CAPSTONE PROJECT

INSURANCE COST-HEALTHCARE PROJECT

PROJECT NOTES 1

—

VIDYA V

—

PGPDSBA.O.2023B

CONTENTS

LIST OF FIGURES

INTRODUCTION OF BUSINESS PROBLEM

Need for the study:

Health Insurance is one of the fields of focus in recent times. With the rise in diseases and treatment costs, more and more people are inclined towards securing a health insurance policy that covers all expected and unexpected medical costs. Insurance companies, thus need a meticulous evaluation of various parameters to determine the premium, so that the risk is rightly assessed and priced appropriately. Incorrect or inaccurate predictions might lead to losses for the companies and insufficient coverage might lead to client dissatisfaction.

Problem Statement:

The dataset here contains various lifestyle such as the amount of exercise, smoking habits etc., and medical parameters like heart diseases history and weight etc., and competitive parameters like whether or not they have been covered by other insurance companies and the target variable is the insurance cost, determined by the predictors.

Objective/ Business Opportunity:

- Parametric Evaluation:
    - To identify and understand the correlation between the various predictors, and the target variable, and to estimate the significance of the predictors on the target variable.
- Risk Assessment:
    - To explore the parameters and compare them against the insurance cost, and identify potential areas of risk, if any, and to develop strategies to avert the same
- Cost Prediction:
    - To build prediction models and optimize their performance by tuning
    - Identify the best model for the given dataset by means of accuracy scores and RMSE scores

Steps done at the current stage:

- Exploring the given dataset, understand the various predictor variables and their nature

- Performing necessary cleaning and treatment to make the data optimal for analysis
- Performing Exploratory Data Analysis by breaking down the dataset into Uni, Bi and Multivariate combinations and seeking insights from observations
- Exploring the data by segmentation, identify segments and relate them to insurance cost

Future Steps:

- Building various predictive models and evaluating their performance against the train and the test sets
- Tuning the models in order to achieve optimum performance
- Comparing and evaluating the performance of models on train and test sets by metrics like RMSE and accuracy
- Identifying the best model for the dataset and the problem statement
- Fitting the model to the test data and determining the insurance costs for the same

Software Used/ Tools Used:

- Jupyter notebook- Python Kernel
- Numpy version 1.24.4
- Pandas Version 1.4.4
- Seaborn Version 0.13.0
- Matplotlib version 3.5.2

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

- This dataset contains 25000 applicants' information from ID 5000-24999 along with their insurance costs
- There are 24 variables in the dataset, including the target variable
- There are no time/date related variables in the given dataset
- The variables of the dataset can be classified as below:
    - Personal demographic information:
        - Age, Gender, Location, Occupation, Applicant ID
    - Biomedical indices:
        - Weight, Cholesterol Level, Fat Percentage, BMI, Average glucose level
    - Lifestyle Parameters:
        - Exercise, smoking and alcohol habits, Walking related counts, Whether involved in Adventurous sports
    - Medical History:
        - Heart disease history, Other major disease history, number of regular checkups and doctor visits in the past year, year in which the person was last admitted, weight change in the past year
    - Insurance company related variables:
        - Number of years of insurance with us, Whether covered by other insurance companies
    - Target Variable:
        - Insurance cost
- Thus, a comprehensive coverage of most of the parameters that could be involved are captured

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

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

Fig.1 Dataset Info

	count	mean	std	min	25%	50%	75%	max
applicant_id	25000.0	17499.50	7217.02	5000.0	11249.75	17499.5	23749.25	29999.0
years_of_insurance_with_us	25000.0	4.09	2.61	0.0	2.00	4.0	6.00	8.0
regular_checkup_last_year	25000.0	0.77	1.20	0.0	0.00	0.0	1.00	5.0
adventure_sports	25000.0	0.08	0.27	0.0	0.00	0.0	0.00	1.0
visited_doctor_last_1_year	25000.0	3.10	1.14	0.0	2.00	3.0	4.00	12.0
daily_avg_steps	25000.0	5215.89	1053.18	2034.0	4543.00	5089.0	5730.00	11255.0
age	25000.0	44.92	16.11	16.0	31.00	45.0	59.00	74.0
heart_decs_history	25000.0	0.05	0.23	0.0	0.00	0.0	0.00	1.0
other_major_decs_history	25000.0	0.10	0.30	0.0	0.00	0.0	0.00	1.0
avg_glucose_level	25000.0	167.53	62.73	57.0	113.00	168.0	222.00	277.0
bmi	24010.0	31.39	7.88	12.3	26.10	30.5	35.60	100.6
Year_last_admitted	13119.0	2003.89	7.58	1990.0	1997.00	2004.0	2010.00	2018.0
weight	25000.0	71.61	9.33	52.0	64.00	72.0	78.00	96.0
weight_change_in_last_one_year	25000.0	2.52	1.69	0.0	1.00	3.0	4.00	6.0
fat_percentage	25000.0	28.81	8.63	11.0	21.00	31.0	36.00	42.0
insurance_cost	25000.0	27147.41	14323.69	2468.0	16042.00	27148.0	37020.00	67870.0

Fig.2. Data Description- Numerical Variables

	count	unique	top	freq
Occupation	25000	3	Student	10169
cholesterol_level	25000	5	150 to 175	8763
Gender	25000	2	Male	16422
smoking_status	25000	4	never smoked	9249
Location	25000	15	Bangalore	1742
covered_by_any_other_company	25000	2	N	17418
Alcohol	25000	3	Rare	13752
exercise	25000	3	Moderate	14638

Fig.3. Data Description- Categorical Variables

```
applicant_id                      0.000000
years_of_insurance_with_us       -0.075217
regular_checkup_last_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                               1.056428
Year_last_admitted                0.013532
weight                            0.109077
weight_change_in_last_one_year    0.068026
fat_percentage                   -0.363262
insurance_cost                    0.331650
dtype: float64
```

Fig.4. Skewness of numerical variables

Observations:

- The shape of the dataset is (25000,24)
- There is one ID variable, one target variable(insurance_cost) and 22 predictor variables
- Of these 2 are of float type, 14 are of int type and 8 are of object types
- 'year_last_admitted' and 'bmi' fields have missing values
- There are no duplicate records
- The five-point summary of the numerical variables are available in Fig.2
- The target variable is insurance cost, having values ranging from 2468 to 67870, with the median value being 27148
- There are only 54 unique values of the 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
```

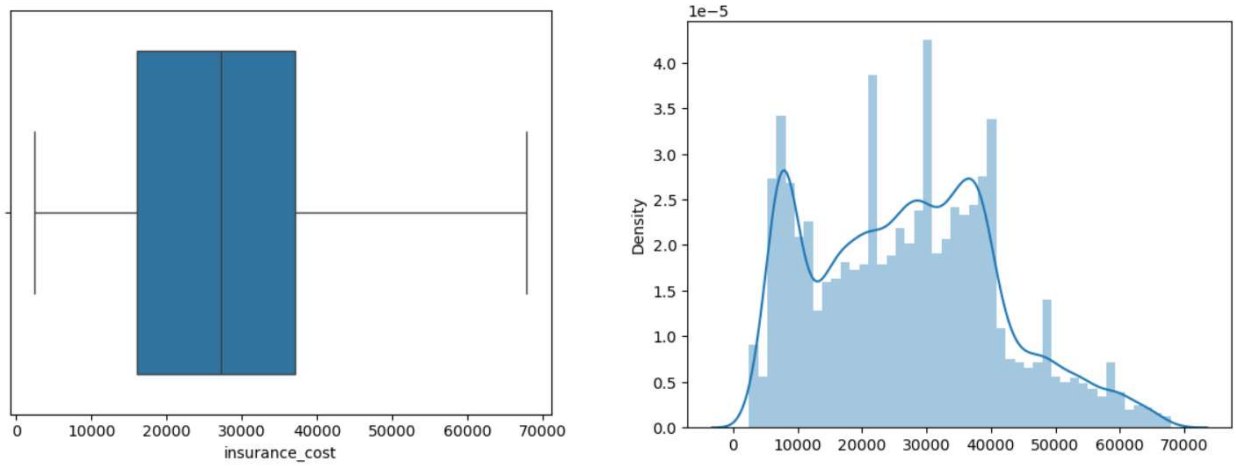


Fig.5. Inspection of target variable

Understanding of attributes (Variable info, renaming if required):

- Among the variables, the following is the categorization for the purpose of analysis:
  - Numerical Variables:
    - Age, Weight, daily_avg_steps, avg_glucose_level, fat_percentage, insurance_cost
  - Binary Categorical Variables:
    - Encoded as 0 or 1: Adventure_sports, covered_by_any_other_company,heart_decs_history, other_major_decs_history
    - Object type: Gender
  - Categorical Variables:
    - Numerically encoded: 'years_of_insurance_with_us', 'regular_checkup_lasy_year', 'visited_doctor_last_1_year', 'weight_change_in_last_one_year'
    - Object type Variables: 'Occupation', 'cholesterol_level', 'smoking_status', Location', 'Alcohol', 'exercise'
- Among these, one of the columns has a typo, and has been renamed from ‘regular_checkup_lasy_year’ to ‘regular_checkup_last_year’

EXPLORATORY DATA ANALYSIS

Removal of unwanted variables:

- The ID variable in this dataset is 'applicant_id'
- It has 25000 continuous values ranging from 5000 to 24999
- As this variable is a unique ID, it was useful in determining that there were no duplicates in the given dataset
- It was also used to ensure that there were no gaps in the data
- However, this variable is no longer needed for further exploration as it does not have any kind of correlation with any of the other variables.
- Hence, this variable is removed and is not a part of further analysis
- Thus, after this step, the dataset shape is (25000,23)

Missing Value Treatment:

Observations:

- From Fig.1, we find that there are missing values in the fields 'year_last_admitted' and 'bmi'
- In the 'year_last_admitted' field has 11881 null values, which is about 47.5% of the size of the field
- The bmi field has 990 missing values, which is about 3.96 the size of the field
- A visual representation of the missing parts of the dataset is given in Fig 1.6.

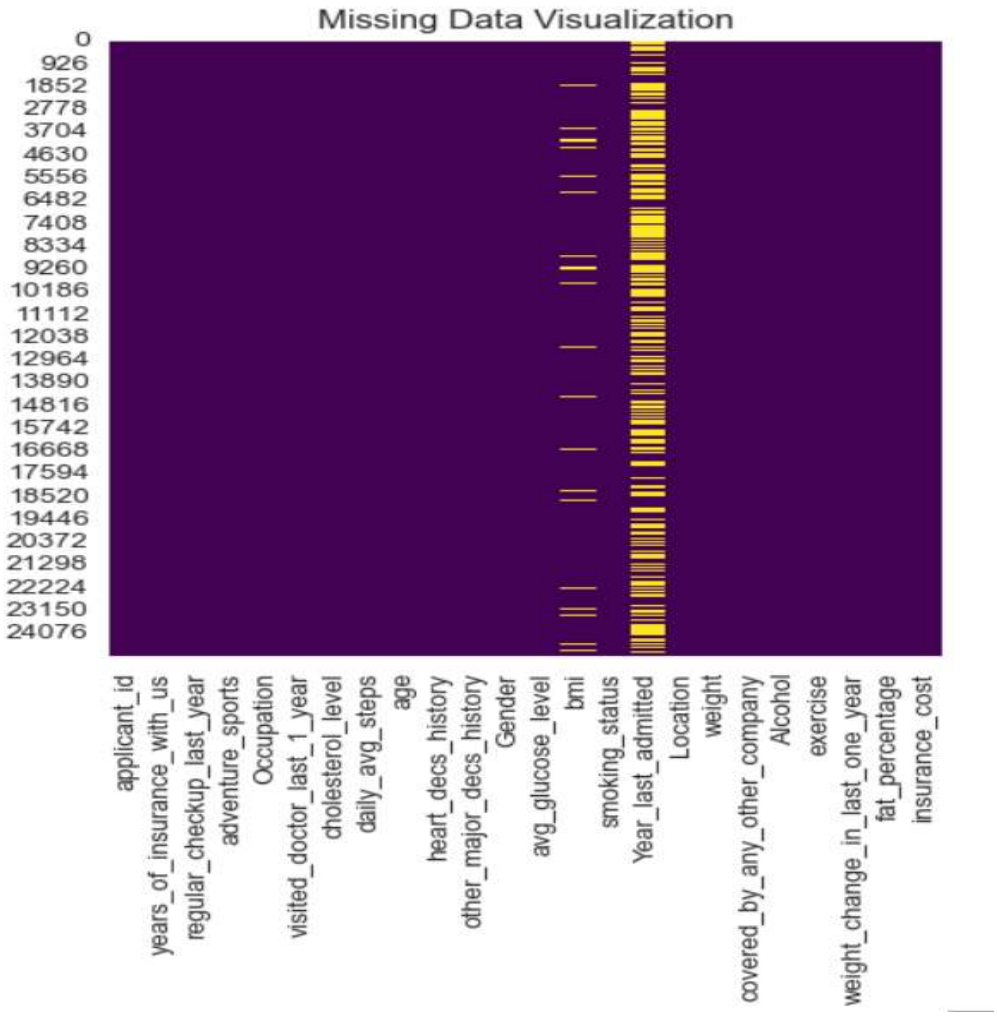


Fig.6. Missing Data Visualization

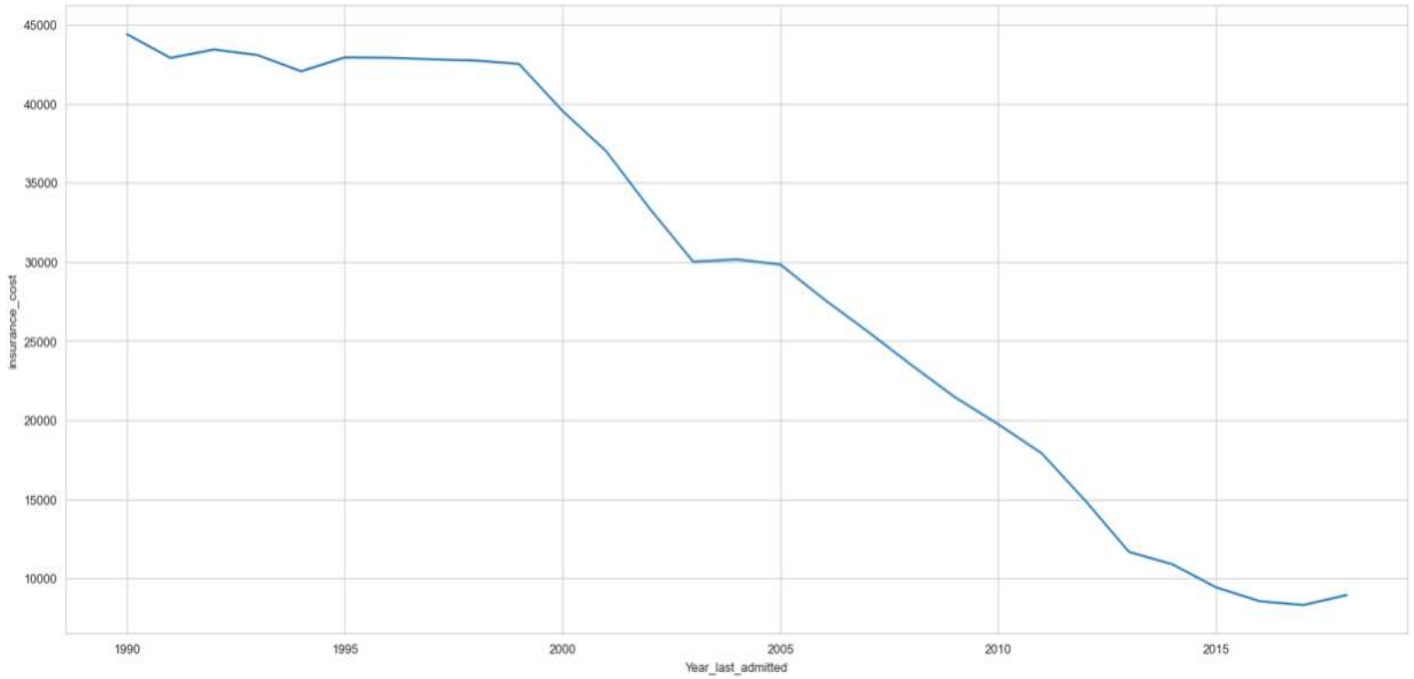


Fig.7. Year_last_admitted vs insurance_cost

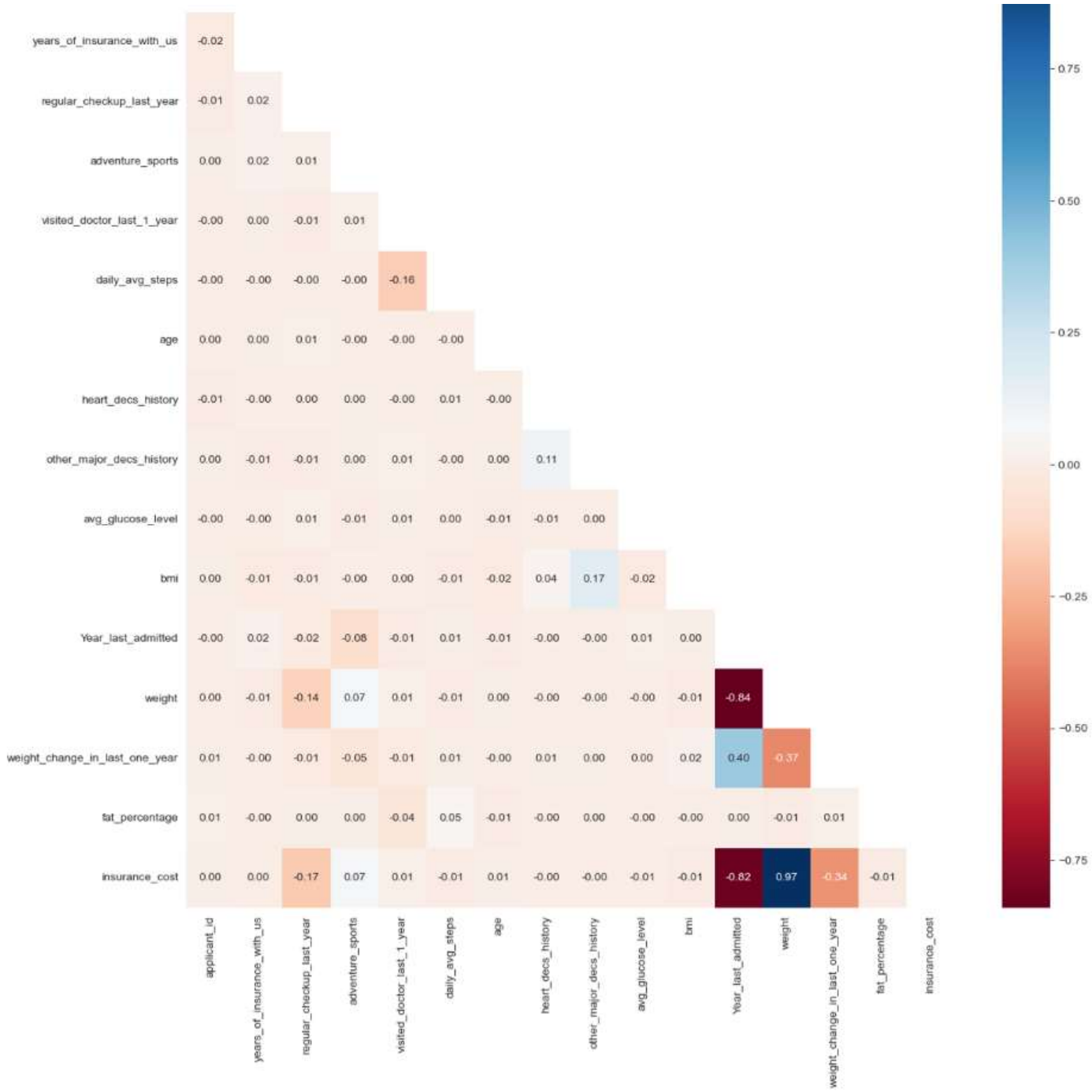


Fig.8. Correlation Heatmap of Dataset

Missing Value Treatment:

- 47.5 % is a very high percentage, and in a business scenario, efforts have to be made to obtain the actual data
- However, in this case, a decision has to be made whether to drop this variable or impute missing values
- The field in general has a high correlation with the target variable and a couple of predictor variables

- A t-test was done to ascertain the significance, and resulted in a p-value<0.05, which indicated that the variable is significant
- Hence, instead of making the data entirely synthetic, a better option would be to drop the column entirely
- Also, the 'bmi' field has very few missing values (3.96%). Thus, here as well, the better option would be to just drop the rows that contain null values in the field instead of imputing
- Hence, dropping rows with null values in this field and dropping the field' year_last_admitted' entirely will result in 8.13% data loss, which is a small percentage
- Hence, after the missing value treatment by dropping the field 'year_last_admitted' and dropping the rows that contained null values in the 'bmi' field, the shape of the dataset is (24010,22)

Bad Value Correction:

- The field 'Occupation' has 3 unique values- 'Salried', 'Student' and 'Business'
- As observed, the 'Salried' value has been misspelt and has been changed to 'Salaried'

Addition of New Variables:

New categorical features have been created for the purpose of EDA using the numerical variables as follows:

- 'avg_step_cat'- Based on 'daily_avg_steps'. The categories are as follows:
    - Sedentary- 'daily_avg_steps' less than 4500
    - Lightly active- 'daily_avg_steps' between 4500-6000
    - Active- 'daily_avg_steps' between 6000-10000
    - Very Active- 'daily_avg_steps' greater than 10000
- 'weight_cat'- Based on Weight.
    - Underweight: 'weight' less than 60
    - Normal weight: 'weight' between 60-70
    - Overweight: 'weight' between 70-80
    - Obese: 'weight' between 80-90
    - Very Obese: 'weight' greater than 90
- 'fat_percentage_cat': Based on fat_percentage
    - Low Body Fat: 'fat_percentage' less than 20
    - Moderate Body Fat: 'fat_percentage' between 20-30
    - High Body Fat: 'fat_percentage' between 30-40

    - Very High Body Fat: ‘fat_percentage’ greated than 40
- ‘insurance_cost_cat’: Based on ‘insurance_cost’
    - ‘1st Tier’: ‘insurance_cost’ less than 10000
    - ‘2nd Tier’: ‘insurance_cost’ between 10000-20000
    - ‘3rd Tier’: ‘insurance_cost’ between 20000-30000
    - ‘4th Tier’: ‘insurance_cost’ between 30000-40000
    - ‘5th Tier’: ‘insurance_cost’ between 40000-50000
    - ‘6th Tier’: ‘insurance_cost’ greater than 50000

Category Aggregation and Value Transformation:

- ‘visited_doctor_last_1_year’ has values ranging from 0-8
- Since the amount of data in each a category above 5 is very little, all values above 5 have been brought to 5.
- Also, the encoded categorical and binary categorical variables were treated as objects for the purpose of EDA

Univariate Analysis:

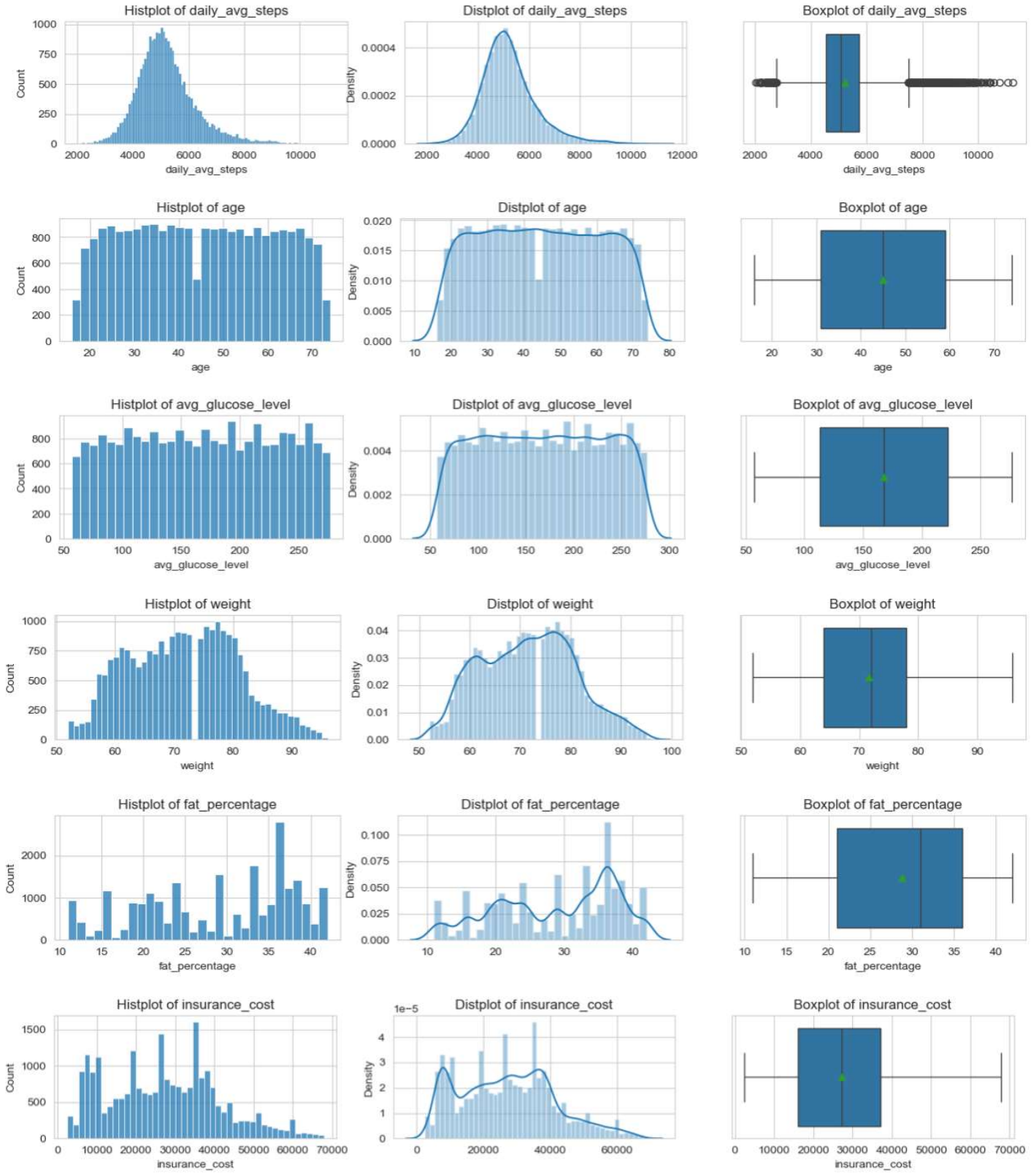
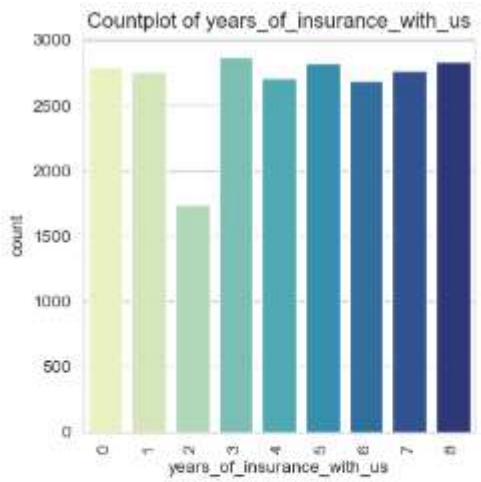


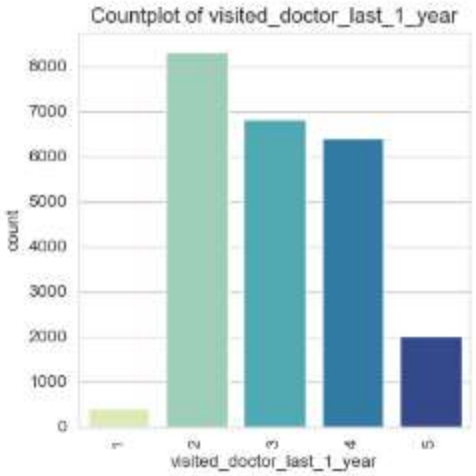
Fig.9 Univariate Analysis- Numerical Variables

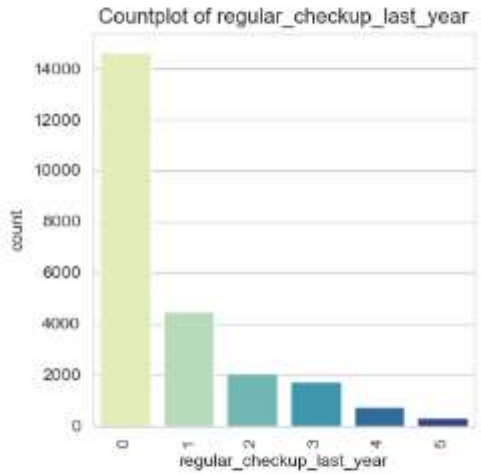
Observations:

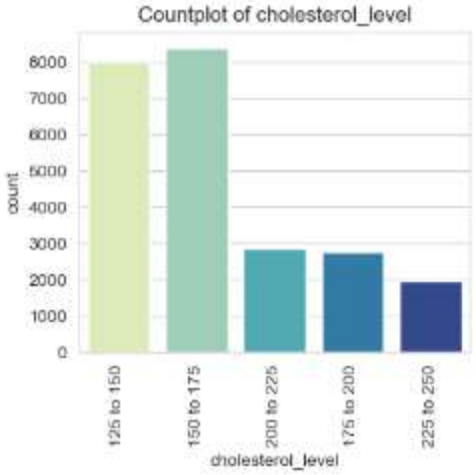
From Fig.9, the following can be observed:

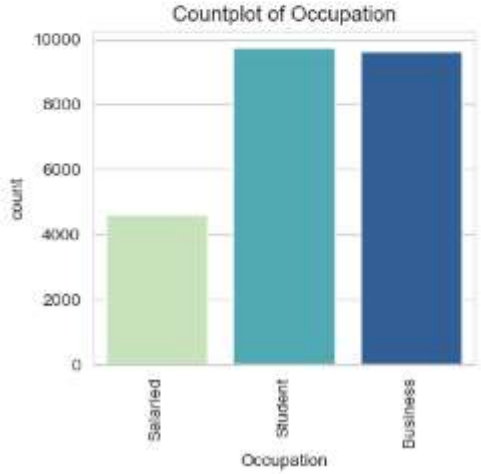
- Columns with outliers: daily_avg_steps
- age,avg_glucose_level have a uniform distribution
- weight variable has a multimodal distribution
- fat_percentage has 32 unique values ranging from 11-42. Hence the distribution is multimodal

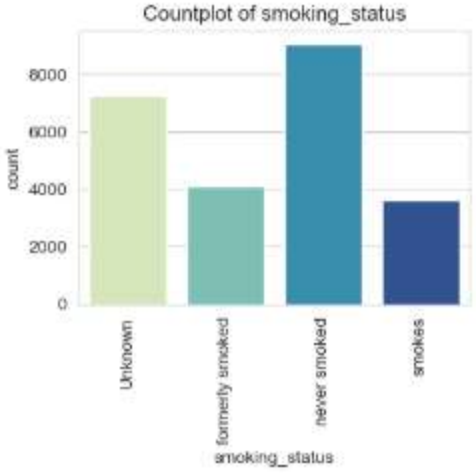

Countplot of years_of_insurance_with_us
count
years_of_insurance_with_us

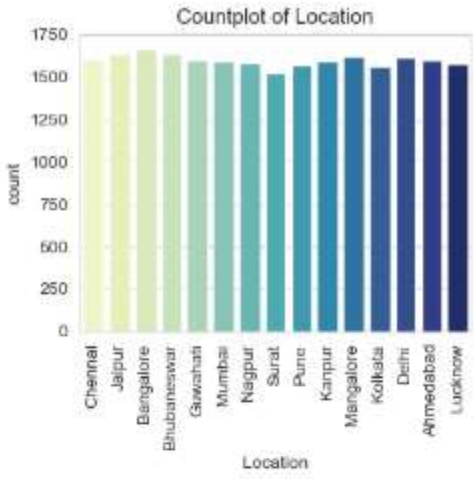

Countplot of visited_doctor_last_1_year
count
visited_doctor_last_1_year

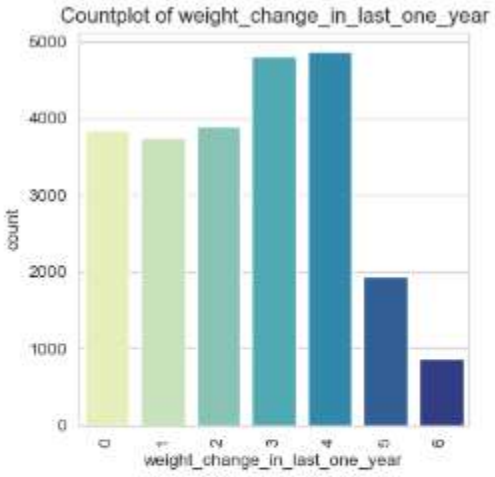

Countplot of regular_checkup_last_year
count
regular_checkup_last_year

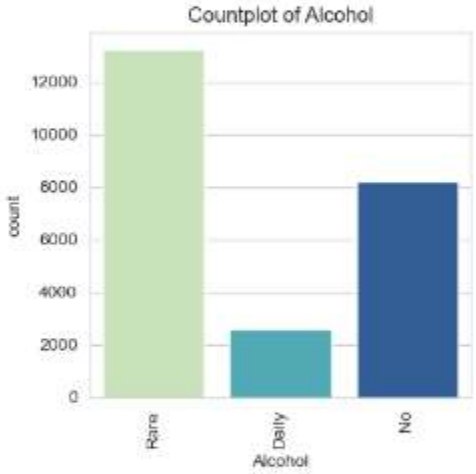

Countplot of cholesterol_level
count
125 to 150
150 to 175
200 to 225
175 to 200
225 to 250
cholesterol_level

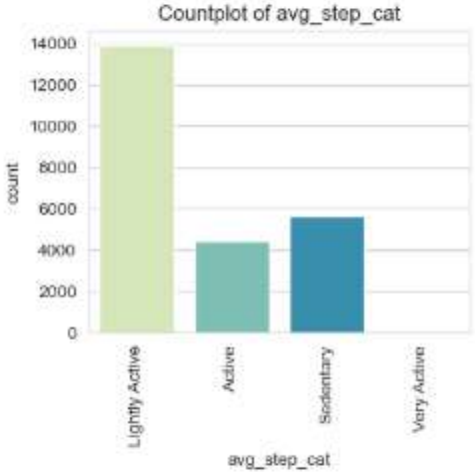

Countplot of Occupation
count
Salaried
Student
Business
Occupation

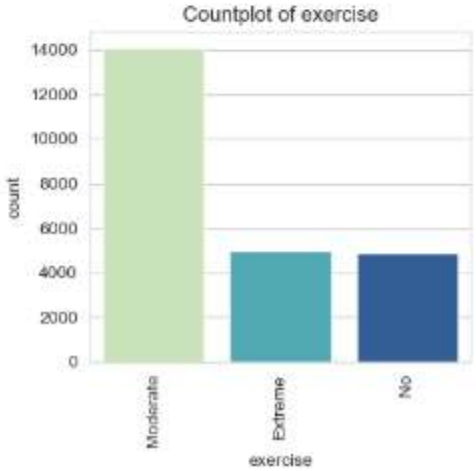

Countplot of smoking_status
count
Unknown
formerly smoked
never smoked
smokes
smoking_status

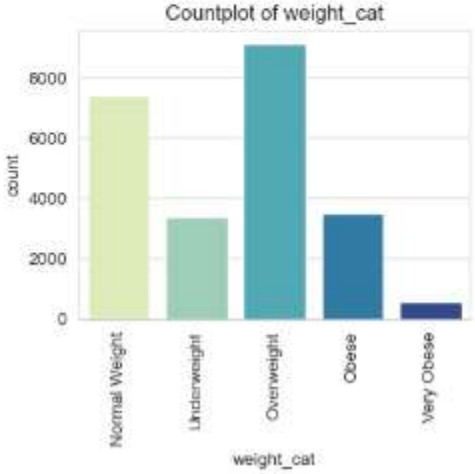

Countplot of Location
count
Location
Chennai
Jaipur
Bangalore
Bhubaneswar
Guwahati
Mumbai
Nagpur
Surat
Pune
Kanpur
Mangalore
Kolkata
Delhi
Ahmedabad
Lucknow


Countplot of weight_change_in_last_one_year
count
weight_change_in_last_one_year


Countplot of Alcohol
count
Alcohol
Rare
Daily
No


Countplot of avg_step_cat
count
avg_step_cat
Lightly Active
Active
Sedentary
Very Active


Countplot of exercise
count
exercise
Moderate
Extreme
No


Countplot of weight_cat
count
weight_cat
Normal Weight
Underweight
Overweight
Obese
Very Obese

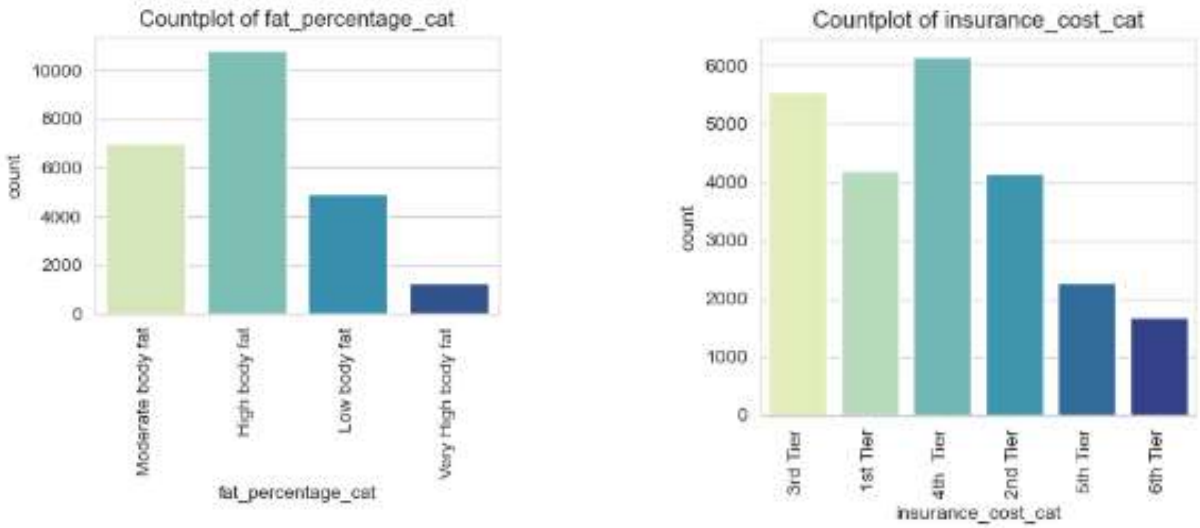


Fig.10. Univariate Analysis- Non-binary Categorical Variables

Observations:

- The count of customers having 2 years of insurance with us is lower than the rest
- Most of the people have not had any regular checkup last year
- There is almost an equal number of Students and Business professionals in the dataset. However, the count of Salaried professionals is almost half than that of the other two categories in Occupation field
- Most of the people seem to have visited the doctor 2-4 times last year
- Most of the people have either 125-150 or 150-175 cholesterol level. Those of them who have higher cholesterol levels are very low in count
- Most of the people in the dataset have reported to have never smoked. However, there is an Unknown category in the smoking_status field. For the purpose of this project, it is treated as Unknown. However, if there is a way to get this information, it would help in analysis
- There are 15 unique locations in the dataset, where people are from, and they are almost equally distributed amongst these locations.
- Most of the people have reported Rare or no alcohol consumption. Very few have reported daily alcohol consumption
- Most of the people have reported moderate exercise
- Most of the customers have reported a weight change of 3-4 in the past year
- Most of the customers are classified as lightly active, i.e. they take 4500-6000 steps everyday
- Most of the people are overweight, ie. have weight ranging from 70-80 . Very few are classified as very obese i.e. >90.
- Most of the people have High body fat.ie, body fat_percentage ranging from 30-40
- Based on the classification of insurance cost, most of the people belong to the 4th tier, i.e., 30000-40000

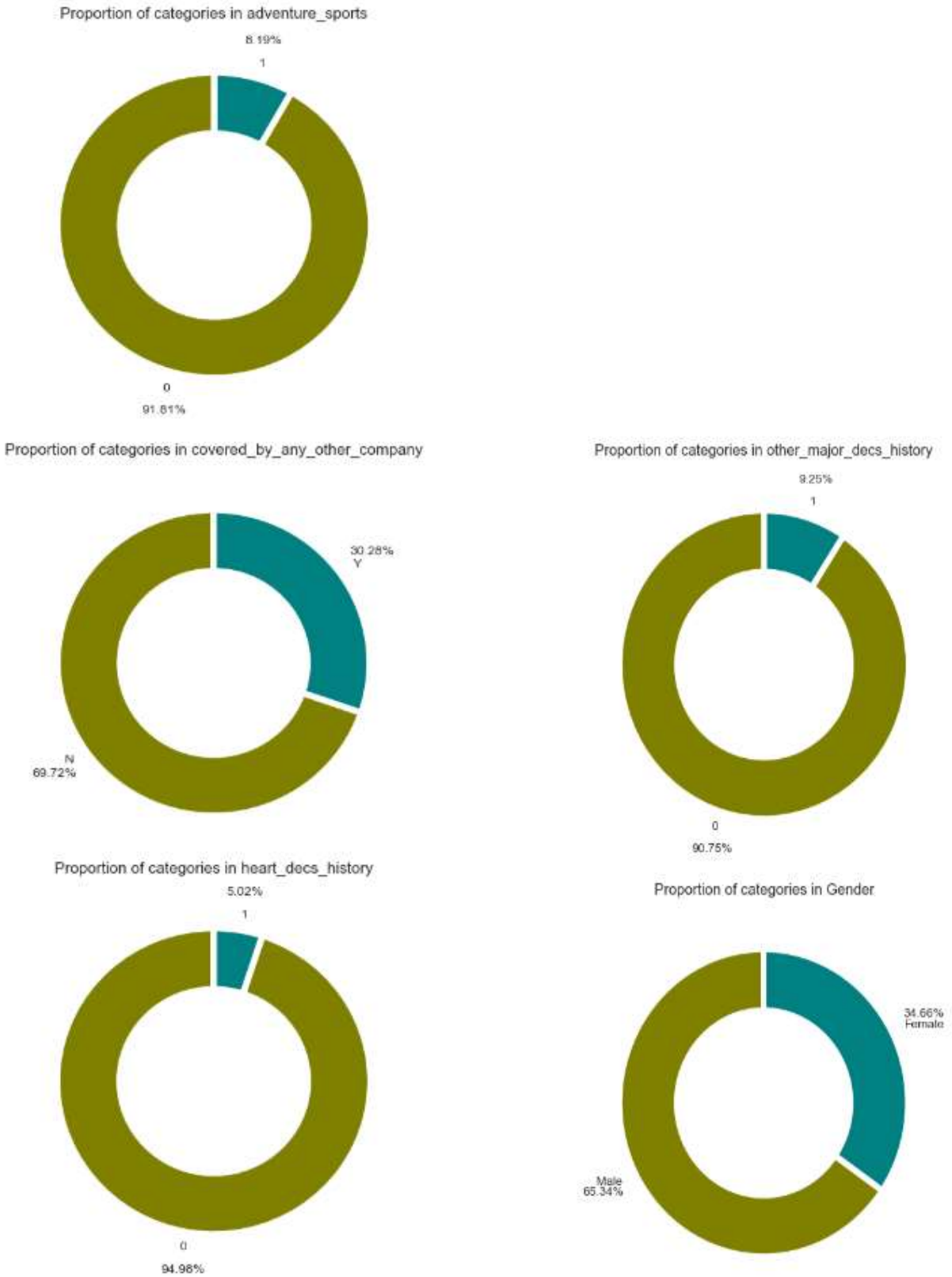


Fig.11. Univariate Analysis- Binary Categorical Variables

Observations:

- Among the categorical variables, Gender, adventure_sports, covered_by_other_company, heart_decs_history, other_decs_history, are binary in nature
- About 2/3rd of the people are males
- Less than 10% have had other major diseased, and less than 5% have had heart diseases
- About 8% of the people engage in adventure sports
- And about 30% are covered by other insurance companies

Bivariate Analysis:

Numerical variables vs numerical variables:

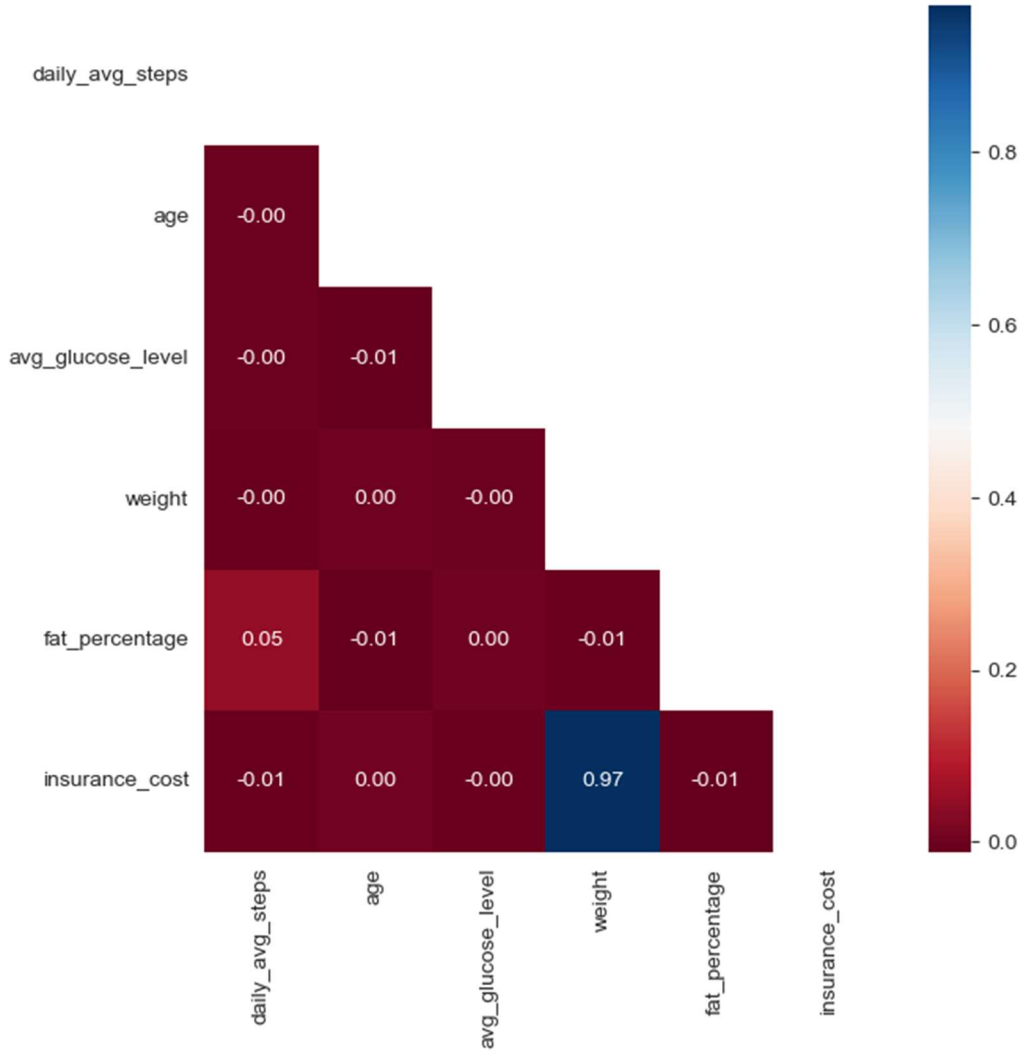


Fig.12. Correlation heatmap of numerical Variables

Of all the numerical variables, only the weight variable seems to have a high positive correlation with the target variable.

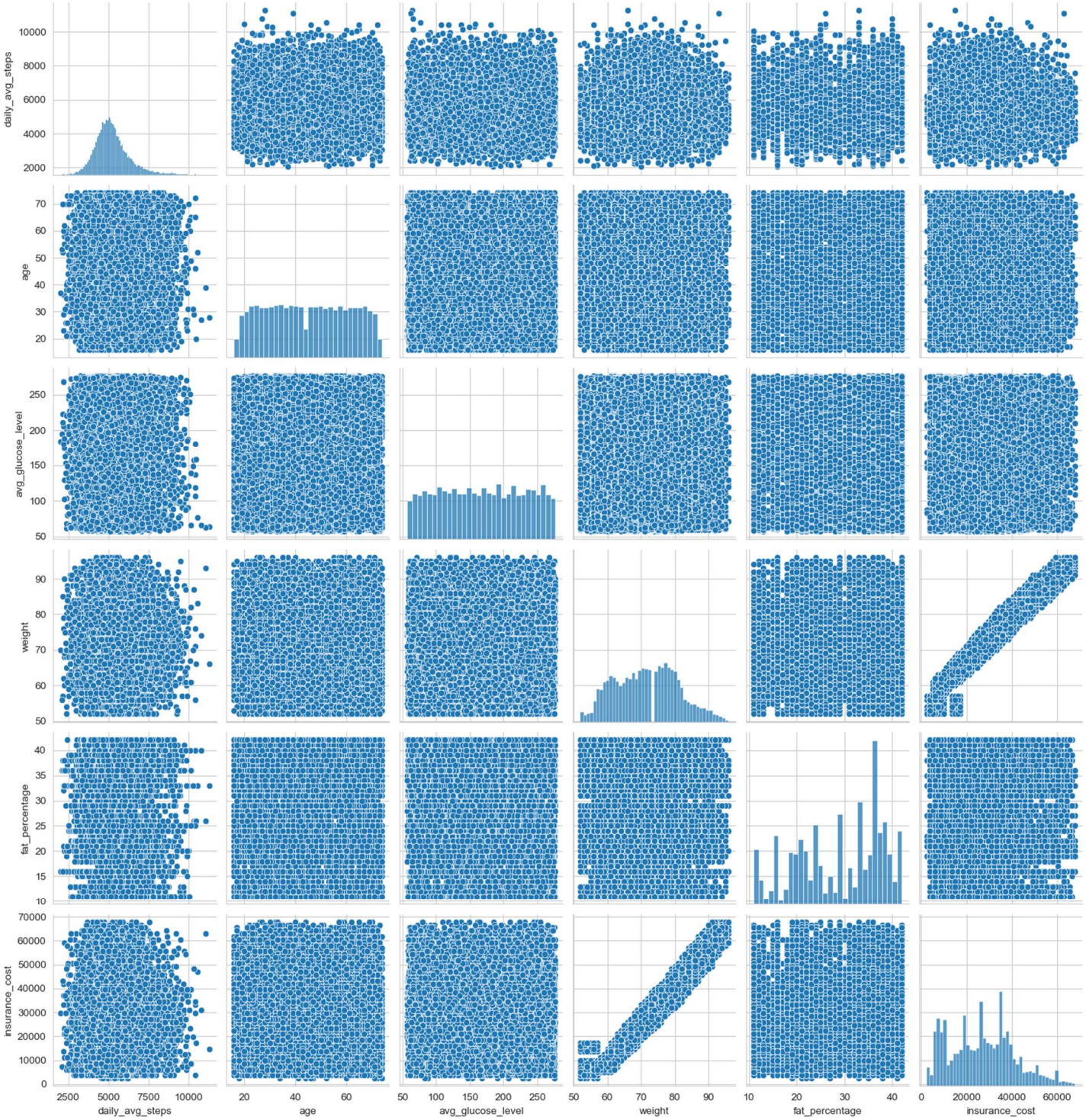


Fig.13. Pairplots of numerical variables

Observations:

- From the pairplots above, a linear relation can be observed only between weight and insurance cost
- No other relation between any of the variables in discernable

Categorical Variables variables vs target variable:

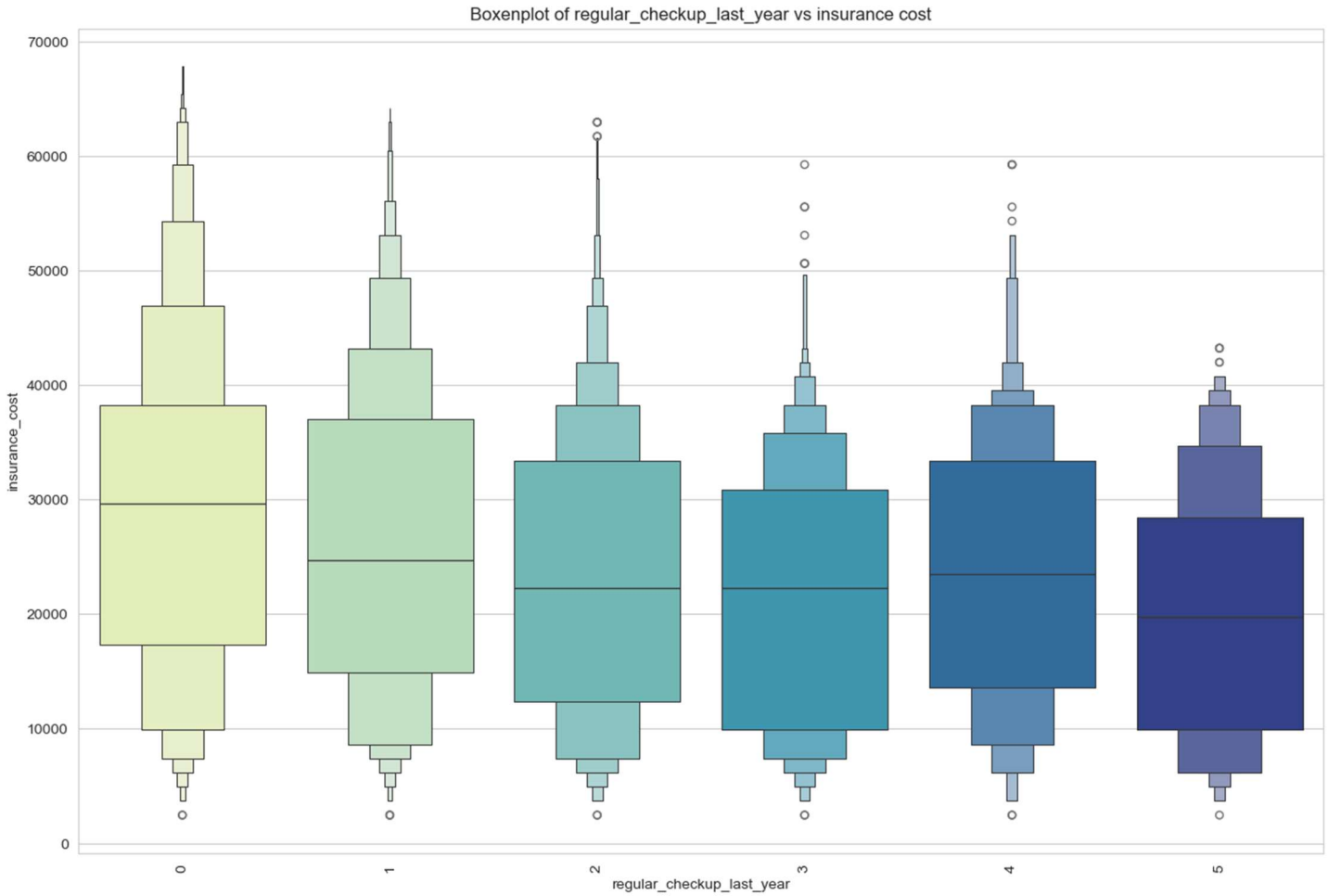


Fig.14 Boxenplot - regular_checkup_last_year vs insurance_cost

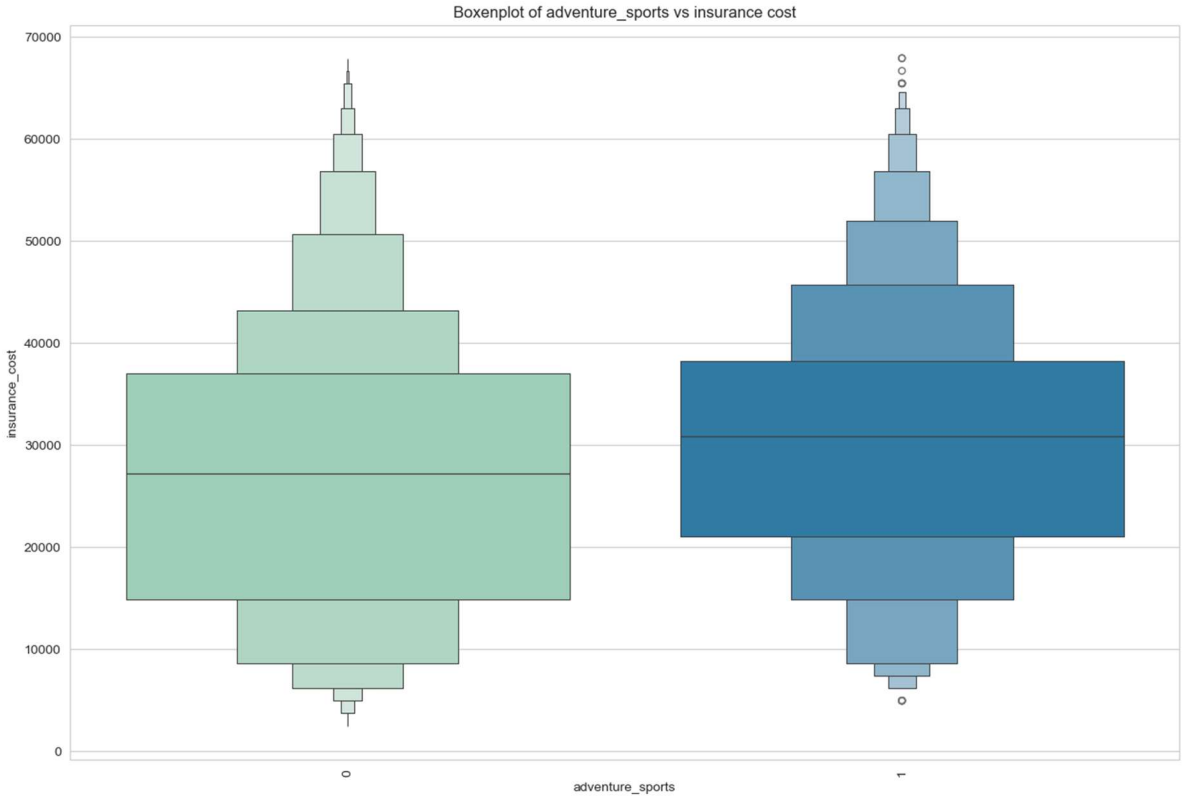


Fig.15 Boxenplot- adventure_sports vs insurance_cost

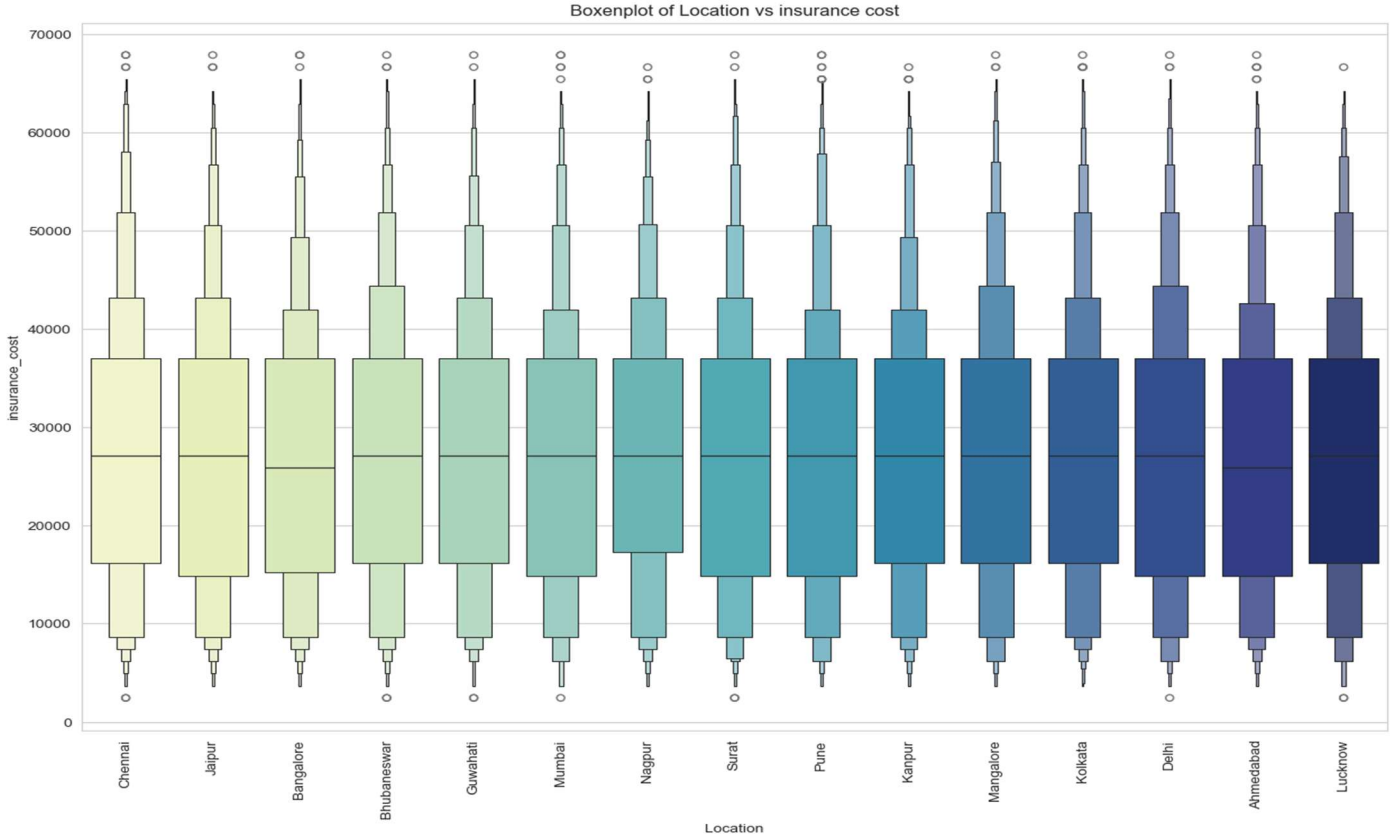


Fig.16. Boxenplot- Location vs insurance_cost

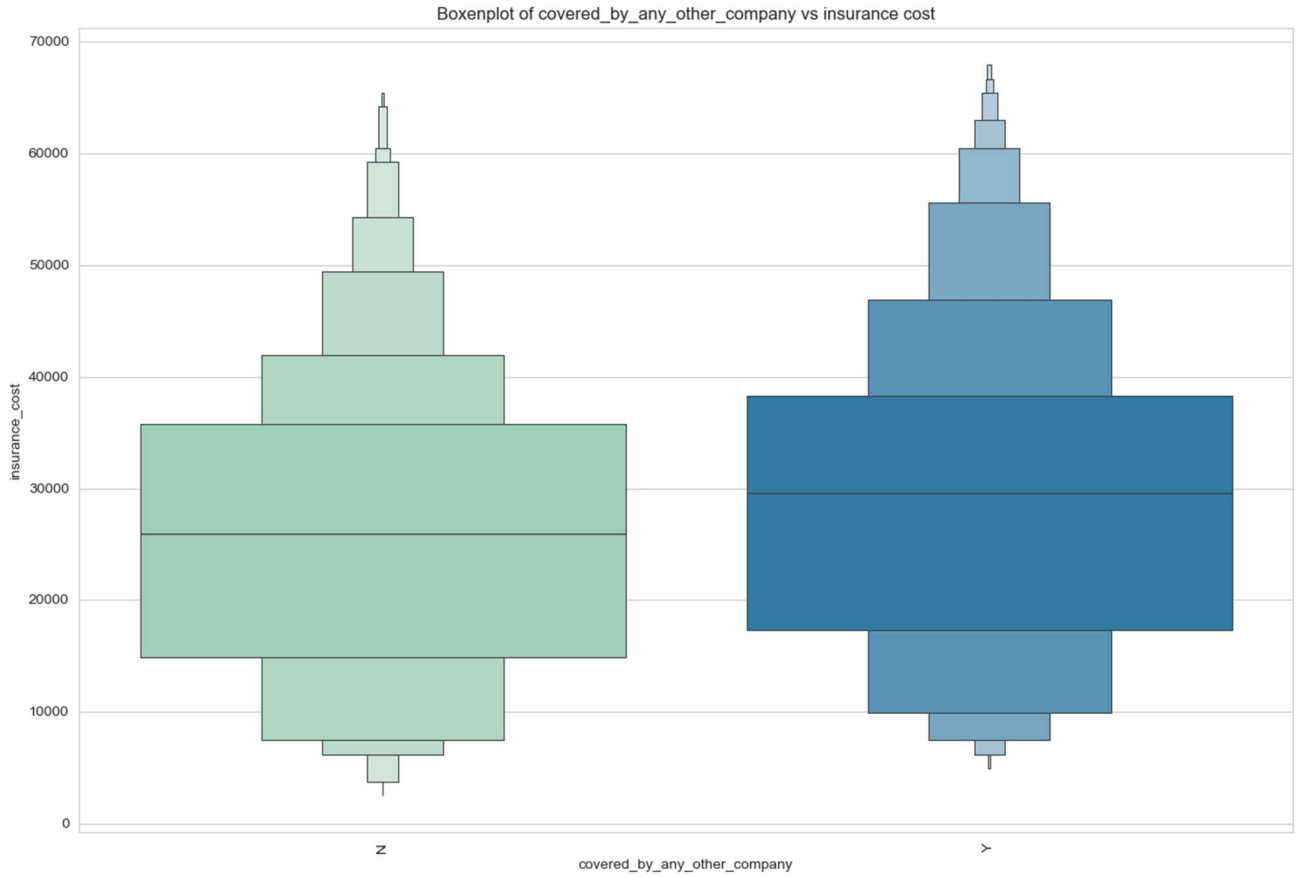


Fig.17. Boxenplot - covered_by_any_other_company vs insurance_cost

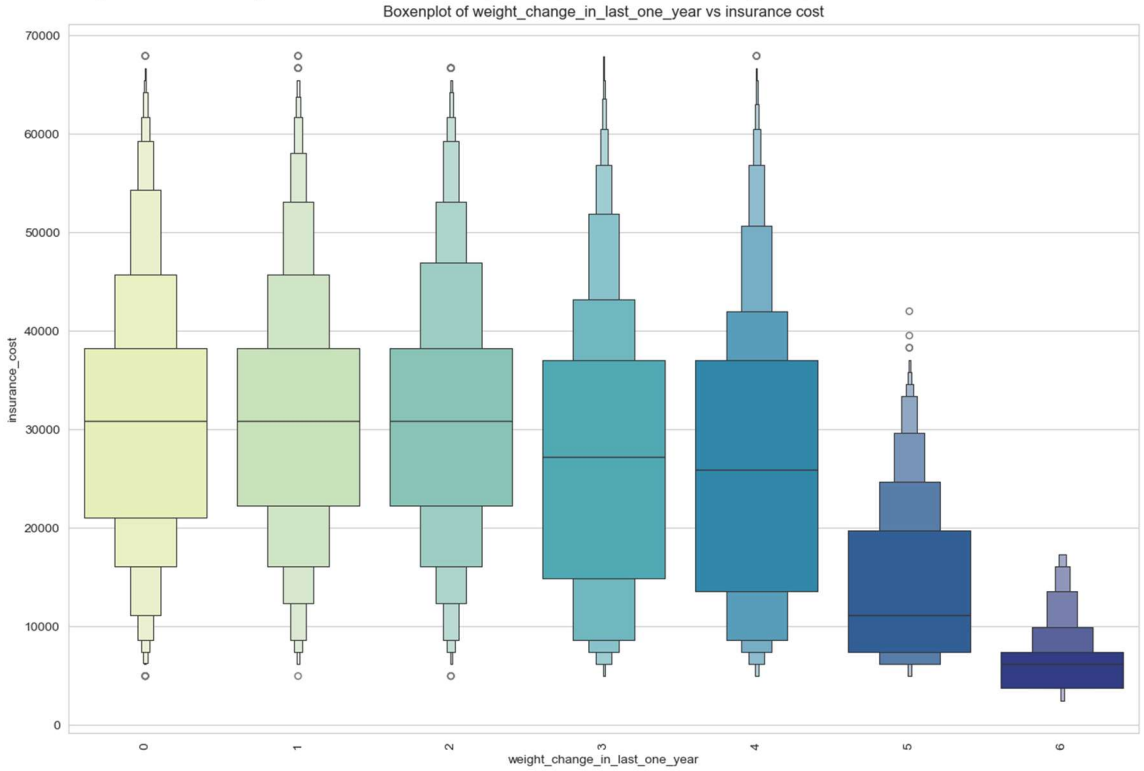


Fig.18. Boxenplot- weight_change_in_last_one_year vs insurance_cost

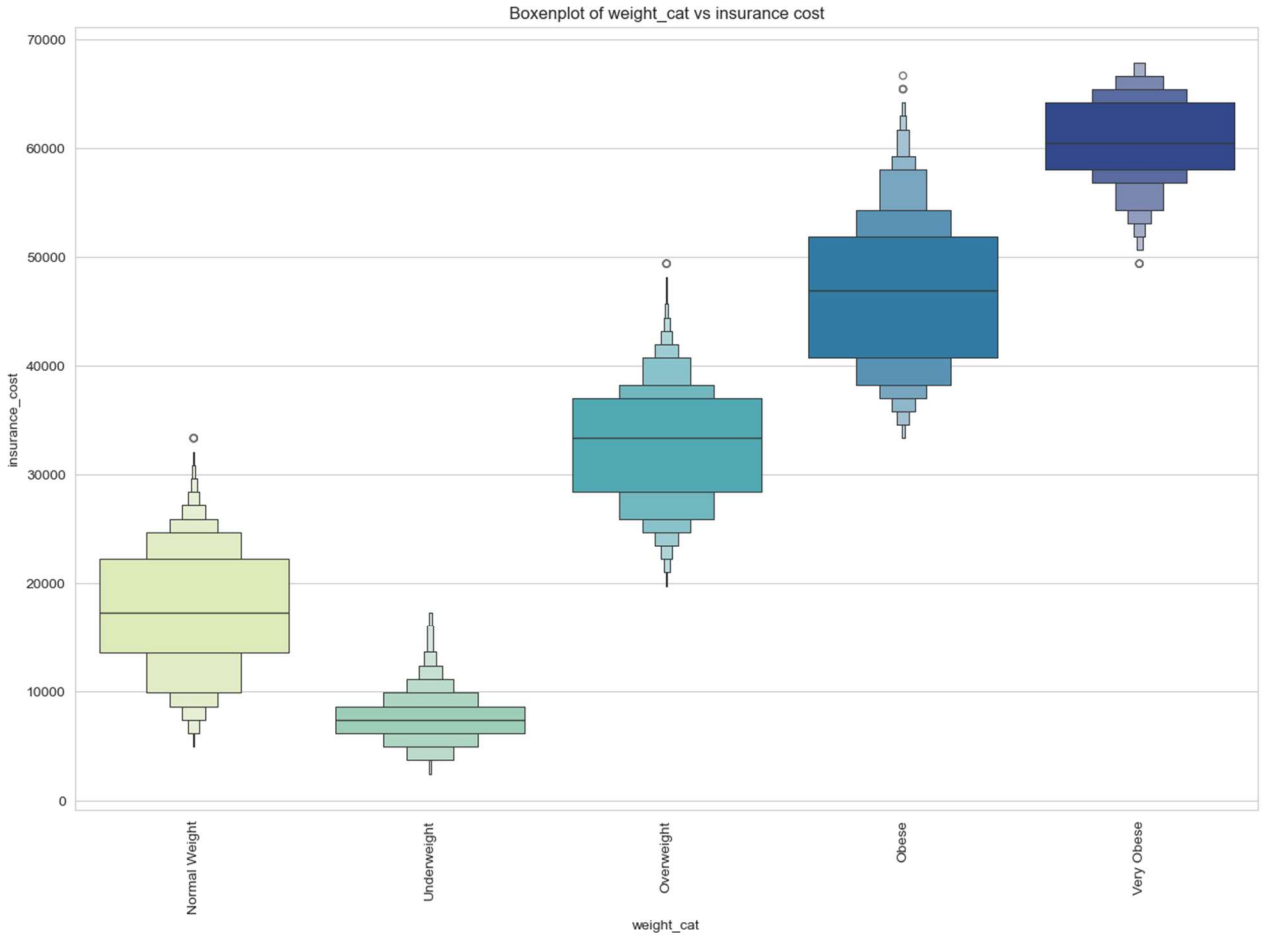


Fig.19. Boxenplot – weight_cat vs insurance_cost

Observations:

From the bivariate analysis of the target variable,i.e., insurance_cost, with the other categorical variables, the following are the observations. The plots for the same are above:

- Those who have had 0 regular checkups last year have the highest median insurance cost (Fig.14)
- Those who have had 5 or more than 5 regular checkups have the lowest median insurance cost (Fig.14)
- Those who engage in adventure sports have higher median insurance cost than those who don't (Fig.15).
- The median insurance cost is slightly lower in Ahmedabad and Bangalore when compared to that in other cities (Fig.16)
- Those who have coverage from other insurance companies tend to have a higher median insurance cost than those who don't (Fig 1.17)
- As the weight change in the past year increases, the median insurance cost decreases (Fig.18)
- As the weight category increases, median insurance cost also increases (Fig.19)

Categorical Variables vs categorical variables:

Heatmaps were plotted for a combination of each categorical variable with another. The following are some of the heatmaps relating to the observations below.

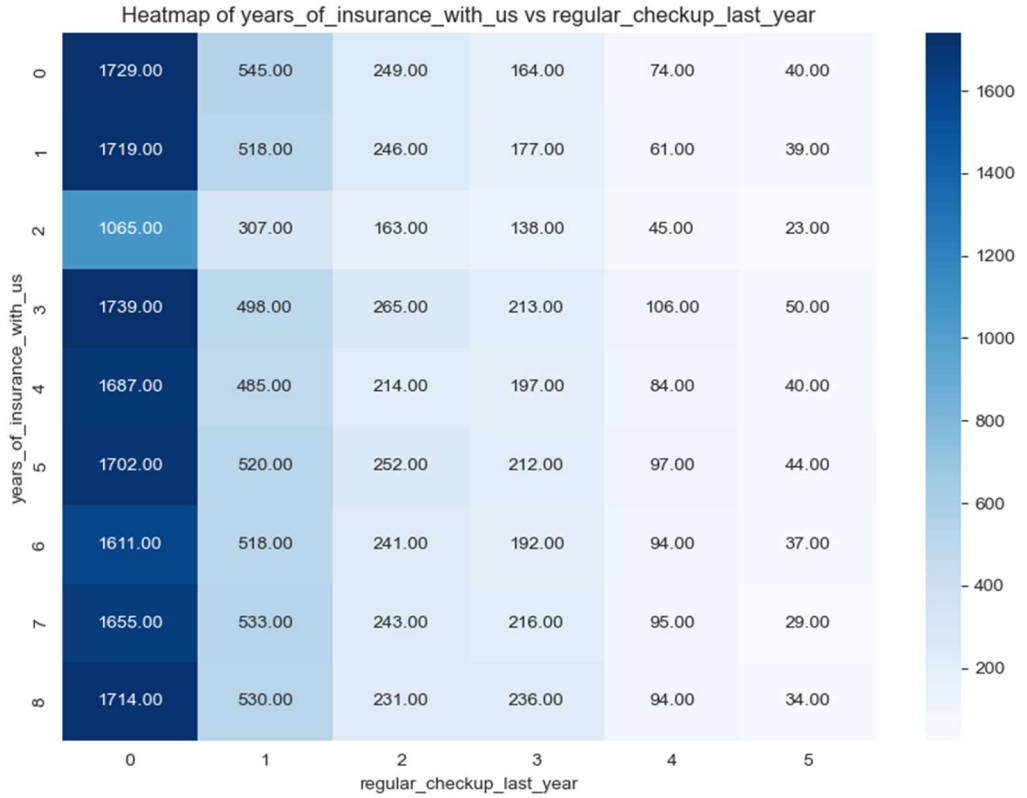


Fig.20 Heatmap- years_of_insurance vs regular_checkup_last_year

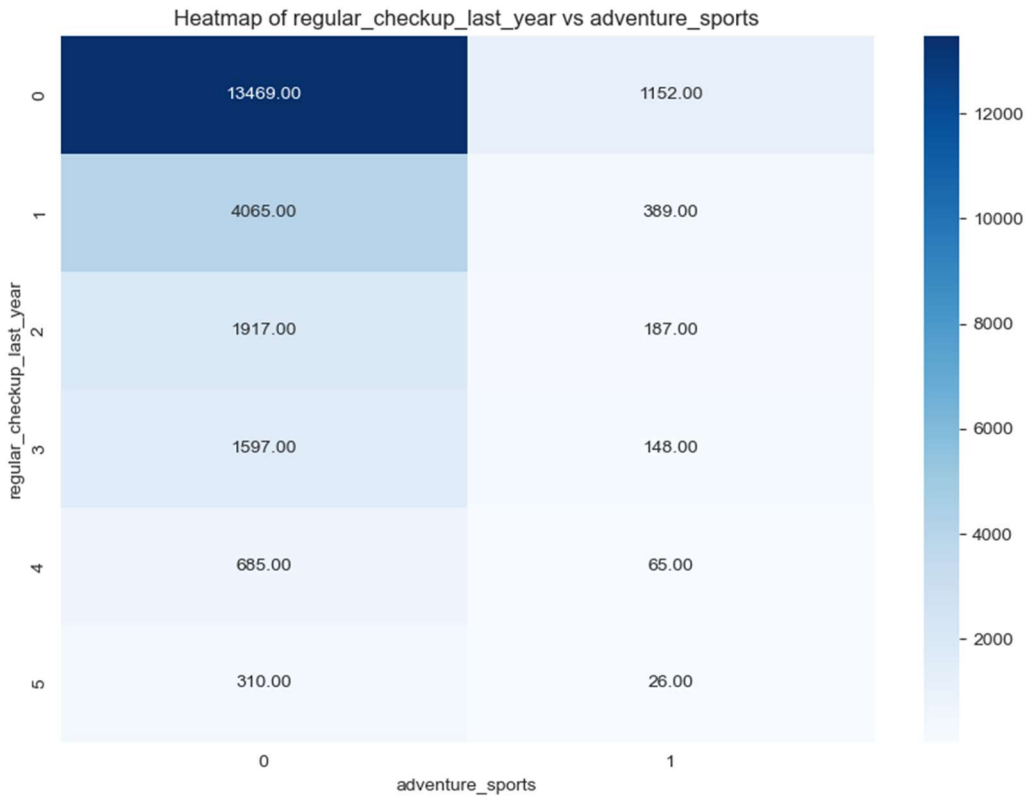


Fig.21- Heatmap- regular_checkup_last_year vs adventure_sports

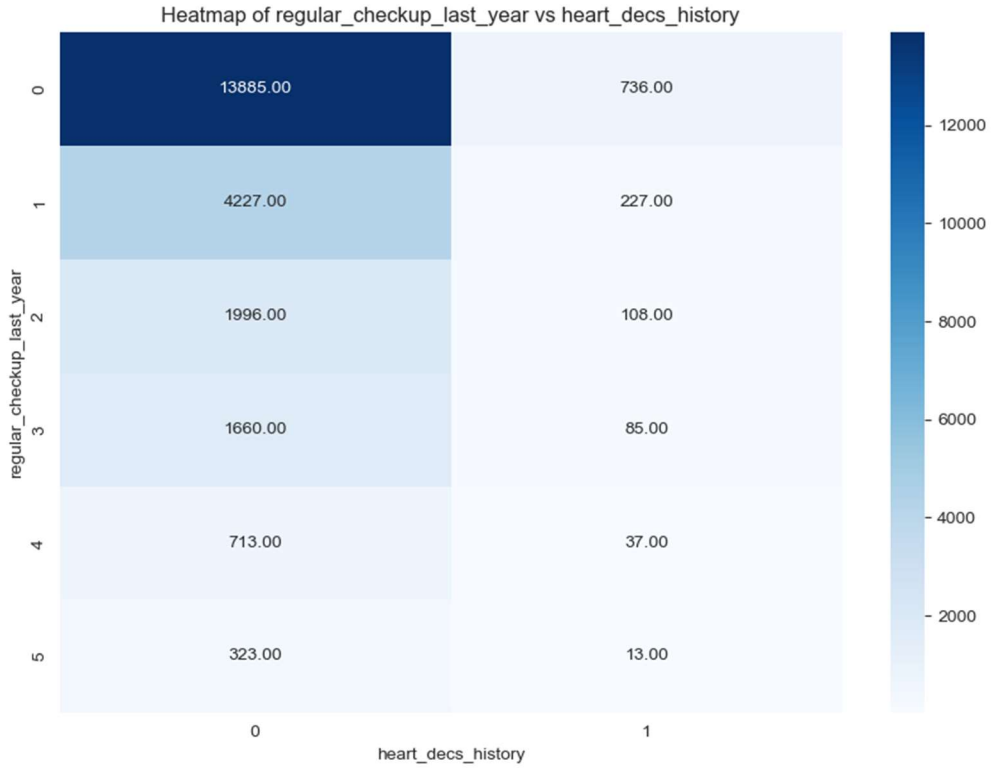


Fig.22. Heatmap- regular_checkup_last_year vs heart_decs_history

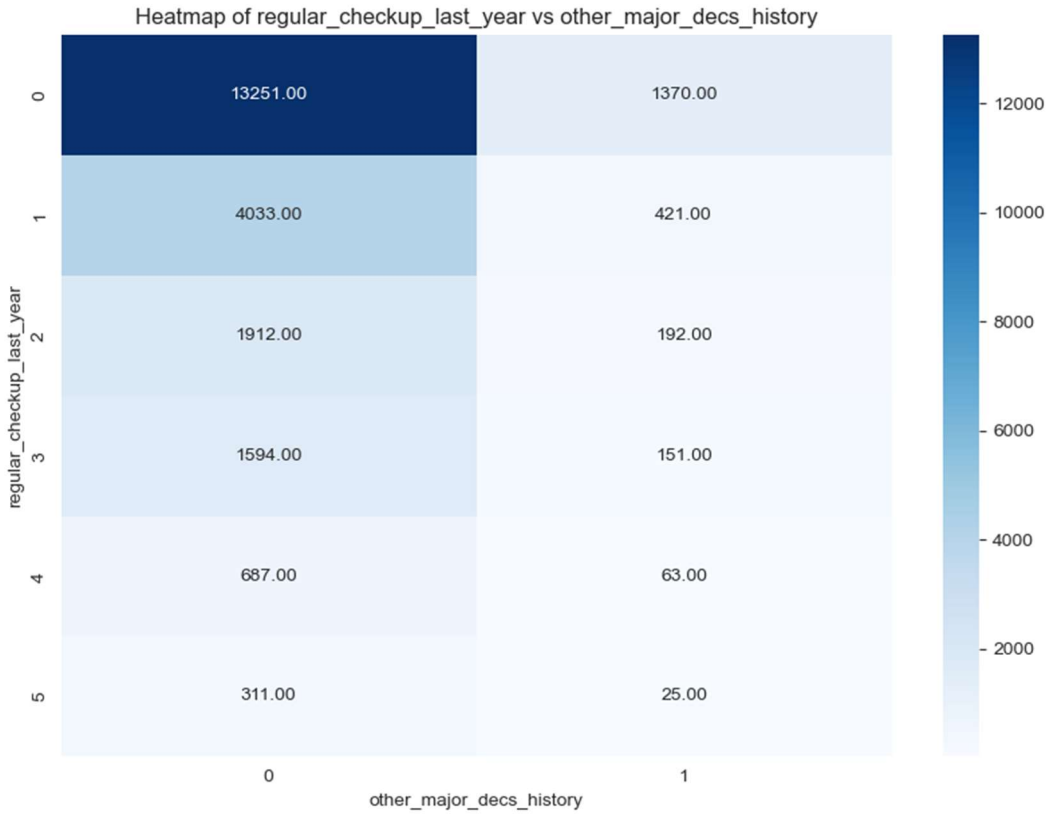


Fig.23. Heatmap- regular_checkup_last_year vs other_major_decs_history

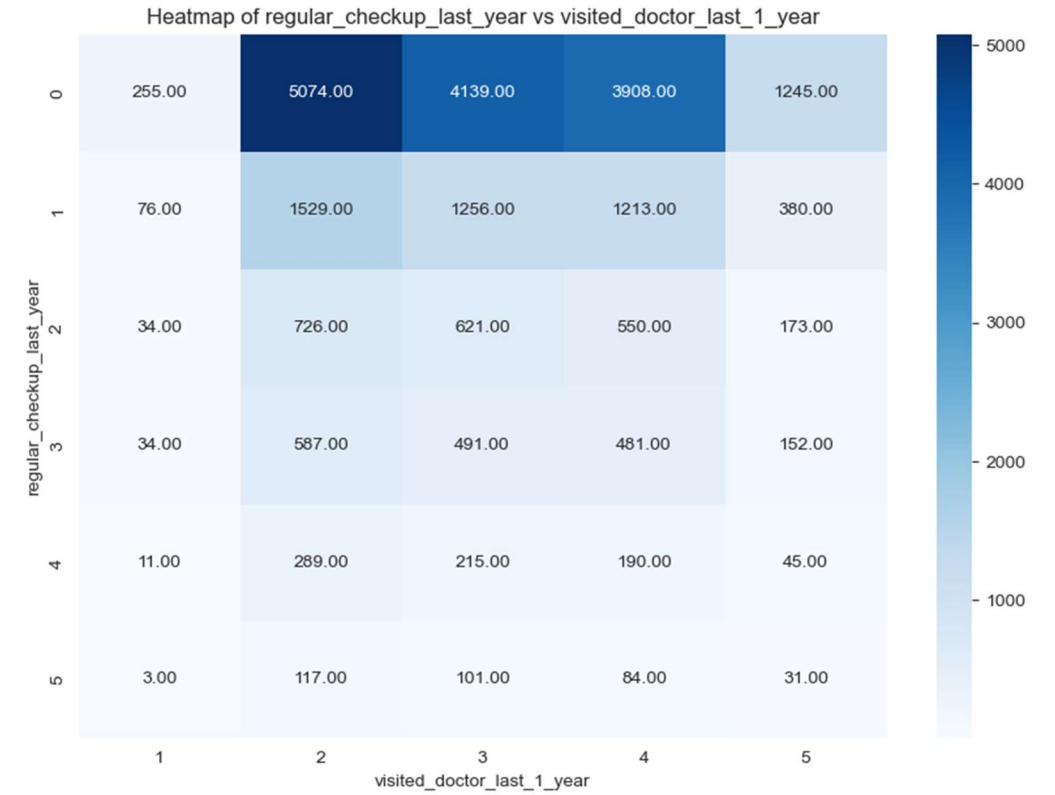


Fig.24. Heatmap- regular_checkup_last_year vs visited_doctor_last_1_year

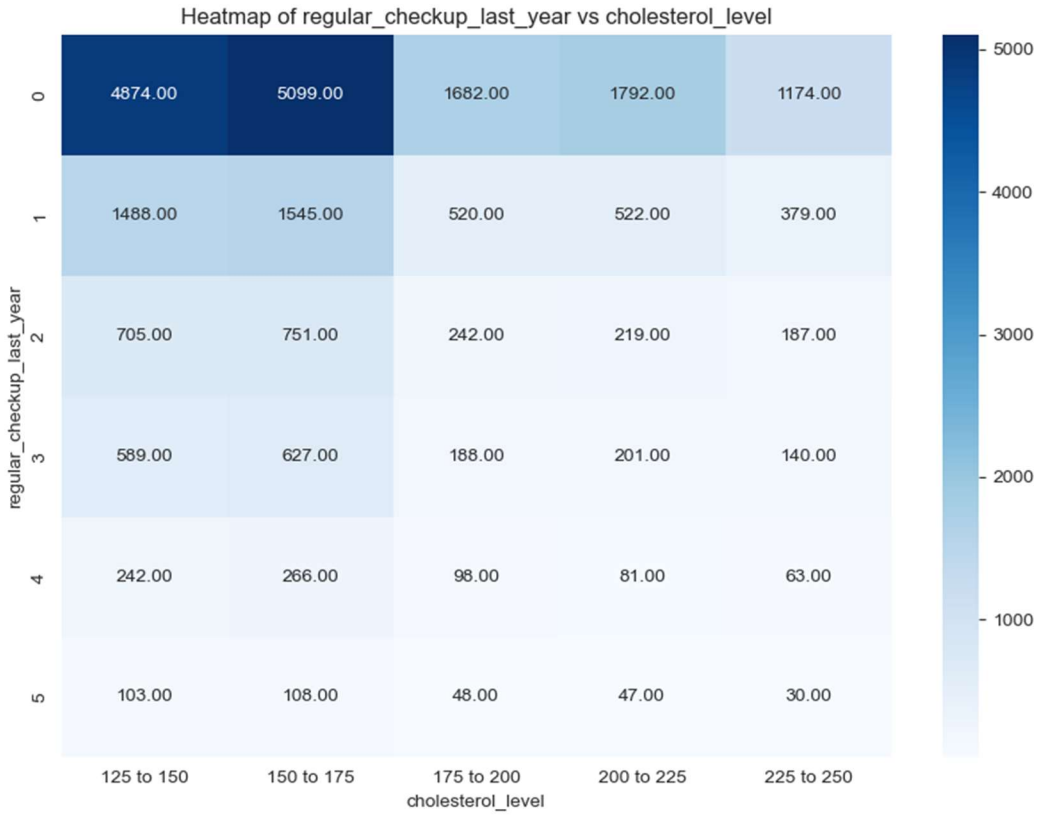


Fig.25. Heatmap-regular_checkup_last_year vs cholesterol_level

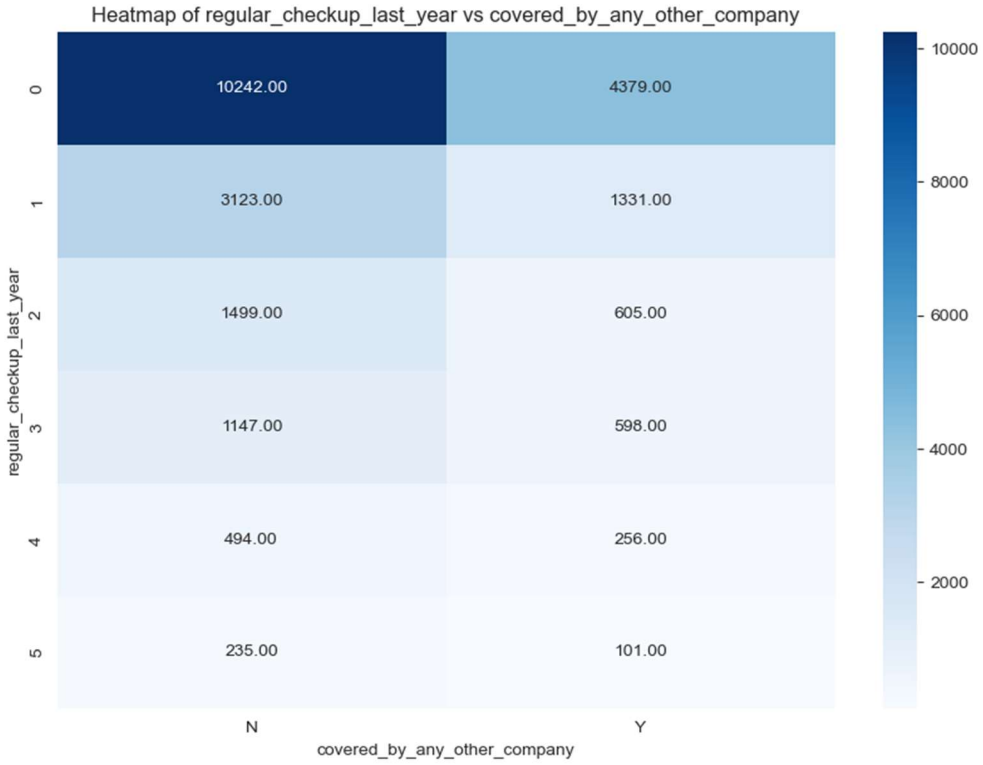


Fig.26. Heatmap- regular_checkup_last_year vs covered_by_any_other_company

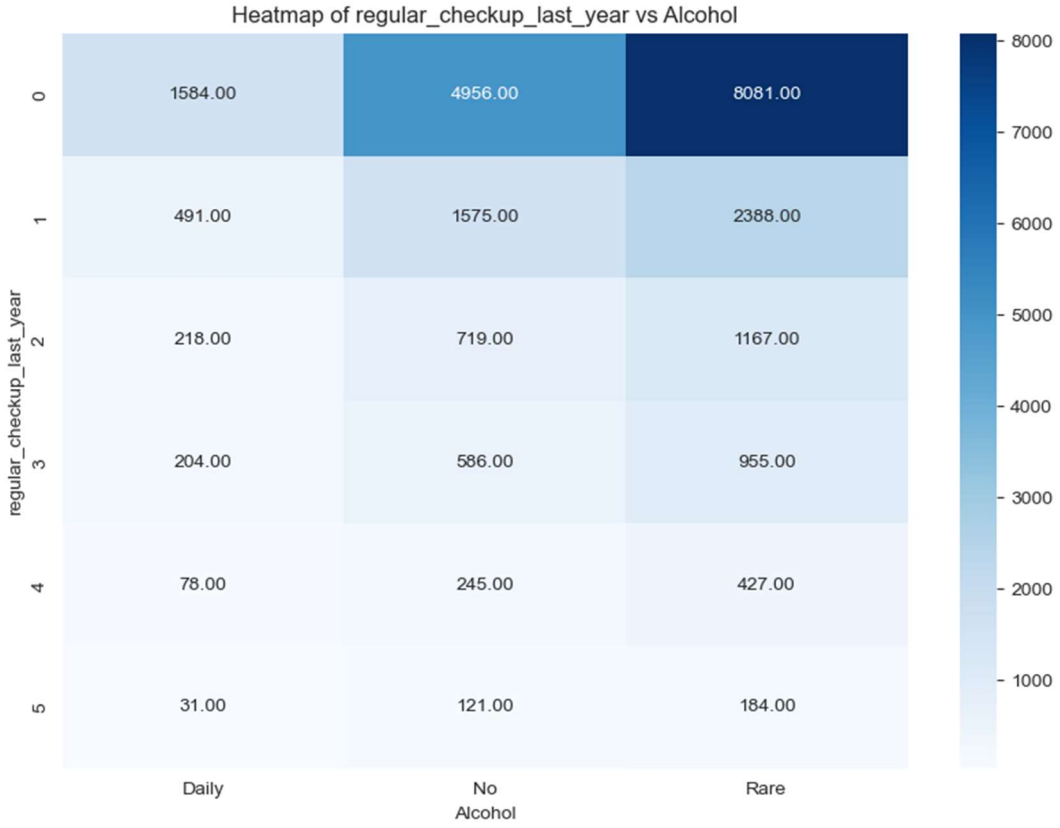


Fig.27. Heatmap- regular_checkup_last_year vs Alcohol

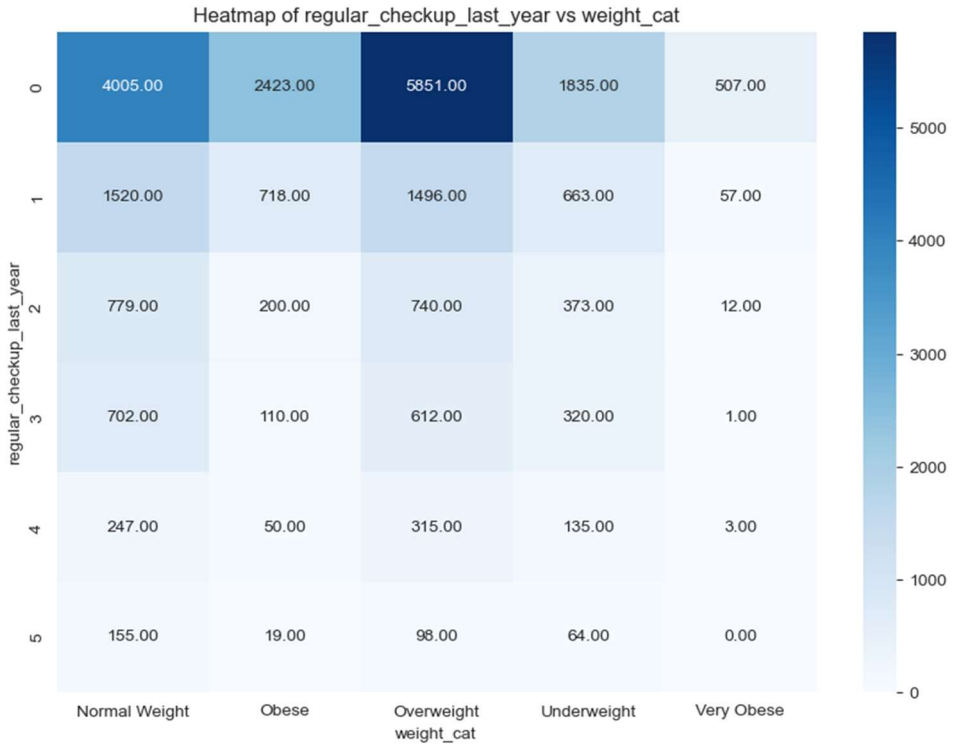


Fig.28. Heatmap- regular_checkup_last_year vs weight_cat

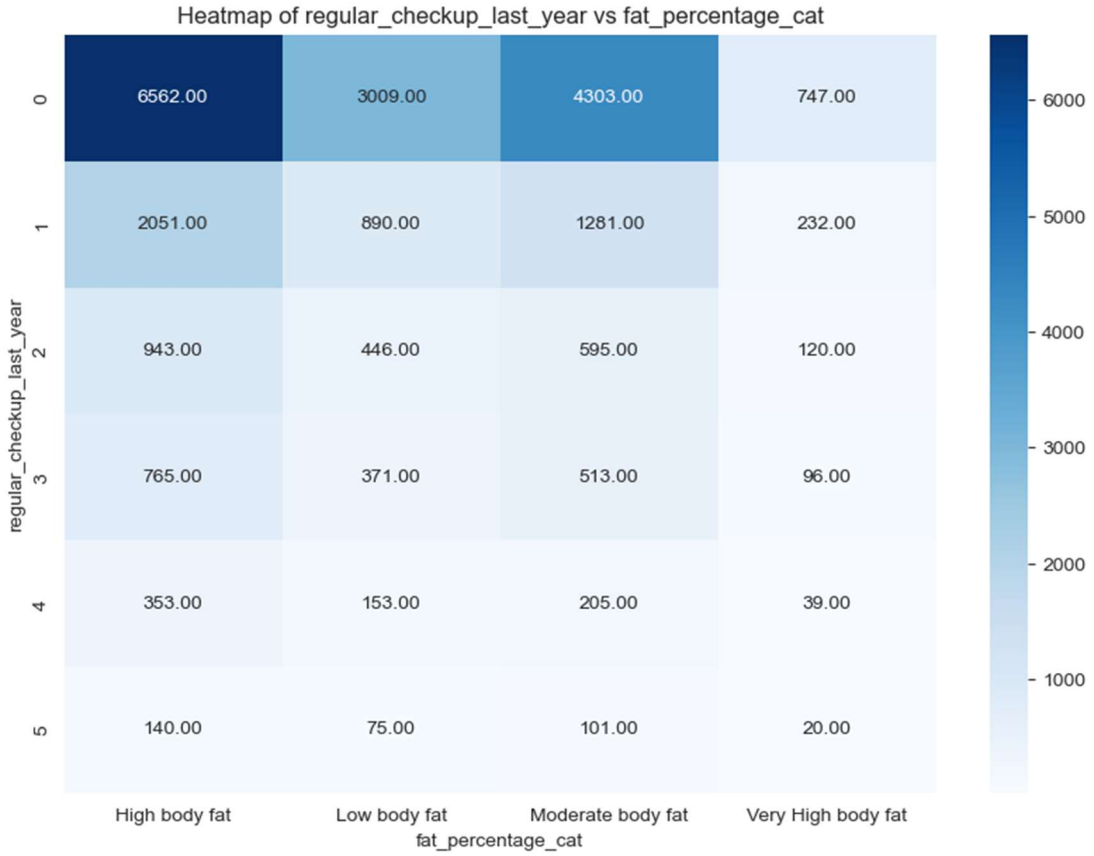


Fig.29. Heatmap- regular_checkup_last_year vs fat_percentage_cat

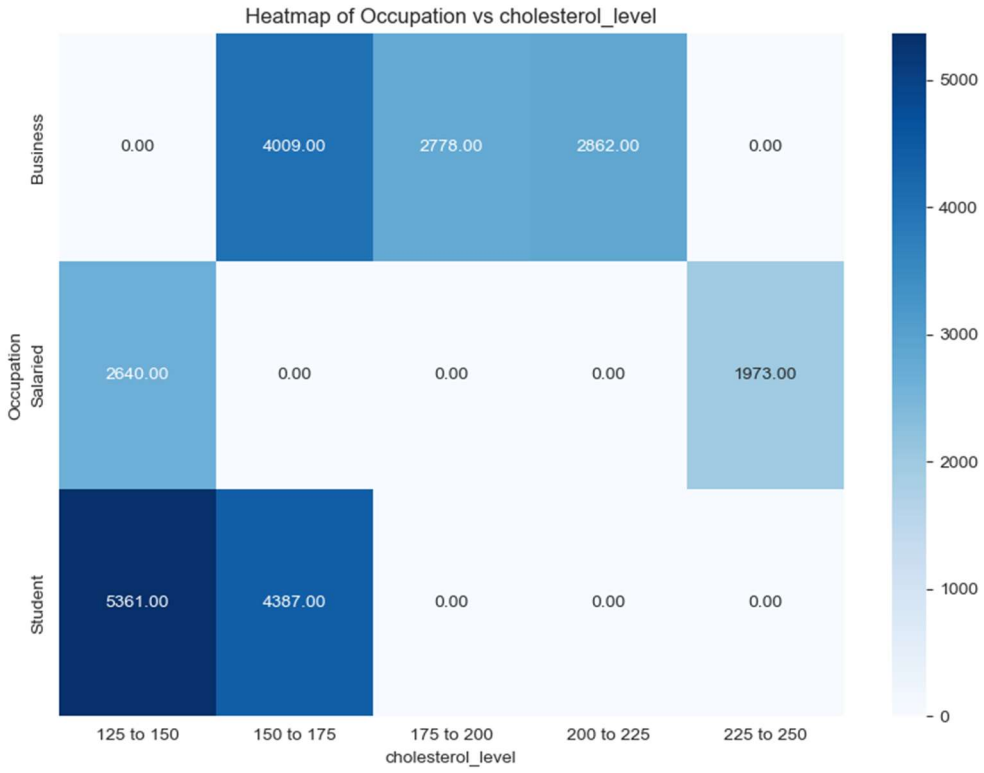


Fig.30. Heatmap- Occupation vs cholesterol_level

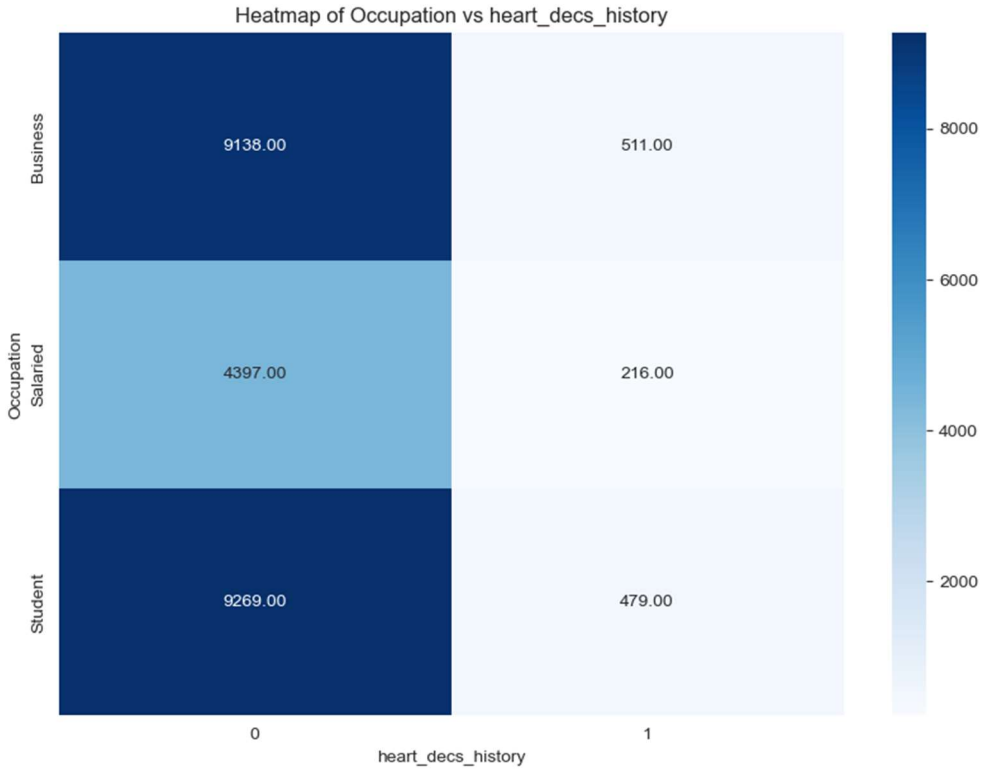


Fig.31. Heatmap- Occupation vs heart_decs_history

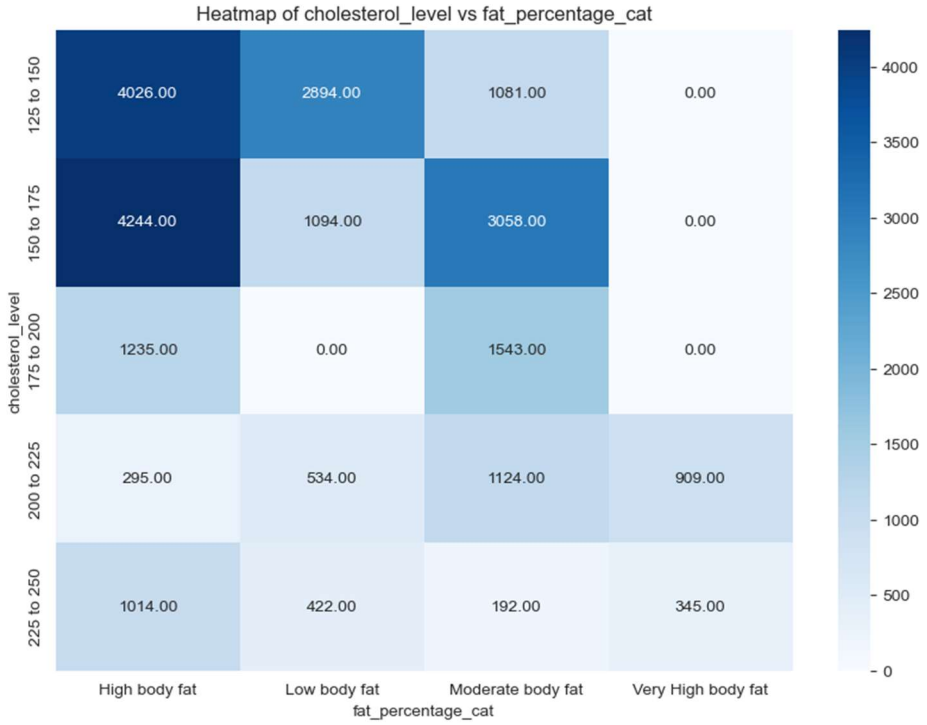


Fig.32. Heatmap- cholesterol_level vs fat_percentage_cat

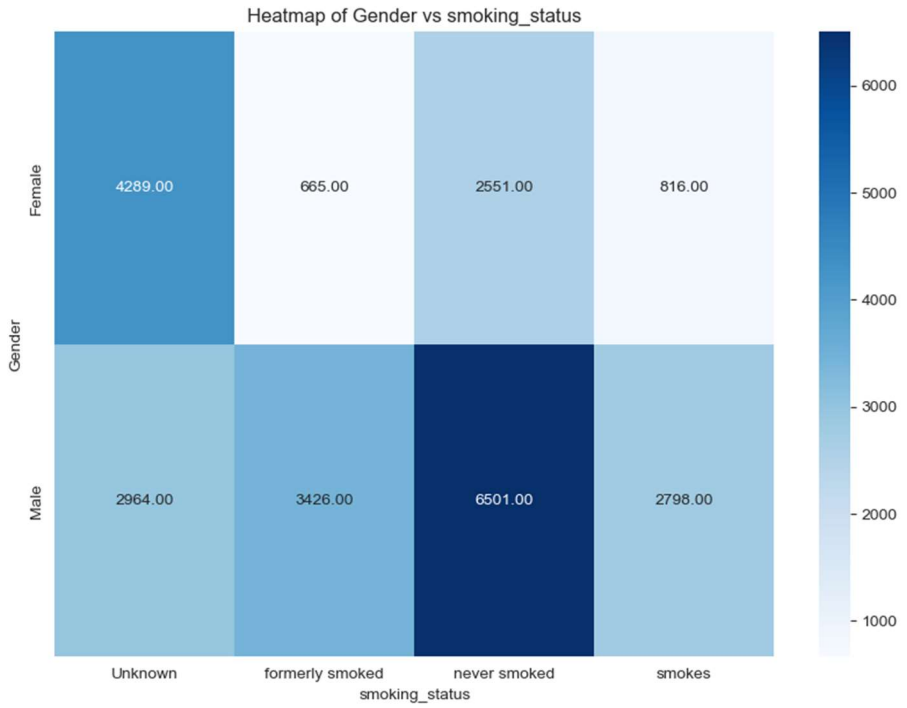


Fig.33. Heatmap- Gender vs smoking_status

Observations:

- 2801 people have less than 0 years insurance with us and are not being covered by any other company, which means they are new to the health insurance field. (Fig.20.)
- Around 13469 people have not had a regular checkup last year and are not into adventure sports. This is more than half the dataset size. (Fig.21.)
- Around half the population have visited the doctor last year and have not had any regular checkups (Fig.24)
- Most of the people who have not had regular checkup last year have a cholestrol level below 175 (Fig.25)
- More than half the population have had no major disease or heart disease and have not had regular checkup- these might be at risk of unknown medical issues. This might be a potential risk. Fig (1.22. and 1.23.)
- Around 10242 people have not had regular checkup and are not covered by any other company (Fig.26)
- 20% of the people have reported rare or daily intake of alcohol, but have not had regular checkup (Fig.27)
- 5851 people are overweight, and have not had regular checkup. This might be a potential risk. (Fig.28.)
- 6562 people have high body fat and have not had regular checkup. (Fig.29)
- No student has a cholesterol level over 175 (Fig.30.)
- Business people are prone to high cholesterol levels 150-225 (Fig.30.)

- Business people have reported the highest number of heart diseases, followed by students (Fig.31)
- People with very High body fat, i.e. fat_percentage from have cholesterol level greater than 200 (Fig.32.)
- Almost half the females have reported an Unknown smoking status (Fig.33)

Multivariate Analysis:

From the observations drawn from univariate and bivariate analysis, multivariate analysis was done on the dataset. The following are some of the plots and their corresponding observations.

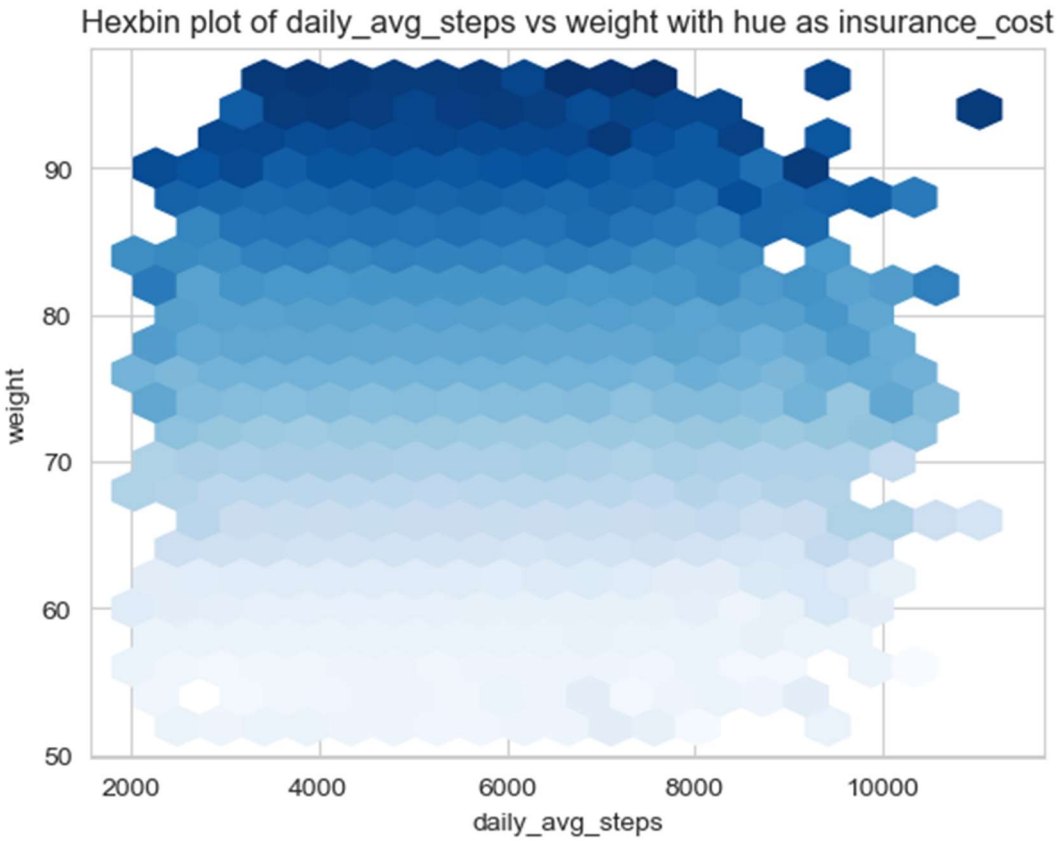


Fig.34. Hexbin plot- weight vs daily_avg_steps with hue as insurance_cost

Observations:

- Weight clearly seems to have a positive correlation with insurance cost
- Other than this, the numerical variables do not seem to be correlated

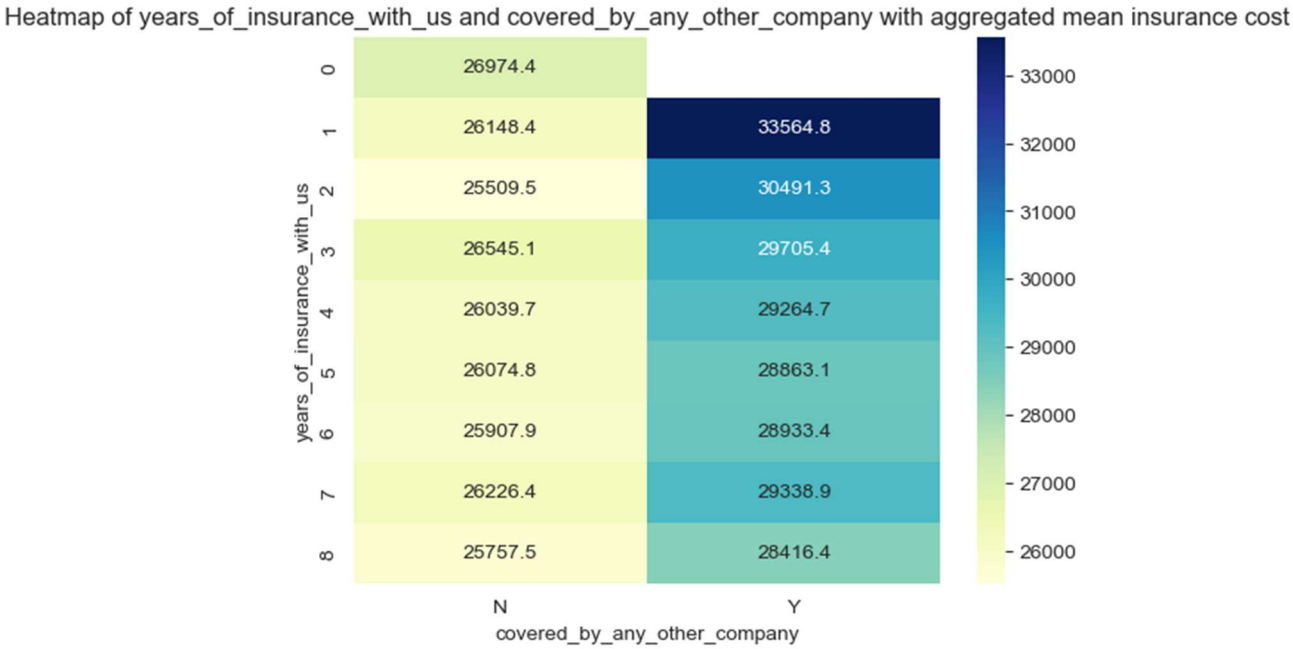


Fig.35. Heatmap- years_of_insurance vs covered_by_other_company vs insurance_cost

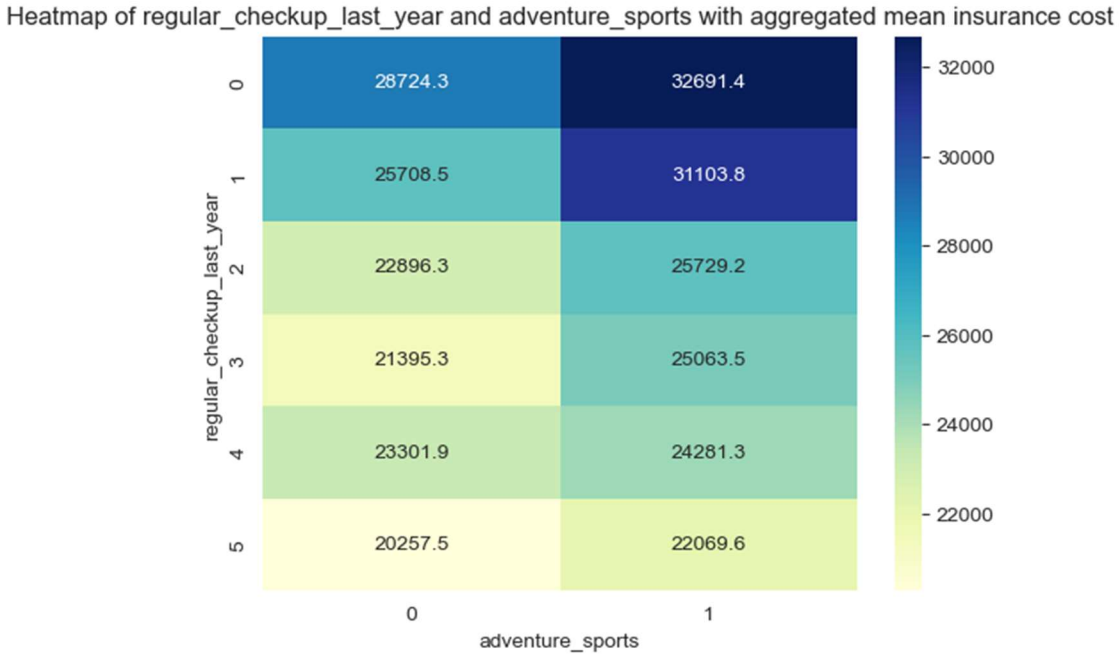


Fig.36. Heatmap- regular_checkup_last_year vs adventure_sports vs insurance_cost

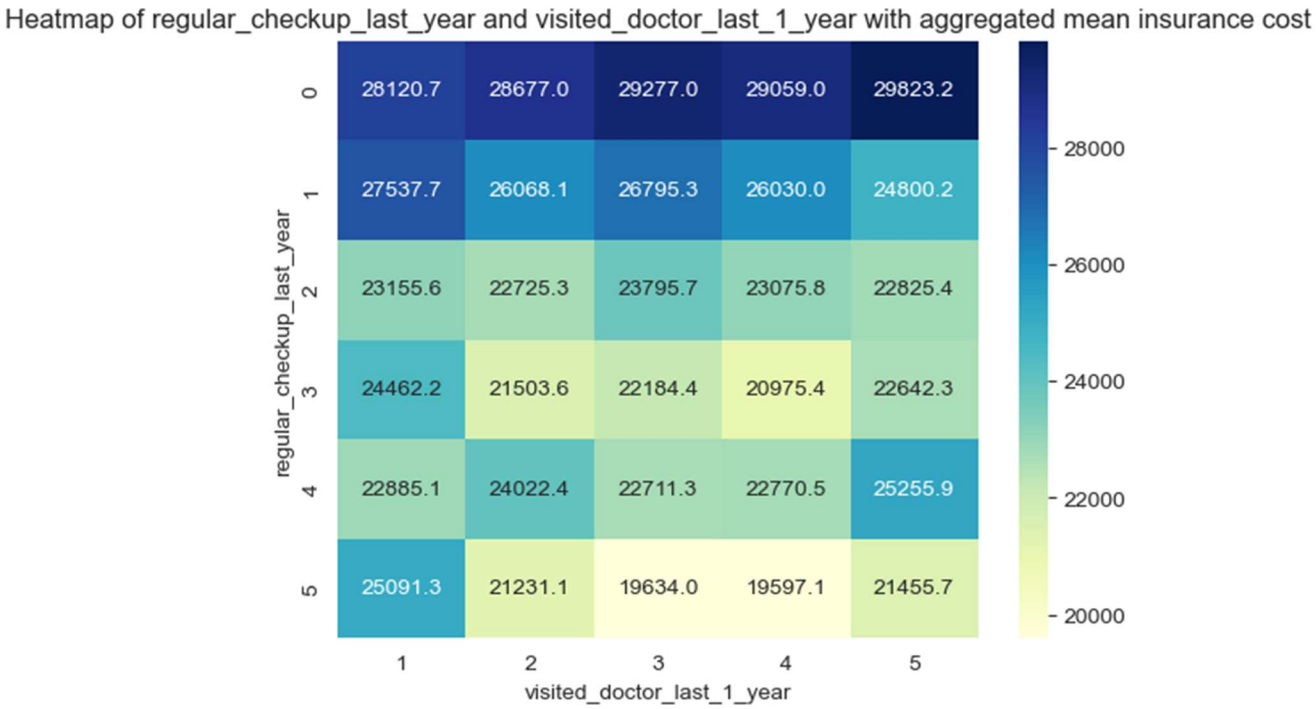


Fig.37. Heatmap- regular_checkup_last_year vs visited_doctor_last_year vs insurance_cost

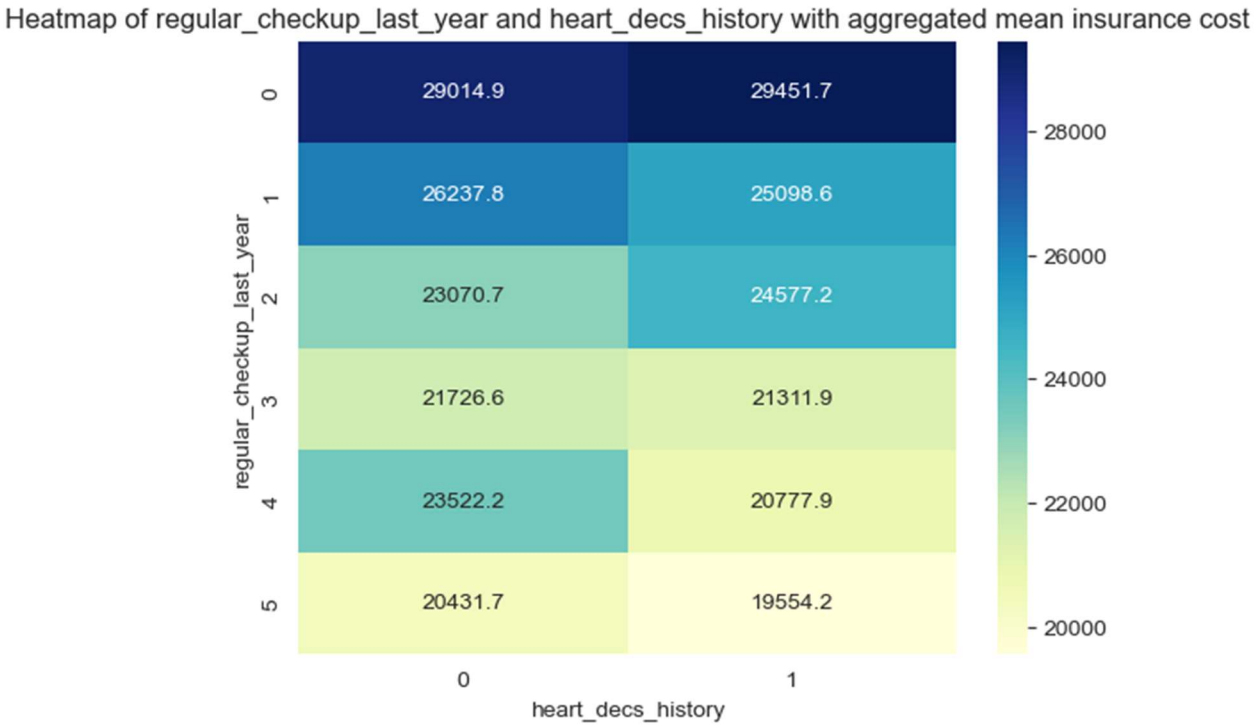


Fig. 38. Heatmap- regular_Checkup_last_year vs heart_decs_history vs insurance_cost

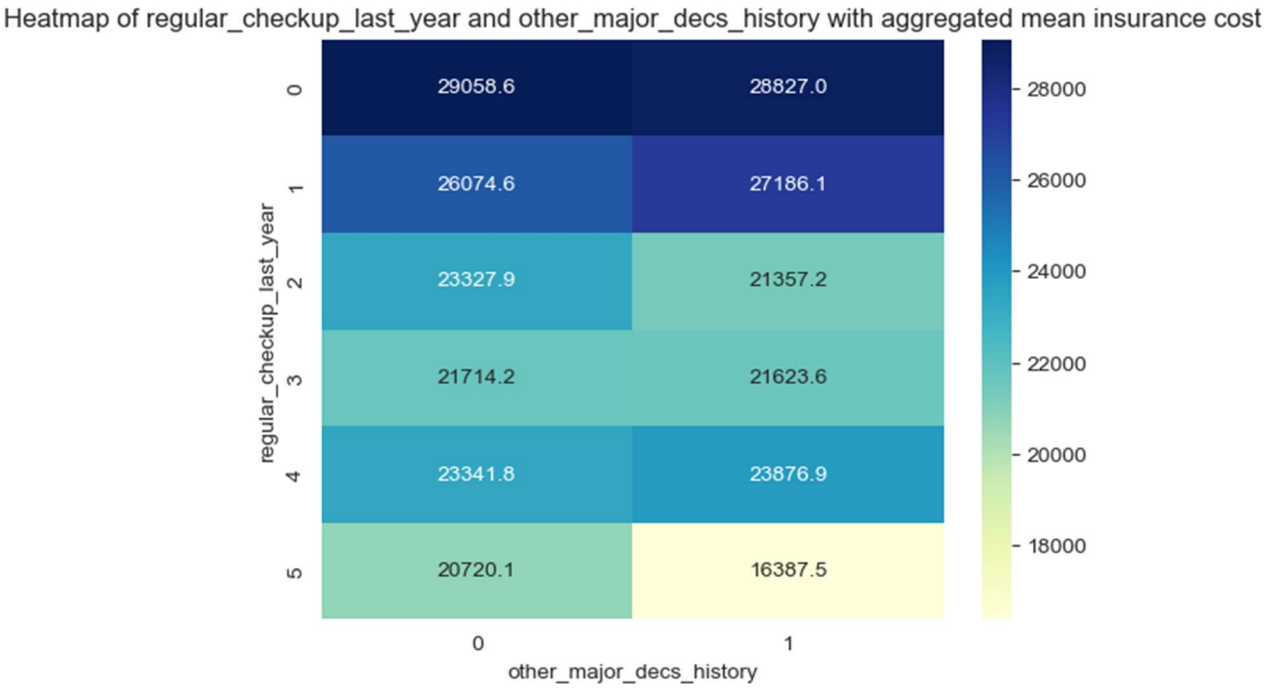


Fig. 39. Heatmap- regular_Checkup_last_year vs other_major_decs_history vs insurance_cost

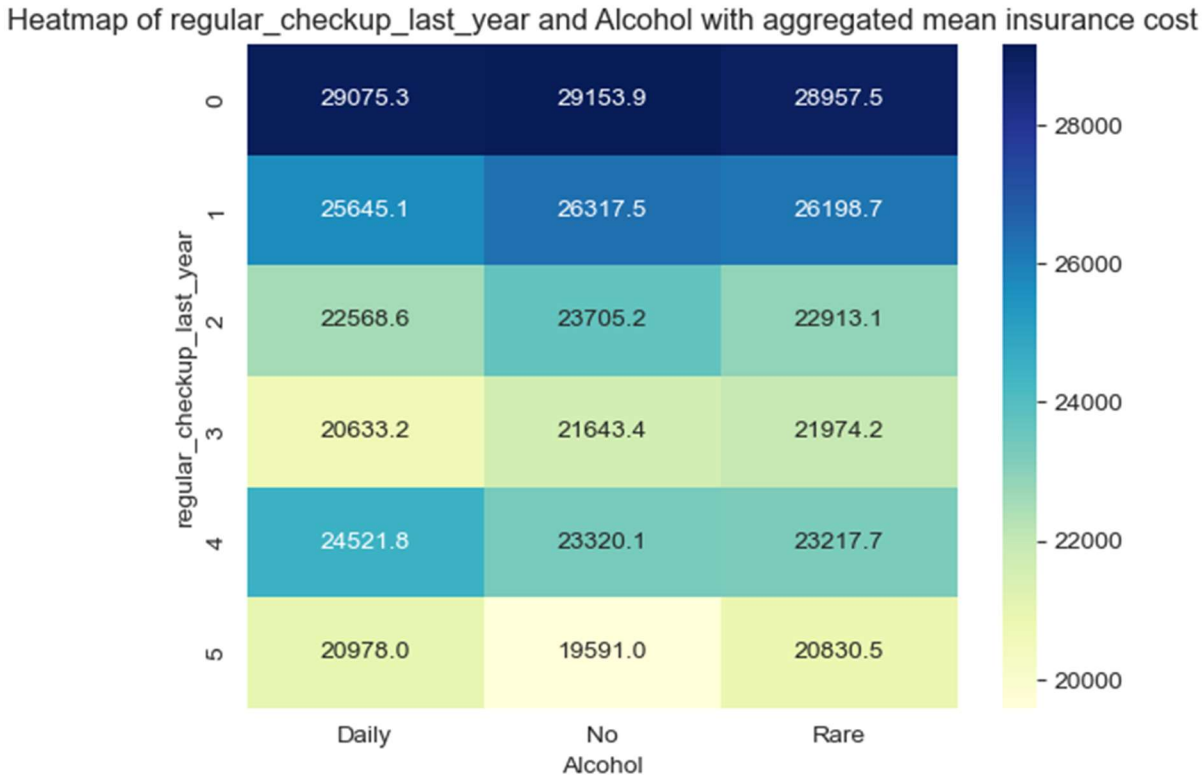


Fig.40. Heatmap- regular_Checkup_last_year vs Alcohol vs insurance_cost

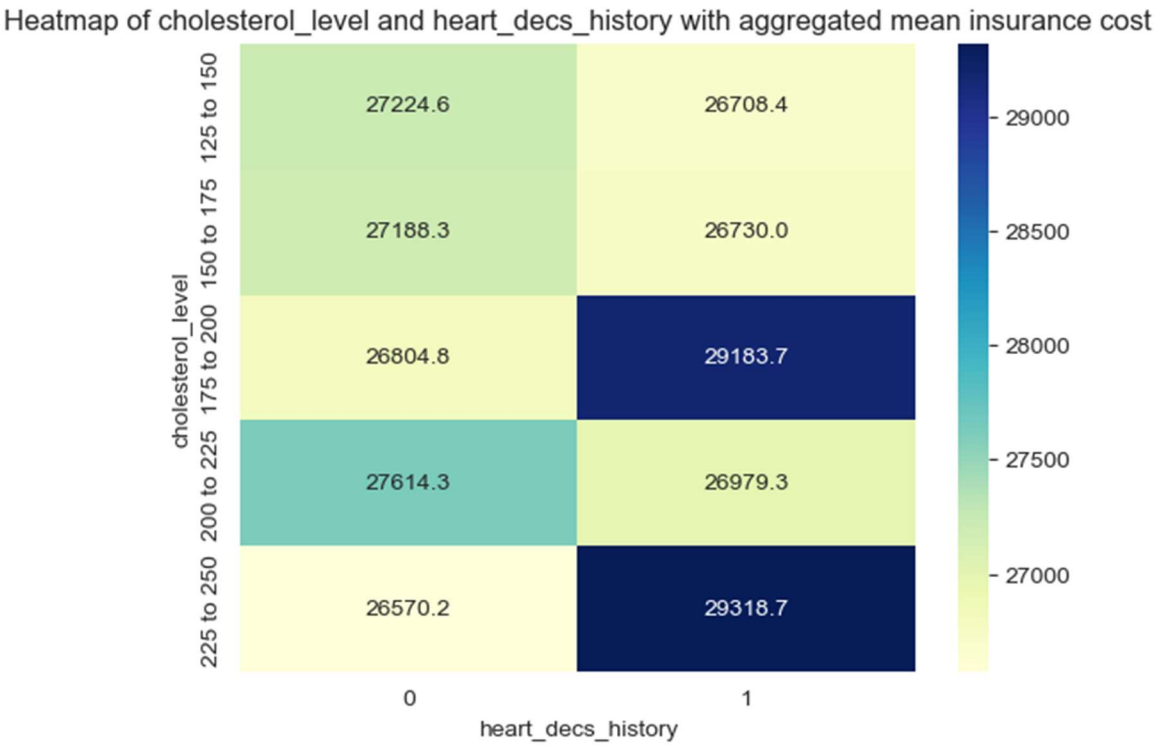


Fig.41. Heatmap- cholesterol_level vs heart_decs_history_vs insurance_Cost

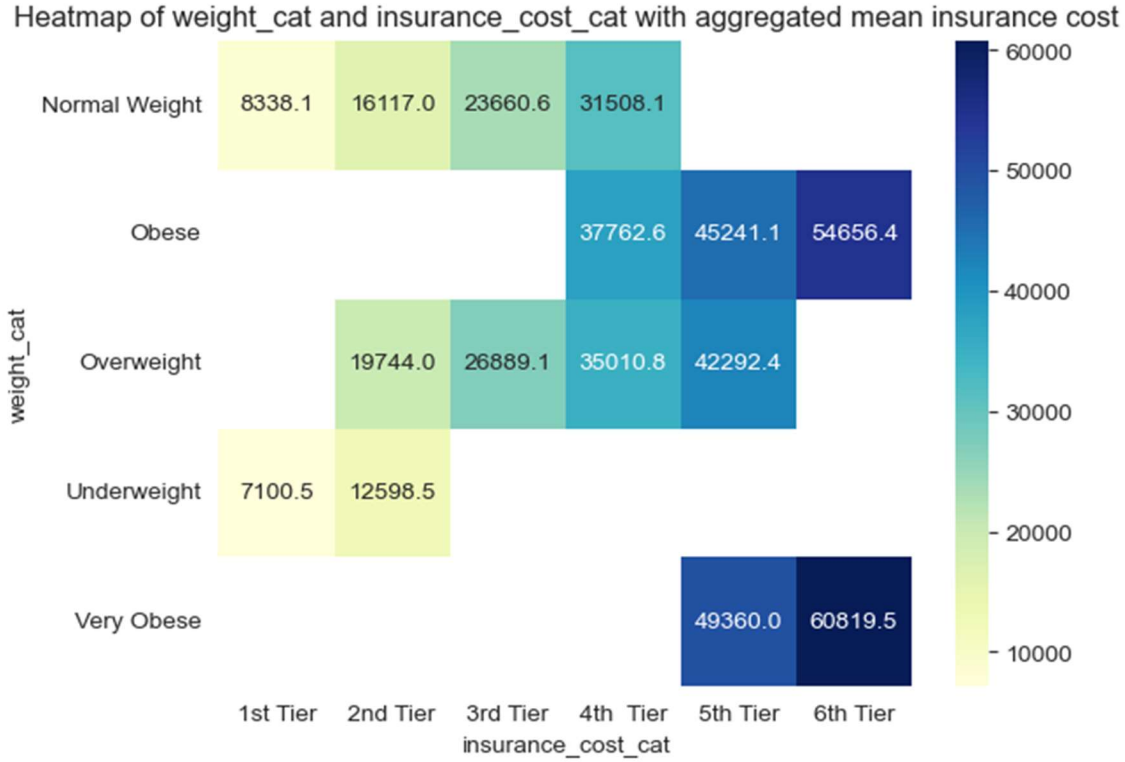


Fig.42. Heatmap- weight_cat vs insurance_cost_cat vs insurance_cost

Observations:

- People who are covered by other insurance companies, and have lesser number of years of insurance with us tend to have higher insurance costs (Fig.35.)
- People who are into adventure sports and have no regular checkups have high insurance cost (Fig.36.)
- People who have not had regular checkups, but have visited doctor for 5 times have the highest mean insurance cost (Fig.37.)
- People who have heart and other diseases, but have had 5 or more regular checkups have low insurance cost (Fig. 38 & 39)
- People who do not consume alcohol, and have regular checkups have low insurance cost (Fig.40)
- People who have a cholesterol level of 225-250 and have had heart disease have higher insurance cost (Fig.41)
- When people belong to the very obese weight category, the insurance cost tier is 5 or above, i.e. insurance cost is greater than 50000. (Fig.42)
- When they are underweight, their insurance tier is 2 or below, i.e., insurance cost is less than 20000. (Fig.42)

Outlier Treatment and Scaling:

- After the EDA, in order to make the data ready for modelling, scaling and outlier treatments were done
- For scaling, StandardScaler() function in sklearn preprocessing library was used
- For outlier treatment, the outliers were cut off at the whiskers by bringing the outlier values to 1.5 times the inter-quartile range of the distribution of the said variable
- Only one column, i.e., daily_avg_steps had outliers and was treated

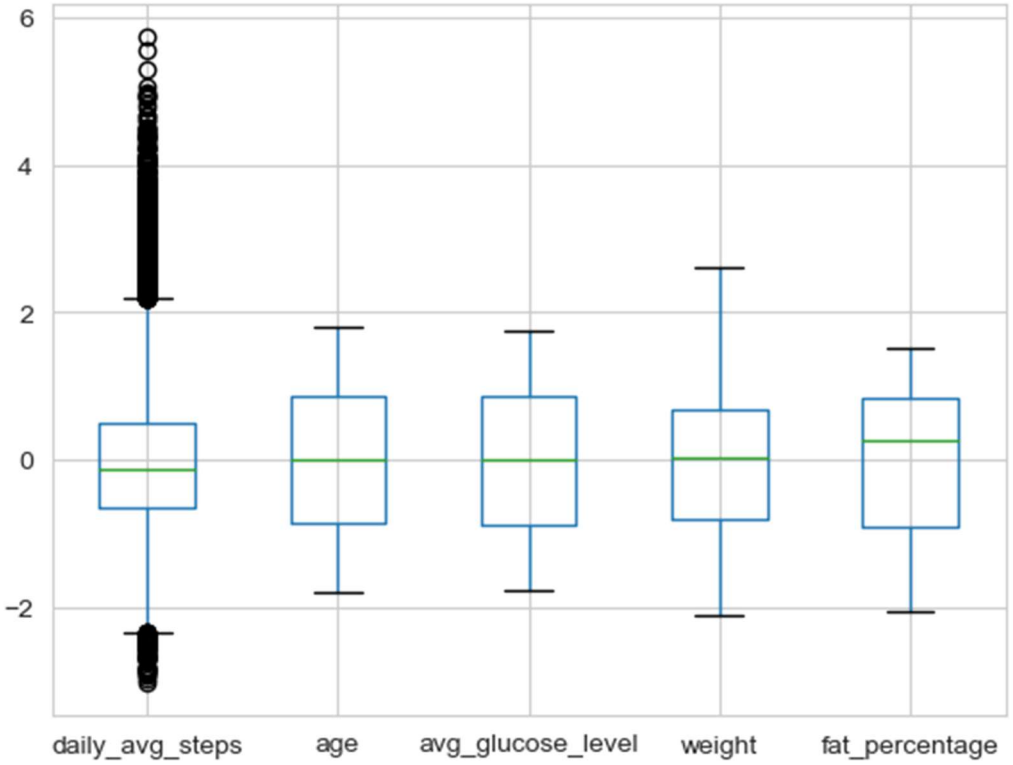


Fig.43. Boxplot of numerical variables before outlier treatment

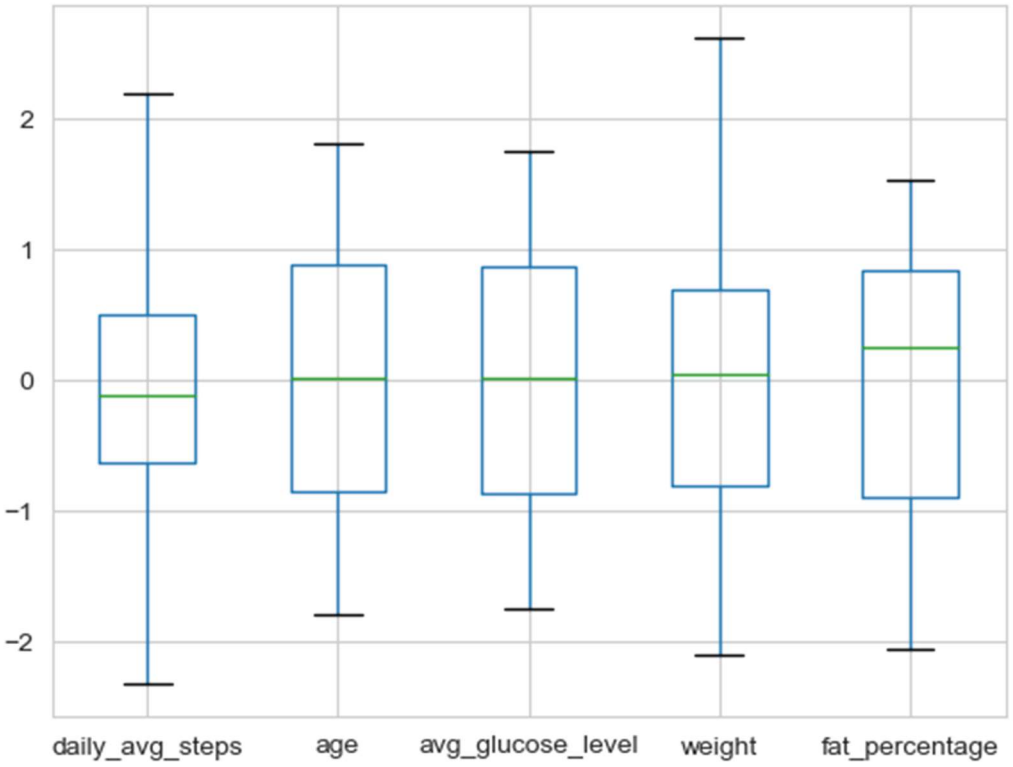


Fig.44. Boxplot of numerical variables after outlier treatment

	count	mean	std	min	25%	50%	75%	max
years_of_insurance_with_us	24010.0	4.09	2.61	0.00	2.00	4.00	6.00	8.00
regular_checkup_last_year	24010.0	0.77	1.20	0.00	0.00	0.00	1.00	5.00
adventure_sports	24010.0	0.08	0.27	0.00	0.00	0.00	0.00	1.00
visited_doctor_last_1_year	24010.0	3.11	1.14	0.00	2.00	3.00	4.00	12.00
daily_avg_steps	24010.0	-0.03	0.92	-2.33	-0.64	-0.12	0.49	2.18
age	24010.0	-0.00	1.00	-1.80	-0.87	0.01	0.88	1.81
heart_decs_history	24010.0	0.05	0.22	0.00	0.00	0.00	0.00	1.00
other_major_decs_history	24010.0	0.09	0.29	0.00	0.00	0.00	0.00	1.00
avg_glucose_level	24010.0	-0.00	1.00	-1.76	-0.87	0.01	0.87	1.75
bmi	24010.0	31.39	7.88	12.30	26.10	30.50	35.60	100.60
weight	24010.0	0.00	1.00	-2.10	-0.82	0.04	0.68	2.61
weight_change_in_last_one_year	24010.0	2.52	1.69	0.00	1.00	3.00	4.00	6.00
fat_percentage	24010.0	0.00	1.00	-2.06	-0.91	0.25	0.83	1.53
insurance_cost	24010.0	27160.64	14332.04	2468.00	16042.00	27148.00	37020.00	67870.00

Fig.45. Data Description after scaling and outlier treatments

Encoding Data:

- In order to enable further predictive modelling, the categorical columns needed to be converted to numerical variables
- Hence one-hot encoding was performed on columns- Location, Occupation, cholesterol_level, smoking_status, Alcohol, Gender, covered_by_any_other_company and exercise
- After encoding the shape of the data is (24010,43)

Data Balancing:

The given problem statement involves building predictive models to estimate the insurance cost, which is a numerical target variable. Hence, data balancing is not required as it is applicable only to classification models

Feature Elimination and Clustering:

- Recursive Feature elimination was done based on a Random Forest Regresser model and 15 top features were identified
- KMeans Clustering was done for the identified features
- After optimizing the number of clusters by silhouette scores and elbow plots, the optimum number of clusters were identified to be 3
- However, the cluster profiling did not reveal any distinctions between the clusters
- Hence, the clustering was inconclusive

Business Insights from EDA:

Identification of Business Opportunities:

- Around 3000 people in the dataset are entirely new to the health insurance stream. Hence, they can be targeted and possible add-ons to existing policies can be sourced

Risk Estimation:

- From the EDA, it is evident that the situations where the risk is high lead to higher insurance costs
- That is, when people do not have regular checkups, or participate in adventure sports, or have a higher weight or an unknown smoking status, their insurance costs tend to be higher
- This is a sound policy, which follows risk aversion
- However, there are certain steps that can be recommended.
- Encouraging the people to have regular checkups might lead to the updation of certain biomedical indices, like cholesterol level, which need to be updated frequently
- This also avoids surprises relating to undiagnosed medical issues
- This can be accomplished by rewarding the people with regular checkups by lowering the insurance costs, and penalizing those that don't
- Also, emphasizing the disclosure of smoking_status can lead to better pricing, and risk mitigation

Identification of significant variables:

- Based on the EDA, the following variables were found to be significant in determining the insurance costs:
    - Weight
    - Regular checkups
    - Participation in Adventure sports
    - Fat percentage