Optimizing the Medical Insurance Cost Per Individual

Capstone Healthcare Project: Final Report

Submitted by:

Akansha Pruthi

PGP – DSBA Jan 2021-22

Project Mentor: Prakasha YB

GREAT LAKES
INSTITUTE OF MANAGEMENT, GURGAON

https://www.mygreatlearning.com

Acknowledgement

We wish to place on record our deep appreciation for the guidance and help provided to us by our Mentor Prakasha YB. Who helped us narrow down on the choice of the Project as well as the scope and focus area of the Project. He gave us valuable feedback at every stage to enhance the process and the outputs.

We would also like to place on record our appreciation for the guidance provided by Dr. Rishabh Pandey for giving us valuable feedback in Presentation and being a source of inspiration in helping us to work on this project.

We certify that the work done by us for conceptualizing and completing this project is original and authentic.

Date: 5th February 2022 Name: Akansha Pruthi

Place: Delhi, India

Table of Contents

Introduction

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 must always be proactive in this 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 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.

Need of the Study

- **a.** Understands how the various variables and perimeters affect the status of prediction model.
- **b.** Understands the behaviour of the individual by analysing the patterns of their activities.
- **c.** Identify the scope of reducing risk by identifying the right segment.
- **d.** Understands the data to study the need of adding variables (if applicable).

Understanding business/social opportunity

a. Good balance life: Proper diet and eat healthy will reduce the health risk and impact the Insurance cost also per individual.
b. Financial assistance.
c. Need good care, less costs.
d. Challenges face by Health care industry.
e. To maximize the value of health care money invested in the workforce.

Understanding/Visual Inspection of the Data

Data set contains-

a. Mutually non-overlapping response categories.
b. Mixed data (quantitative and qualitative)
c. BMI and Customers exercise record
d. 25,000 rows and 24 columns
e. Data Type – 8 object and 16 float/int
f. There are 990 missing values in BMI column and 11,881 in year last admitted.
g. Insurance cost is the target variable.

Data Dictionary

Variable	Description
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_last_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 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 (Target Variable)

Descriptive Summary of the data:

- For Continuous variables

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

- For Categorical Variables (Value Count)

Variables	Count
Occupation	
Student	10,619
Business	10,020
Salaried	4,811
Exercise	
Moderate	14,638
Extreme	5,248
No	5,114
Cholesterol Level	
125 to 150	8,339
150 to 175	8,763
175 to 200	2,881
200 to 225	2,963
225 to 250	2,054
Gender	
Male	16,422
Female	8,578
Alcohol	
Rare	13,752
No	8,541
Daily	2,707
Covered by other company	
Yes	7,582
No	17,418

Variables	Count
Smoking Status	
Never smoked	9,249
Unknown	7,555
Formerly smoked	4,329
Smokes	3,867
Location	
Bangalore	1,742
Jaipur	1,706
Bhubaneshwar	1,704
Mangalore	1,697
Delhi	1,680
Ahmedabad	1,677
Guwahati	1,672
Chennai	1,669
Kanpur	1,664
Nagpur	1,663
Mumbai	1,658
Lucknow	1,637
Pune	1,622
Kolkata	1,620
Surat	1,589

Data Info:

```
<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_last_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
```

Findings:

As, we can see in data summary that the maximum age respondent is 74 and the minimum is 16. Some respondents have 0 years of insurance which means that they are not covered under any insurance policy along with they don't even went for regular check up in the past years. Which can be led to the respondent's medication errors. Average daily steps are between 5000 – 6000, which is a low active as per the internet source (WHO report) whereas average 10,000 steps per day is considered for a healthy and active adult.

we can observe that there are spelling, typos and formatting errors in the data info. After cleaning the data formatting errors and dropping of column year last admitted has high percentage of missing values, dropping missing values will affect the data accuracy. We are dropping them from our dataset to make sure that other valid observations do not get eliminated when we remove or impute the 'na' values.

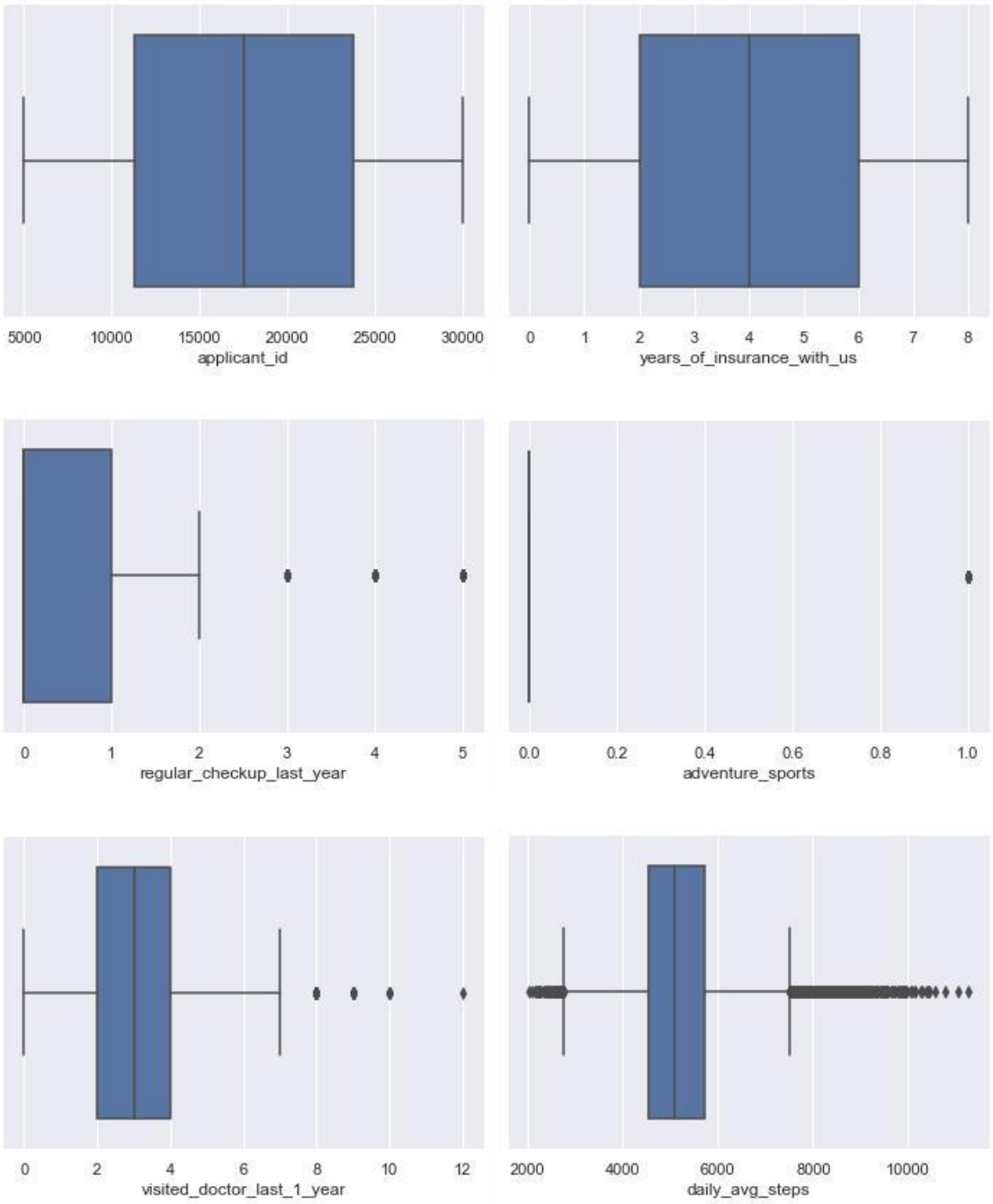
Descriptive Summary of target variable:

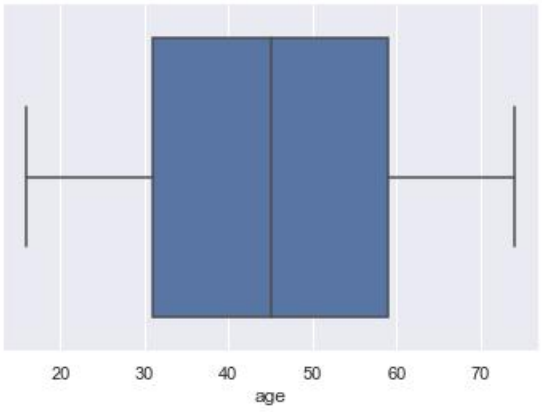
```
count     24010.000000
unique             NaN
top                NaN
freq               NaN
mean      27160.643232
std       14332.038022
min        2468.000000
25%       16042.000000
50%       27148.000000
75%       37020.000000
max       67870.000000
Name: insurance_cost, dtype: float64
```

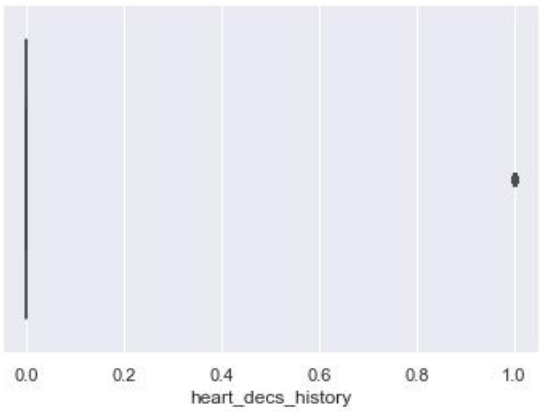
Minimum insurance cost per individual is 2,468.00 and maximum is 67,870.00
Average insurance cost per individual is 27,160.00

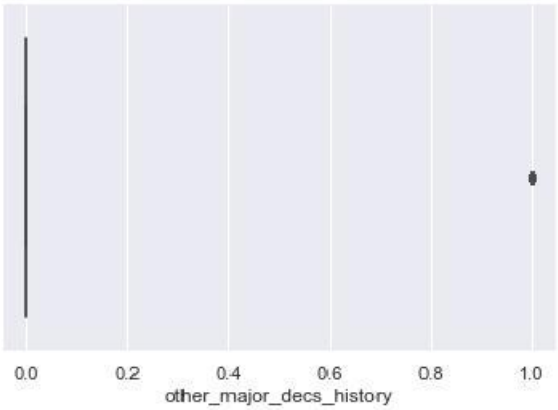
Exploratory Data Analysis (EDA)

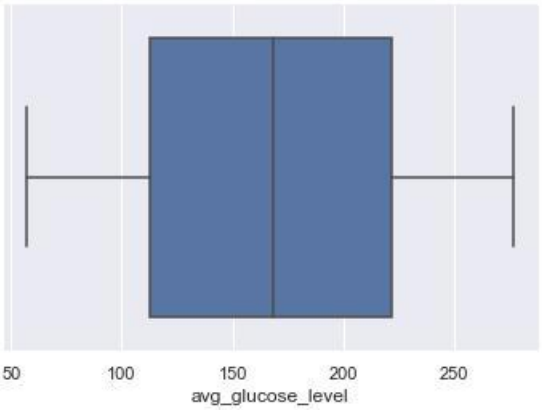
Univariate Analysis of Numerical Variables (Box Plots)

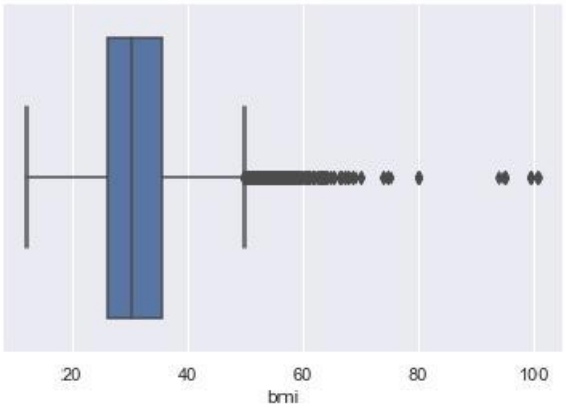


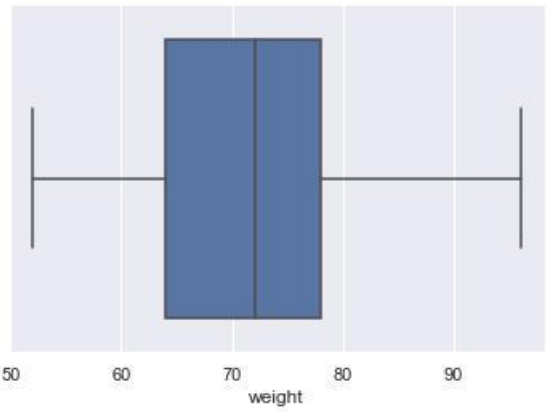

20
30
40
50
60
70
age

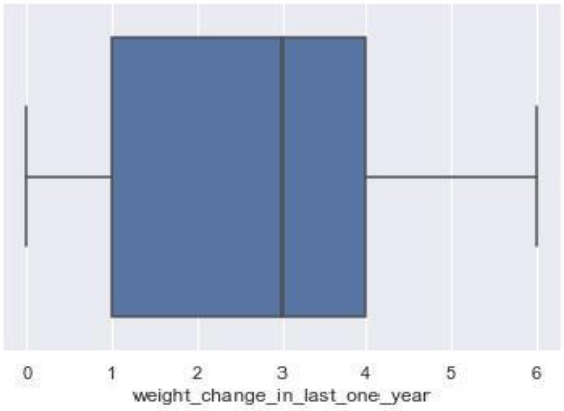

0.0
0.2
0.4
0.6
0.8
1.0
heart_decs_history

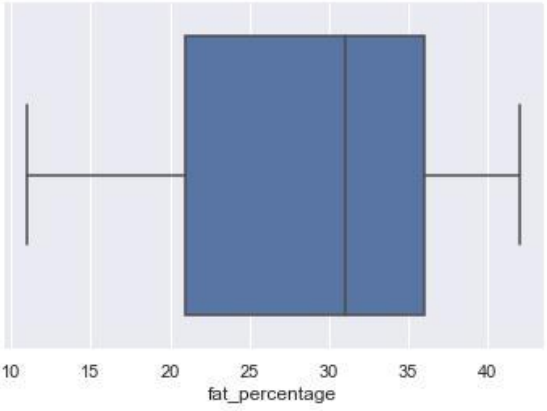

0.0
0.2
0.4
0.6
0.8
1.0
other_major_decs_history

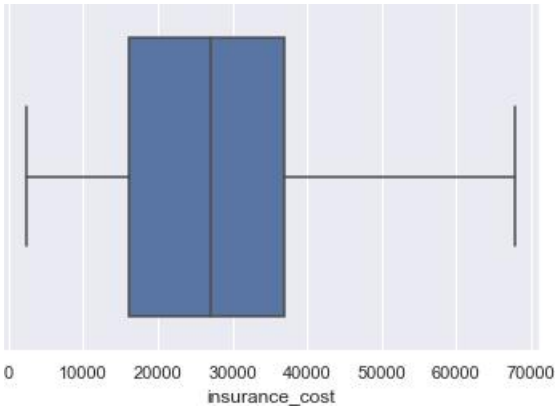

50
100
150
200
250
avg_glucose_level


20
40
60
80
100
bmi


50
60
70
80
90
weight


0
1
2
3
4
5
6
weight_change_in_last_one_year

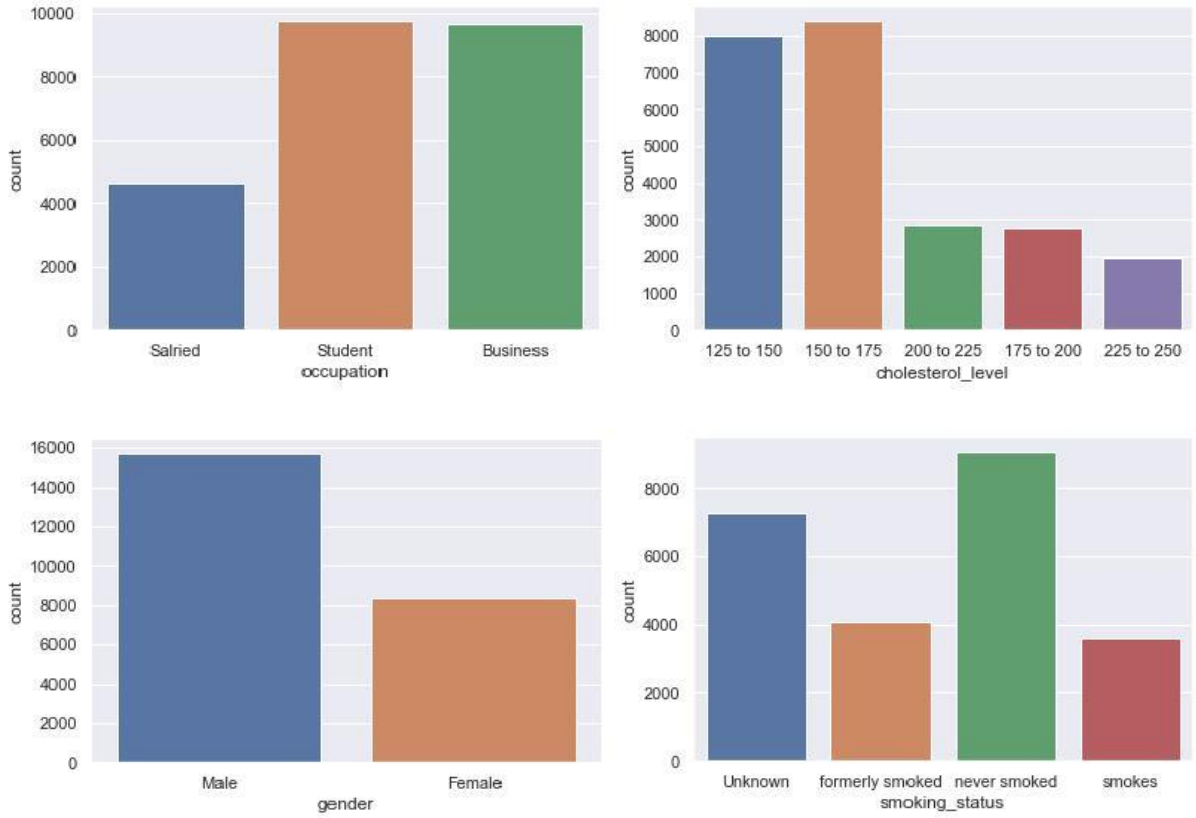

10
15
20
25
30
35
40
fat_percentage

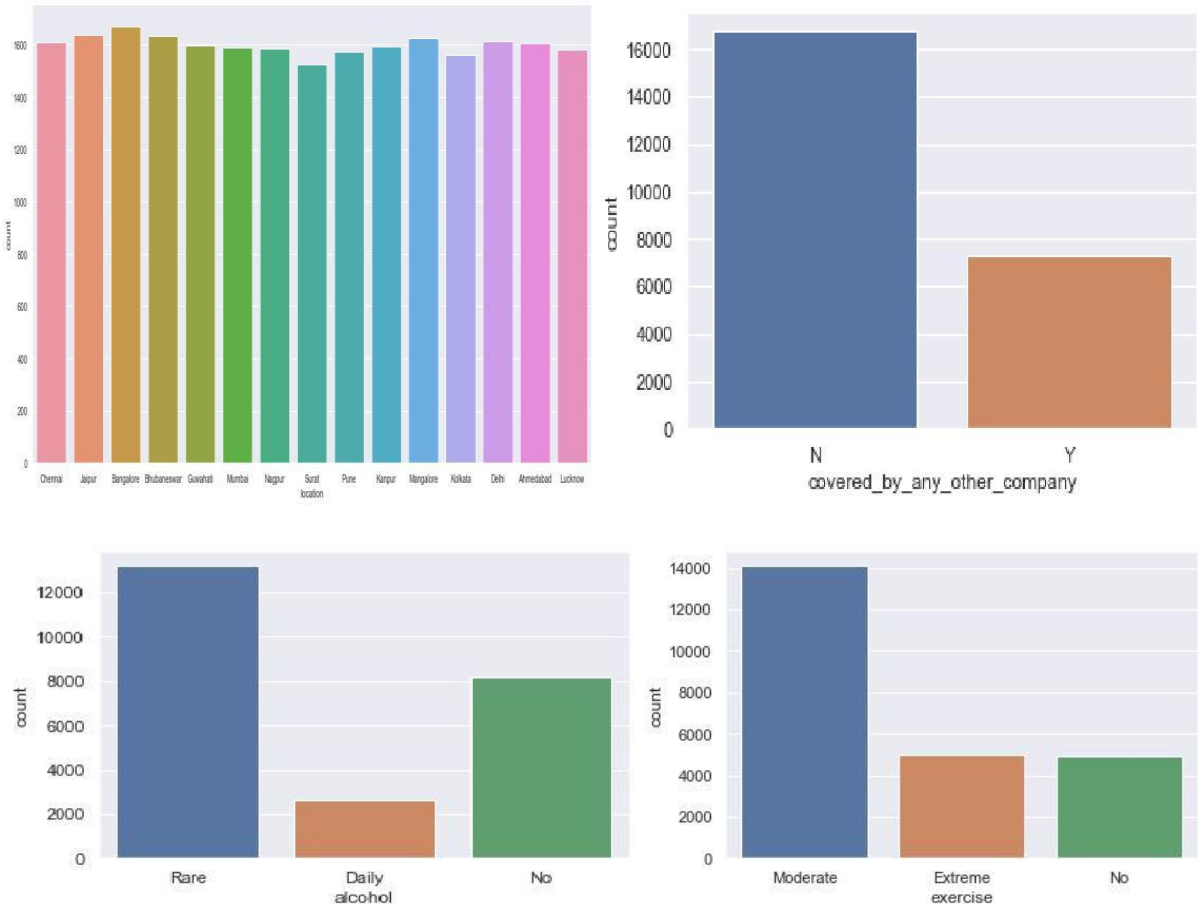


Findings:

There are outliers present in 7 variables (Regular checkups, adventure sports, visited doctor last one year, daily average steps, heart diseases history, other major diseases history and BMI).

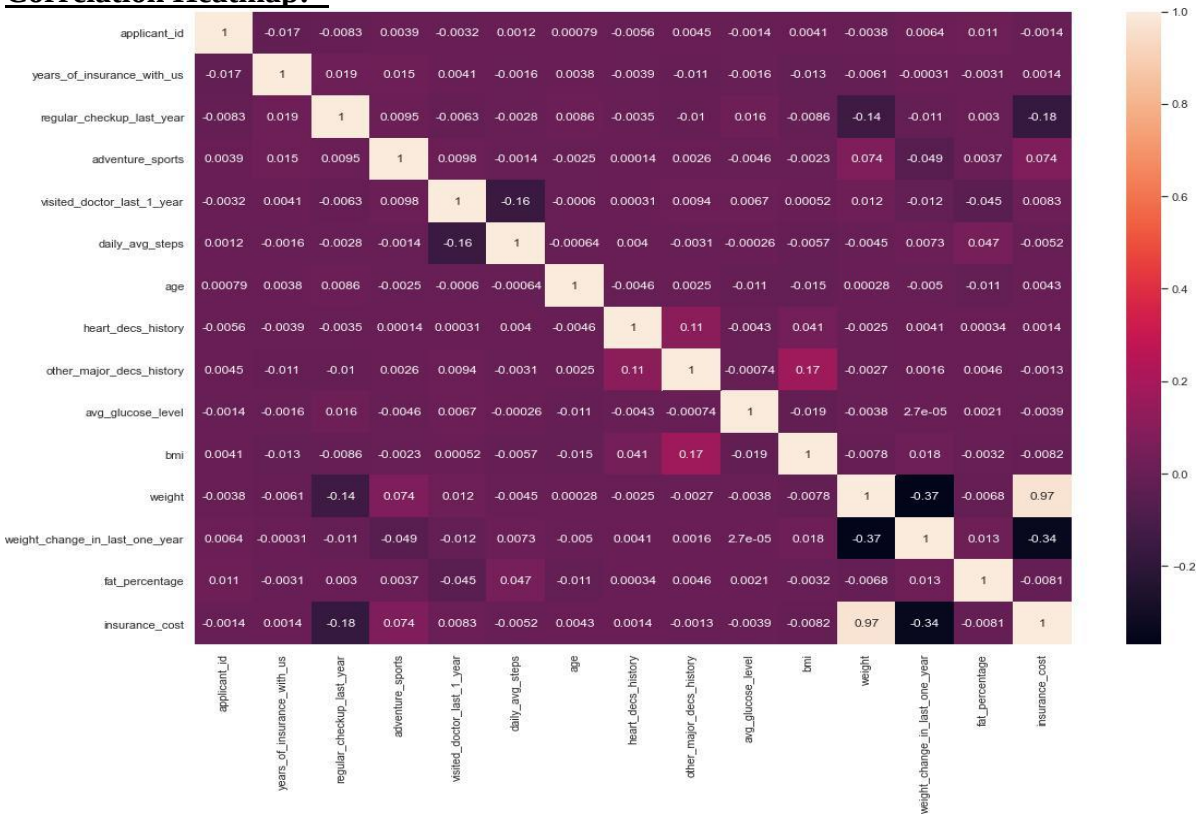
Univariate Analysis of Categorical variables (Bar Plot)





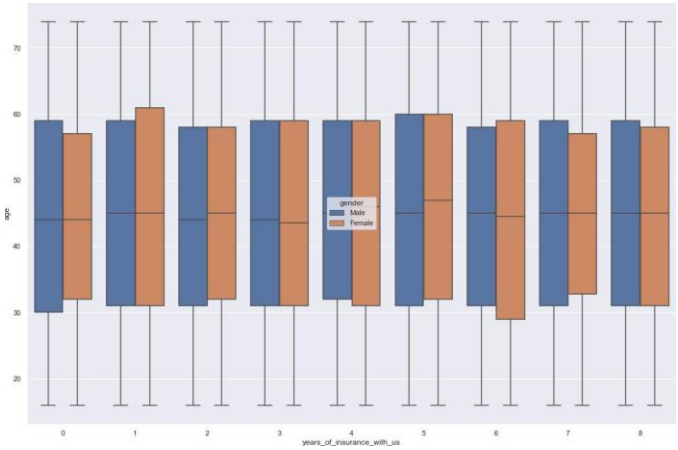
Bivariate Analysis

Correlation Heatmap: -

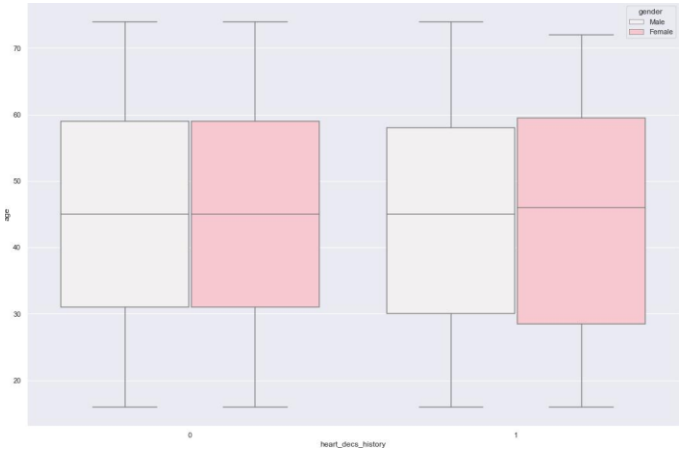


Categorical variables X continuous Variables (Box Plots)

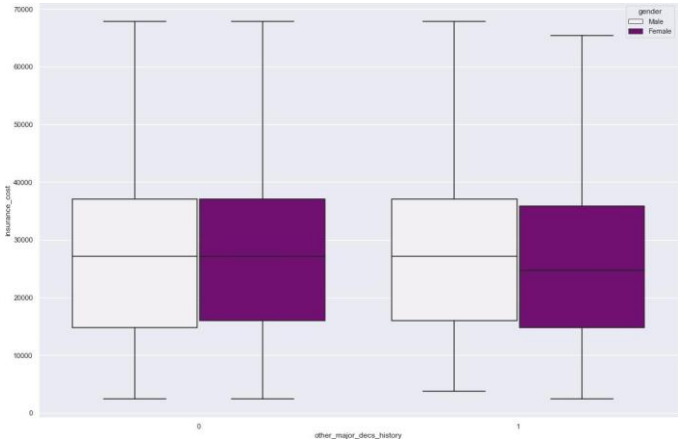
- Years of insurance with us Vs Age (Hue = Gender)



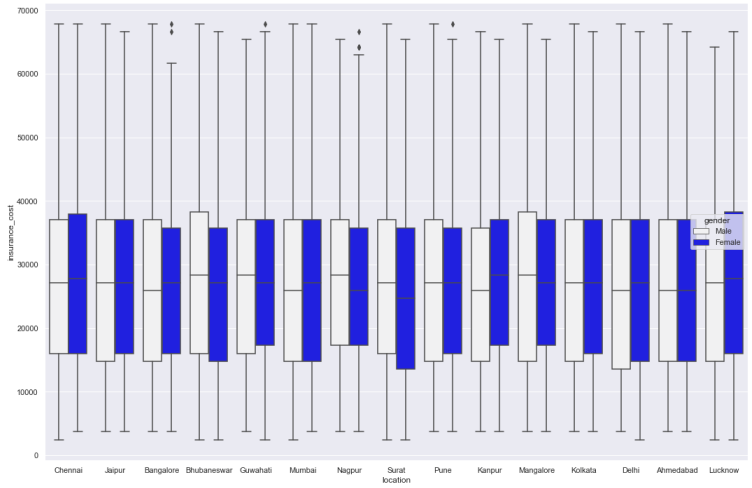
- Heart Diseases History Vs Age (Hue = Gender)



- Other major diseases history Vs Insurance Cost (Hue = Gender)



- Location Vs Insurance Cost (Hue = Gender)



Multivariate Analysis (Pair Plot)

Table (In %)

	Location													
	Ahmeda..	Bangalo..	Bhuban..	Chennai	Delhi	Guwaha..	Jaipur	Kanpur	Kolkata	Lucknow	Mangal..	Mumbai	Nagpur	Pune	Surat
% of Total Adventure Sports ..	7.49%	6.85%	7.00%	6.31%	6.95%	6.56%	7.68%	5.29%	6.22%	6.07%	6.12%	6.61%	7.49%	5.73%	7.64%
% of Total Avg Glucose Level alon..	6.73%	6.96%	6.81%	6.68%	6.71%	6.82%	6.73%	6.60%	6.55%	6.57%	6.72%	6.72%	6.62%	6.43%	6.35%
% of Total Daily Avg Steps along Locati..	6.72%	6.96%	6.84%	6.65%	6.71%	6.68%	6.89%	6.65%	6.48%	6.53%	6.72%	6.66%	6.63%	6.55%	6.34%
% of Total Heart Decs History along..	6.81%	6.30%	7.61%	5.78%	5.42%	6.66%	7.47%	6.37%	7.54%	6.81%	7.10%	6.52%	6.37%	6.88%	6.37%
% of Total Other Major Decs Histor..	6.60%	6.97%	7.21%	6.85%	6.52%	6.28%	6.64%	7.25%	6.89%	6.81%	7.05%	6.44%	5.75%	5.91%	6.85%
% of Total Fat Percentage along ..	6.63%	7.04%	6.81%	6.62%	6.78%	6.69%	6.83%	6.65%	6.40%	6.57%	6.76%	6.64%	6.64%	6.53%	6.39%

Missing Value Treatment: There 11,881 missing values in last year admitted and 990 values missing under BMI. As last year admitted values are more than 50 percentage. Dropping them will highly impact the data accuracy and results. So, for 11,881 missing values we dropped that whole column and dropped 990 missing values from the data.

Business Insights from EDA

a. There is no strong correlation found in the heatmap of the given dataset.
b. Target variable is Insurance cost. So will build the model using that target variable.
c. There are outliers present in the data regular check-up, adventure sports, average daily steps.
d. There is negative correlation between weight & weight changed in last year. Which means there is no such good progress in respondents' health which led to increase in the risk of life.
e. From the insurance company's perspective, they need to introduce such scheme with insurance policies so that it would benefit both parties.
f. Is the data unbalanced? Yes, the data is not correlated and there are missing values in last year admitted in hospital which highly impact the data accuracy

Model building and interpretation

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

Model Tuning

1. Ensemble modelling, wherever applicable
2. Any other model tuning measures (if applicable)
3. Interpretation of the most optimum model and its implication on the business

Linear regression model, ANN, Decision Tree and Random Forest model is created using Scikit and stats model packages.

Output from Stats Model and using comparison tables is as follows:

Models	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3375.862926	3339.684420	0.944401	0.945956
Decision Tree Regressor	0.000000	4356.047598	1.000000	0.908056
Random Forest Regressor	1166.608857	3092.671229	0.993360	0.953655
ANN Regressor	3009.981169	3093.412815	0.955800	0.953632

Values for RMSE models is extreme high, we cannot consider, we will try to perform model tuning see the results and interpret the best model to optimize the insurance cost per individual. Values for R-square models are good score. We can see and compare their scores to interpret the business model.
For now, let's do the grid search for each model and find the best parameters to identify the accurate model.

Model Tuning

Grid Search on Decision Tree
{'max_depth': 10, 'min_samples_leaf': 30, 'min_samples_split': 15}

Grid Search for Random Forest
GridSearchCV(cv=3, estimator=RandomForestRegressor(random_state=150),
param_grid={'max_depth': [7, 10], 'max_features': [4, 6],
'min_samples_leaf': [3, 15, 30],
'min_samples_split': [30, 50, 100],
'n_estimators': [300, 500]})

Grid Search for ANN
As per the consolidated comparison table Decision tress has the best r-square test score.
As all the values in the performances are too high, one should try to add more features in it to improve the model performances

Business Insights & Recommendations

General Talk - Why Health Insurance?

a. Health insurance could be a way of removing financing barriers and improving access to healthcare.
b. Health insurance could be a way of providing financial protection against high medical
c. expenses.
d. Health insurance could be a way of negotiating with the providers for better quality health care.

Recommendations:

a. They must introduce the Skills upgrading scheme which may help individual to organize their day-to- day activities in a structured way. Such as it should include managerial, administrative, technical and social skills.
   Managerial skills – to manage the entire program
   Administrative skills – to manage finances and the funds
   Technical skills – to understand the complexities of health insurance
   Social skills – to understand the community's needs

b. To optimize the best insurance cost per individual they must execute the 15 to 30 days routine check-up by organizing medical camps and awareness program.

c. There must be a network of health care providers (public or private). Without this, it is not wise to talk about health insurance.

d. The people must have the capacity to pay the premium. Unless there will be no takes for health insurance.

e. There are many more recommendations which needs to be fulfilled but at least you should focus on executing the three recommendations to introduce the healthcare insurance programme.

List Of Tables & Charts

The End….