Amrita Jena

Capstone Project Final Report

PGP - Data Science and Business Analytics. PGPDSBA

INDEX

LIST OF TABLES -

. LIST OF FIGURES FOR PROBLEM 1

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.

DataDictionary:

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

Problem Understanding

a)Defining problem statement

b) Need of the study/project

c) Understanding business/social opportunity

2. Data Report

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

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

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

3-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)

4-Model Tuning

a. Ensemble modelling, wherever applicable

b. Any other model tuning measures (if applicable)

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

<u>Introduction - What did you wish to achieve while doing the project ?</u>

The saying “Our body is our temple” is a statement of careful consideration. How a temple is kept clean, worshiped and is kept closed to all negative entities we should also treat our bodies same way.

a) Defining problem statement –

With disease burden spiking up, medical expenses are also increasing day by day, it’s important for us to be conscious about preventive healthcare, which includes keeping ourselves fit. Not taking any preventive measure could open path for disease like obesity, diabetes, and high/low Blood pressure making us high prone to critical health illnesses. Not keeping fit don’t just play havoc with our health, they can severely increase our health insurance premium by several thousand

Need of the study/project –

This project will help us understand how health if not taken proper care of can make us pay heavy price for it and how it is important to prioritize our health and take all kind of preventive healthcare measures in our day to day life .

By leading a healthy lifestyle health insurance will no longer be viewed as an important measure to secure oneself against unforeseen illnesses; rather it will become a part of one’s daily health needs.

Unhealthy lifestyle like smoking,drinking,drugs,minimum sleep and junk eating adds feather to critical disease as well as insurance cost. Insurance companies may pay special attention to your lifestyle and profession. All information shared plays a key role in determining your suitability for the coverage and insurance costs.

b) Understanding business/social opportunity.

This project will give us chance to understands benefits of leading healthy lifestyle to prevent us from critical diseases by reducing our insurance cost.

.

2-Data Report-a) Understanding how data was collected in terms of time, frequency and methodology b) Visual inspection of data (rows, columns, descriptive details) c) Understanding of attributes (variable info, renaming if required)

applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level
5000	3	1	1	Salried	2	125 to 150
5001	0	0	0	Student	4	150 to 175
5002	1	0	0	Business	4	200 to 225
5003	7	4	0	Business	2	175 to 200
5004	3	1	0	Student	2	150 to 175
5005	8	0	0	Salried	2	225 to 250
5006	8	0	0	Student	4	125 to 150
5007	1	0	0	Student	4	150 to 175
5008	8	1	0	Salried	4	125 to 150
5009	4	3	0	Salried	3	125 to 150

daily_avg_steps	age	heart_decs_history	...	smoking_status	Year_last_admitted	Location	weight	covered_by_any_other_company	Alcohol	exercise	v
4866	28	1	...	Unknown	NaN	Chennai	67	N	Rare	Moderate	
6411	50	0	...	formerly smoked	NaN	Jaipur	58	N	Rare	Moderate	
4509	68	0	...	formerly smoked	NaN	Jaipur	73	N	Daily	Extreme	
6214	51	0	...	Unknown	NaN	Chennai	71	Y	Rare	No	
4938	44	0	...	never smoked	2004.0	Bangalore	74	N	No	Extreme	
5306	39	0	...	Unknown	2003.0	Bhubaneswar	78	Y	Rare	No	
4676	40	0	...	never smoked	2004.0	Guwahati	81	N	No	Moderate	
7448	46	0	...	smokes	NaN	Chennai	72	N	Rare	Moderate	
5632	45	0	...	smokes	2007.0	Mumbai	67	Y	Rare	No	
4130	38	0	...	formerly smoked	NaN	Nagpur	63	N	Daily	Moderate	

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
3	13	39488
3	16	37020
0	34	29616
1	12	22212
0	12	8638

TABLE-1-DATA

We can see from the above data that there is an ‘applicant_id’ and ’Location’ which is not of a great use ,therefore we can drop those column.

Checking the shape of the data: –

Previously we were having 25000 rows and 24 columns but after dropping applicant_id' column now we have 25000 rows and 22 columns.

Checking the info of the data: –

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   years_of_insurance_with_us      25000 non-null  int64  
 1   regular_checkup_lasy_year       25000 non-null  int64  
 2   adventure_sports                25000 non-null  int64  
 3   Occupation                      25000 non-null  object 
 4   visited_doctor_last_1_year      25000 non-null  int64  
 5   cholesterol_level               25000 non-null  object 
 6   daily_avg_steps                 25000 non-null  int64  
 7   age                             25000 non-null  int64  
 8   heart_decs_history              25000 non-null  int64  
 9   other_major_decs_history        25000 non-null  int64  
 10  Gender                          25000 non-null  object 
 11  avg_glucose_level               25000 non-null  int64  
 12  bmi                             24010 non-null  float64
 13  smoking_status                  25000 non-null  object 
 14  Year_last_admitted              13119 non-null  float64
 15  weight                          25000 non-null  int64  
 16  covered_by_any_other_company    25000 non-null  object 
 17  Alcohol                         25000 non-null  object 
 18  exercise                        25000 non-null  object 
 19  weight_change_in_last_one_year  25000 non-null  int64  
 20  fat_percentage                  25000 non-null  int64  
 21  insurance_cost                  25000 non-null  int64  
dtypes: float64(2), int64(13), object(7)
memory usage: 4.2+ MB
```

TABLE-2- INFO TABLE

There is total 7 object data types,2 float data type and 13 int data type.

'insurance_cost' is our target variable.

This table shows no of categoerical variables.

adventure_sports,other_major_decs_history and heart_decs_history are also categorical variables,hence we converted them into categorical.

	Occupation	cholesterol_level	Gender	smoking_status	covered_by_any_other_company	Alcohol	exercise
0	Salried	125 to 150	Male	Unknown	N	Rare	Moderate
1	Student	150 to 175	Male	formerly smoked	N	Rare	Moderate
2	Business	200 to 225	Female	formerly smoked	N	Daily	Extreme
3	Business	175 to 200	Female	Unknown	Y	Rare	No
4	Student	150 to 175	Male	never smoked	N	No	Extreme

TABLE-3- Categorical variable info table

Unique counts of each categorical variables-

```
ADVENTURE_SPORTS :  2
1      2043
0     22957
Name: adventure_sports, dtype: int64

OCCUPATION :  3
Salried      4811
Business    10020
Student     10169
Name: Occupation, dtype: int64

CHOLESTEROL_LEVEL :  5
225 to 250     2054
175 to 200     2881
200 to 225     2963
125 to 150     8339
150 to 175     8763
Name: cholesterol_level, dtype: int64

HEART_DECS_HISTORY :  2
1      1366
0     23634
Name: heart_decs_history, dtype: int64

OTHER_MAJOR_DECS_HISTORY :  2
1      2454
0     22546
Name: other_major_decs_history, dtype: int64
```

```
GENDER :  2
0     8578
1    16422
Name: Gender, dtype: int64

SMOKING_STATUS :  4
3    3867
1    4329
0    7555
2    9249
Name: smoking_status, dtype: int64
```

```
COVERED_BY_ANY_OTHER_COMPANY :  2
Y     7582
N    17418
Name: covered_by_any_other_company

ALCOHOL :  3
Daily      2707
No         8541
Rare      13752
Name: Alcohol, dtype: int64

EXERCISE :  3
No           5114
Extreme      5248
Moderate    14638
Name: exercise, dtype: int64
```

TABLE-4- Unique count table

<u>Now we will check for duplicates:-</u>

There are no duplicates in the dataset

<u>Descriptive Statistics of the data set: -</u>

	years_of_insurance_with_us	regular_checkup_lasy_year	visited_doctor_last_1_year	daily_avg_steps	age	avg_glucose_level	bmi
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000
mean	4.089040	0.773680	3.104200	5215.889320	44.918320	167.530000	31.357952
std	2.606612	1.199449	1.141663	1053.179748	16.107492	62.729712	7.720963
min	0.000000	0.000000	0.000000	2034.000000	16.000000	57.000000	12.300000
25%	2.000000	0.000000	2.000000	4543.000000	31.000000	113.000000	26.300000
50%	4.000000	0.000000	3.000000	5089.000000	45.000000	168.000000	30.500000
75%	6.000000	1.000000	4.000000	5730.000000	59.000000	222.000000	35.300000
max	8.000000	5.000000	12.000000	11255.000000	74.000000	277.000000	100.600000

Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost
25000.000000	25000.000000	25000.000000	25000.000000	25000.000000
2003.892217	71.610480	2.517960	28.812280	27147.407680
5.491979	9.325183	1.690335	8.632382	14323.691832
1990.000000	52.000000	0.000000	11.000000	2468.000000
2003.000000	64.000000	1.000000	21.000000	16042.000000
2003.892217	72.000000	3.000000	31.000000	27148.000000
2004.000000	78.000000	4.000000	36.000000	37020.000000
2018.000000	96.000000	6.000000	42.000000	67870.000000

TABLE-5- Descriptive Summary Table

The mean age here is 44.4 with 16 as the minimum age and 74 as the maximum age. The mean BMI is 31.3 with 100 as maximum.

The mean glucose here 167.53 with 57 as minimum and 277 as maximum. The mean weight here is 71.61 with 52 kgs as min weight and 96 as maximum weight, maximum weight loss or weight gain an individual has experienced the previous year ids 6kgs.

The highest fat percentage is 42.00 and 28.81 as the mean fat percentage.

3. Exploratory Data Analysis

Univariate \ Bivariate Analysis:-

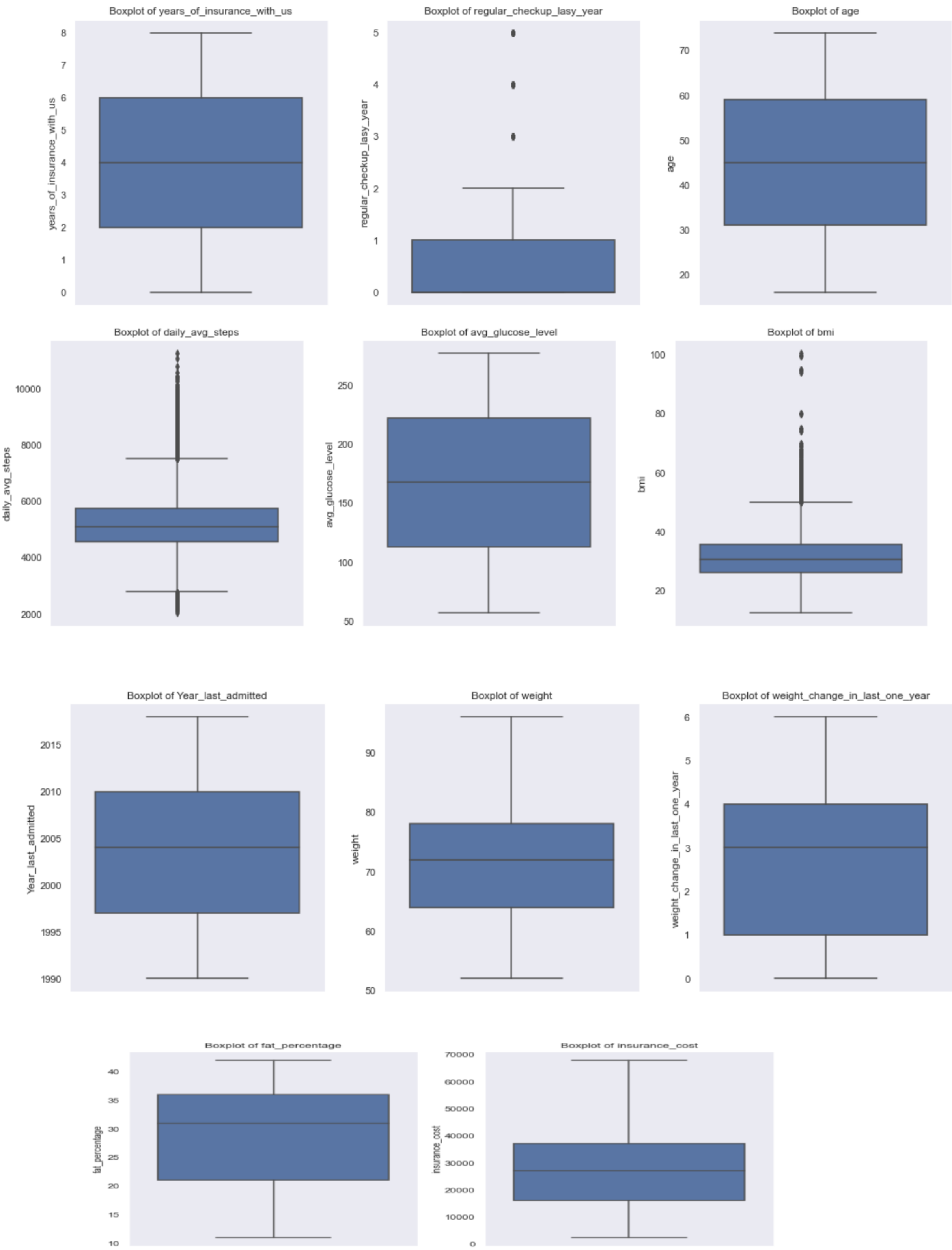
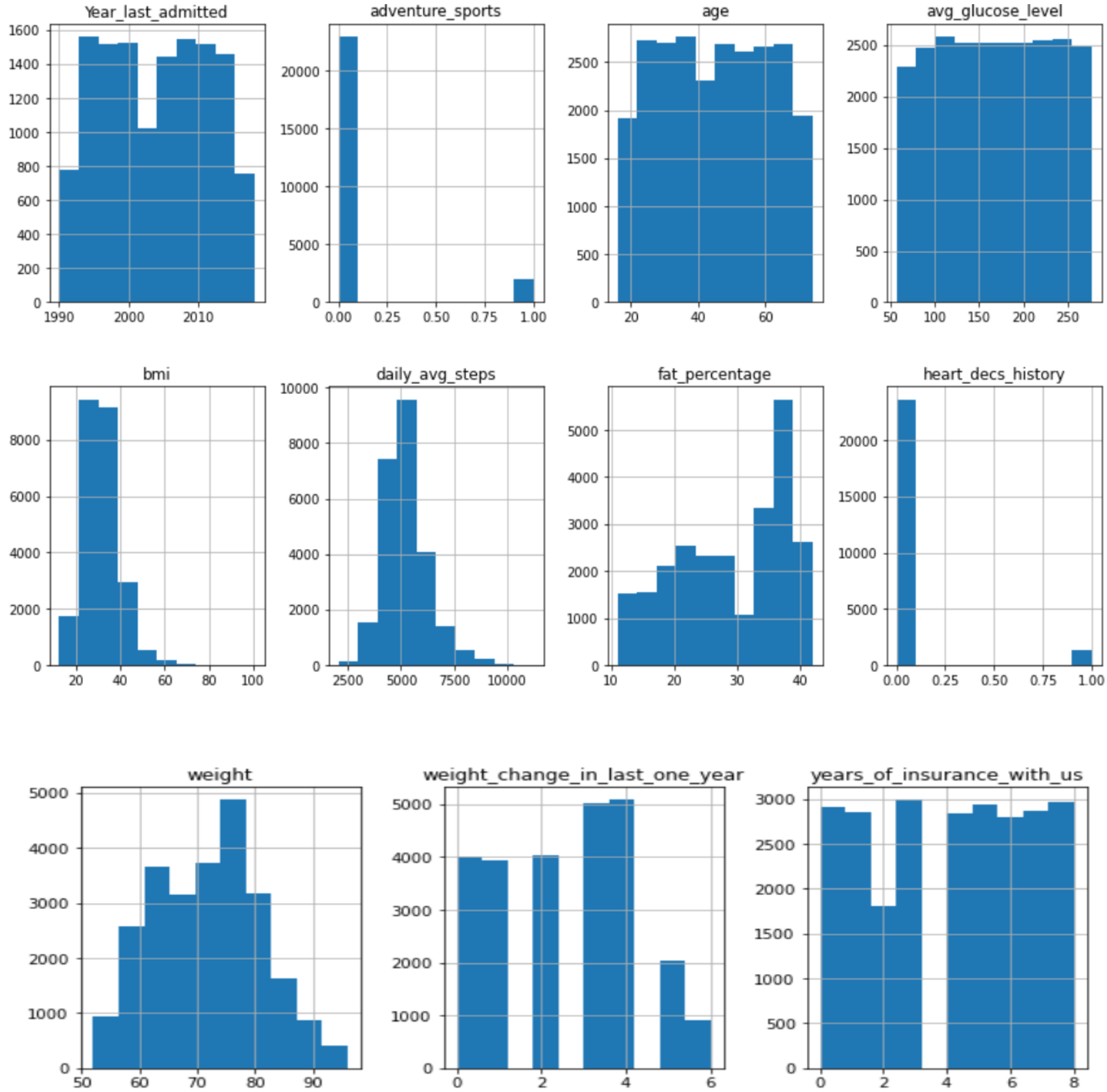


Fig-1-Boxplots shoeing outliers

From the boxplots we could infer that regular_checkup_lasy_year, daily_avg_steps and bmi have outliers.



rs the
ds
neans that

Fig-2-Histogram

```
years_of_insurance_with_us       -0.075217
regular_checkup_lasy_year         1.610907
adventure_sports                  3.054017
visited_doctor_last_1_year        0.978456
daily_avg_steps                   0.908867
age                               0.013860
heart_decs_history                3.919343
other_major_decs_history          2.701327
avg_glucose_level                -0.006389
bmi                               1.090847
Year_last_admitted                0.018679
weight                            0.109077
weight_change_in_last_one_year    0.068026
fat_percentage                   -0.363262
insurance_cost                    0.331650
dtype: float64
```

Table-6-Skewness

From above we could say that data is not highly skewed. we use skewness to understand which data set is normally distributed and which is not. If the skewness =0, It is said to be normally distributed, if it is >0 it is left skewed and if it<0 it is right skewed.

The outliers were removed after treating:-

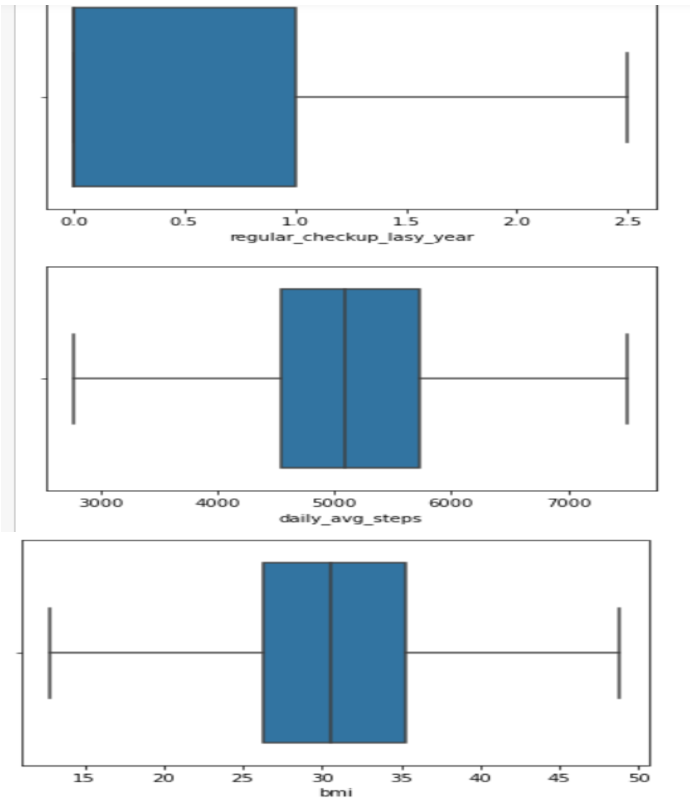
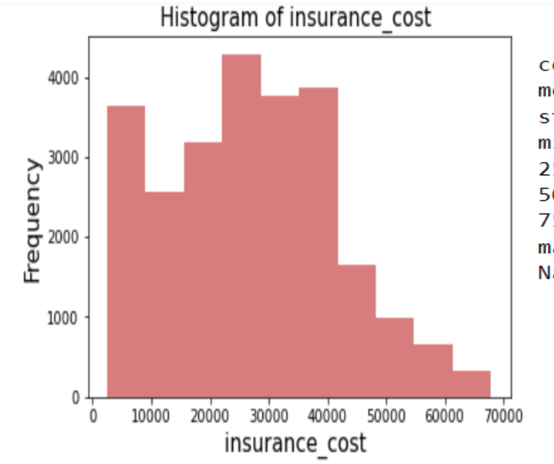


Fig-3-Outlier treated Boxplot

TARGET VARIABLE-




```
count     25000.000000
mean      27147.407680
std       14323.691832
min        2468.000000
25%       16042.000000
50%       27148.000000
75%       37020.000000
max       67870.000000
Name: insurance_cost, dtype: float64
```

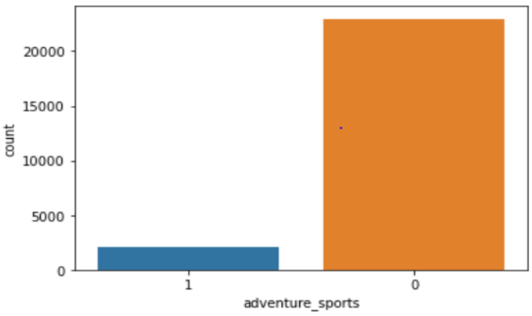
Fig-4-Target Variable Histogram

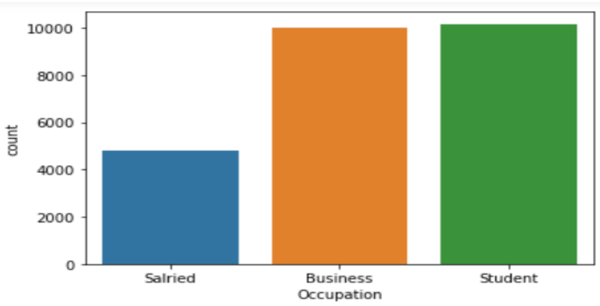
SKEWNESS=0.3316500625115993- The target variable is mostly left skewed with mean cost of 27147.40 rupees with 2468 minimum cost to 67870.00 as maximum cost.

As age was having large number of variables, I grouped age into "age_group" for easy analysis –
Youth (15-24 years)
Adults (25-64 years)
Elderly (65 years and over)

Adult	17992
Elderly	3725
youth	3283

Categorical Variables-


20000
15000
10000
5000
0
count
1
0
adventure_sports


10000
8000
6000
4000
2000
0
count
Salried
Business
Occupation
Student

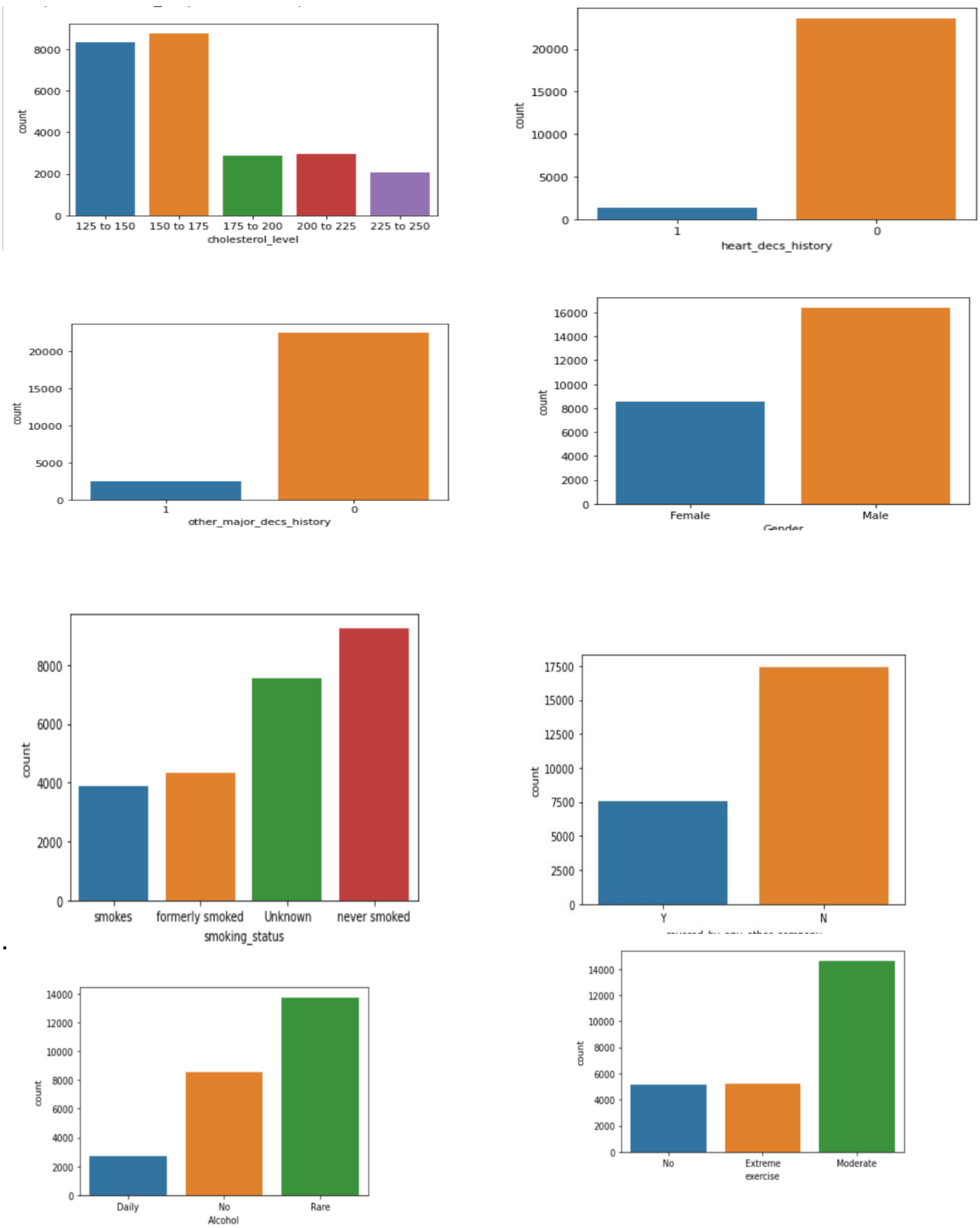


Fig-5-Barplot of Categorical Variable analysis

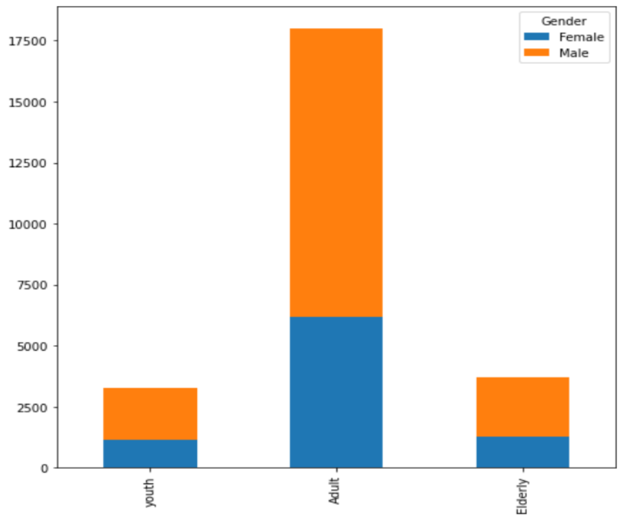
Adventurous sport is not much popular.

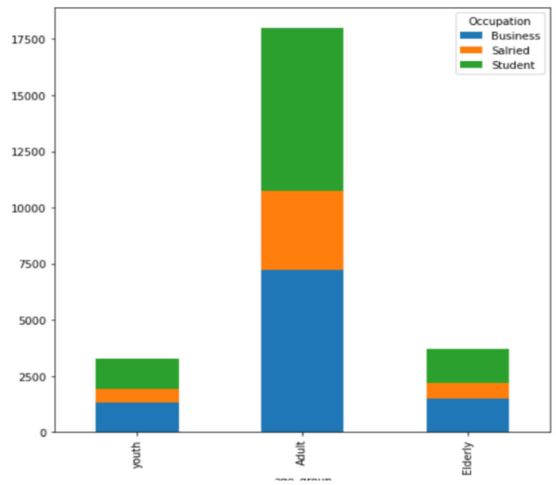
Business and Student occupation is more than salaries occupation.

Maximum insurance holders have normal cholesterol level.

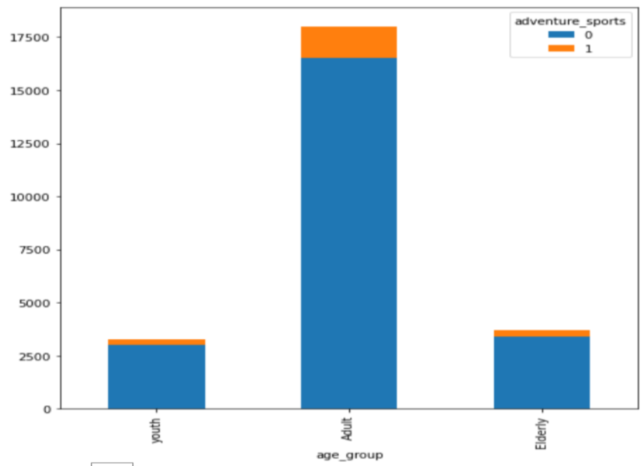
History of any other disease and heart disease is on lower side.

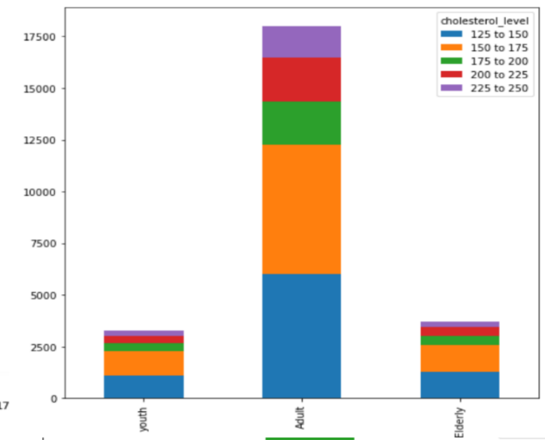
Male population are maximum insurance holders than females. Smokers and daily consumption of alcohol is very low, which is good. Maximum applicants follow a moderate exercise routine.

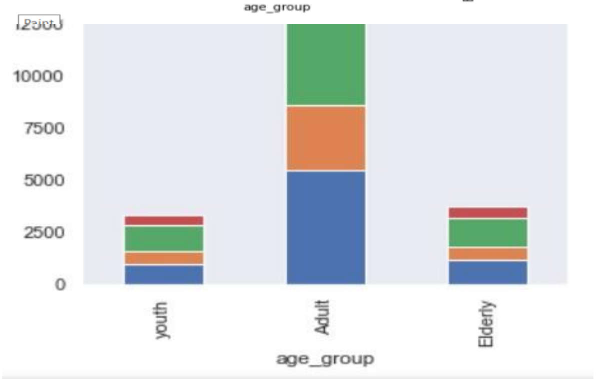


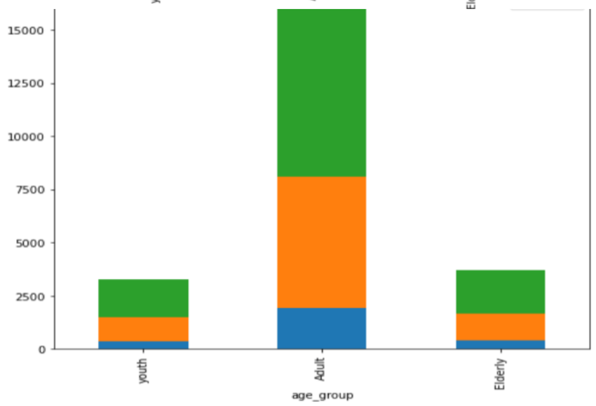


<matplotlib.axes._subplots.AxesSubplot at 0x1fc4e81e348>









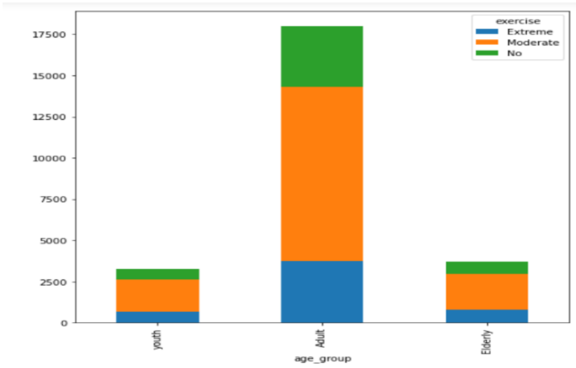


Fig-6-Stackbar of Age analysis with other variables

Maximum are male in adult age group . There are more business holders and students in adult age group.

Most of the variables shows healthy habits with all age group.

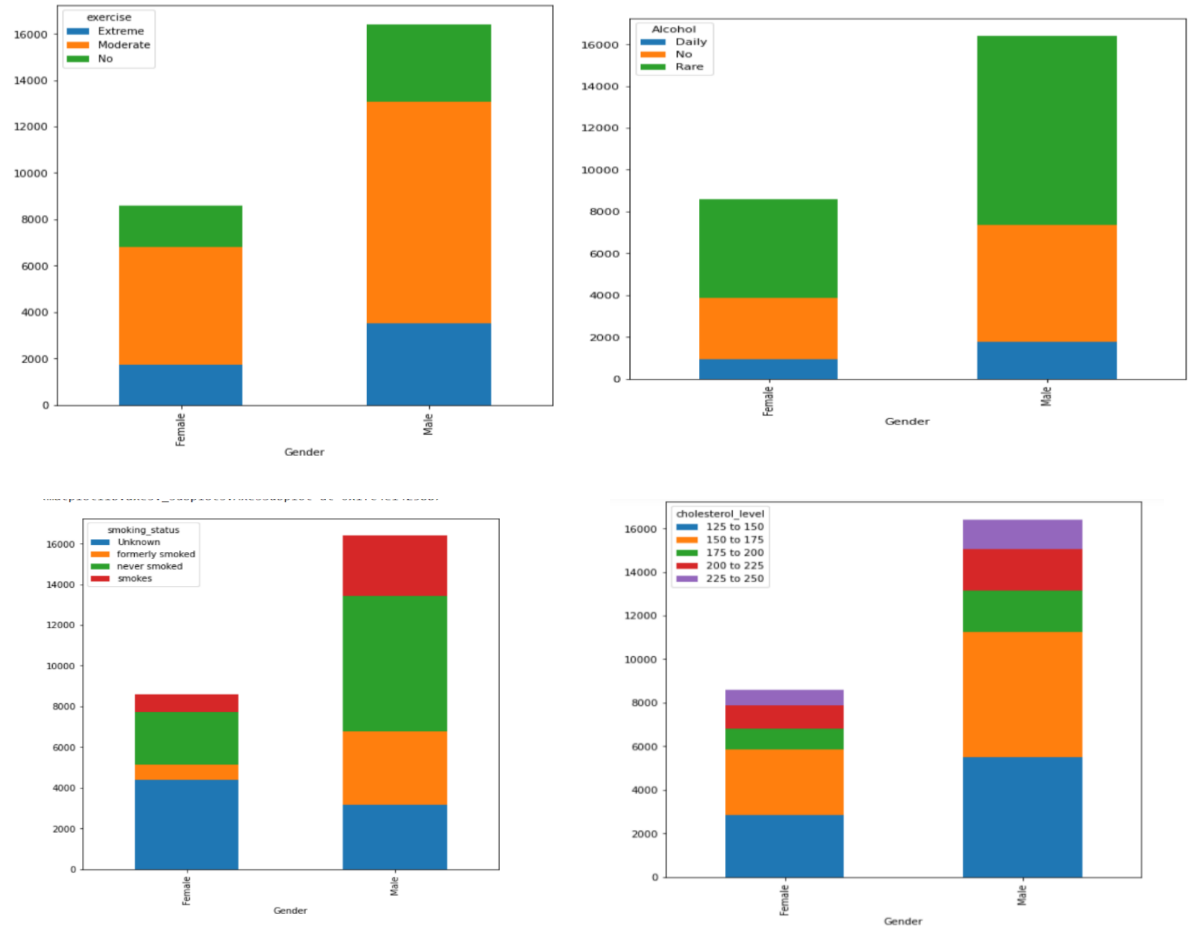


Fig-7-Stackbar of Gender analysis with other variables

As male population is on higher side with no extreme unhealthy habits

MULTIVARIATE ANALYSIS-

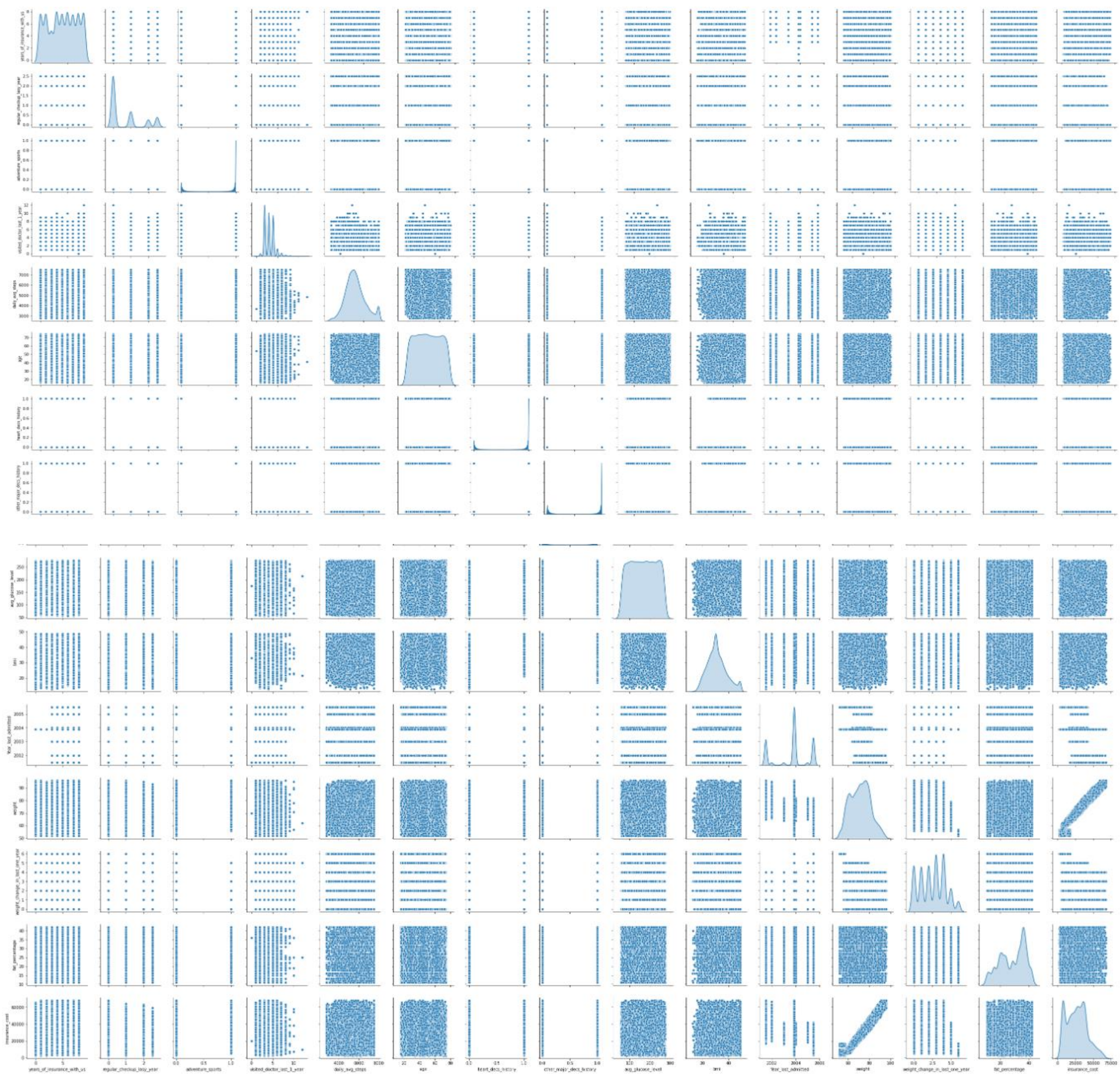
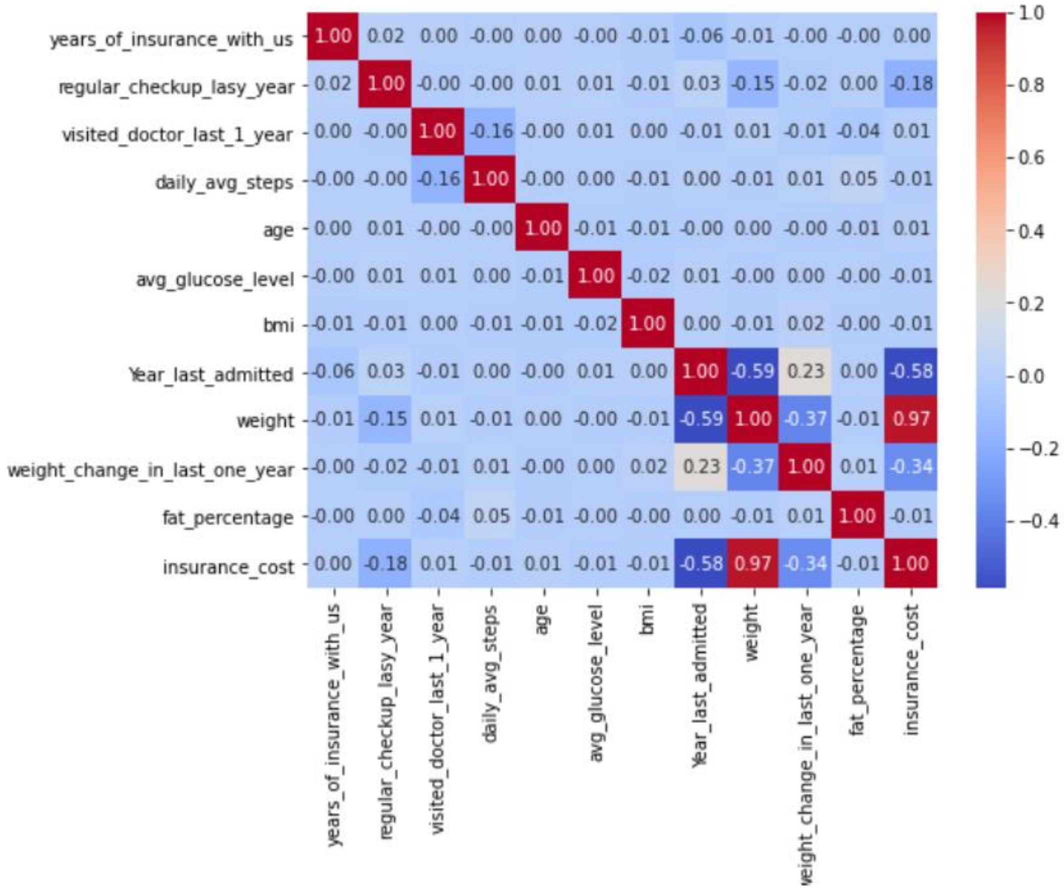


Fig-8-Pairplot



		correlation
insurance_cost	**weight**	0.970357
weight	**Year_last_admitted**	0.585724
insurance_cost	**Year_last_admitted**	0.584723
weight	**weight_change_in_last_one_year**	0.370670
weight_change_in_last_one_year	**insurance_cost**	0.342710

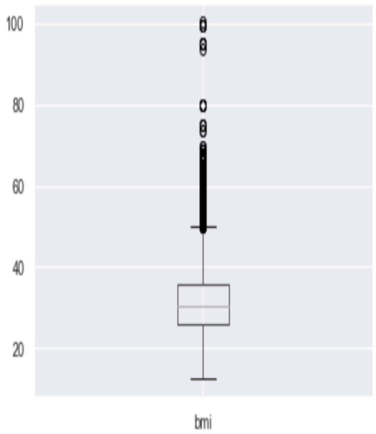
Fig-9-Heatmap

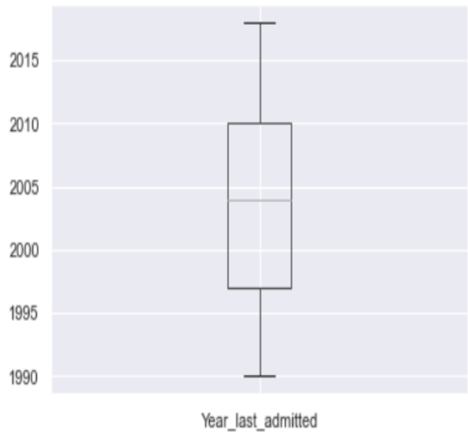
From the heatmap and pair plot the presence of no multicollinearity is visible. Except insurance cost and weight we no strong correlation amongst the variable is observed.

Check for the 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
```

BMI AND Year _last _admitted showed 990 and 11881 respectively. Median imputation was applied for BMI as outliers were present and mean imputation was doe for Year _last _admitted as no outliers were seen.






```
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                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
```

After treating no null values were manifested.

And as **Year_last_admitted** contains 40% of missing value and keeping it in the dataset increases the noise we dropped it .

Table-6-Null Value Treatment

As linear regression analysis does not accept any object data type, all data types were converted into integer.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 22 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   years_of_insurance_with_us      25000 non-null  int64  
 1   regular_checkup_lasy_year       25000 non-null  float64
 2   adventure_sports                25000 non-null  int8   
 3   Occupation                      25000 non-null  int8   
 4   visited_doctor_last_1_year      25000 non-null  int64  
 5   cholesterol_level               25000 non-null  int8   
 6   daily_avg_steps                 25000 non-null  float64
 7   age                             25000 non-null  int64  
 8   heart_decs_history              25000 non-null  int8   
 9   other_major_decs_history        25000 non-null  int8   
 10  Gender                          25000 non-null  int8   
 11  avg_glucose_level               25000 non-null  int64  
 12  bmi                             25000 non-null  float64
 13  smoking_status                  25000 non-null  int8   
 14  weight                          25000 non-null  int64  
 15  covered_by_any_other_company    25000 non-null  int8   
 16  Alcohol                         25000 non-null  int8   
 17  exercise                        25000 non-null  int8   
 18  weight_change_in_last_one_year  25000 non-null  int64  
 19  fat_percentage                  25000 non-null  int64  
 20  insurance_cost                  25000 non-null  int64  
 21  age_group                       25000 non-null  int8   
dtypes: float64(3), int64(8), int8(11)
memory usage: 2.4 MB
```

Table -7-Coversion of object variables into integer

One-Hot-Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.

This is required to do since the machine learning algorithms only works on the numerical data. That is why there is a need to convert the categorical column into numerical one.

get_dummies is the method which creates dummy variable for each categorical variable.

It is considered a good practice to set parameter drop_first as True whenever get dummies is used. It reduces the chances of multicollinearity which will be covered in coming courses and the number of features are also less as compared to drop_first=False.

Train/Test is a method to measure the accuracy of your model.

It is called Train/Test because we split the data set into two sets: a training set and a testing set.80% for training, and 20% for testing or we can divide our data into 60:40,70:30 as well depending on the dataset demand. We train the model using the training set and test the model using the testing set. Train the model means create the model. Test the model means test the accuracy of the model.

Checking the dimensions of the training and test data

```
X_train (17500, 21)
X_test (7500, 21)
y_train (17500, 1)
y_test (7500, 1)
```

```
X.shape
(25000, 21)
y.shape
(25000, 1)
```

We have 25000 rows and 21 columns in train data and 25000 rows and 1 column in test data.

Normalizing and Scaling

Often the variables of the data set are of different scales i.e., one variable is in millions and other in only 100. For e.g., in our data set **avg_glucose_level** is having values in hundreds and **age** in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.

Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data pre-processing while using machine learning algorithms.

In this method, we convert variables with different scales of measurements into a single scale. StandardScaler normalizes the data using the formula (x-mean)/standard deviation.

Z scores also used to address the problem of different scales.

I have applied both for the numerical variables.

Building A Linear Regression Model-

As objective of our project is to build a model, using data that provide the optimum insurance cost for an individual so to produce meaningful continuous output we will opt for regression analysis.

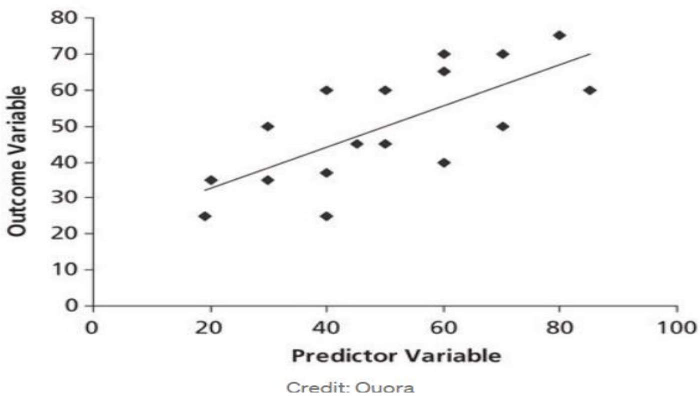
Regression analysis is a statistical method or supervised learning technique that helps us to understand the relationship between dependent and one or more independent variables**.**

In Machine Learning Linear Regression (LR) means finding the best fitting line that explains the variability between the dependent and independent features very well or we can say it describes the linear relationship between independent and dependent features, and in linear regression, the algorithm predicts the continuous features(e.g. Age, Price), rather than deal with the categorical features (e.g. cat, dog)

Linear equation is based on the equation given below-

$$Y = b_0 + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

where y is the dependent variable (target value), x1, x2, … xn the independent variable (predictors), b0 the intercept, b1, b2, ... bn the coefficients and n the number of observations



(image from Quora)

In the picture, you can see a linear relationship. That is, if one independent variable increases or decreases, the dependent variable will also increase or decrease.

<u>Evaluation Metrics for Your Regression Model:</u>

1**- Mean Absolute Error (MAE)**

MAE is a very simple metric which calculates the absolute difference between actual and predicted values.

2-Mean Squared Error (MSE)-

MSE is a most used and very simple metric with a little bit of change in mean absolute error. Mean squared error states that finding the squared difference between actual and predicted value.

3-Root Mean Squared Error(RMSE)-

As RMSE is clear by the name itself, that it is a simple square root of mean squared error

4-R Squared (R2)-

R2 score is a metric that tells the performance of your model, not the loss in an absolute sense that how many wells did your model perform.

In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

So, with help of R squared we have a baseline model to compare a model which none of the other metrics provides. The same we have in classification problems which we call a threshold which is fixed at 0.5. So basically, R2 squared calculates how must regression line is better than a mean line.

MODEL PERFORMANCE

First the LinearRegression function is invoked to find the best fit model on training data.

Then we explore the coefficients for each of the independent attributes

```
The coefficient for years_of_insurance_with_us is -13.393487452772435
The coefficient for regular_checkup_lasy_year is -624.2772509820664
The coefficient for adventure_sports is 129.86549154409138
The coefficient for Occupation is 44.76096252030365
The coefficient for visited_doctor_last_1_year is -35.18716641927201
The coefficient for cholesterol_level is 39.59992963778481
The coefficient for daily_avg_steps is -0.029500264443245907
The coefficient for age is -1.2481182166553348
The coefficient for heart_decs_history is 95.32826701775872
The coefficient for other_major_decs_history is 65.25097732133932
The coefficient for Gender is 38.84196132876468
The coefficient for avg_glucose_level is 0.3586207378092543
The coefficient for bmi is -0.6575834782733294
The coefficient for smoking_status is -4.01838635550217
The coefficient for weight is 1489.1089108243764
The coefficient for covered_by_any_other_company is 1209.1636378498201
The coefficient for Alcohol is 5.656498696088941
The coefficient for exercise is 3.1174217095794408
The coefficient for weight_change_in_last_one_year is 171.965725457744
The coefficient for fat_percentage is -1.0225234354516104
The coefficient for age_group is 155.23519219368364
```

Table-8-coefficient table

•The intercept for our model is -79810.39797063825

•After applying R^ on training data we get value of 0.9447053314178462

•After applying R^ on testing data we get value of 0.9449362616822526

•By applying RMSE on Training data and Testing data we get values 3378.9170619883917 and 3335.746359823853.

Generally, it can be said that RMSE values between 0.2 and 0.5 shows that the model can relatively predict the data accurately. In addition, Adjusted R-squared more than 0.75 is a ve ry good value for showing the accuracy. In some cases, Adjusted R-squared of 0.4 or more is acceptable as well.

R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable.
Instead, we use adjusted R^2 which removes the statistical chance that improves R^2.
Scikit does not provide a facility for adjusted R^2... so we use statsmodel, a library that gives results similar to what we obtain in R language .This library expects the X and Y to be given i n one single data frame.

- We merged X and Y.
- Then we obtained the lml summary

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.164
Model:                            OLS   Adj. R-squared:                  0.163
Method:                 Least Squares   F-statistic:                     201.9
Date:                Sat, 04 Jun 2022   Prob (F-statistic):               0.00
Time:                        20:49:06   Log-Likelihood:            -1.9079e+05
No. Observations:               17500   AIC:                         3.816e+05
Df Residuals:                   17482   BIC:                         3.818e+05
Df Model:                          17
Covariance Type:            nonrobust
==============================================================================================
                                coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
Intercept                  3.596e+04    980.474     36.676      0.000      3.4e+04    3.79e+04
years_of_insurance_with_us -136.0487     39.505     -3.444      0.001    -213.482     -58.615
regular_checkup_lasy_year -2879.9495    108.474    -26.550      0.000   -3092.570   -2667.329
adventure_sports           3102.3008    364.463      8.512      0.000    2387.917    3816.684
Occupation                  -64.3428    129.301     -0.498      0.619    -317.786     189.101
visited_doctor_last_1_year   95.5513     89.171      1.072      0.284     -79.232     270.335
cholesterol_level           -28.5928     92.182     -0.310      0.756    -209.279     152.093
daily_avg_steps              -0.0379      0.105     -0.360      0.719      -0.244       0.169
age                          -0.9641      6.170     -0.156      0.876     -13.058      11.129
heart_decs_history          221.1287    445.029      0.497      0.619    -651.172    1093.430
other_major_decs_history     33.4275    344.177      0.097      0.923    -641.194     708.049
bmi                           3.9904     14.500      0.275      0.783     -24.431      32.412
smoking_status              -60.2924     95.868     -0.629      0.529    -248.203     127.618
covered_by_any_other_company 2877.0316  223.912     12.849      0.000    2438.143    3315.921
Alcohol                     103.7145    147.551      0.703      0.482    -185.500     392.929
exercise                   -225.5871    155.579     -1.450      0.147    -530.538      79.363
weight_change_in_last_one_year -2874.5645 58.856   -48.840      0.000   -2989.929   -2759.200
fat_percentage               -5.7467     11.612     -0.495      0.621     -28.507      17.014
```

Table-9-OLS STATS MODEL SUMMARY TABLE

R2 is the coefficient of determination that tells us that how much percentage variation independent variable can be explained by independent variable. Here, **0.1%** variation in Y can be explained by X. The maximum possible value of R2 can be 1, means the larger the R2 value better the regression.

R-squared_adj :
Represents adjusted R^2 (R^2 corrected according to the number of input features) which is here is same as normal R^2

- Let us check the sum of squared errors by predicting value of y for test cases and subtracting from the actual y for the test cases. This could be pertaining to low p value of all dependent variables.

- We applied Underroot of mean_sq_error which is the standard deviation i.e. avg variance between predicted and actual values - 3335.746359823854

- We obtained # Model score - R2 or coeff of determinant (formula R^2=1–RSS / TSS)-

  0.9449362616822526

From the above observation we could see that P values are 0 where ever T stats are on higher sides

ITERATION 2-Scaled data

To check if regression analysis on scaled data gives a better performance I did a second iteration using scaled data.

The independent attributes have different units and scales of measurement

It is always a good practice to scale all the dimensions using z scores or some other method to address the problem of different scales

.

- **Coefficient and intercept after scaling –**

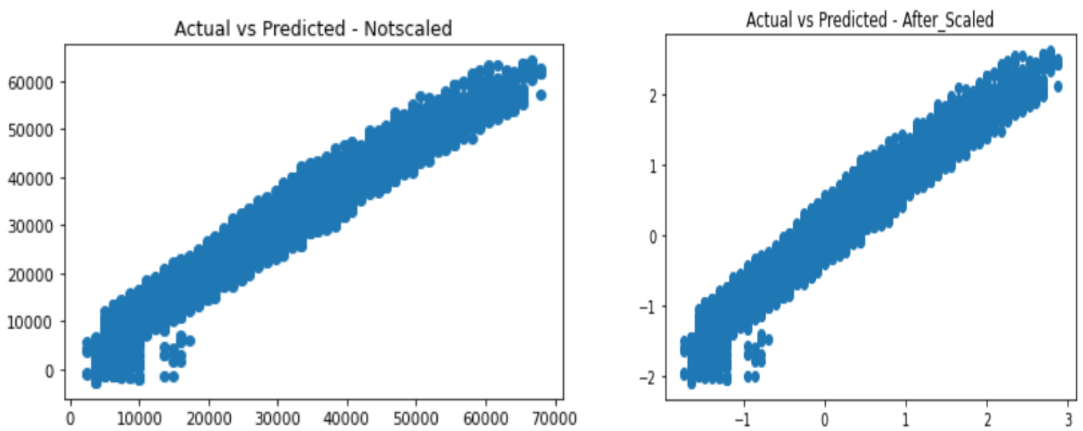
```
The coefficient for years_of_insurance_with_us is -0.0024299085713388904
The coefficient for regular_checkup_lasy_year is -0.03981573702745902
The coefficient for adventure_sports is 0.002469366904307662
The coefficient for Occupation is 0.002800188525640823
The coefficient for visited_doctor_last_1_year is -0.0027796536595197565
The coefficient for cholesterol_level is 0.0034711235096384317
The coefficient for daily_avg_steps is -0.001984846236783699
The coefficient for age is -0.0013991760401877195
The coefficient for heart_decs_history is 0.0014919328369526683
The coefficient for other_major_decs_history is 0.0013366694289673085
The coefficient for Gender is 0.0012828181795454848
The coefficient for avg_glucose_level is 0.0015657187146731143
The coefficient for bmi is -0.00032575286536155087
The coefficient for smoking_status is -0.0002995090657443387
The coefficient for weight is 0.9692636698697614
The coefficient for covered_by_any_other_company is 0.038776352036258156
The coefficient for Alcohol is 0.0002679685695008582
The coefficient for exercise is 0.00013942578610909816
The coefficient for weight_change_in_last_one_year is 0.020273895711055104
The coefficient for fat_percentage is -0.0006138432367906821
The coefficient for age_group is 0.005712004932312524
```

Table-10 -Coefficient table after scaling

- The intercept for our model after scaling is **5.944828137605343e-16**
- R^ after scaling is **0.9449532998861595**
- mean_sq_error after scaling is **0.23462033184240585**

We could observe that regression analysis with scaled data is slightly better then the unscaled data performance.

Figure-10-Actual vs predicted linear model (before and after scaling)



We can clearly see that except for the scale we don't see any change in the relationship or model performance.

Checking for other regressor models for improvement:

In the graph, we can see a that only weight manifests linear relationship with insurance cost while the other shows non linear relationship. Hence along with Linear Regression analysis we will also do Decision Tree, Random Forest and Neural Network regressor analysis

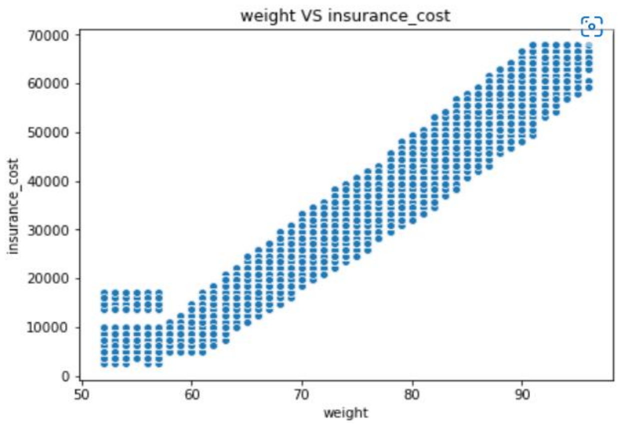


Fig-11-Linear relationship-weight and Insurance Cost

I used Decision tree regressor, Random Forest regressor and ANN Regressor for building regression models. XGBOOST also could have been applied but as its tuning tales a longer time I excluded it.

I used unscaled data for Linear Regression, Decision Tree and Random Forest whereas we used scaled data for ANN model alone

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3378.917062	3335.746360	0.944705	0.944936
Decision Tree Regressor	0.000000	4329.704905	1.000000	0.907232
Random Forest Regressor	1167.993129	3134.142080	0.993393	0.951391
ANN Regressor	7870.203370	7872.391472	0.700015	0.693315

Table-11-Summary RMSE AND RSQUARE BEFORE MODEL TUNING

On the basis of RMSE scores we could say that Decision Tree and Random Forest have very large gap between Train and Test RMSE. So we will go for a model with low RMSE. Hence we will compare both ANN regressor and Linear Regressor .

After comparing them with RMSE and the Test and Train scores of ANN and Linear Regressor we can go for the ANN model first but linear regression model is good at r square score. Random forest also performs well with unscaled data.

Decision tree is overfitting. Hence, we ca go for model tuning. we can also perform same in Random Forest for better performance.

After model tuning-

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3378.917062	3335.746360	0.944705	0.944936
Decision Tree Regressor	2876.723424	3180.641595	0.959920	0.949938
Random Forest Regressor	3186.000003	3399.249038	0.950839	0.942820
ANN Regressor	7870.203370	7872.391472	0.700015	0.693315

Table-12-Summary RMSE AND RSQUARE AFTER MODEL TUNING

Decision tree model tuning-

Hyperparameters used for DT model tuning are

max_depth:int, default=None.

The maximum depth of the tree. If None, then nodes are expanded until all leaves are pure or until all leaves contain less than min_samples_split samples.

min_samples_split:int -The minimum number of samples required to split an internal node:

If int, then consider min_samples_split as the minimum number.

If float, then min_samples_split is a fraction and ceil(min_samples_split * n_samples) are the minimum number of samples for each split.

min_samples_leaf-

The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least min_samples_leaf training samples in each of the left and right branches. This may have the effect of smoothing the model, especially in regression.

My model tuning resulted in -
{'max_depth': 10, 'min_samples_leaf': 30, 'min_samples_split': 15}
which I applies to get better result

RF-model tuning

n_estimators = number of trees in the forest

max_features = max number of features considered for splitting a node

max_depth = max number of levels in each decision tree min_samples_split = min number of data points placed in a node before the node is split

min_samples_leaf = min number of data points allowed in a leaf node

bootstrap = method for sampling data points (with or without replacement)

My model tuning resulted in

```
GridSearchCV(cv=3, estimator=RandomForestRegressor(random_state=40),
             param_grid={'max_depth': [7, 10], 'max_features': [4, 6],
                         'min_samples_leaf': [5, 10, 20],
                         'min_samples_split': [2, 4, 8],
                         'n_estimators': [10, 20, 30]})
```

After tuning Linear regression and Decision Tree model seems to perform betetr than any other models.

<u>MODEL SELECTION</u>

Model for prediction-

If we compared all the Linear Regression Model and Decision Tree model are doing better with respect to prediction. Random forest performed well with unscaled data.

To select best it would be better to have more data for training, validating and testing.

As of now, Decision Tree model and Linear Regression Model looks to be more balance.

However Random Forest regression model is also not bad but requires further analysis.

Model performance very close to full model.

It is the simplest model with no transformation and with least variable. There is no multicollinearity, as the independent variable in the model are not con-elated among each

6-Data Insights-

- The important feature for Insurance cost of individual from the data set provided is coming out to be weight followed by by- weight_change_in_last_one_year
- Variables like eating habit, sleep cycle and frequent pill popping habit which attributes to complications like renal failure should have been included.
- Rather than daily and rare amount of alcohol and no of cigarettes consumed per day should have been included.
- As BMI includes muscles and bone density. Visceral fat content should have been included as it is more reliable than BMI .
- Applicant with unhealthy lifestyle should also have been included properly so that insurance cost on higher side could also have been studied significantly
- Weight is dominating in insurance cost.
- Would advise to work with more variable and data to get better and stable model.
- Before using this model, full fledge testing is of the model is advised.
- Looking at the heat map, except weight other variables are not playing any role in insurance cost determination.
- **Location** wise there was no much difference, so have been dropped.
- **Year_last_admitted** contains 40% of missing value and keeping it in the dataset increases the noise we dropped it
- High dependency of price on weight also needs to be analysed
- For prediction we will choose Decision Tree model and Linear Regression Model

7-Recommendation –

- As cost of health care continues to rise. We have to learn how to take steps to limit your out-of-pocket health care costs.
- From our analysis it is clear that weight is highly corelated with insurance cost and weight_change_in_last_one_year also shows some correlation with our target variable. Insurance holders should religiously follow weight management therapy to reduce insurance cost.
- Increased number of hospitalizations also adds up to medical expenses hence one can get routine health screenings. These tests can catch health problems early, when they may be more easily treated. And one do not have to pay a huge sum when condition worsens .
- Depending on your health coverage, you may have the choice to see providers who are in-network or out-of-network. You pay less to see providers who are in-network, because they have a contract with your health plan. This means they charge lower rates
- A simple way to save money on health care is to stay healthy. Of course, that is sometimes easier said than done. But staying at a healthy weight, getting regular exercise, and not smoking lowers your risk for health problems. Staying healthy helps you avoid costly tests and treatments for ongoing conditions such as diabetes or heart disease.
- Insurance companies may pay special attention to your lifestyle and profession. All information shared plays a key role in determining your suitability for the coverage and insurance costs.