Health Care Project_Life Insurance Cost

Problem Statement:

We all know that Health care is very important domain in the market. It is directly linked with the life of the individual; hence we have to be always be proactive in this particular domain. Money plays a major role in this domain, because sometime treatment becomes super costly and if any individual is not covered under the insurance then it will become a pretty tough financial situation for that individual. The companies in the medical insurance also want to reduce their risk by optimizing the insurance cost, because we all know a healthy body is in the hand of the individual only. If individual eat healthy and do proper exercise the chance of getting ill is drastically reduced.

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

Scoring guide (Rubric) - Project Note 1 (1)	
Criteria	**Points**
1. Problem Understanding	4
a) Defining problem statement b) Need of the study/project c) Understanding business/social opportunity	
2. Data Report	2
a) Understanding how data was collected in terms of time, frequency and methodology b) Visual inspection of data (rows, columns, descriptive details) c) Understanding of attributes (variable info, renaming if required)	
3. Exploratory Data Analysis	10

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones) b) Bivariate analysis (relationship between different variables , correlations) a) Removal of unwanted variables (if applicable) b) Missing Value treatment (if applicable) d) Outlier treatment (if required) e) Variable transformation (if applicable) f) Addition of new variables (if required)	
4. Business insights from EDA	4
a) Is the data unbalanced? If so, what can be done? Please explain in the context of the business b) Any business insights using clustering (if applicable) c) Any other business insights	
Points	20

List of Figures:

List of Tables:

Solution:

1. Problem Understanding

- In the dynamic landscape of healthcare, where the well-being of individuals is intricately intertwined with financial considerations, the challenge of establishing optimal insurance costs emerges as a pivotal concern.
- Healthcare stands as a paramount domain, directly influencing the lives of individuals, and demands proactive solutions that align economic feasibility with comprehensive coverage.

a) Defining problem statement:

- The healthcare sector in India has witnessed substantial growth, encompassing various areas such as hospitals, medical devices, clinical trials, outsourcing, telemedicine, medical tourism, health insurance, and medical equipment. However, despite its significant revenue and employment contribution, there are certain challenges that need to be addressed to ensure the continued progress of the sector. One of the key concerns is the relatively low public expenditure on healthcare, standing at 1.2% of the GDP in Budget 2021. This raises questions about the accessibility and quality of healthcare services for the population.
- The problem revolves around determining an optimal insurance cost for individuals in the healthcare domain.
- This involves utilizing health and lifestyle-related data to predict the most suitable insurance premium. The aim is to strike a balance between affordability for individuals and risk management for insurance companies.

b) Need of the Study/Project:

- The study or project is necessary due to the critical importance of healthcare and the financial implications it carries. Health-related issues can lead to significant financial burdens, especially when not covered by insurance.
- By developing a predictive model to estimate insurance costs based on health and habit parameters, individuals can make informed decisions about coverage, and insurance companies can better assess risks and optimize their pricing strategies.
- Health crises can not only endanger personal well-being but also lead to exorbitant medical expenses that are aggravated when lacking insurance coverage.
- By crafting a predictive model grounded in health data, we equip individuals with insights into insurance costs, facilitating informed decisions, and offer insurance companies a more refined approach to risk assessment and premium calculation.

c) Understanding Business/Social Opportunity:

- The healthcare sector presents both a significant business opportunity and a social responsibility. As the sector is projected to grow three-fold to Rs. 8.6 trillion by 2022, businesses have a chance to capitalize on this growth by providing innovative solutions, technologies, and services that can improve the overall healthcare experience for patients and medical professionals.
- From a business perspective, this project presents an opportunity for insurance companies to enhance their competitiveness by offering personalized insurance plans. By leveraging data analytics and predictive modeling, insurance companies can attract more customers and reduce the likelihood of adverse selection.
- This approach aligns with the growing demand for tailored solutions in the insurance industry.
- On the social front, the project contributes to better healthcare access and financial security for individuals.
- It empowers people to take charge of their health and lifestyle choices by demonstrating the direct impact on insurance costs. Ultimately, this can lead to healthier lifestyles, reduced medical expenses, and improved overall well-being.
- In summary, this project addresses the need for personalized insurance pricing in the healthcare domain, benefiting both insurance companies and individuals, while also promoting healthier lifestyles and financial stability.

2. Data Report:

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

- The dataset contains information from 25,000 observations, with each observation having 24 variables or columns.
- It's essential to note that this data provides valuable insights into the healthcare sector. The data appears to encompass diverse attributes, including applicant demographics, health history, lifestyle factors, insurance-related details, and more.

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

Table 1 Data decription

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	visited_doctor_last_1_year	daily_avg_steps	age	heart_decs_history	other_major_decs_history	avg_glucose_level	bmi	Year_last_admitted	weight	weight_change_in_last_one_year	fat_percentage	insurance_cost
count	25000	25000	25000	25000	25000	25000	25000	25000	25000	25000	24010	13119	25000	25000	25000	25000

mean	17499.5	4.08904	0.77368	0.08172	3.1042	5215.889	44.91832	0.05464	0.09816	167.53	31.39333	2003.892	71.61048	2.51796	28.81228	27147.41
std	7217.023	2.606612	1.199449	0.273943	1.141663	1053.18	16.10749	0.227281	0.297537	62.72971	7.876535	7.581521	9.325183	1.690335	8.632382	14323.69
min	5000	0	0	0	0	2034	16	0	0	57	12.3	1990	52	0	11	2468
25%	11249.75	2	0	0	2	4543	31	0	0	113	26.1	1997	64	1	21	16042
50%	17499.5	4	0	0	3	5089	45	0	0	168	30.5	2004	72	3	31	27148
75%	23749.25	6	1	0	4	5730	59	0	0	222	35.6	2010	78	4	36	37020
max	29999	8	5	1	12	11255	74	1	1	277	100.6	2018	96	6	42	67870

- Upon visually inspecting the dataset, several key observations can be made regarding the structure and content of the data. This inspection provides a preliminary understanding of the dataset's characteristics, which is crucial for further analysis and interpretation.

Data Size and Shape:

- The dataset contains a total of 25,000 observations (rows) and 24 variables (columns). This indicates a substantial amount of data to work with, which is promising for conducting meaningful analyses.

Data Types and Variables:

- The variables within the dataset are of various data types, including integers, floats, and objects (categorical). These variable types suggest that the dataset captures a wide range of information, from numerical measurements to categorical characteristics.

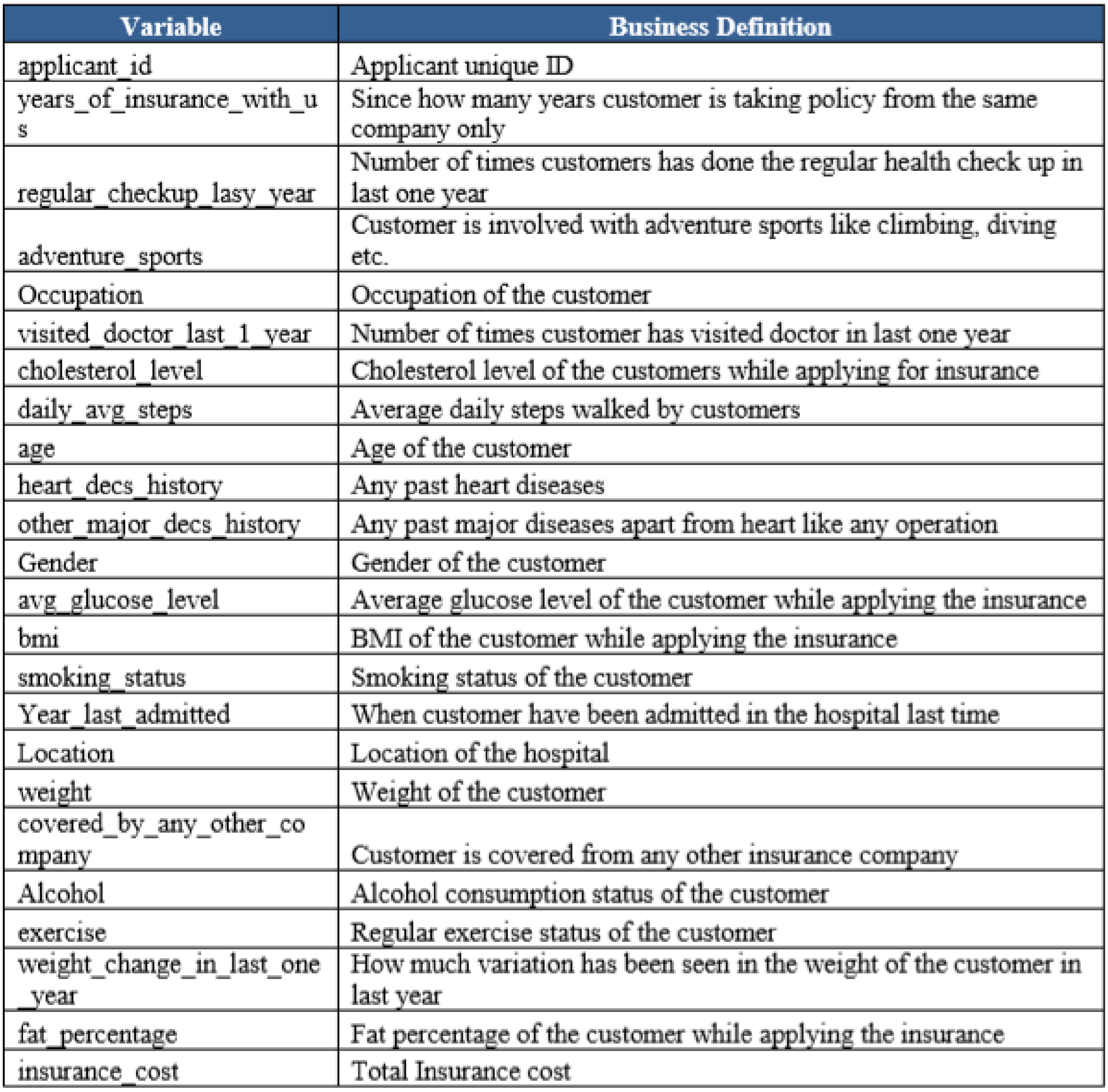
Missing Values:

- Missing values are present in the dataset, particularly in the 'bmi' and 'Year_last_admitted' columns. A total of 990 missing values are observed in the 'bmi' column, while the 'Year_last_admitted' column has 11,881 missing values. This emphasizes the need to handle missing data appropriately during analysis, as these columns could significantly impact any conclusions drawn from the data.

Descriptive Statistics:

- The summary statistics provided for each numerical variable offer valuable insights into the central tendencies, dispersions, and ranges of the data. For instance:
- The average age of applicants is approximately 44.92 years, with a standard deviation of around 16.11 years.
- The average daily average steps taken by applicants is approximately 5215.89, with a standard deviation of around 1053.18 steps.
- The average BMI (Body Mass Index) is approximately 31.39, with a standard deviation of around 7.88.
- These statistics provide a snapshot of the distribution and variability of the data, aiding in identifying potential outliers or anomalies.

c) Understanding of attributes (variable info, renaming if required)



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

Figure 1 Variable Description

- In summary, understanding the attributes involves comprehending the purpose and nature of each variable.
- The variable names generally provide meaningful information. Hence renaming is not necessary for this data.
- Additionally, the presence of missing values in certain variables, such as "bmi" and "Year_last_admitted," underscores the importance of handling them appropriately during analysis.

3. Exploratory data analysis:

a) Univariate analysis (distribution and spread for every continuous attribute, distribution of data in categories for categorical ones)

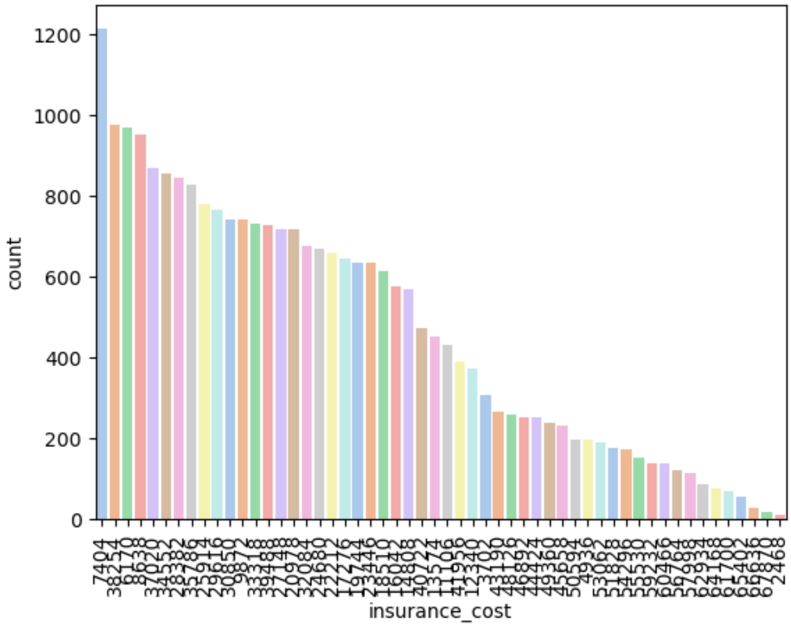


Figure 2 univariate of Insurance cost

- The distribution of insurance costs highlights the frequency of each value. From the distribution, it can be inferred that certain cost values appear more frequently than others.
- An insurance cost of 7404 appears 1214 times.
- A cost of 38254 is observed 977 times.
- The cost 6170 is present 970 times.
- This pattern suggests that there might be specific cost ranges that are more common or preferred among the applicants.

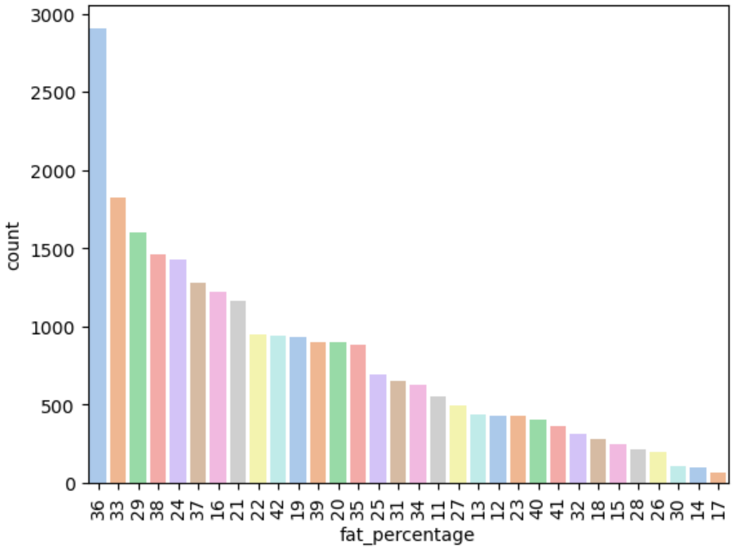


Figure 3 univariate of fat percent

- The distribution of fat percentages reveals the frequency of occurrence for each value. From the distribution, it can be inferred that certain fat percentage values are more prevalent than others.
- A fat percentage of 36 is observed 2908 times.
- A percentage of 33 appears 1828 times.
- The value 29 is present 1604 times.
- This pattern suggests that certain fat percentage ranges might be more common within the dataset.

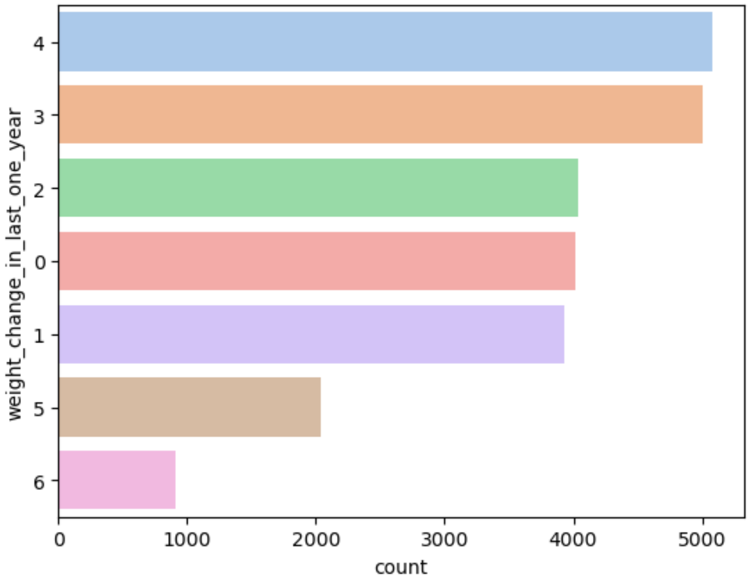


Figure 4 univariate of weight change

- The distribution of weight changes over the past year reveals the frequency of each value. From the distribution, it's evident that certain weight change levels are more common than others.
- A weight change of 4 is observed 5076 times.
- A change of 3 is present 5006 times.
- The value 2 appears 4037 times.
- This pattern suggests that individuals experienced specific levels of weight change more frequently.

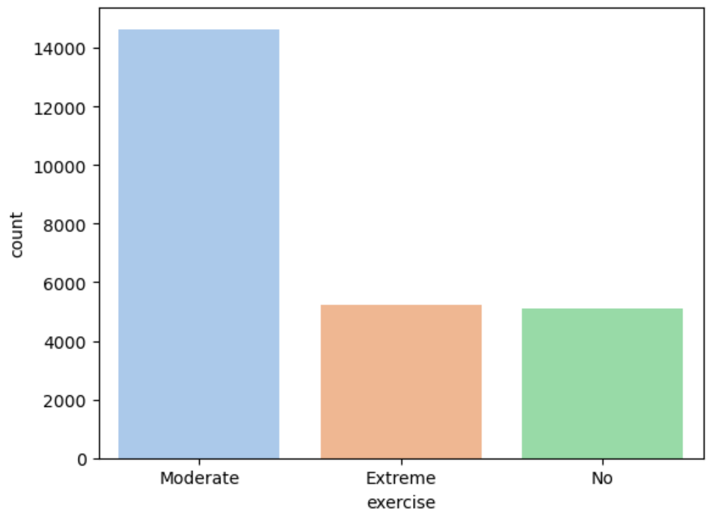


Figure 5 univariate of exercise

- There are 3 unique values within the exercise variable: "Moderate," "Extreme," and "No."
- These unique values represent different categories or levels of exercise routines observed among the individuals.
- The distribution of exercise routines indicates the frequency of each exercise level.
- "Moderate" exercise routines are observed 14,638 times.
- "Extreme" exercise routines are present 5,248 times.
- "No" exercise routines are found 5,114 times.

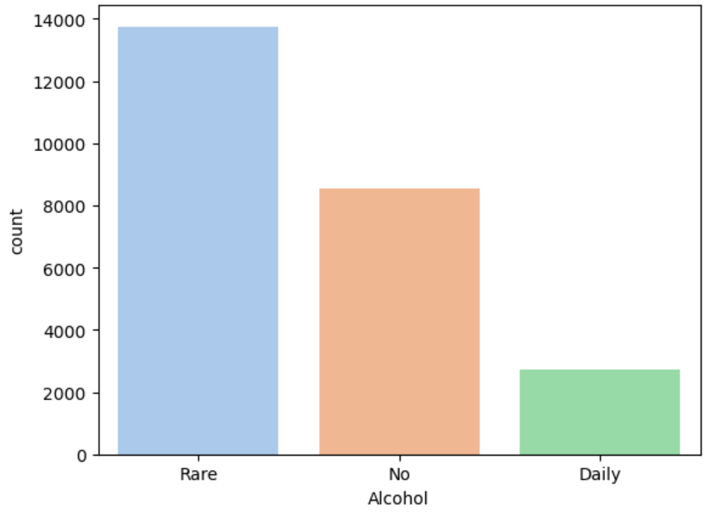


Figure 6univariate of alcohol

- The distribution of alcohol consumption habits indicates the frequency of each category.
- "Rare" alcohol consumption habits are observed 13,752 times.
- "No" alcohol consumption habits are present 8,541 times.
- "Daily" alcohol consumption habits are reported 2,707 times.

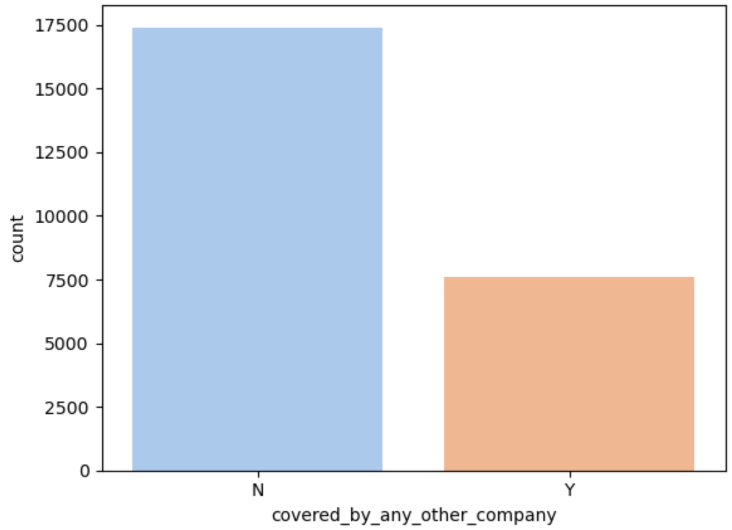


Figure 7 univariate of covered by other

- There are 2 unique values within the covered_by_any_other_company variable: "N" (No) and "Y" (Yes).
- These unique values represent whether individuals are covered by any other insurance company.
- The distribution of coverage status indicates the frequency of each category.
- "N" (No coverage by any other company) is observed 17,418 times.
- "Y" (Covered by another company) is reported 7,582 times.

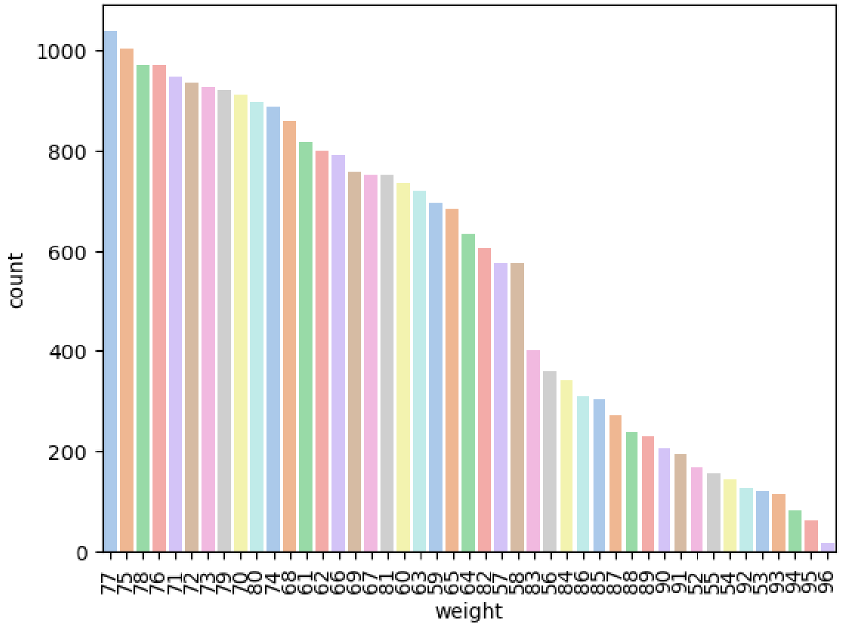


Figure 8 univariate of weight

- The distribution of body weights indicates the frequency of each weight value.
- Certain weight values are more common than others.
- A weight of 77 is observed 1038 times.
- A weight of 75 is reported 1003 times.

- The value 78 appears 970 times.

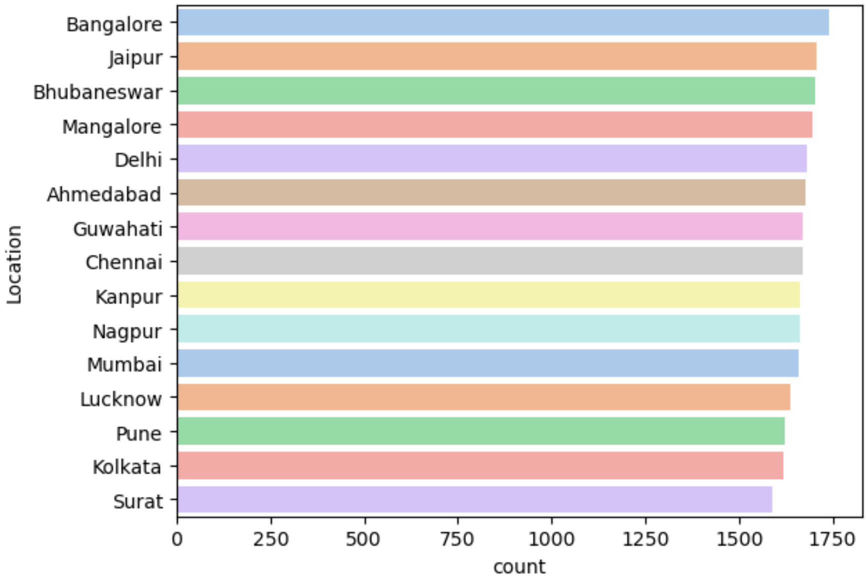


Figure 9 univariate of location

- The distribution of locations indicates the frequency of each location value.
- Certain locations are more prevalent than others.
- "Bangalore" has a frequency of 1742.
- "Jaipur" is observed 1706 times.
- "Bhubaneswar" is present 1704 times.

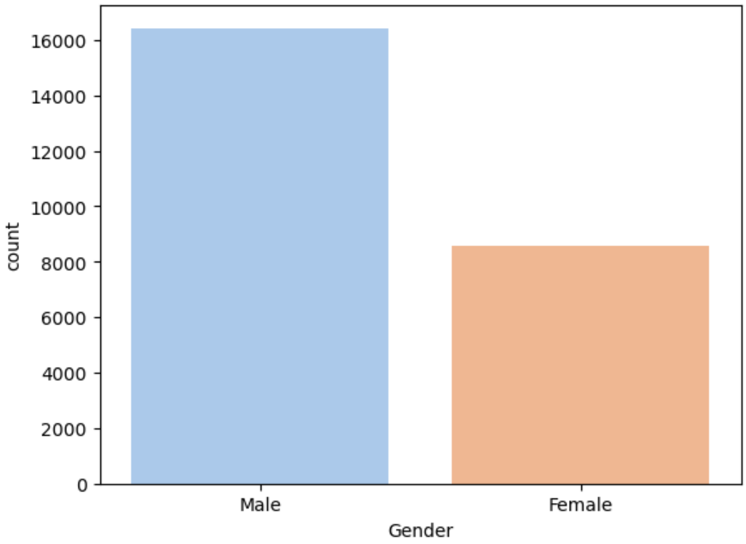


Figure 10 univariate of gender

- The distribution of genders indicates the frequency of each gender category.
- The dataset includes 16,422 instances of "Male" and 8,578 instances of "Female."

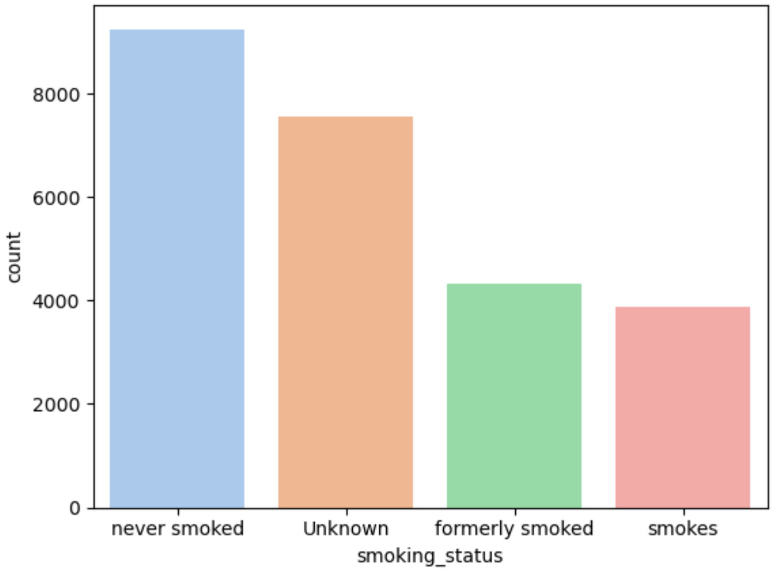


Figure 11 univariate of status of smoking

- The distribution of smoking status indicates the frequency of each smoking habit category.
- Certain smoking statuses are more prevalent than others.
- "never smoked" is observed 9,249 times.
- "Unknown" status is reported 7,555 times.
- The distribution of smoking status provides insights into the smoking behaviors within the dataset.
- It helps in understanding the proportion of individuals with different smoking habits.

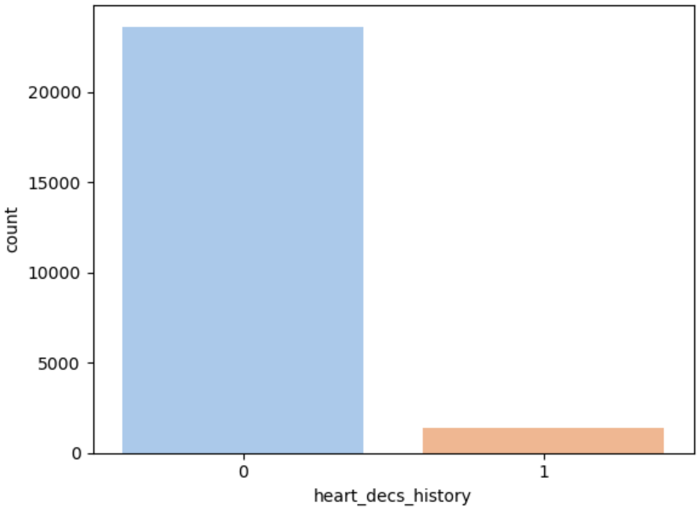


Figure 12 univariate of heart decs

- The distribution of heart disease history indicates the frequency of each category.
- The dataset includes 23,634 instances of individuals without a history of heart disease (0) and 1,366 instances of individuals with a history of heart disease (1).
- The distribution of heart disease history provides insights into the prevalence of heart disease among the individuals.

- It helps in understanding the proportion of individuals with and without a history of heart disease.

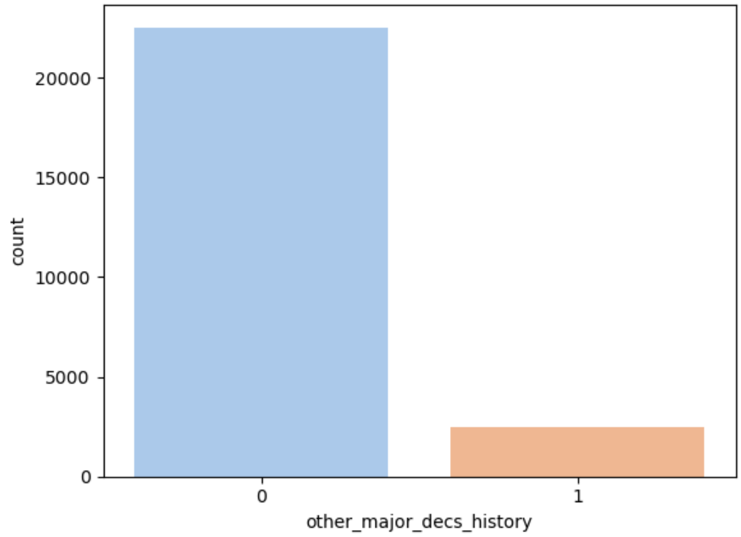


Figure 13 univariate of other decs

- The distribution of other major disease history indicates the frequency of each category.
- The dataset includes 22,546 instances of individuals without a history of other major diseases (0) and 2,454 instances of individuals with a history of other major diseases (1).
- The distribution of other major disease history provides insights into the prevalence of such diseases among the individuals.
- It helps in understanding the proportion of individuals with and without a history of other major diseases.

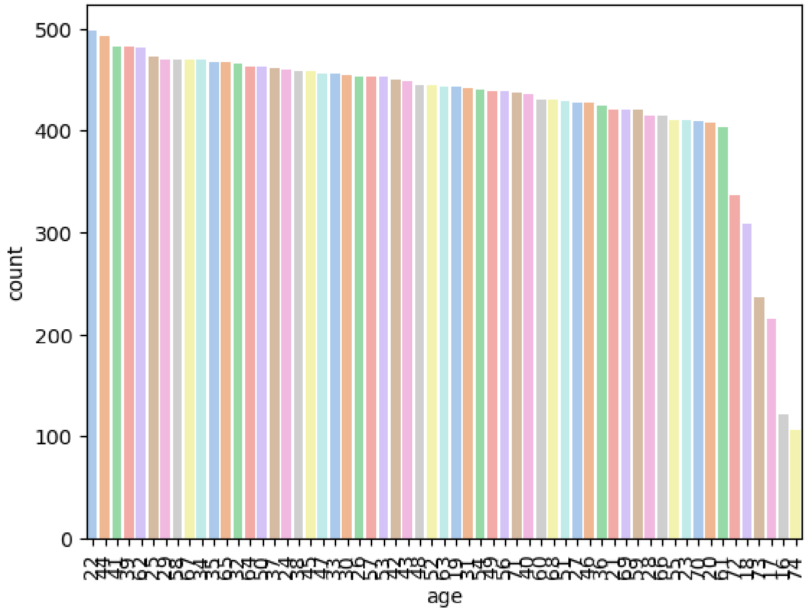


Figure 14 univariate of age

- The distribution of ages indicates the frequency of individuals within each age group.
- Certain age groups are more common than others.
- The age of 22 is observed 498 times.
- The age of 44 appears 493 times.
- The age of 41 is reported 482 times.
- The presence of multiple unique age values reflects the diversity in ages among individuals in the dataset.
- The age range spans from 16 to 74, representing a broad spectrum of life stages.

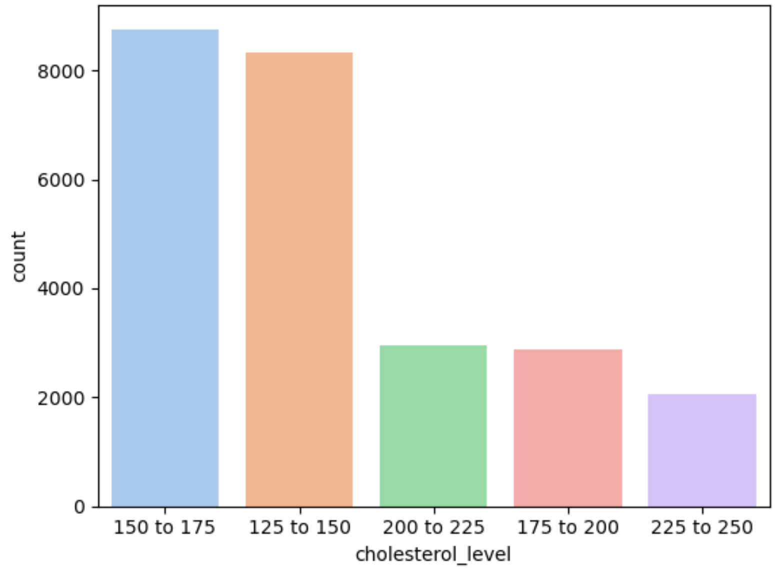


Figure 15 univariate of cholesterol level

- The distribution of cholesterol levels indicates the frequency of individuals within each cholesterol range.
- Cholesterol levels within the ranges are observed with varying frequencies.
- The range "150 to 175" has a frequency of 8,763.
- The range "125 to 150" appears 8,339 times.
- The distribution of cholesterol levels provides insights into the cholesterol makeup of the dataset.
- It helps in understanding the proportion of individuals within different cholesterol ranges.

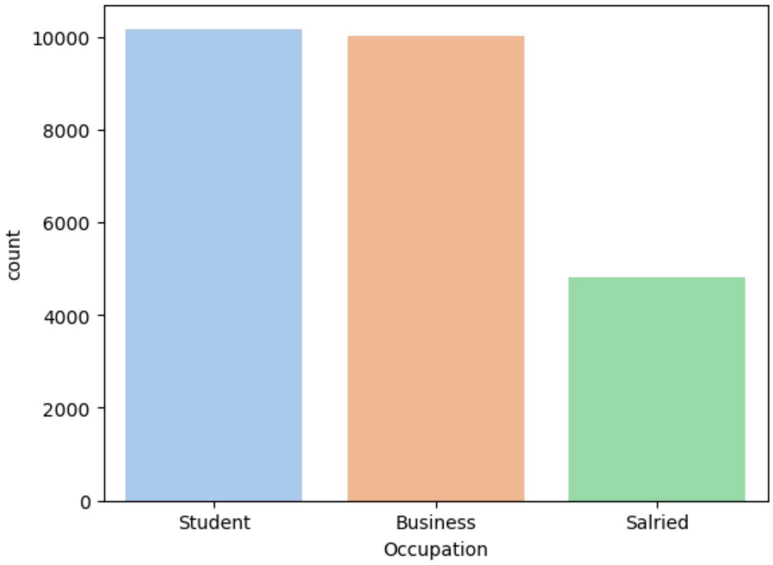


Figure 16 univariate of occupation

- The distribution of occupations indicates the frequency of individuals within each occupational category.
- Among the categories, "Student" is the most common, with 10,169 instances.
- "Business" follows closely with 10,020 instances, while "Salried" is reported 4,811 times.
- The distribution of occupations provides insights into the occupational makeup of the dataset.
- It helps in understanding the proportion of individuals within different occupational categories.

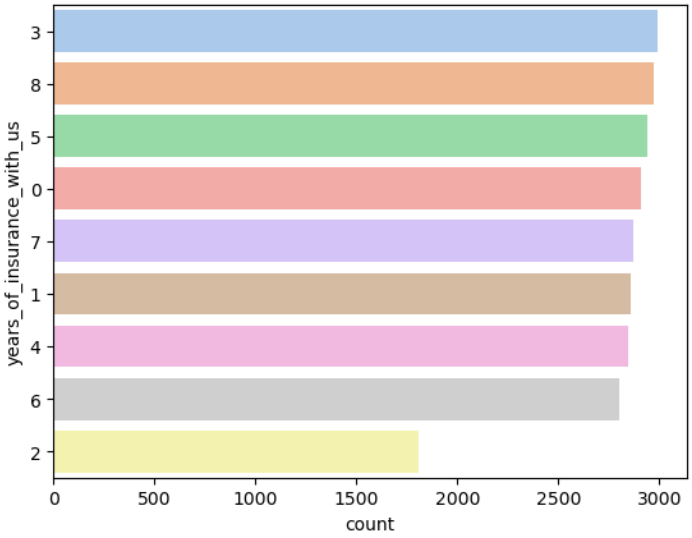


Figure 17 univariate of years of insurance with us

- The distribution of years of insurance indicates the frequency of individuals within each time range.
- The range with the highest frequency is 3 years, with 2,990 instances.

- The range with the second-highest frequency is 8 years, with 2,970 instances.

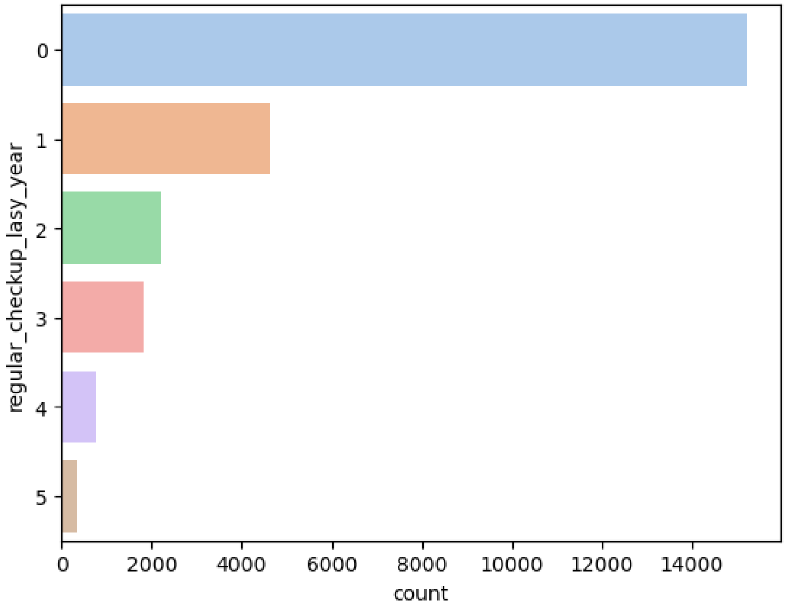


Figure 18 univariate of regular checkup last year

- The distribution of regular checkups indicates the frequency of individuals within each checkup level.
- The level with the highest frequency is "0," indicating no regular checkup, with 15,215 instances.
- The level with the second-highest frequency is "1," indicating one regular checkup, with 4,644 instances.

b) Bivariate analysis (relationship between different variables , correlations)

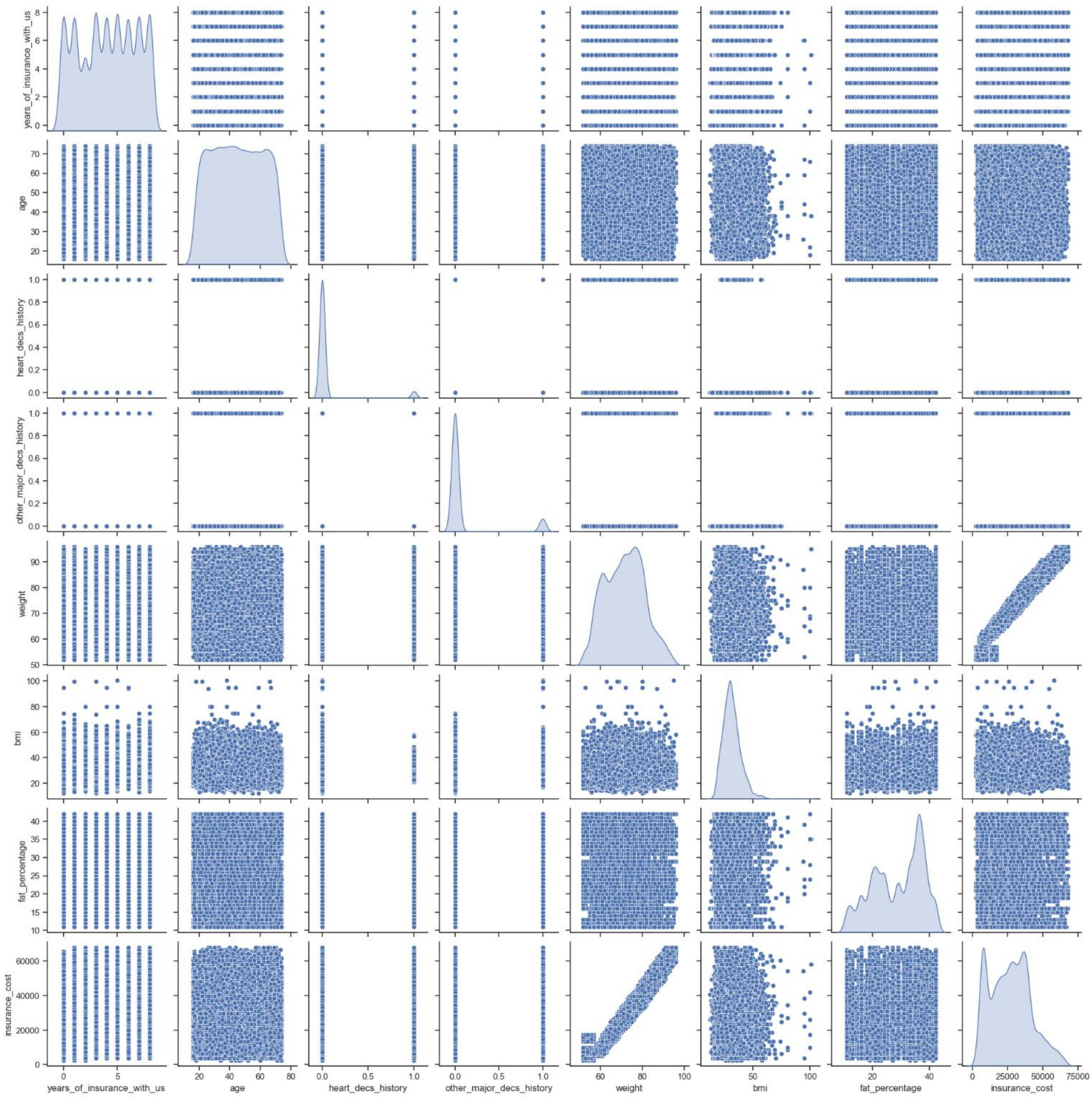


Figure 19 Bivariate Analysis

- Insurance Cost and Count Relationship:
- The observation suggests that there's a slight decrease in the count of applicants as the insurance cost increases.
- This relationship could indicate that higher insurance costs might deter some applicants or lead to a decrease in the number of applicants at higher premium levels.

- It's important to explore the reasons behind this trend, such as affordability concerns or other factors influencing insurance decisions.
- Fat Percentage and Insurance Cost:
- The observation highlights that most applicants have a fat percentage ranging from 20% to 40%, with a significant portion at 40%.
- However, the relationship between fat percentage and insurance cost is not explicitly mentioned. It might be valuable to analyze whether there's any correlation or pattern between fat percentage and insurance cost.
- Past Heart Disease and Other Major Decs:
- The observation indicates that the majority of applicants have a history of no past heart disease.
- Additionally, it's mentioned that the frequency of "Other major decs" is almost equal to the frequency of past heart disease.
- This information raises questions about the nature of these "Other major decs" and whether they are contributing to the health profile of the applicants.
- Years of Insurance with Us and Age:
- The observation suggests that the "years of insurance with us" variable is similar across different age groups.
- This could indicate that the length of the insurance relationship is not significantly impacted by the age of the applicants.
- It might be interesting to explore whether there's a correlation between the years of insurance and other variables like insurance cost or health indicators.
- Weight and Insurance Cost:
- The observation indicates a positive relationship between weight and insurance cost, stating that when the weight increases, the insurance cost also increases.
- This relationship could imply that higher weight might lead to higher insurance premiums, possibly due to the increased health risks associated with obesity.
- It's important to further analyze the strength and significance of this relationship and whether other factors also play a role.

a) Removal of unwanted variables (if applicable)

- Based on the provided data and the context of analysis, “Year_last_admitted” variable can be droped from the above data
- This variable has a significant number of missing values (around 47% missing) and might not be as informative due to its incomplete nature. Unless you have a specific research question related to this variable, it might not provide accurate insights.

b) Missing Value treatment (if applicable)

- bmi missing values are filled with the mean of the available values in that column. This approach can be useful if the variable has a continuous distribution and the mean is an appropriate representative value.

d) Outlier treatment (if required)

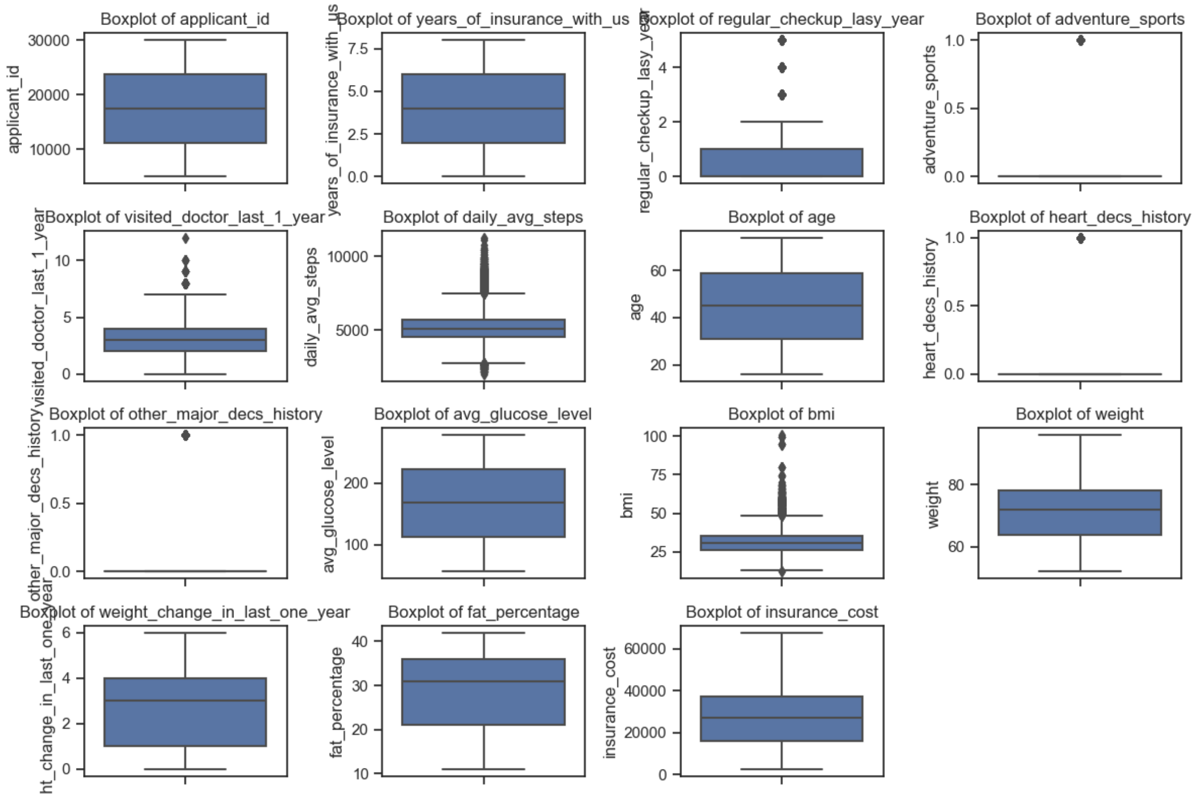


Figure 20 Before treating outliers

- The observation suggests that several variables in the dataset have outliers, as indicated by the presence of data points outside the typical range. These variables include:
- Regular Checkup Last Year: Outliers might indicate instances where the regular checkup frequency significantly deviates from the norm. It's important to investigate these outliers to understand the reasons behind such deviations.
- Adventure Sports: The presence of outliers might indicate individuals with an unusually high or low participation rate in adventure sports. Exploring these outliers could provide insights into extreme behavior patterns.
- Visited Doctor Last 1 Year: Outliers in this variable might point to individuals who either visited the doctor extremely frequently or not at all in the past year. Understanding these cases could reveal specific health conditions or habits.
- Daily Average Steps: Outliers in this variable could represent individuals with an exceptionally high or low activity level. These outliers might be due to specific activities or circumstances affecting step count.
- Heart Decs History and Other Major Decs: Outliers in these variables could indicate individuals with unique medical histories or conditions that deviate from the norm.
- BMI: Outliers in BMI might indicate extreme body mass index values. Investigating these outliers could provide insights into the health profile of individuals with unusual BMI values.

- The approach taken to address these outliers involved outlier treatment, which likely includes transformations or removals. The effectiveness of the treatment can be observed through box plots, showing the distribution before and after outlier treatment.
- This intervention can enhance the data quality and the validity of subsequent analyses, contributing to more accurate insights and conclusions. However, it's crucial to perform in-depth analysis and consider the domain knowledge when handling outliers to ensure that they are addressed appropriately without removing valuable information.

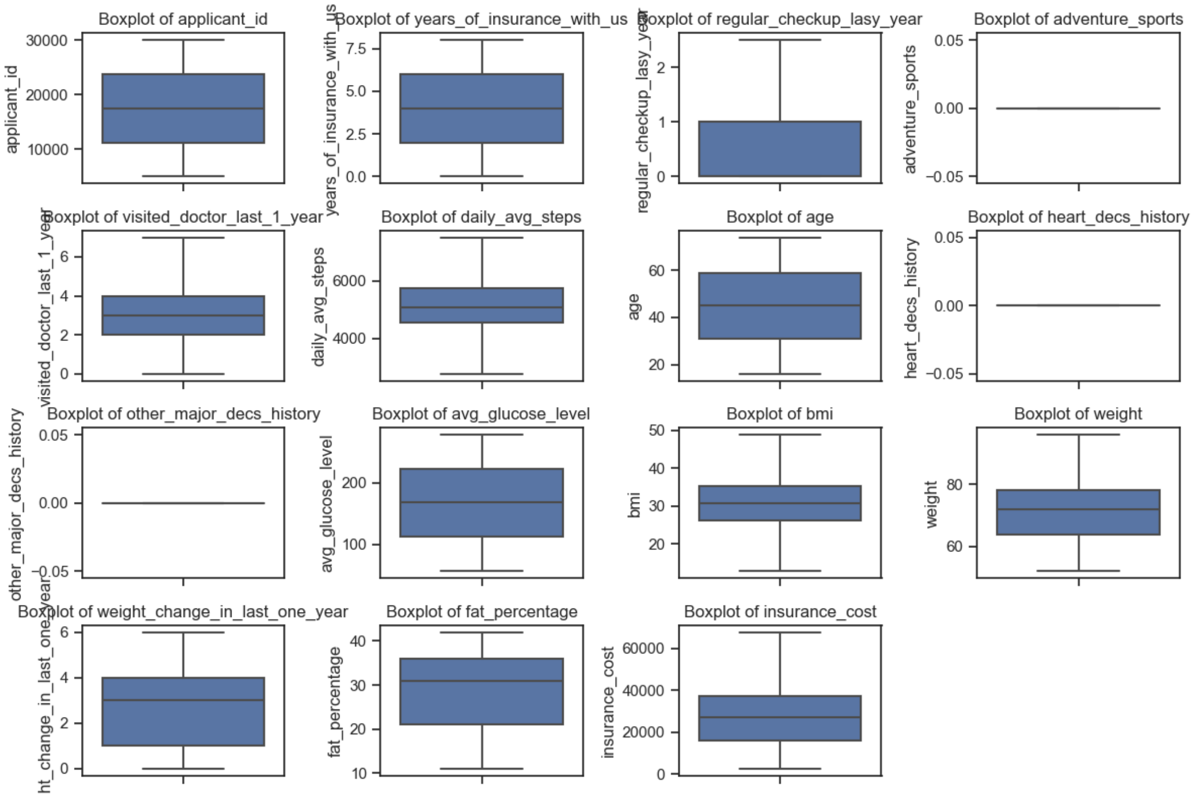


Figure 21 After treating outliers

e) Variable transformation (if applicable)

- The observation that certain variables like "Daily Average Steps, Age, Heart Decs History, Other Major Decs History, Avg Glucose Level, BMI, Weight, Weight Change in Last One Year, Fat Percentage, Insurance Cost" are already in numerical format signifies that these variables are directly quantifiable and measured on a numeric scale. As such, they inherently possess the attributes required for various analytical methods and models without needing any specific variable transformation.

f) Addition of new variables (if required)

- The absence of any mention or suggestion for adding new variables in the dataset implies that the current variables available in the dataset are deemed sufficient for the analysis and objectives at hand. This decision could indicate that the existing variables provide comprehensive information and adequately capture the relevant aspects of the domain being studied.

4. Business insights from EDA:

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

- The concept of data balance pertains to the distribution of target or outcome categories within a dataset. If the data is unbalanced, it means that one category significantly outweighs the others. In the context of the business, an unbalanced dataset can lead to skewed analyses and predictions, particularly if the minority class is of high interest.
- If the dataset contains health insurance data, and the target variable represents whether a policyholder claimed insurance benefits (binary outcome: claimed or not claimed), an imbalanced dataset with a majority of non-claimed cases might lead to a model that performs well in predicting non-claims but poorly in predicting claims. This scenario could have significant business implications, as the goal is likely to identify factors that lead to claims.
- To address data imbalance, techniques such as oversampling, undersampling, or generating synthetic samples using SMOTE (Synthetic Minority Over-sampling Technique) could be employed. This rebalancing helps ensure that the model is exposed to sufficient examples of the minority class, improving its ability to generalize to both classes and make accurate predictions.

b) Any business insights using clustering (if applicable)

- Clustering involves grouping similar data points together. In a business context, clustering can provide valuable insights by identifying customer segments, market segments, or other meaningful groupings within the data. For example, in healthcare, clustering could help identify distinct patient groups based on health metrics and behaviors.
- Using clustering algorithms like k-means or hierarchical clustering on health data, the business could discover different health profiles among customers, allowing for targeted marketing campaigns or customized treatment approaches.

c) Any other business insights

- Exploratory Data Analysis (EDA) may reveal patterns or relationships that lead to meaningful business insights.
- Correlations between variables might unveil relationships between health metrics and insurance claims, influencing policy pricing strategies.
- Age distribution might highlight a significant portion of elderly customers, leading to the development of tailored senior-focused insurance products.
- Insights from variable distributions could guide decision-making in designing fitness and wellness programs for different age groups.
- Overall, EDA goes beyond mere data exploration; it serves as a foundation for informed business decisions by uncovering trends, identifying outliers, revealing patterns, and suggesting hypotheses that can drive strategies and actions.