HEALTHCARE CAPSTONE PROJECT

Anushree Pandey
PGP-DSBA Online
January' 21
Date: 10/04/2022

Table of Contents

List of Figures

List of Tables

Review Parameters :

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

The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual. We have to use the health and habit related parameters for the estimated cost of insurance.

c) Understanding business/social opportunity

Anyone with a healthcare policy pays a monthly insurance premium. A health insurance company gathers the premiums it collects from thousands of customers into a pool. When one of those customers needs coverage for medical care, the insurance company uses money from this pool to pay for it in the form of a claim. A health insurer will also use premiums to pay for the costs of doing business. With the passing of the ACA, the law requires insurance companies to spend 80/85% on claims and 20/15% on administrative costs. The law regulates the amount of income based on the premium charged. Other costs that you pay for your health services (such as copayments and coinsurance) are paid to your healthcare provider (doctors and hospitals), NOT to the insurance company.

So if the company can manage to lower the insurance cost, it will help to take more people under the umbrella. It will be feasible for the general community to get their risks covered from the insurance policies.

2) Data Report

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

This data was collected year-wise from the year 1990 to 2010. Surveys conducted by government agencies like Centers for Disease Control and Prevention/National Center for Health Statistics/ National Health Interview Survey (NHIS) help maintain various databases with lots of patient records. With the digitisation of the data, it's a lot easier to maintain and access these records gathered over the years.

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

Shape attribute tells us number of observations and variables we have in the data set. It is used to check the dimension of data. The data set has 25000 observations and 24 variables in the data set. This dataset contains
no. of rows: 25000
no. of columns: 24

Table 1 - Descriptive statistics of the dataset

	applicant_id	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	Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost
count	25000.00	25000.00	25000.00	25000.00	25000.00	25000.00	25000.00	25000.00	25000.00	25000.00	24010.00	13119.00	25000.00	25000.00	25000.00	25000.00
mean	17499.50	4.089040	0.773680	0.081720	3.104200	5215.889	44.91832	0.054640	0.098160	167.5300	31.39332	2003.892	71.61048	2.517960	28.81228	27147.40
std	7217.022	2.606612	1.199449	0.273943	1.141663	1053.179	16.10749	0.227281	0.297537	62.72971	7.876535	7.581521	9.325183	1.690335	8.632382	14323.69
min	5000.000	0.000000	0.000000	0.000000	0.000000	2034.000	16.00000	0.000000	0.000000	57.00000	12.30000	1990.000	52.00000	0.000000	11.00000	2468.000
25%	11249.75	2.000000	0.000000	0.000000	2.000000	4543.000	31.00000	0.000000	0.000000	113.0000	26.10000	1997.000	64.00000	1.000000	21.00000	16042.00
50%	17499.50	4.000000	0.000000	0.000000	3.000000	5089.000	45.00000	0.000000	0.000000	168.0000	30.50000	2004.000	72.00000	3.000000	31.00000	27148.00
75%	23749.25	6.000000	1.000000	0.000000	4.000000	5730.000	59.00000	0.000000	0.000000	222.0000	35.60000	2010.000	78.00000	4.000000	36.00000	37020.00
max	29999.00	8.000000	5.000000	1.000000	12.00000	11255.00	74.00000	1.000000	1.000000	277.0000	100.6000	2018.000	96.00000	6.000000	42.00000	67870.00

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

applicant_id is the applicant's unique ID
years_of_insurance_with_u s is since how many years customer is taking policy from the same company only
regular_checkup_lasy_year is the number of times customers has done the regular health check up in last one year
adventure_sports shows whether Customer is involved with adventure sports like climbing, diving etc. or not.
Occupation is the occupation of the customer whether they are student or salaried or business person
visited_doctor_last_1_year is the number of times customer has visited doctor in last one year
cholesterol_level is the cholesterol level of the customers while applying for insurance
daily_avg_steps is average daily steps walked by customers
age is age of the customer
heart_decs_history is whether the person has had any past heart diseases
other_major_decs_history is whether the person has had any past major diseases apart from heart like any operation
Gender is gender of the customer
avg_glucose_level is average glucose level of the customer while applying the insurance
bmi is the BMI of the customer while applying the insurance
smoking_status is the smoking status of the customer
Year_last_admitted is record of when customer have been admitted in the hospital last time
Location is the location of the hospital
weight is weight of the customer
covered_by_any_other_company is whether the customer is covered from any other insurance company
Alcohol is alcohol consumption status of the customer
exercise is regular exercise status of the customer
weight_change_in_last_one _year is how much variation has been seen in the weight of the customer in last year
fat_percentage is the fat percentage of the customer while applying the insurance

insurance_cost is the total Insurance cost

Here, is the information about the dataset and it's memory usage.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   applicant_id                    25000 non-null  int64
 1   years_of_insurance_with_us      25000 non-null  int64
 2   regular_checkup_lasy_year       25000 non-null  int64
 3   adventure_sports                25000 non-null  int64
 4   Occupation                      25000 non-null  object
 5   visited_doctor_last_1_year      25000 non-null  int64
 6   cholesterol_level               25000 non-null  object
 7   daily_avg_steps                 25000 non-null  int64
 8   age                             25000 non-null  int64
 9   heart_decs_history              25000 non-null  int64
 10  other_major_decs_history        25000 non-null  int64
 11  Gender                          25000 non-null  object
 12  avg_glucose_level               25000 non-null  int64
 13  bmi                             24010 non-null  float64
 14  smoking_status                  25000 non-null  object
 15  Year_last_admitted              13119 non-null  float64
 16  Location                        25000 non-null  object
 17  weight                          25000 non-null  int64
 18  covered_by_any_other_company    25000 non-null  object
 19  Alcohol                         25000 non-null  object
 20  exercise                        25000 non-null  object
 21  weight_change_in_last_one_year  25000 non-null  int64
 22  fat_percentage                  25000 non-null  int64
 23  insurance_cost                  25000 non-null  int64
dtypes: float64(2), int64(14), object(8)
memory usage: 4.6+ MB
```

This data has 24 columns among which 22 have non-null values and 2 have null values. The 2 columns with null values are bmi and year_last_admitted.
3 data types are present: Integer, float, and object.
The dataset is using 4.6MB memory.

Duplicate records

Number of duplicate rows = 0

Renamed the column regular_checkup_lasy_year to regular_checkup_last_year
Renamed the column cholesterol_level features as follows :

Cholesterol level	Replaced value
125 to 150	1
150 to 175	2
175 to 200	3
200 to 225	4
225 to 250	5

3) Exploratory data analysis

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations.

a) Univariate Analysis

Univariate analysis is the simplest form of analyzing data. “Uni” means “one”, so in other words the data has only one variable. It doesn’t deal with causes or relationships (unlike regression) and its major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

Histogram plot of Numerical Data

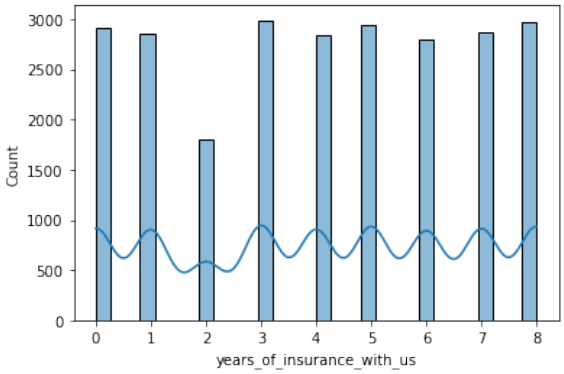


Fig 1 - Histogram plot of years_of_insurance_with_us

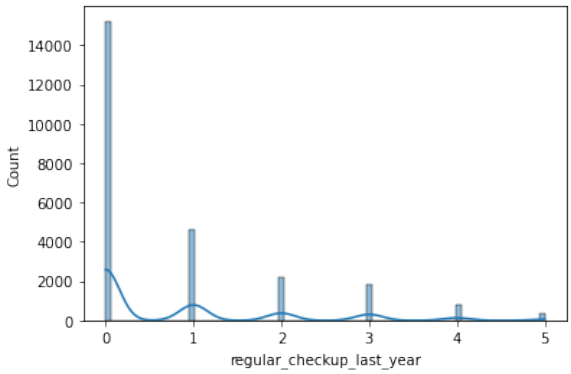


Fig 2 - Histogram plot of regular_checkup_last_year

This distribution in Fig 1 is multimodal. **Multimodal** is what we call a distribution when it has more than two prominent peaks. Most people have had 3 years of insurance with the company. Fig 2 is right skewed. With right-skewed distribution (also known as "positively skewed" distribution), most data falls to the right, or positive side, of the graph's peak. Thus, **the histogram skews in such a way that it’s right side (or "tail") is longer than its left side.** Here, most people have had no checkups in last year and least number of people have gotten checkups upto 5 times.

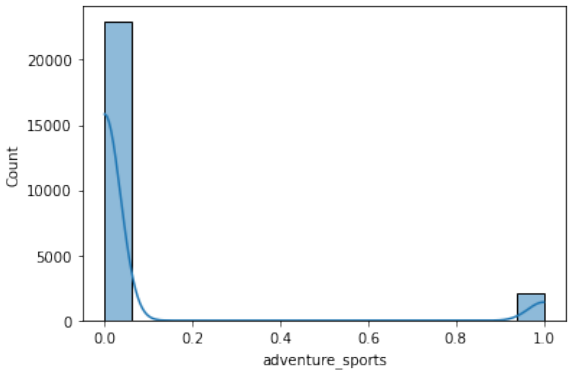


Fig 3 - Histogram plot of adventure_sports

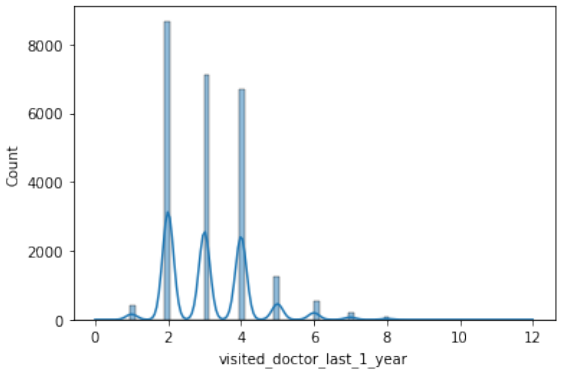


Fig 4 - Histogram plot of visited_doctor_last_1_year

Fig 3 shows that very few people play any adventure sports. Fig 4 is a right skewed distribution graph. Peak is at two visits which means mostly people have visited the doctor twice in last one year. Majority of people have visited the doctor more than once in the past year.

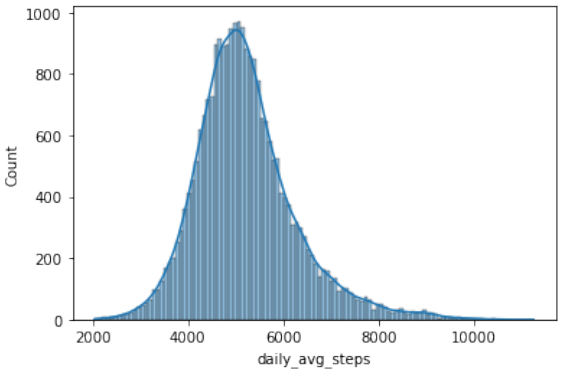


Fig 5 - Histogram plot of daily_avg_steps

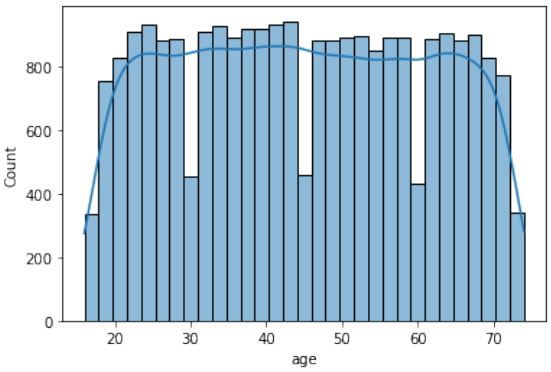


Fig 6 - Histogram plot of age

Fig 5 is a **unimodal** histogram with one prominent peak. It is right skewed which means most of the people walk more than 5000 steps daily on an average. Fig 6 is a uniform model with no prominent peaks. In this dataset, the number of 22 years old people is highest.

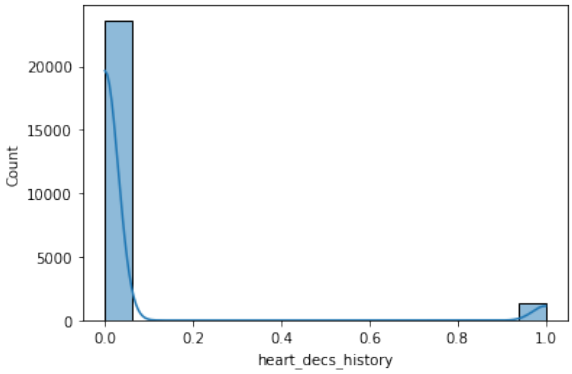


Fig 7 - Histogram plot of heart_decs_history

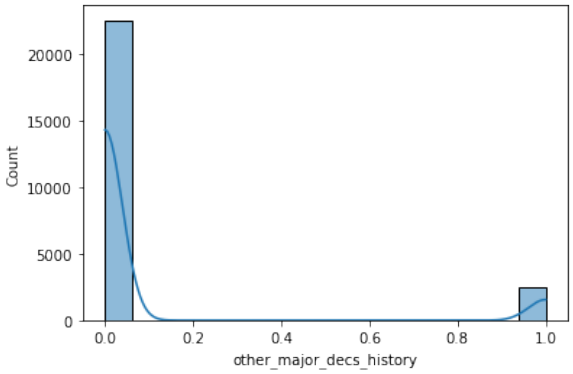


Fig 8 - Histogram plot of other_major_decs_history

Fig 7 is histogram of heart disease history which shows that most people have not had any heart attacks. Fig 8 is histogram of any other major disease which shows that mostly people are healthy and have not had any major diseases.

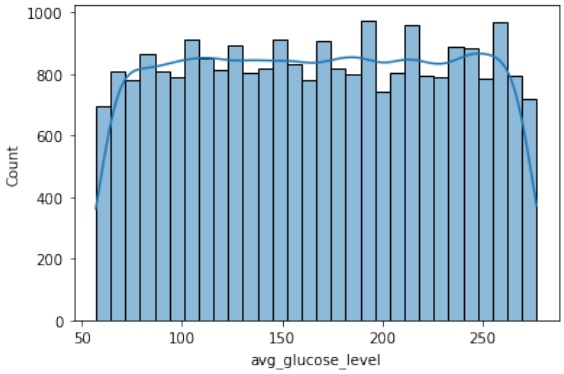


Fig 9 - Histogram plot of avg_glucose_level

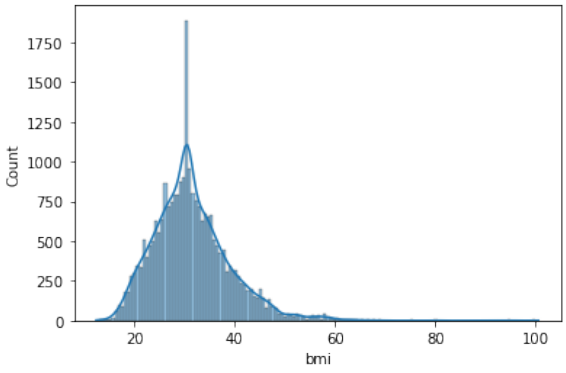


Fig 10 - Histogram plot of bmi

Fig 9 is also a uniform model with no prominent peak. It shows that maximum people have an average glucose level of 243 mg/dL. The range of reported average glucose levels is 57 to 277 mg/dL. Fig 10 is a unimodal histogram with right skewed tail. Mostly people have a bmi of 30.5.

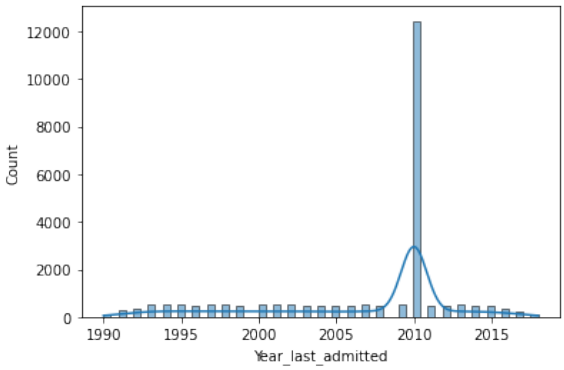


Fig 11 - Histogram plot of Year_last_admitted

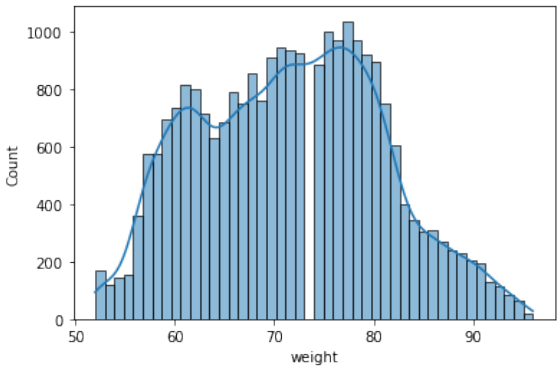


Fig 12 - Histogram plot of weight

Fig 11 is a unimodal with one prominent peak which is left skewed. It means that most number of people were admitted into the hospital before 2010. Fig 12 is a bimodal histogram. Most people have 77 kg weight in this dataset. The weight range is from 52 to 96 kg.

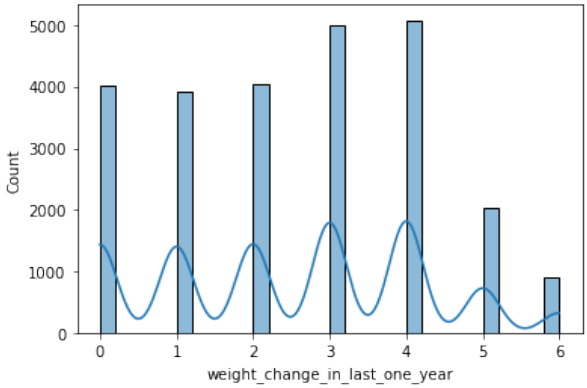


Fig 13 - Histogram plot of weight_change_in_last_one_year

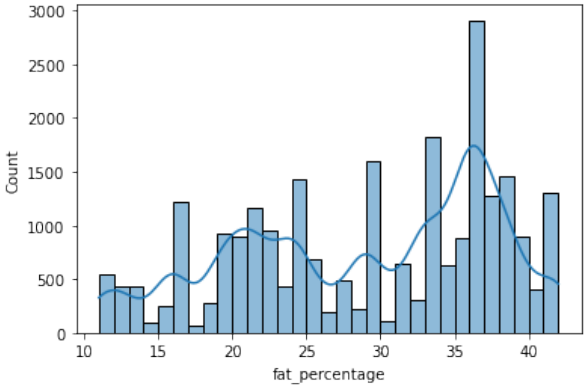


Fig 14 - Histogram plot of fat_percentage

Fig 13 is a multimodal histogram which shows the weight change in last one year. Most people have had 4 weight fluctuations in last one year. Fig 14 is a multimodal histogram with left skewness. It means that most people have fat percentage of less than 36%.

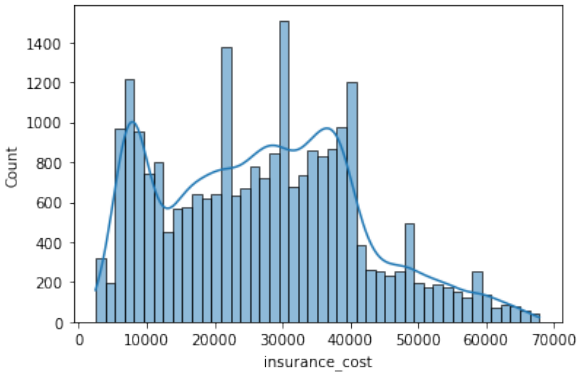


Fig 15 - Histogram plot of insurance_cost

Fig 15 is a multimodal histogram representing insurance cost data. It is right skewed which means that the insurance cost is mostly higher than 7000 and most number of people can afford to pay the insurance cost of 7404.

Histogram plot of Categorical data

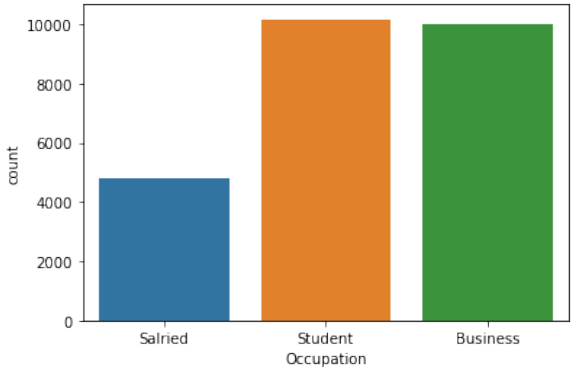


Fig 16 - Countplot of Occupation

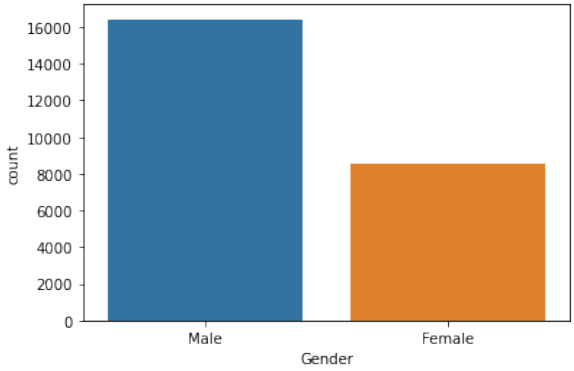


Fig 17 - Countplot of Gender

Fig 16 is a countplot of occupation of people whose data has been collected for this dataset. Among these people maximum number people are students. Fig 17 is countplot of gender which shows that the number of males is higher than the number of females.

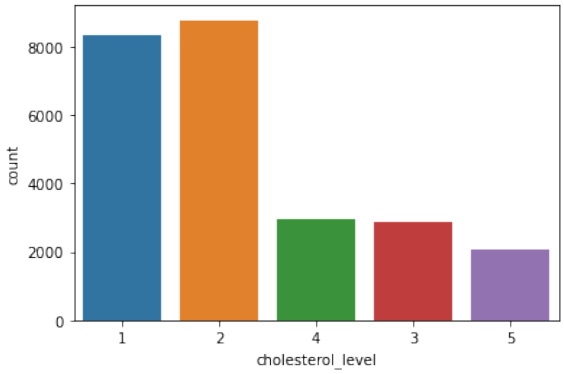


Fig 18 - Countplot of cholesterol_level

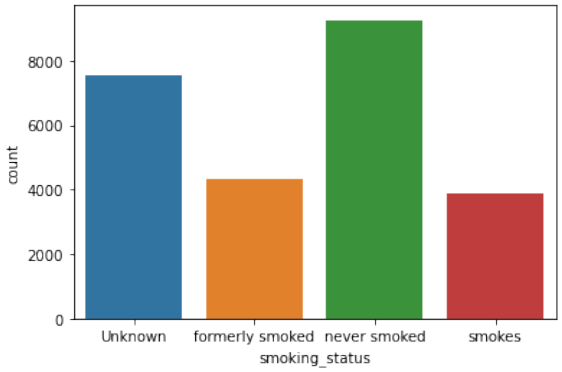


Fig 19 - Countplot of smoking_status

Fig 18 is a countplot of cholesterol level which shows that maximum number of people have a cholesterol level of 150-175 mg/dL. Fig 19 is the countplot of smoking status of the people. Maximum number of people have never smoked.

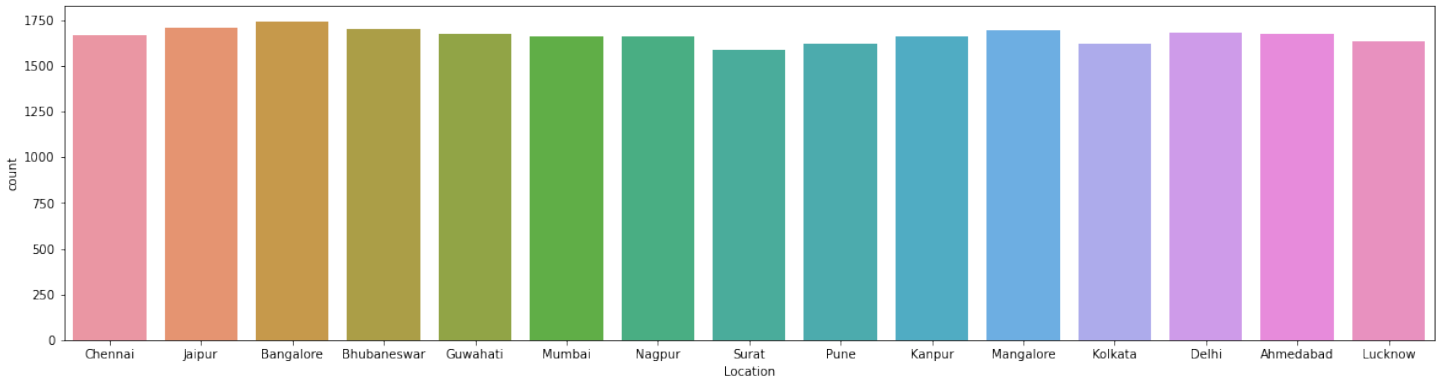


Fig 20 - Countplot of Location

Fig 20 is the countplot of location with the hospitals. Maximum number of hospitals where people were admitted were in Bangalore.

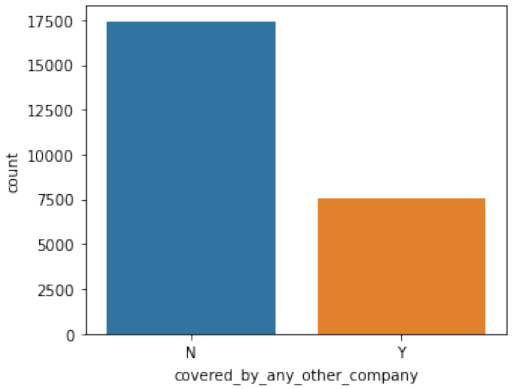


Fig 21 - Countplot of covered_by_any_other_company

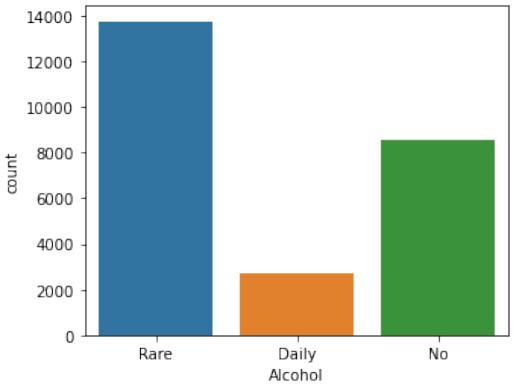


Fig 22 - Countplot of alcohol

Fig 21 is countplot of whether people were covered by any other insurance company or not. Mostly people were not covered by other companies. Fig 22 is the count plot of alcohol use. Mostly people rarely drink.

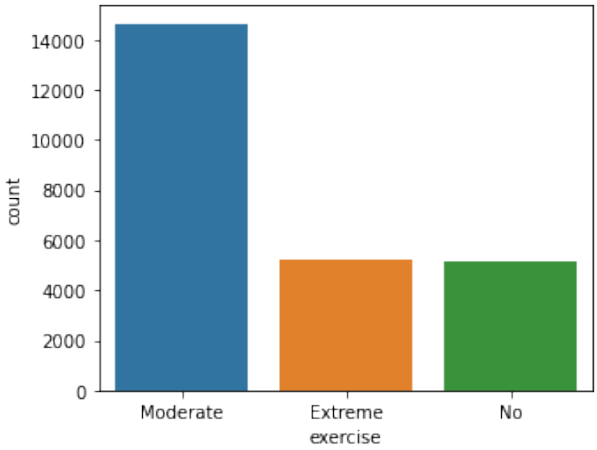


Fig 23 - Countplot of exercise

Fig 23 is the countplot of the exercise status of people. Mostly people are doing moderate exercise.

b) Bivariate analysis

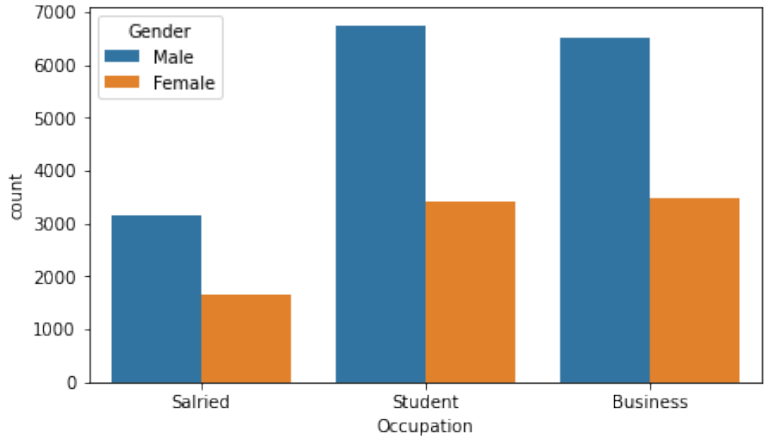


Fig 24 - Countplot of occupation based on gender

In Fig 24 maximum number of people are student among which number of males is higher than the females.

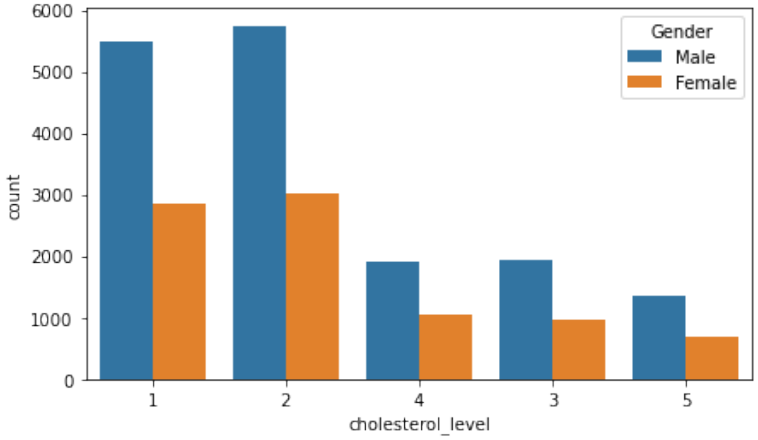


Fig 25 - Countplot of cholesterol level based on gender

In Fig 25 maximum number of people have 150-175 mg/dL cholesterol level among which number of males is higher than the females.

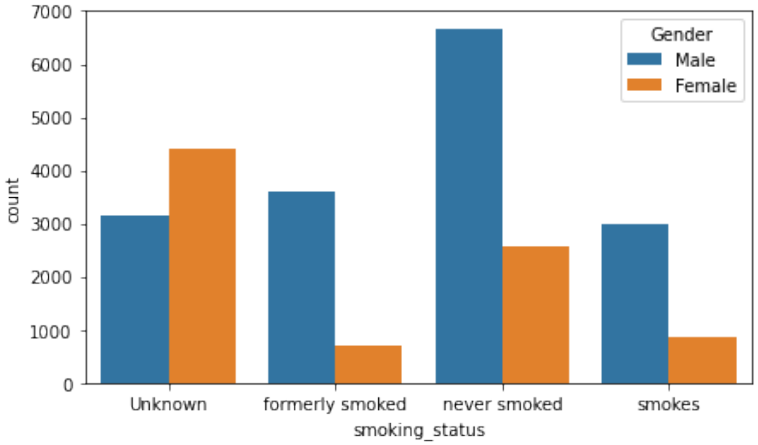


Fig 26 - Countplot of smoking status based on gender

In Fig 26 maximum number of people have never smoked among which number of males is higher than the females. The number of unknown status is 2nd highest in females while in males it is at 3rd position.

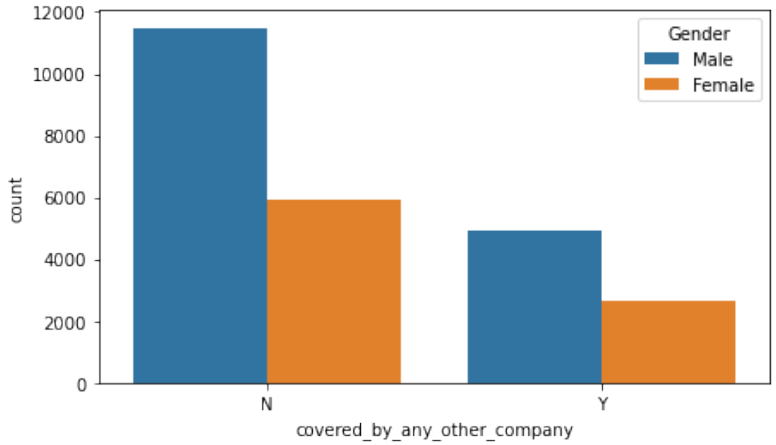


Fig 27 - Countplot of covered by any other company based based on gender

In Fig 27 maximum number of people are not covered by any other company among which number of males is higher than the females.

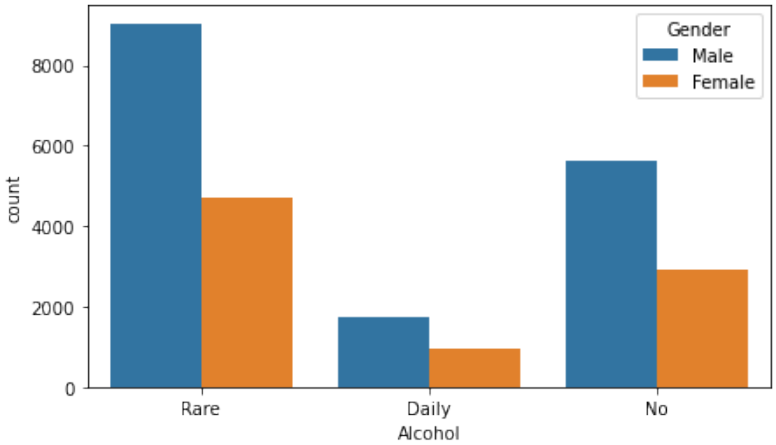


Fig 28 - Countplot of alcohol based on gender

In Fig 28 maximum number of people rarely drink among which number of males is higher than the females.

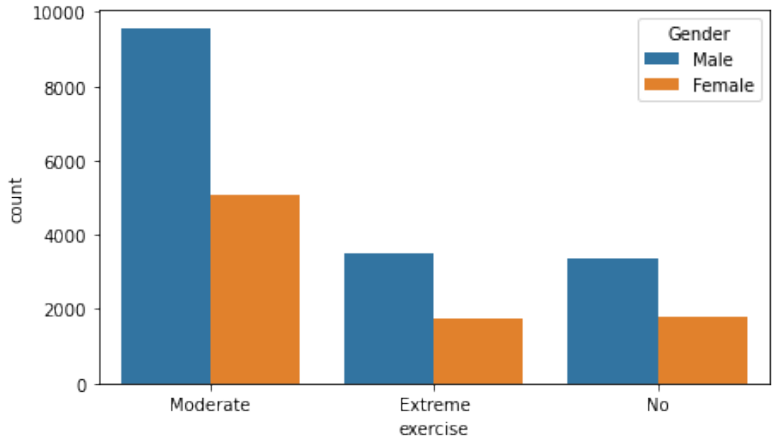


Fig 29 - Countplot of exercise based on gender

In Fig 29 maximum number of people exercise in moderate quantity among which number of males is higher than the females.

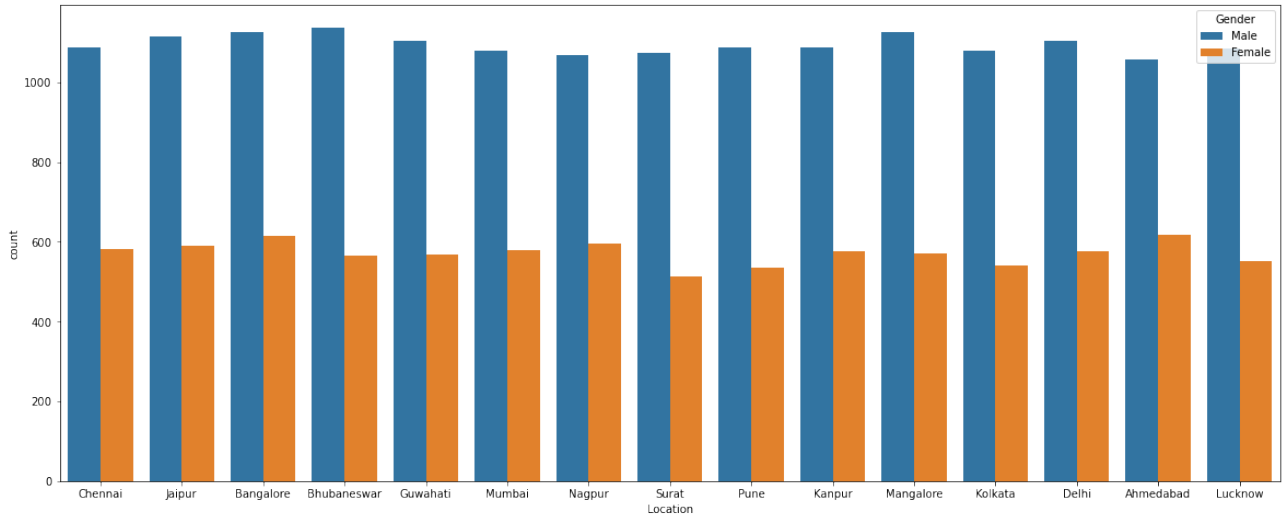


Fig 30 - Countplot of location based on gender

In Fig 30 maximum number of people were admitted in hospitals in Bangalore among which number of males is higher than the females.

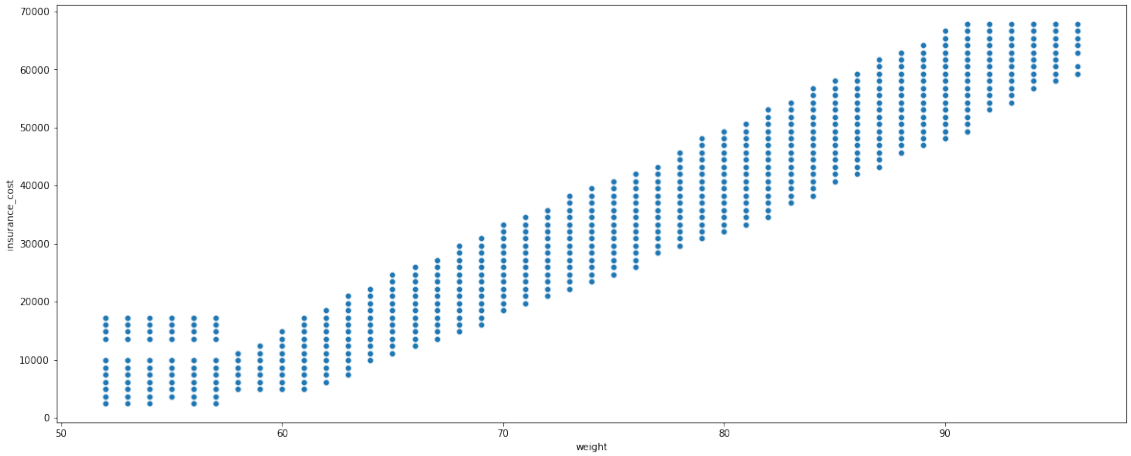


Fig 31 - Scatterplot between weight and insurance cost

In Fig 31 the insurance cost is increasing with an increase in the weight of customers. There is an upward trend in the graph.

Crosstabs for bivariate analysis

Table 2 - Between heart_decs_history and gender

Gender	Female	Male	All
heart_decs_history			
0	0.33676	0.60860	0.94536
1	0.00636	0.04828	0.05464
All	0.34312	0.65688	1.00000

The percentage of males is 65.68% while females have a percentage of 34.31%. 5.4% people have had heart disease while 94.53 have not.

Table 3 - Between years_of_insurance_with_us and gender

Gender	Female	Male	All
years_of_insurance_with_us			
0	0.03848	0.07800	0.11648
1	0.03756	0.07668	0.11424
2	0.02456	0.04776	0.07232
3	0.04336	0.07624	0.11960
4	0.04100	0.07284	0.11384
5	0.03888	0.07876	0.11764
6	0.03880	0.07336	0.11216
7	0.03892	0.07600	0.11492
8	0.04156	0.07724	0.11880
All	0.34312	0.65688	1.00000

11.9% people have had 3 years of insurance with company which is the highest.

Table 4 - Between regular_checkup_last_year and gender

Gender	Female	Male	All
regular_checkup_l			
0	0.20596	0.40264	0.60860
1	0.06460	0.12116	0.18576
2	0.03160	0.05632	0.08792
3	0.02544	0.04728	0.07272
4	0.01096	0.02012	0.03108
5	0.00456	0.00936	0.01392
All	0.34312	0.65688	1.00000

60.08% customers have had no checkups last year.

Table 5 - Between adventure_sports and gender

Gender	Female	Male	All
adventure_sports			
0	0.31448	0.60380	0.91828
1	0.02864	0.05308	0.08172
All	0.34312	0.65688	1.00000

91.8% customers don’t play any adventure sports.

Table 6 - Between visited_doctor_last_1_year and gender

Gender	Female	Male	All
visited_doctor_last			
0	0.00000	0.00004	0.00004
1	0.00588	0.01140	0.01728
2	0.11956	0.22720	0.34676
3	0.09796	0.18580	0.28376
4	0.09012	0.17820	0.26832
5	0.01832	0.03228	0.05060
6	0.00736	0.01448	0.02184
7	0.00256	0.00500	0.00756
8	0.00120	0.00184	0.00304
9	0.00008	0.00044	0.00052
10	0.00008	0.00016	0.00024
12	0.00000	0.00004	0.00004
All	0.34312	0.65688	1.00000

34.67% customers have visited the doctor twice last year.

Table 7 - Between occupation and gender

Occupation	Business	Salaried	Student	All
Gender				
Female	0.13976	0.06656	0.13680	0.34312
Male	0.26104	0.12588	0.26996	0.65688
All	0.40080	0.19244	0.40676	1.00000

40.67% customers are student, 40.08% customers are business people and 19.24% are salaried.

Boxplot

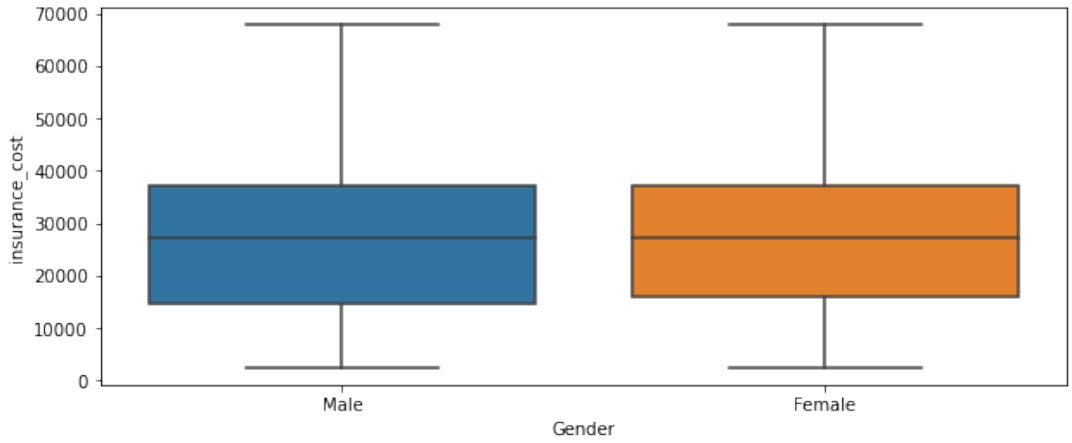


Fig 32 - Boxplot between gender and insurance cost

Fig 32 is box plot between gender and insurance cost in which there are no outliers and the median is also same.

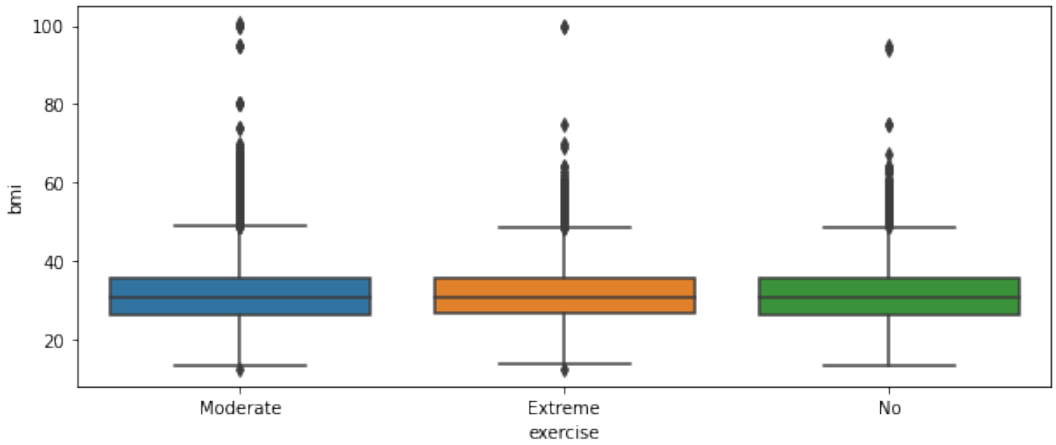


Fig 33 - Boxplot between exercise and bmi

Fig 33 is box plot between exercise and bmi in which there are outliers present and the median is same for all.

Pairplot

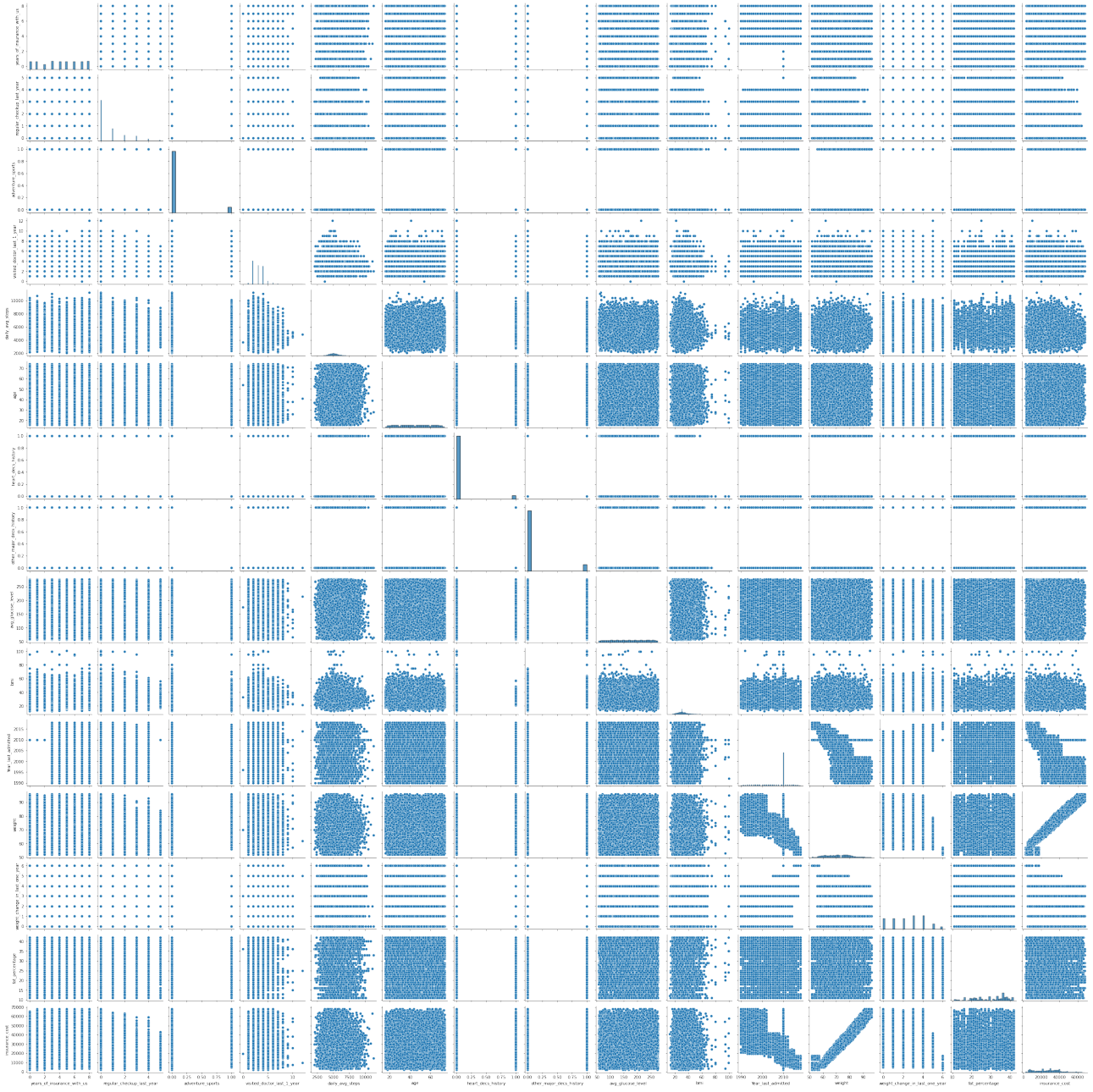


Fig 34 - pairplot for bivariate analysis

Fig 34 is a pairplot used for the bivariate analysis of the dataset. In the above plot scatter diagrams are plotted for all the numerical columns in the dataset. A scatter plot is a visual representation of the degree of correlation between any two columns. The pair plot function in seaborn makes it very easy to generate joint scatter plots for all the columns in the data.

Table 8 - Correlation table of variables among themselves

	years_of_insurance_with_us	regular_checkup_last_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	other_major_decs_history	avg_glucose_level	bmi	Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost
years_of_insurance_with_us	1.000000	0.018549	0.016419	0.002985	-0.001480	0.004100	-0.001123	-0.010548	-0.000174	-0.013185	-0.266589	-0.006285	-0.000608	-0.003302	0.001404
regular_checkup_last_year	0.018549	1.000000	0.009298	-0.005826	-0.001928	0.009033	0.000022	-0.009596	0.014459	-0.008421	0.130511	-0.142492	-0.012990	0.002190	-0.174005
adventure_sports	0.016419	0.009298	1.000000	0.011143	-0.000671	-0.002440	0.002808	0.003661	-0.005933	-0.002196	-0.062013	0.073938	-0.047270	0.003053	0.074561
visited_doctor_last_1_year	0.002985	-0.005826	0.011143	1.000000	-0.156888	-0.001247	-0.003444	0.009338	0.008147	0.000630	-0.010612	0.012098	-0.012733	-0.043455	0.008890
daily_avg_steps	-0.001480	-0.001928	-0.000671	-0.156888	1.000000	-0.000313	0.007256	-0.003661	0.000482	-0.005585	0.004260	-0.005768	0.008348	0.045827	-0.006565
age	0.004100	0.009033	-0.002440	-0.001247	-0.000313	1.000000	-0.003545	0.004386	-0.011551	-0.014744	-0.008515	0.001676	-0.004235	-0.007946	0.005195
heart_decs_history	-0.001123	0.000022	0.002808	-0.003444	0.007256	-0.003545	1.000000	0.107015	-0.005188	0.036496	0.002583	-0.004490	0.005151	-0.003600	-0.000445
other_major_decs_history	-0.010548	-0.009596	0.003661	0.009338	-0.003661	0.004386	0.107015	1.000000	0.000937	0.157325	0.001698	-0.003851	0.001744	0.000587	-0.002268
avg_glucose_level	-0.000174	0.014459	-0.005933	0.008147	0.000482	-0.011551	-0.005188	0.000937	1.000000	-0.018889	0.005617	-0.004684	0.000669	-0.000498	-0.005007
bmi	-0.013185	-0.008421	-0.002196	0.000630	-0.005585	-0.014744	0.036496	0.157325	-0.018889	1.000000	0.008875	-0.007550	0.017853	-0.002963	-0.007966
Year_last_admitted	-0.266589	0.130511	-0.062013	-0.010612	0.004260	-0.008515	0.002583	0.001698	0.005617	0.008875	1.000000	-0.607948	0.252589	0.004089	-0.610714
weight	-0.006285	-0.142492	0.073938	0.012098	-0.005768	0.001676	-0.004490	-0.003851	-0.004684	-0.007550	-0.607948	1.000000	-0.370670	-0.007377	0.970357
weight_change_in_last_one_year	-0.000608	-0.012990	-0.047270	-0.012733	0.008348	-0.004235	0.005151	0.001744	0.000669	0.017853	0.252589	-0.370670	1.000000	0.013273	-0.342710
fat_percentage	-0.003302	0.002190	0.003053	-0.043455	0.045827	-0.007946	-0.003600	0.000587	-0.000498	-0.002963	0.004089	-0.007377	0.013273	1.000000	-0.008486
insurance_cost	0.001404	-0.174005	0.074561	0.008890	-0.006565	0.005195	-0.000445	-0.002268	-0.005007	-0.007966	-0.610714	0.970357	-0.342710	-0.008486	1.000000

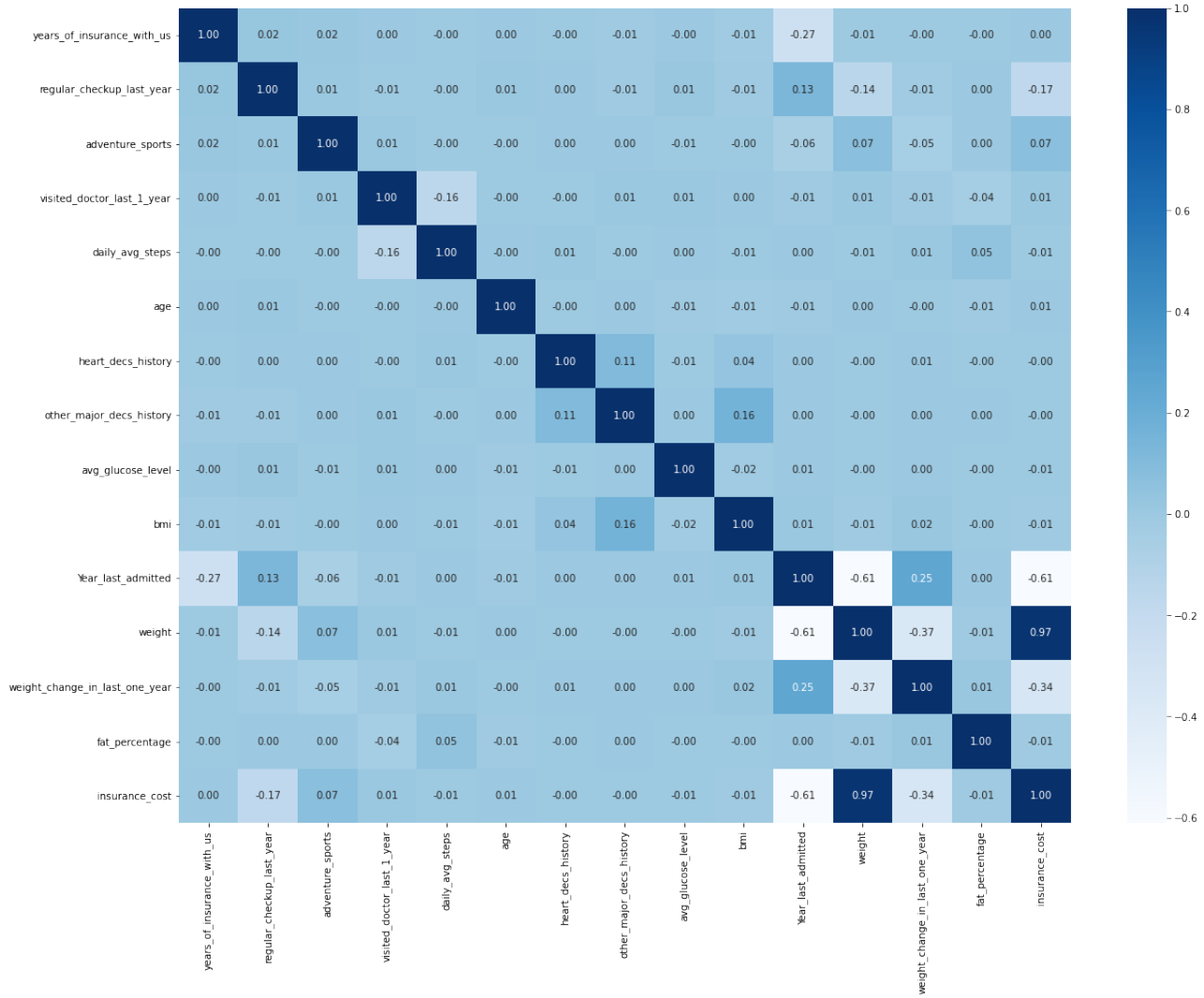


Fig 35 - Correlation heatmap

Fig 35 is a correlation heatmap. It shows that there is a very high positive correlation between weight of the customer and the insurance cost. Then there is some correlation between weight change in last one year and the number of times the customer was admitted in a hospital. There is high negative correlation between weight and the number of times admitted last year.

Multivariate Analysis

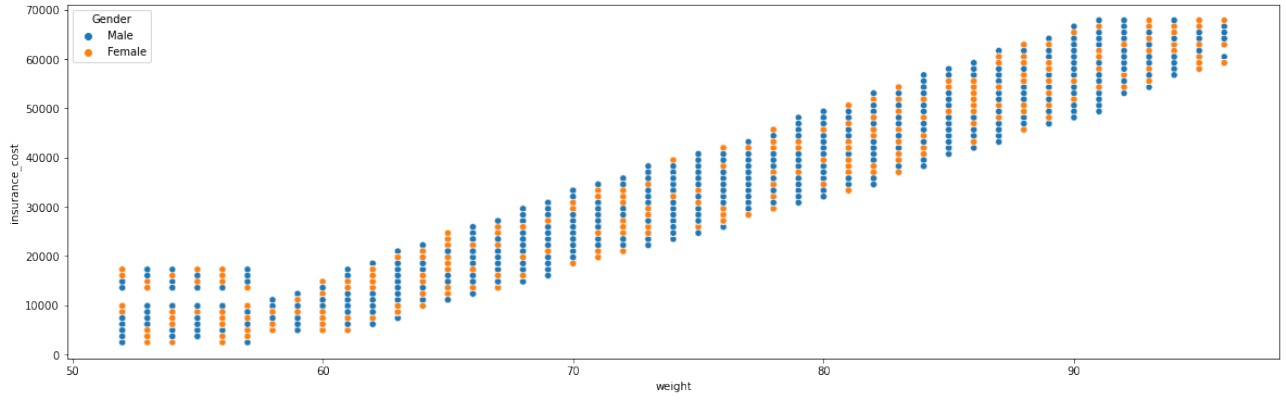


Fig 36 - Scatterplot between weight and insurance cost based on gender

Fig 36 is a scatterplot between weight and insurance cost based on gender. Here, there is an upward trend which means that with increase in weight, the insurance cost is also increasing.

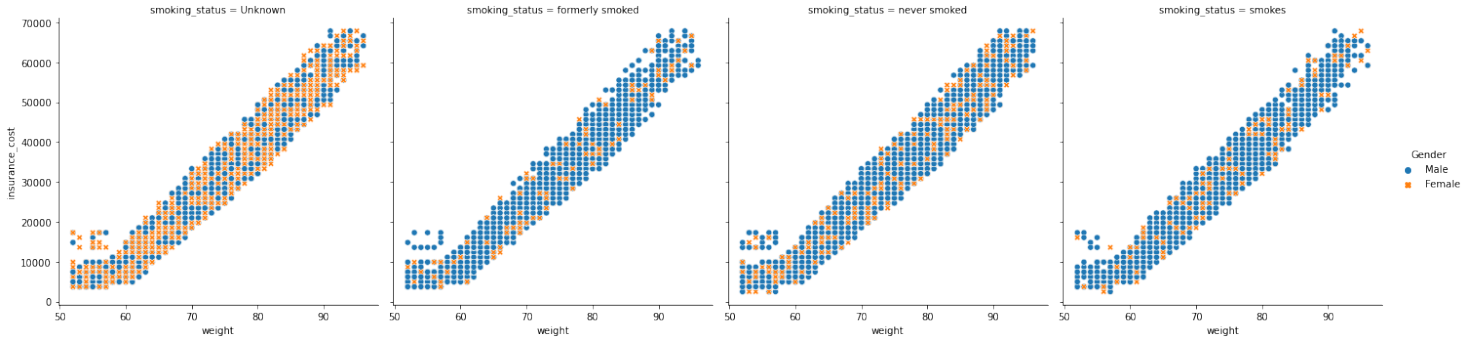


Fig 37 - Scatterplot between weight and insurance cost based on gender, columns are separated by smoking status

Fig 37 is a scatterplot between weight and insurance cost based on gender, columns are separated by smoking status. Here, we can see that number of females is higher in unknown category while number of males is higher in never smoked category.

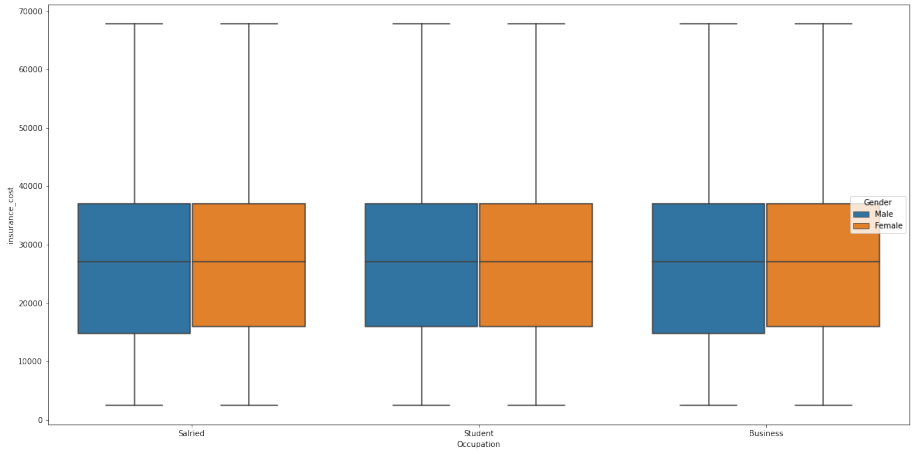


Fig 38 - Boxplot between occupation and insurance cost based on gender

Fig 38 is Boxplot between occupation and insurance cost based on gender which shows that the insurance cost on male and female is same for student category. Insurance cost slightly more for males in other occupations.

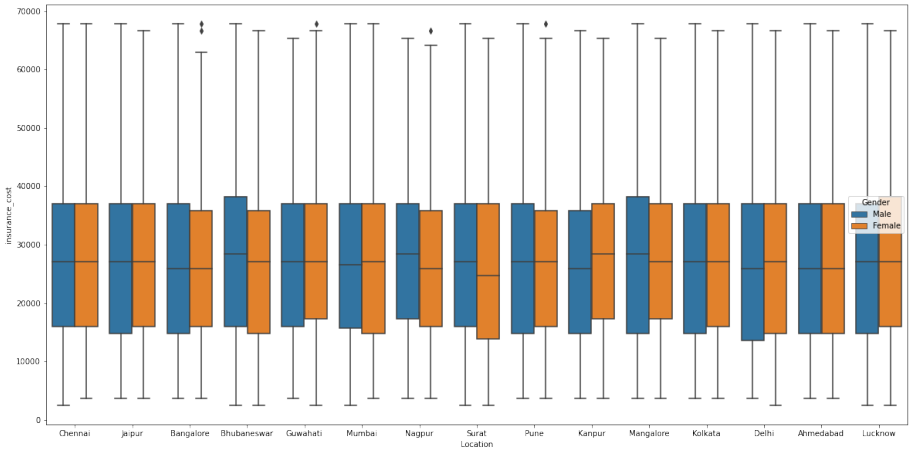


Fig 39 - Boxplot between location and insurance cost based on gender

Fig 39 is Boxplot between location and insurance cost based on gender which shows that insurance cost is high in Bhubaneswar. Males have higher insurance cost than females.

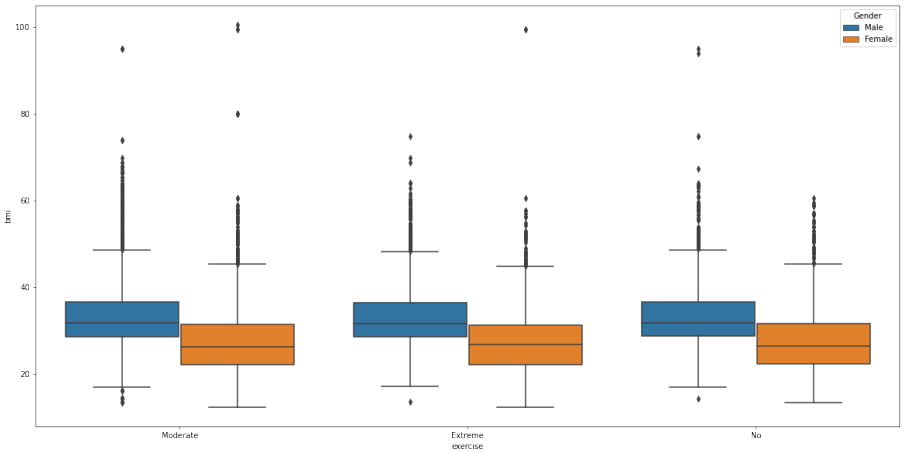


Fig 40 - Boxplot between exercise and bmi based on gender

Fig 40 is box plot between exercise and bmi based on gender which shows that exercise is not effecting bmi much and bmi is higher in males than females.

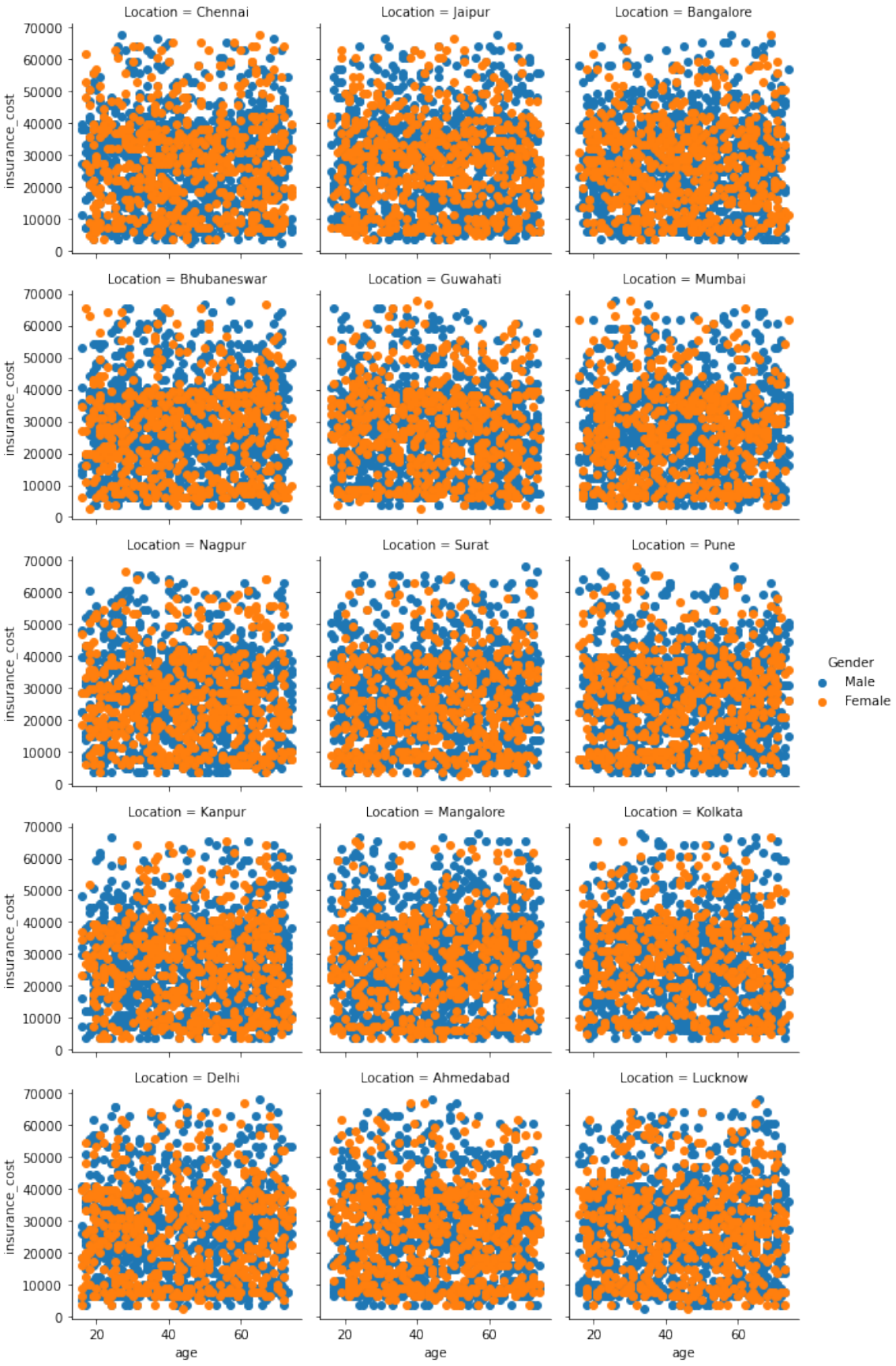


Fig 41 - Facetgrid between insurance cost and age with respect to various location

Fig 41 is a facetgrid between insurance cost and age with respect to various locations of the hospital. The data is differentiated on the basis of gender When exploring multi-dimensional data,

a useful approach is to draw multiple instances of the same plot on different subsets of your dataset. This technique is sometimes called either “lattice” or “trellis” plotting, and it is related to the idea of “small multiples”. The FacetGrid class is useful when we want to visualize the distribution of a variable or the relationship between multiple variables separately within subsets of the dataset.

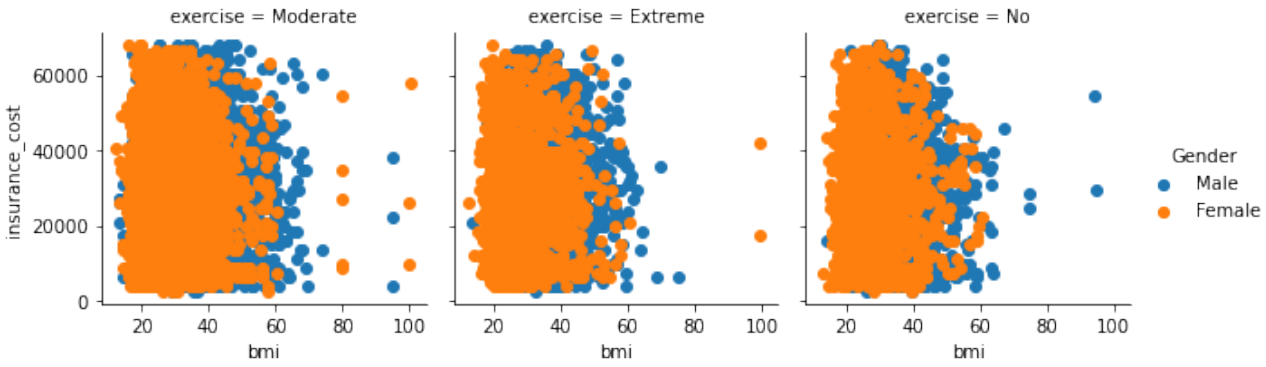


Fig 42 - Facetgrid between insurance cost and bmi with respect to exercise based on gender

Fig 42 is a facetgrid between insurance cost and bmi with respect to exercise based on gender. Here, moderate exercise people have more insurance cost among which cost on males is higher.

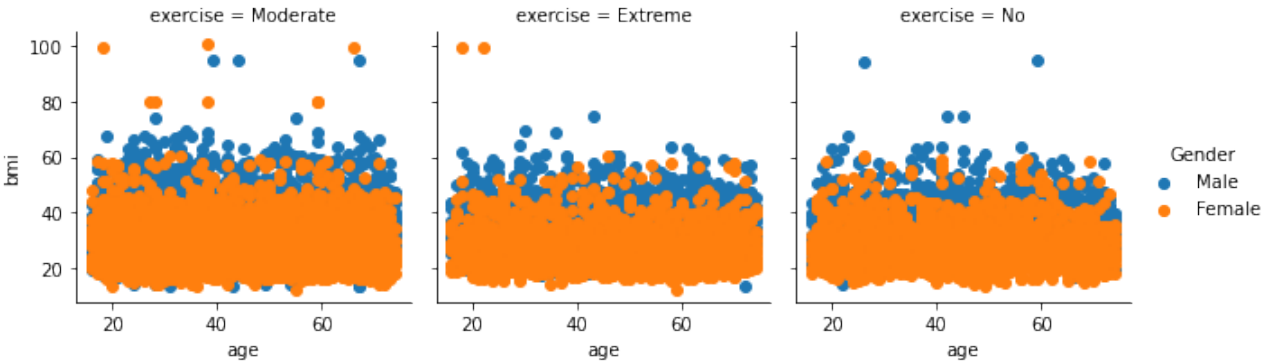


Fig 43 - Facetgrid between age and bmi with respect to exercise based on gender

Fig 43 is a facetgrid between age and bmi with respect to exercise based on gender. People with moderate exercising habits have higher bmi which is decreasing as the age increases. Males have higher bmi than females irrespective of the age.

c) Removal of unwanted variables

The variable applicant_id had no further use in data processing so it was removed.

d) Missing Value treatment

Following columns had missing values :

```
bmi                    990
Year_last_admitted    11881
dtype: int64
```

Among the total values of 25000, 990 is missing in bmi column which is 3.96%. So should not drop the column. We needed to impute this column for missing data.
Among the total values of 25000, 11881 is missing in Year_last_admitted column which is 47.524%. Since the percentage of missing values in the column 'Year_last_admitted' is approximately half of the total values, we will have to decide whether to impute this column or not. Because approximately half of the data in this column will be artificially provided which will not give accurate models to be applied in real time data.
Thus, we need to perform statistical analysis with the help of ANOVA between the columns 'Year_last_admitted' and 'insurance_cost' to find out the significance of this column in the dataset.

Table 9 - ANOVA table

	df	sum_sq	mean_sq	F \
C(Year_last_admitted)	28.0	1.990551e+12	7.109109e+10	1169.434999
Residual	13090.0	7.957538e+11	6.079097e+07	NaN
	PR(>F)			
C(Year_last_admitted)	0.0			
Residual	NaN			

F-Distribution or F-Statistic is the ratio of MSB to MSW.
It gives the degree of how relatively greater the difference is 'between group means' (MSB) compared to 'within group variance' (MSW).
If the ratio is greater than expected, will mean that not all the group means are the same and the mean values are substantially different which is significant here.
Also with the correlation plot it is clear that year_last_admitted has highly negative correlation with the insurance_cost. Thus, we need to impute the null values in this column.
We can see that we have various missing values in respective columns. There are various ways of treating your missing values in the data set. And which technique to use when is actually dependent on the type of data you are dealing with.

- Drop the missing values : In this case we drop the missing values from those variables. In case there are very few missing values you can drop those values.
- Impute with mean value : For numerical column, you can replace the missing values with mean values. Before replacing with mean value, it is advisable to check that the variable shouldn't have extreme values .i.e. outliers.
- Impute with median value : For numerical column, you can also replace the missing values with median values. In case you have extreme values such as outliers it is advisable to use median approach.
- Impute with mode value : For categorical column, you can replace the missing values with mode values i.e the frequent ones.

In this exercise, we will replace the numerical columns with median values and for categorical columns we will replace the missing values with mode values.

Missing values after the replacement of the values with the median :

```
applicant_id                    0
years_of_insurance_with_us      0
regular_checkup_last_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
```

There are no null values present after the imputation.
Information regarding the dataset after null values removal :

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   years_of_insurance_with_us      25000 non-null  int64
 1   regular_checkup_last_year       25000 non-null  int64
 2   adventure_sports                25000 non-null  int64
 3   Occupation                      25000 non-null  object
 4   visited_doctor_last_1_year      25000 non-null  int64
 5   cholesterol_level               25000 non-null  object
 6   daily_avg_steps                 25000 non-null  int64
 7   age                             25000 non-null  int64
 8   heart_decs_history              25000 non-null  int64
 9   other_major_decs_history        25000 non-null  int64
 10  Gender                          25000 non-null  object
 11  avg_glucose_level               25000 non-null  int64
 12  bmi                             25000 non-null  float64
 13  smoking_status                  25000 non-null  object
 14  Year_last_admitted              25000 non-null  float64
 15  Location                        25000 non-null  object
 16  weight                          25000 non-null  int64
 17  covered_by_any_other_company    25000 non-null  object
 18  Alcohol                         25000 non-null  object
 19  exercise                        25000 non-null  object
 20  weight_change_in_last_one_year  25000 non-null  int64
 21  fat_percentage                  25000 non-null  int64
 22  insurance_cost                  25000 non-null  int64
dtypes: float64(2), int64(13), object(8)
memory usage: 4.4+ MB
```

e) Outlier treatment

Presence of outliers is generally detected with the box plots.

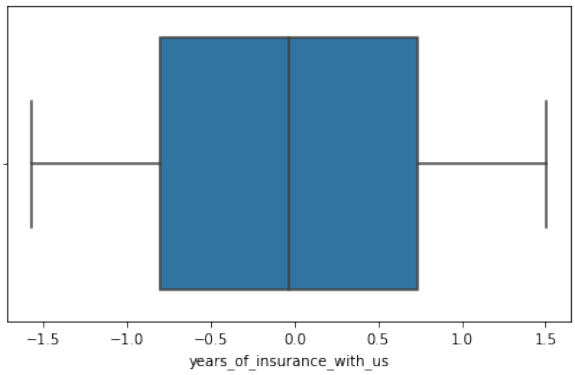


Fig 44 - boxplot of years_of_insurance_with_us

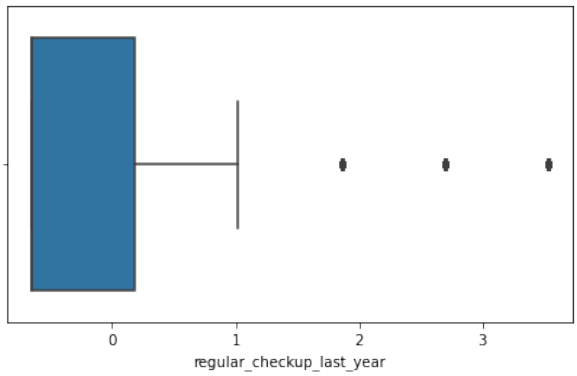


Fig 45 - boxplot of regular_checkup_last_year

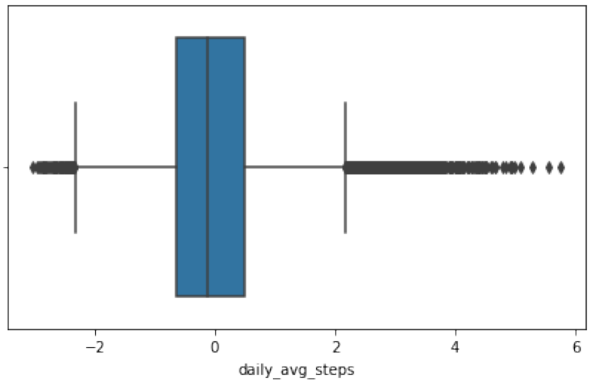


Fig 46 - boxplot of daily_avg_steps

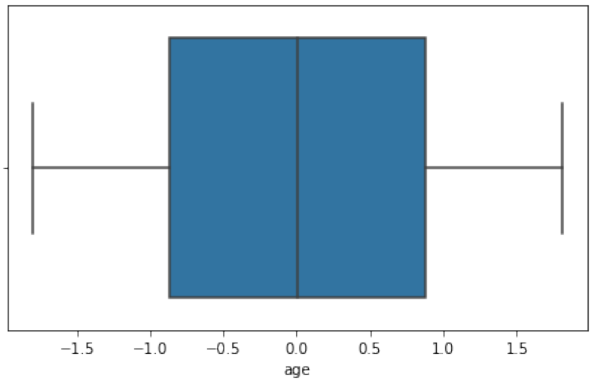


Fig 47 - boxplot of age

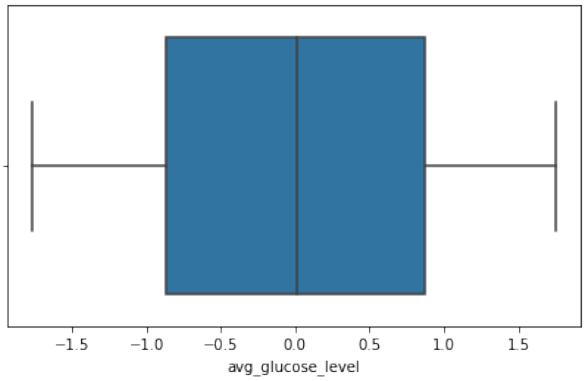


Fig 48 - boxplot of avg_glucose_level

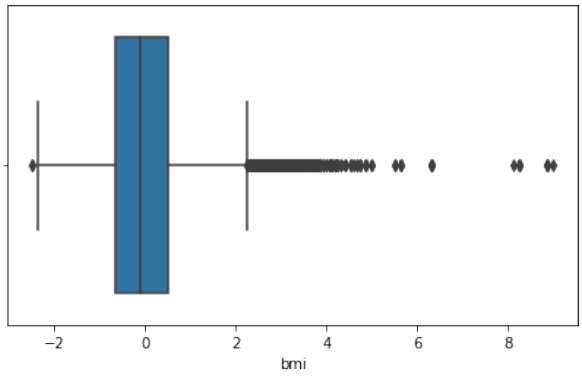


Fig 49 - boxplot of bmi

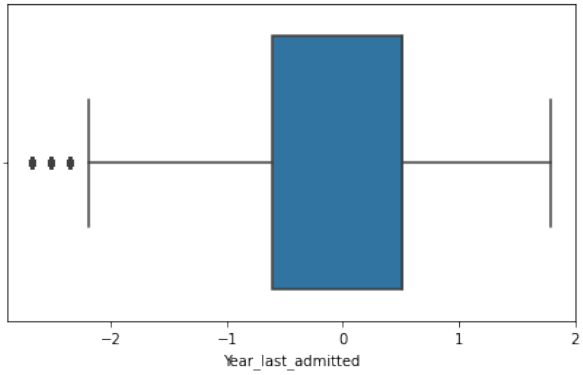


Fig 50 - boxplot of year_last_admitted

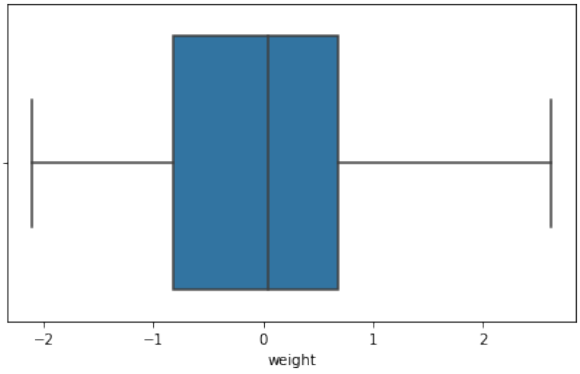


Fig 51 - boxplot of weight

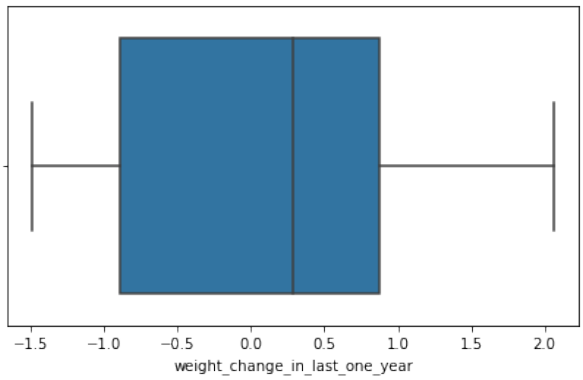


Fig 52 - boxplot of weight_change_in_last_one_year

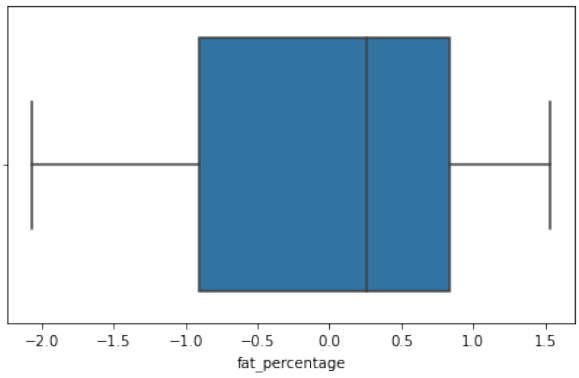


Fig 53 - boxplot of fat_percentage

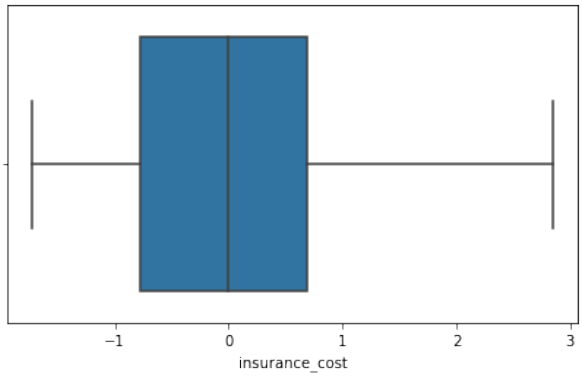


Fig 54 - boxplot of insurance cost

Looking at the box plot, it seems that the four variables regular_checkup_last_year, daily_avg_steps, Bmi, year_last_admitted have outlier present in the variables.
These outliers value needs to be teated and there are several ways of treating them:

- Drop the outlier value
- Replace the outlier value using the IQR

We have replaced the outlier values using IQR method.

Boxplots after the outlier treatment are as follows :

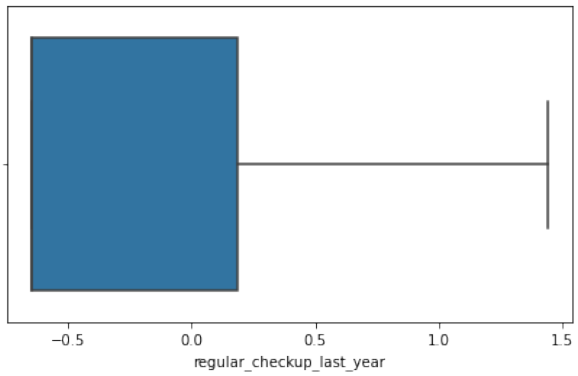


Fig 55 - boxplot of regular_checkup_last_year

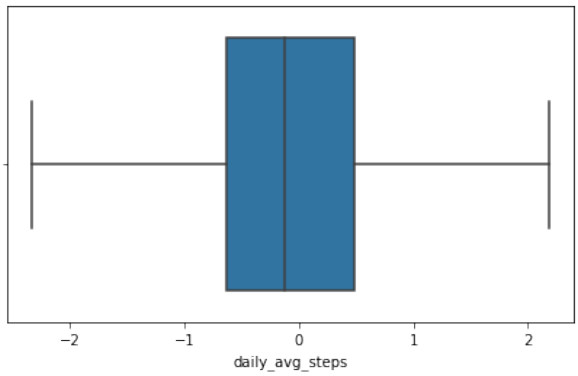


Fig 56 - boxplot of daily_avg_steps after outlier treatment

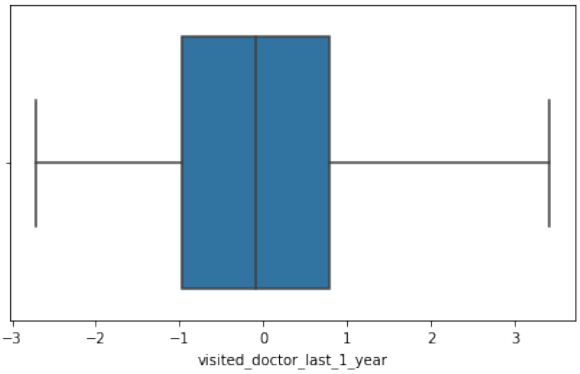


Fig 57 - boxplot of visited_doctor_in_last_1_year

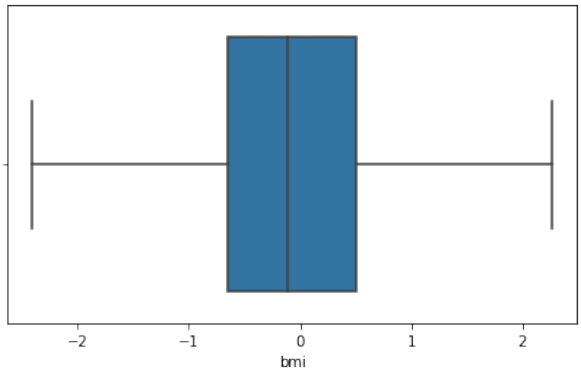


Fig 58 - boxplot of bmi after outlier treatment

Data is free from outliers.

f) Variable transformation

For variable transformation we need to check the skewness and kurtosis of the data.

skewness of insurance_cost is 0.33165006251159934
kurtosis of insurance_cost is -0.5020550403211033

Skewness assesses the extent to which a variable's distribution is symmetrical
Kurtosis is a measure of whether the distribution is too peaked
For an ideal normal distribution (theoretical) Skewness and Kurtosis have to be between -1 to +1
Or we can say that if we are able to reduce the skewness and kurtosis from a very high value to lower values we are able to get the data distirbuted more normally.
Thus, we not need to transform this data.

g) Addition of new variables

No new variable is added.

h) Scaling

Often the variables of the data set are of different scales i.e. one variable is in millions and other in only 100. For e.g. in our data set Income is having values in thousands and age in just two digits. Since the data in these variables are of different scales, it is tough to compare these variables.
Feature scaling (also known as data normalization) is the method used to standardize the range of features of data. Since, the range of values of data may vary widely, it becomes a necessary step in data preprocessing while using machine learning algorithms.
In this method, we convert variables with different scales of measurements into a single scale.
StandardScaler normalizes the data using the formula (x-mean)/standard deviation.
We will be doing this only for the numerical variables.

Separating categorical and numerical columns :

Categorical data
['Occupation', 'cholesterol_level', 'Gender', 'smoking_status', 'Location', 'covered_by_any_other_company', 'Alcohol', 'exercise']

Numerical data
['years_of_insurance_with_us', 'regular_checkup_last_year', 'adventure_sports', 'visited_doctor_last_1_year', 'daily_avg_steps', 'age', 'heart_decs_history', 'other_major_decs_history', 'avg_glucose_level', 'bmi', 'Year_last_admitted', 'weight', 'weight_change_in_last_one_year', 'fat_percentage', 'insurance_cost']

Table 10 - Dataset after scaling

	years_of_insurance_with_us	regular_checkup_last_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	other_major_decs_history	avg_glucose_level	bmi	Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost
0	-0.417807	0.188690	3.352150	-0.967205	-0.332228	-1.050360	4.159520	-0.329915	-1.124370	-0.020458	0.510204	-0.494422	-0.898041	-0.441634	-0.430722
1	-1.568750	-0.645043	-0.298316	0.784661	1.134787	0.315492	-0.240412	-0.329915	0.708929	0.368102	0.510204	-1.459569	0.285180	-0.209944	-1.464554
2	-1.185102	-0.645043	-0.298316	0.784661	-0.671209	1.433007	-0.240412	-0.329915	-0.024391	1.171127	0.510204	0.149010	-1.489652	0.369282	0.086194
3	1.116783	2.689890	-0.298316	-0.967205	0.947731	0.377576	-0.240412	-0.329915	-0.933069	-1.095475	0.510204	-0.065467	0.285180	0.948508	0.000041
4	-0.417807	0.188690	-0.298316	-0.967205	-0.263863	-0.057013	-0.240412	3.031081	-0.789594	-0.629202	-0.444900	0.256249	-1.489652	0.600972	0.172347

Table 11 - Descriptive statistics of the scaled data

years_of_insurance_with_us	regular_checkup_last_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	other_major_decs_history	avg_glucose_level	bmi	Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost	
count	2.500000E+	2.500000E+	2.500000E+	2.500000E+	2.5000	2.50000	2.500000E+	2.500000E+	2.50000	2.50000	2.50000	2.500000E+04	2.500000E+04	2.5000	2.5000
mean	-6.197443E-	-1.056470E	-1.265019E	4.659961E-	-2.032	-8.55937	-1.325162E	3.912026E-	-8.4288	3.06172	1.17133	4.625011E-16	-8.179235E-1	-3.092	3.6823
std	1.000020E+	1.000020E+	1.000020E+	1.000020E+	1.0000	1.00002	1.000020E+	1.000020E+	1.00002	1.00002	1.00002	1.000020E+00	1.000020E+00	1.0000	1.0000
min	-1.568750E-	-6.450427E	-2.983160E	-2.719070E	-3.021	-1.7953	-2.404123E	-3.299154E	-1.7620	-2.4683	-2.6734	-2.103001E+0	-1.489652E+0	-2.063	-1.723
25%	-8.014550E-	-6.450427E	-2.983160E	-9.672049E	-6.389	-8.6410	-2.404123E	-3.299154E	-8.6930	-6.5510	-6.0408	-8.161376E-0	-8.980412E-0	-9.050	-7.753
50%	-3.415997E-	-6.450427E	-2.983160E	-9.127219E	-1.204	5.07103	-2.404123E	-3.299154E	7.49261	-1.1112	5.10203	4.177160E-02	2.851800E-01	2.5343	4.1353
75%	7.331350E-(	1.886905E-	-2.983160E	7.846605E-	4.8816	8.74249	-2.404123E	-3.299154E	8.68345	5.10574	5.10203	6.852035E-01	8.767906E-01	8.3266	6.8926
max	1.500430E+	3.523623E-	3.352150E-	7.792122E-	5.7342	1.80551	4.159520E-	3.031081E-	1.74514	8.96823	1.78367	2.615499E+00	2.060012E+00	1.5277	2.8430

Applying zscore or using StandardScalar give us the same results. It scales the data in such a way that the mean value of the features tends to 0 and the standard deviation tends to 1. Min-Max method ensure that the data scaled to have values in the range 0 to 1.
If we look at the variables, all have been normalized and scaled in one scale now.

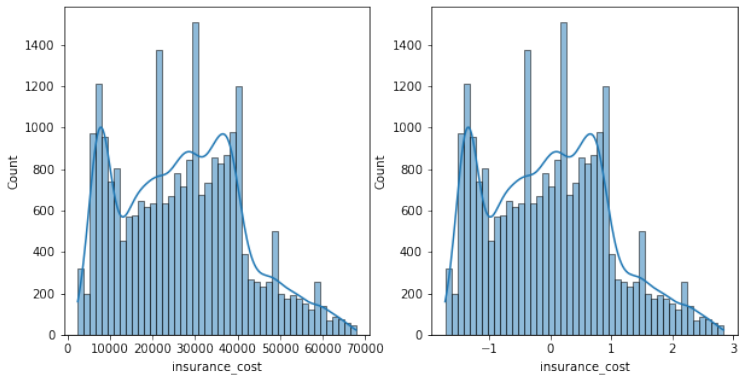


Fig 59 - histogram of insurance cost before and after scaling

i) Checking for multicollinearity

Multicollinearity in a dataset is detected by **Principal component analysis** (PCA) which is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

There are Statistical tests which are done before performing PCA.
Bartletts Test of Sphericity : Bartlett's test of sphericity tests the hypothesis that the variables are uncorrelated in the population.

H0: All variables in the data are uncorrelated
Ha: At least one pair of variables in the data are correlated

If the null hypothesis cannot be rejected, then PCA is not advisable.

Here, the p-value is small, then we can reject the null hypothesis and agree that there is atleast one pair of vairbales in the data wihich are correlated hence PCA is recommended

P-value of Bartlett's test is 0.0 which is less than 0.5, thus null hypothesis is rejected. This means that at least one pair of variables in the data are correlated.

KMO Test The Kaiser-Meyer-Olkin (KMO) - measure of sampling adequacy (MSA) is an index used to examine how appropriate PCA is.

Generally, if MSA is less than 0.5, PCA is not recommended, since no reduction is expected. On the other hand, MSA > 0.7 is expected to provide a considerable reduction is the dimension and extraction of meaningful components.

Confirming the adequacy of the sample size :

Kmo = 0.6546199061928036

Thus we need to perform PCA on the dataset.

Eigen vectors :

```
array([[ 7.32491853e-02, -8.92846658e-02,  7.63074145e-02,
         1.32306464e-02, -7.80785904e-03,  5.38601507e-03,
        -3.74206548e-03, -3.78170312e-03, -5.64363998e-03,
        -1.00023683e-02, -4.76341855e-01,  5.73443882e-01,
        -3.12567503e-01, -9.63335710e-03,  5.71737760e-01],
       [-1.27495437e-01, -2.00173950e-02,  2.26144908e-02,
         5.92473555e-02, -3.69766343e-02, -2.08794551e-02,
         5.53799974e-01,  7.08580768e-01, -5.07634533e-02,
         4.02781469e-01,  3.26903790e-02,  2.25158399e-02,
        -5.03514257e-03, -4.15186304e-02,  2.45682921e-02],
       [ 6.32787752e-02,  1.24443594e-02,  3.72000095e-02,
         6.92612527e-01, -4.93518452e-01,  3.26988537e-02,
        -7.41577730e-02, -2.72603472e-02,  6.02277648e-02,
        -1.85597360e-02, -1.77352823e-03, -2.76308710e-02,
        -2.27680156e-02, -5.07694407e-01, -2.93658086e-02],
       [ 8.96684733e-01,  2.34903710e-02,  5.89675117e-02,
        -3.64831816e-02,  2.51721205e-02,  4.66877359e-02,
         1.13251907e-01,  8.33811114e-02, -2.65482469e-02,
         3.18809920e-02, -3.06944470e-01, -1.40936059e-01,
         1.86863828e-01,  2.78852424e-02, -1.28491063e-01],
       [-4.34489468e-02, -5.96757702e-03, -1.74439304e-02,
        -6.37790769e-02,  1.49619168e-02,  7.10889839e-01,
        -2.35443863e-02, -2.01029141e-02, -6.83807898e-01,
         5.21016399e-04,  1.40261197e-02, -1.19535318e-03,
        -7.53835323e-03, -1.38951544e-01,  1.34062040e-03],
       [-2.78894281e-02,  3.34659065e-02,  7.75807063e-01,
         1.01852754e-01, -5.24951343e-02, -3.73705321e-01,
        -4.17040046e-02, -4.91751500e-02, -4.49904943e-01,
         1.58138546e-02,  4.48692805e-02, -3.19616179e-02,
        -1.55445650e-02,  1.79192231e-01, -3.50996126e-02],
       [-4.45318232e-02,  6.40803366e-02,  5.68592821e-01,
         8.25431443e-03,  3.05079912e-02,  5.73722092e-01,
         2.76092785e-02,  5.46696569e-02,  5.58001903e-01,
        -6.38715724e-02,  4.37820158e-02, -2.49302015e-02,
```

```
        -5.39480882e-02,  1.22949014e-01, -2.63617271e-02],
       [-2.85557662e-03,  1.97739842e-02,  2.28662084e-01,
        -4.19526125e-01,  2.45324964e-01, -1.33903602e-01,
         1.52771996e-01, -9.51714826e-02,  6.18847369e-02,
        -5.42331272e-02,  2.27579275e-02, -1.38465489e-02,
        -3.62240755e-02, -8.07361141e-01, -1.32707444e-02],
       [-3.10038669e-02,  1.40442033e-02, -4.72886318e-02,
         1.22813974e-01, -5.40698411e-02,  1.00761084e-02,
         7.97802351e-01, -4.18664922e-01, -4.02350575e-02,
        -3.83321427e-01,  1.91259805e-02,  2.44405015e-03,
        -3.01972509e-02,  1.29920125e-01,  3.31428201e-03],
       [ 1.66663139e-01,  2.85707623e-01, -8.26384199e-02,
        -2.50331998e-02,  3.90012798e-03, -3.20181882e-02,
        -2.45232226e-02,  1.04097503e-01, -3.94331173e-02,
        -7.71225575e-02,  1.21648689e-01, -1.54506328e-01,
        -8.86366916e-01,  3.67833642e-02, -1.93618013e-01],
       [-2.61057182e-02, -4.69542547e-02, -6.43412883e-03,
         8.56725092e-02,  1.47966513e-01, -5.25141093e-02,
        -1.07024858e-01,  5.30345598e-01, -6.43928439e-02,
        -8.08971292e-01, -1.23847011e-02,  1.41675374e-02,
         1.09263873e-01, -3.29143779e-02,  2.00743042e-02],
       [ 8.07976555e-03, -1.32514523e-03, -1.85530528e-02,
         5.49190972e-01,  8.15712380e-01,  9.61439895e-03,
        -6.38288232e-05, -8.77995750e-02, -1.16034028e-02,
         1.47341786e-01, -1.34916920e-04, -5.80251761e-03,
        -3.06115888e-02, -4.32324292e-02, -1.02999713e-02],
       [-4.92144040e-02,  9.47833132e-01, -3.97798872e-02,
         9.28205131e-03,  7.29424559e-03, -1.25623874e-02,
        -3.94363905e-03,  1.41021272e-02, -1.99585553e-02,
         1.04210511e-03, -4.20325437e-02,  1.46477433e-01,
         2.46568909e-01, -9.01436320e-03,  1.12236220e-01],
       [-3.65695024e-01,  4.33234071e-02,  2.08020941e-03,
        -2.62615387e-03, -2.93342111e-03, -6.03475643e-03,
         1.85933533e-03, -6.58526190e-03,  1.48517469e-03,
        -2.33674455e-03, -8.09989222e-01, -3.17579875e-01,
        -4.03748536e-02, -4.41340852e-03, -3.25100557e-01],
       [-5.94534313e-03,  3.11593651e-02, -1.42051781e-03,
         2.45366506e-03,  1.28418042e-03, -2.78278390e-03,
        -2.79042564e-03, -8.31058052e-04, -1.67302950e-04,
         1.05561440e-03,  2.65169309e-04, -7.08355360e-01,
        -2.14123429e-02,  9.76698073e-04,  7.04797750e-01]])
```

pca variance

```
array([2.70205434, 1.16716936, 1.10356787, 1.0918125 , 1.01241631,
       0.99717427, 0.98594637, 0.96198157, 0.92259565, 0.825542  ,
       0.76286892, 0.74411948, 0.54274455, 0.40330117, 0.02860308])
```

pca variance ratio

```
array([0.1895926 , 0.08189572, 0.07743305, 0.07660822, 0.0710373 ,
       0.06996783, 0.06918001, 0.06749849, 0.06473493, 0.05792506,
       0.05352753, 0.05221196, 0.03808227, 0.02829807, 0.00200697])
```

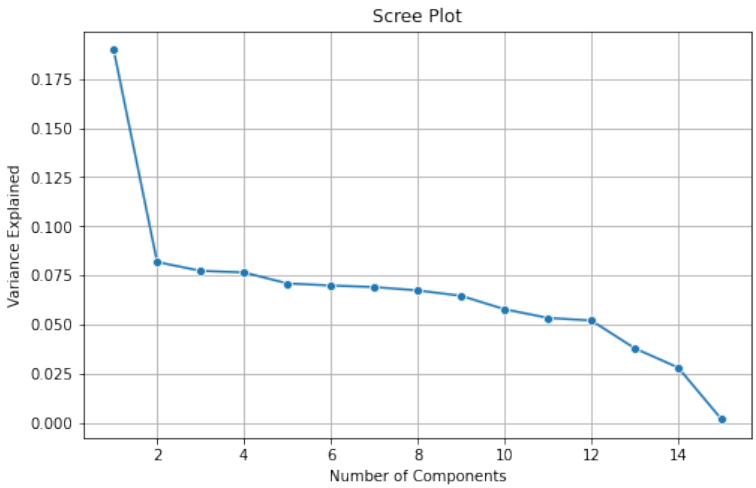
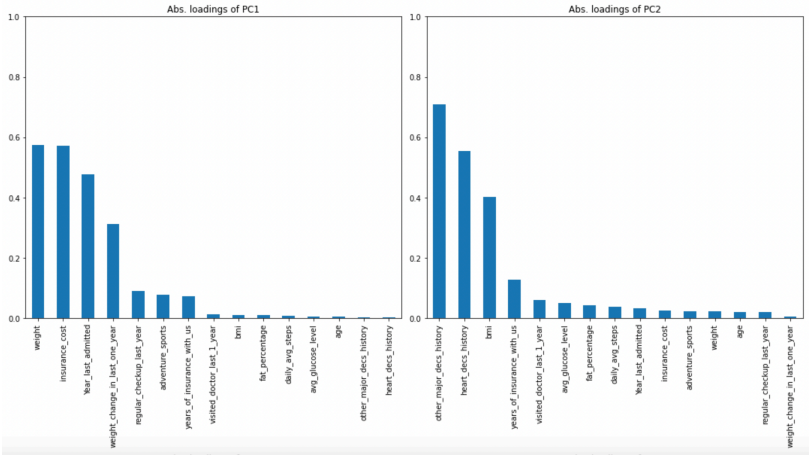
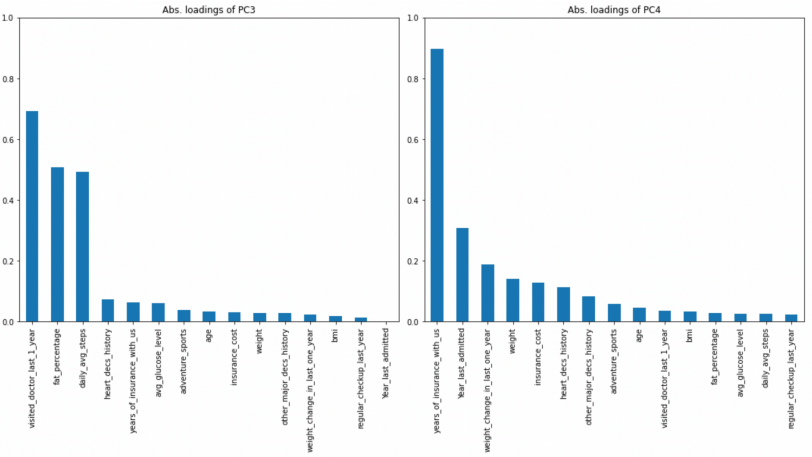


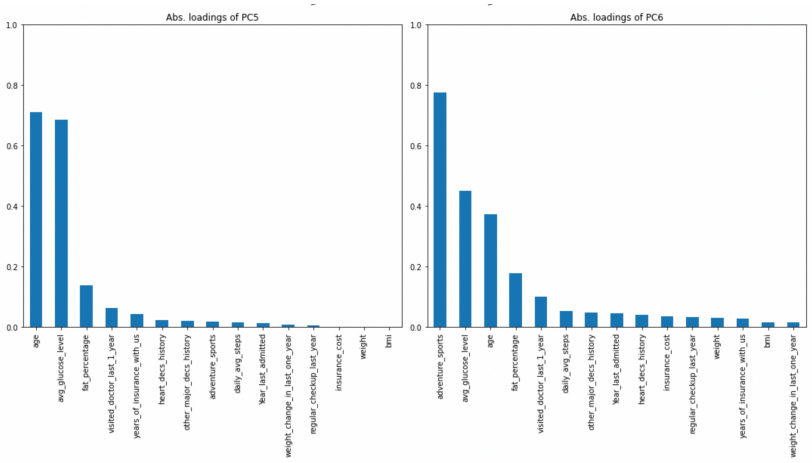
Fig 60 - Scree plot

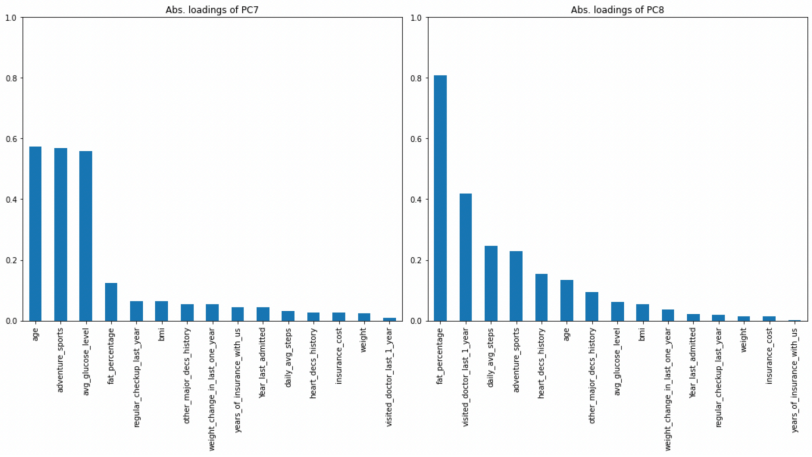
Cumulative pca variance ratio

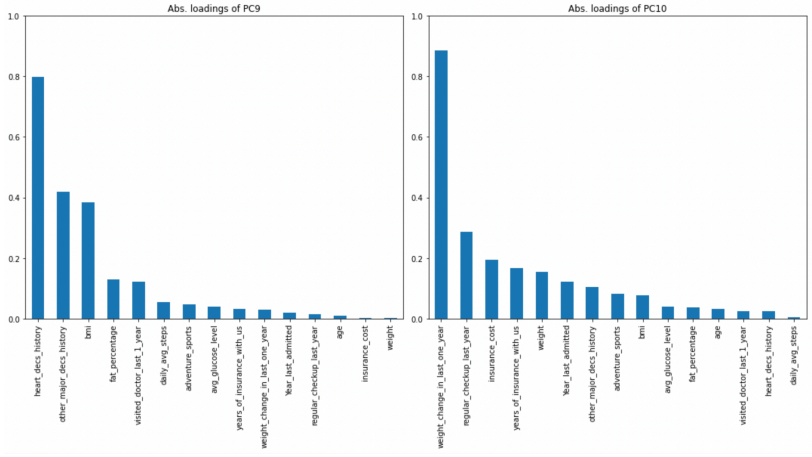
array([0.1895926 , 0.27148832, 0.34892137, 0.42552959, 0.49656689,
0.56653472, 0.63571472, 0.70321321, 0.76794815, 0.82587321,
0.87940074, 0.9316127 , 0.96969497, 0.99799303, 1.])






Abs. loadings of PC5
age
avg_glucose_level
fat_percentage
visited_doctor_last_1_year
years_of_insurance_with_us
heart_decs_history
other_major_decs_history
adventure_sports
daily_avg_steps
Year_last_admitted
weight_change_in_last_one_year
regular_checkup_last_year
insurance_cost
weight
bmi
Abs. loadings of PC6
adventure_sports
avg_glucose_level
age
fat_percentage
visited_doctor_last_1_year
daily_avg_steps
other_major_decs_history
Year_last_admitted
heart_decs_history
insurance_cost
regular_checkup_last_year
weight
years_of_insurance_with_us
bmi
weight_change_in_last_one_year


Abs. loadings of PC7
age
adventure_sports
avg_glucose_level
fat_percentage
regular_checkup_last_year
bmi
other_major_decs_history
weight_change_in_last_one_year
years_of_insurance_with_us
Year_last_admitted
daily_avg_steps
heart_decs_history
insurance_cost
weight
visited_doctor_last_1_year
Abs. loadings of PC8
fat_percentage
visited_doctor_last_1_year
daily_avg_steps
adventure_sports
heart_decs_history
age
other_major_decs_history
avg_glucose_level
bmi
weight_change_in_last_one_year
Year_last_admitted
regular_checkup_last_year
weight
insurance_cost
years_of_insurance_with_us


Abs. loadings of PC9
heart_decs_history
other_major_decs_history
bmi
fat_percentage
visited_doctor_last_1_year
daily_avg_steps
adventure_sports
avg_glucose_level
years_of_insurance_with_us
weight_change_in_last_one_year
Year_last_admitted
regular_checkup_last_year
age
insurance_cost
weight
Abs. loadings of PC10
weight_change_in_last_one_year
regular_checkup_last_year
insurance_cost
years_of_insurance_with_us
weight
Year_last_admitted
other_major_decs_history
adventure_sports
bmi
avg_glucose_level
fat_percentage
age
visited_doctor_last_1_year
heart_decs_history
daily_avg_steps

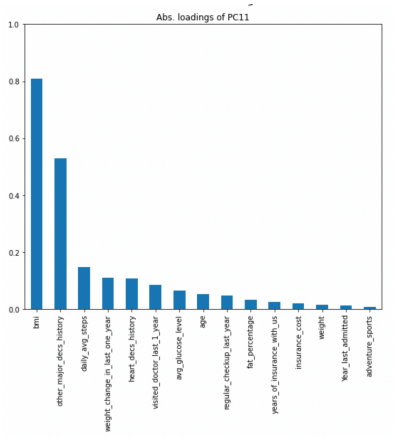


Fig 61 - how original features matter to PCs

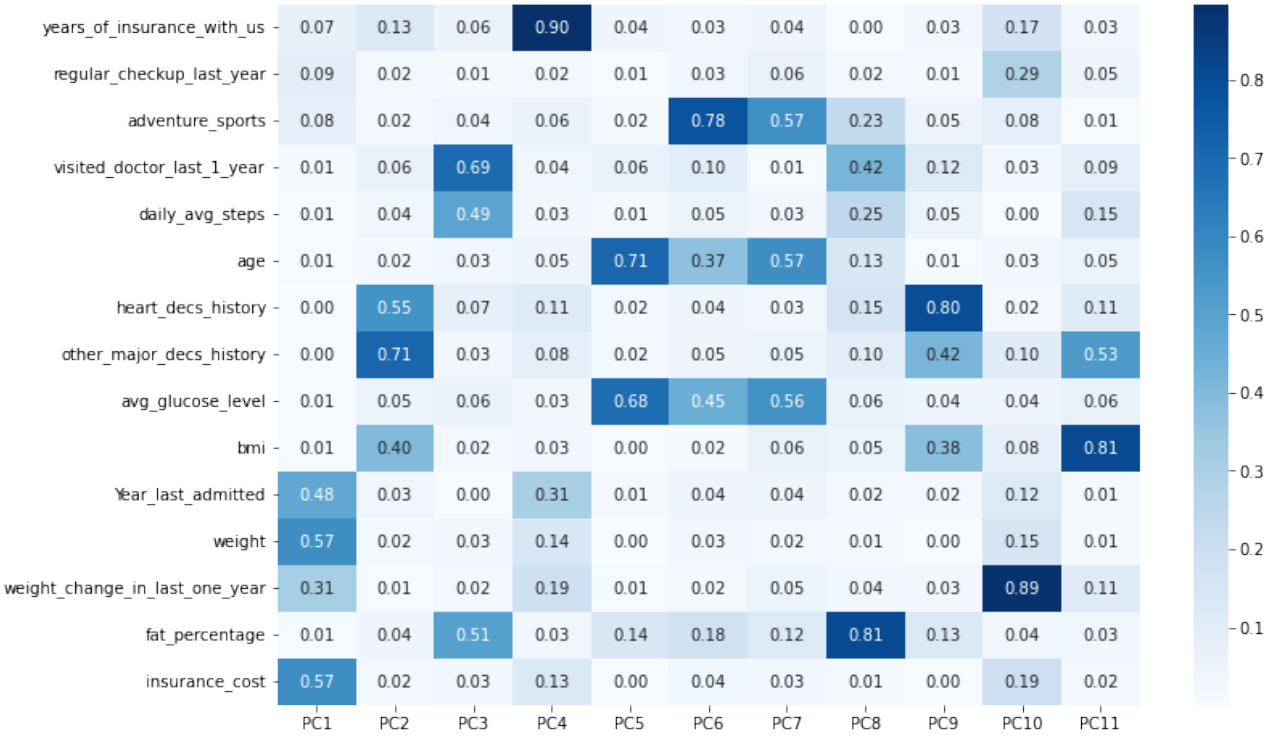


Fig 62 - Influence of original features on PCs

pc score

```
-0.303146 2.23864 -0.536412 0.066401 0.021426 3.266968
0.82685 2.247992 3.232222 0.674977 -0.733131
```

Table 12 - Final PCA dataframe

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
0	-0.312280	2.245061	-0.545164	0.069908	0.020954	3.269717	0.832519	2.253133	3.224128	0.701285	-0.751810
1	-2.095999	-0.088984	0.111812	-1.198345	-0.170372	-0.512059	0.490209	0.064619	-0.174075	-0.194250	-0.237952
2	0.306556	0.325424	0.661206	-1.544015	1.015287	-0.477412	0.684667	-1.064309	-0.217451	0.839521	-1.335353
3	-0.457218	-1.058147	-1.552730	0.965366	0.814952	0.118203	-0.234057	-0.217353	0.353376	0.611243	0.713793
4	0.827028	1.782095	-0.968705	-0.315204	0.444706	0.028834	-0.299492	-0.496620	-1.155580	1.676733	1.889532
5	1.216602	-0.150311	0.222082	1.294813	0.138725	-0.540008	-0.986206	1.765406	-0.805705	-0.600196	-0.782930
6	1.270379	-0.456380	1.502491	1.149709	0.832457	0.230681	-1.498384	0.587589	0.000662	-0.492525	0.314497
7	0.274487	-0.574319	-0.767422	-1.619928	-0.808036	-0.667664	0.836945	-0.204429	0.312278	0.996726	0.569588
8	-0.113237	-0.793061	1.462721	1.133239	0.640412	-0.202749	-0.701953	1.323140	0.127963	1.275268	0.634720
9	-1.194578	0.534754	1.423229	-0.205180	0.905812	0.460208	-1.326245	1.284710	-0.863740	2.050606	-1.961911

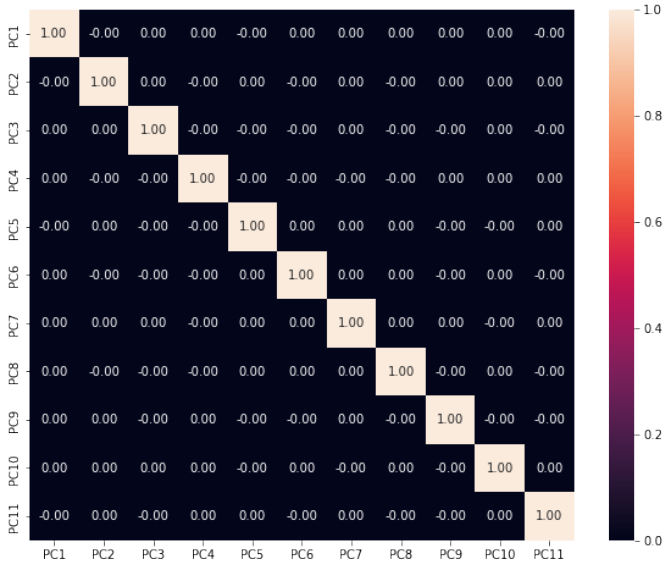


Fig 63 - Presence of correlation among final PCs

j) Encoding

One-Hot-Encoding is used to create dummy variables to replace the categories in a categorical variable into features of each category and represent it using 1 or 0 based on the presence or absence of the categorical value in the record.This is required to do since the machine learning algorithms only works on the numerical data. That is why there is a need to convert the categorical column into numerical one.

get_dummies is the method which creates dummy variable for each categorical variable. It is considered a good practice to set parameter drop_first as True whenever get_dummies is used. It reduces the chances of multicollinearity which will be covered in coming courses and the number of features are also less as compared to drop_first = False

Table 13 - Final dataframe after encoding

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	Occupation_Salried	Occupation_Student	cholesterol_level_2	cholesterol_level_3	cholesterol_level_4	cholesterol_level_5	Gender_Male	smoking_status_former smoked	smoking_status_never smoked	smoking_status_smokes	Location_Bangalore	Location_Bhubaneswar	Location_Chennai	Location_Delhi	Location_Guwahati	Location_Jaipur	Location_Kanpur	Location_Kolkata	Location_Lucknow	Location_Mangalore	Location_Mumbai	Location_Nagpur	Location_Pune	Location_Surat	covered_by_any_other_company_Y	Alcohol_No	Alcohol_Rare	exercise_Moderate	exercise_No
0	-0.312	2.245	-0.545	0.069	0.020	3.269	0.832	2.253	3.224	0.701	-0.751	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0
1	-2.095	-0.08	0.1118	-1.198	-0.170	-0.51	0.490	0.064	-0.174	-0.19	-0.237	0	1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	1	0
2	0.3065	0.325	0.6612	-1.544	1.015	-0.47	0.684	-1.064	-0.217	0.839	-1.335	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	-0.457	-1.05	-1.552	0.965	0.814	0.118	-0.23	-0.217	0.353	0.611	0.7137	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1
4	0.8270	1.782	-0.968	-0.315	0.444	0.028	-0.29	-0.496	-1.155	1.676	1.8895	0	1	1	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

In the data set, each Category in all of the categorical columns have been added as columns with values 0 and 1 Example: Occupation_Salried, Gender_MaleIf Gender_Male =1, then it means its a Male and Gender_Male=0 means its a Female**

4) Business insights from EDA

a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business

Imbalanced data refers to those types of datasets where the target class has an uneven distribution of observations, i.e one class label has a very high number of observations and the other has a very low number of observations. In rare cases like fraud detection or disease prediction, it is vital to identify the minority classes correctly. So model should not be biased to detect only the majority class but should give equal weight or importance towards the minority class too.

Since the y variable or dependent variable (insurance cost) is a continuous variable and has no class like defaulters and non defaulters so, we will have to do it at individual categorical variable level.

```
OCCUPATION :  3
Salaried      4811
Business     10020
Student      10169
Name: Occupation, dtype: int64

This category is unbalanced


CHOLESTEROL_LEVEL :  5
5    2054
3    2881
4    2963
```

```
1    8339
2    8763
Name: cholesterol_level, dtype: int64

This category is unbalanced


GENDER :  2
Female     8578
Male      16422
Name: Gender, dtype: int64

This category is unbalanced


SMOKING_STATUS :  4
smokes              3867
formerly smoked     4329
Unknown             7555
never smoked        9249
Name: smoking_status, dtype: int64

This category is unbalanced


LOCATION :  15
Surat           1589
Kolkata         1620
Pune            1622
Lucknow         1637
Mumbai          1658
Nagpur          1663
Kanpur          1664
Chennai         1669
Guwahati        1672
Ahmedabad       1677
Delhi           1680
Mangalore       1697
Bhubaneswar     1704
Jaipur          1706
Bangalore       1742
Name: Location, dtype: int64

This category is balanced


COVERED_BY_ANY_OTHER_COMPANY :  2
Y     7582
N    17418
Name: covered_by_any_other_company, dtype: int64

This category is unbalanced


ALCOHOL :  3
Daily     2707
No        8541
Rare     13752
Name: Alcohol, dtype: int64

This category is unbalanced


EXERCISE :  3
No           5114
Extreme      5248
```

```
Moderate    14638
Name: exercise, dtype: int64

This category is unbalanced
```

What to do when data is unbalanced

Choose Proper Evaluation Metric : The accuracy of a classifier is the total number of correct predictions by the classifier divided by the total number of predictions. This may be good enough for a well-balanced class but not ideal for the imbalanced class problem. The other metrics such as precision is the measure of how accurate the classifier's prediction of a specific class and recall is the measure of the classifier's ability to identify a class.
Resampling (Oversampling and Undersampling) : This technique is used to upsample or downsample the minority or majority class. When we are using an imbalanced dataset, we can oversample the minority class using replacement. This technique is called oversampling. Similarly, we can randomly delete rows from the majority class to match them with the minority class which is called undersampling. After sampling the data we can get a balanced dataset for both majority and minority classes. So, when both classes have a similar number of records present in the dataset, we can assume that the classifier will give equal importance to both classes.
SMOTE: Synthetic Minority Oversampling Technique or SMOTE is another technique to oversample the minority class. Simply adding duplicate records of minority class often don't add any new information to the model. In SMOTE new instances are synthesized from the existing data. If we explain it in simple words, SMOTE looks into minority class instances and use k nearest neighbor to select a random nearest neighbor, and a synthetic instance is created randomly in feature space.
BalancedBaggingClassifier : When we try to use a usual classifier to classify an imbalanced dataset, the model favors the majority class due to its larger volume presence. A BalancedBaggingClassifier is the same as a sklearn classifier but with additional balancing. It includes an additional step to balance the training set at the time of fit for a given sampler. This classifier takes two special parameters "sampling_strategy" and "replacement". The sampling_strategy decides the type of resampling required (e.g. 'majority' – resample only the majority class, 'all' – resample all classes, etc) and replacement decides whether it is going to be a sample with replacement or not.
Threshold moving : In the case of our classifiers, many times classifiers actually predict the probability of class membership. We assign those prediction's probabilities to a certain class based on a threshold which is usually 0.5, i.e. if the probabilities < 0.5 it belongs to a certain class, and if not it belongs to the other class. For imbalanced class problems, this default threshold may not work properly. We need to change the threshold to the optimum value so that it can efficiently separate two classes. We can use ROC Curves and Precision-Recall Curves to find the optimal threshold for the classifier. We can also use a grid search method or search within a set of values to identify the optimal value.

b) Any business insights using clustering

Clustering

Cluster analysis or clustering is an unsupervised machine learning algorithm that groups unlabeled datasets. It aims to form clusters or groups using the data points in a dataset in such a way that there is high intra-cluster similarity and low inter-cluster similarity. In, layman terms clustering aims at forming subsets or groups within a dataset consisting of data points which are really similar to each other and the groups or subsets or clusters formed can be significantly differentiated from each other. A clustering algorithm can discover groups of objects where the average distances

between the members/data points of each cluster are closer than to members/data points in other clusters.

K-Means clustering with 2 clusters
array([1, 1, 0, ..., 0, 0, 1], dtype=int32)

the within cluster sum of squares for 2 clusters for the K-Means algorithm
357598.8098610467

K-Means clustering with 3 clusters and the within cluster sum of squares for 3 clusters
334666.58050971554

Within Sum of Squares (WSS) for 2 to 15 clusters
[403277.0861948237,
357598.8098610467,
334666.58050971554,
317869.0479431501,
307557.02979533677,
288867.7069537262,
273808.9281934849,
268012.4538565579,
263617.7379769511,
259731.14221520748,
256526.70640403844,
253009.10953251788,
249953.2249468235,
246990.8848361045]

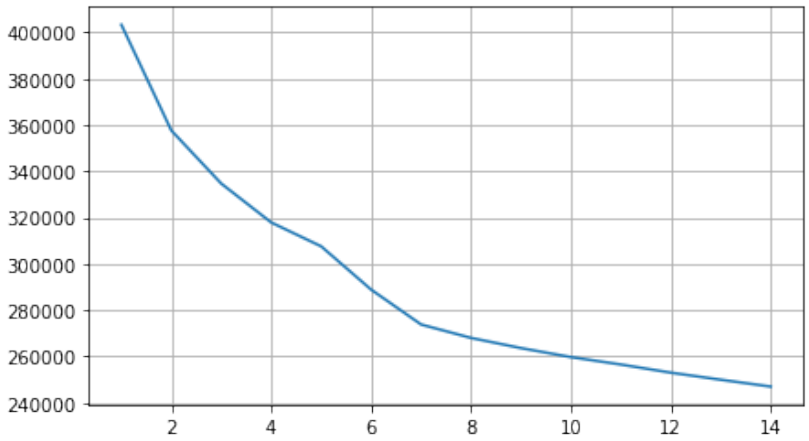


Fig 64 - Within Sum of Squares (WSS) plot

optimum number of clusters from the WSS plot

Firstly, we will check with 2 clusters.

array([1, 1, 0, ..., 0, 0, 1], dtype=int32)

Let us check the silhouette score and silhouette width for 2 clusters.

silhouette score 0.10620928658615003
silhouette width -0.020030057130769306

Now, let us check with 7 clusters.
array([0, 6, 2, ..., 5, 2, 0], dtype=int32)

Let us check the silhouette score and silhouette width for 7 clusters.

silhouette score 0.09226448951145735
silhouette width -0.025087099708634833

Table 14 - The final dataset after data pre-processing

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
0	-0.312280	2.245061	-0.545164	0.069908	0.020954	3.269717	0.832519	2.253133	3.224128	0.701285	-0.751810
1	-2.095999	-0.088984	0.111812	-1.198345	-0.170372	-0.512059	0.490209	0.064619	-0.174075	-0.194250	-0.237952
2	0.306556	0.325424	0.661206	-1.544015	1.015287	-0.477412	0.684667	-1.064309	-0.217451	0.839521	-1.335353
3	-0.457218	-1.058147	-1.552730	0.965366	0.814952	0.118203	-0.234057	-0.217353	0.353376	0.611243	0.713793
4	0.827028	1.782095	-0.968705	-0.315204	0.444706	0.028834	-0.299492	-0.496620	-1.155580	1.676733	1.889532

smoking_status_formerly smoked	smoking_status_never smoked	smoking_status_smokes	Location_Bangalore	Location_Bhubaneswar	Location_Chennai	Location_Delhi
0	0	0	0	0	1	0
1	0	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	1	0
0	1	0	1	0	0	0

Location_Pune	Location_Surat	covered_by_any_other_company_Y	Alcohol_No	Alcohol_Rare	exercise_Moderate	exercise_No	Kmeans_clusters
0	0	0	0	1	1	0	1
0	0	0	0	1	1	0	1
0	0	0	0	0	0	0	0
0	0	1	0	1	0	1	1
0	0	0	1	0	0	0	0

Location_Guwahati	Location_Jaipur	Location_Kanpur	Location_Kolkata	Location_Lucknow	Location_Mangalore	Location_Mumbai	Location_Nagpur
0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Occupation_Salried	Occupation_Student	cholesterol_level_2	cholesterol_level_3	cholesterol_level_4	cholesterol_level_5	Gender_Male
1	0	0	0	0	0	1
0	1	1	0	0	0	1
0	0	0	0	1	0	0
0	0	0	1	0	0	0
0	1	1	0	0	0	1

Business insights from clustering

After the clustering data seems to be more balanced for the gender and covered by any other company column. Since most customers have level 2 cholesterol that is 150-175 mg/dL thus, insurance cost can be raised for them. Insurance cost can be decreased for students by giving student discounts to make it more affordable for them. Insurance cost can be kept high for Bangalore as there is more % of hospitals.

c) Any other business insight

1. Most people have had 3 years of insurance with the company.
2. Most people have had no checkups in last year and least number of people have gotten checkups upto 5 times.
3. Very few people play any adventure sports
4. Mostly people have visited the doctor twice in last one year. Majority of people have visited the doctor more than once in the past year.
5. Most of the people walk more than 5000 steps daily on an average.
6. In this dataset, the number of 22 years old people is highest.
7. Most people have not had any heart attacks.
8. Mostly people are healthy and have not had any major diseases.
9. Maximum people have an average glucose level of 243 mg/dL. The range of reported average glucose levels is 57 to 277 mg/dL.
10. Mostly people have a bmi of 30.5.
11. Most number of people were admitted into the hospital before 2010.
12. Most people have 77 kg weight in this dataset. The weight range is from 52 to 96 kg.
13. Most people have had 4 weight fluctuations in last one year.
14. Most people have fat percentage of less than 36%.
15. The insurance cost is mostly higher than 7000 and most number of people can afford to pay the insurance cost of 7404.

16. Among these people maximum number people are students.
17. The number of males is higher than the number of females.
18. Maximum number of people have a cholesterol level of 150-175 mg/dL.
19. Maximum number of people have never smoked.
20. Maximum number of hospitals where people were admitted were in Bangalore.
21. Mostly people were not covered by other companies.
22. Mostly people rarely drink.
23. Maximum number of people are student among which number of males is higher than the females.
24. Maximum number of people have 150-175 mg/dL cholesterol level among which number of males is higher than the females.
25. Maximum number of people have never smoked among which number of males is higher than the females.
26. The number of unknown status is 2nd highest in females while in males it is at 3rd position.
27. The second highest category is of formerly smoked among males.
28. Maximum number of people are not covered by any other company among which number of males is higher than the females.
29. Maximum number of people rarely drink among which number of males is higher than the females.
30. Maximum number of people exercise in moderate quantity among which number of males is higher than the females.
31. Maximum number of people were admitted in hospitals in Bangalore among which number of males is higher than the females.
32. The insurance cost is increasing with an increase in the weight of customers. There is an upward trend in the graph.
33. The percentage of males is 65.68% while females have a percentage of 34.31%.
34. 5.4% people have had heart disease while 94.53 have not.
35. 11.9% people have had 3 years of insurance with company which is the highest.
36. 60.08% customers have had no checkups last year.
37. 91.8% customers don't play any adventure sports.
38. 34.67% customers have visited the doctor twice last year.
39. 40.67% customers are student, 40.08% customers are business people and 19.24% are salaried
40. There is a very high positive correlation between weight of the customer and the insurance cost. Then there is some correlation between weight change in last one year and the number of times the customer was admitted in a hospital. There is high negative correlation between weight and the number of times admitted last year.
41. Number of females is higher in unknown category while number of males is higher in never smoked category.
42. insurance cost on male and female is same for student category. Insurance cost slightly more for males in other occupations.
43. Insurance cost is high in Bhubaneswar. Males have higher insurance cost than females.
44. Moderate exercise people have more insurance cost among which cost on males is higher.
45. People with moderate exercising habits have higher bmi which is decreasing as the age increases. Males have higher bmi than females irrespective of the age.

Recommendations

After analysing the data, we can say that most of the people are healthy and have healthy lifestyle. So the company can focus in that area. They can try to attract more people to get insurance by giving discounts.

Since 22 year olds number is high, so online advertising can be done to attract their attention as the younger generation is highly active on internet. Social media can be used for advertising.
Referral programs can be set up to increase awareness about the companies policies thus, will help in selling more insurance plans.

Since most people are students so Student discounts can be set up to attract them.
Insurance cost can be raised for the people in business. This can increase the profits.

Since female population is low in purchasing insurance so attractive discounts can be given to females to increase their participation.

Insurance cost of people with Cholesterol level of 150-175 mg/dL can be raised as it increases the risks of getting sick.

In metropolitan areas like Bangalore and Bhubaneswar insurance cost can be raised.

Loyalty schemes can be started for customers who are not covered by other companies.

Insurance cost of old age people can be raised as their risk of getting diseases increases.

Insurance cost of people with higher weight can be raised as the insurance cost increases with increase in weight.

Most of the customers belong to categories like Do not smoke, Rarely drink, Moderate exercise so insurance cost can be increased.