Amrita Jena

Capstone Project-Week 1

PGP - Data Science and Business Analytics. PGPDSBA Online May_B 2021

INDEX

LIST OF TABLES -

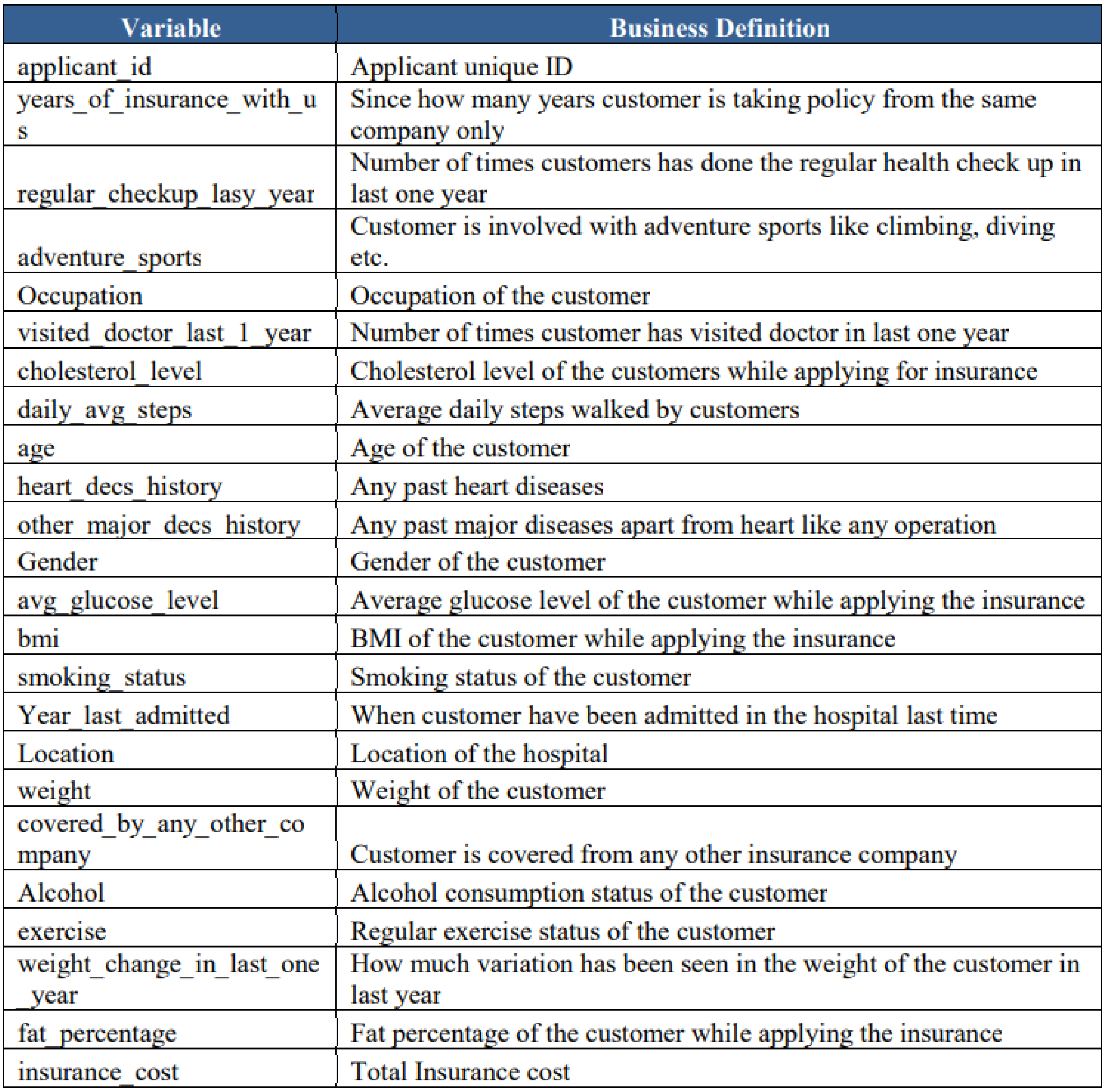
. LIST OF FIGURES FOR PROBLEM 1

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.

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

DataDictionary:



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

1.1. Problem Understanding-

The saying “Our body is our temple” is a statement of careful consideration. How a temple is kept clean, worshiped and is kept closed to all negative entities we should also treat our bodies same way.

a) Defining problem statement –

With disease burden spiking up, medical expenses are also increasing day by day, it’s important for us to be conscious about preventive healthcare, which includes keeping ourselves fit. Not taking any preventive measure could open path for disease like obesity, diabetes, and high/low Blood pressure making us high prone to critical health illnesses.

Not keeping fit don’t just play havoc with our health, they can severely increase our health insurance premium by several thousand

Need of the study/project –

This project will help us understand how health if not taken proper care of can make us pay heavy price for it and how it is important to prioritize our health and take all kind of preventive healthcare measures in our day to day life .

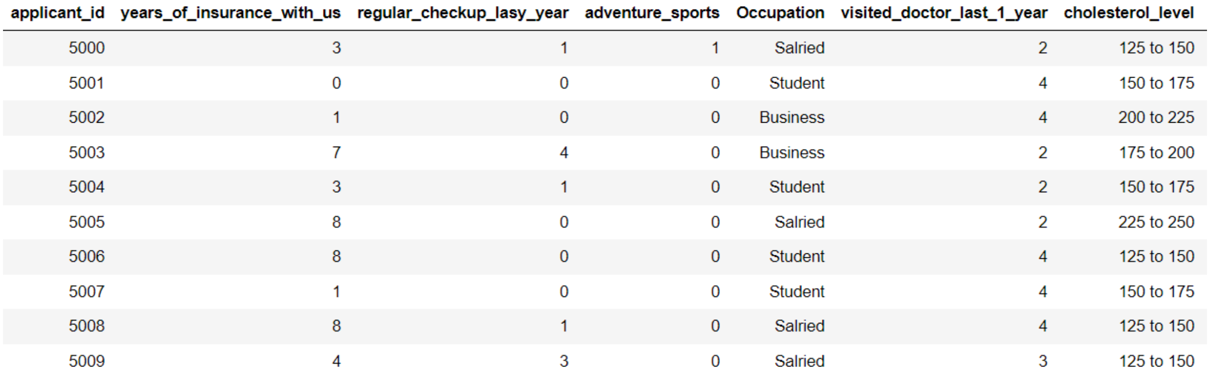
By leading a healthy lifestyle health insurance will no longer be viewed as an important measure to secure oneself against unforeseen illnesses; rather it will become a part of one’s daily health needs.

Unhealthy lifestyle like smoking,drinking,drugs,minimum sleep and junk eating adds feather to critical disease as well as insurance cost. Insurance companies may pay special attention to your lifestyle and profession. All information shared plays a key role in determining your suitability for the coverage and insurance costs.

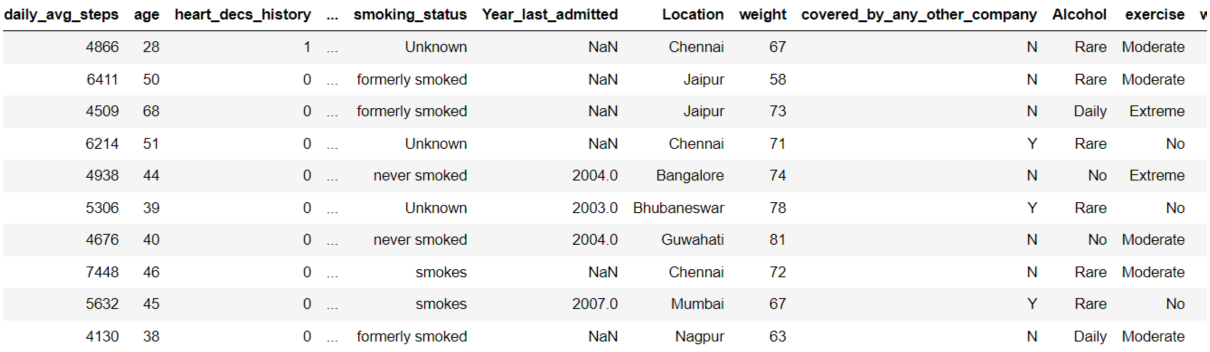
b) Understanding business/social opportunity..

This project will give us chance to understands benefits of leading healthy lifestyle to prevent us from critical diseases by reducing our insurance cost.

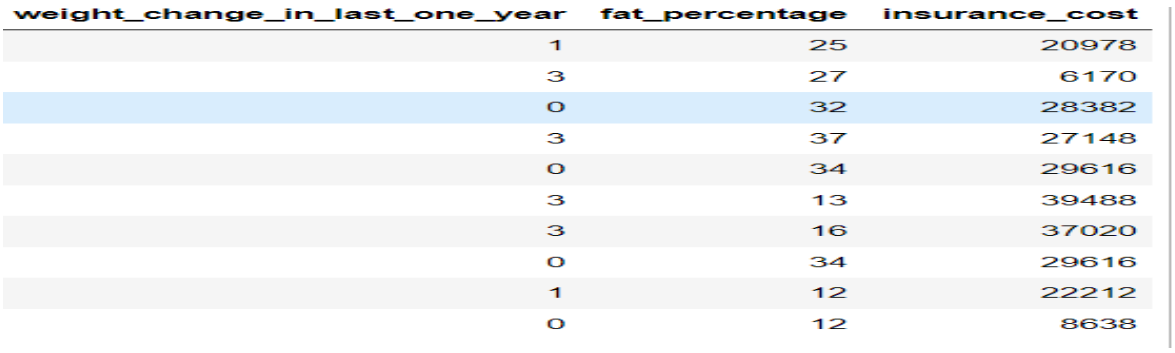
.2. Data Report-a) Understanding how data was collected in terms of time, frequency and methodology b) Visual inspection of data (rows, columns, descriptive details) c) Understanding of attributes (variable info, renaming if required)



applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level
5000	3	1	1	Salried	2	125 to 150
5001	0	0	0	Student	4	150 to 175
5002	1	0	0	Business	4	200 to 225
5003	7	4	0	Business	2	175 to 200
5004	3	1	0	Student	2	150 to 175
5005	8	0	0	Salried	2	225 to 250
5006	8	0	0	Student	4	125 to 150
5007	1	0	0	Student	4	150 to 175
5008	8	1	0	Salried	4	125 to 150
5009	4	3	0	Salried	3	125 to 150



daily_avg_steps	age	heart_decs_history	...	smoking_status	Year_last_admitted	Location	weight	covered_by_any_other_company	Alcohol	exercise	v
4866	28	1	...	Unknown	NaN	Chennai	67	N	Rare	Moderate	
6411	50	0	...	formerly smoked	NaN	Jaipur	58	N	Rare	Moderate	
4509	68	0	...	formerly smoked	NaN	Jaipur	73	N	Daily	Extreme	
6214	51	0	...	Unknown	NaN	Chennai	71	Y	Rare	No	
4938	44	0	...	never smoked	2004.0	Bangalore	74	N	No	Extreme	
5306	39	0	...	Unknown	2003.0	Bhubaneswar	78	Y	Rare	No	
4676	40	0	...	never smoked	2004.0	Guwahati	81	N	No	Moderate	
7448	46	0	...	smokes	NaN	Chennai	72	N	Rare	Moderate	
5632	45	0	...	smokes	2007.0	Mumbai	67	Y	Rare	No	
4130	38	0	...	formerly smoked	NaN	Nagpur	63	N	Daily	Moderate	



weight_change_in_last_one_year	fat_percentage	insurance_cost
1	25	20978
3	27	6170
0	32	28382
3	37	27148
0	34	29616
3	13	39488
3	16	37020
0	34	29616
1	12	22212
0	12	8638

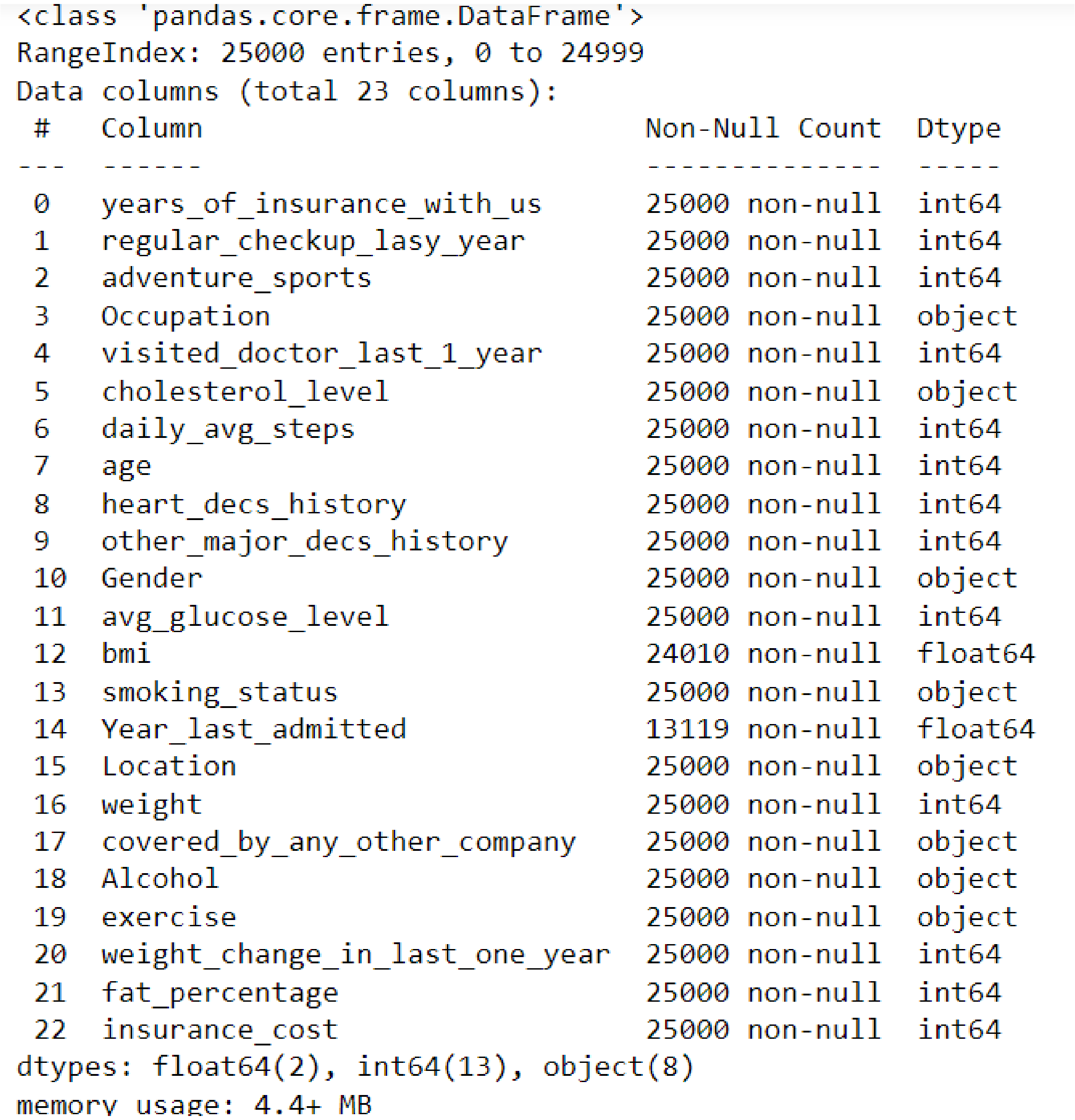
TABLE-1-DATA

We can see from the above data that there is an ‘applicant_id’ which is not of a great use ,therefore we can drop that column.

Checking the shape of the data: –

Previously we were having 25000 rows and 24 columns but after dropping applicant_id’ column now we have 25000 rows and 23 columns.

<u>Checking the info of the data</u>: –



```
<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_lasy_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  
dtypes: float64(2), int64(13), object(8)
memory usage: 4.4+ MB
```

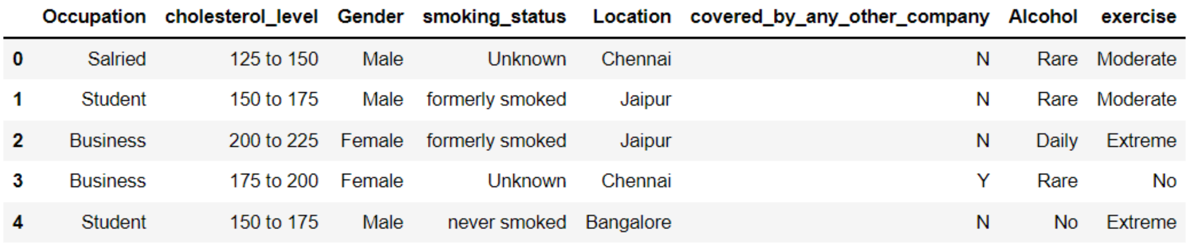
TABLE-2- INFO TABLE

There is total 8 object data types,2 float data type and 13 int data type.

‘insurance_cost’ is our target variable.

This table shows no of categoerical variables.

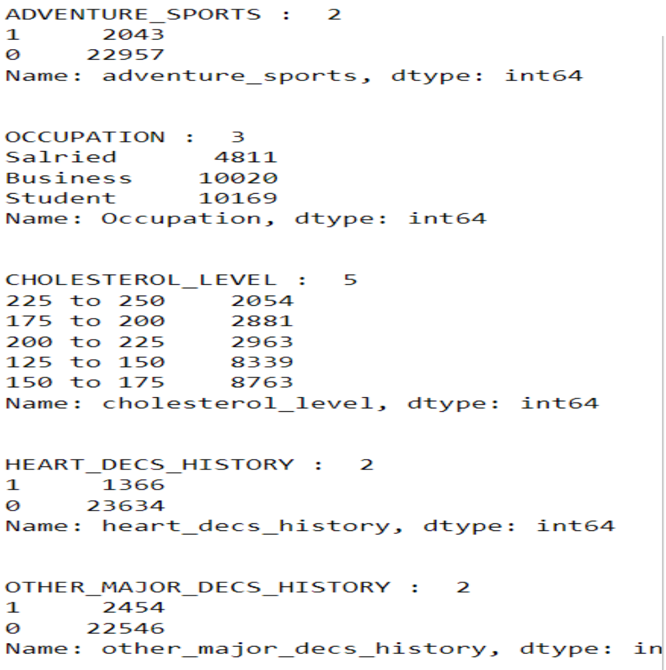
adventure_sports,other_major_decs_history and heart_decs_history are also categorical variables,hence we converted them into categorical.



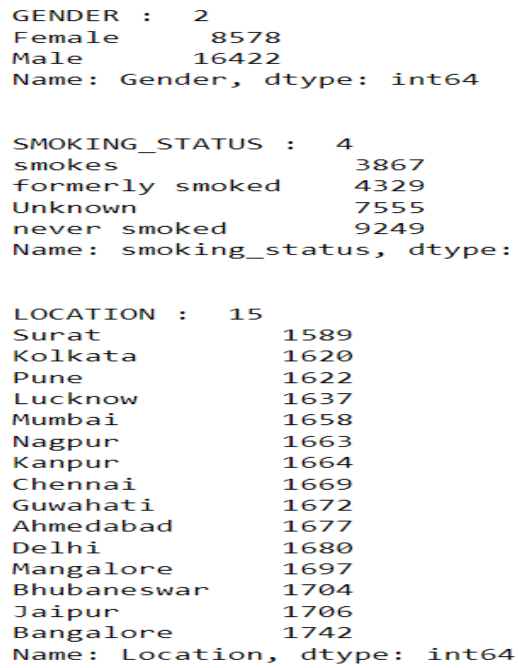
	Occupation	cholesterol_level	Gender	smoking_status	Location	covered_by_any_other_company	Alcohol	exercise
0	Salried	125 to 150	Male	Unknown	Chennai	N	Rare	Moderate
1	Student	150 to 175	Male	formerly smoked	Jaipur	N	Rare	Moderate
2	Business	200 to 225	Female	formerly smoked	Jaipur	N	Daily	Extreme
3	Business	175 to 200	Female	Unknown	Chennai	Y	Rare	No
4	Student	150 to 175	Male	never smoked	Bangalore	N	No	Extreme

TABLE-3- Categorical variable info table

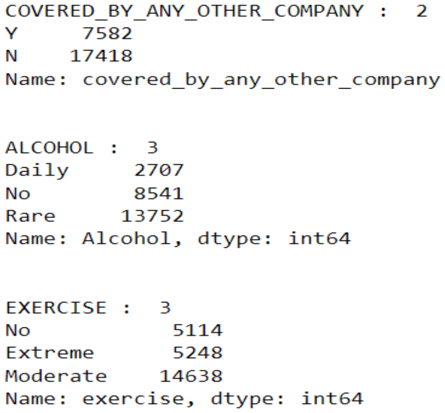
Unique counts of each categorical variables-



```
ADVENTURE_SPORTS :  2
1     2043
0    22957
Name: adventure_sports, dtype: int64


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


CHOLESTEROL_LEVEL :  5
225 to 250    2054
175 to 200    2881
200 to 225    2963
125 to 150    8339
150 to 175    8763
Name: cholesterol_level, dtype: int64


HEART_DECS_HISTORY :  2
1     1366
0    23634
Name: heart_decs_history, dtype: int64


OTHER_MAJOR_DECS_HISTORY :  2
1     2454
0    22546
Name: other_major_decs_history, dtype: in
```



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


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


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
```



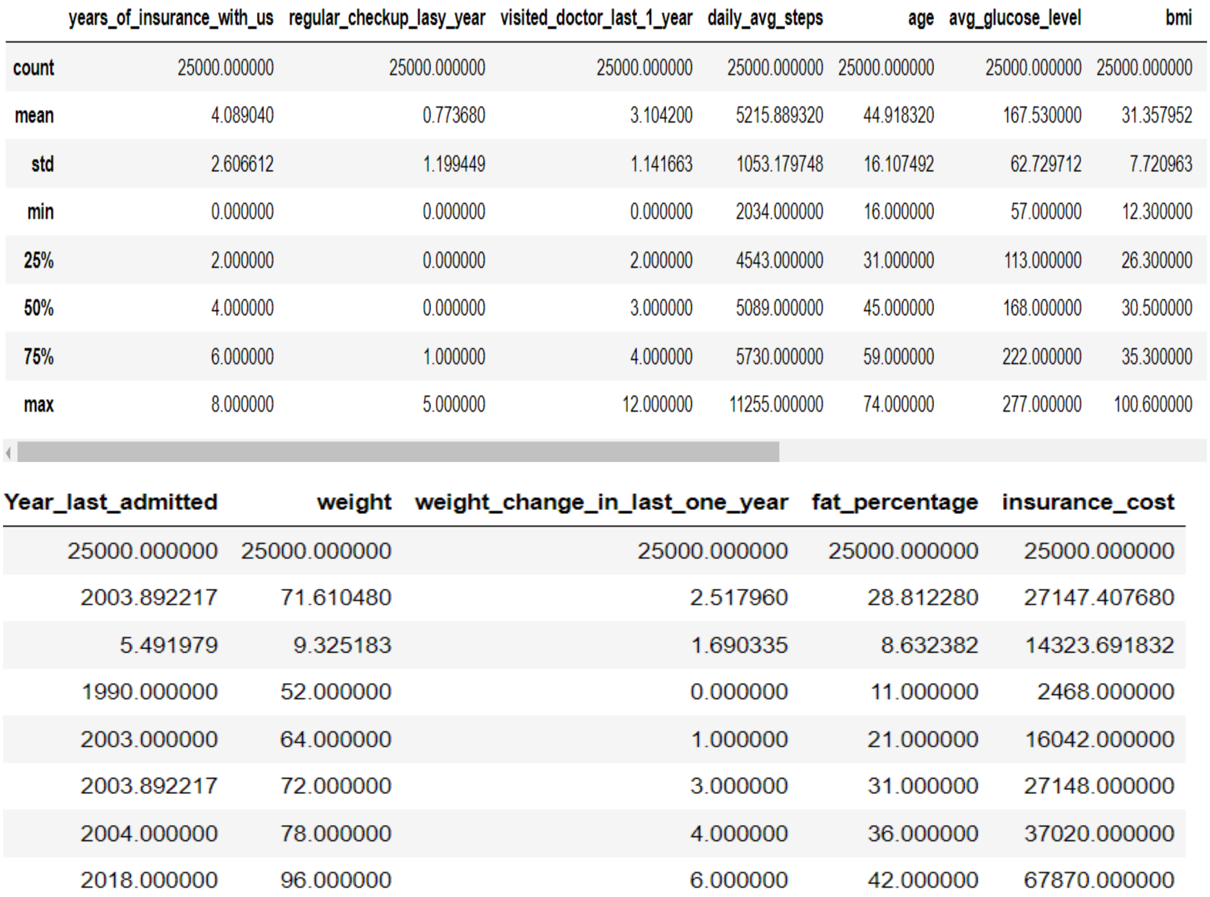
```
COVERED_BY_ANY_OTHER_COMPANY :  2
Y     7582
N    17418
Name: covered_by_any_other_company


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


EXERCISE :  3
No           5114
Extreme      5248
Moderate    14638
Name: exercise, dtype: int64
```

TABLE-4- Unique count table

Now we will check for duplicates:-

There are no duplicates in the dataset

Descriptive Statistics of the data set: -



	years_of_insurance_with_us	regular_checkup_lasy_year	visited_doctor_last_1_year	daily_avg_steps	age	avg_glucose_level	bmi
count	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000	25000.000000
mean	4.089040	0.773680	3.104200	5215.889320	44.918320	167.530000	31.357952
std	2.606612	1.199449	1.141663	1053.179748	16.107492	62.729712	7.720963
min	0.000000	0.000000	0.000000	2034.000000	16.000000	57.000000	12.300000
25%	2.000000	0.000000	2.000000	4543.000000	31.000000	113.000000	26.300000
50%	4.000000	0.000000	3.000000	5089.000000	45.000000	168.000000	30.500000
75%	6.000000	1.000000	4.000000	5730.000000	59.000000	222.000000	35.300000
max	8.000000	5.000000	12.000000	11255.000000	74.000000	277.000000	100.600000

Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost
25000.000000	25000.000000	25000.000000	25000.000000	25000.000000
2003.892217	71.610480	2.517960	28.812280	27147.407680
5.491979	9.325183	1.690335	8.632382	14323.691832
1990.000000	52.000000	0.000000	11.000000	2468.000000
2003.000000	64.000000	1.000000	21.000000	16042.000000
2003.892217	72.000000	3.000000	31.000000	27148.000000
2004.000000	78.000000	4.000000	36.000000	37020.000000
2018.000000	96.000000	6.000000	42.000000	67870.000000

TABLE-5- Descriptive Summary Table

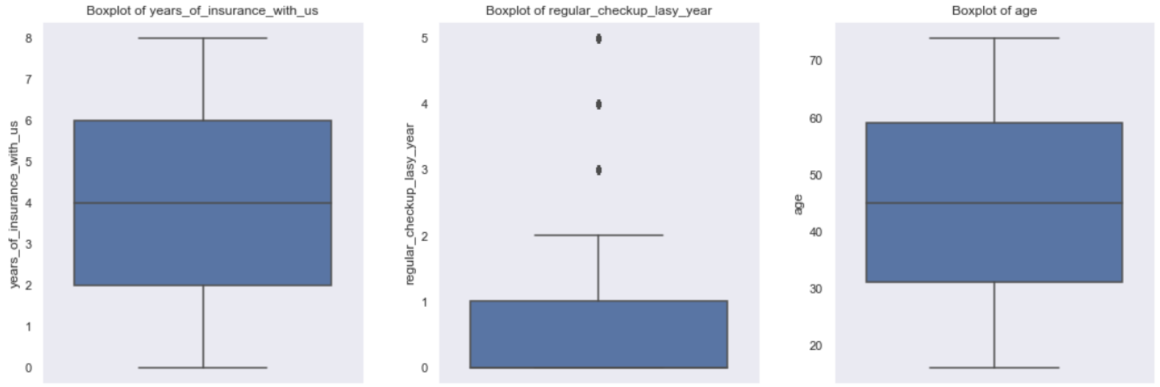
The mean age here is 44.4 with 16 as the minimum age and 74 as the maximum age. The mean BMI is 31.3 with 100 as maximum.

The mean glucose here 167.53 with 57 as minimum and 277 as maximum. The mean weight here is 71.61 with 52 kgs as min weight and 96 as maximum weight, maximum weight loss or weight gain an individual has experienced the previous year ids 6kgs.

The highest fat percentage is 42.00 and 28.81 as the mean fat percentage.

3. Exploratory Data Analysis

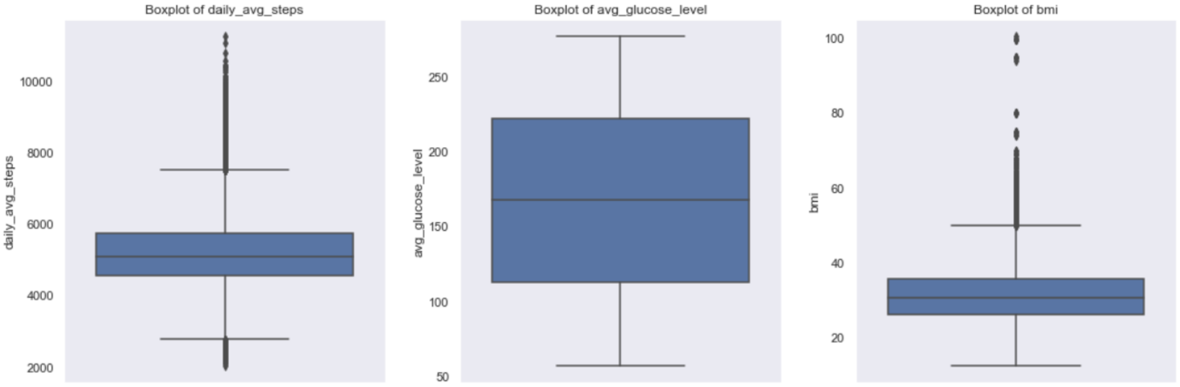
Univariate \ Bivariate Analysis:-



Boxplot of years_of_insurance_with_us

Boxplot of regular_checkup_lasy_year

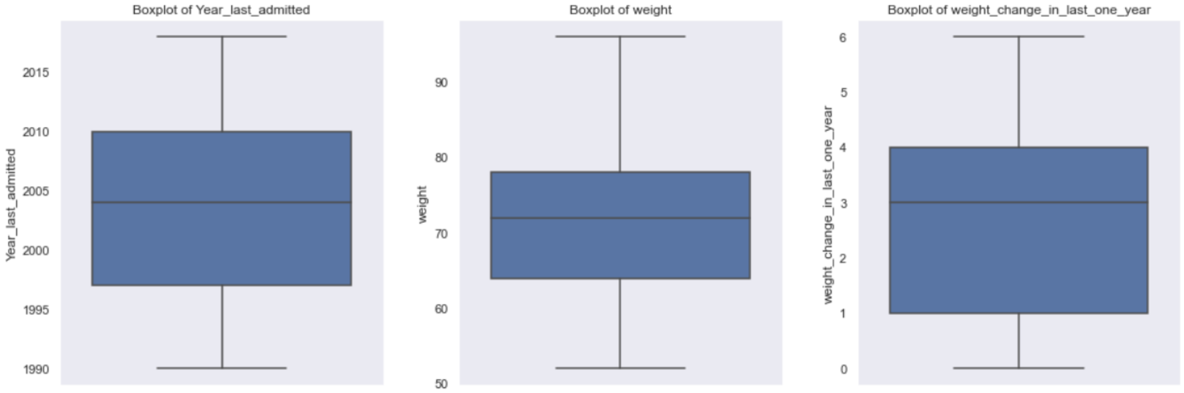
Boxplot of age



Boxplot of daily_avg_steps

Boxplot of avg_glucose_level

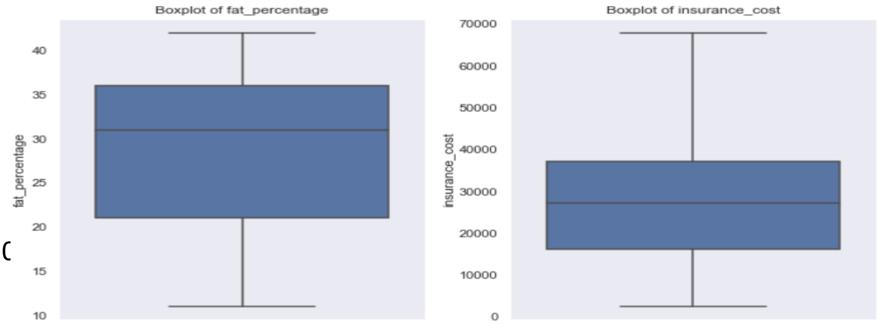
Boxplot of bmi



Boxplot of Year_last_admitted

Boxplot of weight

Boxplot of weight_change_in_last_one_year

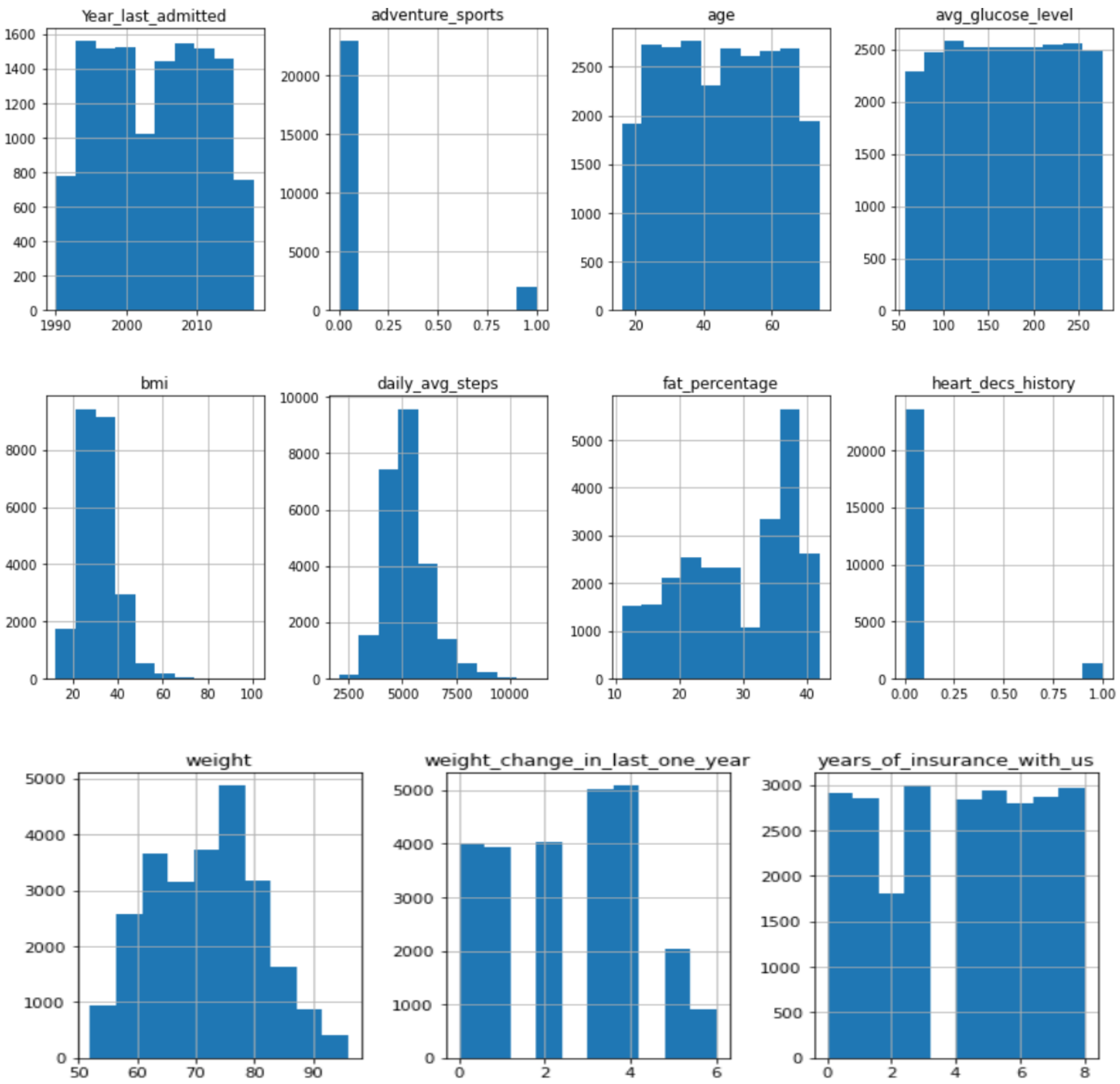


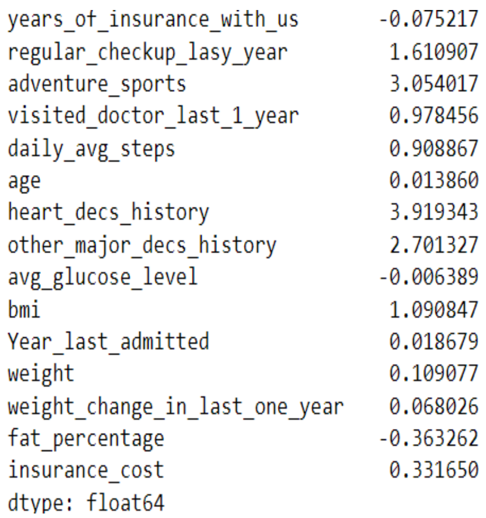
Boxplot of fat_percentage

Boxplot of insurance_cost

Fig-1-Outlier Boxplot

From the boxplots we could infer that regular_checkup_lasy_year, daily_avg_steps and bmi have outliers.





```
years_of_insurance_with_us      -0.075217
regular_checkup_lasy_year        1.610907
adventure_sports                 3.054017
visited_doctor_last_1_year       0.978456
daily_avg_steps                  0.908867
age                              0.013860
heart_decs_history               3.919343
other_major_decs_history         2.701327
avg_glucose_level               -0.006389
bmi                              1.090847
Year_last_admitted               0.018679
weight                           0.109077
weight_change_in_last_one_year   0.068026
fat_percentage                  -0.363262
insurance_cost                   0.331650
dtype: float64
```

Fig-2-Histogram

From above we could say that data is not highly skewed. we use skewness to understand which data set is normally distributed and which is not. If the skewness =0, It is said to be normally distributed, if it is >0 it is left skewed and if it<0 it is right skewed.

The outliers were removed after treating:-

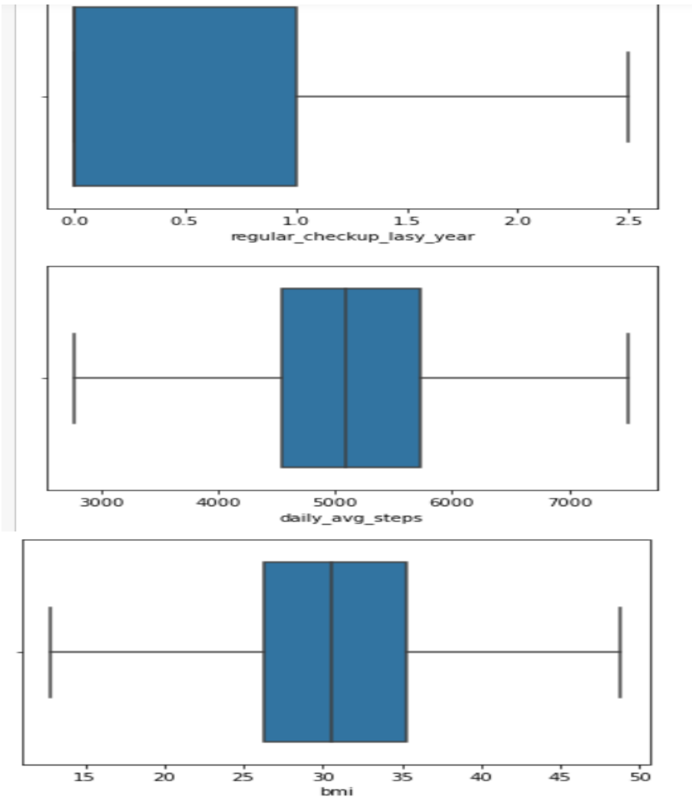
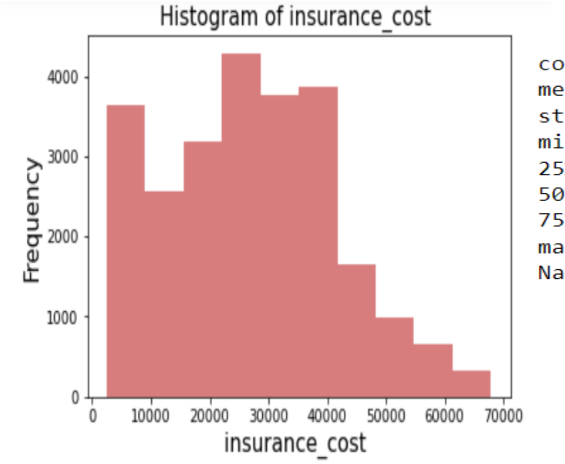
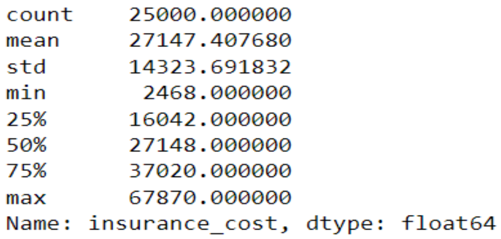


Fig-3-Outlier treated Boxplot

TARGET VARIABLE-



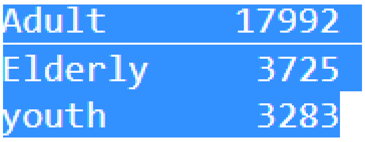


```
count     25000.000000
mean      27147.407680
std       14323.691832
min        2468.000000
25%       16042.000000
50%       27148.000000
75%       37020.000000
max       67870.000000
Name: insurance_cost, dtype: float64
```

Fig-4-Target Variable Histogram

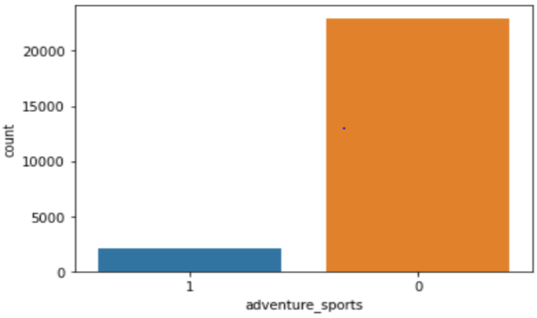
SKEWNESS=0.3316500625115993- The target variable is mostly left skewed with mean cost of 27147.40 rupees with 2468 minimum cost to 67870.00 as maximum cost.

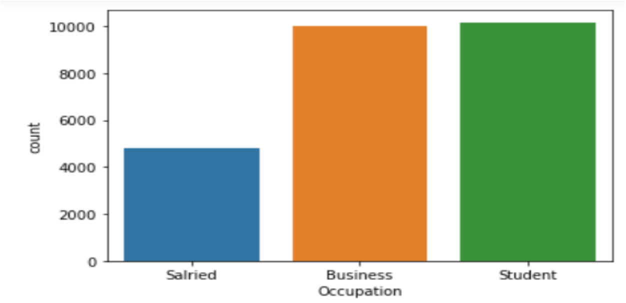
As age was having large number of variables, I grouped age into “age_group” for easy analysis –
Youth (15-24 years)
Adults (25-64 years)
Elderly (65 years and over)



```
Adult       17992
Elderly      3725
youth        3283
```

Categorical Variables-





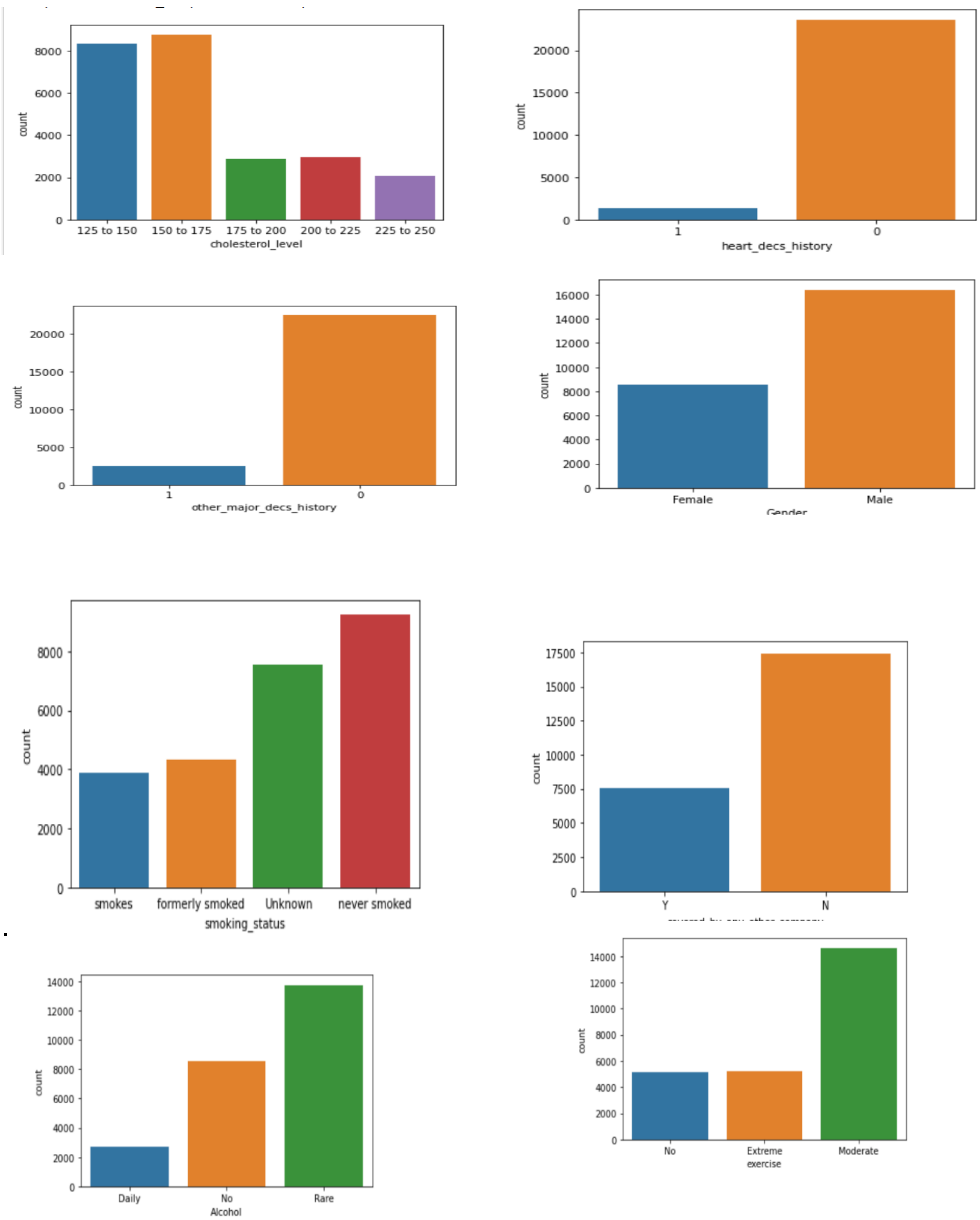


Fig-5-Barplot of Categorical Variable analysis

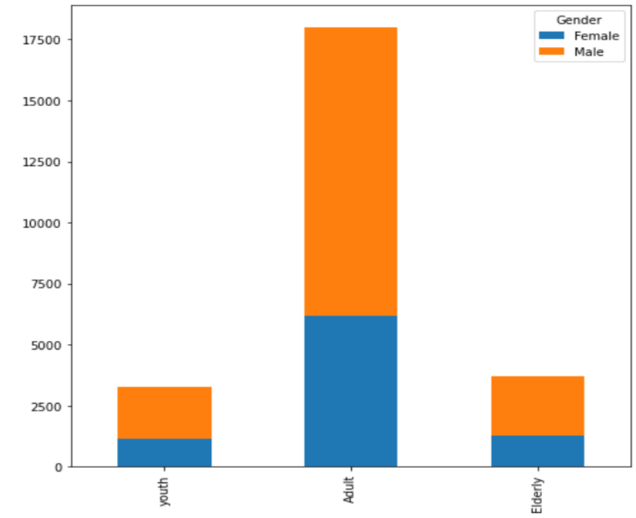
Adventurous sport is not much popular.

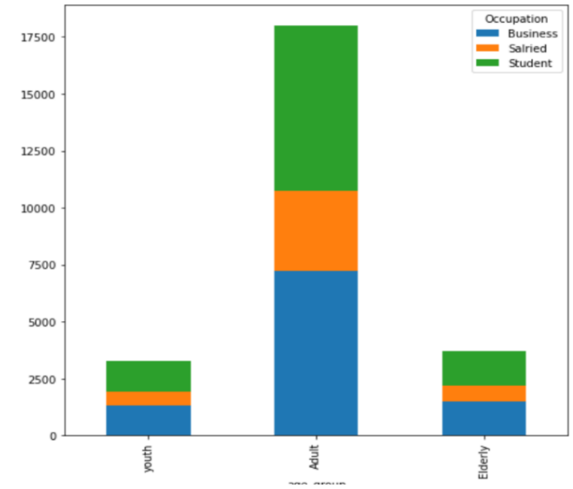
Business and Student occupation is more than salaries occupation.

Maximum insurance holders have normal cholesterol level.

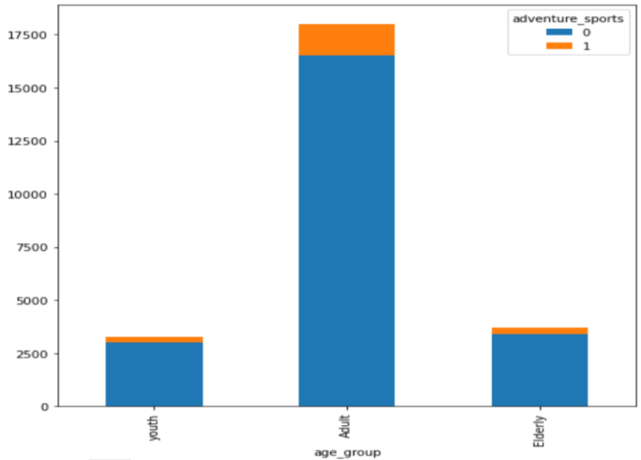
History of any other disease and heart disease is on lower side.

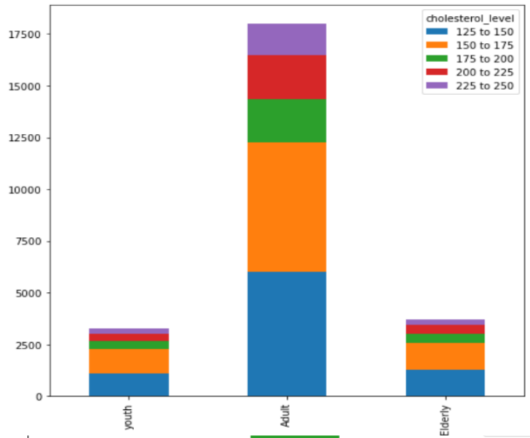
Male population are maximum insurance holders than females. Smokers and daily consumption of alcohol is very low, which is good. Maximum applicants follow a moderate exercise routine.



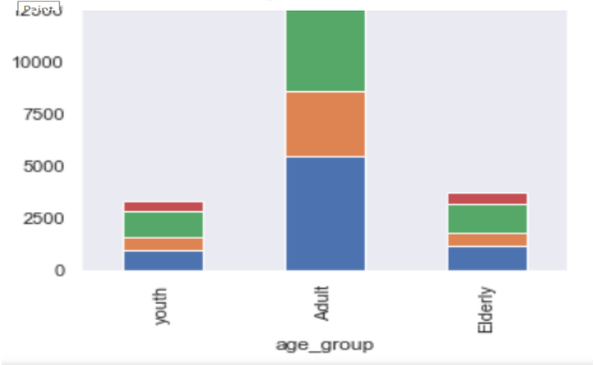


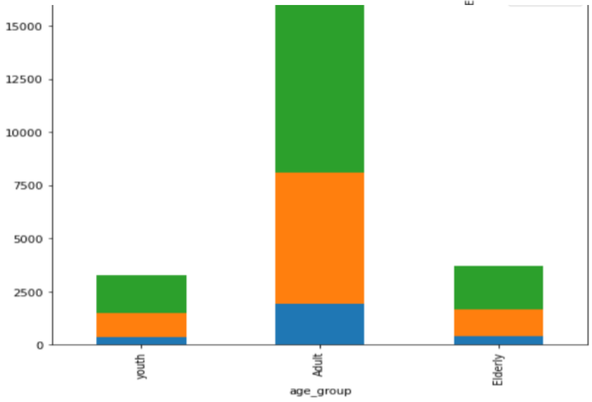
```
<matplotlib.axes._subplots.AxesSubplot at 0x1fc4e81e348>
```





17





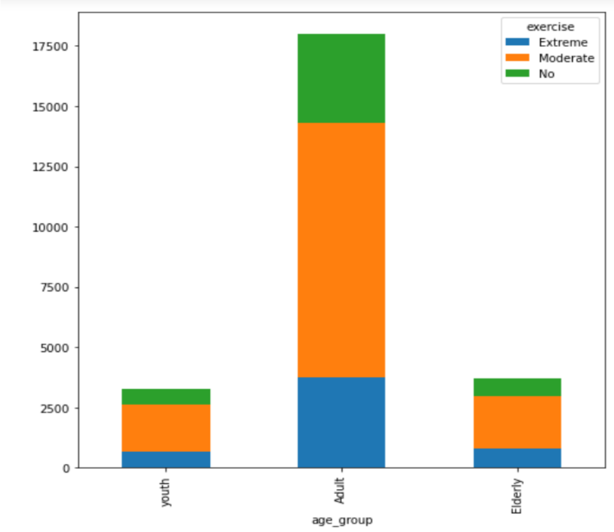
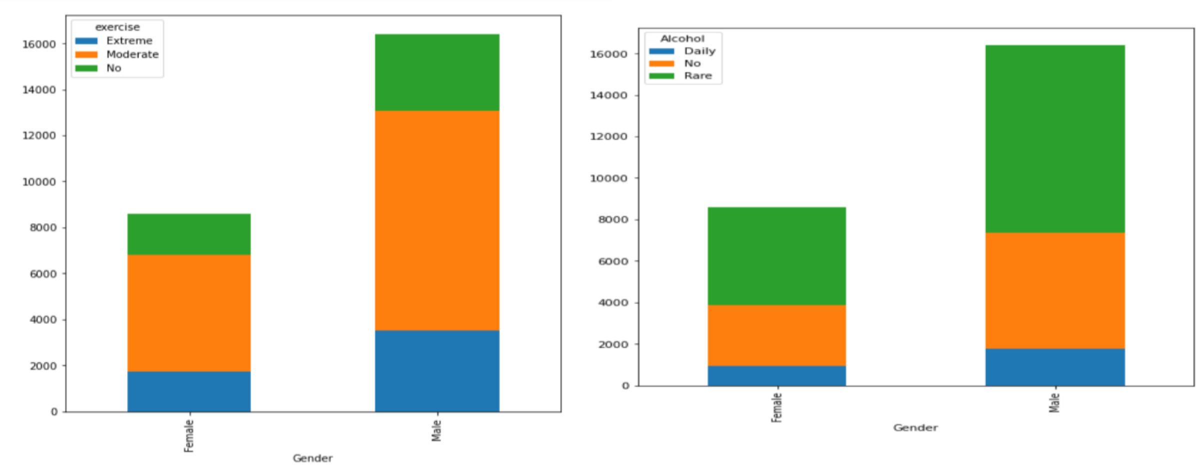


Fig-5-Stackbar of Age analysis with other variables

Maximum are male in adult age group . There are more business holders and students in adult age group.

Most of the variables shows healthy habits with all age group.



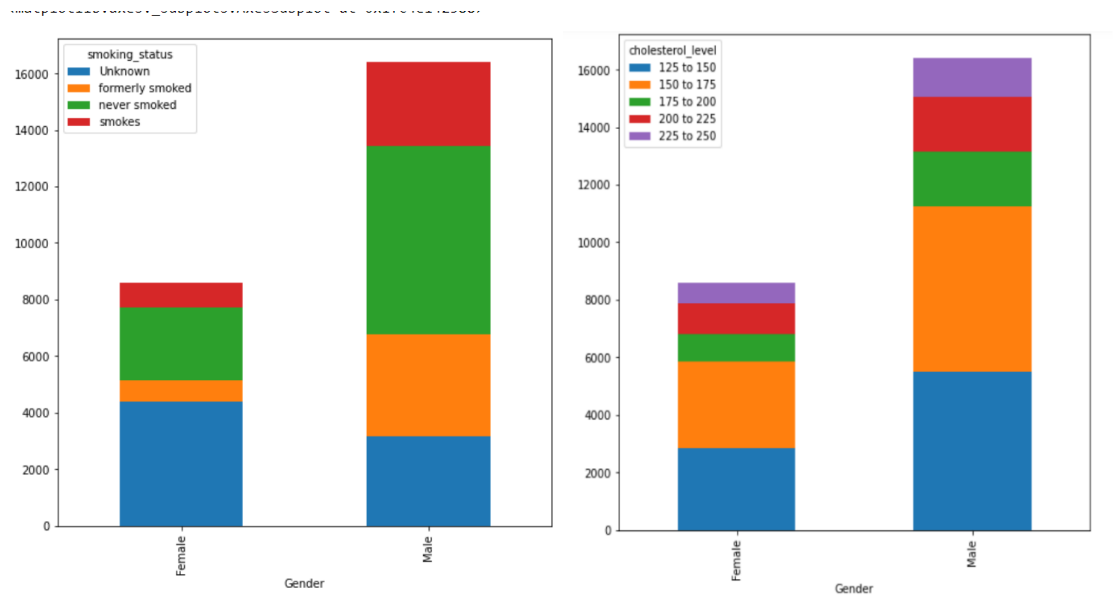


Fig-6-Stackbar of Gender analysis with other variables

As male population is on higher side with no extreme unhealthy habits

MULTIVARIATE ANALYSIS-

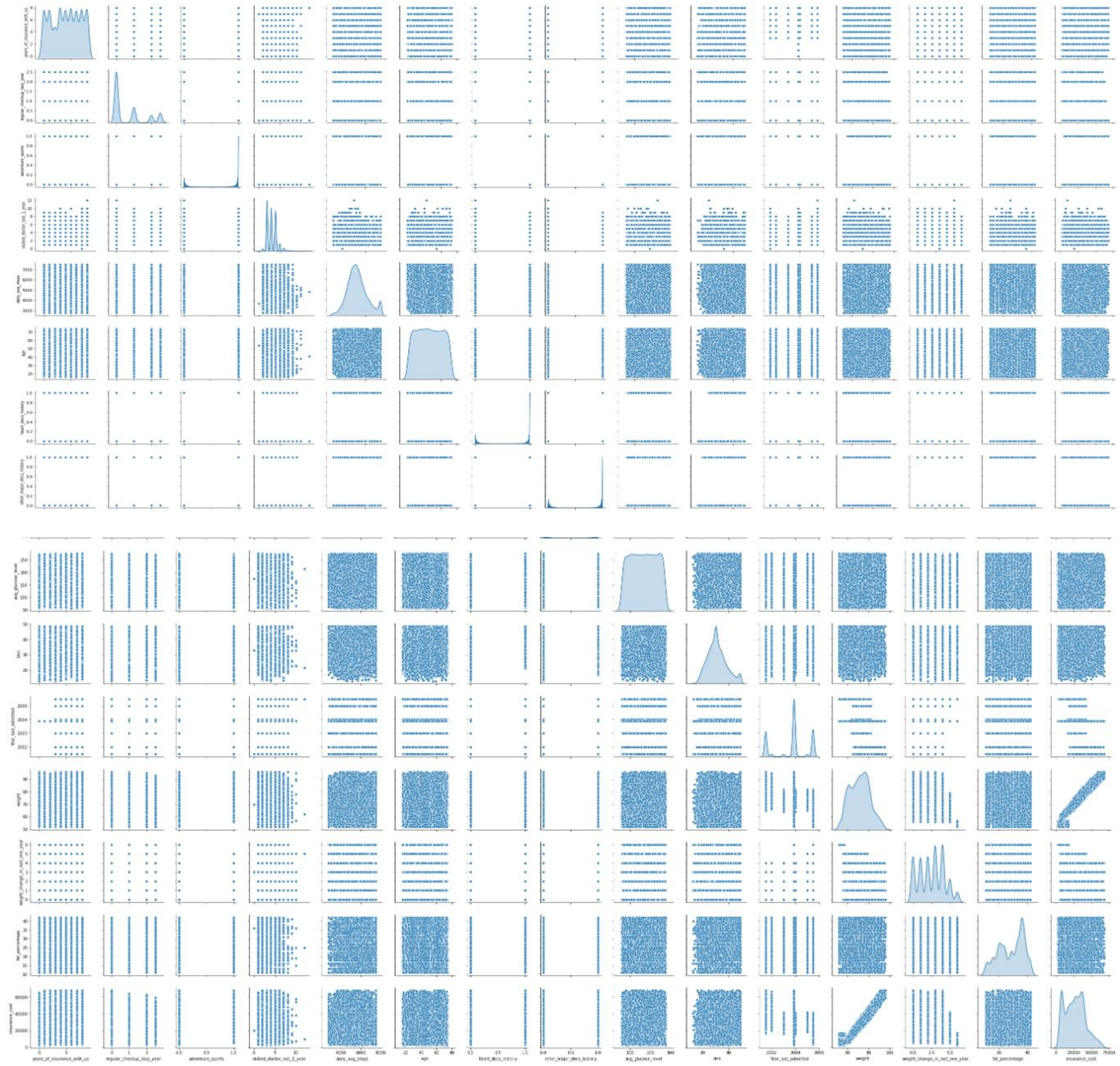
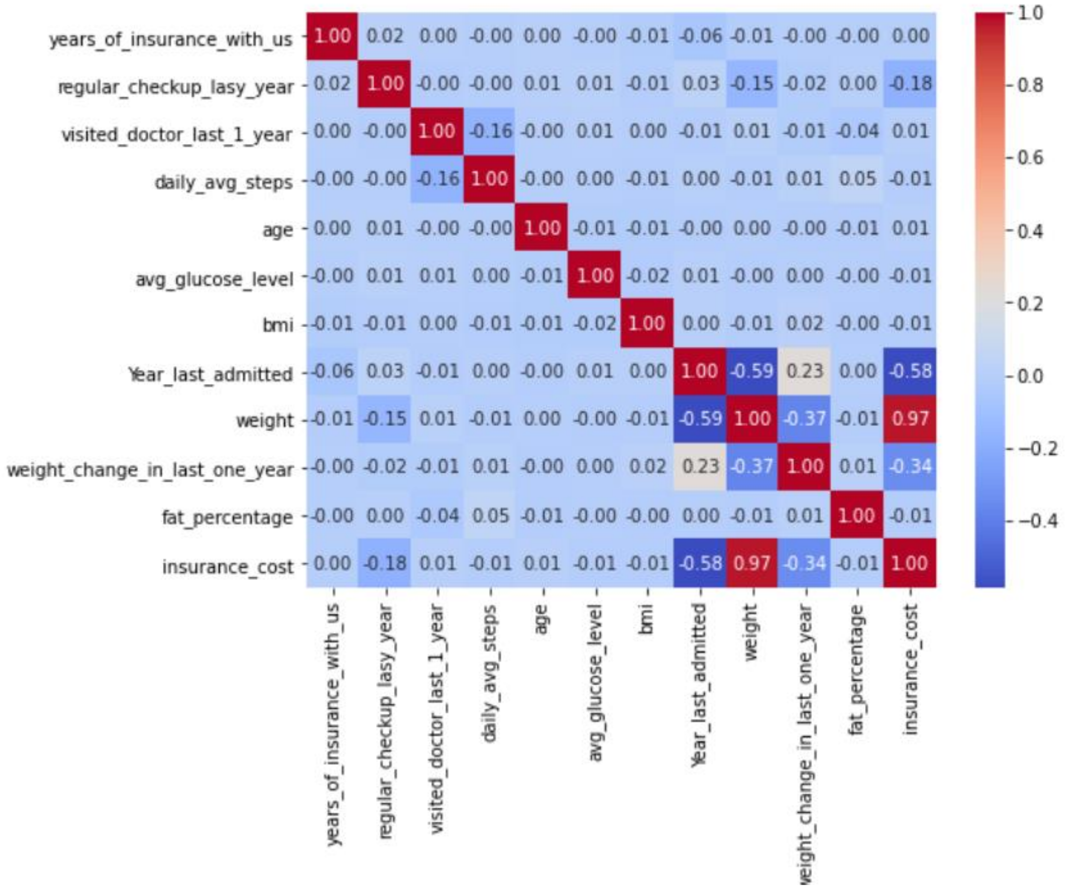
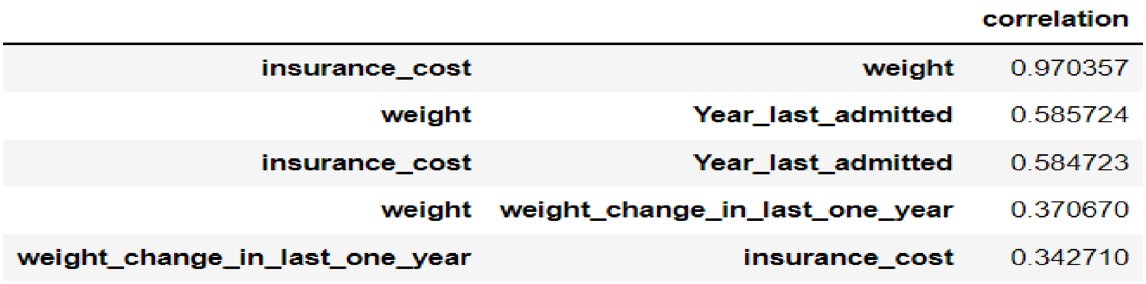


Fig-7-Pairplot



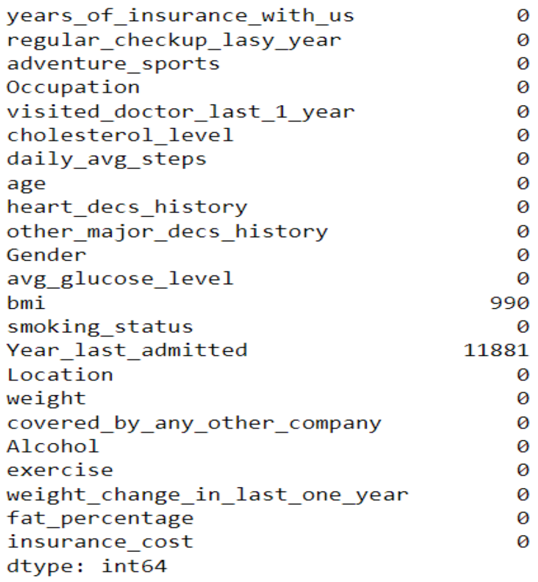


		correlation
insurance_cost	weight	0.970357
weight	Year_last_admitted	0.585724
insurance_cost	Year_last_admitted	0.584723
weight	weight_change_in_last_one_year	0.370670
weight_change_in_last_one_year	insurance_cost	0.342710

Fig-8-Heatmap

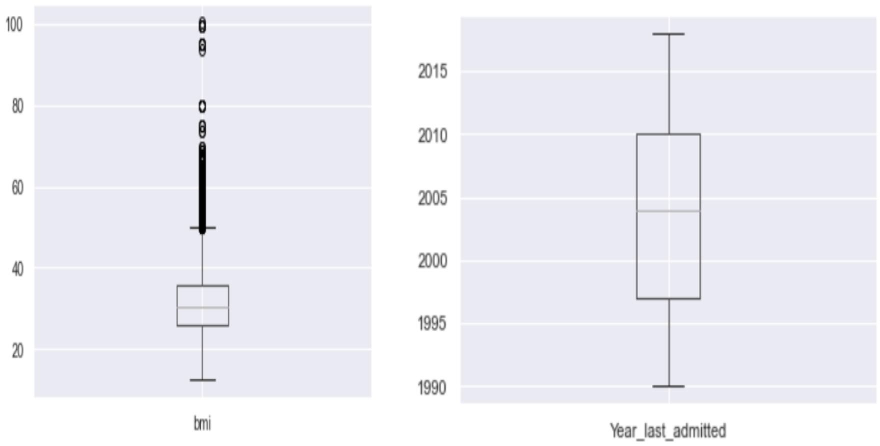
From the heatmap and pair plot the presence of no multicollinearity is visible. Except insurance cost and weight we no strong correlation amongst the variable is observed.

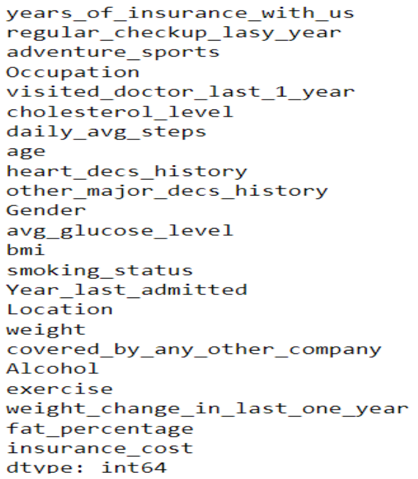
Check for the null values –



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

BMI AND Year _last _admitted showed 990 and 11881 respectively. Median imputation was applied for BMI as outliers were present and mean imputation was doe for Year _last _admitted as no outliers were seen.





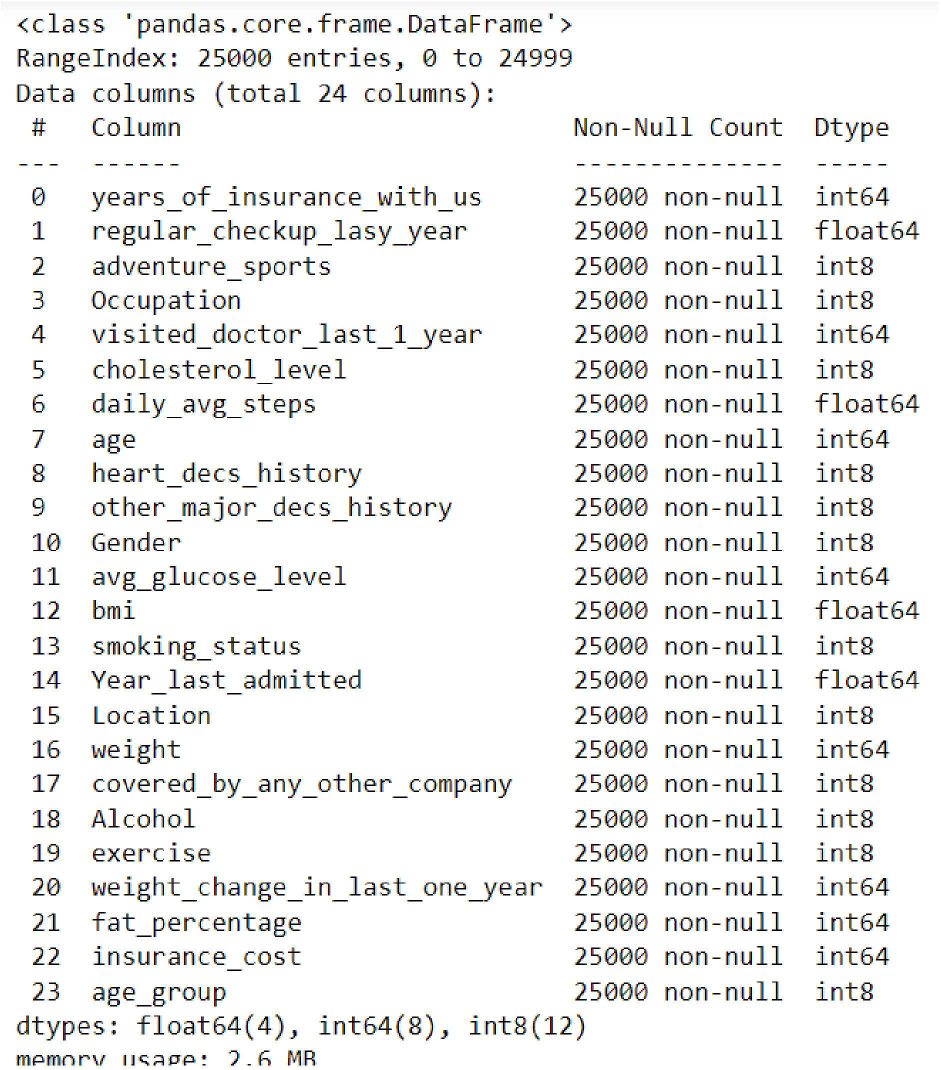
```
years_of_insurance_with_us        0
regular_checkup_lasy_year         0
adventure_sports                  0
Occupation                        0
visited_doctor_last_1_year        0
cholesterol_level                 0
daily_avg_steps                   0
age                               0
heart_decs_history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

After treating no null values were manifested.

Table-6-Null Value Treatment

As linear regression analysis does not accept any object data type, all data types were converted into integer.

Scaling of data was performed.



```
<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_lasy_year       25000 non-null  float64
 2   adventure_sports                25000 non-null  int8   
 3   Occupation                      25000 non-null  int8   
 4   visited_doctor_last_1_year      25000 non-null  int64  
 5   cholesterol_level               25000 non-null  int8   
 6   daily_avg_steps                 25000 non-null  float64
 7   age                             25000 non-null  int64  
 8   heart_decs_history              25000 non-null  int8   
 9   other_major_decs_history        25000 non-null  int8   
 10  Gender                          25000 non-null  int8   
 11  avg_glucose_level               25000 non-null  int64  
 12  bmi                             25000 non-null  float64
 13  smoking_status                  25000 non-null  int8   
 14  Year_last_admitted              25000 non-null  float64
 15  Location                        25000 non-null  int8   
 16  weight                          25000 non-null  int64  
 17  covered_by_any_other_company    25000 non-null  int8   
 18  Alcohol                         25000 non-null  int8   
 19  exercise                        25000 non-null  int8   
 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  age_group                       25000 non-null  int8   
dtypes: float64(4), int64(8), int8(12)
memory usage: 2.6 MB
```

Clustering was performed on the data

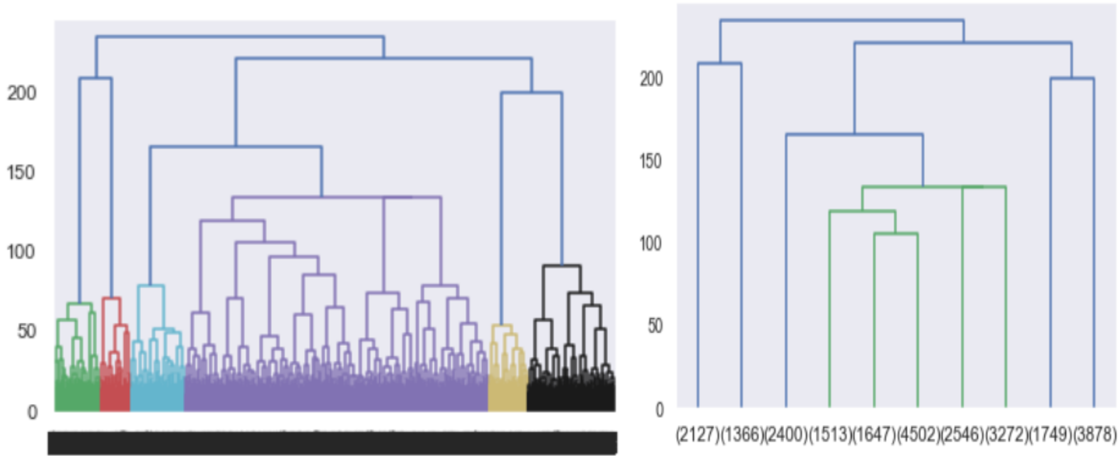
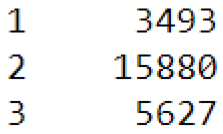


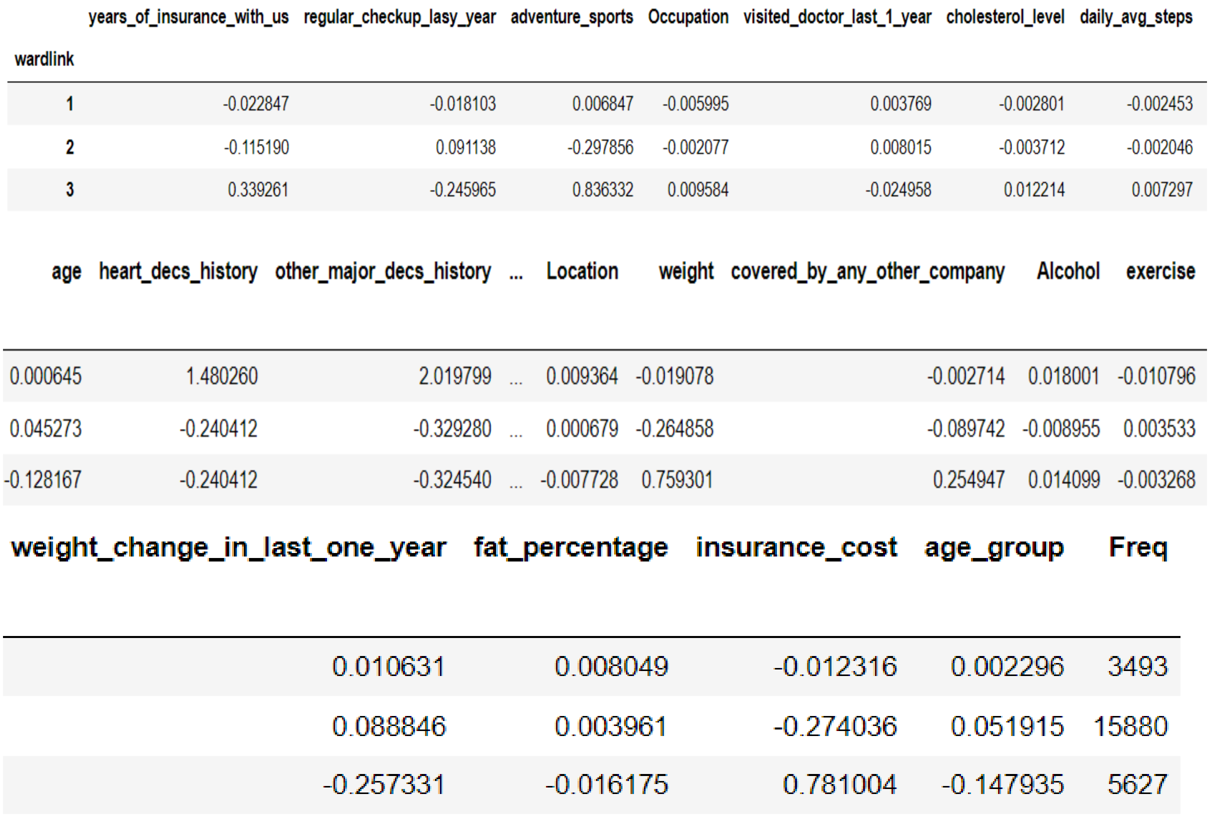
Fig-9-Clustering

Dendrogram 1 indicates all the data points have clustered to different clusters by wards method. To find the optimal number cluster through which we can solve our business objective we use truncate mode = lastp. Wherein we can give last p = 10 according to industry set base value and we get dendrogram 2. Now, we can understand all the data points have clustered into 3 clusters.

Now, we can look at the cluster frequency in our dataset



```
1     3493
2    15880
3     5627
```



wardlink	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age	heart_decs_history	other_major_decs_history	...	Location	weight	covered_by_any_other_company	Alcohol	exercise	weight_change_in_last_one_year	fat_percentage	insurance_cost	age_group	Freq
1	-0.022847	-0.018103	0.006847	-0.005995	0.003769	-0.002801	-0.002453	0.000645	1.480260	2.019799	...	0.009364	-0.019078	-0.002714	0.018001	-0.010796	0.010631	0.008049	-0.012316	0.002296	3493
2	-0.115190	0.091138	-0.297856	-0.002077	0.008015	-0.003712	-0.002046	0.045273	-0.240412	-0.329280	...	0.000679	-0.264858	-0.089742	-0.008955	0.003533	0.088846	0.003961	-0.274036	0.051915	15880
3	0.339261	-0.245965	0.836332	0.009584	-0.024958	0.012214	0.007297	-0.128167	-0.240412	-0.324540	...	-0.007728	0.759301	0.254947	0.014099	-0.003268	-0.257331	-0.016175	0.781004	-0.147935	5627

Table-7-Clustering Result

4-Business insights from EDA-

- Data is imbalanced as more input should have been collected from female population.
- Variables like eating habit, sleep cycle and frequent pill popping habit which attributes to complications like renal failure should have been included.
- Rather than daily and rare amount of alcohol and no of cigarettes consumed per day should have been included.
- Rather than BMI, Visceral fat content should have been included as it is more reliable than BMI because BMI also incudes muscles and bone density.
- Applicant with unhealthy lifestyle should also have been included properly so that insurance cost on higher side could also have been studied significantly.
- As maximum applicants are on healthy lifestyle side not much difference is expected.
- Location wise there was no much difference, it could have been dropped.
- After performing clustering we got three clusters with frequency- 3493,15880 and 5627