Business Report

Capstone: Health Care Project

Final Report Submission

Submitted by: Monica A

Contents

Introduction:....5

Problem Statement: 5

Need of the study/project:....5

Exploratory Data Analysis 5

Dataset Description:....5

Data Dictionary: 6

Data Information:....7

Data Skewness: 7

5-Point Summary or Descriptive Statistics of the data:....8

Data Cleaning and Preparation: 9

Univariate Analysis:....14

Bivariate Plot: 16

Correlation Plot:....18

Data Encoding: 19

Renaming and dropping unnecessary columns for Model Building:....20

Principal Component Analysis 21

Significance of Correlation: 21

Confirmation of Adequacy of Sample Size: 21

PC Analysis:....21

Scree Plot:....22

Principal Component Analysis (PCA) Insights:....24

KMeans Clustering....24

Clustering: 24

KMeans Clustering:....24

WSS Plot:....25

KMeans Clustering Insights:....26

Model Building....26

Train-Test Data Split....26

Understanding the data and metrics:....26

Linear Regression Model:....27

KNN Regressor Model:....32

Decision Tree Regressor Model:....32

Random Forest Regressor Model: 33

Gradient Boosting Regressor Model:....33

XGB Regressor Model: 34
Final Model Comparison 34
Performance of Regression Models: 34
Model Comparison: 35
Model Tuning: 35
Implications of Gradient Boosting 35
Conclusion: 36
Recommendations: 36
Implementation Plan: 37
Outcomes of the Implementation Plan: 38

List of all figures in the document:

Figure Sl. No.	Exploratory Data Analysis
1	Dataset (first 5) Sample
2	Dataset Info
3	Data Skewness
4	Descriptive Statistical data
5	Missing Values before imputation
6	Missing Values after imputation
7	Outlier before treatment
8	Outlier after treatment
9	Midpoint Treatment for Cholesterol
10	Location Count plot
11	Occupation Count plot
12	Gender Count plot
13	Correlation Plot
14	Negative Plot for 'Year_last_admitted'
15	Negative Plot for 'regular_checkup_lasy_year'
16	Negative Plot for 'weight_change_in_last_one_year'
17	Positive Plot for 'weight'
18	Positive Plot for 'adventure_sports'
19	Encoded Data Snippet
20	Final Treated Data Snippet
21	Sample Snippet of Eigenvalues
22	Sample Snippet of Eigenvectors
23	Scree Plot
24	PC Analysis Component Selected
25	WSS Plot
26	First Linear Regression Model
27	Normality Test
28	QQ Plot
29	Final Linear Regression Model

List of all figures in the document:

Table Sl. No.	Table Name
1	Data Dictionary
2	Shape of Test and Train data
3	Model Metrics

Introduction:

This report provides an in-depth exploratory data analysis (EDA) of the dataset and the performance of different Machine Learning Models on the dataset with the aim of gaining insights and understanding the relationships among variables and prediction of insurance cost.

The purpose of the EDA is to identify patterns, trends, and anomalies in the data that can inform decision-making while the machine learning models predict the insurance cost and detect the variables affecting the insurance cost. The dataset analyzed in this report is sourced from a healthcare insurance company and contains information about the insured individuals, including their demographics, medical history, and insurance costs.

Problem Statement:

The healthcare insurance company aims to identify the factors that influence insurance costs and develop strategies to optimize their pricing strategy. To achieve this, the company needs to understand the relationships among variables that affect insurance costs, such as age, gender, medical history, and lifestyle factors. Additionally, the company needs to identify the factors that contribute to customer churn, as this has a significant impact on their revenue and profitability.

Need of the study/project:

The study is necessary because medical treatment can be very expensive, and insurance coverage can help individuals in tough financial situations. However, insurance companies also want to reduce their risk and optimize the insurance cost. By building a model that can accurately estimate the insurance cost based on an individual's health and habit-related parameters, insurance companies can offer customized insurance plans that are tailored to the specific needs of each individual. This will reduce the financial risks for both individuals and insurance companies.

Furthermore, this study can help to promote healthy habits and lifestyles. By incentivizing healthy behaviors such as regular exercise, healthy eating, and not smoking, individuals can reduce their risk of getting ill and thereby reduce their insurance cost. This can help to improve public health and reduce the burden on the healthcare system.

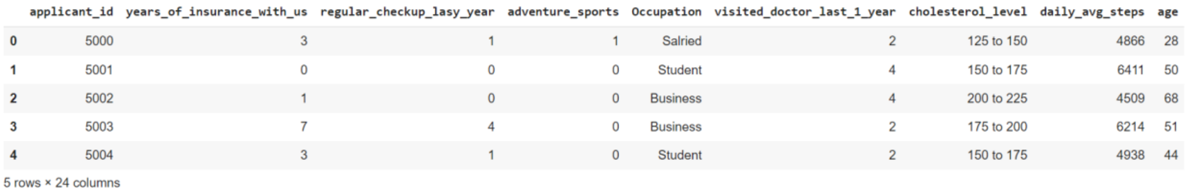
Exploratory Data Analysis

Dataset Description:

The dataset contains information about insured individuals, including demographic information such as age, gender, location, and occupation. It also contains information about their medical history, including cholesterol levels, BMI, and past medical conditions. Furthermore, the dataset contains information about the insured individuals' lifestyles, such as exercise habits, smoking status, and alcohol consumption. Finally, the dataset contains information about insurance costs and whether the insured individuals were covered by any other insurance company.

There are 25000 records (rows) and 24 variables (columns)

The first 5 rows of the dataset is as below:



	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age
0	5000	3	1	1	Salried	2	125 to 150	4866	28
1	5001	0	0	0	Student	4	150 to 175	6411	50
2	5002	1	0	0	Business	4	200 to 225	4509	68
3	5003	7	4	0	Business	2	175 to 200	6214	51
4	5004	3	1	0	Student	2	150 to 175	4938	44

5 rows × 24 columns

Fig 1. Dataset (first 5) Sample

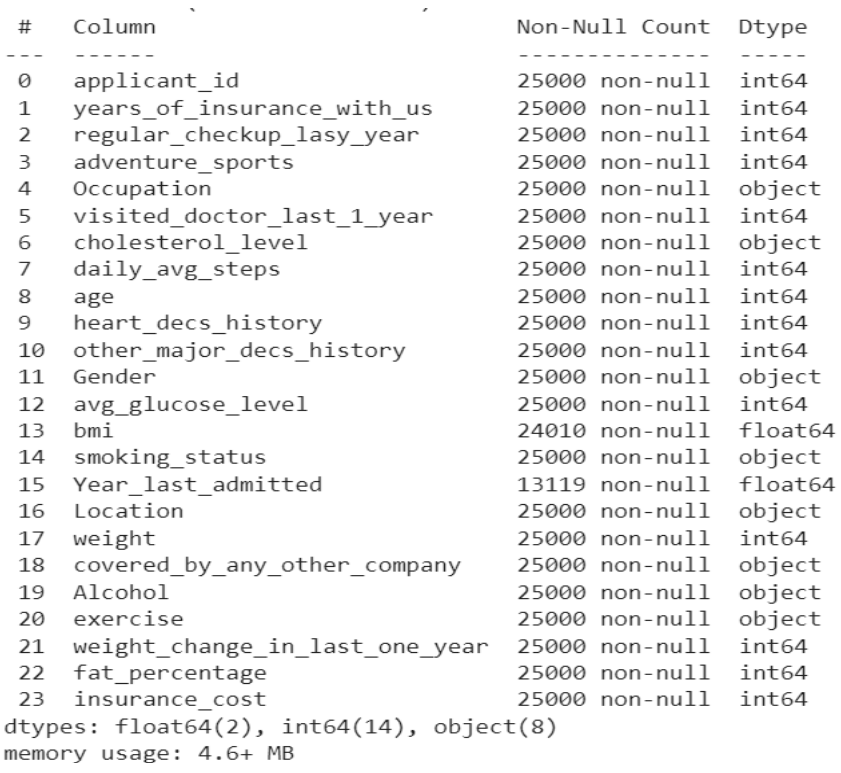
Data Dictionary:

Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check-up in last one year
adventure_sports	Customer is involved with adventure sports like climbing, diving etc.
Occupation	Occupation of the customer
visited_doctor_last_1_year	Number of times customer has visited doctor in last one year
cholesterol_level	Cholesterol level of the customers while applying for insurance
daily_avg_steps	Average daily steps walked by customers
age	Age of the customer
heart_decs_history	Any past heart diseases
other_major_decs_history	Any past major diseases apart from heart like any operation
Gender	Gender of the customer
avg_glucose_level	Average glucose level of the customer while applying the insurance
bmi	BMI of the customer while applying the insurance
smoking_status	Smoking status of the customer
Year_last_admitted	When customer have been admitted in the hospital last time
Location	Location of the hospital
weight	Weight of the customer
covered_by_any_other_company	Customer is covered from any other insurance company
Alcohol	Alcohol consumption status of the customer
exercise	Regular exercise status of the customer
weight_change_in_last_one_year	How much variation has been seen in the weight of the customer in last year
fat_percentage	Fat percentage of the customer while applying the insurance
insurance_cost	Total Insurance cost

Table 1. Data Dictionary

Data Information:

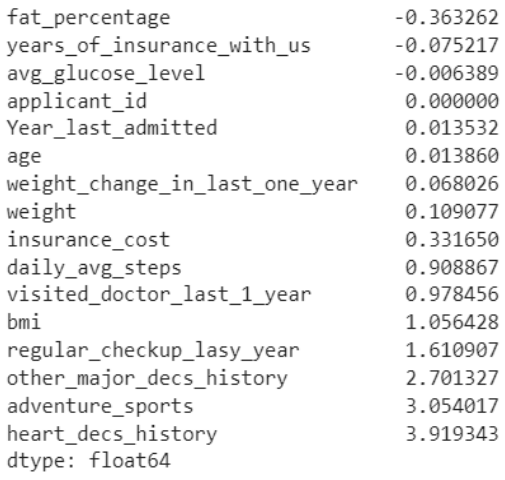
Out of 24 variables in the dataset, there are 2 float64, 14int64 and 8 object.



```
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    25000 non-null  int64
 1   years_of_insurance_with_us      25000 non-null  int64
 2   regular_checkup_lasy_year       25000 non-null  int64
 3   adventure_sports                25000 non-null  int64
 4   Occupation                      25000 non-null  object
 5   visited_doctor_last_1_year      25000 non-null  int64
 6   cholesterol_level               25000 non-null  object
 7   daily_avg_steps                 25000 non-null  int64
 8   age                             25000 non-null  int64
 9   heart_decs_history              25000 non-null  int64
 10  other_major_decs_history        25000 non-null  int64
 11  Gender                          25000 non-null  object
 12  avg_glucose_level               25000 non-null  int64
 13  bmi                             24010 non-null  float64
 14  smoking_status                  25000 non-null  object
 15  Year_last_admitted              13119 non-null  float64
 16  Location                        25000 non-null  object
 17  weight                          25000 non-null  int64
 18  covered_by_any_other_company    25000 non-null  object
 19  Alcohol                         25000 non-null  object
 20  exercise                        25000 non-null  object
 21  weight_change_in_last_one_year  25000 non-null  int64
 22  fat_percentage                  25000 non-null  int64
 23  insurance_cost                  25000 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

Fig 2. Dataset Info

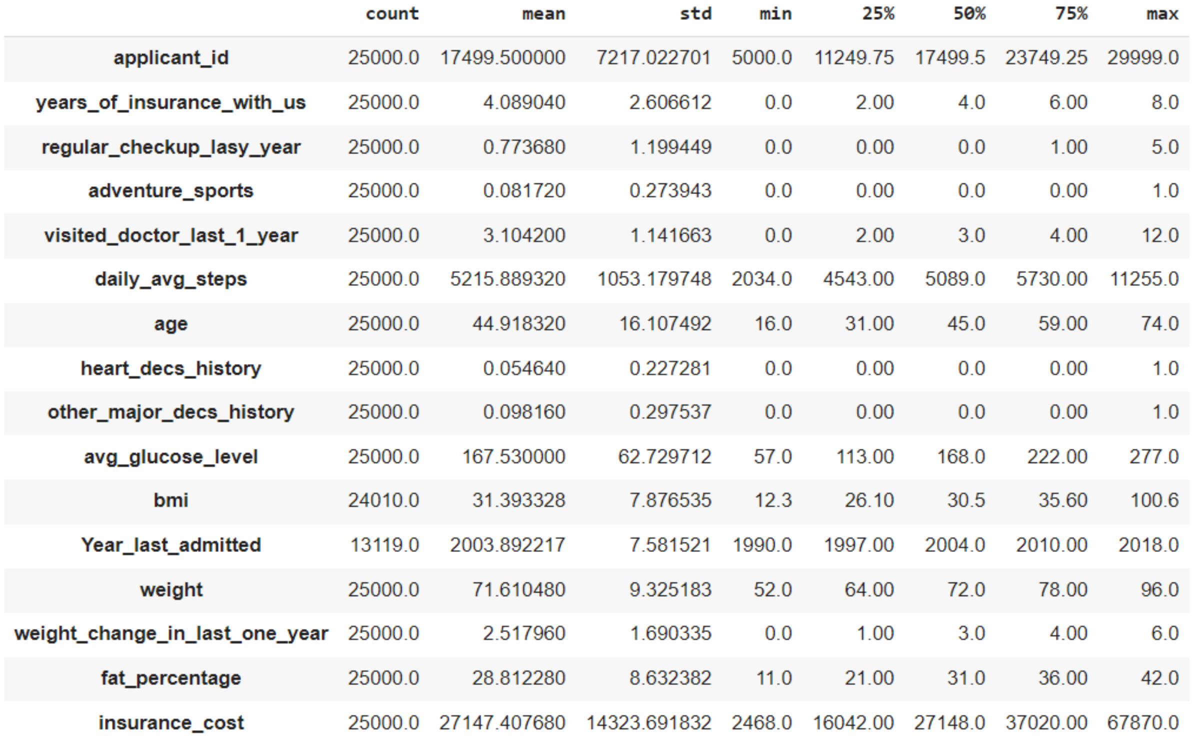
Data Skewness:



```
fat_percentage                   -0.363262
years_of_insurance_with_us       -0.075217
avg_glucose_level                -0.006389
applicant_id                      0.000000
Year_last_admitted                0.013532
age                               0.013860
weight_change_in_last_one_year    0.068026
weight                            0.109077
insurance_cost                    0.331650
daily_avg_steps                   0.908867
visited_doctor_last_1_year        0.978456
bmi                               1.056428
regular_checkup_lasy_year         1.610907
other_major_decs_history          2.701327
adventure_sports                  3.054017
heart_decs_history                3.919343
dtype: float64
```

Fig 3. Data Skewness

5-Point Summary or Descriptive Statistics of the data:



	count	mean	std	min	25%	50%	75%	max
applicant_id	25000.0	17499.500000	7217.022701	5000.0	11249.75	17499.5	23749.25	29999.0
years_of_insurance_with_us	25000.0	4.089040	2.606612	0.0	2.00	4.0	6.00	8.0
regular_checkup_lasy_year	25000.0	0.773680	1.199449	0.0	0.00	0.0	1.00	5.0
adventure_sports	25000.0	0.081720	0.273943	0.0	0.00	0.0	0.00	1.0
visited_doctor_last_1_year	25000.0	3.104200	1.141663	0.0	2.00	3.0	4.00	12.0
daily_avg_steps	25000.0	5215.889320	1053.179748	2034.0	4543.00	5089.0	5730.00	11255.0
age	25000.0	44.918320	16.107492	16.0	31.00	45.0	59.00	74.0
heart_decs_history	25000.0	0.054640	0.227281	0.0	0.00	0.0	0.00	1.0
other_major_decs_history	25000.0	0.098160	0.297537	0.0	0.00	0.0	0.00	1.0
avg_glucose_level	25000.0	167.530000	62.729712	57.0	113.00	168.0	222.00	277.0
bmi	24010.0	31.393328	7.876535	12.3	26.10	30.5	35.60	100.6
Year_last_admitted	13119.0	2003.892217	7.581521	1990.0	1997.00	2004.0	2010.00	2018.0
weight	25000.0	71.610480	9.325183	52.0	64.00	72.0	78.00	96.0
weight_change_in_last_one_year	25000.0	2.517960	1.690335	0.0	1.00	3.0	4.00	6.0
fat_percentage	25000.0	28.812280	8.632382	11.0	21.00	31.0	36.00	42.0
insurance_cost	25000.0	27147.407680	14323.691832	2468.0	16042.00	27148.0	37020.00	67870.0

Fig 4. Descriptive Statistical data

Considering the above descriptive statistics:

- The dataset includes data on 25,000 candidates.
- With a standard deviation of 2.61 years and a mean of 4.09 years, applicants have been insured with the company on average for that length of time.
- The applicant's average annual number of routine check-ups was 0.77, with a standard deviation of 1.20.
- Just 8.17% of those who applied have engaged in adventure sports.
- The average number of doctor visits made by applicants in the previous year was 3.10, with a standard deviation of 1.14.
- With a standard deviation of 1053, the applicants' average daily step count is 5215.
- The standard deviation of the applicant population's ages is 16.11, with an average age of 44.92.
- Just 5.46% of candidates had a history of cardiovascular illness.
- 9.82% of the applicants have a background in serious illnesses.
- The applicants' mean blood glucose level is 167.53 with a standard deviation of 62.73.
- The standard deviation of the applicants' BMI, which is 31.39, is 7.88.
- The standard deviation for the average year of last admittance is 7.58, and it is 2003.89 on average.
- With a standard deviation of 9.33 kg, candidates weigh on average 71.61 kg.
- The average weight change over the past year has been 2.52 kg, with a 1.69 kg standard deviation.
- With a standard deviation of 8.63%, applicants' average fat percentage is 28.81%.
- Insurers' average premiums are Rs.27,147.41, with a standard deviation of Rs.14,323.69

Inferences on Statistics:

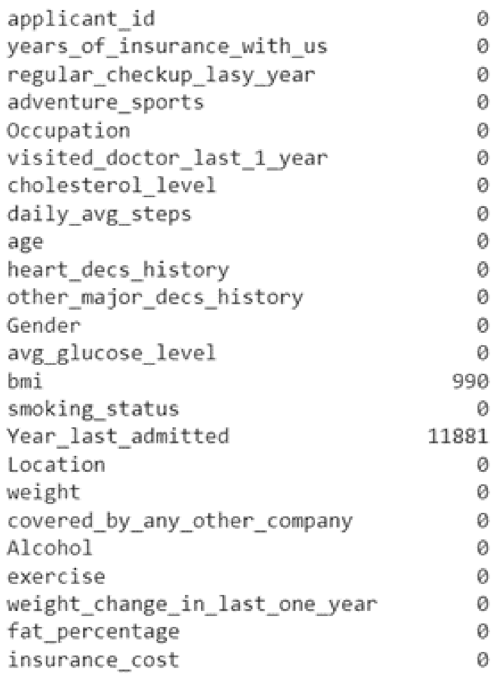
Descriptive statistics can be used to draw conclusions that can be utilised to guide decisions about product price, marketing, and underwriting, among other things. For instance, product pricing can be changed to better reflect the applicants' risk profile using information on applicants' ages and BMI. To reach applicants who are more concerned about their health, marketing efforts might be modified using information on regular checks and doctor visits. The information on participation in adventure sports and weight changes can be utilised to guide underwriting judgements and modify insurance pricing as necessary.

Data Cleaning and Preparation:

- The data cleaning and preparation process started by analyzing missing values, outliers to ensure data quality and consistency.
- For missing values, techniques such as imputation were used to fill in the missing values. In cases where there were good amount of missing values, 990 in bmi and 11881 in Year_last_admitted.
- Visualization techniques were used to identify outliers and anomalies, and statistical methods were used to remove them.
- Data quality and consistency were ensured by standardizing the data, correcting any data entry errors, and verifying the accuracy of the data.
- Overall, the data cleaning and preparation process was an essential step to ensure that the final dataset was ready for use in data analysis and modelling.

Missing Values:

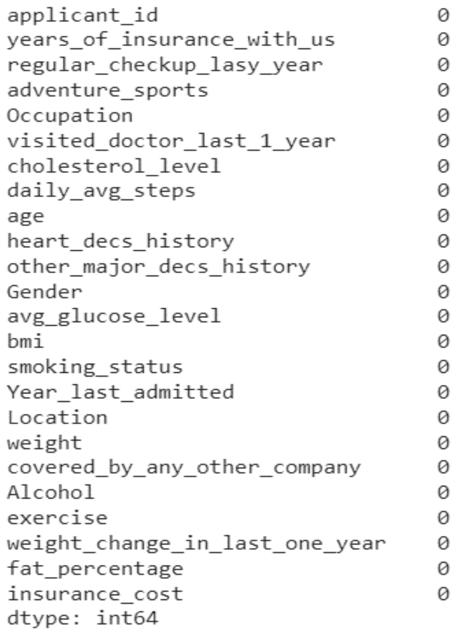
We identified missing values in the dataset by checking for null values in the columns. We used techniques such as imputation to fill in the missing values.



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

Fig 5. Missing Values before imputation

We imputed the missing values for ‘bmi’ and ‘Year_last_admitted’ using the mean and mode respectively.



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

Fig 6. Missing Values after imputation

Outliers:

We used visualization techniques such as box plots to identify the outliers in certain variables. We then used statistical methods i.e., IQR to remove the outliers.

The Interquartile Range (IQR) is a measure of statistical dispersion that describes the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset. The IQR is calculated by subtracting the first quartile from the third quartile and can also be used to create box plots that provide a visual representation of the spread of the data.

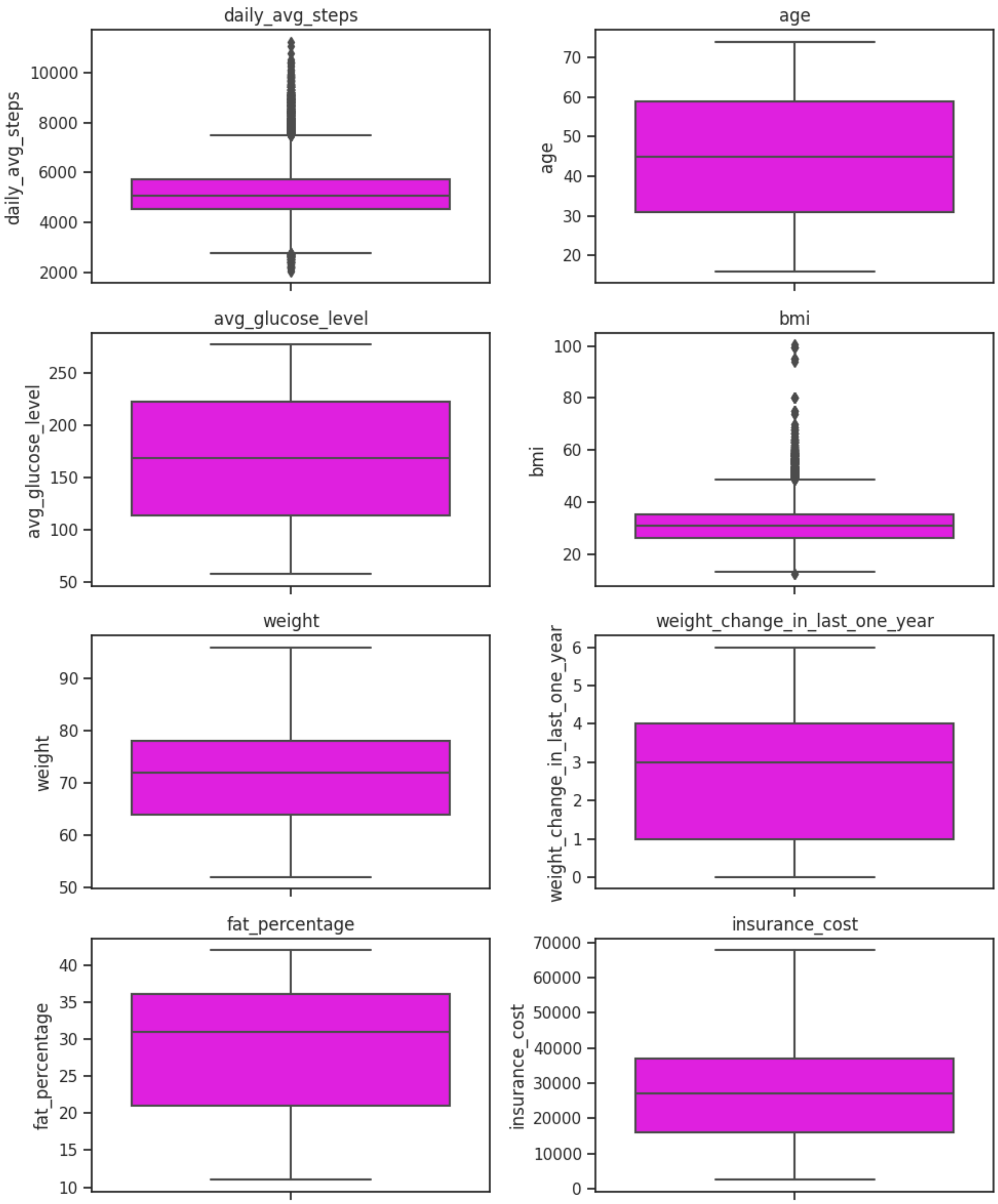


Fig 7. Outlier before treatment

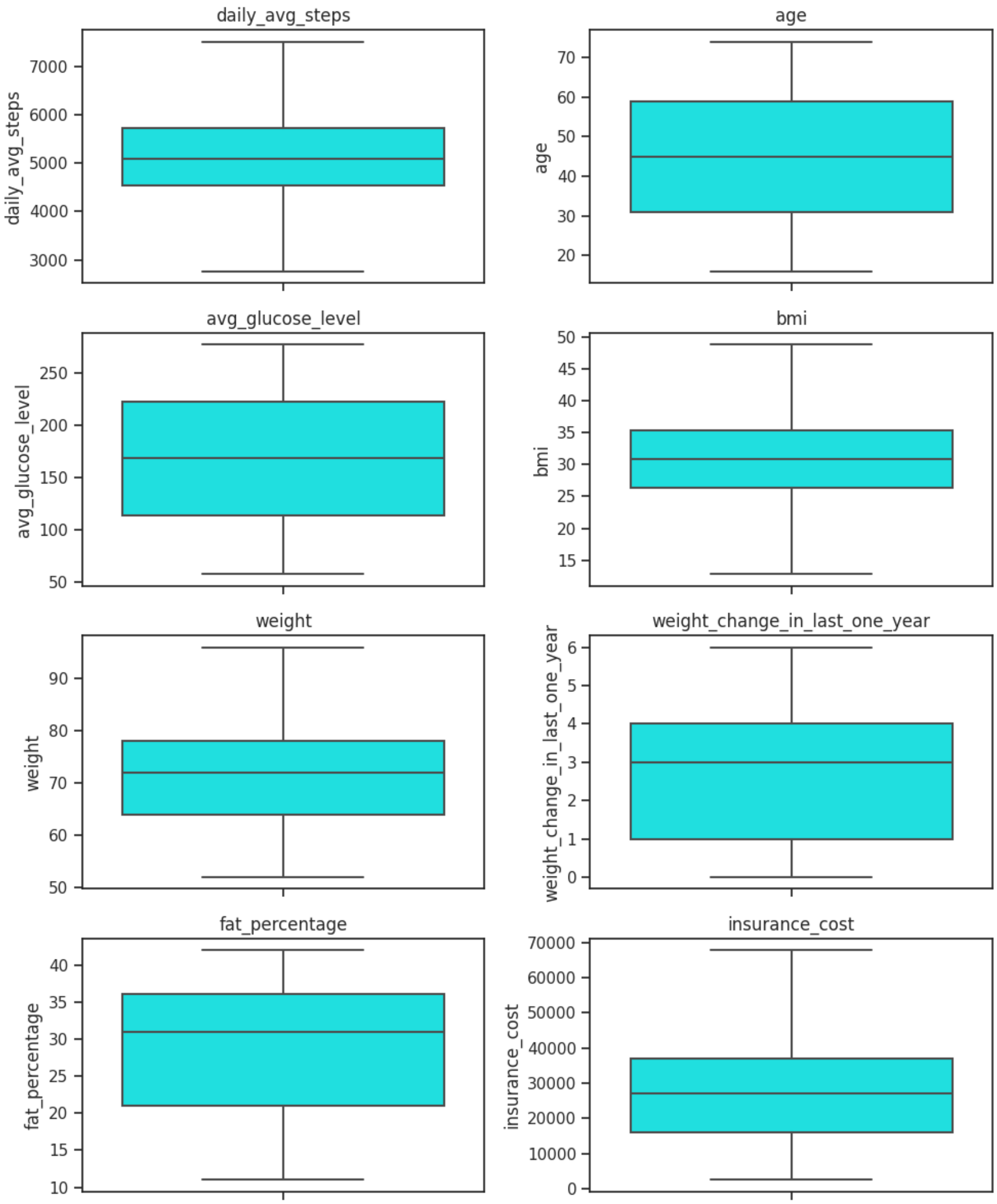


Fig 8. Outlier after treatment

Data Standardization:

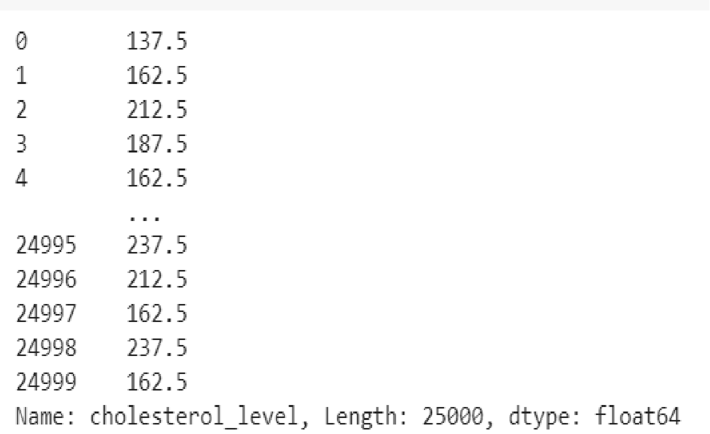
We standardized the data by converting all the necessary columns to a consistent format. In this dataset, we have used the MinMax Scaler to standardize or scale our data to get optimum results.

MinMax scaler is a data pre-processing technique used to transform numeric features by scaling them to a specified range, usually between 0 and 1.

Data Pre-Processing:

We are all set for the visual analysis of the data now that the cleaning and scaling is done. However, there is one more step to be performed but diving into visualizations.

The column 'cholesterol_level' has value like 125 to 150. This cannot be analysed correctly if not treated. So, it was decided that we take the midpoint of the given range.

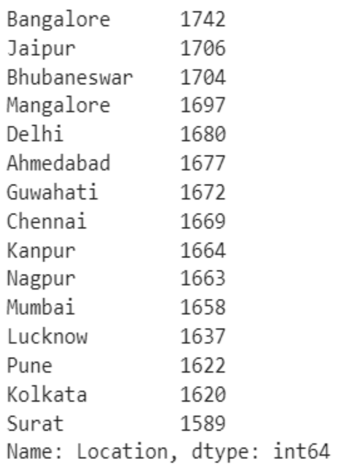


```
0        137.5
1        162.5
2        212.5
3        187.5
4        162.5
         ...
24995    237.5
24996    212.5
24997    162.5
24998    237.5
24999    162.5
Name: cholesterol_level, Length: 25000, dtype: float64
```

Fig 9. Midpoint Treatment for Cholesterol

Univariate Analysis:

1. Location Count Plot:



```
Bangalore      1742
Jaipur         1706
Bhubaneswar    1704
Mangalore      1697
Delhi          1680
Ahmedabad      1677
Guwahati       1672
Chennai        1669
Kanpur         1664
Nagpur         1663
Mumbai         1658
Lucknow        1637
Pune           1622
Kolkata        1620
Surat          1589
Name: Location, dtype: int64
```

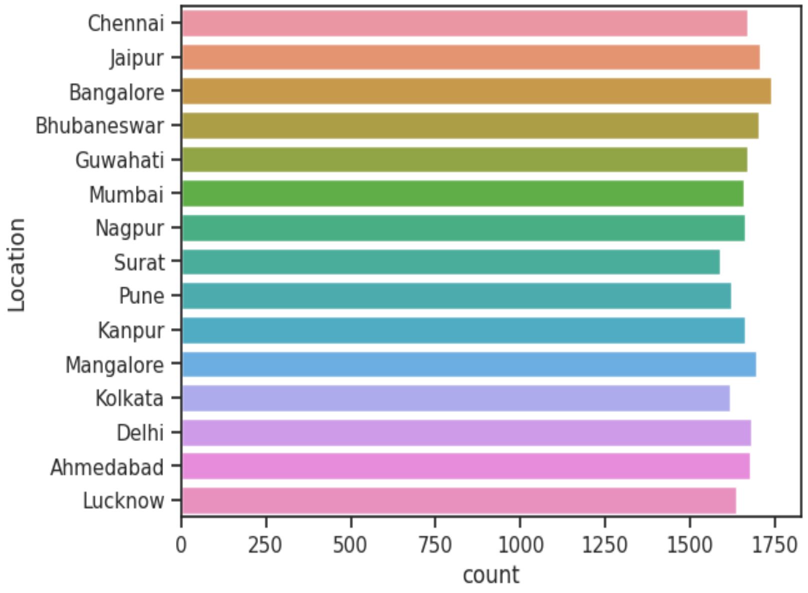


Fig 10. Location Count plot

The above data shows the number of insurances held in various locations in India. The data lists 15 cities along with the corresponding number of insurances held in each of those cities.

Business Inference on Location Count Plot:

- The highest number of insurances is held in Bangalore with a count of 1742, followed closely by Jaipur with 1706 insurances. The lowest count is in Surat with 1589 insurances.
- Focus on maintaining our existing customers in Bangalore and Jaipur, which have the highest number of insurances. Offer better policies to attract new customers in these cities.
- Investigate the reasons for the lower number of insurances in Surat and take measures to increase our market share in this region.

2. Occupation Plot:

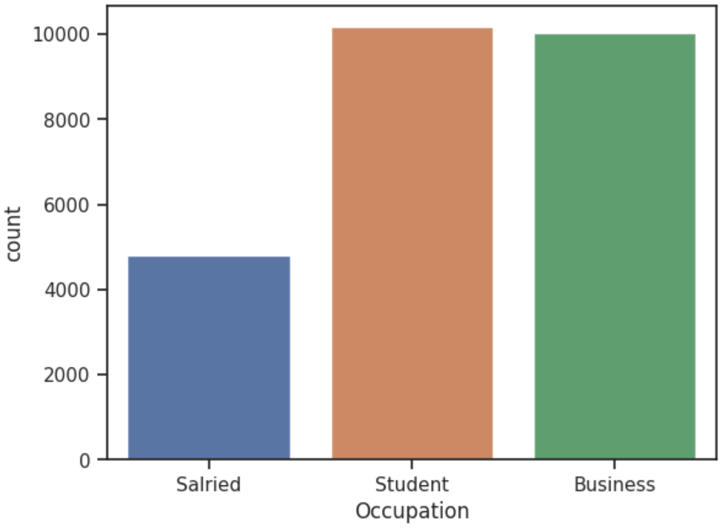


Fig 11. Occupation Count plot

Business Inference on Occupation Count Plot:

The Occupation Plot suggests that students and business owners are the major stakeholders of the insurances.

Based on the above analysis we can:

- Design policies that cater specifically to the needs of segments like Business owners and students to maintain the customers in these segments.
- Target salaried individuals by offering policies that are tailored to their specific needs. Offer discounts or other incentives to attract more customers from these groups.

3. Gender Plot:

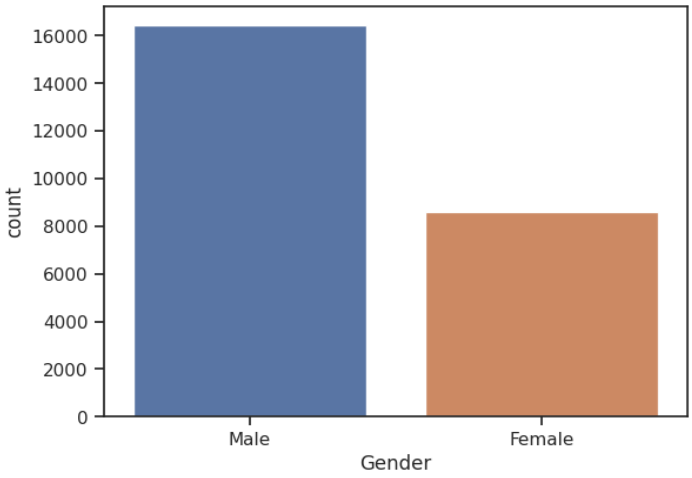


Fig 12. Gender Count plot

Business Inference on Gender Count Plot:

- From the above plot we can see that the major stake holders in opting for insurance are males.

- The count of males who opt for insurance is over 16000 which is double the number of females.
- Investigate the reasons and conduct a market analysis for this gender bias and take measures to attract more female customers.
- Design policies that cater specifically to the needs of women. Offer discounts and other incentives to attract more female customers.

Bivariate Plot:

- The correlated variables mentioned above are plotted against the insurance_cost variable with Occupation and Gender as filters.

Negative Trend or Correlation:

Negative trend or correlation refers to the relationship between two variables where an increase in the value of one variable is associated with a decrease in the value of the other variable. This means that as one variable increases, the other variable decreases, and vice versa.

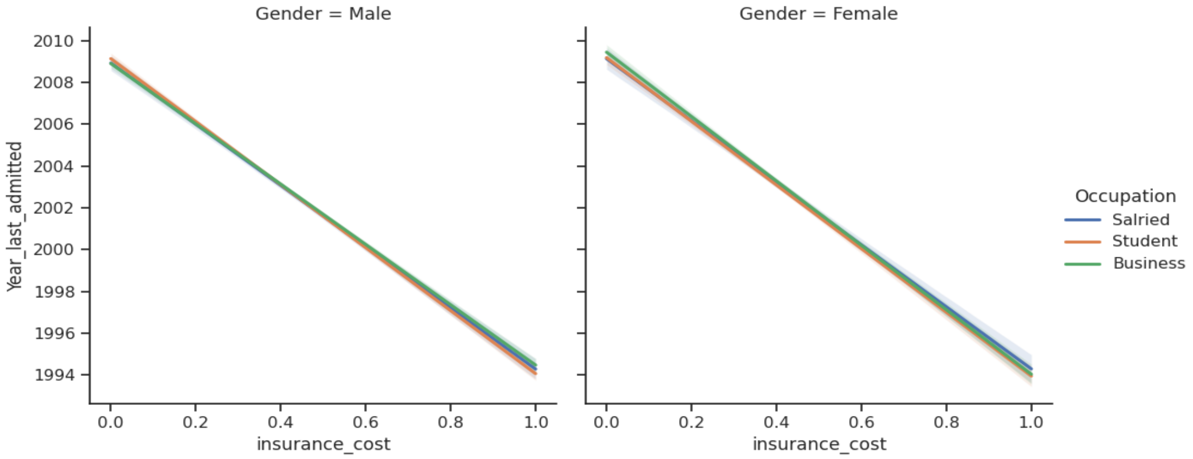


Fig 14. Negative Plot for 'Year_last_admitted'

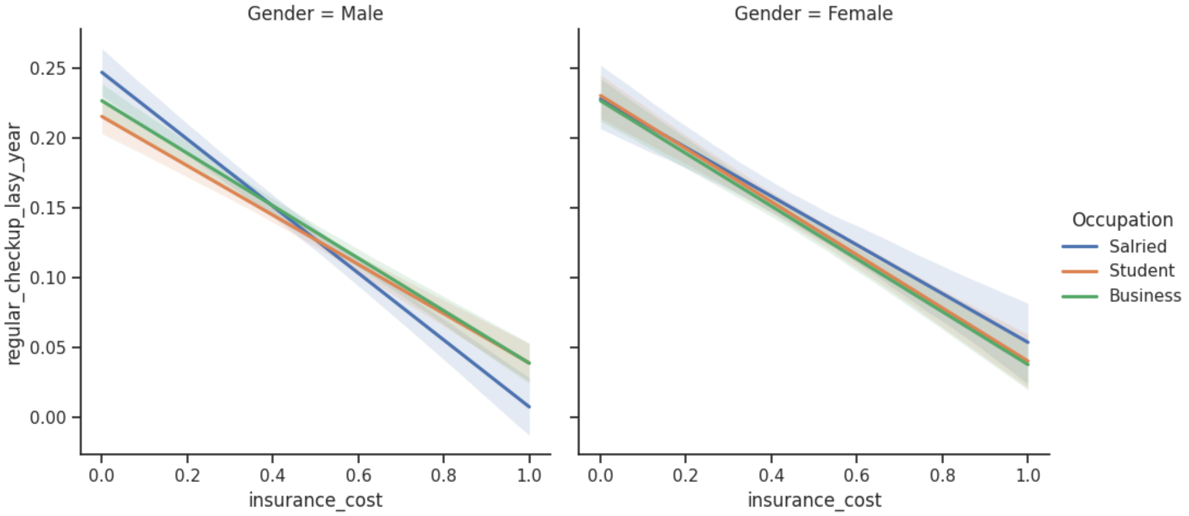


Fig 15. Negative Plot for 'regular_checkup_lasy_year'

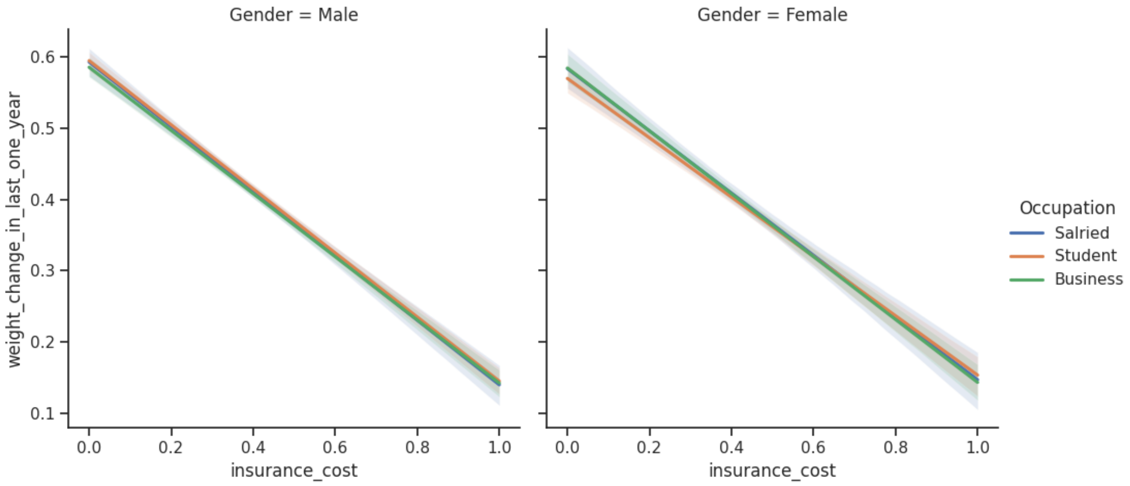


Fig 16. Negative Plot for ‘weight_change_in_last_one_year’

Business Inference on Negative Trend Analysis:

We can analyze the negatively correlated variables such as ‘Year_last_admitted’, ‘regular_checkup_lasy_year’, and ‘weight_change_in_last_one_year’ to identify underlying patterns. Specifically, we can:

- From the above analysis we see that the individuals who have a regular health check up and have a good weight loss or balance tend to opt out of the insurance.
- Though it is understandable reason, we need to target these individuals and offer them specifically tailored policies and conduct a market analysis to understand their needs in term of insurance.
- Offer better value to our customers.

Positive Trend or Correlation:

A positive trend or correlation refers to a pattern where the values of two variables tend to increase or decrease together. In other words, if the value of one variable increases, the value of the other variable also tends to increase. This pattern is known as a positive correlation.

A positive correlation between two variables suggests that they are related and may influence each other.

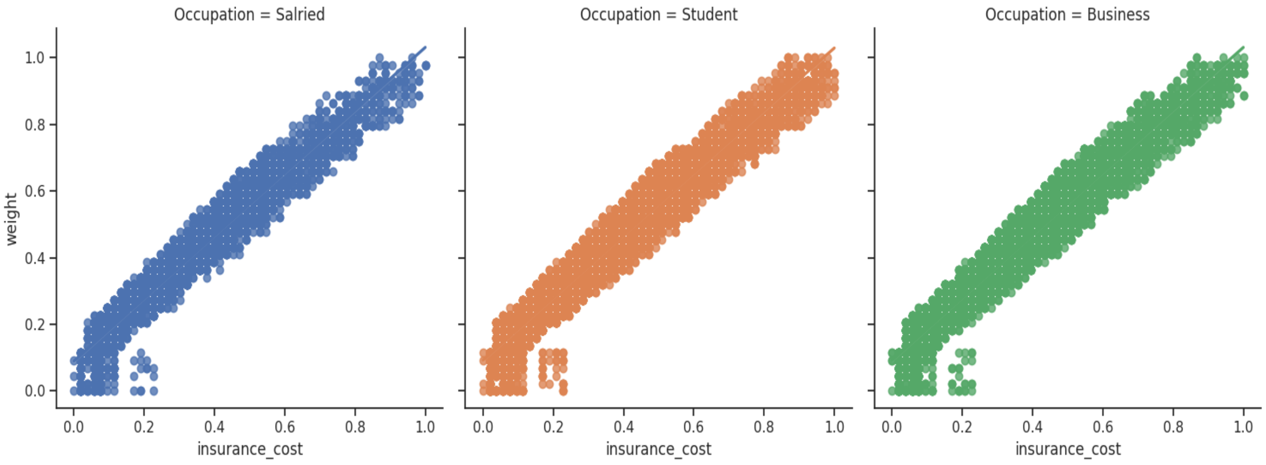


Fig 17. Positive Plot for 'weight'

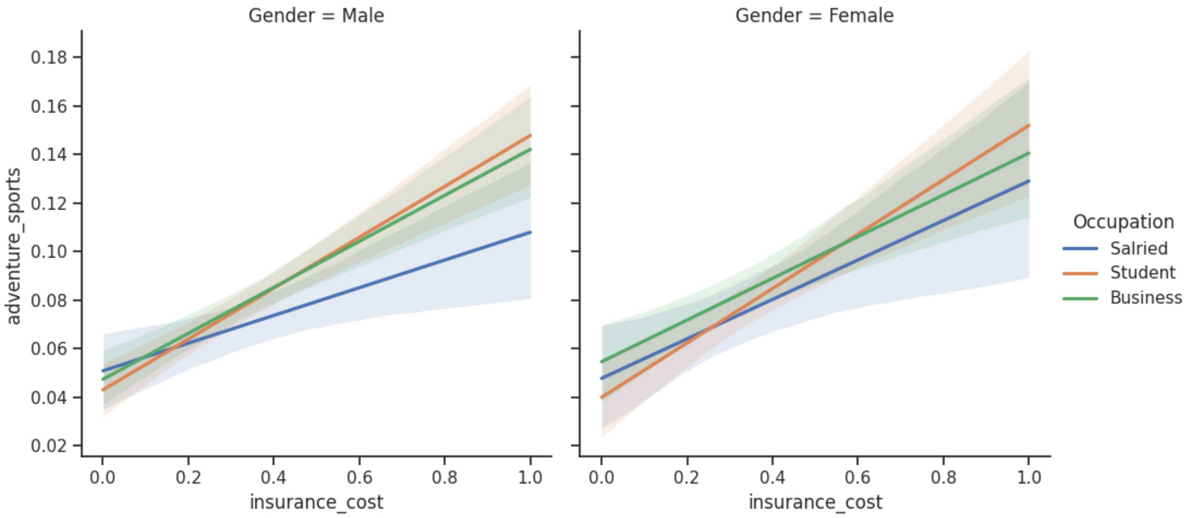


Fig 18. Positive Plot for 'adventure_sports'

Business Inference on Positive Trend Analysis:

Finally, we can identify the variables with a positive correlation with the target variable such as 'weight' and 'adventure_sports'. We can:

- Design policies that cater specifically to customers interested in adventure sports.
- Offer discounts or other incentives to attract customers who are health-conscious and interested in maintaining a healthy weight.

Correlation Plot:

- While analyzing the variables from the above plot, most of them have no relation to the target variable 'insurance_cost'.
- The negatively correlated variables are 'Year_last_admitted' , 'regular_checkup_lasy_year', 'weight_change_in_last_one_year'
- The variables 'weight' and 'adventure_sports' have a positive correlation with the target variable.

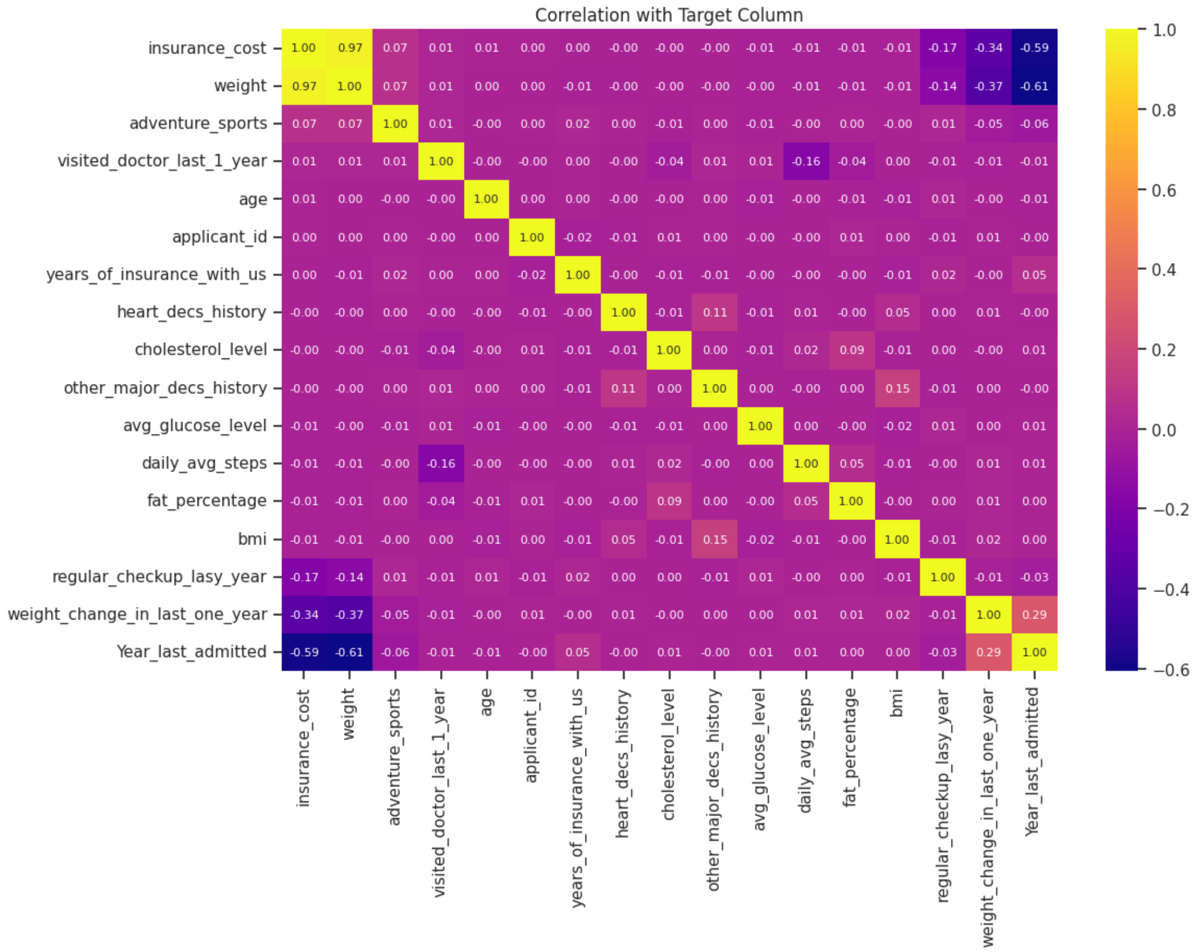


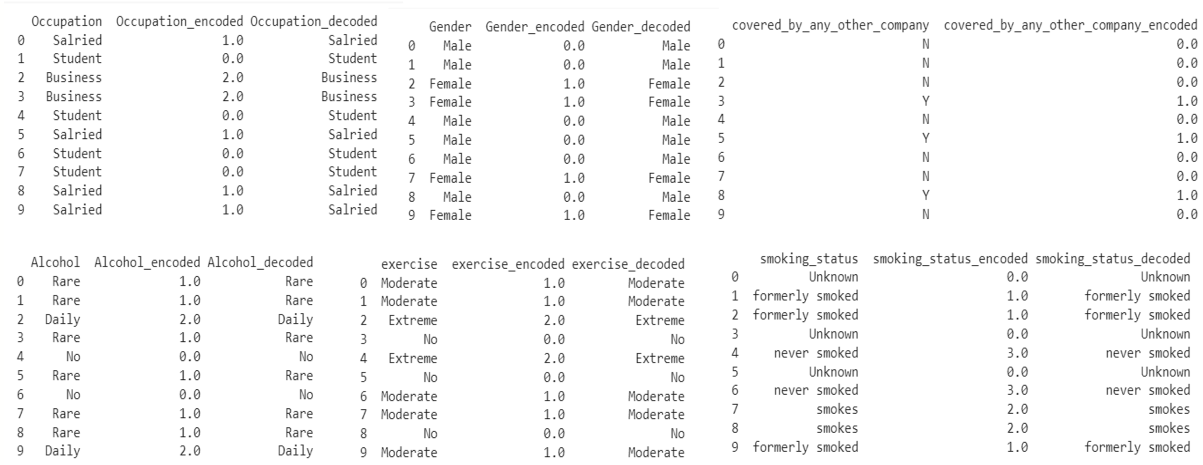
Fig 13. Correlation Plot

Data Encoding:

We did not encode the data for prior analysis just for the easiness of identifying the categorical values. Moving forward we will have to encode the object variables present in the dataset for the proper functioning of Model that are about to build. The below mentioned columns were encoded using Ordinal Encoder:

- Alcohol
- exercise
- smoking_status
- covered_by_any_other_company
- Gender
- Occupation

The below is the snippet of data samples that were encoded.



```
  Occupation  Occupation_encoded Occupation_decoded
0    Salried                 1.0            Salried
1    Student                 0.0            Student
2   Business                 2.0           Business
3   Business                 2.0           Business
4    Student                 0.0            Student
5    Salried                 1.0            Salried
6    Student                 0.0            Student
7    Student                 0.0            Student
8    Salried                 1.0            Salried
9    Salried                 1.0            Salried
```

```
   Gender  Gender_encoded Gender_decoded
0    Male             0.0           Male
1    Male             0.0           Male
2  Female             1.0         Female
3  Female             1.0         Female
4    Male             0.0           Male
5    Male             0.0           Male
6    Male             0.0           Male
7  Female             1.0         Female
8    Male             0.0           Male
9  Female             1.0         Female
```

```
  covered_by_any_other_company  covered_by_any_other_company_encoded
0                            N                                   0.0
1                            N                                   0.0
2                            N                                   0.0
3                            Y                                   1.0
4                            N                                   0.0
5                            Y                                   1.0
6                            N                                   0.0
7                            N                                   0.0
8                            Y                                   1.0
9                            N                                   0.0
```

```
  Alcohol  Alcohol_encoded Alcohol_decoded
0    Rare              1.0            Rare
1    Rare              1.0            Rare
2   Daily              2.0           Daily
3    Rare              1.0            Rare
4      No              0.0              No
5    Rare              1.0            Rare
6      No              0.0              No
7    Rare              1.0            Rare
8    Rare              1.0            Rare
9   Daily              2.0           Daily
```

```
   exercise  exercise_encoded exercise_decoded
0  Moderate               1.0         Moderate
1  Moderate               1.0         Moderate
2   Extreme               2.0          Extreme
3        No               0.0               No
4   Extreme               2.0          Extreme
5        No               0.0               No
6  Moderate               1.0         Moderate
7  Moderate               1.0         Moderate
8        No               0.0               No
9  Moderate               1.0         Moderate
```

```
    smoking_status  smoking_status_encoded smoking_status_decoded
0          Unknown                     0.0                Unknown
1  formerly smoked                     1.0        formerly smoked
2  formerly smoked                     1.0        formerly smoked
3          Unknown                     0.0                Unknown
4     never smoked                     3.0           never smoked
5          Unknown                     0.0                Unknown
6     never smoked                     3.0           never smoked
7           smokes                     2.0                 smokes
8           smokes                     2.0                 smokes
9  formerly smoked                     1.0        formerly smoked
```

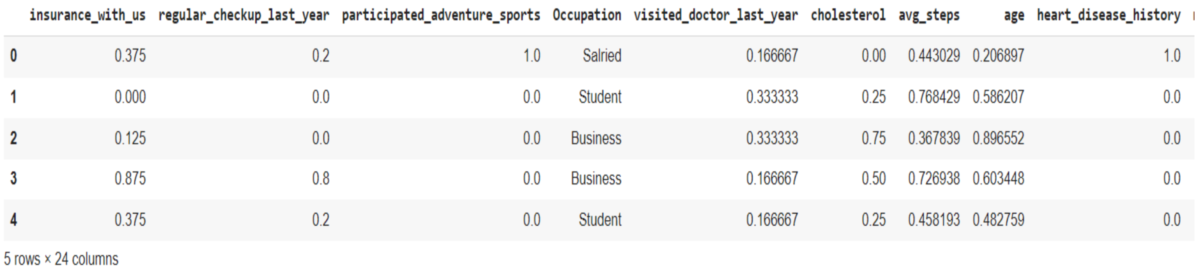
Fig 19. Encoded Data Snippet

Renaming and dropping unnecessary columns for Model Building:

When building a predictive model, it is important to ensure that the data is in the right format and contains the necessary features to make accurate predictions. This includes renaming columns to be more descriptive, dropping unnecessary columns that do not contribute to the model, and ensuring that all columns are in the appropriate format for analysis.

Dropping unnecessary columns involves removing columns that do not contribute to the model. This could include columns that are not relevant to the problem being solved, or columns that contain duplicate or highly correlated information. By removing unnecessary columns, the data becomes more focused and easier to work with.

In summary, renaming and dropping unnecessary columns are important steps in preparing data for model building. These steps help ensure that the data is in the right format and contains the necessary features to make accurate predictions.



	insurance_with_us	regular_checkup_last_year	participated_adventure_sports	Occupation	visited_doctor_last_year	cholesterol	avg_steps	age	heart_disease_history
0	0.375	0.2	1.0	Salried	0.166667	0.00	0.443029	0.206897	1.0
1	0.000	0.0	0.0	Student	0.333333	0.25	0.768429	0.586207	0.0
2	0.125	0.0	0.0	Business	0.333333	0.75	0.367839	0.896552	0.0
3	0.875	0.8	0.0	Business	0.166667	0.50	0.726938	0.603448	0.0
4	0.375	0.2	0.0	Student	0.166667	0.25	0.458193	0.482759	0.0

5 rows × 24 columns

Fig 20. Final Treated Data Snippet

Principal Component Analysis

Principal Component Analysis (PCA) is a statistical technique used for dimensionality reduction and data exploration. It helps in identifying patterns and relationships in high-dimensional datasets by transforming the data into a lower-dimensional space.

Steps Involved in PCA Analysis:

1. Standardization of Data
2. Generate Covariance Matrix for all the dimensions.
3. Perform Eigen Decomposition
4. Select the components based on the above decomposition.

Before performing PCA it is important that we check the correlation and the significance of the correlation. Further we also check the adequacy of the sample size.

Significance of Correlation:

First, we plot a heat map using Seaborn Library. Next up is the analysis of the existing significance using factor analyser library's calculate_bartlett_sphericity.

Bartlett's test of sphericity is a statistical test used in factor analysis to determine whether the correlation matrix of the observed variables is significantly different from the identity matrix, indicating whether the variables are suitable for factor analysis.

The p-value obtained for Bartlett's test is 0.0 which is lesser than 0.05 meaning, there is a significant amount of correlation within the data.

```
p-value: 0.0
```

Confirmation of Adequacy of Sample Size:

The adequacy confirmation is done by factor analyser library's calculate_kmo. If this value is equal to or greater than 0.7 the sample size is deemed adequate for PC Analysis.

The Kaiser-Meyer-Olkin (KMO) test is another measure used in factor analysis to assess the suitability of data for factor analysis. It evaluates the sampling adequacy for each variable and the overall adequacy of the correlation matrix.

The value from kmo test is 0.632 which is remarkably close to 0.7 to confirm the adequacy of the sample set.

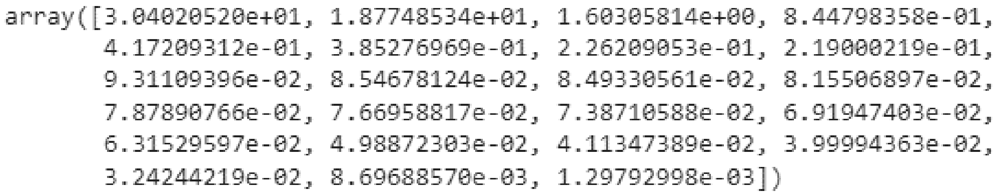
PC Analysis:

Now that we have tested the correlation and adequacy, we build the PCA model using sklearn decomposition library's PCA.

Eigenvalues and eigenvectors are concepts used in linear algebra to analyze linear transformations and matrices. In the context of PCA (Principal Component Analysis), eigenvalues and eigenvectors play a crucial role in determining the principal components of a dataset.

Eigenvalues:

Eigenvalues represent the scaling factors by which the eigenvectors are stretched or compressed when transformed by a given matrix. They provide information about the amount of variance explained by each eigenvector (principal component) in the dataset. Larger eigenvalues indicate that the corresponding eigenvectors capture more variance in the data.

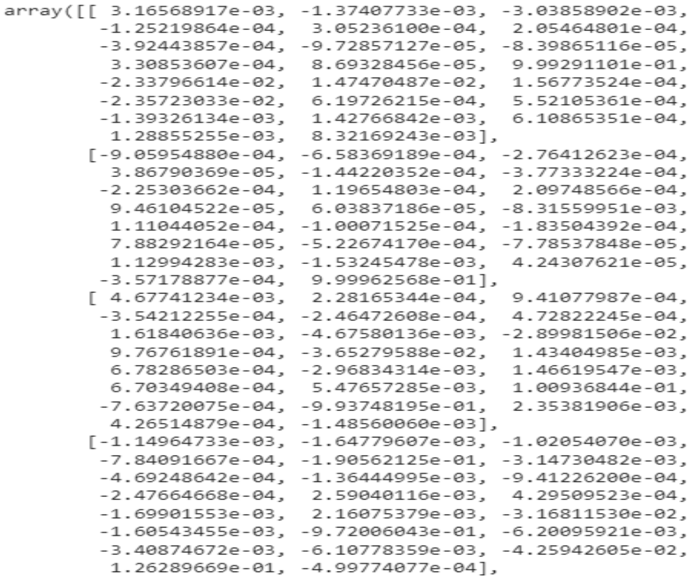


```
array([3.04020520e+01, 1.87748534e+01, 1.60305814e+00, 8.44798358e-01,
       4.17209312e-01, 3.85276969e-01, 2.26209053e-01, 2.19000219e-01,
       9.31109396e-02, 8.54678124e-02, 8.49330561e-02, 8.15506897e-02,
       7.87890766e-02, 7.66958817e-02, 7.38710588e-02, 6.91947403e-02,
       6.31529597e-02, 4.98872303e-02, 4.11347389e-02, 3.99994363e-02,
       3.24244219e-02, 8.69688570e-03, 1.29792998e-03])
```

Fig 21. Sample Snippet of Eigenvalues

Eigenvectors:

Eigenvectors are the non-zero vectors that, when transformed by a given matrix, retain their direction but may be scaled. Each eigenvector corresponds to an eigenvalue and represents a principal component. Eigenvectors are orthogonal to each other, meaning they are linearly independent and perpendicular.



```
array([[ 3.16568917e-03, -1.37407733e-03, -3.03858902e-03,
        -1.25219864e-04,  3.05236100e-04,  2.05464801e-04,
        -3.92443857e-04, -9.72857127e-05, -8.39865116e-05,
         3.30853607e-04,  8.69328456e-05,  9.99291101e-01,
        -2.33796614e-02,  1.47470487e-02,  1.56773524e-04,
        -2.35723033e-02,  6.19726215e-04,  5.52105361e-04,
        -1.39326134e-03,  1.42766842e-03,  6.10865351e-04,
         1.28855255e-03,  8.32169243e-03],
       [-9.05954880e-04, -6.58369189e-04, -2.76412623e-04,
         3.86790369e-05, -1.44220352e-04, -3.77333224e-04,
        -2.25303662e-04,  1.19654803e-04,  2.09748566e-04,
         9.46104522e-05,  6.03837186e-05, -8.31559951e-03,
         1.11044052e-04, -1.00071525e-04, -1.83504392e-04,
         7.88292164e-05, -5.22674170e-04, -7.78537848e-05,
         1.12994283e-03, -1.53245478e-03,  4.24307621e-05,
        -3.57178877e-04,  9.99962568e-01],
       [ 4.67741234e-03,  2.28165344e-04,  9.41077987e-04,
        -3.54212255e-04, -2.46472608e-04,  4.72822245e-04,
         1.61840636e-03, -4.67580136e-03, -2.89981506e-02,
         9.76761891e-04, -3.65279588e-02,  1.43404985e-03,
         6.78286503e-04, -2.96834314e-03,  1.46619547e-03,
         6.70349408e-04,  5.47657285e-03,  1.00936844e-01,
        -7.63720075e-04, -9.93748195e-01,  2.35381906e-03,
         4.26514879e-04, -1.48560060e-03],
       [-1.14964733e-03, -1.64779607e-03, -1.02054070e-03,
        -7.84091667e-04, -1.90562125e-01, -3.14730482e-03,
        -4.69248642e-04, -1.36444995e-03, -9.41226200e-04,
        -2.47664668e-04,  2.59040116e-03,  4.29509523e-04,
        -1.69901553e-03,  2.16075379e-03, -3.16811530e-02,
        -1.60543455e-03, -9.72006043e-01, -6.20095921e-03,
        -3.40874672e-03, -6.10778359e-03, -4.25942605e-02,
         1.26289669e-01, -4.99774077e-04],
```

Fig 22. Sample Snippet of Eigenvectors

Scree Plot:

Once we the above values, we form a data frame with these values and plot a scree plot to select the prominent features.

From the below plot, we could clearly see that the first 4 Principal Components are quite important compared to others. The next 2 components show minor significance. So, it was decided to select first 6 PCs.

However, the insights were limited to only 4, as the next 2 components would need understanding the data more that was done during KMeans Clustering and Model Building.

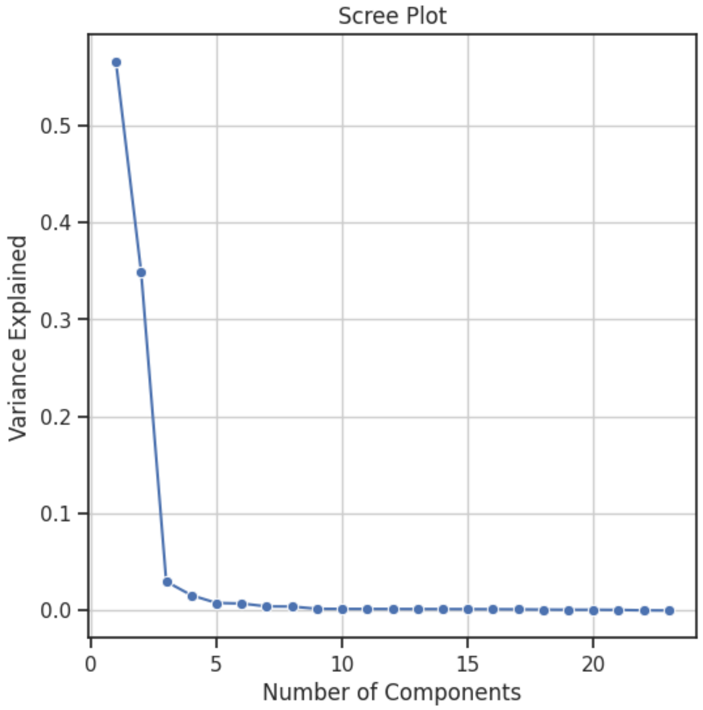
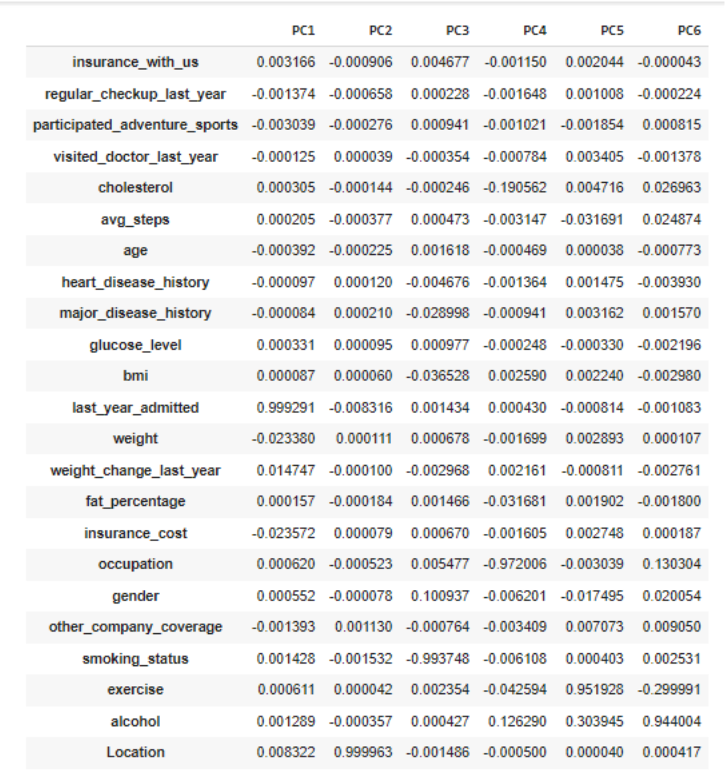


Fig 23. Scree Plot



	PC1	PC2	PC3	PC4	PC5	PC6
insurance_with_us	0.003166	-0.000906	0.004677	-0.001150	0.002044	-0.000043
regular_checkup_last_year	-0.001374	-0.000658	0.000228	-0.001648	0.001008	-0.000224
participated_adventure_sports	-0.003039	-0.000276	0.000941	-0.001021	-0.001854	0.000815
visited_doctor_last_year	-0.000125	0.000039	-0.000354	-0.000784	0.003405	-0.001378
cholesterol	0.000305	-0.000144	-0.000246	-0.190562	0.004716	0.026963
avg_steps	0.000205	-0.000377	0.000473	-0.003147	-0.031691	0.024874
age	-0.000392	-0.000225	0.001618	-0.000469	0.000038	-0.000773
heart_disease_history	-0.000097	0.000120	-0.004676	-0.001364	0.001475	-0.003930
major_disease_history	-0.000084	0.000210	-0.028998	-0.000941	0.003162	0.001570
glucose_level	0.000331	0.000095	0.000977	-0.000248	-0.000330	-0.002196
bmi	0.000087	0.000060	-0.036528	0.002590	0.002240	-0.002980
last_year_admitted	0.999291	-0.008316	0.001434	0.000430	-0.000814	-0.001083
weight	-0.023380	0.000111	0.000678	-0.001699	0.002893	0.000107
weight_change_last_year	0.014747	-0.000100	-0.002968	0.002161	-0.000811	-0.002761
fat_percentage	0.000157	-0.000184	0.001466	-0.031681	0.001902	-0.001800
insurance_cost	-0.023572	0.000079	0.000670	-0.001605	0.002748	0.000187
occupation	0.000620	-0.000523	0.005477	-0.972006	-0.003039	0.130304
gender	0.000552	-0.000078	0.100937	-0.006201	-0.017495	0.020054
other_company_coverage	-0.001393	0.001130	-0.000764	-0.003409	0.007073	0.009050
smoking_status	0.001428	-0.001532	-0.993748	-0.006108	0.000403	0.002531
exercise	0.000611	0.000042	0.002354	-0.042594	0.951928	-0.299991
alcohol	0.001289	-0.000357	0.000427	0.126290	0.303945	0.944004
Location	0.008322	0.999963	-0.001486	-0.000500	0.000040	0.000417

Fig 24. PC Analysis Component Selected

Principal Component Analysis (PCA) Insights:

The dataset has been reduced to four principal components (PC1, PC2, PC3, and PC4) using PCA. Let's explore the insights from each component:

PC1:

- PC1 shows a high negative loading for the first feature and a moderate negative loading for the second feature. This suggests that PC1 captures the variation in these two features, indicating a potential relationship between them.

- The scores for PC1 range from -0.53 to 3.51. Higher positive scores indicate a higher contribution of the first feature, while lower negative scores indicate a higher contribution of the second feature.

PC2:

- PC2 exhibits a strong negative loading for the third feature and a moderate negative loading for the fourth feature. This implies that PC2 captures the variation in these two features, potentially representing a different underlying relationship.

- The scores for PC2 range from -6.96 to -0.96. Higher negative scores indicate a higher contribution of the third feature, while lower negative scores indicate a higher contribution of the fourth feature.

PC3 and PC4:

- PC3 and PC4 show a mix of positive and negative loadings for distinctive features, indicating a more complex relationship and variation captured by these components.

- The scores for PC3 and PC4 represent the contribution of the corresponding features to the overall dataset, with higher absolute scores indicating a stronger influence.

KMeans Clustering

Clustering:

Clustering is a technique in unsupervised machine learning that aims to group similar data points together based on their features or characteristics. It helps to discover patterns, similarities, and structures in the data without any predefined labels or target variables.

KMeans Clustering:

K-means clustering is one of the most commonly used clustering algorithms. It partitions the data into a specified number of clusters (K) based on the distance between data points. The algorithm iteratively assigns data points to the nearest centroid (mean) of a cluster and updates the centroids until convergence.

K-means clustering can be used for various applications such as customer segmentation, image compression, anomaly detection, and more. It is important to note that the quality of the clustering results can be influenced by the choice of the number of clusters (K) and the initial placement of centroids. It is often necessary to experiment with different K values and initialization strategies to find the most appropriate clustering solution.

We perform KMeans clustering using sklearn.cluster library's KMeans. The KMeans Clustering is done by interpreting the number of clusters by using WSS plot.

WSS Plot:

The Within-Cluster Sum of Squares (WSS) plot is a graphical tool used to determine the optimal number of clusters in K-means clustering. It helps in selecting the appropriate number of clusters by analyzing the trade-off between the number of clusters and the WSS.

The WSS measures the sum of the squared distances between each data point and its assigned centroid within a cluster. A lower WSS indicates that the data points within each cluster are closer to their respective centroids, suggesting better clustering.

To create a WSS plot, you perform K-means clustering for a range of different cluster numbers (K) and calculate the WSS for each K. Then, you plot the number of clusters (K) on the x-axis and the corresponding WSS on the y-axis. The plot will typically show a decreasing trend of WSS as the number of clusters increases.

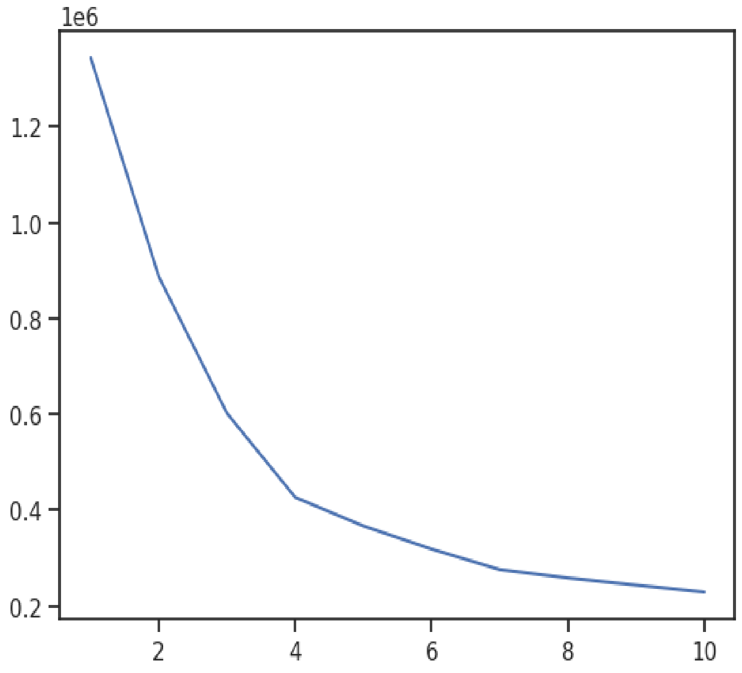


Fig 25. WSS Plot

From the above WSS Plot, we conclude that the number of clusters to be used in the KMeans is 4. The important metrics we need to understand in KMeans are:

Silhouette Score:

The silhouette score is a single value that represents the average silhouette width across all data points. It provides an overall assessment of the clustering quality, where higher values indicate better clustering.

Silhouette Width:

Silhouette width, also known as silhouette coefficient, measures the average distance between each data point and other points in the same cluster compared to the nearest cluster. It ranges from -1 to 1, where a higher value indicates better-defined and well-separated clusters.

KMeans Clustering Insights:

The clustering analysis resulted in the following clusters:

Cluster 0: This cluster of applicants with low years of insurance, regular check-ups, average cholesterol levels, moderate daily steps, no history of heart disease, different smoking statuses, different weights, alcohol consumption, exercise frequency, and variation in fat percentage and insurance cost.

Cluster 1: This cluster of applicants with moderate years of insurance, regular check-ups, diverse occupations, cholesterol levels ranging from moderate to high, and average daily average steps. They are predominantly male, have higher average glucose levels and higher BMI, and are located in different cities.

Cluster 2: This cluster represents applicants with low years of insurance, no regular check-ups, and high cholesterol levels. They are mostly males, have higher average glucose levels and higher BMI, and are not covered by any other insurance company. They have experienced weight change in the last year.

Cluster 3: This cluster comprises applicants with high years of insurance, regular check-ups, diverse occupations, low cholesterol levels, low daily average steps, higher average glucose levels, higher BMI, smokers, and moderate alcohol consumption. They have experienced weight change in the last year.

Model Building

Train-Test Data Split

Separating independent (train) and dependent (test)variables for the linear regression model.

X = independent variables

Y = dependent variable

Shape of Test and Train data:

The training set for the independent variables	(20000, 23)
The training set for the dependent variable	(20000, 1)
The test set for the independent variables	(5000, 23)
The test set for the dependent variable	(5000, 1)

Table 2. Shape of Test and Train data

Understanding the data and metrics:

According to our problem statement, the target variable is "insurance_cost" which is a continuous variable. We cannot use classifier algorithms for the continuous variables instead we build various Regression Model to find the optimal performing model to identify the influencing factors and predict the insurance cost.

- It is not possible to build a classification model for the given dataset like Logistic Regression, LDA, CART, Random Forest Classifier etc.

- So, we build Regressor Models on the dataset with metrics such as MSE, RMSE, MAE and R-squared Scores.
- In Regression Models, accuracy cannot be directly calculated. Hence, we use the above metrics to evaluate the performance of the models.

Mean Squared Error (MSE):

- MSE measures the average squared difference between the predicted values and the actual values.
- It is calculated by taking the average of the squared residuals.
- A lower MSE indicates better model performance, with a value of 0 indicating a perfect fit.

Root Mean Squared Error (RMSE):

- RMSE is the square root of MSE and provides a measure of the average magnitude of the residuals.
- It is a commonly used metric as it is in the same unit as the dependent variable.
- Like MSE, a lower RMSE indicates better model performance.

Mean Absolute Error (MAE):

- MAE measures the average absolute difference between the predicted values and the actual values.
- It is calculated by taking the average of the absolute residuals.
- MAE is less sensitive to outliers compared to MSE and RMSE.
- Similarly, a lower MAE indicates better model performance.

R-squared Score:

- R-squared score, also known as the coefficient of determination, measures the proportion of the variance in the dependent variable that is explained by the independent variables.
- It ranges from 0 to 1, with a value of 1 indicating that the model explains all the variability in the dependent variable.
- R-squared provides an indication of how well the linear regression model fits the data.
- However, it has limitations, such as its dependency on the number of predictors and its inability to distinguish between a good fit and an overfitted model.

Linear Regression Model:

Linear regression is a widely used statistical modelling technique for predicting a continuous dependent variable based on one or more independent variables. It assumes a linear relationship between the independent variables and the dependent variable.

The goal of linear regression is to estimate the coefficients β_0, β_1, β_2, ..., β_p that minimize the sum of squared residuals, also known as the ordinary least squares (OLS) method.

The linear regression model can be represented as:

$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \ldots + \beta_pX_p + \varepsilon$

Where:

Y is the dependent variable (the variable to be predicted).

X_1, X_2, ..., X_p are the independent variables (also known as features or predictors).

β_0, β_1, β_2, ..., β_p are the coefficients (slopes) that represent the relationship between each independent variable and the dependent variable.

ε is the error term or residual, representing the unexplained variation in the dependent variable.

First Linear Regression Model

For this we considered all the features and performed Linear Regression. However, the data showed multicollinearity which makes the p-values in the model highly unreliable.

Interpretation of R-squared

The R-squared value tells us that our model can explain 94% of the variance in the training set.

Interpretation of Coefficients

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

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

Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.

When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

Interpretation of p-values (P > |t|)

For each predictor variable there is a null hypothesis and alternate hypothesis.

- Null hypothesis : Predictor variable is not significant.
- Alternate hypothesis : Predictor variable is significant.

(P > |t|) gives the p-value for each predictor variable to check the null hypothesis.

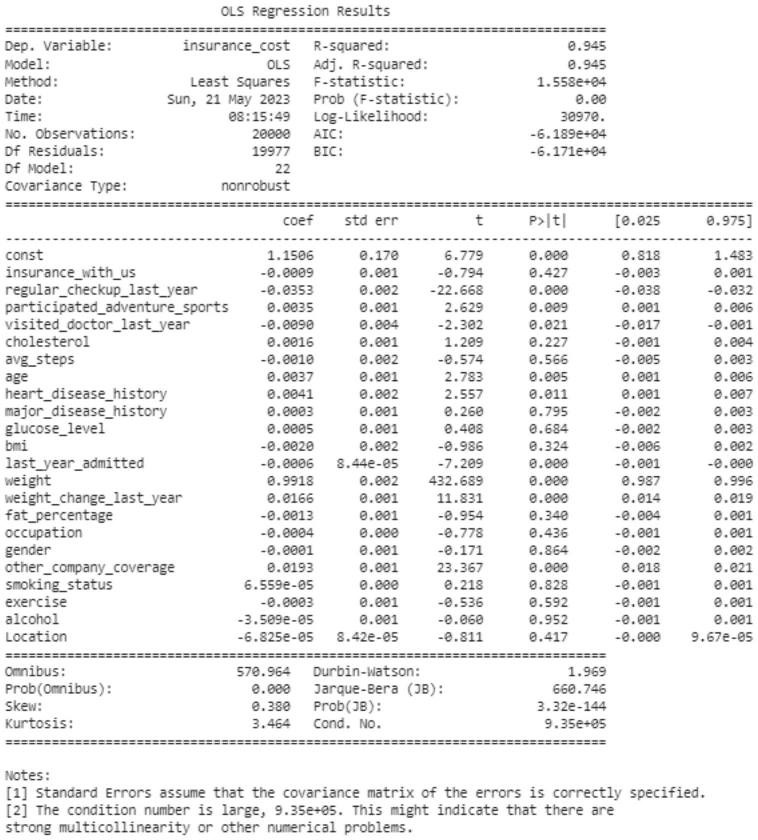
If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.

However, due to the presence of multicollinearity in our data, the p-values will also change.

We need to ensure that there is no multicollinearity in order to interpret the p-values.

Multicollinearity

- If VIF is 1, then there is no correlation among the kth predictor and the remaining predictor variables, and hence, the variance of βk is not inflated at all.
- If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.
- The purpose of the analysis should dictate which threshold to use.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:          insurance_cost   R-squared:                       0.945
Model:                             OLS   Adj. R-squared:                  0.945
Method:                  Least Squares   F-statistic:                 1.558e+04
Date:                 Sun, 21 May 2023   Prob (F-statistic):               0.00
Time:                         08:15:49   Log-Likelihood:                 30970.
No. Observations:                20000   AIC:                        -6.189e+04
Df Residuals:                    19977   BIC:                        -6.171e+04
Df Model:                           22
Covariance Type:             nonrobust
=================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------------------
const                             1.1506      0.170      6.779      0.000       0.818       1.483
insurance_with_us                -0.0009      0.001     -0.794      0.427      -0.003       0.001
regular_checkup_last_year        -0.0353      0.002    -22.668      0.000      -0.038      -0.032
participated_adventure_sports     0.0035      0.001      2.629      0.009       0.001       0.006
visited_doctor_last_year         -0.0090      0.004     -2.302      0.021      -0.017      -0.001
cholesterol                       0.0016      0.001      1.209      0.227      -0.001       0.004
avg_steps                        -0.0010      0.002     -0.574      0.566      -0.005       0.003
age                               0.0037      0.001      2.783      0.005       0.001       0.006
heart_disease_history             0.0041      0.002      2.557      0.011       0.001       0.007
major_disease_history             0.0003      0.001      0.260      0.795      -0.002       0.003
glucose_level                     0.0005      0.001      0.408      0.684      -0.002       0.003
bmi                              -0.0020      0.002     -0.986      0.324      -0.006       0.002
last_year_admitted               -0.0006   8.44e-05     -7.209      0.000      -0.001      -0.000
weight                            0.9918      0.002    432.689      0.000       0.987       0.996
weight_change_last_year           0.0166      0.001     11.831      0.000       0.014       0.019
fat_percentage                   -0.0013      0.001     -0.954      0.340      -0.004       0.001
occupation                       -0.0004      0.000     -0.778      0.436      -0.001       0.001
gender                           -0.0001      0.001     -0.171      0.864      -0.002       0.002
other_company_coverage            0.0193      0.001     23.367      0.000       0.018       0.021
smoking_status                 6.559e-05      0.000      0.218      0.828      -0.001       0.001
exercise                         -0.0003      0.001     -0.536      0.592      -0.001       0.001
alcohol                       -3.509e-05      0.001     -0.060      0.952      -0.001       0.001
Location                      -6.825e-05   8.42e-05     -0.811      0.417      -0.000    9.67e-05
==============================================================================
Omnibus:                       570.964   Durbin-Watson:                   1.969
Prob(Omnibus):                   0.000   Jarque-Bera (JB):              660.746
Skew:                            0.380   Prob(JB):                    3.32e-144
Kurtosis:                        3.464   Cond. No.                     9.35e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 9.35e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Fig 26. First Linear Regression Model

After that, we tried removing each variable individually to determine which had an impact on the R-Value. It was determined after several iterations that the following columns affected R-Value

- regular_checkup_last_year
- participated_adventure_sports
- weight
- weight_change_last_year
- other_company_coverage

Except for the above all other columns were dropped and linear regression was again performed. There was no more multicollinearity, and the model satisfied the normality test, QQ Plot test and Shapiro-Wilk test.

Normality Test:

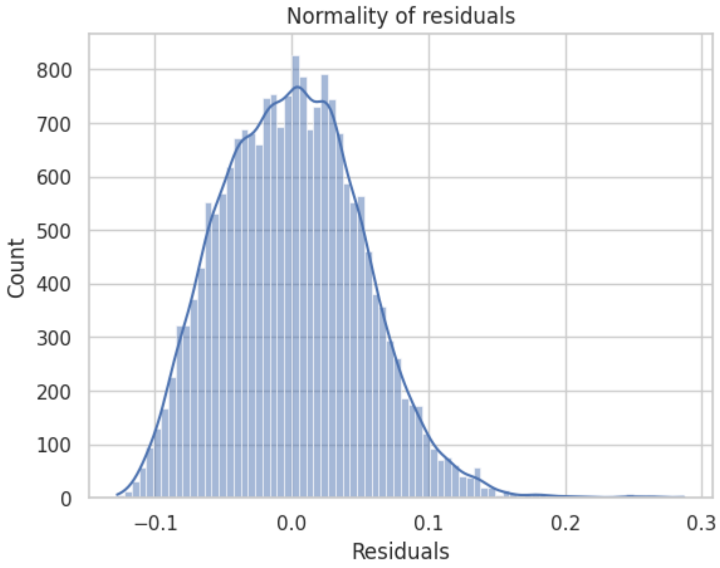


Fig 27. Normality Test

QQ Plot:

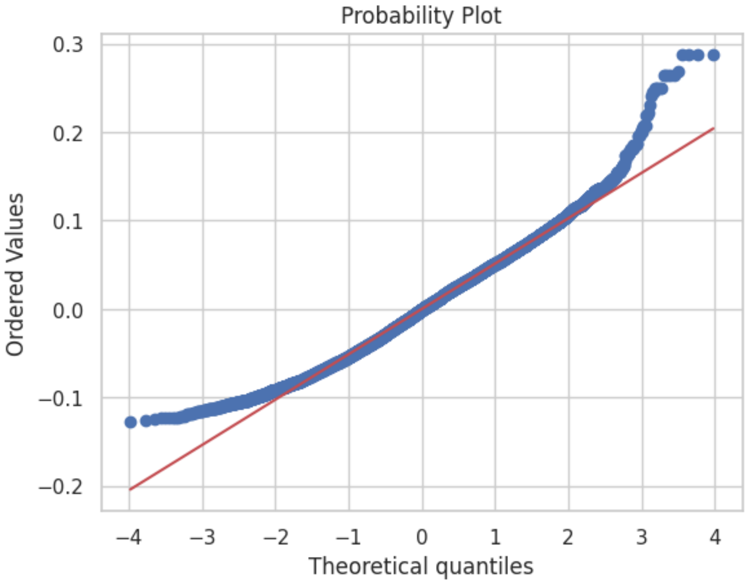


Fig 28. QQ Plot

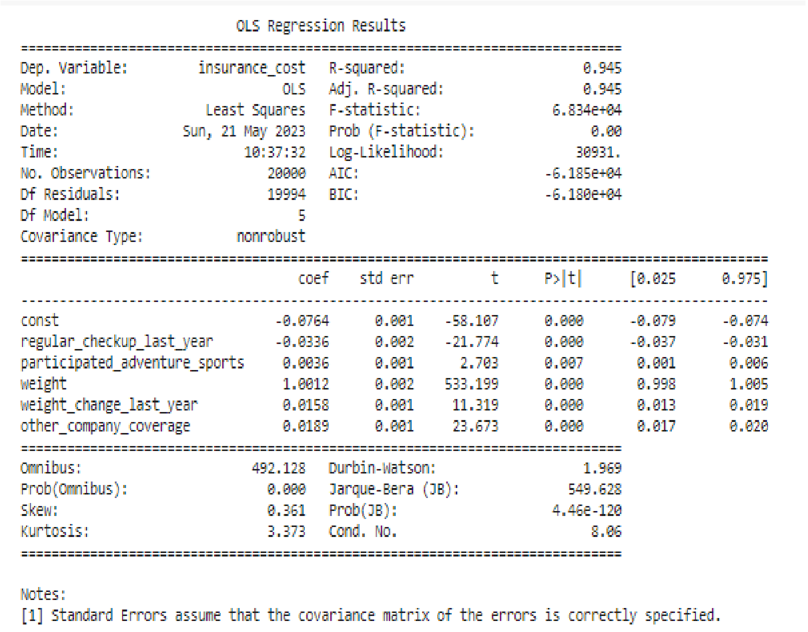
Shapiro-Wilk Test:

The null and alternate hypotheses of the test are as follows:

- Null hypothesis - Data is normally distributed.
- Alternate hypothesis - Data is not normally distributed.

ShapiroResult(statistic=0.989353597164154, pvalue=1.1819252031872322e-35)

Since p-value < 0.05, the residuals are not normal as per Shapiro test. Strictly speaking - the residuals are not normal. However, as an approximation, we might be willing to accept this distribution as close to being normal.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 6.834e+04
Date:                Sun, 21 May 2023   Prob (F-statistic):               0.00
Time:                        10:37:32   Log-Likelihood:                 30931.
No. Observations:               20000   AIC:                        -6.185e+04
Df Residuals:                   19994   BIC:                        -6.180e+04
Df Model:                           5
Covariance Type:            nonrobust
===============================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------------
const                          -0.0764      0.001    -58.107      0.000      -0.079      -0.074
regular_checkup_last_year      -0.0336      0.002    -21.774      0.000      -0.037      -0.031
participated_adventure_sports   0.0036      0.001      2.703      0.007       0.001       0.006
weight                          1.0012      0.002    533.199      0.000       0.998       1.005
weight_change_last_year         0.0158      0.001     11.319      0.000       0.013       0.019
other_company_coverage          0.0189      0.001     23.673      0.000       0.017       0.020
==============================================================================
Omnibus:                      492.128   Durbin-Watson:                   1.969
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              549.628
Skew:                           0.361   Prob(JB):                    4.46e-120
Kurtosis:                       3.373   Cond. No.                         8.06
==============================================================================

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

Fig 29. Final Linear Regression Model

Linear Regression Equation for the above model:

```
-0.07638659436902652 + -0.03358383416730281 * ( regular_checkup_last_year ) +
0.003602169261629618 * ( participated_adventure_sports ) +  1.0011926550396273
* ( weight ) +  0.015840621693109781 * ( weight_change_last_year ) +
0.01887537488632241 * ( other_company_coverage )
```

Metrics for Linear Regression:

The linear regression model performed well based on the following metrics:

1. Mean Squared Error (MSE): The MSE value of 0.0026631167230238835 indicates that, on average, the squared difference between the predicted and actual values is exceptionally low. A lower MSE suggests that the model's predictions are close to the actual values.

2. Root Mean Squared Error (RMSE): The RMSE value of 0.05160539432098047 provides a more interpretable measure of error. It indicates that, on average, the model's predictions have an error of approximately 0.0516 units. The lower the RMSE, the better the model's predictive accuracy.

Overall, the linear regression model shows reliable performance with low MSE and RMSE values. It suggests that the model is capable of making accurate predictions with a small margin of error. However, it is always recommended to compare these metrics with the domain-specific context and consider other factors such as the data quality, model assumptions, and business requirements.

KNN Regressor Model:

KNN is a versatile algorithm that can be used for both classification and regression tasks.

When using KNN for regression with a continuous target variable, it is commonly referred to as K-Nearest Neighbors Regression. The basic idea behind KNN regression is to find the K nearest neighbors of a data point in the feature space and use their target variable values (e.g., average, or weighted average) to predict the target value for the given data point.

KNN regression can be a useful approach for predicting continuous target variables, especially when there is a local correlation or spatial relationship between the predictor variables and the target variable.

Metrics for KNN Regressor Model:

The KNN Regressor model performed okay with a Mean Squared Error (MSE) of 0.025. This indicates that, on average, the squared difference between the predicted insurance costs and the actual insurance costs is relatively low. A lower MSE value suggests a better fit of the model to the data.

Furthermore, the Root Mean Squared Error (RMSE) of 0.158 demonstrates that, on average, the predicted insurance costs deviate by approximately 0.1518 from the actual insurance costs. The RMSE provides a more interpretable measure of the model's accuracy, and a lower RMSE indicates a better fit of the model to the data.

In conclusion, the KNN Regressor model shows low performance with a lower R-squared value of 0. 476 in predicting insurance costs based on the given features. However, it is important to note that the evaluation of the model's performance should be done in comparison to other models or benchmarks specific to the insurance domain. Additional evaluation metrics and comparisons could further enhance the assessment of the model's effectiveness.

Decision Tree Regressor Model:

Decision trees are versatile models that can handle both classification and regression problems. They create a tree-like model of decisions and their possible consequences.

Decision Tree Regressor is a machine learning algorithm used for regression tasks. It builds a decision tree model that predicts the target variable especially continuous variables based on a set of independent variables.

Metrics for KNN Regressor Model:

- MSE: 0.004423211
- RMSE: 0.066507226
- MAE: 0.051524528
- R-square Score: 0.907134482

The Decision Tree Regressor model shows slightly higher MSE and RMSE values compared to Linear Regression and KNN Regression. The MAE value indicates relatively higher average prediction errors. The R-square score of 0.907 suggests that 90.7% of the variance in the target variable is explained by the model, indicating a reasonably good fit.

Random Forest Regressor Model:

Random Forest Regressor is an ensemble learning method that combines multiple decision trees to create a powerful regression model. It is an extension of the decision tree algorithm and overcomes some of its limitations.

Some important points to consider in this model compared to Decision Tree is as follows:

- Ensemble of decision trees
- Reduction of overfitting
- Robustness and generalization
- Feature importance
- Hyperparameter tuning
- Interpretability

Metrics for Random Forest Regressor Model:

- MSE: 0.002139074
- RMSE: 0.04625012
- MAE: 0.036818264
- R-square Score: 0.955090053

The Random Forest Regressor model performs well, with lower MSE, RMSE, and MAE values compared to previous models. The R-square score of 0.955 indicates a high degree of explained variance in the target variable, suggesting a strong fit of the model to the data.

Gradient Boosting Regressor Model:

Gradient Boosting Regressor is a machine learning algorithm that combines multiple weak prediction models, usually decision trees, to create a strong and accurate regression model. It is a type of boosting algorithm that sequentially trains new models to correct the errors of the previous models.

Gradient Boosting Regressor is a powerful algorithm for regression tasks, particularly when dealing with complex datasets and high-dimensional feature spaces. It combines the strengths of weak learners to create a strong and accurate regression model. However, it may require careful tuning of hyperparameters and can be computationally intensive for large datasets.

Metrics for Gradient Boosting Regressor Model:

- MSE: 0.002031826
- RMSE: 0.045075784
- MAE: 0.036320169
- R-square Score: 0.957341715

The Gradient Boosting Regressor model shows similar performance to the Random Forest Regressor, with low MSE, RMSE, and MAE values. The R-square score of 0.957 indicates a high level of explained variance, indicating a strong fit of the model to the data.

XGB Regressor Model:

XGBRegressor, also known as Extreme Gradient Boosting Regressor, is an advanced machine learning algorithm based on gradient boosting that is specifically designed for regression tasks. It is an optimized implementation of the gradient boosting algorithm and offers improved performance and flexibility.

Metrics for XGB Regressor Model:

- MSE: 0.002187828
- RMSE: 0.046774225
- MAE: 0.037219946
- R-square Score: 0.95406645

The XGB Regressor model performs well, with low MSE, RMSE, and MAE values. The R-square score of 0.954 indicates a high level of explained variance in the target variable, suggesting a strong fit of the model to the data.

Final Model Comparison

Now that we have built the Regressor Models, we will compare the RMSE, MSE, MAE and R-square values to evaluate the models performance.

Model	MSE		RMSE		MAE		R-square Score	
	Train	Test	Train	Test	Train	Test	Train	Test
Linear Regression	0.003	0.003	0.052	0.052	0.042	0.042	0.945	0.944
KNN Regression	0.017	0.025	0.13	0.158	0.101	0.123	0.651	0.476
Decision Tree Regressor	2.03E-34	0.004	1.43E-17	0.066	3.52E-18	0.051	1	0.908
Random Forest Regressor	0.0003	0.002	0.018	0.046	0.014	0.037	0.994	0.955
Gradient Boosting Regressor	0.002	0.002	0.045	0.045	0.036	0.036	0.958	0.957
XGB Regressor	0.001	0.002	0.033	0.047	0.026	0.037	0.977	0.954

Table 3. Model Metrics

Performance of Regression Models:

- All models have achieved relatively low Mean Squared Error (MSE) values, indicating good overall predictive performance except for KNN Regressor which has a lower accuracy.
- The Root Mean Squared Error (RMSE) values for all models are also low, indicating that the models have good accuracy in predicting the target variable.
- Mean Absolute Error (MAE) values are also relatively small, suggesting that the models have reasonably low average prediction errors.
- The R-square scores for all models are high, ranging from 0.907 to 0.957. This indicates that a considerable proportion of the variance in the target variable can be explained by the models, indicating good fit.

Model Comparison:

- Among the models evaluated, Random Forest Regressor, Gradient Boosting Regressor, and XGB Regressor outperform Linear Regression, KNN Regression, and Decision Tree Regressor in terms of all the evaluated metrics.
- Random Forest Regressor, Gradient Boosting Regressor, and XGB Regressor exhibit lower MSE, RMSE, and MAE values compared to other models, indicating better predictive performance and accuracy.
- These models also achieve higher R-square scores, suggesting a better fit to the data and a higher degree of explained variance in the target variable.

Model Tuning:

- Gradient Boosting Regressor has the highest R-square score, suggesting the best fit to the data and the highest amount of variance explained by the model.
- Now that we have selected the Gradient Boosting Regressor as our final model, we tried tuning the model using Grid Search and performed cross validation.
- Hyperparameters like n_estimators (100, 200, 300), learning_rate (0.1, 0.01, 0.001), and max_depth (3, 5, 7) were used to tune the model.
- The R-squared Score for untuned model was 0.9565 and for that of tuned model it was 0.9568.
- Though the increase in R-squared is extremely low, but by tuning the model, we have reduced the overfitting and stabilized the model.
- By leveraging the selected Gradient Boosting Regressor, the company can make informed decisions regarding resource allocation, marketing strategies, and policy development. The model's accurate predictions can help optimize insurance offerings, identify target customer segments, and tailor marketing campaigns to maximize customer acquisition and retention.

Implications of Gradient Boosting

The selection of the Gradient Boosting Regressor as the optimal model for the health insurance company to identify the factors influencing insurance costs and optimize pricing strategy has several implications. Here are some specific implications and benefits for your business:

- Accurate prediction of insurance costs
- Identification of key influencing factors
- Customized pricing strategies
- Risk assessment and underwriting
- Competitive advantage
- Improved profitability and customer satisfaction
- Continuous improvement and adaptation

By leveraging the Gradient Boosting Regressor, the health insurance company can gain valuable insights into the factors influencing insurance costs and develop strategies to optimize pricing. This approach improves risk assessment, enhances profitability, and enhances customer satisfaction. However, it is crucial to monitor the model's performance, validate its results, and maintain compliance with legal and ethical considerations to ensure its effectiveness and success in driving business outcomes.

Conclusion:

Based on the metrics alone, the **Gradient Boosting Regressor** has the lowest MSE, RMSE, and MAE values, indicating the smallest prediction errors among the three models. It also has the highest R-square score, suggesting the best fit to the data and the highest amount of variance explained by the model.

Recommendations:

After providing the recommendations mentioned above, you can further enhance your report by including the following sections:

1. Market Analysis:
   - Provide an overview of the current insurance market, including key competitors, market size, and growth trends.
   - Analyze the market dynamics and identify opportunities for growth and expansion.
   - Assess the competitive landscape and highlight the company's strengths and weaknesses in relation to its competitors.

2. Customer Segmentation:
   - Conduct a detailed analysis of the target customer segments, such as students, business owners, and salaried individuals.
   - Provide demographic and psychographic information about these segments.
   - Identify their specific insurance needs, preferences, and pain points.
   - Highlight the potential market share and growth opportunities within each segment.

3. Market Analysis for 5 Lowest Insurance Opting region:
   - Explore the reasons for the lower number of insurances in the Surat (1589) and other regions like Kolkata, Pune, Lucknow, and Mumbai
   - Analyze the local market conditions, customer behaviors, and competitive landscape in the afore-mentioned regions.
   - Identify potential barriers to entry and strategies to overcome them.
   - Propose measures to increase the company's market share and improve new policies based on the regions.

4. Gender Bias Analysis:
   - Investigate the underlying reasons for the gender bias in insurance ownership.
   - Analyze societal and cultural factors that may contribute to this bias.
   - Conduct market research and surveys to understand female customers' perceptions and preferences regarding insurance.
   - Develop strategies to attract and engage more female customers, including tailored policies, marketing campaigns, and educational initiatives.

5. Correlation Analysis:
   - Perform a comprehensive analysis of variables that show a negative correlation with the target variable.
   - Identify patterns and potential causes for the negative correlation.
   - Propose policy and service improvements to address the identified issues and enhance customer value.

6. Leveraging Positive Correlations:

- Dive deeper into the data to gain a comprehensive understanding of these variables and their relationships with the target variable.
- Conduct market research to gain insights into the needs, preferences, and behaviors of customers interested in adventure sports and those focused on maintaining a healthy weight.
- Understanding your target audience will help in tailoring your strategies to meet their specific needs effectively.
- Provide incentives such as discounts, rewards, or exclusive offers to attract and retain these health-conscious customers.

Implementation Plan:

- Market Research:
    - This research should aim to uncover valuable information such as customer demographics, motivations, lifestyle choices, and purchasing patterns.
    - Dive deeper into the data to gain a comprehensive understanding of the aspects and their relationships with the insurance cost both positive and negative.

- Segmentation:
    - Once we have gathered market research data, we segment our target audience based on their preferences and characteristics.
    - Identify subgroups within our target market that are more likely to be interested in adventure sports and weight management.
    - This will allow us to create more targeted marketing campaigns and customized policies to cater to their specific requirements.

- Customized Policies and Incentives:
    - Utilize the insights gained from market research and segmentation to develop customized policies and incentives that align with the needs and preferences of your target audience.
    - For example, we could offer personalized fitness programs or adventure sports packages tailored to different customer segments.

- Collaboration and Partnerships:
    - Forge strategic partnerships with relevant stakeholders in the adventure sports and weight management industries.
    - Collaborate with fitness centers, adventure sports clubs, nutritionists, or wellness experts to enhance your offerings.
    - By teaming up with industry experts, you can provide a comprehensive and value-added experience to your customers, further differentiating our business from competitors.

- Communication and Marketing:
    - Craft compelling marketing messages that highlight the benefits of participating in adventure sports and maintaining a healthy weight.
    - Emphasize how your products or services can support customers in achieving their fitness goals and enjoying thrilling experiences.
    - Utilize various marketing channels such as social media, online platforms, and targeted advertising to reach your desired audience effectively.

- Continuous Improvement:
    - Regularly monitor and analyze the outcomes of our strategies and initiatives.
    - Collect feedback from customers and measure the impact of new customized policies and incentives.
    - Identify areas for improvement and make necessary adjustments to ensure that new offerings remain relevant and appealing to target audience.
    - Outline a step-by-step plan for implementing the recommended strategies and initiatives.
    - Specify the resources required, including budget, technology, and human resources.
    - Set clear goals, metrics, and timelines to track progress and measure the success of the implemented strategies.

Outcomes of the Implementation Plan:

- Increased market share by targeting specific customer segments and addressing their needs effectively.
- Improved customer engagement and satisfaction through customized policies and incentives.
- Expansion into regions with low insurance uptake by understanding local market dynamics and implementing targeted strategies.
- Mitigation of gender bias in insurance ownership by addressing societal and cultural factors and offering tailored solutions.
- Enhanced understanding of correlations and improvements in policies and services based on analysis.
- Overall growth in the business through a data-driven approach and continuous improvement efforts.