Capstone Project

S Nitin Kumar

[Date]

DSBA

Contents

1. Introduction of the business problem

a) Defining problem statement

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance, then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

b) Need of the study/project

Based on the study, we must determine each person's health, offer a reasonable level of insurance, and determine which health-related factors have an impact on the insurance policy for incurable diseases.

c) Understanding business/social opportunity

- Business understanding in this context refers to the ability to analyse and interpret data related to healthcare and insurance in order to make informed decisions and strategies.
- This may include conducting studies to determine an individual's health status, offering appropriate levels of insurance coverage based on this information, and identifying factors that have an impact on insurance policy for incurable diseases.
- This understanding can aid medical insurance companies in developing policies that effectively minimize risk and optimize costs while also providing necessary coverage for individuals.
- It also can help individuals and society to have better access to the health care services they need and to have a fair and transparent pricing for the services.

2. EDA and Business Implication

a) Understanding how data was collected in terms of time, frequency and methodology

The data set being referred to in this statement is likely one that has been compiled by an insurance company and contains information about the health and habits of individuals. This data can be used to estimate the cost of insurance for those individuals. This could include information such as an individual's age, gender, medical history, and habits such as smoking and exercise. By analysing this data, the insurance company can determine the level of risk associated with insuring a particular individual and use that information to set appropriate insurance rates. This can help the company to optimize costs and minimize risk while also ensuring that individuals are offered fair and accurate insurance rates based on their health and habits. b) Visual inspection of data

The data's shape contains 25000 columns and 24 rows

b) Understanding of attributes

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

Understanding how data was collected in terms of time, frequency and methodology

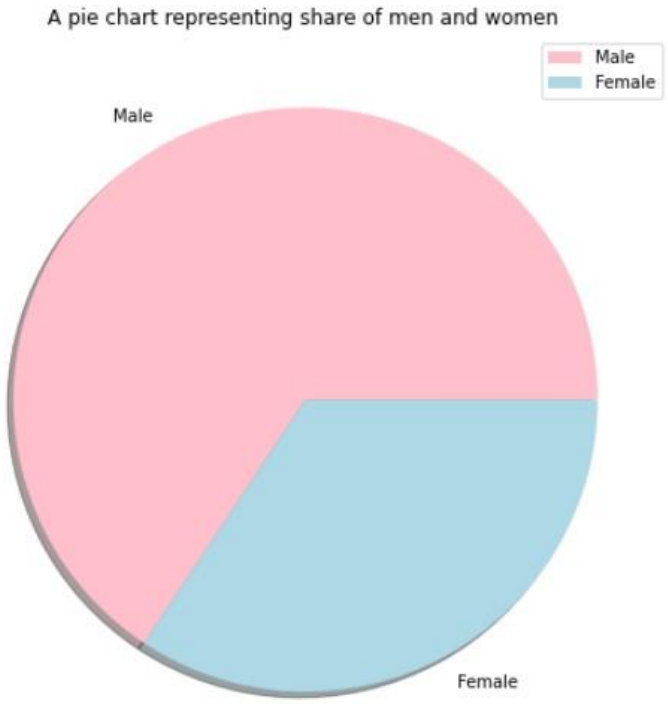
- The data set being referred to in this statement is likely one that has been compiled by an insurance company and contains information about the health and habits of individuals.
- This data can be used to estimate the cost of insurance for those individuals.
- This could include information such as an individual's age, gender, medical history, and habits such as smoking and exercise.
- By analysing this data, the insurance company can determine the level of risk associated with insuring a particular individual and use that information to set appropriate insurance rates.
- This can help the company to optimize costs and minimize risk while also ensuring that individuals are offered fair and accurate insurance rates based on their health and habits.
- Visual inspection of data (rows, columns, descriptive details)

c) Understanding of attributes

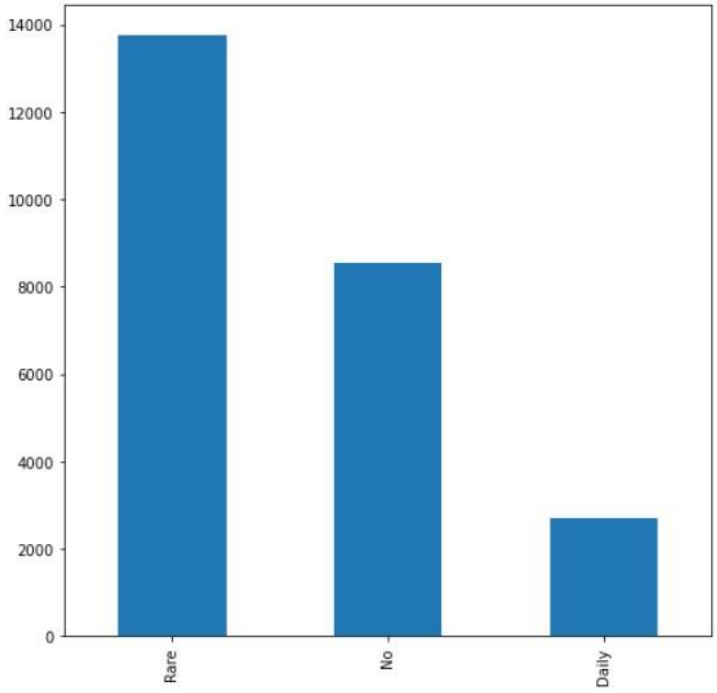
Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
adventure_sports	Customer is involved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

Understanding of attributes (variable info, renaming if required)

a) Univariate analysis

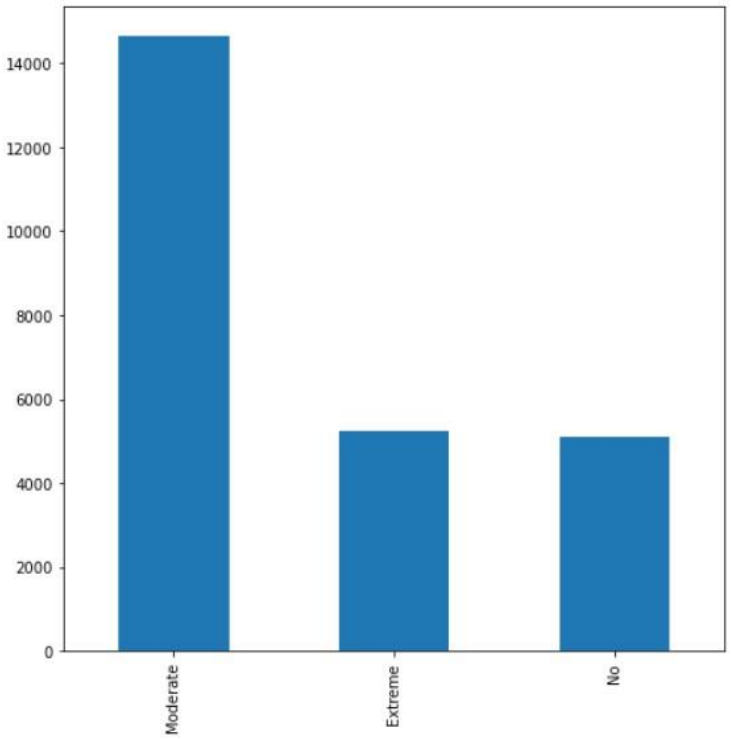


"In above graph male population is high and female population is less in insurance" suggests that there is a greater number of males who have taken insurance policies compared to females in the population being referred to. This could be due to a variety of factors such as differences in income, occupation, lifestyle, and health status between the sexes. For instance, males may be more likely to work in high-risk occupations, which would make them more likely to purchase insurance. Additionally, men might be more inclined to take out insurance policies than women, or they might be offered better rates or terms. However, this could also be a reflection of discrimination against women in terms of access to health care and insurance, which is a serious concern that should be addressed.

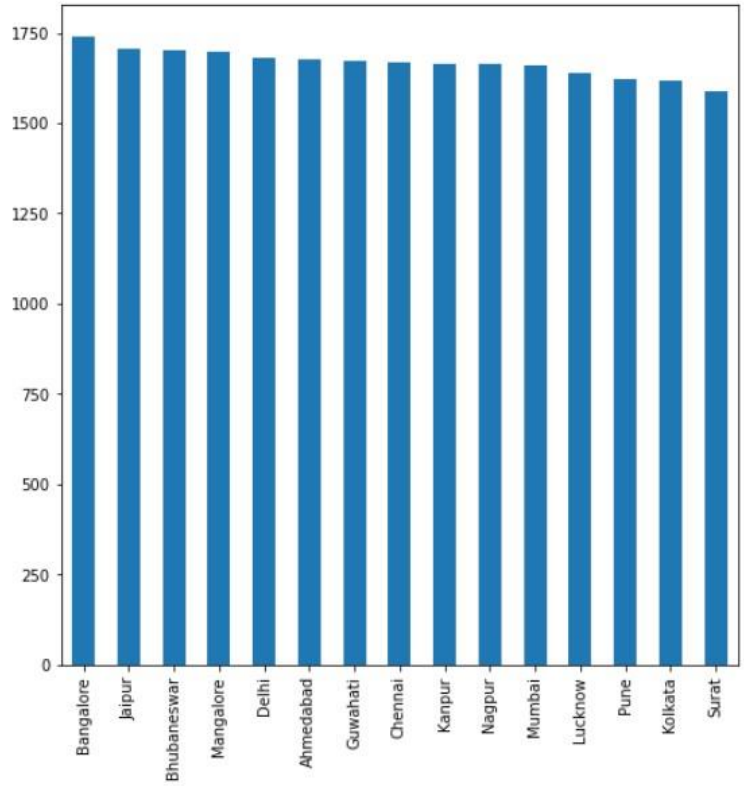


"Above is the consumption of alcohol scenario and rare consumption is high" suggests that the majority of individuals in the population being referred to consume alcohol infrequently or in small amounts. This could be due to a variety of factors such as personal preference, cultural or religious beliefs, or health concerns. It could also be the result of government policies or social norms that discourage excessive alcohol consumption. Additionally, it may

be also that this population is a part of a specific group of people with a high standard of living, who may avoid alcohol consumption or only consume it on special occasions, this can be determined by the context and data provided.

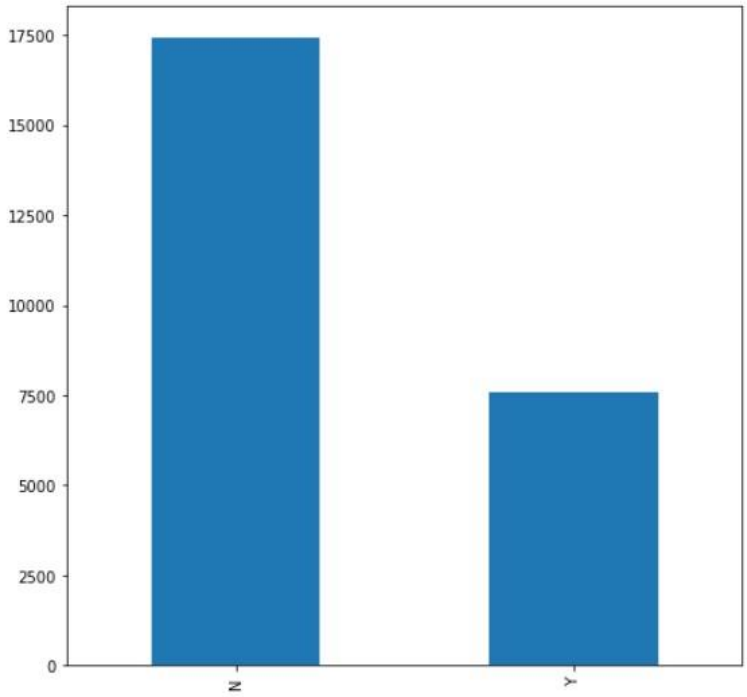


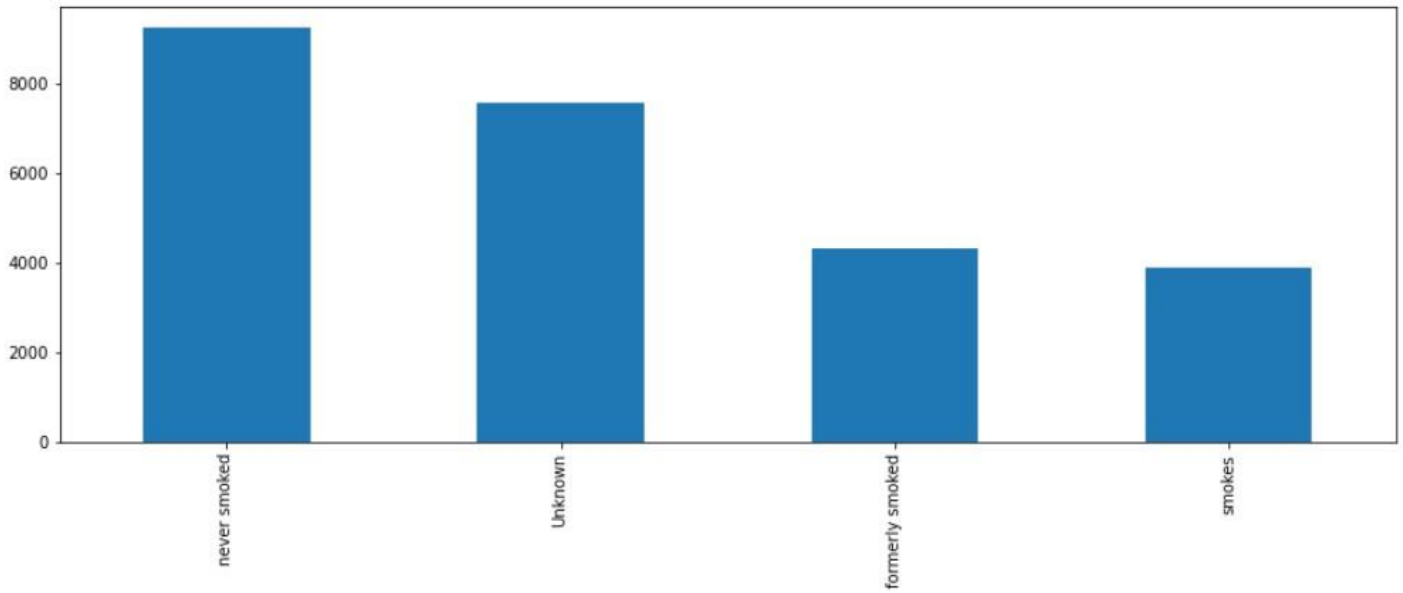
"Above are the people who do exercise and moderate is high" suggests that the majority of individuals in the population being referred to engage in moderate levels of physical exercise. This could be due to a variety of factors such as awareness of the importance of regular exercise for maintaining good health, access to physical activity opportunities, and personal motivation or discipline. It could also be the result of government policies or social norms that encourage regular exercise. Additionally, it may be also that this population is a part of a specific group of people with a high standard of living, who may be more health conscious and engage in moderate exercise regularly.



"Above is the city where insurance policy applied is more in Bangalore compared to other cities" suggests that a higher proportion of individuals in the city of Bangalore have applied for insurance policies compared to other cities.

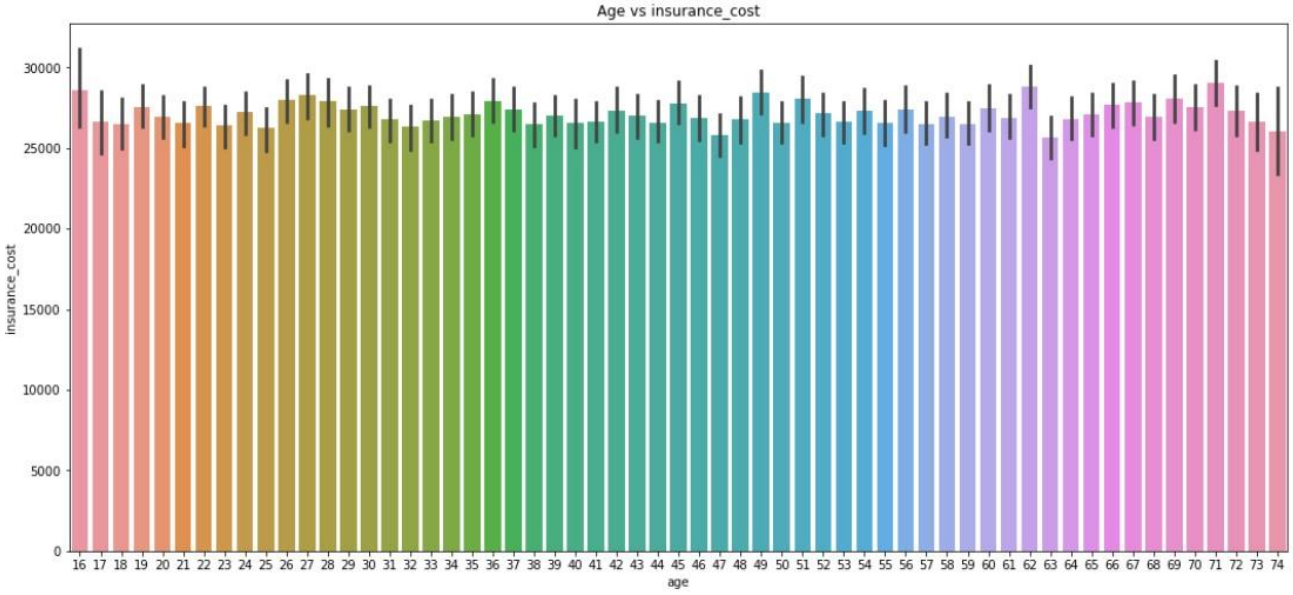
This could be due to a variety of factors such as the economic prosperity of the city, the presence of a large number of insurance companies and their marketing campaigns, or the awareness of the population about the importance of having an insurance policy. Additionally, it could be also that the population in Bangalore is more educated and has a better understanding of insurance policies and its benefits, and they are more likely to take an insurance policy. It is also possible that Bangalore has a larger population than other cities that are being compared, which would naturally result in a higher number of insurance policy applications. More information about the context of the data is needed to determine the exact cause of this trend.





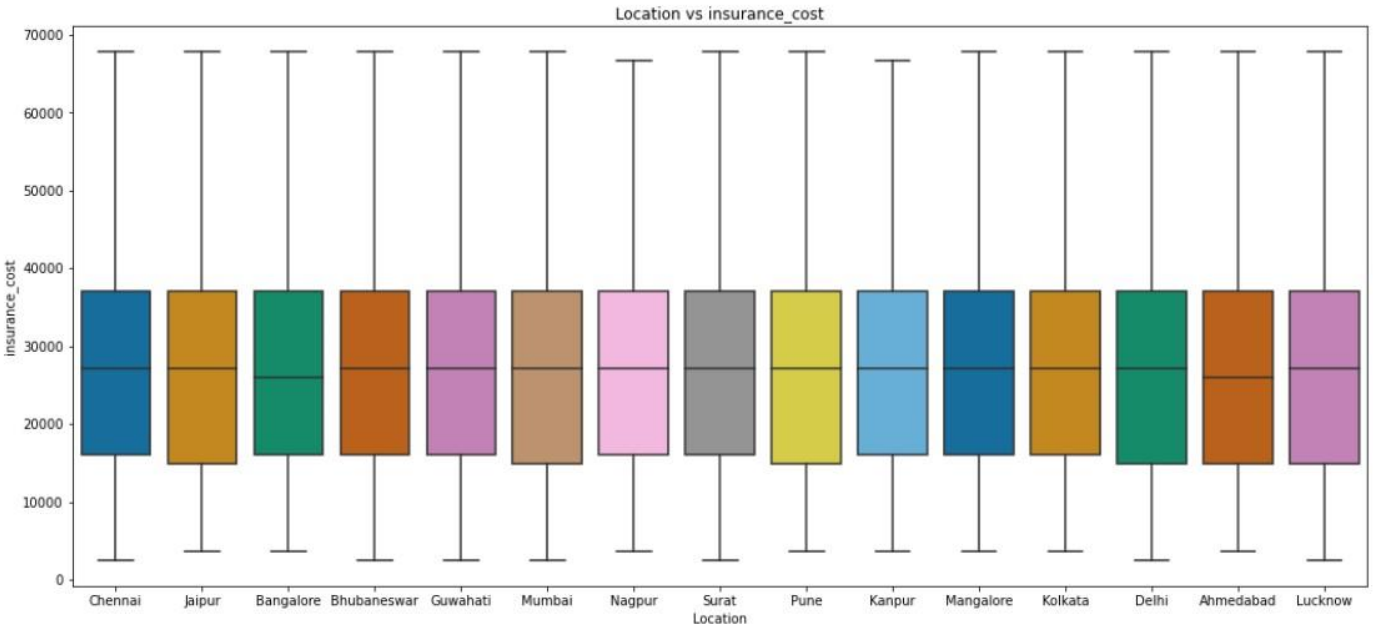
"People who have insurance cost depends on people who smoke and in above figure in the data we see that people never smoked is more which affects insurance cost" suggests that the cost of insurance is affected by whether or not an individual smokes. This likely means that individuals who have never smoked are considered less of a risk for insurance companies, and as a result, their insurance rates may be lower than those of individuals who smoke. This is because smoking has been proven to be a major risk factor for many health issues such as lung cancer, heart disease, stroke, chronic obstructive pulmonary disease, and many other diseases. Insurance companies use this information and many other factors to determine the risk of insuring an individual and to set the premiums. This data can also be useful for individuals to understand how their smoking habits can affect their insurance rates and to make informed decisions about their health.

b) Bivariate analysis (relationship between different variables, correlations)

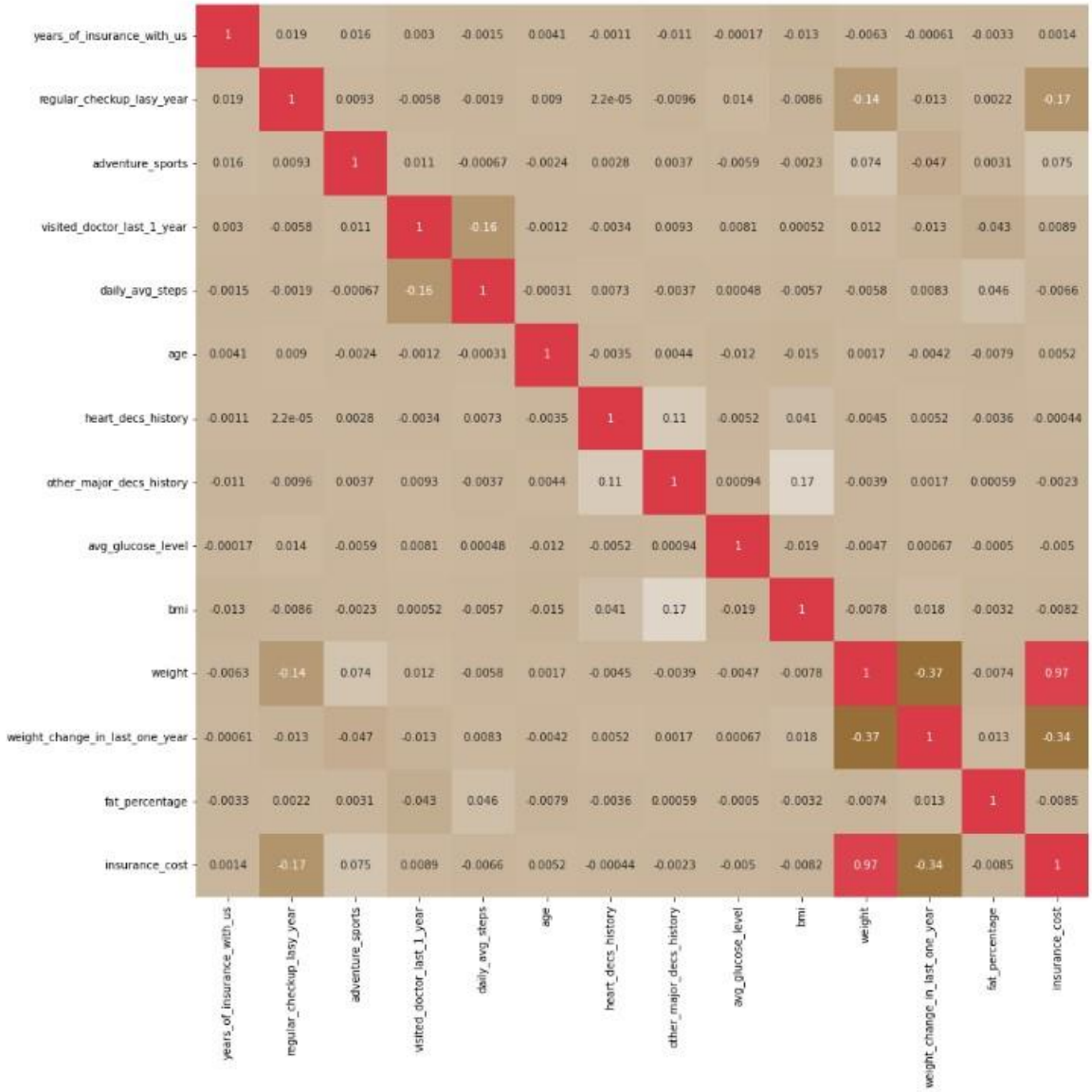


"Insurance cost is similar to all ages" suggests that the cost of insurance is consistent across different age groups. This could mean that the insurance company does not take age into account when determining insurance rates, or that any differences in rates are minimal. This is different from some insurance companies that charge higher rates to older individuals because they are considered to be at a higher risk of certain health conditions. However, there could be other factors such as gender, occupation, income, lifestyle and health status that may have a greater impact on the insurance rates. It is important to note that this statement is dependent on the context and the specific data being analysed, as insurance rates and pricing can vary greatly depending on the company and the policy.

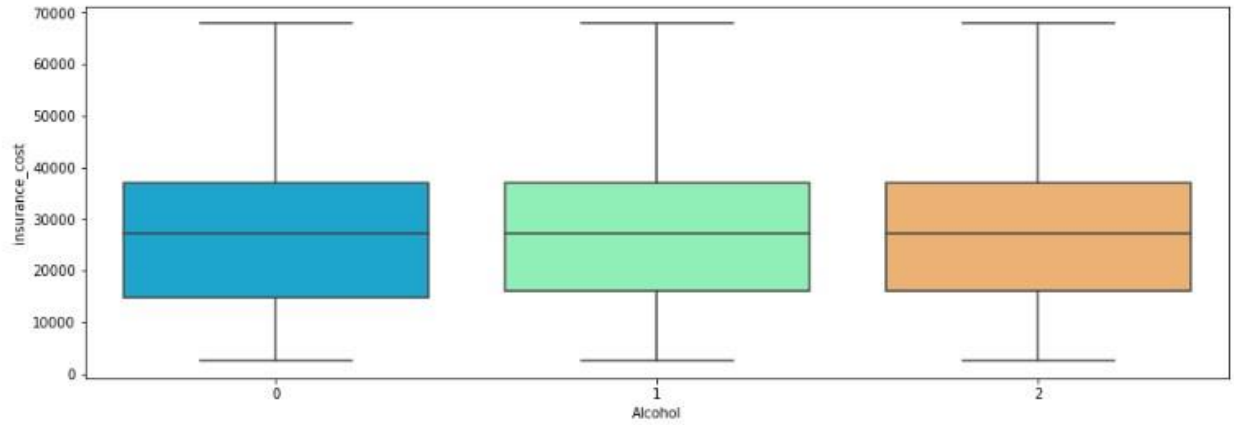
"Based on gender insurance cost is same for all" suggests that the cost of insurance is consistent across different genders. This means that the insurance company does not take into account an individual's gender when determining insurance rates, or that any differences in rates are minimal. This is different from some insurance companies that charge different rates based on gender, with women being typically charged higher rates than men because they are considered to be at a higher risk of certain health conditions. However, there could be other factors such as age, occupation, income, lifestyle and health status that may have a greater impact on the insurance rates. It is important to note that this statement is dependent on the context and the specific data being analysed, as insurance rates and pricing can vary greatly depending on the company and the policy.



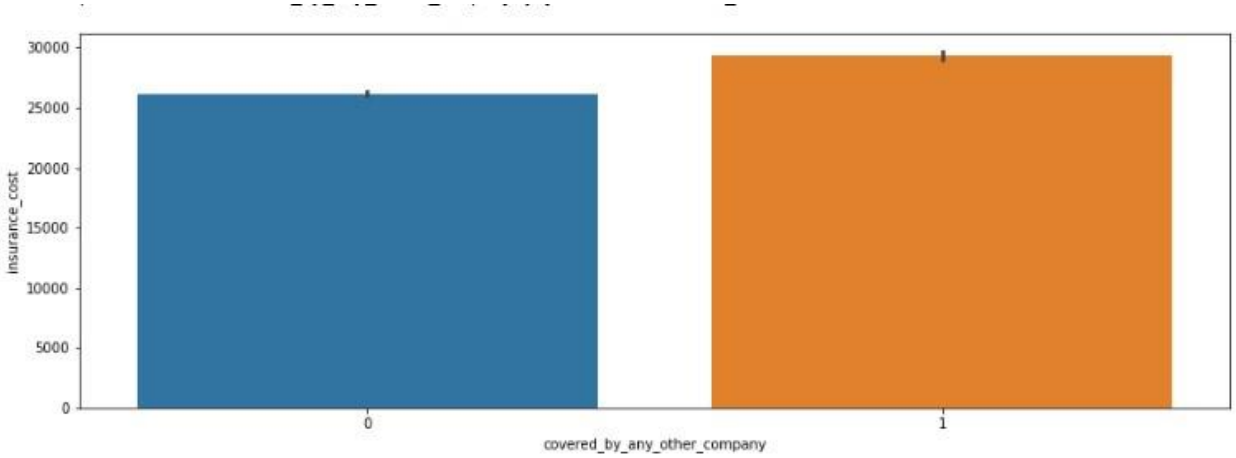
"Insurance cost is same in all city" suggests that the cost of insurance is consistent across different cities. This means that the insurance company does not take into account an individual's location when determining insurance rates, or that any differences in rates are minimal. This could be due to a variety of factors such as the company's pricing strategy, the level of competition in different cities, or the availability of certain types of policies in different areas. However, there could be other factors such as cost of living, crime rate, or access to healthcare facilities that may have an impact on the insurance rates. It is important to note that this statement is dependent on the context and the specific data being analysed, as insurance rates and pricing can vary greatly depending on the company and the policy.



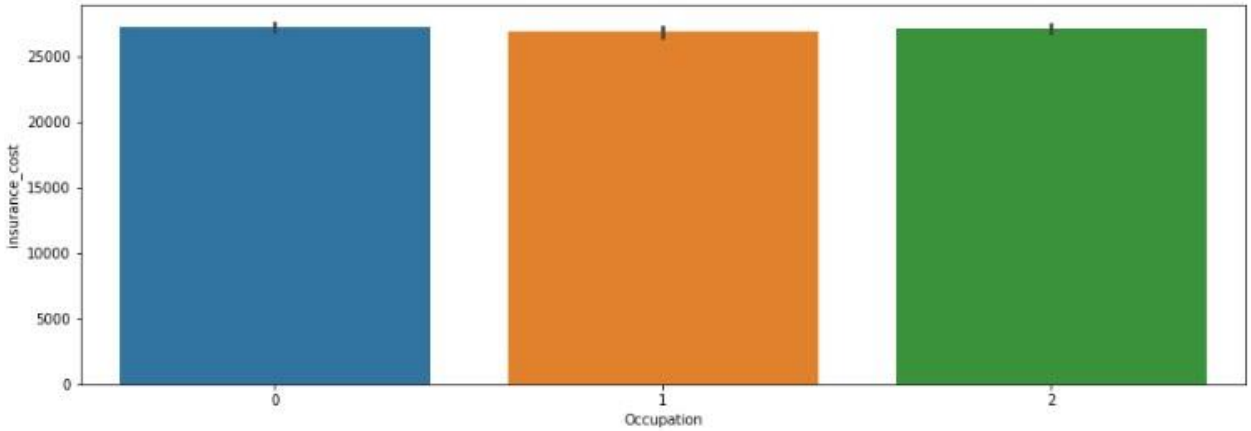
The above diagram represents the correlations of the variables seems there no major correlation between variables



The above diagram represents the consumption of alcohol and consumption is same for all



From the above diagram we can say that the people whose insurance is covered by other company is more



From the above diagram we can say that the issuance cost does not depend on the occupation of the people

3. Data Cleaning and Pre-processing

A. Removal of unwanted variables

"From the given data I have removed applicant_id as it does not give any valuable information and does not help in model building" suggests that the person analyzing the data has determined that the variable "applicant_id" is not useful for their analysis and has therefore removed it from the data set. This is a common practice when working with large data sets, as removing unnecessary variables can make the data easier to work with and can improve the performance of the model being built. The applicant_id is usually a unique identifier for each individual in the data set, it does not give any valuable information in terms of the analysis or the model building, so it can be removed without affecting the final outcome. This decision can be useful to simplify the data analysis process, and also to improve the performance and the accuracy of the model built.

B. Missing Value treatment

Before and after missing value treatment. We ha amputated avg_glucose_level to mean values

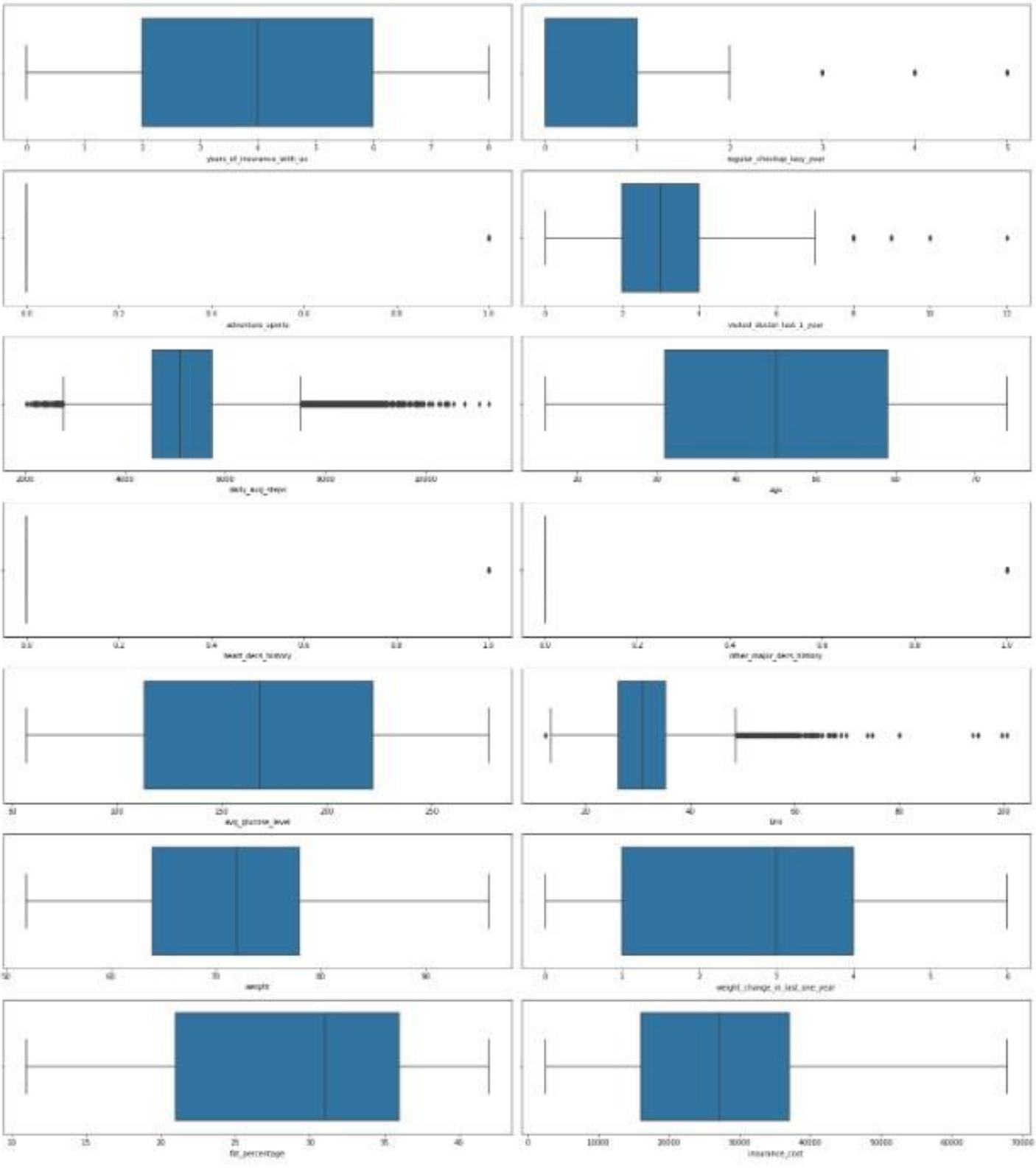
This is a common practice in data cleaning and pre-processing, where outliers or missing values are replaced by the mean value to prevent them from skewing the results. The mean value is a commonly used measure of central tendency, it is calculated by summing up all the values and dividing by the number of values, this allows to have a general idea of the data and to represent it by a single value. Replacing the variable with its mean value can be useful to fill in missing data or to remove outliers, which can improve the performance and accuracy of the model being built. However, it is important to note that replacing values with mean can lead to loss of information and data integrity.

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

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

C. Outlier treatment

"Even we have the outliers we have not imputed the outliers because if we impute missing values then the data set will be imbalanced and will affect the data" suggests that the person analysing the data has chosen not to replace the outliers in the variable "avg_glucose_level" in order to avoid distorting the data set. Imputing missing values or outliers is a common practice in data cleaning and pre-processing, but it can also have a negative impact on the integrity of the data if not done correctly. Replacing the outliers with mean or median values can skew the data set and affect the results of any analysis or modelling that is performed on it. By not imputing the outliers, the data set will retain its natural distribution and variability, which can lead to more accurate and reliable results. However, it is important to note that in some cases, the outliers can be a very important data point and imputing them might lead to loss of information, so it is always recommended to analyse the data set and the specific problem before making a decision on how to deal with outliers


years_of_insurance_with_us
regular_checkup_lasy_year
adventure_sports
visited_doctor_last_1_year
daily_avg_steps
age
heart_decs_history
other_major_decs_history
avg_glucose_level
bmi
weight
weight_change_in_last_one_year
fat_percentage
insurance_cost

D. Variable transformation

	count	mean	std	min	25%	50%	75%	max
years_of_insurance_with_us	25000.0	-6.197443e-16	1.00002	-1.568750	-0.801455	-0.034160	0.733135	1.500430
regular_checkup_lasy_year	25000.0	-1.056470e-15	1.00002	-0.645043	-0.645043	-0.645043	0.188690	3.523623
adventure_sports	25000.0	-1.265019e-15	1.00002	-0.298316	-0.298316	-0.298316	-0.298316	3.352150
Occupation	25000.0	3.785861e-16	1.00002	-1.119445	-1.119445	-0.006632	1.106181	1.106181
visited_doctor_last_1_year	25000.0	4.659961e-16	1.00002	-2.719070	-0.967205	-0.091272	0.784661	7.792122
cholesterol_level	25000.0	-5.448442e-16	1.00002	-1.002742	-1.002742	-0.210186	0.582370	2.167482
daily_avg_steps	25000.0	-2.032907e-16	1.00002	-3.021282	-0.638925	-0.120485	0.488161	5.734284
age	25000.0	-8.559375e-17	1.00002	-1.795369	-0.864107	0.005071	0.874249	1.805511
heart_decs_history	25000.0	-1.325162e-17	1.00002	-0.240412	-0.240412	-0.240412	-0.240412	4.159520
other_major_decs_history	25000.0	3.912026e-16	1.00002	-0.329915	-0.329915	-0.329915	-0.329915	3.031081
Gender	25000.0	6.233680e-16	1.00002	-1.383630	-1.383630	0.722737	0.722737	0.722737
avg_glucose_level	25000.0	-8.428813e-18	1.00002	-1.762039	-0.869302	0.007493	0.868346	1.745141
bmi	25000.0	-3.622880e-16	1.00002	-2.473600	-0.659856	-0.076867	0.506121	8.965938
smoking_status	25000.0	9.904699e-16	1.00002	-1.284920	-1.284920	0.581177	0.581177	1.514226
Location	25000.0	7.472245e-17	1.00002	-1.606745	-0.911425	0.015668	0.942761	1.638081
weight	25000.0	4.625011e-16	1.00002	-2.103001	-0.816138	0.041772	0.685203	2.615499
covered_by_any_other_company	25000.0	1.074723e-15	1.00002	-0.659770	-0.659770	-0.659770	1.515679	1.515679
Alcohol	25000.0	1.703526e-16	1.00002	-2.118525	-0.649164	0.820197	0.820197	0.820197
exercise	25000.0	-8.366641e-18	1.00002	-1.545002	0.008326	0.008326	0.008326	1.561654
weight_change_in_last_one_year	25000.0	-8.179235e-17	1.00002	-1.489652	-0.898041	0.285180	0.876791	2.060012
fat_percentage	25000.0	-3.092238e-16	1.00002	-2.063467	-0.905015	0.253437	0.832663	1.527734
insurance_cost	25000.0	3.682388e-17	1.00002	-1.723013	-0.775333	0.000041	0.689263	2.843080

.Yes, scaling is necessary for clustering in this case as the variance of variables differ from each other. Standard deviation of the columns is also different with spending having the maximum standard deviation. Clustering is a distance-based technique. The units of all the columns are also different. There are 6 amount columns where some are in 100s, some in 1000s and some in 10000s and one is a probability column. So, we need to scale all the features so that we can have a uniform unit for all the columns.

4. Model building

- Linear regression
  - ✓ Linear Regression is a statistical method for modelling the relationship between a dependent variable (output) and one or more independent variables (inputs). The goal of linear regression is to find the best line of fit that minimizes the difference between the observed and predicted values. The line of fit is represented by an equation of the form
  - ✓ A coefficient in linear regression is a numerical value that represents the contribution of an independent variable (predictor) to the prediction of the dependent variable (outcome).
  - ✓ The coefficient of the variables based on the importance is shown below

```
The coefficient for years_of_insurance_with_us is -19.331933363114945
The coefficient for regular_checkup_lasy_year is -410.97977524913523
The coefficient for adventure_sports is 186.77018224478573
The coefficient for Occupation is 4.552955594932844
The coefficient for visited_doctor_last_1_year is -51.39831247157354
The coefficient for cholesterol_level is 22.13627924521841
The coefficient for daily_avg_steps is -0.0241544872620407
The coefficient for age is 2.089441414651534
The coefficient for heart_decs_history is 196.5677470451858
The coefficient for other_major_decs_history is 72.26988345908502
The coefficient for Gender is 22.075954598797807
The coefficient for avg_glucose_level is -0.21265990317474956
The coefficient for bmi is -3.7777055104966806
The coefficient for smoking_status is -3.661267193622765
The coefficient for Location is 7.8266492771426
The coefficient for weight is 1490.7651807331329
The coefficient for covered_by_any_other_company is 1239.1472537020952
The coefficient for Alcohol is 55.275409322244755
The coefficient for exercise is 27.588044827561752
The coefficient for weight_change_in_last_one_year is 157.39109500858117
The coefficient for fat_percentage is -3.9115757861908413
```

- ✓ The intercept in linear regression is the value of the dependent variable (outcome) when all the independent variables (predictors) are equal to zero. The intercept of our model is

```
The intercept for our model is -79773.05899871822
```

- ✓ Using intercept and coefficient data to build our model.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.944
Model:                            OLS   Adj. R-squared:                  0.944
Method:                 Least Squares   F-statistic:                 2.458e+04
Date:                Sun, 29 Jan 2023   Prob (F-statistic):               0.00
Time:                        12:43:26   Log-Likelihood:            -1.6709e+05
No. Observations:               17500   AIC:                         3.342e+05
Df Residuals:                   17487   BIC:                         3.343e+05
Df Model:                          12
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                    -8.147e+04    258.072   -315.706      0.000    -8.2e+04    -8.1e+04
adventure_sports               133.7482     93.648      1.428      0.153     -49.812     317.308
age                              1.8816      1.595      1.179      0.238      -1.246       5.009
cholesterol_level               21.9519     20.418      1.075      0.282     -18.070      61.973
weight                        1500.3236      2.964    506.250      0.000    1494.515    1506.133
covered_by_any_other_company  1181.8058     55.828     21.169      0.000    1072.377    1291.235
weight_change_in_last_one_year 178.2785     16.325     10.920      0.000     146.279     210.278
heart_decs_history             199.3866    115.658      1.724      0.085     -27.315     426.088
other_major_decs_history        75.8379     87.731      0.864      0.387     -96.123     247.799
Gender                           3.6811     55.030      0.067      0.947    -104.182     111.545
Location                        14.5518      5.925      2.456      0.014       2.937      26.166
Alcohol                         58.2363     37.807      1.540      0.123     -15.868     132.341
exercise                        30.7805     39.903      0.771      0.440     -47.433     108.994
==============================================================================
Omnibus:                      453.899   Durbin-Watson:                   1.989
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              503.196
Skew:                           0.377   Prob(JB):                    5.40e-110
Kurtosis:                       3.347   Cond. No.                        864.
==============================================================================

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

The following list contains the variables that affect how much insurance costs.

```
(-81474.76) * Intercept + (133.75) * adventure_sports + (1.88) * age + (21.95) * cholesterol_level + (1500.32) * weight + (118
1.81) * covered_by_any_other_company + (178.28) * weight_change_in_last_one_year + (199.39) * heart_decs_history + (75.84) * ot
her_major_decs_history + (3.68) * Gender + (14.55) * Location + (58.24) * Alcohol + (30.78) * exercise +
```

This is the equation for a multiple linear regression model, where the response variable is represented by the left hand side of the equation (-81474.76),

and the predictor variables are represented by the right hand side of the equation

(adventure_sports, age, cholesterol level, weight, covered_by_any_other_company, weight_change_in_last_one_year, heart_decs_history, other_major_decs_history, Gender, Location, Alcohol, exercise).

The coefficients in front of each predictor variable represent the change in the response variable per unit change in the predictor variable, while holding all other predictor variables constant.

For example, the coefficient in front of the variable "adventure_sports" is 133.75, which means that for every unit increase in adventure sports, the response variable is expected to increase by 133.75 units, holding all other predictor variables constant.

Similarly, for every unit increase in age, the response variable is expected to decrease by 1.88 units, holding all other predictor variables constant.

The intercept is the expected value of the response variable when all predictor variables are equal to zero, in this case -81474.76

- K-Nearest Neighbours (KNN):
  simple and widely used machine learning algorithm for both classification and regression tasks. It is considered an instance-based learning algorithm, as it operates by comparing an unknown example to the nearest examples in the training set.
  Training score is

  ```
  0.8327209956449816
  ```

  Testing score is

  ```
  0.7648587703052009
  ```

- Model building test score and train score and comparison of the below models
  - ❖ Linear Regression
  - ❖ Decision Tree Regressor
  - ❖ Random Forest Regressor
  - ❖ ANN Regressor

	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3353.844295	3406.938203	0.945243	0.943255
Decision Tree Regressor	0.000000	4352.715262	1.000000	0.907377
Random Forest Regressor	1166.999192	3125.393933	0.993370	0.952246
ANN Regressor	2989.885760	3145.103369	0.956483	0.951642

  RMSE is a measure of the difference between the observed values and the predicted values. In other words, it measures the accuracy of the model's predictions. A lower RMSE score indicates that the model's predictions are closer to the actual values and the model is more accurate

  Random forest regression is the better go to model

  Building models using ensemble technique

- Ensemble Learning – Gradient Boost
  Train score

  ```
  0.9556681127914469
  ```

  Test Score

  ```
  0.9542241171265452
  ```

- Ensemble Learning – Bagging
  Train score

  ```
  1.0
  ```

  Test Score

```
0.1624
```

- Naive Bayes Model

  Train score

```
0.18782857142857143
```

  Test Score

```
0.15786666666666666
```

- AdaBoostRegresson

  Train score

```
0.9483660802284467
```

  Test Score

```
0.9472235665438947
```

5. Model validation

If a comparison of multiple models using evaluation metrics such as Root Mean Squared Error (RMSE) or Mean Absolute Error (MAE) shows that the random forest model has the best fit, it may indeed be the best fit model for the problem at hand. However, it's important to consider other factors before making a final decision on which model to use.

Random forest is a popular and powerful machine learning algorithm that is widely used in regression problems. It is an ensemble method that creates multiple decision trees and combines their predictions to make the final prediction. This allows the model to capture complex relationships in the data and handle non-linearity.

However, it's important to consider other factors such as the interpretability of the model, the computational time required to train and make predictions, and the generalization performance of the model. In some cases, simpler models such as linear regression may be more interpretable and easier to use, even though they may not perform as well as the random forest model.

Additionally, it's important to validate the model on a separate test set to ensure that it generalizes well to unseen data. Overfitting, where a model fits the training data too well but performs poorly on test data, can be a concern with complex models such as random forest.

Ultimately, the best model will depend on the specific problem and the data, and it is important to carefully evaluate different models and choose the one that best fits the needs of the problem.

6. Final interpretation / recommendation

a) The data being used is well-distributed and can be used to create a user interface. In other words, the data is balanced in terms of its distribution and can be used to create a visual representation for the user to interact with.

b) Any This statement suggests that the data has been analysed and grouped into 4 clusters based on certain characteristics, such as weight and silhouette score.

- Cluster 1 is described as having low insurance costs, and certain attributes that are associated with this low cost. Specifically,

- it states that people in this cluster have a high rate of regular check-ups, do not participate in adventure sports, have low cholesterol levels,
- no history of heart disease, and are of lower weight. Additionally,
- it is stated that these individuals are not covered by other insurance companies and are considered to be health-conscious. This suggests that the data analysis has found a correlation between these characteristics and low insurance costs,
- and that people with these characteristics are placed in Cluster 1.
- Cluster 4 is described as having high insurance costs, and certain attributes that are associated with this high cost.
- Specifically, it states that people in this cluster have high cholesterol levels, high age, high weight, and they are covered by other insurance companies and are considered not to be health-conscious.
- This suggests that the data analysis has found a correlation between these characteristics and high insurance costs, and that people with these characteristics are placed in Cluster 4. These attributes such as High cholesterol, age, weight and being covered by other company are considered as risk factors that may lead to higher insurance cost.

c) Business insights

- Network of providers and facilities play a crucial role in determining the cost and coverage of a health insurance policy. It would be wise to have a broad network of providers and facilities to offer more options to policy holders.
- Utilizing data and analytics can help identify patterns and trends in healthcare costs and utilization, which can be used to predict future costs and design more effective insurance products.
- Building a strong digital platform for policyholders to access their insurance information, claims, and network providers can improve customer experience and help reduce administrative costs.
- Collaborating with other health-related companies like hospitals, clinics, and pharmacies can help improve the overall health of the population, which can lead to lower healthcare costs.
- It's important to comply with the changing regulations in the health insurance industry and keep up with the market trends and emerging technologies to stay competitive in the market