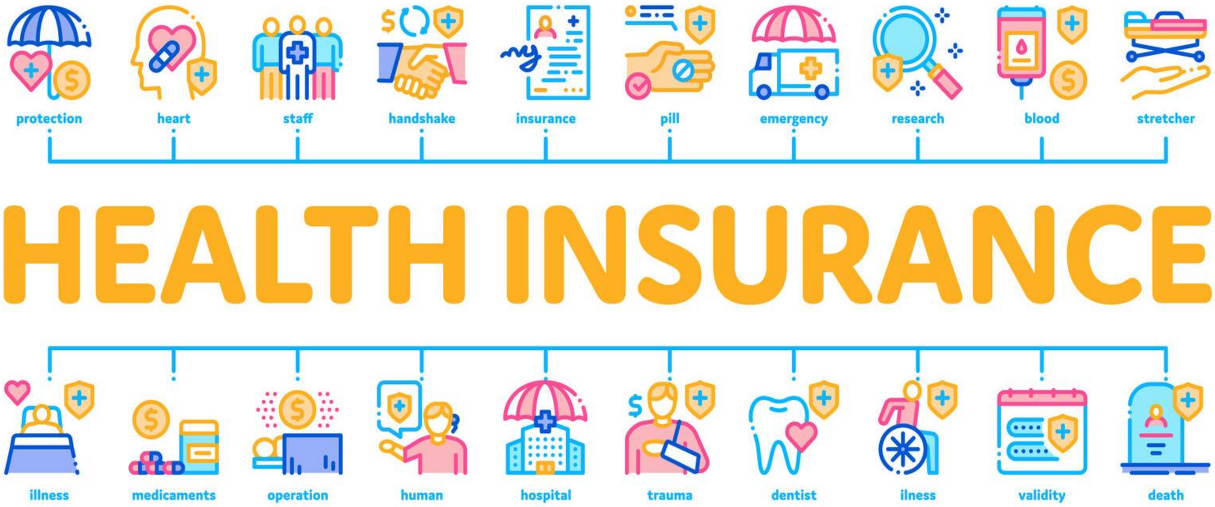


PROJECT

GREAT LEARNING

Authored By,
Shrinidhi CG

Contents

Introduction of the business problem

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. 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.

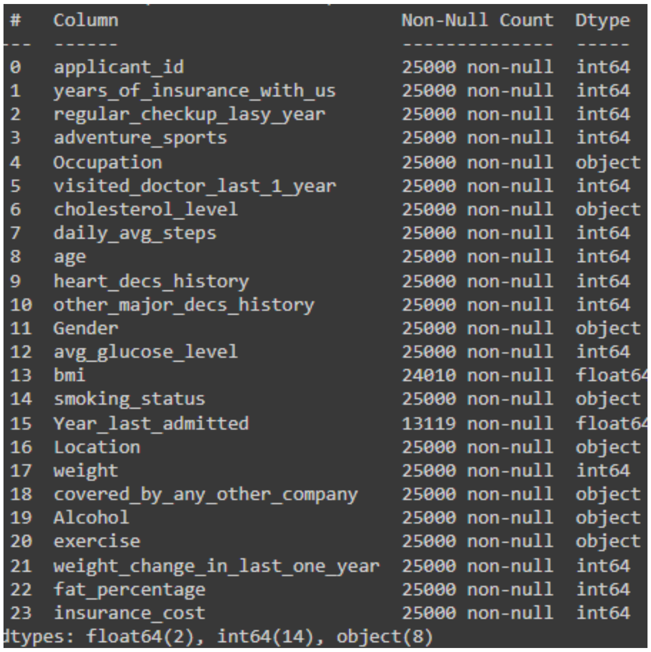
Need for the Study/Project:

This study presents a significant business opportunity for insurance companies to enhance profitability by accurately pricing premiums based on individual health data. Personalized insurance costs can attract more customers, reduce claim risks, and lower medical expenses.

From a social perspective, the model fosters fairness in insurance pricing, making healthcare more accessible and affordable. It also encourages individuals to adopt healthier habits, benefiting both their well-being and financial security.

Data Report

Visual inspection of data & understanding of attributes



```
#   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)
```

- The dataset contains 25,000 rows and 24 columns, capturing various features related to health and insurance, such as applicant demographics, health history, and lifestyle habits.

- There are both numerical and categorical variables, including information on age, bmi, exercise habits, and insurance costs.

- Some columns have missing values (e.g., bmi and Year last admitted), indicating the need for data cleaning or imputation.

Data Cleaning:

Handling Missing Values:

BMI: To handle the missing values in the bmi column, which constituted approximately 3.96% of the data, we imputed these missing values using KNN imputation method. This method helps to fill the gaps in a way that minimally affects the overall distribution of the data. Following this imputation, the dataset no longer has missing values in the bmi column.

Removing unwanted variables:

- Year last admitted: Given that approximately 47.52% of the Year last admitted data is missing, it is reasonable to drop this column from the dataset.
- In the data cleaning process, the '**applicant id**' column was dropped because it contained unique identifiers that were insignificant for analysis.
- Location variable is removed after categorizing the cities into Southwest, Southeast, Northwest, and Northeast based on their geographical positions in India

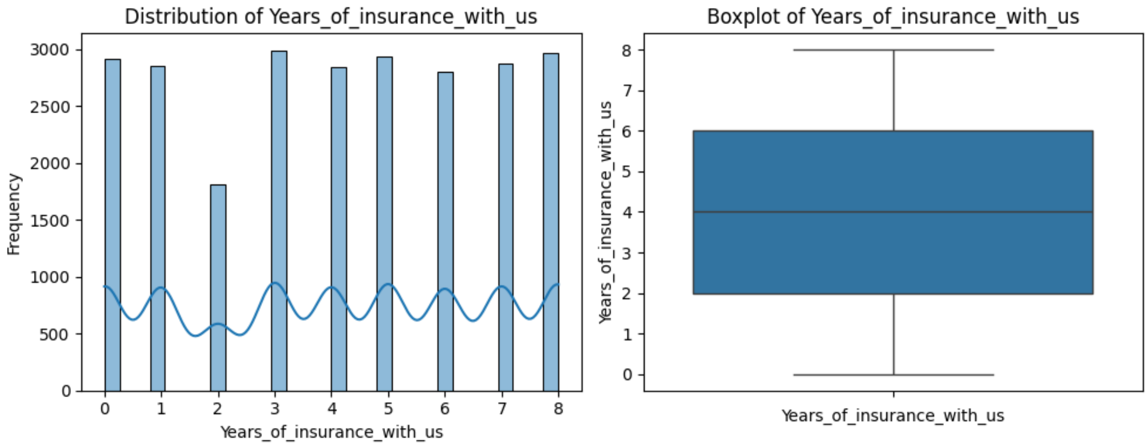
Changing Incorrect Datatype & Spelling

- Several columns ('Adventure sports,' 'Heart decs history,' and 'Other major decs history') were converted to categorical data types to ensure accurate analysis.
- Furthermore, a misspelling in the 'Occupation' column was corrected by replacing 'Salried' with 'Salaried.'

Exploratory data analysis

Univariate analysis of Numerical Variable

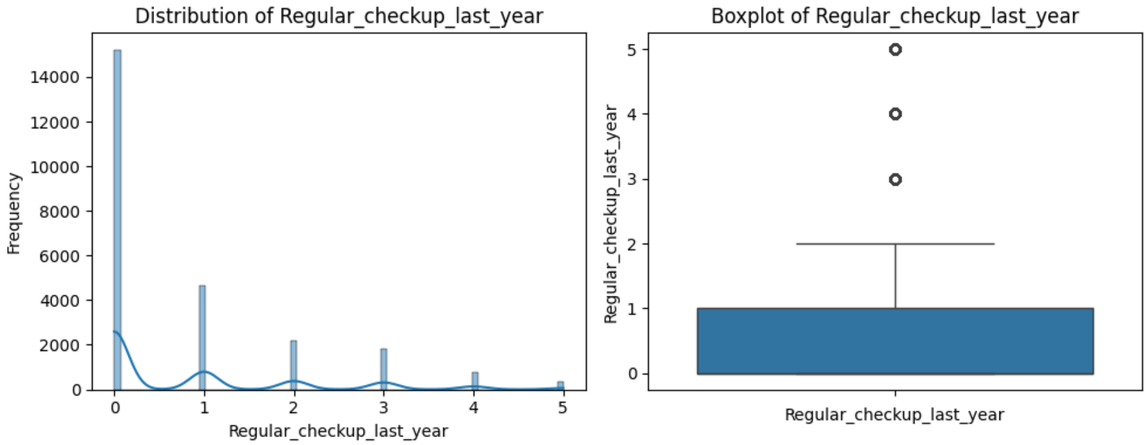
Distribution of Years of insurance with us:



The frequency is nearly the same for all values (0 to 8) with consistent peaks. Each bar represents around 3000 customers across all years except for year 2, which has a lower frequency.

There is a noticeable drop in the number of customers with **2 years of insurance** compared to other years. This dip could suggest that customers are less likely to maintain their insurance around this point. The distribution for other years (0, 1, 3, 4, 5, 6, 7, 8) is relatively consistent. This could imply that customers maintain their insurance across these years.

Distribution of Regular checkup last year

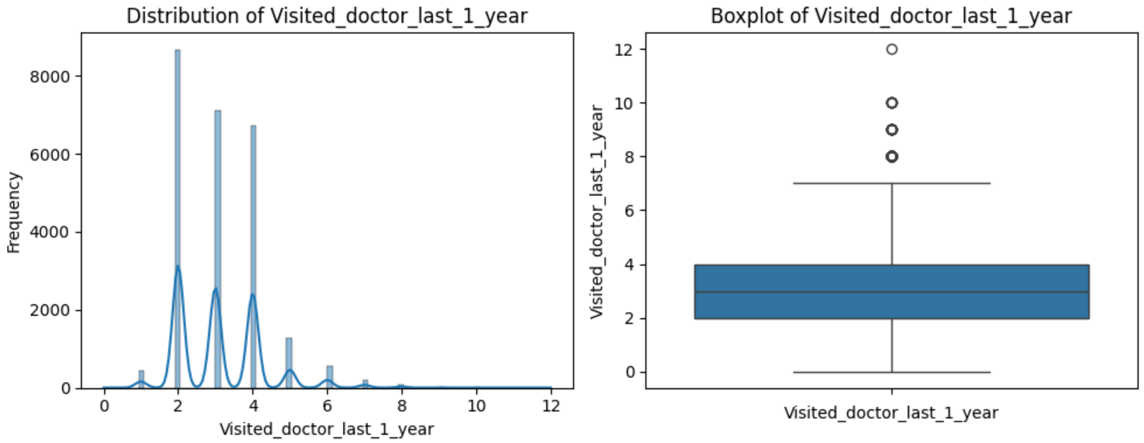


Spike at Zero**:** A majority of customers (spike at 0 checkups) did not attend a regular check-up last year.

Usage Pattern: A small portion of customers had 1 or 2 checkups, with almost negligible numbers having more than 3. The overall low frequency of checkups points to a potential underutilization of available check-up services by customers.

Outliers**:** There are a few outliers, where some customers had 3 or more checkups, but they are rare.

Distribution of Visited doctor last 1 year



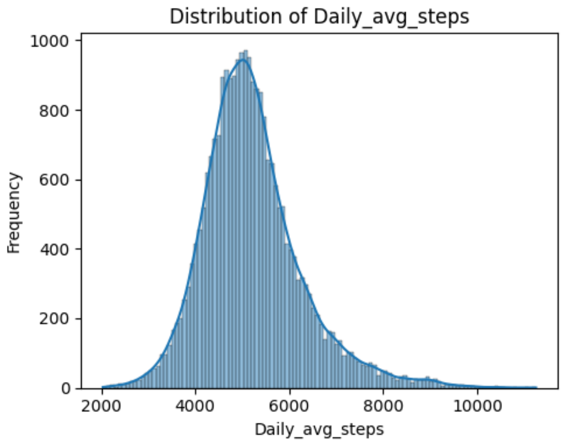
Doctor Visits: The X-axis represents the number of doctor visits, ranging from 0 to 12.

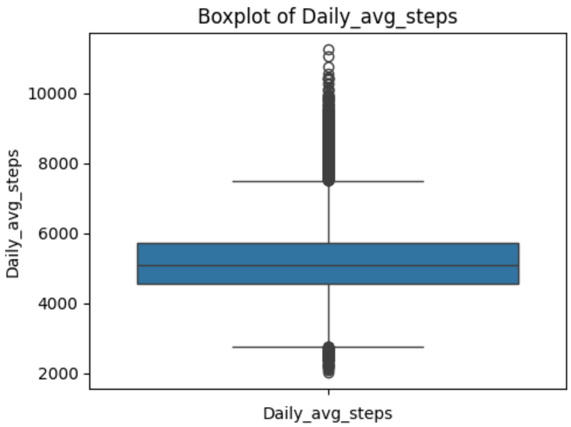
Peaks**:** There are clear peaks at 2, 3, and 4 visits, indicating these are the most common frequencies.

Skewed Distribution**:** The data is heavily skewed to the left, meaning most individuals visited the doctor fewer times (between 0 and 5), with very few people visiting more than 6 times.

Whiskers and Outliers: The whiskers show the range of non-outlier data, extending from 0 to 5. Several outliers are present above the whiskers.

Distribution of Daily Average Steps

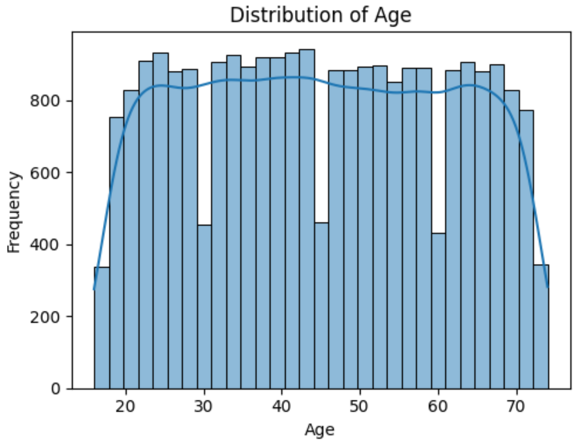


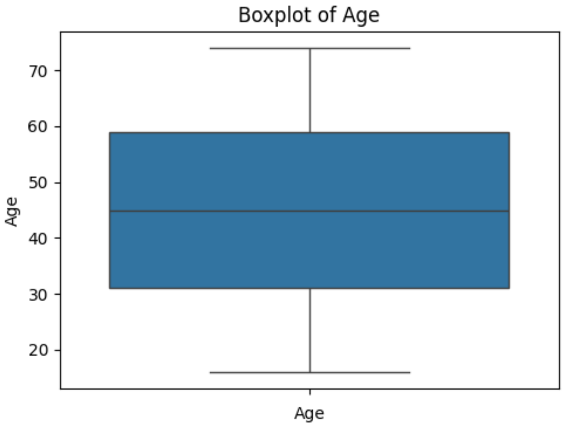


Normal Distribution: The distribution is bell-shaped, suggesting a normal distribution centred around 5,000 to 6,000 steps.

Whiskers and Outliers: The whiskers extend from approximately 3,500 to 7,500 steps, capturing most of the data. Numerous outliers are visible above the top whisker, indicating that some individuals average significantly more than 8,000 steps per day.

Distribution of Age



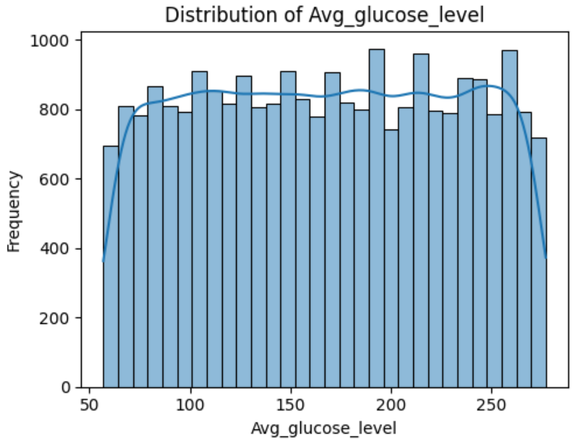


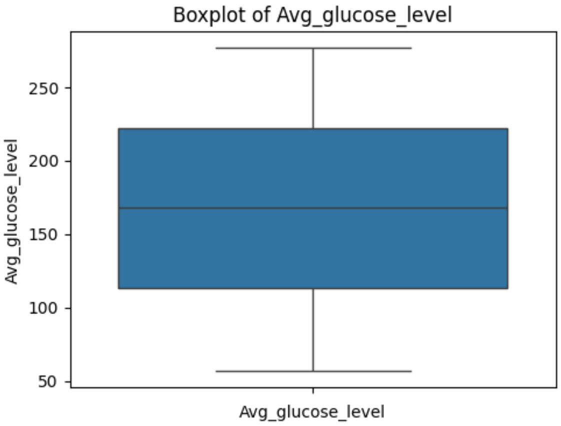
Age Range: The X-axis represents age in years. The distribution appears fairly uniform from age 20 to about 70, with a slight drop-off in between.

Uniform Distribution: The distribution is close to uniform, meaning individuals across different age groups are relatively evenly distributed, with only minor dips between certain age ranges.

Spread: The middle 50% of the data lies between ages 35 and 60, suggesting a concentration of individuals within this range.

Distribution of Avg Glucose Level

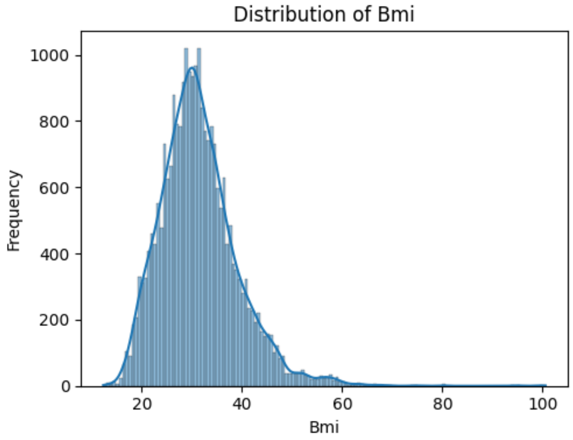


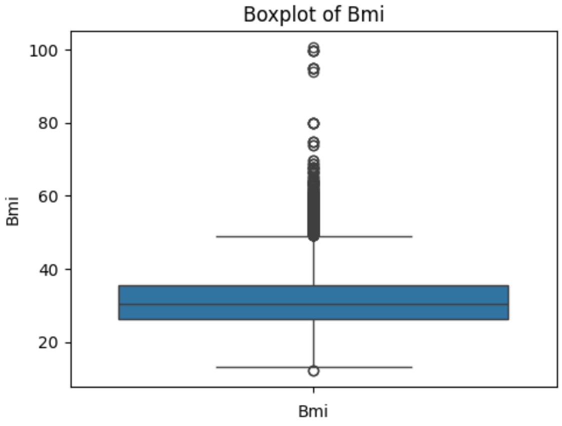


Avg Glucose Level: The X-axis represents glucose levels, ranging from 50 to over 250.

Avg Glucose Level Distribution: The glucose level data is fairly evenly distributed, with the majority of individuals having glucose levels between 80 and 250. There are no significant outliers, and the median glucose level is around 140.

Distribution of BMI

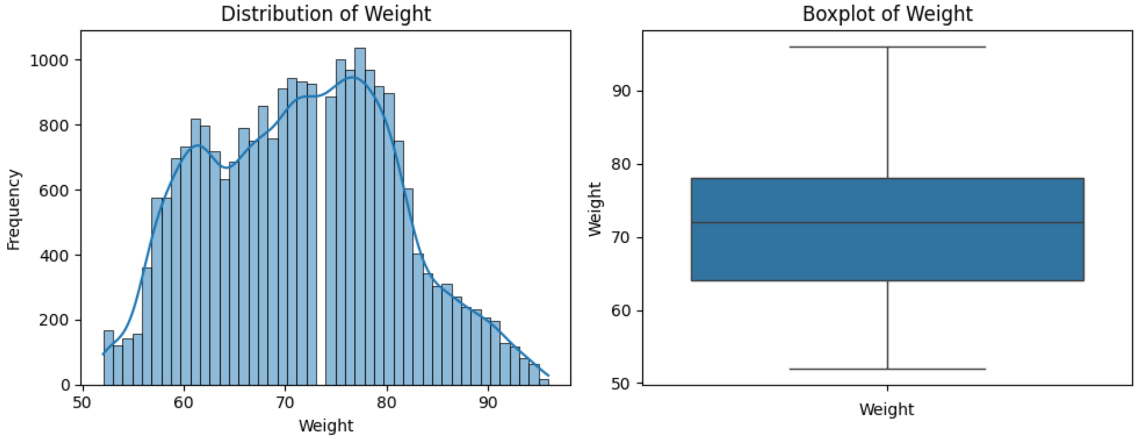




BMI Range: Noticeable peak around 30-35 shows that many individuals have BMIs within that range. Right Skew: The right tail extends to very high BMI values, indicating a small number of individuals with extremely high BMIs (over 50), suggesting obesity or severe obesity.

Outliers: Several dots above the box represent outliers, indicating there are extreme BMI values higher than typical ranges.

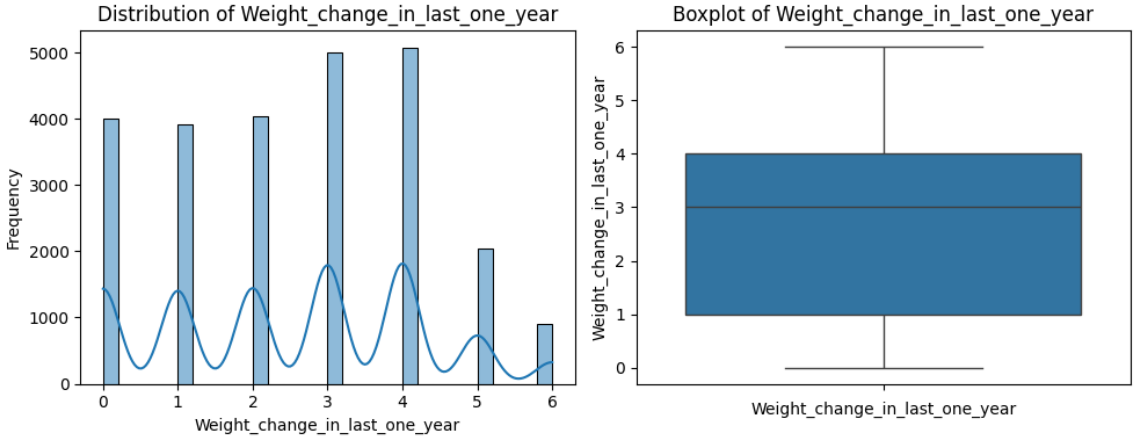
Distribution of Weight



Histogram: Shows the distribution of weight, with most values falling between 60 and 90 kg. The distribution appears relatively normal with peaks around 70-75 kg.

Weight Distribution: More normally distributed, with a central tendency around 70-75 kg, without significant outliers.

Distribution of Weight change in last one year

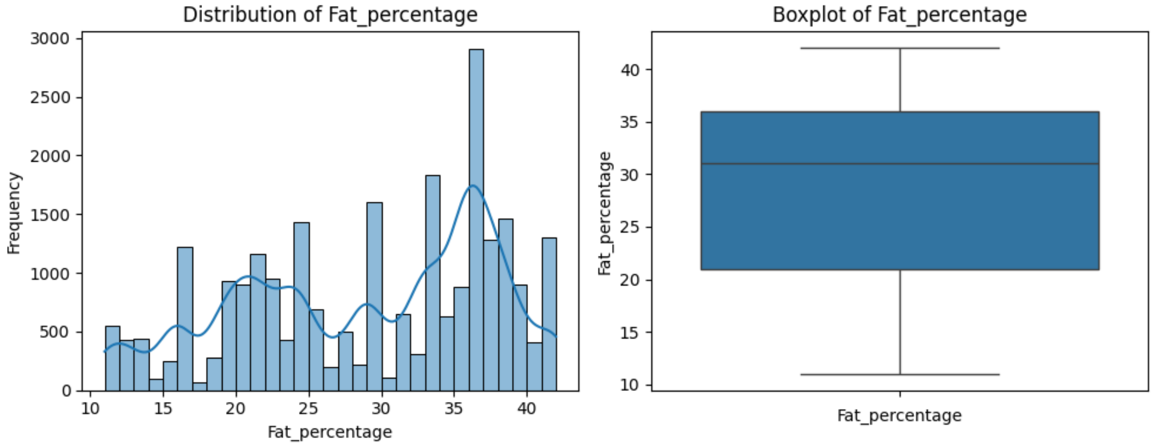


Histogram: The highest frequencies are observed for the values 3 and 4, indicating that most individuals experienced a weight change of around 3 to 4 units in the last year.

Distribution: The distribution is almost uniform for values from 0 to 4, but there is a notable decrease in frequencies beyond 5, indicating fewer occurrences of larger weight changes.

The absence of outliers suggests that extreme weight changes (either very low or very high) were uncommon or not present in the dataset.

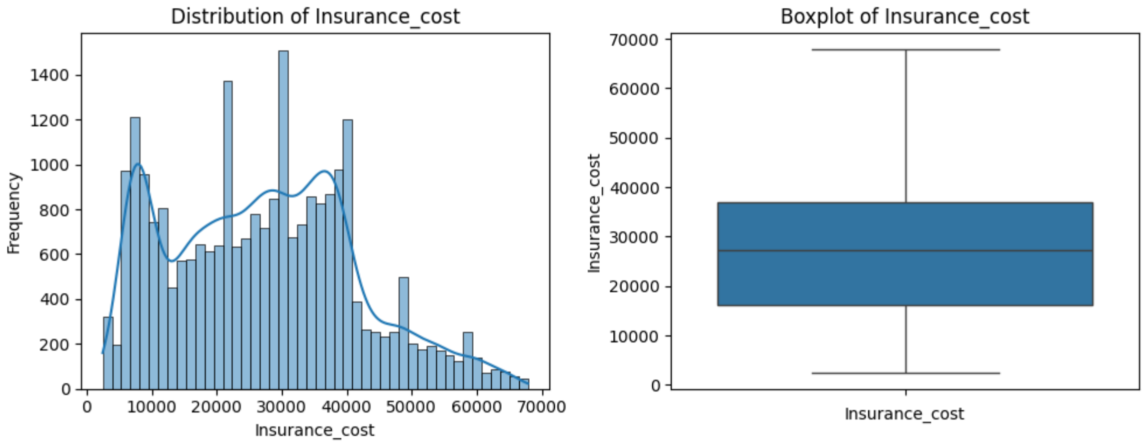
Distribution of Fat percentage



Multiple Peaks: The distribution has multiple peaks around 20%, 30%, and 35%, indicating that the population is segmented into different groups.

The highest concentration is in the 35-40% range, implying that most data points cluster in higher fat percentages. The dip around 25-30% indicates a lower frequency of occurrences in that range.

Distribution of Insurance cost



Histogram: The costs are spread widely, with most data points falling between 10,000 and 50,000. There are notable peaks around 10,000, 20,000, 30,000, and 40,000, suggesting that certain price points may be more common, perhaps due to pricing strategies or policy categories.

Skewed Distribution: The distribution is right-skewed, meaning that most customers have insurance costs between 10,000 and 40,000, with a tail extending to higher values.

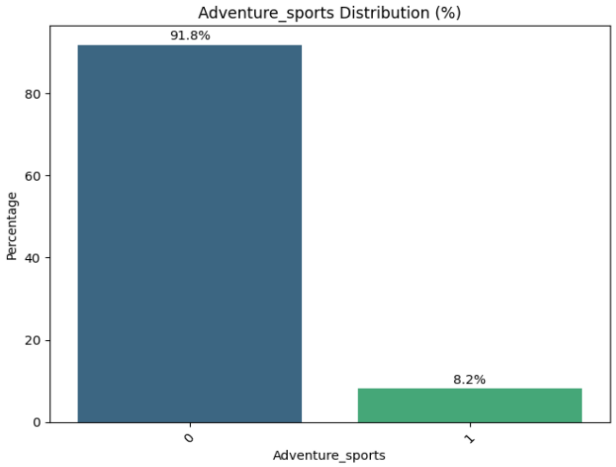
Low Costs: Very few customers have insurance costs beyond 50,000, highlighting that the bulk of the population has mid-range insurance costs.

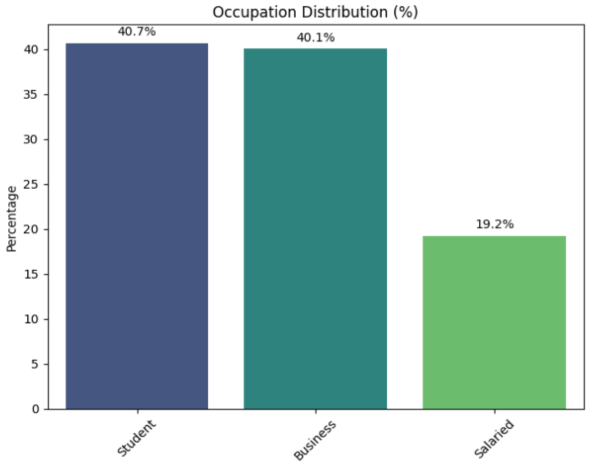
Data Cleaning: Imputing Outlier and why?

2943 outliers in Regular checkup last year, 96 in Visited doctor last 1 year, 952 in Daily avg steps, and 592 in Bmi were replaced with the median.

Imputing outliers is crucial because it helps reduce bias, improves model performance, and maintains data integrity. By addressing outliers, we ensure that statistical analyses and machine learning models yield more accurate and generalizable results.

Exploratory data analysis: Univariate analysis of Categorical Variable





Adventure Sports Distribution

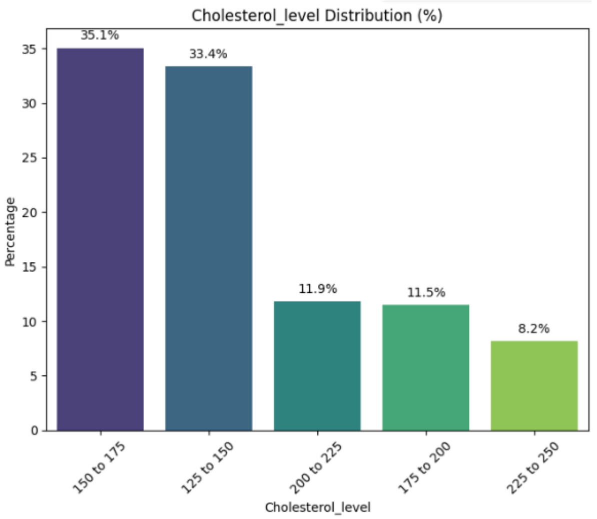
Insight: Majority (91.8%) of individuals do not participate in adventure sports, with only few involved.

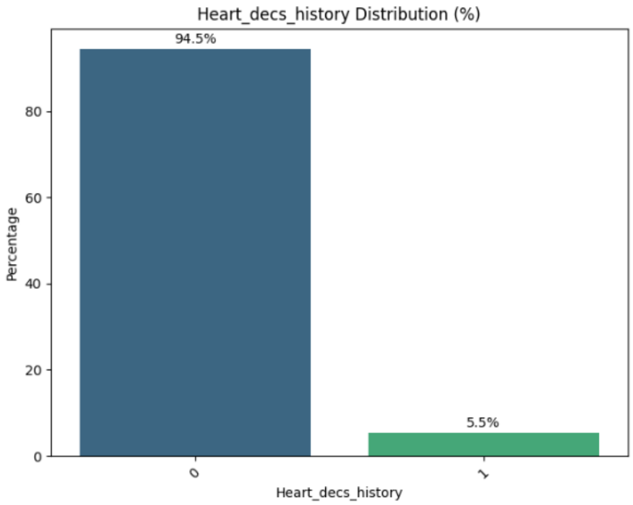
Interpretation: This indicates a low prevalence of adventure sports participation in the population.

Occupation Distribution

Insight: The largest group falls into the "Business" category (40.1%), followed closely by "Student" (40.7%), with a smaller portion (19.2%) in "Salaried" roles.

Interpretation: Business and student populations dominate the data set.





Cholesterol Level Distribution

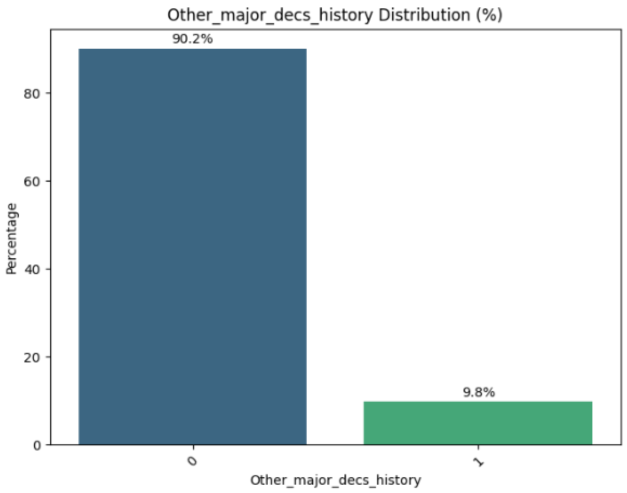
Insight: The highest percentage (35.1%) of individuals have cholesterol levels in the range of 150-175 mg/dL, while smaller percentages are distributed across higher levels.

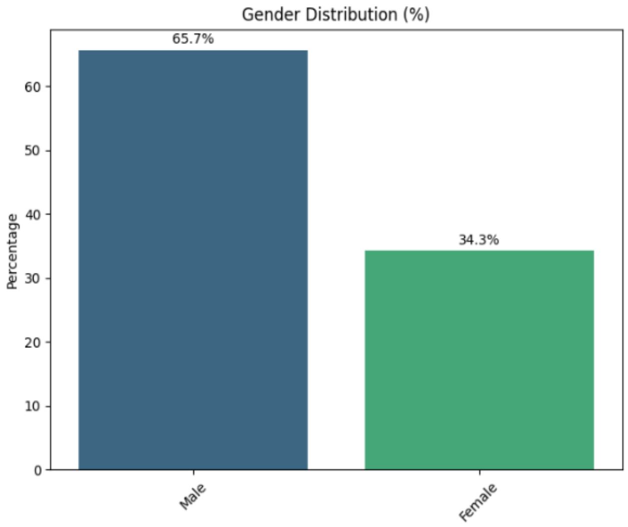
Interpretation: Most people fall into a healthy cholesterol range, with a small fraction at risk due to higher levels.

Heart Disease History Distribution

Insight: A significant majority (94.5%) do not have a history of heart disease, while only 5.5% do.

Interpretation: Heart disease is not common in this sample, but those with heart disease may represent a critical subgroup for targeted interventions.





Other Major Disease History Distribution

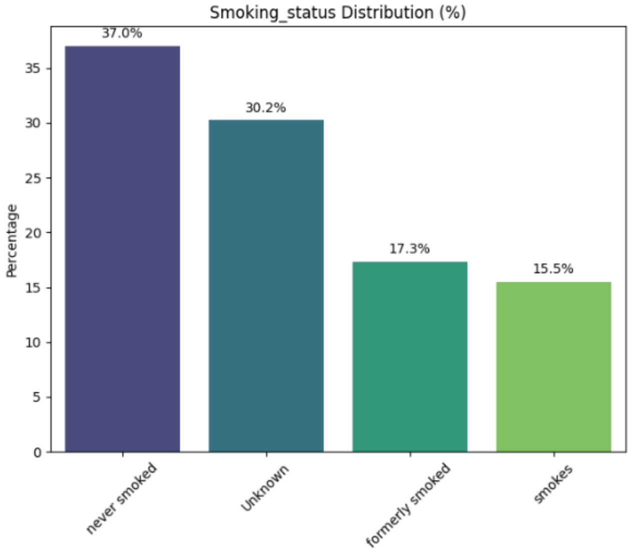
Insight: 90.2% report no major disease history, with 9.8% having experienced another major disease.

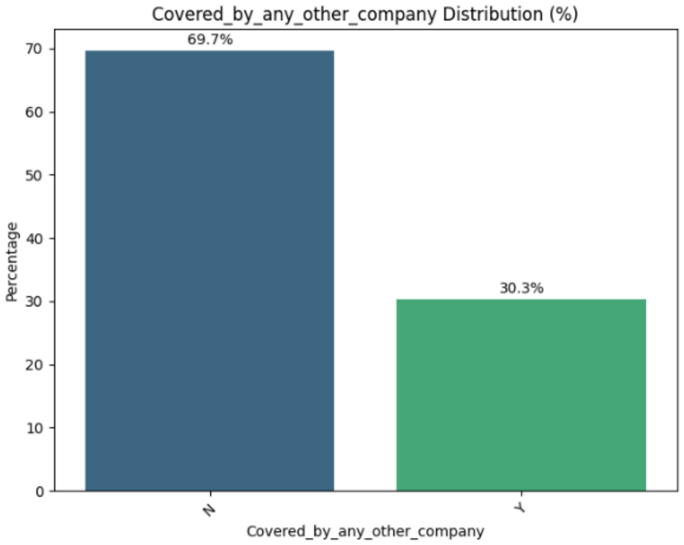
Interpretation: Like heart disease, major diseases are rare in this group, which could suggest a generally healthy population.

Gender Distribution

Insight: The gender distribution is predominantly male (65.7%), with females making up 34.3%.

Interpretation: The male-to-female ratio is skewed, which may have implications for gender-specific analyses.





Smoking Status Distribution

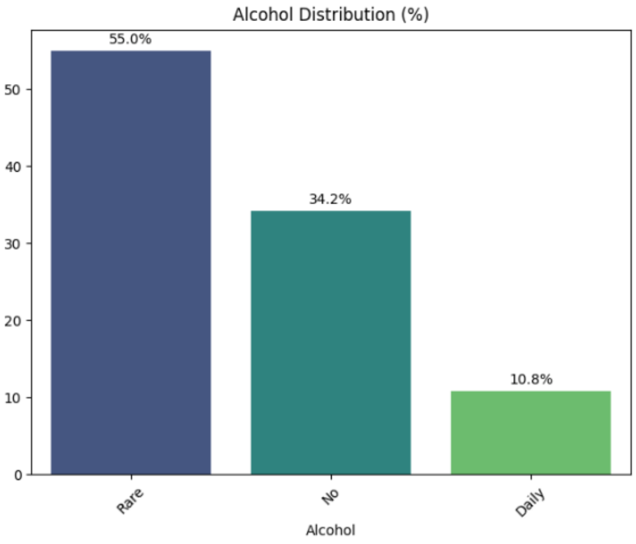
Insight: The largest group is "never smoked" (37%), followed by "Unknown" (30.2%), "formerly smoked" (17.3%), and "smokes" (15.5%).

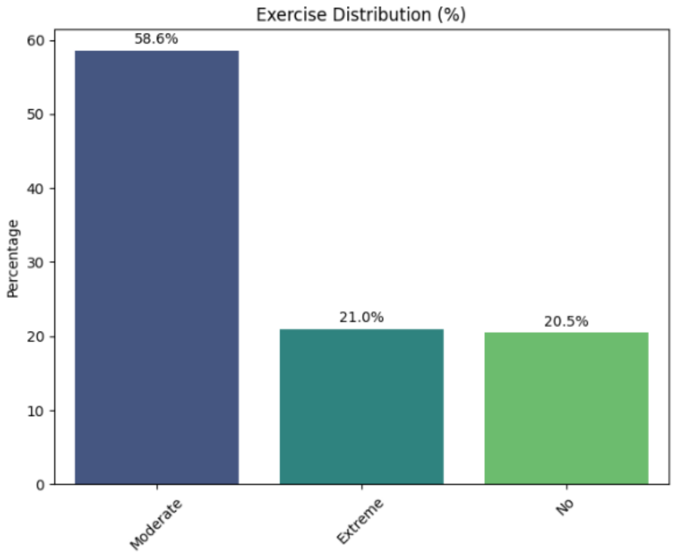
Interpretation: A substantial portion of the population has never smoked, but about one-third either currently or formerly smoked, which could have implications for health outcomes.

Covered by Any Other Company Distribution

Insight: The majority (69.7%) are not covered by another company, while 30.3% are covered.

Interpretation: This suggests that most individuals do not have additional insurance or coverage, which could impact their healthcare options.





Alcohol Consumption Distribution

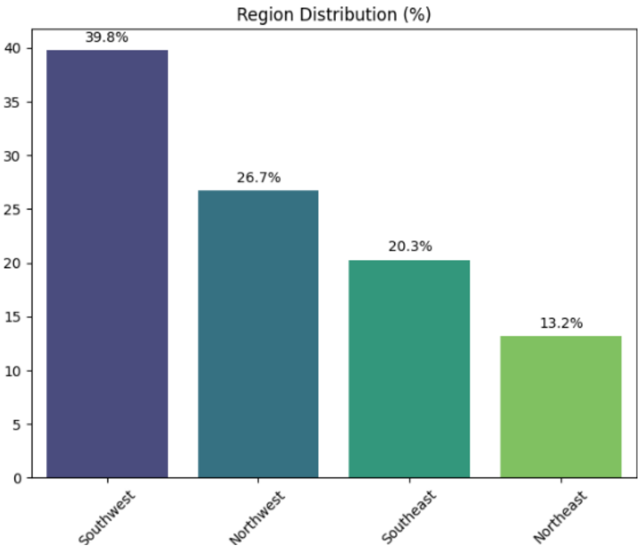
Insight: A significant portion of individuals rarely consume alcohol (55%), with 34.2% abstaining and 10.8% drinking daily.

Interpretation: Alcohol consumption is low overall, though a small group drinks regularly, which may affect health risks.

Exercise Distribution

Insight: Most people engage in moderate exercise (58.6%), while fewer people are extreme exercisers (21%) or do not exercise (20.5%).

Interpretation: Exercise levels are generally good, but there is room for improvement among the non-exercising group.



Region Distribution

Insight: The largest proportion of individuals are from the "Southwest" (39.8%), followed by "Southeast" (26.7%) and smaller percentages from other regions.

Interpretation: The data is skewed towards certain regions, particularly the Southwest, which could influence regional analysis or lifestyle factors.

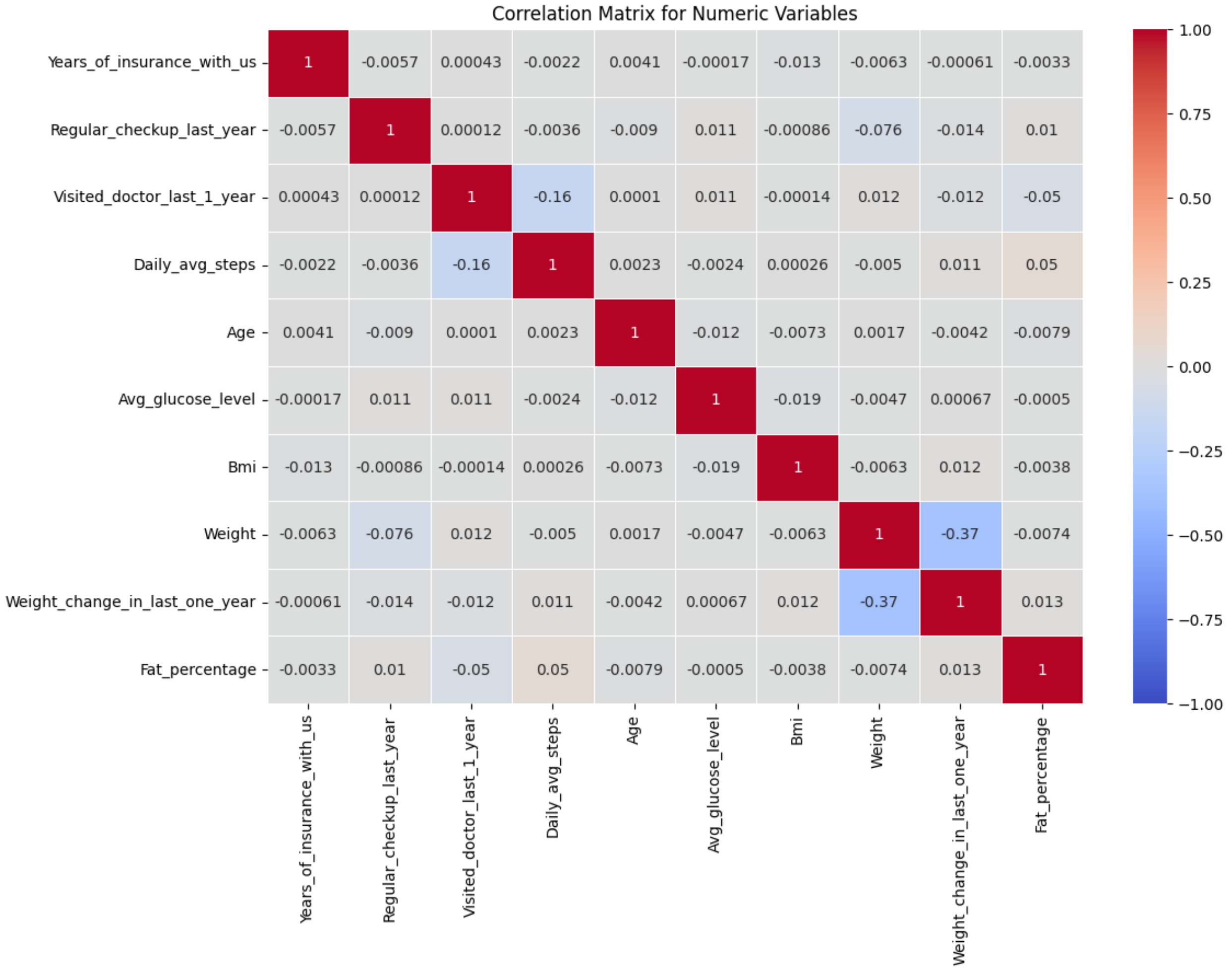
In summary, the data suggests a population with low participation in adventure sports, generally moderate cholesterol levels, and a significant portion engaged in moderate exercise. The sample is predominantly male, with most individuals having no history of major diseases and a relatively healthy lifestyle regarding smoking and alcohol consumption. However, the gender imbalance and smoking history may warrant deeper analysis for targeted health interventions.

Bivariate Analysis & Multivariate Analysis

Correlation matrix and numeric variables & their key Observations:

Weight and Weight Change (Last Year): The correlation between Weight and Weight change in last one year is -0.37. This is a moderate negative correlation, if individuals with higher weights are experiencing lower weight changes (i.e., less fluctuation), it could suggest that heavier individuals might be more stable in their weight compared to those who are lighter. Conversely, lighter individuals might experience more significant changes, possibly due to factors like dieting, lifestyle changes, or metabolic differences.

BMI and Weight: The correlation of -0.0063 between BMI and weight indicates a very weak or almost nonexistent linear relationship between the two variables. This suggests that in this dataset, changes in weight are not strongly associated with changes in BMI.

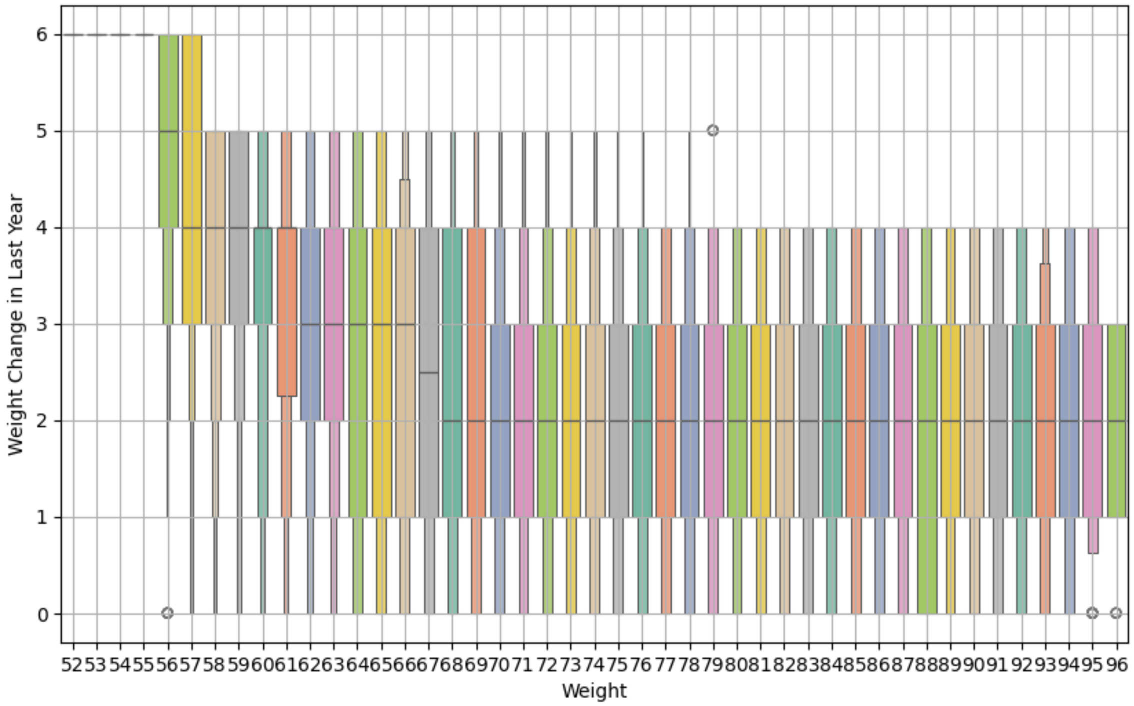


<u>Visited Doctor in the Last Year and Daily Average Steps</u>: The correlation between Visited doctor last 1 year and Daily avg steps is -0.16, which suggests a weak negative correlation. The weak negative correlation between suggests that individuals with health concerns who visit doctors may need encouragement to increase their physical activity levels.

<u>Avg. Glucose Level and Other Variables</u>: The lack of correlation between Avg glucose level and other variables suggests that glucose levels might need to be monitored independently of other factors.

<u>Fat Percentage, Years of Insurance and Other Variables</u>: fat percentage & Years of insurance with us does not seem to have a strong correlation with most variables. It would be beneficial to look deeper into why fat percentage isn't more closely linked with other health metrics like BMI and weight in this dataset.

Weight Vs Weight change in last one year

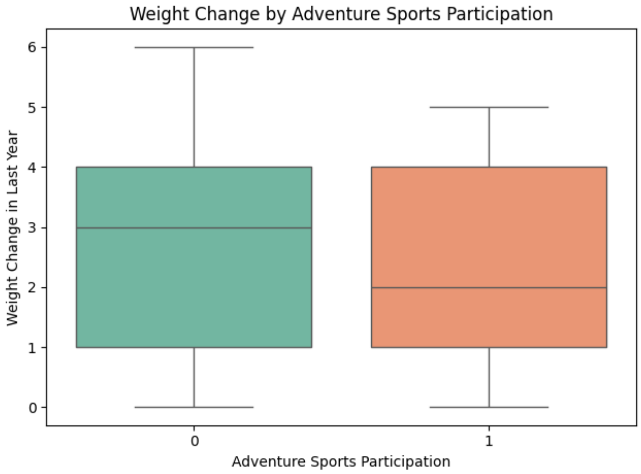


Key Insights: Higher Weight Change for Lower Weights: Individuals with lower body weight exhibit greater weight changes in the last year, with some reaching 5-6 units of change. This suggests that people with lower body weight tend to experience fluctuations in their weight over time.

Stabilization of Weight Change in Higher Weight Ranges:

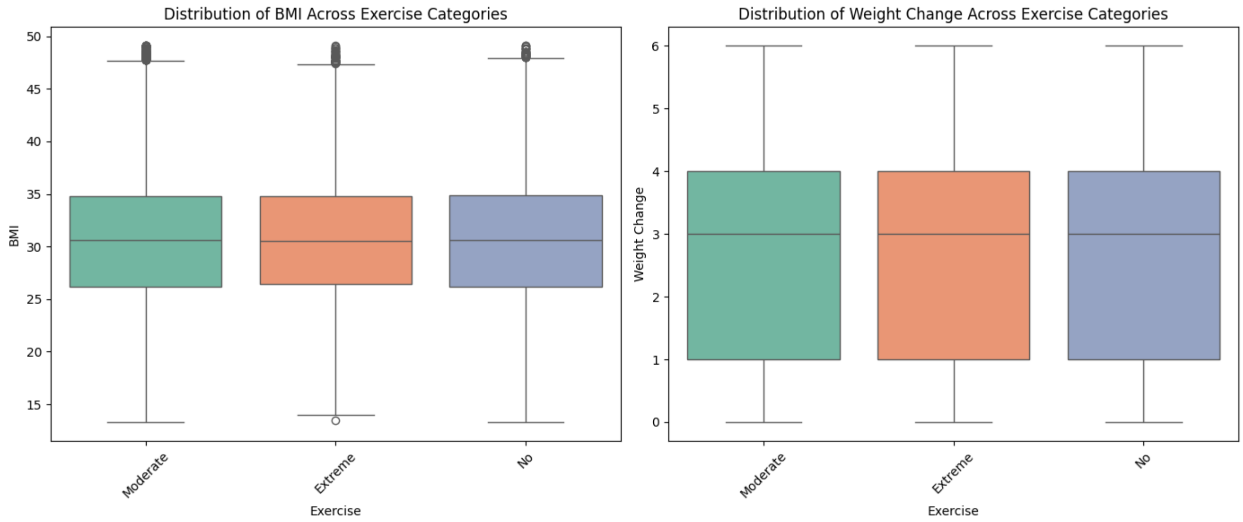
As we move into higher weight categories, the weight change becomes much more consistent. In fact, most individuals in this weight range experience only around 1-3 units of change. This indicates that heavier individuals are more likely to have stable weight with smaller fluctuations year-over-year.

Relationship Between Weight Change & Adventure Sports Participation



Insights: Adventure sports participants tend to experience lower and more consistent weight changes than non-participants. This suggests that engaging in adventure sports may help stabilize weight changes, possibly due to increased physical activity. The variation in weight change for non-participants could be influenced by other factors like lifestyle habits or health conditions.

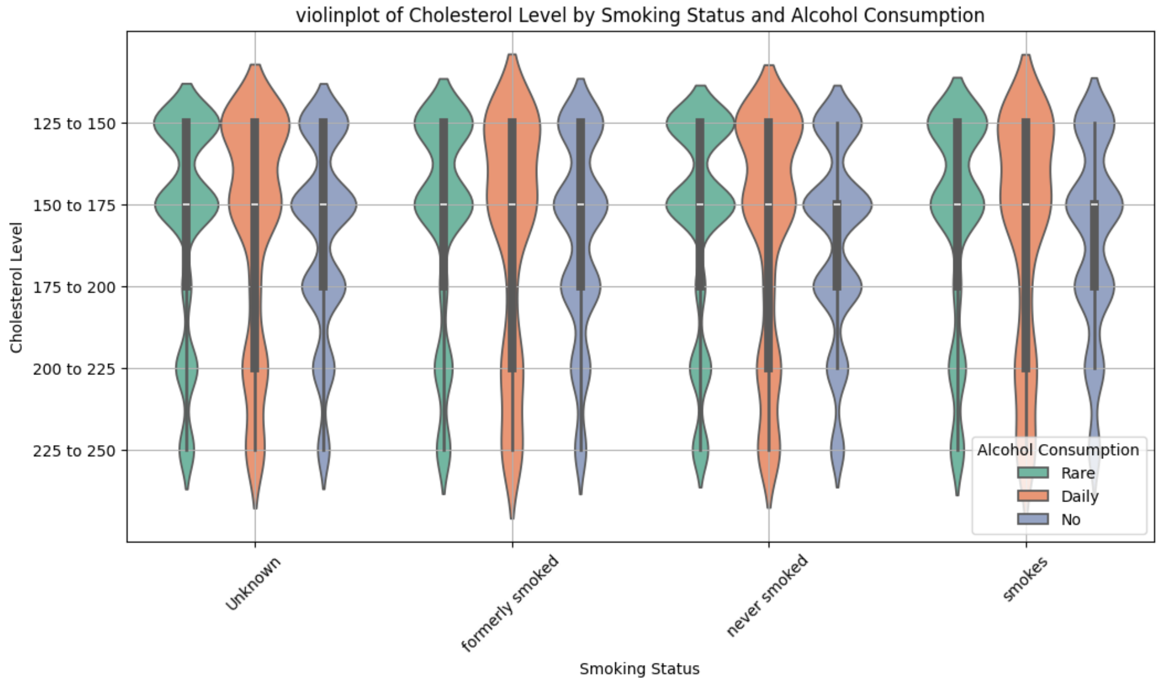
Distribution of BMI and Weight change across Exercise



<u>Exercise Type has No Effect on BMI or Weight Change</u>

The fact that the distributions of BMI and Weight Change are exactly the same for all three exercise types could indicate that exercise type is not a significant factor influencing either BMI or weight change within this dataset. This implies that the level of physical activity (whether moderate, extreme, or none) might not be strongly correlated with changes in body weight or body mass index for the individuals in the dataset.

Cholesterol Levels by Smoking and Alcohol Status



Cholesterol Distribution across Smoking Categories:

The shapes of the violins are similar across all smoking statuses. This suggests that smoking status alone may not be a dominant factor influencing cholesterol levels in this dataset.

There are slightly wider distributions "never smoked" and "formerly smoked" categories, which could imply more variability in cholesterol levels for individuals who have quit or never smoked.

Alcohol Consumption and Cholesterol Levels:

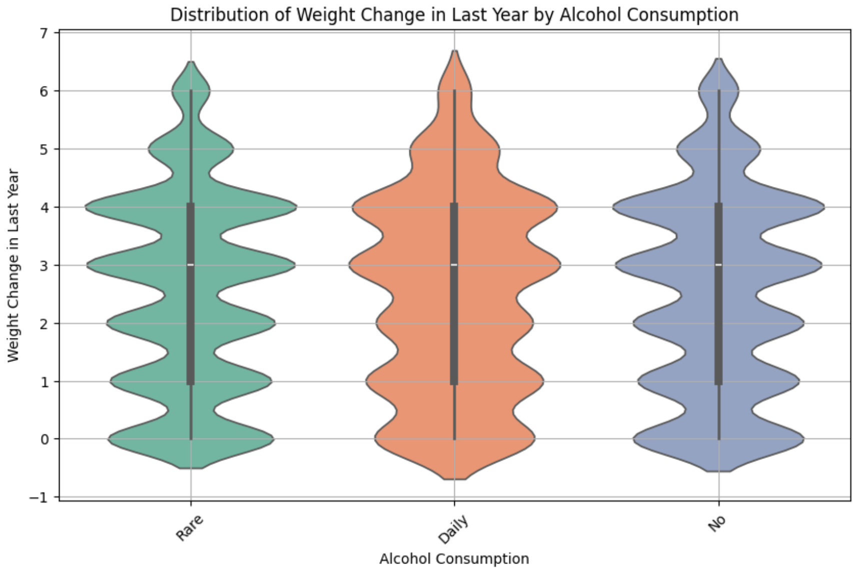
Rare alcohol consumers (green violins) and non-drinkers (blue violins) have similar distributions but with less spread in the higher cholesterol ranges.

Daily alcohol drinkers exhibit slightly higher cholesterol levels compared to rare or non-drinkers, as shown by the thickened areas around the 200 mg/dL mark.

Stable Cholesterol Levels for Non-Smokers with Rare or No Alcohol:

For individuals who never smoked and consume rare or no alcohol, cholesterol levels seem more concentrated in the 150-175 mg/dL range, indicating lower cholesterol levels in this group.

Weight Change in Last Year by Alcohol Consumption



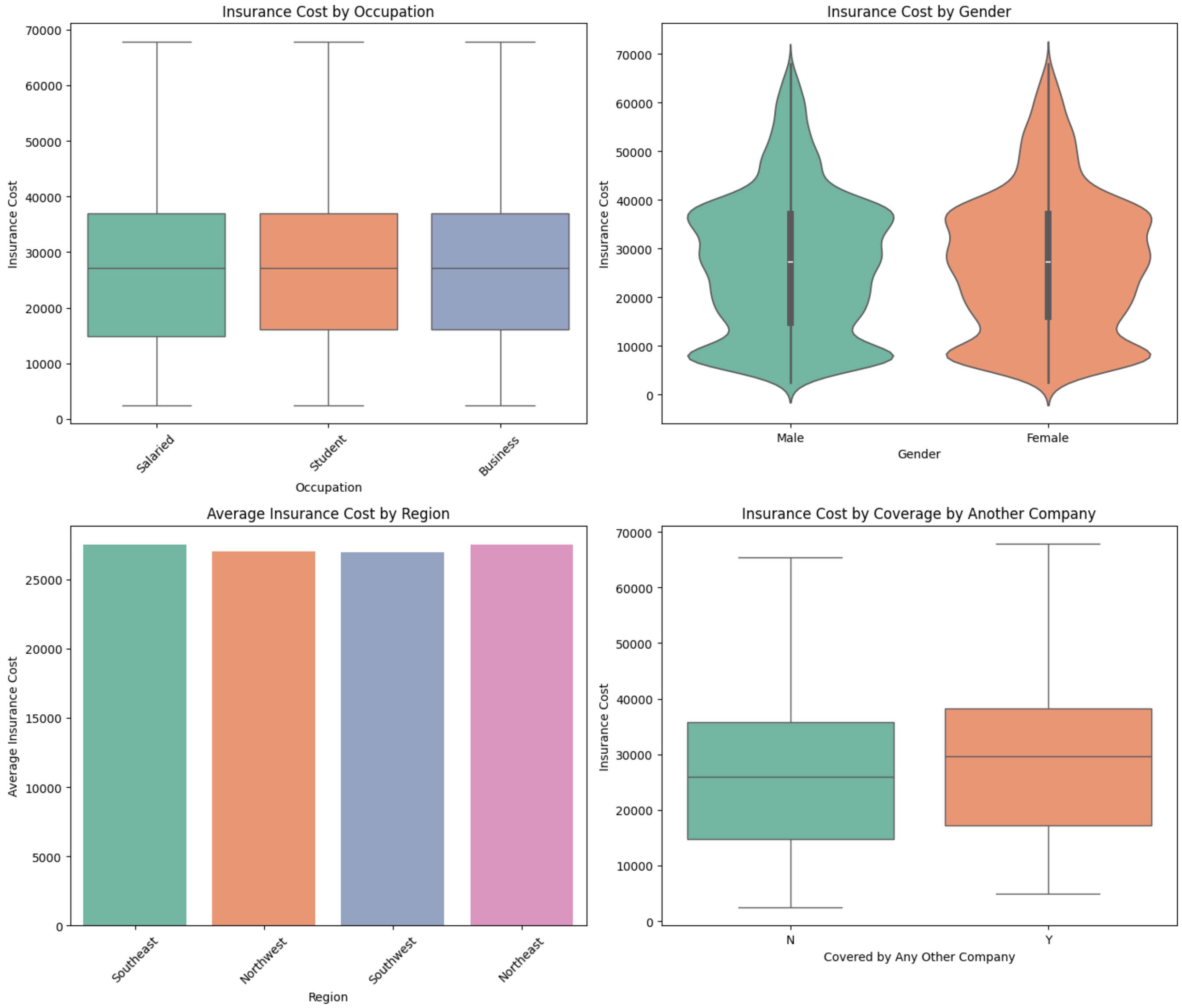
Higher Maximum Weight Gain in Daily Drinkers:

Individuals in the daily alcohol consumption group exhibit a higher upper range of weight change, with some reaching up to +6 units, while the other groups tend to stay just below this level.

This could suggest that daily alcohol consumption might correlate with slightly higher overall weight gain over the year, although the majority still centre around similar weight change values.

These insights suggest that while alcohol consumption may play a role in weight change, the overall patterns of weight gain are similar across different drinking behaviours.

To compare the Insurance cost (target variable) with Occupation, Gender, Region, and Covered by any other company



Insurance Cost by Occupation:

Salaried, Student, and Business occupations have similar distributions of insurance costs.

Median costs are fairly close for all three groups, no clear differences, indicating that Occupation does not seem to play a major role in determining average insurance costs.

Insurance Cost by Gender:

Male and Female genders have almost identical distributions of insurance costs.

There doesn’t seem to be a significant difference in insurance cost between genders.

Average Insurance Cost by Region:

The average insurance cost is nearly the same across all regions (Southeast, Northwest, Southwest, and Northeast).

No clear regional differences, indicating that region does not seem to play a major role in determining average insurance costs.

<u>Insurance Cost by Coverage by Another Company:</u>

Whether someone is covered by another company or not ("N" for No, "Y" for Yes), there is little difference in insurance costs. The plot suggests that individuals who are covered by another company tend to have a slightly higher insurance cost and more variability compared to those without additional coverage. However, the overall range is similar for both groups.

Clustering: K-Means Clustering & Steps to Perform

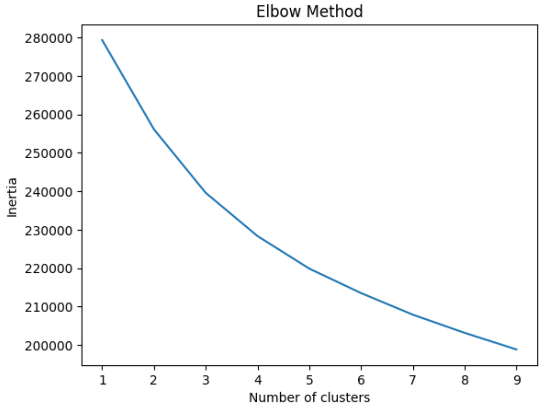
Recursive Feature Elimination (RFE) with Random Forest Regressor

Below are the selected features by applying Recursive Feature Elimination (RFE) with RandomForest.

Selected Features: 'Years of insurance with us', 'Regular checkup last year', 'Visited doctor last 1 year', 'Daily avg steps', 'Age', 'Avg glucose level', 'Bmi', 'Weight', 'Weight change in last one year', 'Fat percentage', 'Smoking status never smoked', 'Covered by any other company Y', 'Alcohol Rare', 'Exercise Moderate', 'Region Southwest'

Elbow Curve: K-Means Clustering with cluster = 3

A silhouette score of around 0.0848 suggests that the clustering might not be very effective. Silhouette scores range from -1 to 1, where values around 0 suggest overlapping clusters, and negative values indicate that data points might be assigned to the wrong clusters.



Customer Profiling

Cluster	Years of insurance with us	Regular checkup last year	Visited doctor last 1 year	Daily avg steps	Age	Avg glucose level	Bmi	Weight	Weight change in last one year	Fat percentage	Insurance cost	Cholesterol level num	freq
0	4.09	0.01	3.13	5080.5	45.09	166.18	30.72	76.86	1.67	28.63	35165	1.25	11740
1	4.1	0.12	3	5175.1	45.02	168.45	30.93	62.98	4.14	29.02	14434	1.28	7385
2	4.07	1.37	3.09	5098.1	44.45	169.06	30.74	71.97	2.17	28.93	27108	1.27	5875

Business Insights and Recommendations using clustering:

For Cluster 0 (Largest, High Costs, Low Engagement in Preventive Care):

Focus on Preventive Health Programs: This cluster shows almost no engagement in regular checkups, yet has the highest insurance costs. Encourage preventive care (e.g., annual checkups) and lifestyle changes to reduce long-term healthcare costs.

Target Chronic Disease Management: Given the elevated BMI, glucose levels, and minimal weight change, this group may benefit from interventions aimed at managing chronic conditions like diabetes and obesity.

Activity Promotion: Incentivize physical activity programs since their daily step count is average, but improvements here can lead to long-term health benefits.

For Cluster 1 (Lower Weight, Moderate Checkup Engagement, Low Costs):

Support for Sustaining Healthy Habits: This group has the lowest insurance costs and relatively lower weight, but higher weight change. Continued support for weight maintenance programs and preventive care can help them maintain their lower risk profile.

Monitor Weight Fluctuations: Given the higher weight change, offer nutrition and lifestyle counselling to prevent weight gain or other potential health risks

For Cluster 2 (Moderate Risk, High Checkup Engagement, Moderate Costs):

Leverage Health Engagement: This cluster engages more with regular checkups, so they are likely receptive to health programs. You can offer personalized wellness programs, such as glucose control or weight management.

Address Rising Health Metrics: Despite regular checkups, this group has high glucose levels and BMI. Continued emphasis on nutrition and activity programs could help prevent or manage chronic diseases.

Business Insights from EDA

Weight & Weight Change in Last Year: Customers who experienced significant weight changes over the past year could indicate potential health risks, especially for those gaining or losing weight drastically. This group may require closer health monitoring, and the insurance company should consider adjusting premiums or offering wellness programs that encourage maintaining a healthy weight.

Regular Health Check-Ups: Customers who frequently undergo regular health check-ups tend to file fewer claims. This emphasizes the need to promote preventive health care by incentivizing regular screenings with premium discounts or additional benefits for policyholders who maintain regular health evaluations.

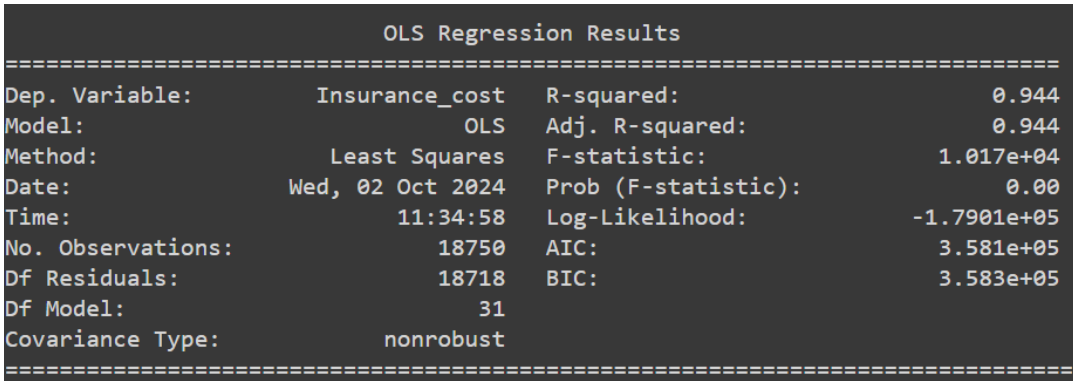
Adventure Sports: A segment of policyholders participating in adventure sports poses higher risk levels. Introducing tailored policies with higher premiums for those engaging in these activities or offering specialized adventure sports coverage can address this risk and cater to their insurance needs.

Covered by Another Company: A significant number of policyholders are also covered by other insurance providers. To retain them and reduce the risk of losing customers, offer competitive pricing, bundled services, or exclusive benefits, particularly for those considering switching providers or consolidating their insurance.

Model building and interpretation

Linear Regression using statsmodel (OLS)

An Ordinary Least Squares (OLS) regression model is a linear regression method used to estimate the relationship between one or more independent variables and a dependent variable. The model minimizes the sum of the squared differences in residuals, providing the best linear fit to the data.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:         Insurance_cost   R-squared:                       0.944
Model:                            OLS   Adj. R-squared:                  0.944
Method:                 Least Squares   F-statistic:                 1.017e+04
Date:                Wed, 02 Oct 2024   Prob (F-statistic):               0.00
Time:                        11:34:58   Log-Likelihood:            -1.7901e+05
No. Observations:               18750   AIC:                         3.581e+05
Df Residuals:                   18718   BIC:                         3.583e+05
Df Model:                          31
Covariance Type:            nonrobust
==============================================================================
```

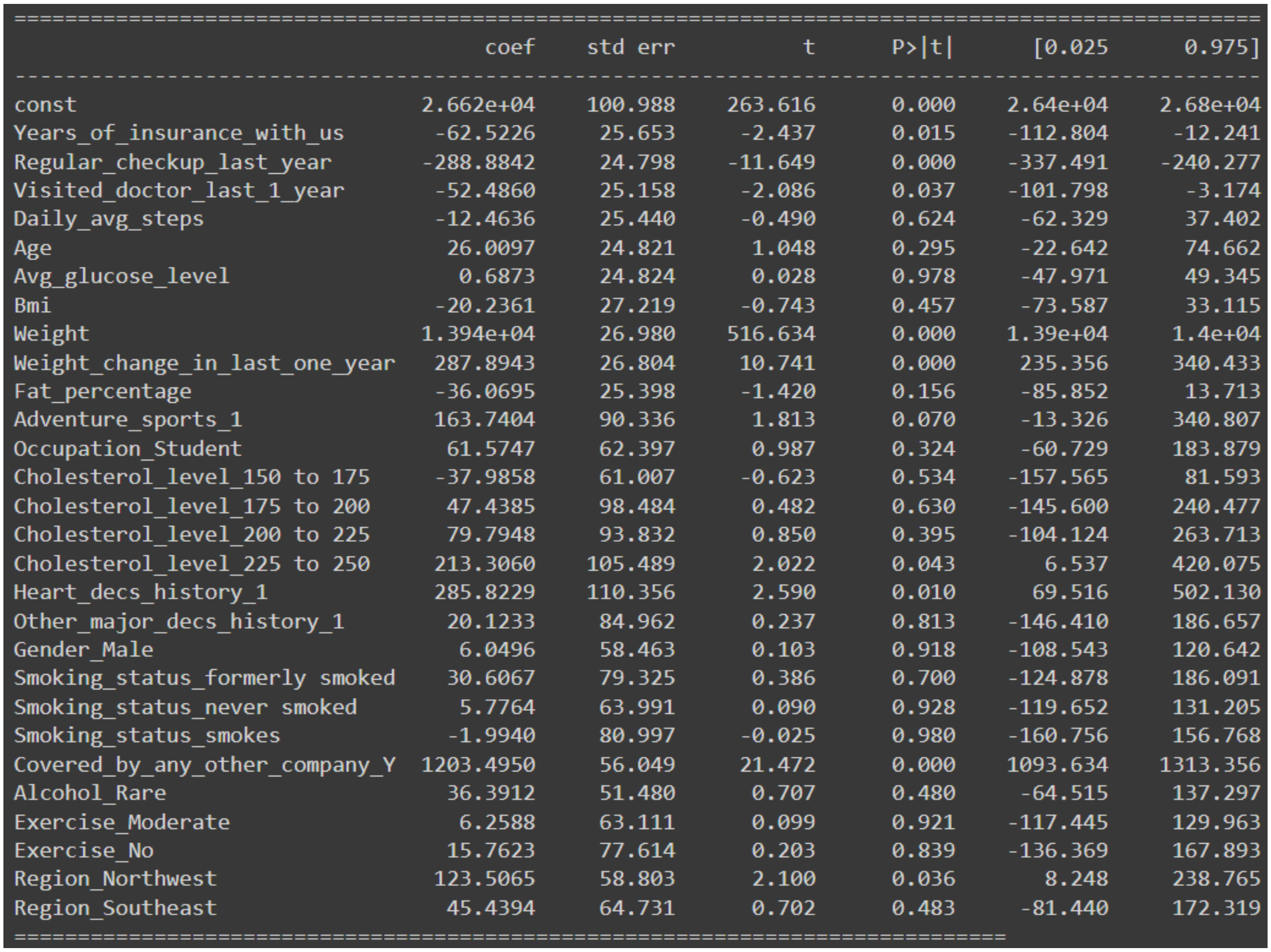
Model Tuning: Variance Inflation Factor (VIF) analysis

We conducted a Variance Inflation Factor (VIF) analysis to check for multicollinearity in the Ordinary Least Squares (OLS) regression model. Multicollinearity occurs when independent variables are highly correlated making it difficult to interpret the significance of each predictor.

To address this issue, we removed variables with high VIF scores:

- Occupation Salaried' (VIF = 4.54)
- Alcohol No (VIF = 2.83)
- Region Southwest (VIF = 2.40)

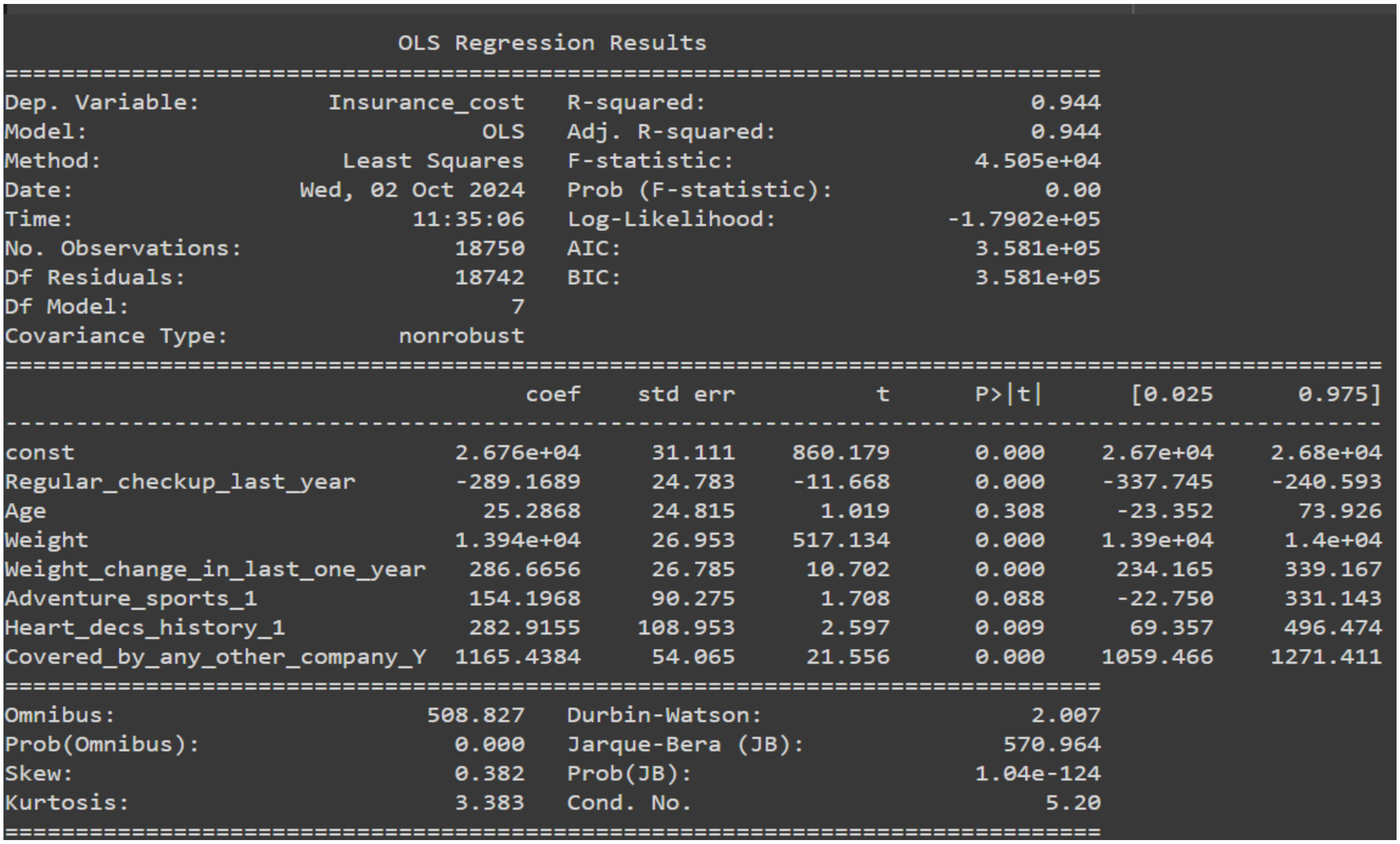
By removing these variables sequentially, we reduce multicollinearity, resulting in a model that is more robust and easier to interpret. This step also improves the precision of the coefficient estimates, ensuring that each variable's contribution is accurately reflected.



```
==================================================================================================
                                  coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                        2.662e+04    100.988    263.616      0.000    2.64e+04    2.68e+04
Years_of_insurance_with_us    -62.5226     25.653     -2.437      0.015    -112.804     -12.241
Regular_checkup_last_year    -288.8842     24.798    -11.649      0.000    -337.491    -240.277
Visited_doctor_last_1_year    -52.4860     25.158     -2.086      0.037    -101.798      -3.174
Daily_avg_steps               -12.4636     25.440     -0.490      0.624     -62.329      37.402
Age                            26.0097     24.821      1.048      0.295     -22.642      74.662
Avg_glucose_level               0.6873     24.824      0.028      0.978     -47.971      49.345
Bmi                           -20.2361     27.219     -0.743      0.457     -73.587      33.115
Weight                       1.394e+04     26.980    516.634      0.000    1.39e+04     1.4e+04
Weight_change_in_last_one_year 287.8943    26.804     10.741      0.000     235.356     340.433
Fat_percentage                -36.0695     25.398     -1.420      0.156     -85.852      13.713
Adventure_sports_1            163.7404     90.336      1.813      0.070     -13.326     340.807
Occupation_Student             61.5747     62.397      0.987      0.324     -60.729     183.879
Cholesterol_level_150 to 175  -37.9858     61.007     -0.623      0.534    -157.565      81.593
Cholesterol_level_175 to 200   47.4385     98.484      0.482      0.630    -145.600     240.477
Cholesterol_level_200 to 225   79.7948     93.832      0.850      0.395    -104.124     263.713
Cholesterol_level_225 to 250  213.3060    105.489      2.022      0.043       6.537     420.075
Heart_decs_history_1          285.8229    110.356      2.590      0.010      69.516     502.130
Other_major_decs_history_1     20.1233     84.962      0.237      0.813    -146.410     186.657
Gender_Male                     6.0496     58.463      0.103      0.918    -108.543     120.642
Smoking_status_formerly smoked 30.6067     79.325      0.386      0.700    -124.878     186.091
Smoking_status_never smoked     5.7764     63.991      0.090      0.928    -119.652     131.205
Smoking_status_smokes          -1.9940     80.997     -0.025      0.980    -160.756     156.768
Covered_by_any_other_company_Y 1203.4950   56.049     21.472      0.000    1093.634    1313.356
Alcohol_Rare                   36.3912     51.480      0.707      0.480     -64.515     137.297
Exercise_Moderate               6.2588     63.111      0.099      0.921    -117.445     129.963
Exercise_No                    15.7623     77.614      0.203      0.839    -136.369     167.893
Region_Northwest              123.5065     58.803      2.100      0.036       8.248     238.765
Region_Southeast               45.4394     64.731      0.702      0.483     -81.440     172.319
==============================================================================
```

Removing the non-significant predictor variables

After reviewing the regression output, it's clear that many variables are statistically insignificant (p-value > 0.05), meaning they do not significantly contribute to explaining the variation in the dependent variable (Insurance Cost). Removing them would streamline the model without losing predictive power, as they are not contributing significantly to explaining insurance cost variations.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:         Insurance_cost   R-squared:                       0.944
Model:                            OLS   Adj. R-squared:                  0.944
Method:                 Least Squares   F-statistic:                 4.505e+04
Date:                Wed, 02 Oct 2024   Prob (F-statistic):               0.00
Time:                        11:35:06   Log-Likelihood:            -1.7902e+05
No. Observations:               18750   AIC:                         3.581e+05
Df Residuals:                   18742   BIC:                         3.581e+05
Df Model:                           7                                         
Covariance Type:            nonrobust                                         
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
const                           2.676e+04     31.111    860.179      0.000    2.67e+04    2.68e+04
Regular_checkup_last_year       -289.1689     24.783    -11.668      0.000    -337.745    -240.593
Age                               25.2868     24.815      1.019      0.308     -23.352      73.926
Weight                          1.394e+04     26.953    517.134      0.000    1.39e+04     1.4e+04
Weight_change_in_last_one_year   286.6656     26.785     10.702      0.000     234.165     339.167
Adventure_sports_1               154.1968     90.275      1.708      0.088     -22.750     331.143
Heart_decs_history_1             282.9155    108.953      2.597      0.009      69.357     496.474
Covered_by_any_other_company_Y  1165.4384     54.065     21.556      0.000    1059.466    1271.411
==============================================================================
Omnibus:                      508.827   Durbin-Watson:                   2.007
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              570.964
Skew:                           0.382   Prob(JB):                    1.04e-124
Kurtosis:                       3.383   Cond. No.                         5.20
==============================================================================
```

Interpretation of Model & It's Feature Importance

In this OLS regression, variables with p-values less than 0.05 are considered statistically significant. Here are the significant variables and their interpretation:

- **Regular checkup last year (p < 0.000):** Individuals who had a regular checkup in the last year tend to have lower insurance costs by $289.17 on average, compared to those who did not. This could be due to early detection of health issues or better health management.

- **Weight (p < 0.000):** Higher weight is associated with a significant increase in insurance costs. For every unit increase in weight, insurance costs rise by $13,940. This suggests that weight is a major factor in determining insurance costs, likely due to higher health risks.

- **Weight change in last one year (p < 0.000):** A positive weight change in the last year leads to a higher insurance cost. For every unit of weight gain, insurance costs increase by $286.67, indicating a potential risk factor related to sudden weight fluctuations.

- **Heart disease history 1 (p = 0.009):** Individuals with a history of heart disease experience a higher insurance cost by $282.92 on average. This aligns with the higher health risks associated with heart conditions.

- **Covered by any other company Y (p < 0.000):** Being covered by another insurance company is associated with an increase in insurance cost by $1,165.44. This may reflect higher overall premiums or coverage costs.

Evaluation metrics for OLS Regression

- **Root Mean Squared Error (RMSE)**: The RMSE values are $3,390.50 for the training data and $3,412.59 for the test data. This small difference indicates the model effectively fits both the training and test data, with an average prediction deviation of around $3,400. Such low errors across datasets confirm that the model captures key patterns in the data without introducing significant error when predicting on unseen cases.

- **Mean Absolute Percentage Error (MAPE)**: The model's MAPE is 15.48% on the training data and 15.02% on the test data. The close match between these values suggests consistent predictive accuracy, with the model's predictions deviating by approximately 15% from actual insurance costs on average. The slightly lower MAPE on the test set suggests stable generalization and that the model does not overfit to the training data.

- **R-squared**: The model achieves an R-squared of 0.9439 on the training data and 0.9434 on the test data, meaning it explains over 94% of the variance in both datasets. This high R-squared demonstrates that the model captures almost all the variability in insurance costs, making it highly effective in modeling the relationship between the predictors and the target variable. The minimal difference between train and test R-squared further affirms the model's robustness.

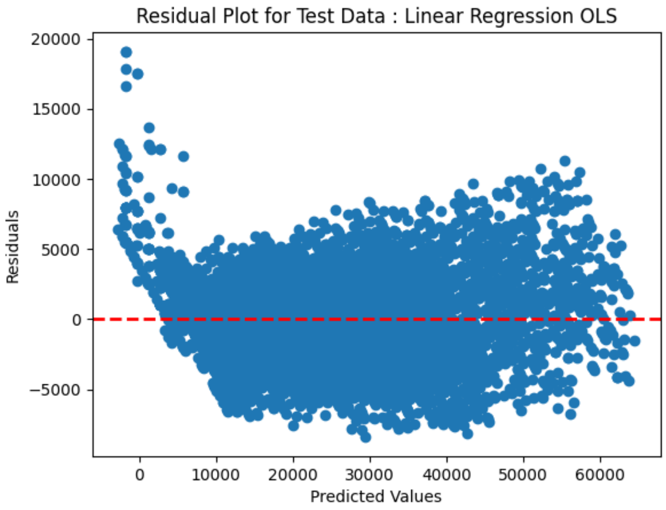
- **Adjusted R-squared**: The Adjusted R-squared is 0.9439 for training and 0.9433 for test data, very close to the R-squared values. This similarity indicates that the model's complexity aligns well with the number of predictors used, enhancing model reliability without unnecessarily complicating the model. A high Adjusted R-squared ensures that the included features contribute meaningfully to the predictive power without introducing noise or overfitting.

Key Takeaways:

- Low RMSE on both train and test sets indicates the model's predictions are quite accurate, with minimal prediction errors.
- Consistent MAPE around 15% shows the model's accuracy remains stable across different datasets.
- High R-squared values (above 0.94) on both train and test sets suggest the model captures almost all of the variability in insurance costs.
- The model provides reliable predictions, with minimal performance degradation when applied to test data, making it suitable for deployment.

Residual Plot Analysis for OLS Model

- The residual plot for the OLS regression shows that the residuals are predominantly centered around zero, indicating that the model's predictions are unbiased.
- However, there is an increasing spread of residuals as the predicted values rise, suggesting larger errors for higher predicted values.
- This pattern implies a potential non-linear relationship that the linear model does not adequately capture.
- Overall, the plot highlights areas where the model's performance may be inconsistent, especially with larger predicted values.



Ridge & Lasso Regression

Ridge and Lasso regression are regularization techniques used to enhance the performance of linear models by addressing multicollinearity and preventing overfitting. Ridge regression applies an L2 penalty, shrinking coefficients to reduce model complexity, while Lasso regression uses an L1 penalty, which can shrink some coefficients to zero, effectively performing variable selection. Both methods improve model stability and generalization on unseen data.

Hyperparameter Tuning for Ridge Regression:

Ridge regression relies on tuning the alpha parameter, which controls the L2 regularization strength. However, unlike Lasso, Ridge shrinks coefficients without setting them exactly to zero. The goal of tuning alpha in Ridge is to balance model complexity and performance by finding the value that minimizes overfitting. The grid search tests a wide range of alpha values from 0.0001 to 10,000, using cross-validation to find the alpha that leads to the lowest mean absolute percentage error (MAPE).

Ridge Regression coefficient Interpretation.

Positive & Negative Coefficients (Top Key Drivers of Insurance Costs):

- **Weight ($13,629.01):** This variable has the largest positive coefficient. For every unit increase in weight, insurance costs rise by approximately $13,629.01. This indicates that weight is a critical factor, likely due to its association with various health risks.

- **Covered by Any Other Company ($1,146.42):** Being covered by another insurance company increases the insurance cost by $1,146.42, indicating that having multiple policies may suggest higher risk factors, thus driving up costs.

- **Weight Change in Last Year ($278.14):** A positive change in weight over the last year also leads to an increase in insurance costs.

- **Regular Checkup Last Year (-$278.14):** Regular medical checkups reduce insurance costs by $278.14, likely because preventive healthcare decreases the likelihood of future medical claims.

- **Years of Insurance (-$71.94):** A longer insurance tenure reduces costs, reflecting loyalty discounts or lower risk associated with extended coverage history.

Evaluation Metrics for Ridge Regression

- **Root Mean Squared Error (RMSE)**: The train RMSE is **3383.33**, indicating that, on average, the model's predictions are 3383.33 units away from actual values on the training data. The test RMSE of **3422.18** is slightly higher, suggesting good generalization without significant overfitting.

- **Mean Absolute Percentage Error (MAPE)**: The train MAPE is **15.44%**, showing that the model's predictions in the training set are, on average, 15.44% off from actual insurance costs. The test MAPE of **15.18%** indicates comparable performance across both datasets, reflecting reliability in predictions.

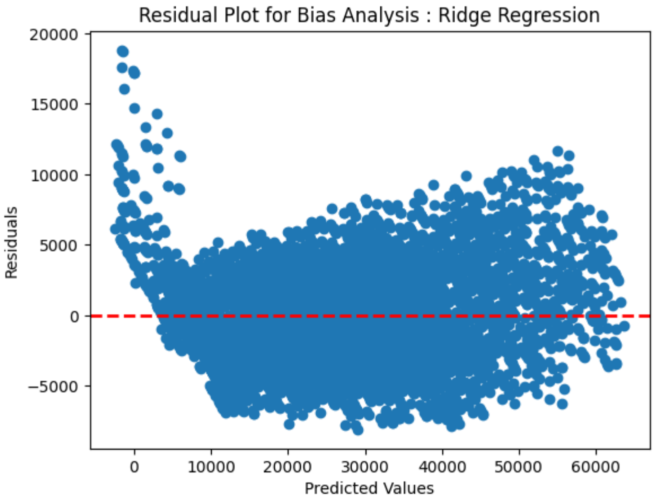
- **R-squared**: The train R-squared value is **0.9442**, meaning approximately 94.42% of the variance in insurance costs is explained by the model on the training data. The test R-squared of **0.9429** indicates that about 94.29% of the variance is explained on the test set, highlighting a strong fit.

- **Adjusted R-squared**: The adjusted R-squared for the training set is **0.9441**, which closely aligns with the regular R-squared. This suggests that the included predictors effectively contribute to the model without introducing unnecessary complexity, indicating robustness in the model's performance.

Key Insights:

- Consistent Performance: The metrics for both the train and test sets are closely aligned, indicating that the Ridge model generalizes well and is not overfitting.
- Strong Fit: With R-squared values above 0.94 for both the train and test sets, the model captures over 94% of the variance in insurance costs.
- MAPE Comparison: Both the train and test MAPE values are around 15%, meaning the model maintains consistent predictive accuracy across datasets. This suggests that even with new, unseen data, the model will perform reliably.
- Low RMSE: The low RMSE values (around 3400) suggest that, while there is some error in the predictions, the model's overall error margin is relatively small given the complexity of the data and the range of potential insurance costs.

Residual Polt for Ridge Regression



- The residual plot for Ridge regression indicates that the residuals are largely distributed around the zero line, suggesting an overall unbiased prediction.
- However, similar to the OLS regression plot, there is a noticeable increase in the spread of residuals as predicted values rise, which may indicate that the model struggles with larger predicted values.
- This suggests that while Ridge regression helps in reducing some issues related to bias, it may still not capture the complexity of the data, particularly for higher values.
- The residuals do not show a clear pattern, implying that the model's assumptions about linearity and homoscedasticity are somewhat met, but there are still areas for improvement

Hyperparameter Tuning for Lasso Regression:

In Lasso regression, the primary hyperparameter to tune is the regularization strength, controlled by the alpha parameter. By testing a range of alpha values, we can identify the optimal amount of regularization to apply. A smaller alpha value results in minimal regularization, allowing the model to retain more complexity, while a larger alpha enforces stronger regularization, which can shrink coefficients to zero and perform feature selection. In this example, a grid search is conducted over a range of alpha values from 0.0001 to 10,000 using cross-validation. This helps determine the alpha that minimizes the mean absolute percentage error (MAPE) and provides a well-generalized model.

Lasso Regression Coefficients Interpretation

Positive Coefficients (Key Drivers of Insurance Costs):

- **Weight (13,798.88):** The variable Weight has the largest positive coefficient. For every unit increase in weight, insurance costs are expected to increase by approximately $13,798.88. This suggests that higher weight is associated with greater health risks, leading to increased insurance premiums.

- **Cholesterol Level (724.22):** The Cholesterol Level variable contributes a positive amount of $724.22 to insurance costs. This indicates that higher cholesterol levels are linked to increased health risks and, therefore, higher insurance costs.

Zero Coefficients (Excluded Features):

- Many features have coefficients of 0, indicating that they do not contribute significantly to predicting insurance costs. These features include Years of insurance with us, Daily avg steps, Age & Various medical history indicators.

Overall Takeaways:

- Key Drivers of Costs: The most impactful factors influencing insurance costs in this model are Weight and Cholesterol Level, suggesting a strong link between these health indicators and insurance premiums.
- Feature Reduction: The Lasso model effectively reduces the number of features by eliminating those that do not significantly impact the outcome, which can simplify the model and improve interpretability.
- Health Implications: The results imply that maintaining a healthy weight and managing cholesterol levels could be critical not only for personal health but also for minimizing insurance costs.

Evaluation Metrics for Lasso Regression

- **Root Mean Squared Error (RMSE)**: The train RMSE is **3398.27**, indicating that, on average, the model's predictions are 3398.27 units away from actual values on the training data. The test RMSE of **3435.58** is slightly higher, suggesting consistent performance across both datasets without significant overfitting.

- **Mean Absolute Percentage Error (MAPE)**: The train MAPE is **15.42%**, indicating that the model's predictions in the training set are about 15.42% off from actual insurance costs. The test MAPE of **15.12%** is comparable, demonstrating the model's reliability and generalization across datasets.

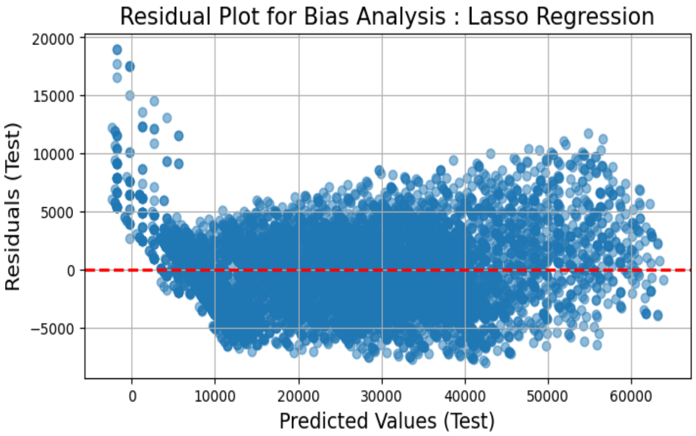
- **R-squared**: The train R-squared value is **0.9437**, which means that approximately 94.37% of the variance in insurance costs is explained by the model on the training data. The test R-squared of **0.9425** indicates that about 94.25% of the variance is explained on the test set, reflecting a strong model fit.

- **Adjusted R-squared**: The adjusted R-squared values are very close, with the train value at **0.9436** and the test value at **0.9422**. This suggests that the model captures significant relationships without overfitting, indicating that the included predictors are effectively contributing to the overall prediction accuracy.

Summary of Insights:

- Consistency: The model performs similarly on both the train and test datasets, indicating that it generalizes well and avoids overfitting.
- Good Fit: With high R-squared values (94%+), the model explains a significant portion of the variance in insurance costs.
- Acceptable Error: The RMSE and MAPE values suggest that the model has relatively low prediction errors, making it suitable for practical use in predicting insurance costs with reasonable accuracy.

Residual Plot for Lasso Regression



- There seems to be a concentration of residuals close to zero, indicating that the model generally performs well.
- A fan-shaped pattern can be observed, with higher residuals at smaller predicted values, which may suggest some heteroscedasticity—where variance increases with predicted values.
- The residuals are fairly symmetrically distributed around the red line but show a slight bias at both the lower and higher ends of the predicted values, indicating that the model may be underfitting certain regions.
- The spread of residuals for very low predicted values is much higher, possibly indicating that the model struggles with low-value predictions.

Random Forest Regressor

Hyperparameter Tuning for Random Forest Regressor

Random Forest Regressor (RFR) hyperparameter tuning, we explore various combinations of model parameters to find the best configuration for performance. In this case, the parameters being tuned are:

- n estimators**:** The number of trees in the forest. In your grid, we're testing 100, which means you'll evaluate the performance of a forest made of 100 trees.
- max depth**:** The maximum depth of each tree. Deeper trees can capture more complex patterns but might lead to overfitting. You're testing with a depth of 10.
- min samples split**:** The minimum number of samples required to split an internal node. A lower value allows the model to make splits even with fewer data points, potentially creating more granular trees. Here, you've set it to 2, allowing a split with at least 2 samples.
- min samples leaf**:** The minimum number of samples required to be at a leaf node. Setting this to 1 ensures that each leaf contains at least 1 data point.

Evaluation Metrics for Random Forest Regressor

- **Root Mean Squared Error (RMSE)**: The train RMSE of **2514.84** indicates that the model's predictions are, on average, 2514.84 units away from the actual values on the training data. The test RMSE of **3150.15** is higher, which suggests a possible overfitting issue or that the test data poses additional challenges compared to the training data.

- **Mean Absolute Percentage Error (MAPE)**: The train MAPE is **10.16%**, indicating that the model's predictions are, on average, 10.16% off from the actual values in the training dataset. The test MAPE of **12.31%** is slightly higher, but this difference remains within an acceptable range, suggesting good generalization capabilities of the model.

- **R-squared**: The train R-squared value of **0.9692** implies that approximately 96.92% of the variance in the target variable is explained by the model on the training data. In contrast, the test R-squared of **0.9516** indicates that about 95.16% of the variance is explained on the test set, reflecting solid performance on unseen data.

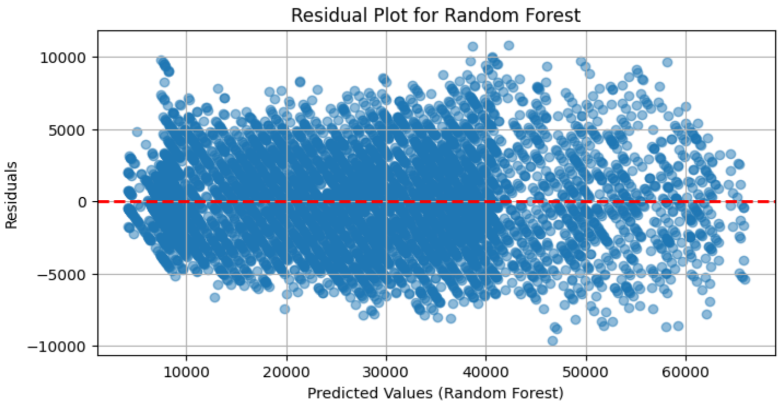
- **Adjusted R-squared**: The adjusted R-squared values are high as well, with the train value at **0.9691** and the test value at **0.9514**. This consistency suggests that the model effectively captures significant relationships in the data while maintaining a low risk of overfitting.

Summary of Insights:

The Random Forest Regressor demonstrates strong performance with both train and test data, evidenced by high R-squared and adjusted R-squared values (95.16% and 96.92%, respectively), indicating the model explains most of the variance in insurance costs. The RMSE and MAPE are reasonably low, with a slight increase on the test set, suggesting the model captures patterns effectively but may have slight overfitting. The close alignment of R-squared and adjusted R-squared values shows the model's complexity is appropriate without overfitting. Overall, the model outperforms alternatives like Lasso and Ridge regression.

Residual Plot for Random Forest Regressor

- The residuals are fairly symmetrically distributed around zero, indicating that the model doesn't exhibit strong bias, performing well across the range of predicted values.
- There are fewer extreme residuals at both the lower and upper ends of predicted values, meaning that Random Forest performs more consistently across different prediction ranges.
- Some residuals slightly above and below the red line indicate a small number of underpredictions and overpredictions.



- The random scatter of residuals suggests Random Forest captures non-linear relationships in the data effectively without obvious patterns in the residuals.

Random Forest Feature Importances

- **Weight:** Importance (%): 98.11%
  The variable "Weight" stands out as the most critical predictor of the target variable, contributing 98.11% to the model's predictions. This suggests that variations in weight are strongly associated with changes in the target outcome (presumably health-related costs, claims, or similar metrics).

- **Covered by any other company (Y):** Importance (%): 0.23%
  This feature has minimal to no effect on the predictions, as indicated by its zero-importance score.

XGBoost Regressor

XGBoost (Extreme Gradient Boosting) is a highly efficient and scalable machine learning algorithm that belongs to the family of gradient boosting methods. It is designed to optimize both speed and performance, making it particularly effective for large-scale and complex datasets. XGBoost builds an ensemble of decision trees sequentially, where each new tree corrects the errors of the previous ones.

Hyperparameter tuning for for XGBoost Regressor

The XGBoost Regressor is a powerful ensemble machine learning model that leverages gradient boosting techniques to improve predictive performance. Tuning its hyperparameters is crucial for optimizing the model's accuracy and preventing overfitting.

- max depth: [3, 4] - This parameter controls the maximum depth of each tree. A shallower tree (depth 3) may prevent overfitting, while a deeper tree (depth 4) may capture more complexity in the data. Testing both values allows for a balance between bias and variance.
- min child weight: [1] **-** This parameter defines the minimum sum of instance weight (hessian) needed in a child. A higher value prevents the model from learning overly specific patterns (overfitting). Here, it is set to 1 to ensure that at least one instance is present in the leaf nodes.
- subsample: [0.8] **-** This parameter indicates the fraction of samples to be used for fitting individual base learners. A value of 0.8 means that 80% of the data will be randomly sampled for training each tree, helping to prevent overfitting.
- colsample bytree: [0.8] - This parameter represents the fraction of features to be randomly sampled for each tree. Setting this value to 0.8 means that 80% of the features will be used in the training process for each tree, promoting diversity and reducing overfitting.
- learning rate: [0.1] - This parameter controls how much the model is updated during each boosting step. A lower learning rate (0.1) can lead to more accurate models but requires more boosting rounds to converge.

Evaluation Metrics for XGBoost Regressor

- **Root Mean Squared Error (RMSE):** The train RMSE of 3007.07 indicates that, on average, the model's predictions are about 3007.07 units away from the actual values on the training data. The test RMSE of 3106.46 is slightly higher, suggesting that the model may have some degree of overfitting or that the test data is inherently more challenging to predict.

- **Mean Absolute Percentage Error (MAPE):** The train MAPE of 12.12% shows that the model's predictions are, on average, off by 12.12% of the actual values in the training dataset. The test MAPE of 12.30% is very close, indicating the model performs similarly on unseen data, which is a positive sign for generalization.

- **R-squared:** The train R-squared of 0.9559 suggests that approximately 95.59% of the variance in the target variable is explained by the model on the training data. A test R-squared of 0.9529 indicates that about 95.29% of the variance is explained on the test set, which is excellent and indicates good model performance.

- **Adjusted R-squared:** The adjusted R-squared values are also very high, suggesting that the model is well-suited for the data and does not overfit excessively. The values (train: 0.9558, test: 0.9528) suggest that the model retains good explanatory power even after accounting for the number of predictors used.

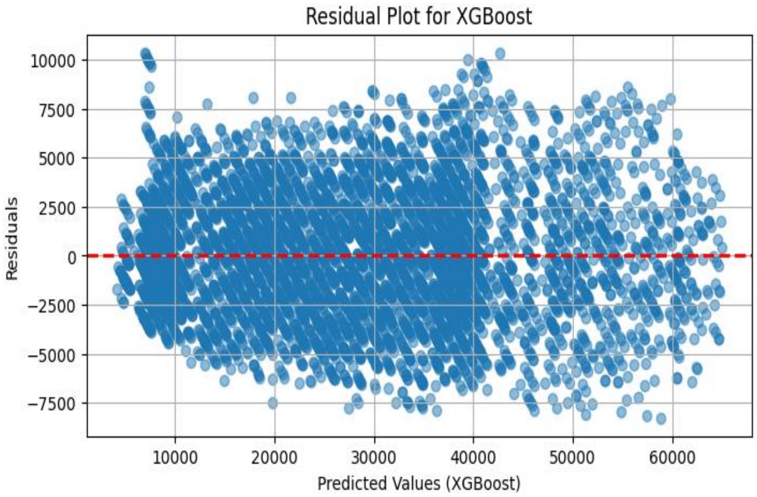
Overall Performance Assessment

Strong Predictive Power: The XGBoost model shows a strong ability to predict the target variable, as indicated by high R-squared values and relatively low RMSE and MAPE metrics for both training and test datasets.

Good Generalization: The close RMSE and MAPE values between the training and testing datasets indicate that the model is generalizing well to unseen data, minimizing concerns about overfitting.

Residual Plot for XGBoost Regressor

The residual plot for the XGBoost model shows a clear spread of residuals across different predicted values. Here are some observations:



1. **Cantered Residuals**: The residuals are fairly centred around the red dashed line (representing zero), which indicates that on average, the model is unbiased and doesn't consistently overpredict or underpredict across the predicted value range.
2. **Spread of Residuals**: The residuals seem to have a relatively uniform spread for most predicted values, but there appears to be more variability at higher predicted values (above 40,000). This could suggest that the model may struggle with predictions at higher costs.
3. **No Clear Pattern**: No strong pattern (like a funnel shape or curved structure) is present, which is a positive sign, as patterns in the residuals could indicate model misspecification or unaddressed non-linearity.
4. **Potential Outliers**: Some residuals fall far from zero, which might be indicative of outliers or specific instances where the model's predictions are particularly off.

Features Importance for XGBoost Regressor

- **Weight:** Importance (%): 77.63%
  Weight is the most critical predictor in the model, accounting for 77.63% of the model's decision-making process. This suggests a strong relationship between weight and the target variable, indicating that changes in weight could be a significant factor in health outcomes or insurance metrics.

- **Weight Change in Last One Year:** Importance (%): 12.26%
  This feature is the second most influential, contributing 12.26% to the predictions. It highlights the importance of tracking weight changes over time, suggesting that both static weight and its changes are crucial indicators of health status or risk assessment.

- **Adventure Sports Participation (Adventure sports 1):** Importance (%): 4.74%
  While this feature has a lower importance score, at 4.74%, it still plays a role in the model. Participation in adventure sports could indicate a higher activity level or risk tolerance, which may relate to health metrics or insurance considerations.

- **Regular Checkup Last Year**: Importance (%): **1.54%**
  This feature's contribution of 3.21% indicates that customers who have had regular health checkups may show different patterns in claims or risks, likely due to the preventive care associated with regular screenings.

- **Covered by Any Other Company** Importance (%): **1.74%**
  This feature has a lower importance yet it suggests that customers with coverage from another insurer may present different risks or behaviors. This insight could help tailor competitive offers to retain customers

Ada Boost Regressor

Hyperparameter tuning for Ada Boost Regressor

In the hyperparameter tuning process for the Ada Boost Regressor, we aim to find the best combination of hyperparameters to improve model performance.

- n estimators: This parameter represents the number of weak learners (decision trees) to be combined in the boosting process. Testing different values like 50, 100, and 200 helps us determine the optimal number of estimators that balances bias and variance. More estimators often reduce bias but can increase the computational cost and risk of overfitting.

- learning rate: The learning rate controls the contribution of each weak learner to the final ensemble. Lower values (like 0.01 or 0.1) make the model more conservative and require more estimators, while higher values (such as 1.0) allow the model to learn faster but may increase the risk of overfitting.

By applying a grid search over these hyperparameters, we can evaluate multiple combinations using cross-validation to select the set that yields the best predictive performance.

Evaluation Metrics for Ada Boost Regressor

- **RMSE (Root Mean Square Error):** The Train RMSE of 3214.34 and Test RMSE of 3270.38 indicate that the model performs well in predicting the target variable, with only a moderate increase in error from training to testing. This suggests that the model generalizes well and does not suffer significantly from overfitting.

- **MAPE (Mean Absolute Percentage Error):** The Train MAPE of 14.32% and Test MAPE of 13.99% reflect the model's prediction accuracy as a percentage. A MAPE below 15% is generally considered acceptable, indicating that the model's predictions are reasonably close to the actual values on average.

- **R-squared:** The Train R-squared of 0.9496 suggests that approximately 94.96% of the variance in the target variable can be explained by the model on the training set. This is a strong indicator of the model's effectiveness. The Test R-squared of 0.9479 is slightly lower but still reflects a high level of explanatory power, indicating that the model remains robust on unseen data.

- **Adjusted R-squared:** The Train Adjusted R-squared of 0.9496 and Test Adjusted R-squared of 0.9476 further confirm the model's suitability, as these values account for the number of predictors in the model, ensuring that the model remains parsimonious while effectively explaining the data.

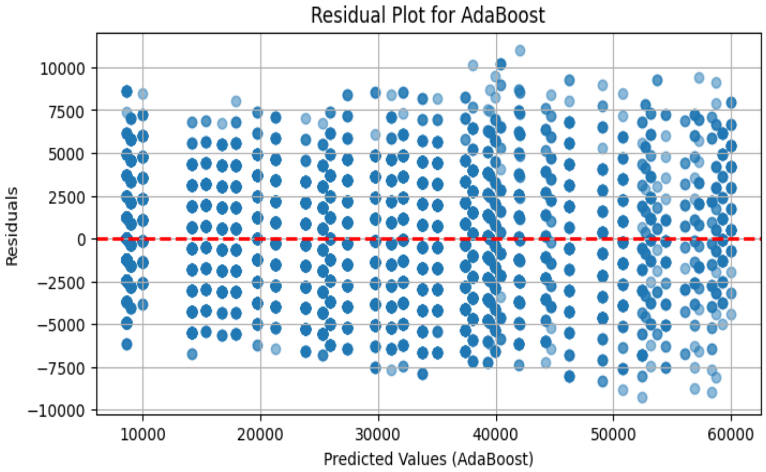
General Insights

- Consistency Across Datasets: The close alignment of the training and testing metrics suggests that AdaBoost has captured the underlying patterns in the data effectively, making it a reliable choice for predictions in this context.
- Room for Improvement: While the performance is commendable, further optimization might be pursued by tuning hyperparameters or experimenting with different base learners to enhance the model's predictive power even further.

Residual Plot for Ada Boost

In the residual plot for the AdaBoost model, here are some key observations:

1. **Segmented Predictions**: The predicted values seem to fall into distinct bands or clusters rather than being continuous, which is typical of AdaBoost. This is because AdaBoost, particularly with weak learners (e.g., decision stumps), tends to make step-like predictions. The segmented nature of the predicted values suggests that the model might be over-simplifying the relationships in the data.



2. **Residual Spread**: Similar to the XGBoost plot, the residuals are centered around zero (the red dashed line), indicating that the model is relatively unbiased. However, the spread of residuals varies across different predicted values, with some groups showing higher variance in residuals.
3. **Higher Variability in Certain Ranges**: There seems to be a higher variability in residuals for predicted values around 30,000 to 50,000. This suggests that the model is struggling more with predictions in this range.
4. **Potential Overfitting or Underfitting**: The "banding" structure could indicate that AdaBoost is either overfitting certain segments or underfitting the overall pattern, especially if the weak learners aren't capturing the full complexity of the data.

Interpretation of Feature Importance for Ada Boost Regressor

- Weight as a Key Driver: The Weight feature has an importance score of 1.00, indicating that it is the sole driver of the model's predictions. This means that the model relies entirely on the Weight variable for making decisions, effectively deeming it the most significant predictor in the dataset.
- Considerations for Other Features: The absence of other significant features may suggest either that they do not contribute meaningfully to the predictions or that they are not necessary for this model. This could prompt further investigation into the relationships between these features and the target variable. It might also indicate that simplifying the model to include only the Weight variable could be sufficient for predictive purposes.

Support Vector Regressor (SVR)

Hyperparameter tuning for SVR

The hyperparameters optimized for the SVR model include C, epsilon, and kernel. After tuning, the best configuration was found to be:

- **C = 10**: Balances the trade-off between margin maximization and training error minimization.
- **Epsilon = 1**: Defines the margin within which no penalty is given to errors, helping the model generalize better.
- **Kernel = 'linear'**: Indicates that a linear kernel provides the best performance in this case, meaning the relationship between features and target is mostly linear.

Model Performance Metrics:

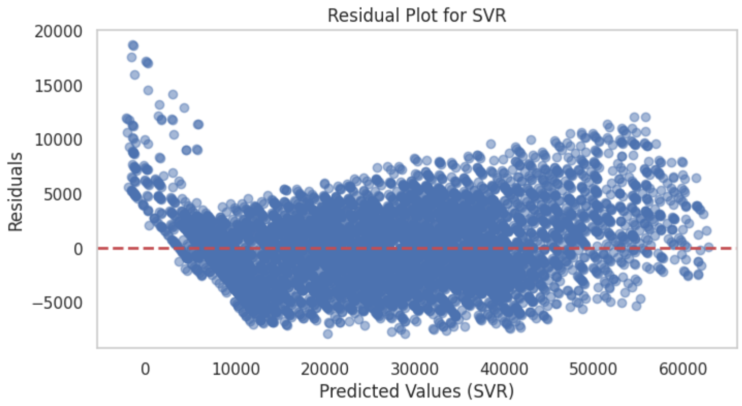
Cross-Validated MAPE (0.15%): Indicates the model is highly accurate across different subsets of the data, with an average error of just 0.15%.

Evaluation Metrics for Support Vector Regression (SVR)

- **RMSE (Training: 3426.61, Test: 3476.00)**: The model's average prediction error is 3426.61 on the training data and 3476.00 on the test data. This indicates a good fit, as the errors are relatively close, suggesting minimal overfitting.
- **MAPE (Training: 15.36%, Test: 15.16%)**: The Mean Absolute Percentage Error indicates that, on average, the model's predictions are about 15.36% off from actual values in the training data and 15.16% off in the test data. This consistency across datasets reflects strong accuracy in predictions.
- **R-squared (Training: 0.9428, Test: 0.9411)**: The R-squared value suggests that approximately 94.28% of the variance in the training data and 94.11% in the test data is explained by the model. This indicates robust predictive power across both datasets.
- **Adjusted R-squared (Training: 0.9427, Test: 0.9408)**: The adjusted R-squared values are very close for both datasets, indicating that the model captures essential relationships without excessive complexity, showcasing good performance overall.

Residual Plot for SVR

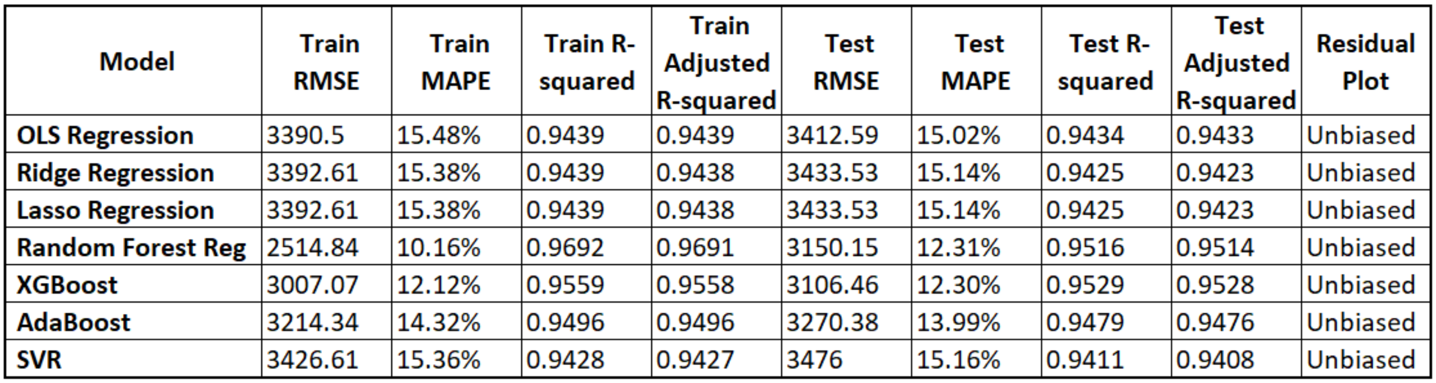
The residual plot for the Support Vector Regressor (SVR) model provides important insights into the model's performance across different predicted values. Here are some key observations:



- **Centered Residuals**: The residuals are fairly centered around the red dashed line (representing zero), suggesting that the model is generally unbiased and does not consistently overpredict or underpredict across the range of predicted insurance costs.
- **Spread of Residuals**: There is a noticeable spread of residuals, especially at lower predicted values (below 10,000). This suggests the model struggles more with smaller insurance costs and has greater variance in its errors at these levels.
- **Pattern in Residuals**: A slight pattern can be observed, with residuals showing higher variability at both low and high predicted values. This pattern might indicate the model's difficulty in capturing certain complex relationships within the data, especially for extreme values.

Interpretation of the most optimum model and its implication on the business

Best Performing Model: XGBoost



Model	Train RMSE	Train MAPE	Train R-squared	Train Adjusted R-squared	Test RMSE	Test MAPE	Test R-squared	Test Adjusted R-squared	Residual Plot
OLS Regression	3390.5	15.48%	0.9439	0.9439	3412.59	15.02%	0.9434	0.9433	Unbiased
Ridge Regression	3392.61	15.38%	0.9439	0.9438	3433.53	15.14%	0.9425	0.9423	Unbiased
Lasso Regression	3392.61	15.38%	0.9439	0.9438	3433.53	15.14%	0.9425	0.9423	Unbiased
Random Forest Reg	2514.84	10.16%	0.9692	0.9691	3150.15	12.31%	0.9516	0.9514	Unbiased
XGBoost	3007.07	12.12%	0.9559	0.9558	3106.46	12.30%	0.9529	0.9528	Unbiased
AdaBoost	3214.34	14.32%	0.9496	0.9496	3270.38	13.99%	0.9479	0.9476	Unbiased
SVR	3426.61	15.36%	0.9428	0.9427	3476	15.16%	0.9411	0.9408	Unbiased

The best-performing model among all is XGBoost. The model is highly accurate at both predicting outcomes on training data and generalizing well to unseen test data.

Why XGBoost is the Best Model

1. Balanced Performance (Generalization):

- XGBoost delivers a low Train RMSE (3007.07) and a similarly low Test RMSE (3106.46), which indicates that the model is generalizing well. The small gap between the training and testing performance shows that the model is not overfitting the training data but is capturing underlying patterns in the data well enough to make accurate predictions on unseen test data.

- This generalization ability is further confirmed by the Test R-squared score of 0.9529, meaning the model can explain over 95% of the variance in the test dataset, a very high level of predictive accuracy.

2. Efficiency: XGBoost uses advanced optimization techniques like shrinkage, regularization, and parallel processing, which make it faster and more efficient. This efficiency is particularly important when working with large datasets or when fast model performance is required for real-time applications. The built-in regularization mechanisms can help reduce overfitting and improve model robustness.

Comparison with Other Models

1. Random Forest Regression

- Although Random Forest has a low Train RMSE, indicating that it fits the training data very well, its Test RMSE (3150.15) is slightly higher than XGBoost's Test RMSE (3106.46).

- Additionally, Random Forest shows signs of overfitting. The large gap between the Train RMSE (2514.84) and Test RMSE (3150.15) indicates that Random Forest may be learning noise from the training data that does not translate to good performance on unseen data.
- While Random Forest performs well, XGBoost outperforms it in terms of generalization and is less prone to overfitting, making it a safer and more reliable choice for production.

2. OLS Regression

- OLS Regression has slightly higher Test RMSE (3412.59) compared to XGBoost (3106.46) shows that XGBoost captures more subtle patterns in the data.

3. Ridge and Lasso Regression

- Their performance metrics (Test RMSEs of 3433.53 and Test R-squared of 0.9425) are slightly worse than XGBoost, and they lack the flexibility to improve on this without feature engineering.

4. AdaBoost Regression

- AdaBoost performs fairly well, but XGBoost shows better results in terms of both Train RMSE (3007.07 vs 3214.34) and Test RMSE (3106.46 vs 3270.38).

6. SVR (Support Vector Regression)

SVR performs reasonably well with a Training RMSE of 3426.61 and a Test RMSE of 3476.00, indicating consistent performance across both training and test data. However, despite capturing much of the variance (R-squared: 0.9411 on the test data), SVR is still not the best model for this dataset compared to alternatives like XGBoost, as its error margins (RMSE) remain higher than ideal for precise prediction, especially for more complex patterns in the data.

Conclusion: XGBoost is the best-performing model due to its ability to generalize well, efficient handling of complex data, and built-in regularization. It outperforms other models like Random Forest and OLS by offering a better balance between bias and variance, ensuring lower overfitting risks, and delivering higher prediction accuracy.

Therefore, XGBoost is recommended for production, as it provides the best performance across multiple metrics and is more robust for unseen data predictions.

Business Implications of using Best Model (XGBoost)

1. **Improved Prediction Accuracy**:
   - XGBoost's low **Test RMSE (3106.46)** and **MAPE (12.30%)** indicate that it makes fewer errors and provides highly accurate predictions.
   - This allows the business to make **precise predictions** for insurance costs, reducing the risk of mispricing. Proper pricing ensures that policies are neither underpriced (which could result in losses) nor overpriced (which could lead to losing customers).

Implication: The business can rely on XGBoost for more accurate cost estimations, improving financial planning, risk management, and pricing strategies.

2. **High R-squared Value (0.9529)**:
   - The high **Test R-squared (0.9529)** shows that XGBoost explains over 95% of the variance in the data. This suggests the model captures the key factors driving insurance costs.

Implication: The business can use XGBoost to identify and focus on important variables, leading to more informed decisions on policy pricing and customer risk segmentation.

3. **Efficient Scalability**:
   - XGBoost handles large datasets and complex relationships well without losing performance. This makes it a strong candidate for handling diverse customer data across different segments and regions.

Implication: The business can scale up its operations by applying XGBoost to a larger customer base or for real-time insurance cost predictions without compromising accuracy.

4. **Cost Savings**:
   - By minimizing prediction errors, XGBoost helps reduce the likelihood of underpricing policies. This, in turn, lowers the risk of financial loss due to unexpected high claims.

Implication: XGBoost can be used to optimize pricing, improve profitability, and reduce the operational costs associated with inaccurate predictions.

5. **Risk Assessment and Customer Segmentation**:
   - With the model's ability to rank feature importance, the business can identify which factors most impact insurance costs and focus on those to refine their risk assessment and offer personalized pricing.

Implication: This leads to better **customer segmentation**—allowing the business to create tailored plans for different customer profiles, improving both customer satisfaction and risk management.

In summary, XGBoost delivers **high accuracy**, **cost savings**, and **scalability**, making it a valuable tool for the business to optimize insurance pricing, risk management, and operational efficiency.

Final Recommendation:

Based on analysis, **XGBoost is the best-performing model** for predicting insurance costs. It achieves accuracy and explains a significant portion of the variance compared to other models.

Key Insights from Feature Importance:

1. **Weight (77.63%)**: This is the most influential factor in determining insurance costs, highlighting the importance of considering weight-related factors in pricing policies.

2. **Weight Change in the Last Year (12.26%)**: Significant weight changes are also a critical factor, suggesting the need to track clients' health trends over time.

3. **Adventure Sports Participation (4.74%)**: Engaging in adventure sports moderately impacts costs, indicating higher risk profiles for such clients.

4. **Coverage by Another Company (1.74%)**: Being covered by another insurance company has a minor influence but should be factored into risk assessments.

5. **Regular Checkups (1.54%)**: Clients who had regular checkups show slightly lower costs, suggesting the importance of preventative care in reducing claims.

Actionable Recommendations:

Based on the analysis, here are actionable recommendations to optimize insurance policy pricing and improve risk management:

1. **Incorporate Weight and Health Trends in Policy Pricing:** Given the substantial influence of weight and weight changes on insurance costs, consider implementing a dynamic pricing model. This model could adjust premiums based on clients’ weight management or health improvements over time, encouraging healthier habits and reducing long-term costs.

2. Promote Wellness Programs to Manage Weight: To manage high-risk factors associated with weight, offer clients access to wellness programs focusing on weight management. Incentivizing these programs could not only improve clients' health but also lower potential claim costs, making policies more attractive.

3. Adjust Premiums for Adventure Sports Participants: Adventure sports participation significantly increases risk profiles. Offer tailored policies for adventure sports enthusiasts that adjust premiums according to risk level, or consider a specific coverage plan for high-risk activities.

4. Develop Partnerships for Multi-Company Coverage: For clients with partial coverage from another insurance provider, explore partnerships or bundled offerings to share risk more effectively. This can help reduce underwriting risk and improve client retention by providing a more comprehensive coverage plan.

5. Incentivize Regular Health Checkups: Clients who undergo regular checkups incur slightly lower insurance costs, indicating the benefit of preventive care. Consider premium discounts or other rewards for clients who complete regular health screenings, as this can lead to early detection of health issues and potentially reduce future claims.