HEALTH INSURANCE CASE STUDY 2023

Submitted By: Hitesh Dadhich

Date: 20-May-2023

Batch: PGPDSBA_May2022

Contents

Executive Summary

In a recent time, health insurance is a need of every individual because treatment is now very costly and individual has to go through tough time if he/she didn’t plan for it therefore having health insurance is very good option as it will cover treatment of illness and individual can focus on health and recovery without thinking much about cost of treatment.

The main objective of this case study is,

→ To find out optimum insurance cost based on different independent variables.
→ To identified different insights/pattern from data visualization and other analysis.
→ Recommendations for business.

Data Description:

There are 25000 Rows and 24 columns in the dataset.

We have 16 numeric and 8 object data types present in the data.

Even though there are some numeric variables which sounds to be a categorical variable such as 'adventure_sports', 'heart_decs_history', 'other_major_decs_history' but it assigned values of 0 and 1 so, it recognized as numeric variables.

We have removed variable 'applicant_id' as it does not provide any significant information & just a continuous number

No duplicate values present.

Please refer to “Appendix for Data head, tail, information and Dictionary” about the dataset.

Predictive Modeling Results:

Out of all the models which have been built, Decision Tree Regressor tune model finds a best fit with an accuracy of about 96% and 95%in train and test data. Also, RMSE value of 0.21 and 0.22 is seen while MAPE score of 10.99% and 10.11% obtained by train and test data set. Overall, 95% variance in Insurance cost has been explained by other variables.

Introduction to business problem

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 ofgetting ill is drastically reduced.

Need of the study/project:

We want to do study on this data because health plan mostly depends on the health history and requirements of individual so with this study, we can derive some useful insights that what is the most common health parameters between customers and what are the cost of different type of plan and as we know that we need to find out the cost of insurance plan which is a regression problem therefore we will try various regression models.

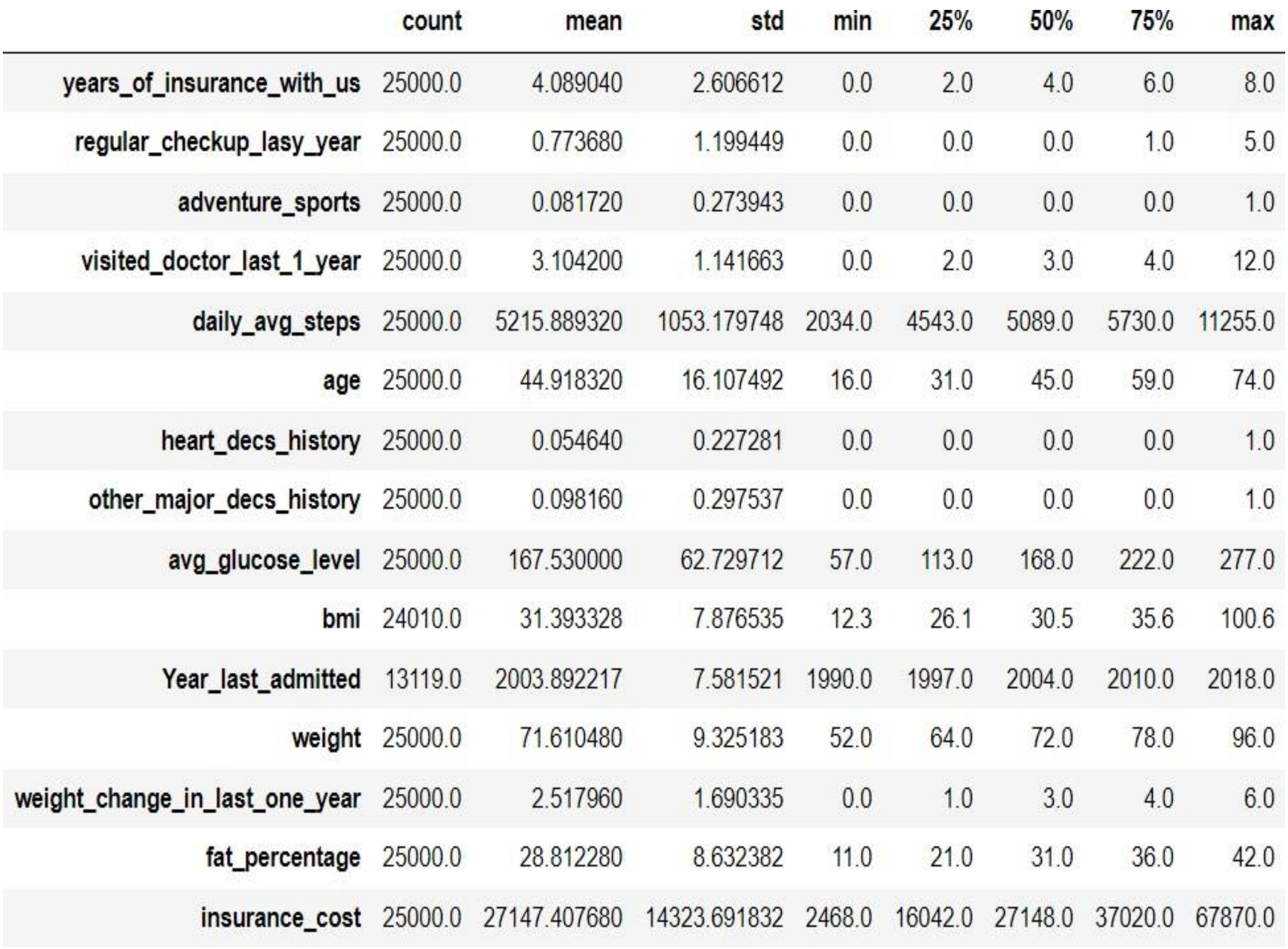
Understanding business/social opportunity:

This case study will be very useful for company to identify most common health condition andby using that company can promote their health plan for that particular health disease/problem.

EDA and Business Implication

In EDA we are going to find out pattern/insights from data by using visualization and non-visualization technique.

Descriptive summary (non-Visual representation):



	count	mean	std	min	25%	50%	75%	max
years_of_insurance_with_us	25000.0	4.089040	2.606612	0.0	2.0	4.0	6.0	8.0
regular_checkup_lasy_year	25000.0	0.773680	1.199449	0.0	0.0	0.0	1.0	5.0
adventure_sports	25000.0	0.081720	0.273943	0.0	0.0	0.0	0.0	1.0
visited_doctor_last_1_year	25000.0	3.104200	1.141663	0.0	2.0	3.0	4.0	12.0
daily_avg_steps	25000.0	5215.889320	1053.179748	2034.0	4543.0	5089.0	5730.0	11255.0
age	25000.0	44.918320	16.107492	16.0	31.0	45.0	59.0	74.0
heart_decs_history	25000.0	0.054640	0.227281	0.0	0.0	0.0	0.0	1.0
other_major_decs_history	25000.0	0.098160	0.297537	0.0	0.0	0.0	0.0	1.0
avg_glucose_level	25000.0	167.530000	62.729712	57.0	113.0	168.0	222.0	277.0
bmi	24010.0	31.393328	7.876535	12.3	26.1	30.5	35.6	100.6
Year_last_admitted	13119.0	2003.892217	7.581521	1990.0	1997.0	2004.0	2010.0	2018.0
weight	25000.0	71.610480	9.325183	52.0	64.0	72.0	78.0	96.0
weight_change_in_last_one_year	25000.0	2.517960	1.690335	0.0	1.0	3.0	4.0	6.0
fat_percentage	25000.0	28.812280	8.632382	11.0	21.0	31.0	36.0	42.0
insurance_cost	25000.0	27147.407680	14323.691832	2468.0	16042.0	27148.0	37020.0	67870.0

Table 1: Descriptive summary of numerical variables

Insights:

1. A customer is linked to the company for 8 years in short, these customer/customersare very happy with their health plan.

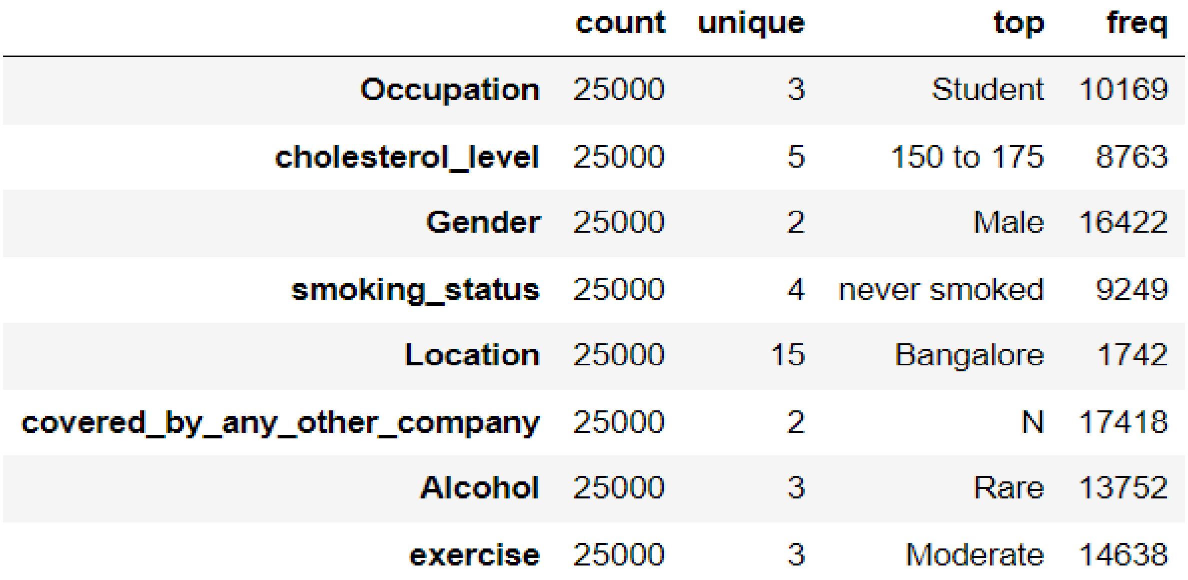
2. Approx 75% or less than 75% customers went for regular check only once in a year.

3. Most of the customers are overweight as per BMI value.

4. Less than or equal to 50% customer's age is 45 and maximum age is 74.

5. Average glucose level of customers is 167 whereas we know normal glucose level is below 100 so are these customers have diabetes.

6. Insurance Cost ranges from approx. 2500 to 68K rupees which shows variety of health plans are available for customers.



	count	unique	top	freq
Occupation	25000	3	Student	10169
cholesterol_level	25000	5	150 to 175	8763
Gender	25000	2	Male	16422
smoking_status	25000	4	never smoked	9249
Location	25000	15	Bangalore	1742
covered_by_any_other_company	25000	2	N	17418
Alcohol	25000	3	Rare	13752
exercise	25000	3	Moderate	14638

Table 2: Descriptive summary of categorical variables

Insights:

1. Mostly customers don't have any job they are students.
2. Most of the customer's cholesterol level is B/W 150 to 175.
3. Most of the customers are male and don't smoke.
4. Most of the customers don't have any other insurance because mostly are non-working so don't have employer's insurance.
5. We have customers from total 15 locations and mostly from Bangalore.

Visual representation

Univariate analysis:

We will do analysis of each individual numeric variable by using boxplot and histogram.

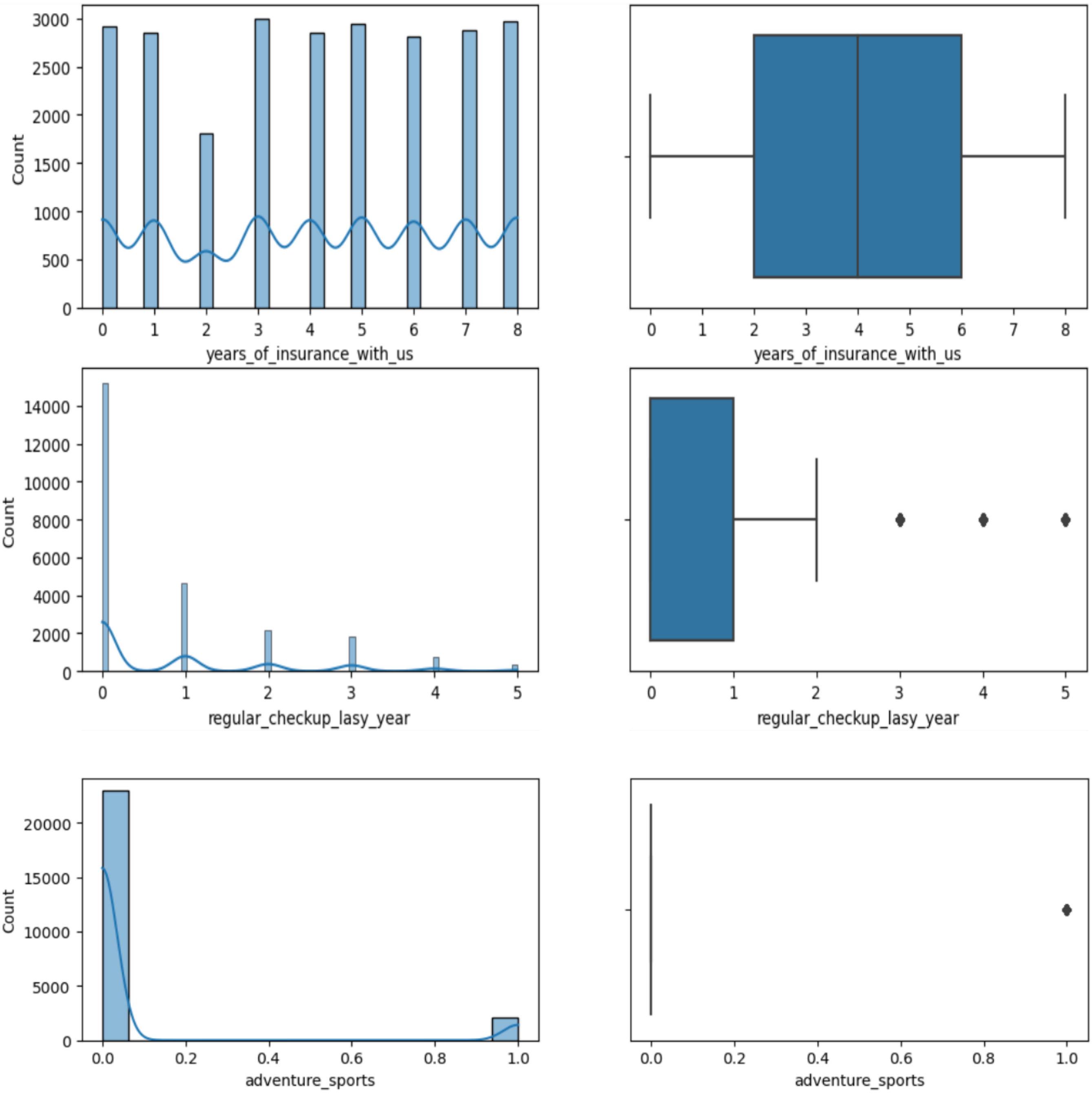


Figure 1: Histogram and Boxplot of numerical variables

Insights:

1. We can see median value of having same insurance is 4 years.

2. Outliers present in the column regular_check_last_year as some of the customers went for check-up more than the other customers.

3. Adventure_sports is just a binary value Yes or No so we cannot say that this is an outlier

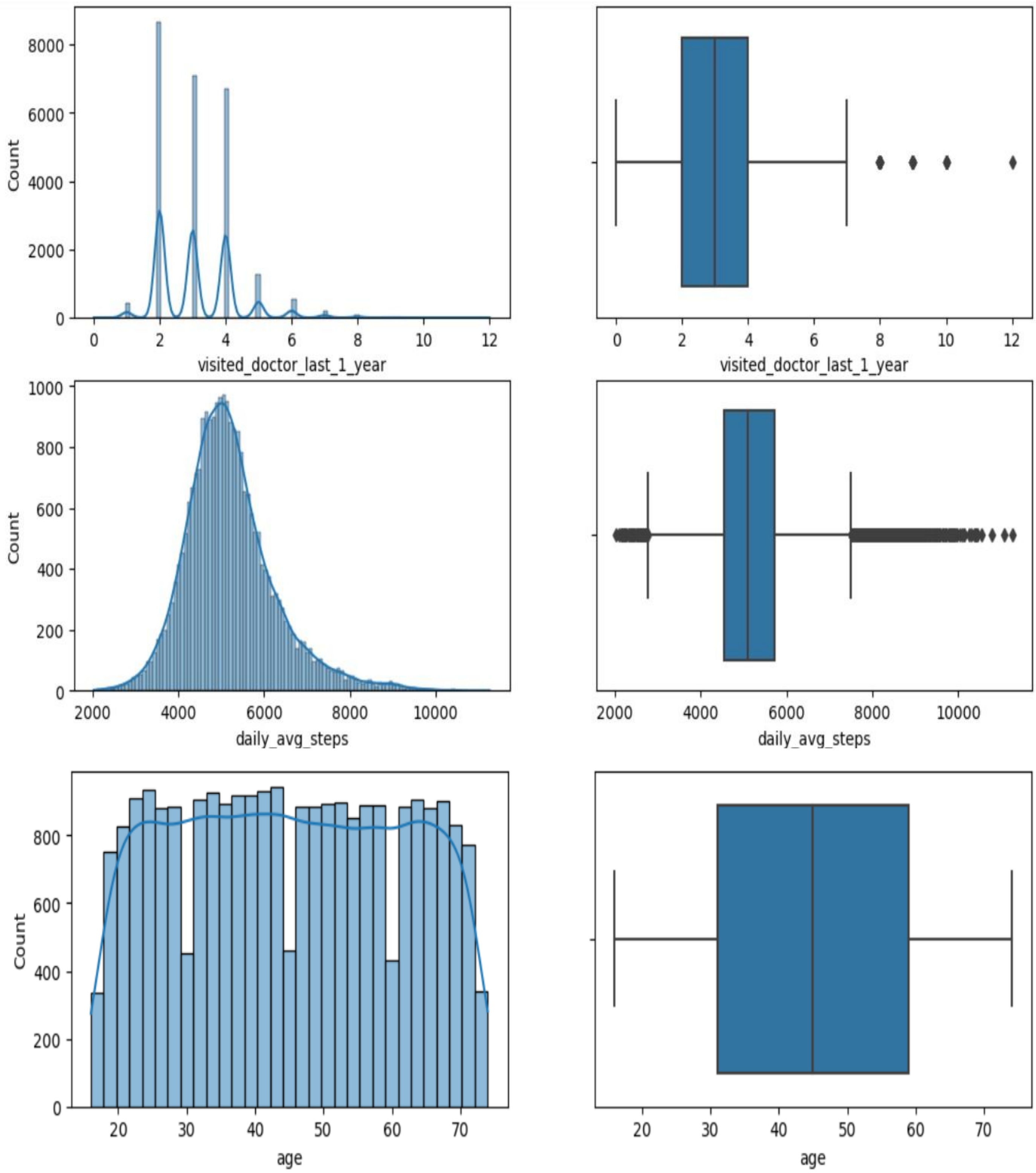


Figure 2: Histogram and Boxplot of numerical variables

<u>Insights</u>:

1. Distribution in the variable 'visited_doctor_last_1_year' & 'cholesterol_level' is positive and median age of customers is 45.

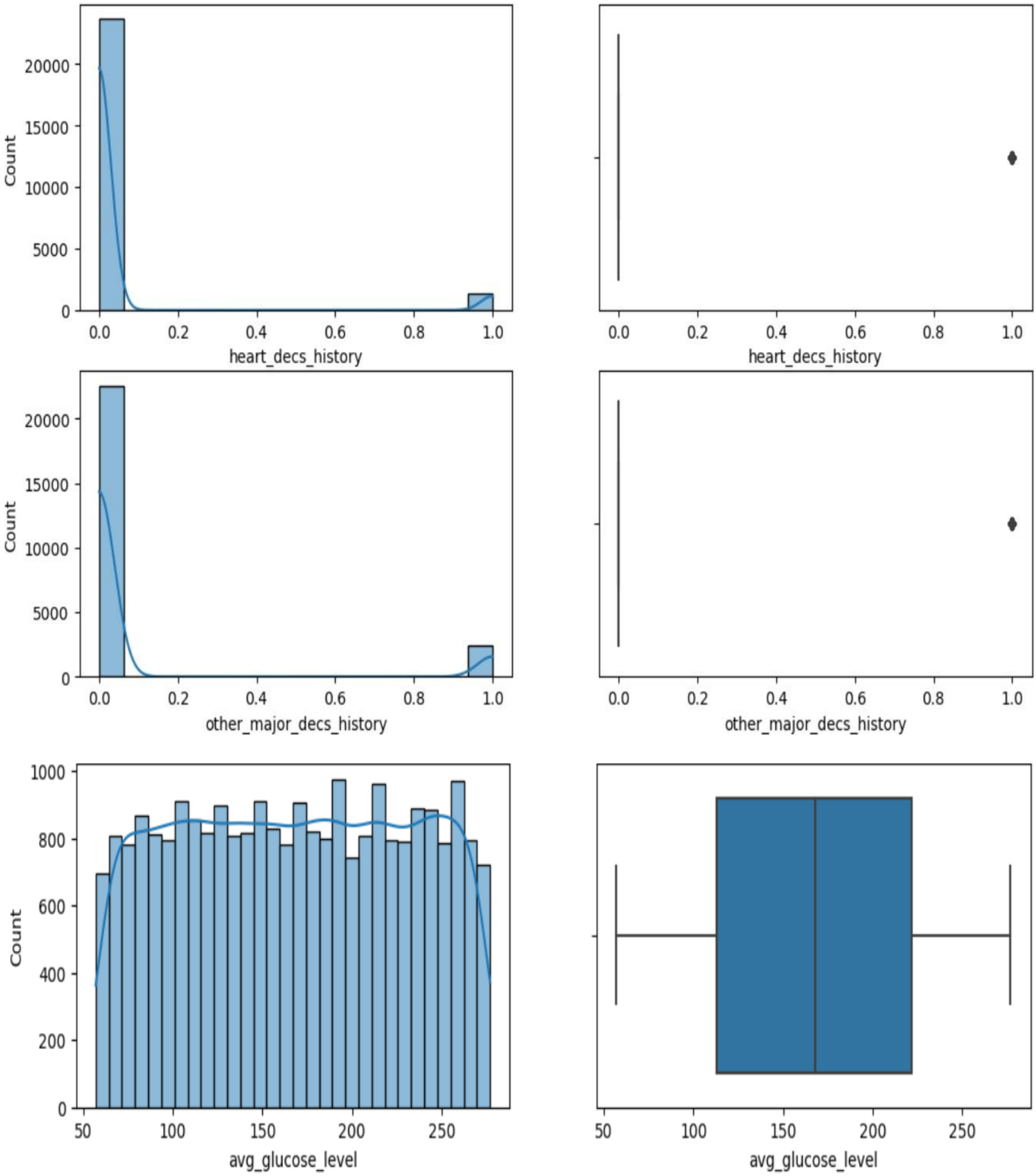


Figure 3: Histogram and Boxplot of numerical variables

<u>Insights</u>:

1. variable heart_decs_history & other_major_decs_history is just a binary value Yes or No so we cannot say that this is an outlier.

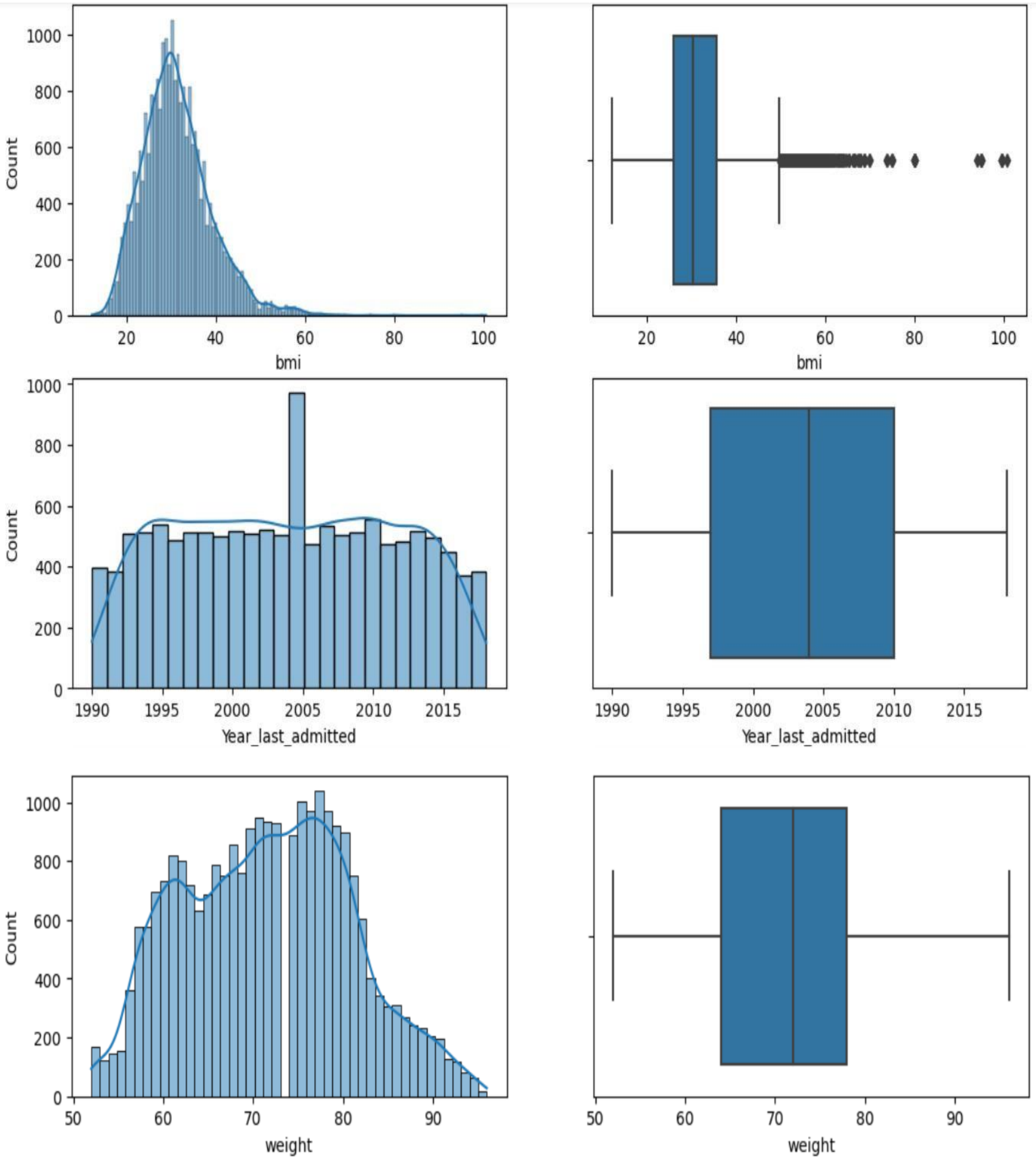


Figure 4: Histogram and Boxplot of numerical variables

Insights:

1. We have outliers in BMI and distribution is positive.

2. We can say that number of customers who were admitted to hospital till 2010 is <= 75%.

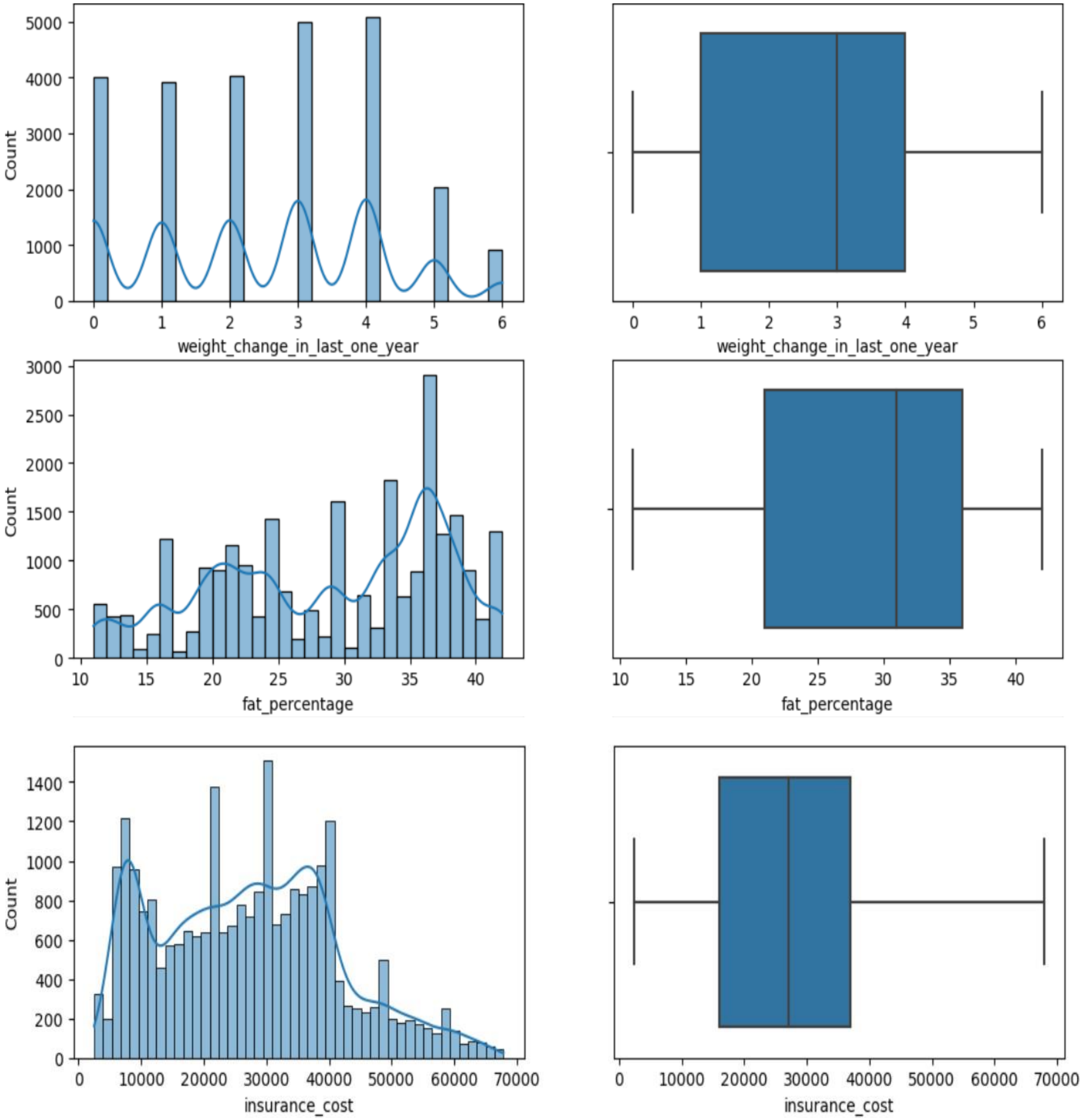


Figure 5: Histogram and Boxplot of numerical variables

<u>Insights</u>:

1. Distribution of fat_percentage is left skewed and it ranges from 10 to 40+

2. Average weight change in last 1 year is less than 3kg.

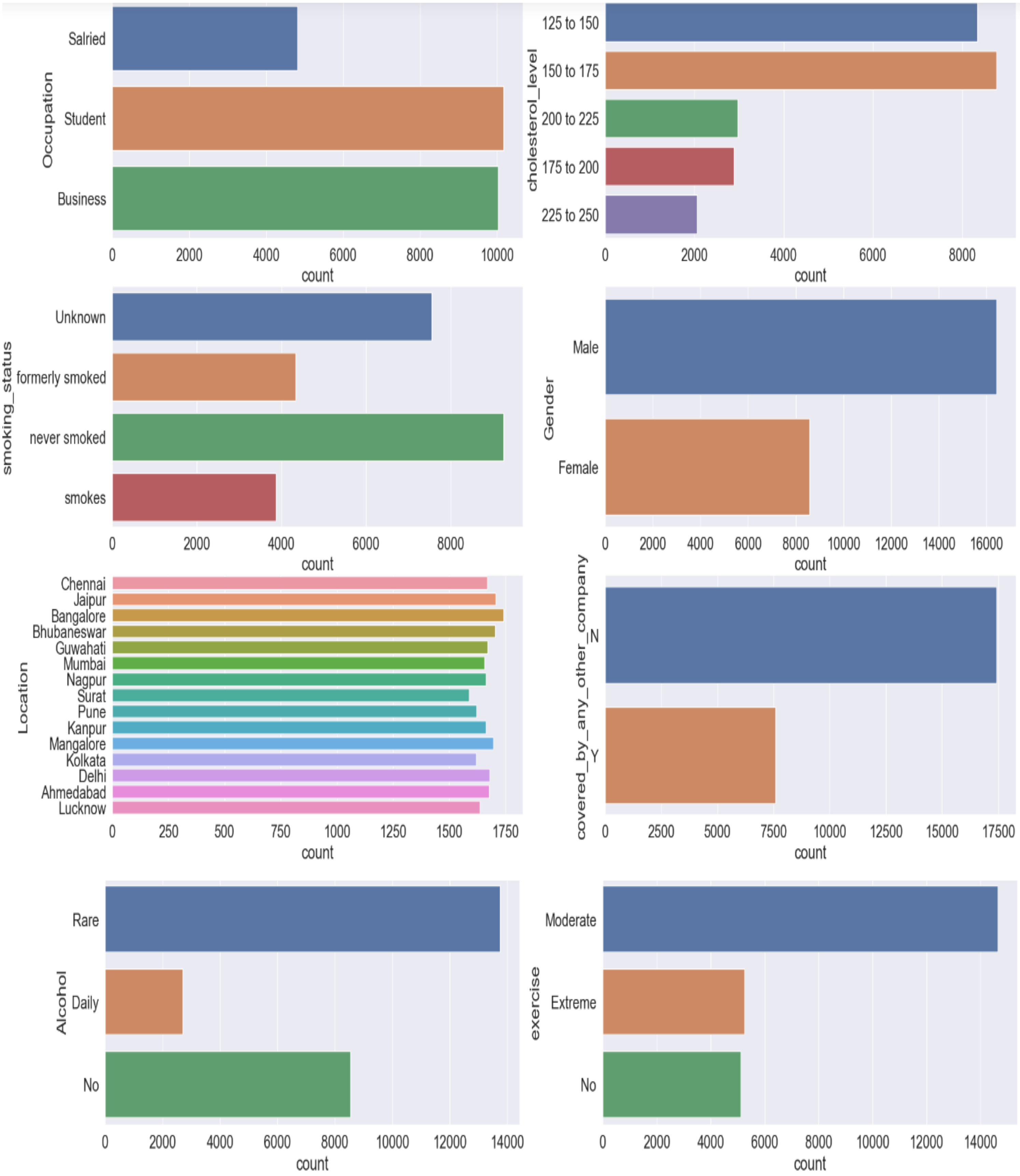


Figure 6: Count plot of Categorical variables

Insights:

1. Most of the customer's cholesterol_level is b/w 125 to 150 and 150 to 175 BUT there are customers whose CL is >200 which is high.

2. We have customer from different location and none of the location is too dominating.

3. Most of the customer do consume alcohol.

4. There are customers whose smoking status is unknown so these could be smokers or non-smokers.

Bivariate analysis:

We are analyzing relation/pattern between 2 variables by using count plot, scatter plot and boxplot.

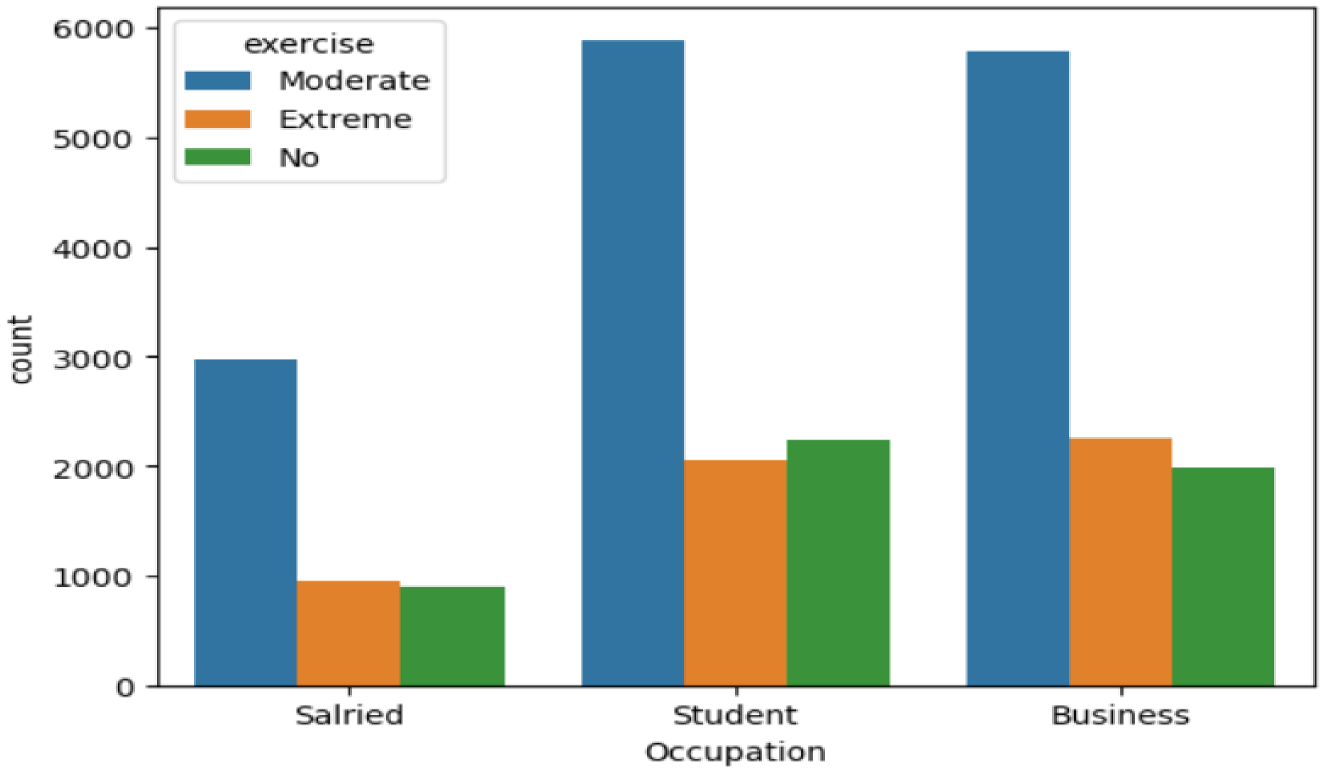


Figure 7: Count plot b/w variable Occupation and exercise

Insights:

1. Customers with occupations as business/student do more exercise compare to salaried person & it can be due to the busy schedule of salaried person.

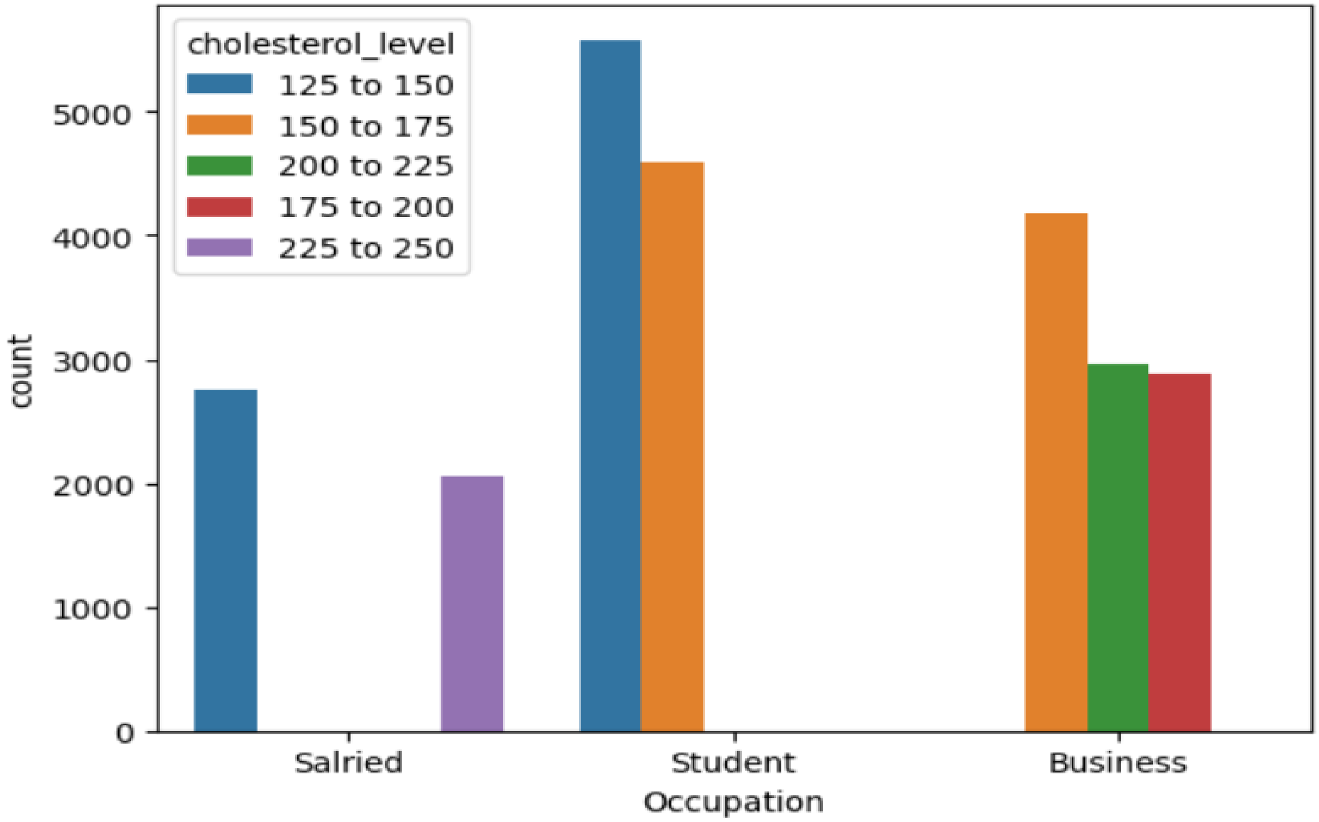


Figure 8: Count plot b/w variable Occupation and Cholesterol_level

<u>Insights:</u>

1. We can see very high cholesterol level (225-250) is only between salaried customers and second highest cholesterol level (200-225) is only in customers whose occupation is business.

2. Customer who doesn’t have any designated occupation(student) have cholesterol level in control.

3. Cholesterol level of salaried customer is either in control OR it's too high.

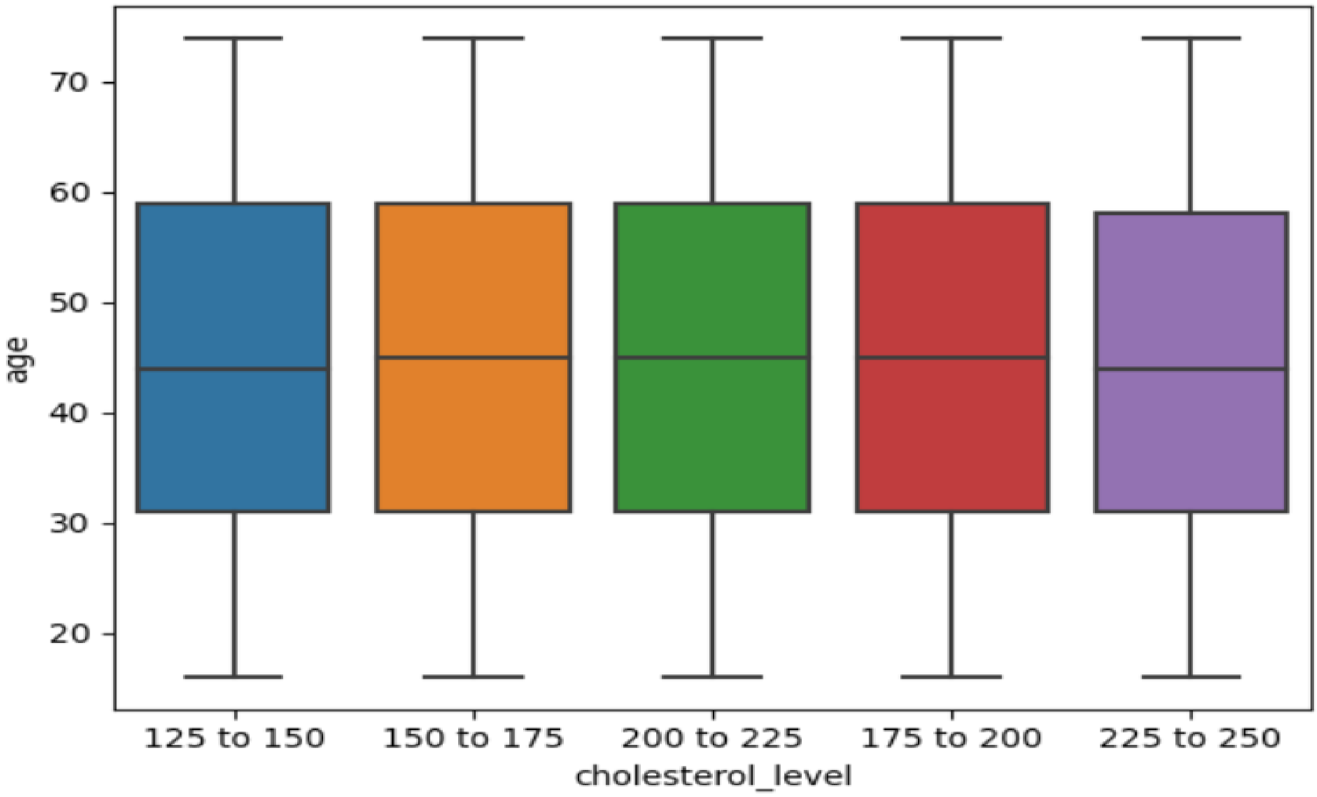


Figure 9: Boxplot b/w variable cholesterol_level and age

Insights:

1. Median age of the customers is almost same across 5 cholesterol level.

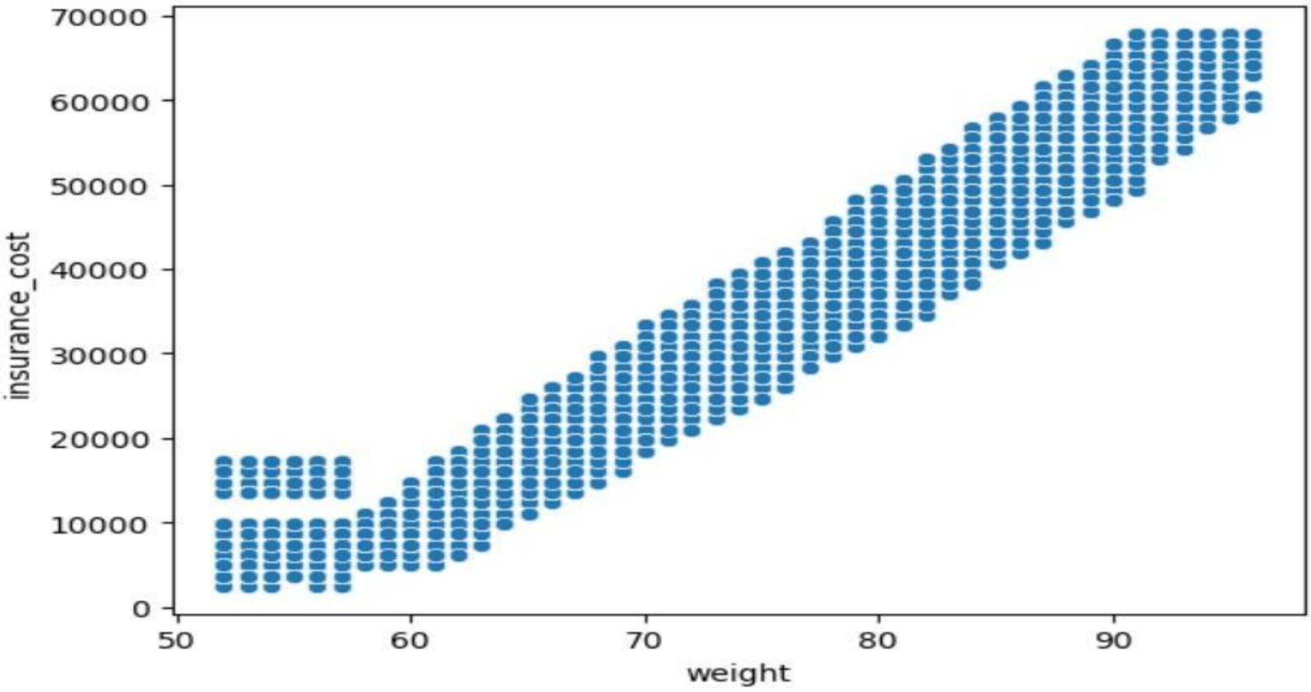


Figure 10: Scatter plot b/w variable weight & insurance cost

Insights:

1. We can see very strong correlation b/w weight and insurance cost that means insurancecost will increase if weight increases.

Observation on Correlation using Heatmap (Multivariate):

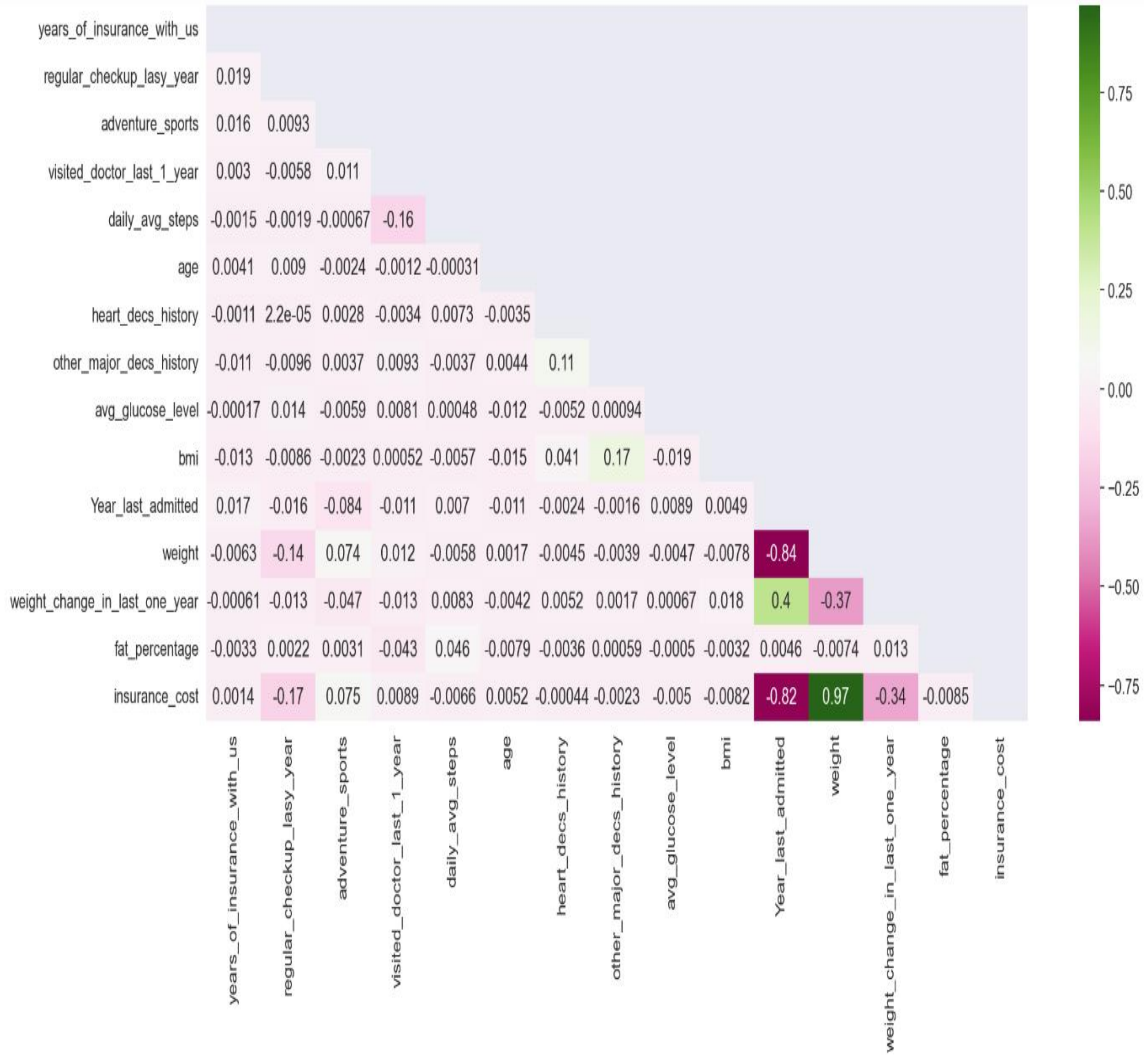


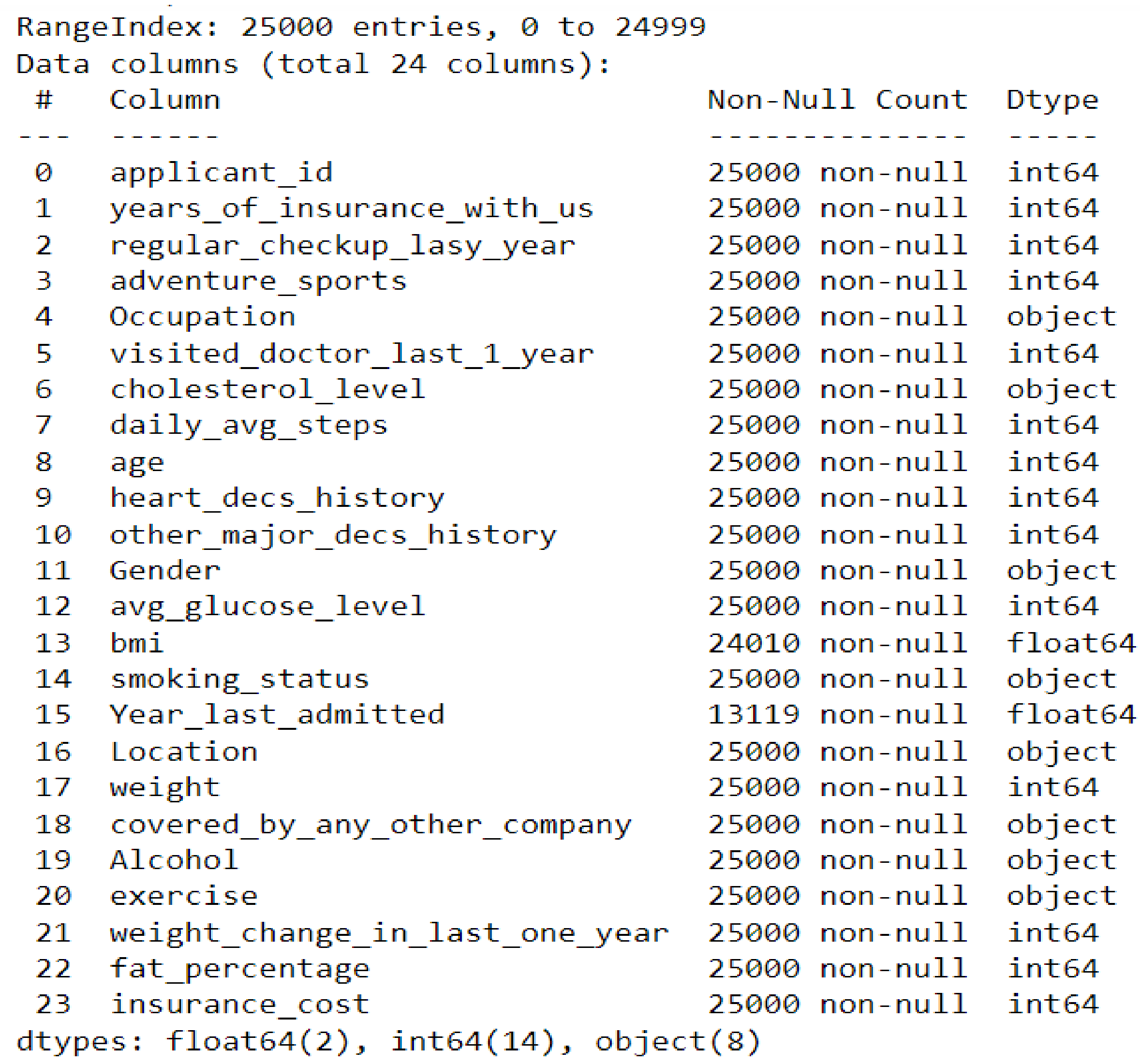
Figure 11: Heatmap

Insights:

1. From above heatmap we can see that only few of the variables are correlated.

Data Cleaning and Pre-processing

Missing Value treatment:



```
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)
```

Table 3: Data Information

From above table we can see that missing values present in dataset and total number of missing values for each variable are:

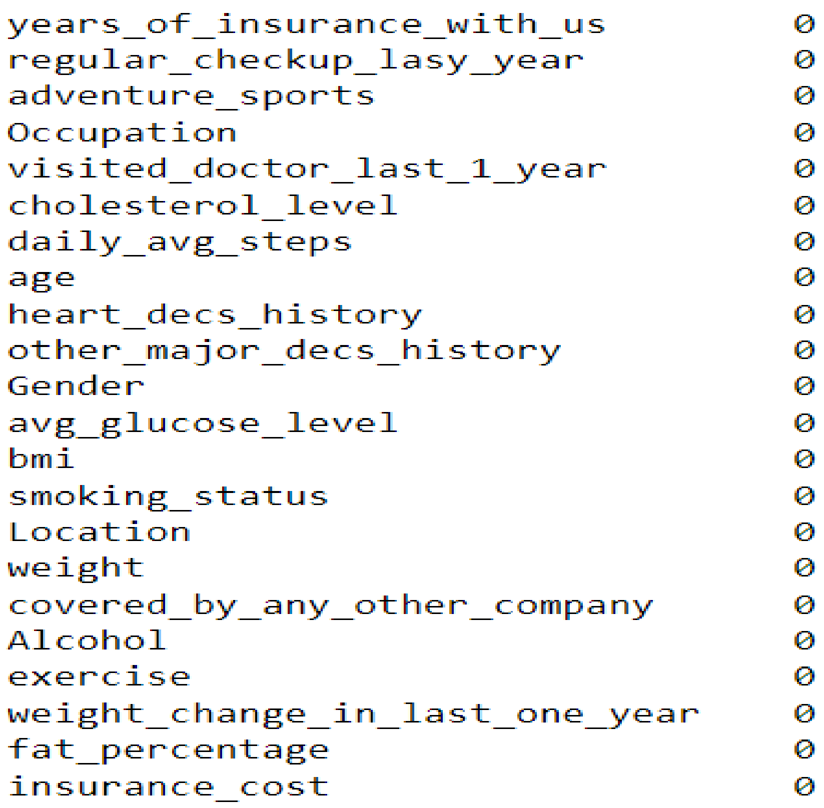
Year_last_admitted	11881
bmi	990

We can see that we have missing value present in these 2 variables and let's check what is the percentage count of missing values.

bmi	3.960
Year_last_admitted	47.524

We can see that in variable 'Year_last_admitted' we have 47% missing values which is more than industry standard of 30% and so we can drop this column.

Whereas for variable 'bmi' we can impute these missing values by using median value as Outliers present in our data for variable 'bmi'.



```
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
```

Table 4: No. of missing values in each variable

We can see that now, no missing value present in the data.

Outlier treatment:

As we have seen in univariate analysis that few of the variables have outliers so let'svisualize boxplot of those variables.

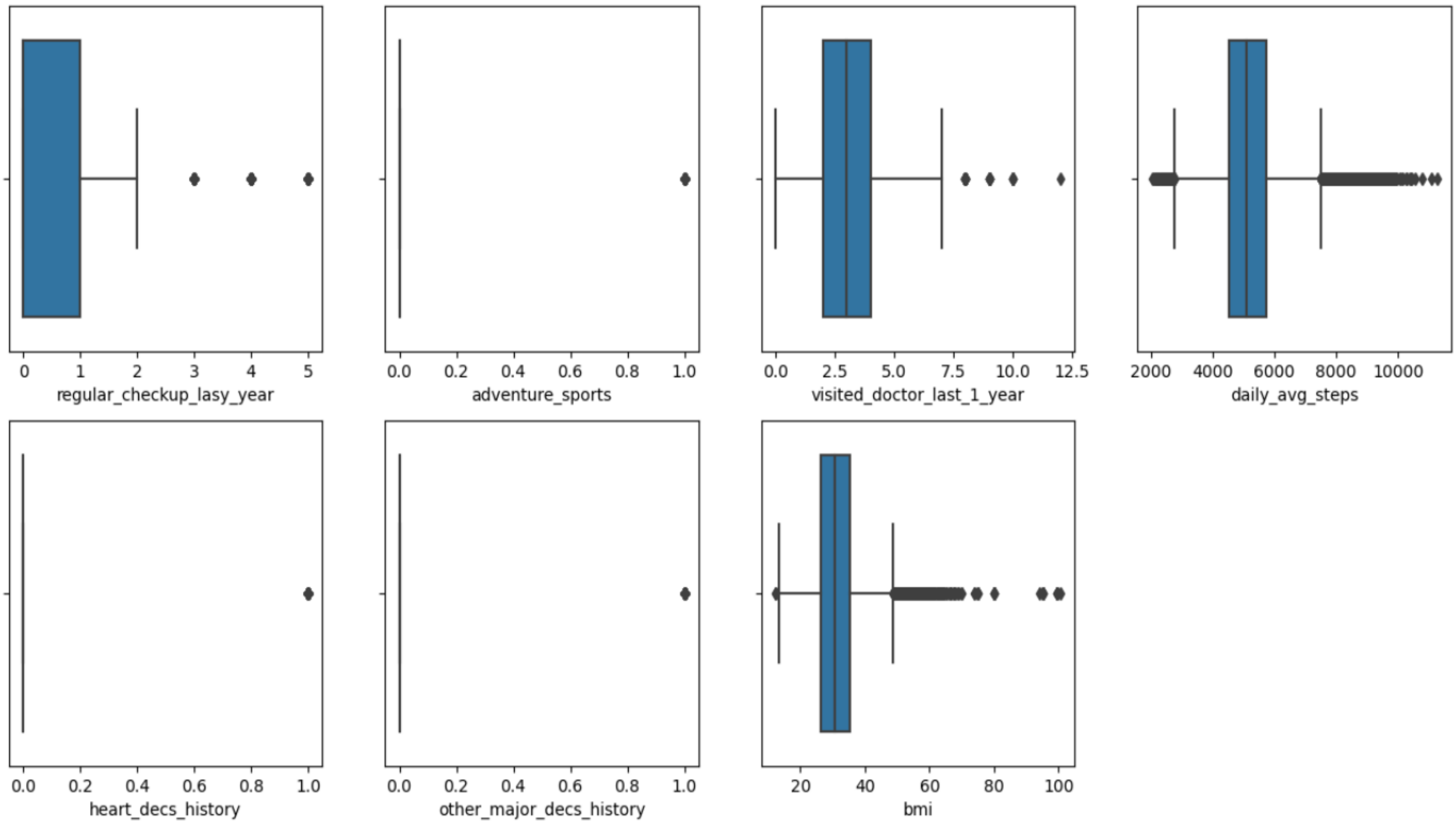


Figure 12: Boxplot of variables containing outliers

1. Variables like 'heart_decs_history', 'other_major_decs_history','adventure_sports' only have values 0 and 1 so this is not an outlier and so we are not treating it.

2. Whereas extreme values of other variables show abnormal behaviour like advisable yearly_regular_checkup is 1 but values such as 4 & 5 looks strange.

3. Similarly for variables 'visited_doctor_last_1_year', 'daily_avg_steps' & 'bmi' extreme values looks strange compare to usual standard values.

We are using IQR method to treat the outliers.

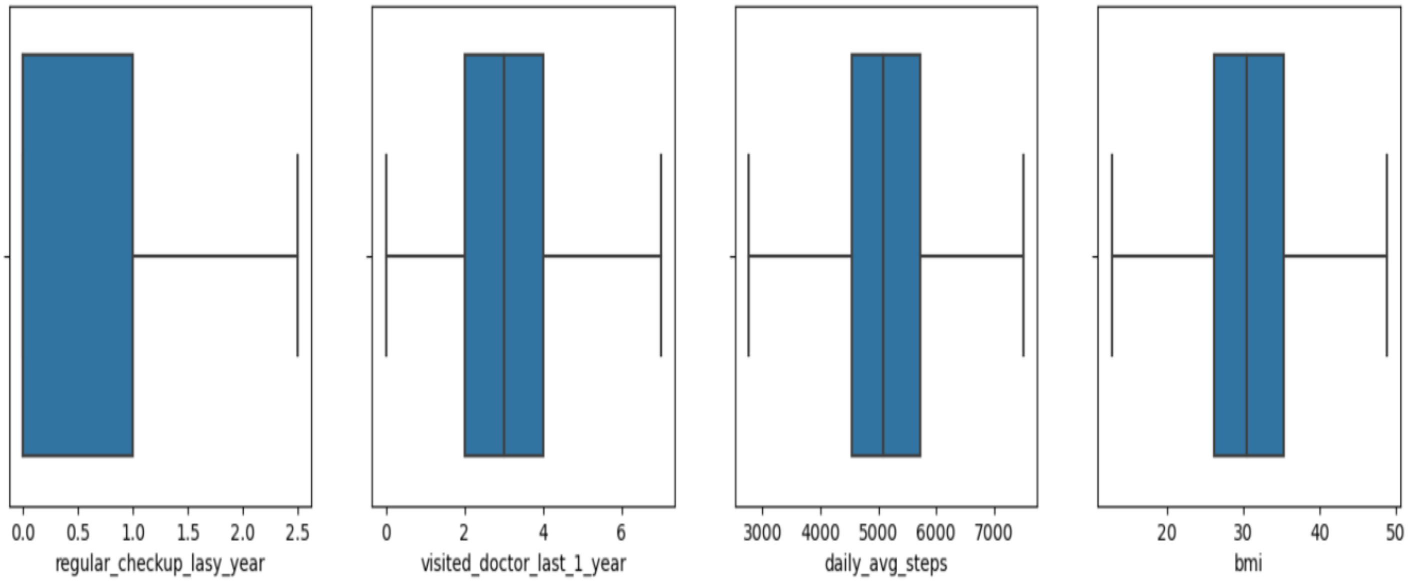


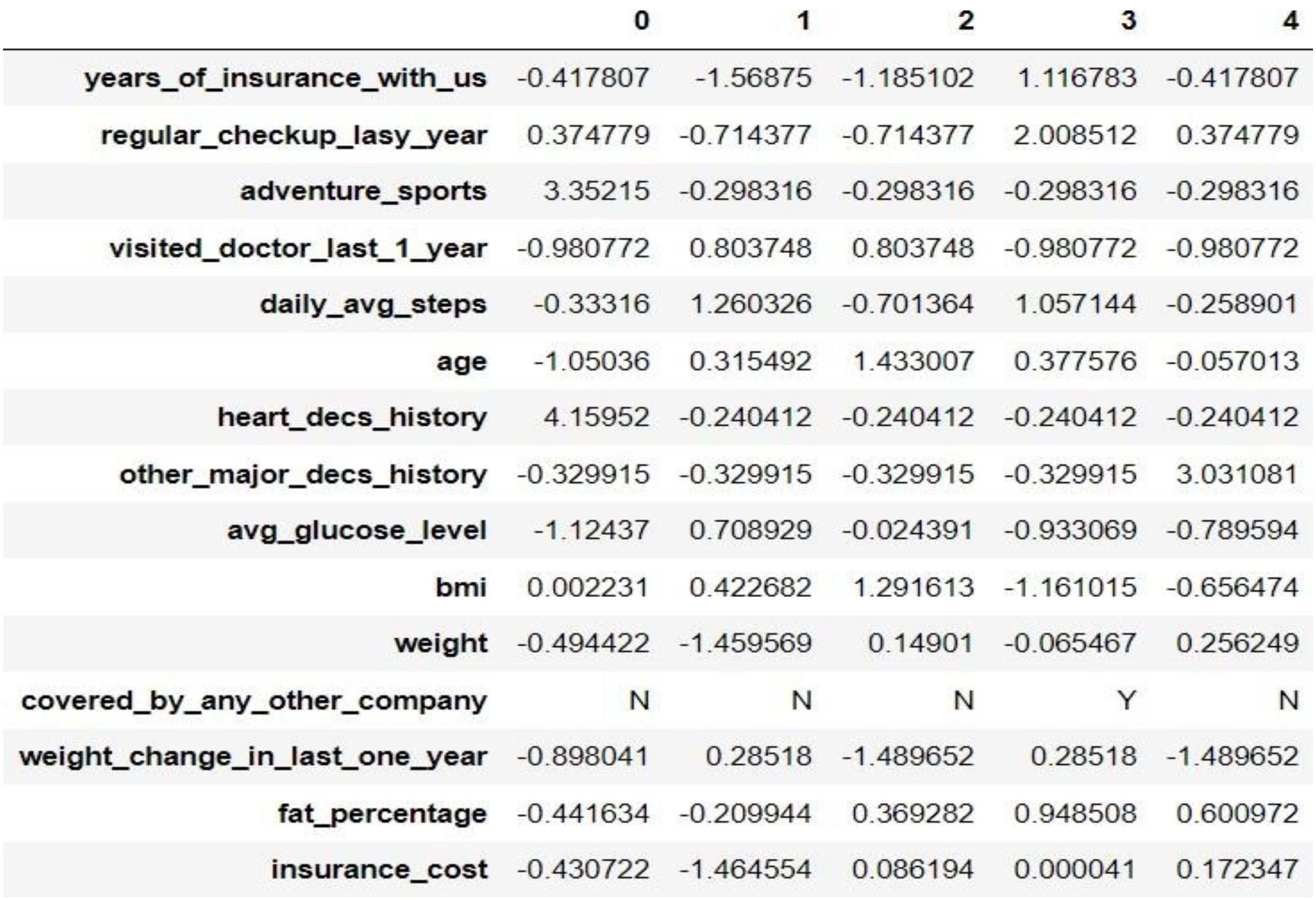
Figure 13: Boxplot after outlier treatments

We have successfully treated the outliers.

Variable transformation:

As we will build different ML models and each algorithm has its own condition so if algorithm demands scaling and we are not using scaled date than it's a problem but there is no problemto use scaled data for an algorithm which does not require scaling therefore it's better to scaled the data and use it for all algorithm & we are using z-score, or standard score for scalingwhich bring mean to 0 and standard deviation to 1.

Below is the sample of scaled data



	0	1	2	3	4
years_of_insurance_with_us	-0.417807	-1.56875	-1.185102	1.116783	-0.417807
regular_checkup_lasy_year	0.374779	-0.714377	-0.714377	2.008512	0.374779
adventure_sports	3.35215	-0.298316	-0.298316	-0.298316	-0.298316
visited_doctor_last_1_year	-0.980772	0.803748	0.803748	-0.980772	-0.980772
daily_avg_steps	-0.33316	1.260326	-0.701364	1.057144	-0.258901
age	-1.05036	0.315492	1.433007	0.377576	-0.057013
heart_decs_history	4.15952	-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	-1.12437	0.708929	-0.024391	-0.933069	-0.789594
bmi	0.002231	0.422682	1.291613	-1.161015	-0.656474
weight	-0.494422	-1.459569	0.14901	-0.065467	0.256249
covered_by_any_other_company	N	N	N	Y	N
weight_change_in_last_one_year	-0.898041	0.28518	-1.489652	0.28518	-1.489652
fat_percentage	-0.441634	-0.209944	0.369282	0.948508	0.600972
insurance_cost	-0.430722	-1.464554	0.086194	0.000041	0.172347

Table 5: Top 5 Rows of scaled data

Removal of unwanted variables:

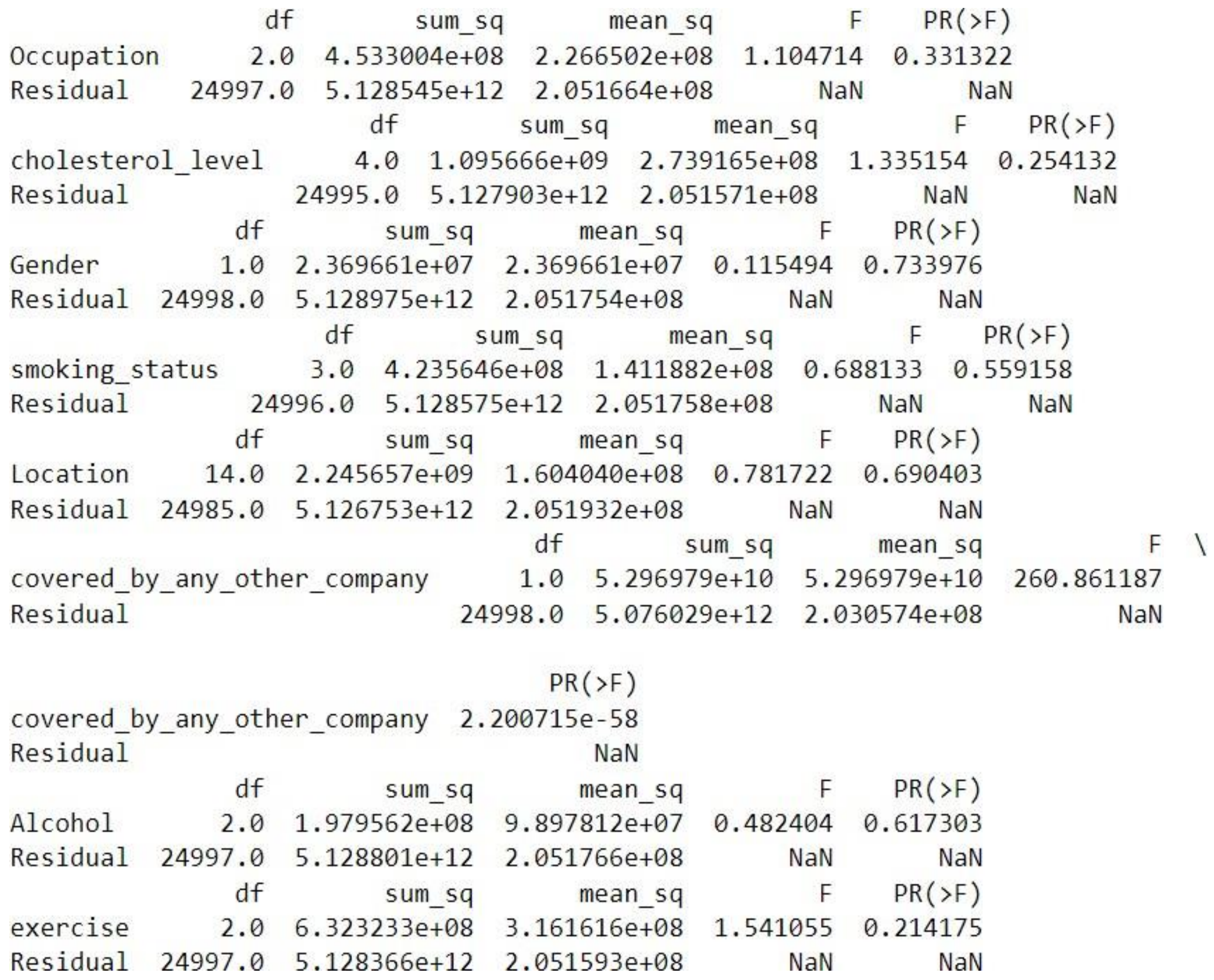
We have already removed 1 unwanted variable 'applicant_id' for now and will remove other variables if it's not significant for our model building.

Test for significance of categorical variables vs insurance_cost by using Anova, it estimates the extent to which a dependent variable is affected by one or more independent categorical data elements.

Null and Alternate hypothesis for anova test is following:

Null hypothesis H0: There is no statistically significant difference in the means of different groups of the independent variable.

Alternate Hypothesis H1: There is a statistically significant difference between the means of different groups of the independent variable.



```
                   df        sum_sq       mean_sq         F    PR(>F)
Occupation        2.0  4.533004e+08  2.266502e+08  1.104714  0.331322
Residual      24997.0  5.128545e+12  2.051664e+08       NaN       NaN
                          df        sum_sq       mean_sq         F    PR(>F)
cholesterol_level        4.0  1.095666e+09  2.739165e+08  1.335154  0.254132
Residual             24995.0  5.127903e+12  2.051571e+08       NaN       NaN
                 df        sum_sq       mean_sq         F    PR(>F)
Gender          1.0  2.369661e+07  2.369661e+07  0.115494  0.733976
Residual    24998.0  5.128975e+12  2.051754e+08       NaN       NaN
                       df        sum_sq       mean_sq         F    PR(>F)
smoking_status        3.0  4.235646e+08  1.411882e+08  0.688133  0.559158
Residual          24996.0  5.128575e+12  2.051758e+08       NaN       NaN
                 df        sum_sq       mean_sq         F    PR(>F)
Location       14.0  2.245657e+09  1.604040e+08  0.781722  0.690403
Residual    24985.0  5.126753e+12  2.051932e+08       NaN       NaN
                                  df        sum_sq       mean_sq           F  \
covered_by_any_other_company     1.0  5.296979e+10  5.296979e+10  260.861187   
Residual                     24998.0  5.076029e+12  2.030574e+08         NaN   

                                    PR(>F)  
covered_by_any_other_company  2.200715e-58  
Residual                               NaN  
                 df        sum_sq       mean_sq         F    PR(>F)
Alcohol         2.0  1.979562e+08  9.897812e+07  0.482404  0.617303
Residual    24997.0  5.128801e+12  2.051766e+08       NaN       NaN
                 df        sum_sq       mean_sq         F    PR(>F)
exercise        2.0  6.323233e+08  3.161616e+08  1.541055  0.214175
Residual    24997.0  5.128366e+12  2.051593e+08       NaN       NaN
```

Table 6: ANOVA test for significance results

Based on this anova test we are failed to reject the null hypothesis (p-Value not less than 0.05) for 7 variables hence these variables are dropped moving forward.

So, we cut down to 16 variables and only 1 categorical variable present now.

Addition of new variables:

As per looking at the variables don't see any variables needs to be added but we need to encode the categorical variable 'covered_by_any_other_company' as model will not read string data therefore, we are using One hot encoding to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1or 0 based on the presence or absence of the categorical value in the record.

In the data set, each Category in categorical column has been added as column with values 0 & 1 Example: covered_by_any_other_company_Y if covered_by_any_other_company_Y = 1,then it means customer is covered by other company and if covered_by_any_other_company_Y = 0 means customer not covered by other company.

Model building

Model selection:

As we know that we are trying to build a model which can provide optimum insurance cost and here targets variable insurance cost is a regression problem therefore we have built different regression models and chose best model after comparing all model performances.

Models that we have built:

- Linear Regression from stats model
- Linear Regression from Sklearn
- Decision Tree Regressor
- Random Forest Regressor
- Ridge Regressor
- Lasso Regressor
- XGBOOST (EXTREME GRADIENT BOOSTING) REGRESSOR

Ridge and Lasso Regression are regularization techniques used to prevent overfitting in linear regression models by adding a penalty term to the loss function (Alpha).

- L1 Regularization, also called a lasso regression, adds the “absolute value of magnitude” of the coefficient as a penalty term to the loss function.

- L2 Regularization, also called a ridge regression, adds the “squared magnitude” of the coefficient as the penalty term to the loss function.

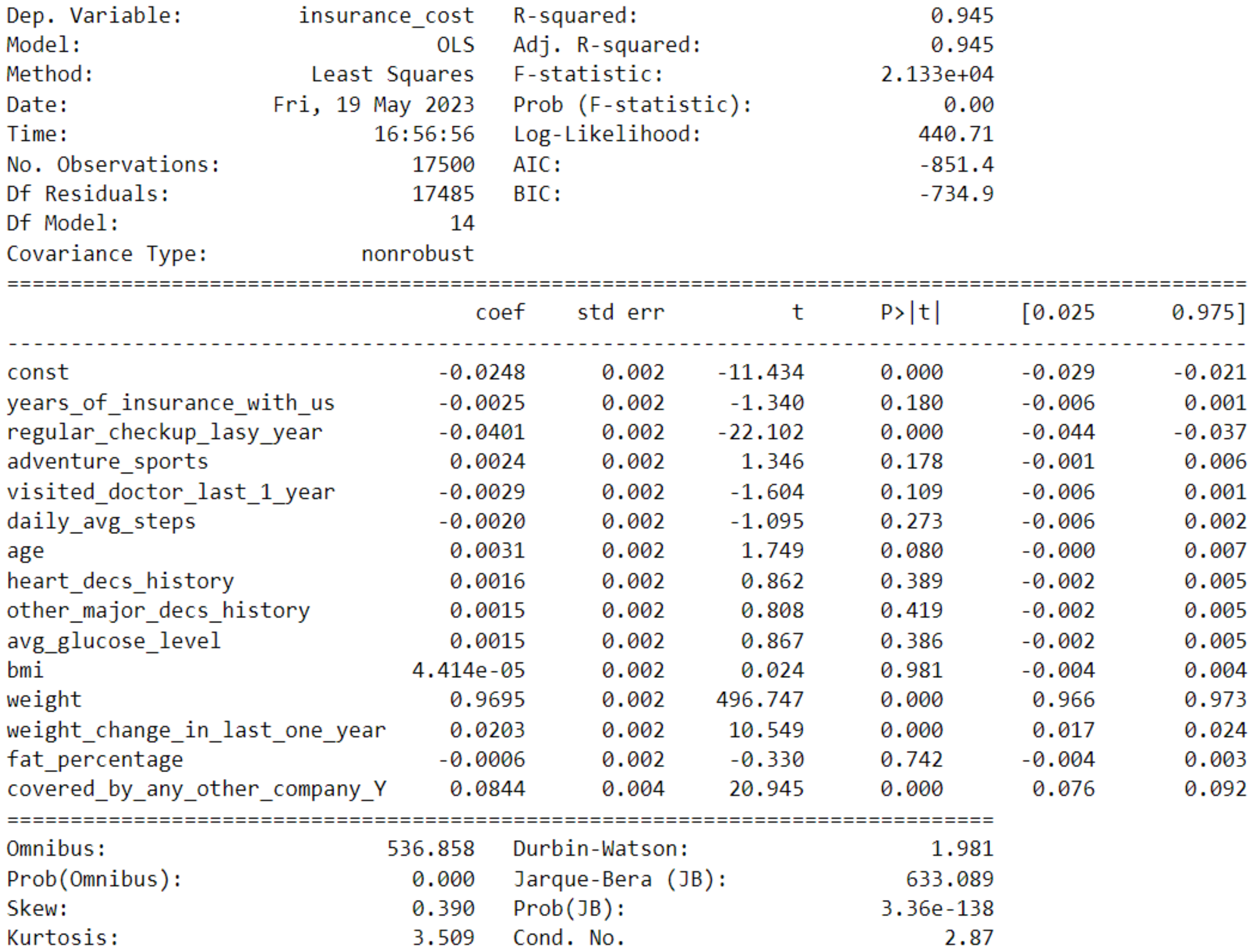
Sometimes, it may not be sufficient to rely upon the results of just one machine learning model. Ensemble learning offers a systematic solution to combine the predictive power of multiple learners. The resultant is a single model which gives the aggregated output from several models and We used 2 ensemble techniques bagging (Random Forest) and Boosting (XGB boost).

we are splitting data into train and test in 70:30 ratio where 70% data will be used to train the models and 30% data will be used to evaluate the model performance.

We have built above mentioned models by using their default/basic parameters and below are the insights/output from these models.

Linear regression from stats models:

We have built LR model with all independent variables



```
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 2.133e+04
Date:                Fri, 19 May 2023   Prob (F-statistic):               0.00
Time:                        16:56:56   Log-Likelihood:                 440.71
No. Observations:               17500   AIC:                            -851.4
Df Residuals:                   17485   BIC:                            -734.9
Df Model:                          14
Covariance Type:            nonrobust
===================================================================================================
                                      coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------------------
const                              -0.0248      0.002    -11.434      0.000      -0.029      -0.021
years_of_insurance_with_us         -0.0025      0.002     -1.340      0.180      -0.006       0.001
regular_checkup_lasy_year          -0.0401      0.002    -22.102      0.000      -0.044      -0.037
adventure_sports                    0.0024      0.002      1.346      0.178      -0.001       0.006
visited_doctor_last_1_year         -0.0029      0.002     -1.604      0.109      -0.006       0.001
daily_avg_steps                    -0.0020      0.002     -1.095      0.273      -0.006       0.002
age                                 0.0031      0.002      1.749      0.080      -0.000       0.007
heart_decs_history                  0.0016      0.002      0.862      0.389      -0.002       0.005
other_major_decs_history            0.0015      0.002      0.808      0.419      -0.002       0.005
avg_glucose_level                   0.0015      0.002      0.867      0.386      -0.002       0.005
bmi                              4.414e-05      0.002      0.024      0.981      -0.004       0.004
weight                              0.9695      0.002    496.747      0.000       0.966       0.973
weight_change_in_last_one_year      0.0203      0.002     10.549      0.000       0.017       0.024
fat_percentage                     -0.0006      0.002     -0.330      0.742      -0.004       0.003
covered_by_any_other_company_Y      0.0844      0.004     20.945      0.000       0.076       0.092
==============================================================================
Omnibus:                      536.858   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              633.089
Skew:                           0.390   Prob(JB):                    3.36e-138
Kurtosis:                       3.509   Cond. No.                         2.87
```

Table 7: OLS Regression Summary for Basic Model

The coefficients tell us how one unit change in X can affect y.

The sign of the coefficient indicates if the relationship is positive or negative

For this model we have calculated the RMSE for training data and it's: 0.23

The lower the RMSE, the better a given model is able to "fit" a dataset.

RMSE tells us the average distance between the predicted values from the model and the actual values in the dataset.

This is the basic Regression model with all independent variables similarly we have built Linear model by using Sklearn library which gives us the same accuracy but we are going With OLS regression because of its statistics advantage and it helps to reduce insignificant variables.

Decision Tree Regressor:

We have built DTR model with default parameters and the accuracy score of training data is 1 whereas RMSE is 0.0 which means this is an overfitted model.

We need to improve the performance of this DTR model and that is by using best parameters which we can obtained from GridsearchCV and will do this in our model tuning exercise.

Random Forest Regressor:

We have built RFR model with default parameters and the accuracy score of training data is 0.99 whereas RMSE is 0.08 which means this is an overfitted model.

We need to improve the performance of this RFR model and that is by using best parameters which we can obtained from GridsearchCV and will do this in our model tuning exercise.

Ridge Regressor:

We have built RR model with default parameters and the accuracy score of training data is 0.94 whereas RMSE is 0.23 both these score looks good but we can try to improve it further by using different value of penalty factor Alpha.

We can try to find out best alpha value out of multiple values and this can be obtained from GridsearchCV and will do this in our model tuning exercise.

Below are the coefficients generated from Ridge model -

```
[-2.47648913e-03 -4.00594630e-02  2.42179351e-03 -2.92394017e-03
 -1.99116430e-03  3.12166795e-03  1.56355554e-03  1.48257243e-03
  1.54686691e-03  4.42300271e-05  9.69434375e-01  2.03082527e-02
 -5.90116595e-04  8.43384507e-02]
```

Lasso Regressor:

We have built Lasso model with default parameters and the accuracy score of training data is 0.0 whereas RMSE is 1 both these scores don’t look good but we can try to improve it further by using different value of penalty factor Alpha.

We can try to find out best alpha value out of multiple values and this can be obtained from GridsearchCV and will do this in our model tuning exercise.

XGBoost (extreme Gradient Boosting) Regressor:

it's an advanced implementation of the gradient boosting algorithm, Extreme gradient boosting is a tree + boosting technique where target variable is predicted by combing the estimates of a set of simpler and weaker models. This gives high priority to weaker models in predicting target variable and gives priority to the weaker models in next iteration to make it stronger. Also, it does L1 and L2 regularization while reducing the complexity and suits high multicollinearity and thus gets a name Extreme gradient boosting.

We have built XGB model with default parameters and the accuracy score of training data is 0.97 whereas RMSE is 0.15 both these score looks good but we can try to improve it further by using best parameters which we can obtained from GridsearchCV and will do this in our model tuning exercise.

Model Performance improvement:

As we have seen from above basic models that there is scope of improvement for these models and that can be performed by using GridSearchCV in which we can use different hyperparameters combination for each algorithm and choose a model with best parameters this is also called hyperparameters tunning.

We have built 2 models which are based on **ensemble learning**

1. Random Forest
2. XGB Regressor

So, let's try to tune these models now

XGB Regressor Tuning:

As a part of tuning of hyperparameters, parameters like n_estimators and max_depth has been passed through a dictionary with these following values,

n_estimators = [401,801]
max_depth = [5,10,15]

And after running GridsearchCV following parameters found to be best parameters

n_estimators = [401]
max_depth = [5]

A model has been built by using these parameters.

Random Forest Regressor Tuning:

As a part of tuning of hyperparameters, parameters like min_samples_split, n_estimators and max_depth has been passed through a dictionary with these following values

min_samples_split = 10, 50
max_depth= 5,10
n_estimators= 301, 501

And after running GridsearchCV following parameters found to be best parameters

min_samples_split = 50
max_depth= 10
n_estimators= 501

A model has been built by using these parameters.

Decision Tree regressor pruning

As a part of tuning of hyperparameters, parameters like criterion, min_samples_leaf, min_samples_split and max_depth has been passed through a dictionary with these following values

Criterion= squared_error, absolute_error
max_depth= 10,20,30,50
min_samples_leaf= 50,100,150
min_samples_split = 150,300,450

And after running GridsearchCV following parameters found to be best parameters

Criterion= squared_error
max_depth= 10
min_samples_leaf= 50
min_samples_split = 150

A model has been built by using these parameters

Ridge Regressor Tuning

Ridge regression a similar form of Lasso regression where both regressions put a toll on coefficients by putting a penalty factor but with a difference where Lasso takes the magnitude of coefficients while ridge takes square on the coefficients. This is L2 regularization technique.

After trying different values of Alpha in GridsearchCV we got best value of alpha as 0.01 and model has been built with this alpha value.

A model has been built by using these parameters

Lasso Regressor Tuning

After trying different values of Alpha in GridsearchCV we got best value of alpha as 0.01 and model has been built with this alpha value.

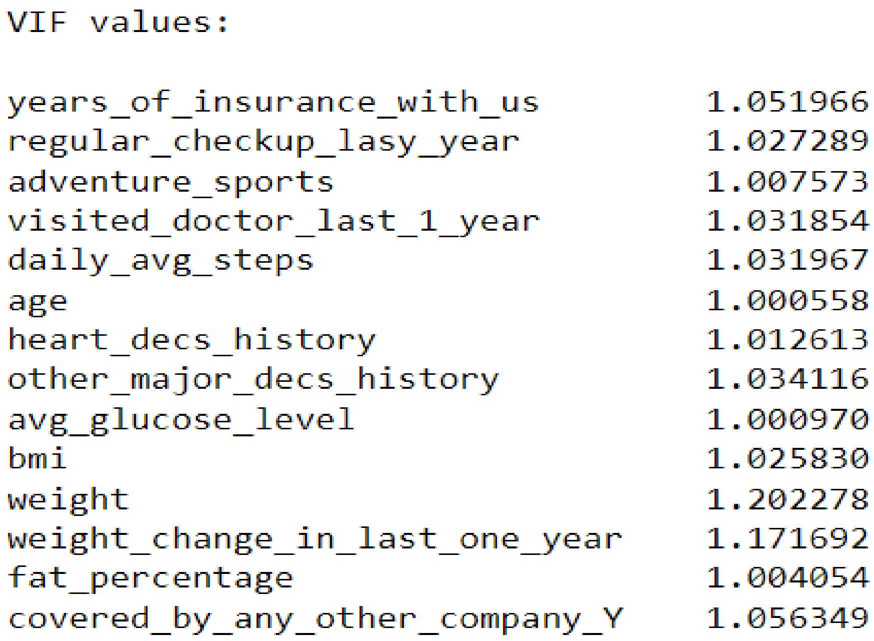
A model has been built by using these parameters.

Linear Regression Tuning

We have tuned all models except Linear regression because we are doing multicollinearity check and if our accuracy doesn't reduce even with lesser number of variables then this will be very good model considering that it giving accuracy of 94% with few variables so let's work on Linear regression now.

As we have seen linear regression model performance from OLS Regression Results, we got R squared value around 94% for both train and test dataset which shows no overfitting in our model and R2 is an increasing function of the number of independent variables i.e, with the inclusion of one more independent variable R2 is likely to increase or at least will not decrease.

We are trying to find out significant variables which really contribute in prediction and to do that we are checking multicollinearity However, there is a definitive tangible method to analyse the variance inflation factor (VIF) used to check multicollinearity and removing them iteratively to avoid heteroscedasticity in the model. Based on the base model, below are the VIF values obtained.



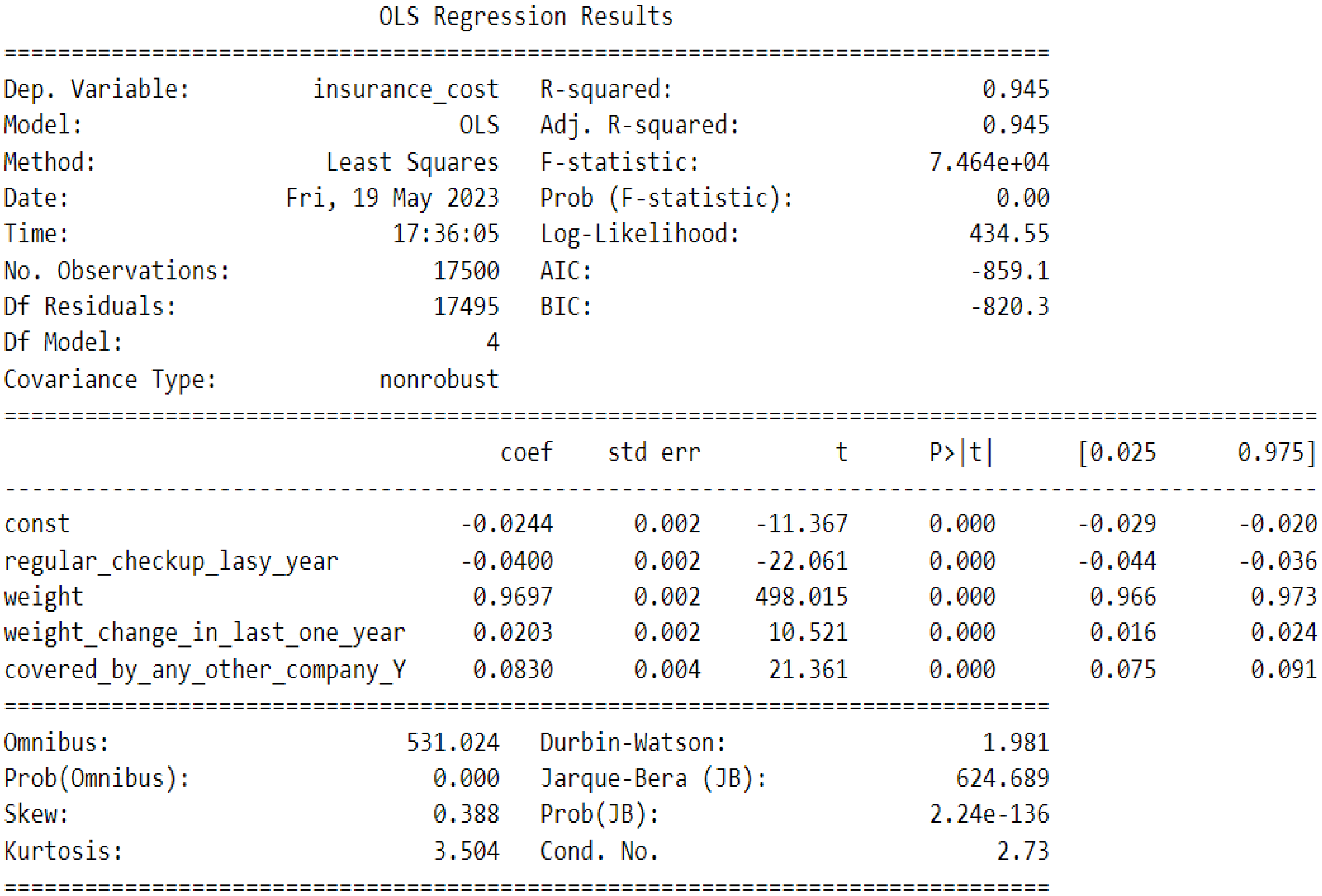
```
VIF values:

years_of_insurance_with_us         1.051966
regular_checkup_lasy_year          1.027289
adventure_sports                   1.007573
visited_doctor_last_1_year         1.031854
daily_avg_steps                    1.031967
age                                1.000558
heart_decs_history                 1.012613
other_major_decs_history           1.034116
avg_glucose_level                  1.000970
bmi                                1.025830
weight                             1.202278
weight_change_in_last_one_year     1.171692
fat_percentage                     1.004054
covered_by_any_other_company_Y     1.056349
```

Table 8: VIF scores of variables

There is no multi collinearity within all the variables subjected to base model. Since all the VIF values are within the range (not more than 5), it is better to check P-values obtained and make decision for elimination of variables in the consecutive model.

Let's review OLS summary from Table 7 and drop variables one by one for which P-Value is greater than 0.05 and first we are dropping variables which has high p-value in short, P-value arranged in Descending Order and then dropped variables one after one.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 7.464e+04
Date:                Fri, 19 May 2023   Prob (F-statistic):               0.00
Time:                        17:36:05   Log-Likelihood:                 434.55
No. Observations:               17500   AIC:                            -859.1
Df Residuals:                   17495   BIC:                            -820.3
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                             -0.0244      0.002    -11.367      0.000      -0.029      -0.020
regular_checkup_lasy_year         -0.0400      0.002    -22.061      0.000      -0.044      -0.036
weight                             0.9697      0.002    498.015      0.000       0.966       0.973
weight_change_in_last_one_year     0.0203      0.002     10.521      0.000       0.016       0.024
covered_by_any_other_company_Y     0.0830      0.004     21.361      0.000       0.075       0.091
==============================================================================
Omnibus:                      531.024   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              624.689
Skew:                           0.388   Prob(JB):                    2.24e-136
Kurtosis:                       3.504   Cond. No.                         2.73
==============================================================================
```

Table 9: OLS results with significant variables

We had 14 variables in First OLS summary and we cut down to 4 variables in Final OLS summary so we dropped 10 insignificant variables.

Model validation

Model Evaluation:

Let’s Evaluate performance of above tunned models.

XGB Regressor Tuning:

```
R Square on training data: 0.9846940262693852
R Square on testing data: 0.947405605312895
RMSE on training data: 0.1241138081708206
RMSE on testing data: 0.227605955899965
MAPE on training data: 8.826275872996527
MAPE on testing data: 10.374904538355796
```

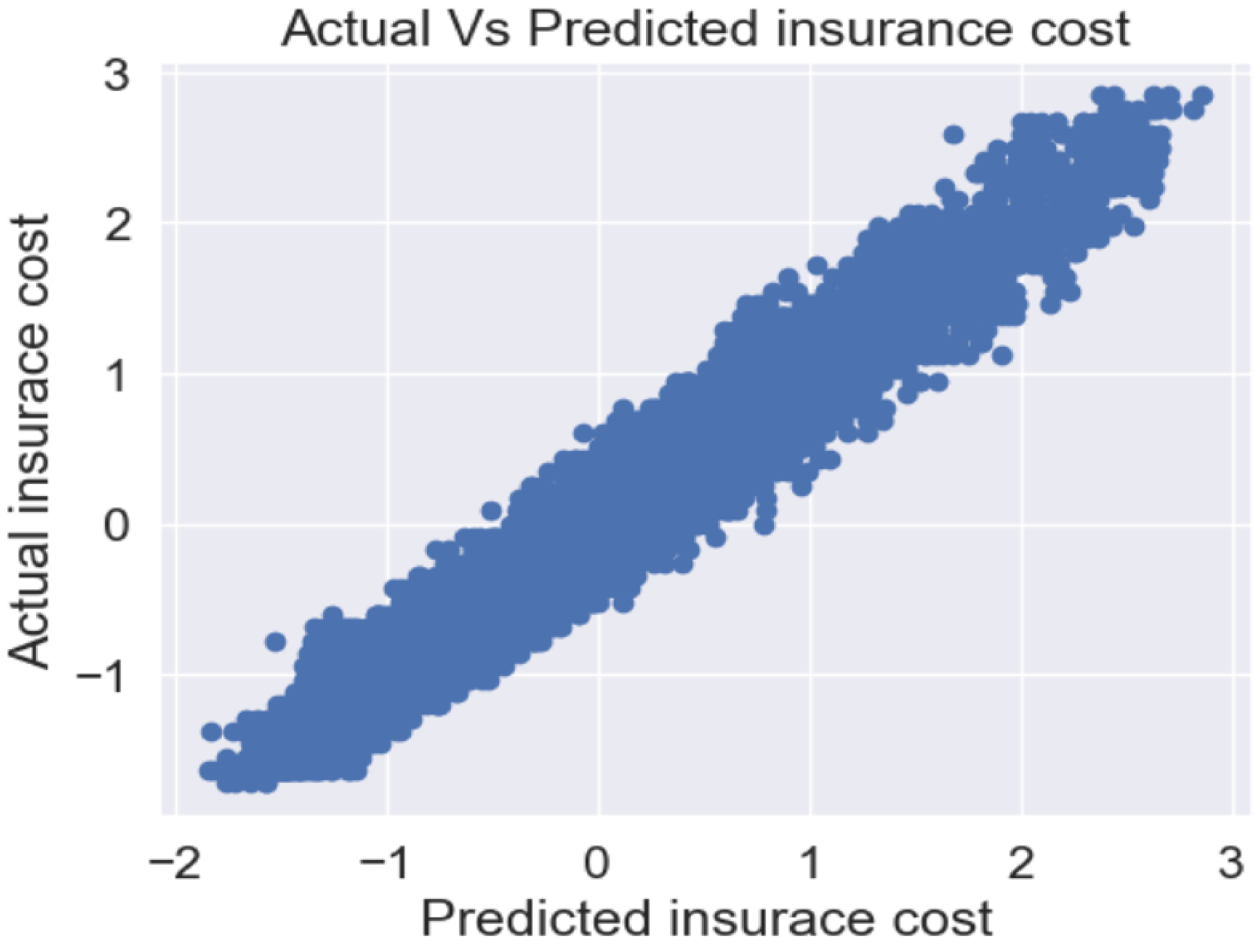


Figure 14: Scatter plot B/w Actual & predicted insurance cost from XGB Model for test data

Based on the results obtained, it is an overfitted model even with pruning of hyperparameters. As per training and test scores there are overfitting issues in this and hence this model is not suitable for the implementation stage. this model is performing very well on training data but will not perform same way in testing/unseen data therefore this model is not suitable. XGB is mostly used for weaker models. But since random forest regressor gave accurate results, XGB model not significant model.

Random Forest Regressor Tuning

```
R Square on training data: 0.9627952244699269
R Square on testing data: 0.9532283325135238
RMSE on training data: 0.19350356192957943
RMSE on testing data: 0.21463736026277902
MAPE on training data: 10.739420822975772
MAPE on testing data: 10.083920367460228
```

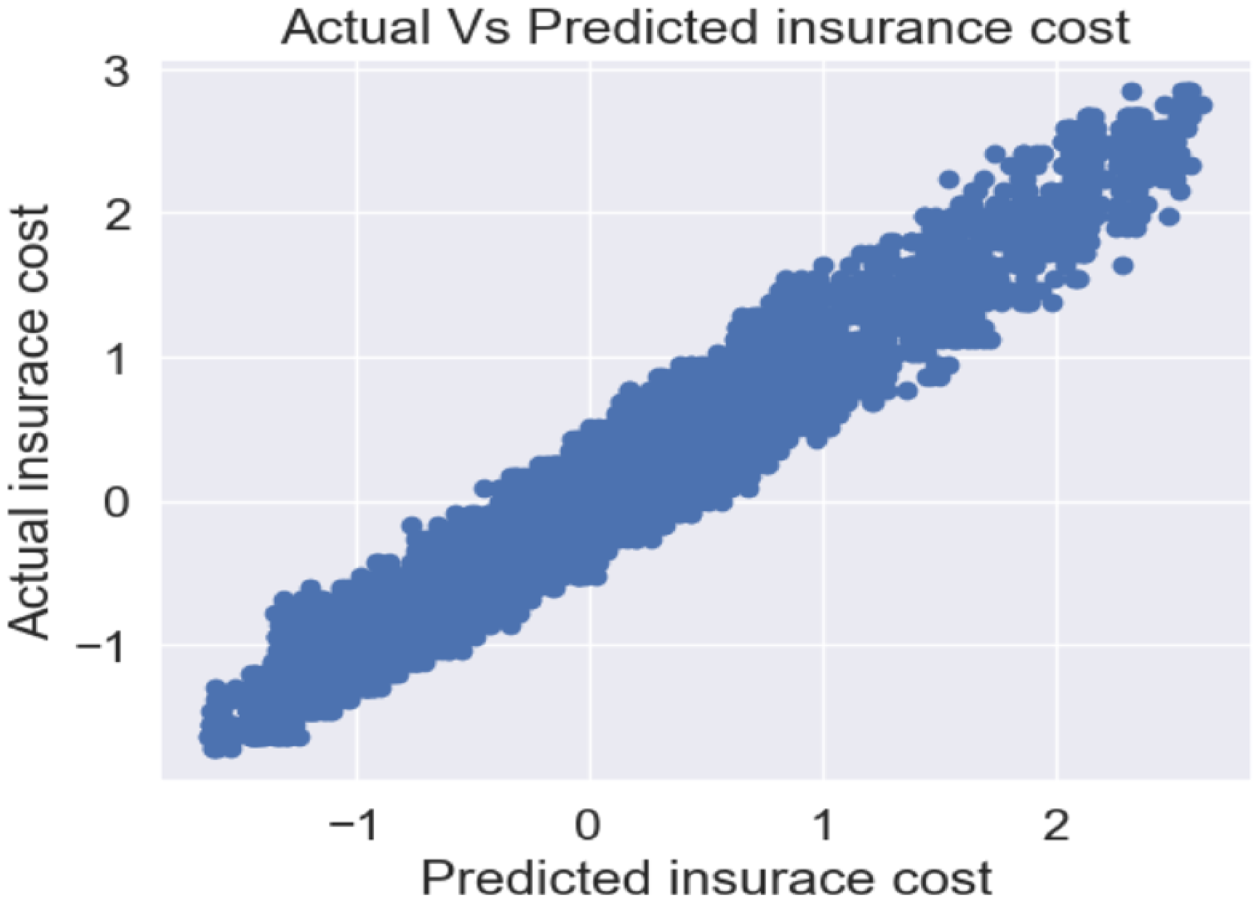


Figure 15: Scatter plot B/w Actual & predicted insurance cost from RFR Model for test data

As we can see from above performance metrics that R square has improved and MAPE also reduced after using different sets of parameters in Random Forest regressor

We have plotted actual and predicted values on scatter plot and we can see linear relation between Actual and predicted insurance costs.

Decision Tree regressor pruning

```
R Square on training data: 0.9570376064848412
R Square on testing data: 0.9519873731369077
RMSE on training data: 0.20793799582429578
RMSE on testing data: 0.21746612985057712
MAPE on training data: 10.990872423273371
MAPE on testing data: 10.110111034548343
```

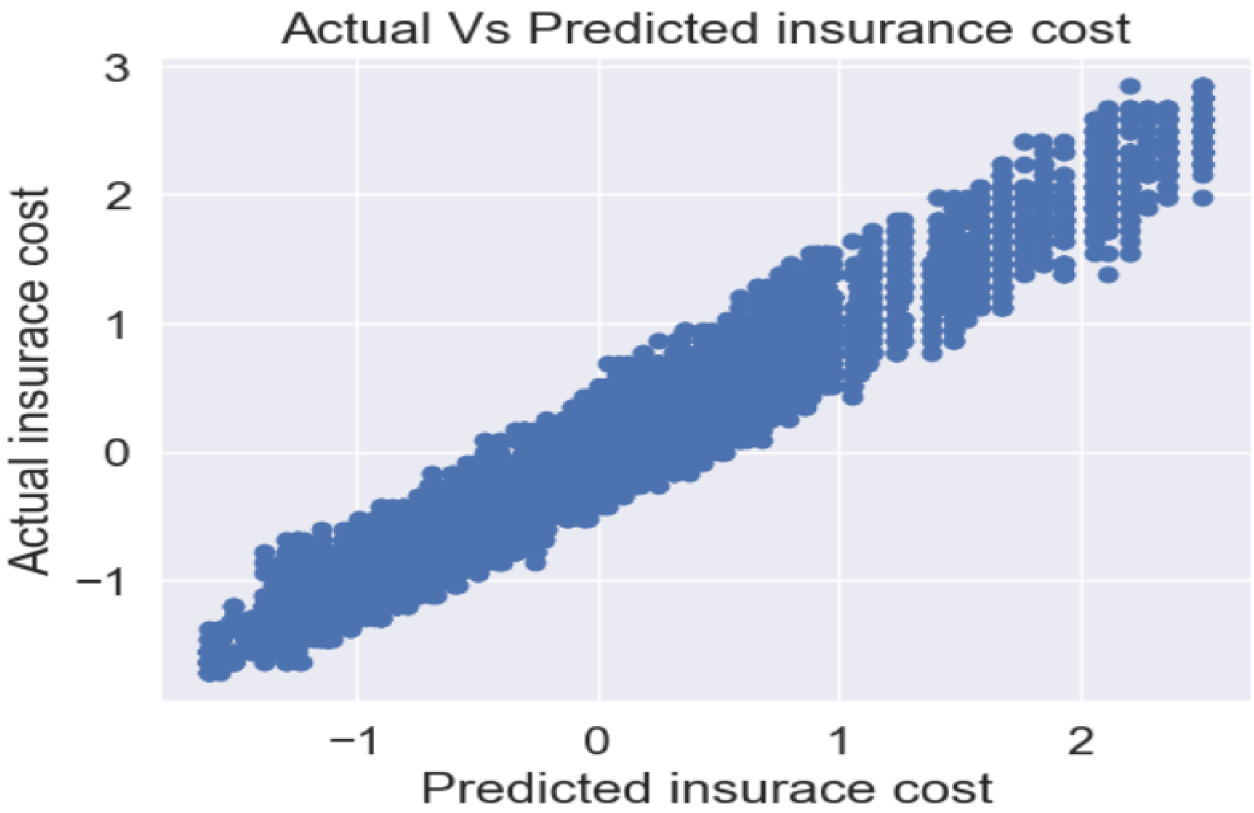


Figure 16: Scatter plot B/w Actual & predicted insurance cost from DTR Model for test data

As we can see from above performance metrics that R square has improved and MAPE also reduced after using different sets of parameters in decision tree regressor.

We have plotted actual and predicted values on scatter plot and we can see linear relation between Actual and predicted insurance costs.

Ridge Regressor Tuning

```
R Square on training data: 0.9446812482764967
R Square on testing data: 0.9449781252252026
RMSE on training data: 0.23595316392813653
RMSE on testing data: 0.23279925373966812
MAPE on training data: 11.184615897593998
MAPE on testing data: 10.111299053324037
```

Coefficients:

```
Ridge model: [-2.47805009e-03 -4.00521978e -02  2.41799555e-03 -2.92496675e-03
              -1.99131508e-03  3.12211234e -03  1.56341369e-03  1.48275746e-03
               1.54727310e-03  4.41359926e -05  9.69499392e-01  2.03341456e-02
              -5.90058605e-04  8.43551672e-02]
```

As we can see from above performance metrics that R square & RMSE has not improved much it almost same compare to the base model of Ridge and MAPE also not reduced even after using best value of alpha in Ridge regressor which indicate that base model already given best possible result on this dataset.

We have plotted actual and predicted values on scatter plot and we can see linear relation between Actual and predicted insurance costs.

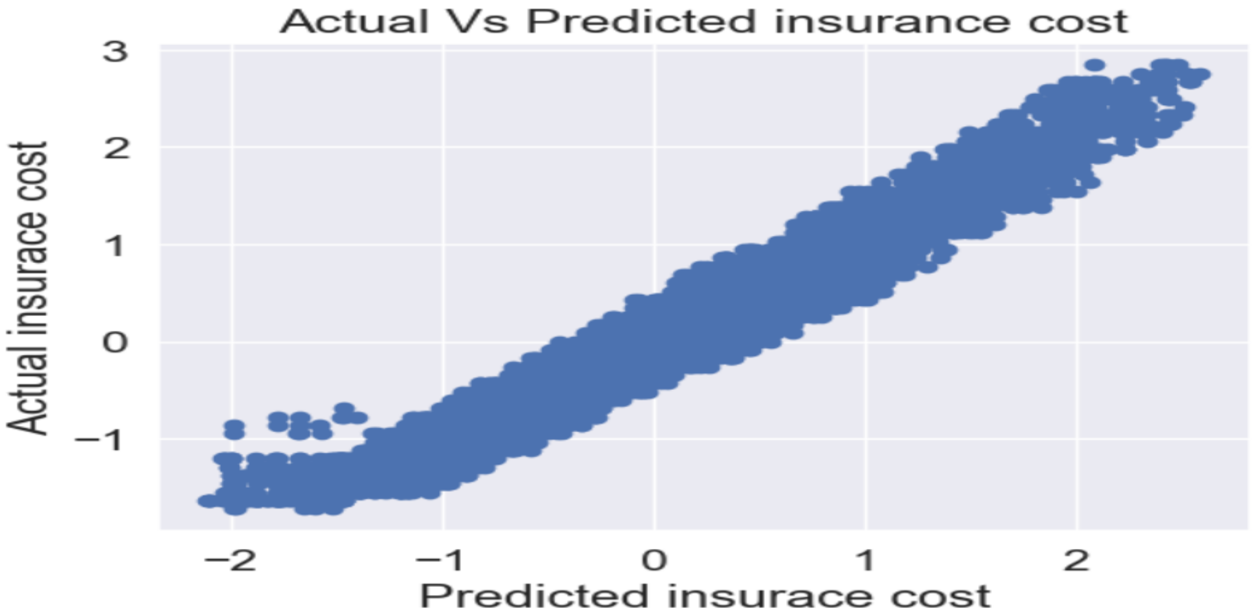


Figure 17: Scatter plot B/w Actual & predicted insurance cost from Ridge Model for test data

Lasso Regressor Tuning

```
R Square on training data: 0.9437953514570824
R Square on testing data: 0.944357953169224
RMSE on training data: 0.23783498439881917
RMSE on testing data: 0.23410756103894342
MAPE on training data: 11.177484525320617
MAPE on testing data: 10.137942966442264
```

Coefficients:

```
Lasso model: [ 0.        -0.03159012       0.          -0.        -0.            0.
               0.         0.               0.           0.         0.95638755   0.00414773
              -0.         0.03554509]
```

From GridsearchCV we got best alpha value 0.01 and built Lasso regression model but there is no improvement in accuracy it's almost same compare to base model but advantage of this tuned model is that it minimizes the coefficient to zero for 10 variables.

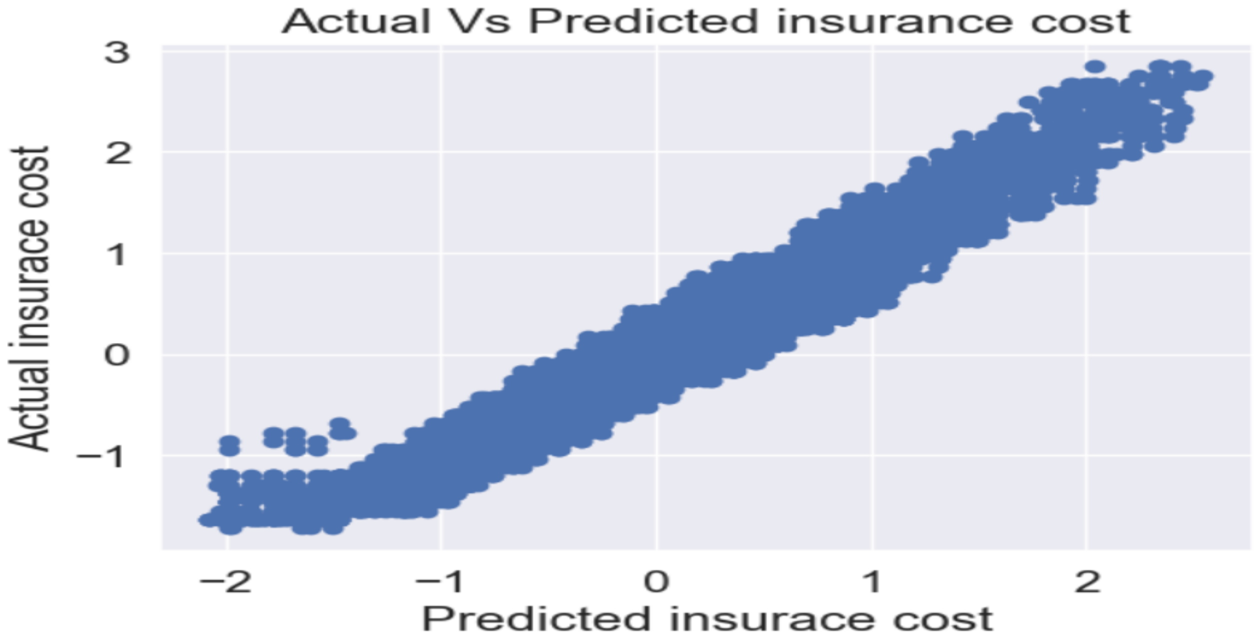


Figure 18: Scatter plot B/w Actual & predicted insurance cost from Lasso Model for test data

Linear Regression Tuning

```
RMSE on training data: 0.23603616946998815
RMSE on testing data: 0.2329569973683786
MAPE on training data: 11.18705465945835
MAPE on testing data: 10.102142070274516
```

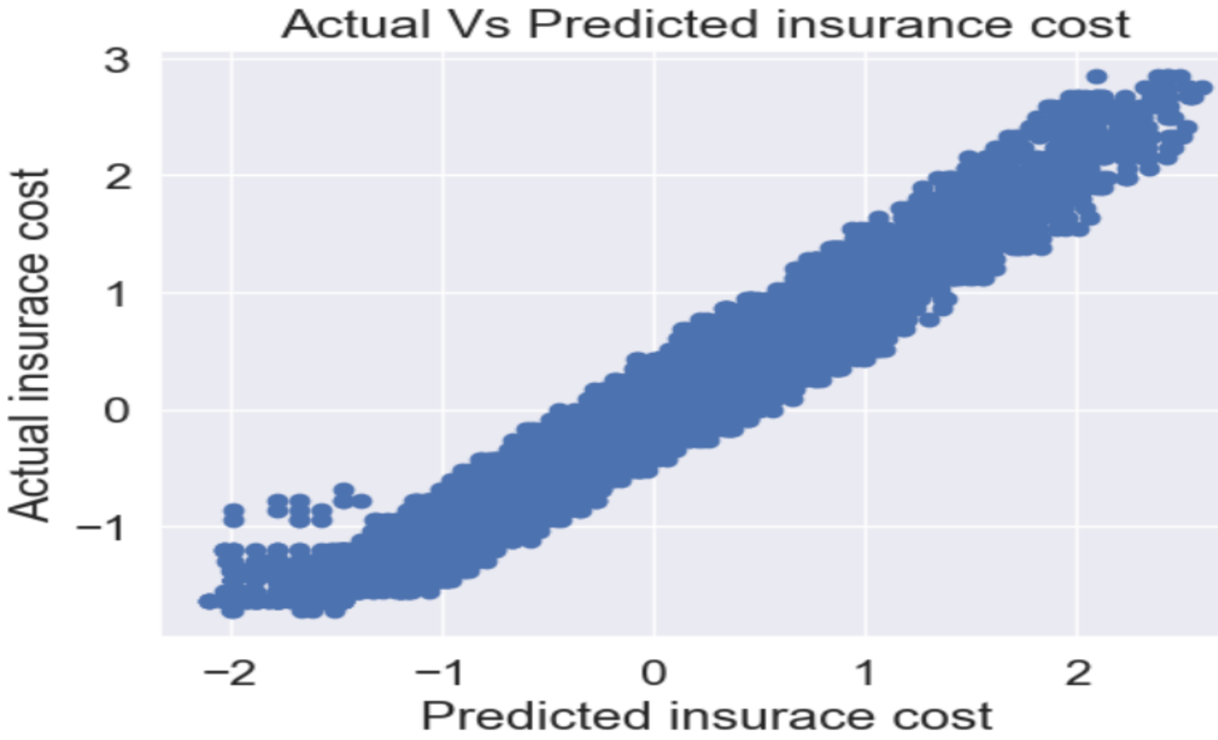
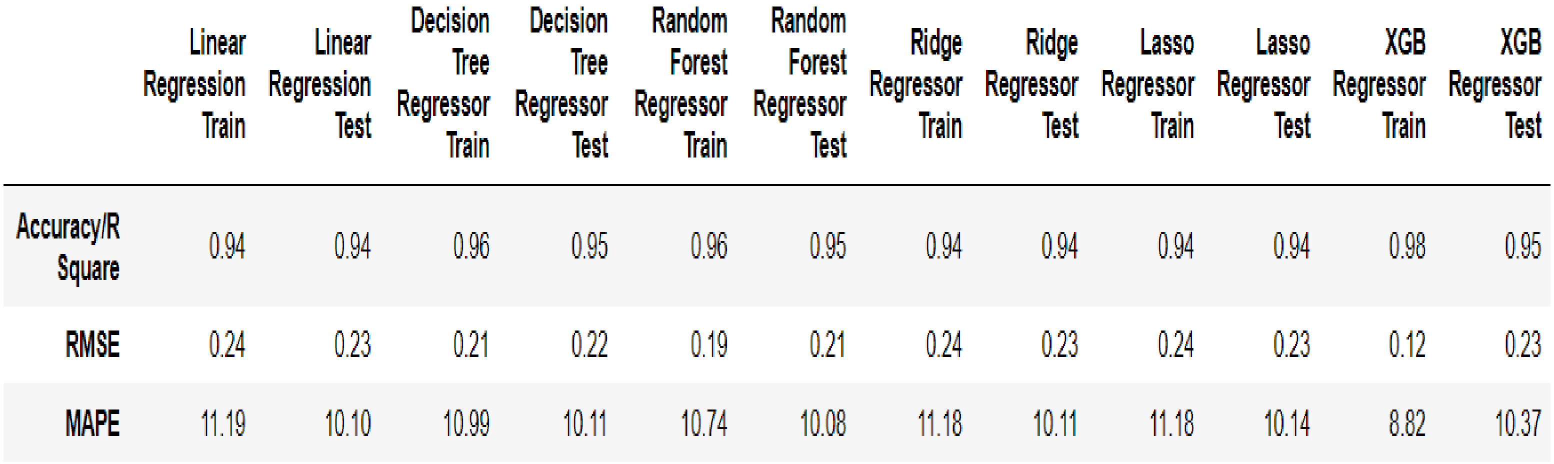


Figure 19: Scatter plot B/w Actual & predicted insurance cost from Linear Model for test data

From above model we got most significant variables with their coefficients and from above scatter plot between actual Vs predicted values a linear relation finds out.

Model’s performance Comparison:



	Linear Regression Train	Linear Regression Test	Decision Tree Regressor Train	Decision Tree Regressor Test	Random Forest Regressor Train	Random Forest Regressor Test	Ridge Regressor Train	Ridge Regressor Test	Lasso Regressor Train	Lasso Regressor Test	XGB Regressor Train	XGB Regressor Test
Accuracy/R Square	0.94	0.94	0.96	0.95	0.96	0.95	0.94	0.94	0.94	0.94	0.98	0.95
RMSE	0.24	0.23	0.21	0.22	0.19	0.21	0.24	0.23	0.24	0.23	0.12	0.23
MAPE	11.19	10.10	10.99	10.11	10.74	10.08	11.18	10.11	11.18	10.14	8.82	10.37

Table 10: Accuracy Comparison of Models

Based on Comparison table we can see that all model's performance is quite good except (XGB Due to overfitting) but if we have to pick one model out of all then it will be Decision Tree Regressor because of it's lowest MAPE and high R Squared.

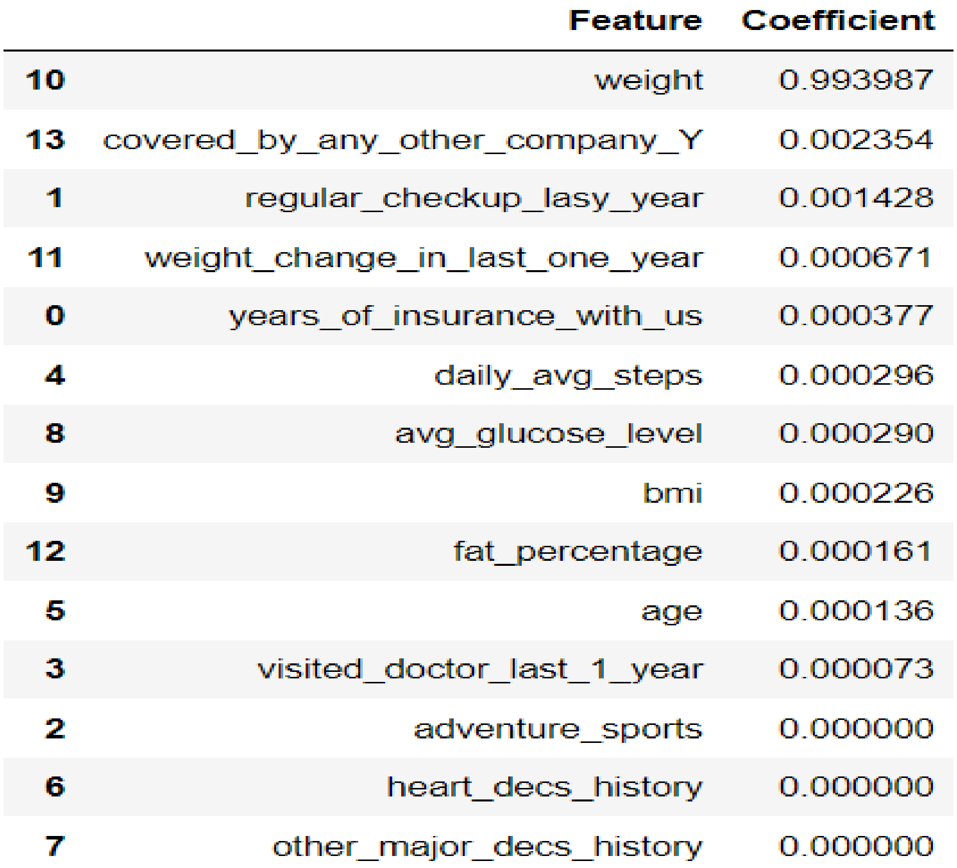
→ **Key takeaways from best DTR model**:

R-squared of 95 indicate that 95% of variance in dependent variable(insurance_Cost) has been explained by independent variables collectively.

RMSE score of 0.22 indicate avg. difference between predicted and actual values as it's an error so we expect it to be lowest.

MAPE score of 10 indicate that DTR model predictions are, on average, off by 10% from the Actual values.

→ **Feature Importance from DTR Model:**



	Feature	Coefficient
10	weight	0.993987
13	covered_by_any_other_company_Y	0.002354
1	regular_checkup_lasy_year	0.001428
11	weight_change_in_last_one_year	0.000671
0	years_of_insurance_with_us	0.000377
4	daily_avg_steps	0.000296
8	avg_glucose_level	0.000290
9	bmi	0.000226
12	fat_percentage	0.000161
5	age	0.000136
3	visited_doctor_last_1_year	0.000073
2	adventure_sports	0.000000
6	heart_decs_history	0.000000
7	other_major_decs_history	0.000000

Table 11: Important Features Based on DTR Model

Final interpretation / recommendation

Insights from this Case study:

1. From EDA we got to know the significance of categorical variables and only 1 variable was significant.

2. From heatmap we don't see a high correlation b/w variable.

3. Most of the customers cholesterol_level is normal and they don't smoke.

4. As we have seen in scatter plot that weight is correlated to insurance cost so we can interpret that weight b/w 78 to 96 leads to the high insurance cost.

5. From DTR feature importance, 3 variables: 'other_major_decs_history', 'heart_decs_history', 'adventure_sports' have 0 Coefficient that means these variables have no contribution in predicting Insurance Cost therefore these are not important so these can be ignored.

6. Weight is the most important variable as it alone has coefficient of 0.99 that means this variable has 99% contribution in predicting insurance cost and same, we observed in Heatmap that this variable was highly correlated with Target variable and our model also proved it.

7. Weight being an important variable sounds logical as we know that overweight has high effect on health it can be reason of many diseases and therefore other variables such as BMI, Fat_percentage, Avg glucose level, heart_decs, other_decs all these are related to weight if weight is in control, then these variables will also be in control in most of the cases except genetics-based disease.

Recommendations:

1. As we have seen that all model performed quite well and Linear regression provided 4 most important variables ('regular_checkup_lasy_year', 'weight', 'weight_change_in_last_one_year', 'covered_by_any_other_company_Y') so these variables need to be checked properly.

2. Out of those 4 variables one variable 'Weight' was found very important in EDA where we observed that target variable is highly correlated with this variable.

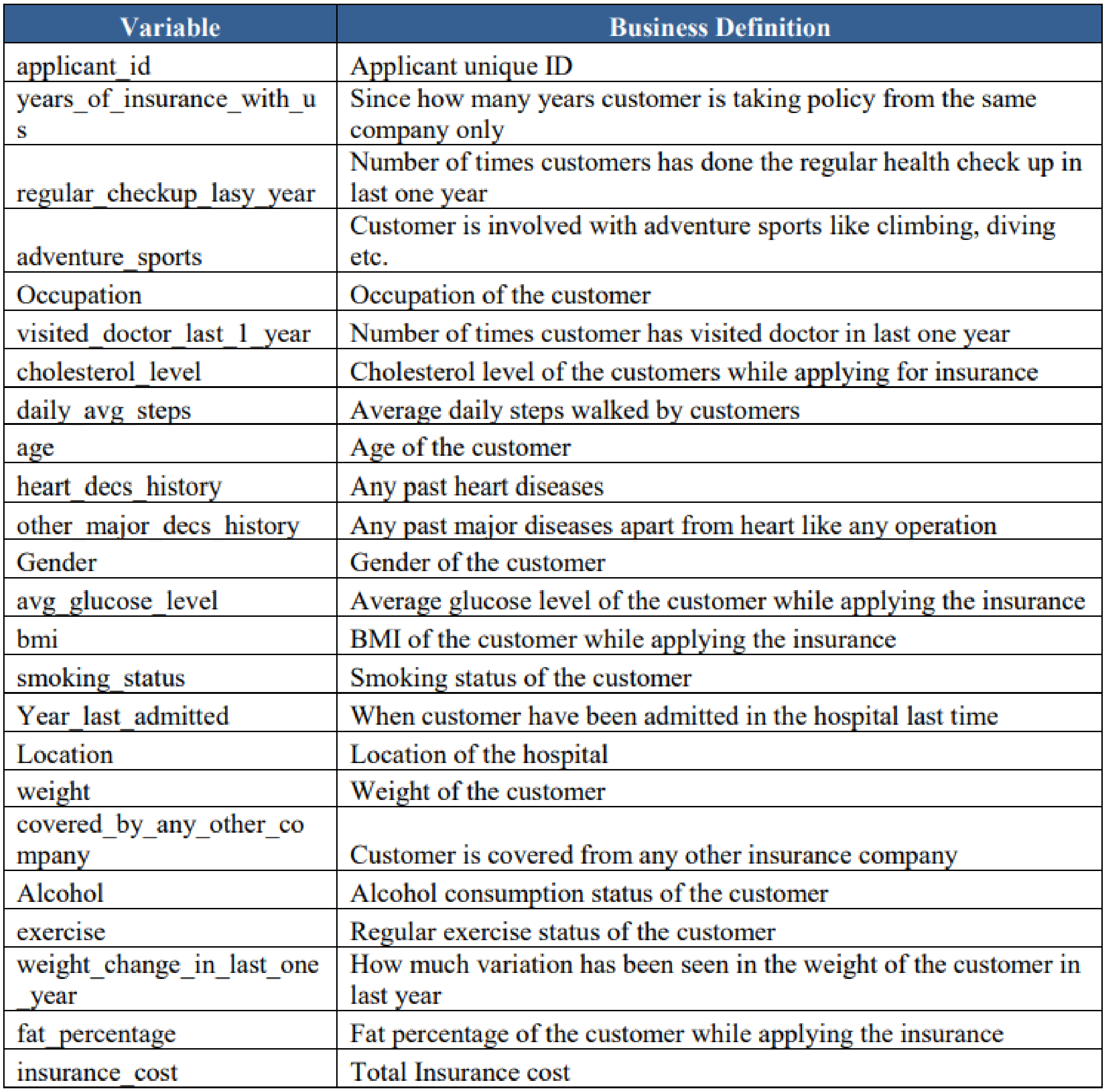
3. These are 3 non important variable as per DTR model adventure_sports, heart_decs_history, other_major_decs_history and same we observed in our linear equation that these variables were removed as these were non-significant so company don't need to give more weightage to these variables.

4. There might be few important variables such as past claim history, past claim settlement amount or any other claim related details because these details can provide additional information about an individual's frequent illness, disease, cause of illness etc.

5. As we have seen that Weight is very important variable for predicting insurance cost and there is similar variable present which is ‘Weight change in a year’ but this variable is not so important for prediction based on feature importance and this might be because change in weight is not specified whether weight has increased or it decreased in last 1 year so this needs to be noted carefully while collecting data whether weight change leading to weight increase or decrease and if this is obtained then this variable might be very helpful to find out individuals recent activities towards his/her health.

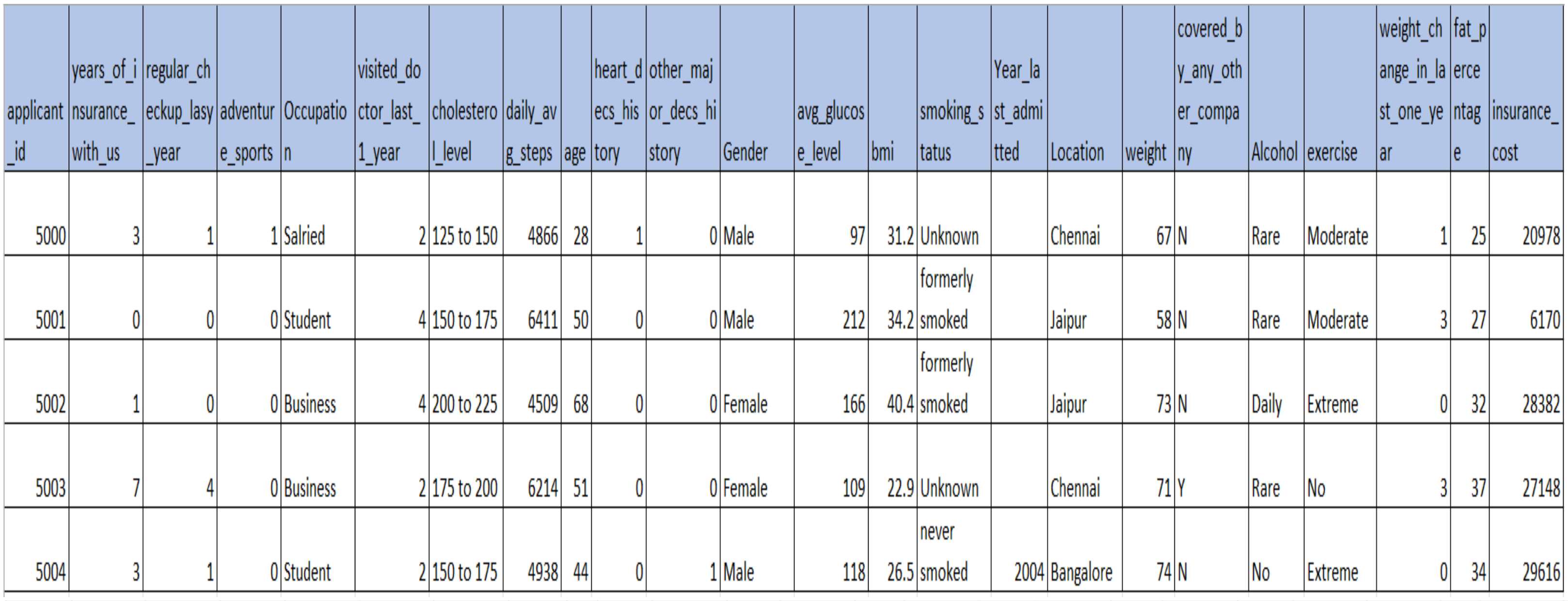
6. We know there are some customers who is associated with the same insurance company for more than 5 years so these are valuable customers and company can use these customer's feedback/survey to identified reason of their trust which can be used while giving plan to other new/not happy customers.

Appendix

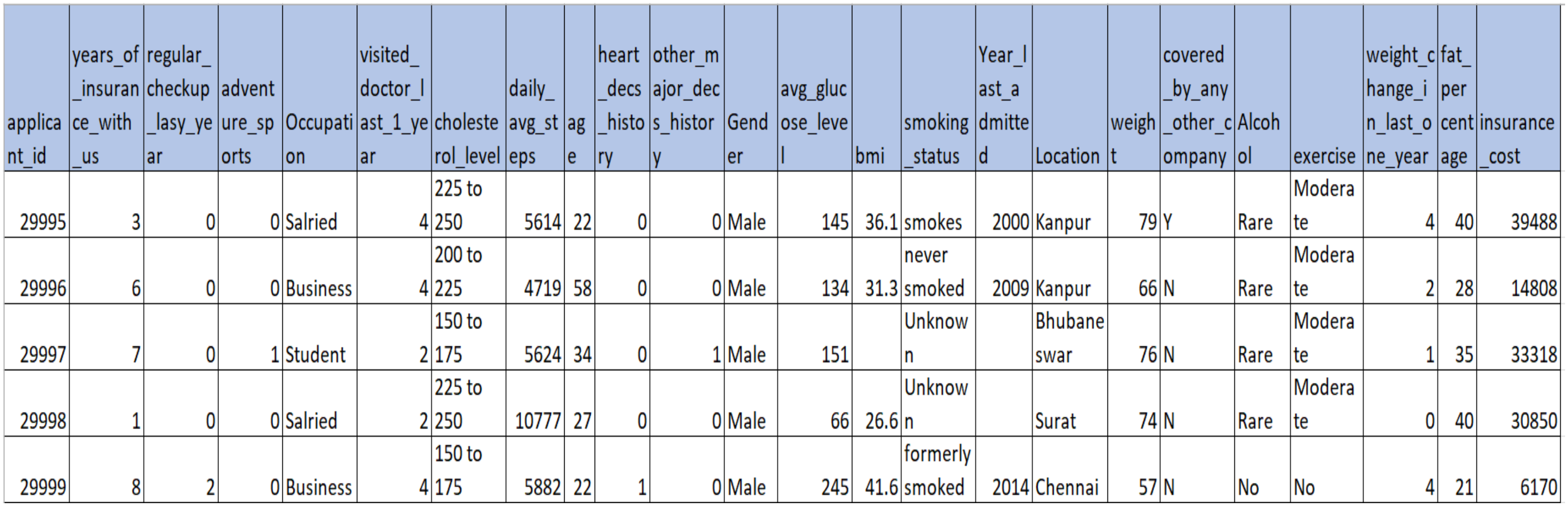


Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
adventure_sports	Customer is involved 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

Table 12: Data Dictionary

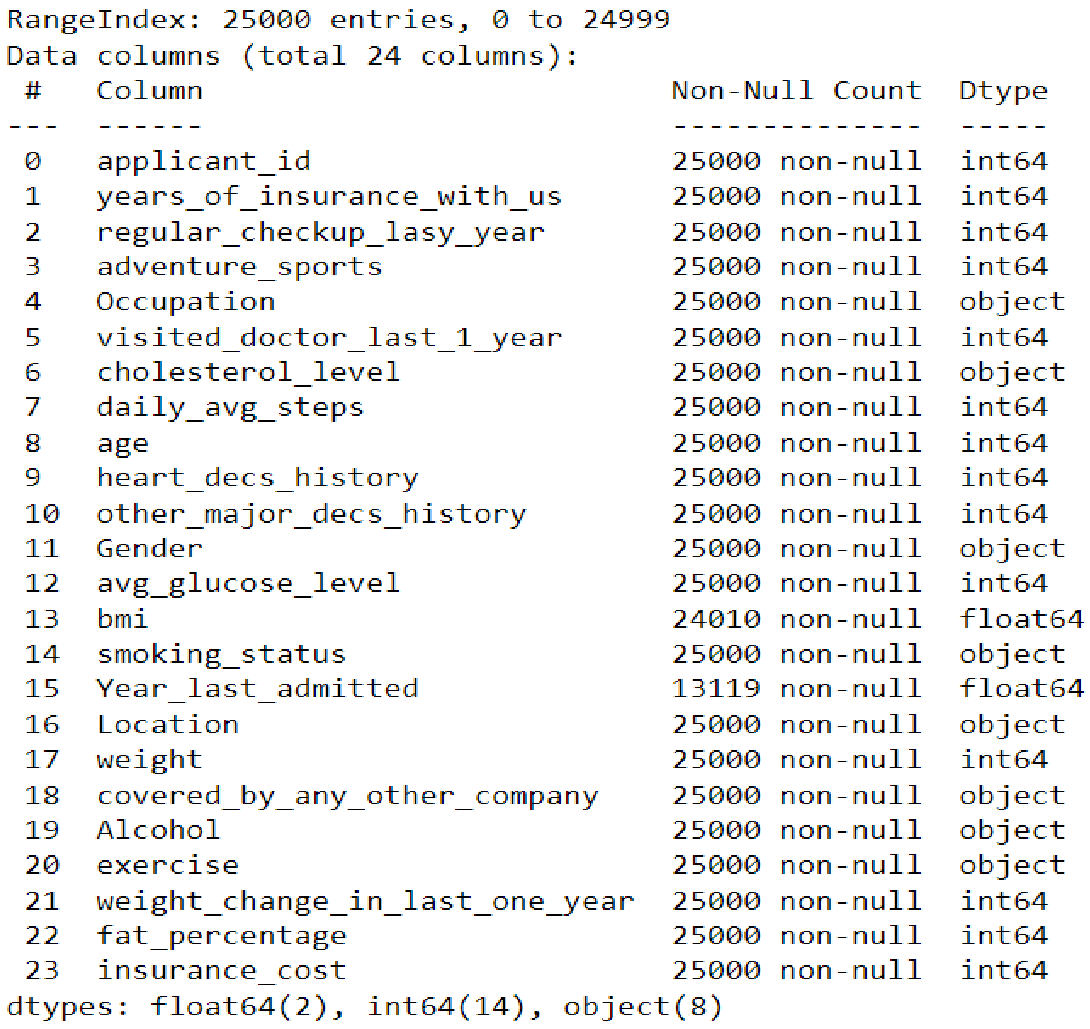


applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age	heart_decs_history	other_major_decs_history	Gender	avg_glucose_level	bmi	smoking_status	Year_last_admitted	Location	weight	covered_by_any_other_company	Alcohol	exercise	weight_change_in_last_one_year	fat_percentage	insurance_cost
5000	3	1	1	Salried	2	125 to 150	4866	28	1	0	Male	97	31.2	Unknown		Chennai	67	N	Rare	Moderate	1	25	20978
5001	0	0	0	Student	4	150 to 175	6411	50	0	0	Male	212	34.2	formerly smoked		Jaipur	58	N	Rare	Moderate	3	27	6170
5002	1	0	0	Business	4	200 to 225	4509	68	0	0	Female	166	40.4	formerly smoked		Jaipur	73	N	Daily	Extreme	0	32	28382
5003	7	4	0	Business	2	175 to 200	6214	51	0	0	Female	109	22.9	Unknown		Chennai	71	Y	Rare	No	3	37	27148
5004	3	1	0	Student	2	150 to 175	4938	44	0	1	Male	118	26.5	never smoked	2004	Bangalore	74	N	No	Extreme	0	34	29616



applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age	heart_decs_history	other_major_decs_history	Gender	avg_glucose_level	bmi	smoking_status	Year_last_admitted	Location	weight	covered_by_any_other_company	Alcohol	exercise	weight_change_in_last_one_year	fat_percentage	insurance_cost
29995	3	0	0	Salried	4	225 to 250	5614	22	0	0	Male	145	36.1	smokes	2000	Kanpur	79	Y	Rare	Moderate	4	40	39488
29996	6	0	0	Business	4	200 to 225	4719	58	0	0	Male	134	31.3	never smoked	2009	Kanpur	66	N	Rare	Moderate	2	28	14808
29997	7	0	1	Student	2	150 to 175	5624	34	0	1	Male	151		Unknown		Bhubaneswar	76	N	Rare	Moderate	1	35	33318
29998	1	0	0	Salried	2	225 to 250	10777	27	0	0	Male	66	26.6	Unknown		Surat	74	N	Rare	Moderate	0	40	30850
29999	8	2	0	Business	4	150 to 175	5882	22	1	0	Male	245	41.6	formerly smoked	2014	Chennai	57	N	No	No	4	21	6170

Table 13: Head and Tail Information of Data



```
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)
```

Table 14: Data Information

THE END