Capstone Project

Insurance Cost Analysis

Contents

PLOTS

Tables

\

1. Problem Understanding

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 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

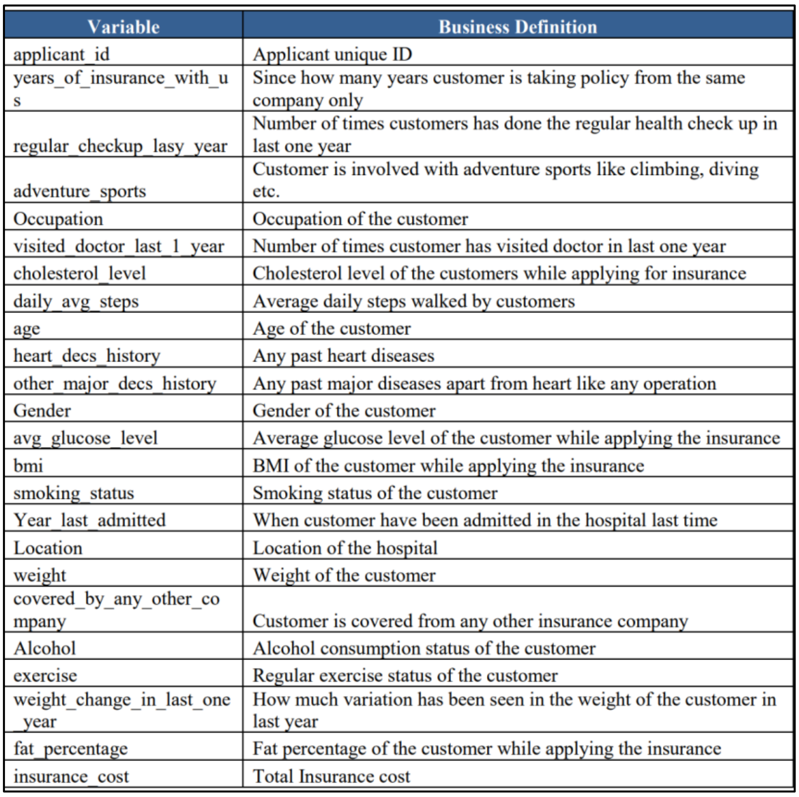
The objective of this exercise is to build a model, using data that provide the optimum insurance cost for an individual. Understanding about the health and habit related parameters through exploratory data analysis can help us in the estimated cost of insurance.

c) Understanding business/social opportunity

The business opportunity lies in accurately measuring the insurance cost which will help in avoiding over or under charge to customers while social opportunity lies in providing suitable type of insurance cover, personalised to consumer's health and economical profile.

2. Data Report

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

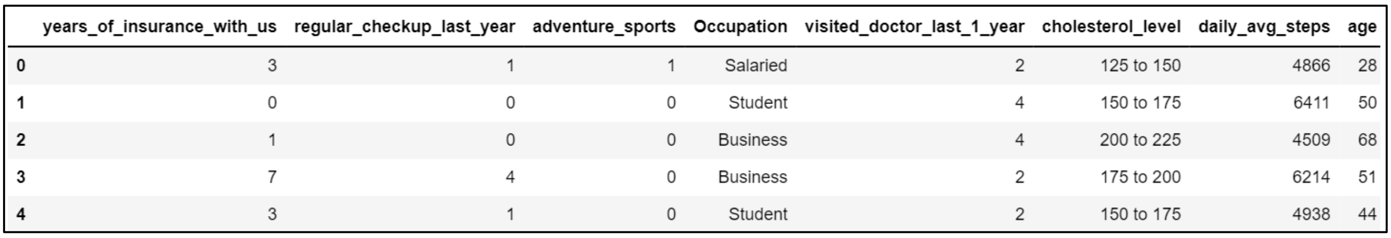


Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the same company only
regular_checkup_lasy_year	Number of times customers has done the regular health check up in last one year
adventure_sports	Customer is involved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

Table 1: Data Dictionary

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

The original dataset has 25000 rows and 24 columns. Then ‘application_id’ was dropped and new column called ‘Cholesterol_degree’ was created. This new column was created because the ‘cholesterol_level’ data was given in range form which would make it difficult interpret the data. So, it was converted into interpretable categorical format. The data head looks as follows:



	years_of_insurance_with_us	regular_checkup_last_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age
0	3	1	1	Salaried	2	125 to 150	4866	28
1	0	0	0	Student	4	150 to 175	6411	50
2	1	0	0	Business	4	200 to 225	4509	68
3	7	4	0	Business	2	175 to 200	6214	51
4	3	1	0	Student	2	150 to 175	4938	44

Table 2: Snapshot of data head

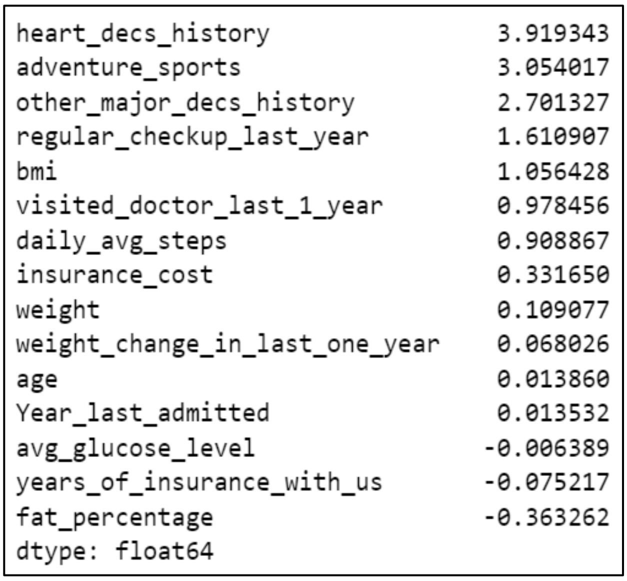
There are no duplicated variables but ‘bmi’ and ‘Year_last_admitted’ has 990 and 11881 null values respectively.

Descriptive statistic summary is as follows:

1. The minimum amount of sample customers having insurance with us is 0 with maximum value is 8, with a mean of 4
2. The sample mostly consist of student population and male gender
3. The age range of the sample population is from 16 to 74 years with a mean of 45 years
4. The range of average glucose level varies from 57 (low level) to 277 (diabetic level), with a mean of 167 (pre-diabetic level)
5. The fat percentage is from 11% to 42% with a mean of 28%. Majority of them have optimal cholesterol degree.
6. The range of BMI is from 12 to 100, with a mean of 31. This data has not been treated hence it cannot be interpreted well. It will be treated in future steps
7. Significant proportion of sample is from Bangalore and have never smoked and rarely drink
8. The weight range is from 52 to 96 kg with a mean of 71. The sample also exercises moderately.
9. Most of them are not covered by other insurance company.
10. The insurance costs range from 2468 to 67,870 with a mean of 27,147.

2.3 Understanding of attributes (variable info, renaming if required).

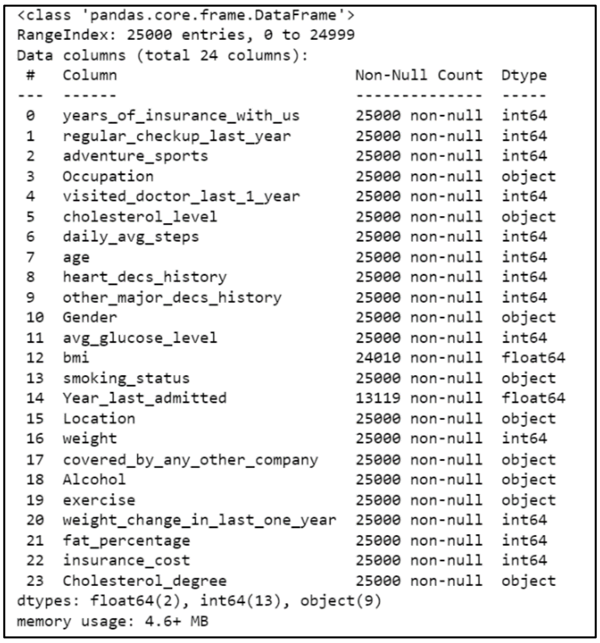
There’s a degree of skewness in numerical variables, as can be interpreted from skewness table below:



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

Table 3: Skewness table

There are 2 float, 13 integer and 9 object type variables. The 'Salried' column was renamed to 'Salaried'.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 24 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                            24010 non-null  float64
 13  smoking_status                 25000 non-null  object 
 14  Year_last_admitted             13119 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  
 23  Cholesterol_degree             25000 non-null  object 
dtypes: float64(2), int64(13), object(9)
memory usage: 4.6+ MB
```

Table 4: Data dtypes

3. Exploratory Data Analysis

3.1 Missing value Treatment and Variable Transformation

Before conducting univariate and bivariate analysis, the data set was treated for missing values, so that we get a better visualisation and interpretation of the variables. Only 'bmi' and 'Years_last_admitted' had 990 and 1181 missing values respectively. These were treated in different ways.

In the 'bmi' variable, the 'nan' values were imputed with median values using 'fillna' function. Median was used because it's a suitable imputing technique when a variable has outliers and 'bmi' has lot of outliers.

The 'Years_last_admitted' variable was renamed as 'admitted_last_30yrs'. This was done because, it is not possible to find actual missing years and imputing it with mean or median, might result in wrong representation of the variable. So, it was converted into a categorical variable 'admitted_last_30yrs', where if the customer was admitted in last 30 years they are labelled 1 and if not, they are labelled 0.

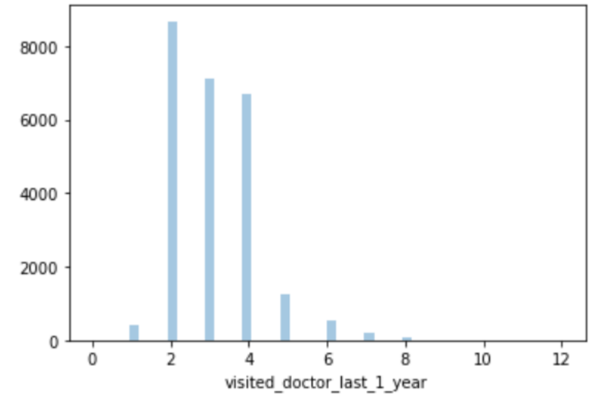
Additionally, as mentioned before, the 'cholesterol_level' data was given in range form which would make it difficult interpret the data. So, it was transformed into a new column called 'Cholesterol_degree' having labels- 'Optimum', 'Intermediate' and 'High' level which indicate different level of cholesterol.

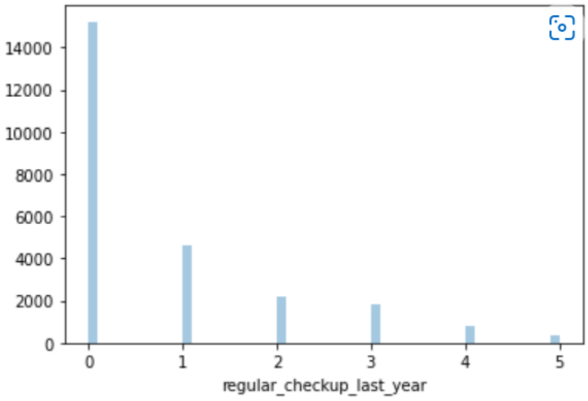
3.2 Univariate analysis and bivariate analysis- Numerical variables

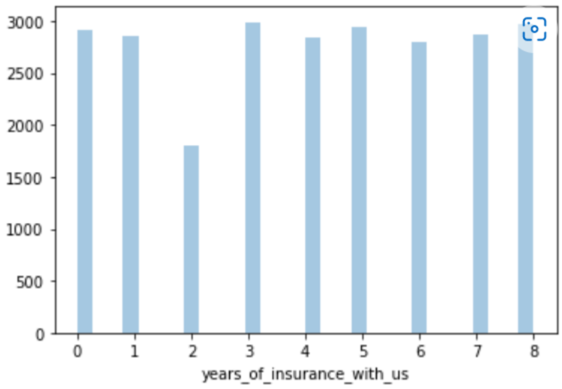
The univariate and bivariate analysis was conducted separately for numeric and categorical variables.

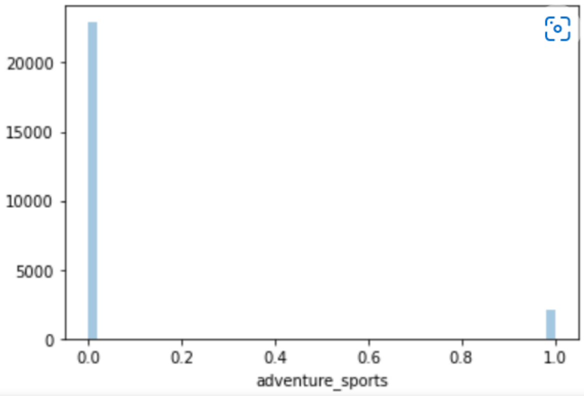
3.2.2 Univariate analysis

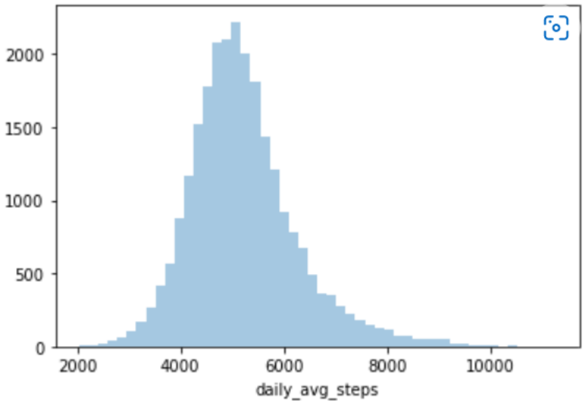
The univariate analysis of numerical variables is as follows:

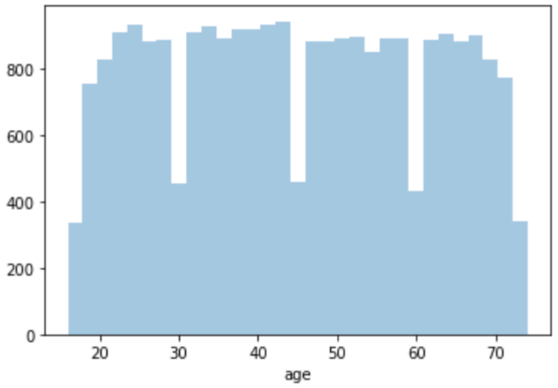


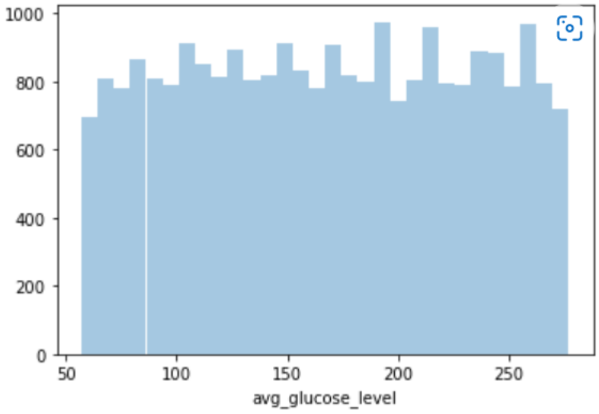


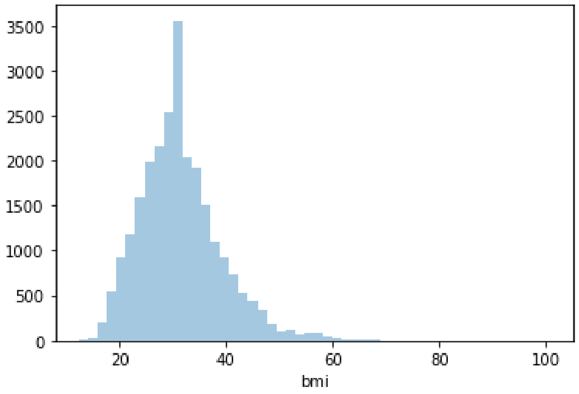

3000
2500
2000
1500
1000
500
0
0
1
2
3
4
5
6
7
8
years_of_insurance_with_us

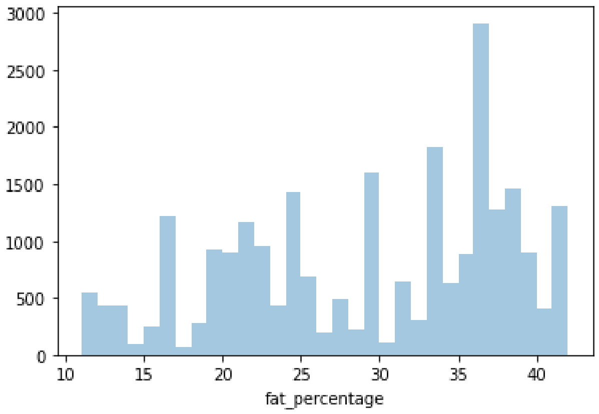

20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
adventure_sports

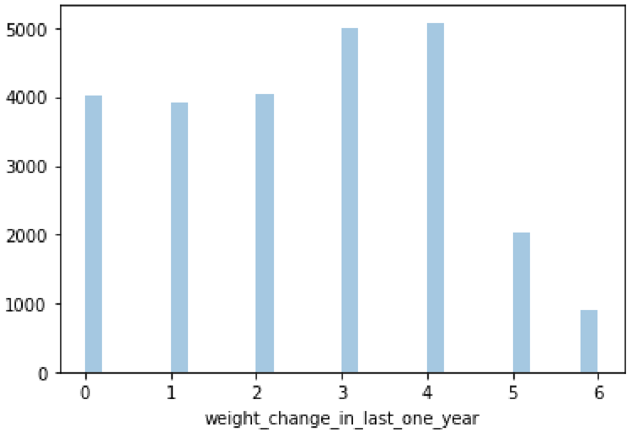

2000
1500
1000
500
0
2000
4000
6000
8000
10000
daily_avg_steps

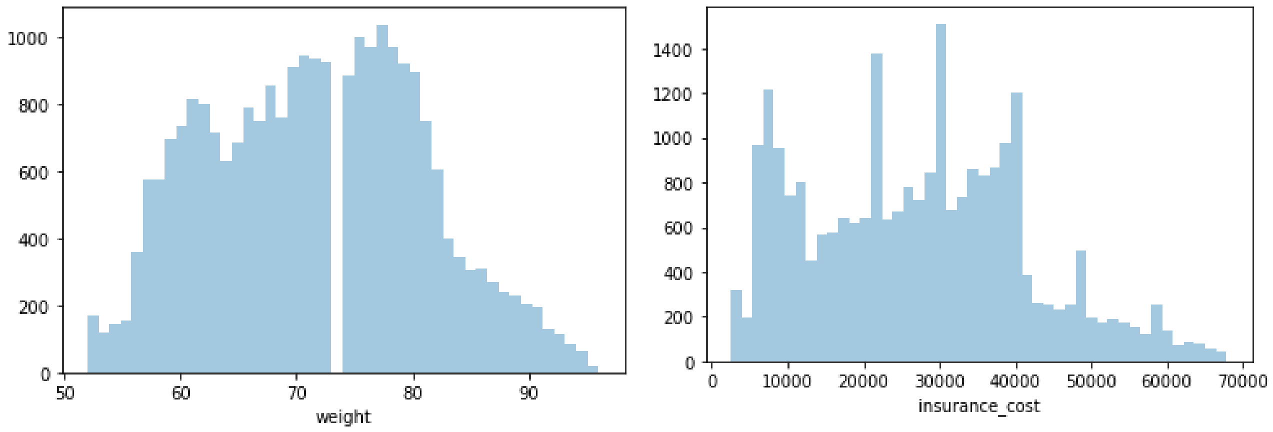

800
600
400
200
0
20
30
40
50
60
70
age


1000
800
600
400
200
0
50
100
150
200
250
avg_glucose_level

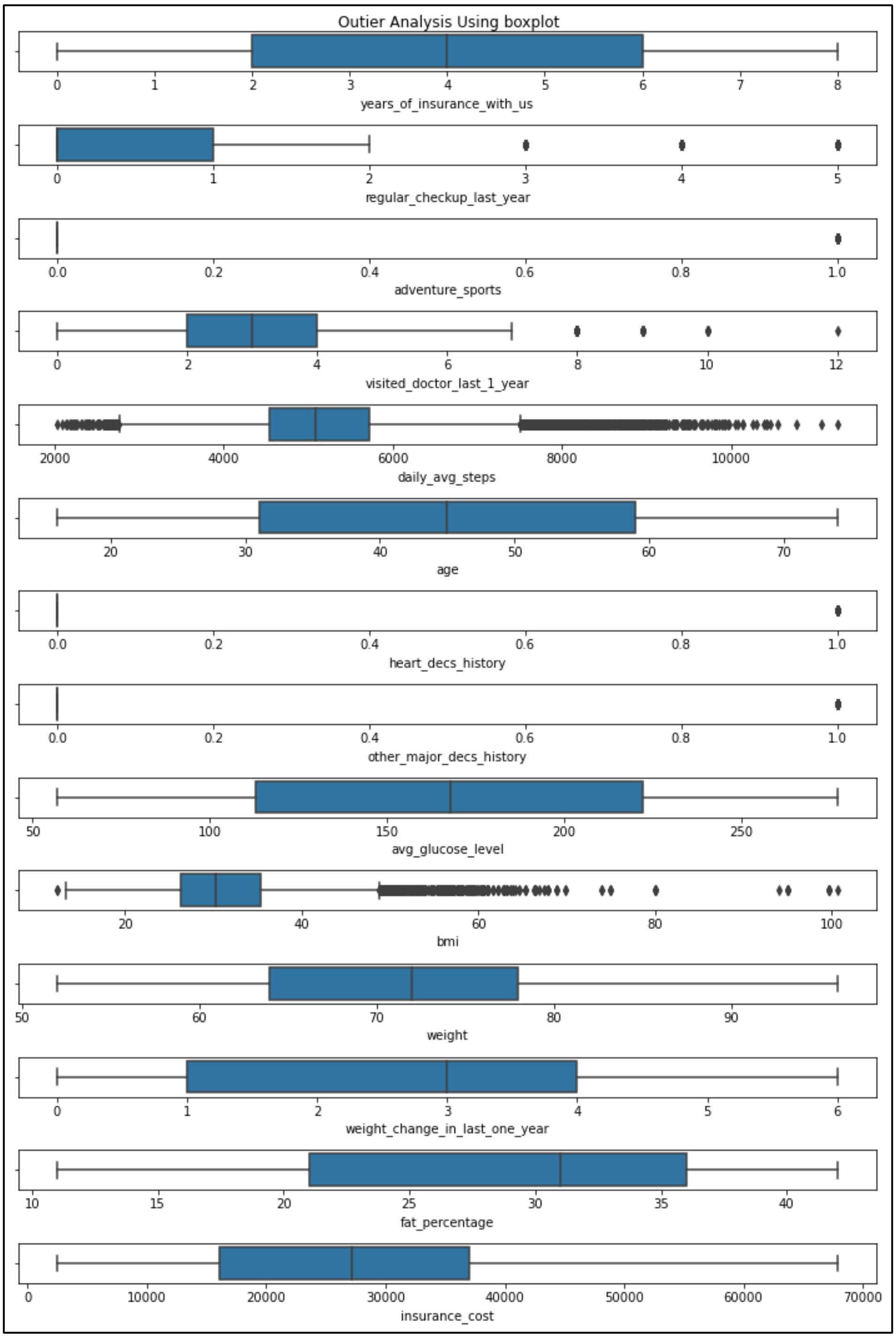

3500
3000
2500
2000
1500
1000
500
0
20
40
60
80
100
bmi


3000
2500
2000
1500
1000
500
0
10
15
20
25
30
35
40
fat_percentage

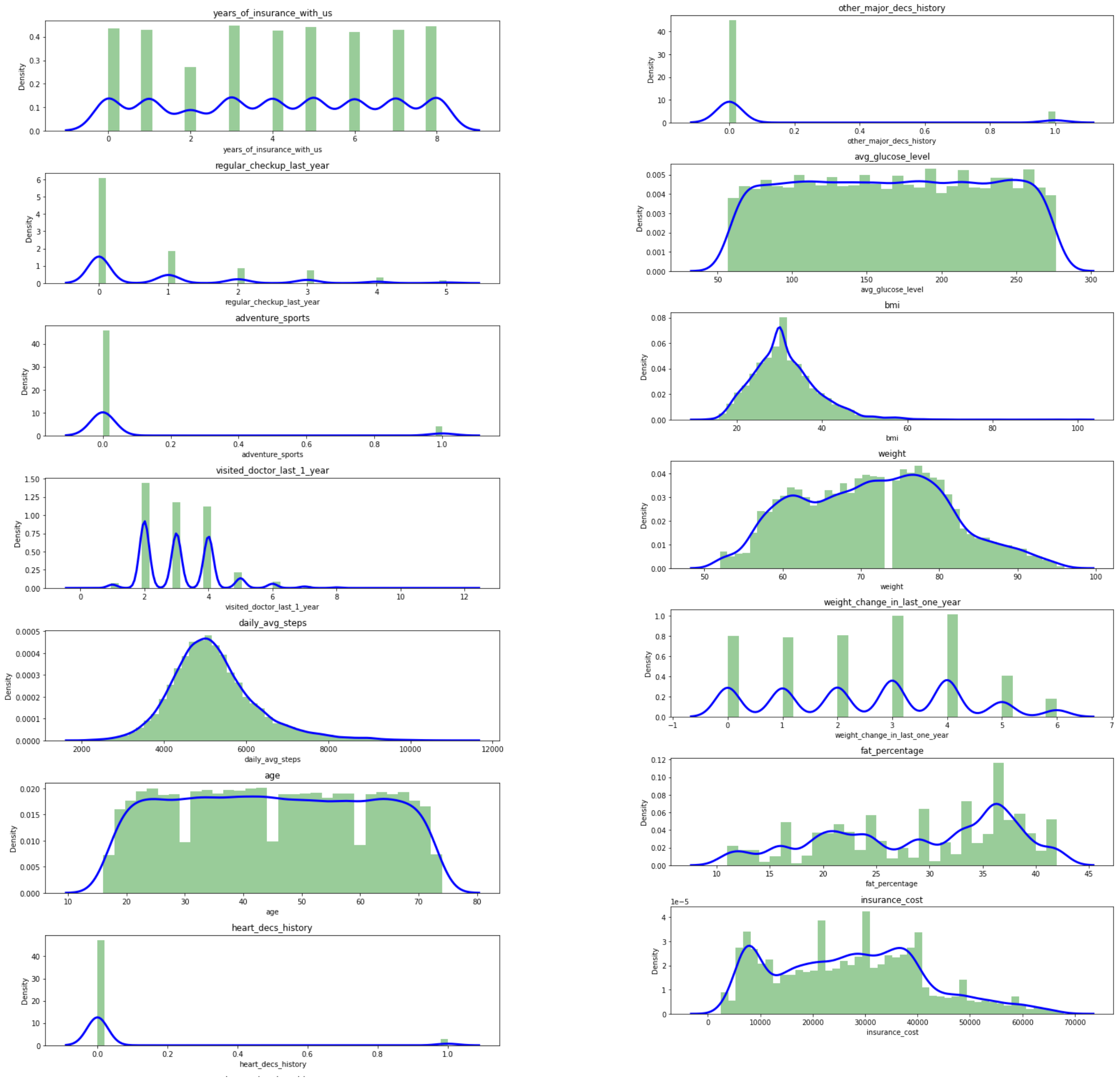

5000
4000
3000
2000
1000
0
0
1
2
3
4
5
6
weight_change_in_last_one_year



Plot 1: Distribution plot of numerical variables



Plot 2: Boxplot showing outliers present in the numerical variables

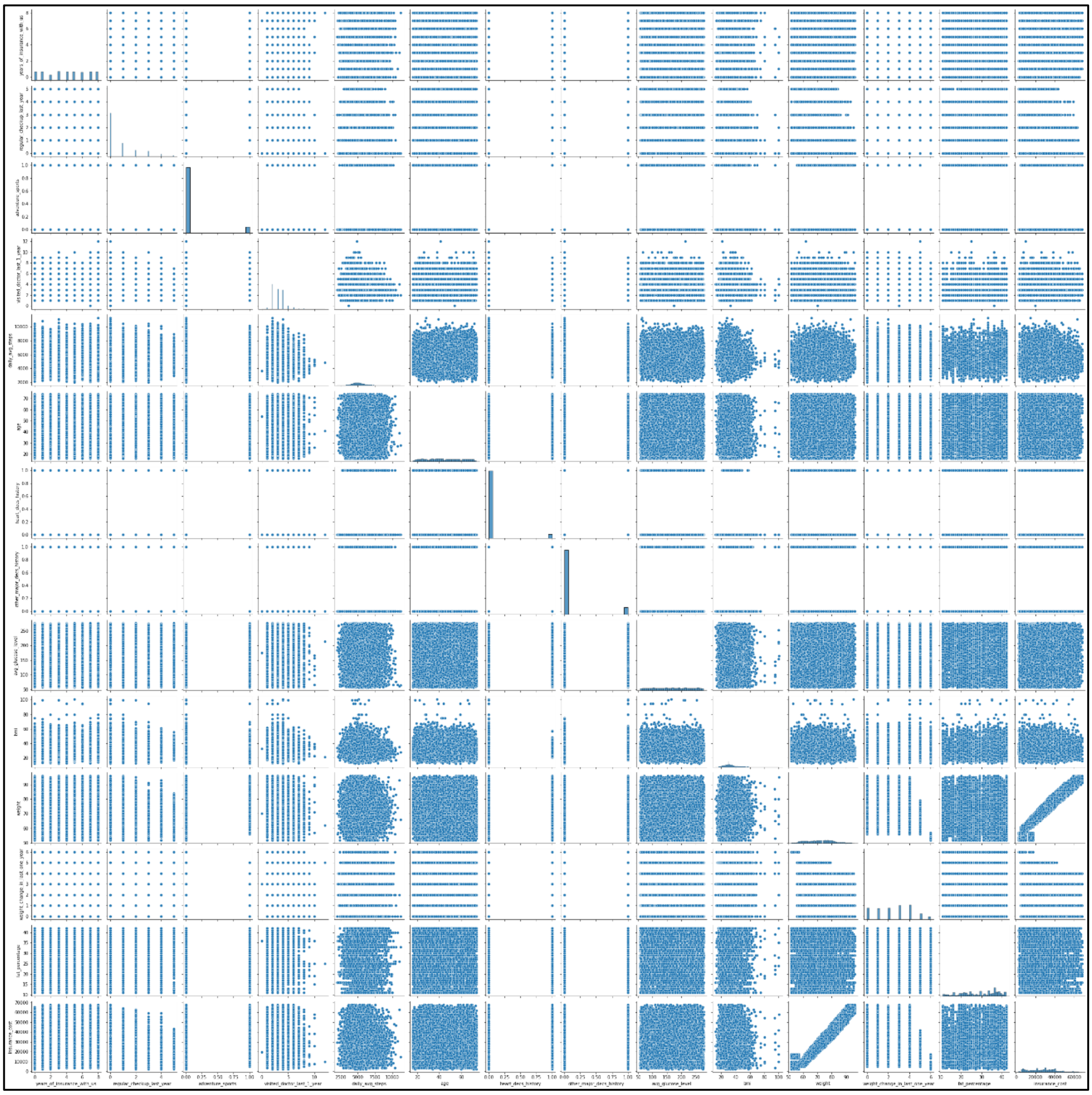


Plot 3: Skewness plot

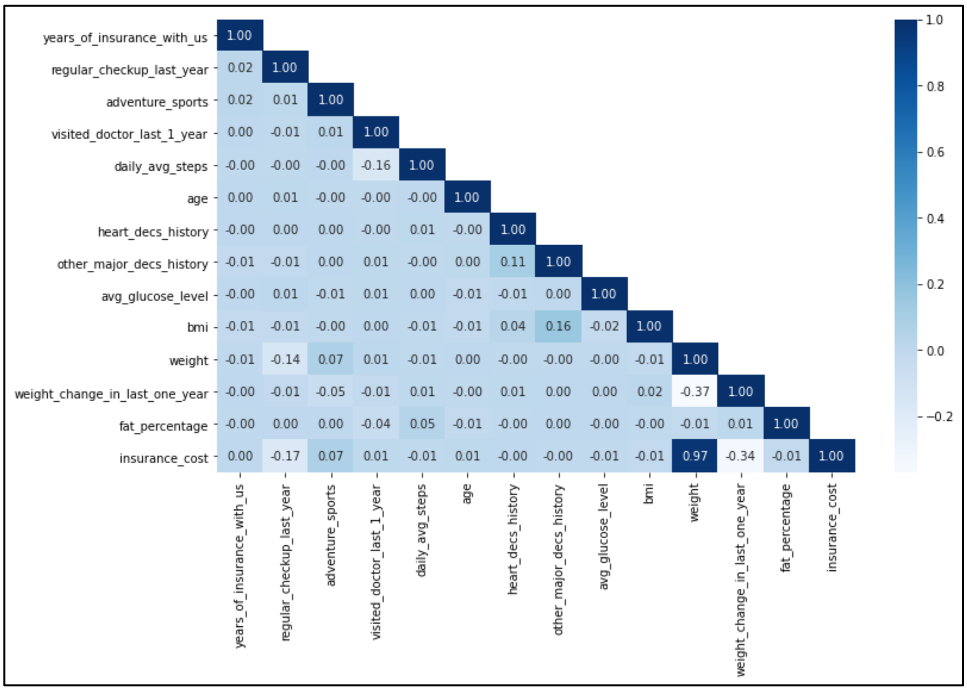
Insights from univariate analysis:

1. No distribution seems normal. As seen from plot 3, the plots are either skewed, multimodal or have a plateau shaped distribution.
2. Outliers are present in ‘bmi’, ‘regular_checkup_last_year’, ‘visited_doctor_last_1_year’ and ‘daily_avg_steps’

3.2.3 Bivariate analysis



Plot 4: Pairplot of numerical variables



Plot 5: Correlation Heatmap

Insights from bivariate analysis:

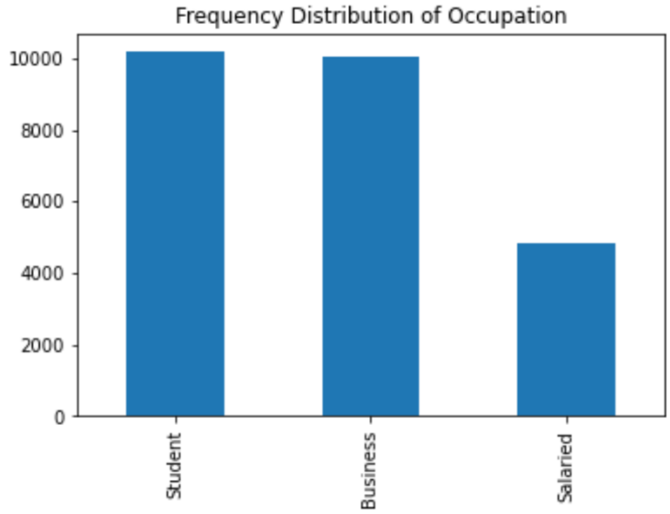
1. As seen from correlation heatmap and Pairplot, there is no evidently correlation among different variables, except for weight and insurance cost.

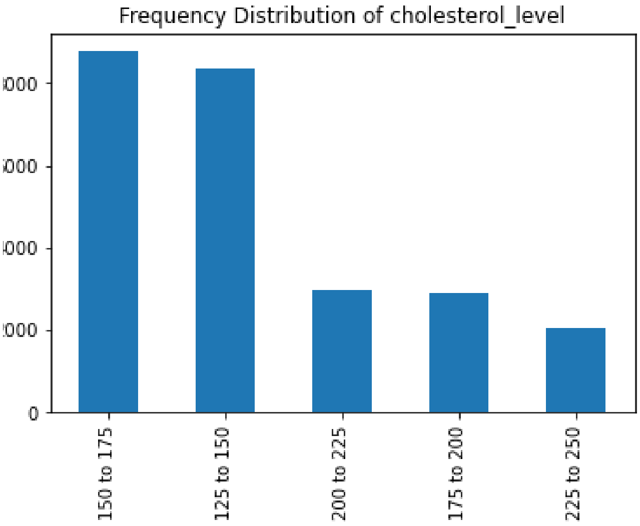
3.3 Univariate analysis and bivariate analysis-Categorical variables

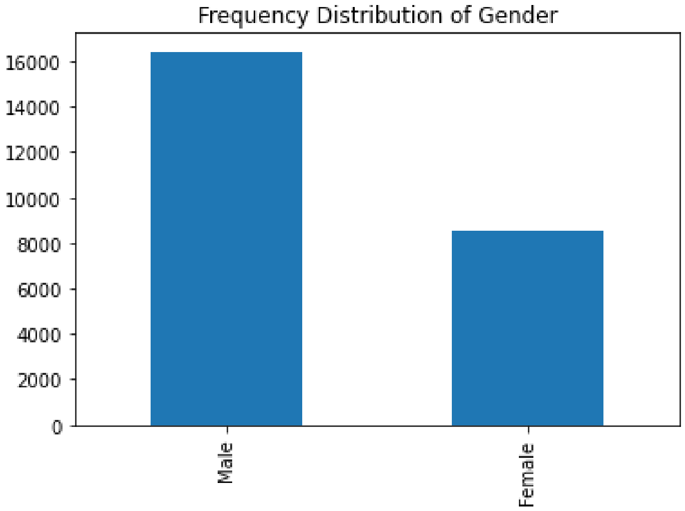
Following ae the categorical variables that are analysed:

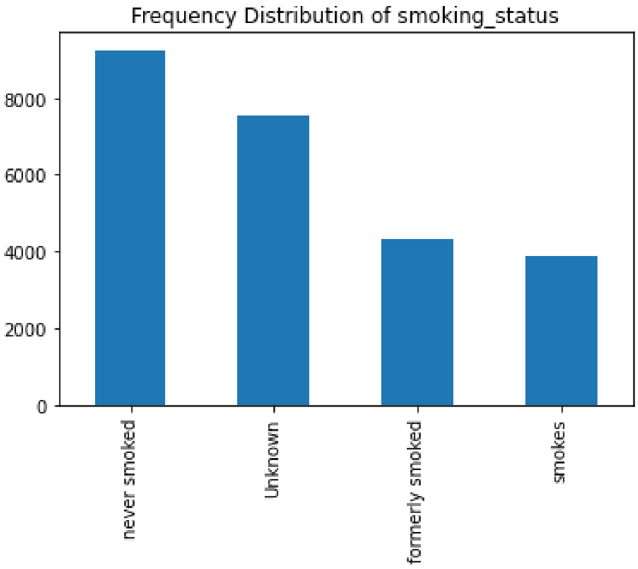
- 'Occupation'
- 'cholesterol_level'
- 'Gender'
- 'smoking_status'
- 'admitted_last_30yrs'
- 'Location'
- 'covered_by_any_other_company'
- 'Alcohol'
- 'exercise'
- 'Cholesterol_degree'

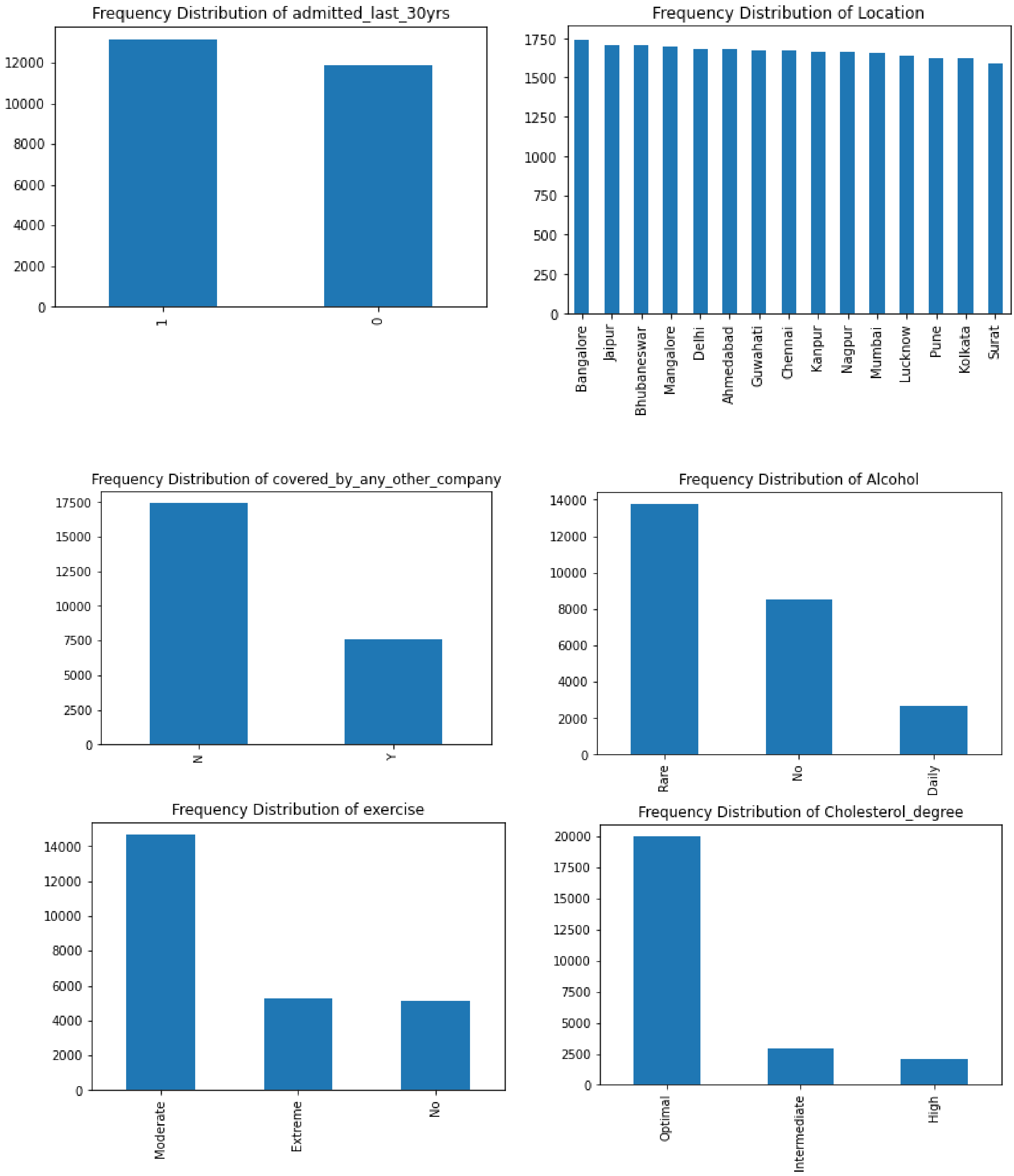
3.3.1 Univariate analysis











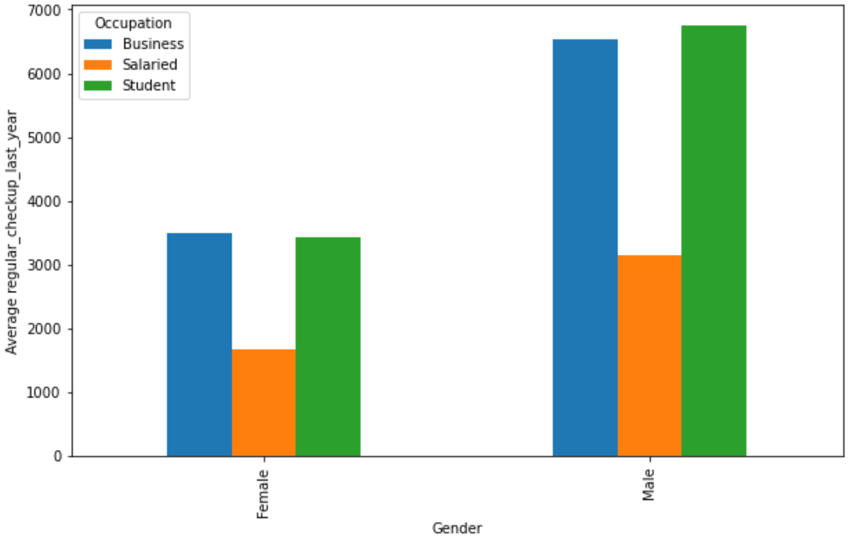
Plot 6: Distribution plot of Categorical variables

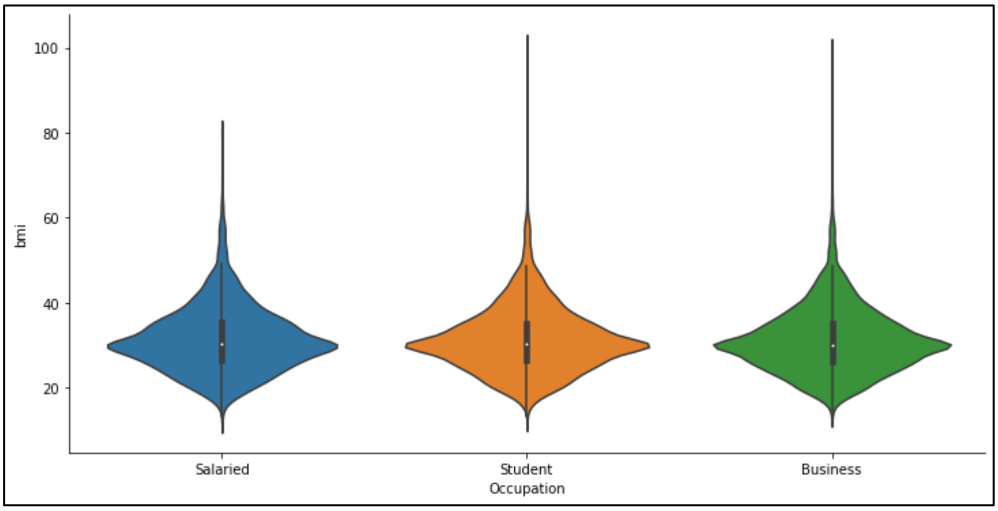
Insights from univariate analysis:

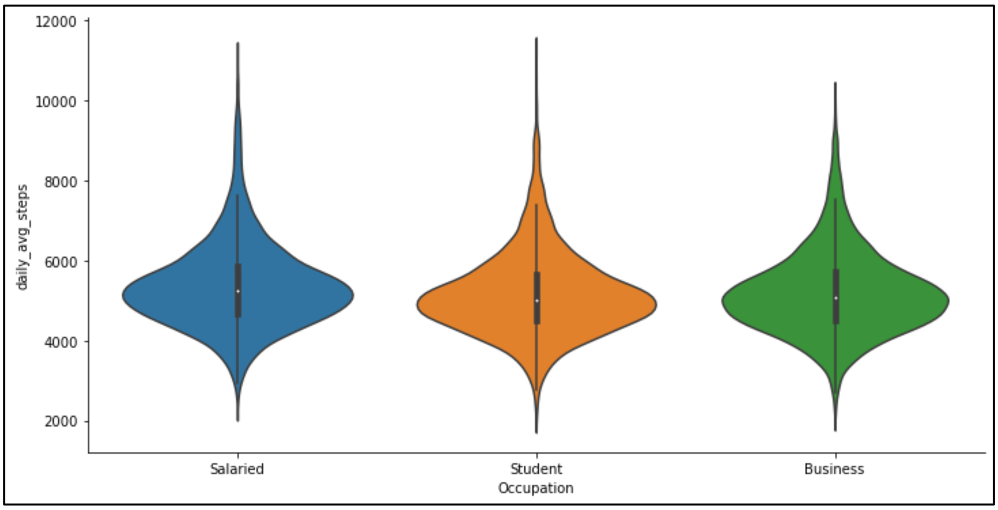
1. The sample mainly consists of Student and businesspeople
2. Most of the sample population lies in optimal cholesterol level of 150 -175 and 125-150. Few lie in intermediate (200-225) and high cholesterol level (225-250)
3. The sample population had more males than females
4. Majority of people never smoked, and the smoking status is unknown for many sample population

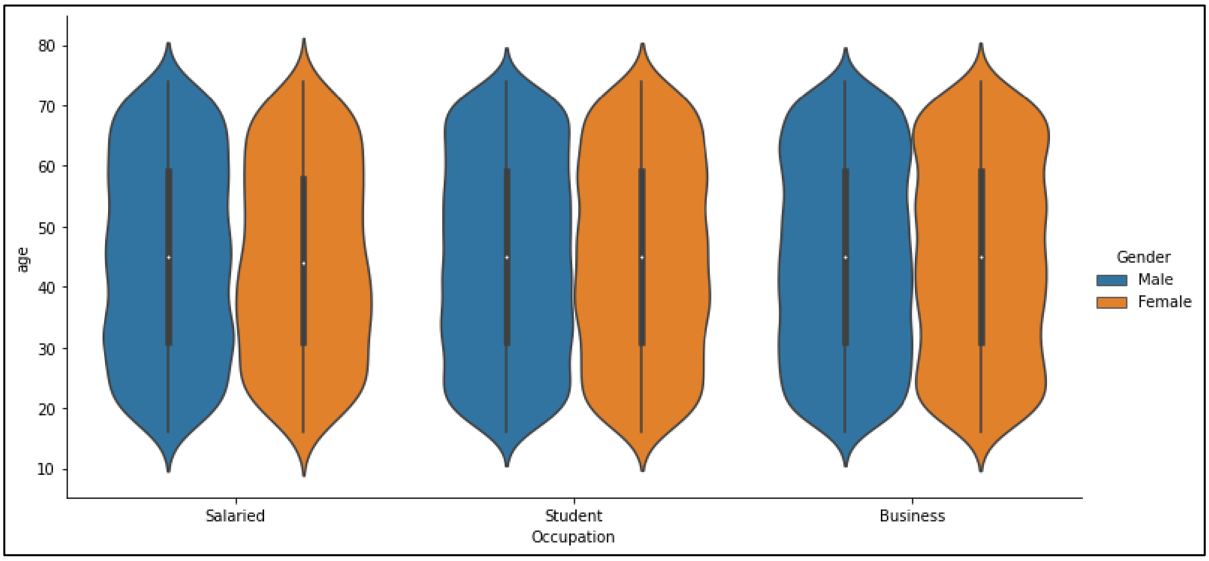
5. Most people were admitted in last 30 years to hospital
6. Majority of people are from Bangalore followed by Jaipur, Bhubaneshwar and Mangalore
7. Many people are not covered by other insurance company. Less than half the people are covered by other companies
8. Most of the people rarely drink alcohol, while half of them dont drink. Less than 10% of people drink daily
9. Around half of the people do moderate exercise, while the other half of people do extreme or no exercise

3.3.2 Bivariate analysis

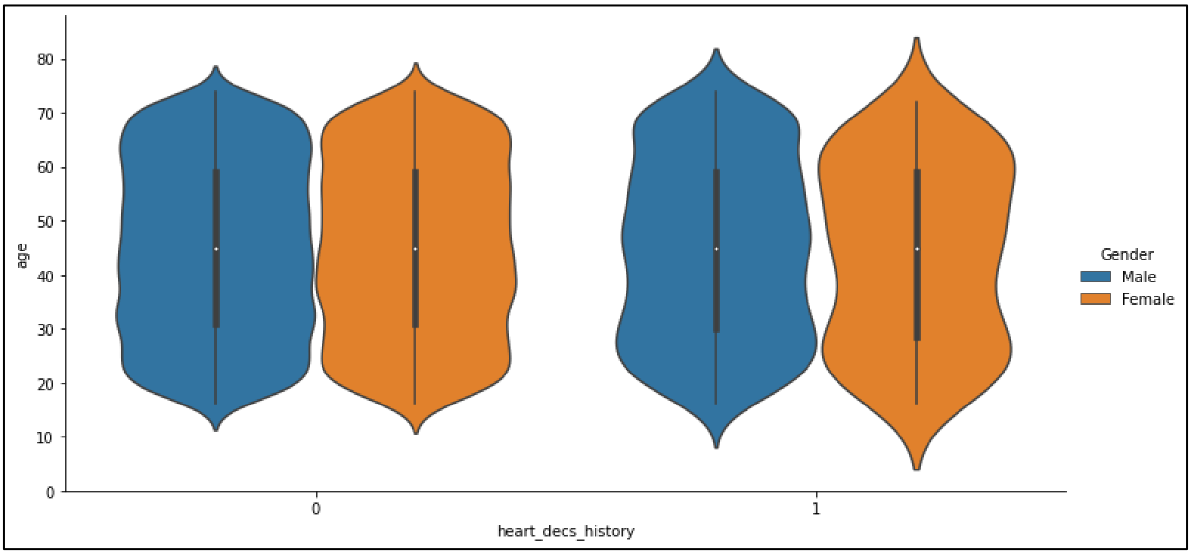


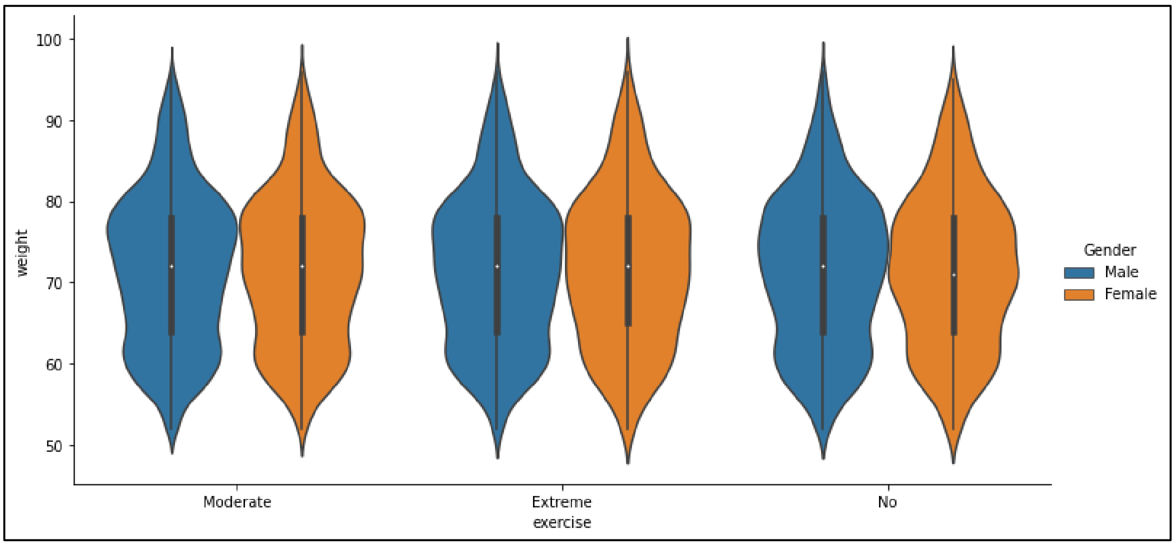




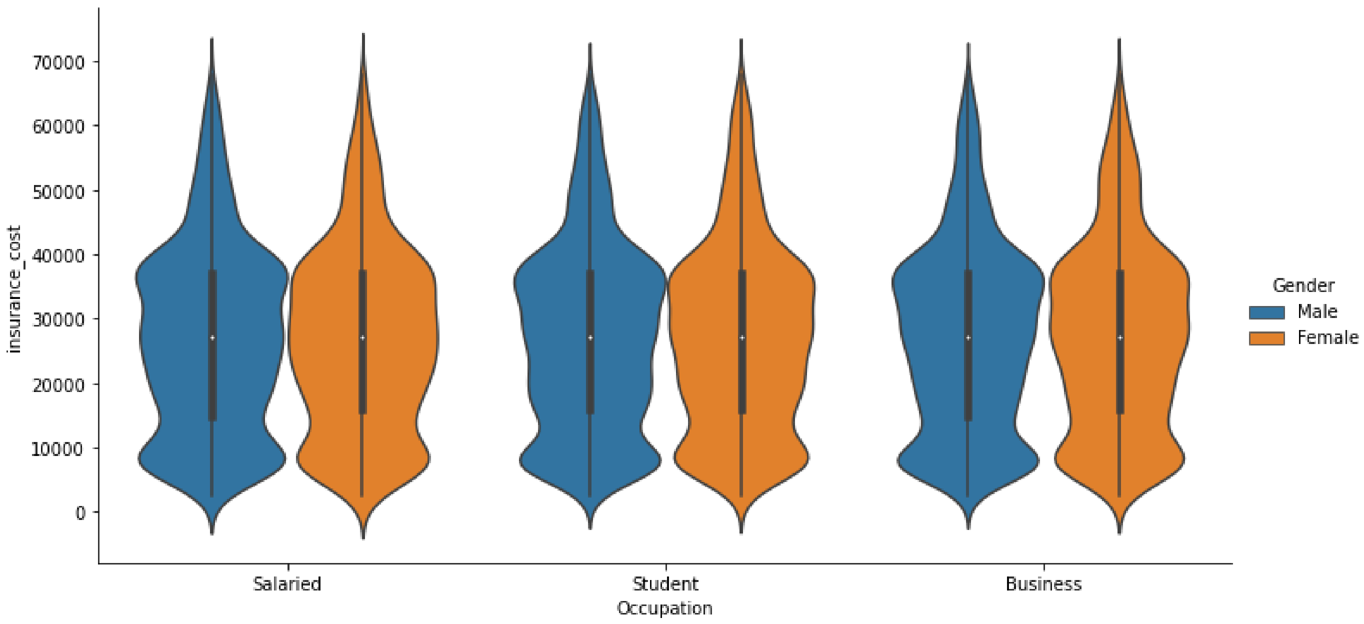


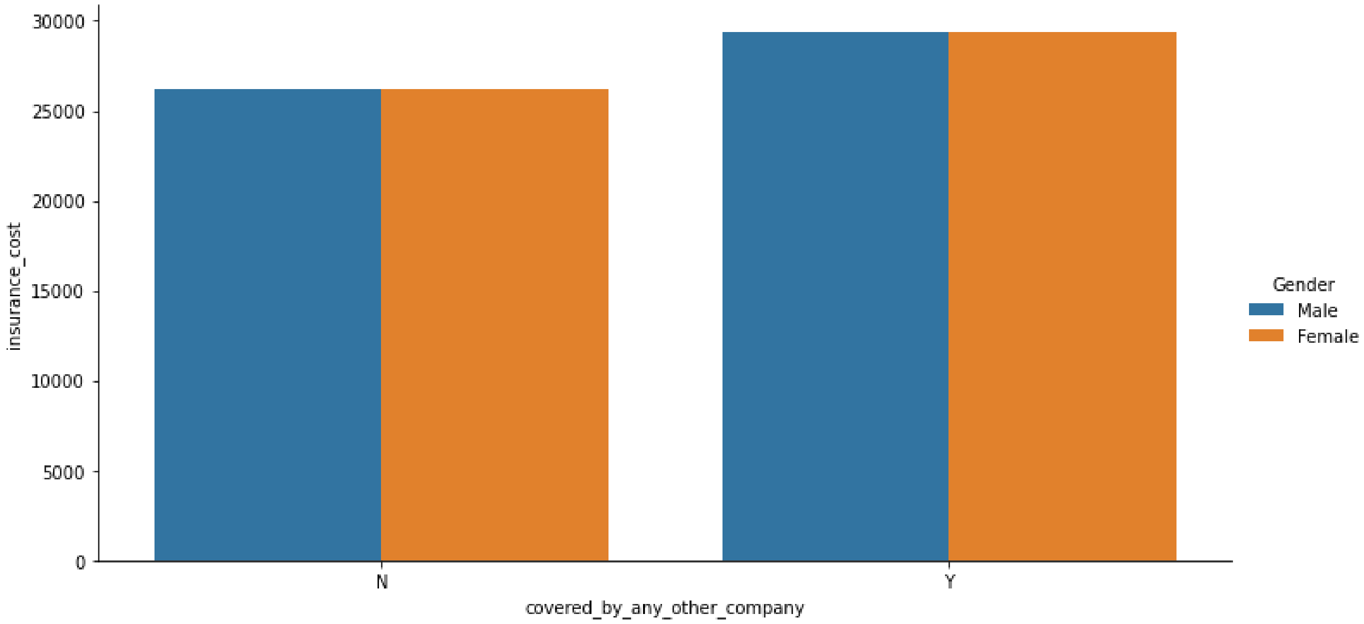
Plot 7: Occupation based analysis

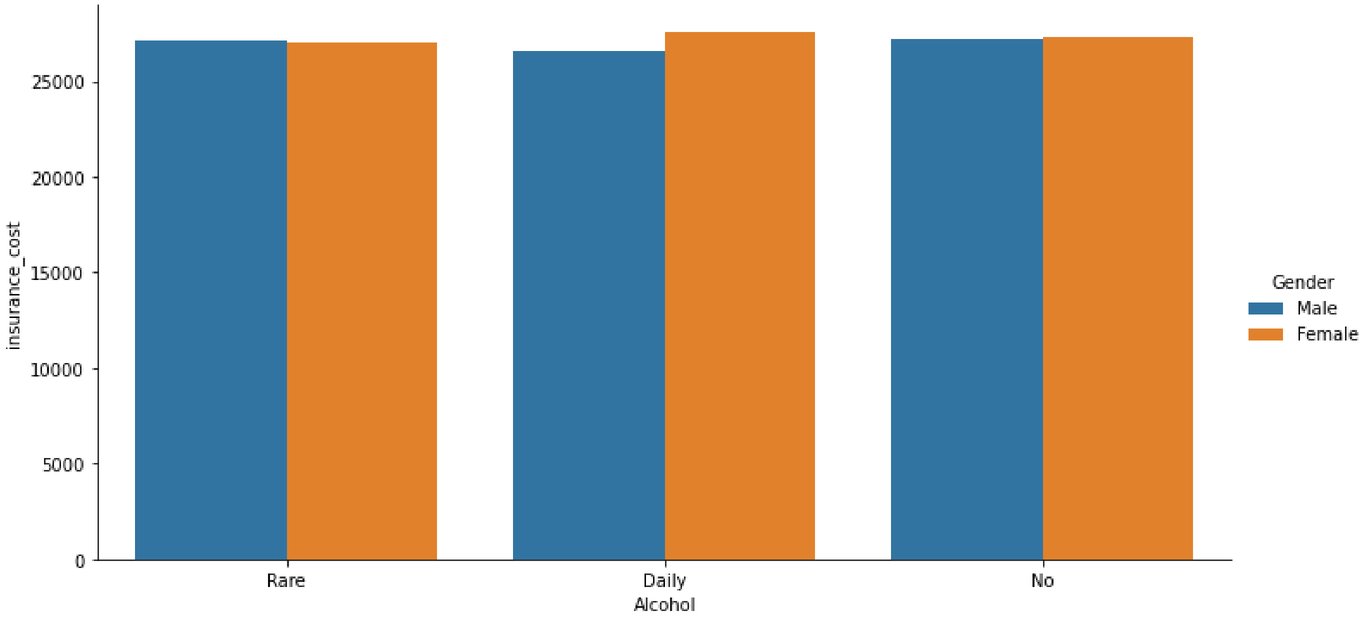


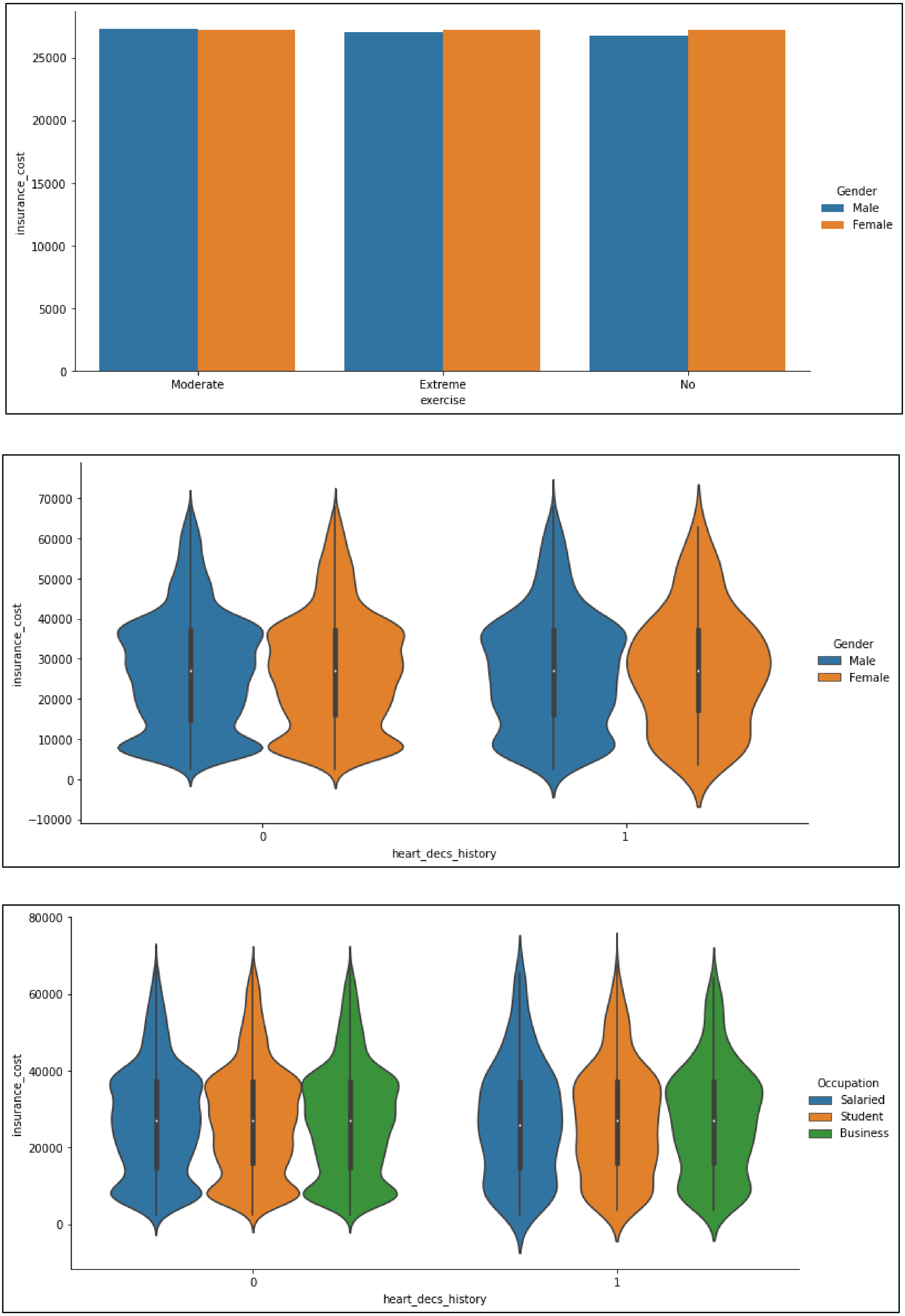


Plot 8: Gender based analysis

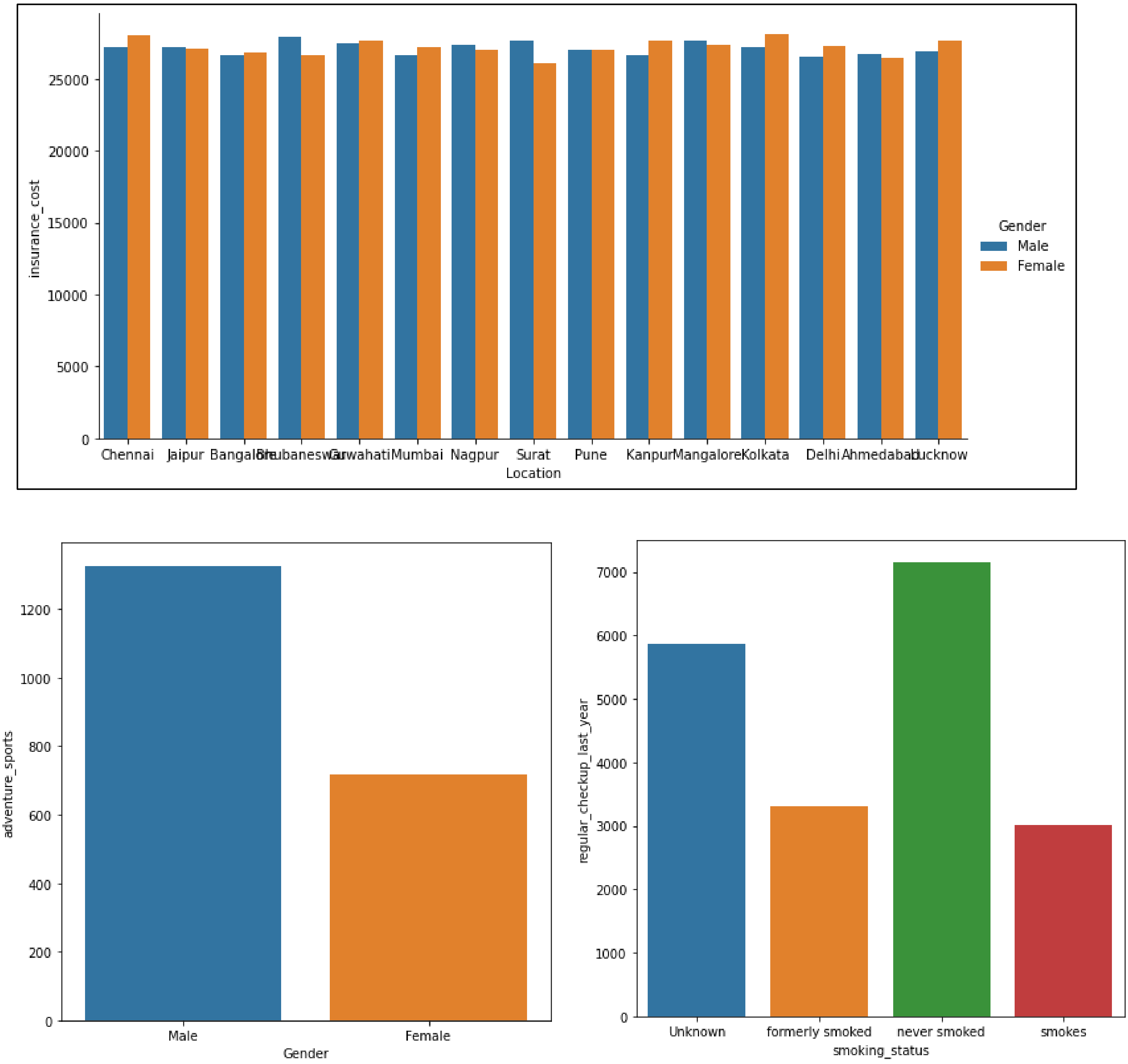

insurance_cost
70000
60000
50000
40000
30000
20000
10000
0
Salaried
Student
Business
Occupation
Gender
Male
Female


insurance_cost
30000
25000
20000
15000
10000
5000
0
N
Y
covered_by_any_other_company
Gender
Male
Female


insurance_cost
25000
20000
15000
10000
5000
0
Rare
Daily
No
Alcohol
Gender
Male
Female



Plot 9: Insurance cost based analysis



Plot 10: Other variables plot analysis

Insights from bivariate analysis:

1. more males got their check-up last year as compared to females. We might be getting this visualisation because there are a smaller number of females than males in the dataset (unbalanced data)

2.the range of 'bmi' of salaried people is less than that of students and businessman and it seems most of them have bmi around 30

3. the age range of salaried, student and businesspeople is same for both males and females

4. the 'daily_average_steps' range is more for salaried and students than for businesspeople. But they all average at above and around 5000 steps

5. According to plot 8, females between the age 25-45 have less heart disease and females above the age of 50 have more heart disease. This might be because after 50 females go through menopause, which might make them more venerable to heart disease.

6. the weight range of people doing no, moderate and extreme exercise is same. However, there are slightly less people who weigh 60-70 kgs and do moderate exercise when compared to people who do no exercise

7. According to plot 9, the insurance cost to all three types of occupation is same. Most common insurance cost comes around to be 10,000 and 40,000 in all three occupation type.

8. The insurance cost is higher for people who are also covered by other companies. And it is slightly higher for females who drink alcohol daily.

9. The insurance cost is slightly in the higher range for people with heart disease history

10. Interestingly, the insurance cost of values 10,000 and 40,000 is more frequent among people with no hear disease than with heart disease in all three occupation type

11. The insurance cost is generally higher for females in most cities except for Surat Bhubaneshwar, Nagpur, Mangalore and Ahmedabad

12. Males go for more adventure sports than females

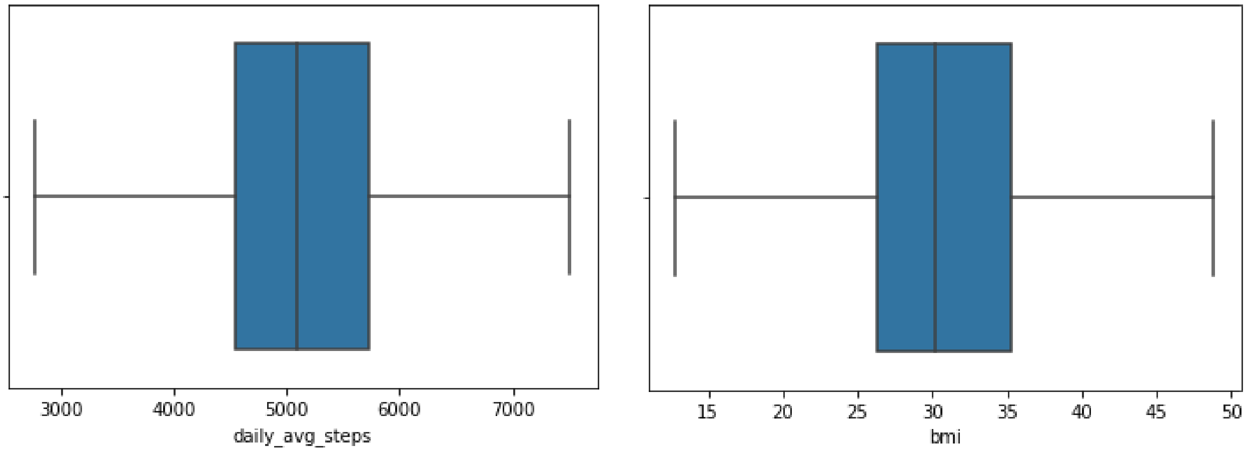
13. People who never smoke or their smoking status is unknown frequented the hospital visit last year as compared to former smokers and smokers

3.4 Removal of unwanted variables (if applicable). Outlier treatment (if required). Addition of new variables (if required)

The only variable which was removed was 'application id' since it had no value. Three variables were added

1. Cholesterol_degree – to convert range format of cholesterol_level into useable categorical form
2. Adimitted_in_past_30_years- since missing values could not be used to fill the missing ears, this column was converted into categorical form denoted as '1' (admitted) and '0' (not admitted)
3. Clus_kmeans- the cluster label that helped in identifying two groups.

Outlier treatment was done for 'bmi' and 'daily_avg_steps' using flooring and capping method. The treated data looks as follows:



Plot 11: Boxplot of 'bmi' and 'daily_avg_steps' showing treated outliers

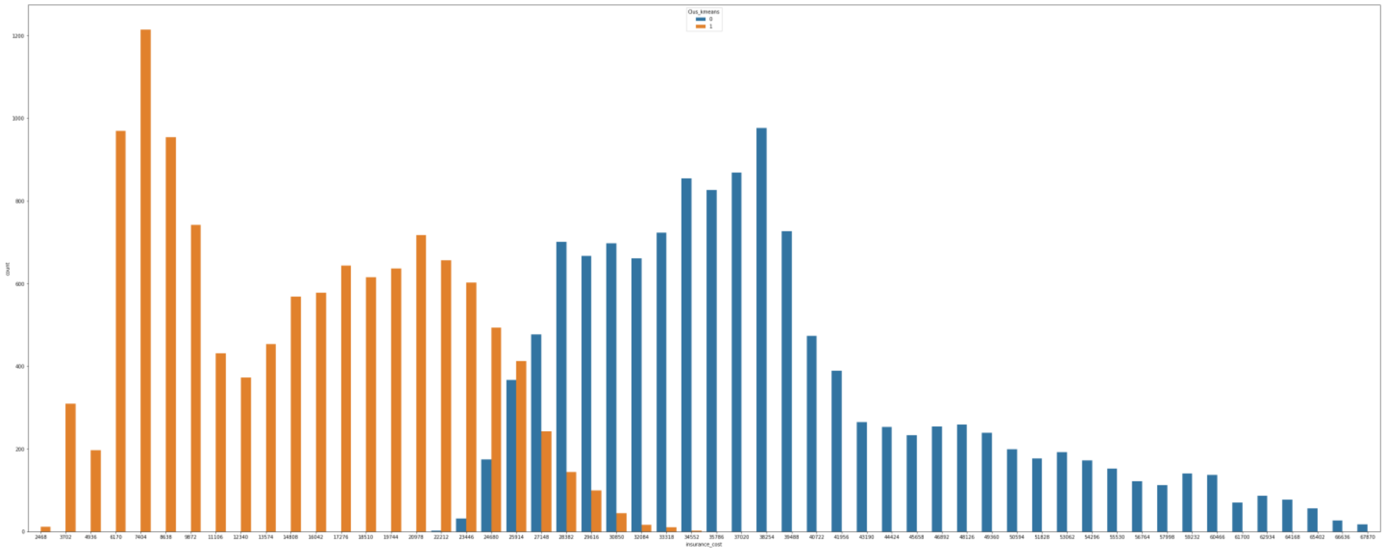
4. Business insights from EDA

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

The data seems unbalanced if we look into the proportion of males and females, where there are more males than females in the given dataset. SMOTE technique can be used to get a more balanced dataset. A balanced representation of gender can help in making better suggestions about the insurance costs for different gender. Right now, due to a smaller number of females the insurance cost determining factor might affect the insurance cost predictability ability.

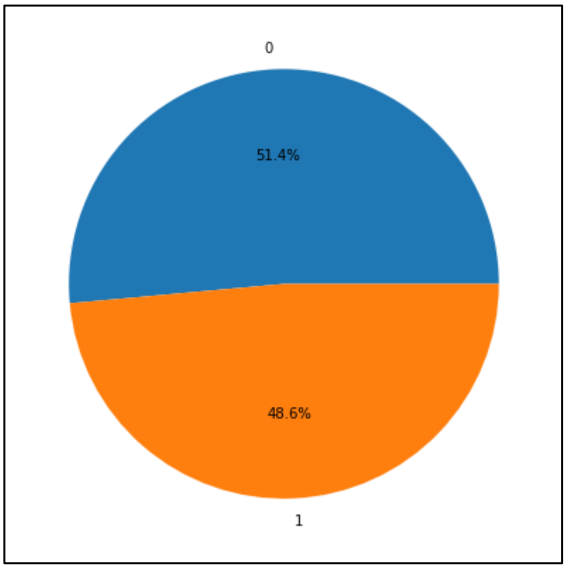
4.2 Any business insights using clustering (if applicable).

Kmeans clustering was used to find clusters on scaled numeric dataset. Then using WSS plot, optimal number of clusters was established by finding the elbow. Here, the elbow was formed at 2, so 2 clusters were constructed, which gave the silhoutte score of 0.13 (nearer the silhoutte score to 1 the better clustered is the data). The two clusters can be visualised as follows:



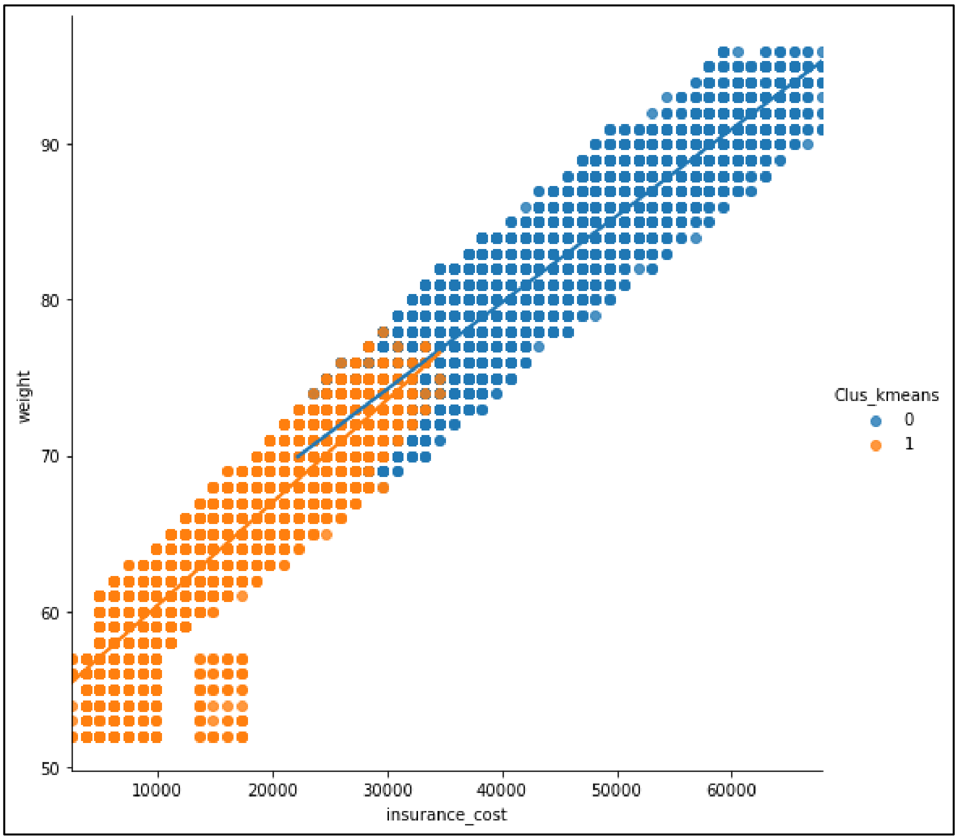
Plot 12: Countplot of two clusters with insurance cost

As it can be seen, the two clusters form two groups, one group whose insurance cost is less (below 35,000) and another whose insurance cost is in the higher range (above 35,000). 51.4% of data belong to group 0, or high insurance payers and 48.6% of them belong to low insurance payers.

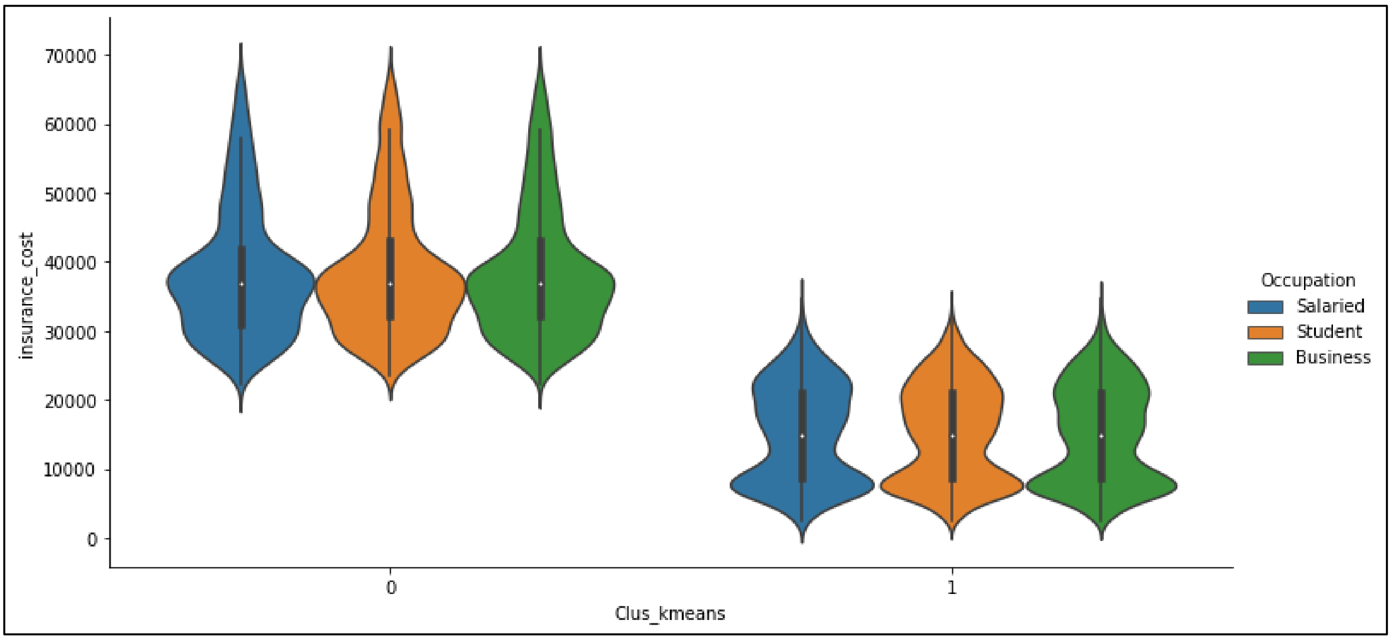


Plot 13: Pie chart showing proportion of two clusters

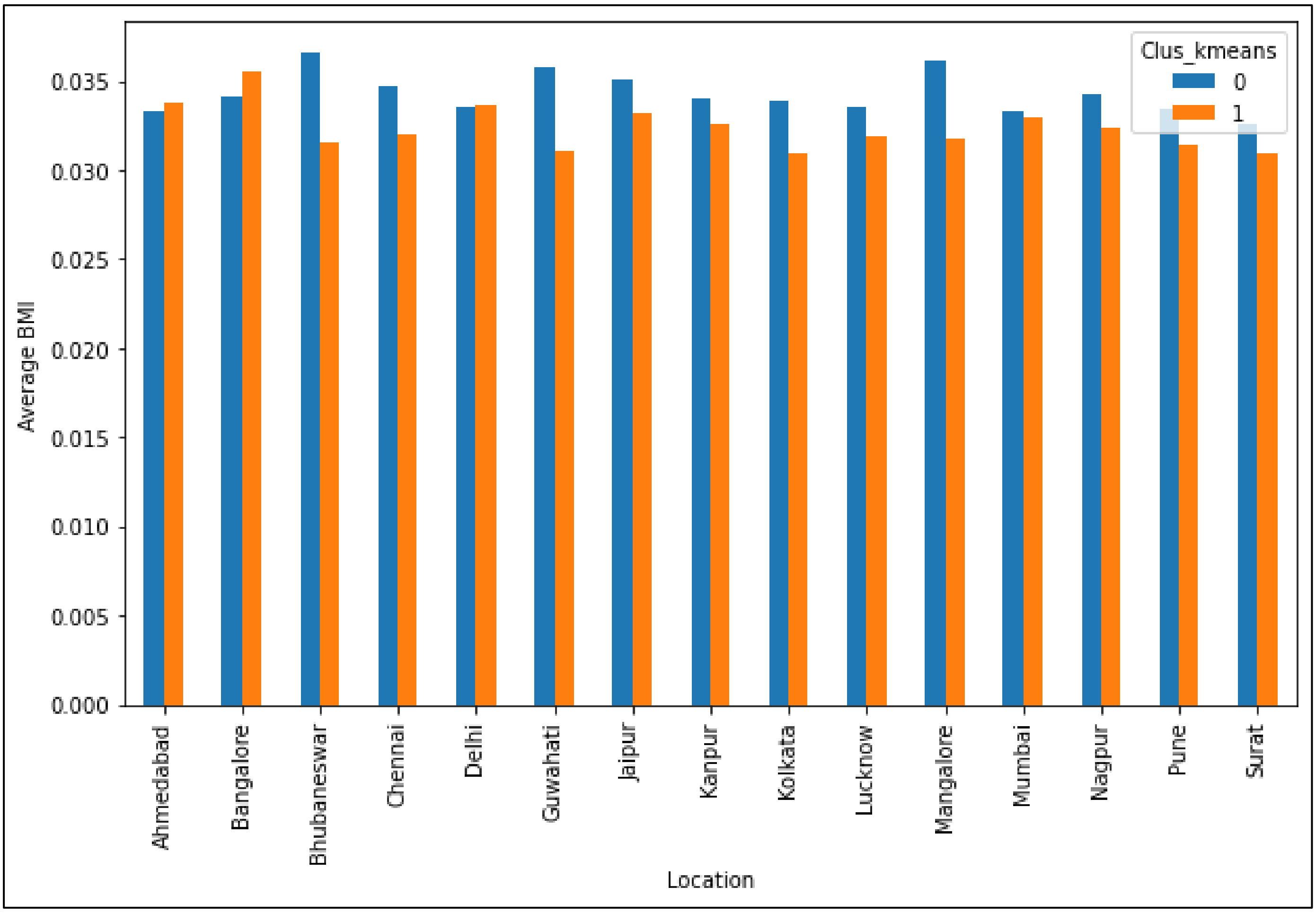
Additionally, group ‘0’ or high payers weight more than low insurance payers of group ‘1’.

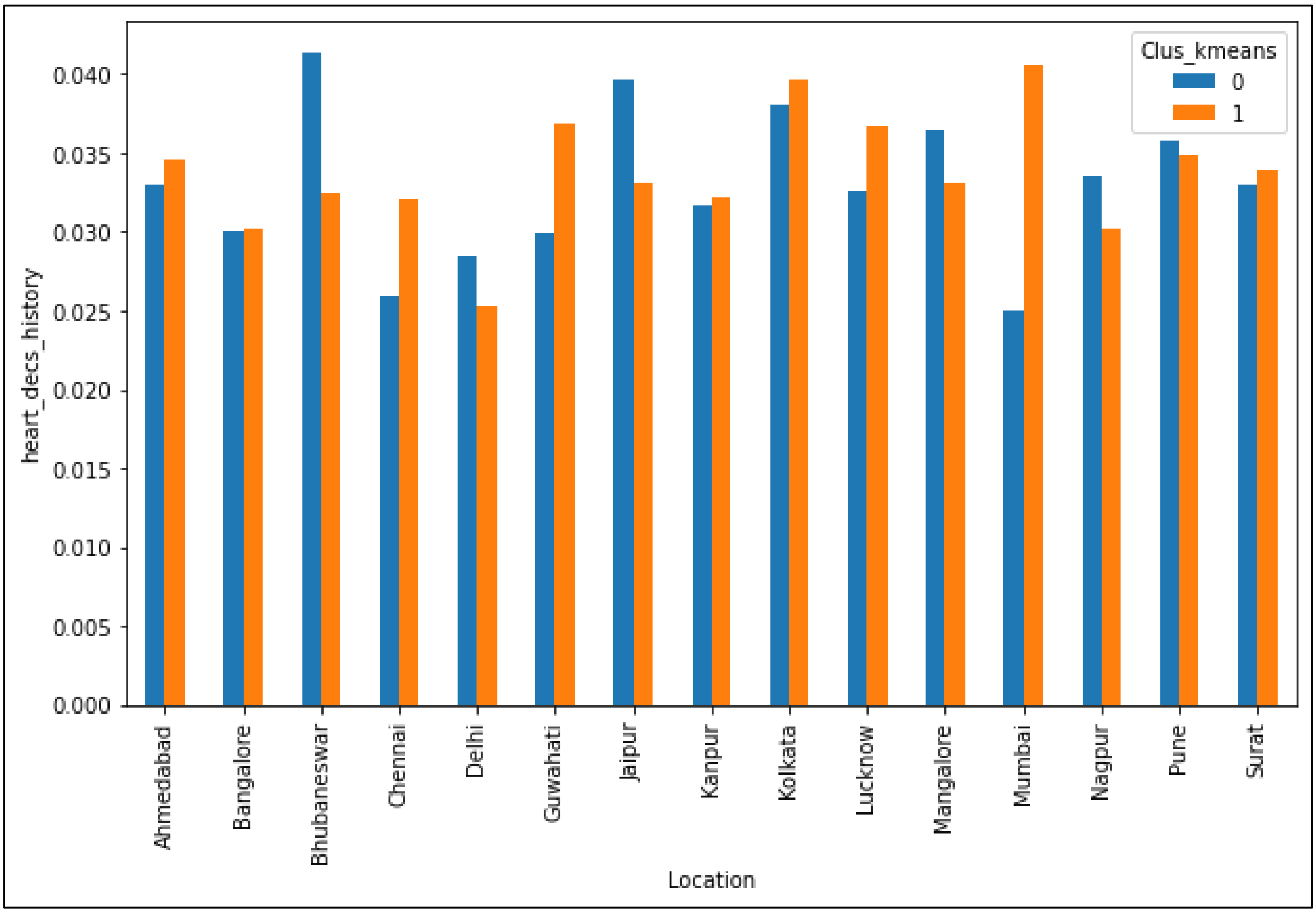


Plot 14: lmplot showing two clusters against weight



Plot 15: Violin plot showing distribution of insurance cost according to Occupation





Plot 16: Barplot showing average bmi and heart disease history for different cities

Plot 15 shows that group ‘0’ people or high insurance cost bearers, more frequently pay around 30,000 to 40,000 and group ‘1’ people pay most around 10,000 or 25,000.
The BMI of high insurance cost bearers (group 0) is high in all the cities except Ahmedabad, Bangalore and Delhi. In these cities, the low cost bearers have more average ‘bmi’.

The heart disease history of high insurance cost bearer is found more in Bhubaneshwar, Delhi, Jaipur, Mangalore, Nagpur and Pune.

4.3 Any other business insights

Given the EDA and cluster analysis, it has been identified that there are two group of people:

- Group 0- High insurance cost payers
- Group 1- Low insurance cost payers

Most distinguishing fact between these group is that people whose weight is more bear more insurance cost.

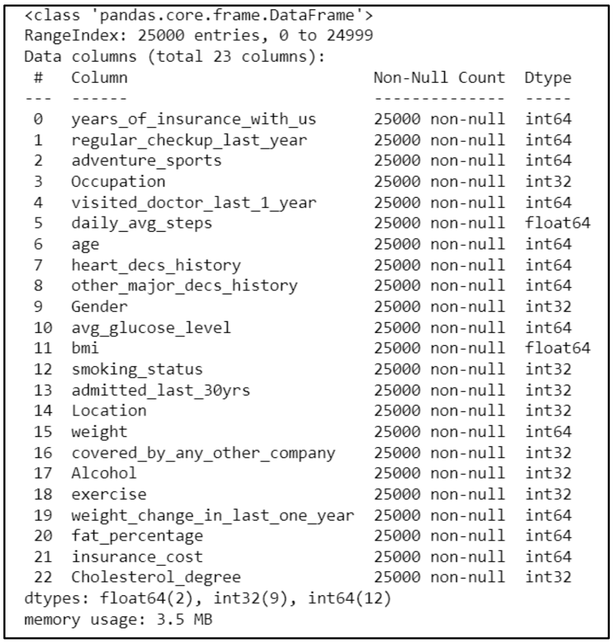
Another thing to notice is that the insurance cost is similar if someone have or does not have heart disease history or if they exercise, more or less. The average insurance cost is also similar if they drink alcohol or not or if they are in different occupation. This approach of might not seem efficient and some customer might be overpaying or underpaying for their insurance. This needs to be worked upon and more segments of insurance should be made according to different health and lifestyle profile of the customer. This is because the risk of health disease varies with diffrent health and lifestyle profile, and different prices should be labelled for customers with different health risk profile.

5. Model building and interpretation.

5.1. Build various models (You can choose to build models for either or all of descriptive, predictive or prescriptive purposes). Test your predictive model against the test set using various appropriate performance metrics.

After pre-processing the data for missing values, outliers and univariate and bivariate analysis, the data was prepared for predictive modelling. Before modelling, ‘cholesterol level’ was dropped since it will now be represented by ‘cholesterol degree’ and ‘Clus_kmeans’ and ‘sil_width’ was dropped since they won’t be useful in predictive analysis.

After that the categorical variables were encoded and converted into integer type. This is because, modelling is usually done in integer type variables.



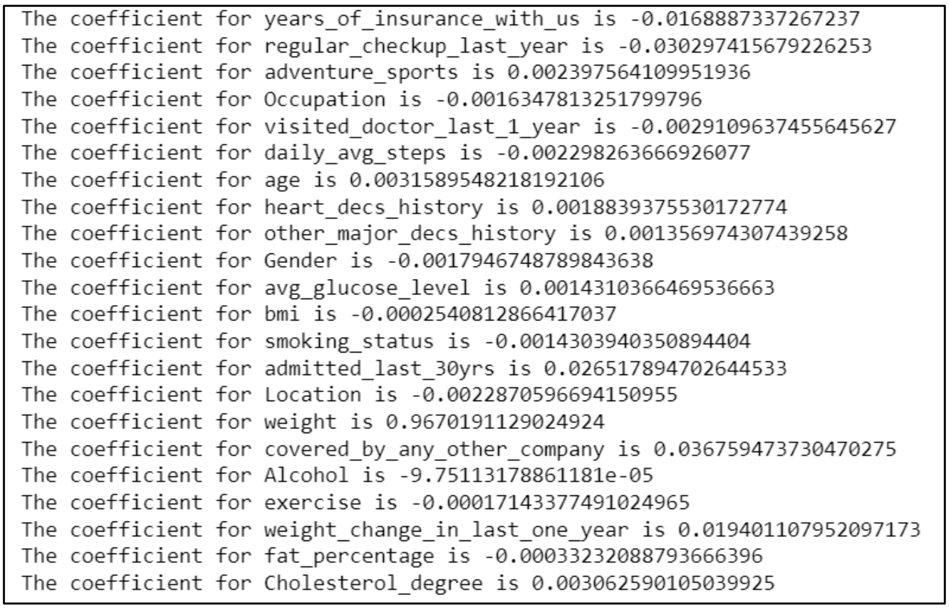
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 25000 entries, 0 to 24999
Data columns (total 23 columns):
 #   Column                          Non-Null Count  Dtype  
---  ------                          --------------  -----  
 0   years_of_insurance_with_us      25000 non-null  int64  
 1   regular_checkup_last_year       25000 non-null  int64  
 2   adventure_sports                25000 non-null  int64  
 3   Occupation                      25000 non-null  int32  
 4   visited_doctor_last_1_year      25000 non-null  int64  
 5   daily_avg_steps                 25000 non-null  float64
 6   age                             25000 non-null  int64  
 7   heart_decs_history              25000 non-null  int64  
 8   other_major_decs_history        25000 non-null  int64  
 9   Gender                          25000 non-null  int32  
 10  avg_glucose_level               25000 non-null  int64  
 11  bmi                             25000 non-null  float64
 12  smoking_status                  25000 non-null  int32  
 13  admitted_last_30yrs             25000 non-null  int32  
 14  Location                        25000 non-null  int32  
 15  weight                          25000 non-null  int64  
 16  covered_by_any_other_company    25000 non-null  int32  
 17  Alcohol                         25000 non-null  int32  
 18  exercise                        25000 non-null  int32  
 19  weight_change_in_last_one_year  25000 non-null  int64  
 20  fat_percentage                  25000 non-null  int64  
 21  insurance_cost                  25000 non-null  int64  
 22  Cholesterol_degree              25000 non-null  int32  
dtypes: float64(2), int32(9), int64(12)
memory usage: 3.5 MB
```

Table 5: Converted Dtypes

5.1.1 Linear Regression

Linear Regressor was imported from sklearn and then the data set was divided into X (consisting of independent variables) and y (consisting of dependent variable 'insurance cost') datasets. These two datasets were further split into train and test data, with 70% data for training with random state of 1. These datasets were then scales by applying zscore, since linear regression is affected by different units and standardizing the units would help in achieving better results.

After that, the coefficients of all the variables were calculated, which came out as follows:



```
The coefficient for years_of_insurance_with_us is -0.0168887337267237
The coefficient for regular_checkup_last_year is -0.030297415679226253
The coefficient for adventure_sports is 0.002397564109951936
The coefficient for Occupation is -0.0016347813251799796
The coefficient for visited_doctor_last_1_year is -0.0029109637455645627
The coefficient for daily_avg_steps is -0.002298263666926077
The coefficient for age is 0.0031589548218192106
The coefficient for heart_decs_history is 0.0018839375530172774
The coefficient for other_major_decs_history is 0.001356974307439258
The coefficient for Gender is -0.0017946748789843638
The coefficient for avg_glucose_level is 0.0014310366469536663
The coefficient for bmi is -0.0002540812866417037
The coefficient for smoking_status is -0.0014303940350894404
The coefficient for admitted_last_30yrs is 0.026517894702644533
The coefficient for Location is -0.0022870596694150955
The coefficient for weight is 0.9670191129024924
The coefficient for covered_by_any_other_company is 0.036759473730470275
The coefficient for Alcohol is -9.75113178861181e-05
The coefficient for exercise is -0.00017143377491024965
The coefficient for weight_change_in_last_one_year is 0.019401107952097173
The coefficient for fat_percentage is -0.0003323208879366396
The coefficient for Cholesterol_degree is 0.003062590105039925
```

Table 6: Coefficient of all the variables

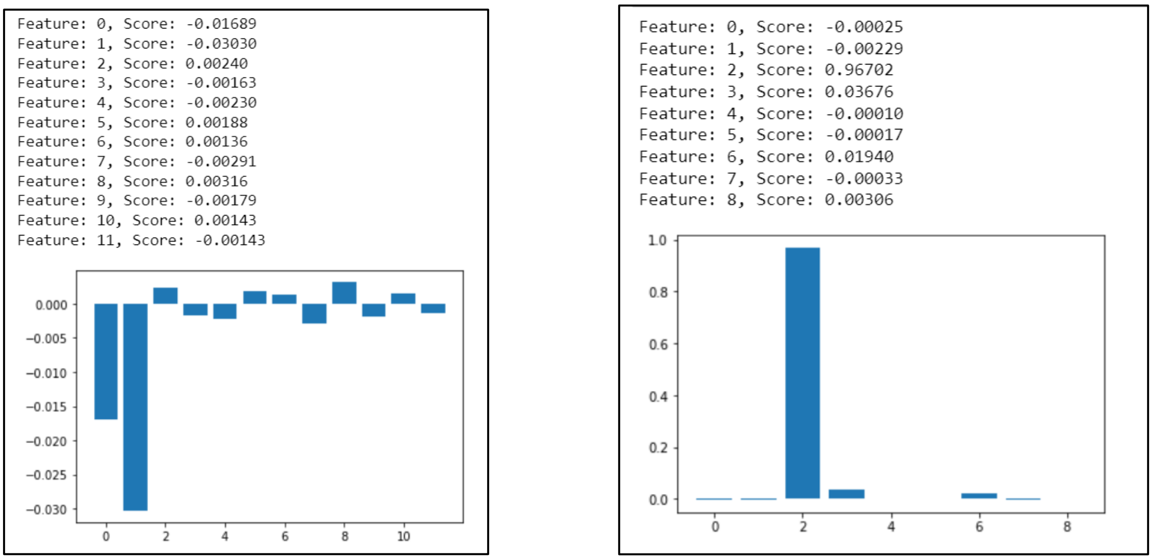
The intercept for the model is 5.984 e-16. The result of linear regression is as follows:

Data set	**R Squared**	**RMSE**
Train	0.945	0.2345
Test	0.945	0.2341

Table 7: Test result for simple Linear regression

5.1.2 OLS Linear Regression

OLS linear regression was also modelled as it gives more indepth analysis by letting us know significance of each predictor variable on predicted variable. For OLS linear regression, firstly, all the variables were used anf their feature importance was calculated:

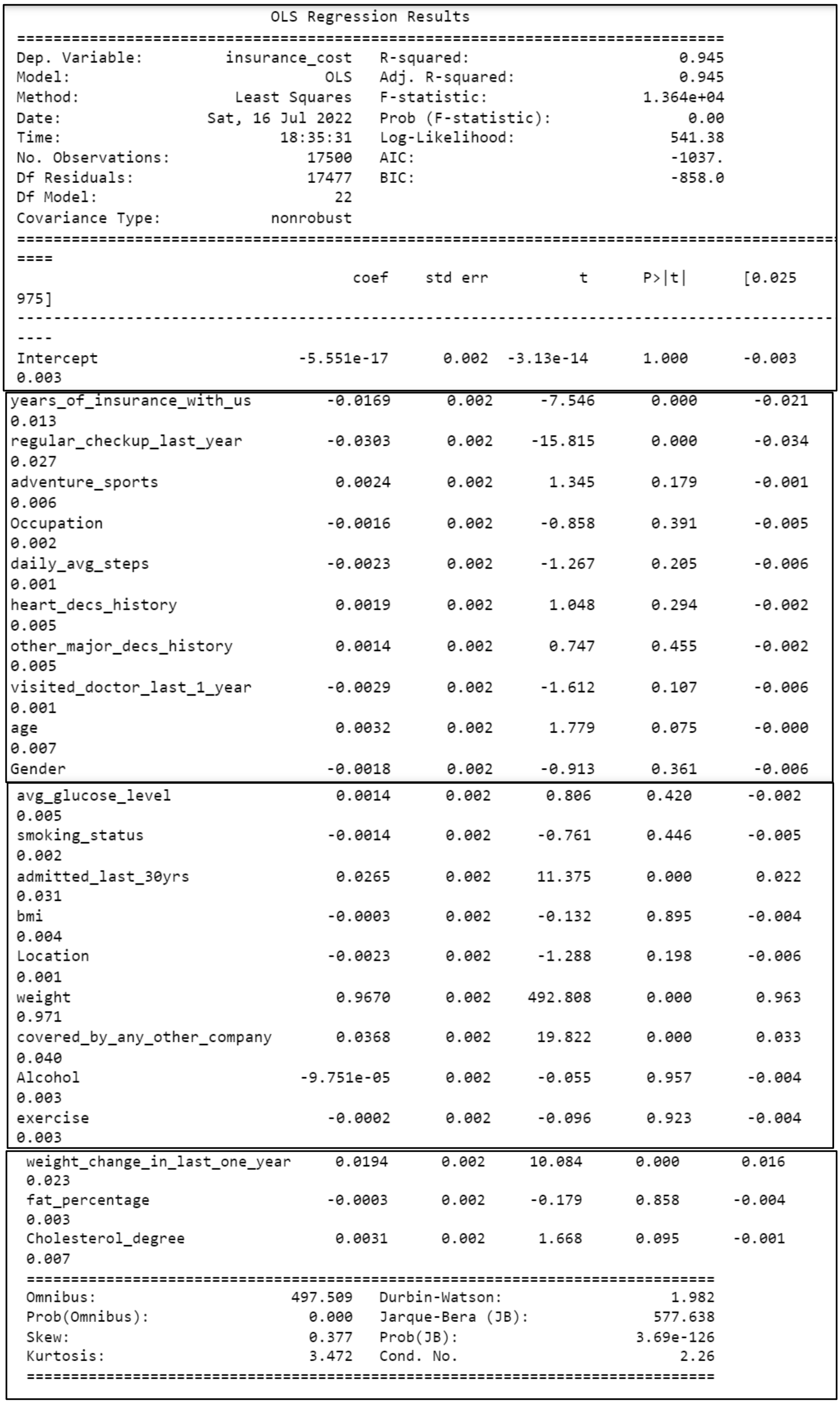


```
Feature: 0, Score: -0.01689
Feature: 1, Score: -0.03030
Feature: 2, Score: 0.00240
Feature: 3, Score: -0.00163
Feature: 4, Score: -0.00230
Feature: 5, Score: 0.00188
Feature: 6, Score: 0.00136
Feature: 7, Score: -0.00291
Feature: 8, Score: 0.00316
Feature: 9, Score: -0.00179
Feature: 10, Score: 0.00143
Feature: 11, Score: -0.00143
```

```
Feature: 0, Score: -0.00025
Feature: 1, Score: -0.00229
Feature: 2, Score: 0.96702
Feature: 3, Score: 0.03676
Feature: 4, Score: -0.00010
Feature: 5, Score: -0.00017
Feature: 6, Score: 0.01940
Feature: 7, Score: -0.00033
Feature: 8, Score: 0.00306
```

Table 8: Feature importance of all the variables

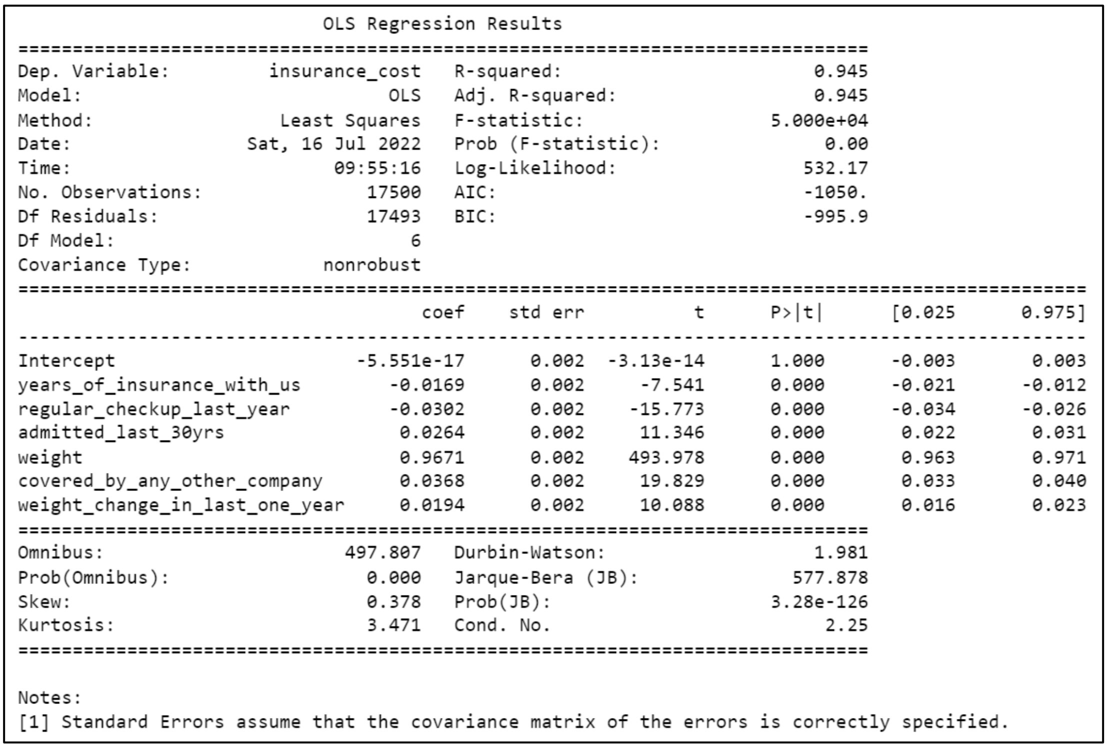
From here it seems that, adventure sports, heart disease history, other major disease, age, average glucose level, admitted in last 30 years, weight, covered by other company, weight change and cholesterol degree seem to have positive impact on insurance cost. Let’s look at the model summary:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 1.364e+04
Date:                Sat, 16 Jul 2022   Prob (F-statistic):               0.00
Time:                        18:35:31   Log-Likelihood:                 541.38
No. Observations:               17500   AIC:                            -1037.
Df Residuals:                   17477   BIC:                            -858.0
Df Model:                          22
Covariance Type:            nonrobust
==================================================================================
====
                                 coef    std err          t      P>|t|      [0.025
975]
----------------------------------------------------------------------------------
----
Intercept                   -5.551e-17      0.002  -3.13e-14      1.000      -0.003
0.003
years_of_insurance_with_us     -0.0169      0.002     -7.546      0.000      -0.021
0.013
regular_checkup_last_year      -0.0303      0.002    -15.815      0.000      -0.034
0.027
adventure_sports                0.0024      0.002      1.345      0.179      -0.001
0.006
Occupation                     -0.0016      0.002     -0.858      0.391      -0.005
0.002
daily_avg_steps                -0.0023      0.002     -1.267      0.205      -0.006
0.001
heart_decs_history              0.0019      0.002      1.048      0.294      -0.002
0.005
other_major_decs_history        0.0014      0.002      0.747      0.455      -0.002
0.005
visited_doctor_last_1_year     -0.0029      0.002     -1.612      0.107      -0.006
0.001
age                             0.0032      0.002      1.779      0.075      -0.000
0.007
Gender                         -0.0018      0.002     -0.913      0.361      -0.006
avg_glucose_level               0.0014      0.002      0.806      0.420      -0.002
0.005
smoking_status                 -0.0014      0.002     -0.761      0.446      -0.005
0.002
admitted_last_30yrs             0.0265      0.002     11.375      0.000       0.022
0.031
bmi                            -0.0003      0.002     -0.132      0.895      -0.004
0.004
Location                       -0.0023      0.002     -1.288      0.198      -0.006
0.001
weight                          0.9670      0.002    492.808      0.000       0.963
0.971
covered_by_any_other_company    0.0368      0.002     19.822      0.000       0.033
0.040
Alcohol                     -9.751e-05      0.002     -0.055      0.957      -0.004
0.003
exercise                       -0.0002      0.002     -0.096      0.923      -0.004
0.003
weight_change_in_last_one_year  0.0194      0.002     10.084      0.000       0.016
0.023
fat_percentage                 -0.0003      0.002     -0.179      0.858      -0.004
0.003
Cholesterol_degree              0.0031      0.002      1.668      0.095      -0.001
0.007
==============================================================================
Omnibus:                      497.509   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              577.638
Skew:                           0.377   Prob(JB):                    3.69e-126
Kurtosis:                       3.472   Cond. No.                         2.26
==============================================================================
```

Table 9: OLS model summary 1

From here, it can be seen that, only, years_of_insurance_with_us + regular_checkup_last_year + admitted_last_30yrs + weight + covered_by_any_other_company + weight_change_in_last_one_year, seem to be significant enough to impact insurance cost. If the OLS modelling is done again taking these variables, the model summary comes as follows:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         insurance_cost   R-squared:                       0.945
Model:                            OLS   Adj. R-squared:                  0.945
Method:                 Least Squares   F-statistic:                 5.000e+04
Date:                Sat, 16 Jul 2022   Prob (F-statistic):               0.00
Time:                        09:55:16   Log-Likelihood:                 532.17
No. Observations:               17500   AIC:                            -1050.
Df Residuals:                   17493   BIC:                            -995.9
Df Model:                           6
Covariance Type:            nonrobust
==================================================================================================
                                    coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
Intercept                     -5.551e-17      0.002  -3.13e-14      1.000      -0.003       0.003
years_of_insurance_with_us       -0.0169      0.002     -7.541      0.000      -0.021      -0.012
regular_checkup_last_year        -0.0302      0.002    -15.773      0.000      -0.034      -0.026
admitted_last_30yrs               0.0264      0.002     11.346      0.000       0.022       0.031
weight                            0.9671      0.002    493.978      0.000       0.963       0.971
covered_by_any_other_company      0.0368      0.002     19.829      0.000       0.033       0.040
weight_change_in_last_one_year    0.0194      0.002     10.088      0.000       0.016       0.023
==============================================================================
Omnibus:                      497.807   Durbin-Watson:                   1.981
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              577.878
Skew:                           0.378   Prob(JB):                    3.28e-126
Kurtosis:                       3.471   Cond. No.                         2.25
==============================================================================

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

Table 10: OLS model summary 2

Here, all the variables come out to be significant, with Adjusted R square of 0.945.

The MSE is 0.05, RMSE is 0.2341.

5.1.3 Random Forest

Random forest regressor is imported from sklearn. Grid search was used to find optimal hyperparameters, which were taken as follows:

'max_depth': [5,8,10,15,20,30],

'max_features': [5,6,9],

'n_estimators': [10, 25, 50,100],

'min_samples_leaf': [50,100,250,500], # 1 to 3% of sample

'min_samples_split': [150, 300, 900] # 3 times of min sample leaf

Random state= 1

CV = 4

On fitting the grid search on train dataset, the best parameters came out as follows:

'max_depth': 15,

'max_features': 9,

'min_samples_leaf': 50,

'min_samples_split': 150,

'n_estimators': 100

On fitting this to train and test data the values were predicted which gave us the MSE of 0.055 and RMSE of 0.235 and MAE of 0.187.

5.1.4 KNN

KNN regressor is imported from sklearn. Grid search was used to find optimal hyperparameters, which were taken as follows:

'n_neighbors': [3, 5, 7, 9, 12,15,17],

'weights': ['uniform', 'distance'],

'metric': ['euclidean', 'manhattan']

CV= 5

N_jobs=-1

The best paramaeters came out as follows:

{'metric': 'euclidean', 'n_neighbors': 12, 'weights': 'distance'}

Using these parameters, the MSE of the model is 1.92, RMSE= 1.38 and MAE =1.168

6. Model Tuning

6.1. Ensemble modelling, wherever applicable. Any other model tuning measures (if applicable)

Two ensemble techniques were used to predict the insurance cost, Adaboost and gradient boost.

6.1.1 AdaBoost

AdaBoost regressor was imported from sklearn library and grid search CV was used to find the optimum hyper parameter. Decision tree regressor was used as the base estimator for Adaboost. The hyperparameters were taken as follows:

base_estimator":[DecisionTreeRegressor(max_depth=1),DecisionTreeRegressor(max_depth=2),DecisionTreeRegressor(max_depth=3)],

"n_estimators": np.arange(10,110,10),

"learning_rate":np.arange(0.1,2,0.1)

CV=3

The MSE of the model came out to be 0.052, RMSE= 0.228 and MAE of 0.185

6.1.2 Gradient Boost

Gradient boost regressor was imported from Sklearn and was fitted into the train dataset with n_estimator =5 and random state 1. The MSE of the model came out to be 0.045, RMSE = 0.213, MAE= 0.172.

6.2. Interpretation of the most optimum model and its implication on the business.

Comparing the model evaluation of all the models used:

Model	Train MSE	Test MSE	Train RMSE	Test RMSE	Train MAE	Test MAE
Linear regression				0.234		
Random Forest	0.05	0.055	0.223	0.235	0.179	0.187
KNN	1.919	1.92	1.38	1.38	1.17	1.168
AdaBoost	0.0508	0.052	0.225	0.228	0.183	0.185
Gradient Descent	0.043	0.045	0.209	0.213	0.169	0.172

Table 11: summary of model evaluation of all models

The models seem to be overfit in all cases except for random forest. The best RMSE value on test data is given by gradient descent method and its MAE is also the lowest from all the variables. Hence, it seems that Gradient can be chosen for predicting the results.

Given the features importance analysis from linear regression, it seems that years_of_insurance_with_us + regular_checkup_last_year + admitted_last_30yrs + weight + covered_by_any_other_company + weight_change_in_last_one_year should be considered while calculating the insurance cost.

Note should be taken that, more the year of insurance the customer has with us and more the regular check-up last year done by the customer, less is the insurance cost. However, more the weight of the customer more is the insurance cost.