greatlearning
Learning for Life

CAPSTONE PROJECT

Health care Life Insurance prediction

BUSINESS REPORT

ROHIT NAGARAHALLI

PGP-DSBA

Contents

Introduction of the Business Problem 6
Problem Statement 6
Need of the study/project 6
Understanding Business/Social Opportunity 7
Data Report 7
Exploratory Data Analysis 9
Univariate Analysis 9
Bivariate Analysis 19
Multivariate Analysis 24
Unwanted Variables 26
Missing Value Treatment 27
Outlier Treatment 27
Business Insights from EDA 28
Clustering Analysis and Insights 28
Business Insights and Recommendations 32
Model Building 33
Ordinary Least Squared Model 33
Inferences from the basic OLS Model: 35
VIF Check 36
OLS Model-2 36
Inferences from OLS Model-2 39
Regularization 40
Ridge Model 40
Lasso Model 41
Decision Tree Regressor 42
Inferences 42
Hyper tuned Decision tree Regressor 43
Inferences 43
Random Forest Regressor (Ensemble learning) 43
Inferences 44
Random Forest Regressor Hyper-tuning parameters 44
Inferences 44
XG Boost (Boosting) 45

Inferences 45
XG Boost (Hyper-tuned parameters) 46
Inferences 46
Cat-Boost Regressor 46
Inferences 47
Cat Boost Regressor (Hyper tuned parameters) 47
Inferences 47
All Model Comparison 48
Actual vs Predicted graphs 48
Inferences 49
Feature Importance 49
Inferences 51
Model Insights 52
Business Recommendations 53

Table of Figures

Figure 1 Years of insurance with company 9
Figure 2 Regular health checkup last year 9
Figure 3 Adventure Sport 10
Figure 4 Occupation 10
Figure 5 Doctors visit last 1 year 11
Figure 6 Cholesterol- level 11
Figure 7 Average steps taken daily 12
Figure 8 Age 12
Figure 9 Heart and other disease history 13
Figure 10 Gender 13
Figure 11 Average glucose level 14
Figure 12 BMI 14
Figure 13 Smoking Status 15
Figure 14 Location 15
Figure 15 Distribution of Weight 16
Figure 16 Alternate Company Insurance 16
Figure 17 Alcohol Consumption 17
Figure 18 Exercise 17
Figure 19 Percentage of Fat 18
Figure 20 Insurance cost 18
Figure 21 Adventure Sports vs Insurance cost 19

Figure 22 Regular check-up vs Insurance cost 19
Figure 23 Location vs Insurance cost 20
Figure 24 Other company vs Insurance cost 20
Figure 25 Alcohol vs Insurance cost 21
Figure 26 Cholesterol level across Occupations 22
Figure 27 Cholesterol level vs Insurance cost 23
Figure 28 Weight variation vs Insurance cost 23
Figure 29 Pair plot 24
Figure 30 Independent variables correlations 25
Figure 31 Target Correlations 25
Figure 32 Weight vs Insurance cost 26
Figure 33 Before and After BMI Outlier treatment 27
Figure 34 Daily average steps 28
Figure 35 Elbow plot 28
Figure 36 Clusters vs Silhouette score 29
Figure 37 Weight cluster analysis 30
Figure 38 Insurance cost cluster analysis 30
Figure 39 Glucose level each cluster 31
Figure 40 age vs clusters 31
Figure 41 BMI cluster-wise 31
Figure 42 homoscedasticity and Normality Plots 34
Figure 43 QQ-Plot for normality 34
Figure 44 Homoscedasticity and Normality plot 37
Figure 45 QQ-Plot Model-2 37
Figure 46 OLS-Model 2 Actual vs Predicted Graph 38
Figure 47 OLS-Model 2 unusual predictions 39
Figure 48 Ridge Actual vs Predicted 40
Figure 49 Lasso predicted unusual values 40
Figure 50 Lasso Actual vs Predicted 41
Figure 51 Lasso unusual prediction 41
Figure 52 Actual vs Predicted graph for XGB Regressor 48
Figure 53 Actual vs Predicted graph for Cat-Boost Regressor 49
Figure 54 Feature Importance XGB-Regressor 49
Figure 55 Feature Importance Cat-Boost Regressor 50

Table of Tables

Table 1 Statistical Summary Continuous variables 8
Table 2 Clusters Statistical Analysis 30
Table 3 Basic OLS Model 33
Table 4 Basic OLS Model Evaluation 34
Table 5 Variance Inflation Factor 36
Table 6 OLS Model-2 36
Table 7 OLS Model-2 Evaluation 37
Table 8 Ridge Model Evaluation 40
Table 9 Lasso Model evaluation 41
Table 10 Decision tree regressor evaluation 42
Table 11 Hyper tuned DT Regressor evaluation 43
Table 12 Random Forest regressor model evaluation 44
Table 13 Random Forest (Hyper-tuned) model evaluation 44
Table 14 XG Boost Model evaluation 45
Table 15 XG-Boost (Hyper-tuned) Model evaluation 46
Table 16 Cat-Boost Regressor Model evaluation 47
Table 17 Cat Boost Regressor (Hyper-tuned) model evaluation 47
Table 18 Model Comparison chart 48
Table 19 Feature importance for XG-Boost Regressor 50
Table 20 Feature importance table for Cat-Boost Regressor 51

Introduction of the Business Problem

Problem Statement

To build a supervised, linear regression predictive model using data that provide the

optimum insurance cost for an individual

We will be using the health and habit related parameters for

the estimated cost of insurance.

By accurately estimating insurance costs, the model aims to provide fair premiums to customers while minimizing the risk for insurance companies.

Data File: Data.csv

Target: insurance_cost (Total insurance cost).

Need of the study/project

- Rising Healthcare Costs: Healthcare expenses are continually increasing, making it crucial to have insurance that provides adequate coverage without being prohibitively expensive.

- Risk Management for Insurance Companies: Insurance companies need to assess and manage risk effectively. By predicting the optimum insurance cost, they can set premiums that reflect the true risk posed by each individual, thereby reducing the likelihood of financial losses.

- Personalized Insurance Plans: With a better understanding of individual health risks, insurance companies can offer more personalized insurance plans that encourage healthy behaviours and provide appropriate coverage based on individual needs.

- Financial Security for Individuals: For individuals, having a fair and accurate insurance premium ensures that they are not overpaying for coverage while still being protected against high medical costs.

- Encouraging Healthy Lifestyles: By linking insurance costs to health and habits, the model can incentivize individuals to adopt healthier lifestyles, thereby reducing their insurance premiums and overall healthcare costs.

Understanding Business/Social Opportunity

- Market Differentiation for Insurance Companies: Developing an advanced model for predicting insurance costs can set a company apart from its competitors by offering more accurate and fair pricing.

- Enhanced Customer Satisfaction: Personalized and fair insurance premiums can lead to higher customer satisfaction and loyalty. Customers are more likely to stick with an insurance provider that offers transparent and reasonable pricing.

- Promoting Public Health: By incentivizing healthy habits through lower premiums, the insurance industry can play a significant role in promoting public health. Healthier individuals mean lower claims and costs for insurance companies in the long run.

- Data-Driven Decisions: The project leverages data analytics to make informed decisions, reflecting a modern, tech-savvy approach that can attract customers who value data-driven insights.

- Policy Development: Insights gained from the model can help in developing policies that are aligned with health trends and risks, enabling better strategic planning and resource allocation for insurance companies.

Data Report

The data consists of 25000 records and 24 variables.

The data was observed to have a good mixture of categorical and continuous types.

The raw data had 16 numerical data type combinations of int64 and float 64. Eight object data types were observed. However, categorical data was observed in the int64 and object data types. The object data types require to be encoded depending upon the ordinal or nominal.

The variable Application ID consisting of unique numbers given to each customer holding the insurance. No duplicates were observed. This will be termed as the irrelevant variable for analysis and modelling.

Irregularities in value counts: Presence of "Unknown" value in the smoking status variable accounting more than 30% of the total records. It suggests a large number of missing or unavailable smoking status information.

The variables BMI and Year last admitted have missing records. The missing values in Year last admitted has been assumed to be customer have never been admitted last time. Missing values in the BMI needs to be handled since the percentage of missing records in BMI is just 3.96.

Statistical Summary

	count	mean	std	min	25%	50%	75%	max
daily_avg_steps	25000	5215.9	1053.2	2034	4543	5089	5730	11255
Age	25000	44.918	16.107	16	31	45	59	74
avg_glucose_level	25000	167.53	62.73	57	113	168	222	277
Bmi	24010	31.393	7.8765	12.3	26.1	30.5	35.6	100.6
Weight	25000	71.61	9.3252	52	64	72	78	96
fat_percentage	25000	28.812	8.6324	11	21	31	36	42
insurance_cost	25000	27147	14324	2468	16042	27148	37020	67870

Table 1 Statistical Summary Continuous variables

Analysis

- BMI variable has only 24010 count indicating missing records.
- By observing the standard deviation and mean it can be noted that daily average steps variable has outliers.
- Similarly, the variable BMI has outliers in the upper range
- The average of insurance cost (target) is 27147. Minimum insurance cost being 2468 and maximum being 67870.

Exploratory Data Analysis

Univariate Analysis

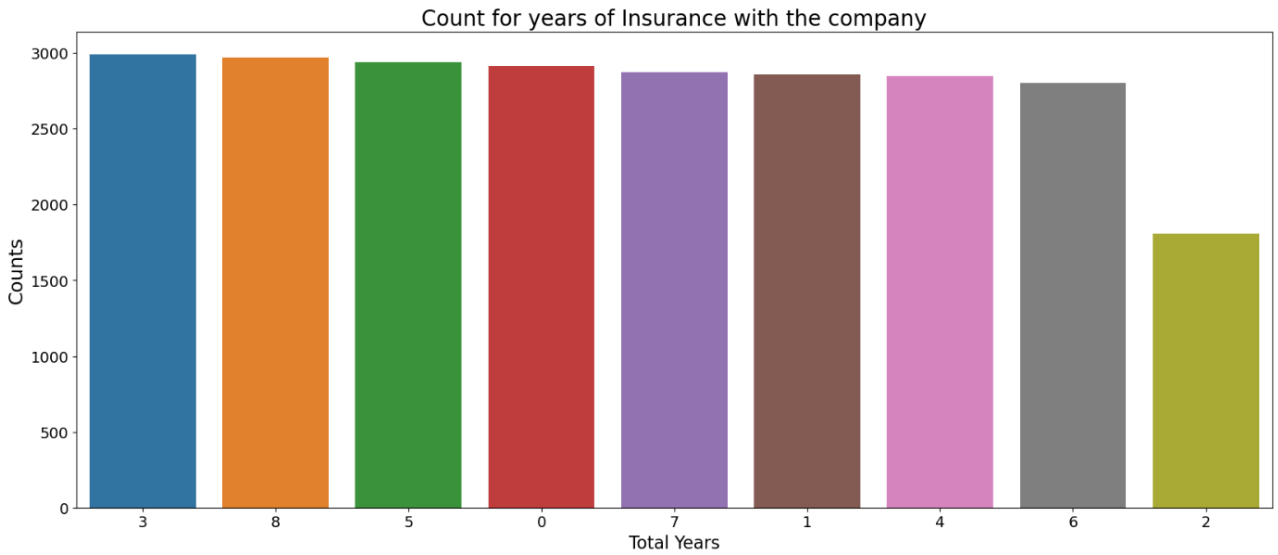


Figure 1 Years of insurance with company

Observation: Majority of the customers in the data have been taking policy from the same company for 3 years followed by 8 and 5 years. Zero might be indicating the customers who had recently purchased the insurance from the company.

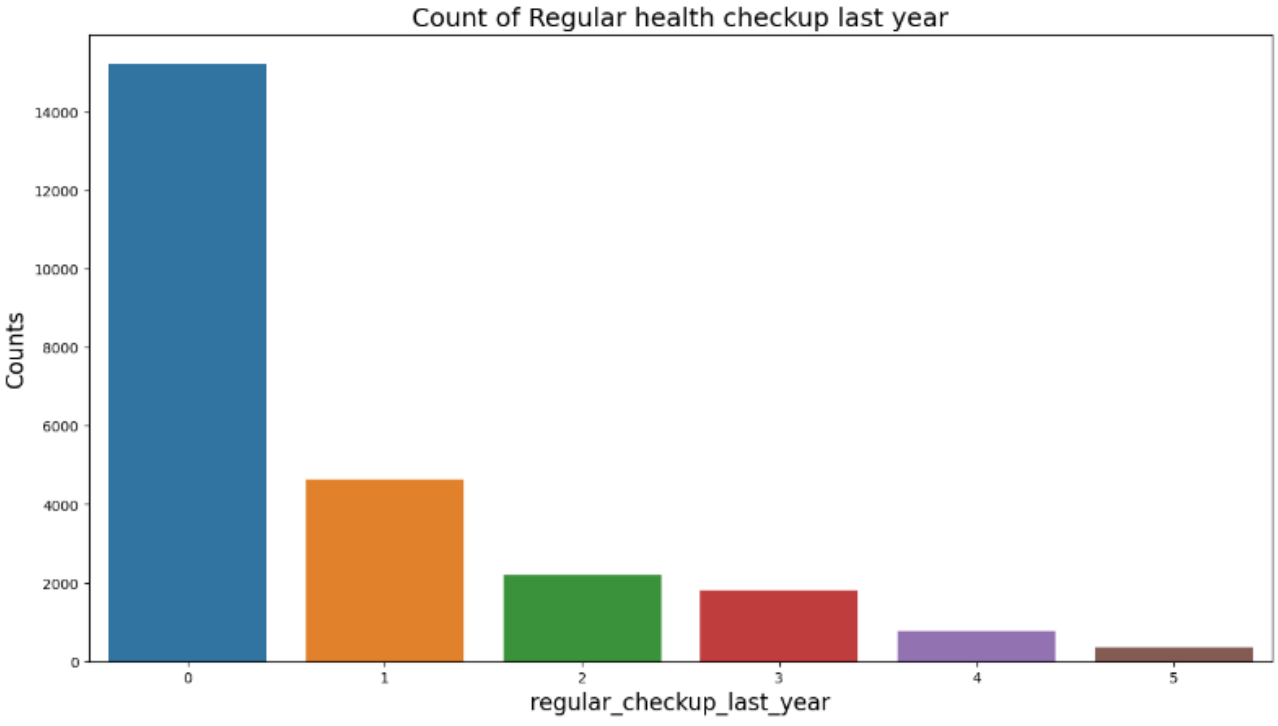


Figure 2 Regular health checkup last year

Observation: Approximately 60% of the current customers have not had health check-up last year. The company can lay down policies in such a way that the customers having regular health checkups could attract lesser insurance prices or depreciation in the current cost.

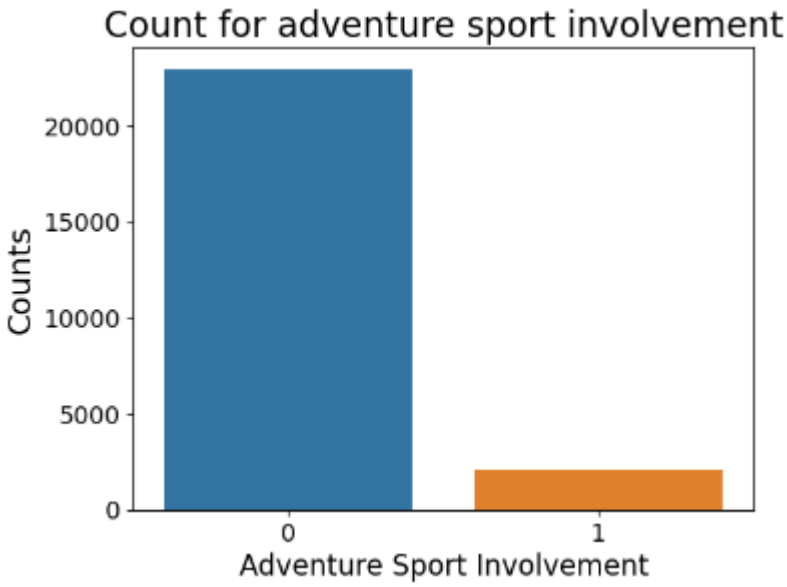


Figure 3 Adventure Sport

Observation: A smaller number of customers have been involved in Adventure sport activities. Higher age categories not involving in an adventure sport is acceptable but it is advisable for the lower age category to involve to stay active.

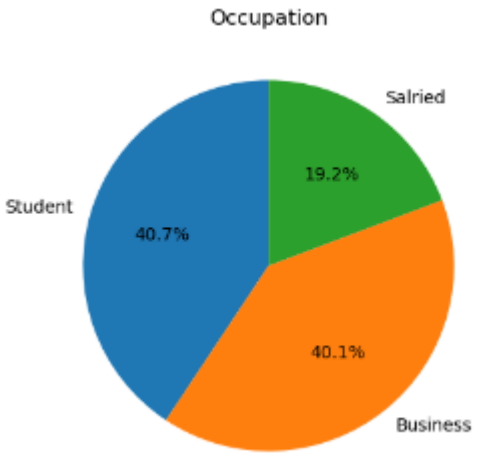


Figure 4 Occupation

Observation: Business and Students have almost equal ratio in choosing this insurance. However, Salaried class have the least accounting to only 19% of the total occupation. More promotions can be initiated by marketing or any other strategies to attract more Salaried class.

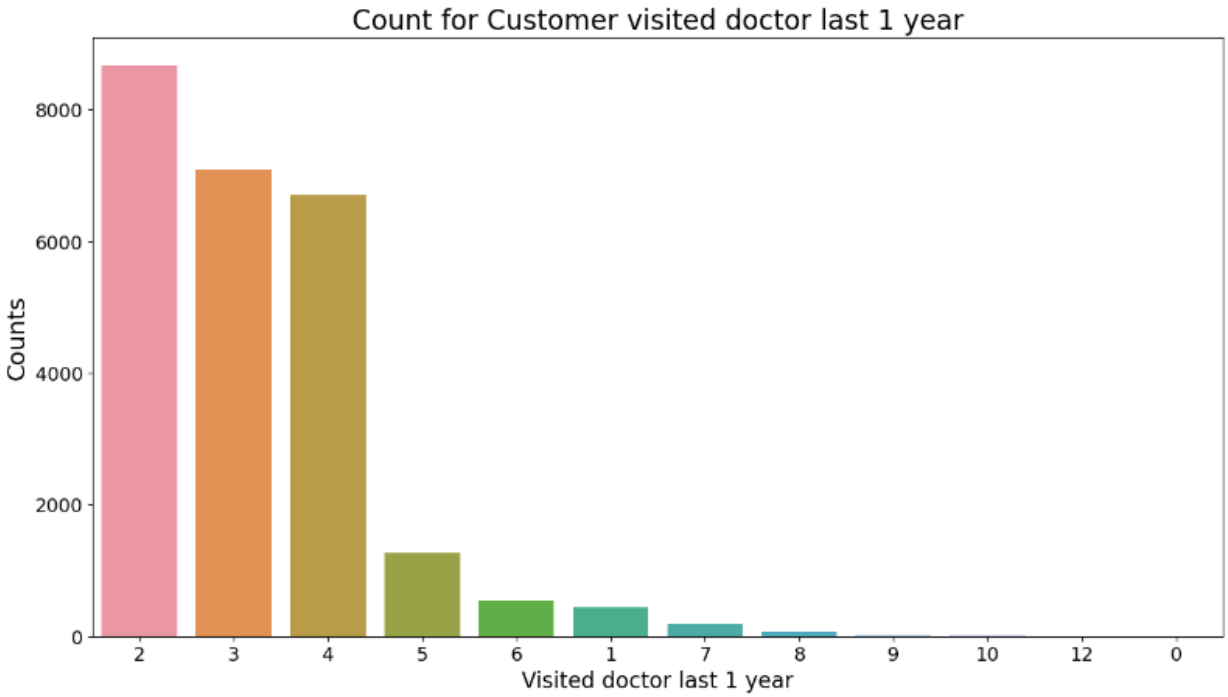


Figure 5 Doctors visit last 1 year

Observation: Majority of the customers visited twice, thrice or four times in the last one year. An insurance holder visiting doctor more than 4 times in last 1 year could indicate he/she suffering from an illness and most like to visit doctor frequently and claim insurance amount too. Single customer was observed to have not visited doctor once and another to have visited 12 times

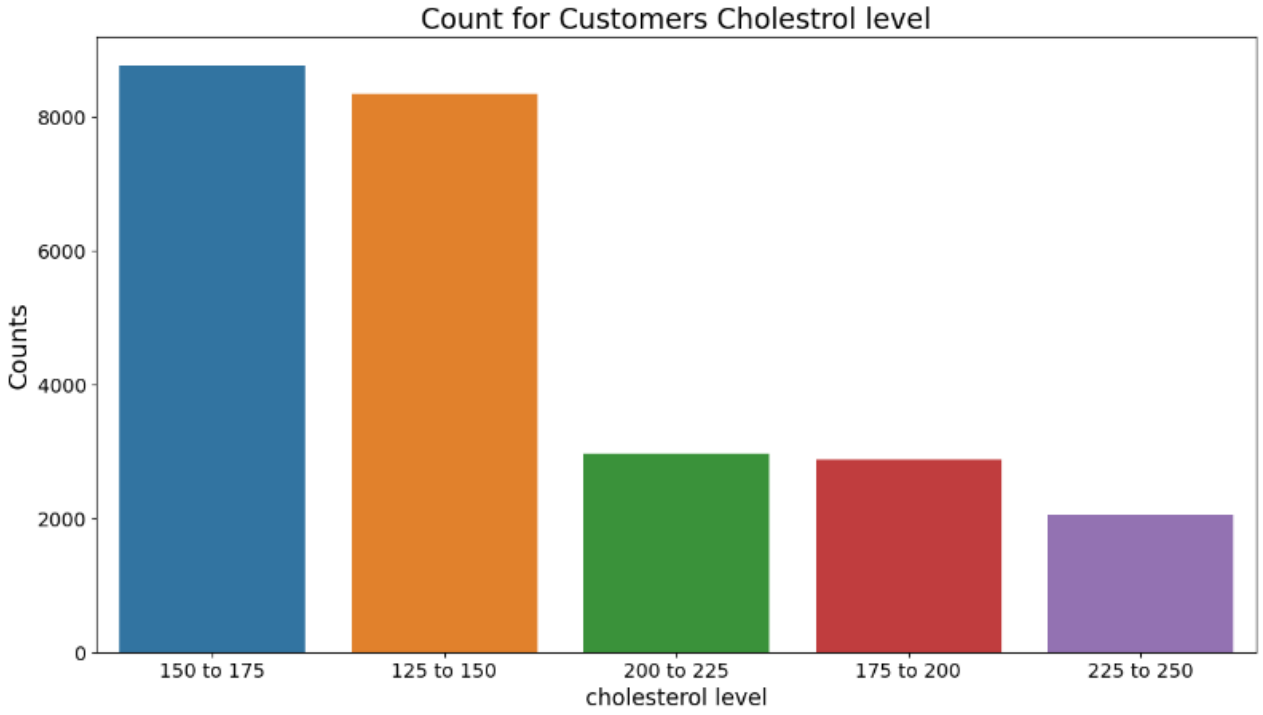


Figure 6 Cholesterol- level

Majority of the customers were in the lower ranges of cholesterol level.

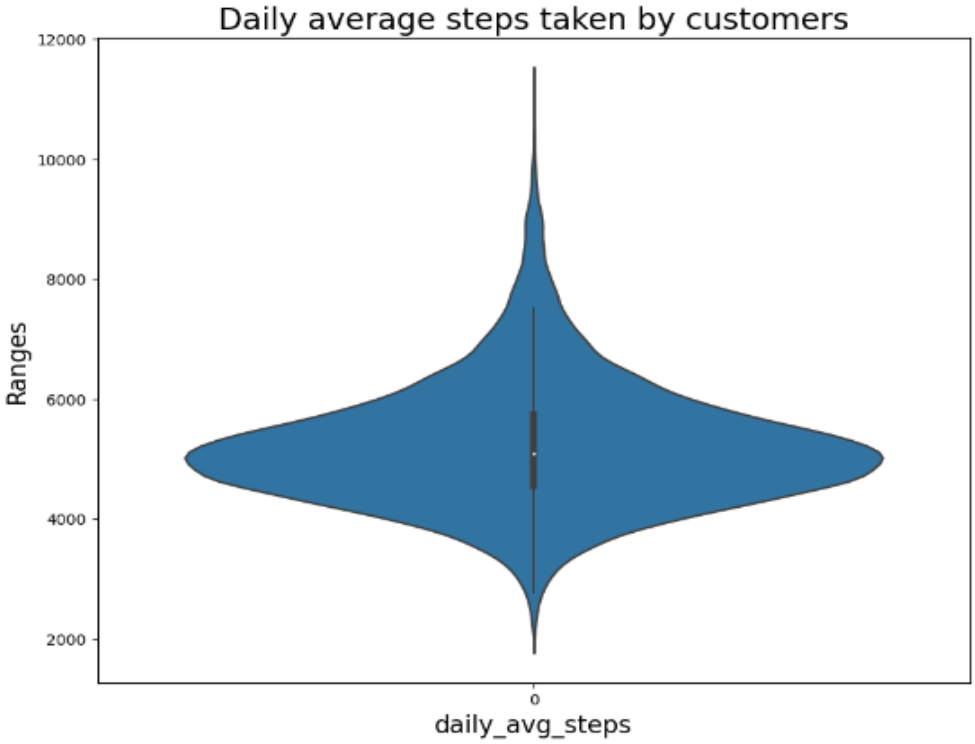


Figure 7 Average steps taken daily

Observation: The data appears to be skewed towards right indicating few customers had taken higher average steps daily. The company can look into lesser and higher average daily steps and its contribution to a person being healthy.

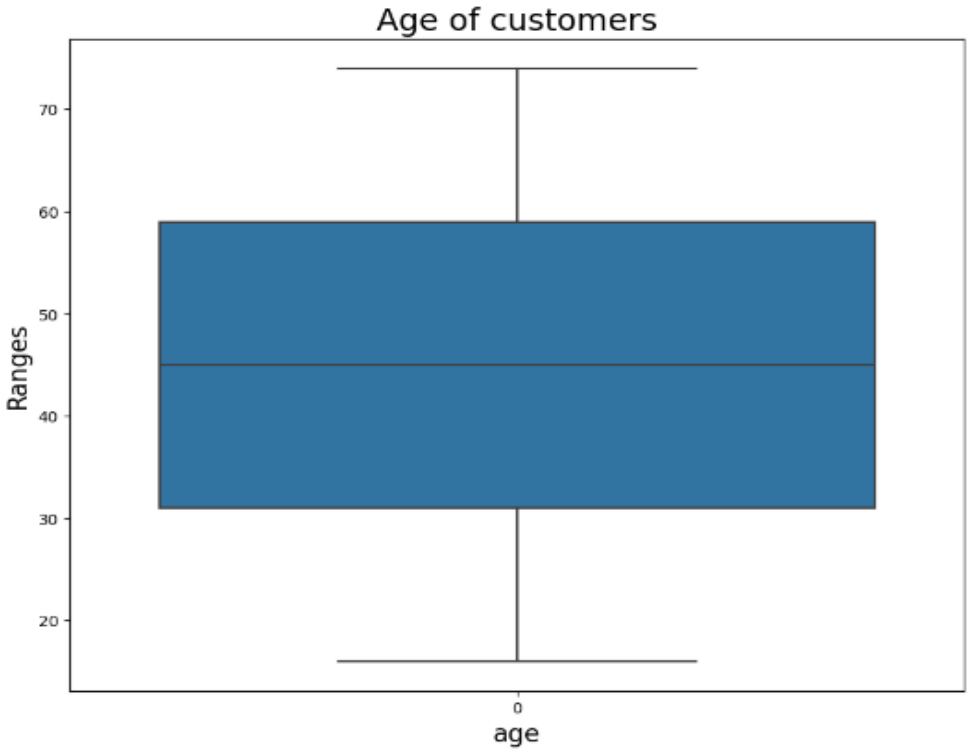


Figure 8 Age

Observation: Customers of all eligible age have purchased the insurance. No skewness was observed for this variable and no irregularities was observed. The company has targeted all the age groups in the right proportions.

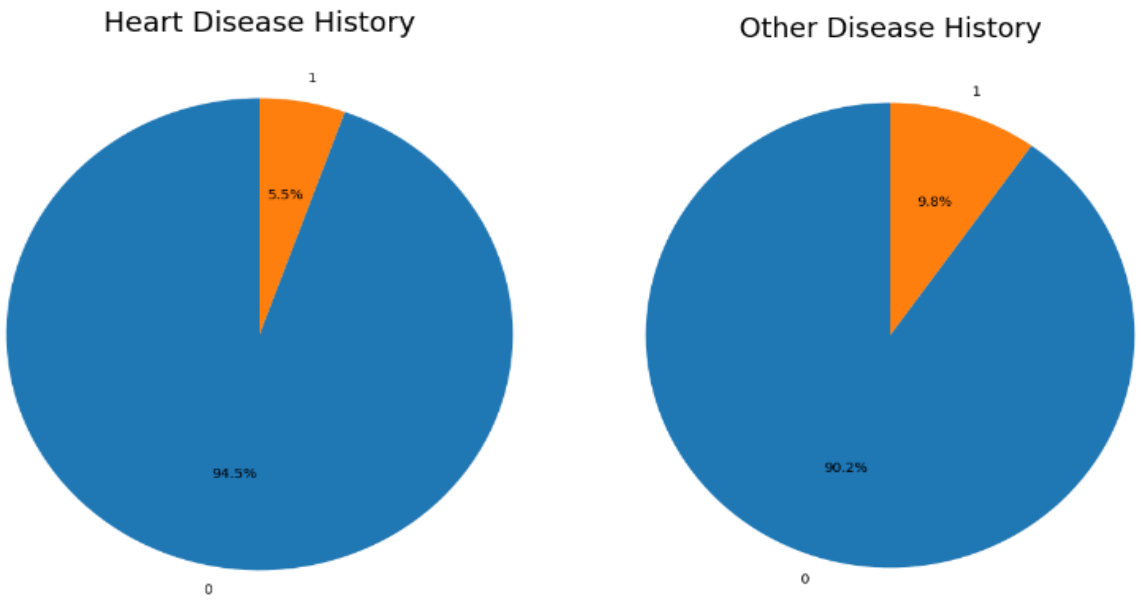


Figure 9 Heart and other disease history

Observations: Only 5.5% of the customers had heart disease history and approximately 10% of the customers had other disease history. Customers with disease history should be charged slightly higher insurance costs than the normal. This is because the one with history of disease could potentially attract more disease or get hospitalized frequently. There is an imbalance observed, the data needs to be split proportionally for model building for such cases.

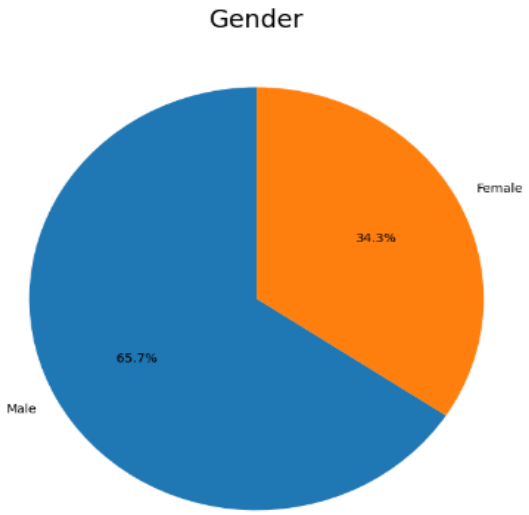


Figure 10 Gender

Observation: 34% of the customers are female. Although this doesn't seem under sampled yet measure needs to be taken to attract female customers to buy the insurance.

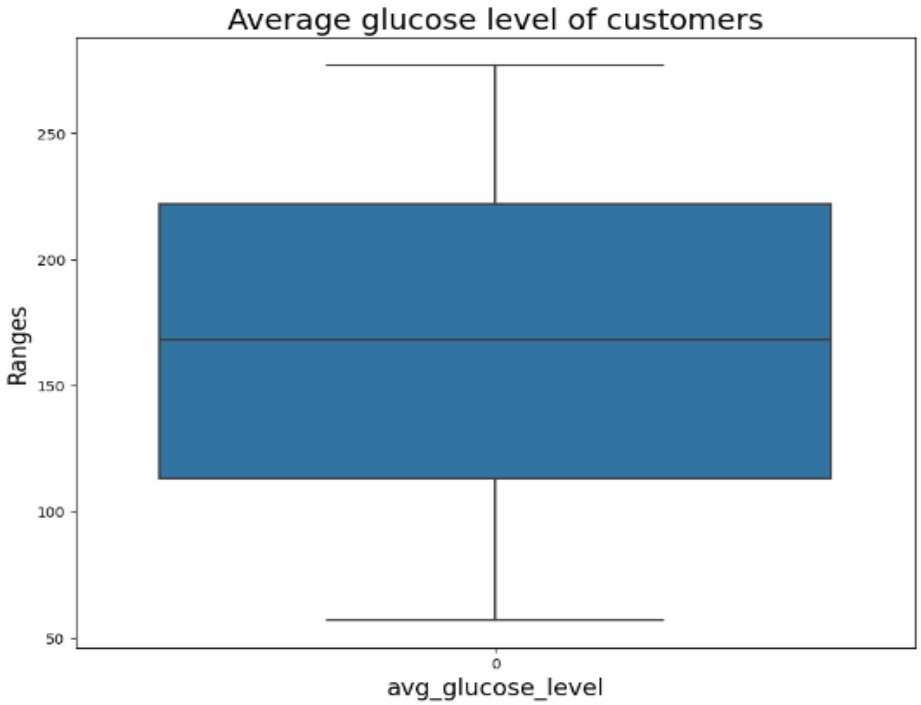


Figure 11 Average glucose level

Observations: The data is evenly distributed with no outliers or skewness. However, the customers have had normal and abnormal glucose levels as per the data.

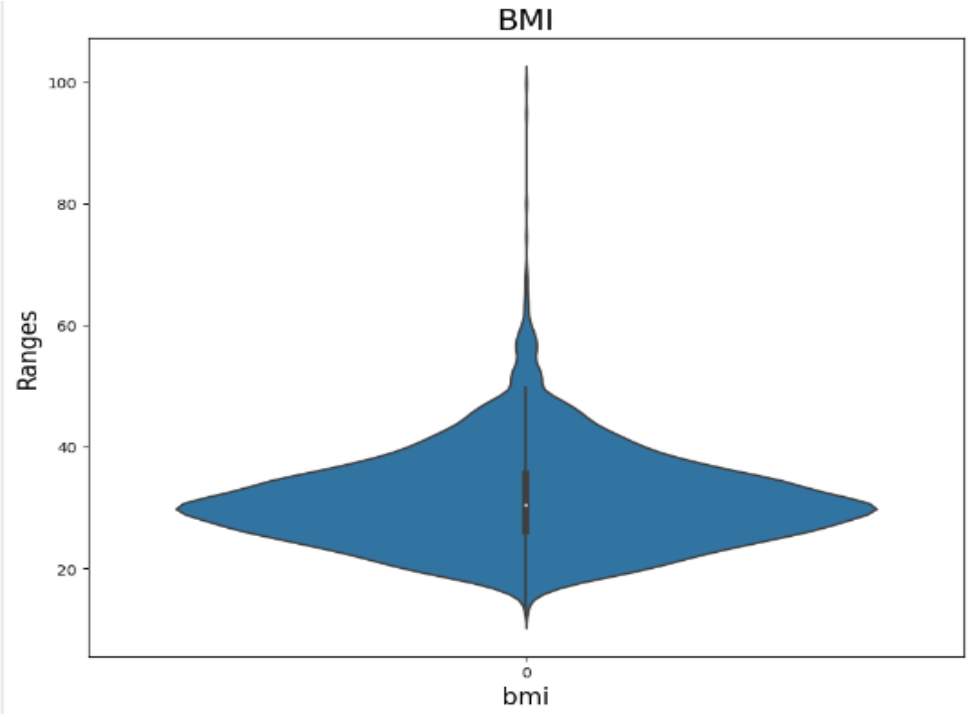


Figure 12 BMI

Observation: The variable BMI has seen unusually higher ranges which seems almost impossible ranges which could be the result of data being skewed towards right and possess outliers.

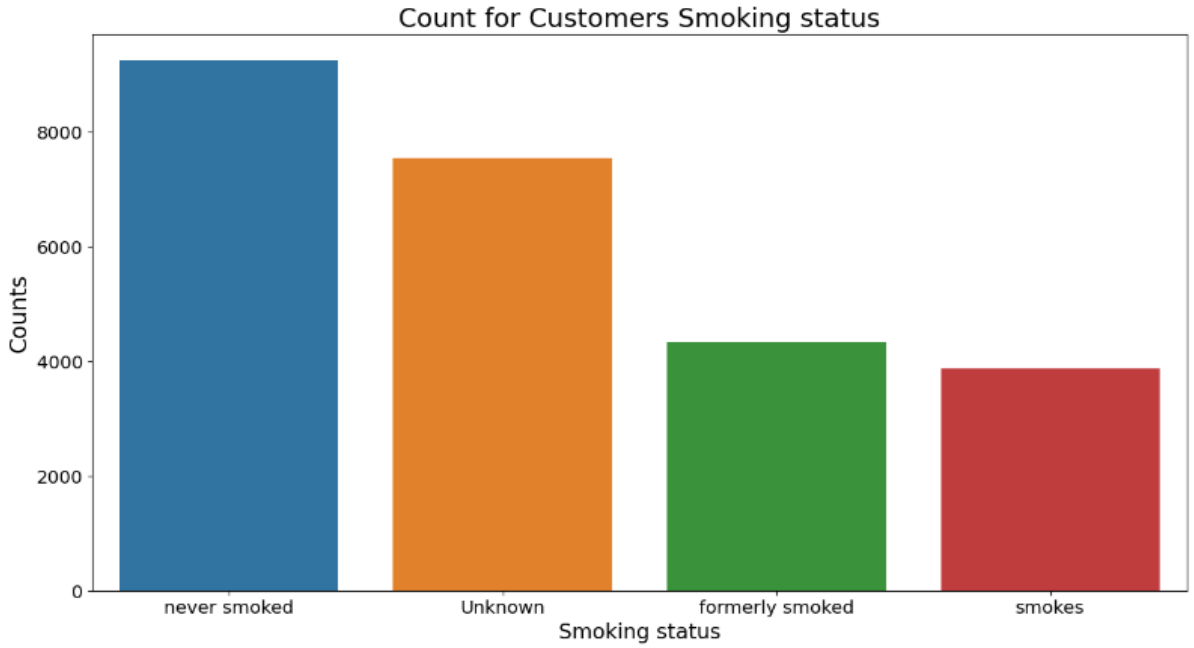


Figure 13 Smoking Status

Observation: Presence of Unknown category indicating missing records. The variable will be dropped if found to have more than or equal 30% of such Unknown category.

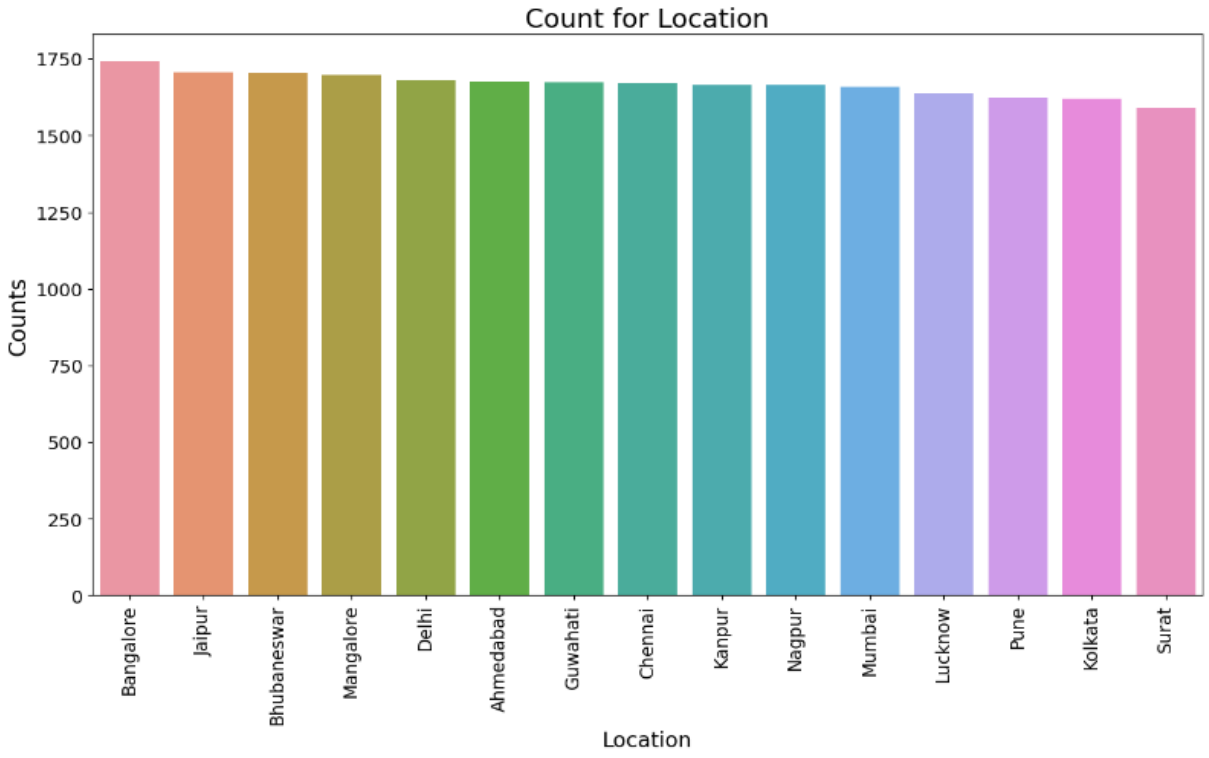


Figure 14 Location

Observation: The company has targeted the locations equally as an outcome the difference in the count of locations being low. The highest is the Bangalore and the least being Surat.

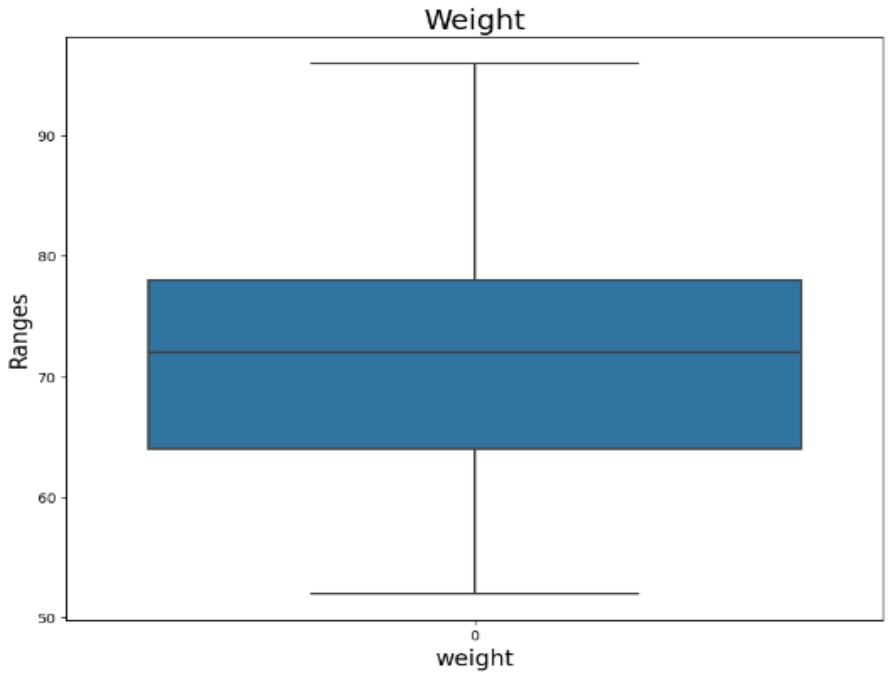


Figure 15 Distribution of Weight

Observations: The weight of the customer has been distributed well with less or no skewness observed.

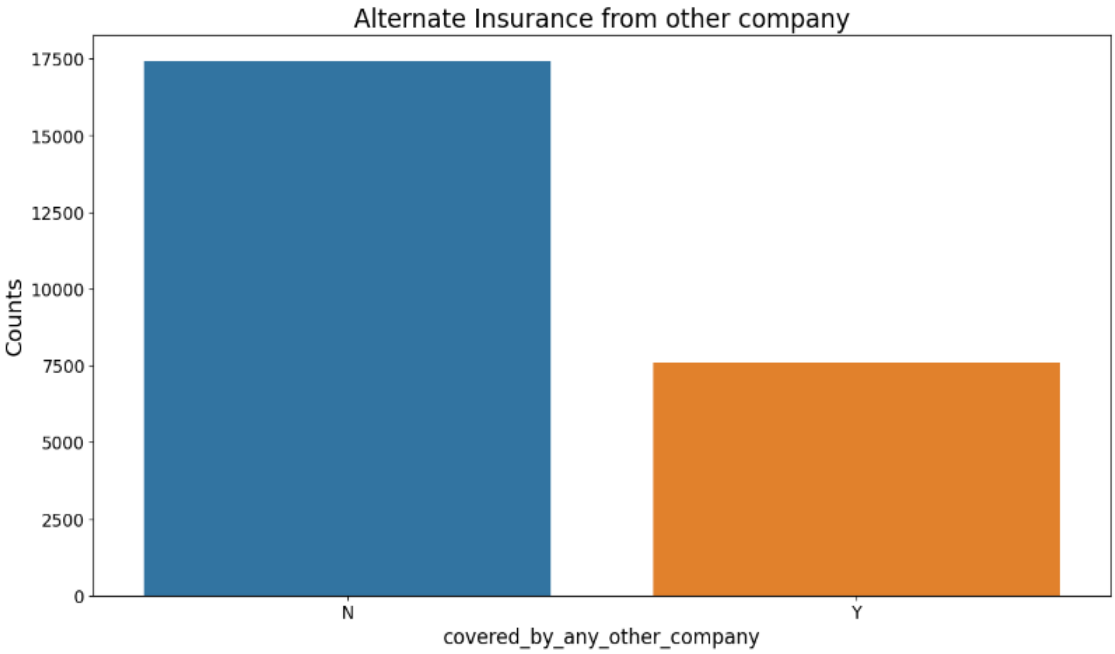


Figure 16 Alternate Company Insurance

Observation: Majority of the customers rely on this company. However, there are customers who have also opted for other company insurance. There is a chance of such customers churning out of this company.

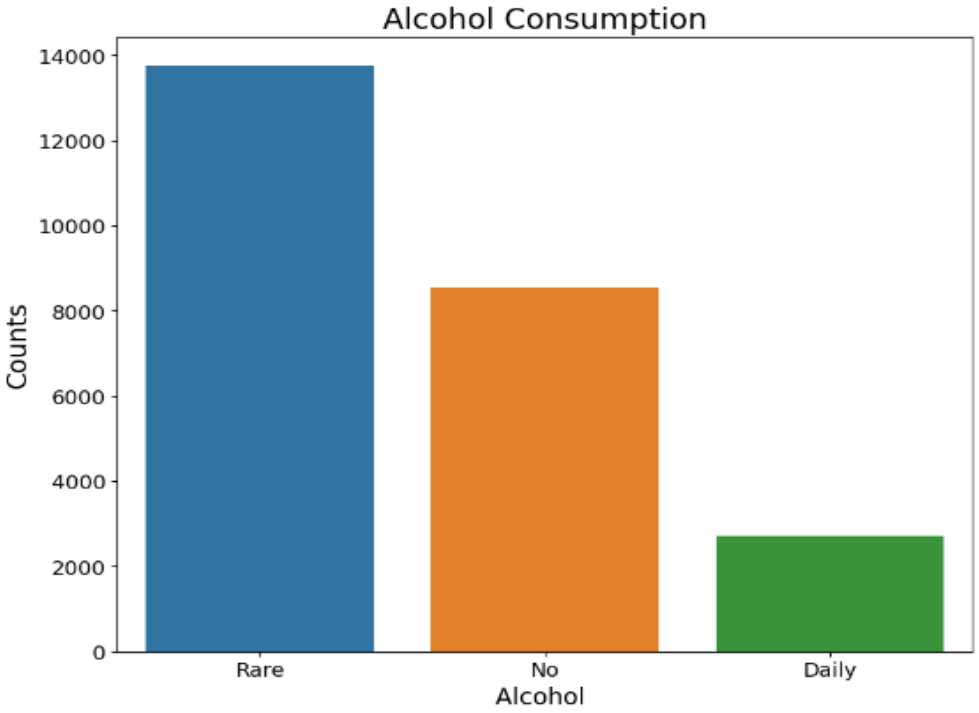


Figure 17 Alcohol Consumption

Observations: Majority of the customers consume alcohol but on rare occasions. However, there are customers who consume it daily. Such customers are prone to have multiple organ disease and are at a higher risk of being hospitalized and claim insurances.

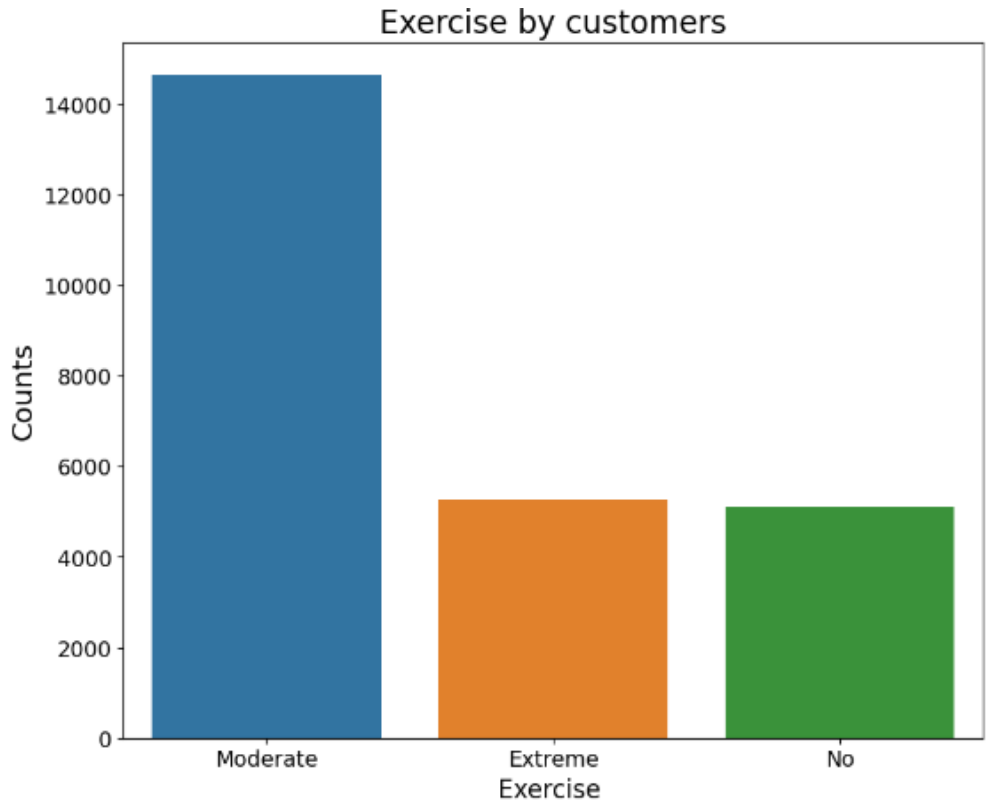


Figure 18 Exercise

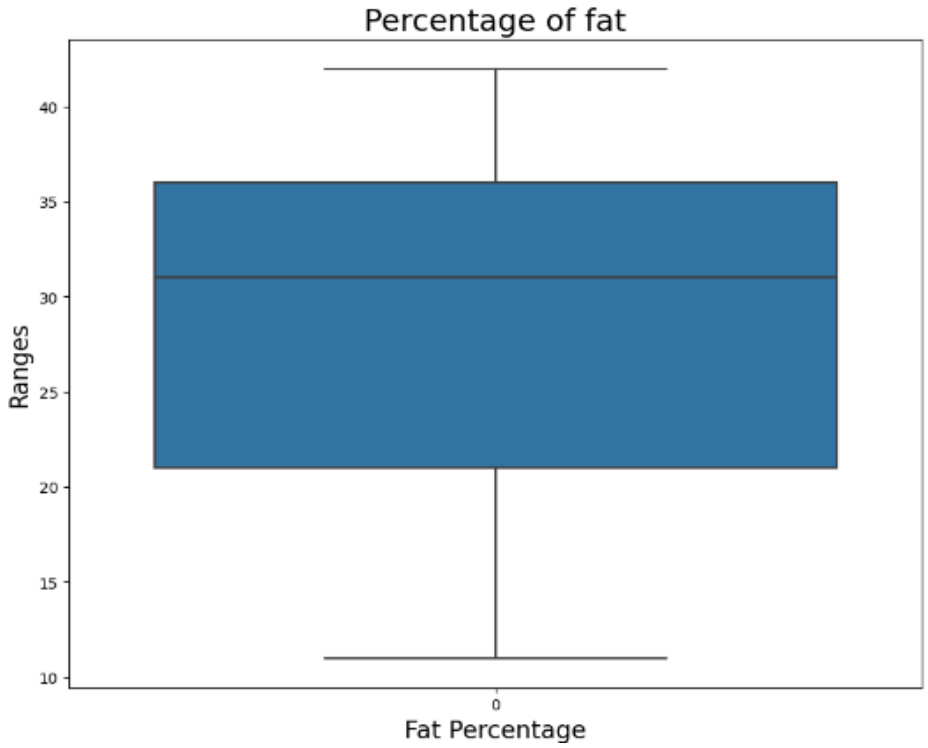


Figure 19 Percentage of Fat

Observation: No outliers observed. Data appears to have skewed on the left at the lower ranges.

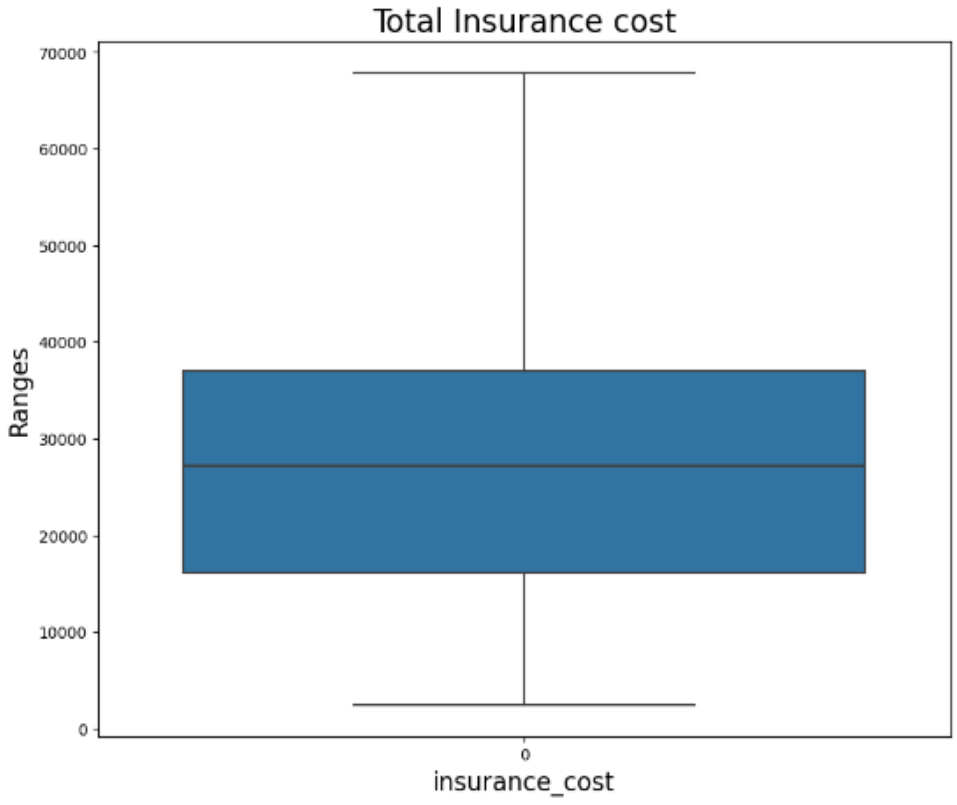


Figure 20 Insurance cost

Observation: The target variable, consists of no outliers but skewed towards its right at the higher insurance cost. Need to observe the factors that are influencing such higher insurance costs.

Bivariate Analysis

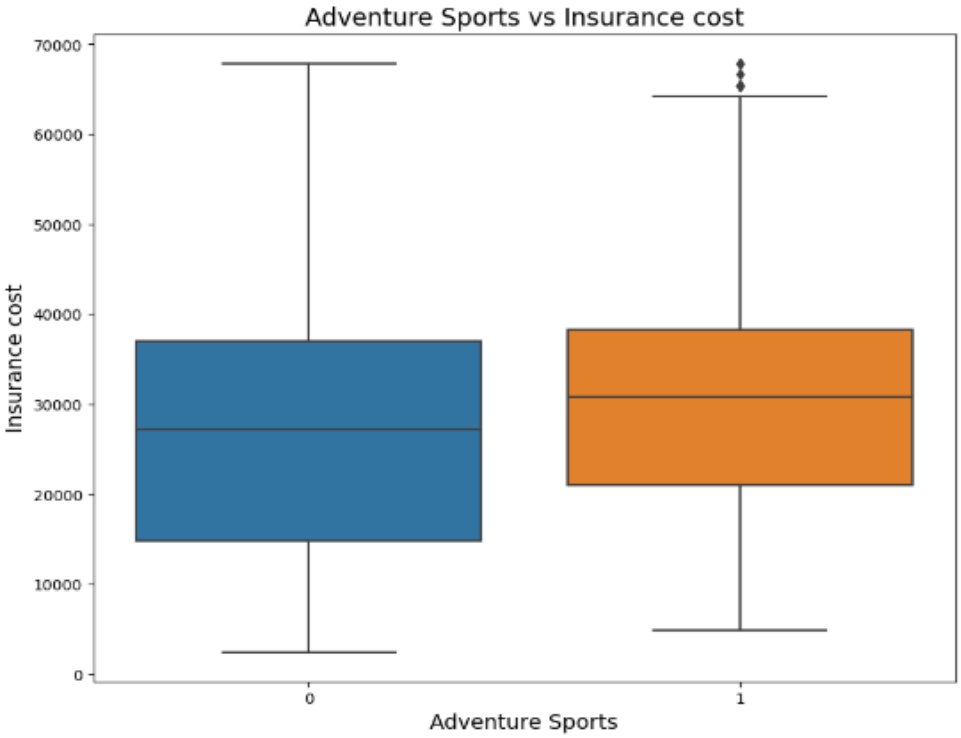


Figure 21 Adventure Sports vs Insurance cost

Observation: It is surprising that the customers involved in adventure sports have attracted higher insurance cost compare to that of customers who did not. The median is higher for the customers indulging in adventure sports.

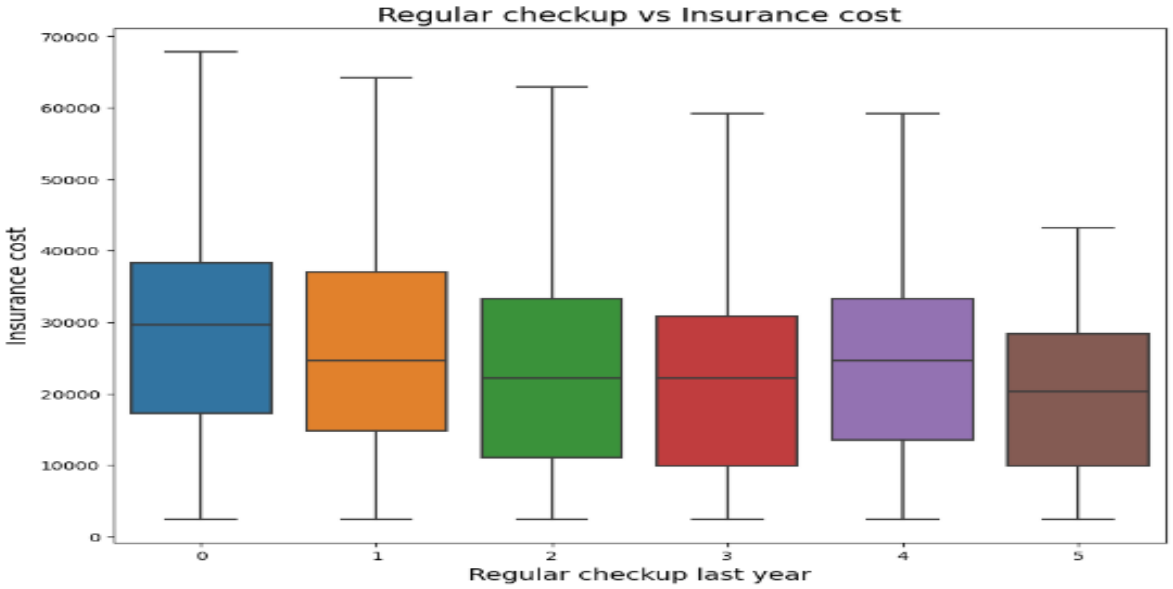


Figure 22 Regular check-up vs Insurance cost

Observation: The plot is straight forward; it indicates more the regular check-up last year lesser the cost of insurance. However, regular check up 4 seems to have higher costs compared to 2 and 3. Needs to verify what went wrong in such scenarios.

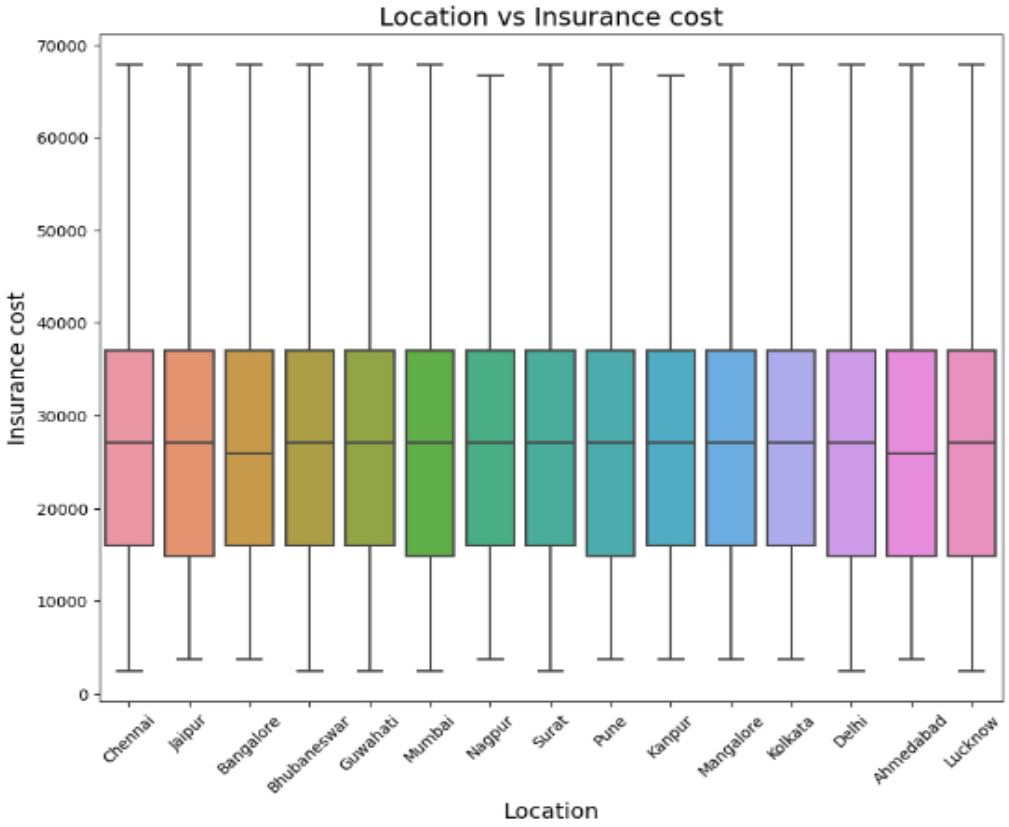


Figure 23 Location vs Insurance cost

The location has almost no impact on the insurance price. The median, min and max all fall under almost same category.

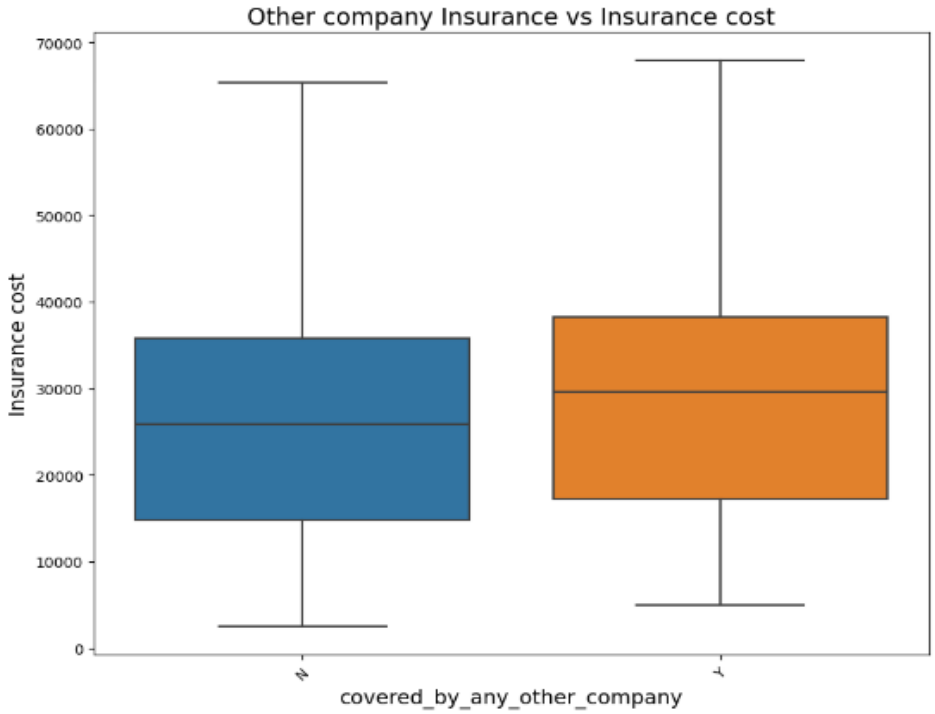


Figure 24 Other company vs Insurance cost

Observation: Customers who had other company insurances attracted higher insurance costs.

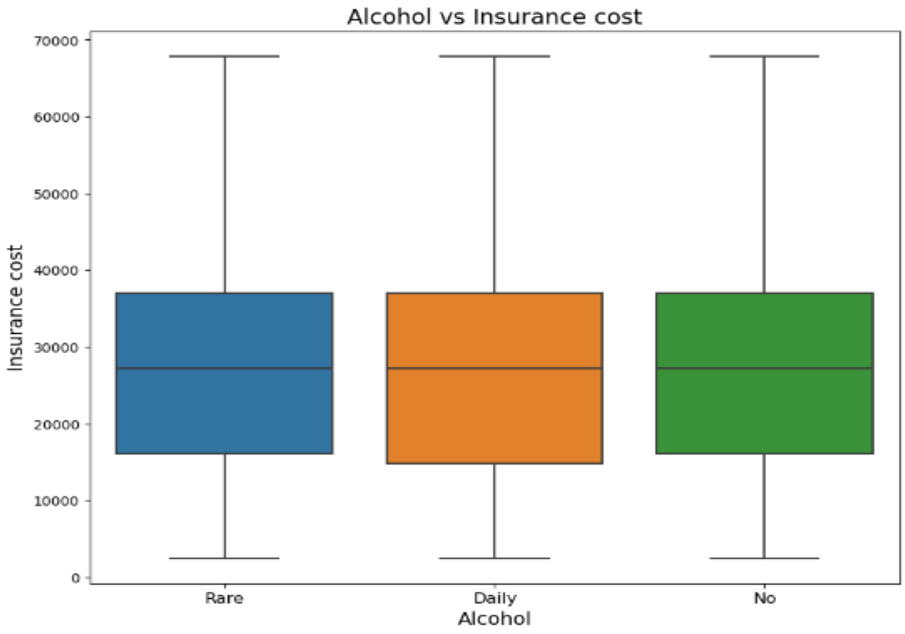
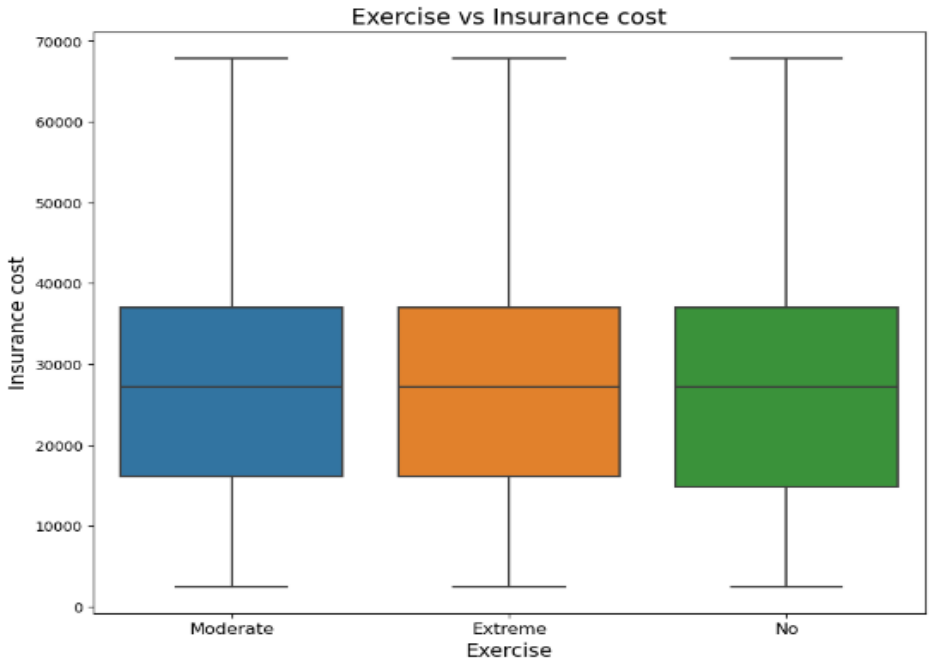


Figure 25 Alcohol vs Insurance cost

Observations: The results are surprising as the customers who consume alcohol daily attract the same insurance prices as the one who don't consume alcohol. The three categories have almost same median, min and max insurance cost.



Observation: Exercise categories seem to have no impact on the price of insurance. The scenario is such that the one who exercises at least moderately are less likely to attract disease and are more likely to stay fit and less hospitalized. But there has been no variation in the insurance among the categories of Exercise.

Occupations and Insights with respect to their cholesterol levels

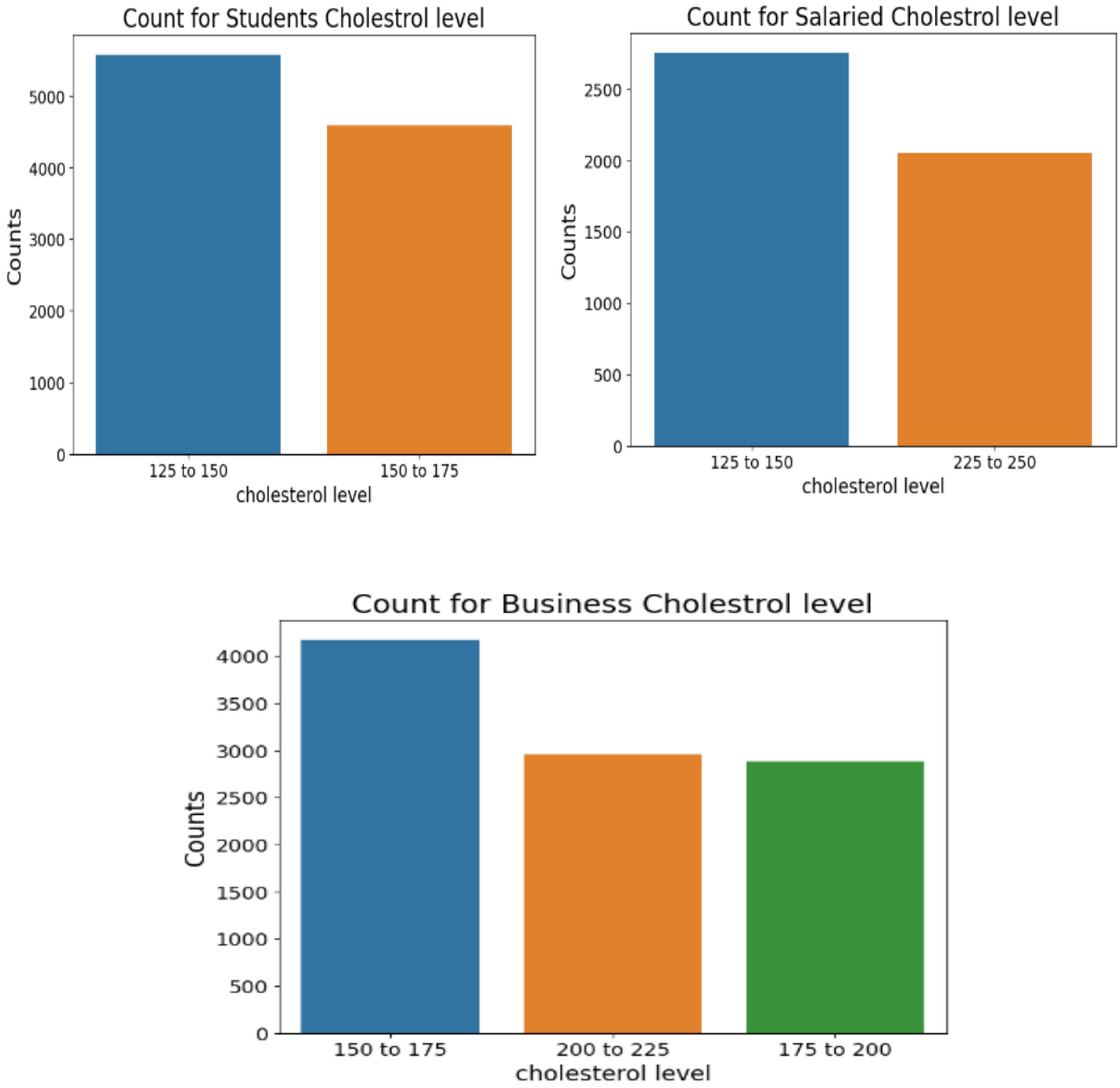


Figure 26 Cholesterol level across Occupations

Observation:

The cholesterol level for students is in the lower range category which is125 – 175, for salaried it is either lower or higher as the ranges says 125-150 and 225-250. For Business category the range is between 150 to 225.

Students did not have higher cholesterol ranges, Salaried either had higher cholesterol or lower ranges whereas, business neither had too high nor too low ranges but were in the mid-range.

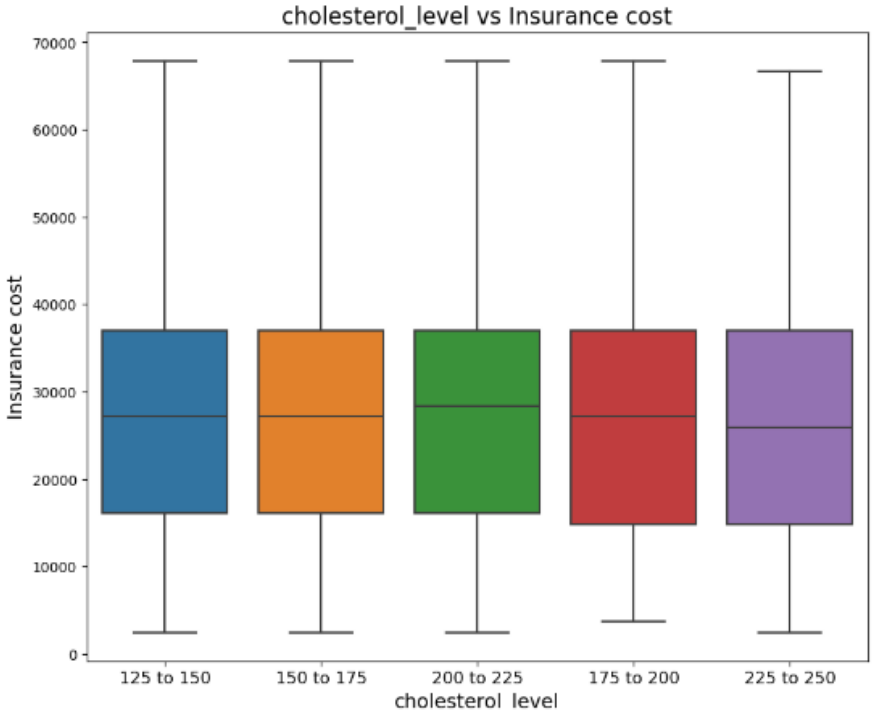


Figure 27 Cholesterol level vs Insurance cost

Observation: Despite students having low cholesterol ranges were not spared from higher insurance cost since the insurance cost for all cholesterol level is distributed evenly as the min, max and median is almost the same across the categories.

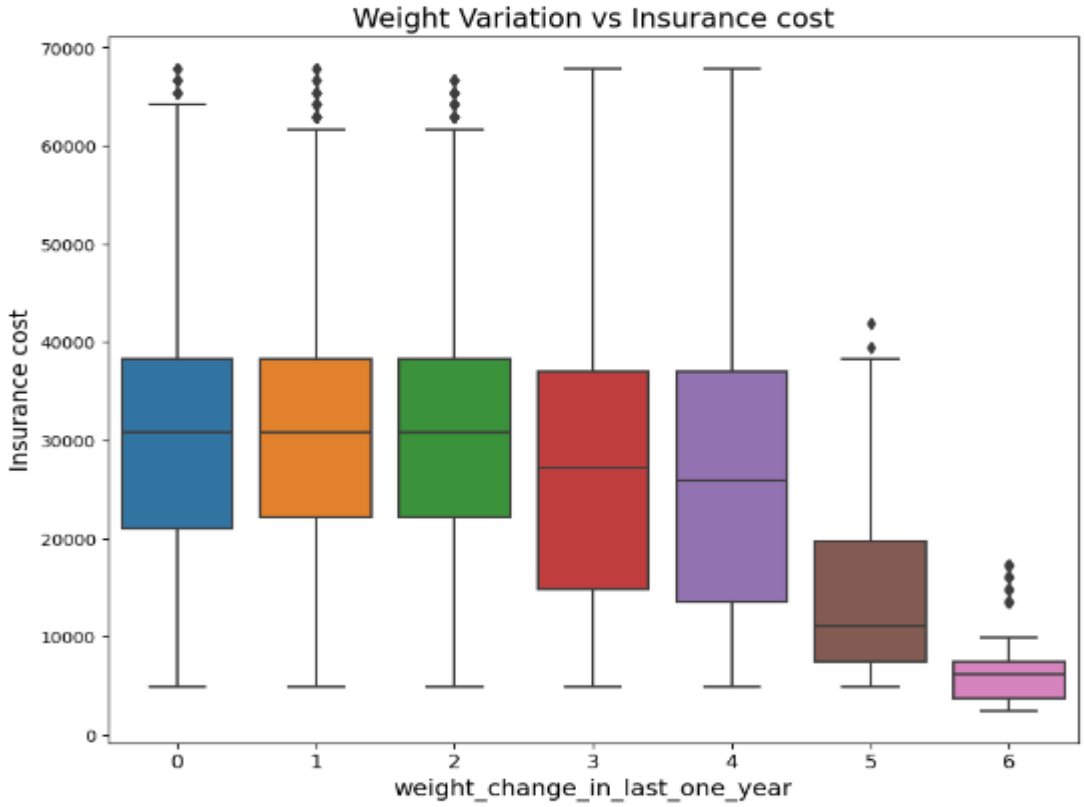


Figure 28 Weight variation vs Insurance cost

Observation: There are some key points in the plot above which indicates that higher the variation in weight lesser the insurance cost. This further indicates that compared to other variables weights and variations associated to that have contributed in depicting the price of Insurance.

Multivariate Analysis

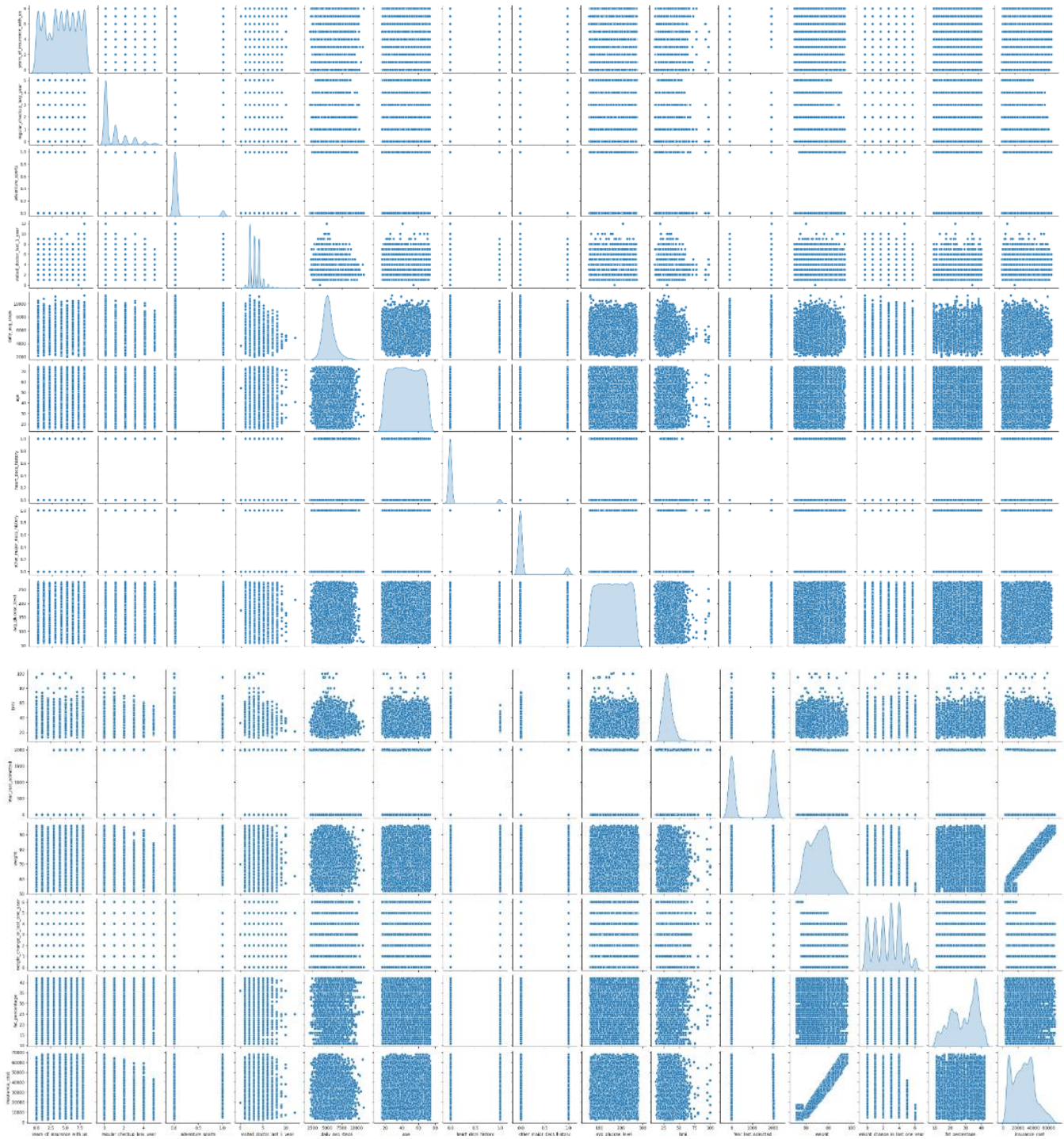


Figure 29 Pair plot

Correlations between the continuous independednt Variables

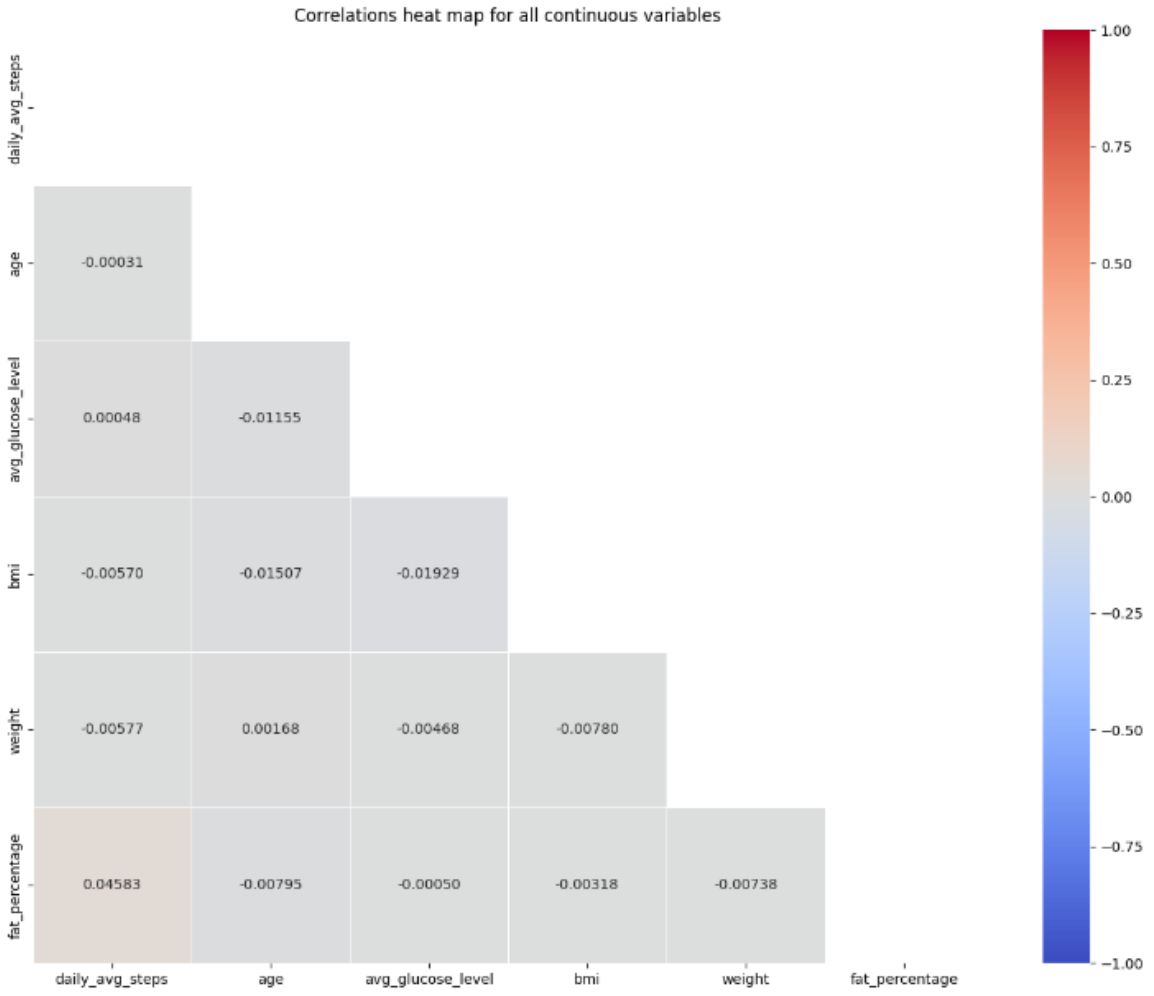


Figure 30 Independent variables correlations

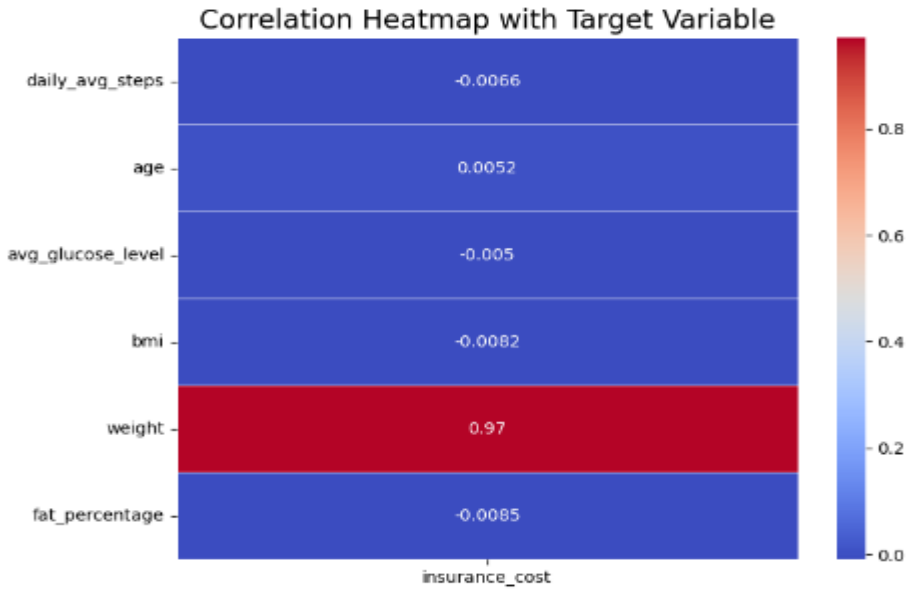


Figure 31 Target Correlations

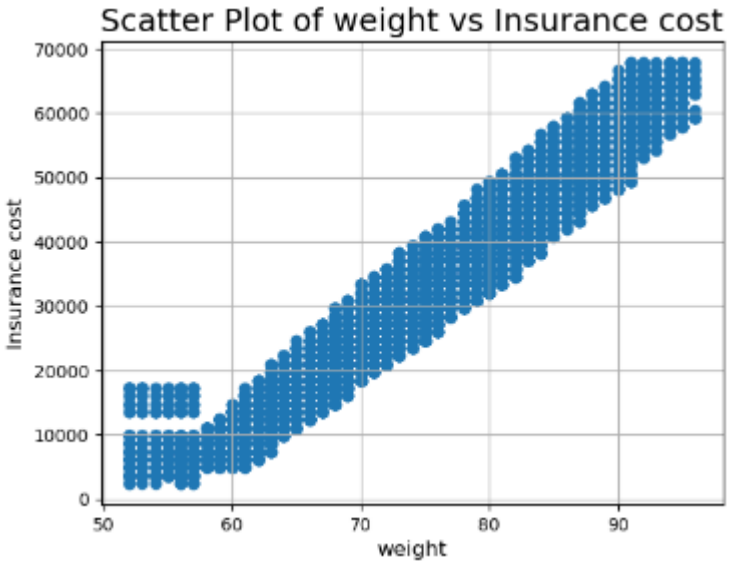


Figure 32 Weight vs Insurance cost

Observations

From the pair plot it was observed that weight exhibited a strong correlation with the target (insurance cost). Indicating when the weight is high the insurance cost is high. Although, correlations do not mean causation but something that has happened in the past. The heat map of target vs independent too depicted that the correlation of weight with respect to target is 0.97 which is very close to 1.

Since we are handling continuous data, the aim is that we will use Linear regression which assumes data to have no or less multicollinearity. From the figure 29 it can be observed that all the continuous independent variables have almost no multicollinearity which is a positive sign.

The pair plot also indicates that most of the continuous variables are almost normally distributed. If at all there exists skewness they do in a small scale and can be negligible of not there is always an option to transform the data.

Unwanted Variables

Post analysing the data it was noted that the variable application Id was of no use and was dropped for further analysis and model building. It was observed, presence of "Unknown" value in the smoking status variable accounting more than 30% of the total records. Since the standard threshold of dropping variables that had more than 30% missing or unknown records was prefixed considering the shape of data which is fairly less compared to the standard data shape.

The variables Application ID and smoking status were excluded from the data for any further analysis and model building.

Missing Value Treatment

Variables identified

1. BMI
2. Year last admitted

It was assumed that the missing values in Year last admitted indicates that the patient was not admitted which is Zero. The missing values were treated as zero indicating the customer was not admitted to hospital. The rest of the values which displayed years were treated as 1 leading the variable to be a binary representing whether the patient was admitted or not (Yes or No).

The missing values in the BMI were about 4% of the entire data and it was decided to be replaced with the values identified by KNN imputer. The data was scaled and the ideal neighbours were assigned as 5.

Post imputing the missing values it is important to check if the imputation has altered the distribution or the statistical summary of data. A boxplot for BMI pre imputing and post imputing was plotted and displayed no such irregularities or constraints.

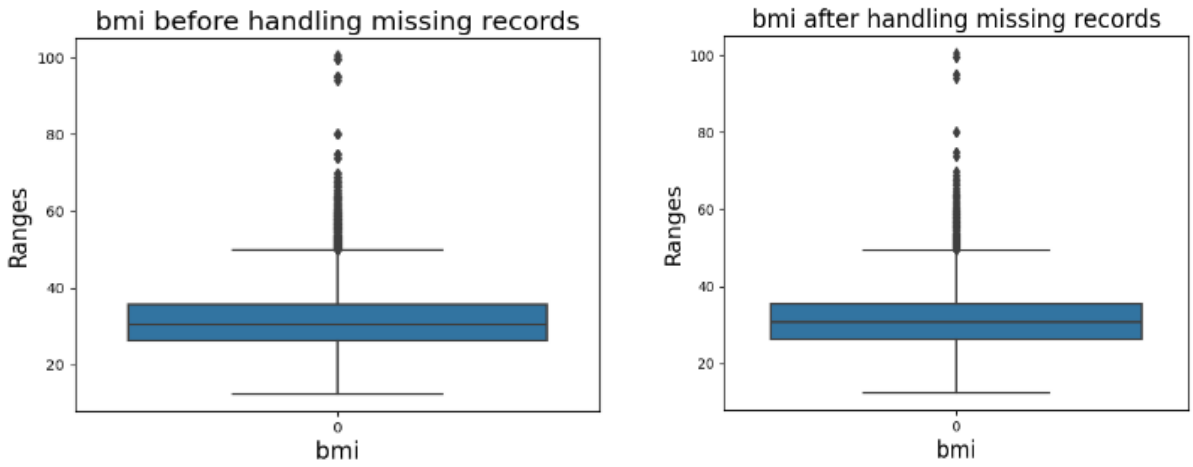


Figure 33 Before and After BMI Outlier treatment

Outlier Treatment

Variables Observed

1. BMI
2. Daily Average steps

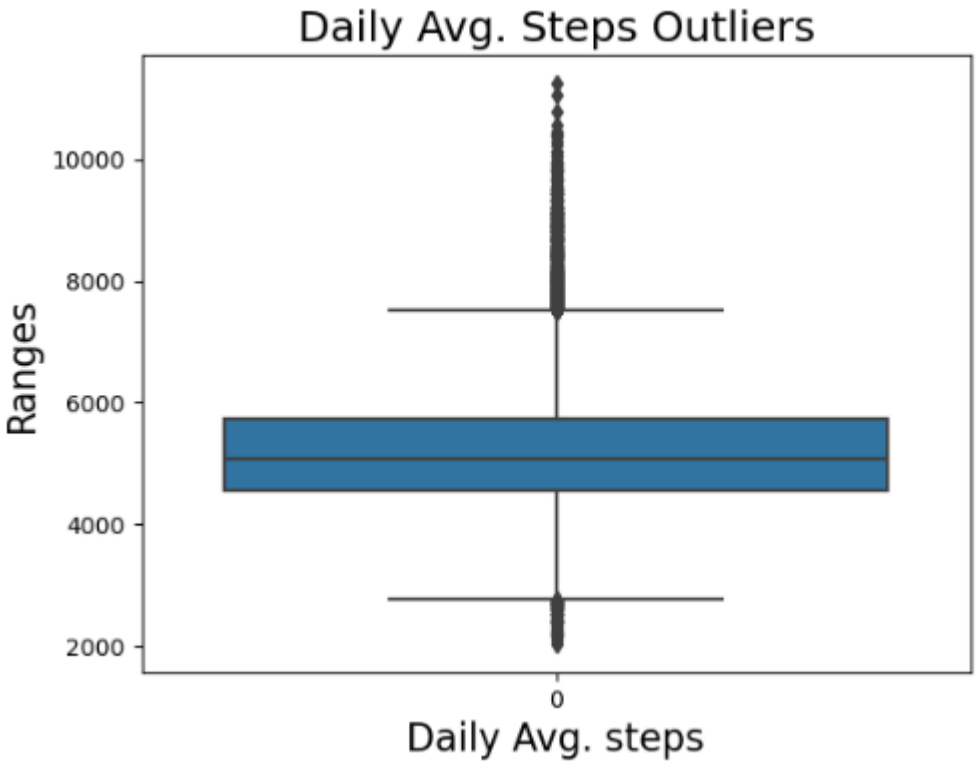


Figure 34 Daily average steps

To identify the outliers 85% and 15% were considered since dragging more values to outliers could potentially harm the legitimate values. These values were converted to nan and imputed using KNN imputer. The distribution of the data post imputing was checked and there were no irregularities observed.

Business Insights from EDA

Clustering Analysis and Insights

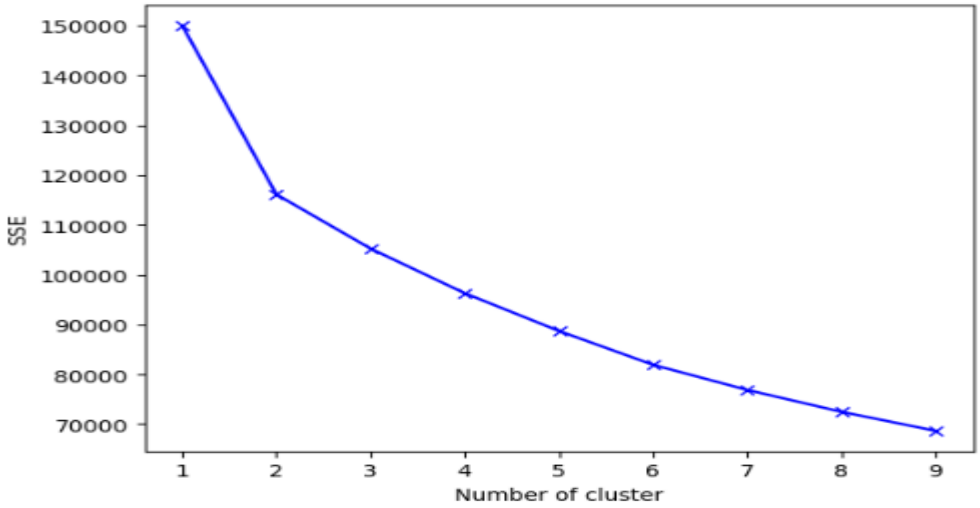


Figure 35 Elbow plot

Clusters vs Silhouette score

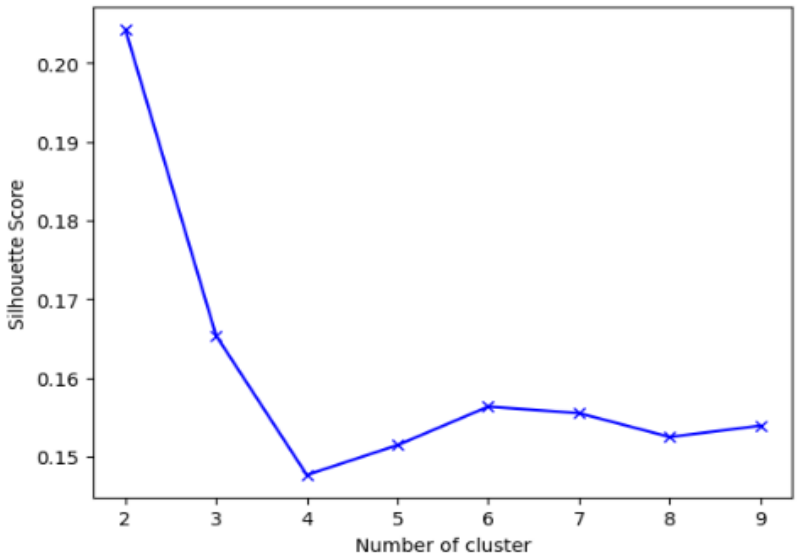


Figure 36 Clusters vs Silhouette score

Observations

- From the elbow plot consistent dip was observed from k = 2.
- To cross check if the selected k was the best a plot of cluster vs Silhouette was analysed and the best score was when the k = 2.
- The optimum cluster for clustering the data was identified to be 2.

Analysis from the two clusters

	group_0 Mean	group_1 Mean	group_0 Median	group_1 Median
age	44.713176	45.114915	44.0	45.0
avg_glucose_level	168.475069	166.624315	169.0	166.0
weight	**63.750777**	**79.142645**	**64.0**	**78.0**
fat_percentage	28.877391	28.749883	31.0	30.0
daily_avg_steps	5198.646085	5194.060207	5087.5	5084.0

	group_0 Mean	group_1 Mean	group_0 Median	group_1 Median
bmi	31.374958	31.284916	30.7	30.5
insurance_cost	**15131.478666**	**38662.594548**	**14808.0**	**37020.0**

Table 2 Clusters Statistical Analysis

Weight and insurance cost distribution for cluster 1 and cluster 2

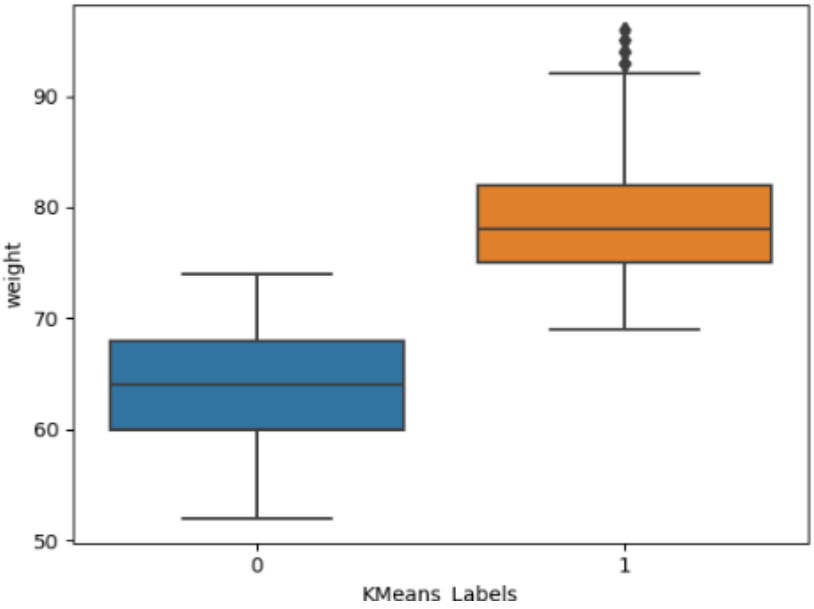


Figure 37 Weight cluster analysis

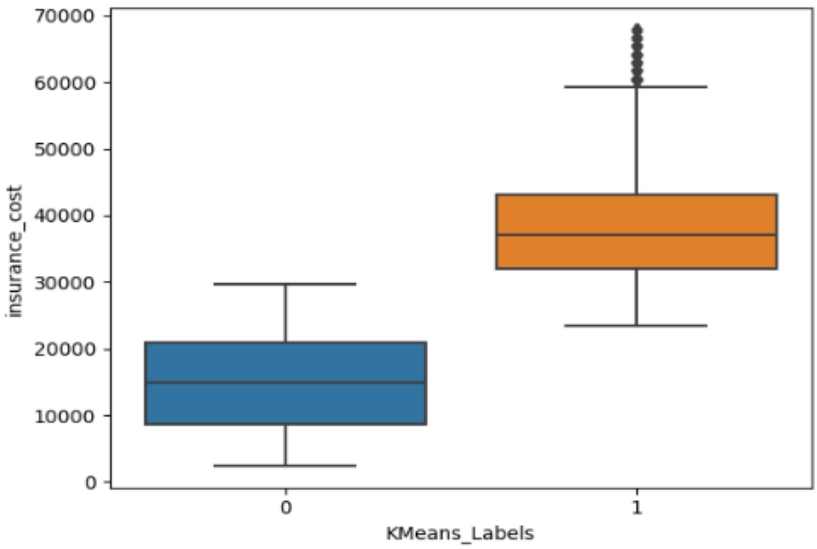


Figure 38 Insurance cost cluster analysis

Contradicting Parameters with respect to each clusters

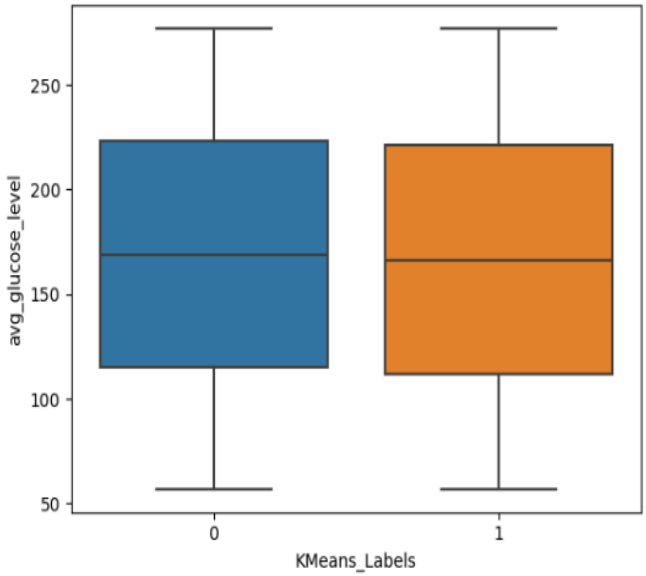


Figure 39 Glucose level each cluster

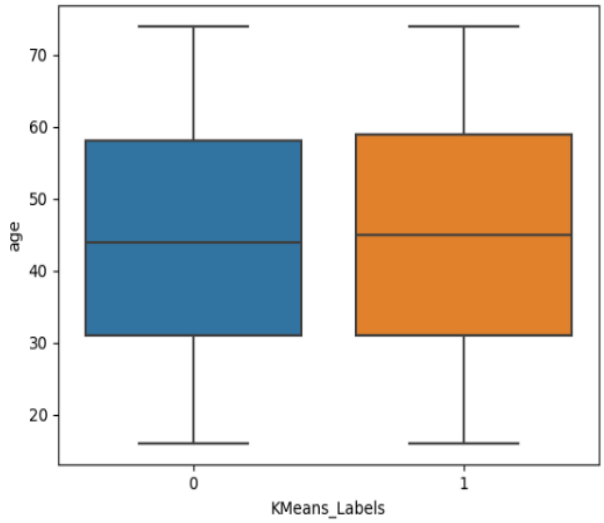


Figure 40 age vs clusters

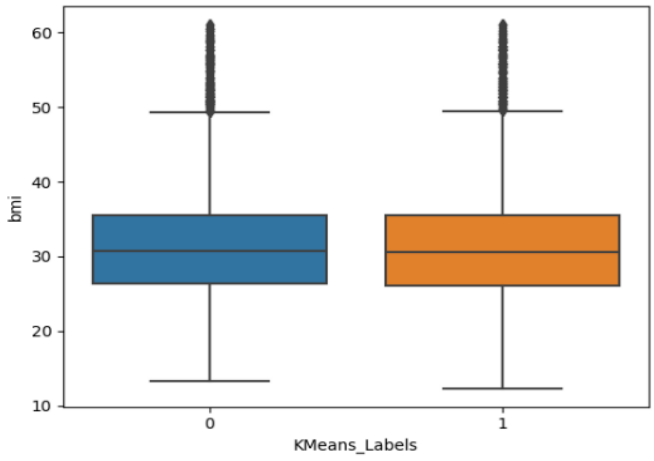


Figure 41 BMI cluster-wise

Business Insights and Recommendations

- Regular health checkup needs to be promoted. Considering how crucial it is to identify a disease to resolve in early stages. The insurance company can come up with discounts for regular health check-ups which will benefit the company and the customer. Customers get benefitted by early detection and early recovery and company benefits by not having to spend too much on claims made by the customer.

- Marketing or other strategies to attract more salaried class since the count is comparatively low. This could be an impact of salaried category already being provided with health benefit from the company they work for. Strategies such as advising such category customers on benefits of having personal health insurance might attract more salaried customers.

- 34% of the customers are female. Although this doesn't seem under sampled yet measure needs to be taken to attract female customers to buy the insurance. This can be done by educating them more about female health issues and how does having a health insurance help prevail such difficulties.

- Customers were observed to have alternate insurance from other company. There could be a chance that customers opt out of either of the companies. It is advisable to the company to not only provide benefits but also to not overcharge.

- It was observed that customers consuming alcohol on a daily basis had similar insurance cost distribution to the one who did not consume. Prices of such customers should be slightly on the upper ranges considering how deadly drinking on a regular basis can be.

- Weight was observed to be the main criteria of insurance cost. Higher the weight indicated higher the insurance costs. Weight variation too displayed similar results, higher the weight variations lower the insurance prices. However, weight variation 4 displayed the non-obvious distribution.

- Clustering was performed on the continuous data. The mean and median were observed for all the records, cluster 0 had all the records with lower insurance costs since the mean was 15131 whereas the cluster 1 had a mean of 38662 indicating higher insurance costs.

- The cluster analysis too displayed the similar scenario when the average of weight was on lower range the insurance cost too was lower and higher when average of weight was higher.

- The parameters such as age, glucose level and BMI displayed no major changes between clusters indicating that these parameters contributed minimal in predicting the insurance price. This could lead to company being bias in assuming only weight as a major criterion.

Model Building

The raw data analysed, cleaned and transformed was further subjected to training and testing to predict the test data based on trained values and evaluate the performance.

Evaluation metrics for Model performing

- **R2-** Score: The percentage of variance explained by the model explaining the relationship between dependent and independent variables.
- **RMSE:** The average difference between the models predicted values and the actual values.
- **MAPE:** The average absolute percentage difference between predicted values and actual values.

Ordinary Least Squared Model

This is a basic model and was built on the cleaned data. Data was subjected to scaling and no other assumptions were taken into account considering it to be a basic model for evaluating.

OLS Regression Results

Dep. Variable:	insurance_cost	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.946
Method:	Least Squares	F-statistic:	9304.
Date:	Wed, 17 Jul 2024	Prob (F-statistic):	0.00
Time:	14:37:14	Log-Likelihood:	-1.7876e+05
No. Observations:	18750	AIC:	3.576e+05
Df Residuals:	18714	BIC:	3.579e+05
Df Model:	35		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-7.981e+04	355.087	-224.754	0.000	-8.05e+04	-7.91e+04
Occupation_Salried	118.2579	71.294	1.659	0.097	-21.484	258.000
Occupation_Student	64.4083	63.802	1.009	0.313	-60.650	189.467
cholesterol_level	22.9886	23.214	0.990	0.322	-22.513	68.490
Alcohol	7.8072	39.390	0.198	0.843	-69.401	85.015
exercise	-52.6508	38.371	-1.372	0.170	-127.861	22.560
years_of_insurance_with_us	-93.7700	11.846	-7.916	0.000	-116.990	-70.550
regular_checkup_lasy_year	-345.8968	22.303	-15.509	0.000	-389.613	-302.180
adventure_sports	213.0472	90.546	2.353	0.019	35.568	390.526
visited_doctor_last_1_year	-57.7557	21.851	-2.643	0.008	-100.586	-14.926
age	3.7973	1.517	2.504	0.012	0.825	6.770
heart_decs_history	298.5270	108.929	2.741	0.006	85.016	512.038
other_major_decs_history	48.9160	83.772	0.584	0.559	-115.285	213.117
Gender	-41.6477	56.170	-0.741	0.458	-151.745	68.450
avg_glucose_level	-0.2781	0.390	-0.713	0.476	-1.043	0.487
Year_last_admitted	722.7582	64.524	11.201	0.000	596.285	849.231
weight	1485.8651	2.890	514.156	0.000	1480.201	1491.530
covered_by_any_other_company	1169.8366	55.602	21.040	0.000	1060.852	1278.821
weight_change_in_last_one_year	159.2809	15.635	10.187	0.000	128.635	189.927
fat_percentage	-2.7271	2.970	-0.918	0.359	-8.549	3.095
Location_Bangalore	407.5585	132.263	3.081	0.002	148.311	666.806
Location_Bhubaneswar	337.6376	133.775	2.524	0.012	75.427	599.848
Location_Chennai	468.2761	134.247	3.488	0.000	205.141	731.411
Location_Delhi	633.7917	134.026	4.729	0.000	371.089	896.494
Location_Guwahati	310.6958	133.829	2.322	0.020	48.388	573.012
Location_Jaipur	368.5436	133.695	2.757	0.006	106.489	630.598
Location_Kanpur	307.0008	134.589	2.281	0.023	43.194	570.808
Location_Kolkata	323.6058	134.263	2.410	0.016	60.437	586.774
Location_Lucknow	471.1241	134.432	3.505	0.000	207.625	734.623
Location_Mangalore	255.6172	133.775	1.911	0.056	-6.593	517.828
Location_Mumbai	372.5870	133.775	2.785	0.005	110.375	634.799
Location_Nagpur	480.2197	134.012	3.583	0.000	217.544	742.896
Location_Pune	365.7864	134.405	2.722	0.007	102.339	629.233
Location_Surat	446.7622	135.301	3.302	0.001	181.561	711.963
daily_avg_steps	-0.0296	0.029	-1.007	0.314	-0.087	0.028
bmi	-1.8893	3.988	-0.474	0.636	-9.707	5.928

Omnibus:	380.206	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	489.713
Skew:	0.337	Prob(JB):	1.08e-89
Kurtosis:	3.263	Cond. No.	8.58e+04

Table 3 Basic OLS Model

Model Evaluation:

Metrics	Training	Testing
R2-Score	94.6	94.6
RMSE	3343.48	3001.39
MAPE	0.152	1.063

Table 4 Basic OLS Model Evaluation

Assumptions check:

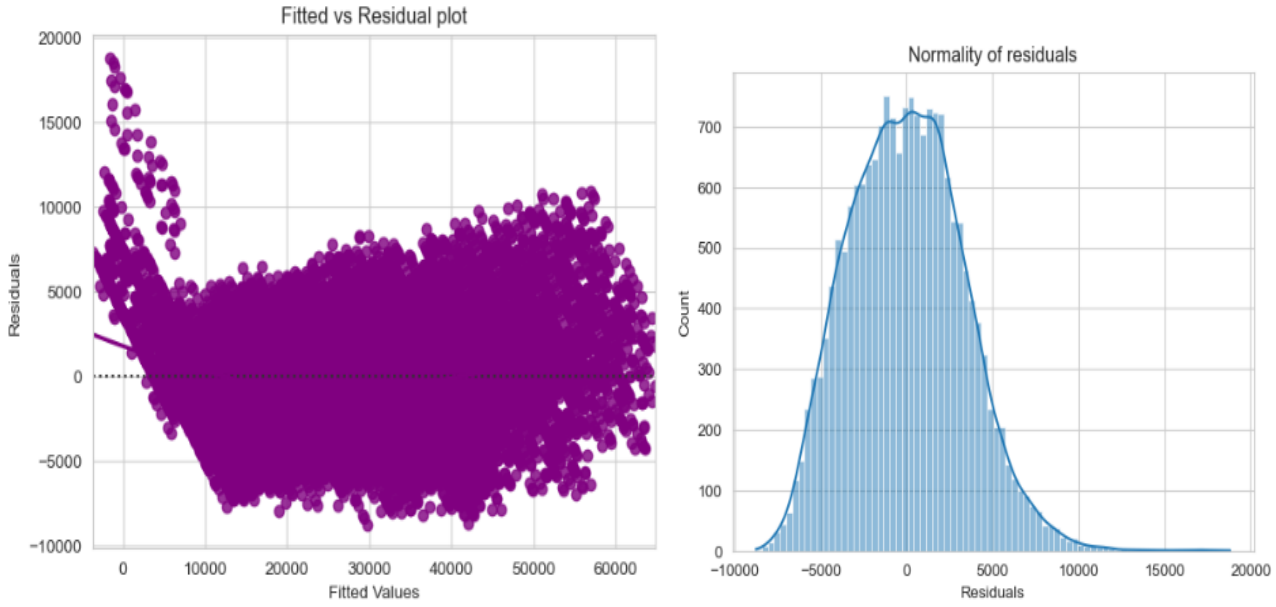


Figure 42 homoscedasticity and Normality Plots

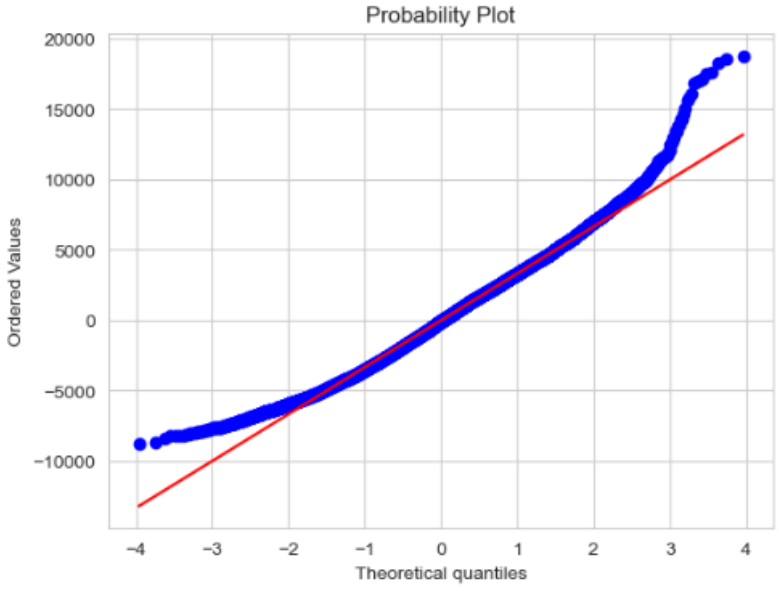


Figure 43 QQ-Plot for normality

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

• Null hypothesis - Data is normally distributed.

• Alternate hypothesis - Data is not normally distributed.

p-value = **8.195023518946873e-33**

Since p-value < 0.05, the **residuals are not normal** as per Shapiro test

Test for Homoscedasticity: The null and alternate hypotheses of the goldfeldquandt test are as follows:

• Null hypothesis: Residuals are homoscedastic.

• Alternate hypothesis: Residuals have heteroscedasticity.

p-value = 0.6524045981425037

Since p-value > 0.05 we can say that the **residuals are homoscedastic.**

Inferences from the basic OLS Model:

- The R2-score (r-squared) is well acceptable with explaining up-to 94.6% of variance. However, the model has not been up-to the mark in the RMSE score.
- The model appears to overfit in explaining the percentage of errors since the MAPE for training is low whereas for testing is high.
- It appears that errors follow no patten and goldfeldquandt test for hypothesis confirms the same.
- The errors when plotted appears to have skewed slightly towards right despite handling outliers in the data. The Shapiro-Wilk test supports the failure of this assumption.
- It was observed that many variables violated P- value feature importance test indicating presence of features contributing 0 or close to zero in predicting the target.
- Checking multicollinearity and eliminating features with very low feature importance one by one and further evaluating the model could improve the model for which VIF was employed and |P| > t was observed.

VIF Check

Occupation_Salried	1.314709	**weight_change_in_last_one_year**	1.172687
Occupation_Student	1.642167	**fat_percentage**	1.099683
cholesterol_level	1.425725	**Location_Bangalore**	1.917865
Alcohol	1.021919	**Location_Bhubaneswar**	1.876517
exercise	1.014815	**Location_Chennai**	1.864681
years_of_insurance_with_us	1.598195	**Location_Delhi**	1.876628
regular_checkup_lasy_year	1.1929	**Location_Guwahati**	1.875264
adventure_sports	1.008306	**Location_Jaipur**	1.877053
visited_doctor_last_1_year	1.038151	**Location_Kanpur**	1.858764
age	1.002025	**Location_Kolkata**	1.863755
heart_decs_history	1.021793	**Location_Lucknow**	1.85863
other_major_decs_history	1.051805	**Location_Mangalore**	1.87514
Gender	1.191267	**Location_Mumbai**	1.879304
avg_glucose_level	1.001959	**Location_Nagpur**	1.866521
Year_last_admitted	1.738491	**Location_Pune**	1.8607
weight	1.219255	**Location_Surat**	1.838639
covered_by_any_other_company	1.089913	**daily_avg_steps**	1.057207
		bmi	1.188256

Table 5 Variance Inflation Factor

Observation: As a predefined threshold any VIF value greater than 5 would be eliminated. However, the model displayed very minimal or almost no multicollinearity. This indicates elimination of features with respect to Multicollinearity is not required. Hence, the model passes the assumption of No Multicollinearity.

OLS Model-2

OLS Regression Results

Dep. Variable:	insurance_cost	R-squared:	0.946
Model:	OLS	Adj. R-squared:	0.946
Method:	Least Squares	F-statistic:	1.303e+04
Date:	Wed, 17 Jul 2024	Prob (F-statistic):	0.00
Time:	14:42:58	Log-Likelihood:	-1.7876e+05
No. Observations:	18750	AIC:	3.576e+05
Df Residuals:	18724	BIC:	3.578e+05
Df Model:	25		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.711e+04	24.439	1109.429	0.000	2.71e+04	2.72e+04
Occupation_Salried	43.0805	24.457	1.761	0.078	-4.857	91.018
years_of_insurance_with_us	-244.7641	30.891	-7.923	0.000	-305.313	-184.215
regular_checkup_lasy_year	-414.6598	26.686	-15.538	0.000	-466.967	-362.352
adventure_sports	58.0365	24.535	2.365	0.018	9.946	106.127
visited_doctor_last_1_year	-62.9889	24.448	-2.576	0.010	-110.910	-15.068
age	61.5214	24.459	2.515	0.012	13.580	109.463
heart_decs_history	66.0847	24.452	2.703	0.007	18.156	114.014
Year_last_admitted	361.0962	32.218	11.208	0.000	297.946	424.247
weight	1.388e+04	26.980	514.278	0.000	1.38e+04	1.39e+04
covered_by_any_other_company	536.6232	25.510	21.036	0.000	486.621	586.625
weight_change_in_last_one_year	268.3128	26.451	10.144	0.000	216.466	320.160
Location_Bangalore	103.7155	33.837	3.065	0.002	37.391	170.040
Location_Bhubaneswar	84.4315	33.471	2.523	0.012	18.825	150.038
Location_Chennai	116.3714	33.363	3.488	0.000	50.977	181.765
Location_Delhi	157.8378	33.472	4.716	0.000	92.230	223.446
Location_Guwahati	77.0939	33.455	2.304	0.021	11.519	142.669
Location_Jaipur	92.3261	33.479	2.758	0.006	26.704	157.948
Location_Kanpur	75.4678	33.304	2.266	0.023	10.188	140.747
Location_Kolkata	80.7037	33.358	2.419	0.016	15.319	146.088
Location_Lucknow	116.5849	33.311	3.500	0.000	51.293	181.877
Location_Mangalore	64.4407	33.455	1.926	0.054	-1.135	130.016
Location_Mumbai	92.9257	33.499	2.774	0.006	27.265	158.587
Location_Nagpur	118.9677	33.382	3.564	0.000	53.536	184.400
Location_Pune	89.8990	33.326	2.698	0.007	24.576	155.222
Location_Surat	109.3005	33.130	3.299	0.001	44.363	174.238

Omnibus:	380.656	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	410.202
Skew:	0.338	Prob(JB):	8.43e-90
Kurtosis:	3.263	Cond. No.	5.04

Table 6 OLS Model-2

The OLS Model-2 was obtained as a result of eliminating non important features. The column P>|t| was observed. Any value greater than 0.05 was considered exhibiting more than 5% of feature unimportance. Features were eliminated one by one and model was evaluated. Table 6 is an outcome of final OLS model post eliminating unimportant features.

Although many features were eliminated the R-squared remained same.

Metrics	Training	Testing
R2-Score	94.6	94.6
RMSE	3344.09	3394.5
MAPE	0.152	0.153

Table 7 OLS Model-2 Evaluation

Assumption check

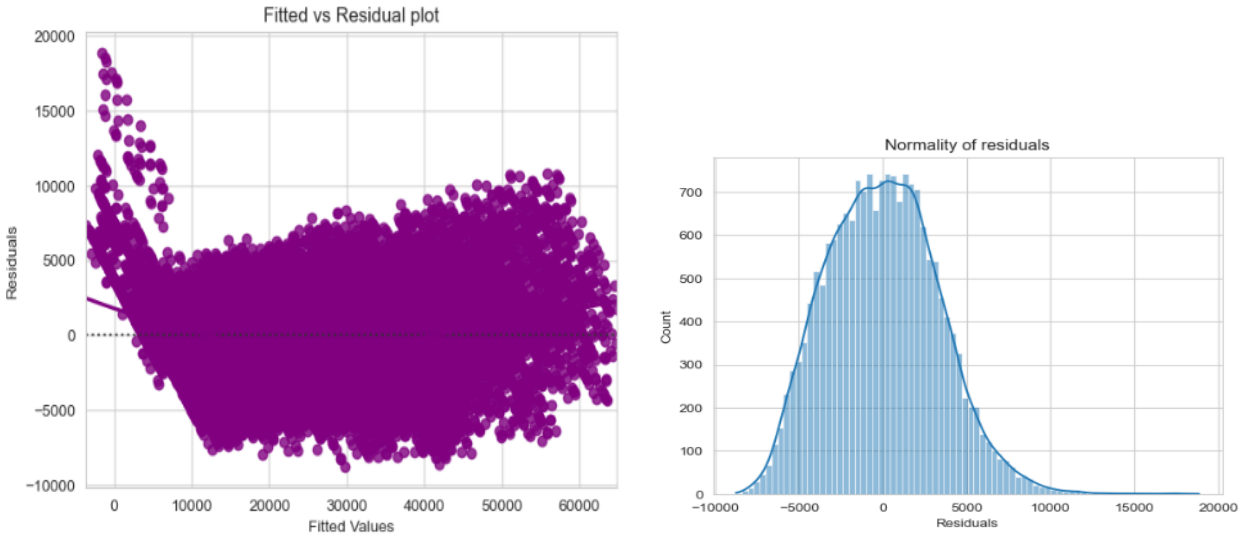


Figure 44 Homoscedasticity and Normality plot

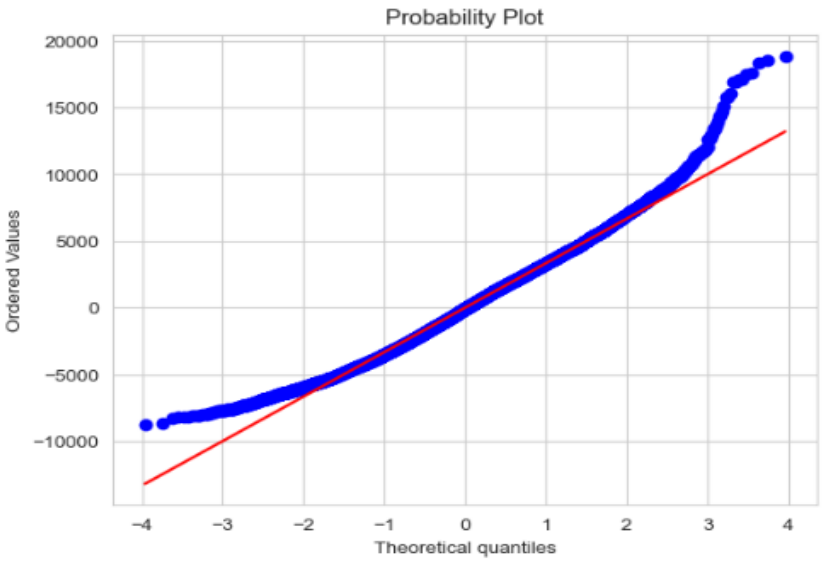


Figure 45 QQ-Plot Model-2

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

• Null hypothesis - Data is normally distributed.

• Alternate hypothesis - Data is not normally distributed.

p-value = **6.92276688126871e-33**

Since p-value < 0.05, the **residuals are not normal** as per Shapiro test. However, the p-values has improved compared to previous model

Test for Homoscedasticity: The null and alternate hypotheses of the goldfeldquandt test are as follows:

• Null hypothesis: Residuals are homoscedastic.

• Alternate hypothesis: Residuals have heteroscedasticity.

p-value = **0.6635874106938857**

Since p-value > 0.05 we can say that the **residuals are homoscedastic.**

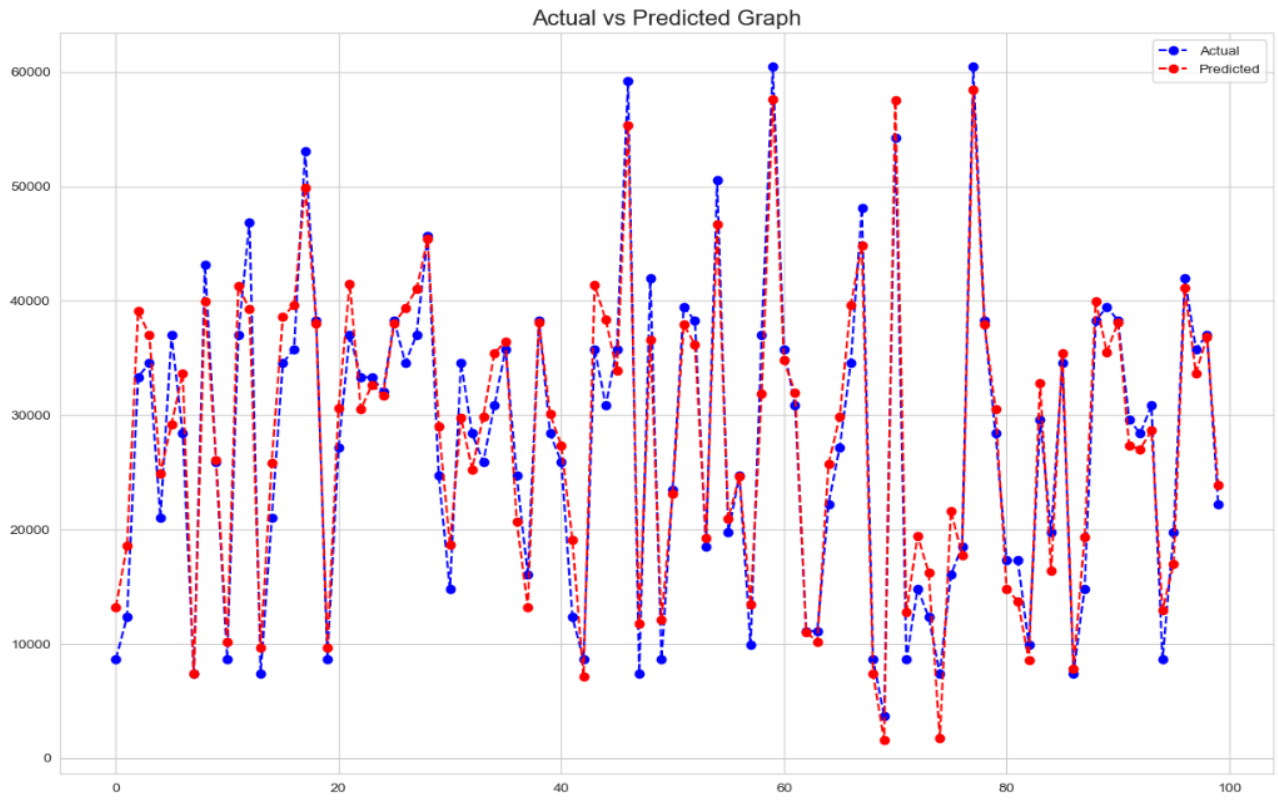


Figure 46 OLS-Model 2 Actual vs Predicted Graph

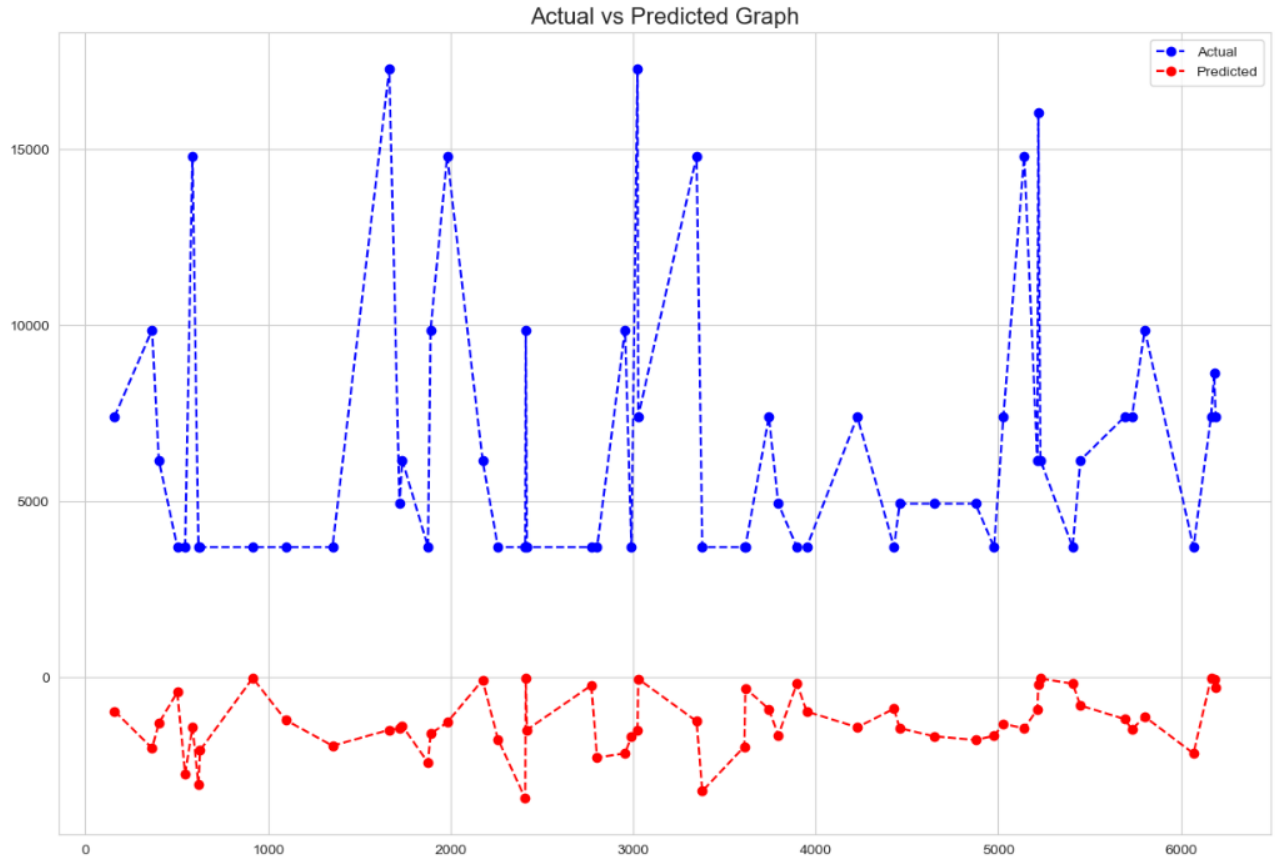


Figure 47 OLS-Model 2 unusual predictions

Inferences from OLS Model-2

- The R2-score (r-squared) is well acceptable with explaining up-to 94.6% of variance. However, the model has not been up-to the mark in the RMSE score.

- The model appears to not overfit in explaining the variance, RMSE and MAPE with minimal difference between training and testing.

- It appears that errors follow no patten and goldfeldquandt test for hypothesis confirms the same.

- The errors when plotted appears to have skewed slightly towards right despite handling outliers in the data. The Shapiro-Wilk test supports the failure of this assumption but has prevailed over the previous model.

- The actual vs predicted graph denoted minimal error yet there have been 56 unusual observations where the values were predicted either zero or less than 0 (negative). Such observations are highly impossible as these observations are a result of model extrapolating.

Regularization

Ridge Model

Metrics	Training	Testing
R2-Score	94.5	94.3
RMSE	3337.4	3401.7
MAPE	0.151	0.152

Table 8 Ridge Model Evaluation

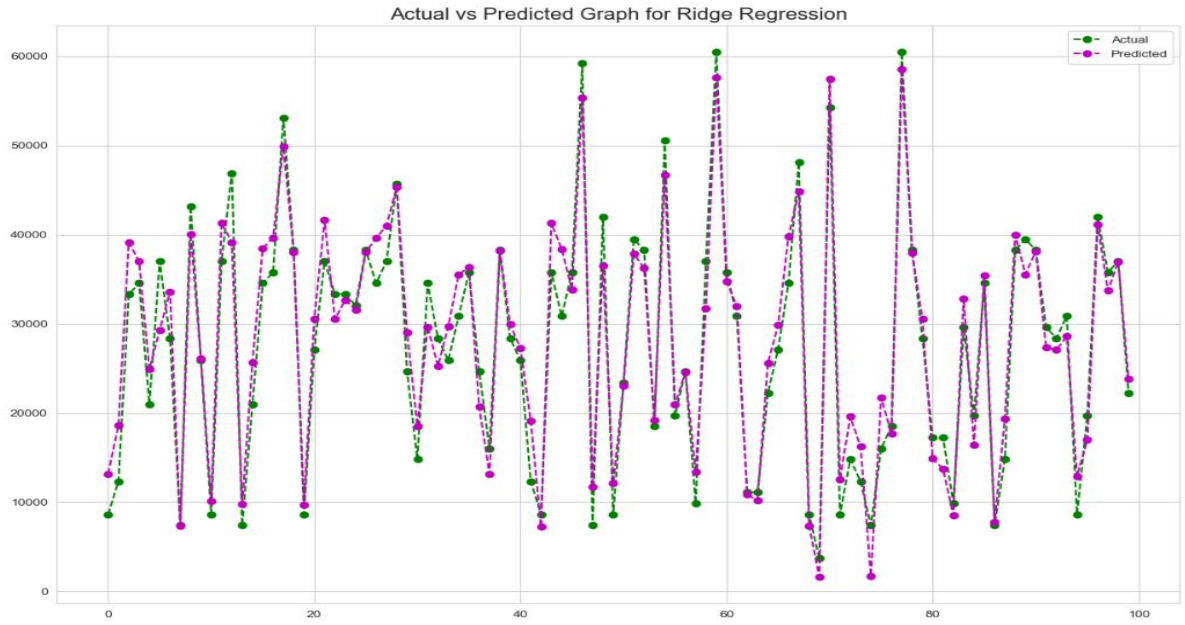


Figure 48 Ridge Actual vs Predicted

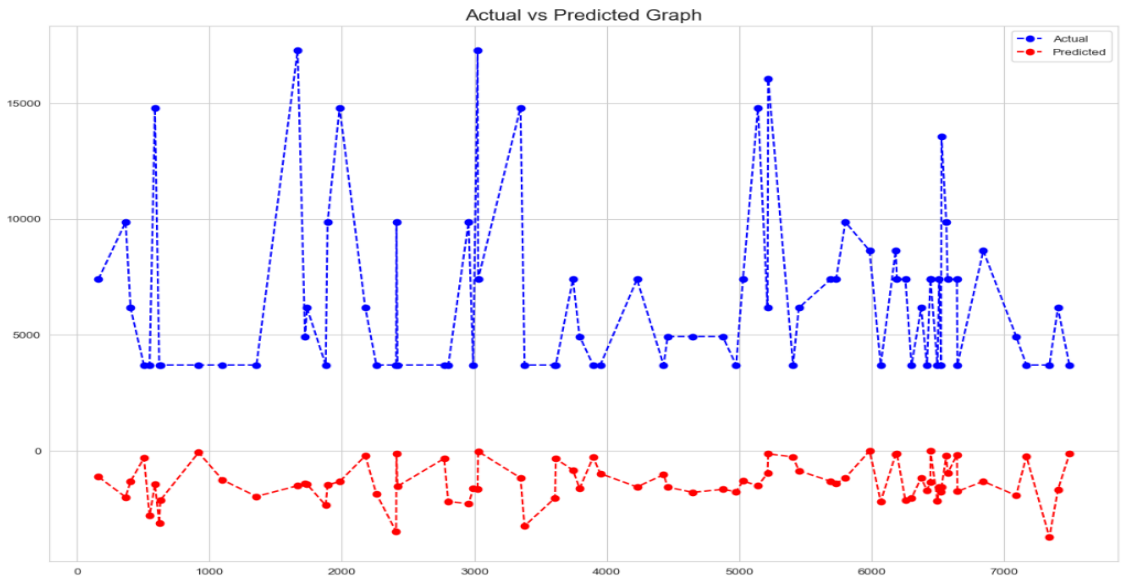


Figure 49 Lasso predicted unusual values

Lasso Model

Metrics	Training	Testing
R2-Score	94.5	94.3
RMSE	3337.4	3401.6
MAPE	0.151	0.155

Table 9 Lasso Model evaluation

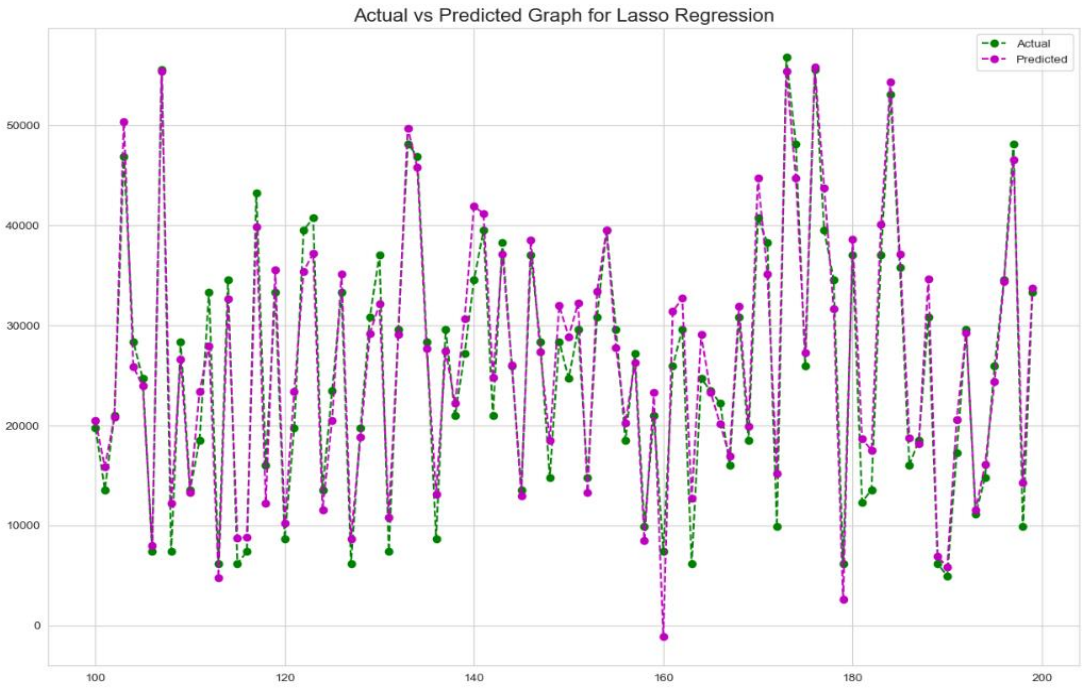


Figure 50 Lasso Actual vs Predicted

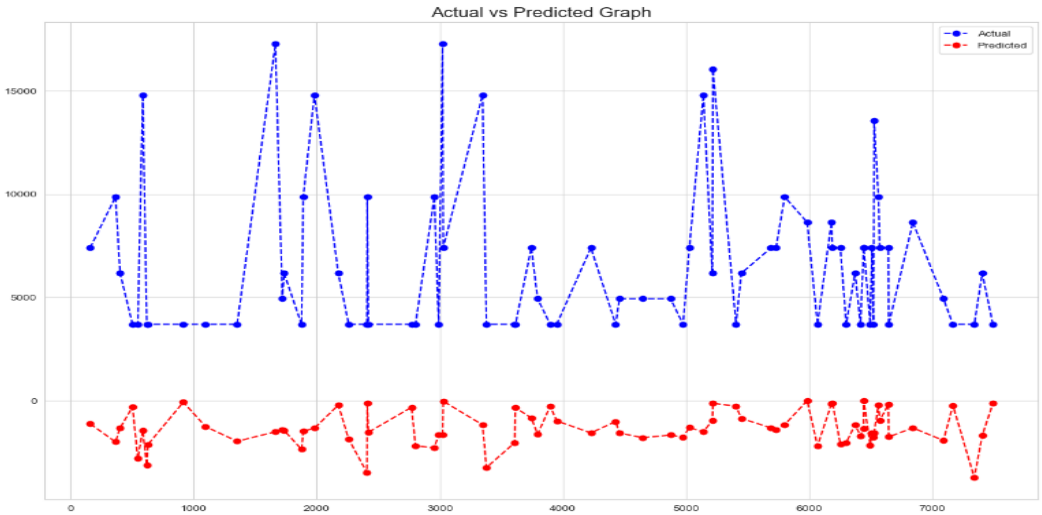


Figure 51 Lasso unusual prediction

Decision Tree Regressor

A basic decision tree regressor was built on the clean data and the model was evaluated.

Mechanism:

- A Decision Tree Regressor splits the data into subsets based on the feature that reduces variance the most at each split.
- It recursively partitions the data until a stopping criterion is met, such as a maximum depth or minimum number of samples per leaf.

Handling Data:

- Handles both categorical and continuous features.
- Requires categorical data to be encoded if not using specialized implementations.

Strengths:

- Simple to understand and interpret.
- Non-parametric and requires little data preprocessing.
- Can capture non-linear relationships.

Weaknesses:

- Prone to overfitting, especially with deep trees.
- Sensitive to small variations in data.

Model evaluation

Metrics	Training	Testing
R2-Score	100	90.8
RMSE	0	4325.94
MAPE	0.0	0.158

Table 10 Decision tree regressor evaluation

Inferences

Over fitting model

- Performs well on training but worst on testing. Hyper tuning parameters could help overcome overfitting issues.
- Overfitting observed in r-squared, RMSE and MAPE.

Hyper tuned Decision tree Regressor

Metrics	Training	Testing
R2-Score	91.92	91.89
RMSE	4075.6	4061.86
MAPE	0.15	0.151

Table 11 Hyper tuned DT Regressor evaluation

Inferences

- Overfitting reduced as the difference between training and testing for all the evaluation metrics is minimal.
- Although the overfitting is reduced and the model's ability to extrapolation is eliminated yet the errors are high. There could be other models that might further reduce the error which shall be evaluated further.
- The hyper tuned model performs exceedingly well when compared to the one with basic decision tree regressor model.

Random Forest Regressor (Ensemble learning)

Mechanism:

- An ensemble method that builds multiple decision trees using different subsets of the data and features.
- Each tree votes for a prediction, and the final prediction is the average of all tree predictions.
- Uses techniques like bagging and feature randomness to improve performance and reduce overfitting.

Advantages:

- Reduces overfitting compared to a single decision tree.
- Handles missing values and maintains accuracy when a large proportion of data is missing.
- Robust to noise in the data.

Random Forest regressor model evaluation

Metrics	Training	Testing
R2-Score	99.34	95.19
RMSE	1160.36	3127.31
MAPE	0.04	0.12

Table 12 Random Forest regressor model evaluation

Inferences

- Model Overfitting.
- Performed well on training but poorly on testing.
- Hyper tuning parameters would help resolve overfitting.

Random Forest Regressor Hyper-tuning parameters

Random Forest (Hyper-tuning parameters) model evaluation

Metrics	Training	Testing
R2-Score	97.54	95.44
RMSE	2248.19	3127.31
MAPE	0.08	0.11

Table 13 Random Forest (Hyper-tuned) model evaluation

Inferences

- Majority of the overfitting observed in the basic random forest model is reduced.
- The difference of variance explained in training and testing is minimal but the gap is slightly larger when compared to the RMSE and MAPE indicating the model's error is comparatively low on training but high on testing
- This model although hyper tuned displays overfitting but not at a larger range. This can still be considered an average model.
- The gap could further be reduced and a better model could be built with boosting algorithms such as XG-Boost and CAT-Boost.

XG Boost (Boosting)

Mechanism:

- An optimized implementation of gradient boosting that is highly efficient and scalable.
- Uses techniques like regularization, sparsity awareness, and parallel processing to improve performance.
- Similar to gradient boosting but with additional enhancements.

Advantages:

- Superior performance and accuracy.
- Handles missing values and sparse data well.
- Provides built-in cross-validation and early stopping.

Disadvantages:

- Complex and less interpretable than simpler models.
- Requires extensive hyperparameter tuning.
- Still limited in extrapolation capabilities.

XG-Boost Model evaluation

Metrics	Training	Testing
R2-Score	97.69	95.04
RMSE	2175.54	3177.04
MAPE	0.08	0.12

Table 14 XG Boost Model evaluation

Inferences

- The model was built on the XGB Regressor with no parameter being tuned.
- Overfitting observed
- The gap between training and testing for RMSE and MAPE was observed indicating lesser errors in training and slightly larger errors in testing.
- Hyper tuning parameters could help overcome overfitting.

XG Boost (Hyper-tuned parameters)

Metrics	Training	Testing
R2-Score	95.7	95.4
RMSE	2972.72	3034.75
MAPE	0.11	0.12

Table 15 XG-Boost (Hyper-tuned) Model evaluation

Inferences

- The model overcame overfitting issues observed with the basic model.
- The error with respect to RMSE and MAPE too reduced.
- The gap between the training and testing was bridged as minimal difference was observed for the r-squared, RMSE and MAPE in training and testing.

Cat-Boost Regressor

Mechanism:

- An advanced gradient boosting method that handles categorical features natively without requiring preprocessing.
- Uses ordered boosting and other techniques to reduce overfitting and improve accuracy.
- Can automatically handle categorical data and missing values.

Advantages:

- Excellent performance on datasets with categorical features.
- Reduces the need for extensive data preprocessing.
- Less prone to overfitting due to techniques like ordered boosting.

Disadvantages:

- More complex and less interpretable than simpler models.
- Requires proper tuning of parameters to achieve optimal performance.
- Computationally intensive compared to simpler models.

Cat-Boost Regressor Model evaluation

Metrics	Training	Testing
R2-Score	96.8	95.39
RMSE	2526.55	3062.99
MAPE	0.098	0.12

Table 16 Cat-Boost Regressor Model evaluation

Inferences

- This boosting method is known to be advanced gradient boosting technique and the results signify the same. The above model built was non hyper tuned parametric method yet the concept of overfitting is minimal but present.

- The model when compared to other tree regressor basic model (non-hyper tuning parameter) has displayed better outcomes in terms of explaining the variance, RMSE and MAPE for both training and testing.

- We can conclude that this model when hyper tuned could provide far better results when compare with other models.

Cat Boost Regressor (Hyper tuned parameters)

Metrics	Training	Testing
R2-Score	95.8	95.5
RMSE	2935.32	3022.51
MAPE	0.11	0.12

Table 17 Cat Boost Regressor (Hyper-tuned) model evaluation

Inferences

- The hyper tuning of parameters for Cat boost regressor has worked well. The overfitting has been eliminated.

- The gap between training and testing has drastically reduced when compare with other models.

- The distance between the R2-score, average RMSE and MAPE for training and testing has been observed to be minimal. This model has prevailed over other hyper-tuned models and is emerging out to be the **best model**.

All Model Comparison

Models	R2-Score		RMSE		MAPE	
	Training	Testing	Training	Testing	Training	Testing
Ordinary Least squared Model-1	94.6	94.6	3343.48	3001.39	15.20%	106.30%
Ordinary Least squared Model-2	94.6	94.6	3344.09	3394.5	15.20%	15.36%
Ridge	94.5	94.3	3337.4	3401.77	15.12%	15.52%
Lasso	94.5	94.3	3337.4	3401.67	15.12%	15.55%
ElasticNet	94.5	94.3	3337.4	3401.62	15.11%	15.54%
Tree Regressor						
Decision Tree	100	90.8	0	4325.94	0.00%	15.89%
Decision Tree (Hyper-tuned)	91.92	91.89	4075.6	4061.86	15.06%	15.17%
Random Forest	99.34	95.19	1160.36	3127.31	4.00%	12.00%
Random Forest (Hyper-tuned)	97.54	95.44	2248.19	3044.65	8.00%	11.90%
XGBoost	97.69	95.04	2175.54	3177.04	8.50%	12.50%
XGBoost (Hyper-tuned)	**95.7**	**95.4**	**2972.72**	**3034.75**	**11.67%**	**12.09%**
CatBoost	96.8	95.39	2526.55	3062.99	9.80%	12.05%
CatBoost (Hyper-tuned)	**95.81**	**95.51**	**2935.32**	**3022.51**	**11.45%**	**12.00%**

Table 18 Model Comparison chart

From the table above it can be observed that **XG-Boost Regressor** and **Cat-Boost Regressor** when hyper tuned the parameters have exhibited better results when compared to other and have also overcome the overfitting and underfitting issues. The root mean squared error and mean absolute percentage error too have been observed to be reduced when compare with other models.

Actual vs Predicted graphs

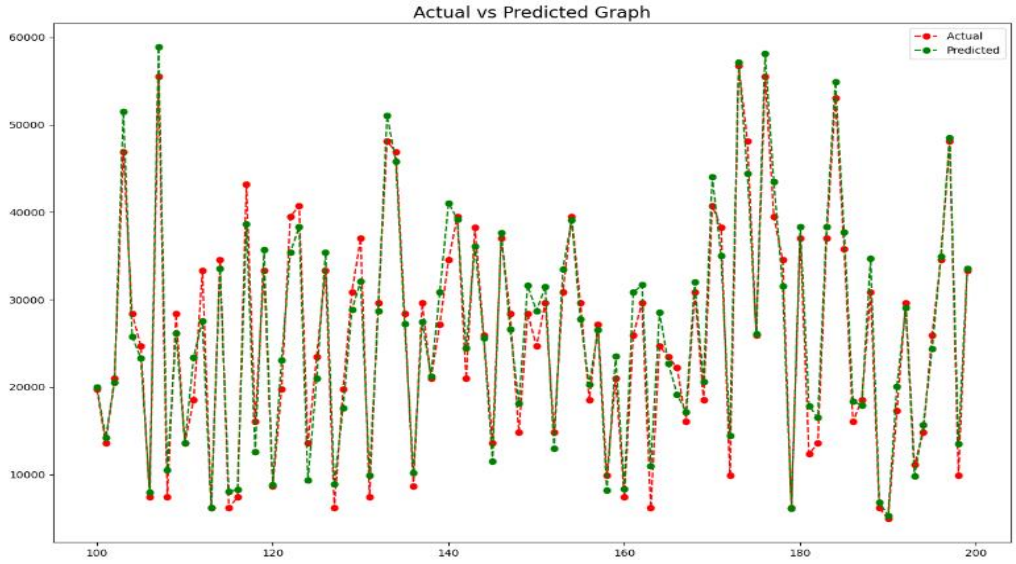


Figure 52 Actual vs Predicted graph for XGB Regressor

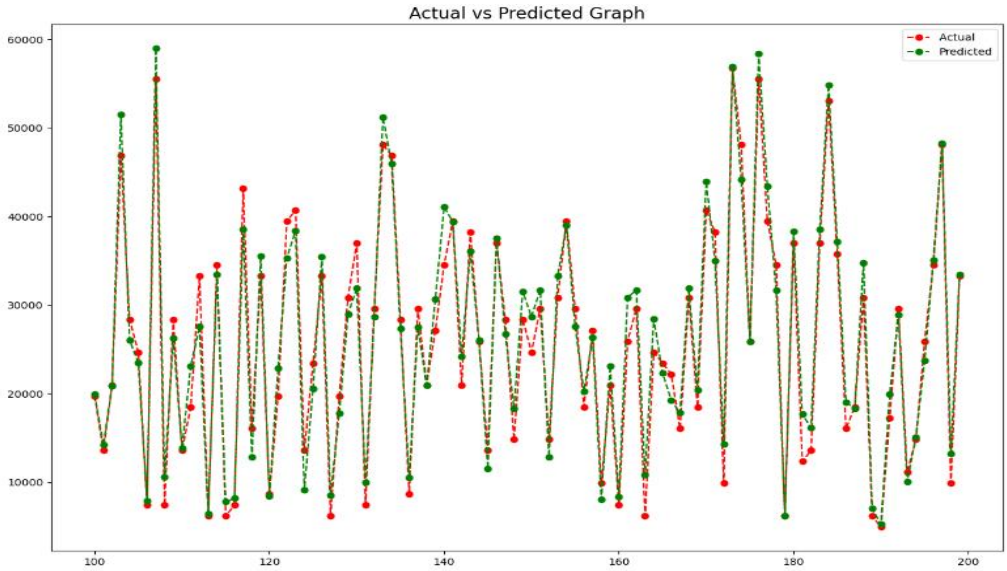


Figure 53 Actual vs Predicted graph for Cat-Boost Regressor

Inferences

- The Actual values most of the times overlap with the Predicted values with minimal or negligible distances.

- When compared to that of Linear regression, Regularization (Lasso, Ridge and Elastic Net) techniques the prediction has been more accurate and the concept of extrapolation has been eliminated further leading to non-predictions of unusual data such as zero or negative.

Feature Importance

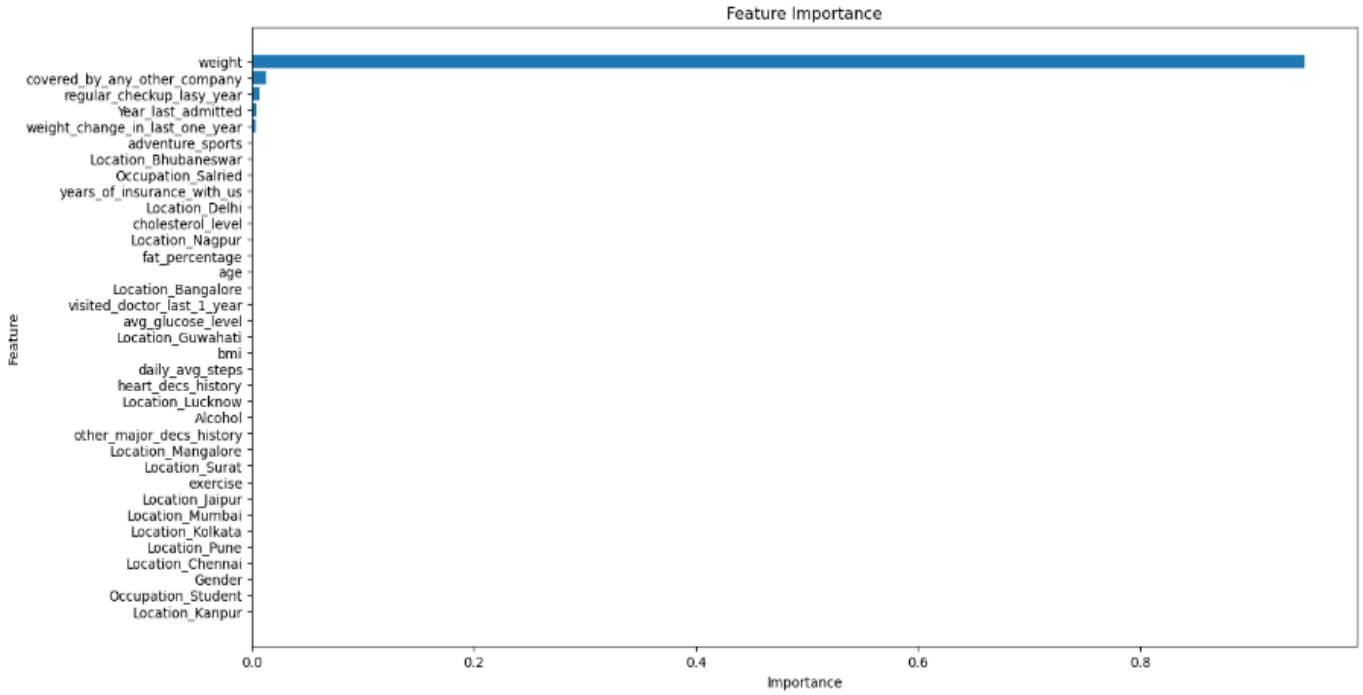


Figure 54 Feature Importance XGB-Regressor

Feature	Importance
weight	0.948191166
covered_by_any_other_company	0.01324499
regular_checkup_lasy_year	0.006913209
Year_last_admitted	0.004160683
weight_change_in_last_one_year	0.003753703
adventure_sports	0.001258778
Location_Bhubaneswar	0.001143344
Occupation_Salried	0.001126966
years_of_insurance_with_us	0.001119938
Location_Delhi	0.001093971
cholesterol_level	0.001030798
Location_Nagpur	0.000933461
fat_percentage	0.00093078
age	0.000921829
Location_Bangalore	0.000902919
visited_doctor_last_1_year	0.000901561
avg_glucose_level	0.000886366
Location_Guwahati	0.000861966
bmi	0.000861745
daily_avg_steps	0.000835132
heart_decs_history	0.000804682
Location_Lucknow	0.00080258
Alcohol	0.000798242
other_major_decs_history	0.00078319
Location_Mangalore	0.000766617
Location_Surat	0.000746916
exercise	0.000669571
Location_Jaipur	0.000659876
Location_Mumbai	0.0006501
Location_Kolkata	0.000593163
Location_Pune	0.000582033
Location_Chennai	0.000579795
Gender	0.000489884
Occupation_Student	0
Location_Kanpur	0

Table 19 Feature importance for XG-Boost Regressor

Feature Importance of Cat-Boost Regressor model

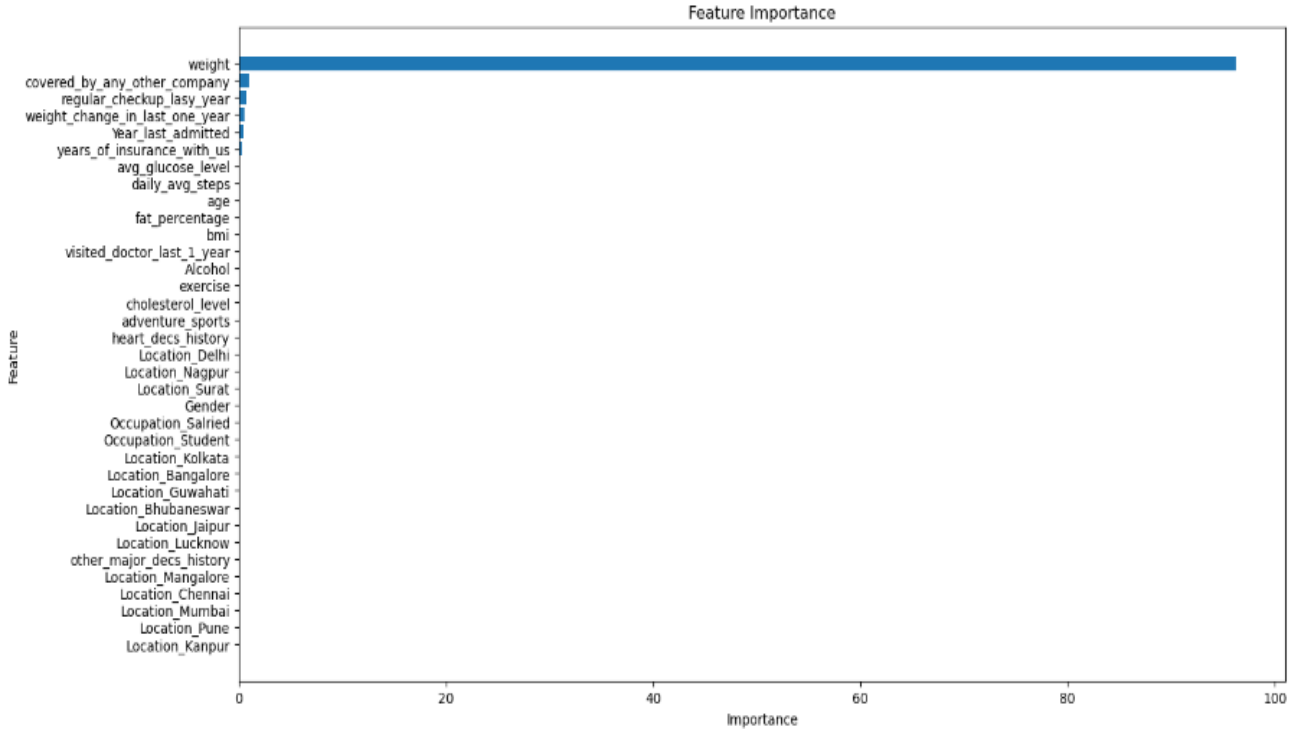


Figure 55 Feature Importance Cat-Boost Regressor

Feature	Importance
weight	96.22457733
covered_by_any_other_company	1.002386681
regular_checkup_lasy_year	0.662617819
weight_change_in_last_one_year	0.547799101
Year_last_admitted	0.455546121
years_of_insurance_with_us	0.242776061
avg_glucose_level	0.104510077
daily_avg_steps	0.094273846
age	0.087927748
fat_percentage	0.084174668
bmi	0.083902813
visited_doctor_last_1_year	0.068272808
Alcohol	0.059928447
exercise	0.055954538
cholesterol_level	0.051318474
adventure_sports	0.049349901
heart_decs_history	0.013934731
Location_Delhi	0.010867048
Location_Nagpur	0.009613397
Location_Surat	0.009098142
Gender	0.008761887
Occupation_Salried	0.007746915
Occupation_Student	0.007530685
Location_Kolkata	0.006847993
Location_Bangalore	0.006641866
Location_Guwahati	0.006399888
Location_Bhubaneswar	0.005545134
Location_Jaipur	0.005455423
Location_Lucknow	0.005058079
other_major_decs_history	0.004714495
Location_Mangalore	0.00456443
Location_Chennai	0.004501028
Location_Mumbai	0.003543509
Location_Pune	0.002463065
Location_Kanpur	0.001395852

Table 20 Feature importance table for Cat-Boost Regressor

Inferences

- The feature importance plot for the best models was plotted and a table for exact importance value was displayed.
- It was noted that weight appears to be the major factor influencing in the prediction of the insurance cost for both the models.
- The second most important feature was whether the insurance was covered by any other company consisting of binary values. However, the importance of second most important feature appears to be far lower than that of the weight feature.
- Other features in the data have exhibited its importance very close to zero when rounded up these are good to consider as zero.
- All the features were included into the model building to display how features that seem important in real world for finding optimal insurance cost have not contributed here.

Model Insights

- Post analysis on raw data, inferences were drawn that would avert the policy makers to design policies in such a way both the makers and the customers would get benefitted. Click here to know more about inferences on the raw data and assumptions made.

- The raw data analysis supports the model building since the feature weight has contributed majorly in predicting the model. This was observed in the EDA when it was projected against the insurance cost.

- All possible regressor models were built. From basic models to building advanced hyper tuned parameter model were evaluated and observations were noted down.

- Few unacceptable observations were identified for models such as Linear regression and Regularization techniques such as Lasso, Ridge and Elastic Net. There were predictions made by the mentioned models predicting zeroes and negatively values which is impossible in the case of finding optimal insurance cost.

- Apart from the zero and negative value predictions, models exhibited higher error percentages despite explaining good variance for both training and testing. The Ordinary least squared model/Linear regression models were subjected to assumptions check and were evaluated.

- All possible tree regressor models were built and evaluated since other models exhibited extrapolation leading models to predict zero and negative insurance prices. Although basic models displayed overfitting, this was resolved post hyper tuning parameters.

- Out of all the models that were built and evaluated, Xtreme Gradient Boosting Regression and Categorical Boosting Regression model was observed to have displayed high performing results. The model explained excellent variance, the average root mean squared error and average percentage error too were observed to be low when compare with rest of the models with no overfitting.

- XG-Boost and Cat-Boost regressor model feature importance and actual vs predicted graphs were plotted. The actual vs predicted graph for both the models exhibited minimal distance between actual and predicted. However, for both the model's weight was observed to be the most important feature with other features being either close to zero or zero itself.

- Important decisions need to be taken by the insurance company considering the data analysis and how models have reacted with the independent features in explaining the variance. Only weight has major weightage in predicting the target whereas age, glucose level, alcohol, BMI, heart and other disease have contributed less.

Business Recommendations

1. Broaden the Risk Assessment Criteria:
    - Holistic Health Evaluation: Incorporate a more comprehensive set of health metrics in the risk assessment process. This includes Alcohol consumption, BMI, Heart disease, other major diseases, exercise habits, daily average steps, glucose levels, and cholesterol levels.

2. Personalized Insurance Plans:
    - Customized Premiums: Develop personalized insurance plans that consider a wider range of health metrics. This can ensure that customers with healthy lifestyles and good health metrics are rewarded with lower premiums.

    - Incentives for Healthy Behaviour: Introduce incentives for customers who engage in healthy behaviours, such as regular exercise, maintaining a healthy diet, and undergoing regular health check-ups.

3. Education and Awareness Programs:
    - Health Literacy: Implement programs to educate customers on the importance of various health metrics beyond weight. This can include the risks associated with high cholesterol, glucose levels, and lack of exercise.

    - Preventive Health: Promote preventive health measures and regular health screenings to help customers maintain overall well-being.

4. Enhanced Data Collection:
    - Wearable Technology: Encourage customers to use wearable devices that track daily activity, heart rate, and other health metrics. This data can provide a more accurate picture of a customer's health and risk profile.

    - Regular Health Updates: Require customers to update their health information regularly, ensuring the insurance company has the most current data for accurate risk assessment.

5. Policy Adjustments:
    - Review Underwriting Policies: Reevaluate underwriting policies to ensure they reflect a balanced approach to health risk assessment.

    - Dynamic Premium Adjustments: Implement dynamic premium adjustments that reflect changes in a customer's health status over time, rewarding positive health improvements.

Impact on Business and Customers:

For the Company:

- Improved Risk Management: A comprehensive approach to risk assessment can lead to more accurate premium pricing and better management of risk.
- Customer Retention: Personalized and fair insurance plans can improve customer satisfaction and retention.
- Market Differentiation: Offering health-oriented incentives and personalized plans can differentiate the company in a competitive market.

For the Customers:

- Fair Pricing: Customers receive insurance premiums that reflect their overall health, not just their weight, leading to fairer pricing.
- Health Incentives: Incentives for maintaining good health can motivate customers to adopt healthier lifestyles.
- Support and Education: Increased support and educational resources can empower customers to make informed health decisions.

By implementing these recommendations, the company can create a more equitable and comprehensive insurance offering that benefits both the business and its customers, while minimizing bias and promoting overall health.