[Healthcare Project]

[Predicting the insurance cost based on several health parameters]

[Submitted by -Jyotiranjan Padhiary]

Table of Contents

List of Tables

List of Figures

Health care

1. Introduction

Health care expenditure is one of the major expenses and everybody wants to avoid it by taking insurance, but nowadays people are not aware of how much they should invest in insurance to properly cover themselves in case of medical emergencies. Insurance companies on the other hand should have mechanism to know how much they should take premium from an individual in order to reduce their risk.

It is very important for an individual to know how much they should invest in insurances to avoid financial crisis during emergencies and insurance companies should also know how much is the risk involved with an individuals life in order to align the cost of insurance for him/her.

It will have a great impact in society where we can tell people how much they should invest in insurance policy according to their medical history and health conditions. Also insurance companies will be benefited from this where they can use this model to predict the insurance premium for an individual by taking info about their medical history and health conditions. They can reduce the risk by doing this.

2. EDA and Business Implication

The quantitative data has been collected for 25000 individuals from an insurance company, in the year of 2018-19.

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

There are 25000 rows and 24 columns.

	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps
count	25000.000000	25000.000000	25000.000000	25000	25000.000000	25000	25000.000000
unique	NaN	NaN	NaN	3	NaN	5	NaN
top	NaN	NaN	NaN	Student	NaN	150 to 175	NaN
freq	NaN	NaN	NaN	10169	NaN	8763	NaN
mean	4.089040	0.773680	0.081720	NaN	3.104200	NaN	0.511062
std	2.606612	1.199449	0.273943	NaN	1.141663	NaN	0.204211
min	0.000000	0.000000	0.000000	NaN	0.000000	NaN	0.000000
25%	2.000000	0.000000	0.000000	NaN	2.000000	NaN	0.375000
50%	4.000000	0.000000	0.000000	NaN	3.000000	NaN	0.489996
75%	6.000000	1.000000	0.000000	NaN	4.000000	NaN	0.625000
max	8.000000	5.000000	1.000000	NaN	12.000000	NaN	1.000000

Table 1- Descriptive details of the data

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

Data Dictionary

Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
adventure_sports	Customer is involved with adventure sports like climbing, diving etc.
Occupation	Occupation of the customer
visited_doctor_last_1_year	Number of times customer has visited doctor in last one year
cholesterol_level	Cholesterol level of the customers while applying for insurance
daily_avg_steps	Average daily steps walked by customers
age	Age of the customer
heart_decs_history	Any past heart diseases
other_major_decs_history	Any past major diseases apart from heart like any operation
Gender	Gender of the customer
avg_glucose_level	Average glucose level of the customer while applying the insurance
bmi	BMI of the customer while applying the insurance
smoking_status	Smoking status of the customer

Table 2-Data Dictionary

```
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-Variable info

There are 25 variables 15 numerical and 10 categorical. From the 15 numerical there are 7 discrete

and 8 continuous variables.

Exploratory Data Analysis

Distribution of continuous attributes.

Histograms

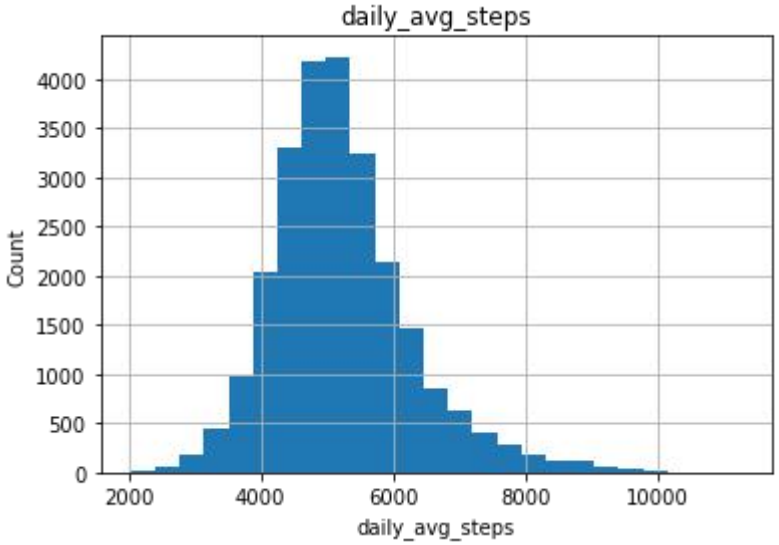


Figure 1- Distribution of daily_avg_steps

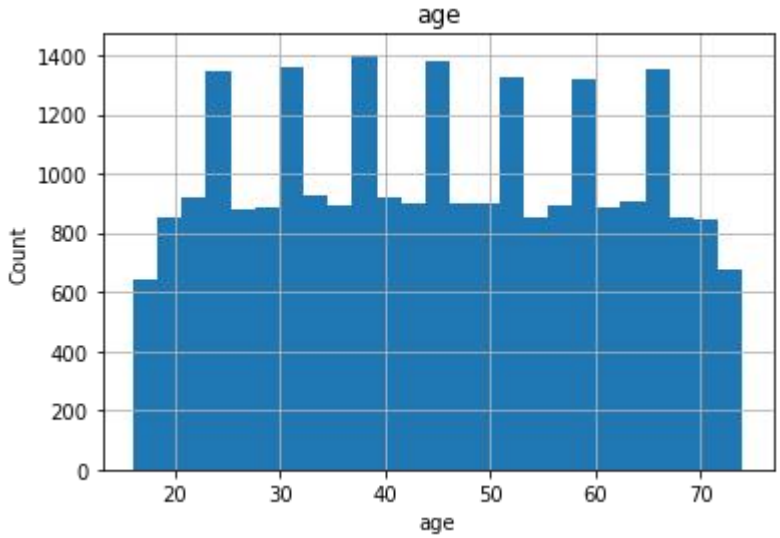


Figure 2 - Distribution of age

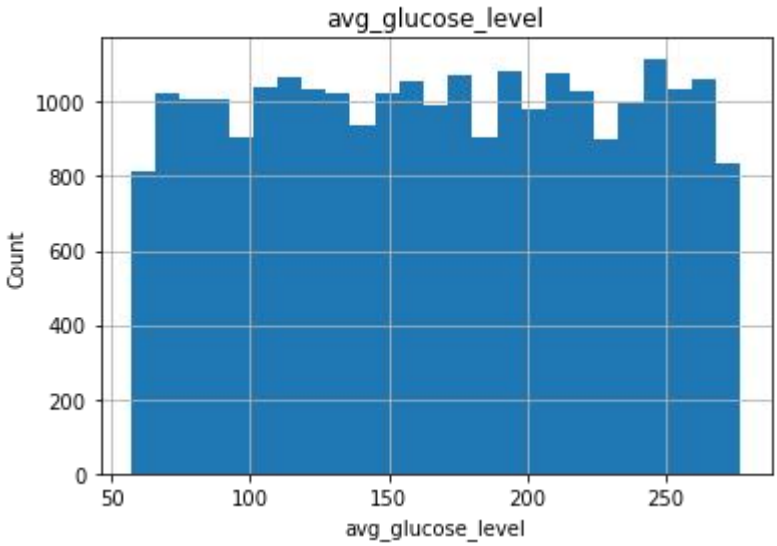


Figure 3-Distribution of avg_glucose_level

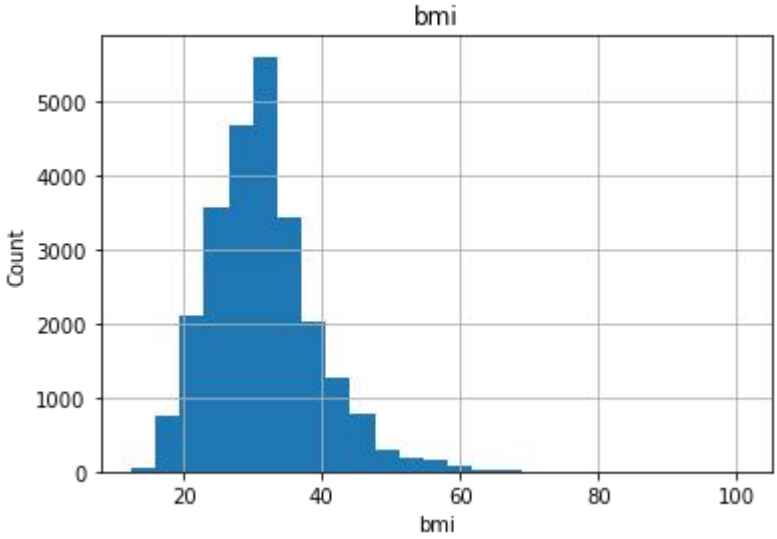


Figure 4- Distribution of BMI

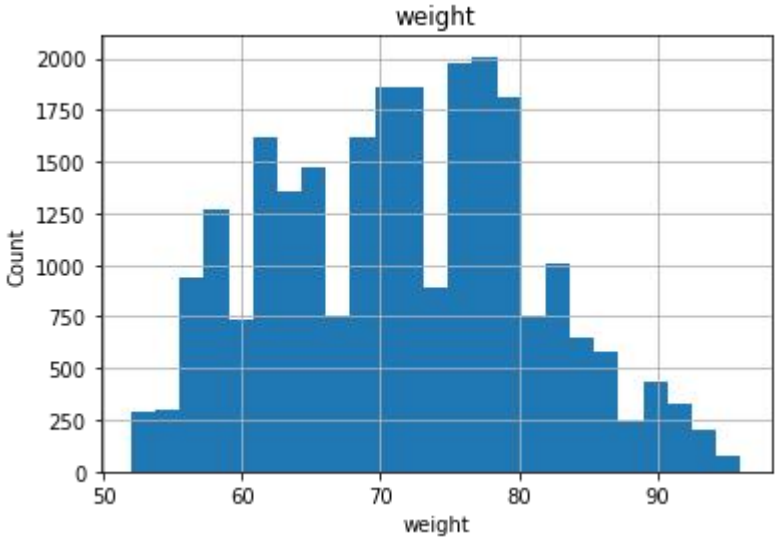


Figure 5-Distribution of weight

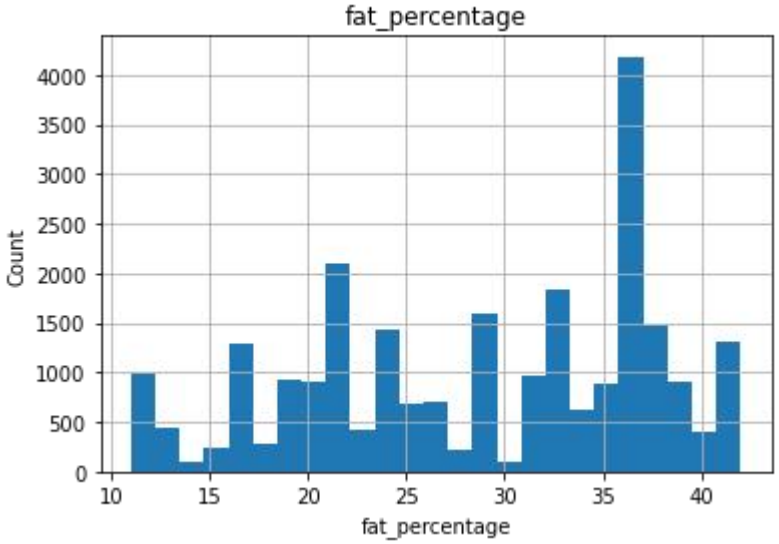


Figure 6-Distribution of fat_percentage

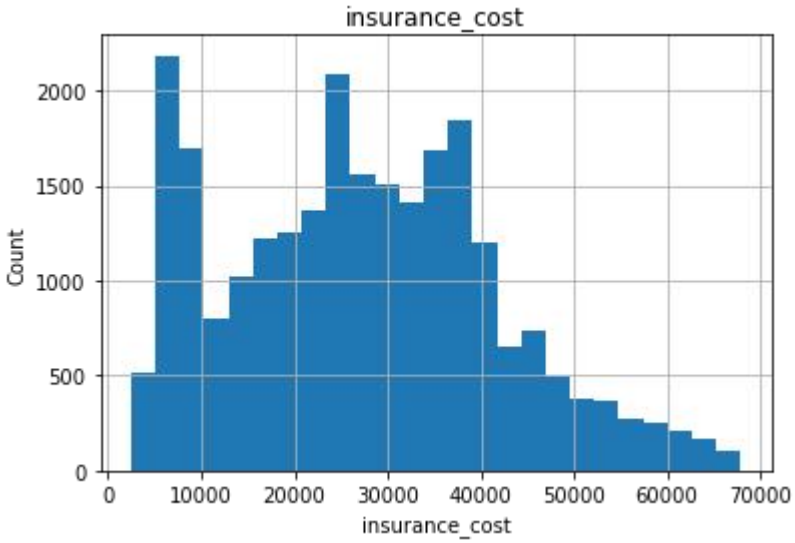


Figure 7- Distribution of insurance cost

From the above distribution plots of continuous variables it can be inferred that distribution of daily average steps and BMI are left skewed.

Box-plots

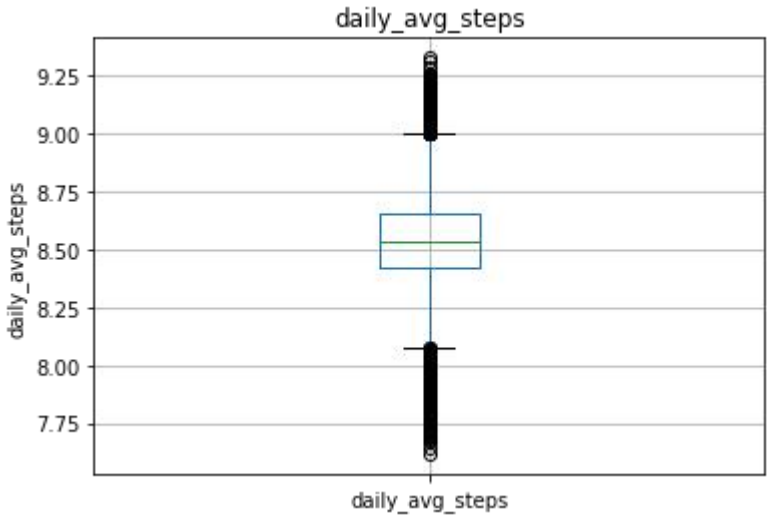


Figure 8-Box plot of daily avg steps

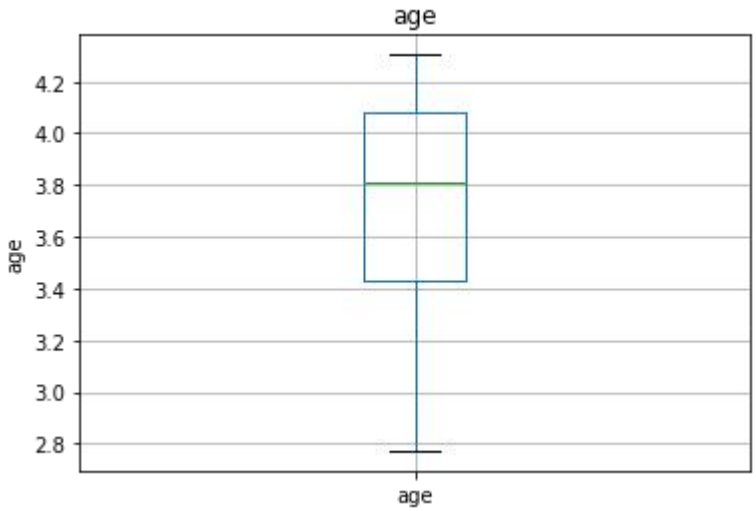


Figure 9-Box plot of age

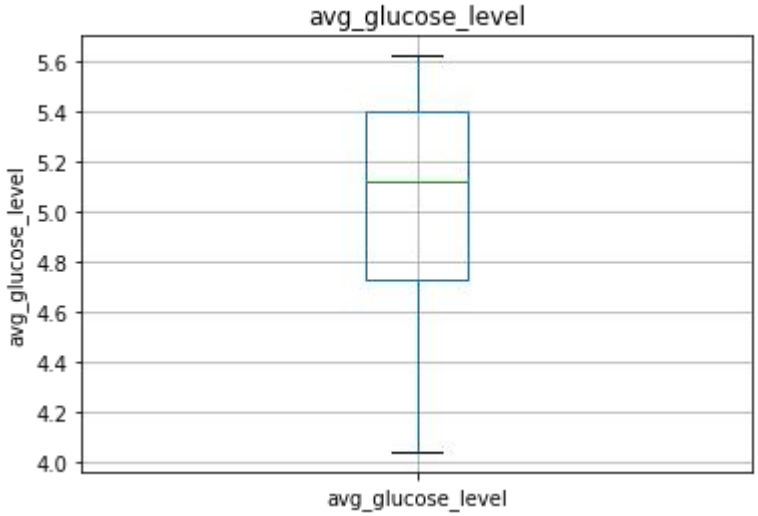


Figure 10-box plot of avg glucose level

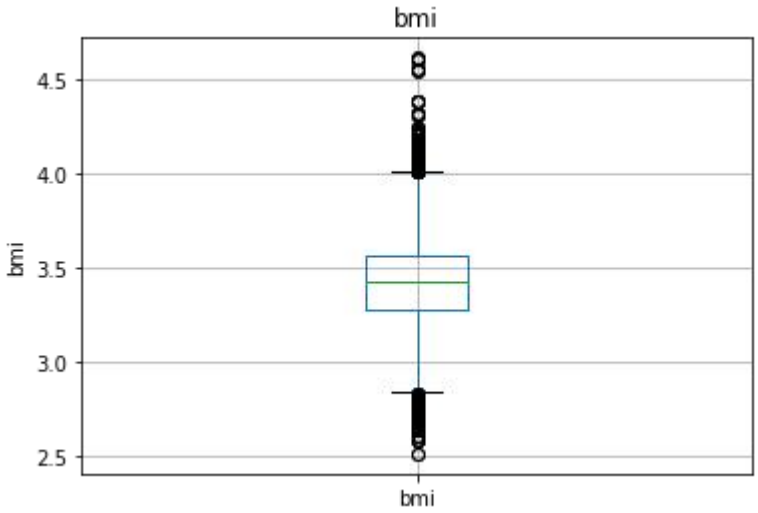


Figure 11-box plot of BMI

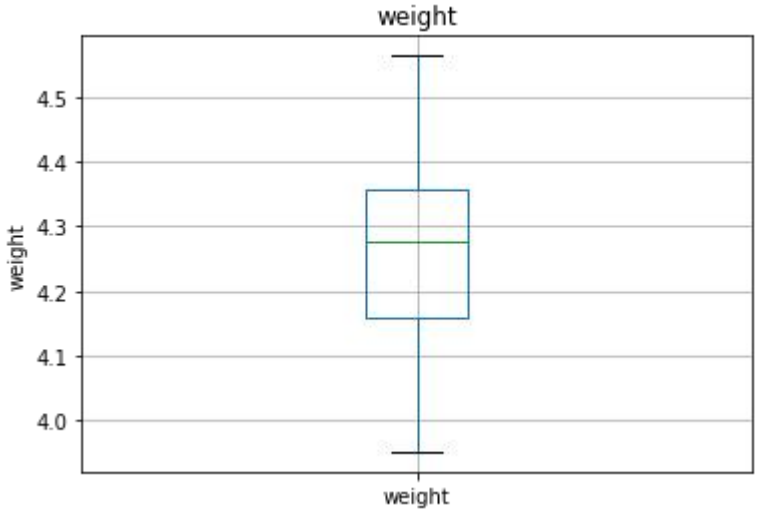


Figure 12-Box Plot of weight

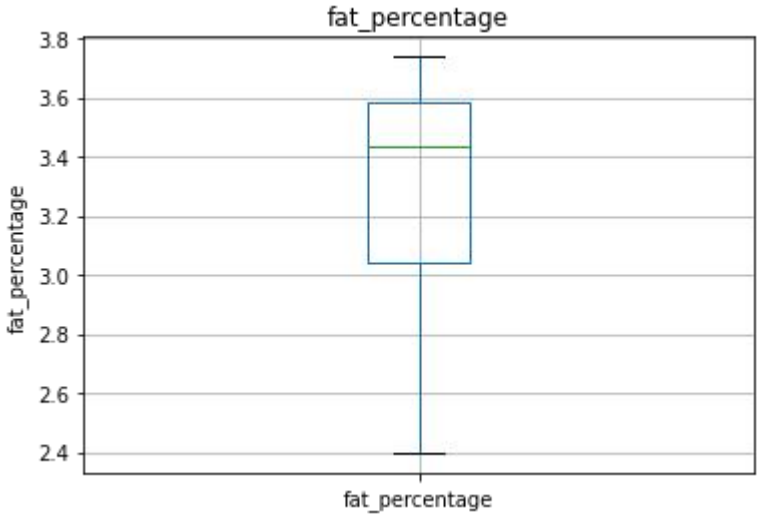


Figure 13-box plot of fat_percentage

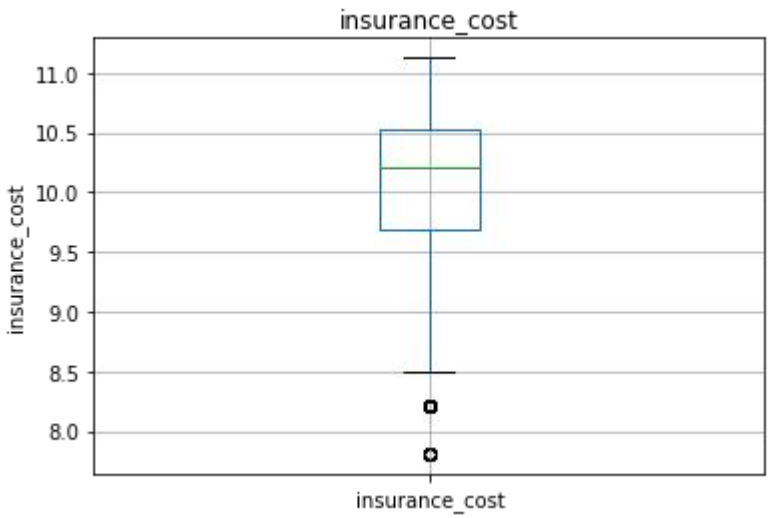


Figure 14-box plot of insurance cost

Inference : from the above box plots it can be observed that no variable is normally distributed.

Discrete attributes :

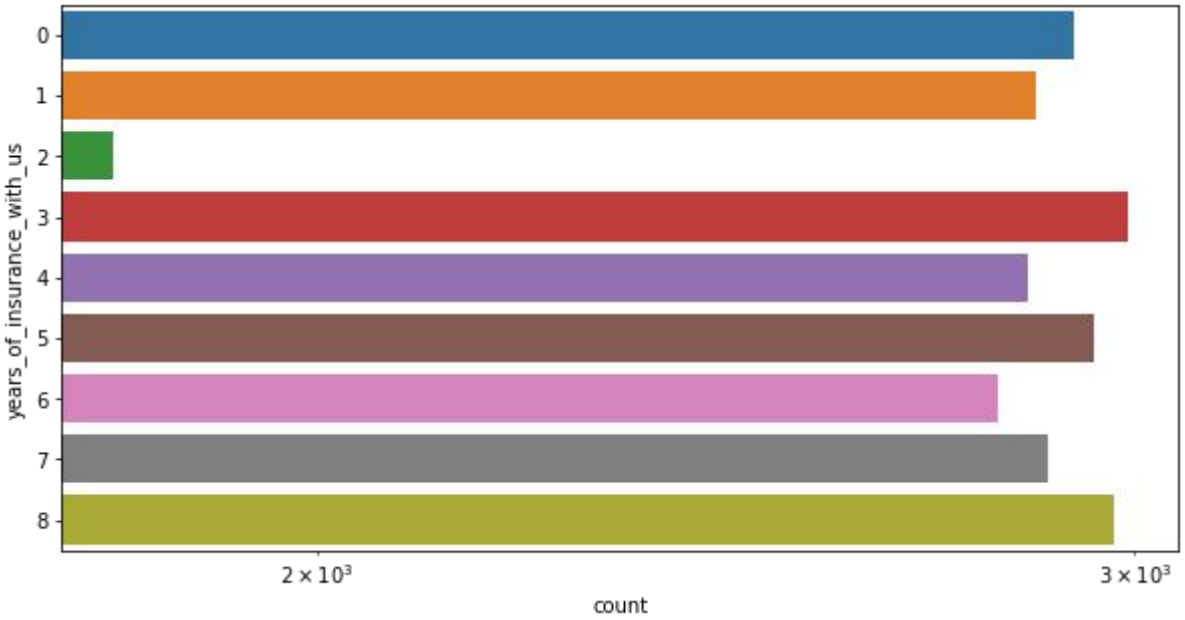


Figure 15-Distribution of years of insurance with us

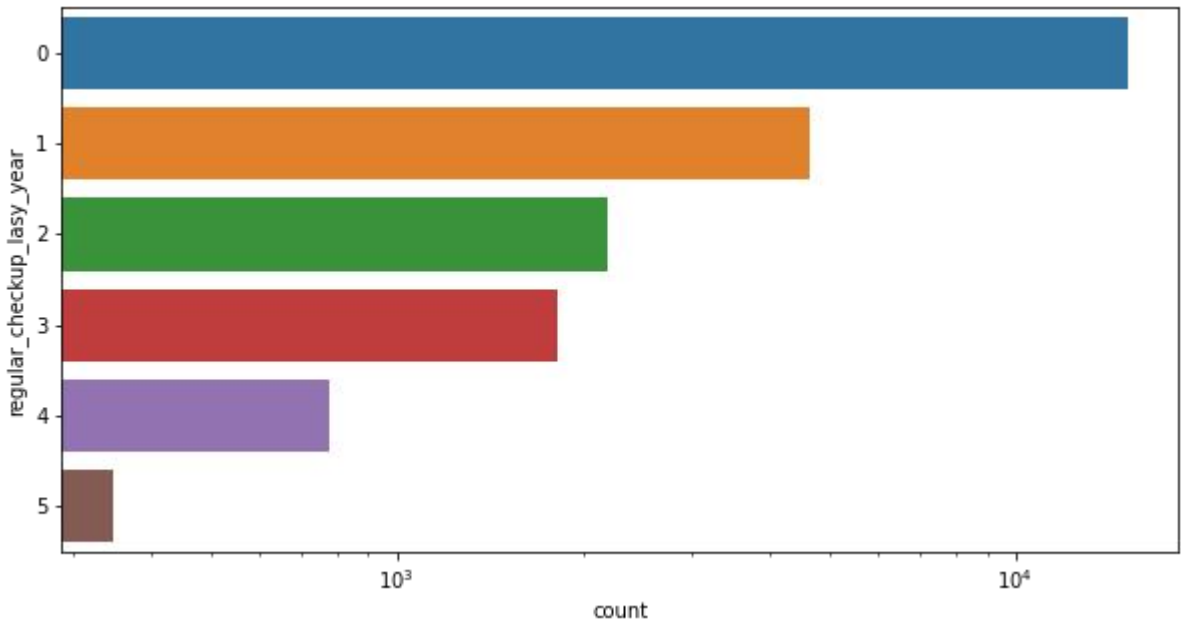


Figure 16-Distribution of regular checkup last year

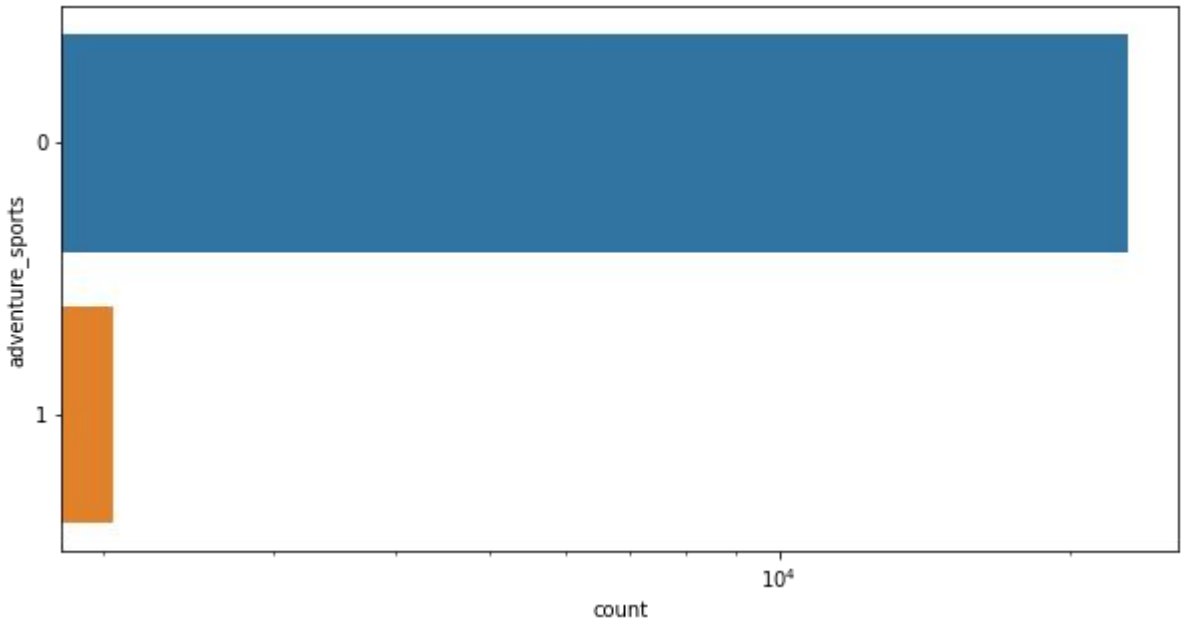


Figure 17-distribution of adventure sports

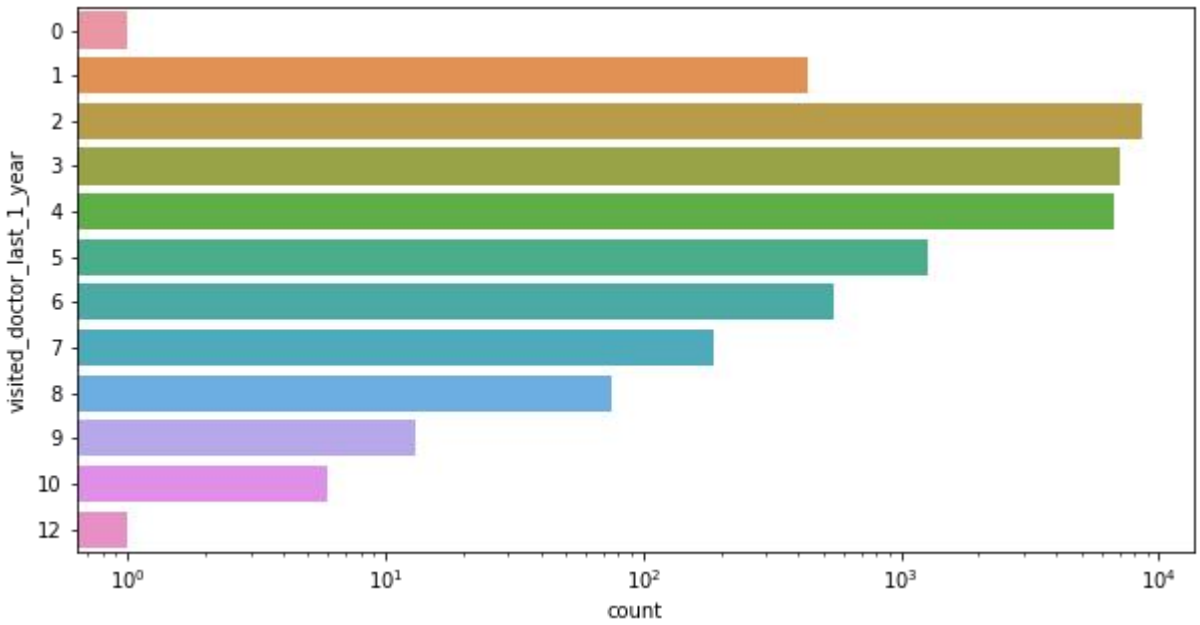


Figure 18-distribution of visited doctor last yr

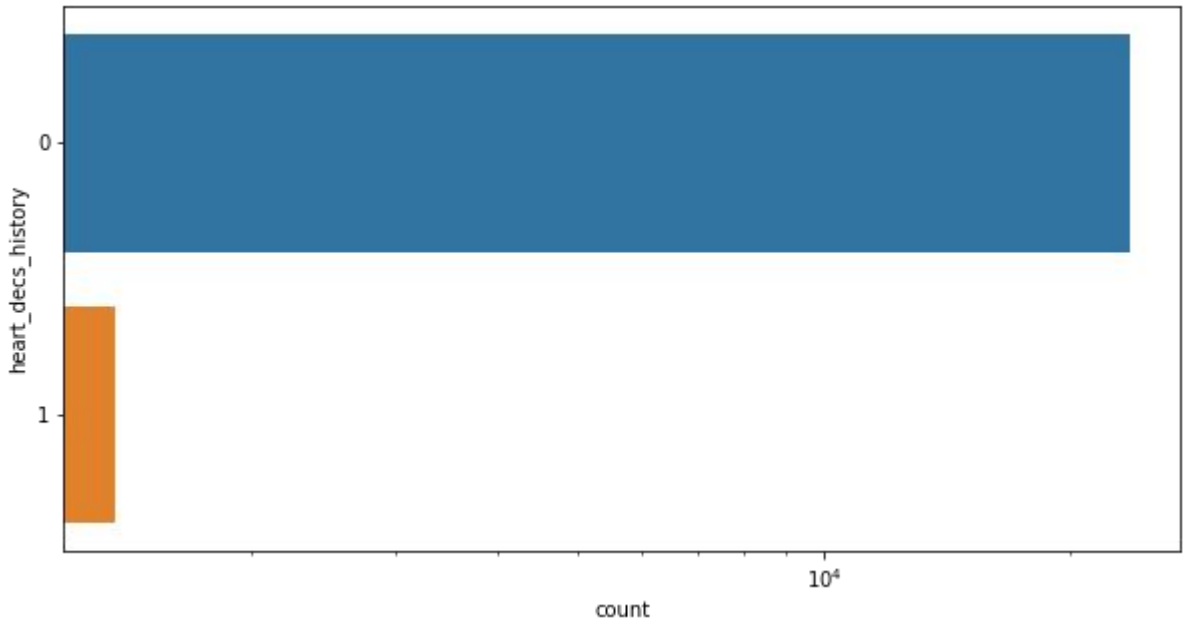


Figure 19-distribution of heart disease history

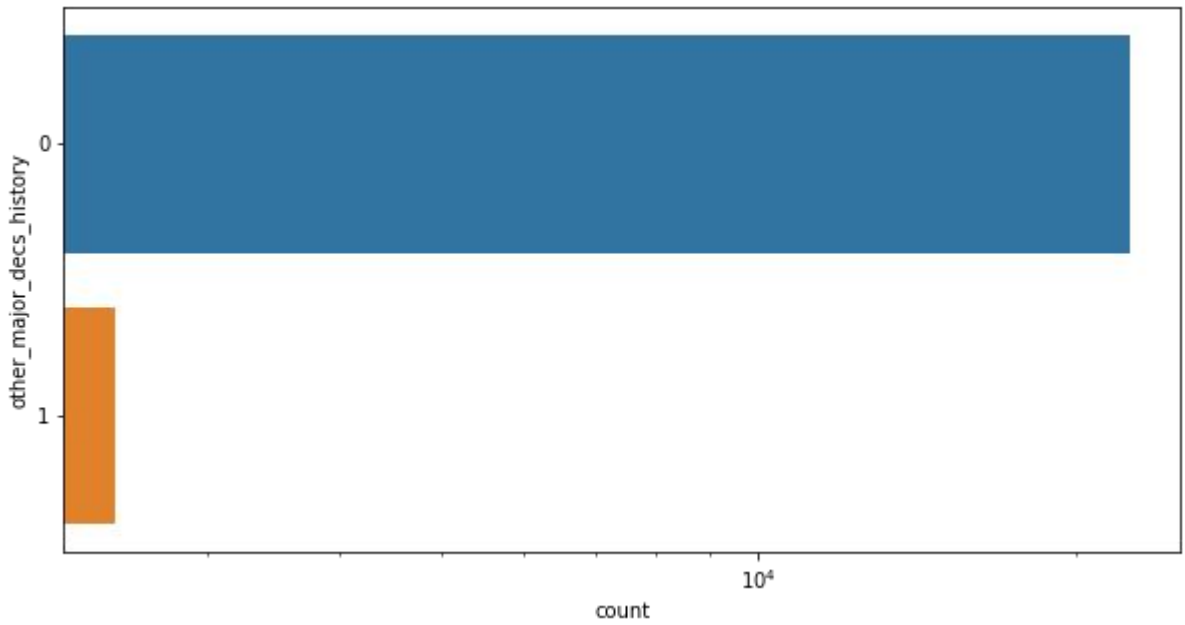


Figure 20-distribution of other major disease history

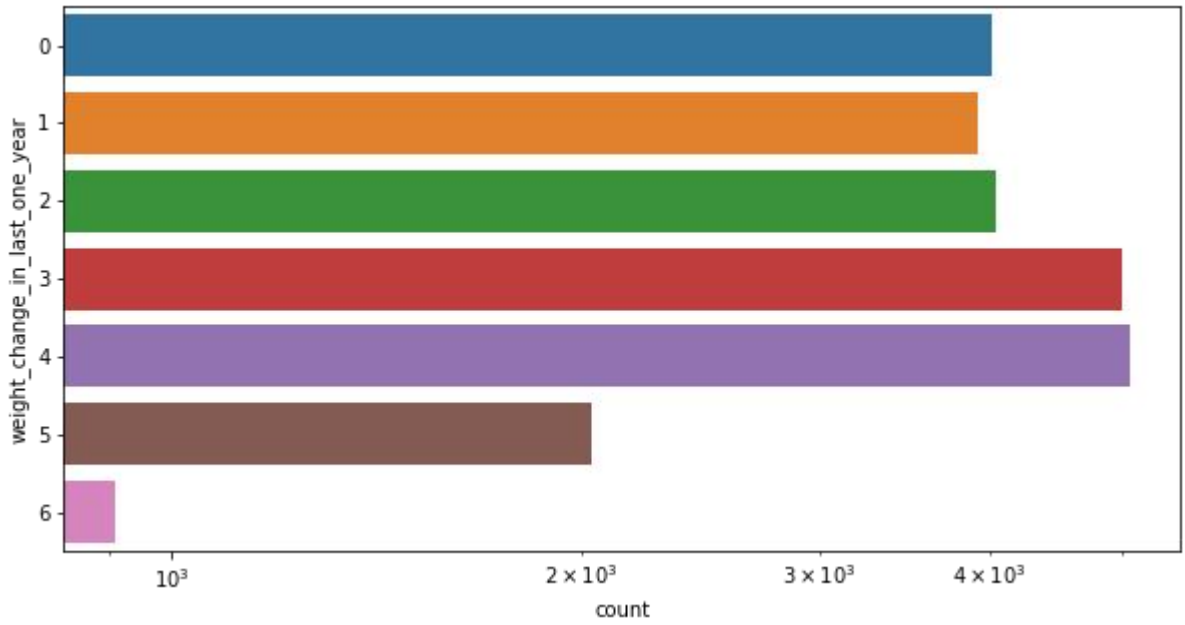


Figure 21-distribution of weight change in last one year

Inference : very less people are having heart and other disease history, and having a max of 3yrs of insurance with the company.

Categorical attributes:

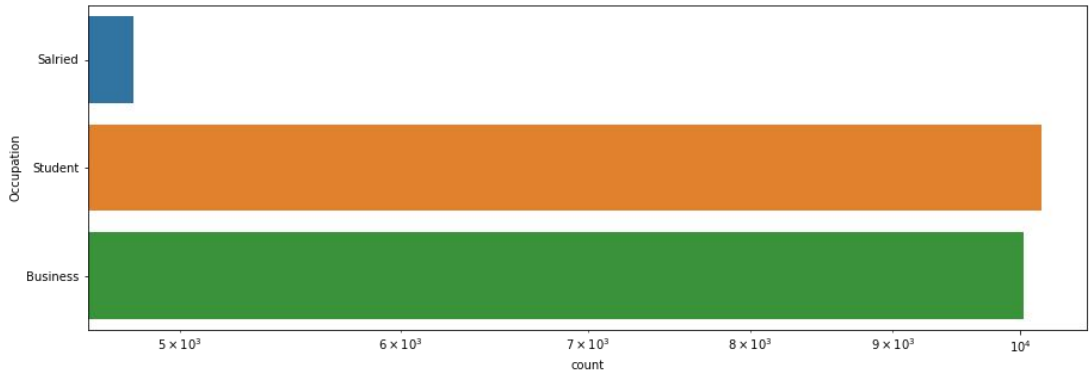


Figure 22-Distribution of occupation

Salaried people are taking very less insurance compared to students and business men. As they will be having a corporate insurance already provided by their employers.

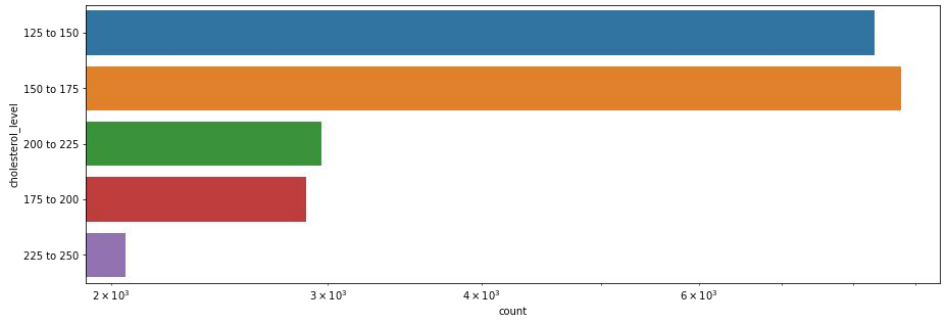
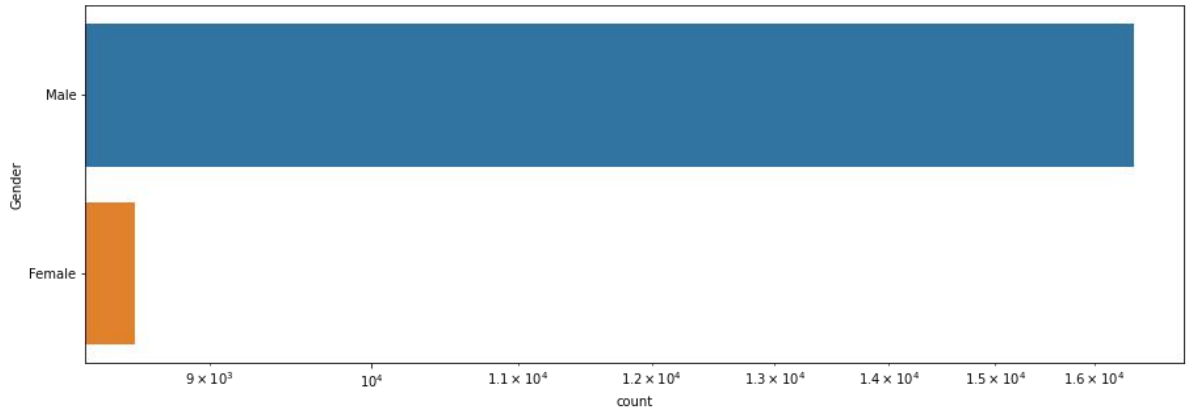


Figure 23-distribution of cholesterol levels

Most people are having cholesterol in the range of 150-175



Females are having very less insurance compared to male.

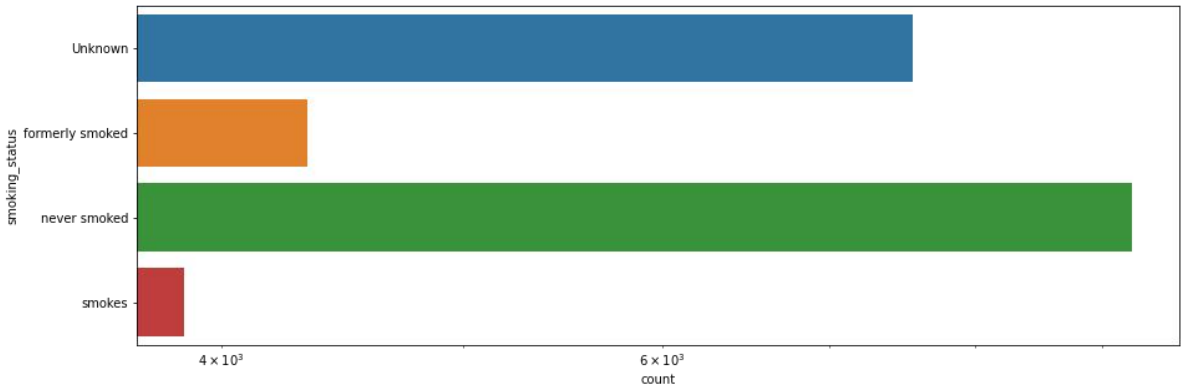


Figure 24-distribution of smoking status

Most people have never smoked

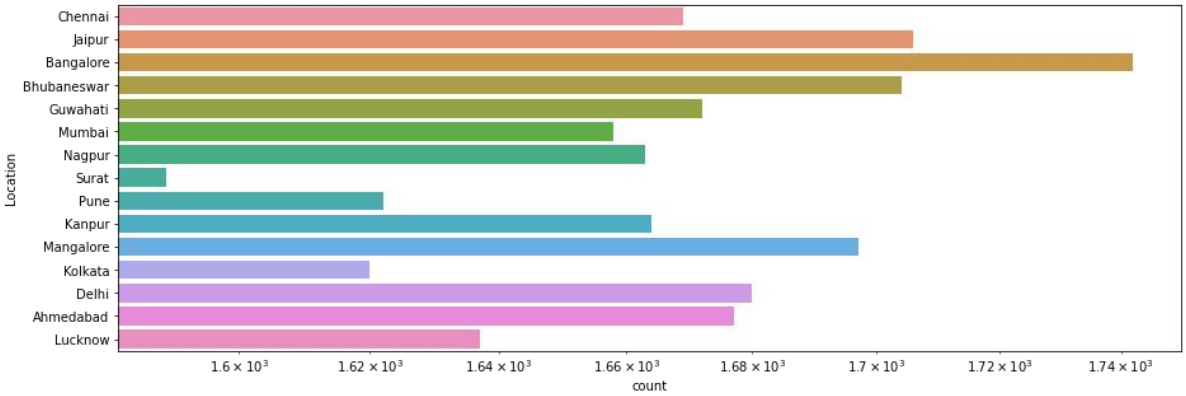


Figure 25-distribution of location

Most of the people from Bangalore are taking insurance

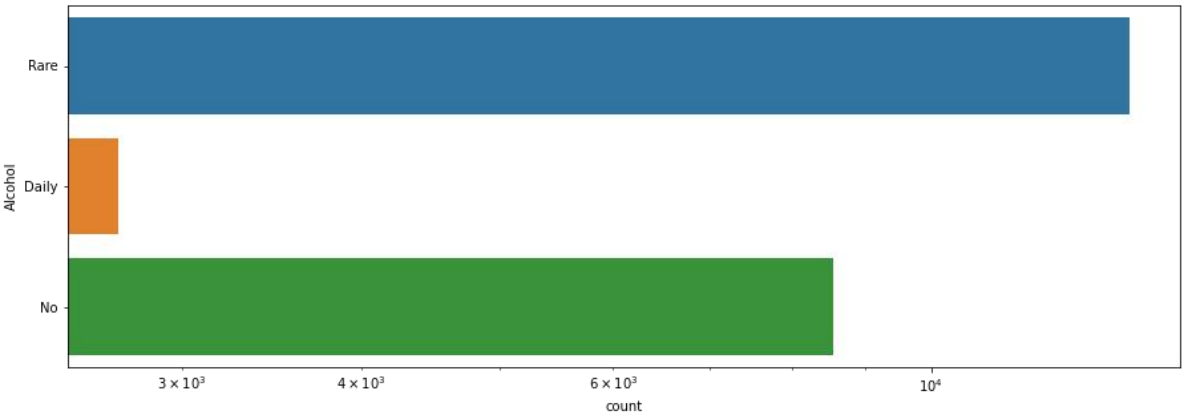


Figure 26-Distribution of alcohol

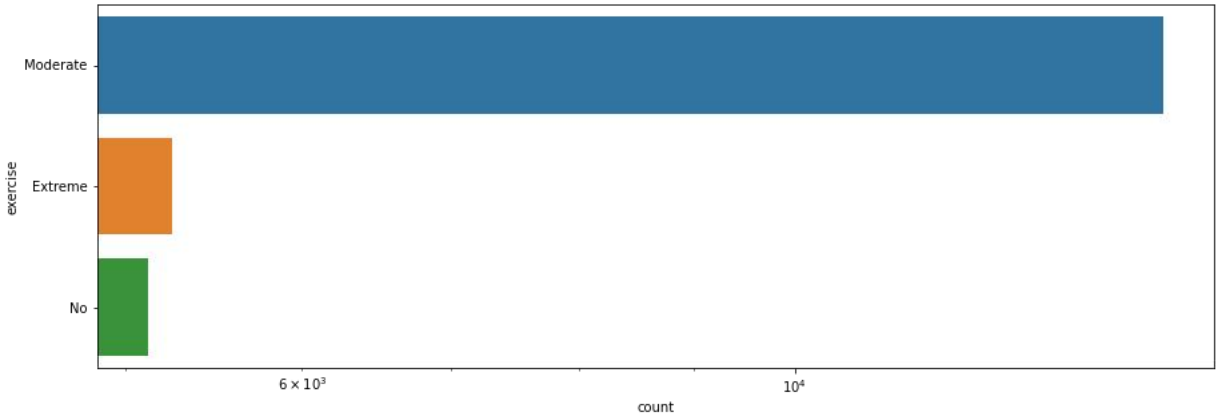


Figure 27- distribution of exercise

Most people are on moderate exercise routine.

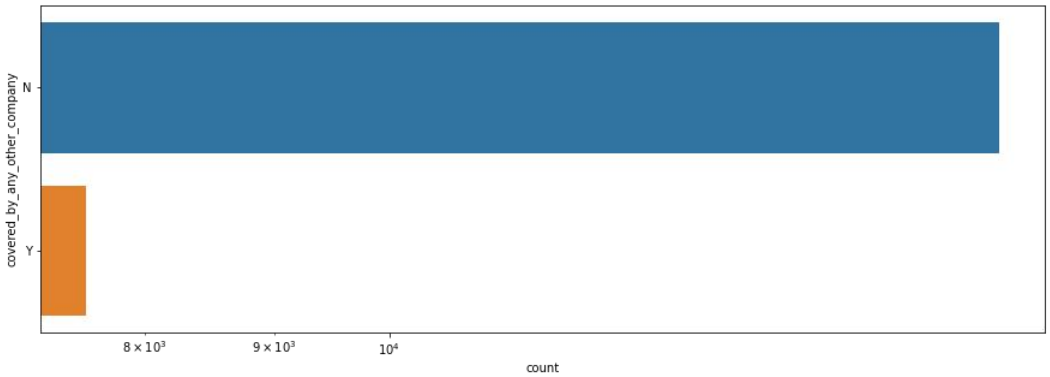


Figure 28-distribution of covered by any other company

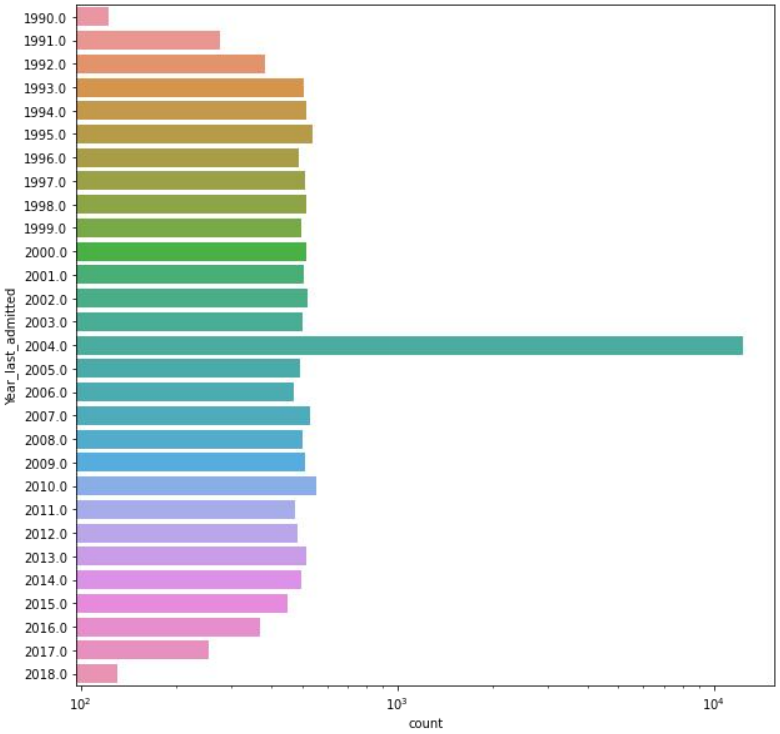


Figure 29-Distribution of year last admitted

We can observe that most people last got admitted in the year of 2004.

Bi-variate analysis

Correlations

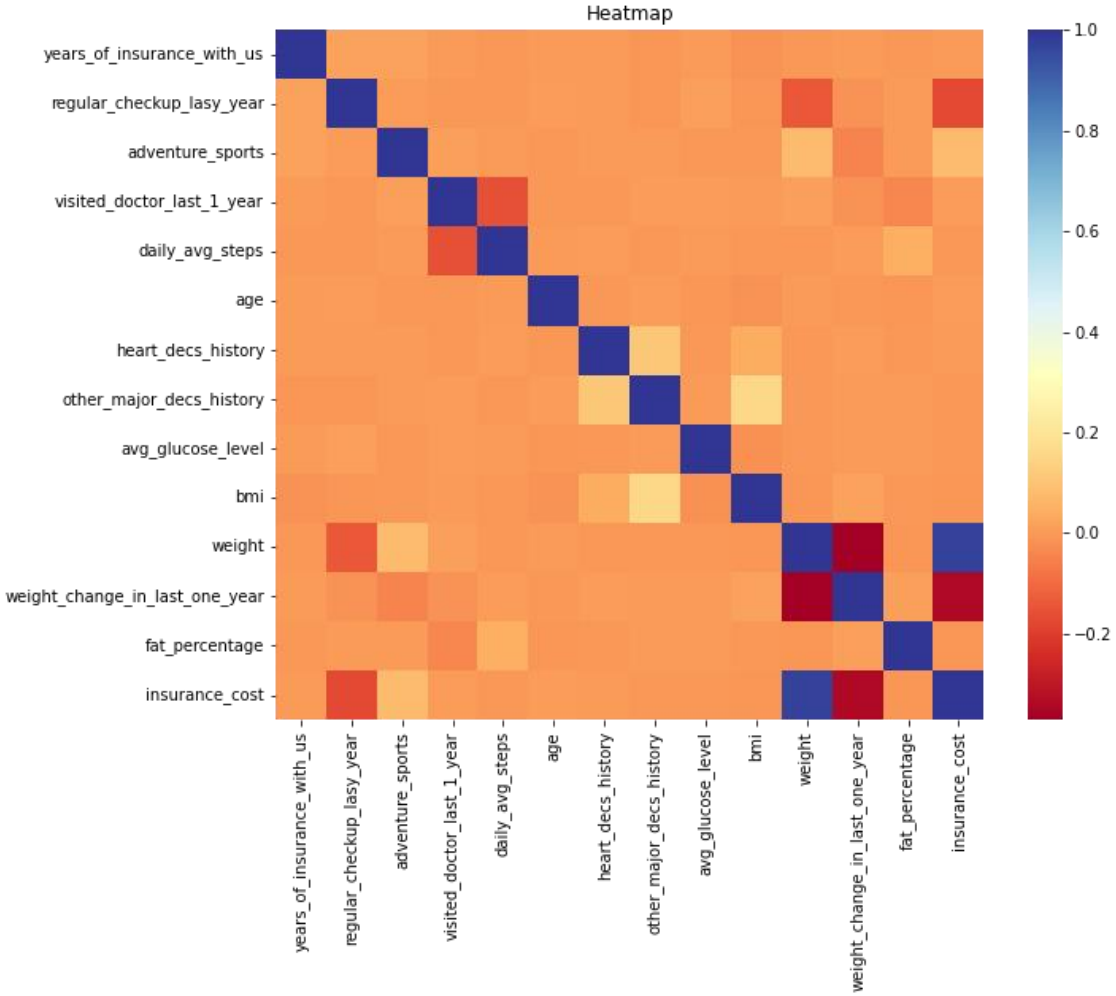


Figure 30-Correlation plot

There are no variables which are highly correlated to each other.

Relation between two variables.

Continuous variables

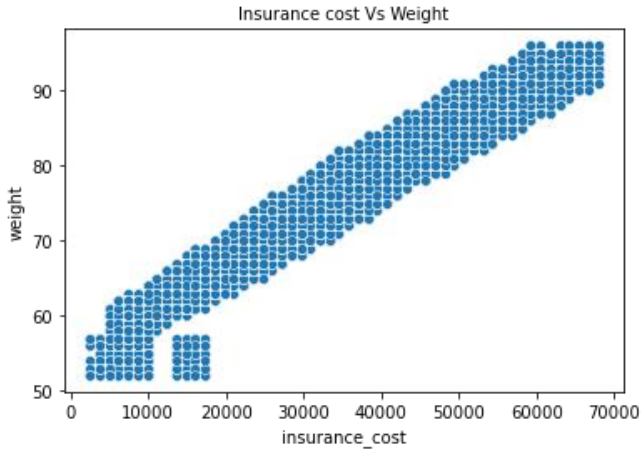


Figure 31- Insurance cost vs weight

It can be observed that there is a linear relationship between weight of a person and insurance cost.

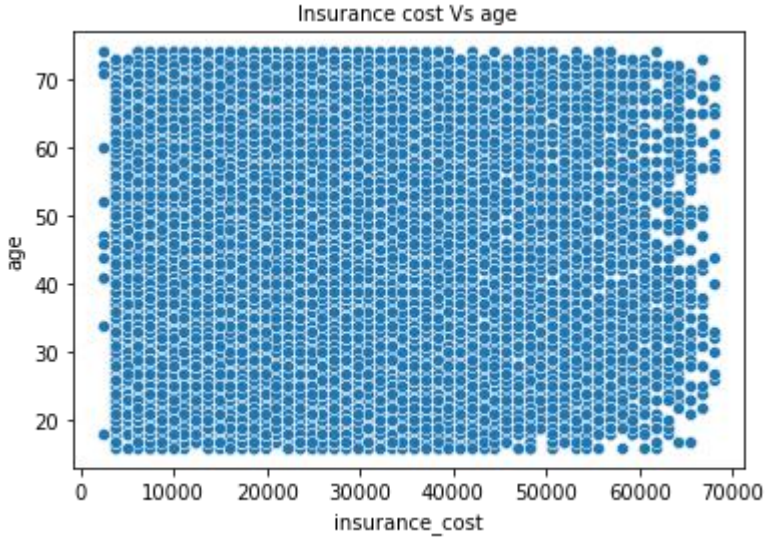


Figure 32-insurance cost vs age

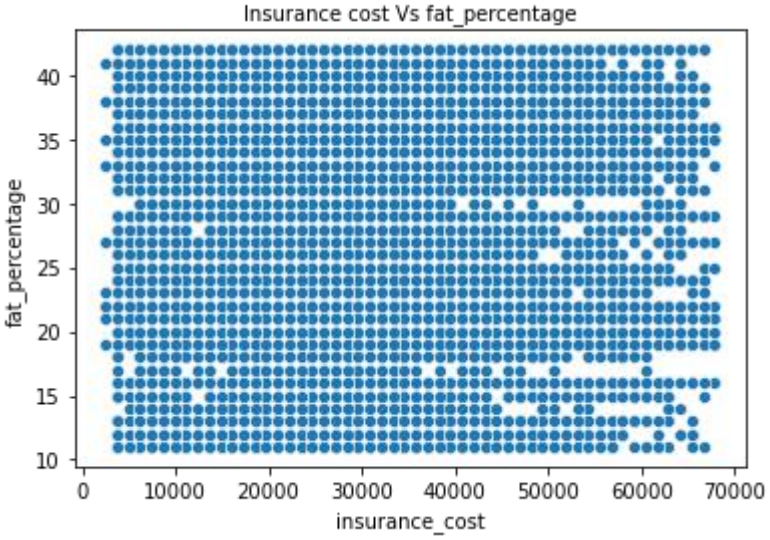


Figure 33-insurance cost vs fat percentage

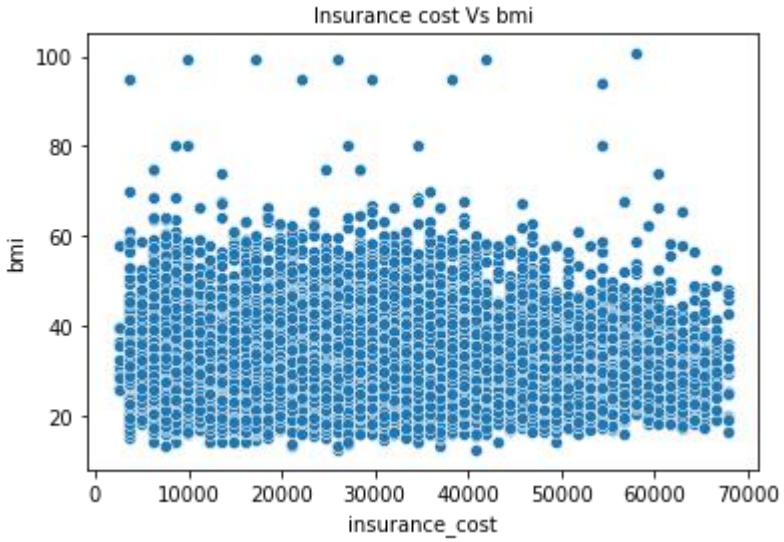


Figure 34-insurance cost vs BMI

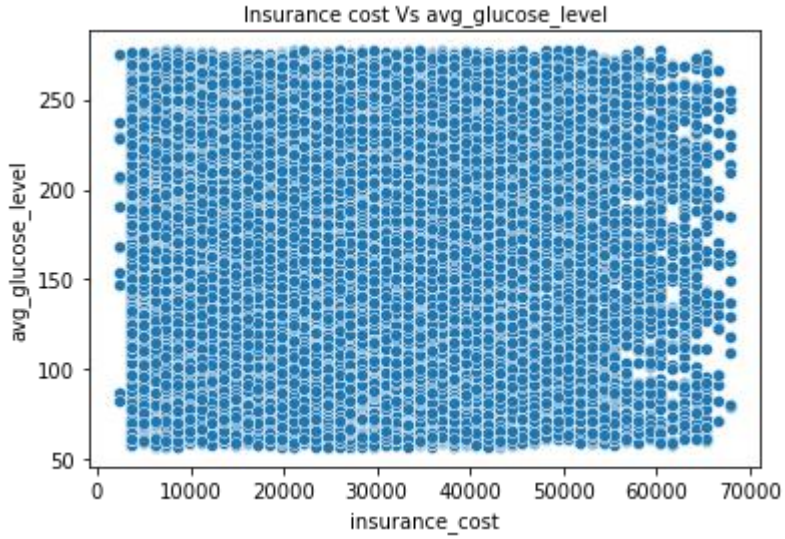


Figure 35-insurance cost vs avg glucose level

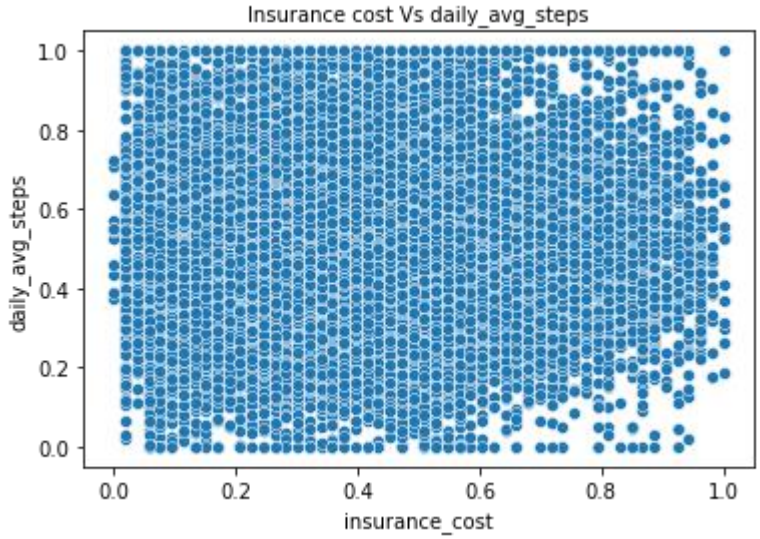


Figure 36- Insurance cost vs daily avg steps

Discrete attributes:

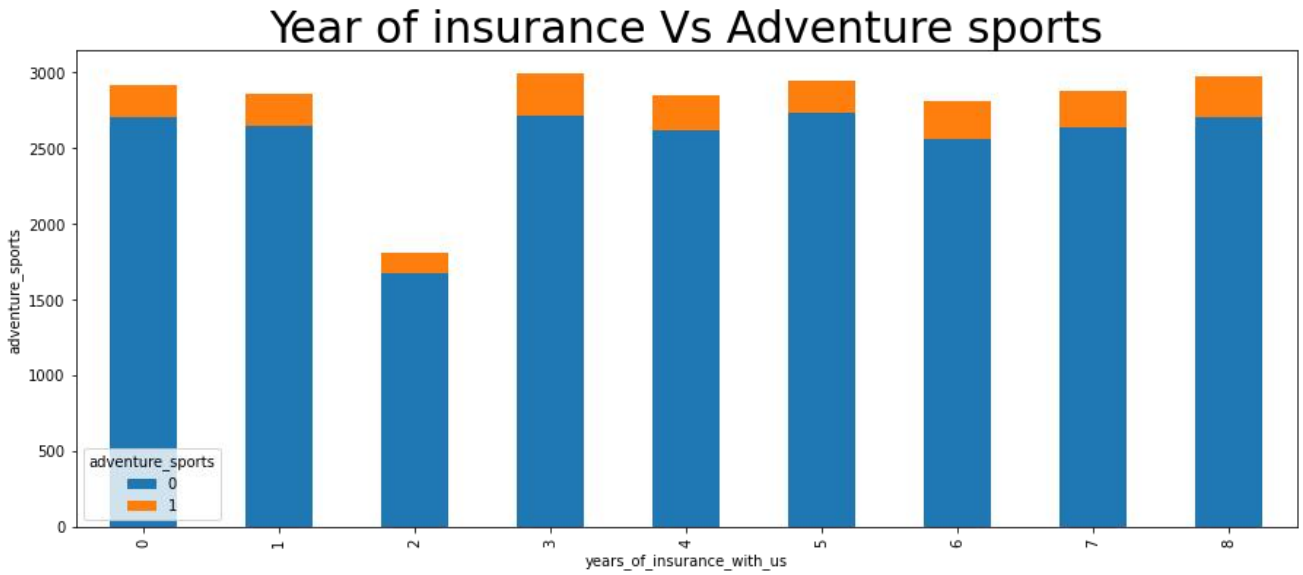


Figure 37- Year of insurance with us vs adventure_sports

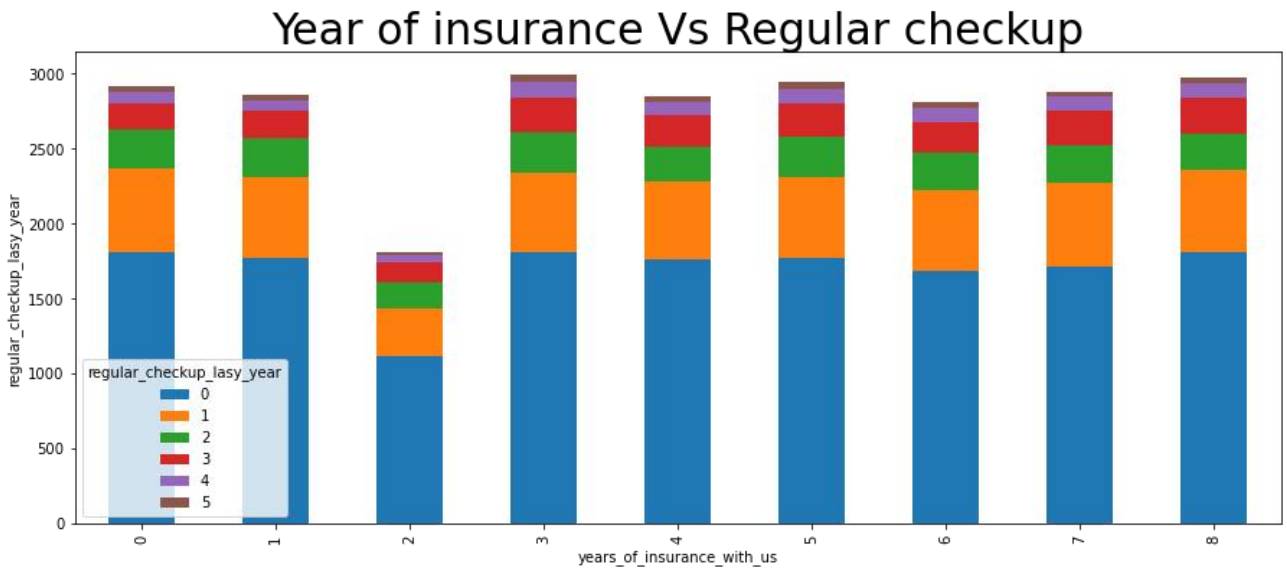


Figure 38-years of insurance with us vs regular checkup last yr

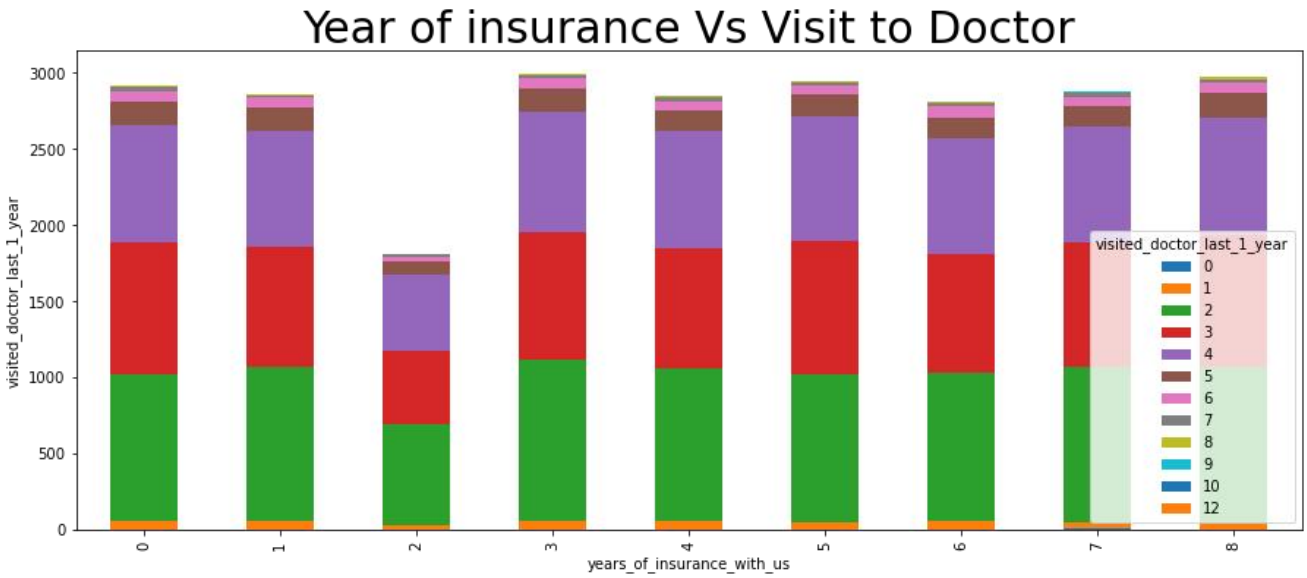


Figure 39-visited doctor last yr vs years of insurance with us

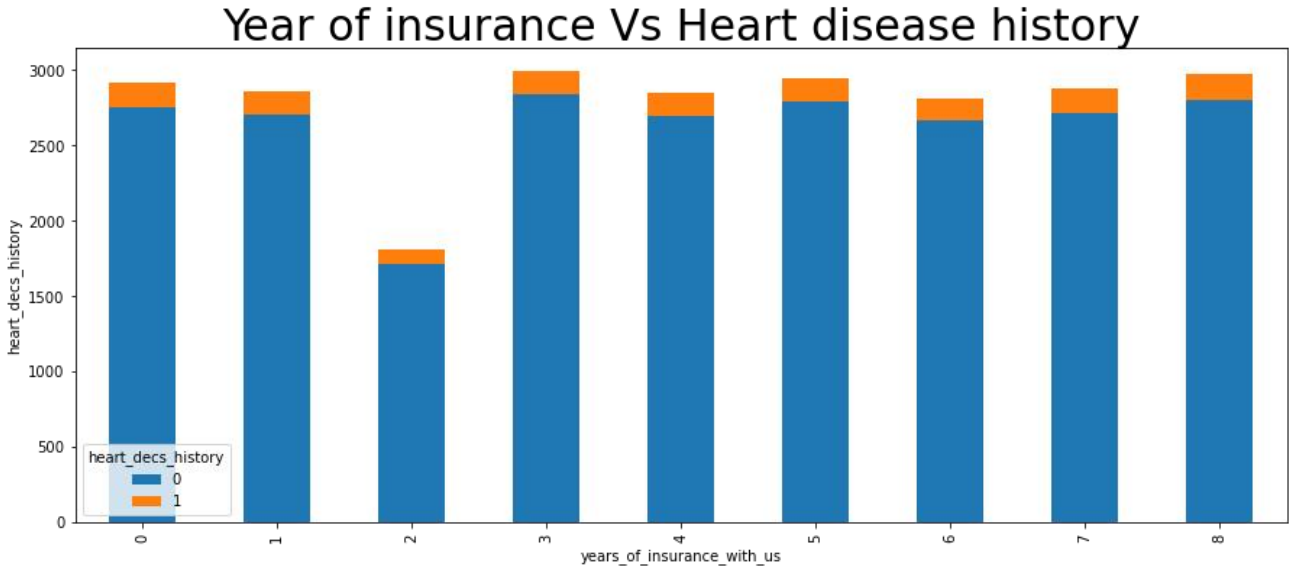


Figure 40-heart disease history vs years of insurance with us

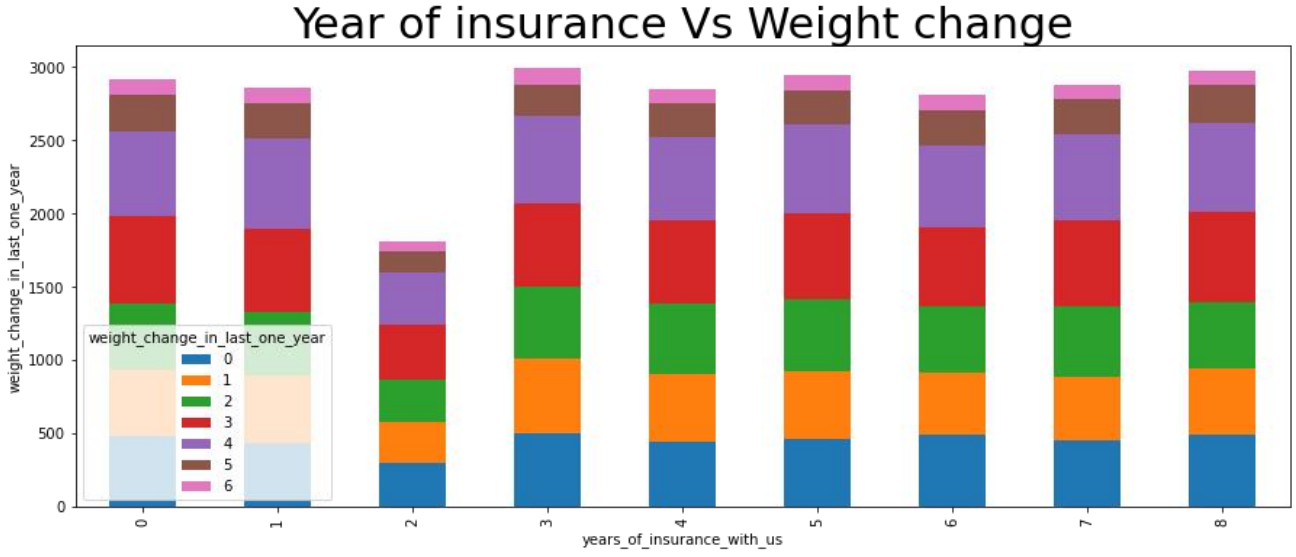


Figure 41-years of insurance with us vs weight change in last one year

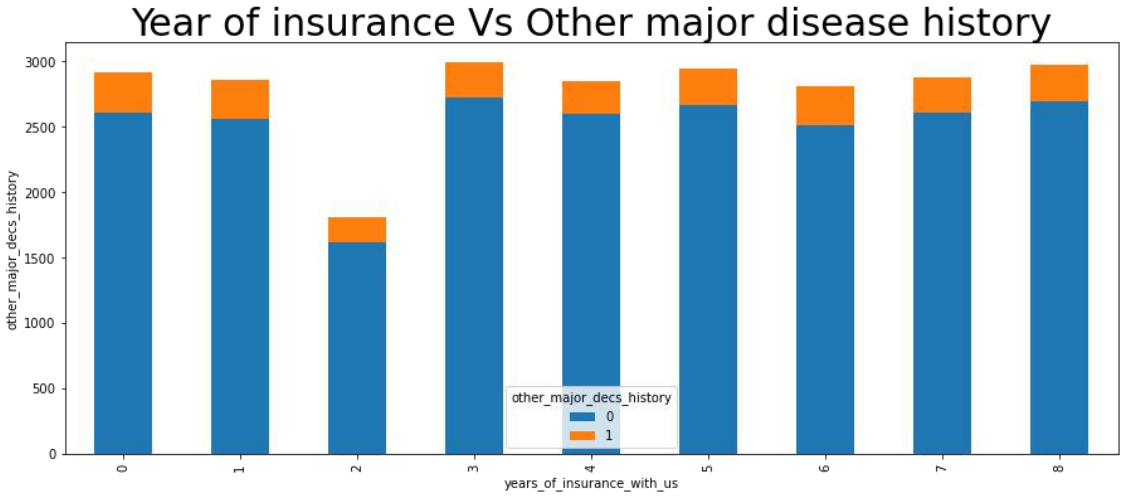


Figure 42-other major disease history vs years of insurance with us

Categorical Variables

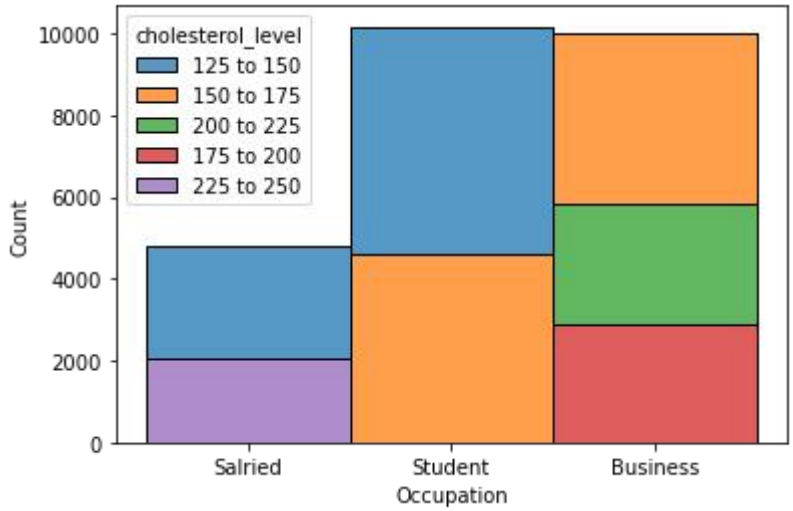


Figure 43-Occupation vs cholesterol variables

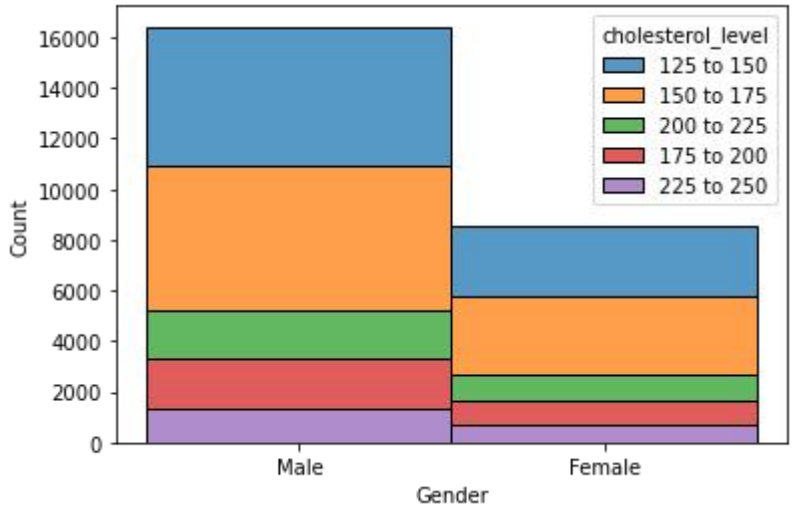


Figure 44-Gender vs cholesterol

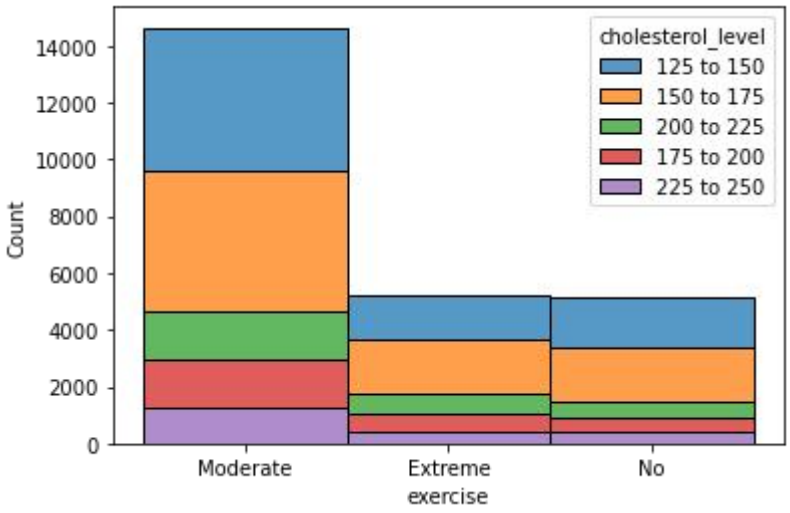


Figure 45-Exercise vs cholesterol

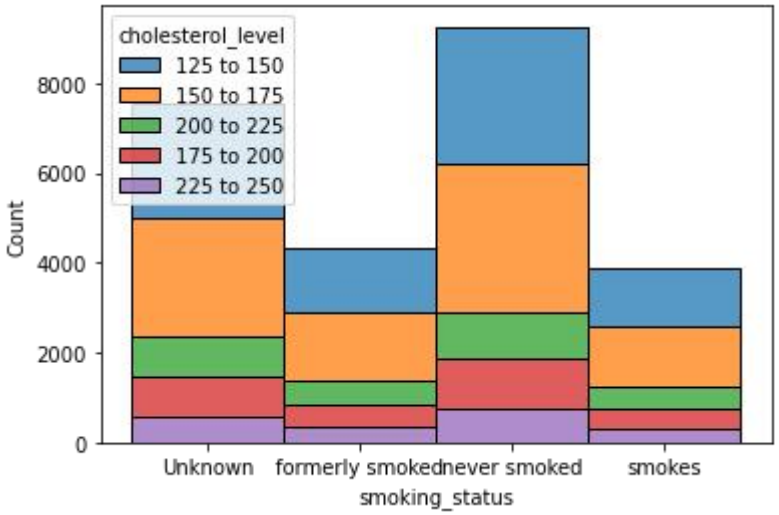


Figure 46-Smoking status vs cholesterol

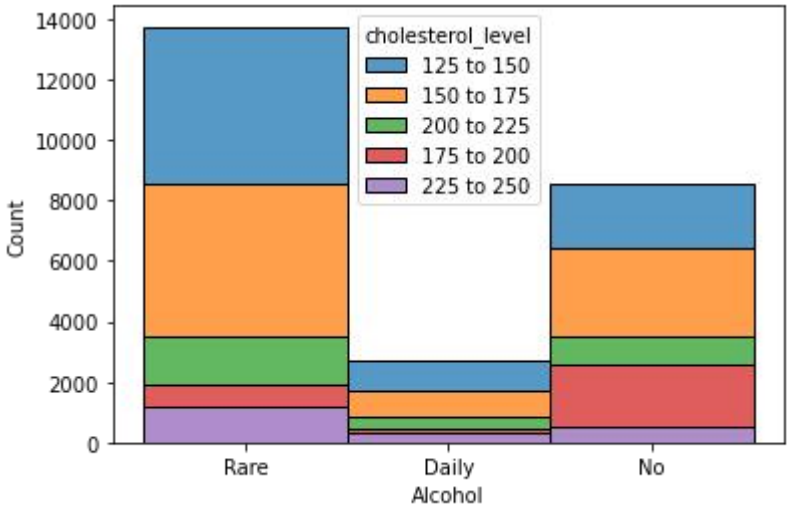


Figure 47-Alcohol vs cholesterol

Inference : men who are salaried professionals and having alcohol rarely and exercise moderately are having high cholesterol of 225 to 250

Business insights from EDA

From our analysis we can give the below insights

1. Men who are salaried professionals and having alcohol rarely and exercise moderately are having high cholesterol of 225 to 250.
2. Females are taking very less insurance as compared to men.
3. Weight is linearly related to insurance cost.
4. Most people are taking insurance for 3 yrs only.
5. The target variable distribution is skewed
6. Insurance cost is lowest for salaried professionals as they have already company provided insurance.
7. Cost of insurance for men are higher than women
8. People in the range of cholesterol level between 150-175 are paying highest insurance cost.
9. Highest no of people have taken an insurance in the range of Rs.5,000 to 10,000
10. Year last admitted , covered by any other company ,regular check up last year, weight and weight change in last one year are significantly contributing in explaining the insurance cost.

3.Data Cleaning and Pre-Processing

a)Removal of unwanted variables (if applicable)

We have removed the applicant_id column as it is not required.

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

	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level
0	3	1	1	Salried	2	125 to 150
1	0	0	0	Student	4	150 to 175
2	1	0	0	Business	4	200 to 225
3	7	4	0	Business	2	175 to 200
4	3	1	0	Student	2	150 to 175

Figure 48- Applicant id column removed

a) Missing Value treatment (if applicable)

There are missing values present which we need to treat here for better model performance.

```
applicant_id                        0
years_of_insurance_with_us          0
regular_checkup_lasy_year           0
adventure_sports                    0
Occupation                          0
visited_doctor_last_1_year          0
cholesterol_level                   0
daily_avg_steps                     0
age                                 0
heart_decs_history                  0
other_major_decs_history            0
Gender                              0
avg_glucose_level                   0
bmi                               990
smoking_status                      0
Year_last_admitted              11881
Location                            0
weight                              0
covered_by_any_other_company        0
Alcohol                             0
exercise                            0
weight_change_in_last_one_year      0
fat_percentage                      0
insurance_cost                      0
dtype: int64

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

Figure 49- Missing values dropped

b) Outlier treatment (if required)

We have treated the outliers as we are dealing with a regression problem where outlier can impact the accuracy of the model badly.

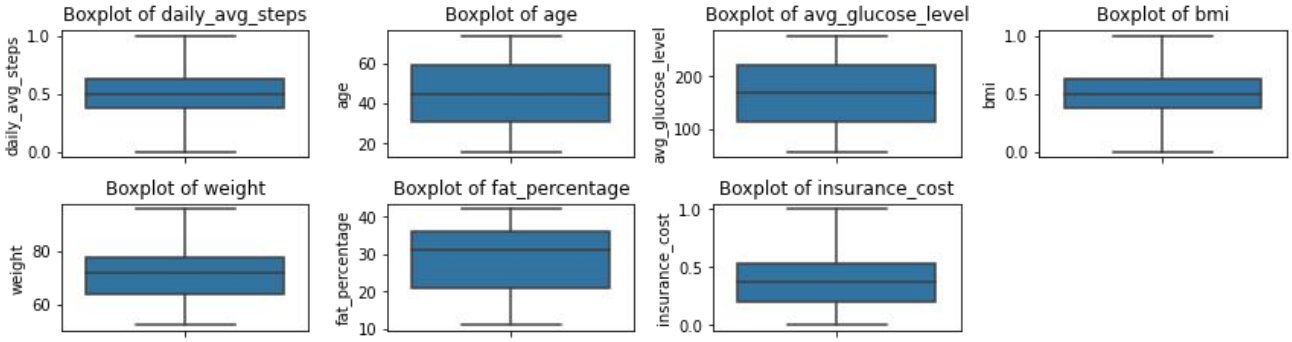


Figure 50- After removal of outlier

c) Variable transformation (if applicable)

We could not get all the outliers removed after the treatment so we had to scale and transform the variables. So we scaled the variable insurance cost, bmi and avg daily steps.

	daily_avg_steps	age	avg_glucose_level	bmi	weight	fat_percentage	insurance_cost
0	0.443029	28.0	97.0	0.511111	67.0	25.0	0.283019
1	0.768429	50.0	212.0	0.594444	58.0	27.0	0.056604
2	0.367839	68.0	166.0	0.766667	73.0	32.0	0.396226
3	0.726938	51.0	109.0	0.280556	71.0	37.0	0.377358
4	0.458193	44.0	118.0	0.380556	74.0	34.0	0.415094

Figure 51- Continuous variables after scaling

d) Clustering

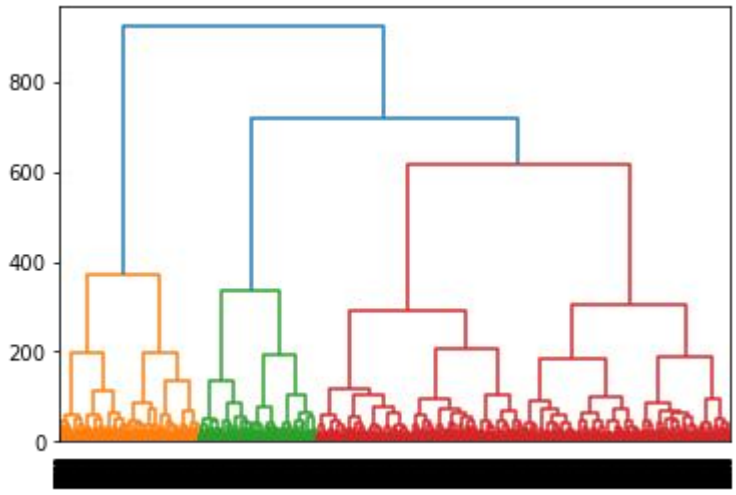


Figure 52-Clusters

clusters	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps
1	5.548282	0.434248	0.057785	0.997504	3.088501	1.280860	0.014927
2	5.472480	0.433749	0.103964	1.022197	3.085844	1.223783	0.010156
3	3.197451	0.986277	0.083442	1.004162	3.114789	1.271787	-0.007973

3 rows × 24 columns

Table 4-Clusters table

Three different clusters have been identified above.

4.Model building.

Why Build various models

Linear Regression :

To understand the strength of relationships between variables.

To know what predictors in a model are statistically significant and which are not.

Decision Tree Regressor :

Easy to understand

Lesser data cleaning is required

Random forest Regressor :

To determine the importance of a given feature and its impact on the prediction.

It computes the score for each feature after training and scales them in a manner that summing them adds to one.

This gives us an idea of which feature to drop as they do not affect the entire prediction process. With lesser features, the model will less likely fall prey to over-fitting.

Artificial Neural Network :

Linear regression can only learn the linear relationship between the features and target and therefore cannot learn the complex non-linear relationship.

In order to learn the complex non-linear relationship between the features and target, we are in need of other techniques. One of those techniques is to use Artificial Neural Networks.

Building Linear Regression

We have separated the dependent and independent variables as x and y and split the data set into train and test set by 70% and 30%

After that the shape of x_train data is (17500,23) and y_test data shape is (7500,1) and y_train shape is (17500,1) and x_test shape is (7500,23)

We have build a basic model at first and found the train data R-Square value to be 0.944 and test data R-Square value to be 0.945

The MSE score is 0.05 and RMSE score is 0.23.

We have also plotted the predicted value against the test data and the plot has come something like below

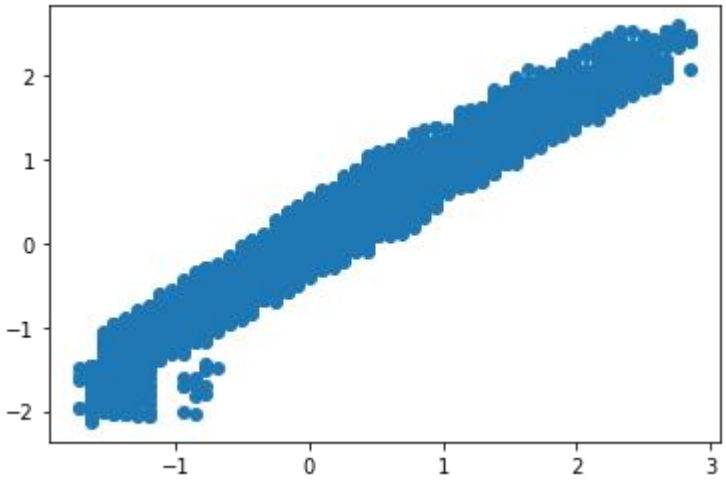


Figure 53-Linear regression prediction against test data

Intercept	0.010988
Occupation	0.002747
cholesterol_level	0.002848
Gender	0.002551
smoking_status	-0.000185
Year_last_admitted	-0.003265
Location	0.000548
covered_by_any_other_company	0.084734
Alcohol	0.000687
exercise	0.000087
years_of_insurance_with_us	-0.000879
regular_checkup_lasy_year	-0.032901
adventure_sports	0.008052
visited_doctor_last_1_year	-0.002545
daily_avg_steps	-0.001995
age	0.003055
heart_decs_history	0.006384
other_major_decs_history	0.004376
avg_glucose_level	0.001603
bmi	-0.000520
weight	0.959220
weight_change_in_last_one_year	0.012979
fat_percentage	-0.000750
dtype: float64	

Table 5-Intercepts of OLS stat model

OLS Regression Results

Dep. Variable:	insurance_cost	R-squared:	0.945
Model:	OLS	Adj. R-squared:	0.945
Method:	Least Squares	F-statistic:	1.358e+04
Date:	Sun, 30 Oct 2022	Prob (F-statistic):	0.00
Time:	19:18:26	Log-Likelihood:	450.41
No. Observations:	17500	AIC:	-854.8
Df Residuals:	17477	BIC:	-676.1
Df Model:	22		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.0110	0.012	0.932	0.351	-0.012	0.034
Occupation	0.0027	0.002	1.183	0.237	-0.002	0.007
cholesterol_level	0.0028	0.002	1.721	0.085	-0.000	0.006
Gender	0.0026	0.004	0.615	0.538	-0.006	0.011
smoking_status	-0.0002	0.002	-0.106	0.916	-0.004	0.003
Year_last_admitted	-0.0033	0.000	-7.773	0.000	-0.004	-0.002
Location	0.0005	0.000	1.328	0.184	-0.000	0.001
covered_by_any_other_company	0.0847	0.004	21.040	0.000	0.077	0.093
Alcohol	0.0007	0.003	0.259	0.795	-0.005	0.006
exercise	8.655e-05	0.003	0.031	0.975	-0.005	0.006
years_of_insurance_with_us	-0.0009	0.001	-1.239	0.215	-0.002	0.001
regular_checkup_lasy_year	-0.0329	0.002	-21.675	0.000	-0.036	-0.030
adventure_sports	0.0081	0.007	1.228	0.219	-0.005	0.021
visited_doctor_last_1_year	-0.0025	0.002	-1.590	0.112	-0.006	0.001
daily_avg_steps	-0.0020	0.002	-1.088	0.277	-0.006	0.002
age	0.0031	0.002	1.712	0.087	-0.000	0.007
heart_decs_history	0.0064	0.008	0.796	0.426	-0.009	0.022
other_major_decs_history	0.0044	0.006	0.705	0.481	-0.008	0.017
avg_glucose_level	0.0016	0.002	0.898	0.369	-0.002	0.005
bmi	-0.0005	0.002	-0.267	0.790	-0.004	0.003
weight	0.9592	0.002	398.079	0.000	0.954	0.964
weight_change_in_last_one_year	0.0130	0.001	11.349	0.000	0.011	0.015
fat_percentage	-0.0007	0.002	-0.417	0.677	-0.004	0.003

Omnibus:	640.409	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	794.665
Skew:	0.418	Prob(JB):	2.76e-173
Kurtosis:	3.626	Cond. No.	118.

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

Table 6-OLS Summary

From the above table we can observe that there are only 5 variables where the p value is less than 0.05 and the null hypothesis is rejected.

Regression equation :

Insurance cost = 0.00*occupation+0.00*cholesterol_level+0.00*gender-0.00*smoking_status-0.00*year_last admitted+0.00*location+0.08*covered_by_any_other_company+0.00*Alcohol+0.00*exercise-0.00*years_of insurance_with_us-0.03*regular_checkup_last_year+0.00*adventure_sports-0.00*visited_doctor_last_1_year-0.00*daily_avg_steps+0.00*age+0.00*heart_decease_history+0.00*other_major_decease_history+0.00*avg*glucose level-0.00*bmi+weight*0.95+0.01*weight_change_in_last_one_year-0.00*fat percentage

Building ANN, Decision Tree and Random Forest Regressor

We have made three models ANN , Decision tree and Random forest and below are the score mapped.

	Train RMSE	Test RMSE	Training Score	Test Score
Decision Tree Regressor	3.111544e-17	0.296986	1.000000	0.910455
Random Forest Regressor	7.969658e-02	0.214129	0.993689	0.953450
ANN Regressor	1.762023e-01	0.257123	0.969151	0.932880

Table 7-Results of ANN , DT and RF

From the above table we can see that the Random forest regression model is giving the best score with RMSE score 0.214

Improving Model Performance

To tune the models we are performing grid search CV to see if we are getting a better score

After performing the grid search CV we have found the below results.

	Train RMSE	Test RMSE	Training Score	Test Score
Decision Tree Regressor	0.197853	0.217529	0.961104	0.951960
Random Forest Regressor	0.217861	0.234351	0.952839	0.944242
ANN Regressor	0.226199	0.226050	0.949160	0.948122

Table 8- Results after Grid search cv

After grid search CV we can see that Decision tree model is giving 0.217 RMSE score which is the lowest.

We have also plotted the feature importance plot

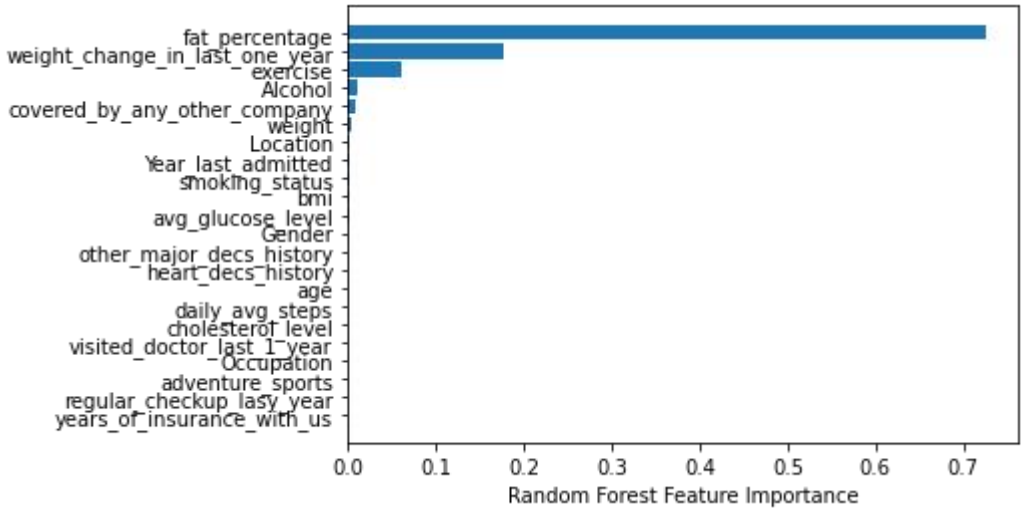


Figure 54-Feature importance plot of Random forest

From the above table we can observe that the most important feature is fat percentage and least one is years of insurance with us.

We will build another model with only the important features.

Below are the important features we are considering for building model.

'fat_percentage','weight_change_in_last_one_year','exercise','Alcohol','covered_by_any_other_company','weight'

We have found the below scores

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	0.239296	0.235292	0.943103	0.943794
Decision Tree Regressor	0.098839	0.297851	0.990293	0.909932
Random Forest Regressor	0.122338	0.243845	0.985129	0.939633
ANN Regressor	0.214520	0.218374	0.954275	0.951586

Table 9-Results of the models built from only important features

From the above table we can observe that ANN model is giving the best result with 0.218 as RMSE score.

Again we would like to do grid search CV to tune the models further and see if we can a better score.

	Train RMSE	Test RMSE	Training Score	Test Score
Linear regression	0.239296	0.235292	0.943103	0.943794
Decision Tree Regressor	0.210464	0.219969	0.955988	0.950876
Random Forest Regressor	0.212880	0.216330	0.954971	0.952488
ANN Regressor	0.214520	0.218374	0.954275	0.951586

Table 10-Results of model built from important features after doing grid search cv

From the above plot we can observe that Random forest regressor is giving the best score with 0.216 RMSE score.

5.Model Validation

Optimum model.

	Model Name	Test RMSE score
Regular	Linear regression	0.232
	Decision Tree	0.296
	ANN	0.257
	Random Forest	0.214
After Grid search cv	Linear regression	0.232
	Decision Tree	0.217
	ANN	0.226
	Random Forest	0.234
Taking the important features only	Linear regression	0.235
	Decision Tree	0.297
	ANN	0.218
	Random Forest	0.243
Taking the important features only and then doing grid search cv	Linear regression	0.235
	Decision Tree	0.219
	ANN	0.218
	Random Forest	0.216

Table-11 - RMSE score on test data of all models.

After mapping all the RMSE scores of the models on test data we can conclude that the Random forest model is performing best with 0.214 score.

Business implications :

The random forest builds on the decision tree model, and makes it more sophisticated. Random forests representing “stochastic discrimination” or the “stochastic guessing” method on data applied to multidimensional spaces. Stochastic discrimination tends to be a way to enhance the analysis of data models beyond what a single decision tree can do.

Basically, a random forest creates many individual decision trees working on important variables with a certain data set applied. One key factor is that in a random forest, the data set and variable analysis of each decision tree will typically overlap. That's important to the model, because the random forest model takes the average results for each decision tree, and factors them into a weighted decision. In essence, the analysis is taking all of the votes of various decision trees and building a consensus to offer productive and logical results.

One way that this could be applied to business is to take various health parameter data from people and use a random forest to indicate how much insurance they should take based on their risk level.

This model needs to be changed time to time as it is a regression model and it has been trained on train data and it will perform in this ambit , any extrapolation might result in poor performance.

6. Final Recommendation

- The insurance premium should increase with the fat percentage, exercise and weight change in a year of an individual.
- Insurance cost should decrease with daily average steps , visited doctor last year , regular check up last year and years of insurance with the same insurance company.
- With one unit increase in weight , insurance cost should increase by 0.95unit.