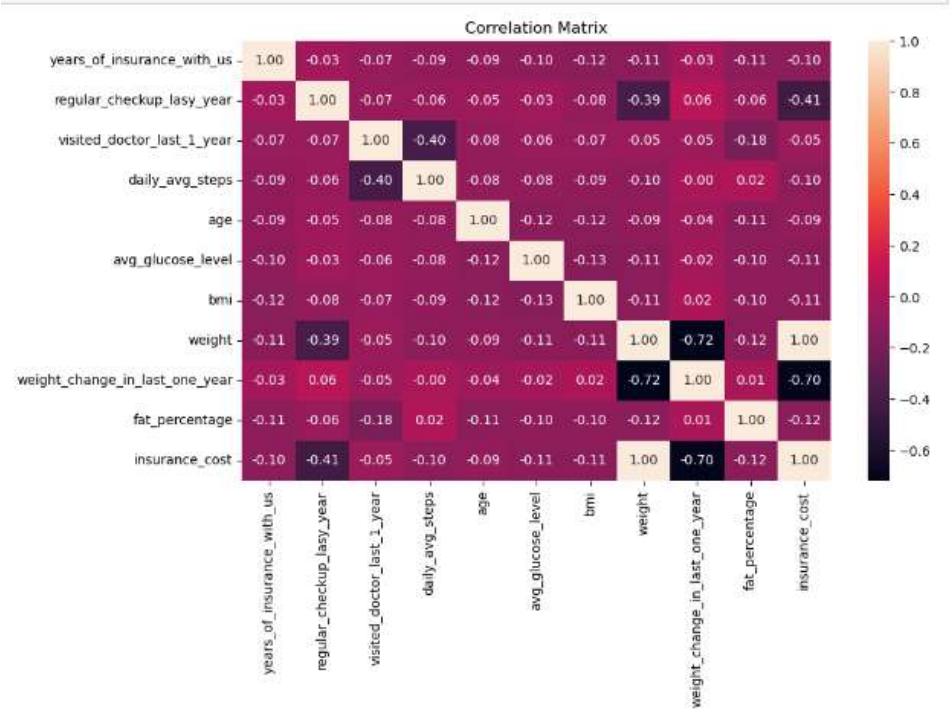

Figure 40 Correlation Matrix

- We have removed variables which are categorical and have high cardinality eg. applicant_id, adventure_sports, heart_decs_history, other_major_decs_history
- Weight & insurance cost have high positive correlation but weight change in last one year has high negative correlation

- We can see positive correlation between bmi and other major decease history. This suggest that people with major decease history have higher bmi.
- We see a high negative correlation between weight change in last one year vs insurance cost. This suggests that increase in weigh change is leading to lower insurance cost.

Business Implication of EDA

3. The analysis suggests that individuals who have been insured for 3 years tend to pay the highest insurance cost, while those insured for 2 years have the lowest insurance cost. Additionally, individuals in the 50–53 age group have varied insurance costs depending on their financial situation (either paying less than 10k or up to 70k).
4. People involved in adventure sports have slightly higher average weight, and fitness-conscious individuals (who are non-smokers and have a lower fat percentage) tend to pay higher insurance premiums while making fewer claims.
5. The BMI comparison (Figure 29) indicates that most women have a lower BMI (~20) compared to men, who tend to have higher BMI (~30). This could point to different health risk profiles between genders.
6. The cholesterol levels (Figure 25) show that a significant portion of the population falls within the 125 to 175 cholesterol range. High cholesterol is a key risk factor for several chronic conditions.
7. The occupation analysis (Figure 26) shows that individuals in business and student roles tend to have higher engagement with fitness, whereas salaried employees are less likely to prioritize fitness.
8. The positive correlation between weight and insurance cost (Figure 40) suggests that individuals with higher weight tend to pay more for insurance. Conversely, weight change in the last year shows a negative correlation with insurance costs, indicating that people who have lost weight may pay lower premiums.

3.Data Cleaning & Preprocessing

Approach used for identifying and treating missing values and outlier treatment (and why)

Missing Values Check:

```
applicant_id                        0
years_of_insurance_with_us          0
regular_checkup_lasy_year           0
adventure_sports                    0
Occupation                          0
visited_doctor_last_1_year          0
cholesterol_level                   0
daily_avg_steps                     0
age                                 0
heart_decs_history                  0
other_major_decs_history            0
Gender                              0
avg_glucose_level                   0
bmi                                 0
smoking_status                      0
Year_last_admitted              11881
Location                            0
weight                              0
covered_by_any_other_company        0
Alcohol                             0
exercise                            0
weight_change_in_last_one_year      0
fat_percentage                      0
insurance_cost                      0
dtype: int64
```

Figure 41 Missing values check

- Missing Values in Year_last_admitted is: 47.52% The missing values is very high and we will drop the column as treating it will lead to bias in model.
- Applicant id is also not redundant variable so we need to drop it.

After dropping unnecessary columns:

```
years_of_insurance_with_us          0
regular_checkup_lasy_year           0
adventure_sports                    0
Occupation                          0
visited_doctor_last_1_year          0
cholesterol_level                   0
daily_avg_steps                     0
age                                 0
heart_decs_history                  0
other_major_decs_history            0
Gender                              0
avg_glucose_level                   0
bmi                               990
smoking_status                      0
Location                            0
weight                              0
covered_by_any_other_company        0
Alcohol                             0
exercise                            0
weight_change_in_last_one_year      0
fat_percentage                      0
insurance_cost                      0
dtype: int64
```

Figure 42 Dropping columns

We have dropped the irrelevant and high missing values column. However we still need to treat the missing values in bmi variable. We will treat them in later section.

Missing Value treatment (if applicable):

- The missing values in in Year_last_admitted was 47.52%. Since it was very high we decided to drop entire column.
- BMI also has missing values and we will use KNN imputer to impute the missing values.
- We have used KNN Imputer it uses the k-nearest neighbors algorithm to fill in missing values based on the values of the nearest neighbors. Unlike simple imputation which imputes mean value for all missing values.
- After imputing the value for BMI mean is same 31.39 and null values are zero.

```
years_of_insurance_with_us          0
regular_checkup_lasy_year           0
adventure_sports                    0
Occupation                          0
visited_doctor_last_1_year          0
cholesterol_level                   0
daily_avg_steps                     0
age                                 0
heart_decs_history                  0
other_major_decs_history            0
Gender                              0
avg_glucose_level                   0
bmi                                 0
smoking_status                      0
Location                            0
weight                              0
covered_by_any_other_company        0
Alcohol                             0
exercise                            0
weight_change_in_last_one_year      0
fat_percentage                      0
insurance_cost                      0
dtype: int64
```

Figure 43 Missing value imputation

Outlier treatment (if required)

Regression Models are sensitive to outliers. We will treat the outliers using IQR method. We will not do any treatment on Target variable. We have checked in Univariate analysis the presence of outliers and will treat them. After treating outliers we can see that there is no outliers in Fig. 45.

Boxplot:

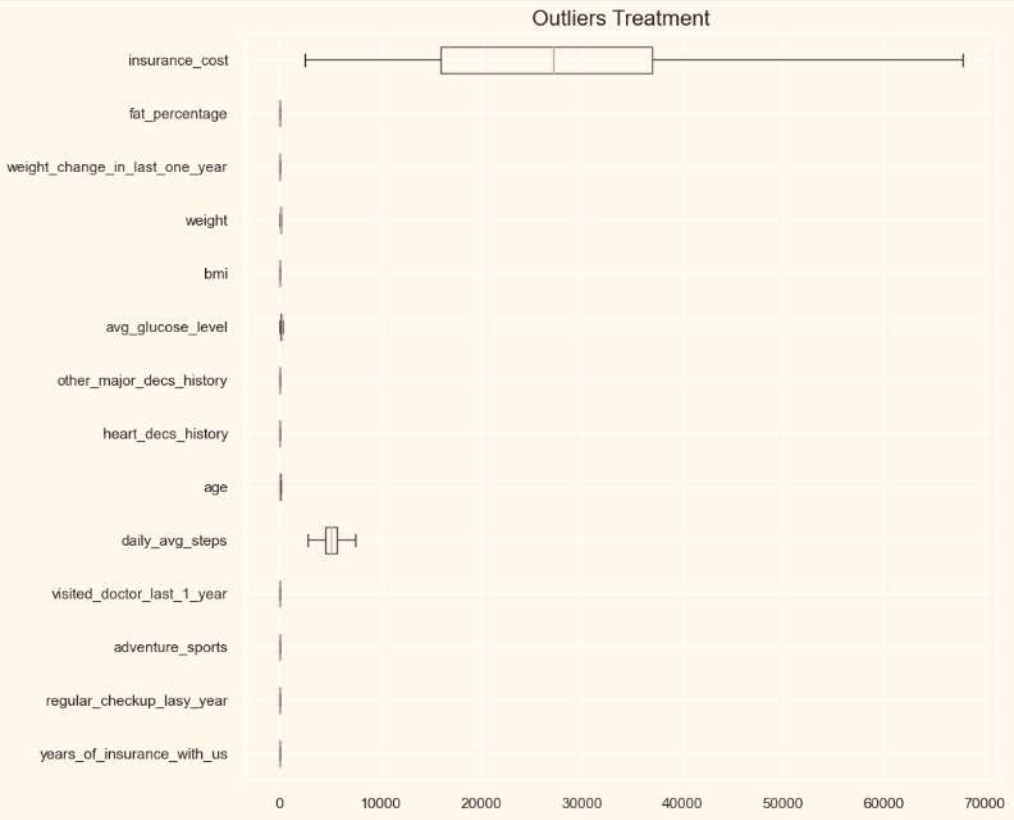

Figure 44 Outliers treatment

Need for variable transformation (if any)

1. We will create categorical and numeric data set.
2. We will do One-hot encoding on categorical data using One Hot Encoder to convert them into numeric field.
3. We will do scaling of numerical data using Standard Scaler method.
4. We will concat the data set of encoded and scaled variables to create final dataset with transformed variables.

Numeric Data:

	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	other_ma
0	3.0	1.0	0.0	2.0	4866.0	28.0	0.0	
1	0.0	0.0	0.0	4.0	6411.0	50.0	0.0	
2	1.0	0.0	0.0	4.0	4509.0	68.0	0.0	
3	7.0	2.5	0.0	2.0	6214.0	51.0	0.0	
4	3.0	1.0	0.0	2.0	4938.0	44.0	0.0	
...	...	...	...	...	...	...	...	
24995	3.0	0.0	0.0	4.0	5614.0	22.0	0.0	
24996	6.0	0.0	0.0	4.0	4719.0	58.0	0.0	
24997	7.0	0.0	0.0	2.0	5624.0	34.0	0.0	
24998	1.0	0.0	0.0	2.0	7510.5	27.0	0.0	
24999	8.0	2.0	0.0	4.0	5882.0	22.0	0.0	

25000 rows × 14 columns

Figure 45 Numerical data

Categorical Data:

	Occupation	cholesterol_level	Gender	smoking_status	Location	covered_by_any_other_company	Alcohol	exercise
0	Salried	125 to 150	Male	Unknown	Chennai	N	Rare	Moderate
1	Student	150 to 175	Male	formerly smoked	Jaipur	N	Rare	Moderate
2	Business	200 to 225	Female	formerly smoked	Jaipur	N	Daily	Extreme
3	Business	175 to 200	Female	Unknown	Chennai	Y	Rare	No
4	Student	150 to 175	Male	never smoked	Bangalore	N	No	Extreme
...	...	...	...	...	...	...	...	...
24995	Salried	225 to 250	Male	smokes	Kanpur	Y	Rare	Moderate
24996	Business	200 to 225	Male	never smoked	Kanpur	N	Rare	Moderate
24997	Student	150 to 175	Male	Unknown	Bhubaneswar	N	Rare	Moderate
24998	Salried	225 to 250	Male	Unknown	Surat	N	Rare	Moderate
24999	Business	150 to 175	Male	formerly smoked	Chennai	N	No	No

25000 rows × 8 columns

Figure 46 Categorical Data

One Hot Encoding:

	Occupation_Salried	Occupation_Student	cholesterol_level_150 to 175	cholesterol_level_175 to 200	cholesterol_level_200 to 225	cholesterol_level_225 to 250	Gender_Male	smoking_
0	1.0	0.0	0.0	0.0	0.0	0.0	1.0	
1	0.0	1.0	1.0	0.0	0.0	0.0	1.0	
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0	
3	0.0	0.0	0.0	1.0	0.0	0.0	0.0	
4	0.0	1.0	1.0	0.0	0.0	0.0	1.0	
...	...	...	...	...	...	...	...	
24995	1.0	0.0	0.0	0.0	0.0	1.0	1.0	
24996	0.0	0.0	0.0	0.0	1.0	0.0	1.0	
24997	0.0	1.0	1.0	0.0	0.0	0.0	1.0	
24998	1.0	0.0	0.0	0.0	0.0	1.0	1.0	
24999	0.0	0.0	1.0	0.0	0.0	0.0	1.0	

25000 rows × 29 columns

Figure 47 Encoded data

Scaling:

	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	othe
0	-0.417807	0.374779	0.0	-0.980772	-0.333160	-1.050360	0.0	
1	-1.568750	-0.714377	0.0	0.803748	1.260326	0.315492	0.0	
2	-1.185102	-0.714377	0.0	0.803748	-0.701364	1.433007	0.0	
3	1.116783	2.008512	0.0	-0.980772	1.057144	0.377576	0.0	
4	-0.417807	0.374779	0.0	-0.980772	-0.258901	-0.057013	0.0	
...	...	...	...	...	...	...	...	
24995	-0.417807	-0.714377	0.0	0.803748	0.438314	-1.422864	0.0	
24996	0.733135	-0.714377	0.0	0.803748	-0.484773	0.812165	0.0	
24997	1.116783	-0.714377	0.0	-0.980772	0.448628	-0.677855	0.0	
24998	-1.185102	-0.714377	0.0	-0.980772	2.394332	-1.112444	0.0	
24999	1.500430	1.463934	0.0	0.803748	0.714725	-1.422864	0.0	

25000 rows × 14 columns

Figure 48 Scaled data

Final Data:

	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	othe
0	-0.417807	0.374779	0.0	-0.980772	-0.333160	-1.050360	0.0	
1	-1.568750	-0.714377	0.0	0.803748	1.260326	0.315492	0.0	
2	-1.185102	-0.714377	0.0	0.803748	-0.701364	1.433007	0.0	
3	1.116783	2.008512	0.0	-0.980772	1.057144	0.377576	0.0	
4	-0.417807	0.374779	0.0	-0.980772	-0.258901	-0.057013	0.0	
...	...	...	...	...	...	...	...	
24995	-0.417807	-0.714377	0.0	0.803748	0.438314	-1.422864	0.0	
24996	0.733135	-0.714377	0.0	0.803748	-0.484773	0.812165	0.0	
24997	1.116783	-0.714377	0.0	-0.980772	0.448628	-0.677855	0.0	
24998	-1.185102	-0.714377	0.0	-0.980772	2.394332	-1.112444	0.0	
24999	1.500430	1.463934	0.0	0.803748	0.714725	-1.422864	0.0	

25000 rows × 43 columns

Figure 49 Concat data final

Variables removed or added and why (if any)

We have 25 features and 45 columns after encoding. Adding new variables can lead to the curse of dimensionality. As the number of variables increases, the volume of the space increases exponentially, which can make data sparsity a significant problem during model building.

- Year_last_admitted: Dropped due to high missing values (47.52%).
- Applicant_id: Dropped as it is a unique identifier without predictive value.
- High Cardinality Features: Any columns with high cardinality and no useful predictive information dropped to avoid the curse of dimensionality and reduce model complexity.

4. Business insights from EDA

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

- In classification problems we can check the unbalance in data by calculating the class variables of target column. If any class variable (0 or 1) is less than 10% we consider it as imbalanced data.
- In regression problems we do not have hard boundaries between the class values. They are continuous and we can check the distribution of dataset, visualize boxplot and check data description to check the target variable.

Lets check the distribution of Target Variable after scaling:

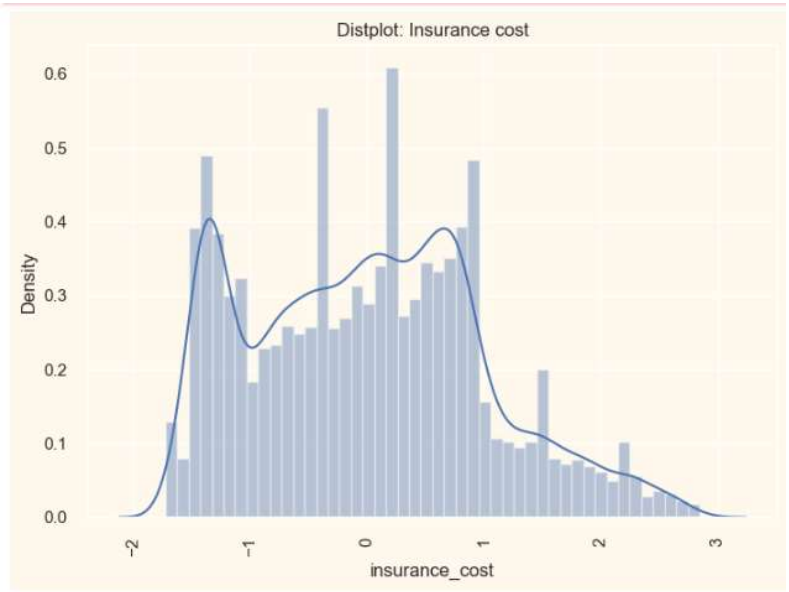

Figure 50 Distplot: Insurance cost

- We see that the data is imbalanced. The data has moderate skewness of 0.33. The data set is unbalanced because data for values from 2 to 3 are very less compared to data for -1 to 1 range.

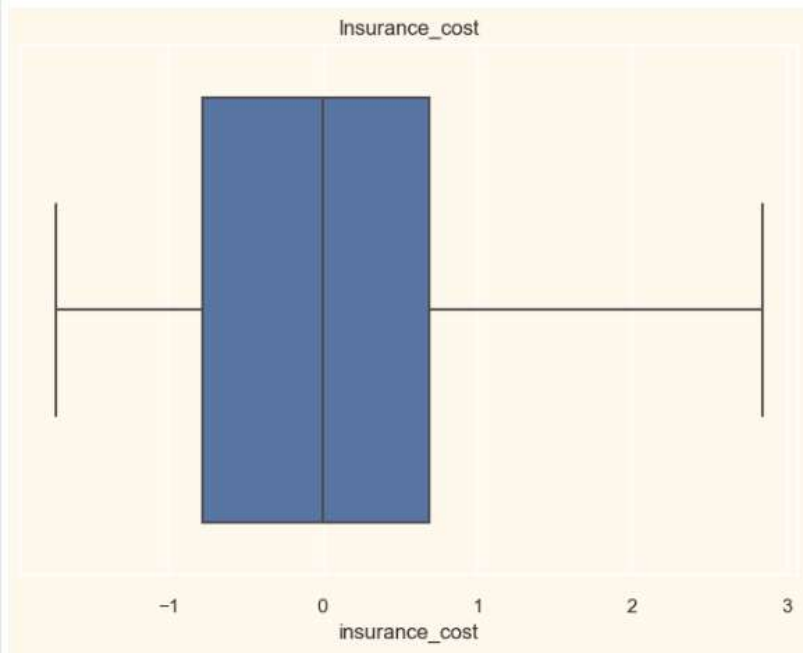

Figure 51 Boxplot imbalanced dataset

- The Boxplot shows clustered values for -1 to 1 range. But very less data available for 1 to 3 range. This shows that that data is imbalanced.

How to deal with imbalanced dataset?

1. **Oversampling:** We can increase instances for data for values which are underrepresented. Thus data for all values will become similar.
2. **Undersampling:** We can minimize the instances for values which are over represented. However, this will reduce the dataset size available if the data set is already small.
3. **SMOTE:** We can use Synthetic Minority Over-sampling Technique for regression to generate synthetic samples for minority class values. This algorithm calculates k-nearest neighbours for minority values to generate synthetic values. SMOTE can be imported from imblearn library.

Business Context:

If we use an imbalanced data set, there will be a higher weightage towards predicting of majority values. This will lead to bias. The predictive power for minority values will be less or not reliable. Therefore, we need to ensure that we have sufficient data for all values by oversampling or undersampling so that there is no bias and high predictability of the model.

Any business insights using clustering (if applicable)

Hierarical Clustering (agglomerative):

Based on Heirarical Clustering we can see the dendogram showing the 12 clusters.

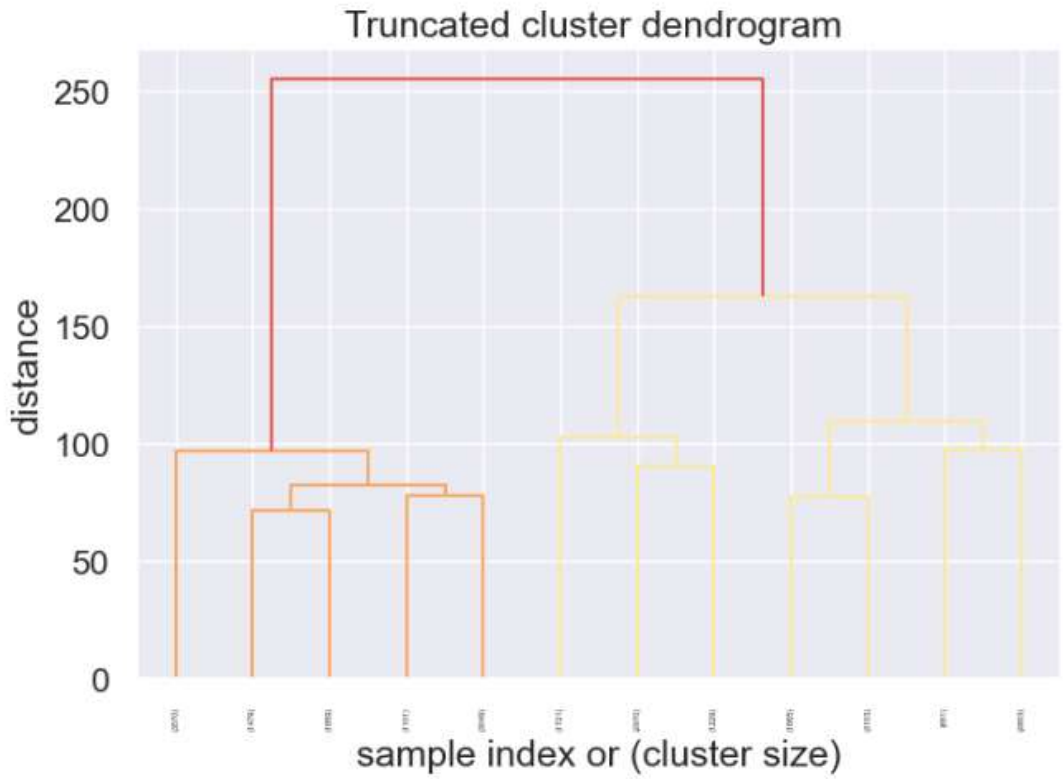

Figure 52 Hierarical Clustering

If we take cutoff line between distance 100-150. We will have atleast 3 clusters.

Lets see the proportion of clusters:

```
HCluster
1    0.43736
0    0.32012
2    0.24252
```

- Hierarical clustering has classified the data into 3 segments where cluster 1 includes 43.73%, cluster 0 with 32.01% and cluster 2 with 24.25% individuals.

K-means Clustering:

We will now use K-means Clustering and try to find the optimal number of clusters using elbow method.

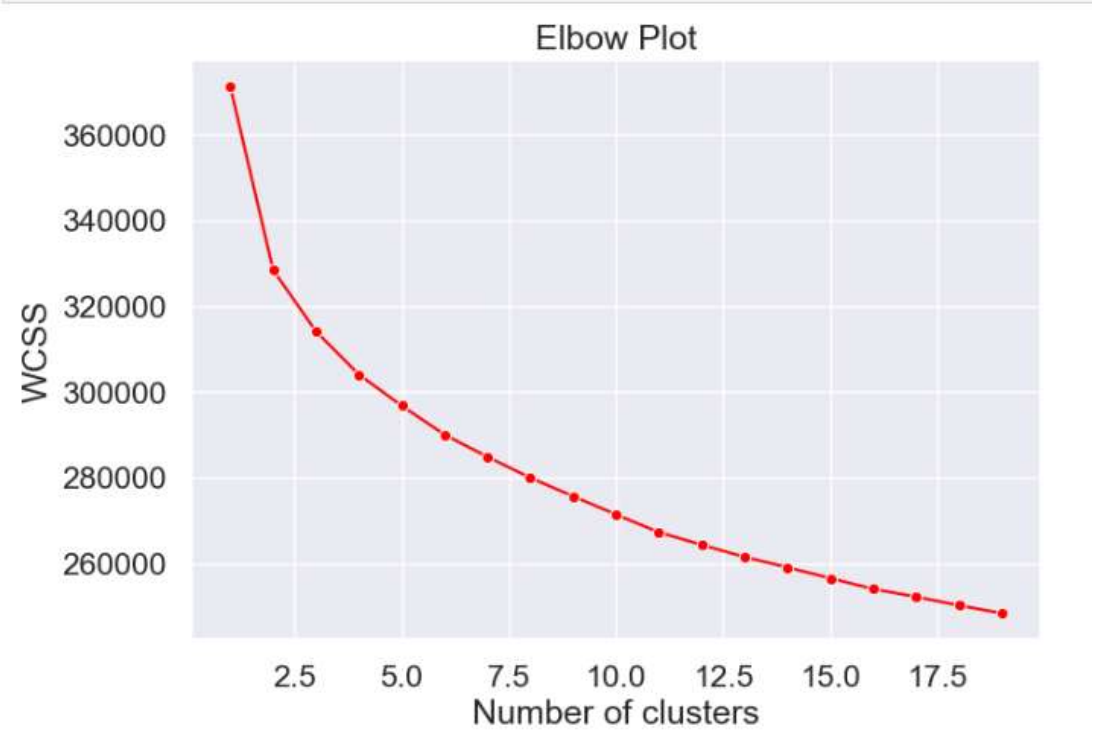

Figure 53 Elbow plot

We can see k=2 has maximum drop. However it is difficult see see any abrupt change in WCSS scores after that for n=19 clusters.

We will check Silhoutte scores to find optimal number of clusters:

```
The Average Silhouette Score for 2 clusters is 0.10725
The Average Silhouette Score for 3 clusters is 0.09203
The Average Silhouette Score for 4 clusters is 0.06888
The Average Silhouette Score for 5 clusters is 0.0564
The Average Silhouette Score for 6 clusters is 0.05596
The Average Silhouette Score for 7 clusters is 0.05378
The Average Silhouette Score for 8 clusters is 0.05307
The Average Silhouette Score for 9 clusters is 0.0541
The Average Silhouette Score for 10 clusters is 0.05131
The Average Silhouette Score for 11 clusters is 0.05085
The Average Silhouette Score for 12 clusters is 0.04706
The Average Silhouette Score for 13 clusters is 0.04637
The Average Silhouette Score for 14 clusters is 0.04662
The Average Silhouette Score for 15 clusters is 0.04446
The Average Silhouette Score for 16 clusters is 0.04633
The Average Silhouette Score for 17 clusters is 0.04685
```

We can see a major drop after k=3 from 0.09 to 0.06. Therefore we will consider k=2 and k=3 for cluster analysis using K-Means method.

Lets see the Silhoette Plot

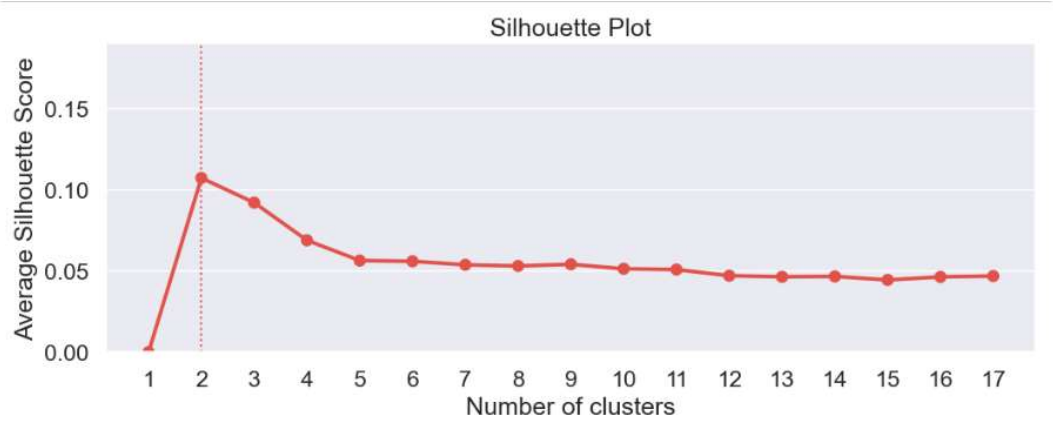

Figure 54 Silhoette Plot

We can see that based on K means clustering the optimal value of K is 2. Lets do clustering for k=2

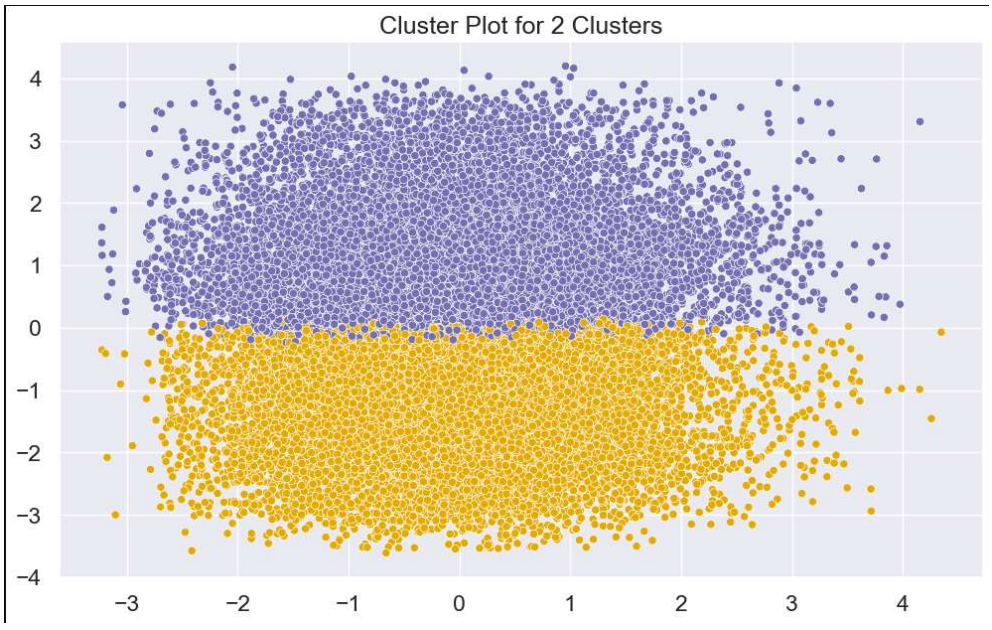

Figure 55 K-means cluster, k=2

Here we can see that if we make 2 clusters there is a clear distinction between two clusters.

Based on Agglomerative Clustering we have 3 clusters. If we take k=3 and build 3 clusters using K Means clustering. We get below plot:

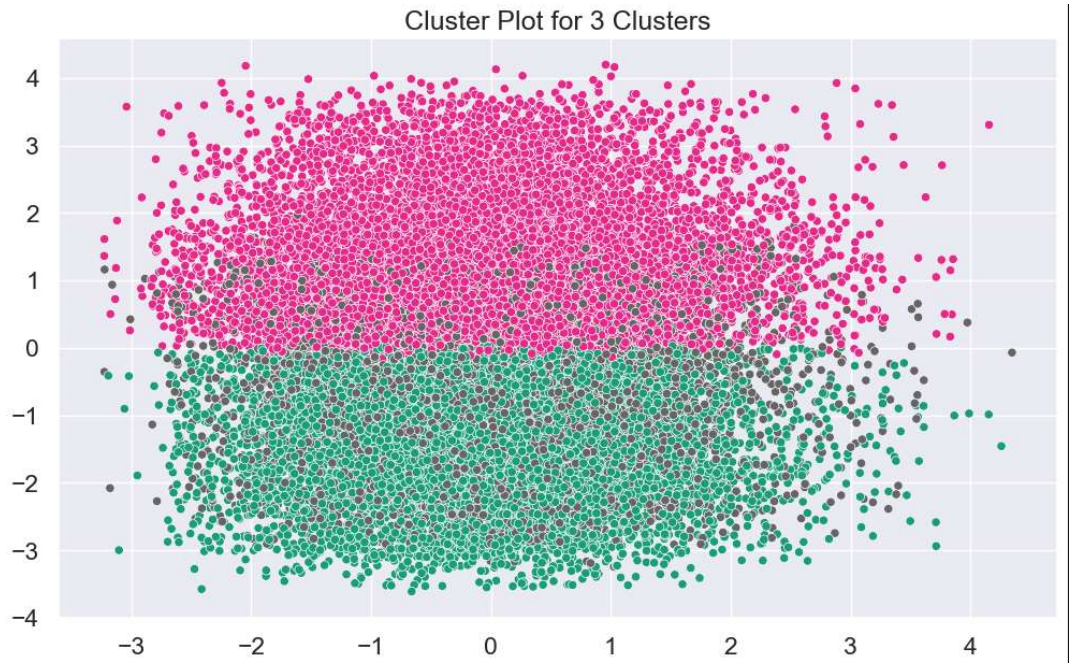

Figure 56 Kmeans Cluster, K=3

```
KCluster
0    0.44832
1    0.31364
2    0.23804
```

Now we have 3 clusters and we will see how it is related to different variables.

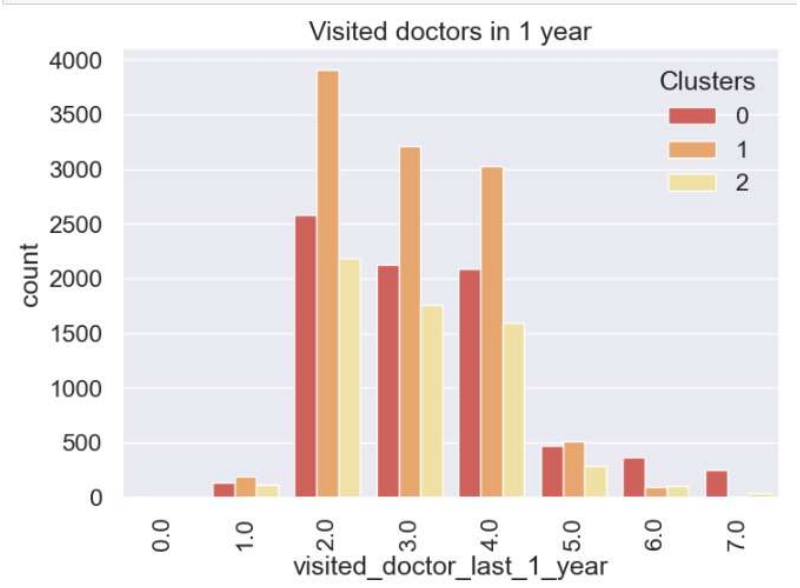

Figure 57 Clusters: Visited doctors in last 1 year

- We cannot see any major effect of clustering on the variable visited doctors in last 1 year. However we can see that cluster 0 individuals are highest among vistor who went 6 times or more.
- Cluster 0 basically represents old age people or peple who are being diagnosed with some disease.
- Cluster 1 people seems to be working population of middle age who has visited atleast 2 to 4 times to doctor in last 1 year.

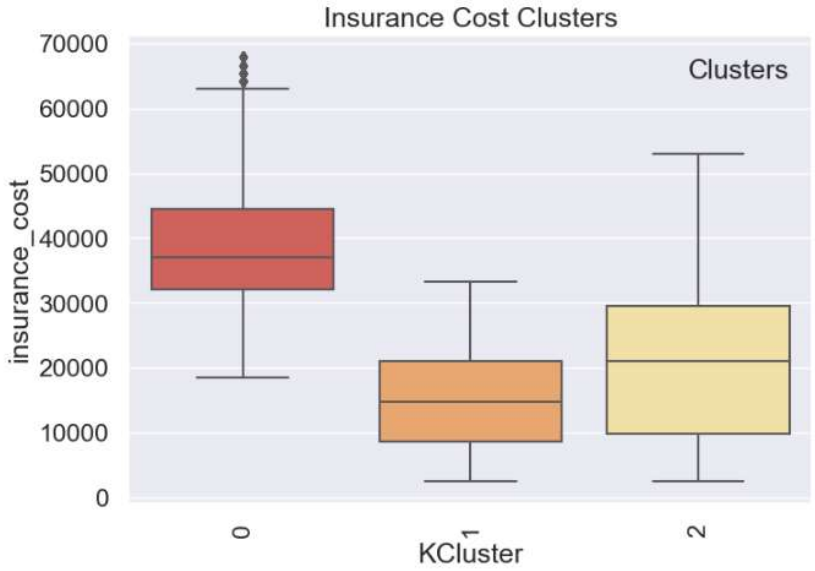

Figure 58 Insurance cost: clusters

- Clusters have a clear variation based on insurance cost.
- Cluster 1 people are among the lowest paying and cluster 0 is among the highest paying for insurance.

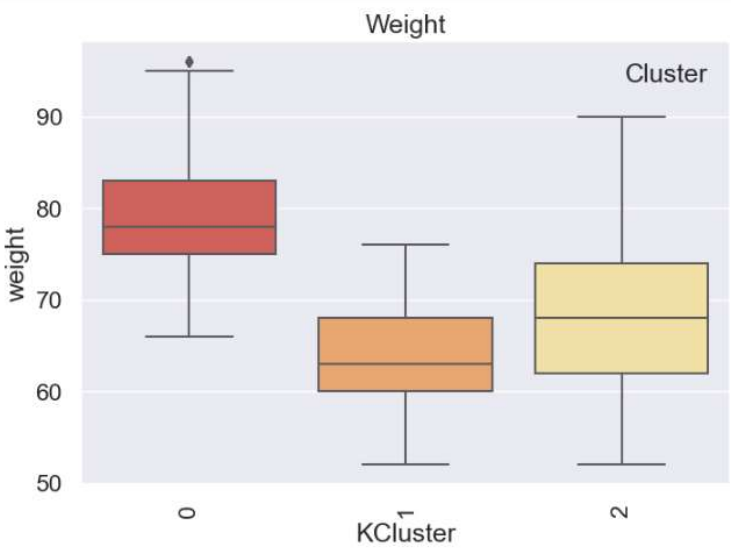

Figure 59 Weight clusters

- Cluster 0 is among the highest weight range group and cluster 0 is among the lowest weight range group. Cluster 0 individuals have high weight and seems to be of higher age and paying highest insurance cost. Cluster 0 is the most important segment among all.

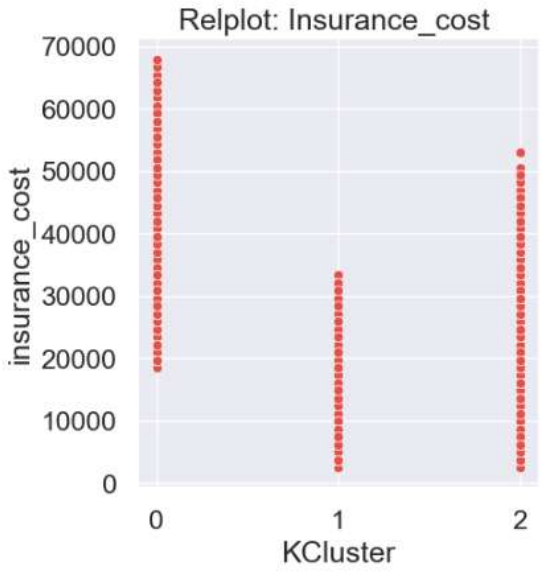

Figure 60 Relplot: Insurance Cost clusters

- Cluster 0 has highest insurance cost of 70k, cluster 1 has highest insurance cost of 35k and cluster 2 has highest cost of 55k. This suggests that the

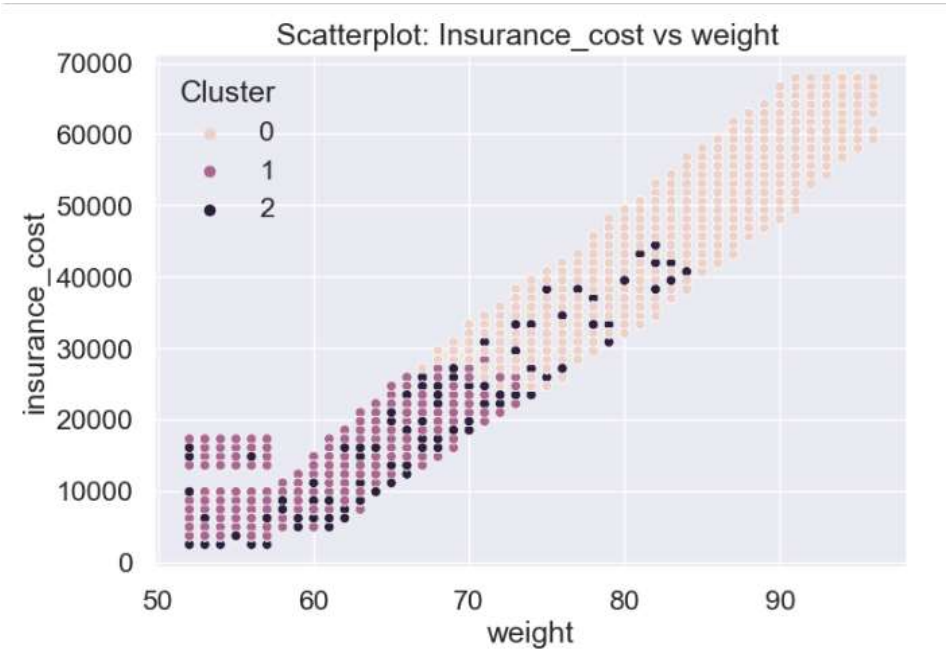

Figure 61 Scatterplot: Insurance vs weight for clusters

- Insurance cost and weight has a strong and almost linear positive correlation.
- Cluster 0 includes people of weight range 70 to 95 who pay the highest insurance cost range of 30 to 70k
- Cluster 1 includes people with weight range of 50 to 70 who are among the lowest insurance cost range of 10k to 30k
- Cluster 2 is difficult to identify with any particular age group. However they are part of age range of 60 to 85 with morderante costs of 10k to 50k.

Any other business insights:

Based on the EDA, Clustering and other visualizations done we can bring following insights which can be helpful in business decisions.

1. Individuals with age range of 70 to 95 are very important group who contribute to maximum insurance costs. Therefore, they need to be given priority in targeting through business development and marketing activities.
2. BMI, Major disease other than heart, weight and age are critical parameters which need to be addressed in any financial modelling for insurance costs.
3. Most of the individuals fall in the range of 25k to 40k of insurance costs. This shows that most of the individuals are comfortable spending within this range and the avg insurance cost for any average healthy individual must fall in this bracket.

4. Model building and interpretation

a. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes)

b. Test your predictive model against the test set using various appropriate performance metrics

c. Interpretation of the model(s)

Test Train Split

We will build multiple models for regression and test the performance for train and test. We will split the data in 80:20 ratio into Train and test set.

The target variable is Insurance Cost for which regression needs to be performed.

```
Train X shape:(20000, 44),
Test X shape: (5000, 44),
Train y shape:(20000,),
Test y shape: (5000,))
```

1.Linear Regression Model

Linear Regression model finds a relationship between a dependent variable and one or more independent variables by fitting a linear equation. The goal is to find the best-fitting line that minimizes the sum of squared differences between the observed values and the predicted values.

Model performance summary after training the model:

Model:	OLS	Adj. R-squared:	1.000
Dependent Variable:	insurance_cost	AIC:	-1322714.4881
Date:	2024-10-17 16:08	BIC:	-1322366.7347
No. Observations:	20000	Log-Likelihood:	6.6140e+05
Df Model:	43	F-statistic:	4.228e+32
Df Residuals:	19956	Prob (F-statistic):	0.00
R-squared:	1.000	Scale:	1.1071e-30

Figure 62 Model 1 Summary

- The AIC score is important for comparing the models. If AIC score is reducing then it is a better model.
- AIC is -1322714 for this model.

Performance metrics train set:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.0	5.853995e-13

Figure 63 Model 1 Train Metrics

We have chosen RMSE, R-squared and Mean Absolute Percentage Error to evaluate the model performance. The Lower the MAPE value of model the better is the performance of the model.

We can see that the RMSE score is 1.05e-15 and MAPE is 5.8e-13. We will see how the model perfroms on test set.

Performance metrics test set:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.0	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.0	5.476335e-13

Figure 64 Model 1 Test Metrics

The metrics have similar performance showing no over fitting of the model. The model is performing better on test set.

RMSE score is 1.03e-15 and MAPE is 5.4e-13. The model performance is better than train performance. The lesser the value of MAPE the better will be the model.

2.Linear Regression Model 2 (removed significant variables)

We will check the significant values and will try to see if the performance improves after removing variable with P value higher than 0.5.

The non-significant variables and their P-values are:

1. regular_checkup_lasy_year(0.5544)
2. avg_glucose_level(0.3982)
3. heart_decs_history (0.5990)
4. Occupation_Salried (0.0860)
5. cholesterol_level_175 to 200 (0.8547)
6. Location_Chennai (1.0)
7. Location_Delhi (0.1275)
8. Location_Guwahati (0.7353)
9. Location_Mangalore (0.1246)
10. Location_Pune (0.3155)

Location_chennai, cholesterol_level_175 to 200, Location_Guwahati have very high P-Values and therefore are non-significant variables.

Weight, weight_change_in_last_one_year, fat_percentage & age are some of the variables which are significant in prediction of the variables. The P-Value for these variables is 0.00.

We will build another Linear Regression Model only with significant variables. Lets see the model summary after fitting the model.

Model:	OLS	Adj. R-squared:	1.000
Dependent Variable:	insurance_cost	AIC:	-1290347.7542
Date:	2024-10-17 16:27	BIC:	-1290079.0356
No. Observations:	20000	Log-Likelihood:	6.4521e+05
Df Model:	33	F-statistic:	1.091e+32
Df Residuals:	19966	Prob (F-statistic):	0.00
R-squared:	1.000	Scale:	5.5878e-30

Figure 65 Model 2 Summary

The AIC value of the model is -1290374. The AIC value is better than Linear Regression model 1. Let us train and test the model.

Model Performance Train set:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.0	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.0	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.0	6.289073e-13

Figure 66 Model 2 Train Metrics

RMSE score is 2.36e-15 and MAPE is 6.28e-13. The model performance is not better than previous model.

Model Performance Test set:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.0	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.0	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.0	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.0	5.772675e-13

Figure 67 Model 2 Test Metrics

RMSE score is 2.33e-15 and MAPE is 5.77e-13. The model performance is not better than previous model although model performs better on test data than train data

3.Linear Regression Ridge Model

Ridge linear regression model adds a penalty term equal to the square of the magnitude of coefficients. This helps to prevent overfitting by discouraging overly complex models.

Train Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.0	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.0	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.0	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.0	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.0	1.087011e-02

Figure 68 Model 3 Train Metrics

RMSE score is 2.12e-05 and MAPE is 1.08e-02. The model performance is not better than previous models.

Test Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.0	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.0	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.0	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.0	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.0	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.0	9.684168e-03

Figure 69 Model 3 Test Metrics

RMSE score is 2.11e-05 and MAPE is 9.68e-03. The model performance is not better than previous models but performs better than train model. Let us see how the Lasso regression model performs.

4. Linear Regression Lasso

Lasso Regression includes a penalty term equal to the absolute value of the magnitude of coefficients, which helps in both regularization and feature selection. It can shrink some coefficients to zero, effectively selecting a simpler model with fewer predictors

We have built Lasso model with alpha =0.1 and max_iterations =500.

Train Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02

Figure 70 Model 4 Train Metrics

RMSE score is 9.9e-03 and MAPE is 3.39e-02. The model performance is not better than previous models.

Test Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02

Figure 71 Model 4 Train Metrics

RMSE score is 9.8e-03 and MAPE is 3.25e-02. The model performance is not better than previous models however performs better on test data than train data.

5.Decision Tree Regression Model

Decision tree algorithm uses a tree-like graph of decisions and their possible consequences. It splits the data into subsets based on feature values, leading to a final decision or output. It's easy to interpret but can be prone to overfitting.

Train performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.479301e-15	1.000000	4.357586e-15

Figure 72 Model 5 Train Metrics

RMSE score is 3.47e-15 and MAPE is 4.35e-15. The **model performance is better** than previous models built till now.

Test Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
5	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
6	Decision Tree (Train set)	3.661853e-15	1.000000	4.389408e-15
7	Decision Tree (Test set)	3.732391e-15	1.000000	4.465949e-15

Figure 73 Model 5 Test Metrics

RMSE score is 3.73e-15 and MAPE is 4.46e-15. The **model performance is better** than previous models built till now. However, model performance is not better than train model performance. The difference between Train RMSE and Test RMSE is comparatively small suggesting no overfitting.

6.Support Vector Machine

SVM algorithm finds the hyperplane that separates data points of different classes in high-dimensional space, maximizing the margin between the closest points (support vectors) of each class.

Train Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
11	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01

Figure 74 Model 6 Train Metrics

RMSE score is 5.43e-02 and MAPE is 2.26e+1. The model performance is not better than previous models built till now.

Test Perfromance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01

Figure 75 Model 6 Train Metrics

RMSE score is 6.51e-02 and MAPE is 2.64e+1. The model performance is not better than previous models built till now. However, test performance is better than train performance.

We have built 6 models till now for regression and found that Decision Trees are giving very good performance than any other model. We will use ensemble techniques to improve the model performance.

Effort to improve Model Performance:

a. Ensemble modelling (if necessary)

b. Any other model tuning measures (if applicable)

c. Interpretation of the most optimum model and its implication on the business

7. Ensembling using Random Forrest Regressor

Ensembling is a powerful technique which uses combines predictions from multiple models to improve model performance and eliminate the weaknesses of the models.

Let us build a Random Forest model along with Bagging and Boosting techniques. Lets build a model with N_estimators =200.

The performance score after building the model is:

- Training set score: 1.000
- Test set score: 1.000

Train Perfomance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15

Figure 76 Model 7 Train Metrics

RMSE score is 3.2e-15 and MAPE is 3.08e-15. The **model performance is better** than previous models built till now based on the MAPE score.

Test Model:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15

Figure 77 Model 7 Test Metrics

RMSE score is 3.2e-15 and MAPE is 3.1e-15. The **model performance is better** than previous models built till now based on the MAPE score. However, test performance is lower than train performance.

8. Bagging with Random Forrest

Bagging is an ensemble technique that creates multiple subsets of the training data through bootstrap sampling and trains a model (usually the same type) on each subset. The final output is an aggregate of the predictions which reduces variance and improves model stability.

Train Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07

Figure 78 Model 8 Train Metrics

RMSE score is 3.06e-05 and MAPE is 5.15e-07. The model performance is not better than previous models built till now based on the MAPE score.

Test Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07

Figure 79 Model 8 Test Metrics

RMSE score is 3.36e-05 and MAPE is 5.75e-07. The model performance is not better than previous models built till now based on the MAPE score. Test performance is also lower than train performance.

9. Bagging with Decision Tree Regressor

Train Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.000000	2.047845e-15

Figure 80 Model 9 Train Metrics

RMSE score is 1.9e-15 and MAPE is 2.04e-15. The **model performance is better** than previous models built till now based on the MAPE score.

Test Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.000000	2.047845e-15
17	Bagging with Decision Tree (Test set)	1.968472e-15	1.000000	2.123253e-15

Figure 81 Model 9 Test Metrics

RMSE score is 1.96e-15 and MAPE is 2.12e-15. The **model performance is better** than previous models built till now based on the MAPE score. However test performance is low compared to train performance.

10. Boosting

Boosting is an ensemble method that sequentially trains models, with each new model focusing on the errors made by the previous ones. It combines weak learners to create a strong predictive model. We will use Gradient Boosting algorithm to train and test the model.

Train Performance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.000000	2.047845e-15
17	Bagging with Decision Tree (Test set)	1.968472e-15	1.000000	2.123253e-15
18	Boosting (Train set)	3.603290e-04	1.000000	6.553084e-03

Figure 82 Model 10 Train Metrics

RMSE score is 3.6e-04 and MAPE is 6.55e-03. The model performance is not better than previous models built till now based on the MAPE score.

Test Perfromance:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.000000	2.047845e-15
17	Bagging with Decision Tree (Test set)	1.968472e-15	1.000000	2.123253e-15
18	Boosting (Train set)	3.603290e-04	1.000000	6.553084e-03
19	Boosting (Test set)	3.467312e-04	1.000000	6.171581e-03

Figure 83 Model 10 Train Metrics

RMSE score is 3.46e-04 and MAPE is 6.17e-03. The model performance is not better than previous models built till now based on the MAPE score. However test performance is better compared to train performance.

Best Model:

The best model with lowest Mean Absolute Percentage Error of 2.04e-15 is Bagging model with Decision Tree Regressor as base model.

	Model Name	RMSE	R-squared	MAPE
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.0	2.047845e-15

Figure 84 Best Model

Model Comparision:

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.000000	2.047845e-15
17	Bagging with Decision Tree (Test set)	1.968472e-15	1.000000	2.123253e-15
18	Boosting (Train set)	3.603290e-04	1.000000	6.553084e-03
19	Boosting (Test set)	3.467312e-04	1.000000	6.171581e-03

Figure 85 Model Comparison

- Bagging with Decision Tree model is the best performing model for both test and train set.
- Bagging with Decision Tree model has very small difference between RMSE scores for Test and Train. This means model is not overfitting.
- Boosting the model does not give any boost in performance of the model.
- The second best performing model is Random Forest model.
- The third best model is Linear Regression model.
- The final model which should be deployed is Bagging with Decision tree model.
- We have used Bagging and boosting for tuning the model for best results and has helped improvise the prediction of target variable

Figure 86 Predicted Values

5.Model Validation

How was the model validated? Just accuracy, or anything else too?

The model validation is an important step to ensure that the model is giving suitable performance after training the data on test set.

	Model Name	RMSE	R-squared	MAPE
0	Linear Regression Model 1 (Train set)	1.051035e-15	1.000000	5.853995e-13
1	Linear Regression Model 1 (Test set)	1.034615e-15	1.000000	5.476335e-13
2	Linear Regression Model 2 (significant values)...	2.361847e-15	1.000000	6.289073e-13
3	Linear Regression Model 2 (significant values)...	2.336508e-15	1.000000	5.772675e-13
4	Linear Regression Ridge (Train set)	2.125460e-05	1.000000	1.087011e-02
5	Linear Regression Ridge (Test set)	2.110553e-05	1.000000	9.684168e-03
6	Linear Regression Lasso (Train set)	9.968682e-03	0.999901	3.395467e-02
7	Linear Regression Lasso (Test set)	9.811903e-03	0.999901	3.255109e-02
8	Decision Tree (Train set)	3.595197e-15	1.000000	4.346328e-15
9	Decision Tree (Test set)	3.663559e-15	1.000000	4.421111e-15
10	SVM (Train set)	5.434540e-02	0.997065	2.665793e+01
11	SVM (Test set)	6.516666e-02	0.995643	2.647130e+01
12	Random Forrest (Train set)	3.203327e-15	1.000000	3.086118e-15
13	Random Forrest (Test set)	3.209199e-15	1.000000	3.104425e-15
14	Bagging with Random Forest (Train set)	3.063573e-05	1.000000	5.150239e-07
15	Bagging with Random Forest (Test set)	3.363267e-05	1.000000	5.750397e-07
16	Bagging with Decision Tree (Train set)	1.921597e-15	1.000000	2.047845e-15
17	Bagging with Decision Tree (Test set)	1.968472e-15	1.000000	2.123253e-15
18	Boosting (Train set)	3.603290e-04	1.000000	6.553084e-03
19	Boosting (Test set)	3.467312e-04	1.000000	6.171581e-03

The model validation steps are explained during the model building. The metrics that are important for measuring the test performance or validation of model are

1. **Mean absolute percentage error (MAPE):** It calculates the average percentage difference between the actual and predicted values, expressed as a percentage. Lower values of MAPE indicate better model performance. This metric is commonly used in business applications, but it can be misleading if the true values have a lot of

zeros or very small values. Here we have used MAPE is most importanct metric for valiadation of the model.

- Bagging with Decision Tree was validated with a MAPE value of 2.12e-15 which has the best MAPE value and also the lowest among all models built.

2. **Root Mean Squared Error (RMSE):** MSE is the square root of MSE and gives an error metric in the same units as the target variable. Like MSE, it penalizes large errors but is more interpretable because the result is in the original units. Lower values of RMSE indicate better model performance. RMSE is sensitive to large deviations and is commonly used when larger errors are particularly undesirable.
    - Bagging with Decision Tree has RMSE value of 1.9e-15 for validation set.
    - Linear Regression Model 1 has RMSE value of 1.03e-15 which is lower than Bagging model but do not have better MAPE value which is only 5.47e-13 which is very high than MAPE of Bagging model 1.9e-15

3. **R-Sqared**: R^2 represents the proportion of variance in the target variable that is explained by the model. It is a measure of how well the model fits the data. R^2 ranges from 0 to 1. A value close to 1 indicates that the model explains most of the variance in the data. A value close to 0 means that the model doesn't explain much of the variance. Negative values indicate that the model performs worse than a simple mean predictor.
    - R-sqaured value for Bagging with Decision tree model is 1. Most of the model have R-sqaured values is 1 meaning that model explain most of the variance and fits the data well.

6. Final interpretation / recommendation

Detailed recommendations for the management/client based on the analysis done.

Implication on Business due to final model:

If we see the actual and predicted value of the model we can see that the predictive power of the model is very high. Model is able to predict correctly upto 13 decimal places. Therefore the model is highly reliable for regression problem.

The model when provided with new parameter values will be able to give the exact cost for insurance for any individual. This will help the business to ensure that the insurance premium being charged is similar for individuals based on the parameters.

	True values	Predicted values	Accuracy
20094	-1.206096	-1.206096	-7.364075e-14
17218	-0.344569	-0.344569	-0.000000e+00
3773	-0.086111	-0.086111	2.900897e-13
20529	-1.119944	-1.119944	-1.784377e-13
18073	0.086194	0.086194	-6.762266e-13

Detailed Business Recommendations

1. **Target elderly (70-95) for high-premium plans**: This age group contributes significantly to higher insurance costs. Tailor products for their needs, such as chronic illness management and personalized care options.
2. **Develop cost-effective insurance products for younger, middle-aged (40-60) groups**: Focus on preventive care and family coverage options that can cater to this demographic's lower premiums but high coverage needs.
3. **Design weight-based insurance plans**: Since weight is strongly correlated with insurance cost, create pricing models based on BMI and weight ranges, offering discounts for healthier individuals.
4. **Use health risk factors like BMI and medical history** in premium calculations: People with a high BMI or significant medical history should be priced higher, while those with better health can be offered lower premiums.
5. **Promote fitness programs** to incentivize healthier behavior: Offer discounts or rewards for clients engaged in fitness or wellness programs, particularly targeting those with a BMI over 30.

6. **Segment products for high-risk individuals (Cluster 0)**: Older adults with high weight and chronic conditions form a critical segment. Focus on offering plans with higher coverage for medical treatments and hospitalization.
7. **Expand product offerings to address gender-specific health needs**: Use gender-based insights (e.g., higher BMI in men) to create personalized products, such as women's health packages or heart disease-focused plans for men.
8. **Leverage regional marketing in high-density areas like Bangalore**: Focus marketing efforts in regions like Bangalore, which have the highest concentration of insured individuals, while considering expansion in lower-represented areas like Surat.
9. **Introduce specialized plans for individuals with frequent doctor visits**: Based on the clustering results, people with higher doctor visits may require more comprehensive health insurance with added benefits for frequent check-ups.
10. **Encourage long-term relationships with clients**: Offer loyalty benefits or reduced premiums for customers who have been with the company for multiple years, especially those with consistent insurance coverage over 3 years.