

**Amrita Jena**

**Capstone Project Final Report**

**PGP - Data Science and Business Analytics. PGPDSBA Online May\_B 2021**

**INDEX**

Problem Statement .................................................................... 3

Data Dictionary and Questions.................................................... 4 -5

**Introduction - …………………………………………………………………………6**

EDA................................................................................................7

Model buiding..............................................................................23-30

Mode selection and conclusion…………………………………………………31

**LIST OF TABLES** -

TABLE 1 ...........................................................................................7

TABLE 2 ...........................................................................................8

Table 3 and Table 4 .........................................................................9

TABLE 5............................................................................................10

TABLE 6............................................................................................13

TABLE 7 ...........................................................................................22

TABLE 8............................................................................................25

Table 9…………………………………………………………………………………………26

Table 10……………………………………………………………………………………….28

Table 11………………………………………………………………………………………29

Table 12……………………………………………………………………………………..30

. LIST OF FIGURES FOR PROBLEM 1

FIG.1 ................................................................................................ 11

FIG.2 ................................................................................................ 12

FIG.3 ................................................................................................ 13

FIG.4 ................................................................................................ 14

Fig.5 ................................................................................................ 16

Fig.6 ................................................................................................ 17

Fig.7 ................................................................................................18

Fig.8 ................................................................................................19

Fig.9 ................................................................................................20

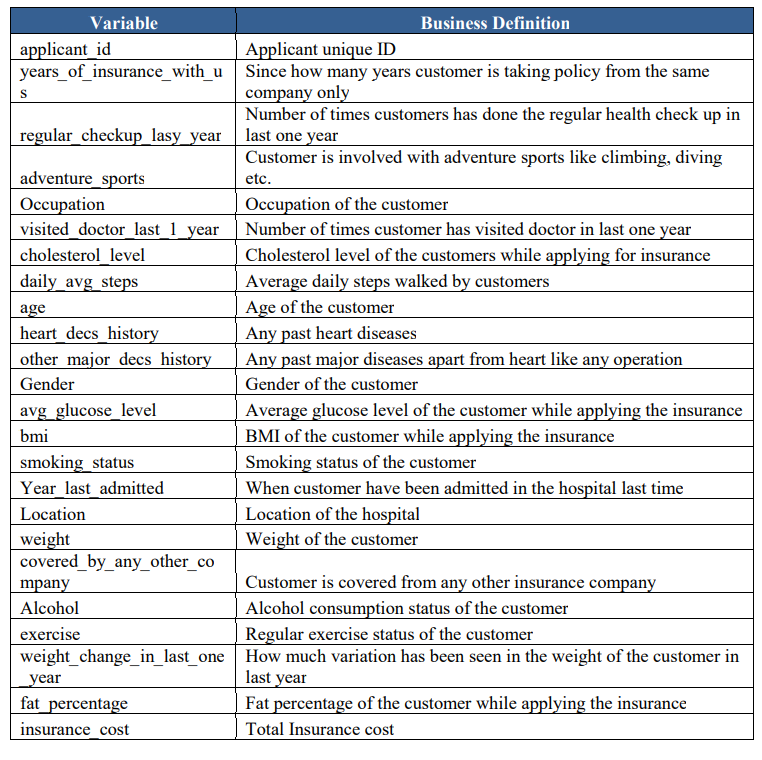
Fig.10...............................................................................................28

Fig.11 ...............................................................................................29

**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:**

###### **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

**Introduction - What did you wish to achieve while doing the project ?**

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.

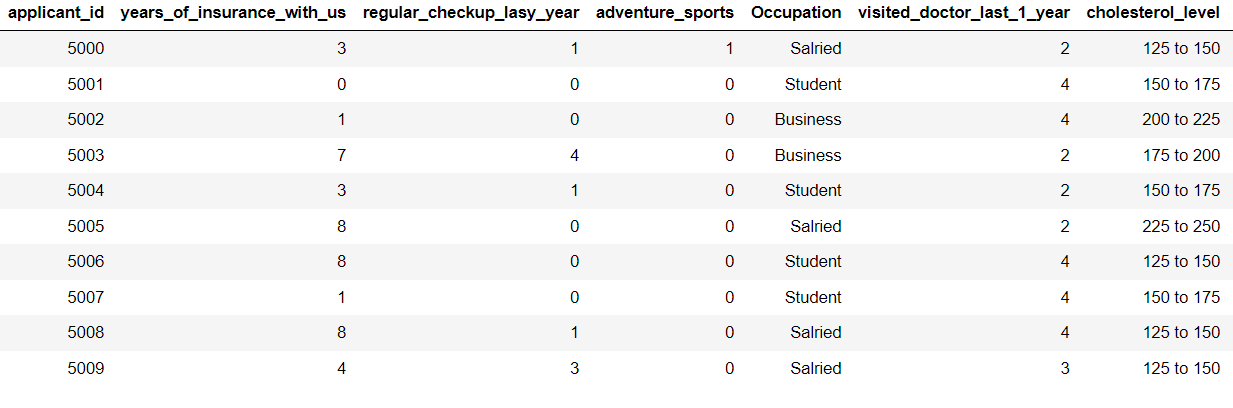
**b) Understanding business/social opportunity.**

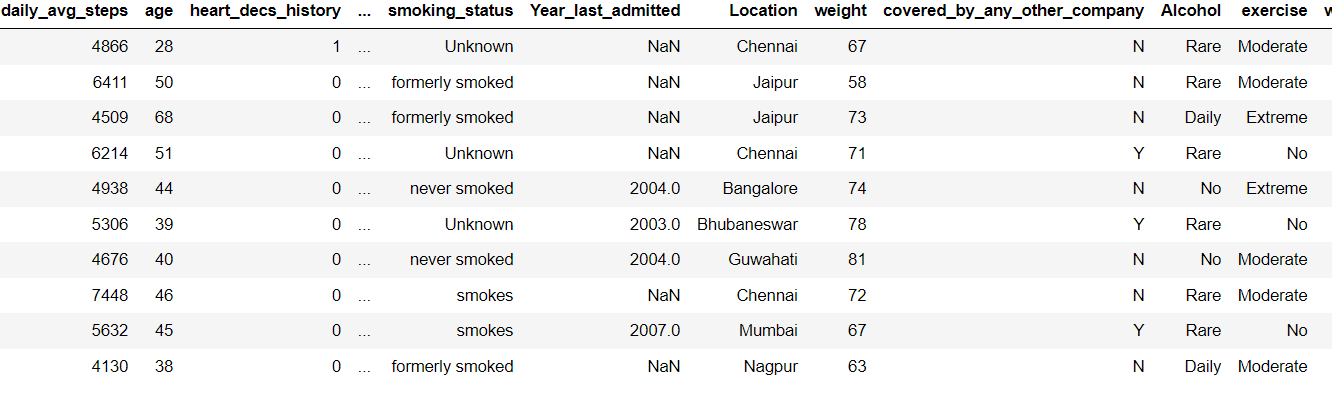
This project will give us chance to understands benefits of leading healthy lifestyle to

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)**





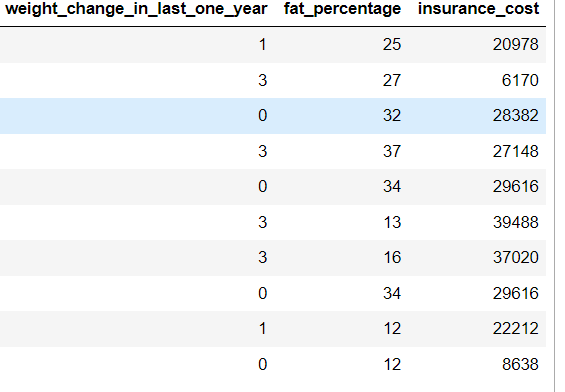


TABLE-1-DATA

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: –**

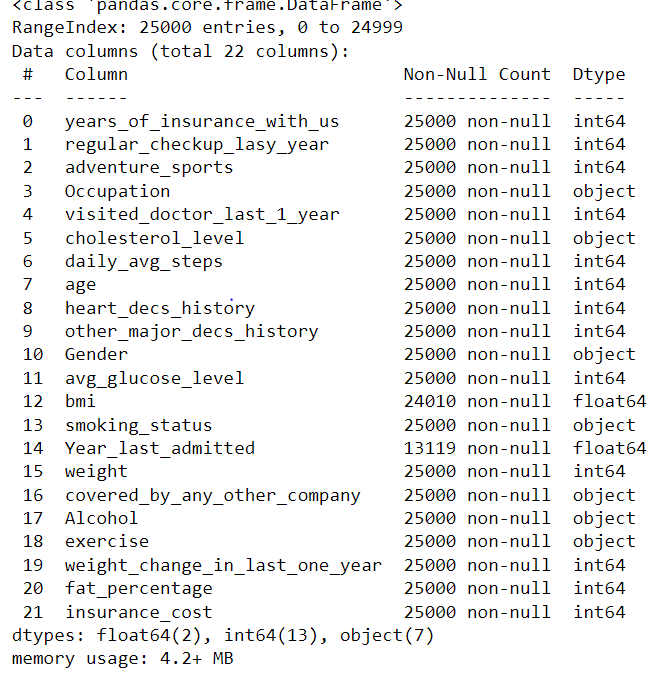


TABLE-2- INFO TABLE

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.

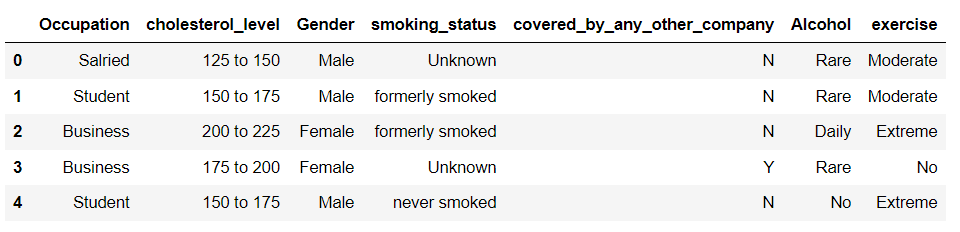
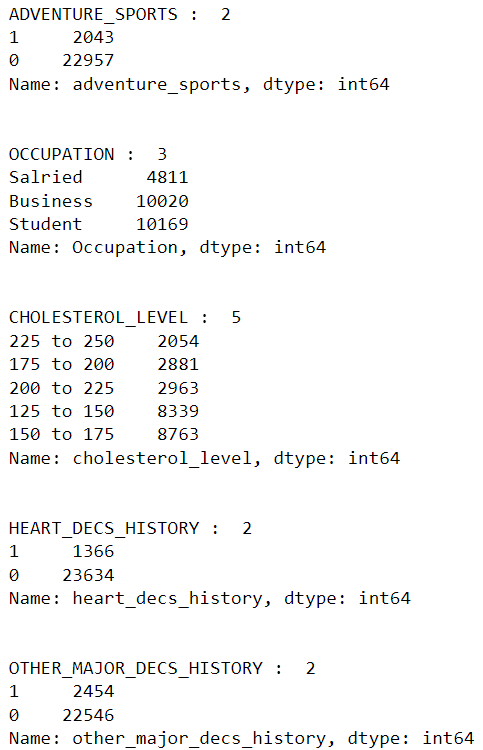
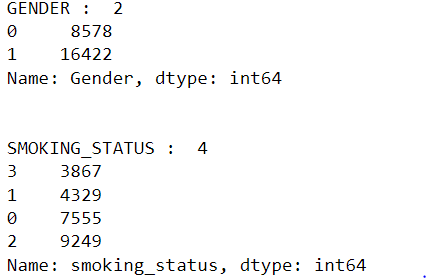


TABLE-3- Categorical variable info table

Unique counts of each categorical variables-



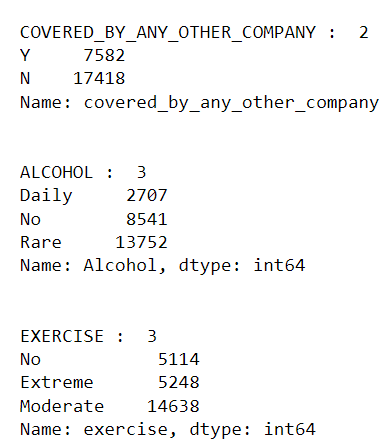
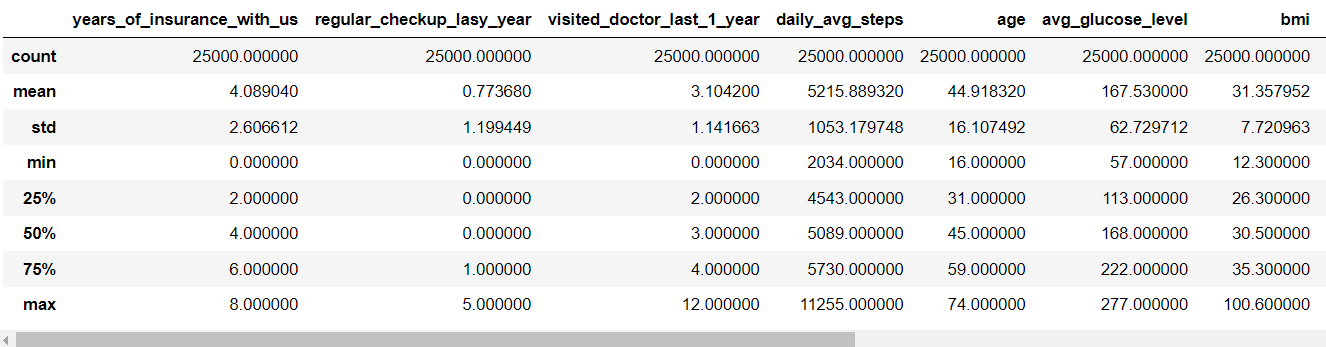


TABLE-4- Unique count table

**Now we will check for duplicates:-**

There are no duplicates in the dataset

**Descriptive Statistics of the data set: -**



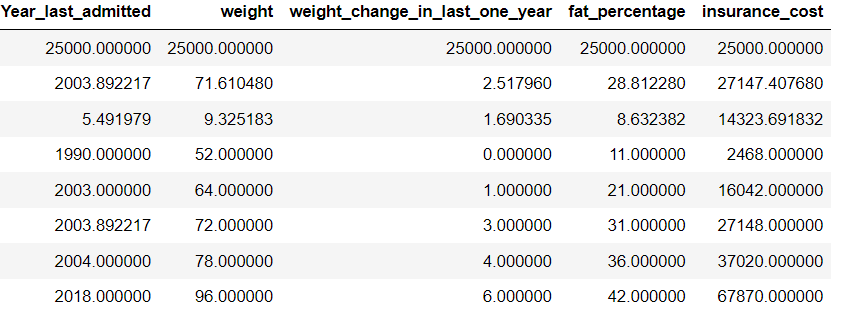


TABLE-5- Descriptive Summary Table

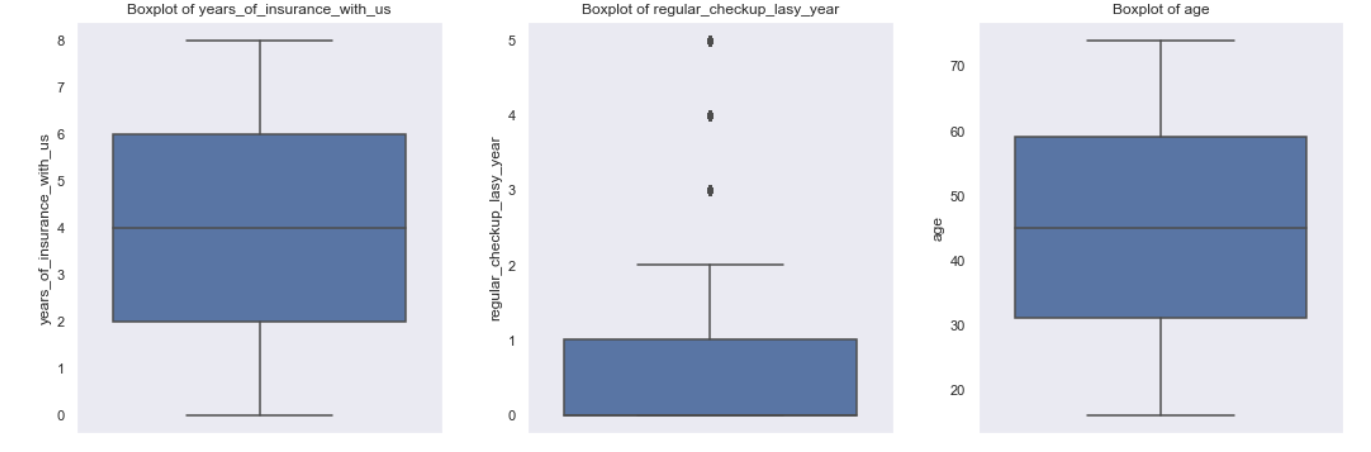
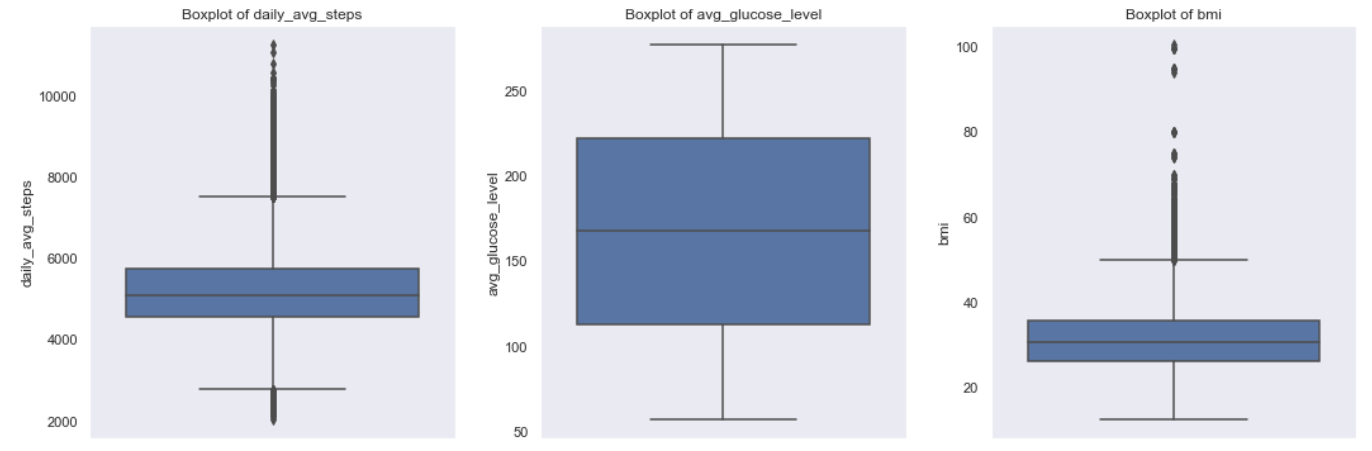
The mean age here is 44.4 with 16 as the minimum age and 74 as the maximum age. The mean BMI is 31.3 with 100 as maximum.

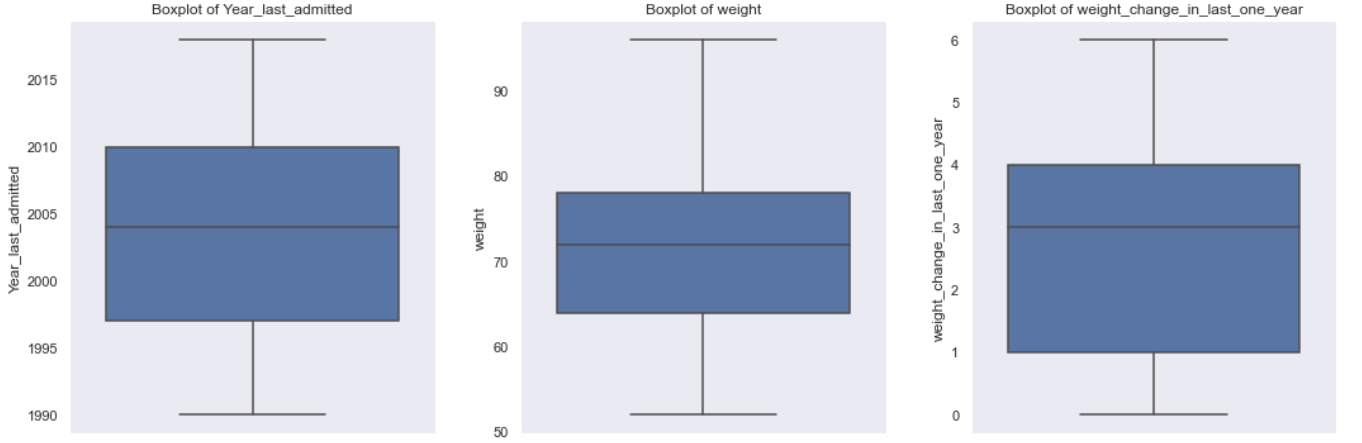
The mean glucose here 167.53 with 57 as minimum and 277 as maximum. The mean weight here is 71.61 with 52 kgs as min weight and 96 as maximum weight, maximum weight loss or weight gain an individual has experienced the previous year ids 6kgs.

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

3. Exploratory Data Analysis

**Univariate \ Bivariate Analysis:-**





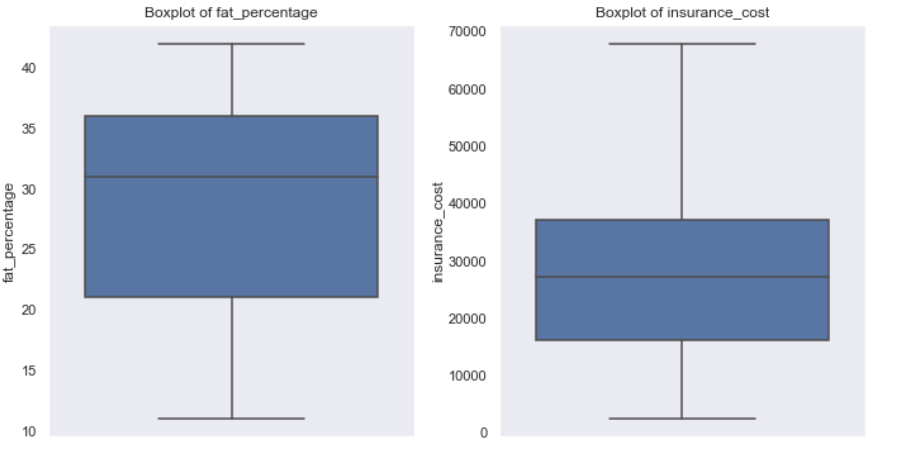
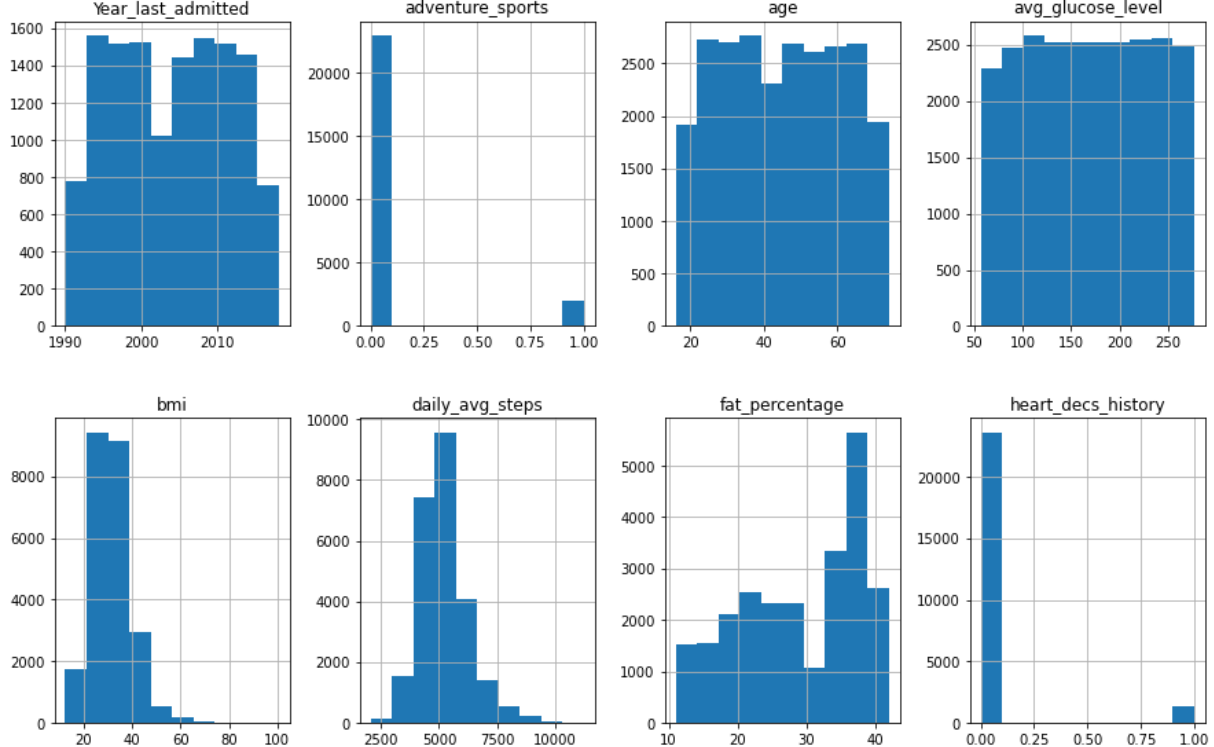
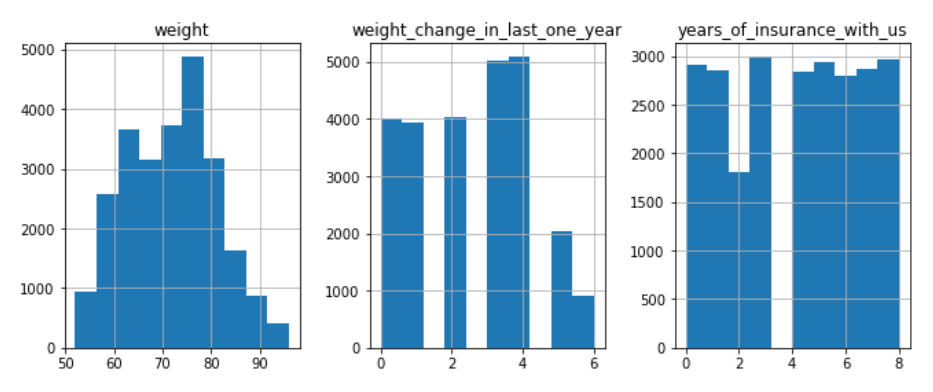


Fig-1-Outlier Boxplot

Fig-1-Boxplots shoeing outliers

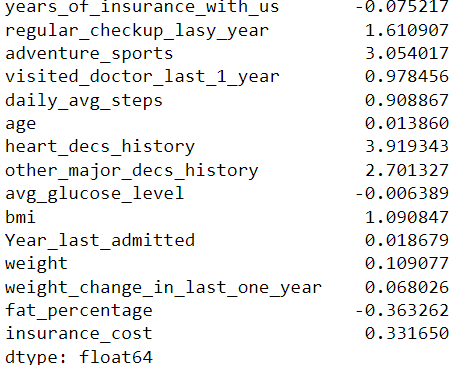
From the boxplots we could infer that regular\_checkup\_lasy\_year, daily\_avg\_steps and bmi have outliers.





Our target variable is also positively skewed and contains outliers. From the above it appears the distribution of prices is skewed to the right i.e., as the prices increase, the number of diamonds decrease. Our skew is a positive value (greater than 1) indicating a positive skewness which means that our mean is greater than the median.

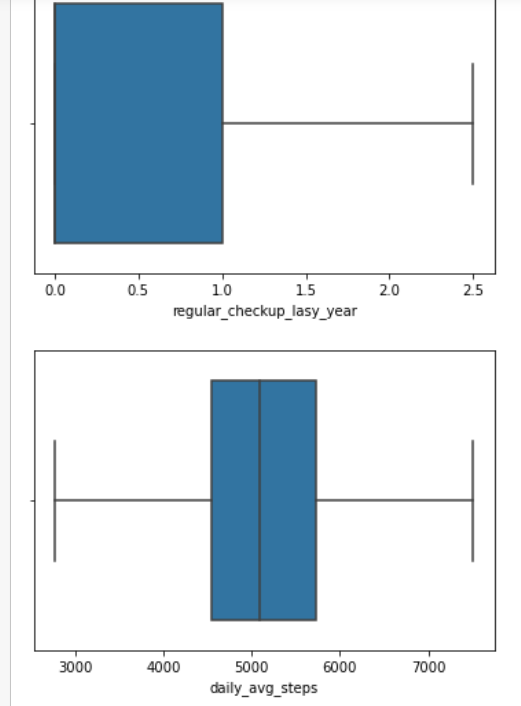
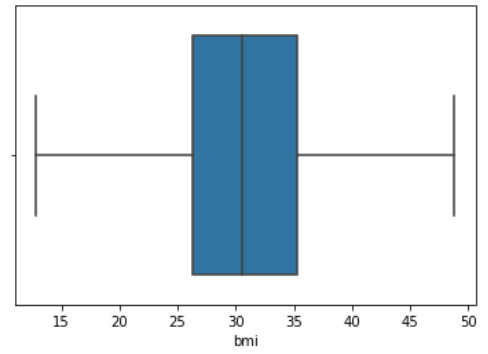
Fig-2-Histogram



**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-**

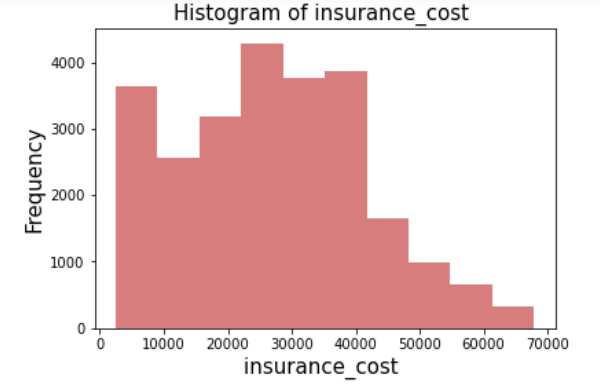
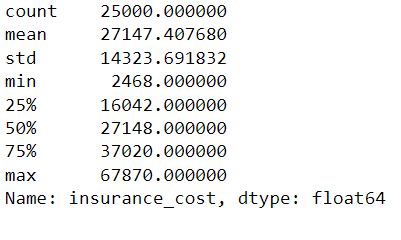


Fig-4-Target Variable Histogram

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.

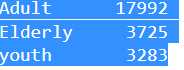
As age was having large number of variables, I grouped age into “age\_group” for easy

analysis –

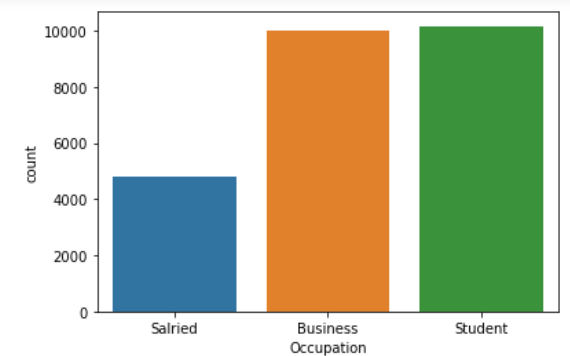
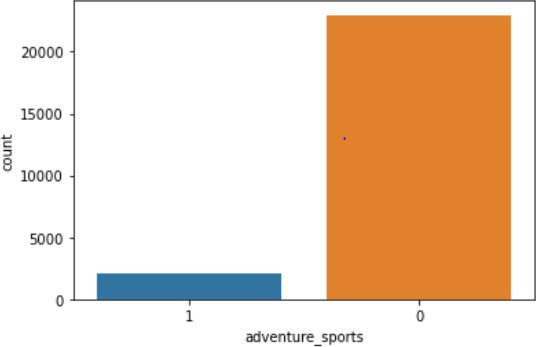
Youth (15-24 years)

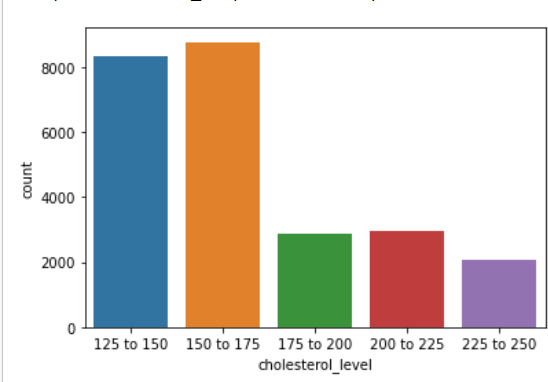
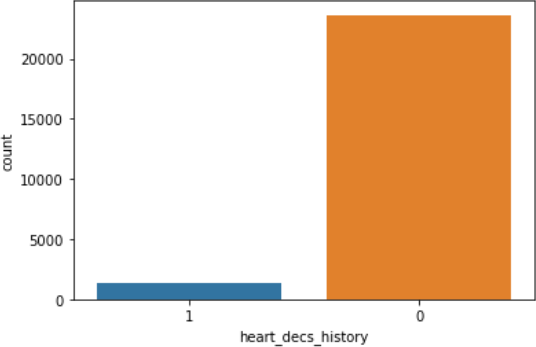
Adults (25-64 years)

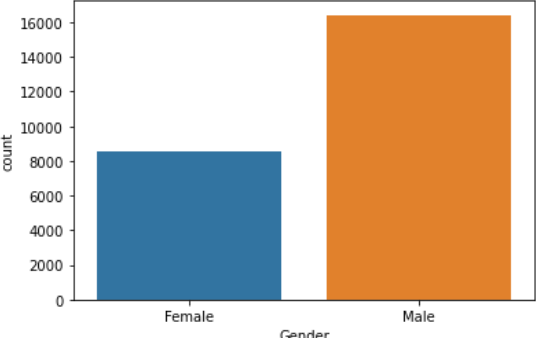
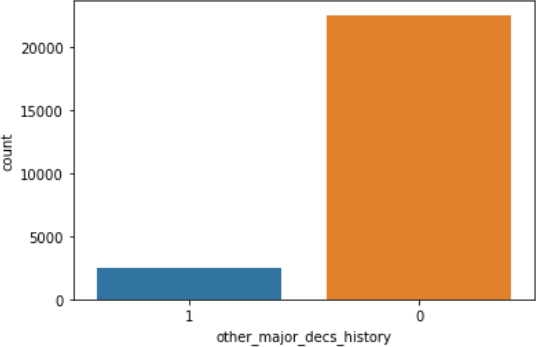
Elderly (65 years and over)

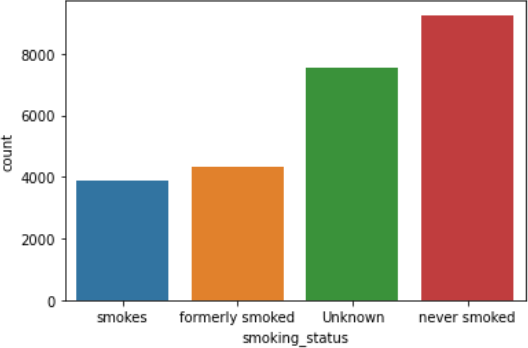


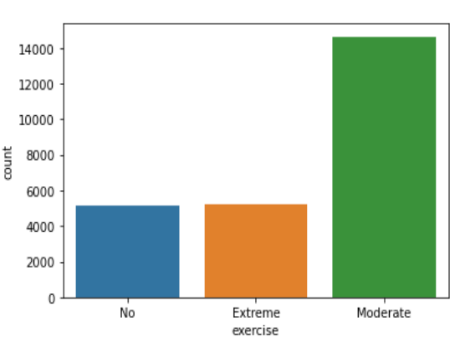
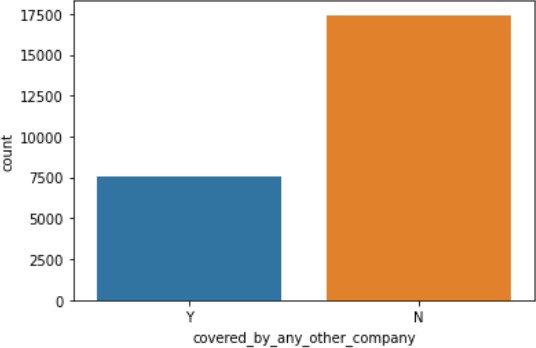
**Categorical Variables-**





.



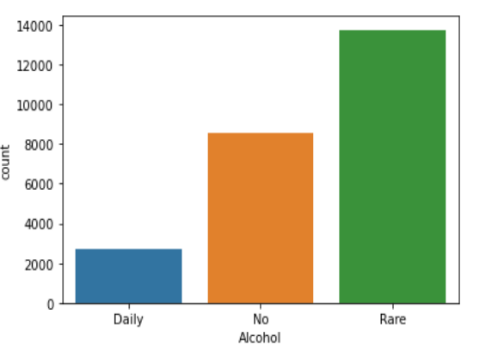


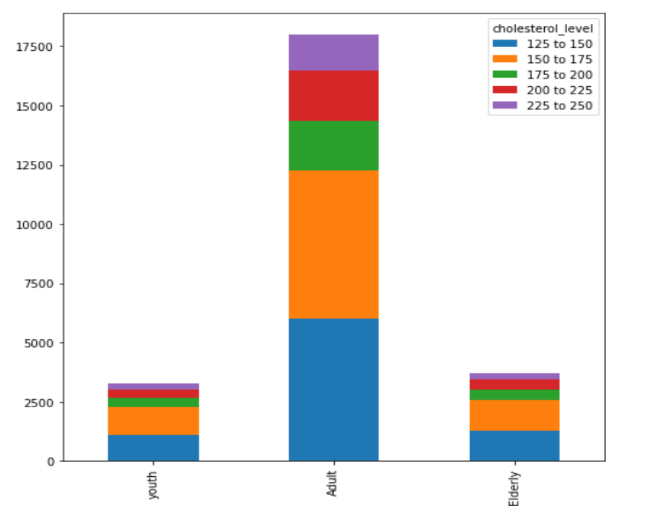
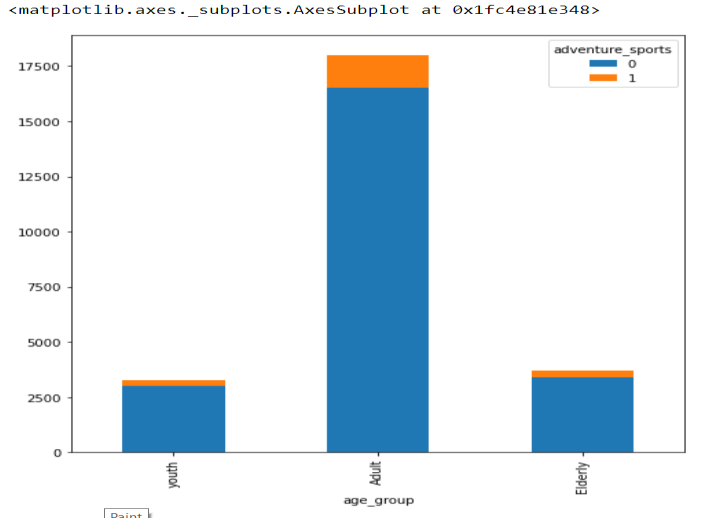
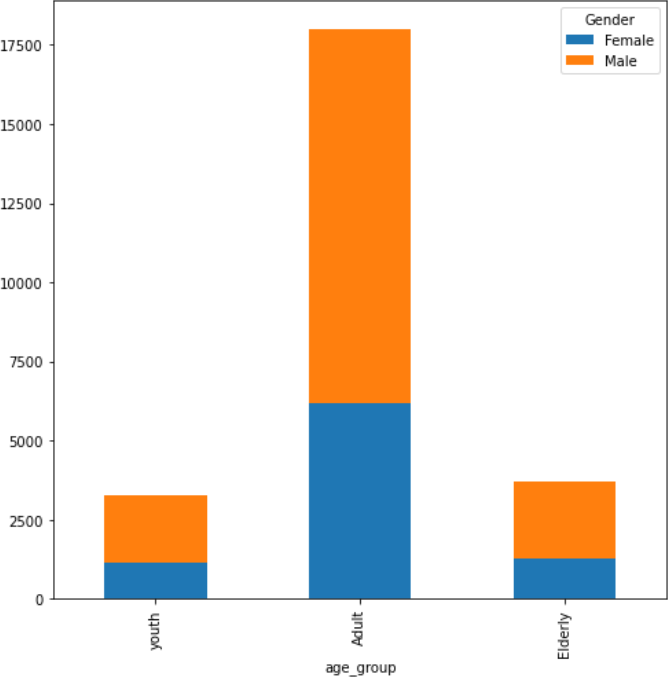
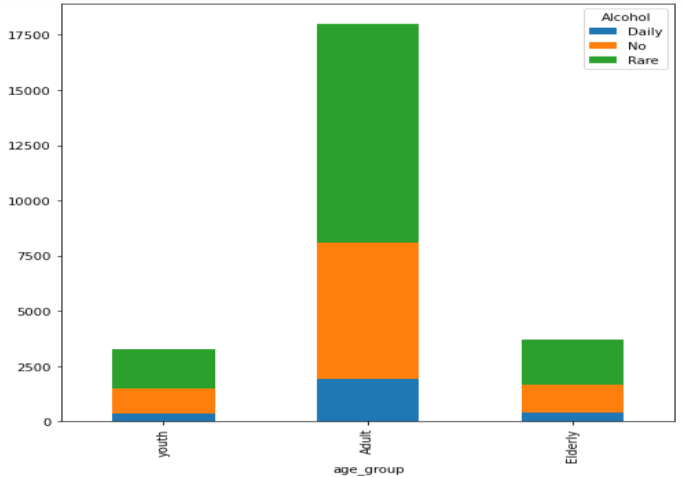
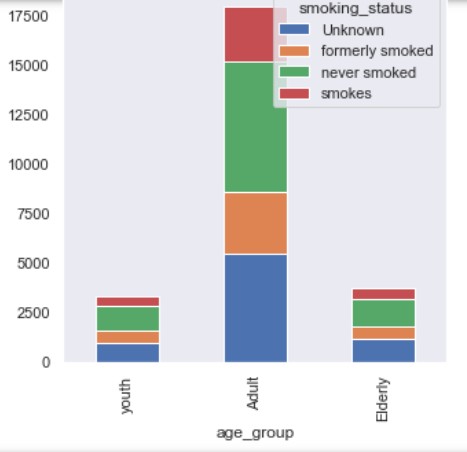
Fig-5-Barplot of Categorical Variable analysis

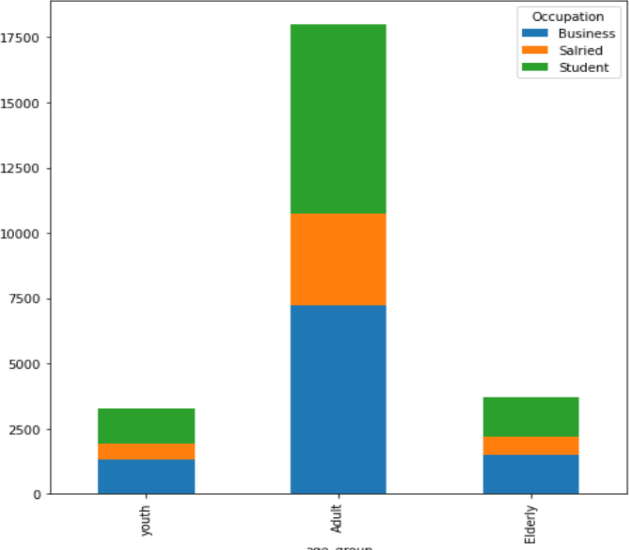
Adventurous sport is not much popular.

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.





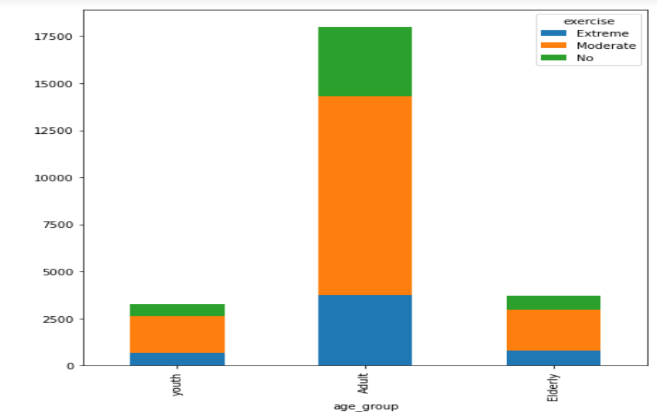
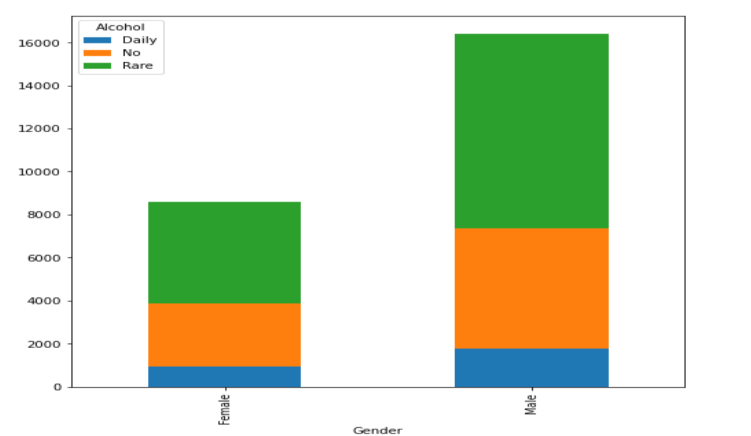
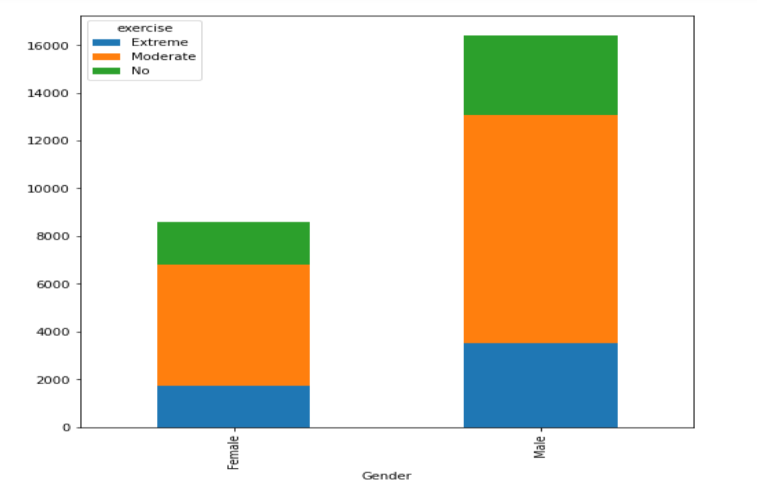


Fig-6-Stackbar of Age analysis with other variables

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.



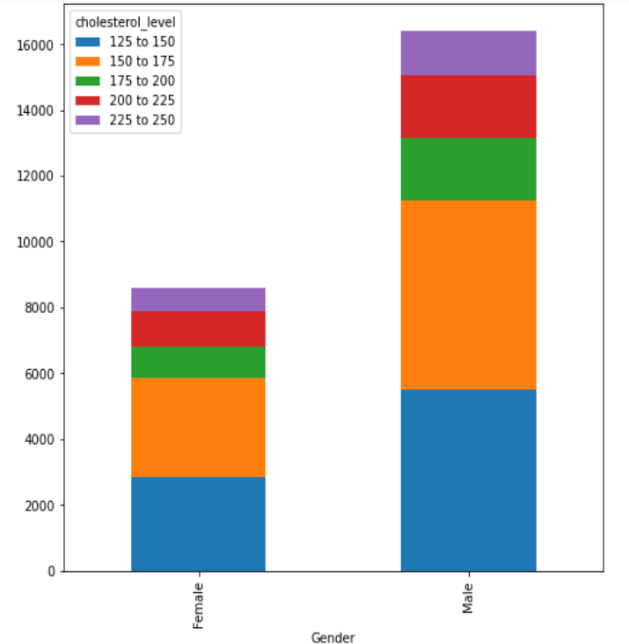
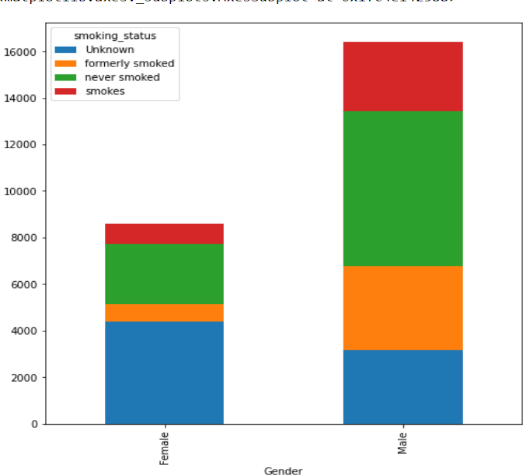
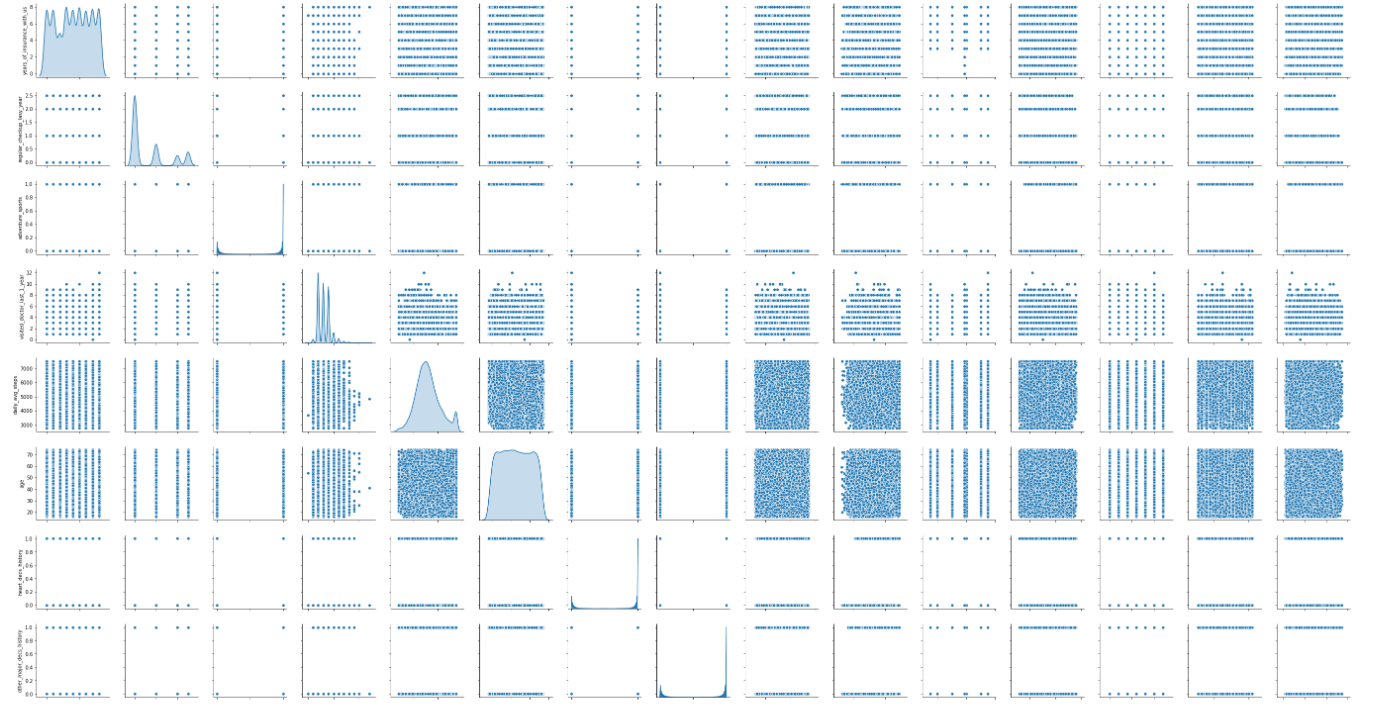


Fig-7-Stackbar of Gender analysis with other variables

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

MULTIVARIATE ANALYSIS-



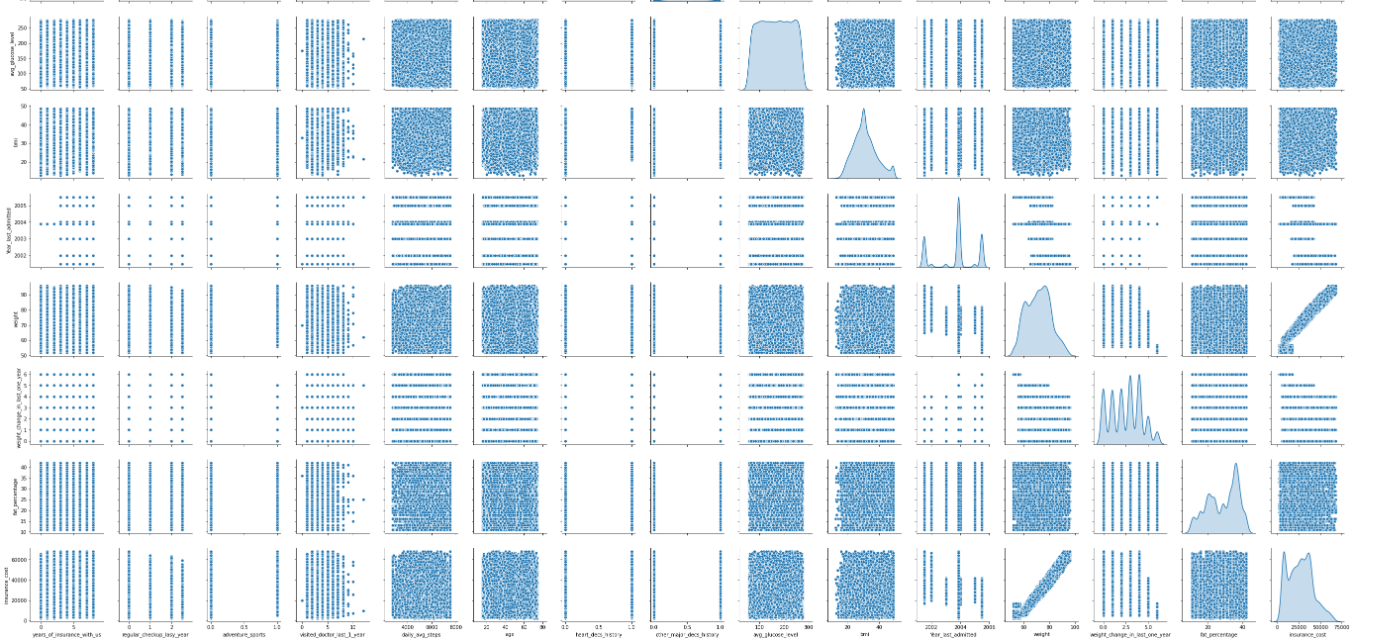
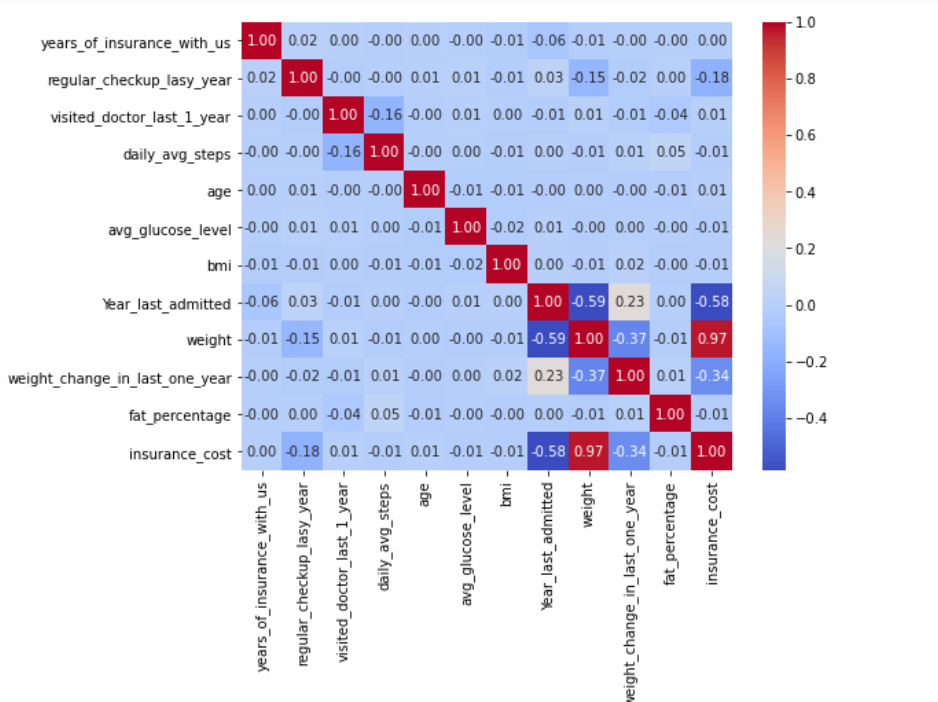


Fig-8-Pairplot



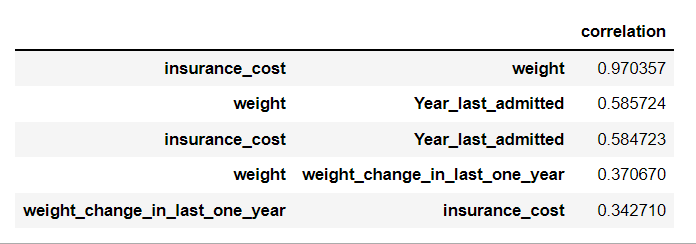
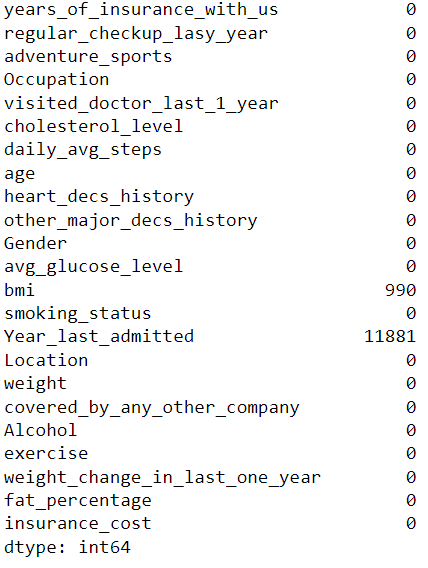


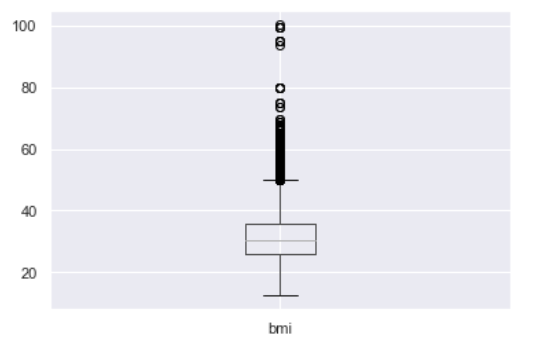
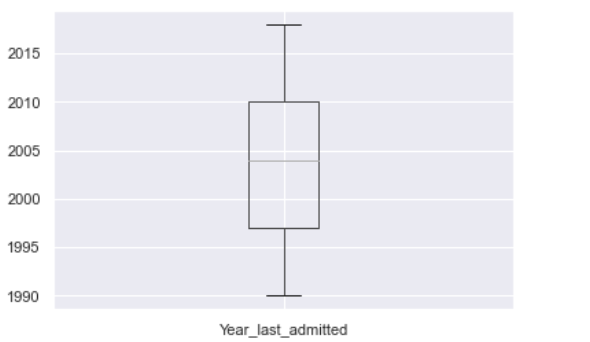
Fig-9-Heatmap

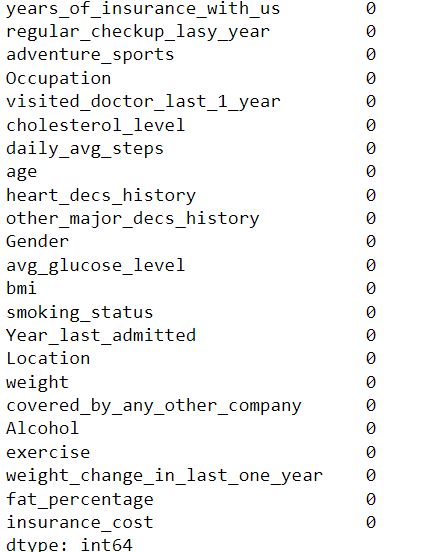
From the heatmap and pair plot the presence of no multicollinearity is visible. Except insurance cost and weight we no strong correlation amongst the variable is observed.

**Check for the null values –**



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.

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.

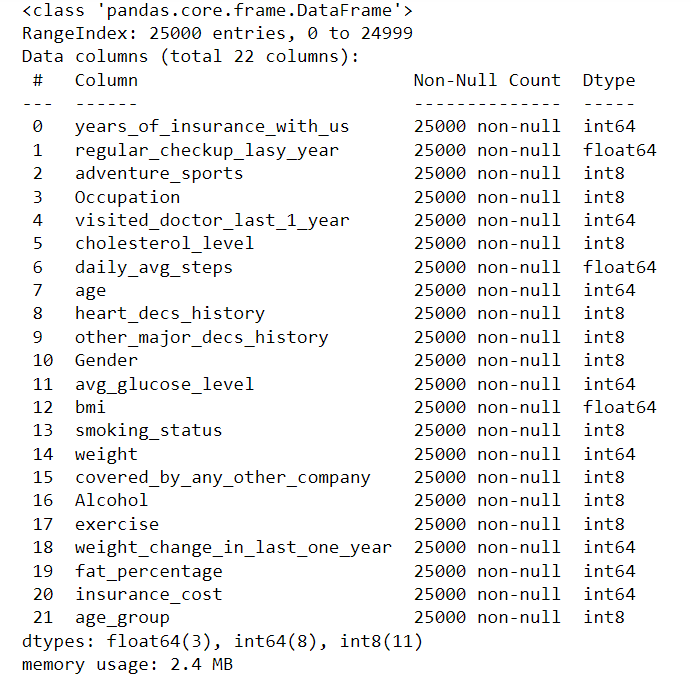


Table -7-Coversion of object variables into integer

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.

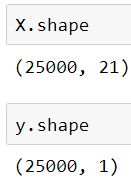
This is required to do since the machine learning algorithms only works on the numerical data. That is why there is a need to convert the categorical column into numerical one.

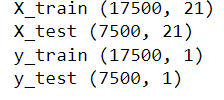
get\_dummies is the method which creates dummy variable for each categorical variable.

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**



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.

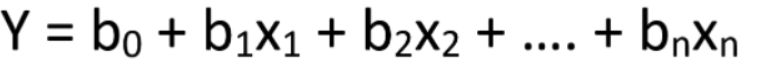
**Building A Linear Regression Model**-

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.

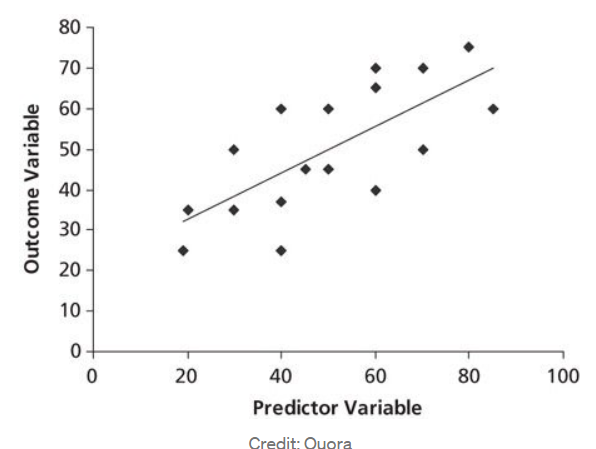
Regression analysis is a statistical method or supervised learning technique that helps us to understand the relationship between dependent and one or more independent variables**.**

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-



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.

**Evaluation Metrics for Your Regression Model:**

## 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.

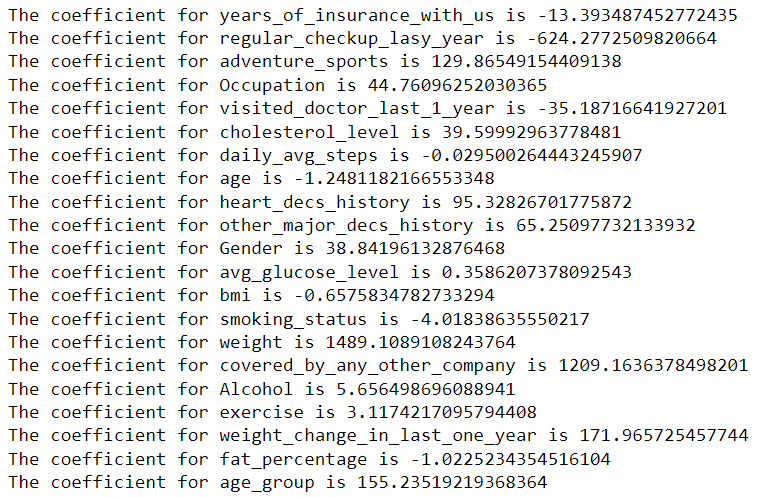
In contrast, MAE and MSE depend on the context as we have seen whereas the R2 score is independent of context.

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



**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 very good value for showing the accuracy. In some cases, Adjusted R-squared of 0.4 or more is acceptable as well.

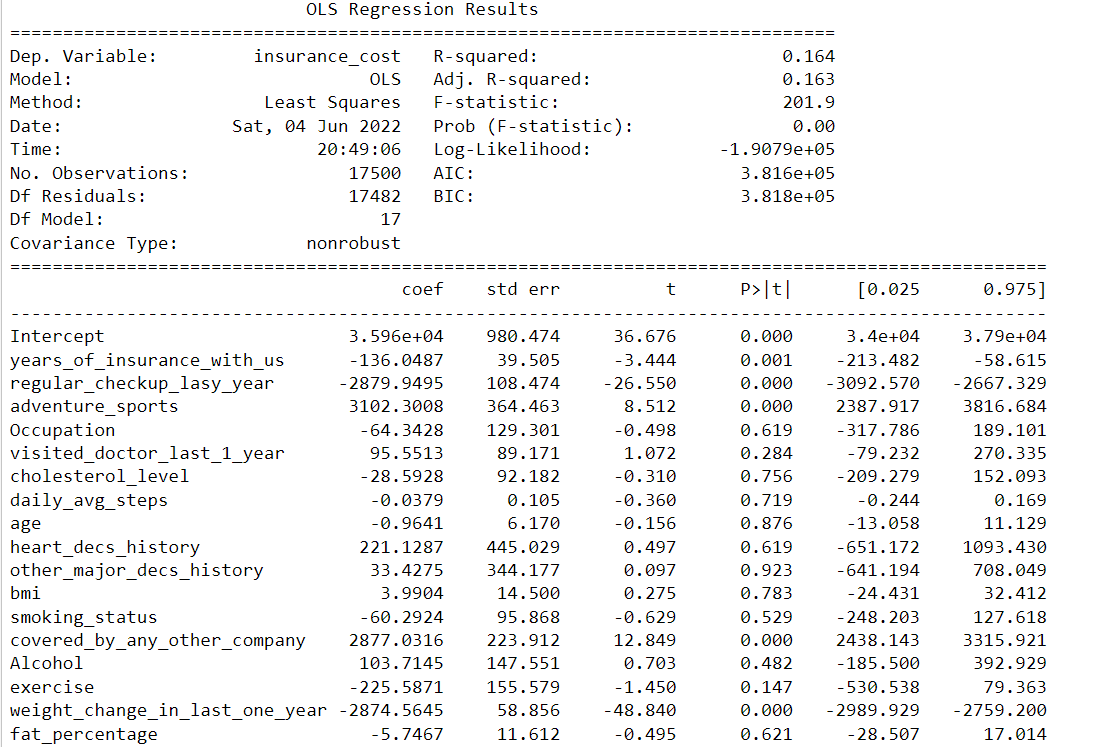
R^2 is not a reliable metric as it always increases with addition of more attributes even if the attributes have no influence on the predicted variable.

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 in one single data frame.

• We merged X and Y.

• Then we obtained the lml summary



**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 𝑅² (𝑅² corrected according to the number of input features) which is here is same as normal 𝑅²

* 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**

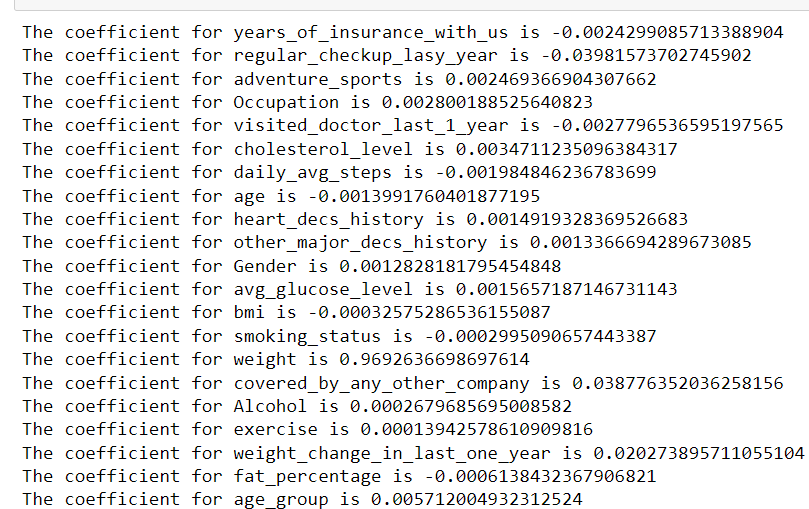
To check if regression analysis on scaled data gives a better performance I did a second iteration using scaled data.

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 –**

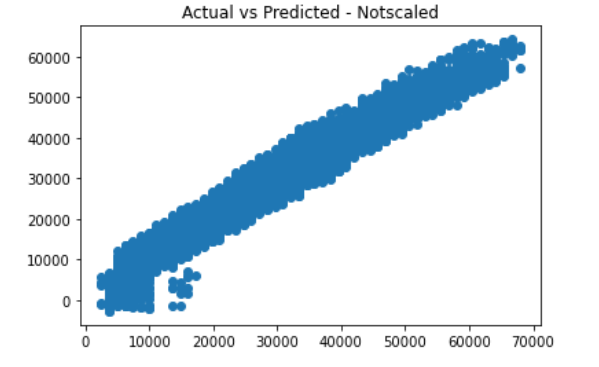
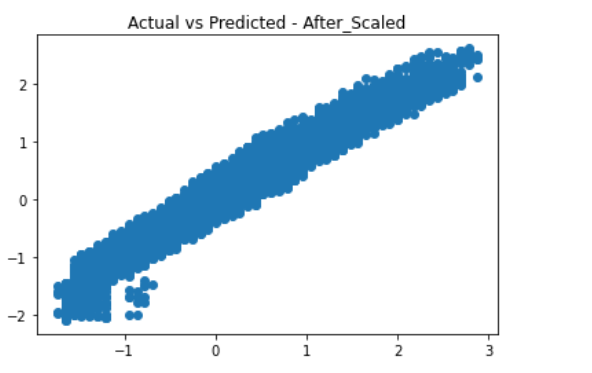


**Table-10 -Coefficient table after scaling**

* 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**

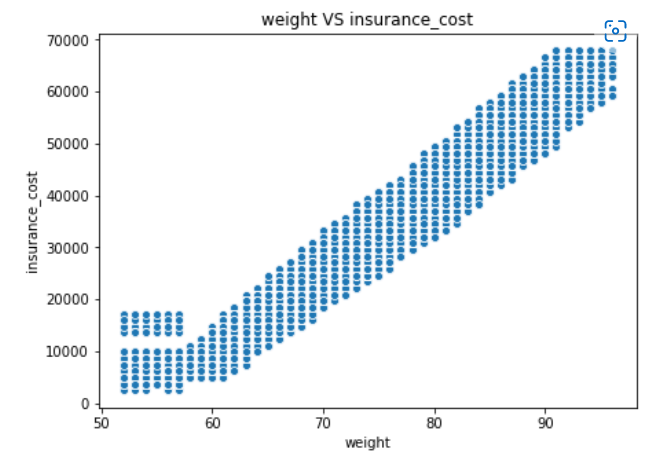
We could observe that regression analysis with scaled data is slightly better then the unscaled data performance.

**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.

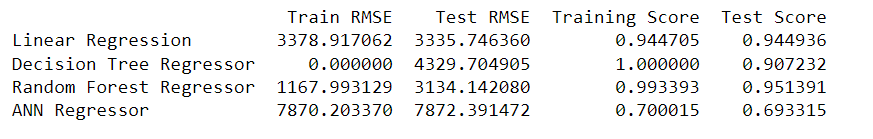
## Checking for other regressor models for improvement:

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



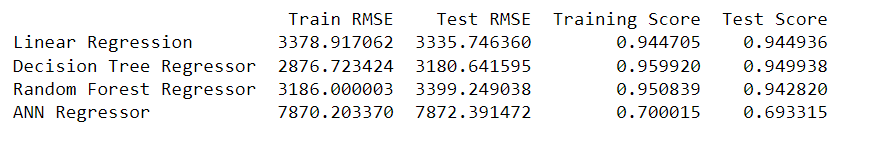
**Table-11-Summary RMSE AND RSQUARE BEFORE MODEL TUNING**

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-



**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

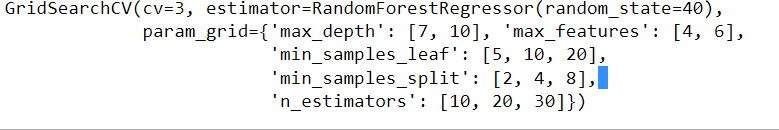
max\_depth = max number of levels in each decision tree

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



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

**MODEL SELECTION**

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.