HEALTH CARE PROJECT

BUSSINESS REPORT

FINAL REPORT

Pavan Ambare
PGP-Data science and Business analytics
PGPDSBA.O.MAY23.A
Date-26/05/2024

Table of Contents

Business Problem:

We all know that Health care is very important domain in the market. It is directly linked with the lifeof 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 forthe estimated cost of insurance

File: Data.csv

Target variable: insurance_cost

Data dictionary:

Variable	Business Definition
applicant_id	Applicant unique ID
years_of_insurance_with_us	Since how many years customer is taking policy from the samecompany only
regular_checkup_lasy_year	Number of times customers has done the regular health check up inlast one year
adventure_sports	Customer is involved with adventure sports like climbing, divingetc.
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 inlast year
fat_percentage	Fat percentage of the customer while applying the insurance
insurance_cost	Total Insurance cost

2) Introduction of the business problem.

a) Defining problem statement.

- The problem statement is to build a model that can provide the optimum insurance cost for an individual based on health and habit-related parameters. This involves predicting the insurance cost using a dataset containing information such as the customer's years of insurance with the company, regular checkup frequency, and involvement in adventure sports, occupation, medical history, gender, BMI, smoking status, and other relevant factors.

b) Need of the study/project.

- The healthcare domain is critical, directly impacting individuals' lives and financial well-being. With rising healthcare costs, it's crucial for insurance companies to optimize insurance costs while ensuring adequate coverage for customers. By developing a model to estimate insurance costs based on various factors, insurance companies can better manage their risk and offer more competitive and tailored insurance plans to customers.

c) Understanding business/social opportunity.

- This project presents a significant business opportunity for insurance companies to enhance their pricing strategies and risk management practices. By leveraging data analytics and predictive modeling, insurance companies can gain deeper insights into their customers' health profiles and behaviors, enabling them to offer personalized and cost-effective insurance solutions. Additionally, by promoting healthy lifestyles and preventive care, this project can contribute to the social goal of improving overall public health and reducing healthcare costs.

3) Basic Analysis of the HealthCare Dataset

1. Performing basic analysis and printing top 5 rows (Head).

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age
0	5000	3	1	1	Salried	2	125 to 150	4866	28
1	5001	0	0	0	Student	4	150 to 175	6411	50
2	5002	1	0	0	Business	4	200 to 225	4509	68
3	5003	7	4	0	Business	2	175 to 200	6214	51
4	5004	3	1	0	Student	2	150 to 175	4938	44

2. Performing basic analysis and printing bottom 5 rows (tail).

	applicant_id	years_of_insurance_with_us	regular_checkup_lasy_year	adventure_sports	Occupation	visited_doctor_last_1_year	cholesterol_level	daily_avg_steps	age
24995	29995	3	0	0	Salried	4	225 to 250	5614	22
24996	29996	6	0	0	Business	4	200 to 225	4719	58
24997	29997	7	0	1	Student	2	150 to 175	5624	34
24998	29998	1	0	0	Salried	2	225 to 250	10777	27
24999	29999	8	2	0	Business	4	150 to 175	5882	22

3. Performing basic analysis and printing shape of dataset.

Number of rows= **25000**

Number of columns= **24**

4. Performing basic analysis and checking for the null values in dataset.

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

- There are 990 null values in bmi and 11881 null values in year last admitted columns.

5. Performing basic analysis & printing summary of dataset for numerical columns.

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

6. Performing basic analysis and checking for the duplicates values in dataset.
    - There are no duplicate rows in dataset.

7. Performing basic analysis and printing information of dataset.

```
 #   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)
```

4) Data Report

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

- Since dataset contains information such as the **year_last_admitted we have Data from 1990 to 2018** and **years_of_insurance_with_us 0 to 8 years** we can assume that the data was collected over a period of time, possibly spanning several years.
- Variables like **regular_checkup_last_year** and **visited_doctor_last_1_year** indicate data collected within the last year, showing a periodic data collection approach.
- It can be inferred that the data collection process involved individuals providing information about their health, habits, and other relevant details during the insurance application process or health check-up visits.

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

- The dataset consists of 25,000 rows and 24 columns.
- Each row represents an individual applicant for insurance
- Each column represents a specific attribute or feature related to the applicant.
- Descriptive statistics for numerical variables provide insights into the distribution of data:
- **years_of_insurance_with_us:** Mean of 4.09 years with a standard deviation of 2.61.
- **regular_checkup_last_year:** Mean of 0.77 checkups with a standard deviation of 1.20.
- **daily_avg_steps:** Mean of 5215.89 steps with a standard deviation of 1053.18.
- **age:** Mean age of 44.92 years with a standard deviation of 16.11.
- **avg_glucose_level:** Mean glucose level of 167.53 with a standard deviation of 62.73.
- **bmi:** Mean BMI of 31.36 with a standard deviation of 7.72.
- **weight:** Mean weight of 71.61 with a standard deviation of 9.33.
- **weight_change_in_last_one_year**: Mean weight change of 2.52 with a standard deviation of 1.69.
- **fat_percentage:** Mean fat percentage of 28.81 with a standard deviation of 8.63.
- **insurance_cost:** Mean insurance cost of 27147.41 with a standard deviation of 14323.69.

- Categorical variables such as Occupation, Gender, smoking_status, Location, covered_by_other_insurance, Alcohol, and exercise have different levels with varying frequencies, showing the diversity of the dataset

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

- **Applicant ID:** This is a unique identifier for each applicant. This column should be dropped as it is not needed for analysis and modeling.
- **Years of Insurance with the Company**: This is numerical variable indicates the number of years the applicant has been with the insurance company.
- **Regular Checkup Lasy Year:** This is numerical variable represents the number of regular health checkups the applicant had in the last year. 6 unique values are present ranging from 0 to 5.we need to convert this column to the categorical which will be helpful for our analysis.
  **If we can predefine it to have a set of values it is always good it will make the model interpretation and analysis better.
- **Adventure Sports:** This binary variable indicates whether the applicant is involved in adventure sports.
- **Occupation:** This categorical variable represents the occupation of the applicant. The values are ‘Salried’,’Student’, ‘Business’.
- **Visited Doctor Last 1 Year:** This is numerical variable indicates the number of times the applicant visited a doctor in the last year. 12 unique values are present ranging from 0 to 12.we need to convert this column to the categorical which will be helpful for our analysis.
- **Cholesterol Level Category:** This categorical variable represents the cholesterol level of the applicant. This column has to be encoded.
- **Daily Average Steps:** This is numerical variable represents the average daily steps walked by the applicant.
- **Age:** This is numerical variable represents the age of the applicant.
- **Heart Disease History**: This binary variable indicates whether the applicant has a history of heart diseases.
- **Other Major Disease History:** This binary variable indicates whether the applicant has a history of major diseases other than heart diseases.

- **Gender:** This categorical variable represents the gender of the applicant.
- **Average Glucose Level:** This is numerical variable represents the average glucose level of the applicant.
- **BMI:** This is numerical variable represents the Body Mass Index (BMI) of the applicant. This column has 3.96 % of the missing values and we need to drop the rows with BMI NA this will not effect on our analysis because we still have 96% of the real data for analysis and model. Imputing Synthetic data is not good for analysis and modeling.
- **Smoking Status:** This categorical variable represents the smoking status of the applicant. 'Unknown', 'formerly smoked', 'never smoked', 'smokes'.
- **Year Last Admitted:** This variable represents the year when the applicant was last admitted to the hospital. This column has 47.52% of missing values.
  If any column has more than 30% missing values when we impute more than 30% the missing values of the data it becomes synthetic data which is not good for analysis and modeling.
- **Location:** This categorical variable represents the location of the hospital where the applicant was last admitted.
- **Weight:** This is numerical variable represents the weight of the applicant.
- **Covered by Other Insurance:** This binary variable indicates whether the applicant is covered by any other insurance company.
- **Alcohol:** This categorical variable represents the alcohol consumption status of the applicant. 'Rare', 'Daily', 'No'.
- **Exercise:** This categorical variable represents the regular exercise status of the applicant. 'Moderate', 'Extreme', 'No'.
- **Weight Change in Last One Year:** This is numerical variable represents the variation in the weight of the applicant in the last year. 7 unique values are present ranging from 0 to 6.we need to convert this column to the categorical which will be helpful for our analysis.
- **Fat Percentage:** This is numerical variable represents the fat percentage of the applicant.
- **Insurance Cost:** This is numerical column represents the total insurance cost of the each customer. This is the target column and when the target column is numerical we can build the linear regression model.

- Renamed column **Regular Checkup Lasy Year = Regular checkup Last Year.**

5) Exploratory data analysis

A. Removal of unwanted variables

- **Applicant ID:** This is a unique identifier for each applicant. This column is dropped as it is not needed for analysis and modeling.
- **Year Last Admitted:** This variable represents the year when the applicant was last admitted to the hospital.
  This column has 47.52% of missing values.
  If any column has more than 30% missing values when we impute more than 30% the missing values of the data it becomes synthetic data which is not good for analysis and modeling. Hence this column is dropped.

B. Missing Value treatment

- **BMI:** This is numerical variable represents the Body Mass Index (BMI) of the applicant. This column has 3.96 % (990) of the missing values and we need to drop the rows with BMI NA this will not effect on our analysis because we still have 96% of the real data for analysis and model. Imputing Synthetic data is not good for analysis and modeling.

Info of dataset and checking for null values in the dataset after dropping the rows with BMI NA & removing the Applicant ID & Year Last Admitted column.

```
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   years_of_insurance_with_us    24010 non-null  int64
 1   regular_checkup_lasy_year     24010 non-null  int64
 2   adventure_sports              24010 non-null  int64
 3   Occupation                    24010 non-null  object
 4   visited_doctor_last_1_year    24010 non-null  int64
 5   cholesterol_level             24010 non-null  object
 6   daily_avg_steps               24010 non-null  int64
 7   age                           24010 non-null  int64
 8   heart_decs_history            24010 non-null  int64
 9   other_major_decs_history      24010 non-null  int64
 10  Gender                        24010 non-null  object
 11  avg_glucose_level             24010 non-null  int64
 12  bmi                           24010 non-null  float64
 13  smoking_status                24010 non-null  object
 14  Location                      24010 non-null  object
 15  weight                        24010 non-null  int64
 16  covered_by_any_other_company  24010 non-null  object
 17  Alcohol                       24010 non-null  object
 18  exercise                      24010 non-null  object
 19  weight_change_in_last_one_year  24010 non-null  int64
 20  fat_percentage                24010 non-null  int64
 21  insurance_cost                24010 non-null  int64
dtypes: float64(1), int64(13), object(8)
```

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

C. "Discretization" or "binning."

Discretization involves dividing a continuous variable into categories or bins. This is done for reasons, such as to simplify the analysis, handle non-linearity, or prepare the data for certain machine learning algorithms that work better with categorical variables.

I. Conversion of 'regular_checkup_last_year' from a numerical to a categorical type

- This is numerical variable represents the number of regular health checkups the applicant had in the last year. 6 unique values are present ranging from 0 to 5.we need to convert this column to the categorical which will be helpful for our analysis.
  **If we can predefine it to have a set of values it is always good it will make the model interpretation and analysis better and if in future any data which comes 5 and greater than 5 we can floor it to 5 that means data greater 5 is also considered as 5.
- Transformed the 'regular_checkup_last_year' variable from a numerical to a categorical type using binning. Binning is a process of dividing a continuous variable into discrete intervals or bins. In this case, we divided the numerical values into six bins ('Very Low', 'Low', 'Medium', 'High', 'Very High', 'Extremely High') based on their value ranges. This transformation allows us to analyze the variable as categories, which can be more interpretable and useful for analysis and modeling.

II. Conversion of 'visited_doctor_last_1_year' from a numerical to a categorical type

- This is numerical variable indicates the number of times the applicant visited a doctor in the last year. 12 unique values are present ranging from 0 to 12.we need to convert this column to the categorical which will be helpful for our analysis 7 visualization.
- Converted the 'visited_doctor_last_1_year' variable from a numerical to a categorical type by casting it to a categorical data type. This transformation categorizes the variable based on the unique numerical values present in the dataset. Each unique numerical value is represented as a category, allowing us to analyze the variable in a categorical context. This conversion enhances the interpretability of the variable and facilitates its use in categorical analysis and modeling.

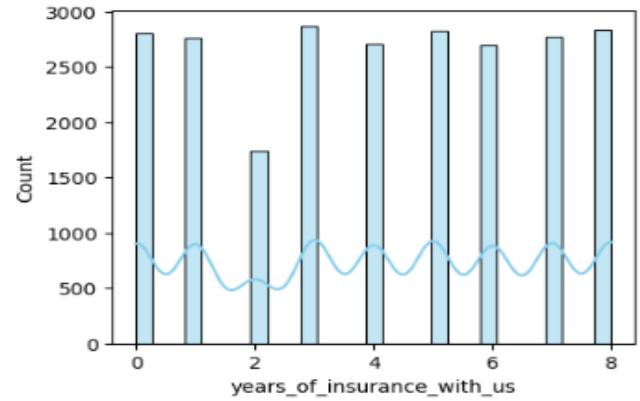
III. Conversion of 'weight_change_in_last_one_year' from a numerical to a categorical type

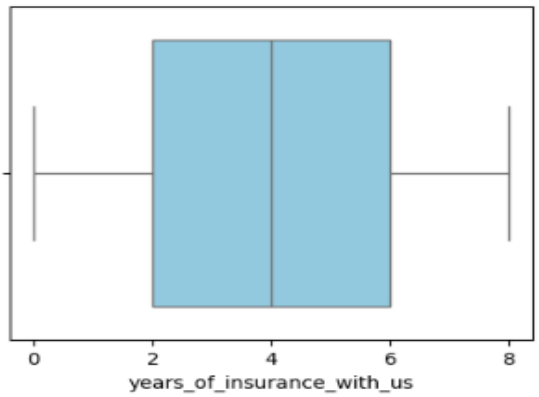
- This is numerical variable represents the variation in the weight of the applicant in the last year. 7 unique values are present ranging from 0 to 6.we need to convert this column to the categorical which will be helpful for our analysis.
- Transformed the 'weight_change_in_last_one_year' variable from a numerical to a categorical type using the pandas 'Categorical' function. This conversion categorizes the variable based on the unique numerical values present in the dataset, with each unique numerical value representing a category. This transformation enhances the interpretability of the variable and enables us to analyze it in a categorical context, which can be beneficial for analysis and modeling.

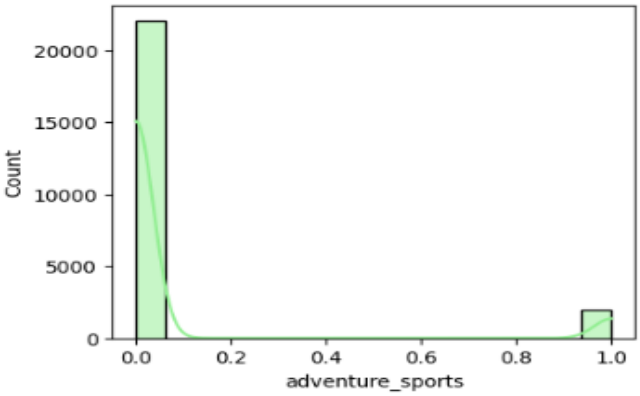
IV. Encoding of the 'cholesterol_level' column:

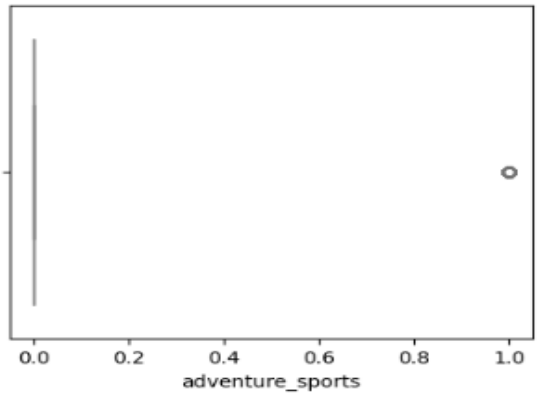
- Encoded the 'cholesterol_level' column to convert the categorical ranges into numerical values. Each categorical range ('125 to 150', '150 to 175', '175 to 200', '200 to 225', '225 to 250') was mapped to a corresponding numerical value (1, 2, 3, 4, 5) using a predefined encoding dictionary. This encoding allows us to represent the cholesterol levels numerically, making it easier to analyze and model the data in a quantitative manner.

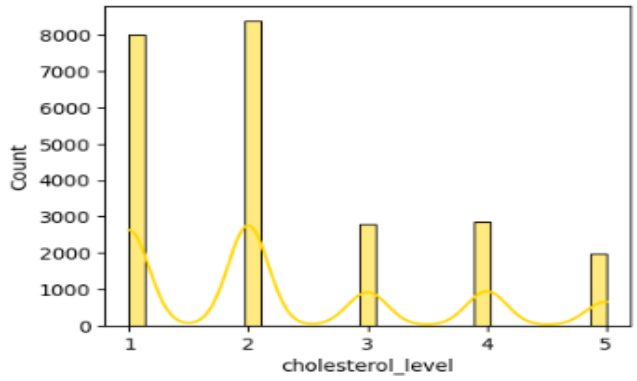
D. Univariate Analysis of continuous variables

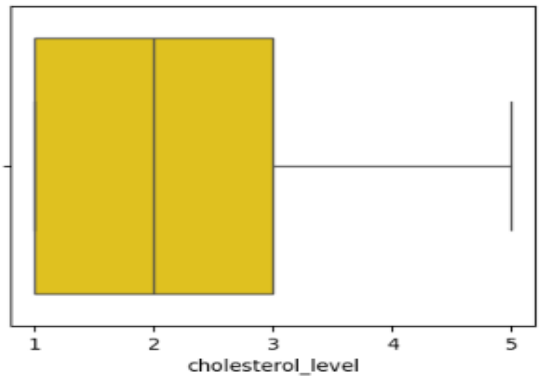


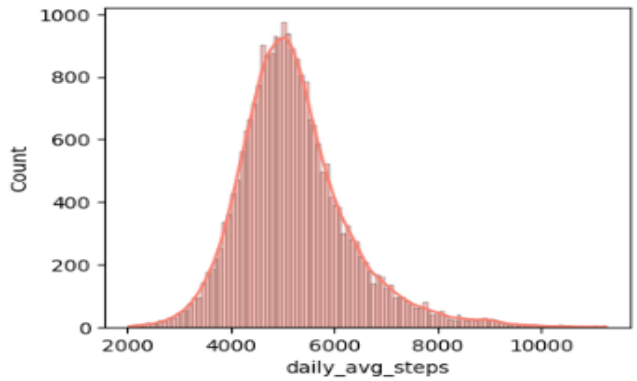


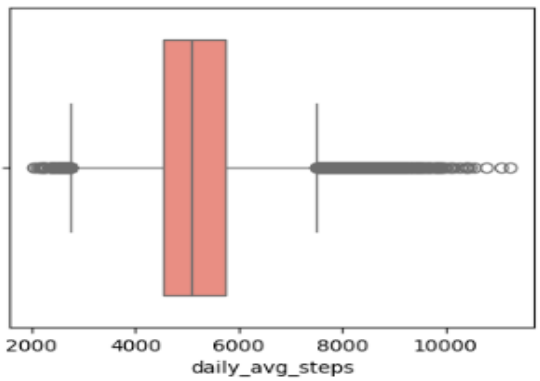


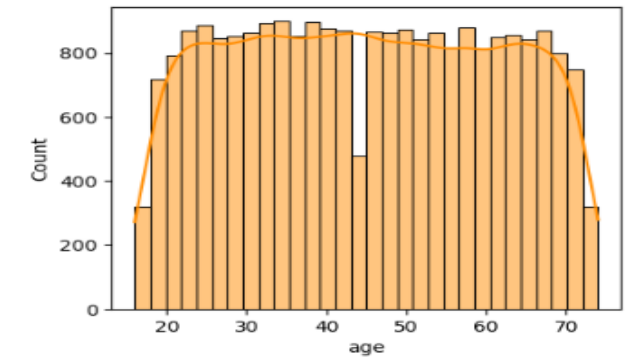


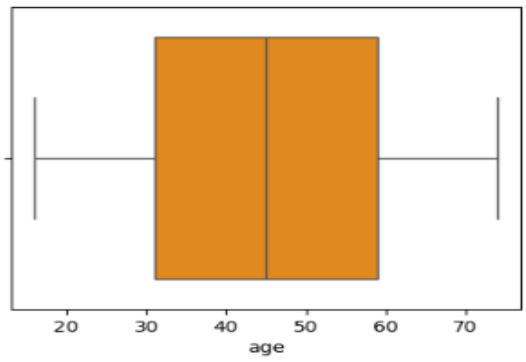


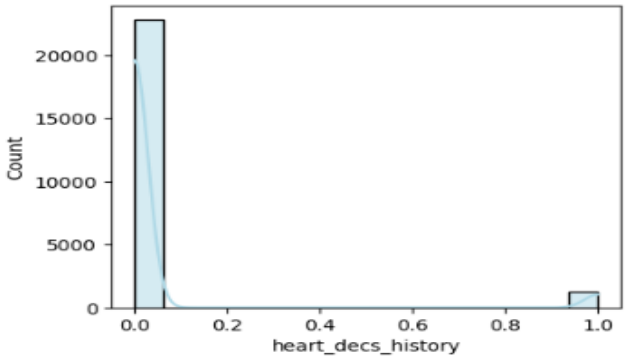


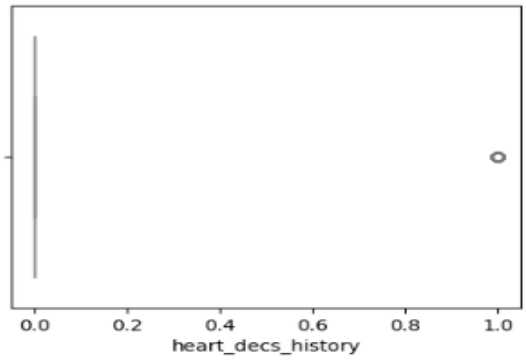


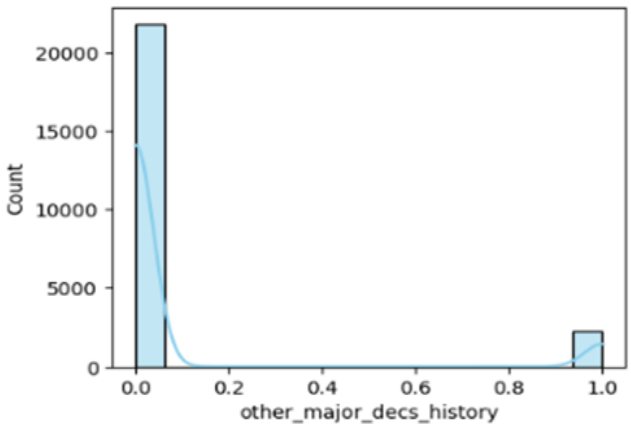


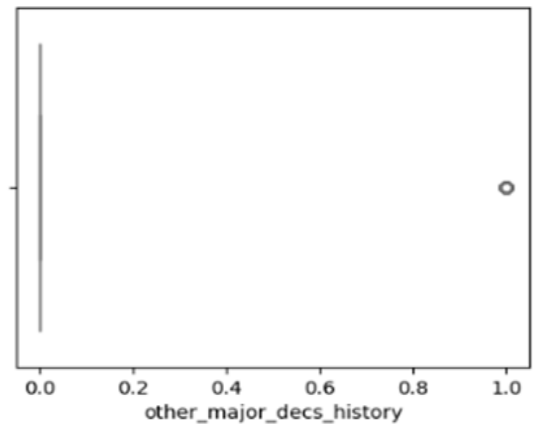

Count
800
600
400
200
0
20
30
40
50
60
70
age

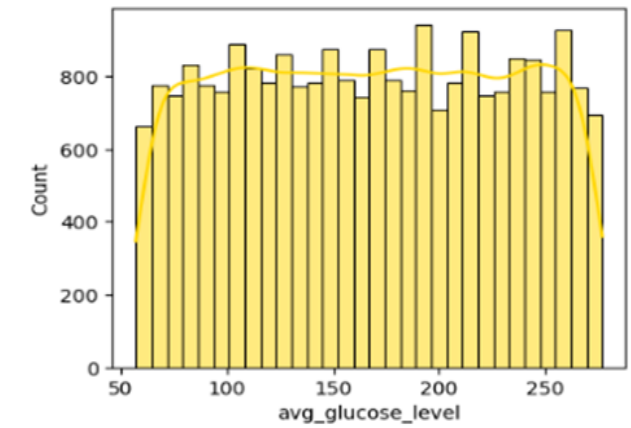

20
30
40
50
60
70
age

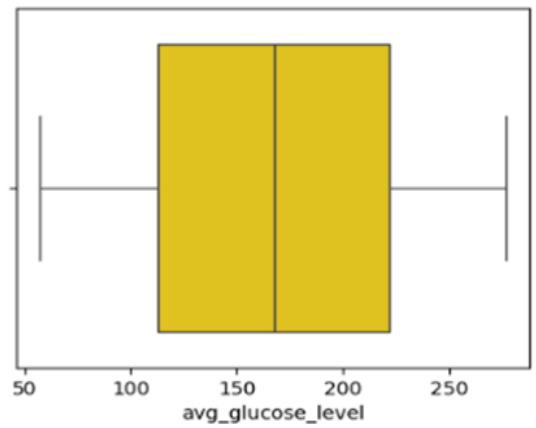

Count
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
heart_decs_history

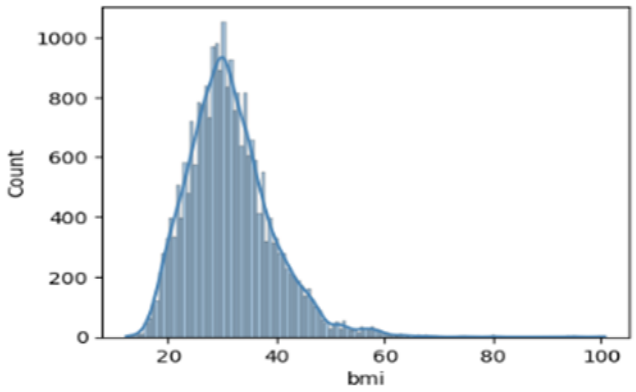

0.0
0.2
0.4
0.6
0.8
1.0
heart_decs_history

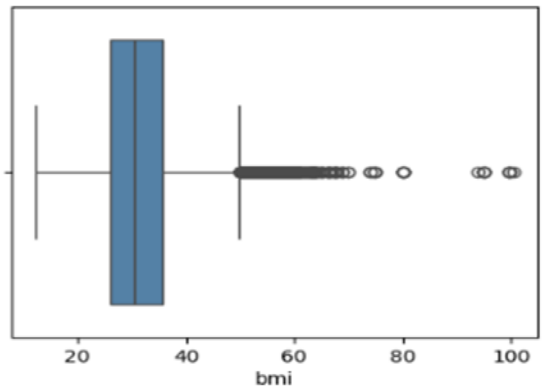

Count
20000
15000
10000
5000
0
0.0
0.2
0.4
0.6
0.8
1.0
other_major_decs_history

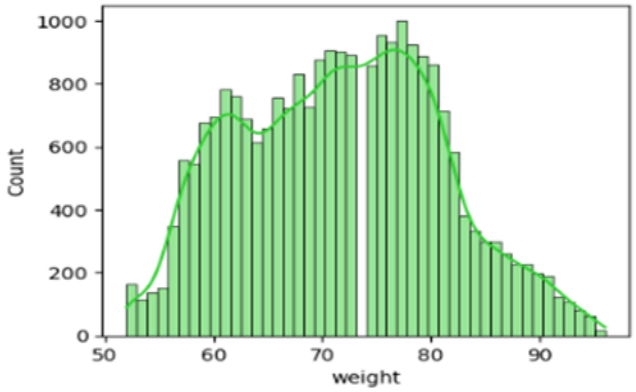

0.0
0.2
0.4
0.6
0.8
1.0
other_major_decs_history

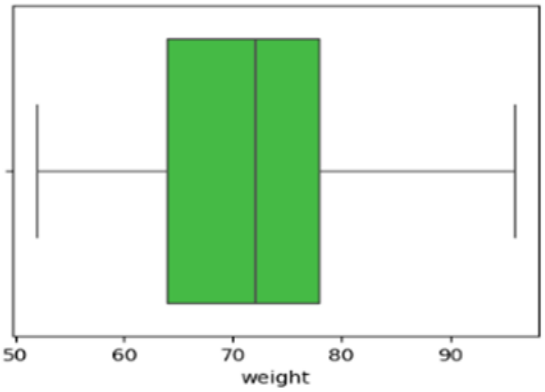

Count
800
600
400
200
0
50
100
150
200
250
avg_glucose_level

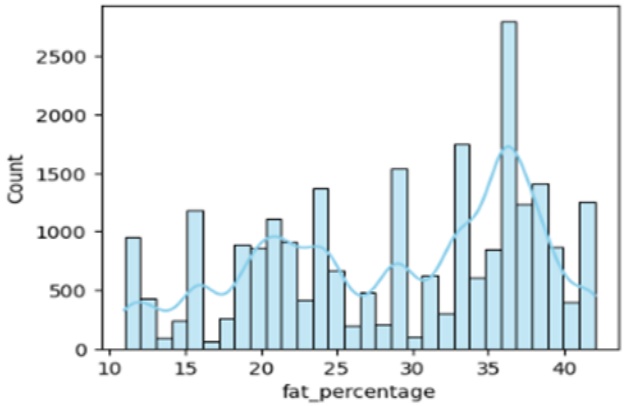

50
100
150
200
250
avg_glucose_level

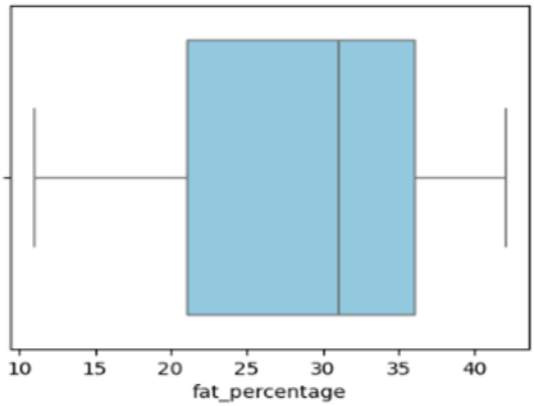

Count
bmi

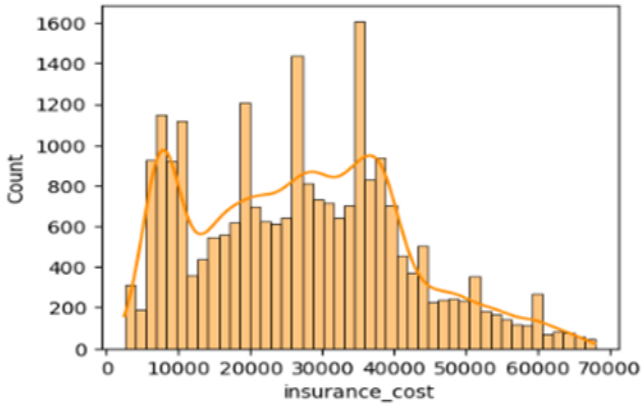

bmi

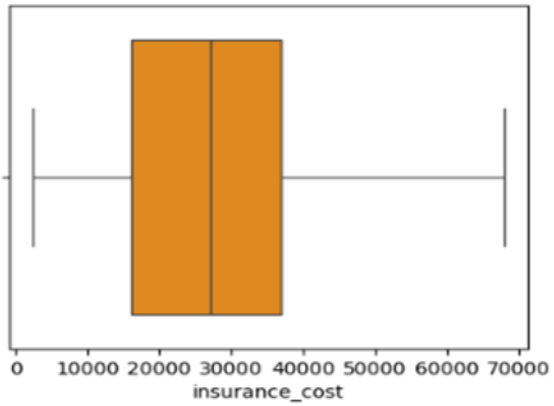

Count
weight


weight


Count
fat_percentage


fat_percentage


Count
insurance_cost


insurance_cost

INFERENCES

1. **Insurance Duration**: Customers have insurance policies with the company ranging from 0 to 8 years, indicating a diverse range of policy durations.

2. **Adventure Sports**: The majority of customers are not participating in adventure sports, suggesting that this activity is not a common part of their lifestyle.

3. **Cholesterol Levels**: Customers with cholesterol levels between 150 to 175 are the most numerous, followed by those with levels between 125 to 150, indicating a range of cholesterol levels among the customer base. This suggests varying levels of cardiovascular risk within the population.

4. **Average Daily Steps:** Customers have a wide range of average daily steps, with a minimum of 2034 and a mean of 5215. This indicates varying levels of physical activity among customers.

5. **Age Distribution:** Customers' ages range from a minimum of 16 years to a maximum of 74 years, indicating a diverse age distribution among the customer base with mean age 44 most customers are between age 30 to 60

6. **Heart Disease History**: Customers with a history of heart disease are relatively few, suggesting that heart disease may not be a common health concern among the customer population.

7. **Other Major Diseases:** Customers with other major diseases are also relatively few, indicating that serious health conditions may not be prevalent among the customer base.

8. **Glucose Levels:** The average glucose level among customers is 167.53, with a minimum of 57 and a maximum of 277, indicating a wide range of glucose levels among customers.

9. **BMI:** Customers' BMI ranges from a minimum of 12.3 to a maximum of 100.6, with a mean of 31.39, indicating a range of body weights and compositions among customers.

10**. Weight:** Customers' weights range from a minimum of 52 to a maximum of 96, with a mean of 71, indicating a range of body weights among customers.

11. **Fat Percentage:** Customers' fat percentages range from a minimum of 11 to a maximum of 42, with a mean of 28, indicating a range of body compositions among customers.

12. **Insurance Cost:** Insurance costs range from a minimum of 2468 to a maximum of 67870, with a mean of 27147, indicating a wide range of insurance costs among customers.

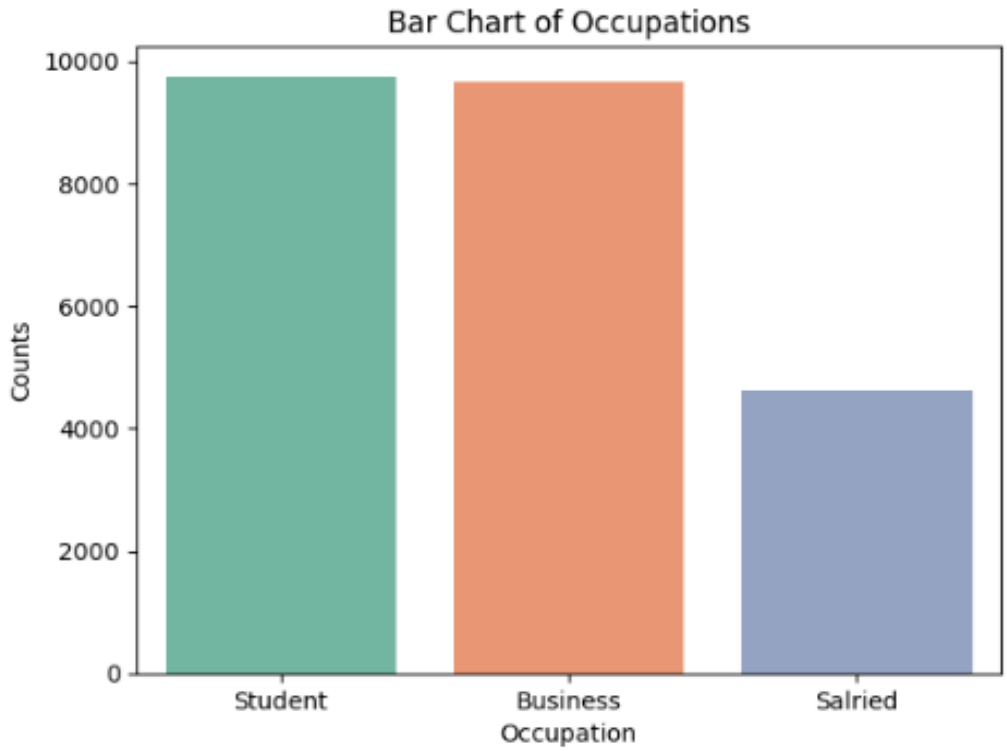
13**. Health Factors:**

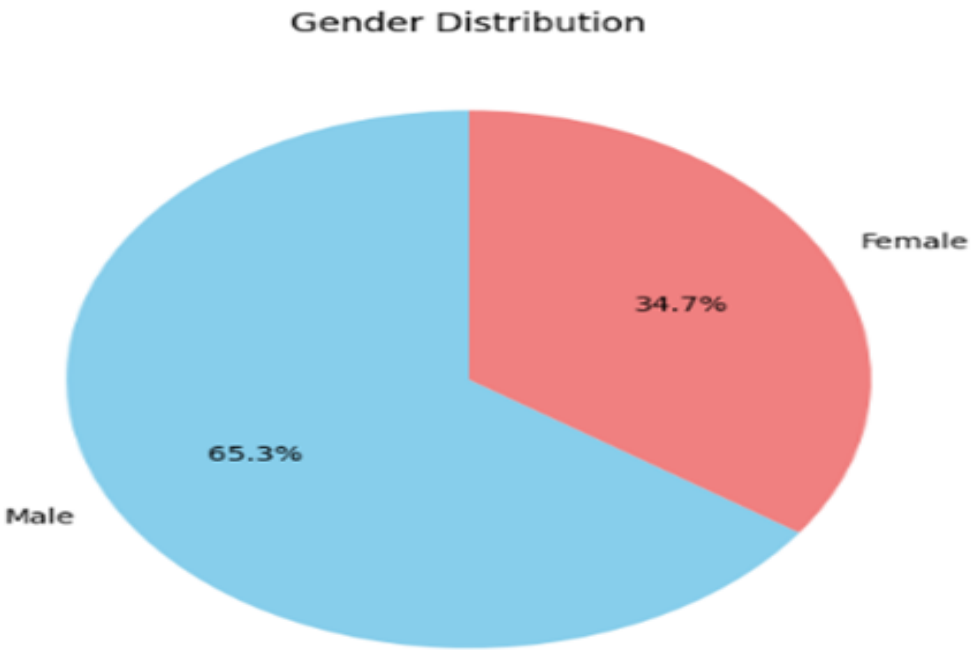
Overweight Customers: The analysis includes overweight customers, indicating a need for health interventions related to weight management.

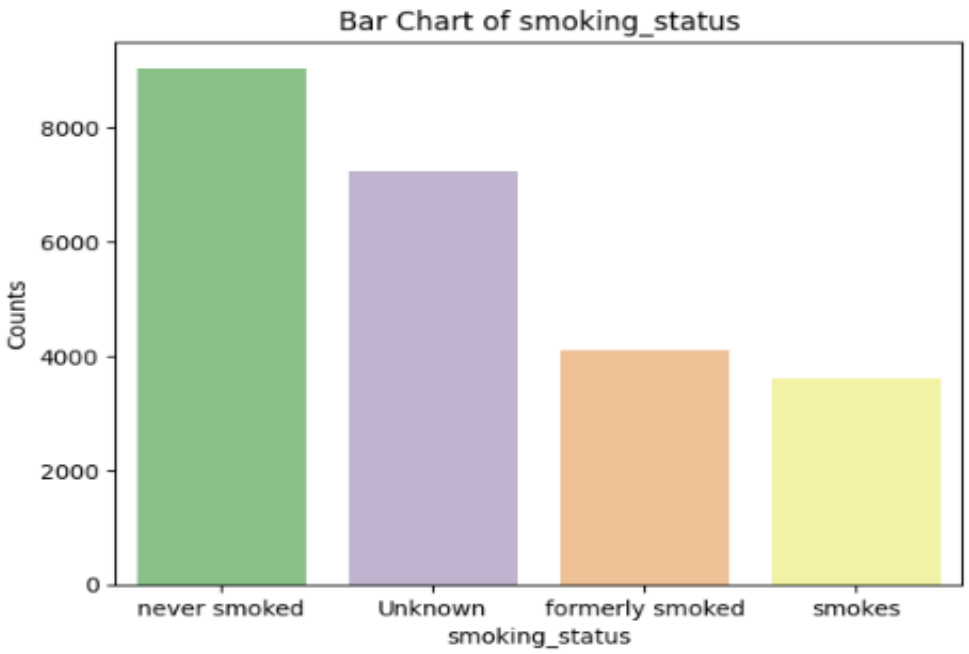
Aged Customers: There is a presence of aged customers in the dataset, highlighting the importance of healthcare services for older individuals which is important for insurance industry.

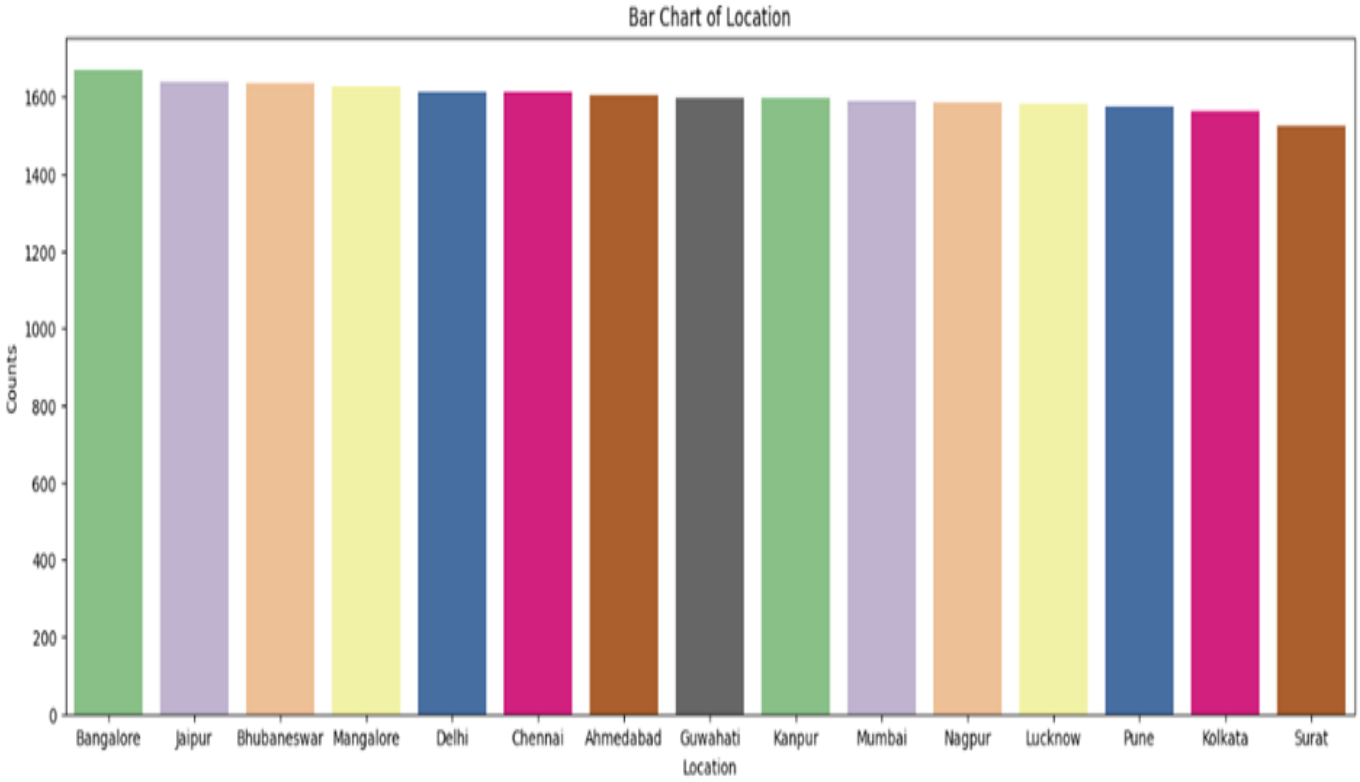
High BMI Customers: Customers with high BMI values are present in the dataset, indicating a potential risk factor for various health conditions.

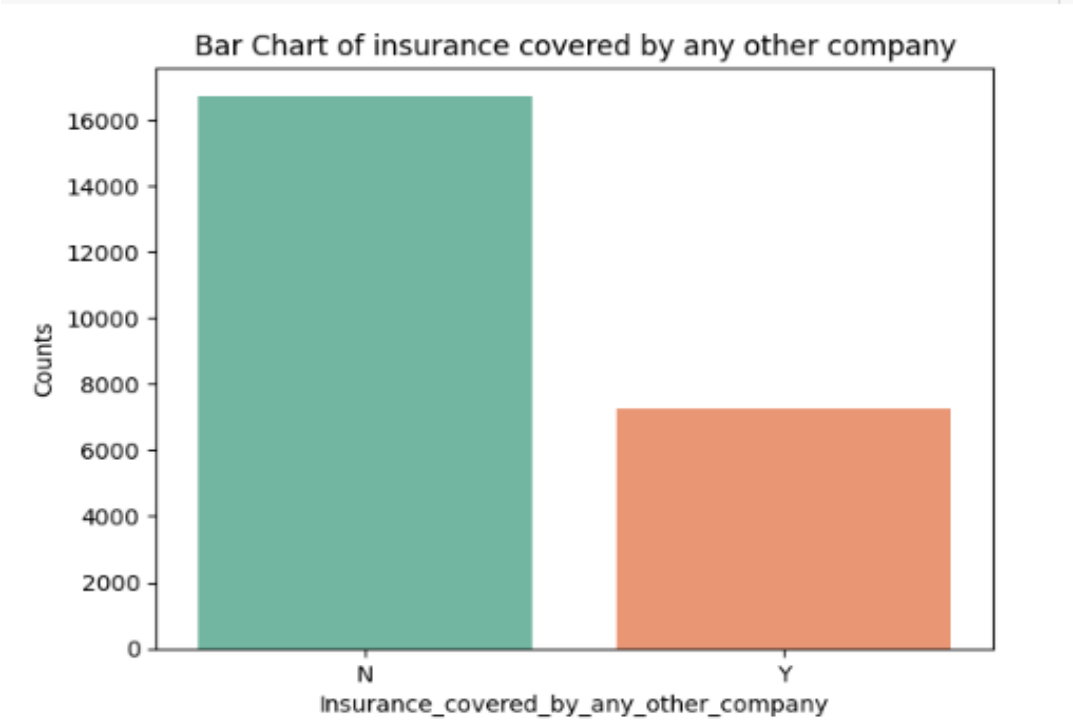
E. Univariate Analysis of Categorical variables.

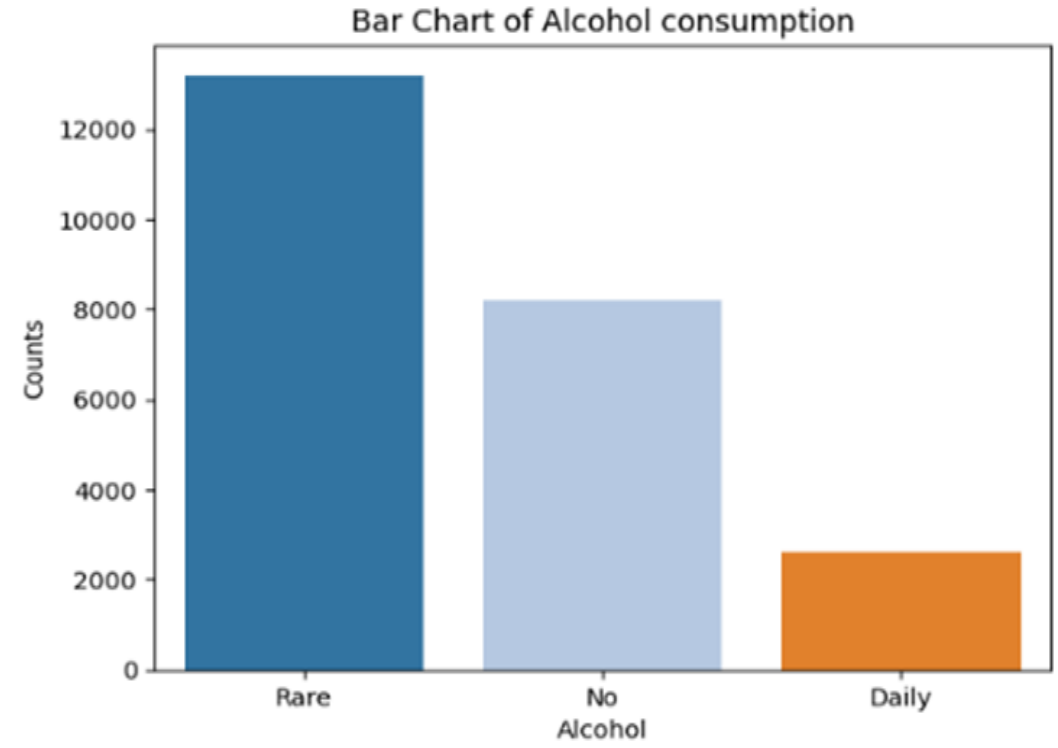


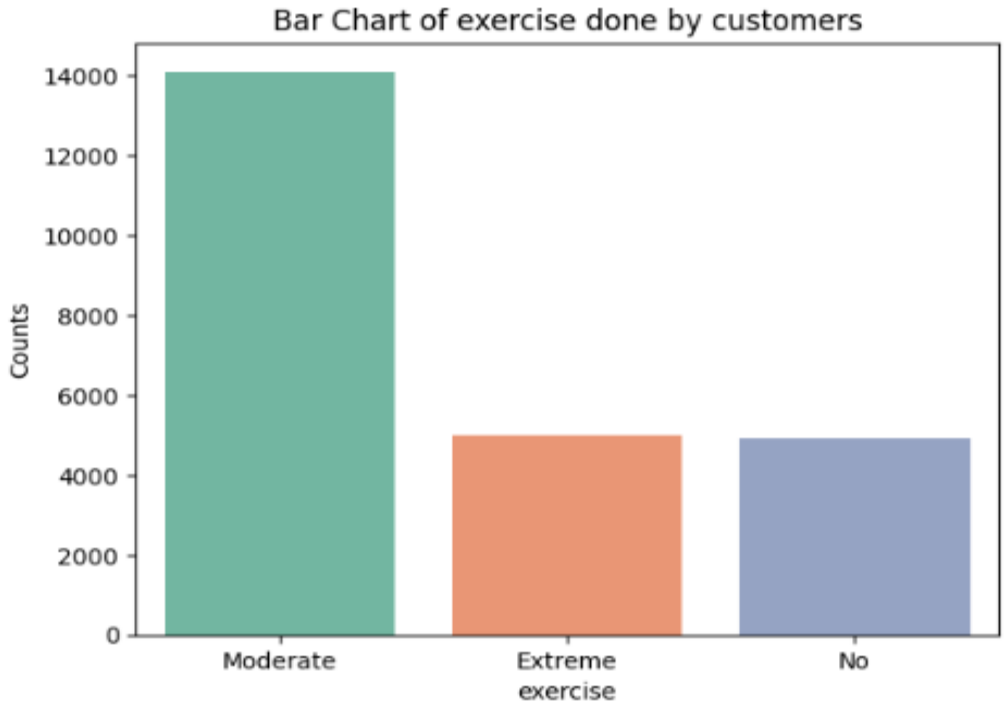


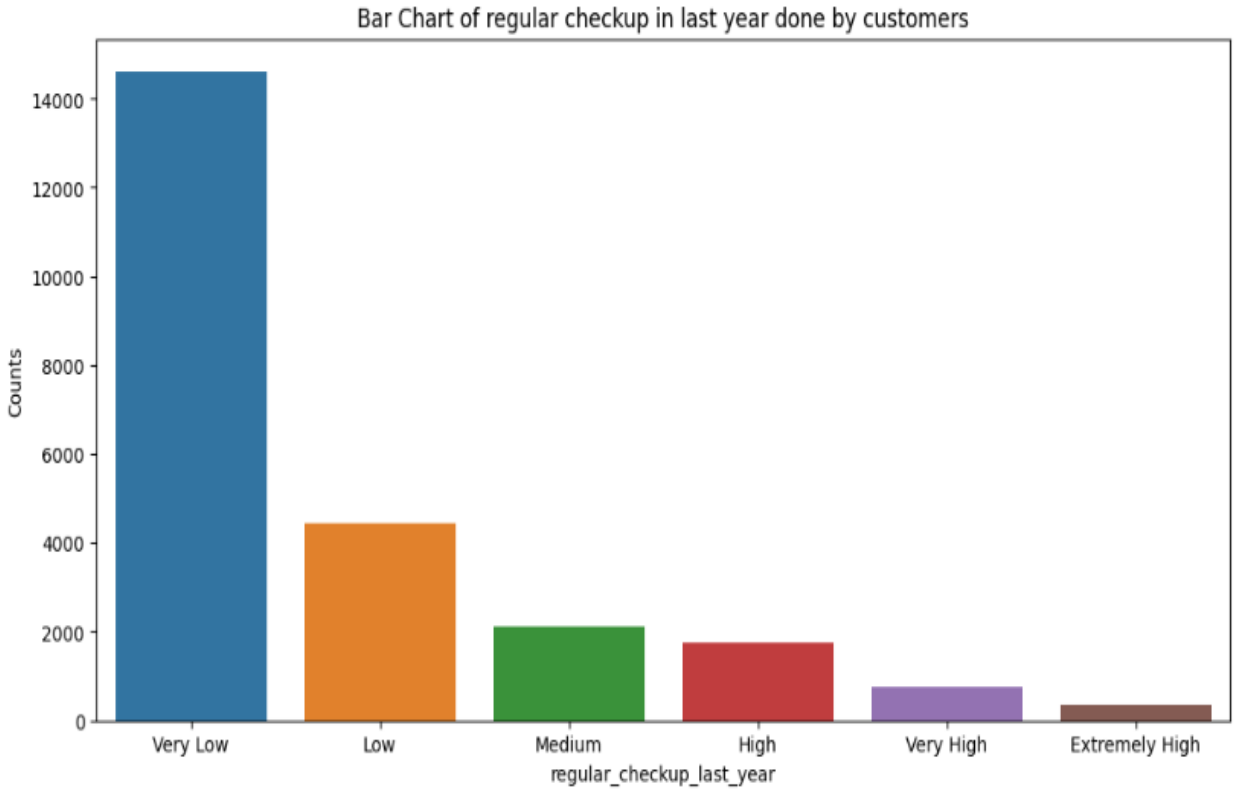

Bar Chart of smoking_status
Counts
8000
6000
4000
2000
0
never smoked
Unknown
formerly smoked
smokes
smoking_status

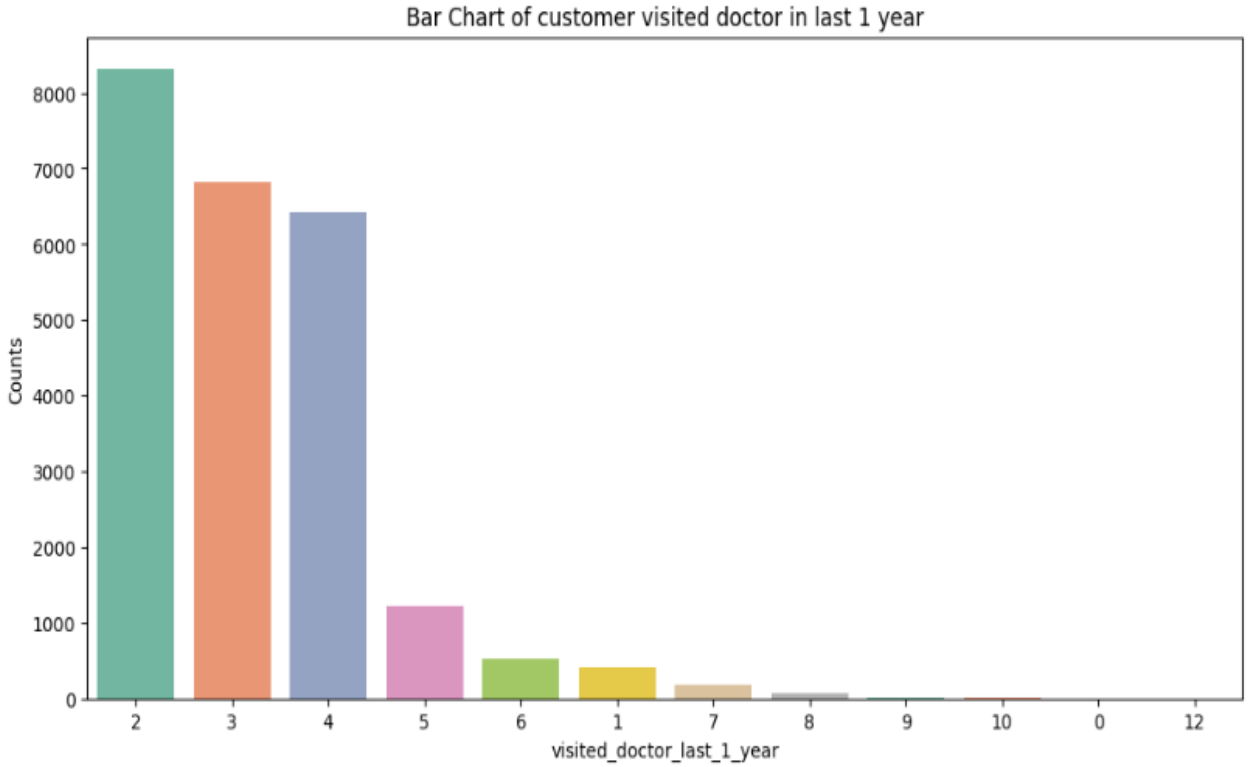

Bar Chart of Location
Counts
1600
1400
1200
1000
800
600
400
200
0
Bangalore
Jaipur
Bhubaneswar
Mangalore
Delhi
Chennai
Ahmedabad
Guwahati
Kanpur
Mumbai
Nagpur
Lucknow
Pune
Kolkata
Surat
Location

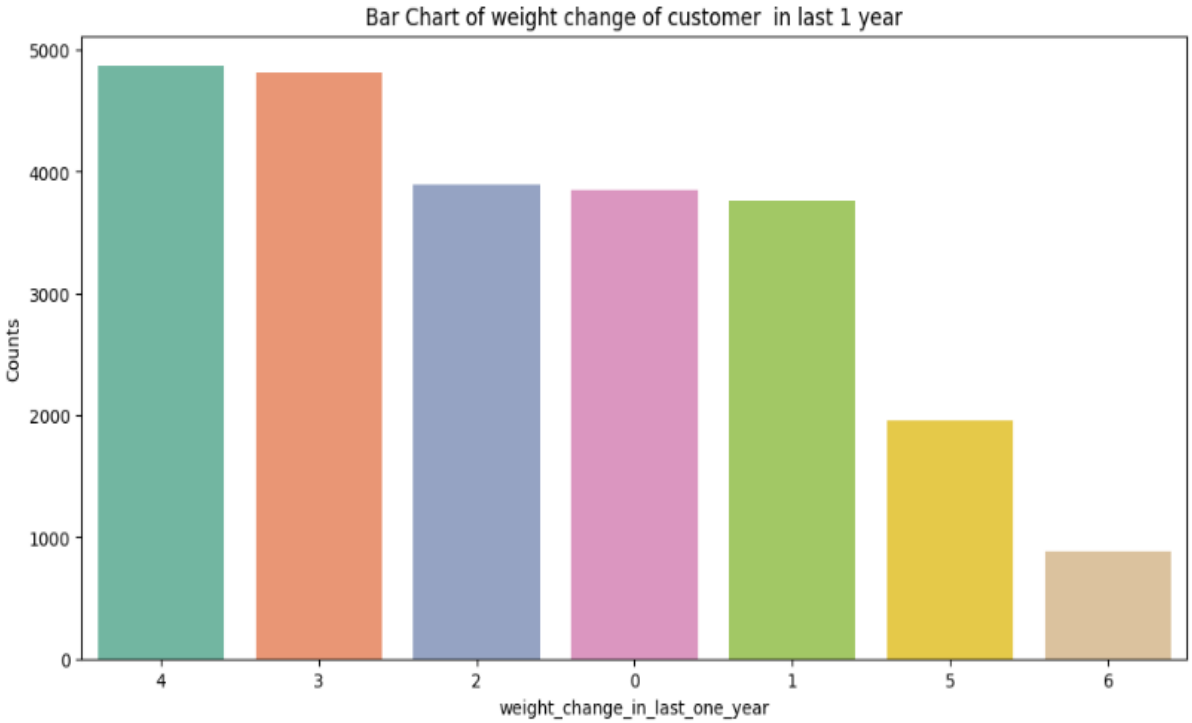

Bar Chart of insurance covered by any other company
Counts
16000
14000
12000
10000
8000
6000
4000
2000
0
N
Y
Insurance_covered_by_any_other_company


Bar Chart of Alcohol consumption
Counts
12000
10000
8000
6000
4000
2000
0
Rare
No
Daily
Alcohol


Bar Chart of exercise done by customers
Counts
14000
12000
10000
8000
6000
4000
2000
0
Moderate
Extreme
No
exercise


Bar Chart of regular checkup in last year done by customers
Counts
14000
12000
10000
8000
6000
4000
2000
0
Very Low
Low
Medium
High
Very High
Extremely High
regular_checkup_last_year


Bar Chart of customer visited doctor in last 1 year
Counts
8000
7000
6000
5000
4000
3000
2000
1000
0
2
3
4
5
6
1
7
8
9
10
0
12
visited_doctor_last_1_year


Bar Chart of weight change of customer in last 1 year
Counts
5000
4000
3000
2000
1000
0
4
3
2
0
1
5
6
weight_change_in_last_one_year

INFERENCES

1. **Occupation Distribution**:
- Students represent the largest customer segment, indicating a younger demographic or a specific marketing focus on students.
- Business occupations show a slightly lower number of customers, followed by salaried occupations, suggesting a diverse customer base with varying income sources.

2. **Gender Distribution**:
- Female customers account for 34.7% of the total, while males account for 65.3%, indicating a gender imbalance in the customer base.

3. **Smoking Status**:
- A higher number of customers have never smoked, suggesting a trend towards non-smoking habits.
- The unknown category and formerly smoked categories follow, indicating a mix of smoking histories among customers.
- The least number of customers are current smokers, indicating a low prevalence of smoking among customers.

4**. City Distribution:**
- Bangalore has the highest number of customers, followed by Jaipur and Bhubaneshwar, indicating a concentration of customers in these cities.
- Surat has the least number of customers, suggesting a lower presence or market penetration in this city.

5. **Insurance Coverage:**
- Most customers do not have insurance with another company, indicating a potential market for additional insurance products.
- Only around 6000 customers have insurance with another company, suggesting a relatively small portion of the customer base.

6**. Alcohol Consumption:**
- Rare alcohol consumption is the most common category among customers, followed by no alcohol consumption and daily consumption, indicating varying alcohol consumption habits among customers.

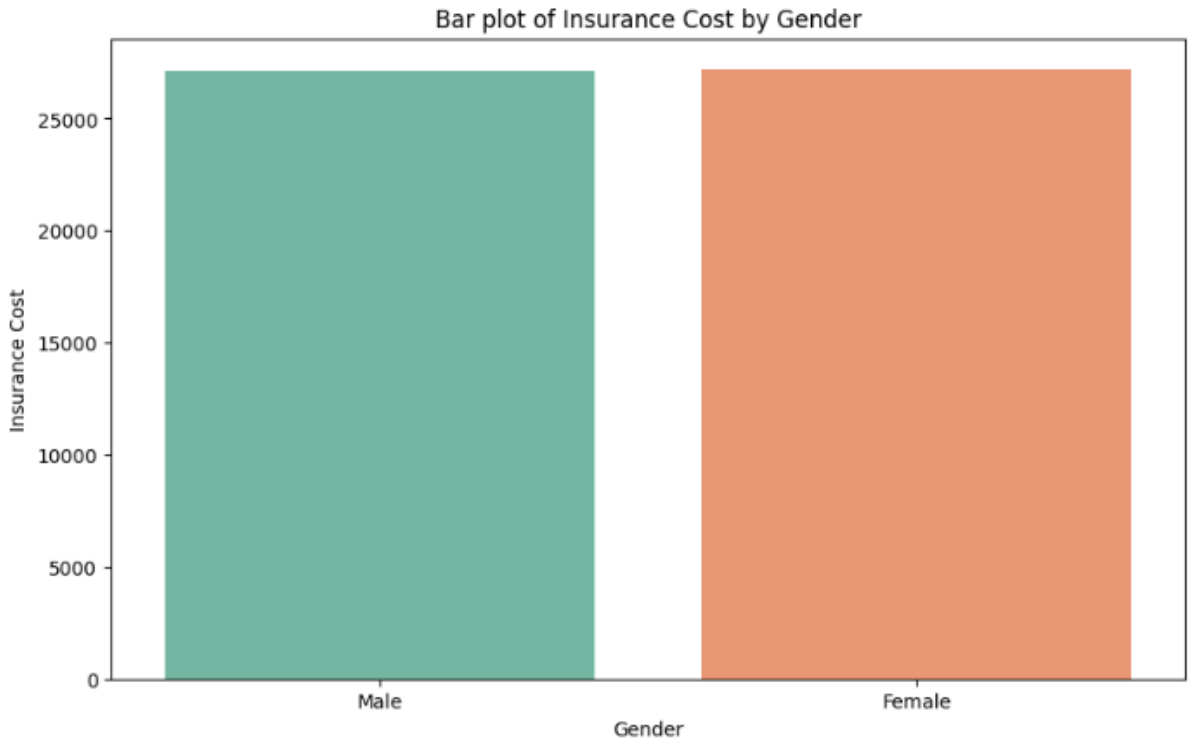
7. **Exercise Habits:**
    - Moderate exercise is the most common category among customers, followed by extreme exercise and no exercise, suggesting a range of physical activity levels among customers.

8. **Regular Checkups:**
    - A very low number of customers have had a regular checkup in the last year, indicating a potential gap in preventive healthcare practices among customers.

9. **Doctor Visits:**
    - The most common number of doctor visits in the last year is 2, followed by 3 visits, suggesting a moderate level of healthcare utilization among customers.

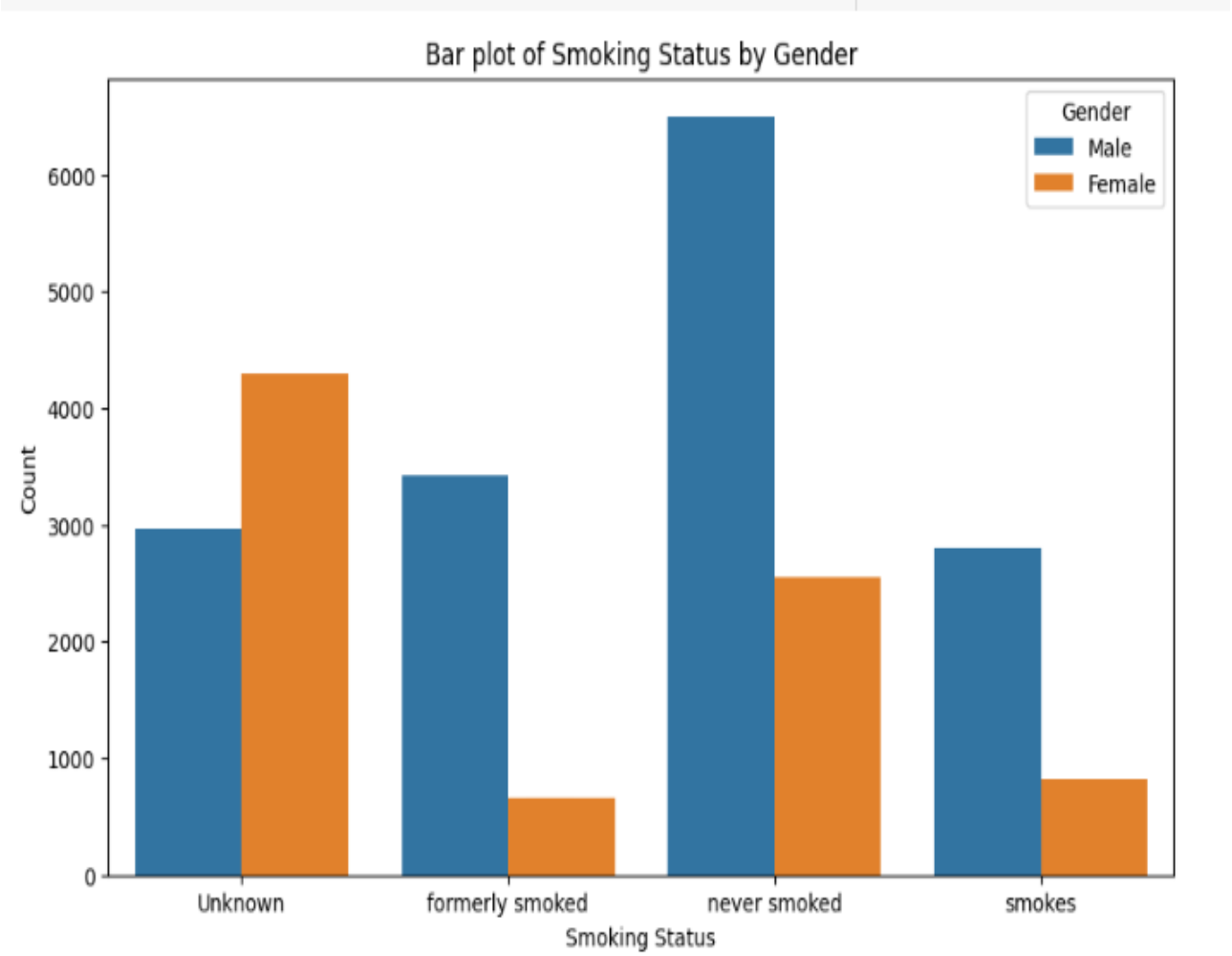
10. **Weight Change:**
    - The most common weight change among customers in the last year is a loss of 4 kg, followed by 3 kg, 2 kg, and no change, indicating a trend towards weight loss among customers.

F. Bivariate Analysis



INFERENCES

- The bar plot compares the insurance costs between males and females.
- Both genders have similar insurance costs, with females having a slightly higher cost.
- For Females Insurance costs = 27205.83
- For males Insurance costs = 27136.67
- This suggests that gender alone may not be a significant factor in determining insurance premiums and there is no much gender bias for health insurance.

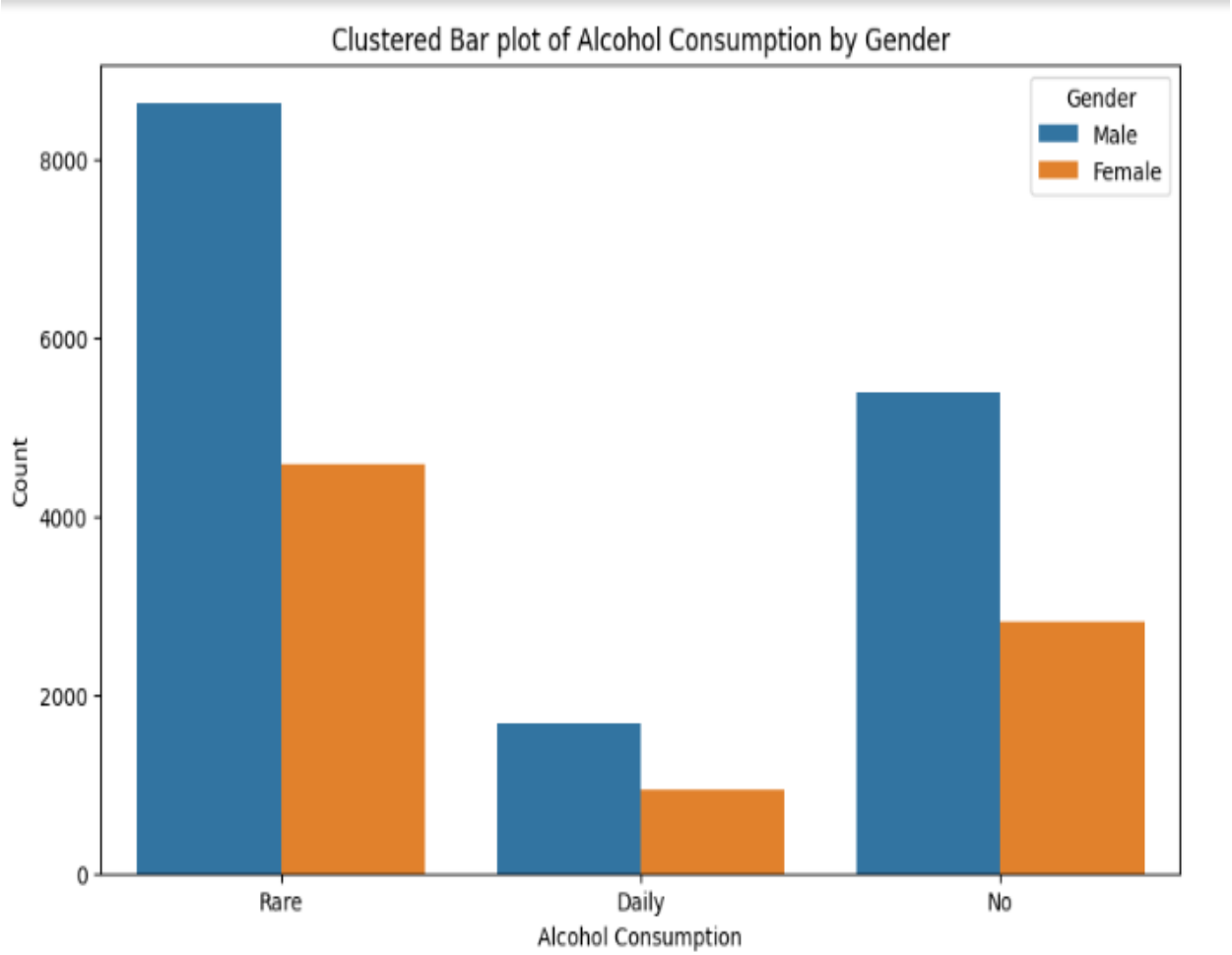


INFERENCES

- The highest count is observed in the “never smoked” category are males.
- The “formerly smoked” category has higher number of males’ gender.
- The “smokes” category shows a higher count for male.
- The “Unknown” category shows a higher count for female.
- Smoking behavior varies by gender, with more males being current smokers.
- The data suggests that gender may play a role in smoking habits.
- The data shows that females also smoke and formerly smoked in considerable amount.

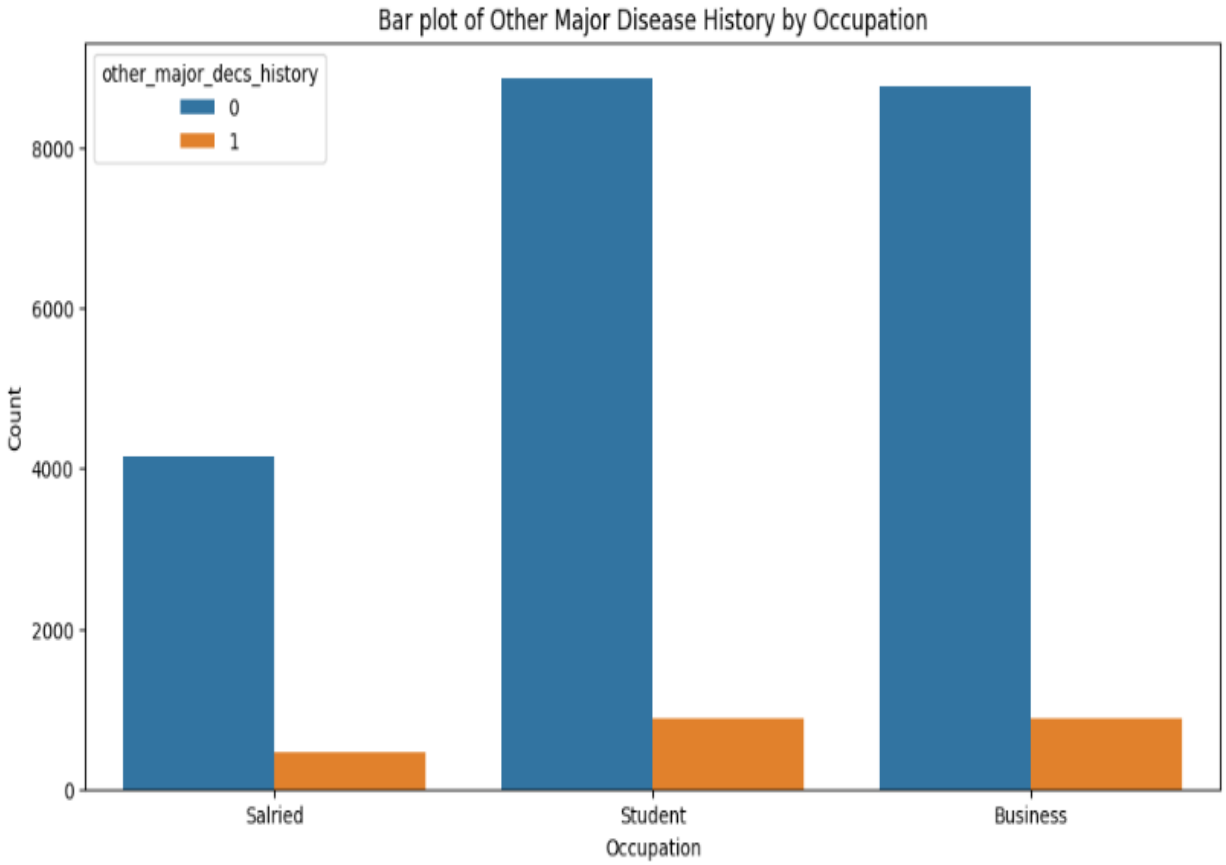
INFERENCES

- The plot compares the distribution of Body Mass Index (BMI) between males and females.
- Both genders have a similar median BMI, indicated by the white dot in the center of each plot.
- Median BMI Males=32.1 and for females=26.2
- The interquartile range (IQR) suggests that the middle 50% of BMIs fall within a similar range for both males and females.
- However, there is a slight difference:
    - Males exhibit a slightly wider distribution of BMIs, as shown by the broader green plot. This implies greater variability in BMI among males.
    - Females have a more symmetrical distribution around the median BMI, as indicated by the narrower orange plot.



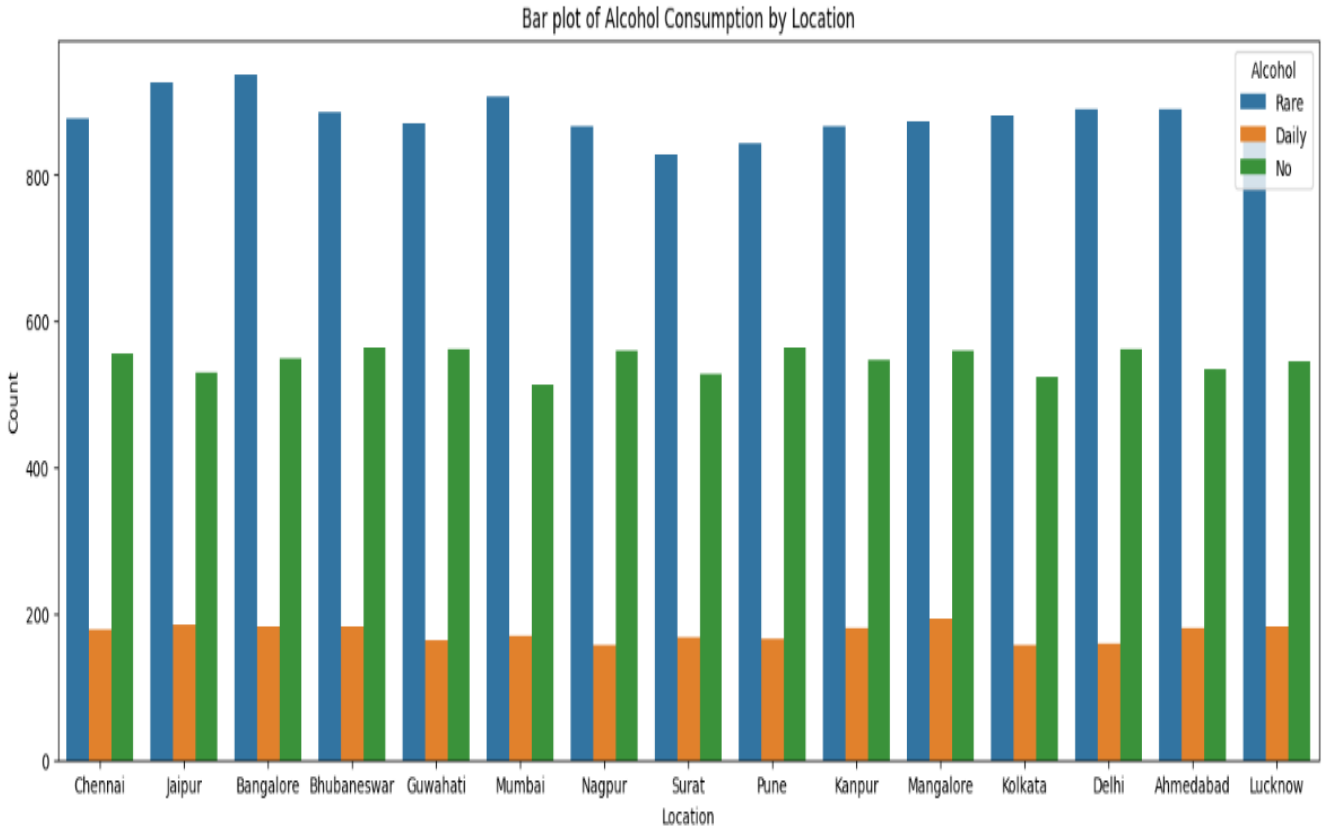
INFERENCES

- Rare Consumption:
    - Over 8000 males fall into this category.
    - Around 4000 females fall into this category.
- Daily Consumption:
    - Fewer than 2000 males consume alcohol daily.
    - Fewer than 1000 females consume alcohol daily.
- No Alcohol Consumption:
    - Approximately 6000 males do not consume alcohol at all.
    - Around 3000 females fall into this category.



INFERENCES

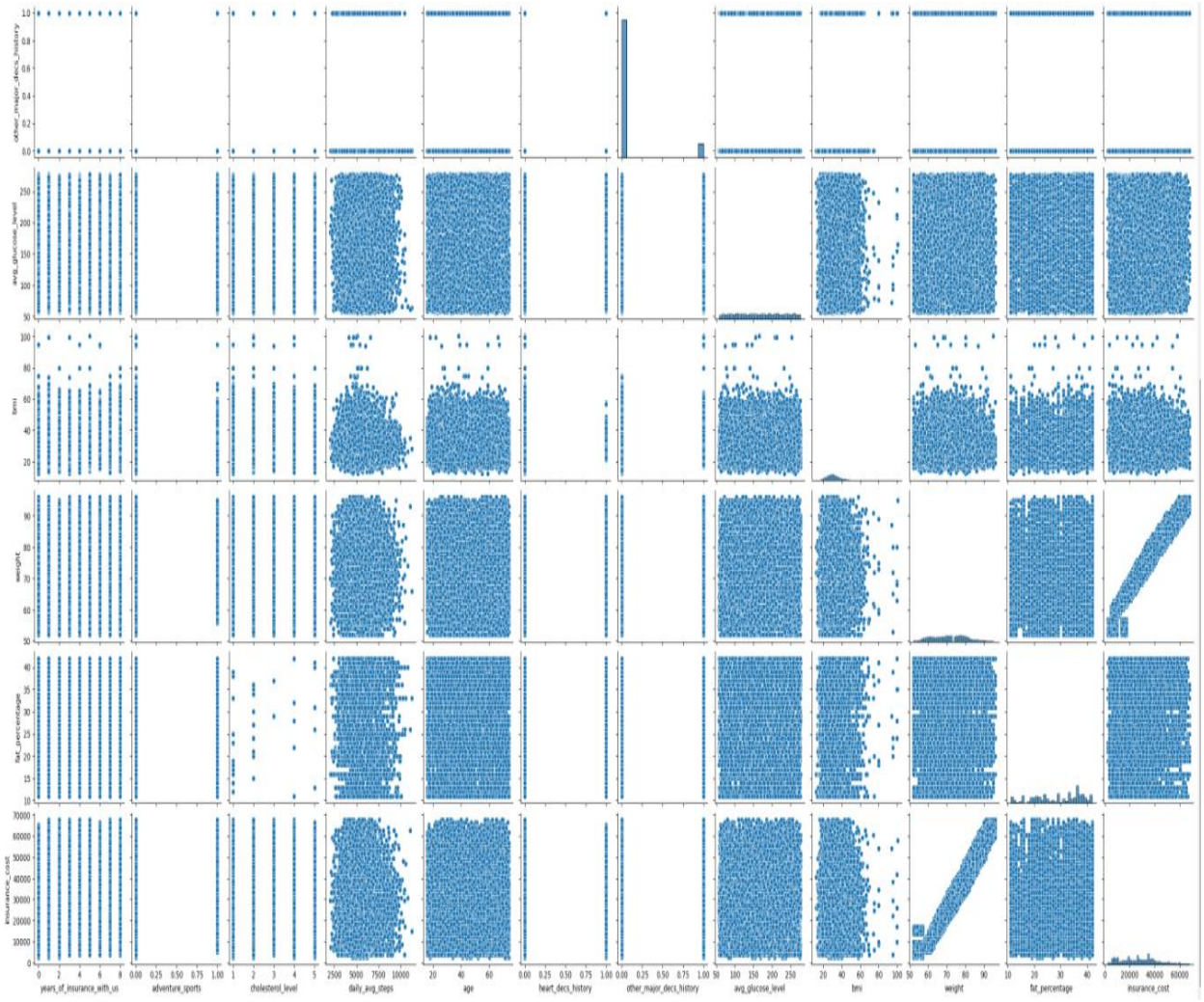
- **Salaried Individuals**:
  Have a lower count of individuals with a history of other major diseases
  The count of salaried individuals without a history of other major diseases is higher
- **Students**:
  Exhibit a significantly higher count of individuals without a history of other major diseases (878). Compared to salaried and business occupation.
  The number of students with a disease history is higher than salaried and business occupation.
- **Business Professionals**:
  Similar to students, business professionals have a higher count of individuals without a history of other major diseases.
  The proportion of business professionals with a disease history is similar to the students.(885)



INFERENCES

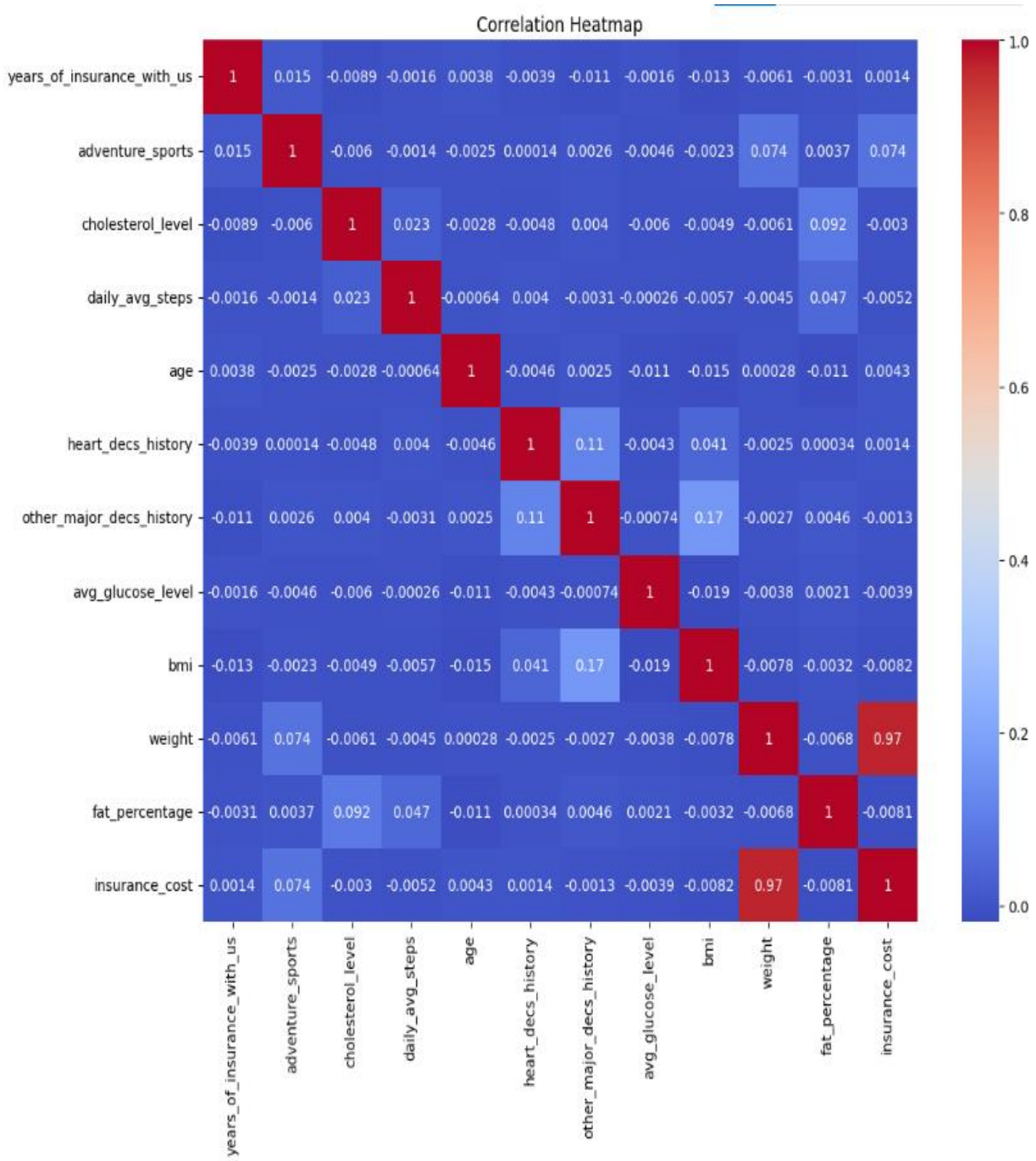
- **Bangalore, Jaipur, Mumbai, Ahmedabad, Delhi:**
  Have a higher number of people who rarely consume alcohol
- **Bhubaneswar, Delhi, Mangalore, Nagpur:**
  Have more non-consumers.
- **Mangalore, Lucknow, Bhubaneswar, Bangalore:**
  Have more Daily consumers.

G. Multivariate Analysis



INFERENCES

Pair plot tells us about the interaction of each variable with every other variable present.
As such there is strong relationship present between the variables.
There is a mixture of positive and negative relationships though which is expected.
Overall, it's a rough estimate of the interactions; clearer picture can be obtained by heat map values and also different kinds of plot.


Correlation Heatmap
years_of_insurance_with_us
adventure_sports
cholesterol_level
daily_avg_steps
age
heart_decs_history
other_major_decs_history
avg_glucose_level
bmi
weight
fat_percentage
insurance_cost
1 0.015 -0.0089 -0.0016 0.0038 -0.0039 -0.011 -0.0016 -0.013 -0.0061 -0.0031 0.0014
0.015 1 -0.006 -0.0014 -0.0025 0.00014 0.0026 -0.0046 -0.0023 0.074 0.0037 0.074
-0.0089 -0.006 1 0.023 -0.0028 -0.0048 0.004 -0.006 -0.0049 -0.0061 0.092 -0.003
-0.0016 -0.0014 0.023 1 -0.00064 0.004 -0.0031 -0.00026 -0.0057 -0.0045 0.047 -0.0052
0.0038 -0.0025 -0.0028 -0.00064 1 -0.0046 0.0025 -0.011 -0.015 0.00028 -0.011 0.0043
-0.0039 0.00014 -0.0048 0.004 -0.0046 1 0.11 -0.0043 0.041 -0.0025 0.00034 0.0014
-0.011 0.0026 0.004 -0.0031 0.0025 0.11 1 -0.00074 0.17 -0.0027 0.0046 -0.0013
-0.0016 -0.0046 -0.006 -0.00026 -0.011 -0.0043 -0.00074 1 -0.019 -0.0038 0.0021 -0.0039
-0.013 -0.0023 -0.0049 -0.0057 -0.015 0.041 0.17 -0.019 1 -0.0078 -0.0032 -0.0082
-0.0061 0.074 -0.0061 -0.0045 0.00028 -0.0025 -0.0027 -0.0038 -0.0078 1 -0.0068 0.97
-0.0031 0.0037 0.092 0.047 -0.011 0.00034 0.0046 0.0021 -0.0032 -0.0068 1 -0.0081
0.0014 0.074 -0.003 -0.0052 0.0043 0.0014 -0.0013 -0.0039 -0.0082 0.97 -0.0081 1
1.0
0.8
0.6
0.4
0.2
0.0

INFERENCES

1. Weight and Insurance Cost: Strong positive correlation: Higher weight means higher insurance cost.
2. **Fat Percentage and Insurance Cost:** Strong positive correlation: Higher fat percentage means higher insurance cost.
3. **Weight and Fat Percentage:** Strong positive correlation: Higher weight is associated with higher fat percentage.
4. **BMI and Insurance Cost:** Moderate negative correlation: Higher BMI slightly reduces insurance cost.
5. **Other Major Diseases and Heart Disease History:** Moderate positive correlation: History of other major diseases is linked to heart disease history.
6. **Cholesterol Level and Fat Percentage:** Moderate positive correlation: Higher cholesterol levels are linked to higher fat percentage.
7. **Age and Fat Percentage**: Slight positive correlation: Older age is slightly associated with higher fat percentage.
8. **Years of Insurance with Us:** Very weak correlations: Duration of insurance doesn't strongly relate to other health metrics.
9. **Adventure Sports**: Weak correlations: Participation in adventure sports doesn’t strongly relate to other health metrics or insurance cost.
10. **Daily Average Steps:** Weak correlations: Number of daily steps doesn't strongly relate to other health metrics or insurance cost.

Overall, weight and fat percentage are the most important factors affecting insurance cost. Other variables like years of insurance, adventure sports, and daily steps have minimal impact.

6) Business insights from EDA

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

Yes the data is unbalanced

1. **Gender Distribution:** There is a gender imbalance in the customer base, with 65.3% male customers and 34.7% female customers. This imbalance could impact the fairness and accuracy of models and analyses that rely on gender-related data.
2. **Occupation Distribution**: The data indicates that students represent the largest customer segment, suggesting a potential imbalance in the representation of different occupation types. This imbalance could skew analyses and models that rely on occupation-related data.
3. **Smoking Status:** The data shows that a higher number of customers have never smoked followed by customers with unknown smoking status, formerly smoked, and currently smoke. This imbalance could affect analyses and models related to smoking habits and health outcomes.
4. **City Distribution**: There are variations in the number of customers across different cities, with some cities having a higher concentration of customers than others. This imbalance could impact the accuracy of regional analyses and models.
5. **regular_checkup_last_year:** There are 6 unique categories, with "Very Low" or 0 being the most frequent. This suggests that a large portion of customers had a very low number of regular checkups last year compared to other categories, indicating potential imbalance.
6. **cholesterol_level:** The mean value is 2.267389, indicating that the majority of customers have a cholesterol level of 2.0 or below. This could suggest an imbalance in the distribution of cholesterol levels among customers.

To address data imbalance, we can consider the following strategies:

1. **Data Collection:** Ensure that future data collection efforts are more balanced to avoid bias in the dataset. This could involve targeted marketing campaigns to attract underrepresented groups or adjusting data collection methods to reach a more diverse customer base.
2. **Data Augmentation:** Augment the existing dataset with additional data to balance out the representation of different groups. For example, additional data on female customers could be collected through targeted surveys or outreach efforts.
3. **Sampling Techniques:** Use sampling techniques, such as oversampling or undersampling, to balance out the data. This involves either increasing the number of samples in underrepresented groups (oversampling) or reducing the number of samples in overrepresented groups (undersampling).
4. **Model Adjustments:** Adjust models to account for the imbalance in the data. This could involve using different evaluation metrics that are less sensitive to imbalance, such as F1-score or AUC-ROC, or using techniques like class weights in classification models.
5. **Business Strategy:** Develop business strategies that are inclusive and address the needs of all customer segments. This could involve offering products and services that cater to underrepresented groups or adjusting marketing strategies to reach a more diverse audience.

b) Business insights using clustering

Cluster 0:

1. Customers in this cluster are relatively new to the insurance company, with a low number of years of association.
2. They have a cautious approach towards adventure sports, indicating a potentially risk-averse behavior.
3. This cluster exhibits lower-than-average cholesterol levels, which might indicate a healthier lifestyle or better health management.
4. The cluster shows a tendency towards lower weights, which could be a factor in their overall health profile.
5. While their exercise habits are not significantly different from other clusters, they may benefit from wellness programs focused on maintaining good health**.**

Cluster 1:

1. Customers in this cluster have a moderate tenure with the insurance company, indicating a somewhat stable relationship.
2. Their engagement in adventure sports is relatively low, suggesting a preference for less risky activities.
3. With average cholesterol levels, this cluster represents a standard health profile.
4. The cluster's weight distribution is around average, indicating a diverse range of body types.
5. While they exhibit normal exercise patterns, promoting healthy lifestyle choices could be beneficial for this group.

Cluster 2:

1. Customers in this cluster have a moderate tenure with the insurance company, indicating a stable but not long-standing relationship.
2. Their moderate engagement in adventure sports suggests a balanced approach to risk-taking activities.
3. With slightly lower cholesterol levels than average, this cluster may have a slightly healthier diet or lifestyle.
4. The cluster's weight distribution is around average, indicating a diverse range of body types.
5. While they exhibit normal exercise patterns, promoting healthy lifestyle choices could be beneficial for this group.

Cluster 3:

1. Customers in this cluster have a slightly longer tenure with the insurance company, indicating a more established relationship.
2. Their moderate engagement in adventure sports suggests a balanced approach to risk-taking activities.
3. With average cholesterol levels, this cluster represents a standard health profile.
4. The cluster's weight distribution is around average, indicating a diverse range of body types.
5. While they exhibit normal exercise patterns, promoting healthy lifestyle choices could be beneficial for this group.

Cluster 4:

1. Customers in this cluster are the newest to the insurance company, with the shortest tenure.
2. They have a slightly higher engagement in adventure sports, indicating a potentially more adventurous or risk-taking behavior.
3. This cluster exhibits the highest cholesterol levels among all clusters, which might indicate a need for health interventions or monitoring.
4. The cluster shows a tendency towards higher weights, which could be a factor in their overall health profile.
5. While their exercise habits are not significantly different from other clusters, they may benefit from wellness programs focused on maintaining good health.

c) Business Implications

1. **Health Intervention Opportunities**: There is a need for targeted health interventions for customers with high cholesterol levels and overweight individuals across all clusters. Programs focusing on diet, exercise, and regular checkups could be beneficial.

2**. Customer Engagement Strategies**: Since most customers are not participating in adventure sports, insurance companies could focus on other engagement strategies, such as wellness programs, to attract and retain customers.

3. **Age-Specific Health Programs:** With a diverse age distribution, insurance companies could develop age-specific health programs. For example, programs targeting older individuals could focus on managing chronic conditions, while programs for younger individuals could focus on preventive care.

4. **Gender-Based Marketing:** There is a gender imbalance in the customer base, with more males than females. Insurance companies could consider gender-specific marketing strategies to attract more female customers.

5. **Regional Variations:** There are regional variations in customer behavior, such as alcohol consumption. Insurance companies could tailor their products and services to meet the specific needs of customers in different regions.

6**. Customer Education:** There is a need for customer education on the importance of regular checkups and doctor visits, as indicated by the low number of customers who have had a regular checkup in the last year.

7. **Health Monitoring Services**: Insurance companies could offer health monitoring services, such as regular health checkups and screenings, to help customers manage their health and reduce the risk of developing chronic conditions.

1) Model building and interpretation.

a) Building various Predictive models.

We chose to build predictive models because target column is continuous. This means that our goal is to predict a numerical value, which aligns with the nature of regression analysis. By building predictive models, we aim to understand the relationships between the input features and the continuous target variable, allowing us to make informed predictions for new data points.

There are 2 main categories of the machine learning model

1) **Parametric Models:**

Parametric models make strong assumptions about the functional form of the relationship between the input features and the target variable. Once the parameters of the model are learned from the training data, the model structure is fixed.

Example: Linear regression is a parametric model that assumes a linear relationship between the input features and the target variable. The model parameters (coefficients) are estimated during training, and the model can then make predictions for new data points based on this linear relationship.

2) **Non-Parametric Models:**

Non-parametric models make fewer assumptions about the underlying data distribution and can therefore be more flexible in modeling complex relationships. These models can adapt to the data, allowing the model complexity to increase with the amount of data available.

Example: Decision trees are a non-parametric model that can capture complex interactions between features. The structure of the tree is not fixed and can vary based on the training data.

I have built 4 parametric models and 4 non parametric models

Parametric models

1. Linear Regression:
    - Linear regression is used to establish a linear relationship between the input features and the target variable.
    - The model is trained using the least squares method to minimize the sum of squared errors. It is evaluated using metrics such as RMSE, R-squared, adjusted R-squared, MAPE (Mean Absolute Percentage Error).

2. Lasso Regression:
    - Lasso regression is used for feature selection and regularization to prevent over fitting.
    - The model is trained using the Lasso algorithm, which adds a penalty term to the least squares objective. It is evaluated using metrics such as RMSE, R-squared, adjusted R-squared, MAPE (Mean Absolute Percentage Error).

3. Ridge Regression:
    - Ridge regression is also used for regularization to prevent over fitting, but it uses a different penalty term than Lasso.
    - The model is trained using the Ridge algorithm, which adds a penalty term to the least squares objective. It is evaluated using metrics such as RMSE, R-squared, adjusted R-squared, MAPE (Mean Absolute Percentage Error).

4. Polynomial Regression:
    - Polynomial regression is used when the relationship between the input features and the target variable is non-linear.
    - The model is trained using polynomial features of the input data and a linear regression algorithm. It is evaluated using metrics such as RMSE, R-squared, adjusted R-squared, MAPE (Mean Absolute Percentage Error).

Non-Parametric models

1. Random Forest Regression:
    - Random forest regression is an ensemble method used for both classification and regression tasks, known for its ability to handle complex relationships in data.
    - The model is trained using an ensemble of decision trees. It is evaluated using metrics such as RMSE, R-squared, and MAPE (Mean Absolute Percentage Error).

2. AdaBoost Regression:
    - AdaBoost regression is another ensemble method that combines multiple weak learners to create a strong learner.
    - The model is trained using a series of weak learners, typically decision trees. It is evaluated using metrics such as RMSE, R-squared, and MAPE (Mean Absolute Percentage Error).

3. XGBoost Regression:
    - XGBoost (Extreme Gradient Boosting) is an optimized distributed gradient boosting library known for its speed and performance.
    - The model is trained using an ensemble of decision trees and is optimized for speed and performance. It is evaluated using metrics such as RMSE, R-squared, and MAPE (Mean Absolute Percentage Error).

4. Gradient Boosting Machine Regression:
    - Gradient Boosting Machine (GBM) regression is similar to XGBoost but may use a different underlying algorithm.
    - The model is trained using an ensemble of decision trees. It is evaluated using metrics such as RMSE, R-squared, and MAPE (Mean Absolute Percentage Error).

b. Testing predictive model against the test set using various appropriate performance metrics & Residual analysis.

Parametric models Performance Metrics & Residual analysis:

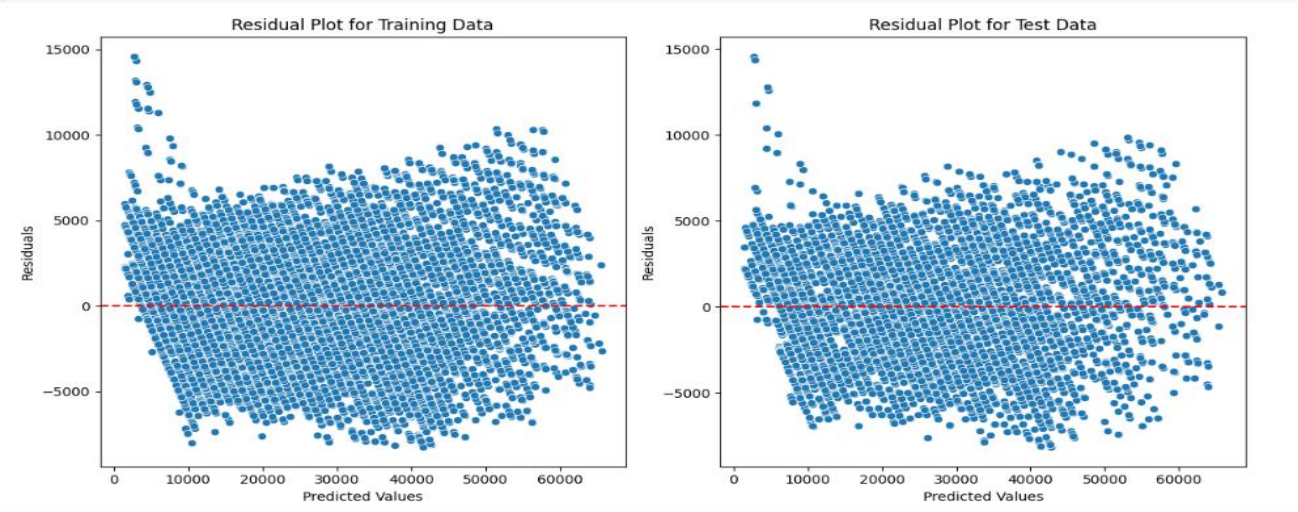
1. Linear Regression:

Training Data:

RMSE: 3220.89
R-squared: 0.94
Adjusted R-squared: 0.94
MAPE for Training Data: 14.50

Test Data:
RMSE: 3178.58
R-squared: 0.95
Adjusted R-squared: 0.95
MAPE for Test Data: 14.18

- Both the training and test data show relatively good performance, with RMSE values around 3200 and R-squared values close to 0.95.
- The model's performance is consistent between training and test datasets, indicating good generalization.



Residual plot was used to visualize the residuals against the predicted values, the residual plot showed random scattering of points around the horizontal line at zero, indicating that the assumptions of linear regression were not violated and the model is unbiased.

2. Lasso Regression:

Training Data:
RMSE: 3220.95
R-squared: 0.94
Adjusted R-squared: 0.94
MAPE: 14.50

Test Data:
RMSE: 3178.45
R-squared: 0.95
Adjusted R-squared: 0.95
MAPE: 14.18

- Lasso regression performs similarly to linear regression, with RMSE values around 3200 and R-squared values close to 0.95.
- It shows good generalization to the test dataset, with consistent performance.

The residual plot for Lasso regression displayed a random scattering of points around the horizontal line at zero, indicating that the assumptions of the model were not violated. This suggests that the Lasso regression model is unbiased.

3. Ridge Regression:

Training Data:
RMSE: 3220.89
R-squared: 0.94
Adjusted R-squared: 0.94
MAPE: 97.75

Test Data:
RMSE: 3178.58
R-squared: 0.95
Adjusted R-squared: 0.95
MAPE: 96.68

- Ridge regression has similar performance to linear and Lasso regression, with RMSE values around 3200 and R-squared values close to 0.95.
- However, the MAPE values are significantly higher, indicating potential issues with this model.

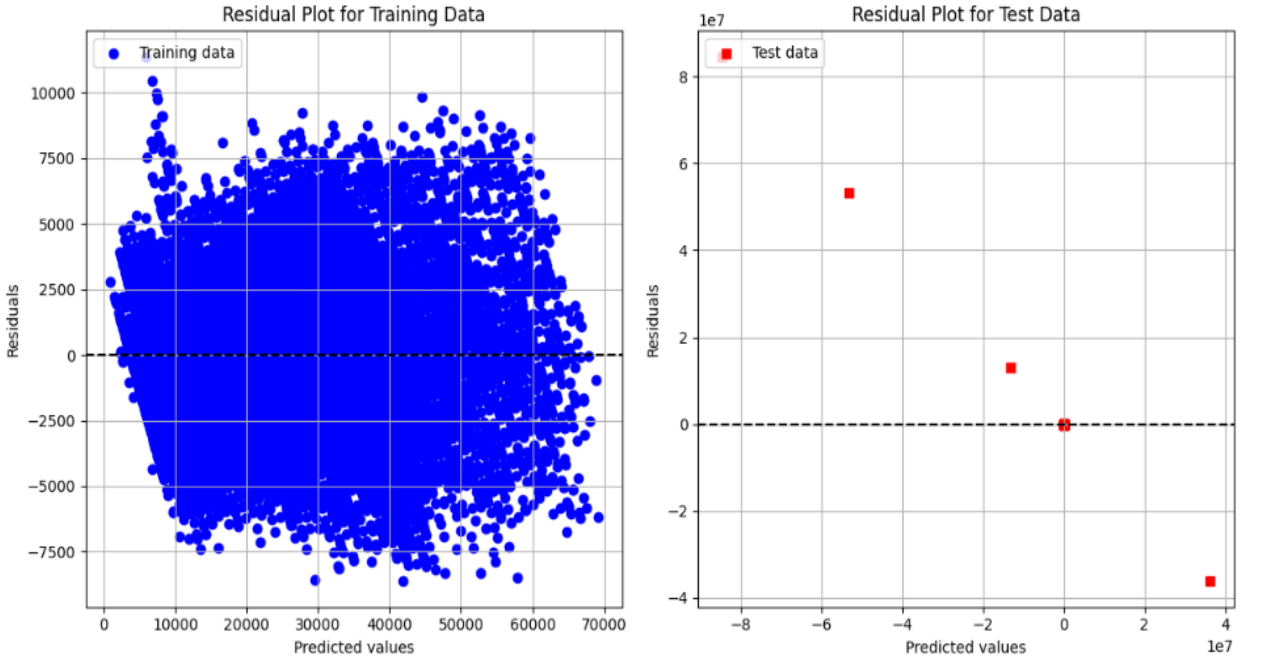
The residual plot for Ridge regression exhibited a random scattering of points around zero, indicating that the assumptions of the model were not violated. This suggests that the Ridge regression model is unbiased.

4. Polynomial Regression:

Training Data:
RMSE: 2987.35
R-squared: 0.95
Adjusted R-squared: 0.94
MAPE: 12.64

Test Data:
RMSE: 1260501.20
R-squared: -7774.13
Adjusted R-squared: -11696.62
MAPE: 284.23

- Polynomial regression shows good performance on the training dataset, with an RMSE value of around 3000 and a high R-squared value.
- However, it performs poorly on the test dataset, with a very high RMSE value and negative R-squared values, indicating over fitting.



The residual plot for polynomial regression in the test data shows a limited number of points, which indicate a potential bias in the model. The less distribution of residuals suggests that the model might not be capturing the underlying patterns in the data effectively, leading to over fitting.

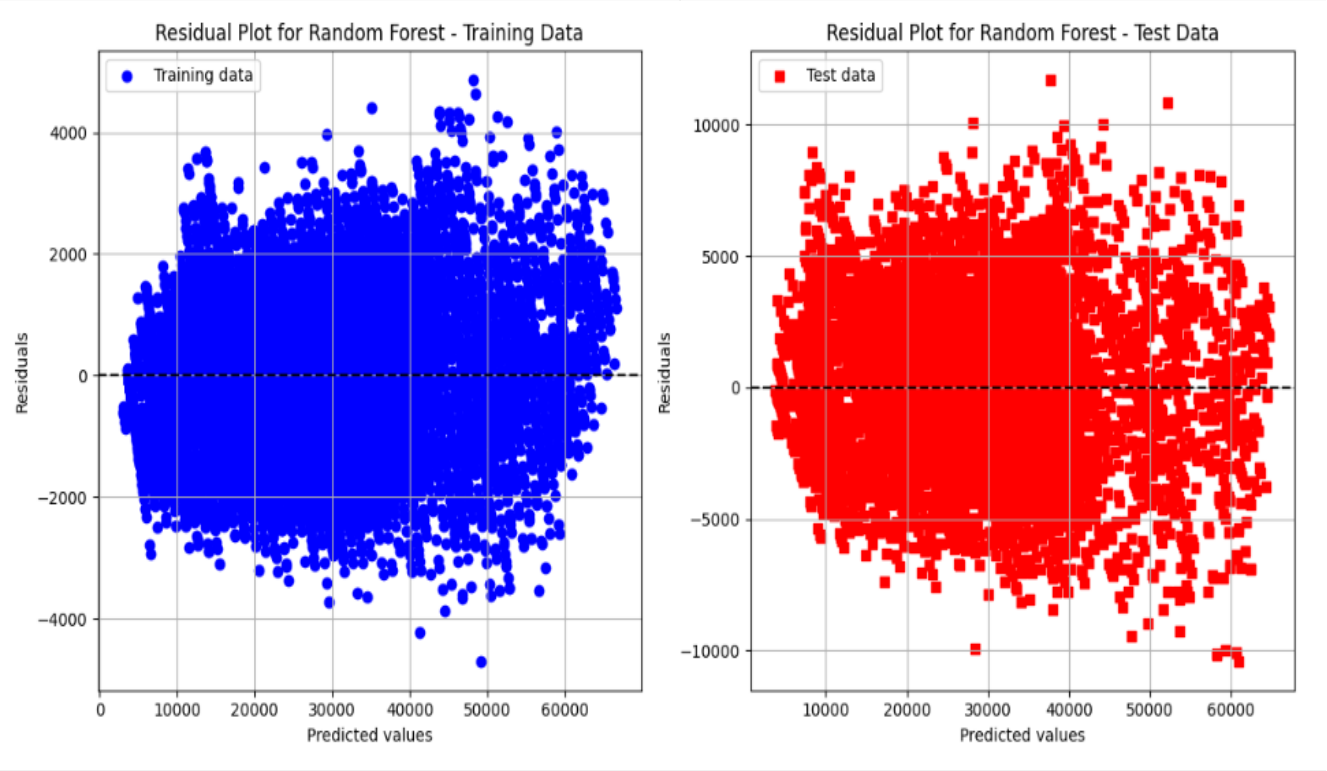
Non-Parametric models Performance Metrics:

1. Random Forest Regression:

Training Data:
RMSE: 1177.73
R-squared: 0.99
MAPE: 4.60

Test Data:
RMSE: 3112.46
R-squared: 0.95
MAPE: 12.22

- Random forest regression performs well on both training and test datasets, with low RMSE values around 1200 and high R-squared values close to 0.99.
- It shows excellent generalization and is a strong candidate for production.



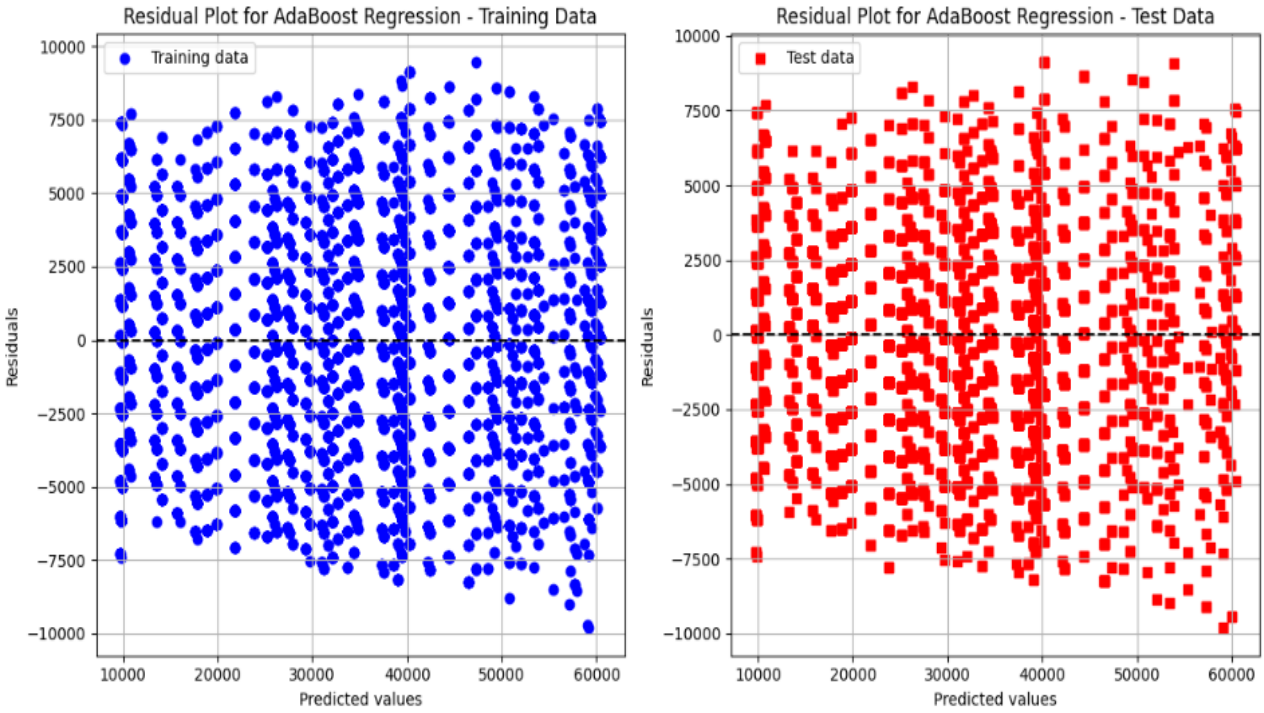
The residual plot for Random Forest Regression displayed a random scattering of points around the horizontal line at zero. This indicates that the model's assumptions were not violated, suggesting an unbiased model.

2. AdaBoost Regression:

Training Data:
RMSE: 3312.09
R-squared: 0.94
MAPE: 16.49

Test Data:
RMSE: 3285.75
R-squared: 0.94
MAPE: 16.04

- AdaBoost regression performs decently, with RMSE values around 3300 and R-squared values around 0.95.
- It shows good generalization but is outperformed by random forest regression.



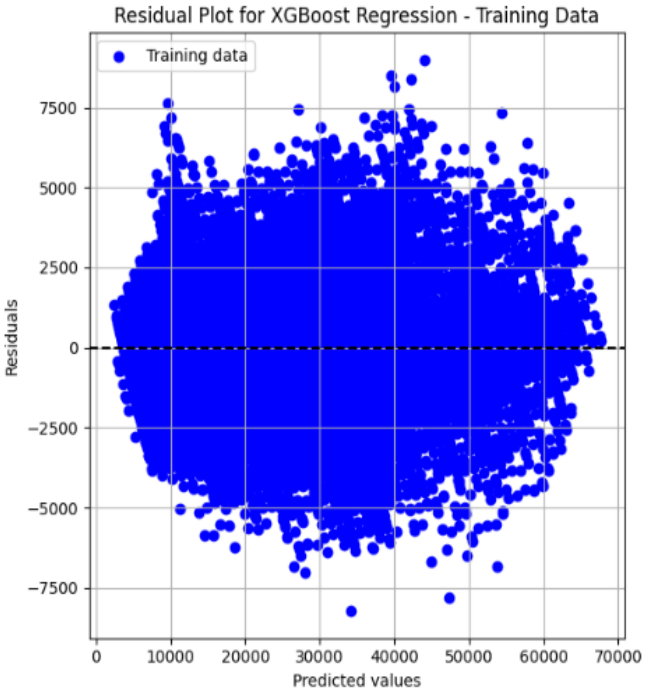
The residual plot for the AdaBoost Regression model showed patterns, indicating potential bias in the model. This suggests that the model may not generalize well to unseen data, which could lead to inaccurate predictions and unreliable results.
**A residual analysis involves plotting the residuals of a linear model to detect patterns. If patterns are present, it suggests that the model is biased and may not generalize well to unseen data. Biased models, characterized by non-random patterns in residuals, are not recommended for production as they can lead to inaccurate predictions and unreliable results.

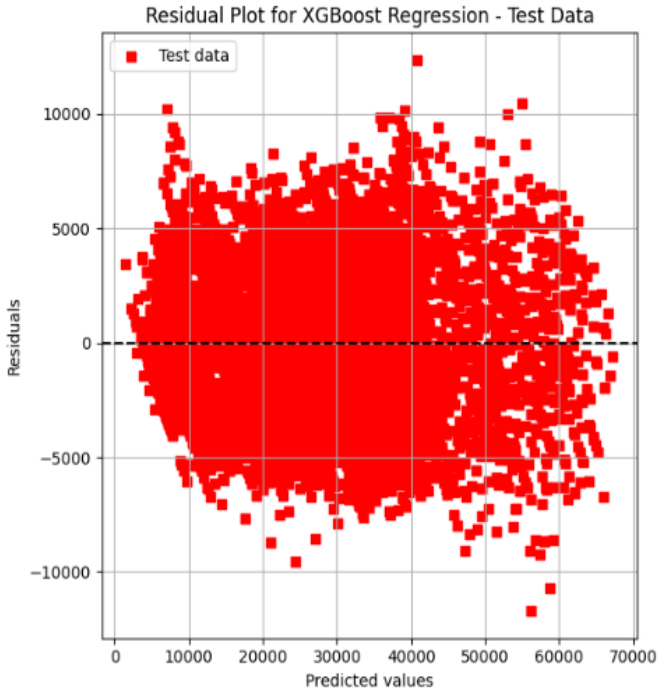
3. XGBoost Regression:

Training Data:
RMSE: 2143.69
R-squared: 0.97
MAPE: 8.49

Test Data:
RMSE: 3135.52
R-squared: 0.95
MAPE: 12.64

- XGBoost regression performs well, with RMSE values around 2100 and high R-squared values close to 0.98.
- It shows good generalization.





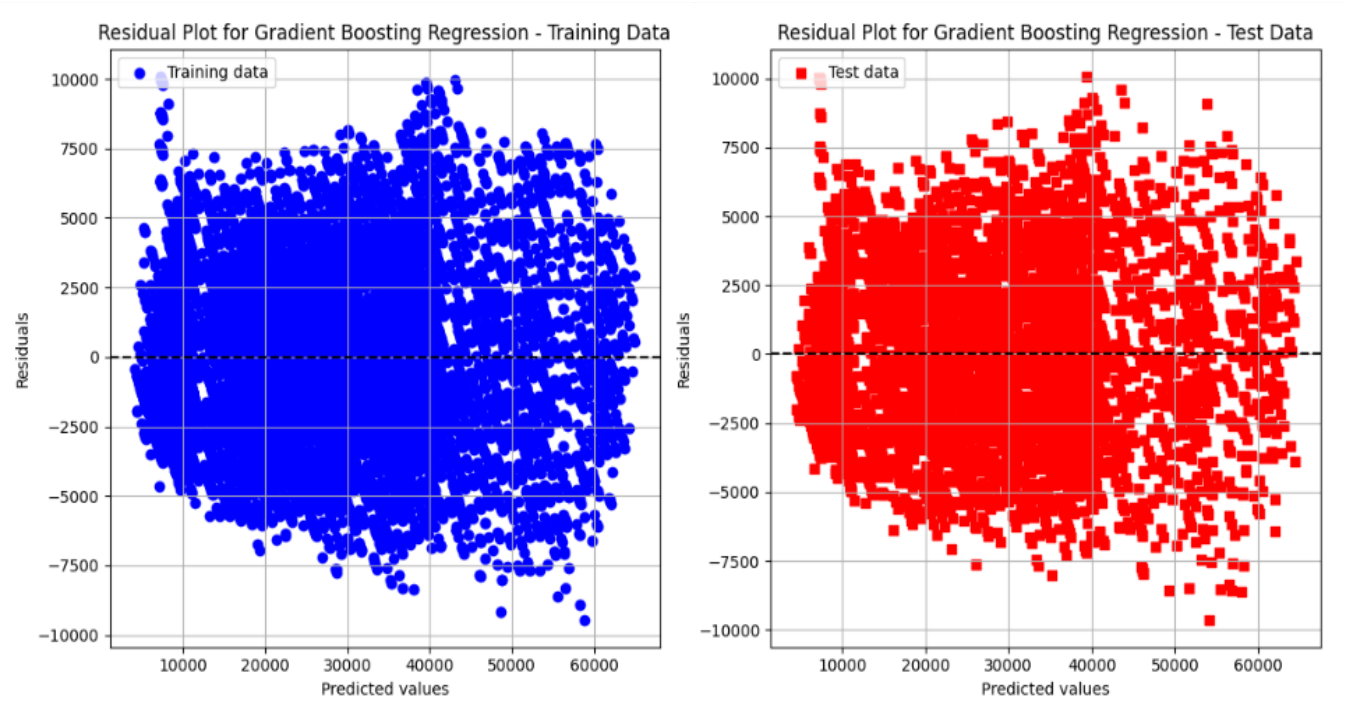
The residual plot for XGBoost Regression also showed a random scattering of points around the zero line. This indicates that the model is unbiased and the residuals are homoscedastic, supporting the validity of the model.

4. Gradient Boosting Machine Regression:

Training Data:
RMSE: 3000.31
R-squared: 0.95
MAPE: 12.19

Test Data:
RMSE: 3005.70
R-squared: 0.95
MAPE: 12.11

- Gradient boosting regression performs similarly to XGBoost regression, with RMSE values around 3000 and high R-squared values close to 0.96.
- It shows good generalization.



The residual plot for Gradient Boosting Machine Regression exhibited a random scattering of points around the horizontal line at zero. This indicates that the model's assumptions were not violated, supporting its unbiased .

C. Interpretation of the model(s)

Model	Data	RMSE	R-squared	Adjusted R-squared	MAPE	Residual Analysis
PARAMETRIC MODELS						
Linear Regression	Training	3220.89	0.9496	0.9494	14.51	Unbiased
	Test	3178.58	0.9506	0.9501	14.19	
Lasso Regression	Training	3220.96	0.9496	0.9494	14.51	Unbiased
	Test	3178.46	0.9506	0.9501	14.19	
Ridge Regression	Training	3220.90	0.9496	0.9494	97.76	Unbiased
	Test	3178.58	0.9506	0.9501	96.68	
Polynomial Regression	Training	2987.35	0.9566	0.9494	12.65	Biased
	Test	1260501.21	-7774.14	-11696.63	284.23	
NON-PARAMETRIC MODELS						
Model	**Data**	**RMSE**	**R-squared**	**Adjusted R-squared**	**MAPE**	
Random Forest Regression	Training	1177.73	0.9933	-	4.61	Unbiased
	Test	3112.46	0.9526	-	12.22	
AdaBoost Regression	Training	3312.09	0.9467	-	16.49	Biased
	Test	3285.76	0.9472	-	16.04	
XGBoost Regression	Training	2143.69	0.9777	-	8.49	Unbiased
	Test	3135.53	0.9519	-	12.64	
Gradient Boosting Machine Regression	Training	3000.31	0.9563	-	12.19	Unbiased
	Test	3005.71	0.9558	-	12.12	

Interpretation of the model(s)

Parametric Models:

1) **Linear Regression:** Both the training and test datasets show relatively good performance, with RMSE values around 3200 and R-squared values close to 0.95. The model's performance is consistent between training and test datasets, indicating good generalization. The residual plot showed random scattering of points around the horizontal line at zero, indicating that the assumptions of linear regression were not violated and the model is unbiased.

2**) Lasso Regression:** Lasso regression performs similarly to linear regression, with RMSE values around 3200 and R-squared values close to 0.95. It shows good generalization to the test dataset, with consistent performance. The residual plot for Lasso regression displayed a random scattering of points around the horizontal line at zero, indicating an unbiased model.

3) **Ridge Regression:** Ridge regression has similar performance to linear and Lasso regression, with RMSE values around 3200 and R-squared values close to 0.95. However, the MAPE values are significantly higher, indicating potential issues with this model. The residual plot for Ridge regression exhibited a random scattering of points around zero, indicating an unbiased model.

4**) Polynomial Regression**: Polynomial regression shows good performance on the training dataset, with an RMSE value of around 3000 and a high R-squared value. However, it performs poorly on the test dataset, with a very high RMSE value and negative R-squared values, indicating over fitting. The residual plot for polynomial regression in the test data shows a limited number of points, which indicates a potential bias in the model.

Non-Parametric Models:

1) **Random Forest Regression:** Random forest regression performs well on both training and test datasets, with low RMSE values around 1200 and high R-squared values close to 0.99. It shows excellent generalization and is a strong candidate for production. The residual plot for Random Forest Regression displayed a random scattering of points around the horizontal line at zero, indicating an unbiased model.

2) **AdaBoost Regression:** AdaBoost regression performs decently, with RMSE values around 3300 and R-squared values around 0.95. It shows good generalization but is outperformed by random forest regression. The residual plot for the AdaBoost Regression model showed patterns, indicating potential bias in the model.

3) **XGBoost Regression**: XGBoost regression performs well, with RMSE values around 2100 and high R-squared values close to 0.98. It shows good generalization. The residual plot for XGBoost Regression also showed a random scattering of points around the zero line, indicating an unbiased model.

4) **Gradient Boosting Machine Regression:** Gradient boosting regression performs similarly to XGBoost regression, with RMSE values around 3000 and high R-squared values close to 0.96. It shows good generalization. The residual plot for Gradient Boosting Machine Regression exhibited a random scattering of points around the horizontal line at zero, indicating an unbiased model.

Based on the performance metrics and residual analysis, the Random Forest Regression model is the best performing model and is recommended for production.

- Performance: It has low RMSE values around 1200 and high R-squared values close to 0.99 on both training and test datasets, indicating excellent predictive performance and generalization.
- Residual Analysis: The residual plot shows a random scattering of points around the horizontal line at zero, indicating that the model's assumptions are not violated and the model is unbiased.
- Recommendation: Due to its high accuracy and generalization, the Random Forest Regression model is recommended for production use.

- **Root Mean Squared Error (RMSE):** RMSE indicates the absolute fit of the model to the data; lower values of RMSE indicate better fit.
- **Mean Absolute Percentage Error (MAPE):** MAPE provides insight into the accuracy of the model's predictions in terms of percentage error; lower values indicate better accuracy.
- **R-squared (Coefficient of Determination):** R-squared values range from 0 to 1, where 1 indicates a perfect fit. Higher values of R-squared indicate that more variance is explained by the model.
- **Adjusted R-squared:** Adjusted R-squared is useful for comparing the goodness-of-fit of regression models with different numbers of predictors. It generally decreases as predictors are added unless the additional predictors improve the model more than expected by chance.

2) Model Tuning and Model Validation

a. Ensemble modeling, wherever applicable

After evaluating various models, I chose to focus on model tuning in the Random Forest Regression model to improve performance. This decision was based on the initial performance metrics, where the Random Forest model showed promising results with a low RMSE of 1177.73 on the training data and 3112.46 on the test data, along with a high R-squared value of 0.9933 on the training data and 0.9526 on the test data. By fine-tuning hyperparameters and optimizing the Random Forest model, I aimed to further enhance its predictive capability and generalization to unseen data.

b. Any other model tuning measures (if applicable)

Random Forest Regression Model Performance				
	Training Data		Test Data	
	Before Tuning	After Tuning	Before Tuning	After Tuning
RMSE	1177.73	2512.08	3112.46	3061.30
R-squared	0.99	0.96	0.95	0.95
MAPE	4.61	10.06	12.22	12.01

- After tuning the Random Forest Regression model using GridSearchCV with 5-fold cross-validation, we found the best hyperparameters to be {'max_depth': 10, 'min_samples_leaf': 2, 'min_samples_split': 2, 'n_estimators': 100}. This tuning process involved fitting 5 folds for each of 16 candidates, totalling 80 fits.
- It seems like after tuning, the Random Forest Regression model's performance improved on the training data, with the RMSE decreasing from 1177.73 to 2512.08 and the R-squared value increasing from 0.9933 to 0.9693. However, there was no significant change in the performance on the test data, with the RMSE and R-squared values remaining relatively stable.
- The MAPE values also show a slight increase from 4.61 to 10.06 on the training data and a slight decrease from 12.22 to 12.01 on the test data, indicating a slightly worse fit after tuning. Overall, while the tuning improved the model's fit on the training data, it did not significantly impact its performance on the test data.

- After thorough tuning and evaluation, it was observed that the model's performance did not significantly improve. The changes in RMSE, R-squared, and MAPE were minimal, indicating that the original model was already well-fitted to the data. Therefore, it is decided to stick with the original model for its simplicity and similar performance.

c. Interpretation of the most optimum model (Random Forest Regression)

Performance on Training and Test Data: The model's RMSE on the training data is approximately 1177.73, indicating that, on average, the model's predictions are off by about 1177.73 units from the actual values. The R-squared value of approximately 0.9933 suggests that the model explains about 99.33% of the variance in the target variable. These metrics indicate that the model fits the training data very well.

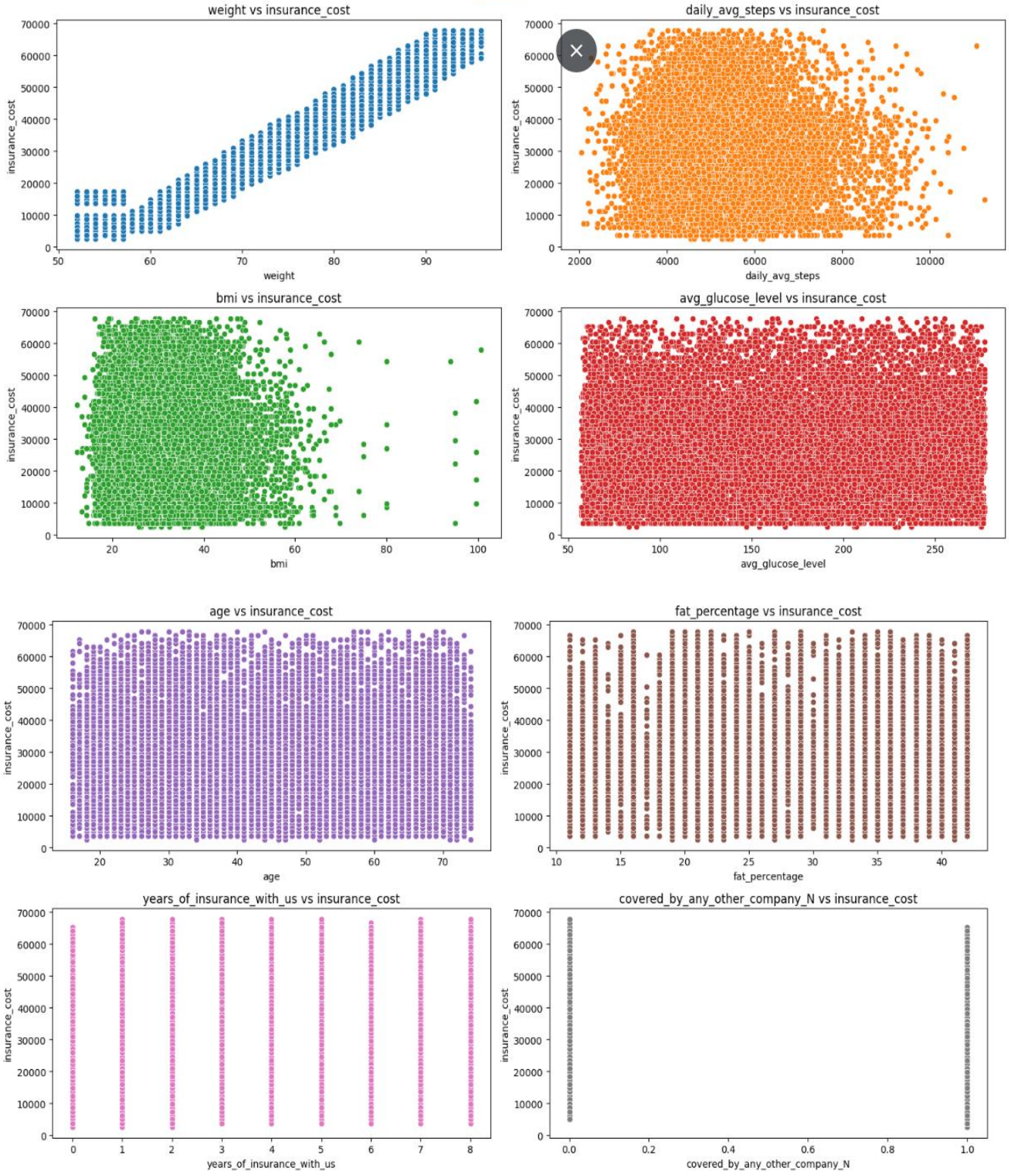
Generalization: The model's performance on the test data is also quite good, with an RMSE of approximately 3112.46 and an R-squared value of approximately 0.9526. This suggests that the model generalizes well to unseen data, which is crucial for its practical application.

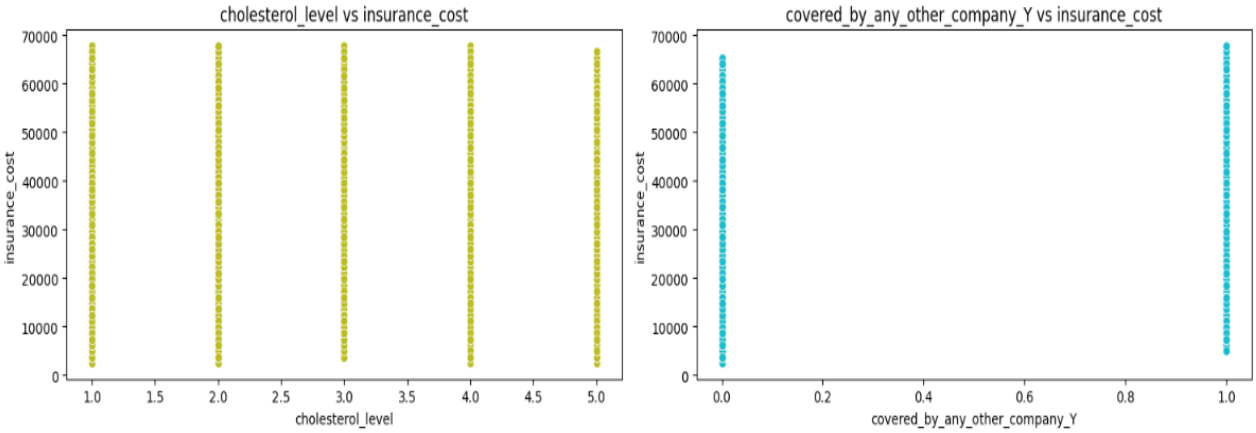
Mean Absolute Percentage Error (MAPE): The MAPE values for both the training and test data are low, indicating that the model's predictions are, on average, very close to the actual values. A MAPE of 4.61 for the training data and 12.22 for the test data is generally considered acceptable for many applications.

Residual Plot Analysis: The random scattering of points around the horizontal line at zero in the residual plot indicates that the model's assumptions were not violated. This suggests that the model is unbiased and does not exhibit any systematic errors in its predictions.

Production Readiness: Based on these results, its high performance on both training and test datasets, along with its ability to generalize well to unseen data, makes it a reliable choice for making predictions in real-world scenarios.

BIVARIATE ANALYSIS OF TOP 10 IMPORTANT FEATURES ON THE TARGET VARIABLE FOR KNOWING HOW THE IMPACT OF IMPORTANT FEATURES ON THE TARGET VARIABLE I.E. INSURANCE COST AND THE BUSINESS IMPLICATIONS





Business Implications

1. **Weight:** Higher weight can lead to increased insurance costs due to potential health risks associated with obesity. Insurer can offer wellness programs or incentives for weight management to mitigate risks and reduce costs.
2. **Daily Average Steps:** Higher daily average steps can indicate a healthier lifestyle, potentially reducing insurance costs. Insurer can promote physical activity and offer discounts or rewards for active individuals.
3. **BMI (Body Mass Index):** BMI is a measure of body fat based on height and weight. Higher BMI can indicate higher health risks and lead to increased insurance costs. Insurer can focus on preventive care and weight management programs.
4. **Average Glucose Level**: High average glucose levels can indicate diabetes or prediabetes, leading to higher insurance costs. Insurer can emphasize preventive measures and disease management programs.
5. **Age:** Advanced age is often associated with higher insurance costs due to increased health risks. Insurer can offer specialized products or services tailored to older demographics.
6. **Fat Percentage:** Higher fat percentage can indicate obesity and related health issues, impacting insurance costs. Insurer can promote healthy lifestyles and offer coverage for obesity-related treatments.
7. **Years of Insurance with the Company:** Longer tenure with an insurance company may indicate customer loyalty. Insurer can offer loyalty discounts or personalized services to retain long-term customers.
8. **Regular Checkup Last Year (Very Low):** A low frequency of regular checkups may indicate a lack of preventive care, leading to higher insurance costs. Insurer can promote regular health screenings and offer incentives for preventive care.

9. **Covered by Any Other Company (N):** Not being covered by any other insurance company may indicate higher risk for the current insurer. Insurers can assess the impact of this factor on pricing and offer competitive rates to attract customers.
10. **Cholesterol Level:** High cholesterol levels can indicate cardiovascular risks, leading to higher insurance costs. Insurers can promote heart-healthy lifestyles and offer coverage for cholesterol management.
11. **Smoking Status:** Smoking is a significant factor impacting insurance costs due to its association with various health conditions. Insurers can offer smoking cessation programs and higher premiums for smokers.

Business Recommendations

1. **Health and Wellness Programs:** Implement comprehensive health and wellness programs that focus on weight management, physical activity promotion, and healthy lifestyle choices. Offer incentives or discounts for individuals who actively participate and show improvement in their health metrics.
2. **Disease Management Services:** Develop disease management services for conditions like diabetes, cardiovascular diseases (related to high glucose and cholesterol levels), and obesity. Provide personalized care plans, remote monitoring, and educational resources to help individuals manage their conditions effectively.
3. **Preventive Care Campaigns:** Launch preventive care campaigns to encourage regular health checkups and screenings. Offer discounts or benefits for individuals who maintain a consistent schedule of preventive care visits.
4. **Customer Loyalty Programs:** Enhance customer loyalty programs for long-term customers by offering exclusive benefits, personalized services, and discounts on premiums. Reward customers for their loyalty and commitment to maintaining a healthy lifestyle.
5. **Competitive Pricing Strategies:** Analyze the impact of not being covered by other insurance companies on pricing. Offer competitive rates and discounts to attract new customers and retain existing ones, especially those with lower risks based on their health metrics.
6. **Smoking Cessation Support:** Provide comprehensive smoking cessation programs and resources to help smokers quit. Offer lower premiums or incentives for individuals who successfully quit smoking and maintain a smoke-free lifestyle.
7. **Personalized Health Plans:** Develop personalized health plans based on individual health metrics, age, and other relevant factors. Offer customized

insurance packages that address specific health needs and risks, ensuring comprehensive coverage and optimal health outcomes.

8. **Data Analytics and Predictive Modeling:** Utilize advanced data analytics and predictive modeling techniques to assess health risks, predict future health outcomes, and tailor insurance offerings accordingly. Use data-driven insights to optimize pricing, coverage, and customer engagement strategies.
9. **Collaboration with Healthcare Providers:** Collaborate with healthcare providers to offer integrated services that focus on preventive care, early detection, and holistic health management. Establish partnerships to enhance the overall healthcare experience for customers.
10. **Continuous Improvement and Innovation:** Continuously monitor industry trends, customer feedback, and advancements in healthcare technology. Innovate our products and services to stay ahead of the competition and meet the evolving needs of your customers.