HEALTHCARE INSURANCE COST PREDICTION

By,

SHANMUGAPRIYA K

CONTENTS

LIST OF FIGURES

1. INTRODUCTION

The goal of this project is to get an idea about the necessary amount required according to individual's health status. Later they can comply with any health insurance company and their schemes & benefits keeping in mind the predicted amount from our project. This can help a person in focusing more on the health aspect of an insurance rather than the futile part.

Health insurance is one of the major contributors of growth of general insurance industry in India. It is an emerging insurance sector after life and automobile insurance sector. It alone accounts for around 29% of total general insurance premium income earned in India. Rise in middle class, higher hospitalization cost, expensive health care, digitization and increase in awareness level are some important drivers for the growth of health insurance market in India.

According to the Economic Times, inflation in healthcare is also growing at a rate of 12 to 18%! This includes overall costs such as cost of medicines, hospital admission charges, cost of various treatments, medical advancements and so on. Due to the rise in these expenses, your insurer too needs to increase your sum insured every year i.e. coverage to be able to cover for these costs when you make a claim. This is primarily why there is consequently an increase in health insurance premium too when renew for the new policy year.

So, providing optimal insurance cost is an important requirement in healthcare industry since there is a need for providing an optimum cost to the needy at the right time to avoid getting into a financial trouble when they are already into a health crisis.

Because, lifestyles have changed and rare non-communicable diseases are now common. Other reasons are incomplete financial planning, Medical care is unbelievably expensive

So, the benefitting individual or brokers or third-party administrators servicing health insurance claims would get the most optimum insurance cost from the prediction where they protect the clients from financial turmoil.

2. DATA REPORT

2.1. VISUAL INSPECTION OF DATA

There are 25000 rows and 24 columns are present in the dataset in which 2 float variables, 18 int variables, and 8 objects are present.

Below are the summary of the data where mean, std, 25%, 50% and 75% are shown.

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level
count	25000.000000	25000.000000	25000.000000	25000.000000	25000	25000.000000	25000
unique	NaN	NaN	NaN	NaN	3	NaN	5
top	NaN	NaN	NaN	NaN	Student	NaN	150 to 175
freq	NaN	NaN	NaN	NaN	10169	NaN	8763
mean	17499.500000	4.089040	0.773680	0.081720	NaN	3.104200	NaN
std	7217.022701	2.606612	1.199449	0.273943	NaN	1.141663	NaN
min	5000.000000	0.000000	0.000000	0.000000	NaN	0.000000	NaN
25%	11249.750000	2.000000	0.000000	0.000000	NaN	2.000000	NaN
50%	17499.500000	4.000000	0.000000	0.000000	NaN	3.000000	NaN
75%	23749.250000	6.000000	1.000000	0.000000	NaN	4.000000	NaN
max	29999.000000	8.000000	5.000000	1.000000	NaN	12.000000	NaN

11 rows × 24 columns

Year_last_admitted	Location	weight	covered_by_any_other_company	Alcohol	exercise	weight_change_in_last_one_year	fat_percentage	insurance_cost
13119.000000	25000	25000.000000	25000	25000	25000	25000.000000	25000.000000	25000.000000
NaN	15	NaN	2	3	3	NaN	NaN	NaN
NaN	Bangalore	NaN	N	Rare	Moderate	NaN	NaN	NaN
NaN	1742	NaN	17418	13752	14638	NaN	NaN	NaN
2003.892217	NaN	71.610480	NaN	NaN	NaN	2.517960	28.812280	27147.407680
7.581521	NaN	9.325183	NaN	NaN	NaN	1.690335	8.632382	14323.691832
1990.000000	NaN	52.000000	NaN	NaN	NaN	0.000000	11.000000	2468.000000
1997.000000	NaN	64.000000	NaN	NaN	NaN	1.000000	21.000000	16042.000000
2004.000000	NaN	72.000000	NaN	NaN	NaN	3.000000	31.000000	27148.000000
2010.000000	NaN	78.000000	NaN	NaN	NaN	4.000000	36.000000	37020.000000
2018.000000	NaN	96.000000	NaN	NaN	NaN	6.000000	42.000000	67870.000000

Looks like there is no much outliers present in the dataset except for ‘bmi’ and ‘year’ columns.

2.2. UNDERSTANDING OF ATTRIBUTES

From the given dataset,

The ‘bmi’ column is defined into four categories under column name (‘cbmi’)

1. Underweight when bmi is below 18.5.
2. Normal weight when bmi is from 18.5 to 23.9
3. Overweight when bmi is between 24 and 29.9
4. Obese when bmi is greater than 30

The ‘age’ column is defined into 10 categories under column name (‘cage’) as

1. 0 for (From new born to 9 years old)- ‘0-9’
2. 1 as ’10-19’ years old
3. 2 as ’20-29’ years old
4. 3 as ’30-39’ years old
5. 4 as ’40-49’ years old
6. 5 as ’50-59’ years old
7. 6 as ’60-69’ years old
8. 7 as ’70-79’ years old
9. 8 as ’80-89’ years old
10. 9 as ’90-99’ years old

The daily_avg_steps column is defined into 11 categories under column name(‘cdailyavgsteps’) as

1. 0 as ‘0-999’ steps
2. 1 as '1000-1999' steps
3. 2 as ‘2000-2999’ steps
4. 3 as ‘3000-3999’ steps
5. 4 as ‘4000-4999’ steps
6. 5 as ‘5000-5999’ steps
7. 6 as ‘6000-6999’ steps
8. 7 as ‘7000-7999’ steps
9. 8 as ‘8000-8999’ steps
10. 9 as ‘9000-9999’ steps
11. 10 as ‘10000-10999’ steps
12. 11 as ‘11000-11999’ steps

The avg_glucose_level column is defined into 3 categories under column name(‘cavgglucoselevel) as

1. Normal when the value is less than 140
2. Pre-diabetic when the value is between 140 and 199
3. Diabetes when the value is greater than 200.

EXPLORATORY DATA ANALYSIS

2.3. UNIVARIATE ANALYSIS

The univariate analysis are represented as below:

Figure 1. Univariate Distribution of Gender

Among the given dataset, 65.34 % are Male and 34.66% are Female.

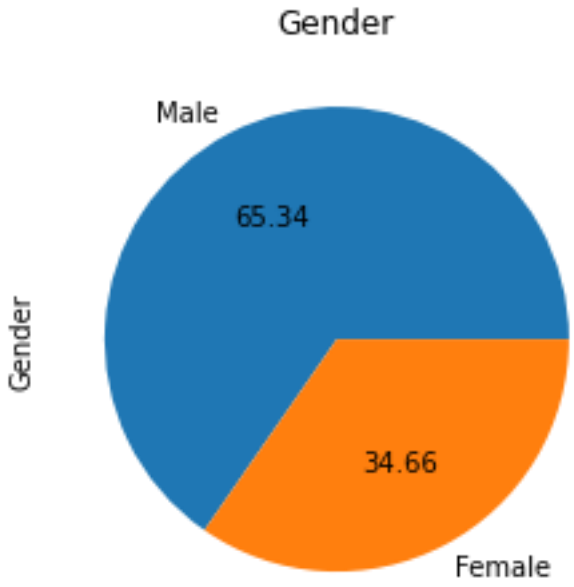


Figure 2. Univariate Distribution of Occupation

Among the given dataset, 40.60% are Student, 40.19% are Business and 19.21% are Salaried

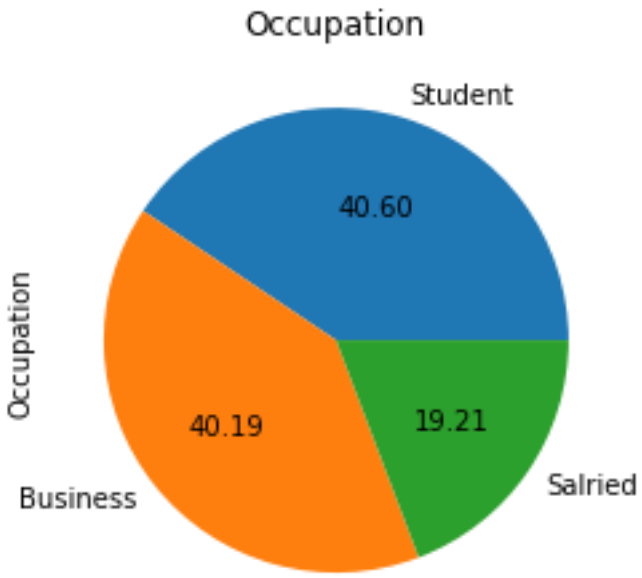


Figure 3. Univariate Distribution of Location

Bangalore is the top most location of the hospitals where individuals get admitted.

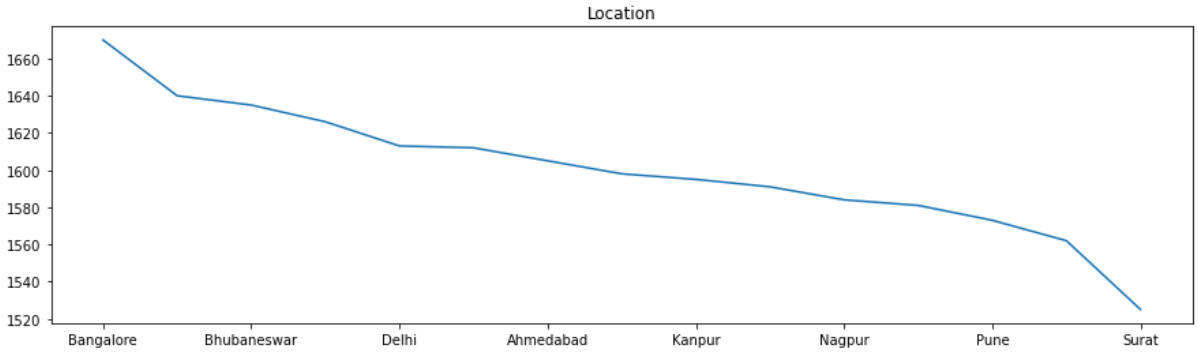


Figure 4. Univariate Distribution of Alcohol

Among the given dataset, only 10.85% of people consume alcohol daily, 54.99% of consumes rare and 34.16% of people doesn't consume alcohol at all.

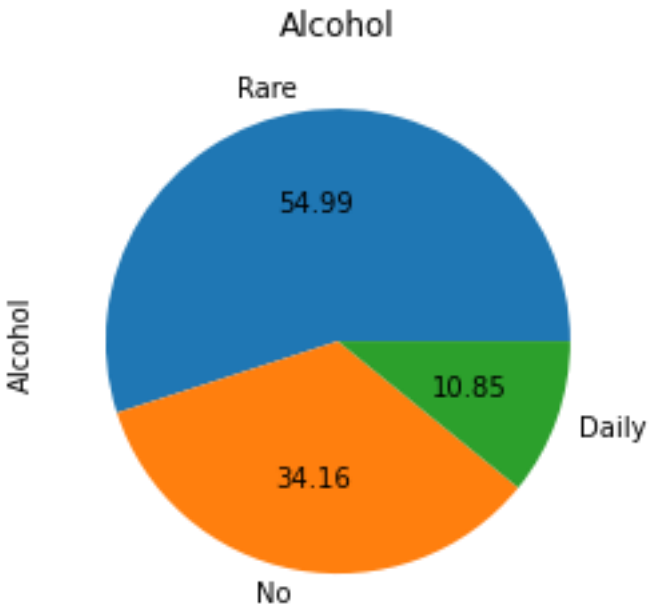
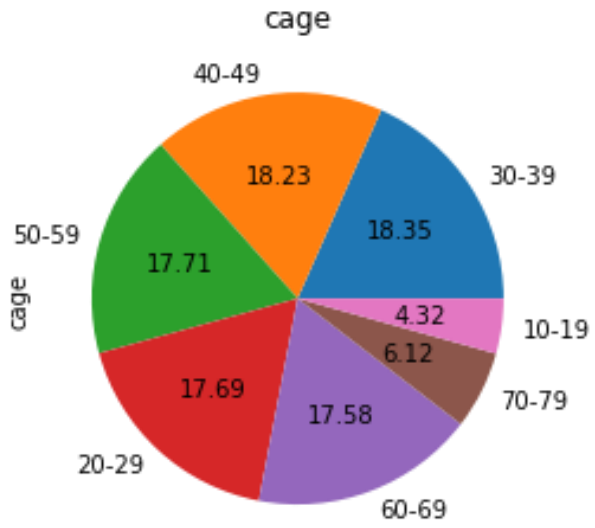


Figure 5. Univariate Distribution of categorised age

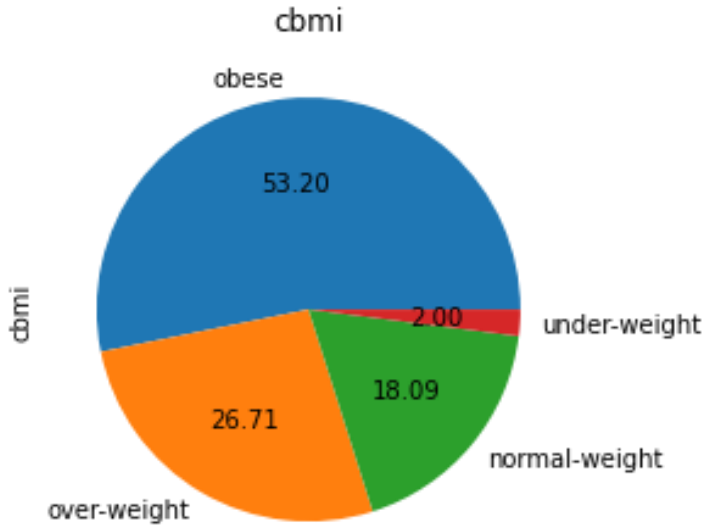
Since we categorized the age column, we are able to see the percentage of people falls into each category.



Only 4.32% of people are under 10-19 years old, most of the people are under 30-39 years old, i.e. 18.35%. Next to that, 18.23% of people are under 40-49 years old category.

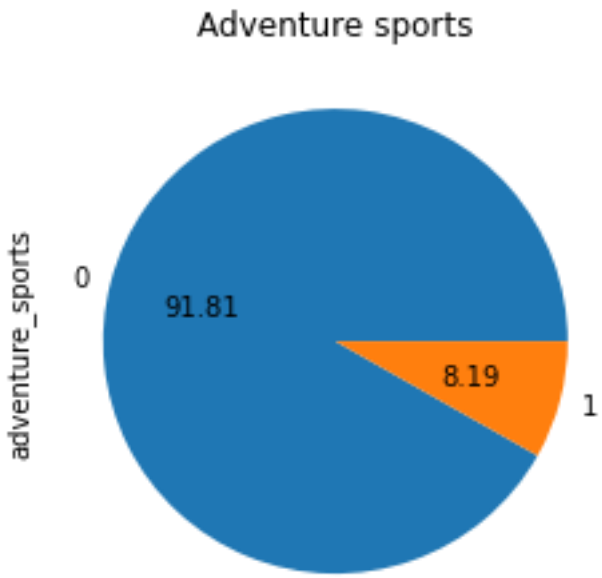
Figure 6. Univariate Distribution of categorised bmi

Since we categorized the bmi column, we are able to see the percentage of people falls into each category.



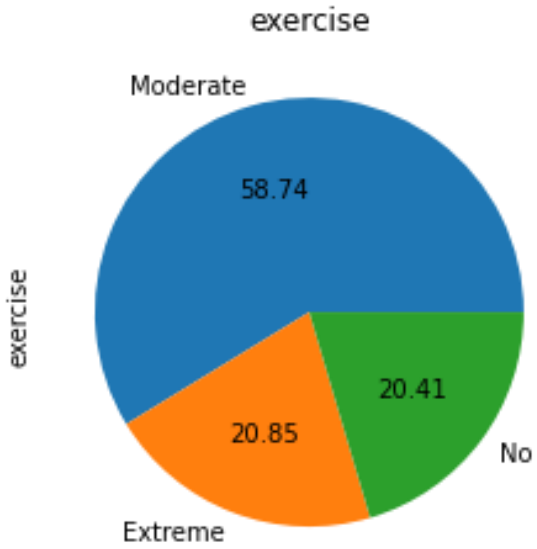
Only 2.0% of people are underweight, 18.09% of people are normal weight, 26.71% of people are overweight. Most of the people in the given dataset are obese i.e. 53.2% of people.

Figure 7. Univariate Analysis of Adventure sports



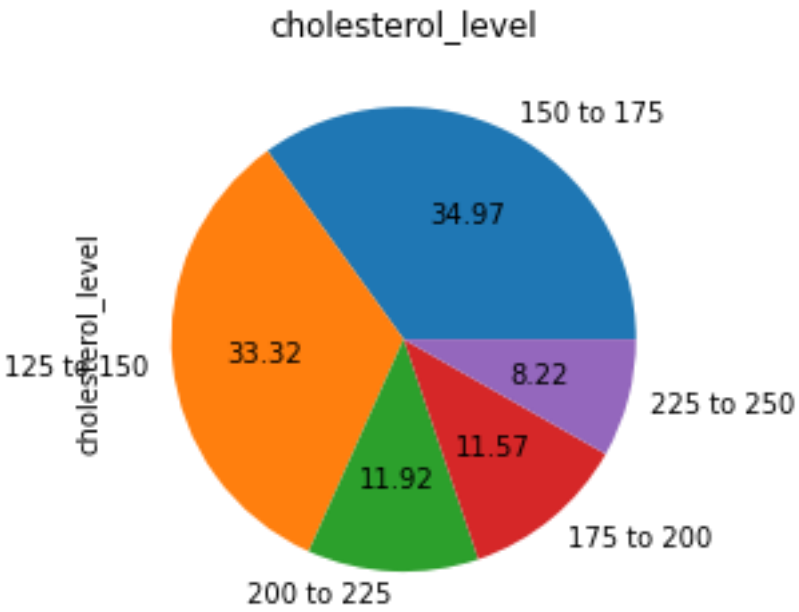
Only 8.19% of people are active in adventure sports. Remaining 91.81% are not active in adventure sports.

Figure 8. Univariate Analysis of Exercise



From the given dataset, only 20.41% of people are not active in doing exercise. 58.74% of people do moderate exercise and 20.85% of people do extreme exercise.

Figure 9. Univariate Analysis of Cholesterol level



From the given dataset,

i. only 33.32% of people have normal cholesterol level of 125 to 150.

ii. 46.54% of people have borderline high cholesterol level of 150 to 200.

iii. 14% of people have very high level of cholesterol level.

Figure 10. Univariate Analysis of History of heart disease

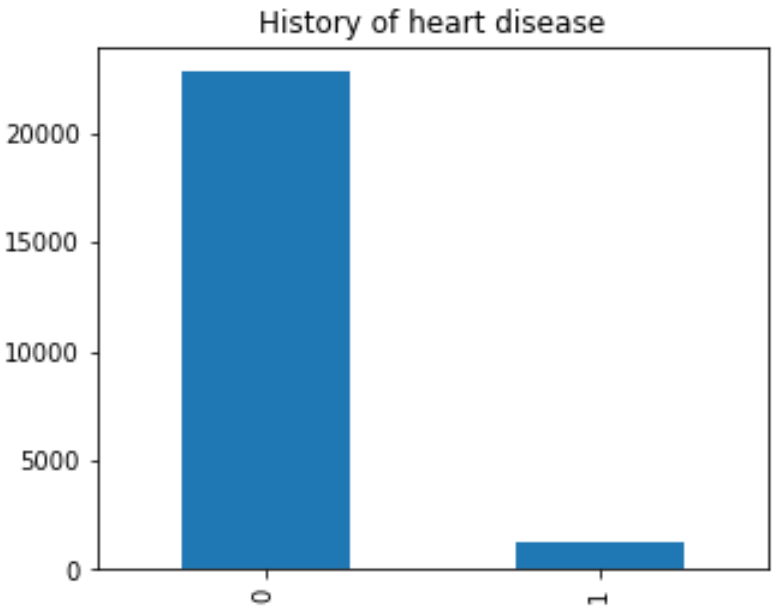


Figure 11. Univariate Analysis of other major disease other than heart disease

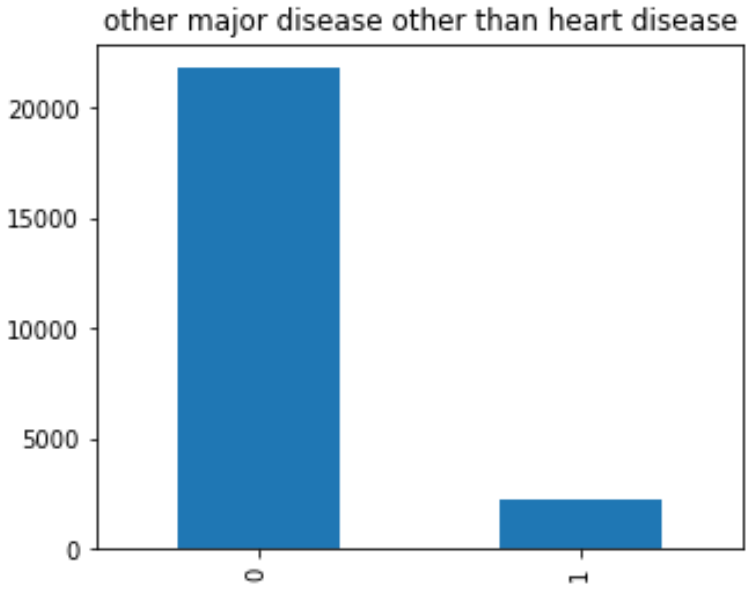
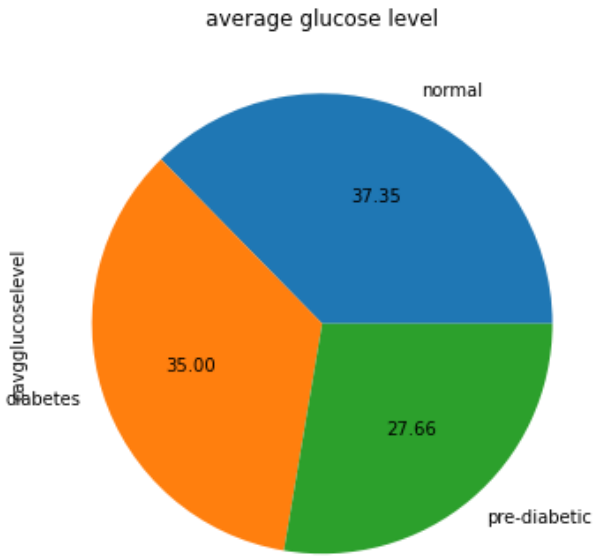


Figure 12. Univariate Analysis of other average glucose level



From the given dataset,

i. only 37.35% of people have normal average glucose level, i.e. less than 140.

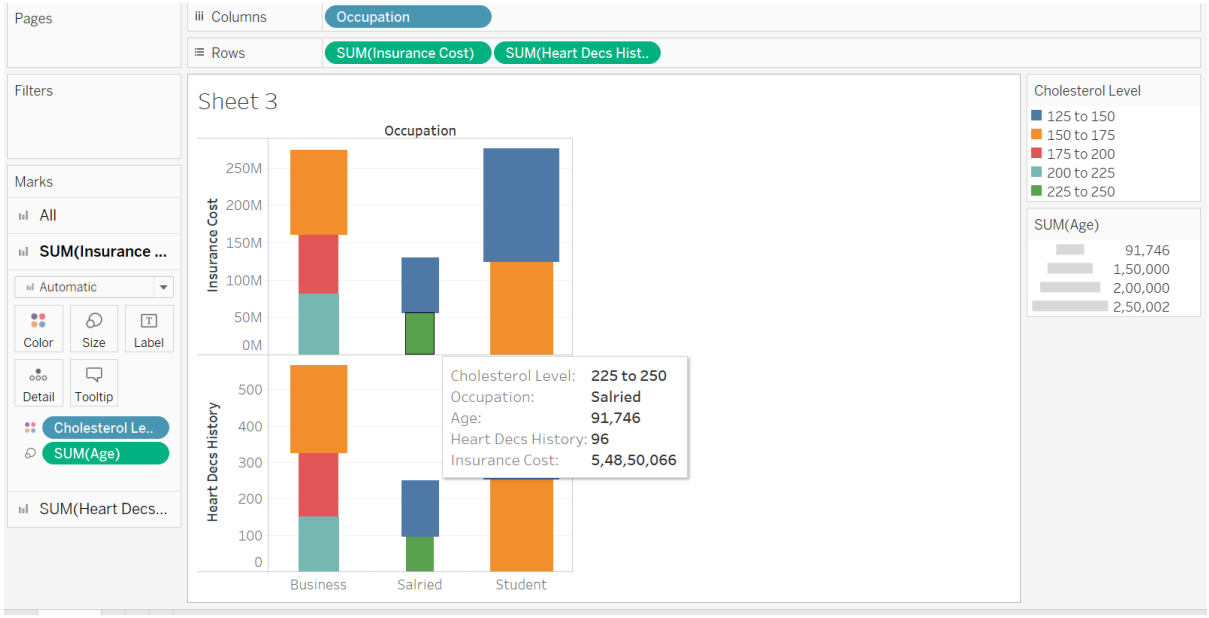
ii. 27.66% of people are pre-diabetic whose average glucose level range from 140 to 199.

iii. 35% of people are diabetic whose average glucose level are greater than 200.

2.4. BIVARIATE ANALYSIS

Below are the bivariate analysis of various features

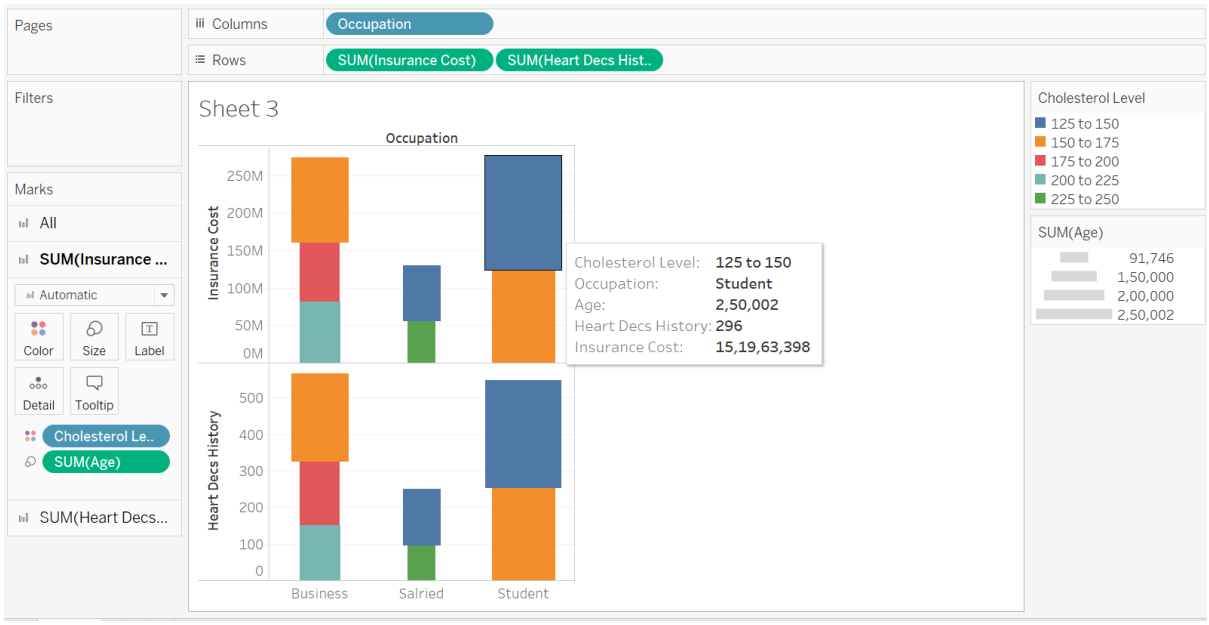
Figure 13. Occupation vs Sum of Insurance cost/ Sum of Heart Disease History



Insights from above graph,

- Young salaried people have less heart disease history and less amount of insurance cost claimed compared to Business and Student.
- On the contrary, young salaried people have very high level of cholesterol range (225 to 250). Next to that, Middle aged business people have high cholesterol level range

Figure 14. Occupation vs Sum of Insurance cost/ Sum of Heart Disease History



Insights from above graph,

- Interesting to note that, more old aged people belongs to student category and they have very high heart disease history and high amount of insurance cost claimed compared to Salaried.
- On the contrary, More old aged people belongs to Student category have normal level of cholesterol range (125 to 150) though they do have very high heart disease history.

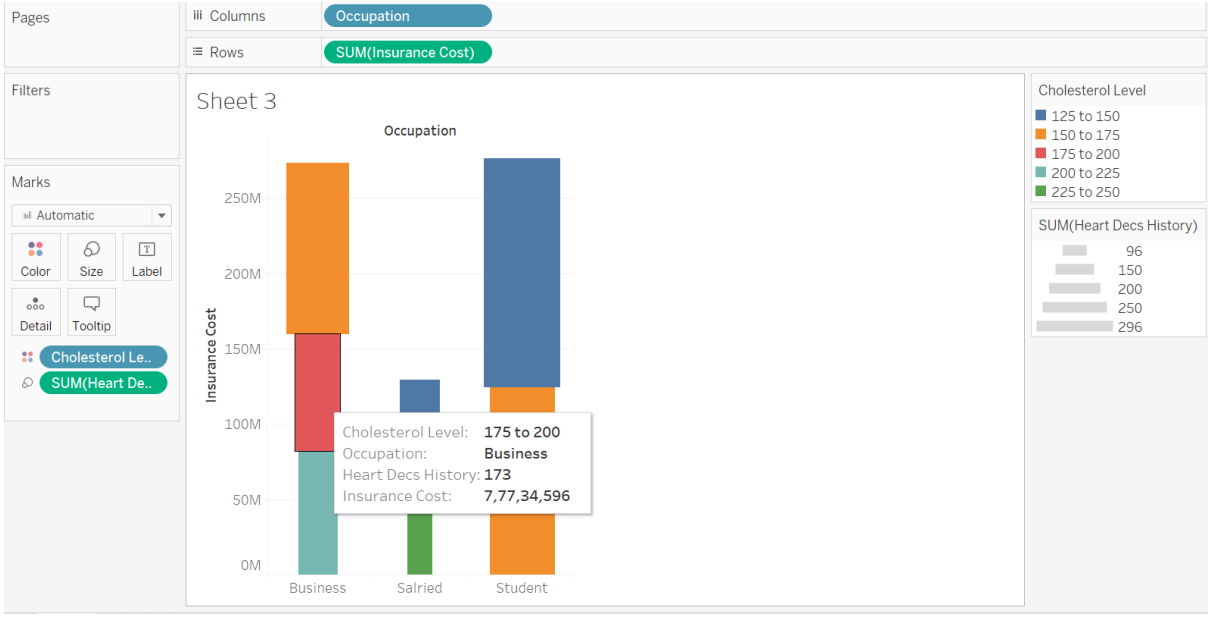


Figure 15. Insurance cost vs gender

From the below chart, no significant change in insurance cost based on gender.

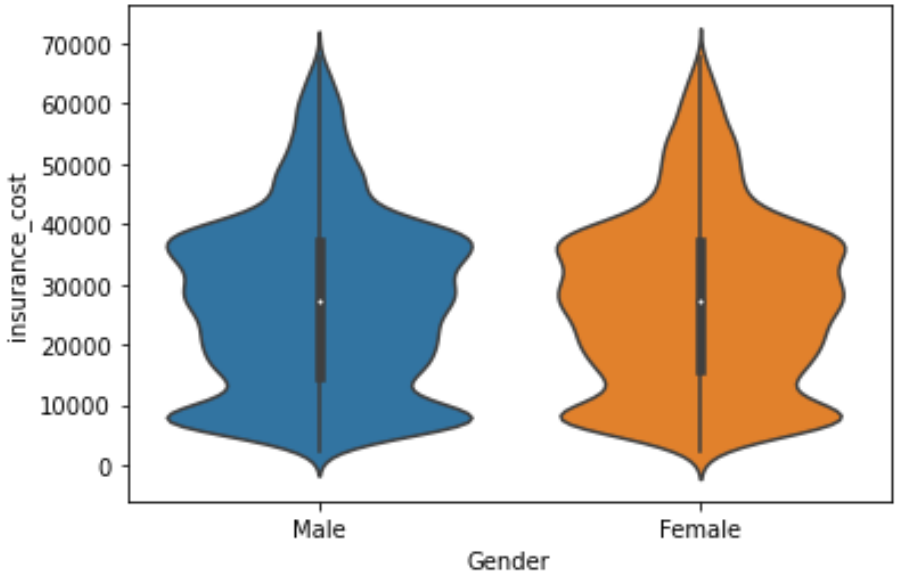


Figure 16. Insurance cost vs Alcohol

From the below chart, no significant change in insurance cost based on drinking alcohol.

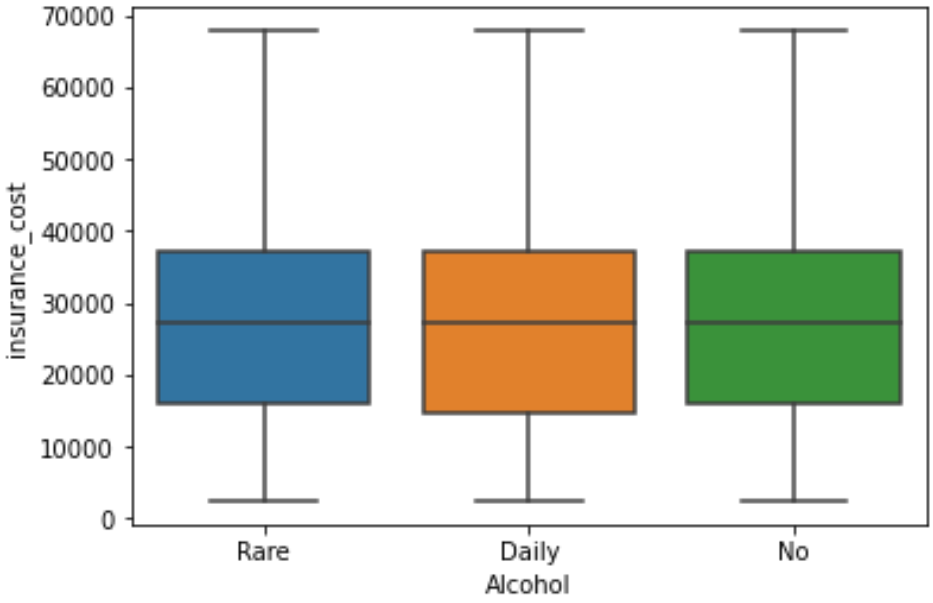


Figure 17. Insurance cost vs Adventure sports

From the below chart, the insurance cost is little higher for the ones who involve in adventure sports.

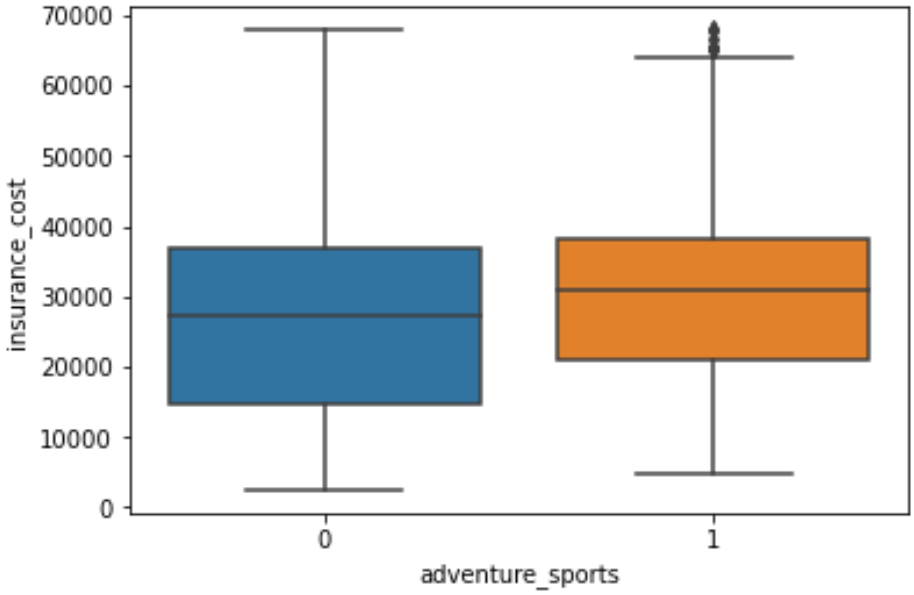


Figure 18. Insurance cost vs Years of insurance with us

From the below chart, there is no big difference in insurance cost for number of years of insurance with that insurance company.

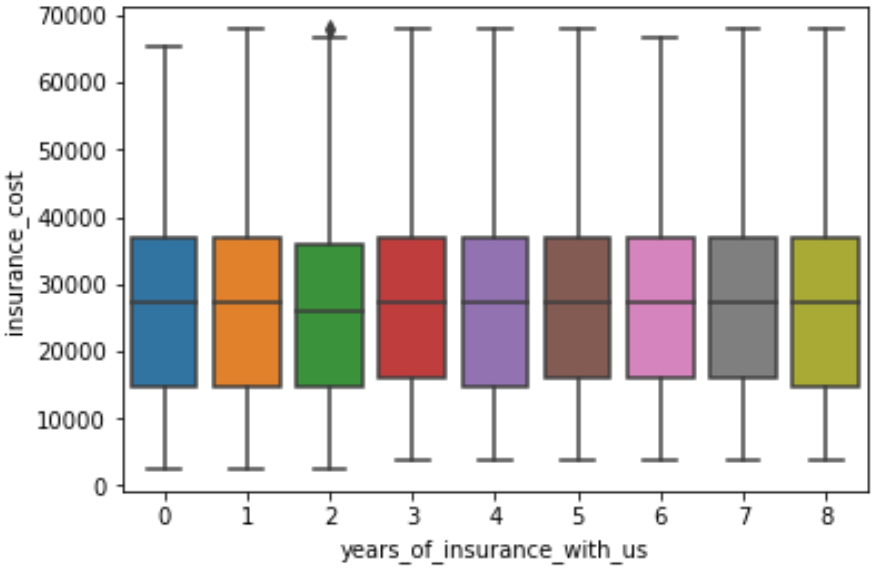


Figure 19. Insurance cost vs Regular checkup last year

From the below chart, there is a decrease in insurance cost when the number of regular checkups increases in last year

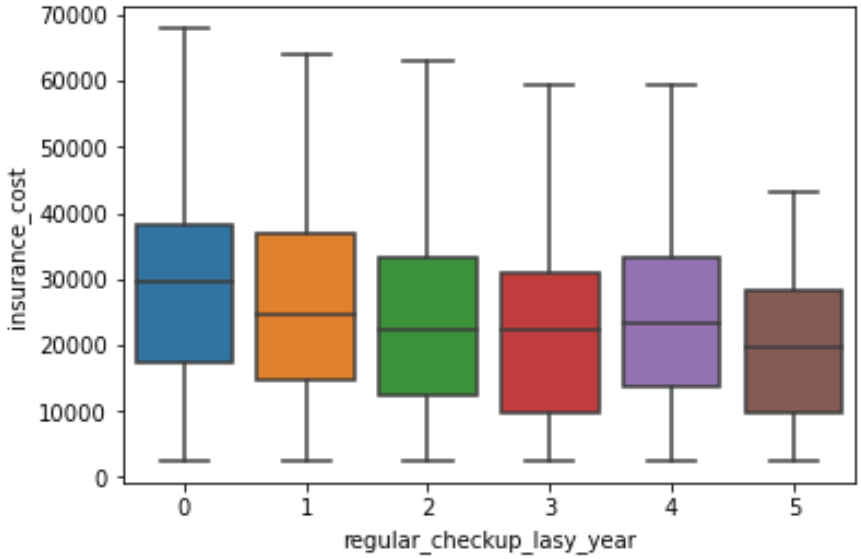


Figure 20. Insurance cost vs daily average steps

From the below chart, there is a increase in insurance cost when the number of average daily steps increases

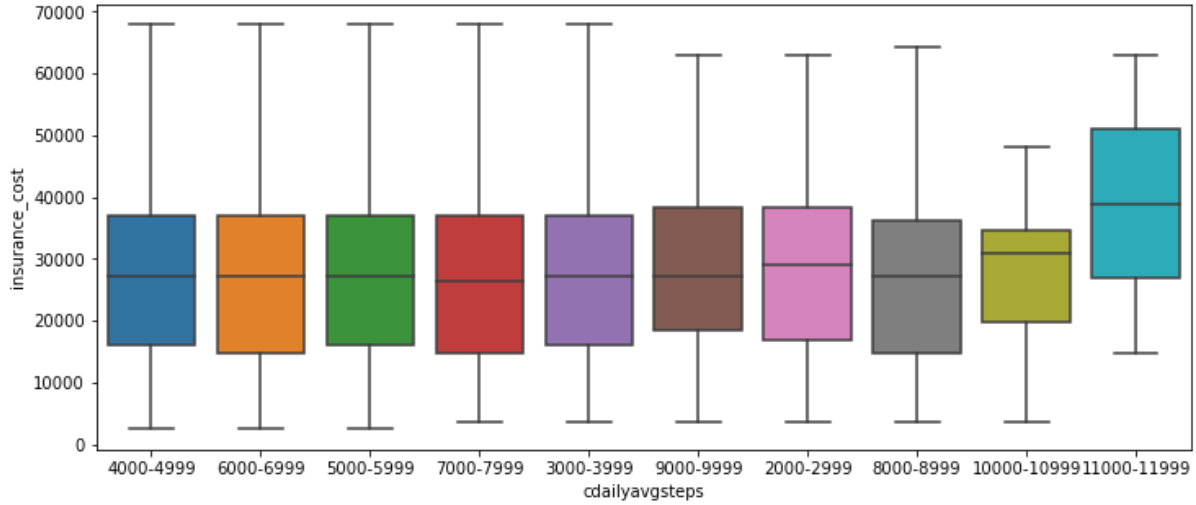


Figure 21. Insurance cost vs Last year admitted

From the below chart, recent admission in hospital decreases in insurance cost

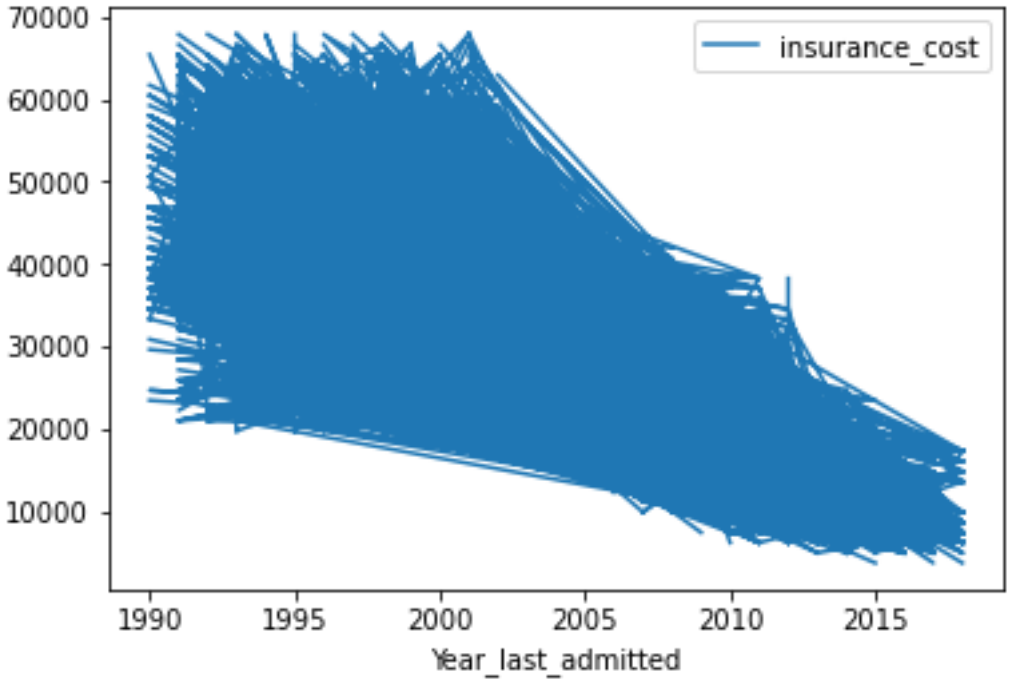


Figure 22. Insurance cost vs Weight change in last one year

From the below chart, higher weight change in last one year has decreased insurance cost

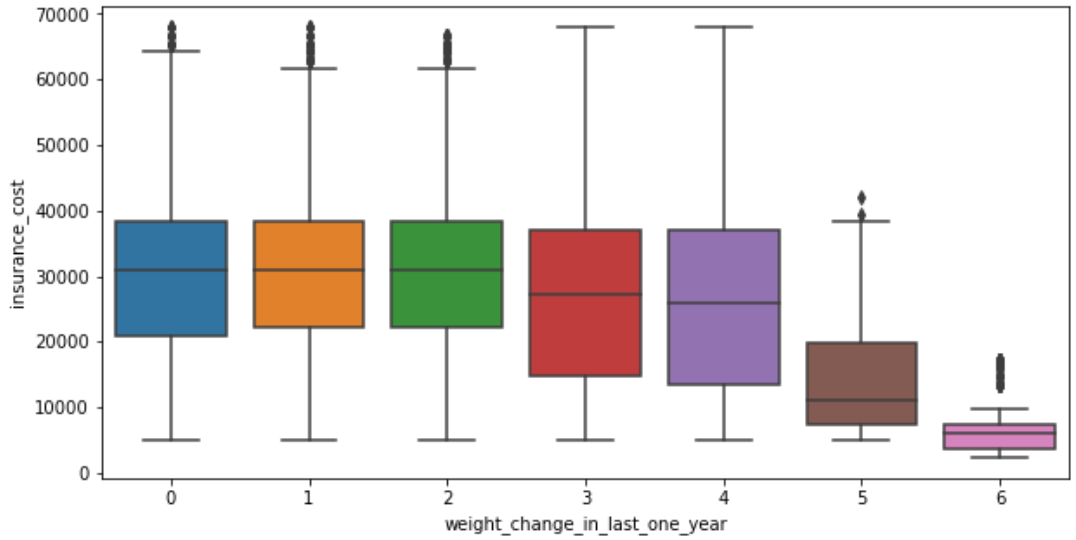
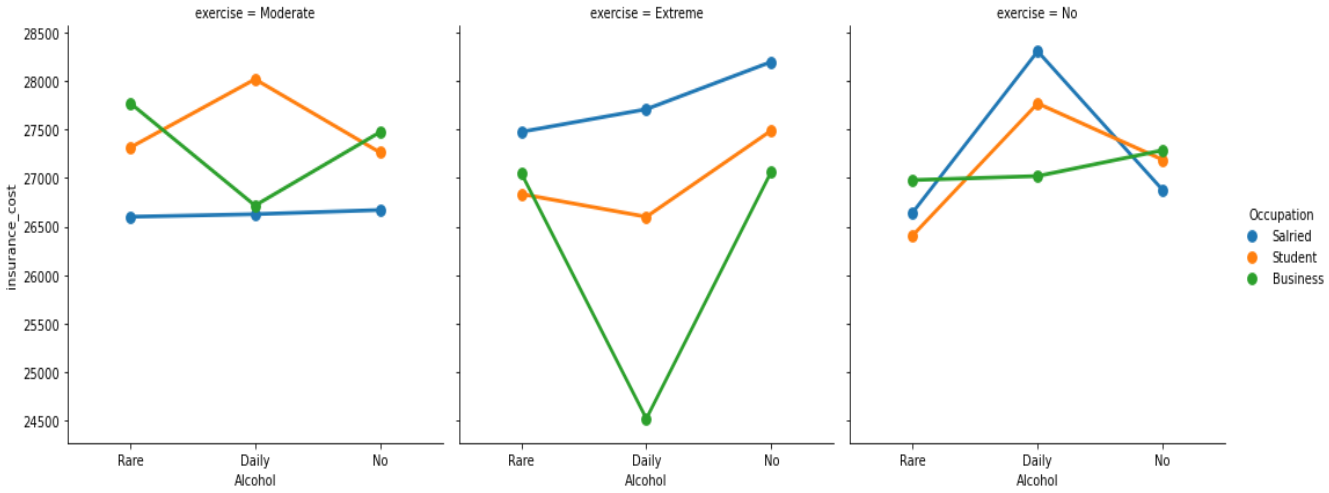


Figure 23. Insurance cost vs Exercise vs Alcohol vs Occupation



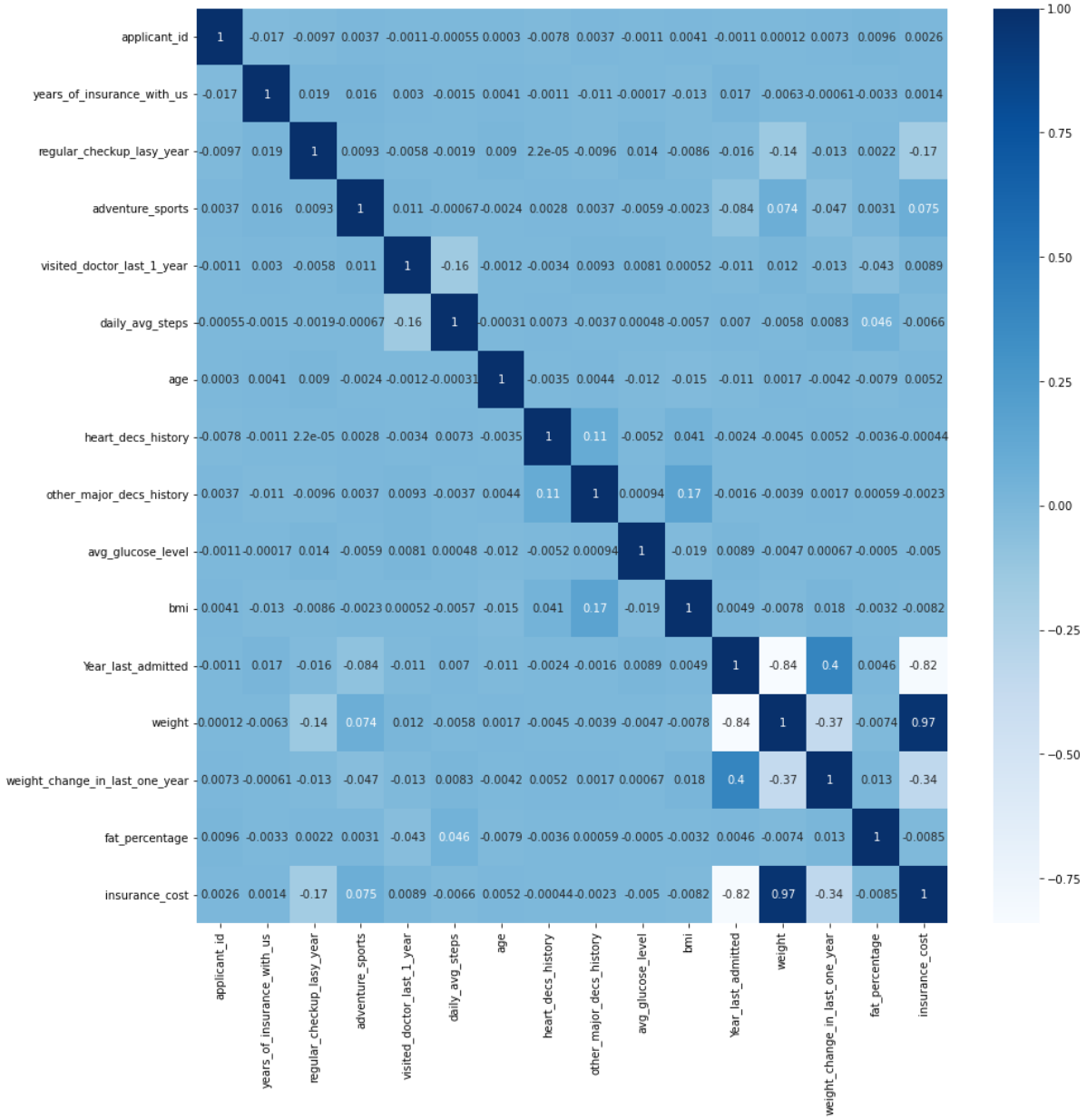
From the above chart,

Though whatever the occupation of people, who drink alcohol daily and do not exercise have very high insurance cost.

Business people who do extreme exercise even they drink alcohol daily have very less insurance cost. Interesting to note that, business people who do extreme exercise and never drink alcohol have very high insurance cost.

People who do moderate exercise and drink alcohol daily have lesser insurance cost than who never drink or drink rare.

Figure 24. Heatmap



It is significant from the graph that, for insurance cost, positive correlation is present with weight.

Last year admitted has positive correlation with weight change in last year

MISSING VALUE TREATMENT

For Missing value ‘Unknown’ in smoking status column, used knn imputer and found the value is closer to never smoked as from below computed value,

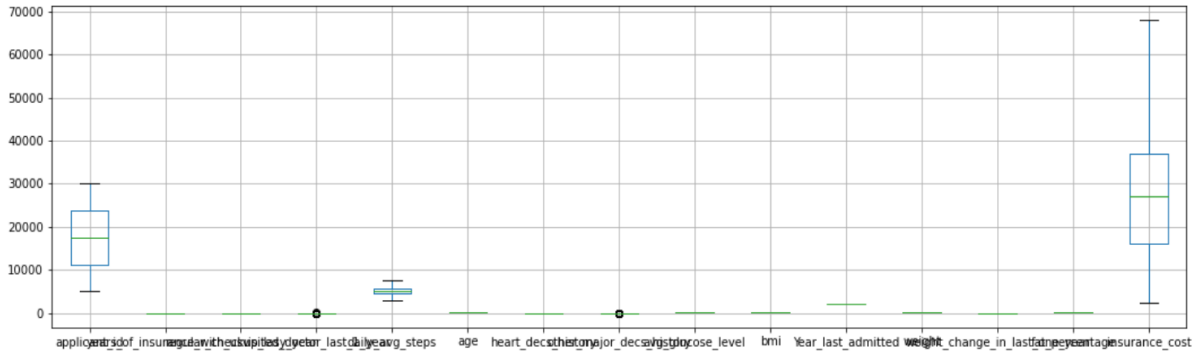
Knn computed value	value of smoking status	Count
1	never smoked	9249
0.973517	Unknown	7555
0	formerly smoked	4329
2	smokes	3867

Year admitted has 40% of missing data, it is nearly half of the data. We might lose some insights behind them. So, we replace them with mean since it doesn’t have outliers in it.

OUTLIER TREATMENT:

After treating outliers, outliers present in daily average steps, bmi are removed.

Figure 25. Boxplot of numerical variables



4. BUSINESS INSIGHTS FROM EDA

From the below major insights,

- Young salaried people have less heart disease history and less amount of insurance cost claimed compared to Business and Student.
- On the contrary, young salaried people have very high level of cholesterol range (225 to 250). Next to that, Middle aged business people have high cholesterol level range
- Interesting to note that, more old aged people belongs to student category and they have very high heart disease history and high amount of insurance cost claimed compared to Salaried.

- On the contrary, More old aged people belongs to Student category have normal level of cholesterol range (125 to 150) though they do have very high heart disease history. Though whatever the occupation of people, who drink alcohol daily and do not exercise have very high insurance cost.
- Business people who do extreme exercise even they drink alcohol daily have very less insurance cost. Interesting to note that, business people who do extreme exercise and never drink alcohol have very high insurance cost.
- People who do moderate exercise and drink alcohol daily have lesser insurance cost than who never drink or drink rare.

From the above insights, the given dataset is bit irrelevant with themselves. Because of this Clustering could not be used to classify the regression type dataset to get a clear picture.

5. MODEL BUILDING AND INTERPRETATION

5.1. BUILD VARIOUS MODELS

Before deploying modelling on given dataset, one-hot encoding is done on categorical variables to make it ready for modelling.

Min max scaling is also done for other numerical variables to get features in the range of (0,1) to match it with encoded categorical variables.

After this, data is split into training and testing set in 75:25 ratio with random state 1

And we have built various models such as Linear regression, Decision tree regressor, Random Forest Regressor and ANN regressor on the training and test set.

5.2, 5.3. TEST PREDICTIVE MODEL AGAINST TEST SET USING VARIOUS PERFORMANCE METRICS AND INTERPRETATION

As below, we performed various predictive models on training set against test dataset where we have got good training score and testing score for Linear regression and ANN regressor though RMSE is not so low value. But looks like Decision Tree regressor, and Random Forest regressor are over-fitting.

Let's Grid Search to get the best parameters.

Figure 26. Comparison table for training and test set using RMSE and model score

VARIOUS MODELS	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3360.677137	3292.165547	0.944958	0.947287
Decision Tree Regressor	0	4245.153729	1	0.912353
Random Forest Regressor	1157.990575	3055.222262	0.993465	0.954602
ANN Regressor	3340.195451	3325.882384	0.945627	0.946202

6. MODEL TUNING AND BUSINESS IMPLICATION

6.1. TUNING

After using param grid values for grid search for Decision Tree regressor,

The best parameters are 'max_depth': 10, 'min_samples_leaf': 30, 'min_samples_split': 15

After using range of param grid values for grid search for Random Forest Regressor,

The best parameters are 'max_depth': 10, ‘'max_features': 4,'min_samples_leaf': 30, 'min_samples_split': 30, ‘n_estimators':300

Figure 27. Comparison table of Training and Test dataset using RMSE and model score after tuning the parameters

VARIOUS MODELS	Train RMSE	Test RMSE	Training Score	Test Score
Linear Regression	3360.677137	3292.165547	0.944958	0.947287
Decision Tree Regressor	2802.939121	3107.456426	0.961712	0.953036
Random Forest Regressor	6086.060224	6342.336294	0.819485	0.804363
ANN Regressor	3340.195451	3325.882384	0.945627	0.946202

After tuning the parameters of grid search of Decision tree regressor and Random Forest regressor, the training score and test score values are improved better.

6.2. INTERPOLATION OF THE MOST OPTIMUM MODEL AND ITS IMPLICATION ON THE BUSINESS

From the above tuned model’s results, Decision tree regressor provides the most optimum model for providing medical insurance premium cost. Next to that, Linear regression model gives better results for the given dataset.

Optimal Insurance cost is predicted as such as below,

Figure 28. Comparison table for Insurance cost with Optimal Insurance cost after model deployment

Insurance cost	Optimal insurance cost after model deployment
6170	7261.466093
51828	50461.5968

From the above result, for low insurance cost the premium is increased compared to high insurance cost claimed. This health insurance sector is prone to claims and it is always under tremendous pressure. In recent times, Insurance Regulatory Development Authority has taken bold step by increasing the premium rate of health insurance products. This will help in the growth of this sector.

With better technological expertise coming in from the foreign partners and involvement by the IRDA the health insurance sector in India must turn around and start to earn profit.

6.3. INSIGHTS OF ANALYSIS

- Symmetric - age, years of insurance with us, average glucose level,year_last_admitted, weight change in last one year, fat_percentage, insurance cost Right skewed - Regular

checkup last year,adventure sports,visited doctor last 1 year,daily avg steps,heart_decs_history,other_major_decs_history, bmi

- Distribution looks like multimodal distribution.The insurance cost ranges from 2468 to 67870.00. Skew is 0.33 which means symmetrically distributed.
- Overall the highest insurance cost is for obese followed by over-weight. Lowest insurance cost for under weight people
- Highest cost for people's age between 30 to 69. Surprisingly for old age people 70-79 has the lowest insurance cost. Teenagers have very low insurance cost claim.
- People who has less daily average stepw count 2000-2999 has the lowest insurance cost
- The highest insurance cost collected in Bhubaneshwar followed by Mangalore and Bangalore The lowest insurance cost collected in Surat
- Highest insurance cost claimed by Student category. This is because due to any diseases or accident happended. Next to that, Business category people claimed high insurance cost.
- Slaried professional claimed very lowest insurance cost compared to other categories.
- Below are the feature importance ranking, thus the least ranking could be removed. The highest ranking provides most significant contribution to the target variable.

Specifications	Score
weight	2269.24233
weight_change_in_last_one_year	903.804244
regular_checkup_lasy_year	537.928831
covered_by_any_other_company_Y	489.007973
adventure_sports	228.884753

Specifications	Score
visited_doctor_last_1_year	3.888938
Year_last_admitted	8.962069
fat_percentage	7.948025

- Young salaried people have less heart disease history and less amount of insurance cost claimed compared to Business and Student.
- On the contrary, young salaried people have very high level of cholesterol range (225 to 250). Next to that, Middle aged business people have high cholesterol level range
- Interesting to note that, more old aged people belongs to student category and they have very high heart disease history and high amount of insurance cost claimed compared to Salaried.
- On the contrary, More old aged people belongs to Student category have normal level of cholesterol range (125 to 150) though they do have very high heart disease history.
- Though whatever the occupation of people, who drink alcohol daily and do not exercise have very high insurance cost.
- Business people who do extreme exercise even they drink alcohol daily have very less insurance cost. Interesting to note that, business people who do extreme exercise and never drink alcohol have very high insurance cost.
- People who do moderate exercise and drink alcohol daily have lesser insurance cost than who never drink or drink rare.
- It is significant from the heatmap that, for insurance cost, positive correlation is present with weight.
- Last year admitted has positive correlation with weight change in last year

6.4. RECOMMEDATIONS

- We recommend that health insurance company deploys the decision tree classifier model with training score(0. 961712, RMSE-2802.939121), testing score(0.953036. RMSE-3107)
- The insurance company provide their services through any mobile application or through their website, they can enter the input value and the output for the same will be optimal insurance cost could be incurred.
- Person's weight, weight_change_in_last_one_year, regular_checkup_lasy_year, covered_by_any_other_company_Y , adventure_sports can have impact on insurance cost and give better insurance cost prediction