Prediction of Insurance Cost

1. Business problem and Objective

Health care is a very important domain and the demand for good health care is continuously increasing in India. A country where around 70% of treatment takes place in private hospitals , money plays an important role because if treatment becomes costly and the individual is not covered by any health insurance company then it will become a very tough financial situation for the individual.
Insurance companies also want to reduce their risk of loss by optimising the insurance cost.
So the objective of this project is to make a model based on data provided by insurance company which will predict appropriate insurance costs. so that insurance does not become very costly for the individual and reduces the risk of loss for insurance companies.

2. EDA and Business implication

Data is collected by insurance companies when a customer has come to buy an Insurance. Data consists details of 25000 individuals

Number of Rows - 25000
Number of columns - 24
Dependent variable - Insurance_cost

Variables are as following -

Variables	Business Definition
1. Applicant id	Applicant unique id
2. Years_of_insurance_with_us	Since how many years individual is taking insurance from same company only
3. Regular_checkup_last_year	Number of times customer has done regular check up last year.
4. Adventure_sports	Customer is involved in adventure sports like climbing, diving etc.
5. Occupation	Occupation of the customer
6. Visited_doctor_ last_1_year	Number of times customer has visited doctor

	Last one year
7. Cholesterol_level	Cholesterol level of customer while applying for insurance
8. Daily_avg_steps	Average daily steps walked by customers
9. Age	Age of the customer
10. Heart_decs_history	Any past heart diseases
11. Other_major_decs_history	Any past major disease apart from heart like any operation
12. Gender	Gender of the customer
13. Avg_glucose_level	Average glucose level of the customer while Applying the insurance
14. Bmi	BMI of the customer while applying the Insurance
15. Smoking_status	Smoking status of the customer
16. Year_last_admitted	When customer have been admitted last time
17. Location	Location of the hospital
18. Weight	Weight of the customer
19. Covered_by_any_other_company	Customer is covered by any other insurance Company
20. Alcohol	Alcohol consumption status of the customer
21. Exercise	Regular exercise status of the customer
22. Weight_change_in_last_1_year	How much variation has been seen in the Weight of customer
23. Fat_percentage	Fat percentage of customer while applying The insurance
24.Insurance_cost	Total cost of insurance

- Variable applicant id has been dropped before proceeding towards EDA because it is just an identification number and is not going to play any role in making predictions

- 5 point summary of some variables

Table -1

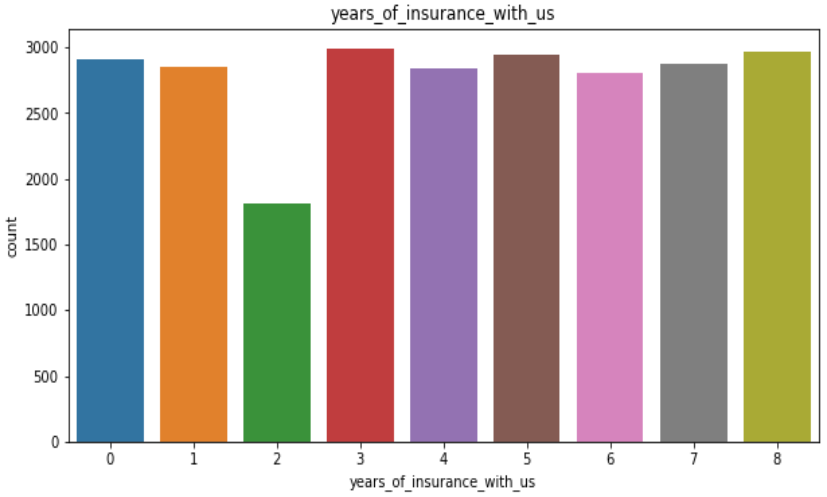
	daily_avg_steps	age	avg_glucose_level	bmi	Year_last_admitted	weight	fat_percentage	insurance_cost
count	25000	25000	25000	24010	13119	25000	25000	25000
mean	5215.89	44.92	167.53	31.39	2003.89	71.61	28.81	27147.41
std	1053.18	16.11	62.73	7.88	7.58	9.33	8.63	14323.69
min	2034.00	16.00	57.00	12.30	1990.00	52.00	11.00	2468.00
25%	4543.00	31.00	113.00	26.10	1997.00	64.00	21.00	16042.00

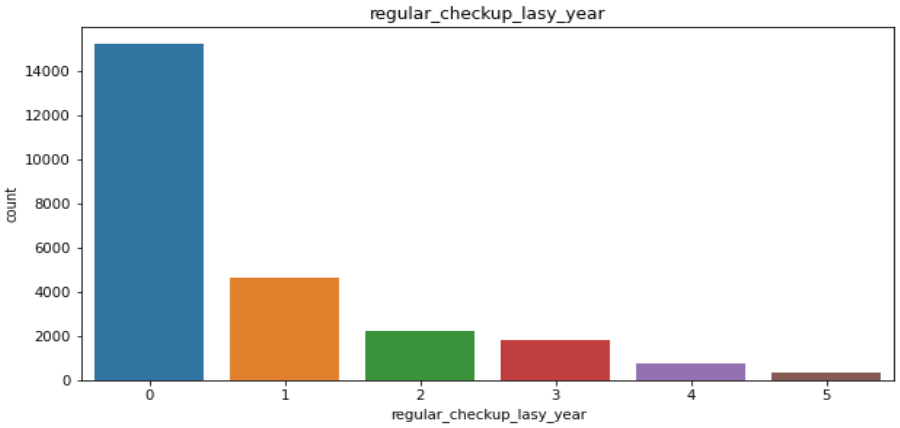
50%	5089.00	45.00	168.00	30.50	2004.00	72.00	31.00	27148.00
75%	5730.00	59.00	222.00	35.60	2010.00	78.00	36.00	37020.00
max	11255.00	74.00	277.00	100.60	2018.00	96.00	42.00	67870.00

Inference of table 1- All variables are normally distributed
- BMI and year_last_admitted have missing values

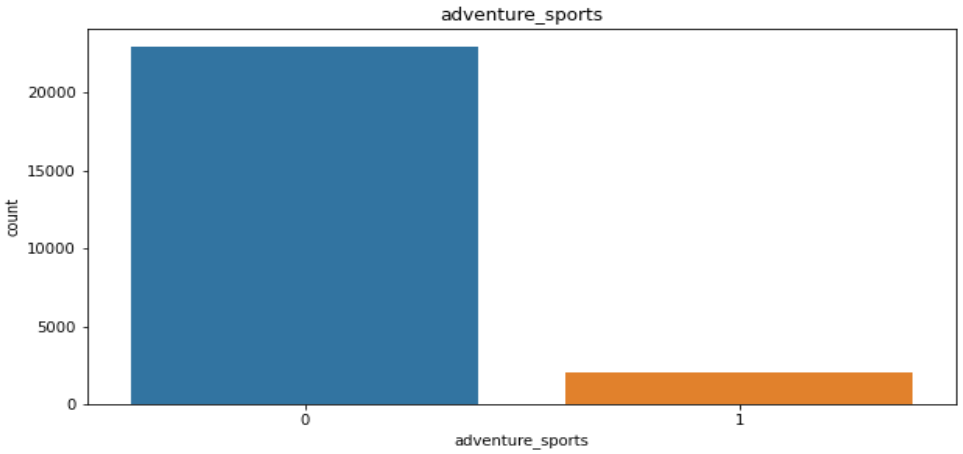
Univariate Analysis

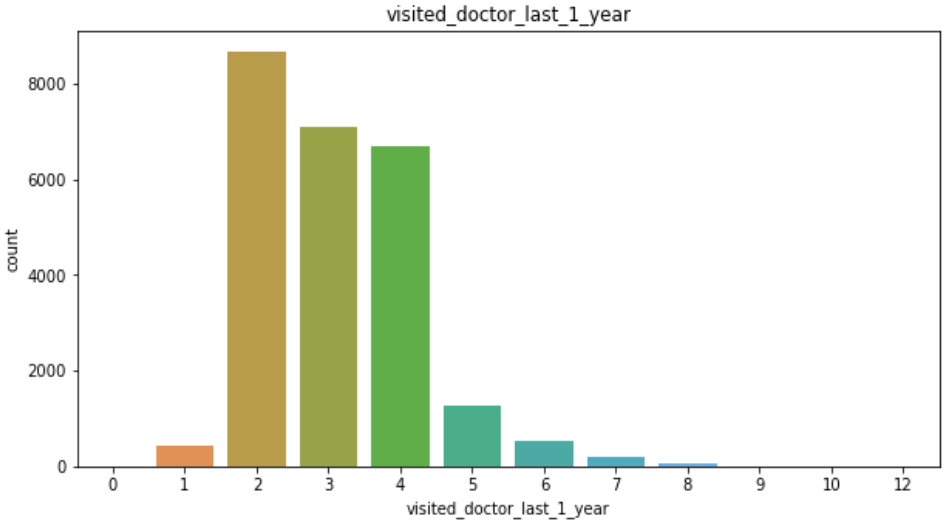
1. Categorical variables - Analysis is done by count plots

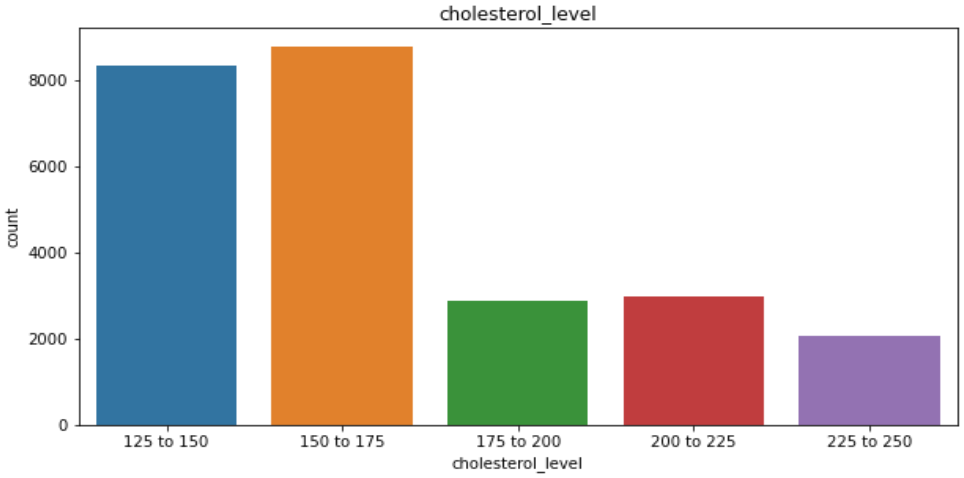


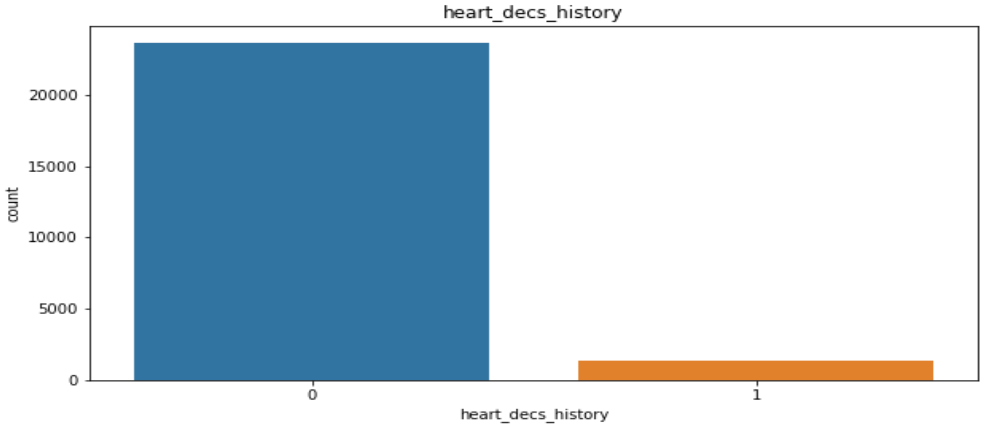


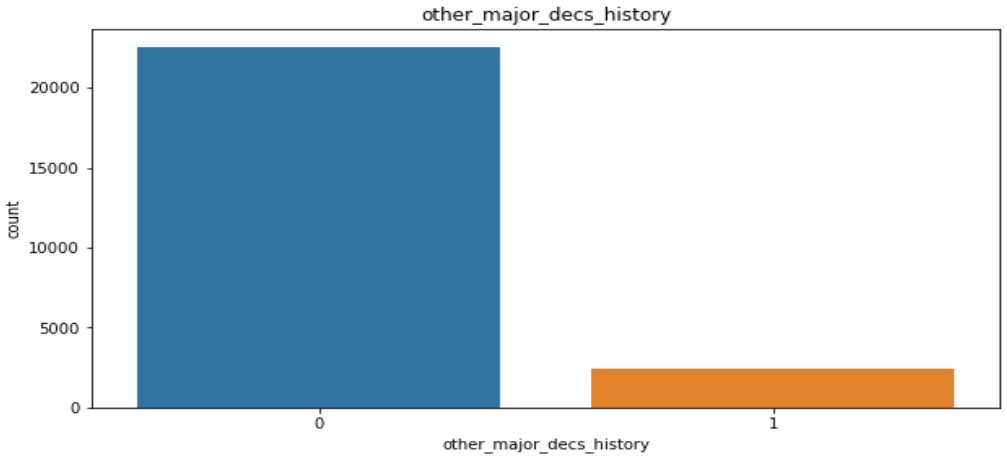

Occupation
10000
8000
6000
4000
2000
0
count
Business
Salried
Student
Occupation

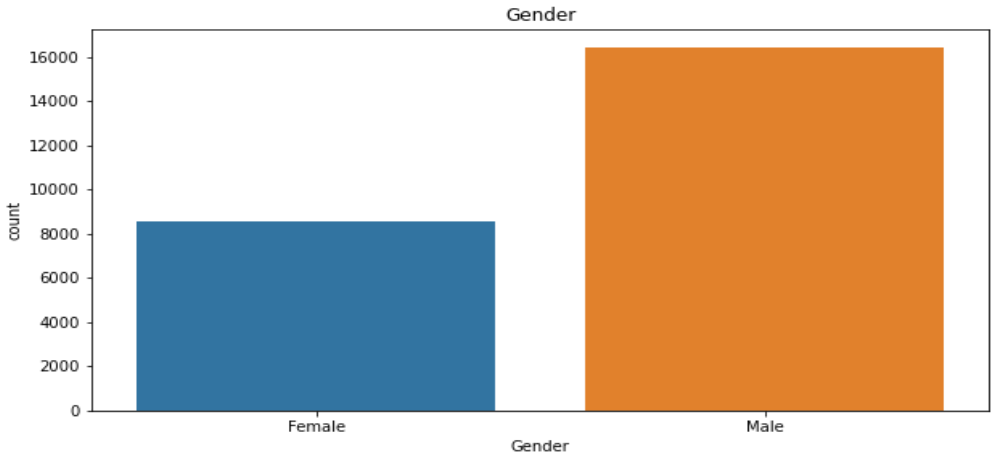

adventure_sports
20000
15000
10000
5000
0
count
0
1
adventure_sports

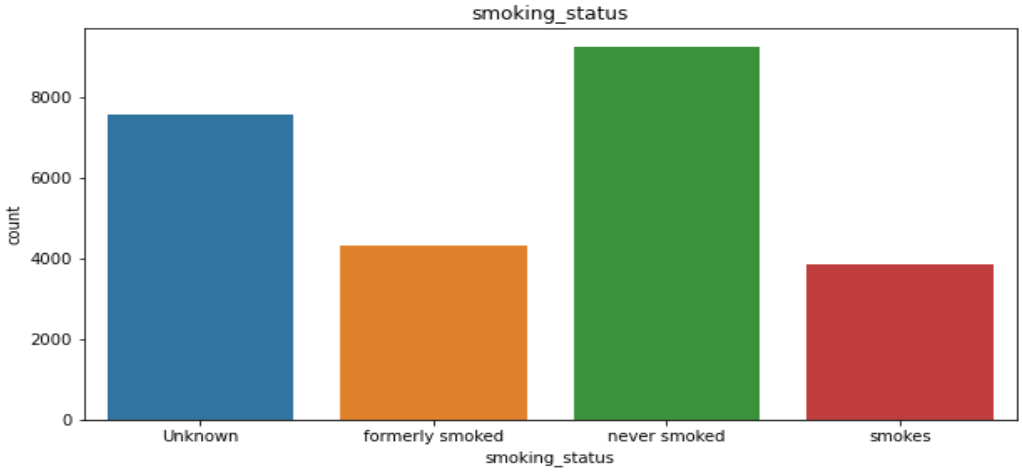

visited_doctor_last_1_year
8000
6000
4000
2000
0
count
0
1
2
3
4
5
6
7
8
9
10
12
visited_doctor_last_1_year

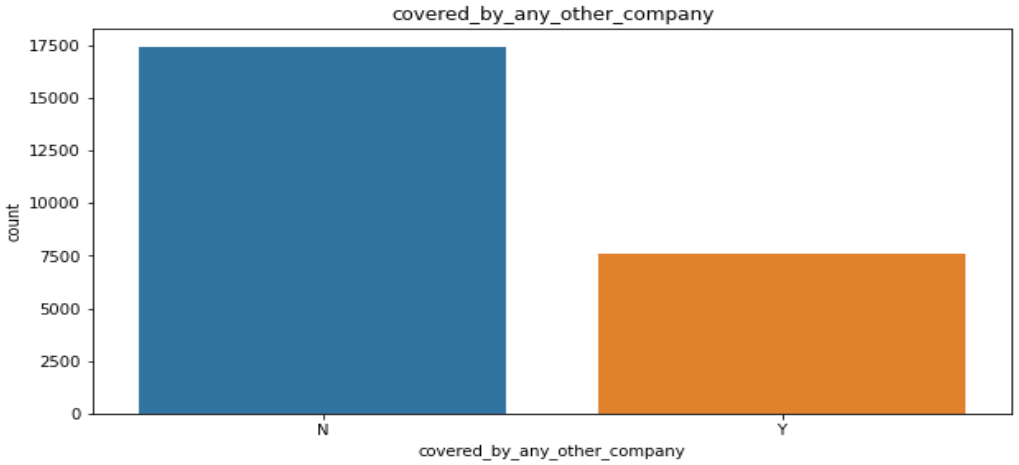

cholesterol_level
count
8000
6000
4000
2000
0
125 to 150
150 to 175
175 to 200
200 to 225
225 to 250
cholesterol_level

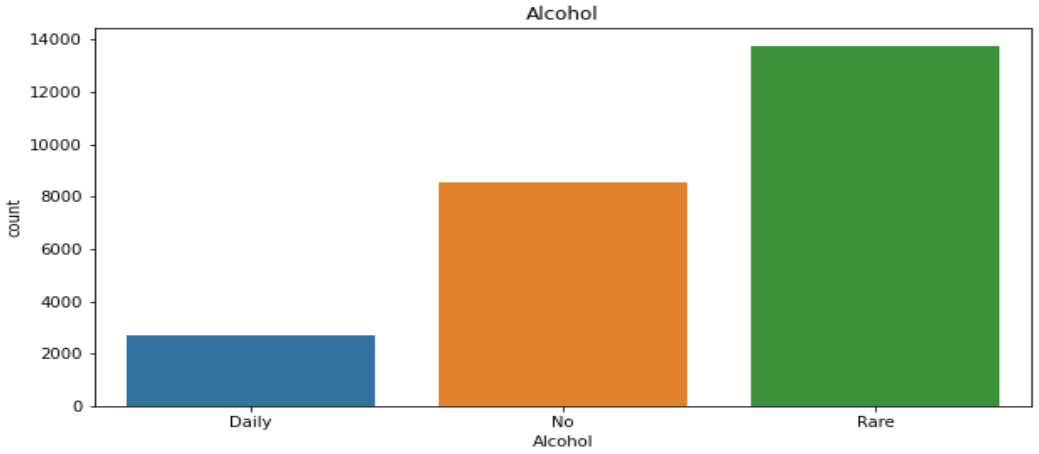

heart_decs_history
count
20000
15000
10000
5000
0
0
1
heart_decs_history

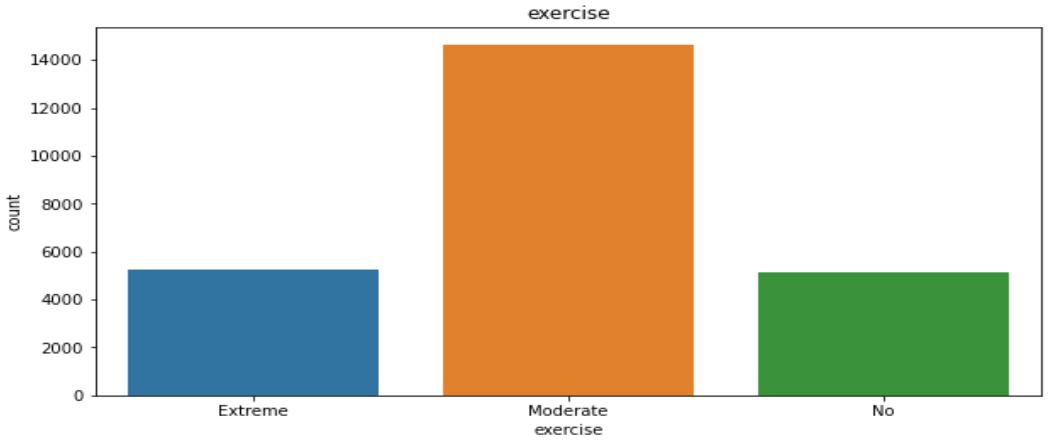

other_major_decs_history
count
20000
15000
10000
5000
0
0
1
other_major_decs_history

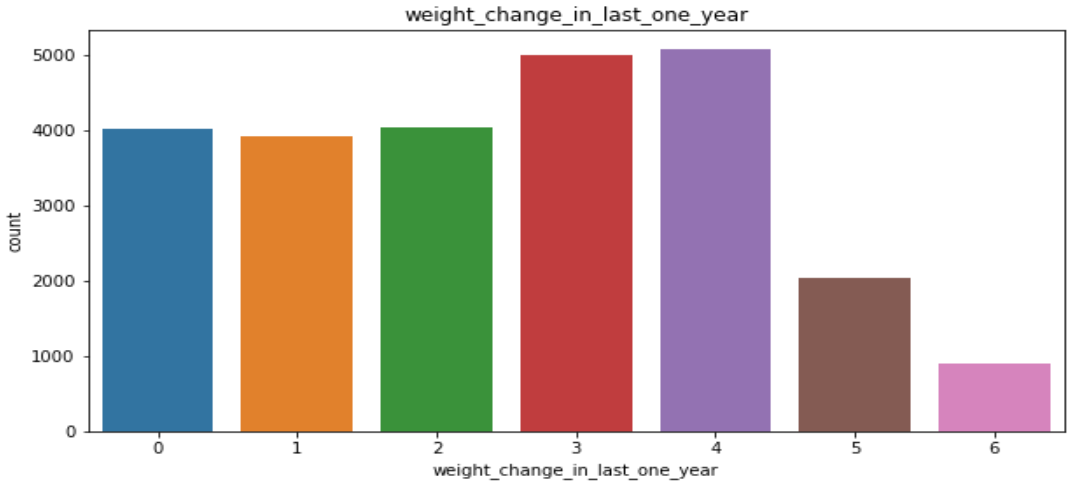

Gender
16000
14000
12000
10000
8000
6000
4000
2000
0
count
Female
Male
Gender


smoking_status
8000
6000
4000
2000
0
count
Unknown
formerly smoked
never smoked
smokes
smoking_status


covered_by_any_other_company
17500
15000
12500
10000
7500
5000
2500
0
count
N
Y
covered_by_any_other_company





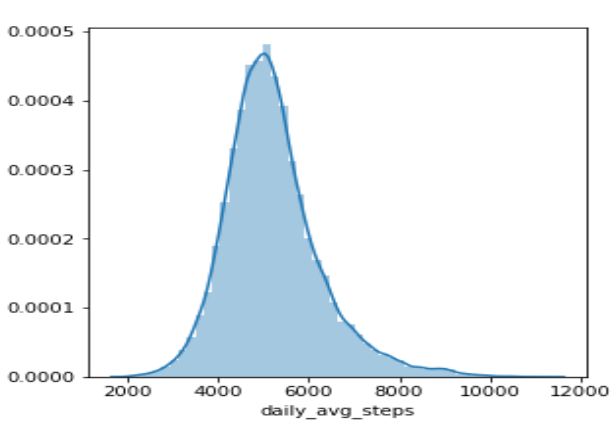


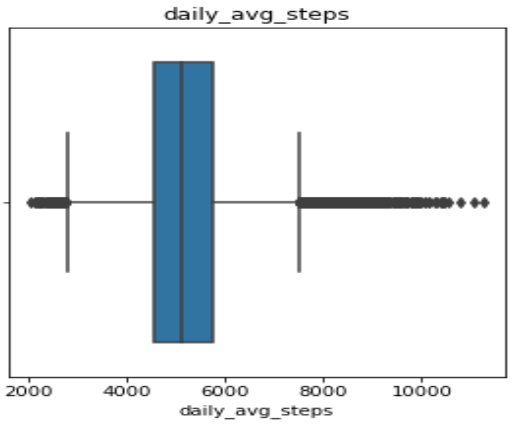
Inference of Categorical univariate analysis -

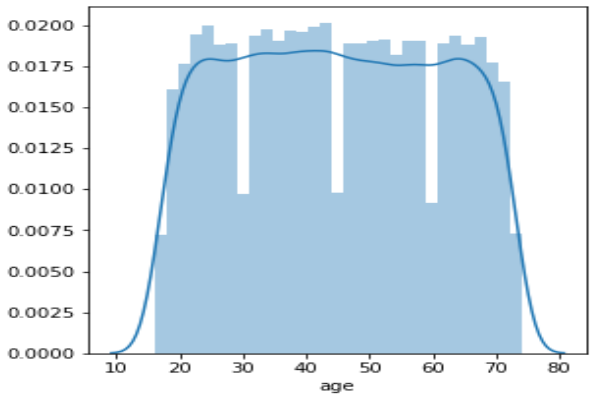
- Around 75% of customers have been taking insurance from the same company for a very long time. Around 15% of customers are new.
- Around 60% of customers don't get regular checkups done and around 22% of customers get regular checkups done 2-5 times a year.
- 40% of customers are students and 60% of customers are in occupation.
- Majority of customers don't take part in adventure sports.
- Majority of customers have visited doctors 2-4 times a year for regular checkups.

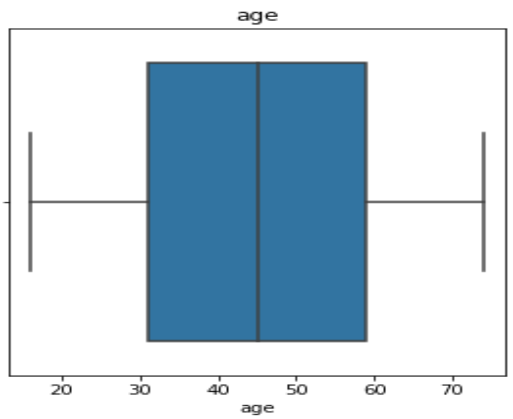
- Around 78% of customers have cholesterol in normal range(125-200) and around 22% of customers have high cholesterol level
- Majority of customers have had no heart disease or any other major diseases in the past.
- 34% of customers are female and 66% of customers are male.
- Around 36% of customers have never smoked.Around 16% of customers are current smokers and around 16% of customers used to smoke.
- Around 70% of customers are not covered by some other insurance company.
- Around 10% of customers consume alcohol daily and around 35% of customers do not consume alcohol.
-Around 56% of customers do regular exercises and 20% do not do any exercises.
- Only 4000 of customers have no change in their weight while majority of the customers have changed in their weight from 1-6 kgs.

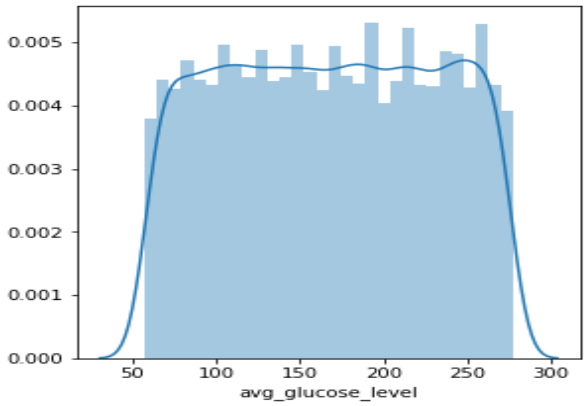
2. Continuous Variables - Analysis is done by distribution plot ant box plot

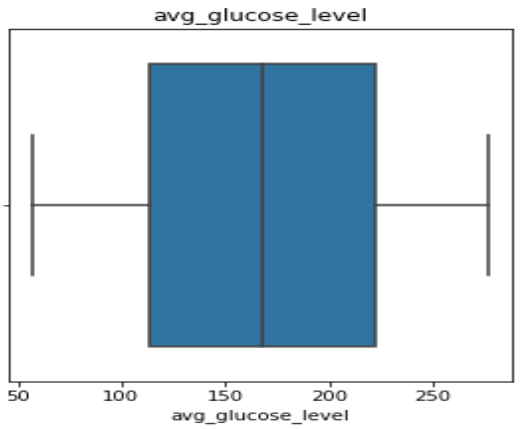


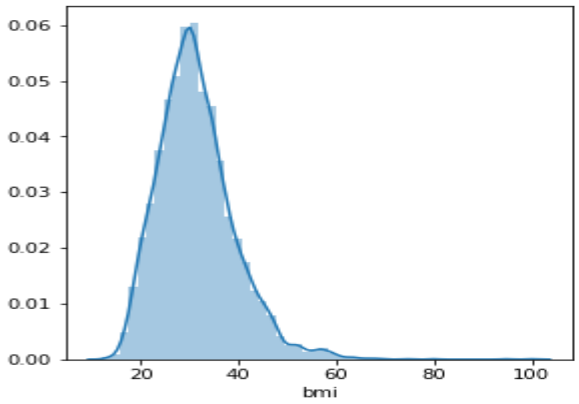


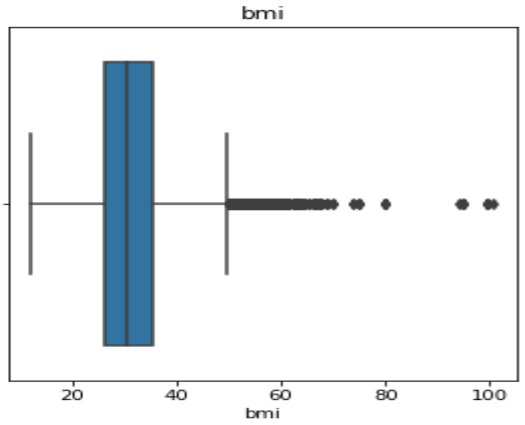


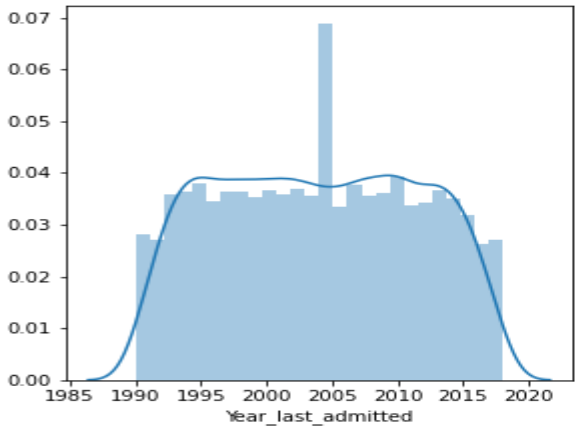


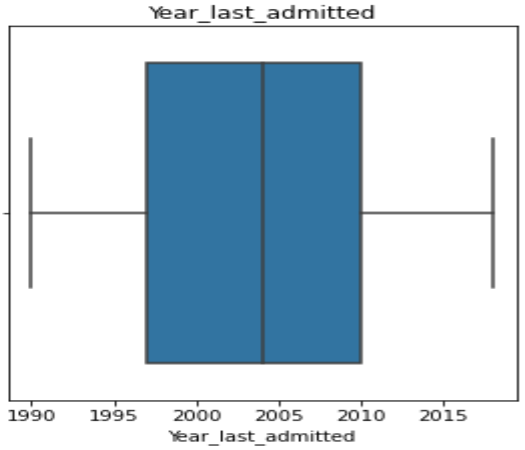

0.005
0.004
0.003
0.002
0.001
0.000
50
100
150
200
250
300
avg_glucose_level

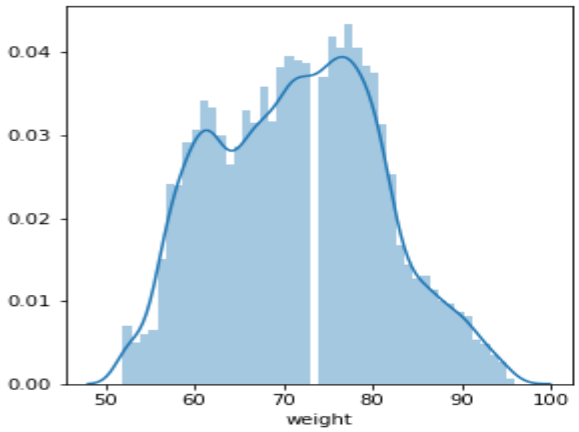

avg_glucose_level
50
100
150
200
250
avg_glucose_level

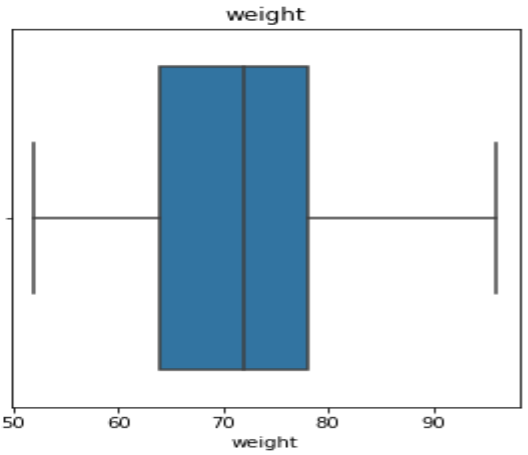

0.06
0.05
0.04
0.03
0.02
0.01
0.00
20
40
60
80
100
bmi

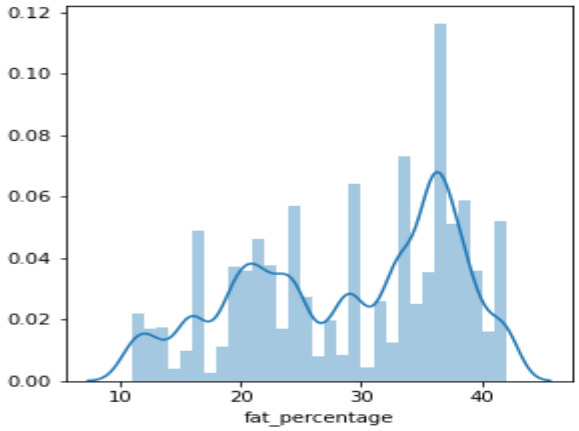

bmi
20
40
60
80
100
bmi

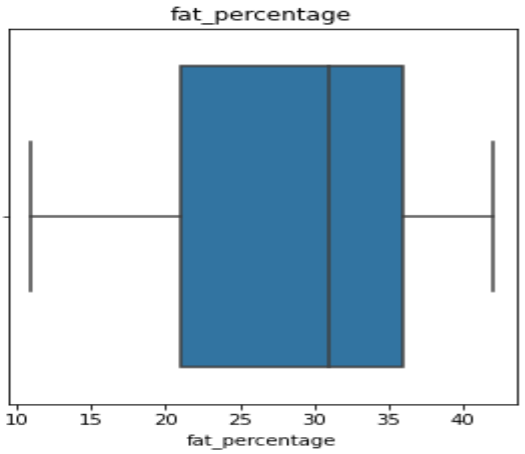

0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
1985
1990
1995
2000
2005
2010
2015
2020
Year_last_admitted

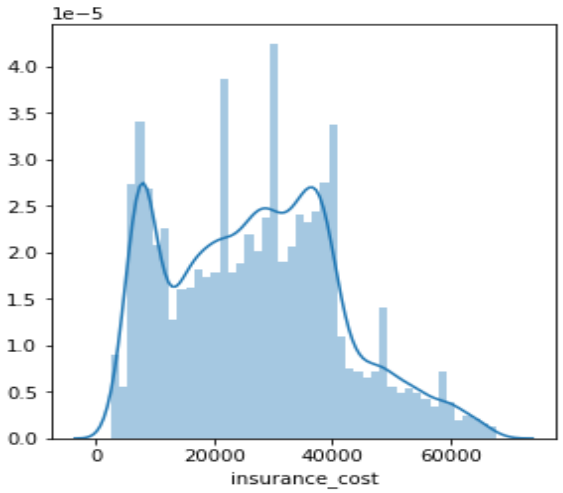

Year_last_admitted
1990
1995
2000
2005
2010
2015
Year_last_admitted

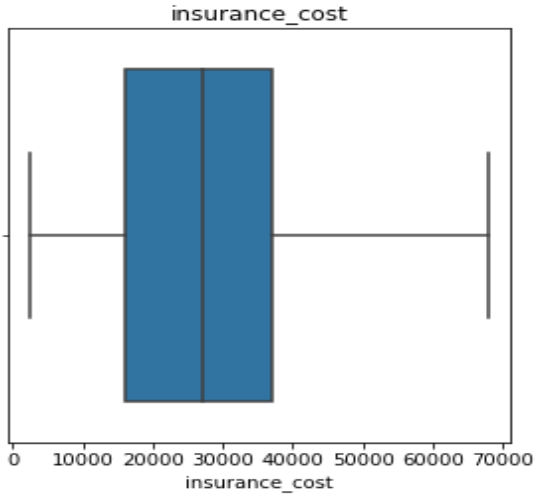

0.04
0.03
0.02
0.01
0.00
50
60
70
80
90
100
weight


weight
50
60
70
80
90
weight









Inference of continuous variables univariate analysis-

- All variables except fat_percentage are normally distributed and not skewed.
- Fat_percentage distribution is negatively skewed.
- Daily_avg_steps have outliers in upper and lower sides while bmi has outliers on Upper side.

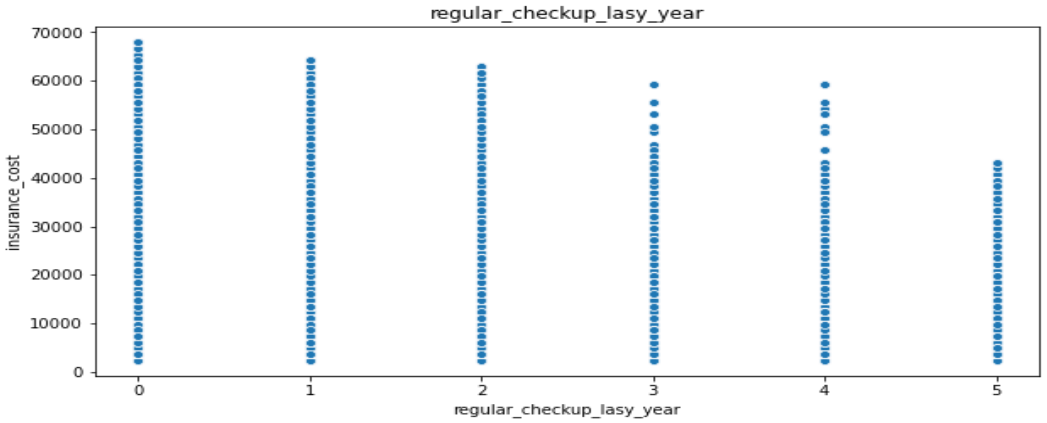
Bivariate Analysis - Here bivariate analysis is done by using ANOVA hypothesis test and Scatter plot.

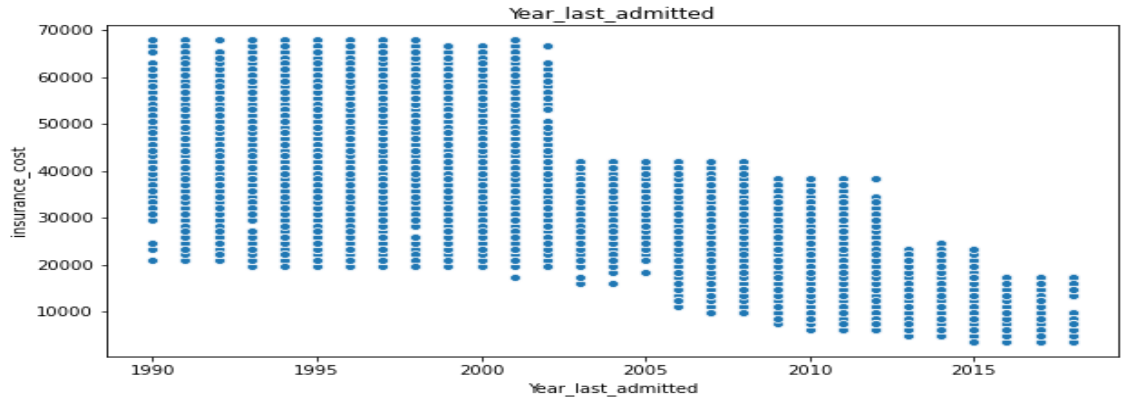
1. ANOVA Hypothesis test -
   - Null Hypothesis : The variables are not correlated to each other
   - P-value : The probability of null hypothesis being true
   - Accept Null hypothesis if P-value > 0.05 which means variables are not correlated to each other.
   - Reject Null hypothesis if P-value < 0.05 which means variables are correlated to each other.
   - Variables having P-value < 0.05 and hence correlated to dependent variables are as followings -

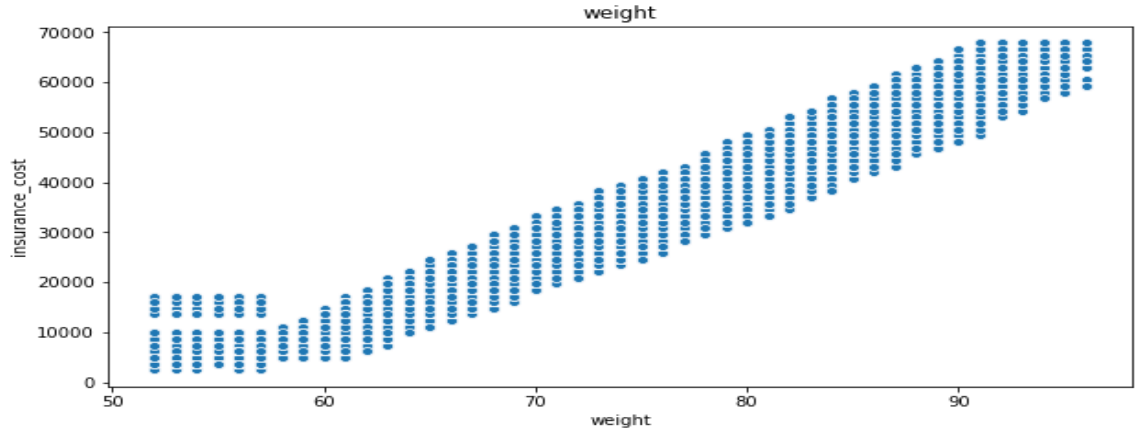
Variables	P-value
1. Regular_checkup_lasy_year	4.8128016959475236e-182
2. Adventure_sports	3.68036823282535e-32

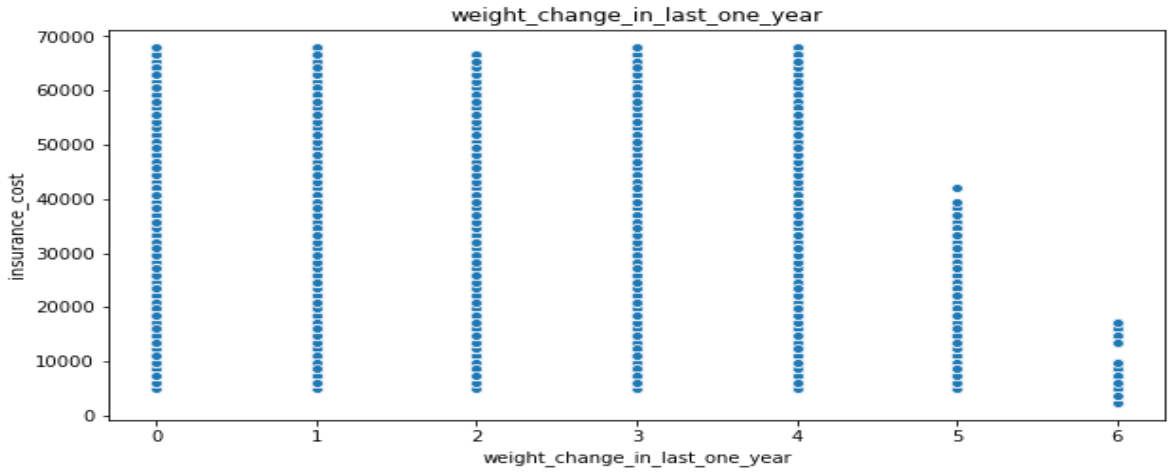
Variables	P-value
3. Year_last_admitted	0.0
4. Weight	0.0
5. Weight_change_in_last_one_year	0.0
6. Covered_by_any_other_company	2.2007147247396044e-58

2. Scatter plots - Scatter plots of some important variables showing correlation with dependent variable is as followings -







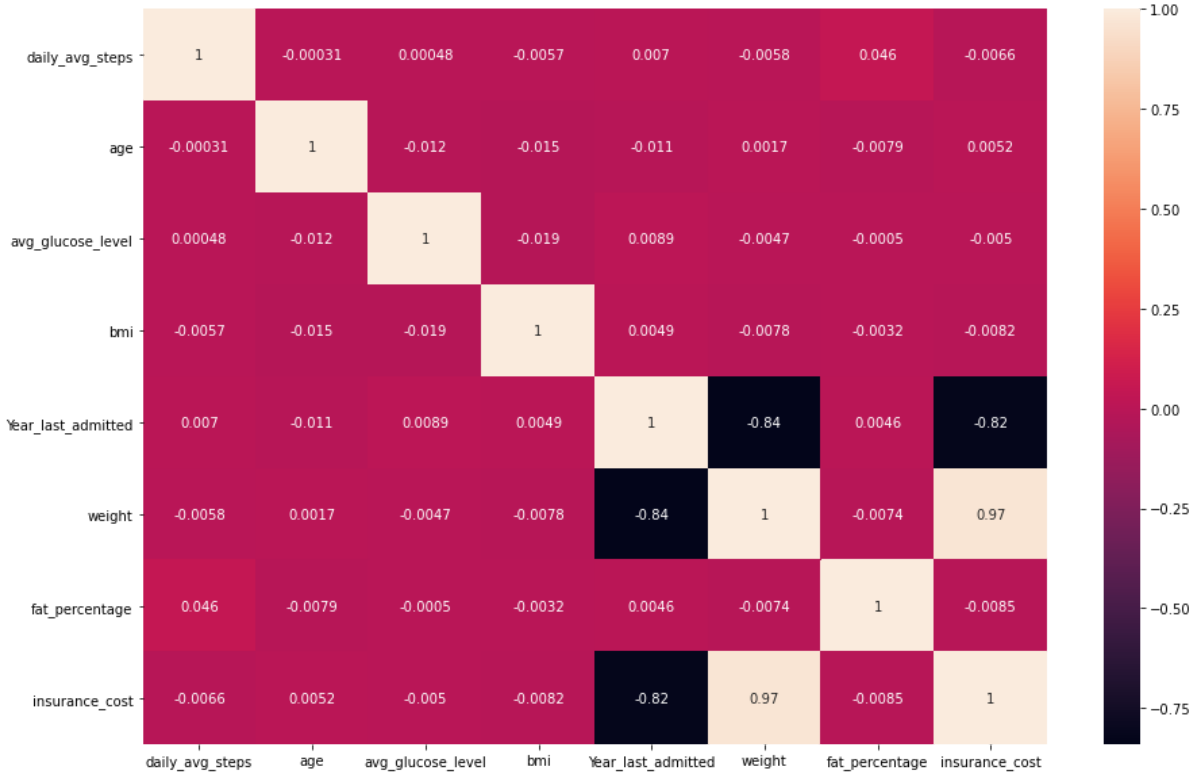


Inference from scatter plot -

- Number of regular checkups last year have negative correlation with insurance cost. As the number of checkups increases, insurance costs are decreasing.
- Year_last admitted has negative correlation with insurance_cost. If a customer was admitted recently insurance cost is less and if the customer was admitted way back in the past insurance cost is less more.
- Weight is showing positive correlation with insurance cost. If weight is increasing, insurance costs are increasing.
- If a customer is not having insurance from some other companies as well then insurance cost is less.
- Weight_change_in_last_one_year is showing negative correlation with insurance_cost.If one is increasing the other one is decreasing.

Multivariate Analysis - Here multivariate analysis is done by pair plot and heat map

Heat map



Inference from heat map -

- Year_last_admitted is negatively correlated to weight and insurance_cost.
- Weight is positively correlated to insurance_cost.

Business implication of EDA

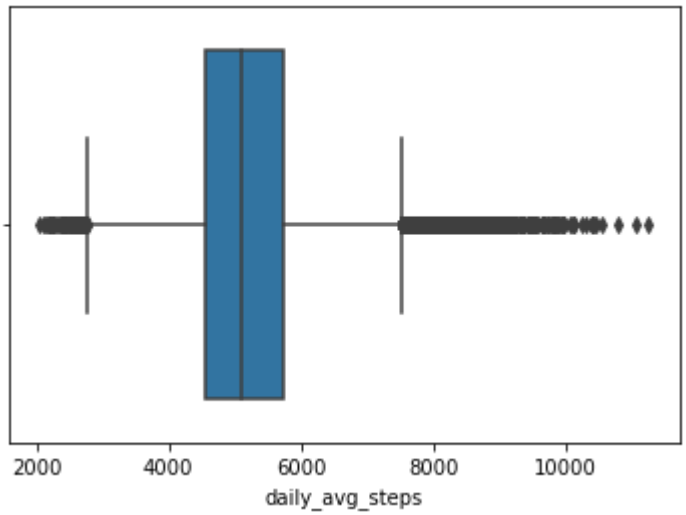
- For customers who are getting regularly checked insurance_cost will be less and are not getting regularly checked insurance_cost will be more.
- If Customer has recently been admitted to a hospital, the insurance cost will be less and if the patient was admitted in long past, the insurance cost will be less.
- If weight is more , insurance cost will be high and aif it is less insurance cost will be less.
- If a customer has taken insurance from another company as well then insurance costs will be high.
- If the weight change in the last one year is more than insurance cost will be less.

3. Data cleaning and pre-processing

1. Missing value treatment -
   - There are 990 missing values in BMI and 11881 missing values in Year_last_admitted.
   - As missing value is around 47% and year_last_admitted is showing strong correlation with dependent variables we are going to impute missing values. We have used K nearest neighbour for imputation because it is a very robust way to impute missing values.KNN algorithm identify 'k' samples in the data set that are close or similar in the space and then we use these 'k' to identify missing data points. Each sample's missing value is then imputed using mean value of the 'k' neighbours found in the data set. Here k=10.
   - As missing values of bmi is just 0.3% and is not showing any correlation with dependent variable, we are going to delete these missing values.

2. Outlier treatment
   - Only daily_avg_steps are having outliers in both upper and lower sides.

Box plot showing outliers



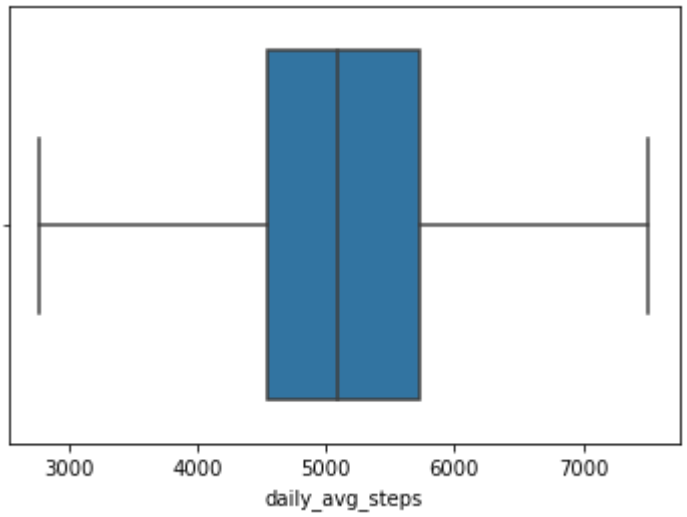
- Outliers are treated by bringing then in normal range where

Lower range = Q1-(1.5*IQR)

Upper range = Q3+(1.5*IQR)

IQR = Q3-Q1

Box plot after outlier treatment



3. Variable Transformation - Last_year_admitted is transformed to Year_since_last_admitted. Values of years are replaced by interval since last admitted keeping 2022 as the current year.

4. Variables removed - Applicant id and bmi are removed.

4. Model building

As we have to predict a continuous variable, we are going to use a regression model. We are going to use both linear and non-linear regression to make the prediction.

Steps used in model buildings are as following -
- Data set is divided into train and test sets in a 70:30 ratio.
- Then models are made and fitted on to train set
- model is then validated on test set.
- Performance of the model is checked.
- If model is not performing well, to improve the performance tuning of model is done.

1. Non-Linear regression model or Ensemble method -
Going to use this because it uses decision trees and gives the average of the predictions produced by trees in the forest.

1) Random forest Regression- This works by constructing a multitude of decision trees and choosing the best prediction from a decision tree.
   - To improve the performance tuning is done by using GridsearchCV and it resulted in slight increase in the performance and hence final model is made by using following parameters-
     Max_depth = 10
     Max_features = 15

Min_samples_leaf = 5
Min_samples_split = 10
n_estimators = 501

2) ADA Boosting - It starts by giving equal weight to each observation. If prediction is incorrect using the first learner,then it gives higher weight to observations which have been predicted incorrectly. It continues to add learners until a limit is reached.
- Model showed improve performance by changing n_estimator and hence model is Made by using following parameters -
n_estimators = 5
Learning rate = 1.0

3) Gradient Boosting - It also work on the basis of training weak learner sequentially and It acts by reducing bias error when the next model is combined with the previous model.
- As changing parameter did not result in improving performance of the model, i have made the model by using default parameters.
N_estimator = 100
Learning_rate = 0.1
Subsample = 1.0
Criterion = friedman_mse
Min_sample_split = 2
Min_sample_leaf = 1
Max_depth = 3

4) Bagging - Here a random sample of data in a training set is selected with replacement. After several data samples are generated, these weak models are then trained in parallel and predictions are made by using average of independent predictions.

2. Linear regression model -
This will help us to understand how dependent variable depends on one or more independent variables and mean change in dependent variable given a unit change in independent variables.

1) Linear regression - On using Variance inflation factor(VIF) and removing independent variables based on high VIF to reduce multicollinearity resulted in decreased model performance and hence made model with default parameters.

2) Lasso regression - It works by obtaining the subset of predictors that minimises the prediction error for a dependent variable. Lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.
- Model is made by using following parameters -
Alpha = 1.0
Max_iter = 1000

Tol = 0.0001

3) Ridge Regression - This is a model tuning method that is used to analyse any data that suffers from more multicollinearity. This method uses the L2 regularisation. When the issue of multicollinearity occurs,least-squares are unbiased,and variances are large. Ridge regression tries to overcome these problems if multicollinearity occurs.
   - Model is made by using following parameters -
     Alpha = 1.0
     Tol = 0.001

4) Elastic net Regression - Elastic net is an extension of linear regression that combines L1 and L2 penalty function to the loss function during training.
   - Model is made by using following parameters -
     Alpha = 1.0
     L1_ratio = 0.5
     max_iter=1000
     Tol = 0.0001

5. Model Validation

To compare performances of different models following metrics are used -

i) Absolute_R2
ii) RMSE - Root mean squared error
iii) MAPE - Mean absolute percentage error

Comparison of different models performances

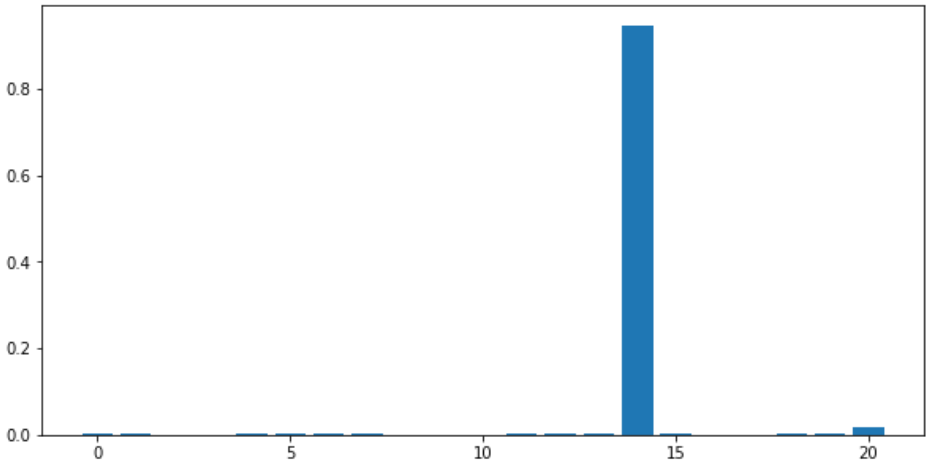
Model	Absolute_R2		RMSE		MAPE	
	Train	Test	Train	Test	Train	Test
1. Random forest	0.97	0.96	2301.92	2701.2	8.28	9.62
2. ADA Boost	0.94	0.94	3214.6	3206.8	13.87	13.87
3. Gradient Boost	0.96	0.96	2715.8	2754.9	10.11	10.27
4.Bagging	0.99	0.96	1219.2	2869.3	4.00	10.02
5.Linear Regression	0.94	0.94	3291	3249.5	14.3	14.1
6.Lasso Regression	0.94	0.94	3291	3249.5	14.3	14.1
7.Ridge Regression	0.94	0.94	3291	3249.5	14.3	14.1
8. Elasticnet Regression	0.94	0.94	3318.9	3274.5	14.1	13.9

Important variables of non-linear regression

- Important variables are determined by coefficients for that particular variable. If the coefficient of the variable is nearing 0 then that variable is not playing any role in making predictions.
- Important variables based on coefficients are -
    1) Feature 0 : years_of_insurance_with_us
    2) Feature 1 : regular_checkup_last_year
    3) Feature 4 : visited_doctor_last_1_year
    4) Feature 5 : cholesterol_level
    5) Feature 6 : daily_avg_steps
    6) Feature 7 : age
    7) Feature 11 : avg_glucose_level
    8) Feature 12 : smoking_status
    9) Feature 13 : Location
    10) Feature 14 : weight
    11) Feature 15 : covered_by_any_other_company
    12) Feature 18 : weight_change_in_last_one_year
    13) Feature 19 : fat_percentage
    14) Feature 20 : years_since_last_admitted

- Out of all these variables weight is most important

Bar Graph showing important variables



Important variables in linear regression -

Important variables in linear regression are determined by P-value,if P-value is less than 0.5 then the variable is important.

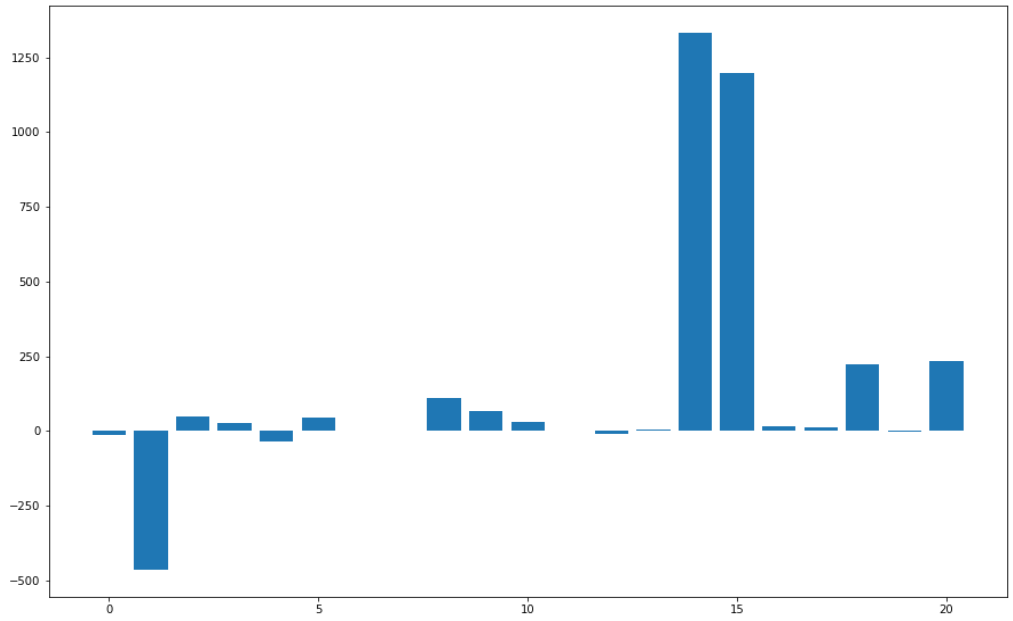
- Following variables have P-value< 0.5 and thus are important for linear regression-

1) regular_checkup_last_year
2) regular_checkup_last_year
3) cholesterol_level
4) cholesterol_level
5) heart_decs_history
6) heart_decs_history
7) covered_by_any_other_company
8) weight_change_in_last_one_yea
9) years_since_last_admitted

- Coefficients of different variables will tell us change in 1 unit of independent variables will produce how much change in dependent variable.

Variable	Coefficient
Feature: 0: years_of_insurance_with_us	Score: -13.51564
Feature: 1: regular_checkup_last_year	Score: -463.88201
Feature: 2: adventure_sports	Score: 49.83166
Feature: 3: Occupation	Score: 27.90398
Feature: 4: visited_doctor_last_1_year	Score: -33.54307
Feature: 5: cholesterol_level	Score: 47.03201
Feature: 6: daily_avg_steps	Score: -0.02100
Feature: 7: age	Score: 2.42726
Feature: 8: heart_decs_history	Score: 111.57213
Feature: 9: other_major_decs_history	Score: 67.00532
Feature: 10: Gender	Score: 30.51770
Feature: 11: avg_glucose_level	Score: 0.41409
Feature: 12: smoking_status	Score: -8.20285
Feature: 13: Location	Score: 5.30186
Feature: 14: weight	Score: 1332.14598
Feature: 15: covered_by_any_other_company	Score: 1198.98624
Feature: 16: Alcohol	Score: 14.66674
Feature: 17: exercise	Score: 13.43045
Feature: 18: weight_change_in_last_one_year	Score: 224.88731
Feature: 19: fat_percentage	Score: -1.50108
Feature: 20: years_since_last_admitted	Score: 235.59674

Graph showing coefficients of different variables



Inference - 1 unit change in weight is increasing cost by 1332 rs.
- 1 unit change in covered_by_another insurance company is increasing cost by 1198
- 1unit change in regular_checkup_last_year is decreasing cost by 463.
- 1 unit change in year_since_last_admitted is increasing cost by 235.5
- 1 unit change in weight_chang_last_year is increasing insurance cost by 224.8

6. Final interpretation and recommendation

- Random forest is the best performing model because-
    1) MAPE = 9.62, this is an absolute value and it is lower than 10 which means the model is performing well.
    2) Absolute_R2 = 0.96, which means model is 96% accurate
    3) RMSE = 2701.2 , this is a relative measure and is lowest for this model.

Recommendation -

-Going to recommend Random forest model because -

1) This model has the best performance based on MAPE,Absolute_R2 and RMSE.
2) Random forest predictions are more accurate because it averages the predictions of all decision trees.
3) Random forest is more robust.

- While using Random forest model following variables are more important in making predictions-

1) years_of_insurance_with_us - if this is more cost will be more
2) regular_checkup_last_year - If patient is getting regular checkup done then cost will be less
3) visited_doctor_last_1_year - More the visits cost will be less
4) cholesterol_level - More cholesterol cost will be more
5) daily_avg_steps - I daily average steps taken is more cost will be less
6) age - older customer more cost
7) avg_glucose_level - If average glucose level, cost will be higher
8)smoking_status - If a customer is a smoker, the cost will be more.
9) weight - If customer is heavy , cost will be more
10) covered_by_any_other_company - If customer has taken insurance from some Other company as well then cost will be more
11) weight_change_in_last_one_year - I customer has gained more weight in last 1 Year then cost will be more
12) fat_percentage - I fat percentage is more, cost will be high
13)years_since_last_admitted - if customer has recently admitted to hospital, then cost Will be less.