Capstone Project

Health Insurance

- R.PRAKASH

Business Problem:

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. We have to use the health and habit related parameters for the estimated cost of insurance

File: Data.csv
Target variable: insurance_cost

Data dictionary:

Variable	Business Definition
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 of the business problem

a) Defining problem statement:

To build various regression models and choose the best model in determining the insurance cost from 23 independent features. The independent features given are related to health, habitual, occupational and few others.

b) Need of the study/project:

1. To gather data.

2. Perform EDA and draw insights.

3.Data cleaning.

4. Undertake various feature engineering techniques.

5. Split the data into train and test data.

6. Build the model and choose the best model by considering multiple performance metrics.

c) Understanding business/social opportunity:

This project will help multiple stakeholders such as insurance companies, customers, government agencies, etc in knowing the parameters involved in finding the insurance cost.

a) On consumers perspective:

- Will help in selecting optimal health insurance plan form multiple companies.
- Maintain better health practices to minimise out of pocket health expenditure.

b) On business perspective:

- Fix better pricing standards to have edge over competitors.
- To optimize the cost of insurance to attract beneficiaries, thereby increasing customer base.
- Provide personalized insurance products with appropriate value-added benefits to customers.
- Affordable healthcare insurance to the downtrodden sections of the society

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

- There are 25000 rows and 24 variables in the dataset.
- In this dataset, there are 8 object datatype, 2 float datatype and 14 integer datatype.

- Year_last_admitted variable is given in float datatype, since it belongs to date format we need to convert it into datetime format.

```
<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                             24010 non-null  float64
 14  smoking_status                  25000 non-null  object
 15  Year_last_admitted              13119 non-null  float64
 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: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Fig:1.1-Information of the dataset

- The description of the dataset is detailed below:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
applicant_id	25000.0	NaN	NaN	NaN	17499.5	7217.022701	5000.0	11249.75	17499.5	23749.25	29999.0
years_of_insurance_with_us	25000.0	NaN	NaN	NaN	4.08904	2.606612	0.0	2.0	4.0	6.0	8.0
regular_checkup_lasy_year	25000.0	NaN	NaN	NaN	0.77368	1.199449	0.0	0.0	0.0	1.0	5.0
adventure_sports	25000.0	NaN	NaN	NaN	0.08172	0.273943	0.0	0.0	0.0	0.0	1.0
Occupation	25000	3	Student	10169	NaN	NaN	NaN	NaN	NaN	NaN	NaN
visited_doctor_last_1_year	25000.0	NaN	NaN	NaN	3.1042	1.141663	0.0	2.0	3.0	4.0	12.0
cholesterol_level	25000	5	150 to 175	8763	NaN	NaN	NaN	NaN	NaN	NaN	NaN
daily_avg_steps	25000.0	NaN	NaN	NaN	5215.88932	1053.179748	2034.0	4543.0	5089.0	5730.0	11255.0
age	25000.0	NaN	NaN	NaN	44.91832	16.107492	16.0	31.0	45.0	59.0	74.0
heart_decs_history	25000.0	NaN	NaN	NaN	0.05464	0.227281	0.0	0.0	0.0	0.0	1.0
other_major_decs_history	25000.0	NaN	NaN	NaN	0.09816	0.297537	0.0	0.0	0.0	0.0	1.0
Gender	25000	2	Male	16422	NaN	NaN	NaN	NaN	NaN	NaN	NaN
avg_glucose_level	25000.0	NaN	NaN	NaN	167.53	62.729712	57.0	113.0	168.0	222.0	277.0
bmi	24010.0	NaN	NaN	NaN	31.393328	7.876535	12.3	26.1	30.5	35.6	100.6
smoking_status	25000	4	never smoked	9249	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Year_last_admitted	13119.0	NaN	NaN	NaN	2003.892217	7.581521	1990.0	1997.0	2004.0	2010.0	2018.0
Location	25000	15	Bangalore	1742	NaN	NaN	NaN	NaN	NaN	NaN	NaN
weight	25000.0	NaN	NaN	NaN	71.61048	9.325183	52.0	64.0	72.0	78.0	96.0
covered_by_any_other_company	25000	2	N	17418	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Alcohol	25000	3	Rare	13752	NaN	NaN	NaN	NaN	NaN	NaN	NaN
exercise	25000	3	Moderate	14638	NaN	NaN	NaN	NaN	NaN	NaN	NaN
weight_change_in_last_one_year	25000.0	NaN	NaN	NaN	2.51796	1.690335	0.0	1.0	3.0	4.0	6.0
fat_percentage	25000.0	NaN	NaN	NaN	28.81228	8.632382	11.0	21.0	31.0	36.0	42.0
insurance_cost	25000.0	NaN	NaN	NaN	27147.40768	14323.691832	2468.0	16042.0	27148.0	37020.0	67870.0

Fig-1.2-Description of the dataset

Head of the dataset:

	0	1	2	3	4
applicant_id	5000	5001	5002	5003	5004
years_of_insurance_with_us	3	0	1	7	3
regular_checkup_lasy_year	1	0	0	4	1
adventure_sports	1	0	0	0	0
Occupation	Salried	Student	Business	Business	Student
visited_doctor_last_1_year	2	4	4	2	2
cholesterol_level	125 to 150	150 to 175	200 to 225	175 to 200	150 to 175
daily_avg_steps	4866	6411	4509	6214	4938
age	28	50	68	51	44
heart_decs_history	1	0	0	0	0
other_major_decs_history	0	0	0	0	1
Gender	Male	Male	Female	Female	Male
avg_glucose_level	97	212	166	109	118
bmi	31.2	34.2	40.4	22.9	26.5
smoking_status	Unknown	formerly smoked	formerly smoked	Unknown	never smoked
Year_last_admitted	NaN	NaN	NaN	NaN	2004.0
Location	Chennai	Jaipur	Jaipur	Chennai	Bangalore
weight	67	58	73	71	74
covered_by_any_other_company	N	N	N	Y	N
Alcohol	Rare	Rare	Daily	Rare	No
exercise	Moderate	Moderate	Extreme	No	Extreme
weight_change_in_last_one_year	1	3	0	3	0
fat_percentage	25	27	32	37	34
insurance_cost	20978	6170	28382	27148	29616

Fig-1.3-Head of the dataset

Exploratory Data Analysis:

a) years_of_insurance_with_us

- This variable denotes for how many years a customer has been taking policy from the same company only.
- The number of years of insurance the customer holds with the same company ranges from 0 to 8 years.
- There are no outliers in the variable and also no skewness can be seen.

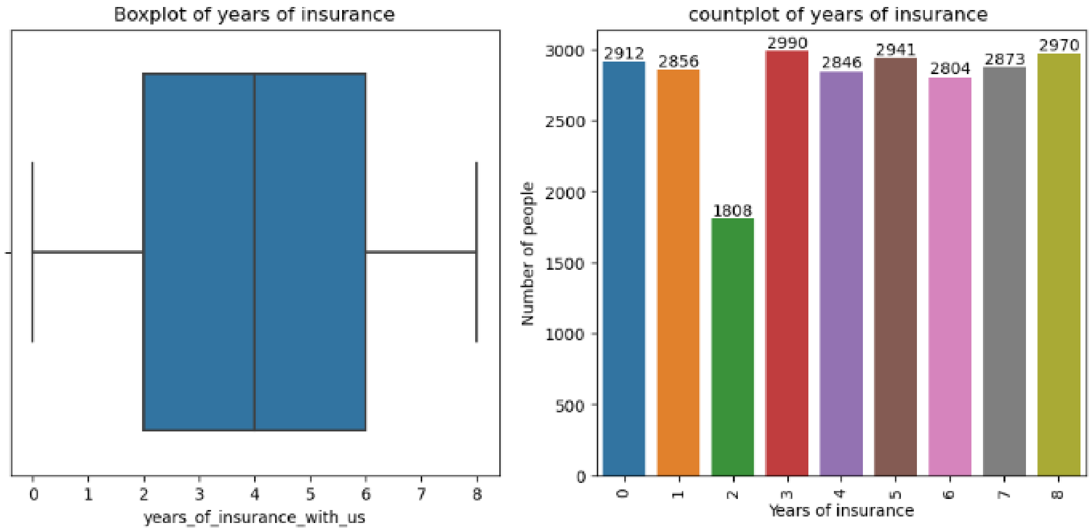


Fig-1.4-Plot of years of insurance

b)regular_checkup_lasy_year:

- It denotes number of times a customer has undergone regular health check-up in the last one year.
- Nearly 60 percent of the customers didn't even get checked once in the last year.

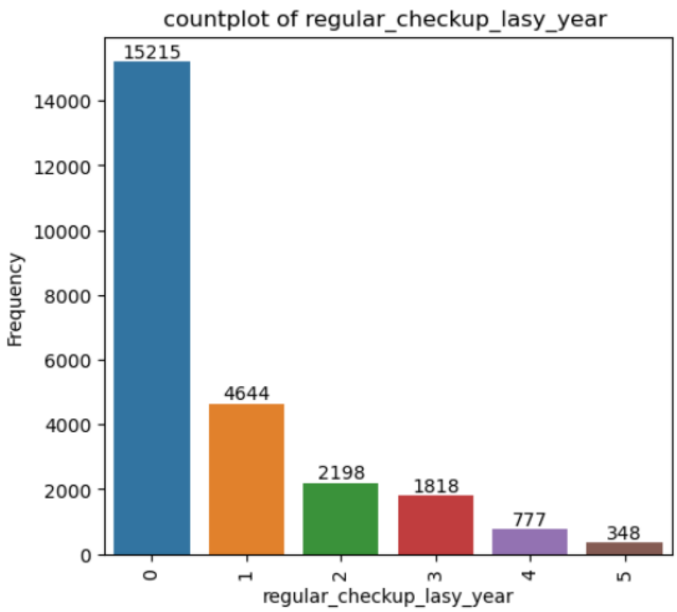


Fig-1.5-Plot of regular check up last year

c)adventure_sports

- It denotes the customers involved with adventure sports like climbing, diving.
- 1-denotes involved in adventure sports

- 0-denotes not involved.
- Almost 8 percent of the customers were involved in adventure sports.

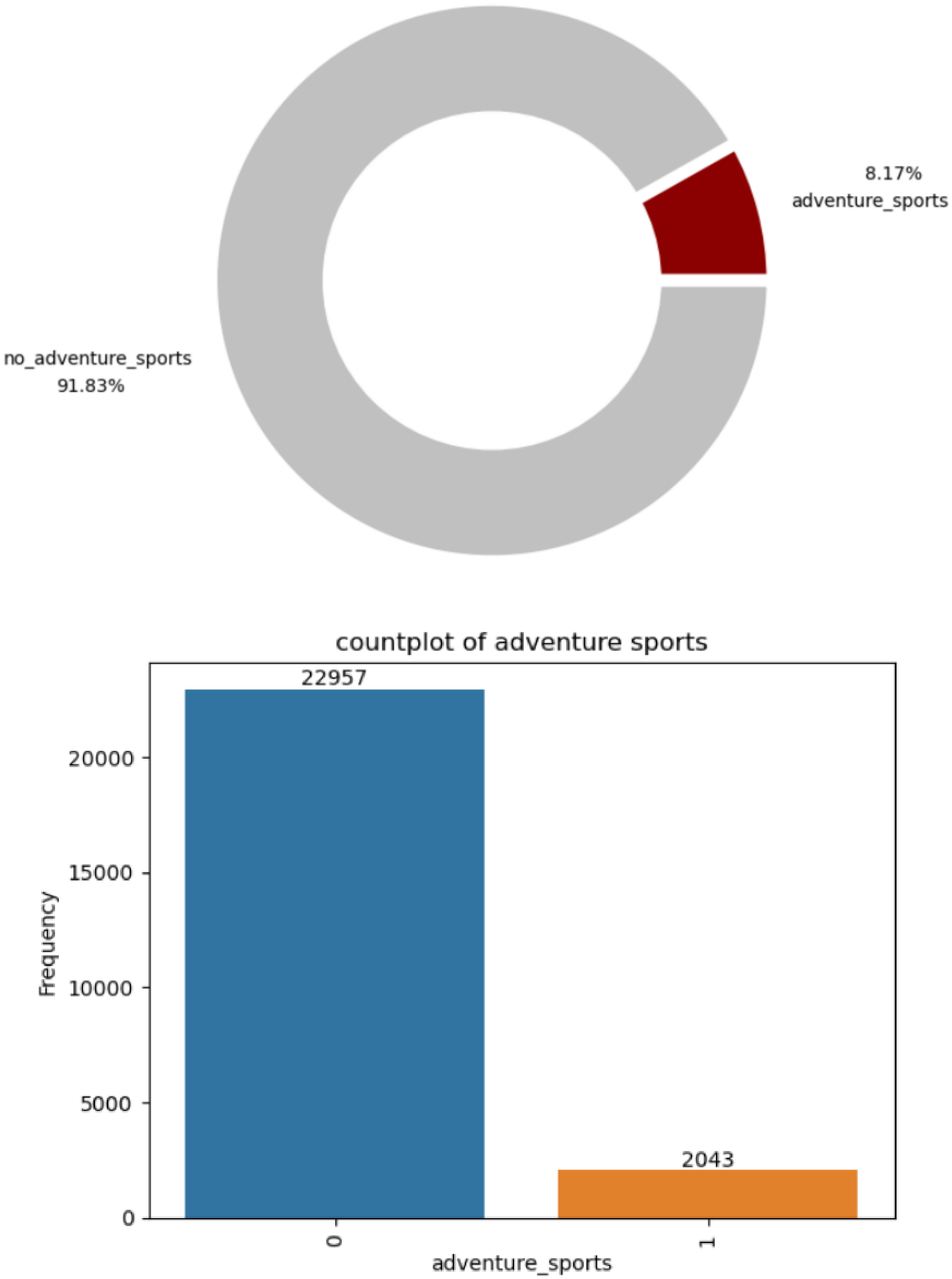


Fig-1.6-Plot of adventure sports

d) Occupation:

- It seems that the customers are less among the salaried class.
- Student and business form major part of customer base.

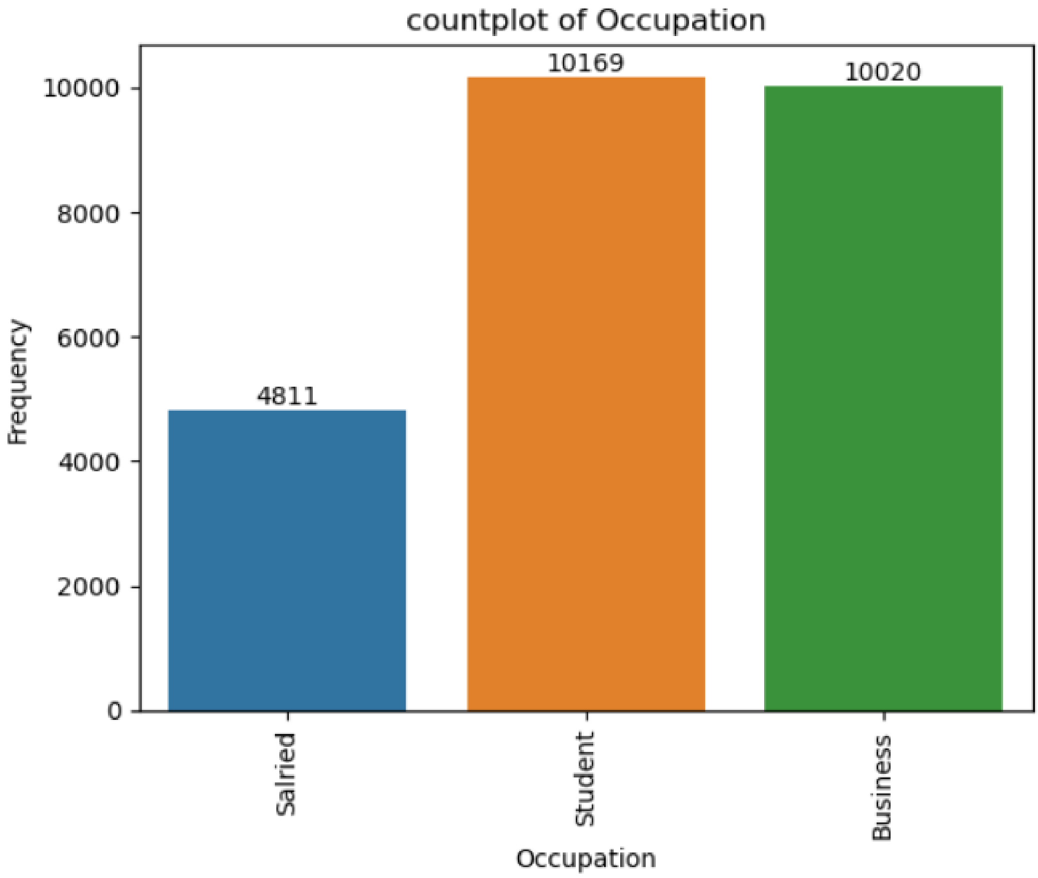


Fig-1.7-Plot of occupation

e) visited_doctor_last_1_year:

- It shows the number of times a customer has visited doctor in last one year.
- Nearly 98 percent of customer had visited doctor at least twice in last one year.

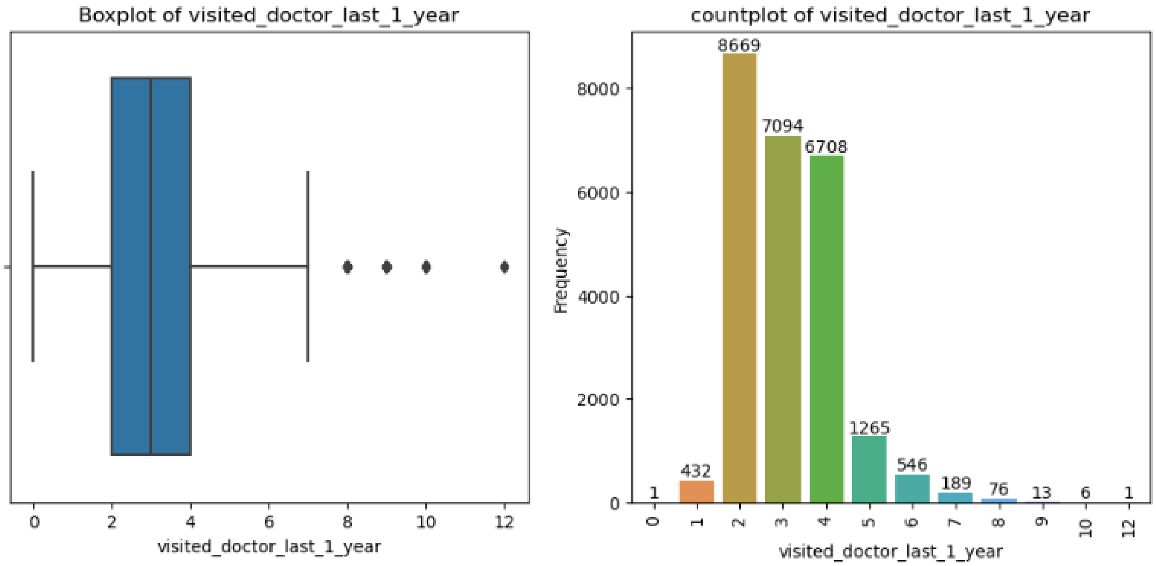


Fig-1.8-Plot of visited doctor last 1 year

f) cholesterol_level:

- Customers cholesterol levels are segregated under 5 categories.
- Most number of customers are in the range of 150 to 175 and least number of customers is in the range of 225 to 250 level.
- 125 to 150 level indicate lower cholesterol level and 225 to 250 level indicate higher cholesterol level.

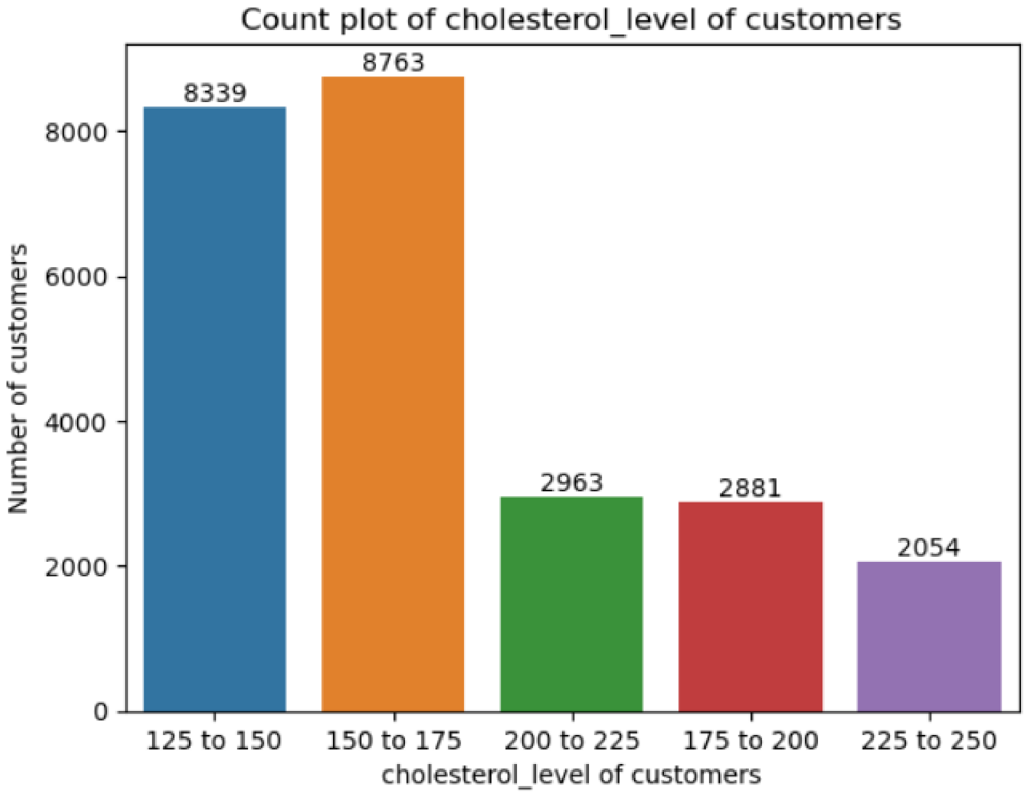
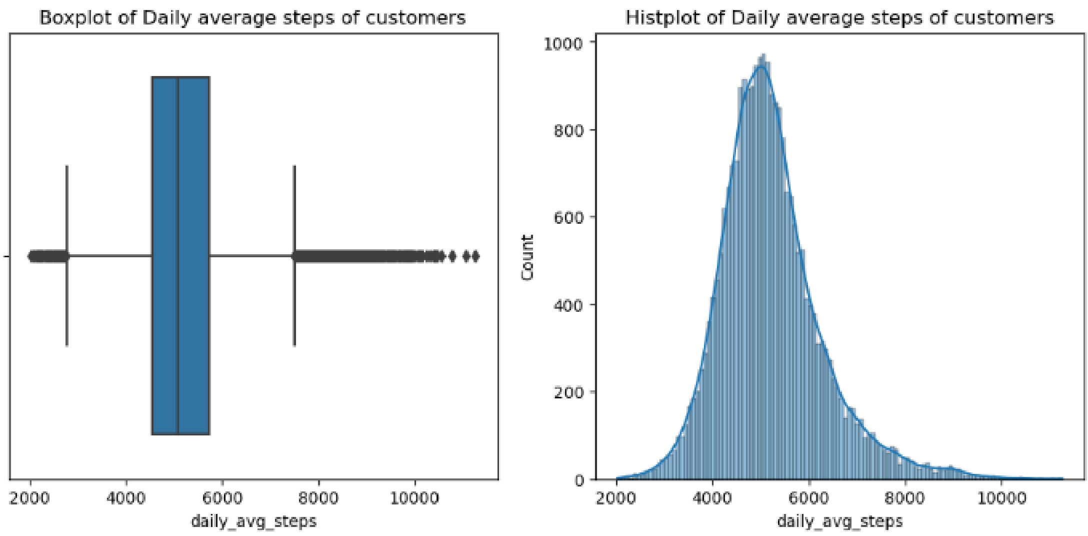


Fig-1.9-Plot of cholesterol level of customer

g)daily_avg_steps:

- It tells the average number of steps walked by the customer daily.
- This feature follows normal distribution.
- There are outliers in the dataset.
- The average number of steps among all the customers is 5215.

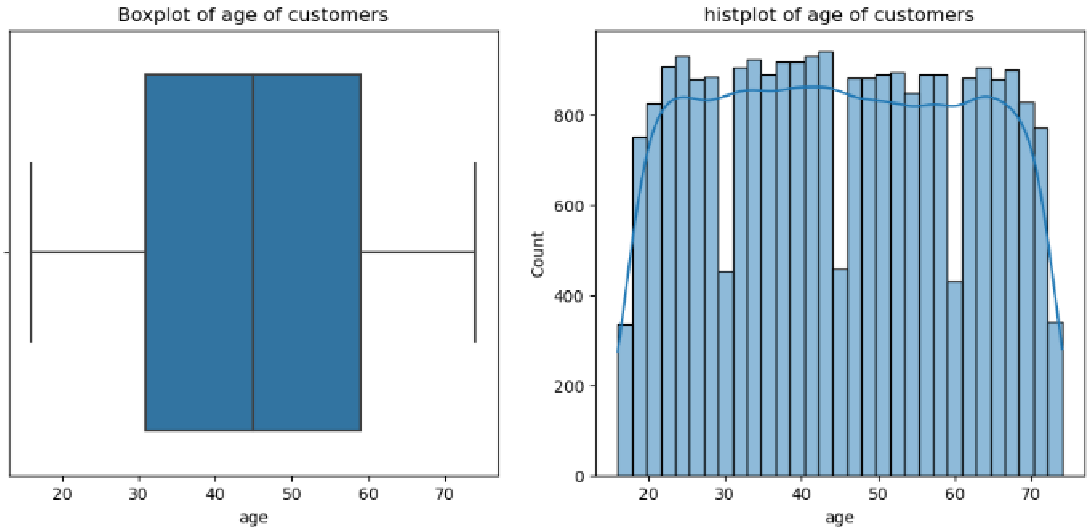



```
count    25000.000000
mean      5215.889320
std       1053.179748
min       2034.000000
25%       4543.000000
50%       5089.000000
75%       5730.000000
max      11255.000000
Name: daily_avg_steps, dtype: float64
```

Fig-1.10-Plot of daily_avg_steps

h) Age of the customer:

- The average age of the customers is 45 years.
- Minimum and maximum age of customers in the dataset is 16 and 74 respectively.




```
count    25000.000000
mean        44.918320
std         16.107492
min         16.000000
25%         31.000000
50%         45.000000
75%         59.000000
max         74.000000
Name: age, dtype: float64
```

Fig-1.11-Plot of age

i) heart_decs_history:

- It tells whether the customer has a history of heart disease or not. 1 denotes yes and 0 denotes No.
- Nearly 94 percent of the customers do not have a history of heart disease.

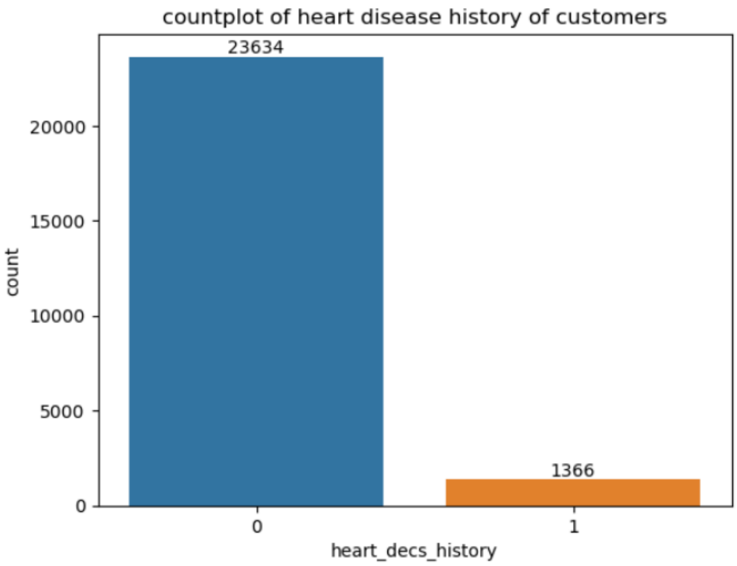


Fig-1.12-Plot of heart_desc_history

j) other_major_decs_history:

- It includes any past major diseases apart from heart like any operation.
- 9.8 percent of the customer have a history of a major disease.

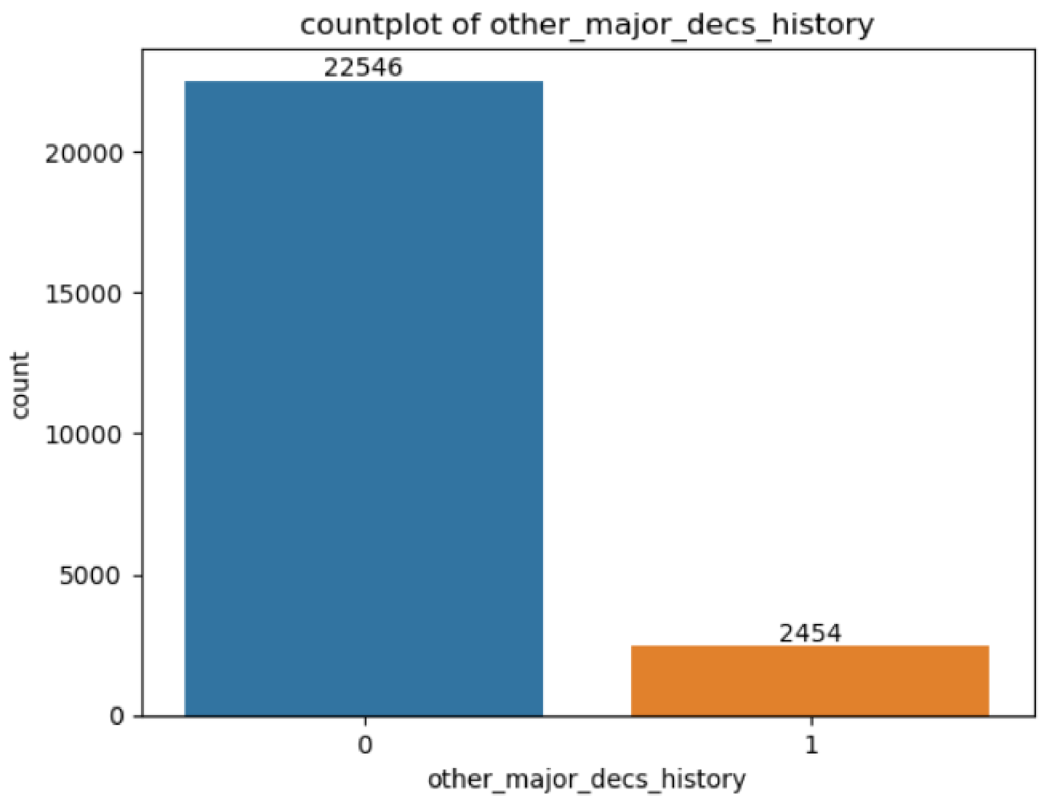


Fig-1.13-Plot of other_major_desc_history

k)Gender:

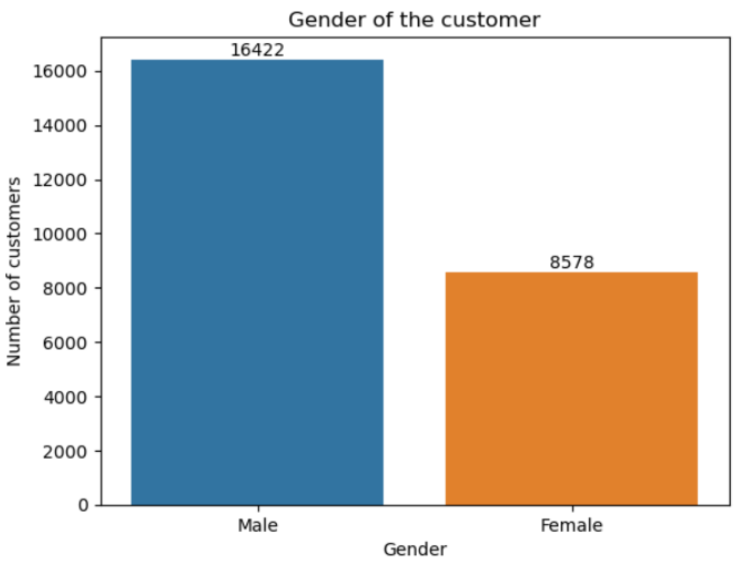


Fig-1.14-Plot of Gender

- Nearly 65 percent of the customers are male.
- From this, we can say that medical insurance penetration is less among the women.

l) avg_glucose_level:

- It denotes the average glucose level of the customers while applying the insurance.
- Average glucose level varies from 57 to 277.

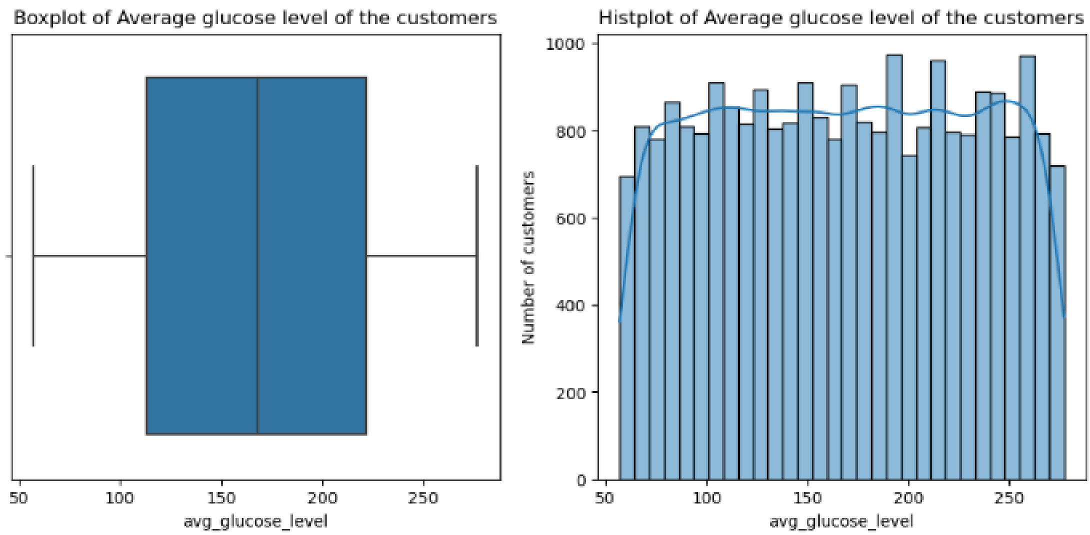


Fig-1.15-Plot of avg_glucose_level

j)Body Mass Index(bmi):

- This variable denotes the body mass index of the customers.
- There are outliers in the dataset.
- It is normally distributed.

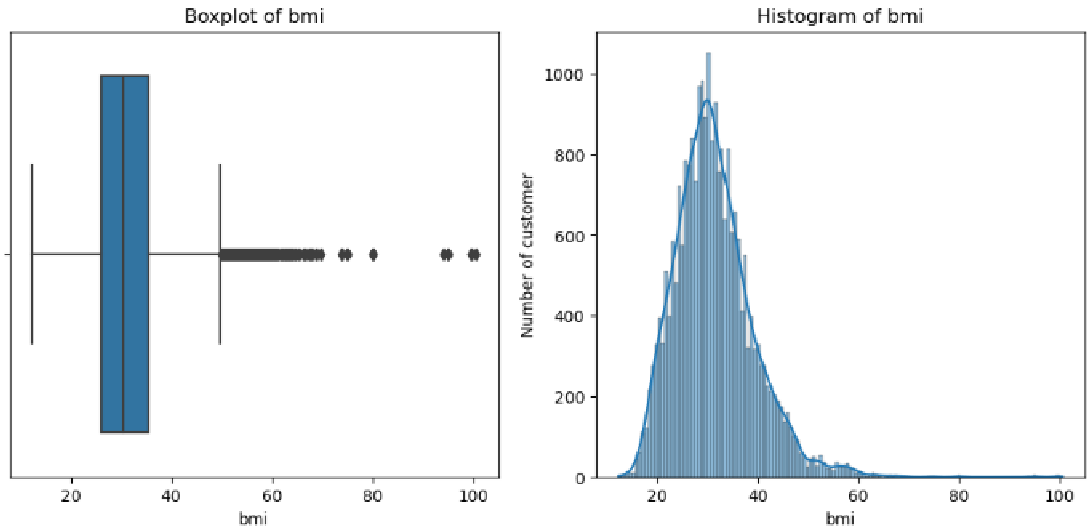


Fig-1.15-Plot of bmi

k) smoking_status:

- This feature tells the Smoking status of the customers.
- Customers who never smoked are more in this dataset.

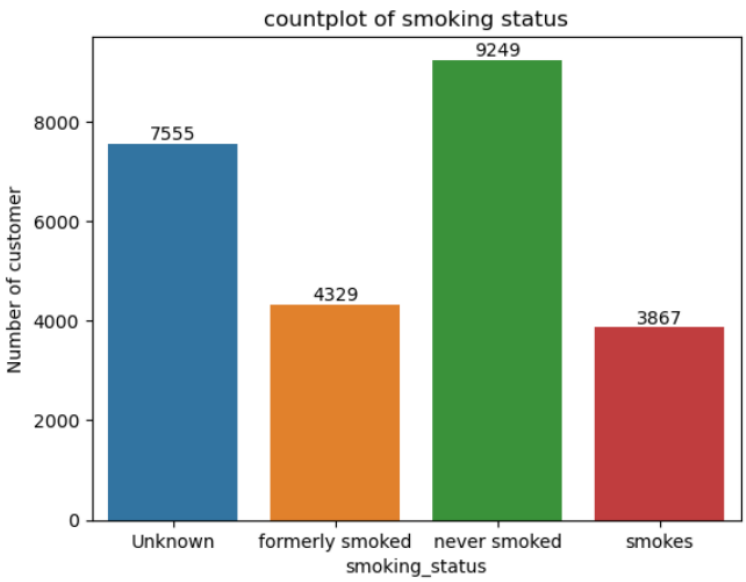


Fig-1.16-Plot of smoking_status

l) Year_last_admitted:

- It tells when customer have been admitted in the hospital last time.
- It varies from 1990 to 2018.

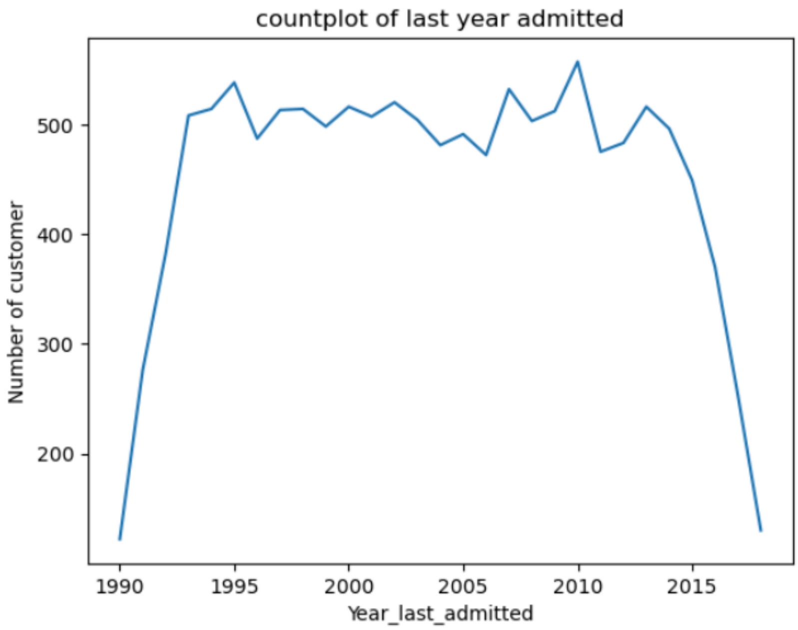


Fig-1.17-Plot of year_last_admitted

m)Location:

- Surat has lowest number of customers and Bangalore has highest number of customers.
- There is no significant difference in number of customers between various cities.

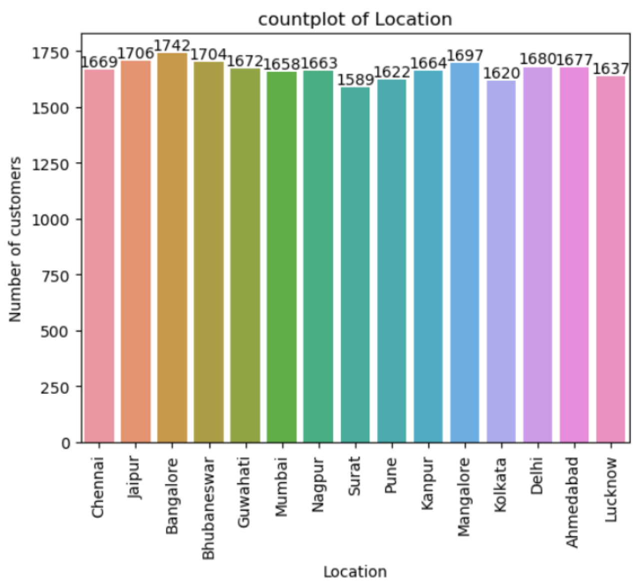
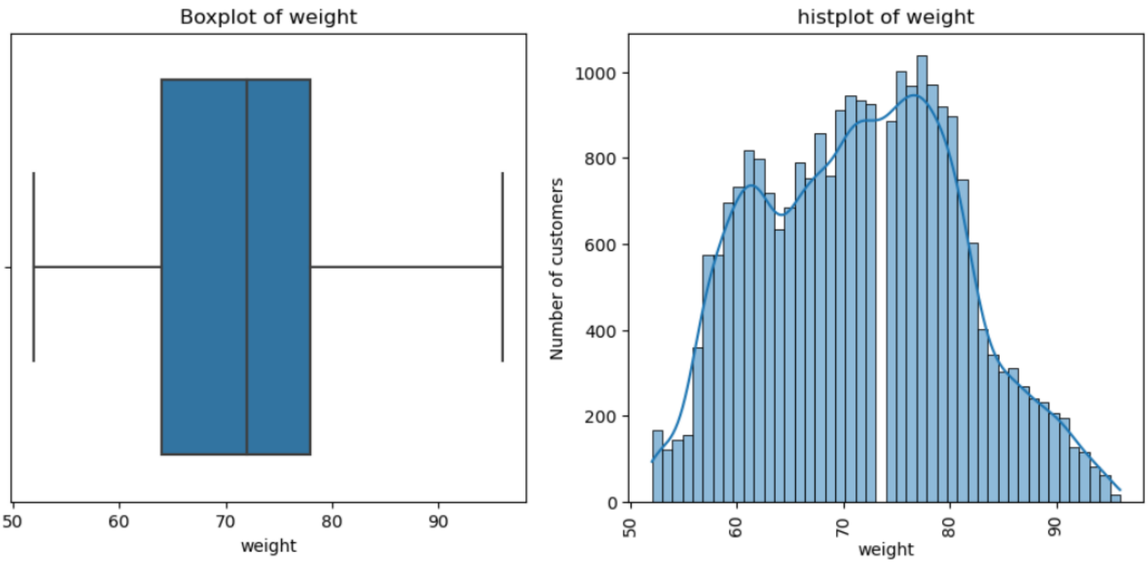


Fig-1.18-Plot of location

n) weight:

- There are no outliers in the dataset.
- The average weight of the customers is 71.6.
- Minimum and maximum weight of the customers is 52 and 96 respectively.




```
count    25000.000000
mean        71.610480
std          9.325183
min         52.000000
25%         64.000000
50%         72.000000
75%         78.000000
max         96.000000
Name: weight, dtype: float64
```

Fig-1.19-Plot of weight

o) covered_by_any_other_company:

- It tells whether customer is covered from any other insurance company or not.
- Nearly 30 percent of the customers have insurance from other company.

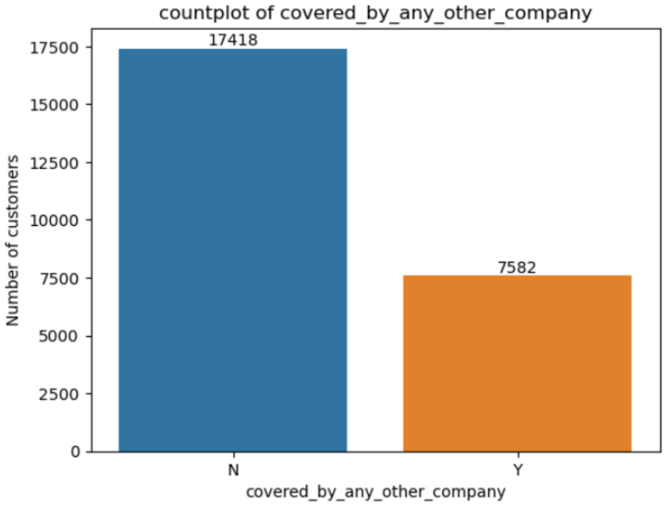


Fig-1.20-covered by any other company

p)Alcohol:

- It denotes the alcohol consumption status of the customer.
- Nearly 8541 have no alcohol consumption status.

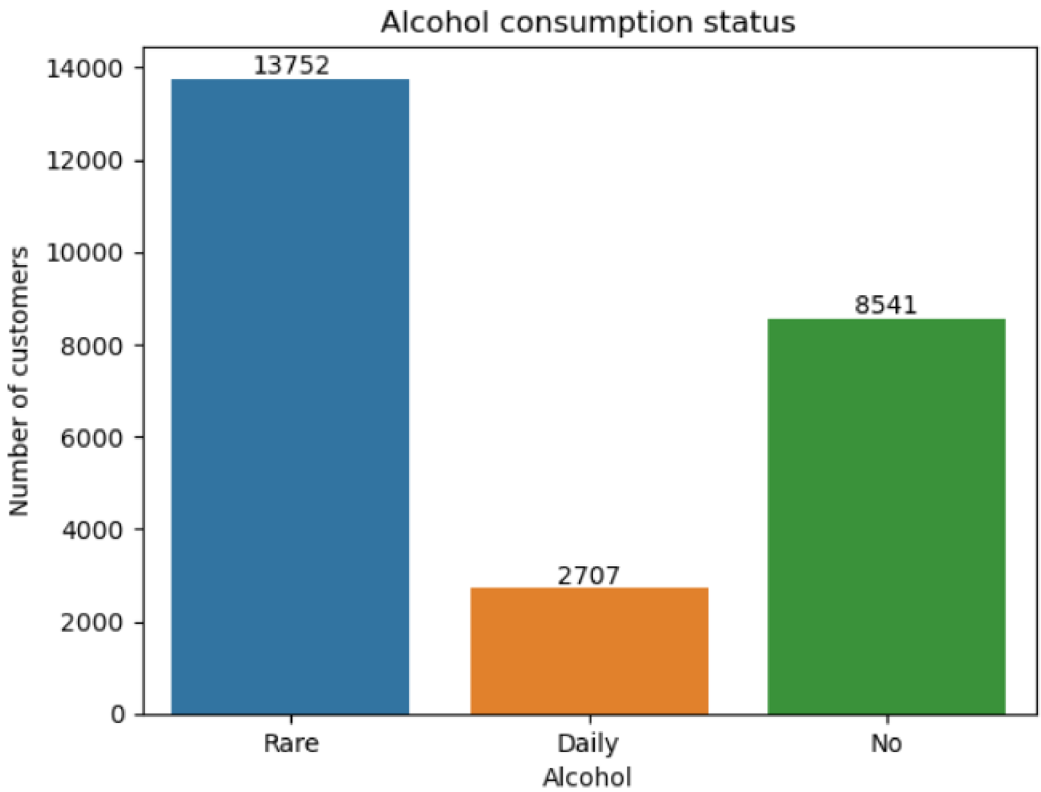


Fig-1.21-Plot alcohol consumption status

q) Exercise:

- It denotes the regular exercise status of the customer.
- Most of the customers had moderate exercising activity.

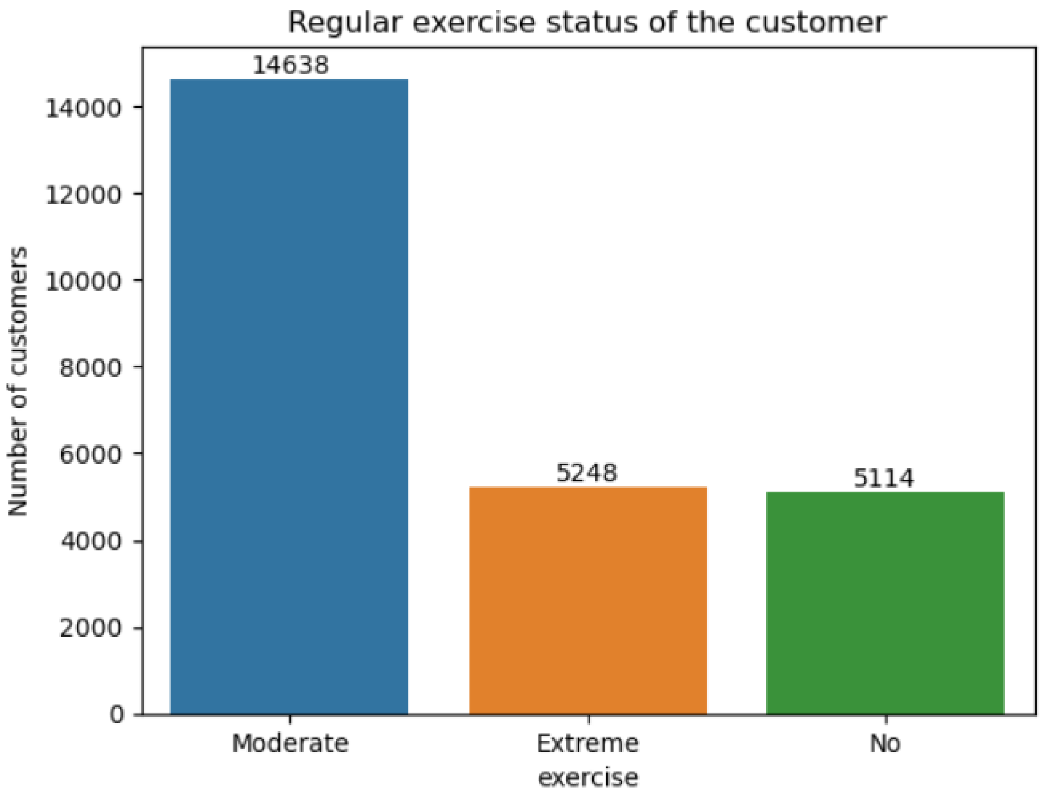
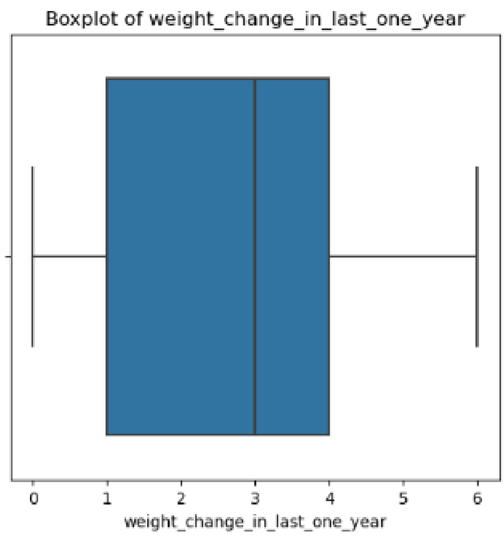


Fig-1.22-Plot of exercise

r)weight_change_in_last_one_year:

- It tells the how much variation has been seen in the weight of the customer in last year.
- Among the all the customers, variation of 4 is highest.



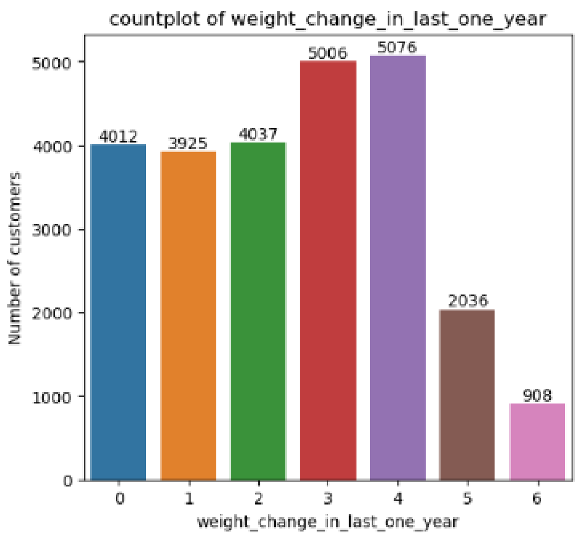


Fig-1.23-Plot of weight_change_in_last_one_year

s)fat_percentage:

- It denotes the fat percentage of the customer while applying the insurance.
- The mean fat percentage of the customer is 28 percentage.
- There are no outliers.
- It seems to be left skewed.

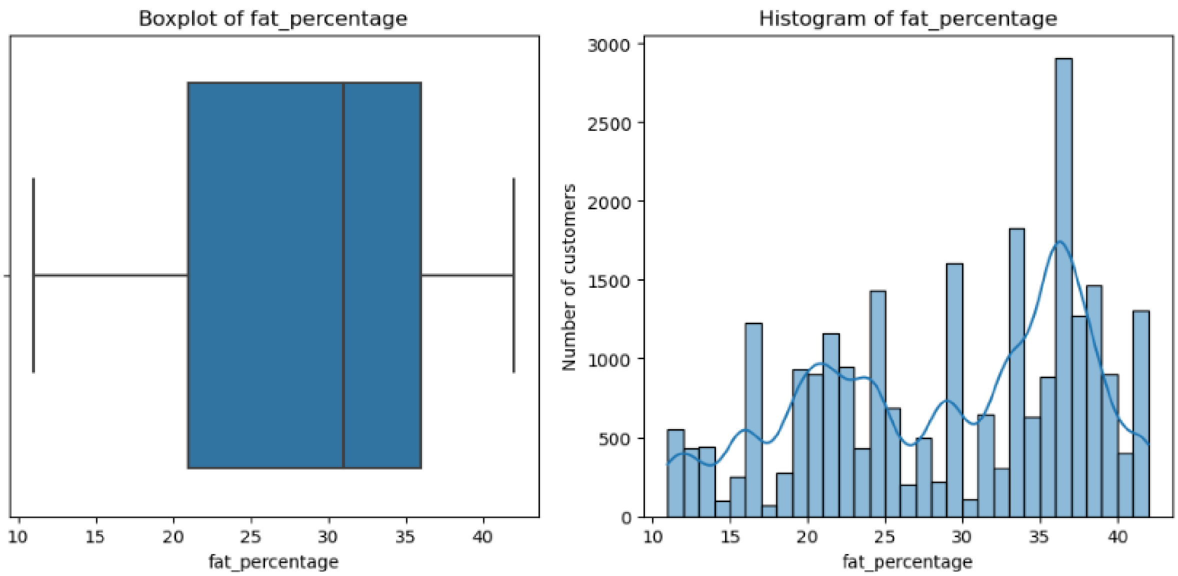


Fig-1.24-Plot of fat_percentage

Target variable-insurance_cost:

- It denotes the insurance cost of various customers.
- It is ***not normally distributed***.
- There are no outliers in the dataset.
- Minimum insurance cost is 2468.
- Maximum insurance cost is 67870.

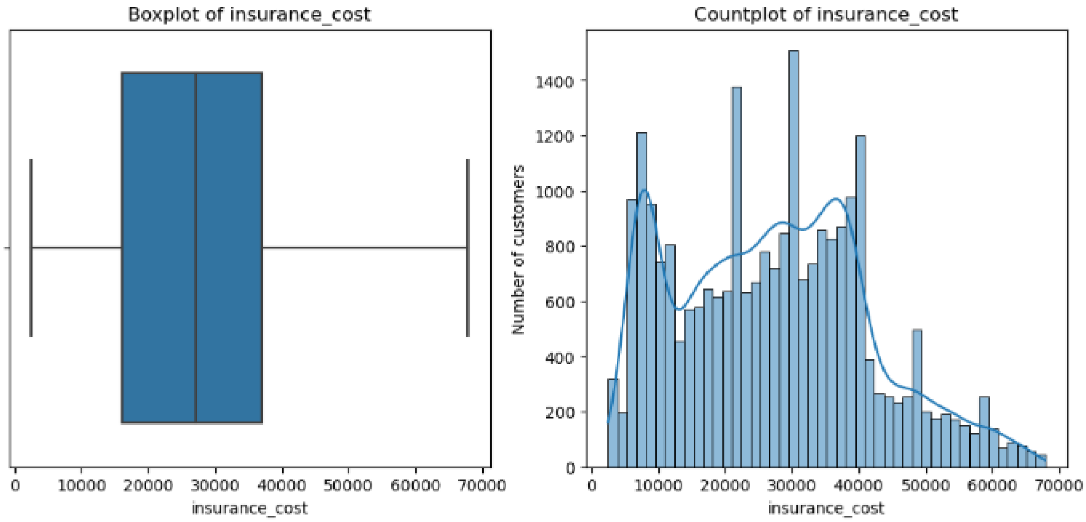


Fig-1.25-Plot of insurance _cost

Bi-Variate analysis:

a)Weight vs Insurance cost:

- There is strong correlation between weight and insurance cost.

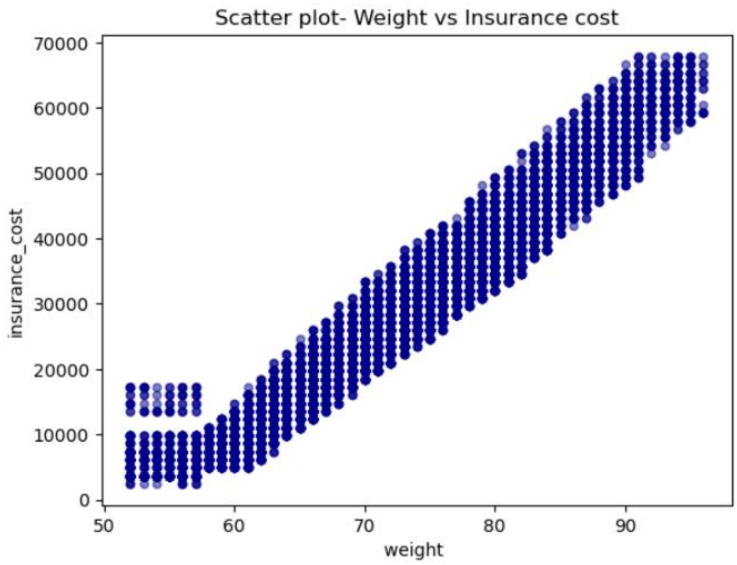


Fig-1.25-Plot of weight vs insurance _cost

b) Bmi vs Insurance cost:

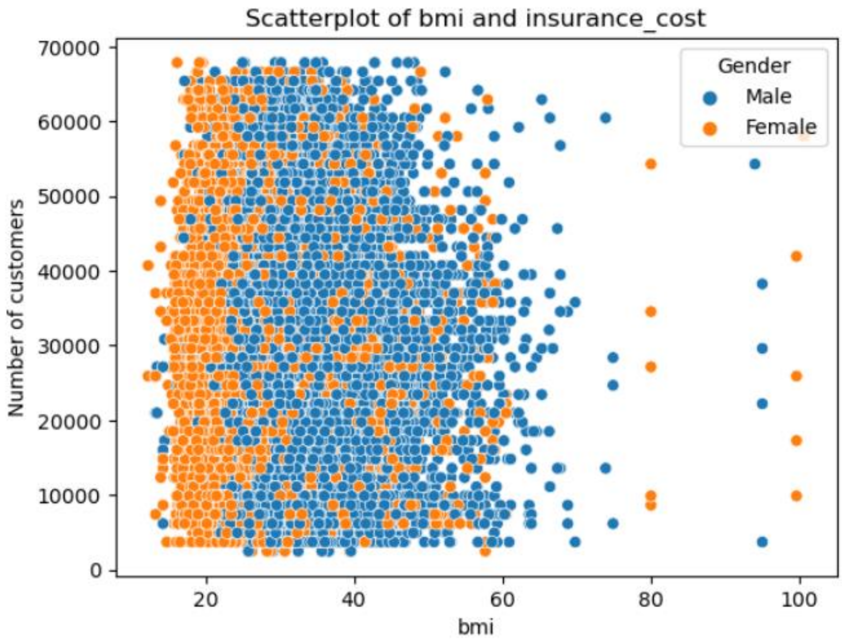


Fig-1.25-Plot of Bmi vs Insurance cost

- There is no significant correlation between bmi and insurance cost.
- But the bmi of female gender is lower compared to Male.

c) weight_change_in_last_one_year vs Insurance cost:

- weight_change_in_last_one_year has moderate correlation with the insurance cost. It is more visible in the persons whose variation is 5 and 6.

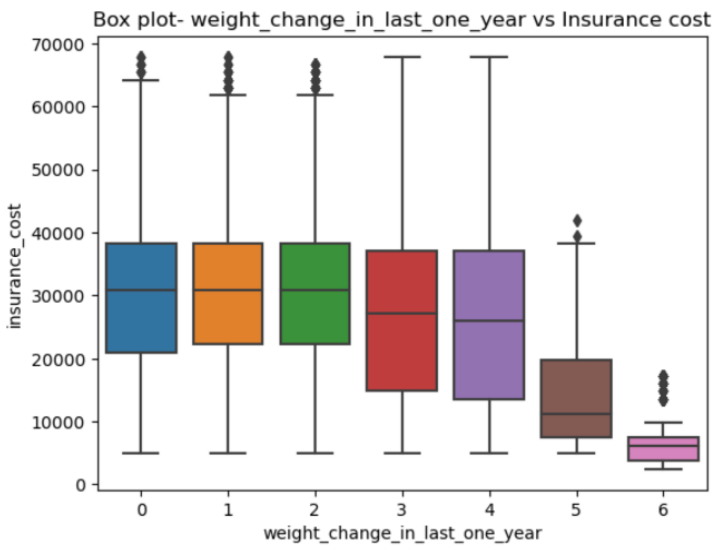


Fig-1.26-Plot of weight_change_in_last_one_year vs Insurance cost

Multivariate analysis:

a)Pair Plot:

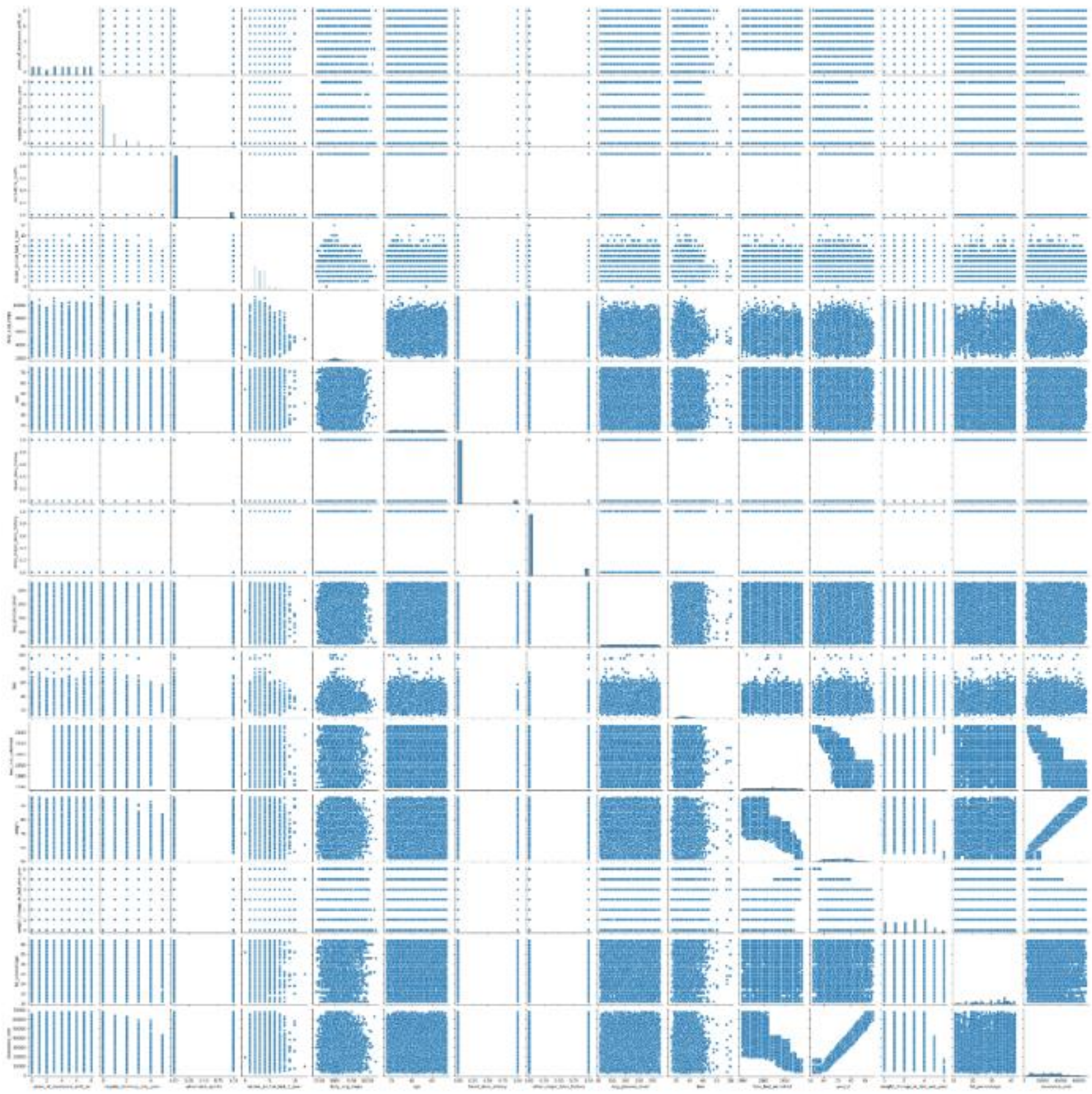


Fig-1.27-Pairplot

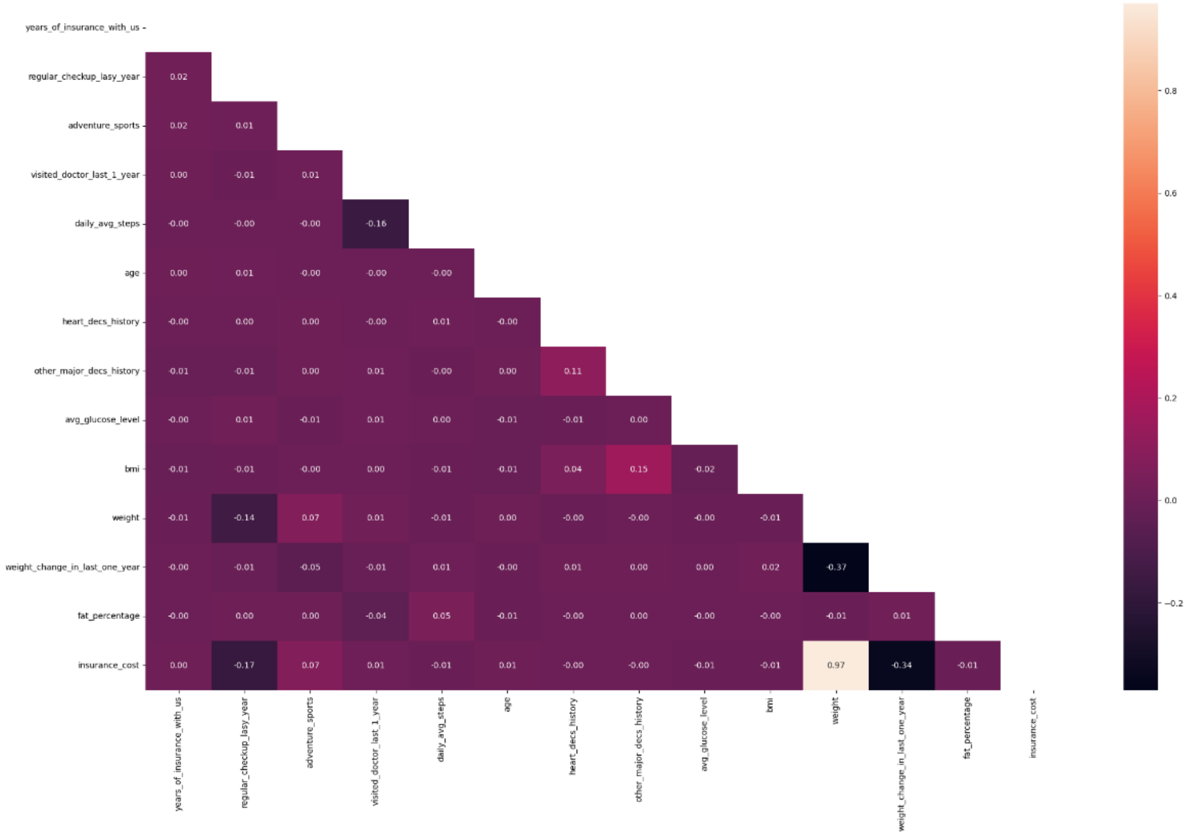


Fig-1.28-Heatmap

- There is strong correlation between weight and insurance_cost.
- There is moderate relationship between insurance_cost and weight_change_in_last_one_year
- High correlation means there's *linear* relationship; it means that there is information in the feature that predicts the target.

Impact of analysis on business:

1. Female customers are less when compared to male customers. So, it indicates there is gender equality in the health insurance sector.

2. Except weight parameter, there is no significant correlation with other features. By this we can say weight is the one of the important parameter in determining the insurance cost.

3. There are some discrepancies in collecting data as there are null values in few variables.

4. Insurance among the salaried class is comparatively less when compared to students and business people. So, the policy can be fine-tuned to suit the salaried class.

Removal of unwanted variables:

- Application_id is removed as it serves no purpose in our prediction.

Null Values:

```
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
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
```

- There are 990 null values in bmi feature and 11,881 null values in the year_last_Admitted variable.
- We remove the year_last_admitted variable as nearly 47.5 percent of the values are missing. Imputing them will only affect the model.
- By using simple imputer, we impute the null values of bmi variable with the median value as it is less sensitive to outliers than the mean.

```
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
```

Duplicates:

- There are no duplicates in the dataset.

<u>Outlier Treatment:</u>

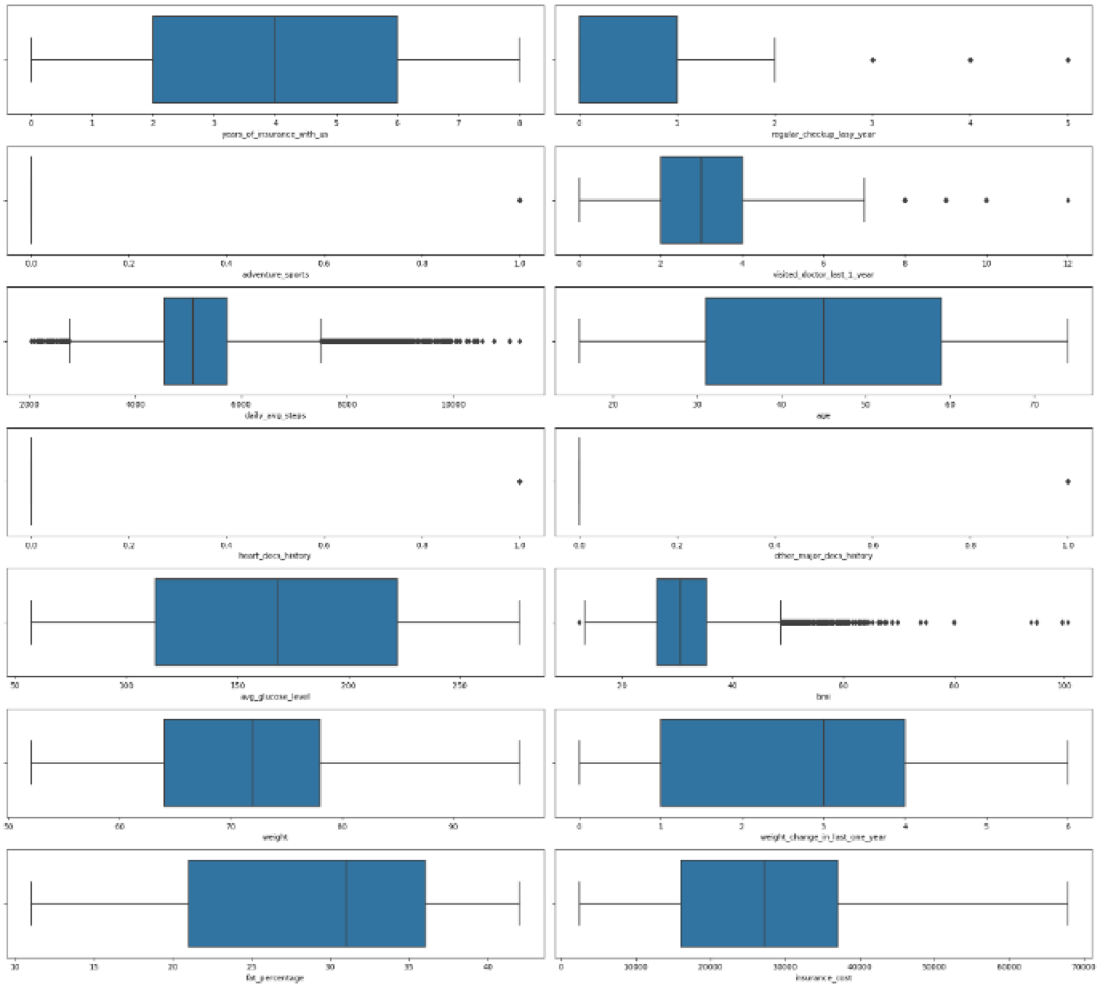


Fig-1.29-Outlier before Treatment

- From this, the variables such as adventure_sports,bmi,daily_avg_steps, other_major decs_history, regular_checkup_lasy_year, visited_doctor_last_1_year, heart_decs_hi story has presence of outliers.
- For binary variables such as adventure_sports, heart_desc_history and other_major_ decs_history, it is not necessary to take outliers.
- To generalize the model and to avoid bias, we do outlier treatment.
- By ***flooring and capping method***, we treat the outliers for the variables bmi and daily avg-steps to make the model more stable.
- For remaining variables, we retain the outliers as it seems like a natural variation in the dataset.

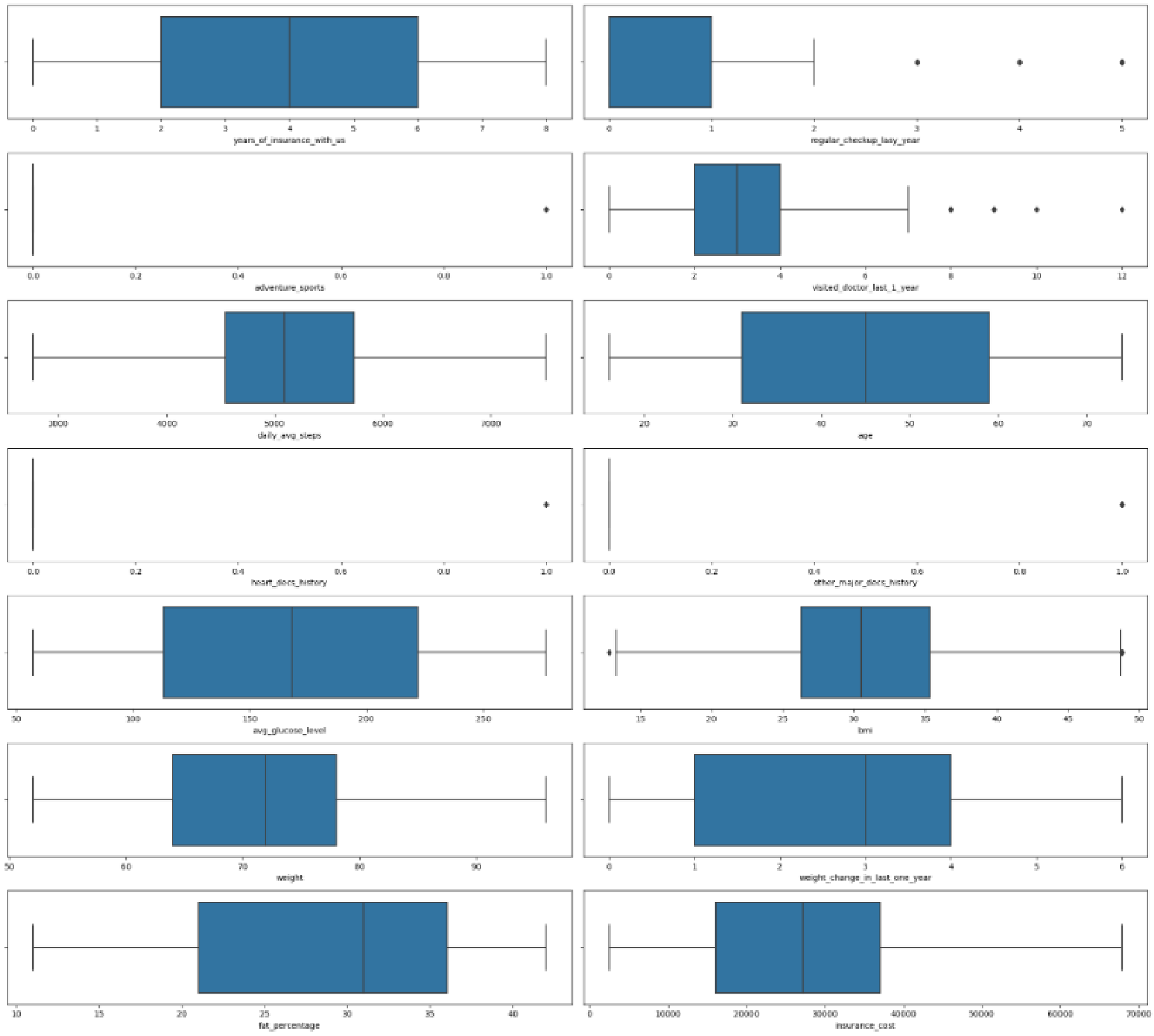


Fig-1.30-Outlier-After treatment

Variable Transformation:

Encoding:

a)One hot encoding:

- For nominal data, we have to do one hot encoding.
- For variables such as 'Gender', 'Occupation', 'smoking_status', 'Location', 'Alcohol' and 'covered_by_any_other_company'

b)Label encoding:

- For ordinal variables such as exercise, cholesterol level we can do label encoding as it will order in rank.

Scaling:

- For clustering and model building process we need to scale the variables. Scaling is the technique to bring the data points closer to each other. As ***clustering is a distance-based algorithm***, scaling is necessary.

Clustering:

- We tried to find cluster by using hierarchical cluster and k-means, but there is no significant pattern involved. Below mentioned elbow plot validates it.

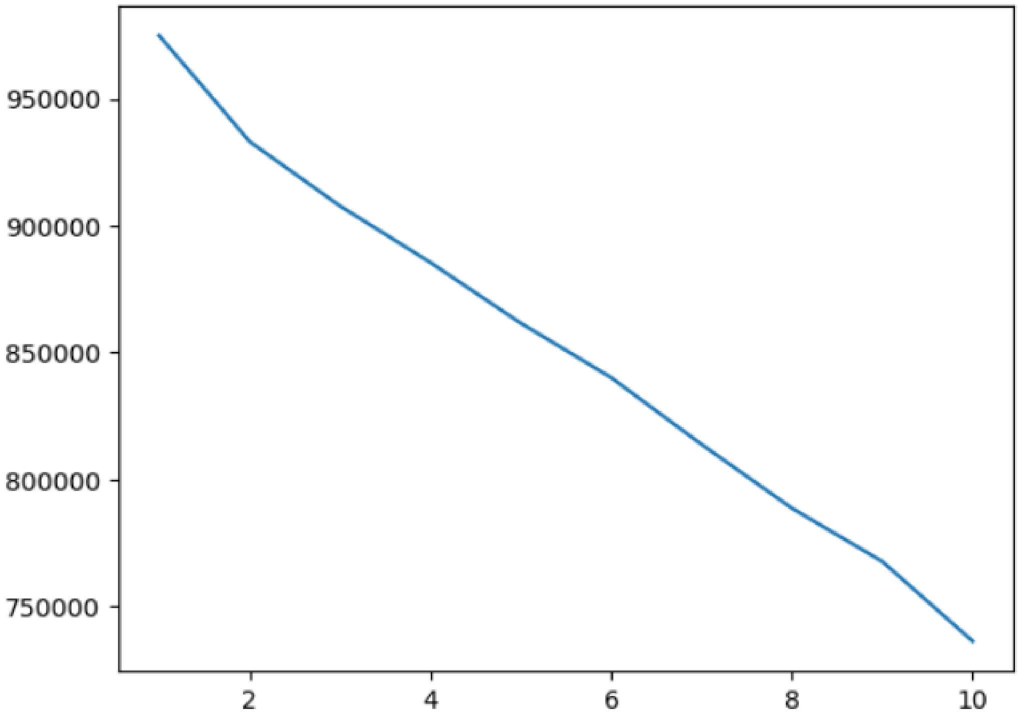


Fig 1.31-Elbow plot

Scaling:

- Depending upon the model we use, we need to scale the variables. Scaling is the technique to bring the data points closer to each other.
- In this dataset, we do scaling because
  - we have different units for different variables. For example, weight in kgs and average number of steps.
  - For gradient descent algorithms such as linear regression, scaling is done for smoother convergence of gradient descent.
- We apply scaling to all the models for better comparison of performance metrics of various models.
- Below let us see the head of few features before and after scaling

Before Scaling:

	0	1	2	3	4
years_of_insurance_with_us	3.0	0.0	1.0	7.0	3.0
regular_checkup_lasy_year	1.0	0.0	0.0	4.0	1.0
adventure_sports	1.0	0.0	0.0	0.0	0.0
visited_doctor_last_1_year	2.0	4.0	4.0	2.0	2.0
cholesterol_level	0.0	1.0	3.0	2.0	1.0
daily_avg_steps	4866.0	6411.0	4509.0	6214.0	4938.0
age	28.0	50.0	68.0	51.0	44.0
heart_decs_history	1.0	0.0	0.0	0.0	0.0
other_major_decs_history	0.0	0.0	0.0	0.0	1.0
avg_glucose_level	97.0	212.0	166.0	109.0	118.0
bmi	31.2	34.2	40.4	22.9	26.5
weight	67.0	58.0	73.0	71.0	74.0
exercise	1.0	1.0	0.0	2.0	0.0
weight_change_in_last_one_year	1.0	3.0	0.0	3.0	0.0
fat_percentage	25.0	27.0	32.0	37.0	34.0
insurance_cost	20978.0	6170.0	28382.0	27148.0	29616.0
Gender_Male	1.0	1.0	0.0	0.0	1.0
Occupation_Salried	1.0	0.0	0.0	0.0	0.0
Occupation_Student	0.0	1.0	0.0	0.0	1.0
smoking_status_formerly smoked	0.0	1.0	1.0	0.0	0.0
smoking_status_never smoked	0.0	0.0	0.0	0.0	1.0
smoking_status_smokes	0.0	0.0	0.0	0.0	0.0
Location_Bangalore	0.0	0.0	0.0	0.0	1.0
Location_Bhubaneswar	0.0	0.0	0.0	0.0	0.0
Location_Chennai	1.0	0.0	0.0	1.0	0.0

After Scaling:

	0	1	2	3	4
years_of_insurance_with_us	-0.417807	-1.568750	-1.185102	1.116783	-0.417807
regular_checkup_lasy_year	0.188690	-0.645043	-0.645043	2.689890	0.188690
adventure_sports	3.352150	-0.298316	-0.298316	-0.298316	-0.298316
visited_doctor_last_1_year	-0.967205	0.784661	0.784661	-0.967205	-0.967205
cholesterol_level	-1.002742	-0.210186	1.374926	0.582370	-0.210186
daily_avg_steps	-0.333160	1.260326	-0.701364	1.057144	-0.258901
age	-1.050360	0.315492	1.433007	0.377576	-0.057013
heart_decs_history	4.159520	-0.240412	-0.240412	-0.240412	-0.240412
other_major_decs_history	-0.329915	-0.329915	-0.329915	-0.329915	3.031081
avg_glucose_level	-1.124370	0.708929	-0.024391	-0.933069	-0.789594
bmi	0.002231	0.422682	1.291613	-1.161015	-0.656474
weight	-0.494422	-1.459569	0.149010	-0.065467	0.256249
exercise	0.008326	0.008326	-1.545002	1.561654	-1.545002
weight_change_in_last_one_year	-0.898041	0.285180	-1.489652	0.285180	-1.489652
fat_percentage	-0.441634	-0.209944	0.369282	0.948508	0.600972
Gender_Male	0.722737	0.722737	-1.383630	-1.383630	0.722737
Occupation_Salried	2.048518	-0.488158	-0.488158	-0.488158	-0.488158
Occupation_Student	-0.828045	1.207664	-0.828045	-0.828045	1.207664
smoking_status_formerly smoked	-0.457628	2.185179	2.185179	-0.457628	-0.457628
smoking_status_never smoked	-0.766290	-0.766290	-0.766290	-0.766290	1.304988
smoking_status_smokes	-0.427766	-0.427766	-0.427766	-0.427766	-0.427766
Location_Bangalore	-0.273677	-0.273677	-0.273677	-0.273677	3.653946
Location_Bhubaneswar	-0.270454	-0.270454	-0.270454	-0.270454	-0.270454

- Now we are going to build different models with our pre-processed dataset.
- We divide the data into Train and test data in the ratio of 75:25.
- X_train and X_test is our independent variable and y_train and y_test is our dependent variable.

Linear Regression Model:

- By using stats model, we build the linear regression model using training dataset.
- We use ***Variable inflation factor*** to detect the ***multicollinearity*** among the predictor variables. When multicollinearity is present, the estimated regression coefficients may become large and unpredictable, leading to unreliable inferences about the effects of the predictor variables on the response variable.
- Generally, VIF above 5 is considered to have multicollinearity.
- Since we have encoded all our categorical variables, we perform VIF for all our variables.

	variables	VIF
36	Alcohol_Rare	2.766929
35	Alcohol_No	2.766611
21	Location_Bangalore	1.903205
26	Location_Jaipur	1.882151
22	Location_Bhubaneswar	1.881714
30	Location_Mangalore	1.878420
24	Location_Delhi	1.874200
25	Location_Guwahati	1.866678
23	Location_Chennai	1.865273
31	Location_Mumbai	1.862933
27	Location_Kanpur	1.862721
32	Location_Nagpur	1.861200
29	Location_Lucknow	1.849182
33	Location_Pune	1.842738
28	Location_Kolkata	1.842730
34	Location_Surat	1.826210
17	Occupation_Student	1.654724
19	smoking_status_never smoked	1.561439
18	smoking_status_formerly smoked	1.466319
4	cholesterol_level	1.429552
20	smoking_status_smokes	1.395288
16	Occupation_Salried	1.319694
15	Gender_Male	1.254432
11	weight	1.198894
10	bmi	1.196552

Fig 1.32-VIF

- Since for no feature VIF is greater than 5, we don’t have to drop the features.

Feature Importance:

- Feature with p-value above .05 is considered to be insignificant in prediction. So, 28 variables are removed from the model as they have p-value greater than .05. Finally, we get 9 variables.

```
                              OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 3.560e+04
Date:                Sat, 08 Apr 2023   Prob (F-statistic):               0.00
Time:                        20:56:28   Log-Likelihood:            -1.7888e+05
No. Observations:               18750   AIC:                         3.578e+05
Df Residuals:                   18740   BIC:                         3.579e+05
Df Model:                           9
Covariance Type:            nonrobust
==================================================================================================
                                 coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                   2.714e+04     24.583   1104.028      0.000    2.71e+04    2.72e+04
regular_checkup_lasy_year   -536.8810     24.990    -21.484      0.000    -585.864    -487.898
adventure_sports              57.6985     24.582      2.347      0.019       9.516     105.881
visited_doctor_last_1_year   -55.4400     24.748     -2.240      0.025    -103.949      -6.931
age                           69.1781     24.653      2.806      0.005      20.856     117.501
heart_decs_history            65.7016     24.535      2.678      0.007      17.610     113.793
weight                      1.387e+04     26.869    516.026      0.000    1.38e+04    1.39e+04
weight_change_in_last_one_year  277.8688  26.624     10.437      0.000     225.683     330.055
Location_Delhi                57.0207     24.993      2.281      0.023       8.031     106.010
covered_by_any_other_company_Y  575.0657  24.744     23.240      0.000     526.565     623.567
==============================================================================
Omnibus:                      437.274   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              484.983
Skew:                           0.352   Prob(JB):                    4.87e-106
Kurtosis:                       3.354   Cond. No.                         1.55
==============================================================================

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

Fig 1.33-Model Summary

- ***Coefficients of top 5 important features*** are in the order as:

  1. Weight.

  2.covered_by_any_other_company_Y

  3.Regular_check_last_year- This variable has negative influence on insurance cost.

  4.Weight_change_in_last_one_year.

  5.Age.

- We build the linear regression model with above mentioned 9 variables and we get the performance metrics as:

Train RMSE	Test RMSE	Train R-squared	Test R-squared	Train MAPE	Test MAPE
3365.1198386330175	3383.0473703755742	0.9447383105399607	0.944412070536664	0.15567769400593012	0.15567769400593012

Linear regression-Interpretation:

- Weight variable coefficient is higher thereby influencing the prediction.
- Variable "Regular_checkup_Last_year" influences negatively the insurance cost.
- This model shows accuracy of 94% and MAPE of 15.56 percentage which is a good model.

  We can do regularization techniques to see if there is any improvement in the model.

Regularization:

Regularization techniques such as Lasso and Ridge are used to avoid the risk of overfitting. By this, the cost function, that is residual sum of squares is minimized by adding a penalty.

1)Ridge:

- By using this, we reduce the coefficients by adding a penalty thereby reducing the va riance and adding some bias.
- We build the ridge model by sklearn. Here, alpha denotes the constant that control t he regression strength.
- By trial-and-error method, we applied multiple values for alpha and arrived at 0.1.
- We got the coefficients of the features as below:

```
Ridge model: [-3.16942379e+01 -5.38799805e+02  5.93733696e+01 -5.63852709e+01
  3.90027470e+01 -1.39580257e+01  6.90372600e+01  6.61390362e+01
  8.15530988e+00  1.73865164e+01 -3.12954764e+01  1.38634574e+04
  2.05806785e+01  2.78148195e+02 -2.07515604e+01  1.85781632e+01
  4.10743729e+01  3.61141348e+01  3.87972889e-01  6.99162371e+00
 -2.89311864e+01  8.26801316e+01  6.78211543e+01  1.11729302e+02
  1.34147266e+02  7.20668748e+01  9.77951167e+01  5.65028341e+01
  8.38069490e+01  9.24208098e+01  8.97688013e+01  9.43983007e+01
  8.51764438e+01  6.76270435e+01  7.31292214e+01  5.57315611e+00
  1.84748031e+01  5.85454180e+02]
```

Fig 1.34-Coefficients Ridge

The performance metrics of the ridge model is below

```
        Train RMSE   Test RMSE  Training Score  Test Score  Train MAPE  \
Ridge  3362.672795 3384.023689        0.944819     0.94438    0.151534

       Test MAPE
Ridge   0.155617
```

- The r-squared value on test data is 94%.

- Here, also there is no significant improvement in r2 value.

2)Lasso:

- It is similar to ridge but it goes one step further by reducing the coefficients of least i mportant variable to zero. So, it can also be called as “feature selection” method.

- By setting alpha at 20, we can see the coefficients were reduced to zero.

```
Lasso model: [-4.24302902e+00 -5.19244254e+02  3.87009153e+01 -3.47451369e+01
  1.15799790e-01 -0.00000000e+00  4.84430534e+01  4.63578971e+01
  0.00000000e+00  0.00000000e+00 -2.04648246e+00  1.38380280e+04
  5.36480403e-01  2.46395462e+02 -2.35659675e+00  0.00000000e+00
  1.48366924e+01  0.00000000e+00  0.00000000e+00  0.00000000e+00
 -1.06091602e+01  0.00000000e+00 -0.00000000e+00  1.49647722e+01
  3.53790164e+01 -0.00000000e+00  1.47501420e+00 -2.76748171e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00 -0.00000000e+00 -0.00000000e+00 -0.00000000e+00
  0.00000000e+00  5.55790027e+02]
```

Fig 1.35-Coefficients Lasso

Performance Metrics of Lasso model:

	Train RMSE	Test RMSE	Training Score	Test Score	Train MAPE	Test MAPE
Lasso	3365.285599	3382.094982	0.944733	0.944443	0.15164	0.155489

Here, we get r-squared value as 94.47, so even in lasso there is no significant difference in lasso model.

3. Elastic Net:

- Elastic Net is the middle ground between Ridge Regression and Lasso Regression. Its **regularization term is a mixture of those of Ridge and Lasso** and the mix ratio can be controlled by r.
- Here, we tried various combinations of Lasso: ridge in the ratio of 10:90,20:80..etc. Finally, we found that if the ratio of the lasso increases the model gives better performance.But even then it is along the similar lines with accuracy of 94.48

Decision Tree regressor:

- A regression tree is basically a decision tree that is used for the task of regression which can be used to predict continuous valued outputs instead of discrete outputs.

- We use gini or entropy for classification to split the node, whereas here we use Mean Squared Error(MSE) as our target variable is continuous.

Train RMSE	Test RMSE	Train R-squared	Test R-squared	Train MAPE	Test MAPE
0.000000	4342.417771	1.000000	0.908414	0.000000	0.162508

Decision Tree interpretation:

- Here we can see that Train RMSE is 0 and also there is difference of nearly 10 percent between train and test accuracy which is clear case of ***overfitting***.
- This can be addressed by hyperparameter tuning

Model tuning-Decision Tree:

- ***GridSearchCV*** is a hyperparameter search procedure that is done over a defined grid of hyperparameters. Each one of the hyperparameter combinations is used for training a new model, while a ***cross-validation*** process is executed to measure the performance of the provisional models. Once the process is done, the hyperparameters and the model with the best performance are chosen.
- **Max_depth**-It denotes maximum depth of the tree.
- **min_samples_split-**The minimum number of samples required to split an internal node.
- **min_samples_leaf-**The minimum number of samples required to be at a leaf node.

- By gridsearch, we are able to get our best parameters as
- 'max_depth': 10, 'min_samples_leaf': 30, 'min_samples_split': 15.

- By building the model using the above parameters, we get our metrics as

Train RMSE	Test RMSE	Training R-squared	Test R-squared	Train MAPE	Test MAPE
2887.768221	3138.698245	0.959304	0.952152	0.112740	0.122560

- We can see that we have addressed the issue of over-fitting.
- Our test accuracy has improved to 95 percent and test MAPE to 12 percent which indicates good model.

Random Forest regressor:

A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy a nd control over-fitting.

Train RMSE	Test RMSE	Training R-squared	Test R-squared	Train MAPE	Test MAPE
1170.428365	3104.849404	0.993315	0.953178	0.045621	0.121848

Interpretation-Random Forest:

- This model addressed the issue of overfitting which we saw in decision tree.
- Our train accuracy is 99.3 percent and test accuracy is 95.3 percent. It also has very MAPE in train and test data. Seems like best model so far.

Model Tuning-Random Forest:

- By grid search, we get the best parameters as

```
{'max_depth': 15, 'max_features': 10, 'min_samples_leaf': 3, 'min_samples_split': 30, 'n_estimators': 500}
```

Train RMSE	Test RMSE	Training R-squared	Test R-squared	Train MAPE	Test MAPE
2892.187308	3341.268215	0.959180	0.945777	0.122403	0.142306

- We can see decrease in r-squared value. Due to more computational time we couldn't further fine tune it.

Artificial Neural network:

Train RMSE	Test RMSE	Training R-squared	Test R-squared
2674.456981	3290.199900	0.965091	0.947436

- This model also shows good accuracy of 94.7%. But in ANN it is difficult the interpret the results and also has longer computational time.It makes the ANN model less attractive.

Ensemble Techniques:

An Ensemble method creates multiple models and combines them to solve it.

a) Bagging Regressor:

It works in the following manner:

1. Create multiple datasets from the train dataset by selecting observations with replacements
2. Run a base model on each of the created datasets independently
3. Combine the predictions of all the base models to each the final output

```
                   Train RMSE    Test RMSE  Training Score  Test Score
Bagging regressor  1161.140136  3099.777605        0.993421    0.953331

                   Train MAPE  Test MAPE
Bagging regressor    0.045477   0.121661
```

This model performs better than previous models.

b)Gradient Boosting:

This estimator builds an additive model in a forward stage-wise fashion; it allows for the optimization of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function.

```
                            Train RMSE    Test RMSE  Training Score
gradient_boosting_regressor  2466.0237  3137.640294        0.970323

                            Test Score  Train MAPE  Test MAPE
gradient_boosting_regressor   0.952184    0.095603   0.126131
```

Other methods:

Tried to find out any change in model performance without outlier treatment and without imputation but there is no significant change.

Overall Model Insights:

	Linear Regression	Lasso	Ridge	Decision Tree	Random Forest	Decision Tree (Tuning)	Random Forest (Tuning)	Bagging	Gradient Boosting
Train RMSE	3365.11	3362.65	3365.28	0.000	1170.42	2887.768221	2892.18	1161.14	2466.02

Test RMSE	3383.047	3384.02	3382.09	4342.41	3104.84	3138.66	3341..26	3099.77	3137.64
Train R 2	94%	94.4%	94.47	100%	99.33 %	95.9%	95.91	99.34	97.03
Test R2	94%	94.4%	94.44	90.8%	95.31 %	95.2%	94.57	95.3	95.21
Train MAPE	15.17	15.15	15.16	0	4.5	11.27	12.24	4.54	9.56
Test MAPE	15.56	15.56	15.58	16.2	12.18	12.25	14.2	12.16	12.61

Fig 1.36-Overall Insights

1. From the above table, we can see that Random Forest and bagging model outperforms all other models in all performance metrics.
2. Tuning in random forest model doesn't improve the model much.
3. Overfitting in decision tree model is addressed by hyper parameter tuning and random forest model.
4. Though Gradient boosting also gives similar results to bagging and random forest in test data its accuracy is less in train data.
5. Almost all models performed well with or without hyperparameter tuning.
6. For all our models MAPE is less than 20% so all models are good in general.

Why we choose random forest as best model?

- ***Random Forest and bagging model*** outperforms all other models in all performance metrics.
- In random forest we take random subsets of features for training the individual trees ; in bagging, we offer each tree with the complete set of features.
- So in random forest, the trees are more independent of every other compared to regular bagging, which frequently leads to better predictive performance (due to raised variance-bias trade-offs) and so it is faster than bagging and very important because each tree learns only from a subset of features.
- So, ***we take random forest for prediction***.

Model validation:

- **R-squared** value tells how well the regression line fits the data. It exists between 0 and 1. Closer to 1, model is said to have good accuracy. R-squared value always increases with the addition of variables. To overcome this, we use adjusted-R2.
- **Adjusted R-Squared** takes into account the number of independent variables you employ in your model and can help indicate if a variable is useless or not. The more variables you add to your model without predictive quality the lower your Adjusted R-Squared will be.

- In the linear regresssion model when we do feature selection, we are not able to find any difference between R-Squared and adjusted R-squared.
- **RMSE(Root Mean Squared Error)** is a standard deviation of prediction errors or residuals. It indicates how spread out the data is around the line of best fit. Lower the RMSE value better the model.
- **MAPE (Mean Absolute Percentage Error)** could be a better way to measure your performance of the model and unlike RMSE it is insensitive to the scale in which the variables are measured.
- Less than 20 % MAPE is considered good. Lower the MAPE better the model.
- So, we take MAPE and R-Squared as our primary performance metric.
- Considering these metrics random forest performs well compared to other models.

Business Recommendations:

1. Weight variable alone contributes significantly to the prediction of insurance_cost. It explains more than 90 percent of our target variable.
2. We can find very little correlation of insurance cost with other similarly important varaibles such as age, adventure_sports. This shows the pricing part of the insurance cost is random to some extent.
3. There are some discrepancies in data collection part as we can see many values are missing in variables such as Year_last_admitted and bmi varaibles. Other than that in variable “smoking_status” for 7555 customers it is marked as unknown.
4. By providing higher premium for senior citizens and adventure sportspersons, company can maximise their profits.
5. Penalise customers who has unhealthy habits such as alcohol consumption and smoking in determination of insurance cost.
6. Company can provide offers for customers who maintain their BMI and doing regular exercise activities. This will reduce the payout.
7. We can see nearly 15215 customers didn’t get check up in last one year. Discount can be provided to customers who do regular check up atleast once in a year, this can help in early detection of diseases.
8. Only 34.3 percent of customers are female. To increase insurance penetration, company can offer women oriented policies.
9.Our data contains only urban centres and even then it didn’t cover north eastern cities other than guwahati. So, company can expand its business in rural areas and North eastern states.
10. Insurance among salaried class is less comparatively with students and business. Company can tie-up with employers in providing insurance to salaried class.