UNIVERSITY OF ECONOMICS AND LAW

FACULTY OF FINANCE AND BANKING

FINAL ESSAY

CASE 4: INSURANCE PREMIUM PREDICTION

LECTURER: Master Phan Huy Tam

STUDENT: Nguyen Thi Tuyet Ngan

STUDENT ID: K214051252

CLASS: K21414

Ho Chi Minh city, January 14, 2023

COMMENTS OF THE SUPERVISOR

LIST OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Data description and problem statement

1.1. Describe the dataset

1.1.1. Basic information

This dataset consists of 7 columns and 1338 rows, representing 7 variables and 1338 observations. There are no null values in this data. The dataset contains 3 categorical variables and 4 numerical variables. Specifically in the numerical variables, there are 2 variables of type float64 and 2 variables of type int64.

Table 1. Data information

#	Column	Non-Null	Count	Dtype
1	age	non-null	1338	int64
2	sex	non-null	1338	object
3	bmi	non-null	1338	float64
4	children	non-null	1338	int64
5	smoker	non-null	1338	object
6	region	non-null	1338	object
7	charges	non-null	1338	float64

Source: Author

1.1.2. The meaning of each variable in the dataset

Table 2. Variables meaning

Features	Description
age	The age of the individual
sex	The gender of the individual
bmi	Stands for Body Mass Index, which is a measure of body fat based on height and weight
children	The number of children the individual has
smoker	Whether the individual is a smoker or not
region	The region where the individual is located

charges	the medical insurance charges for each individual

Source: Author

1.1.3. Characteristics of the quantity measured

Table 3. Characteristics of data

#	age	sex	bmi	children	smoker	region	charges
Attribute measured	Age of individuals	Gender or sex of individuals	Body Mass Index (BMI)	Number of children	Smoking status of individuals	Geographical region of individuals	Insurance charges
Unit of measurement	Years	Categorical (Male or Female)	kg/m^2 (kilograms per square meter)	Count (whole numbers)	Categorical (Yes or No)	Categorical (Northeast, Northwest, Southeast, Southwest)	Currency
Number of observation	1338	1338	1338	1338	1338	1338	1338
How was the attribute measured	Self-reported or obtained from records or surveys	Self-reported or obtained from records or surveys	Self-reported or obtained from records or surveys	Self-reported or obtained from records or surveys	Self-reported or obtained from records or surveys	Self-reported or obtained from records or surveys	Self-reported or obtained from records or surveys

Source: Author

1.2. The problem statement

The objective of this project is to provide individuals with an estimation of the amount they require based on their unique health situation. Subsequently, customers can collaborate with any health insurance provider and explore their plans and benefits, while keeping the projected cost from our study in mind. This can assist individuals in focusing on the health aspect of an insurance policy rather than its inefficiencies.

1.3. The nature of this case

The nature of this case is an analysis of a dataset containing information about individuals' insurance-related attributes. It involves examining the relationships between variables such as age, gender, BMI, number of children, smoking status, region, and insurance charges. The goal is to gain insights into the factors that influence insurance charges and understand how these variables are associated with each other.

The case aims to provide a comprehensive understanding of the dataset and uncover patterns or trends that can inform insurance pricing and decision-making.

2. Data issues and solutions

From this step onwards, the variable 'children' will be converted to the category type for the following reasons:

Accurate representation: 'children' has discrete values that represent different groups (0, 1, 2, 3, 4, 5). By converting 'children' to the category type, we can accurately represent the meaning of the data. We can consider 'children' as a categorical variable with different values representing each group.

Visualization purposes: Using a categorical variable instead of a numerical variable can make data visualization easier. Categorical variables can be used to create meaningful plots and charts that effectively communicate the distribution and patterns within different groups.

Support for machine learning models: In some cases, machine learning models can perform better if the 'children' variable is treated as a categorical variable rather than a numerical variable. Different models may react differently to numerical and categorical variables, and sometimes this conversion can improve the model's performance. Algorithms such as RandomForest, XGBoost, or CatBoost often require categorical variables as inputs to leverage the classification properties of the variable.

Flexibility in analysis: Categorical variables allow for grouping operations and analysis that align with the nature of the data. By converting 'children' to the category type, we can easily group the data by 'children' and calculate group statistics such as counts and percentages.

2.1. Duplicate data

The presence of duplicate data in a dataset can have several effects on data analysis. Firstly, it can introduce bias into the analysis results by disproportionately inflating the importance of certain instances. This can lead to skewed estimates, misleading patterns, and inaccurate conclusions. Secondly, duplicate data can overemphasize certain patterns or relationships, making them appear more significant than they actually are. This can result in an inflated sense of correlation or predictive power. Lastly, duplicate data reduces the representativeness and diversity of the dataset,

potentially compromising the generalizability of the findings. This dataset contains duplicate data:

Table 4. Duplicate data

#	row	age	sex	bmi	children	smoker	region	charges
1	195	19	male	30.59	0	no	northwest	1639.5631
2	581	19	male	30.59	0	no	northwest	1639.5631

Source: Author

There are lots of way to solve this problem: removing, merging, flagging and choosing duplicate. But the simplest and most common way is to remove them from the dataset. This can reduce the size and complexity of the dataset, and improve the quality and performance of the machine learning models. Therefore, duplicate data will be removed from this dataset.

2.2. Outliers

Outliers can distort statistical measures such as the mean and standard deviation, leading to biased estimates of central tendency and variability. They can also impact the results of statistical tests, potentially leading to incorrect conclusions or misleading interpretations. Outliers can also affect the performance of machine learning models by pulling the estimated regression line or decision boundary towards them, resulting in less accurate predictions. The outliers of this dataset have been detected using the z-score:

Table 5. Outliers

#	row	age	sex	bmi	children	smoker	region	charges
1	34	28	male	36.4	1	yes	southwest	51194.55914
2	116	58	male	49.060	0	no	southwest	11381.32540
3	543	54	female	47.410	0	yes	southwest	63770.42801
4	577	31	female	38.095	1	yes	northeast	58571.07448
5	819	33	female	35.530	0	yes	northeast	55135.40209
6	847	23	male	50.380	1	no	southeast	2438.05520
7	1047	22	male	52.580	1	yes	southeast	44501.39820
8	1146	60	male	32.8	0	yes	southeast	52590.82939
9	1230	52	male	34.485	3	yes	southeast	60021.39897

10	1300	45	male	30.360	0	yes	southeast	62592.87309
11	1317	18	male	53.130	0	no	southeast	1163.46270

Source: Author

There are various methods to handle outliers depending on the context, such as removing them, replacing them with the median, replacing them with the mean, etc. In this particular dataset, the author decided to remove the outliers because of their undesirable significance and impact on data analysis. By removing outliers, the author can minimize their influence on the analysis results and improve the accuracy and reliability of statistical measures and predictions. In this dataset, the z-score is still used to remove these outliers as follows:

Step 1: Compute the z-score for each value in the columns of the dataset.

Step 2: Identify the rows that contain outliers by comparing the corresponding z-scores with a threshold (typically 3).

Step 3: Create a mask to specify the rows that do not contain outliers.

Step 4: Optionally, store the outlier rows in a separate DataFrame.

Step 5: Remove the rows containing outliers from the main DataFrame.

3. Special point and potential issue that the analyst must pay attention to

Outliers: Check for outliers in numerical variables like age, BMI, and charges. Outliers can significantly impact statistical measures and may need special consideration.

Data Completeness: Ensure that there are no missing values in critical variables. Missing data can affect the accuracy of analyses and predictions.

Variable Types: Confirm that each variable has the correct data type. For instance, "smoker" should be binary (yes/no) or numerical (1/0), and other variables should be numeric if they are expected to be.

Data Distribution: Examine the distribution of numerical variables to understand their shapes. Skewed distributions may impact the performance of certain statistical methods.

Class Imbalance: Investigate if there is a significant imbalance in the distribution of classes. Class imbalances can affect the performance of machine learning models.

Correlations: Explore correlations between variables. High correlations may indicate multicollinearity, which can affect the interpretability of regression models.

Data Consistency: Ensure consistency in data entry across different categories, regions, or any other categorical variables.

Sampling Bias: Investigate whether the dataset is representative of the entire population. If there's sampling bias, adjustments or weighting may be necessary to improve representativeness.

Confounding Variables: Be aware of potential confounding variables that could influence the relationships observed in the data.

Domain Knowledge: Consider the context of the data and any domain-specific factors that could impact the interpretation of results.

Ethical Considerations: Be mindful of potential biases, ethical concerns, and privacy issues associated with the data and its analysis.

Statistical Significance: Understand the difference between statistical significance and practical significance. A result may be statistically significant but may not have a meaningful impact in the real world.

4. Data exploratory analysis

4.1. Descriptive analysis

Table 6. Descriptive analysis for continuous variables in the model

#	age	bmi	charges
count	1326	1326	1326
mean	39.227753	30.570743	13039.837572
std	14.037019	5.992764	11677.031551
min	18.000000	15.960000	1121.873900
25%	27.000000	26.220000	4740.287150
50%	39.000000	30.300000	9303.297725
75%	51.000000	34.560000	16389.832412
max	64.000000	48.070000	49577.662400

Source: Author

The table presents descriptive statistics for three variables: age, BMI (Body Mass Index), and charges. The dataset contains 1326 observations for each variable.

For the 'age' variable, the mean age is approximately 39.23 years, with a standard deviation of 14.04. The minimum age is 18 years, and the maximum age is 64 years. From these statistics, we can infer that there is a diversity in the ages of individuals in this dataset, ranging from young adults (18 years old) to middle-aged adults (up to 64 years old). Additionally, the mean age of 39.23 suggests that a majority of middle-aged individuals are involved in insurance payments.

Regarding the 'BMI' variable, the mean BMI is approximately 30.57, with a standard deviation of 5.99. The minimum BMI value is 15.96, and the maximum BMI value is 48.07. With an average BMI of 30.57, it can be observed that the BMI of individuals participating in insurance in this dataset is significantly higher than the standard BMI range (approximately 18.5 to 24.9). This suggests that they may be facing a higher prevalence of obesity. The standard deviation of 5.99 indicates a considerable variation in BMI within the dataset, further emphasizing the significant fluctuations in BMI values. These findings support the notion that the individuals in this dataset are more likely to be at risk for obesity-related health issues.

For the 'charges' variable, which represents medical charges, the mean charge is approximately 13,039.84, with a standard deviation of 11,677.03. The minimum charge is 1,121.87, and the maximum charge is 49,577.66. The data above demonstrates a large variation in charges. Some individuals have relatively low charges, while others have much higher charges, potentially due to more extensive or costly medical treatments.

4.2. Univariate Analysis

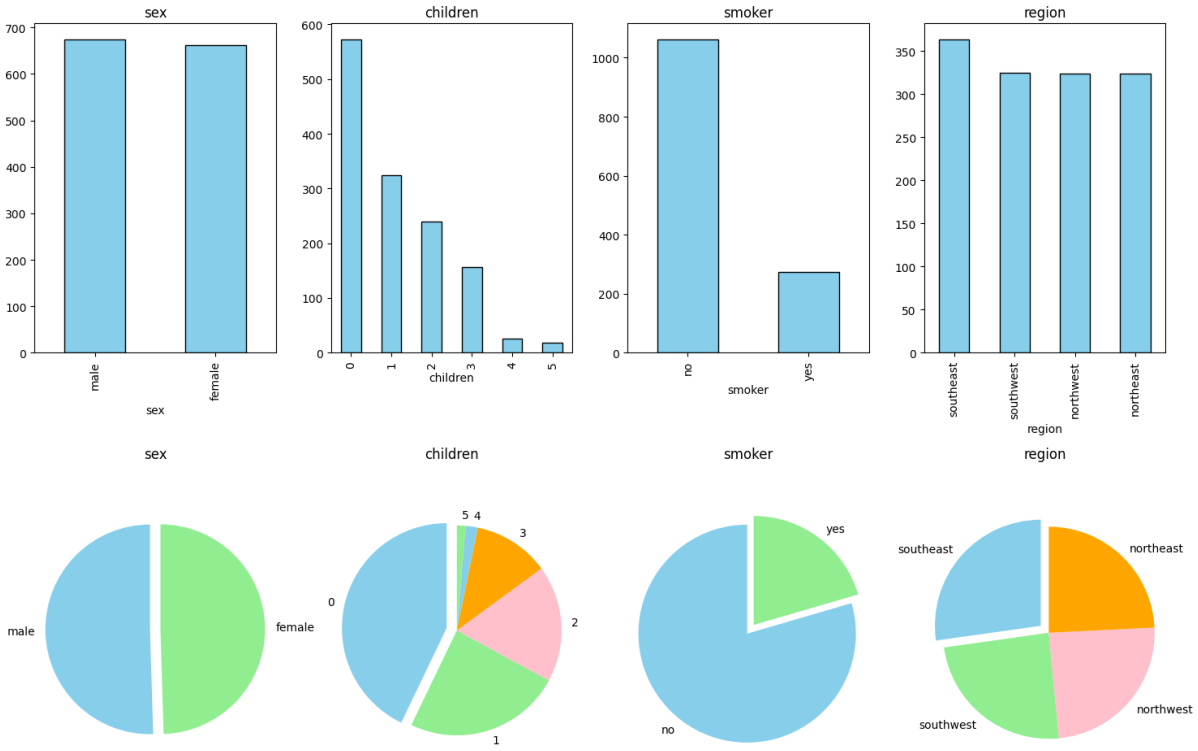

Figure 1. Univariate analysis of categorical variables

From the charts, it is obvious that the number of males and females participating in the insurance is nearly equal, with slightly more males. In terms of children, almost half of the insured individuals do not have any children, a few have 1 to 3 children, and only a few have more than 4 children. Additionally, the majority of insured individuals do not smoke, with around 1000 people. Regarding the region, almost every region has an equal number of insured individuals, except for the southeast which has slightly more than the other regions.

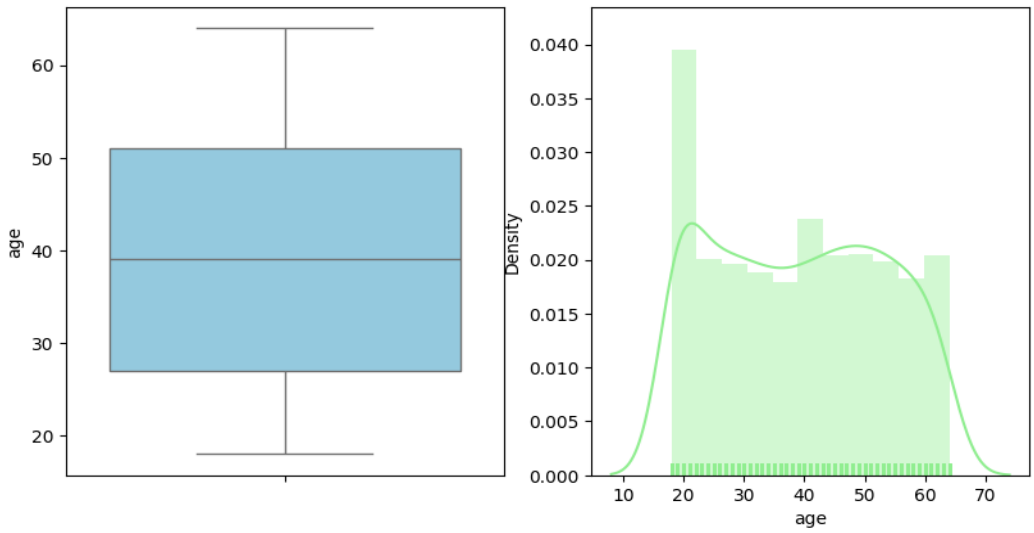

Figure 2. Box plot and distribution plot of age

From the box plot, it is clear that the distribution of the variable "age" is slightly skewed to the right, indicating a higher concentration of individuals participating in insurance at older ages compared to the rest. The plot also reveals that the average age of insurance participants is around 40 years old. Additionally, the plot demonstrates a wide range of ages within the dataset. Regarding the distribution plot, it is obvious that it exhibits a multimodal distribution (specifically, two peaks), indicating higher density for the age groups of 20 and 50-60 compared to the rest of the distribution.

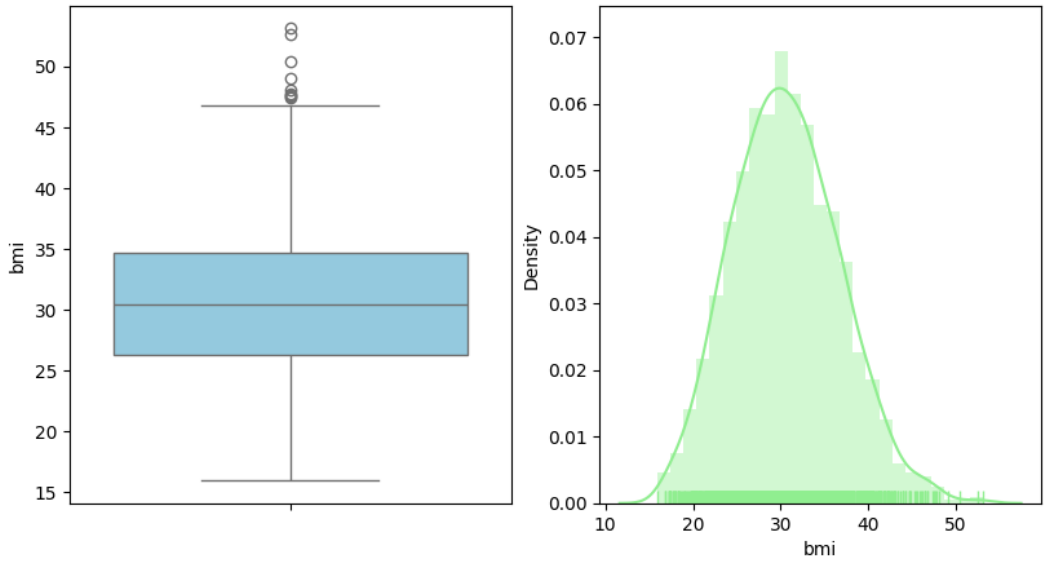

Figure 3. Box plot and distribution plot of bmi

From the box plot, the distribution of the "bmi" variable is slightly skewed to the right. However, when looking at the distribution plot, it appears to have an approximately symmetric shape. Therefore, in this case, there need to consider the mean, median, and mode. The bmi variable has the mean = 30.57, median = 30.3, and mode = 32.3 → mean > median → the graph is skewed to the right → the data aligns with the box plot. This indicates that there are more people with higher BMI compared to the rest. Furthermore, the average BMI in this dataset is around 30, and there are some individuals with BMI values higher than 40.

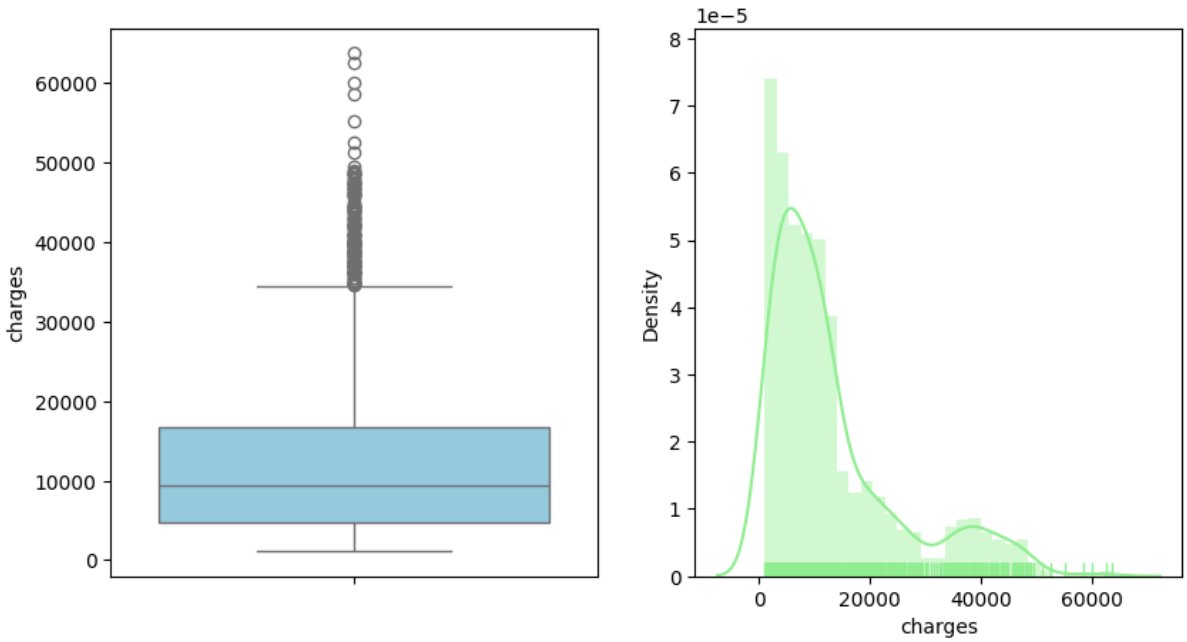

Figure 4. Box plot and distribution plot of charges

From the box plot and distribution plot, it can be observed that the data is right-skewed. To further confirm this, there need to consider the mean, mode, and median. The research finds that mean = 13039.8375 > median = 9303.2977 > mode = 1639.563 → The graph is skewed to the right, similar to the previous two plots. This indicates that the majority of individuals in the dataset have insurance charges above 10000. The average payment is around 10000, and there are also individuals with charges exceeding 40000.

4.2. Bivariate Analysis

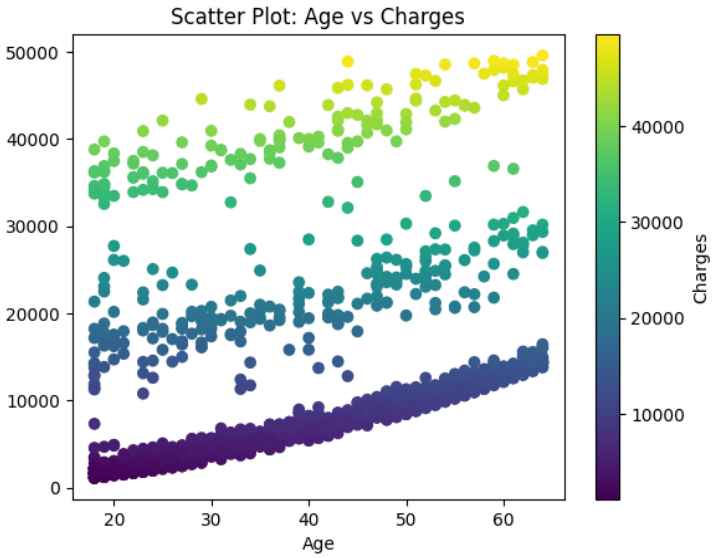

Figure 5. Scatter plot of age and charges

From the scatter plot, it is evident that there is a positive correlation between 'age' and 'charges'. This indicates that as age increases, the charges also tend to increase correspondingly. However, there are also some outliers such as the age of 41 with charges close to 50000, which suggests that there may be other factors beyond age that can contribute to higher charges.

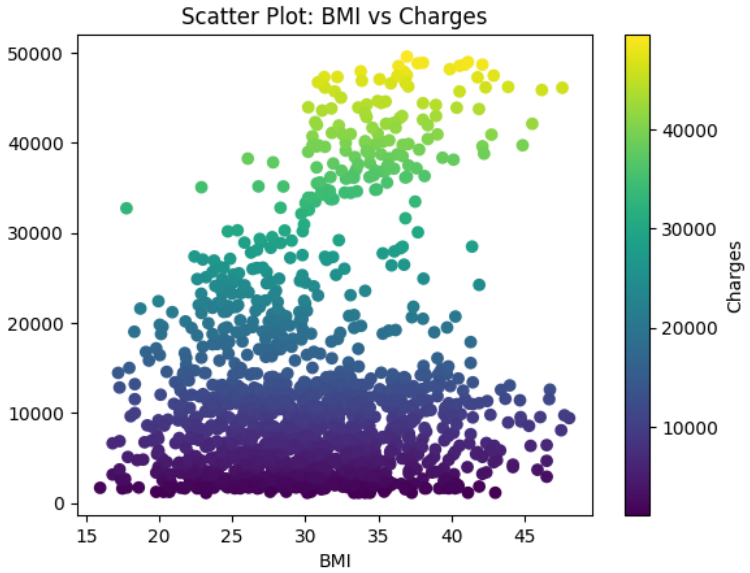

Figure 6. Scatter plot of bmi and charges

From the graph, it can be observed that there is a positive but relatively weak correlation between BMI and charges. People with higher BMI can still have relatively low charges. Most individuals with a BMI in the range of 30-40 tend to have the highest charges. This suggests that while there is a general trend of higher charges with increasing BMI, the relationship is not very strong. Other factors such as age, smoking status, children, region and gender may also play significant roles in determining the charges.

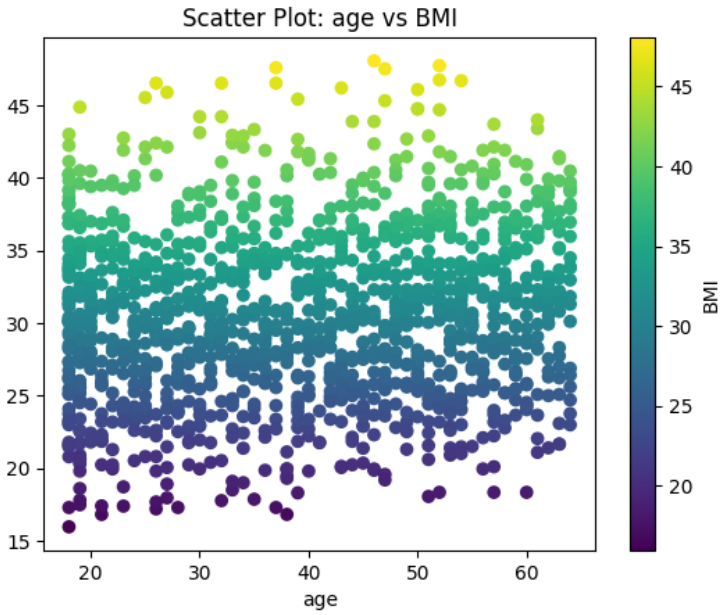

Figure 7. Scatter plot of age and bmi

From the graph, it is evident that there is no clear correlation between age and BMI of an individual. This means that a person of any age can have a BMI that varies depending on their lifestyle and dietary habits. This suggests that age alone is not a determining factor in predicting BMI. Other factors such as genetics, exercise habits, diet, and overall health condition might play a more significant role in influencing an individual's BMI.

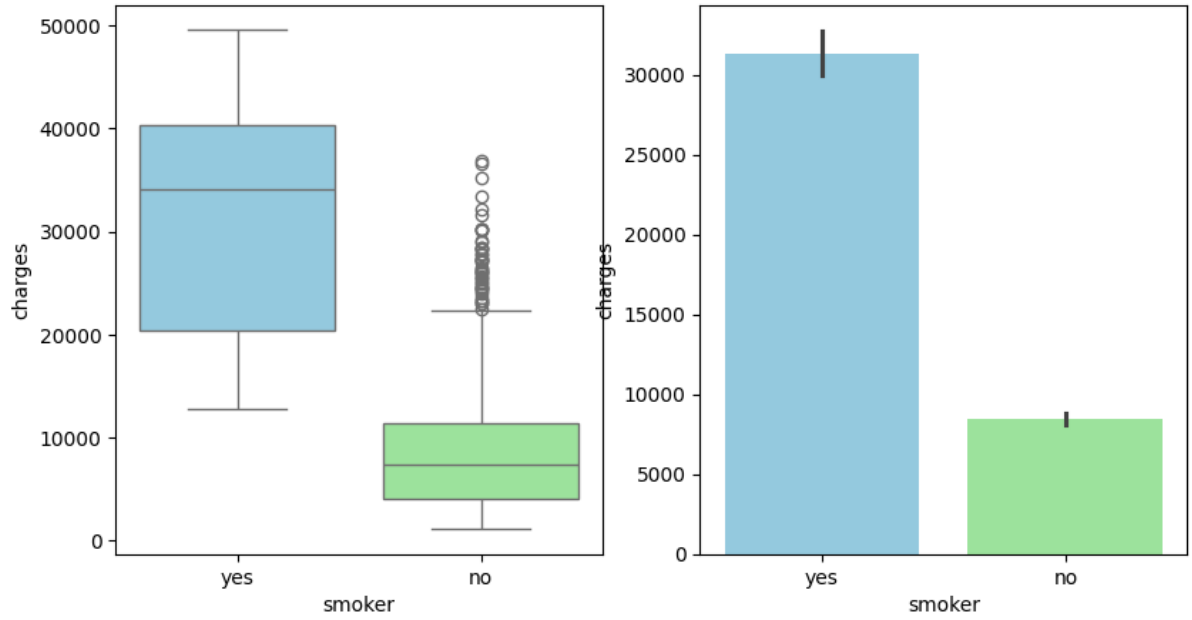

Figure 8. Box plot and bar chart for smoker status and charges

The box plot chart clearly illustrates that, on average, smokers tend to have higher charges than non-smokers. However, there are instances where non-smokers have charges that exceed those of some smokers. This suggests that while smoking is generally associated with higher healthcare costs, there are other factors at play that can influence charges, such as pre-existing conditions or individual healthcare needs.

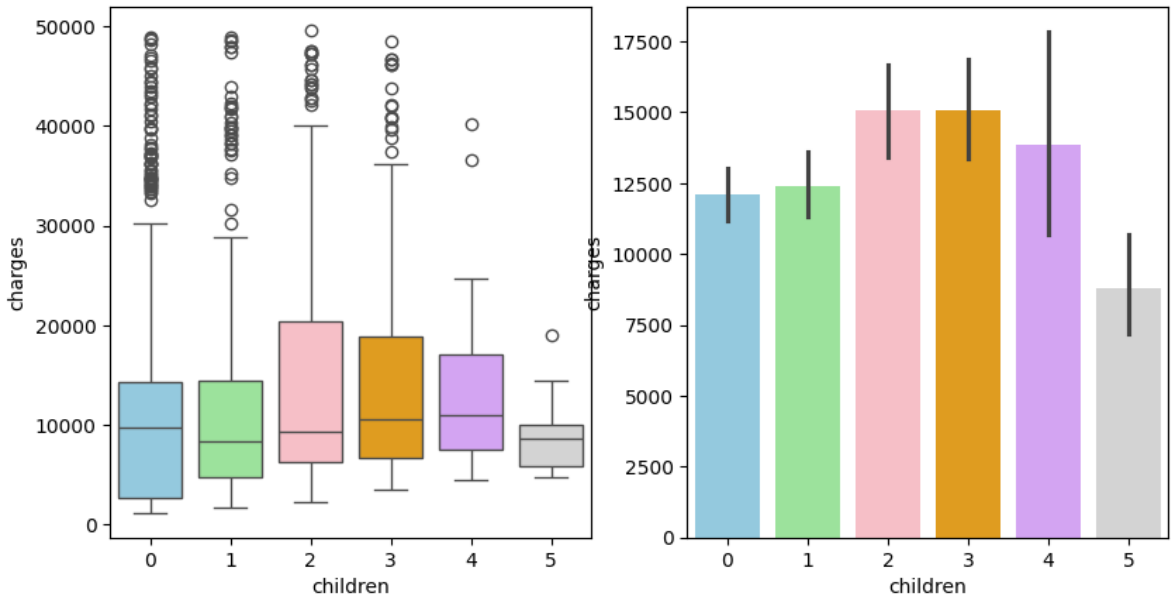

Figure 9. Box plot and bar chart for children and charges

In general, individuals with around 2-3 children tend to have higher charges, while those with 5 children have the lowest charges. However, it is important to note that there are many individuals with only 1 or no children who have significantly higher charges. This observation suggests that while the number of children can have some impact on healthcare charges, it is not the sole determining factor. Other variables such as age, overall health condition, lifestyle choices, and pre-existing medical conditions may also contribute to the variation in charges.

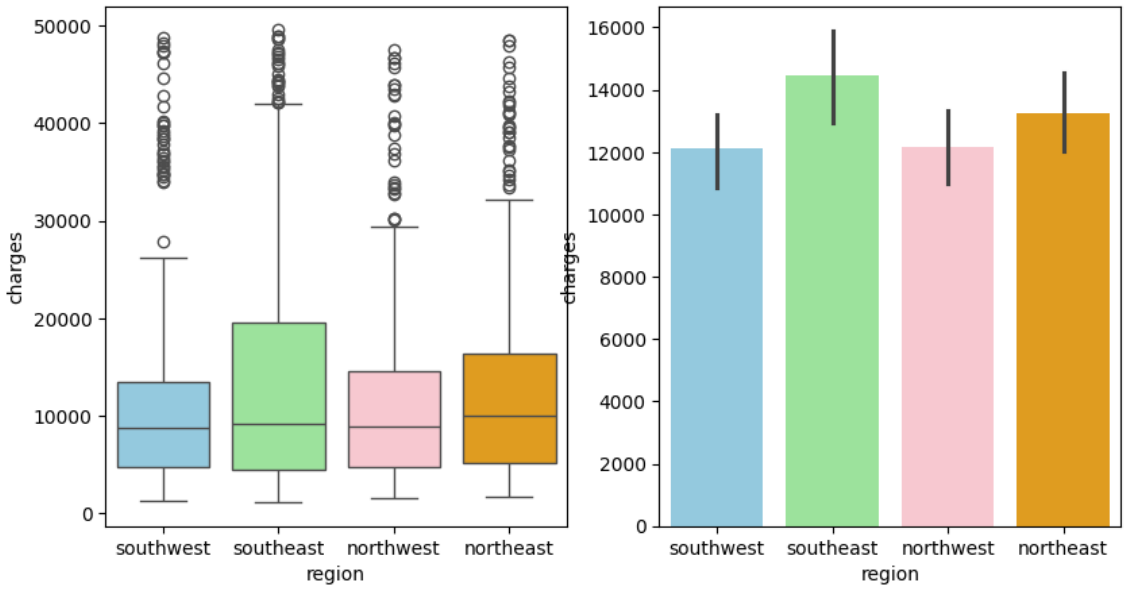

Figure 10. Box plot and bar chart for region and charges

The chart reveals that the average charges across regions are relatively similar, with the southeast region and northeast slightly higher than the others. However, it is important to note that in each region, there are individuals with charges significantly higher than the average. This observation suggests that there are other factors besides region that influence charges.

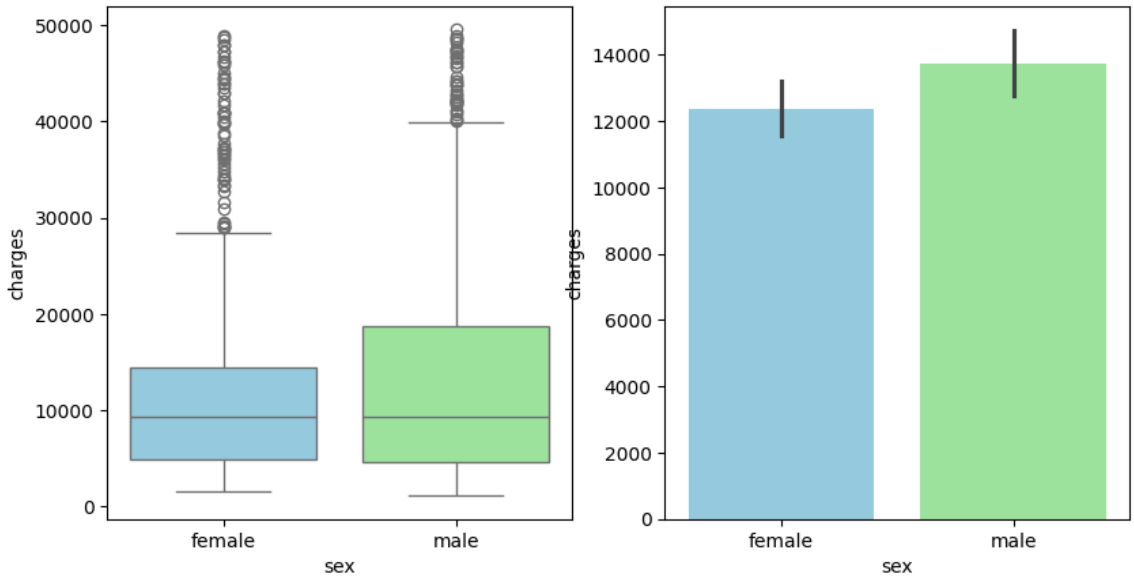

Figure 11. Box plot and bar chart for sex and charges

From the chart, it is evident that males generally have higher insurance charges compared to females. Both males and females exhibit exceptions, with some individuals having charges significantly higher than the average, showcasing there are other factors that could impact the charges.

4.3. Multivariate analysis

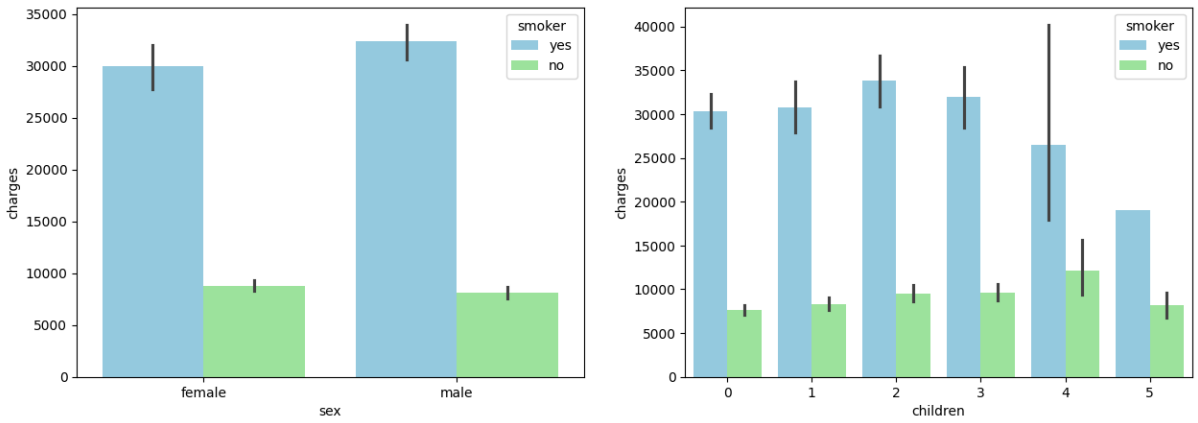

Figure 12. Bar chart for sex, children, smoker and charges

Looking at the sex chart, it can be observed that both males and females who smoke have significantly higher charges compared to the non-smoking group. Regarding the children chart, it is evident that regardless of the number of children, smokers consistently have higher charges than non-smokers. This indicates a correlation between smoking and medical expenses, with smoking potentially being a factor that affects health and thus increases insurance costs. It also suggests that smoking may be a more significant factor than the number of children in influencing medical insurance charges.

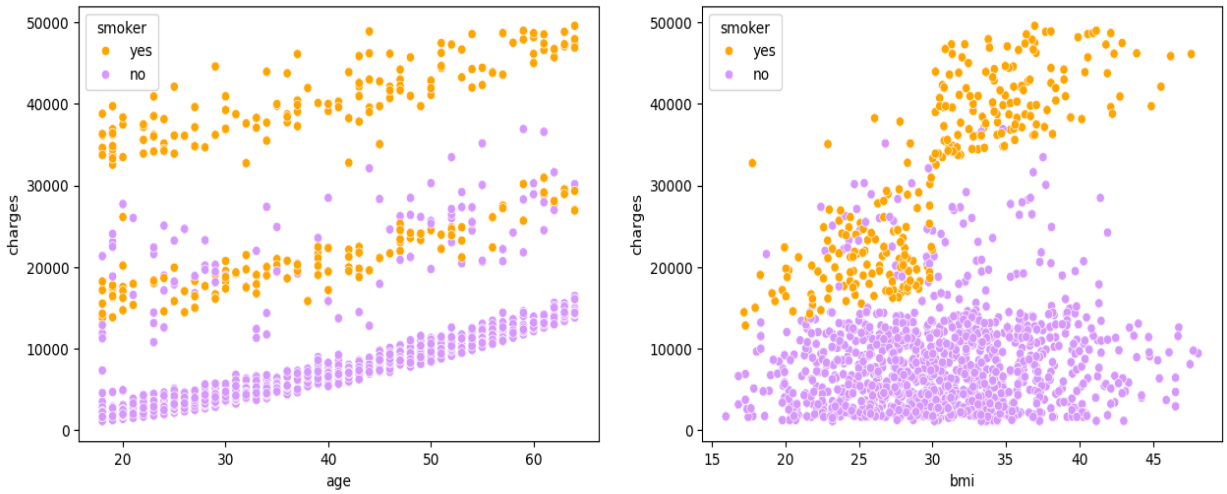

Figure 13. Scatter plot for age, bmi, smoker and charges

From the scatter plots, it is clear that across all ages and BMIs, individuals who smoke tend to have higher charges compared to non-smokers. This further confirms the influence of smoking status on medical insurance charges.

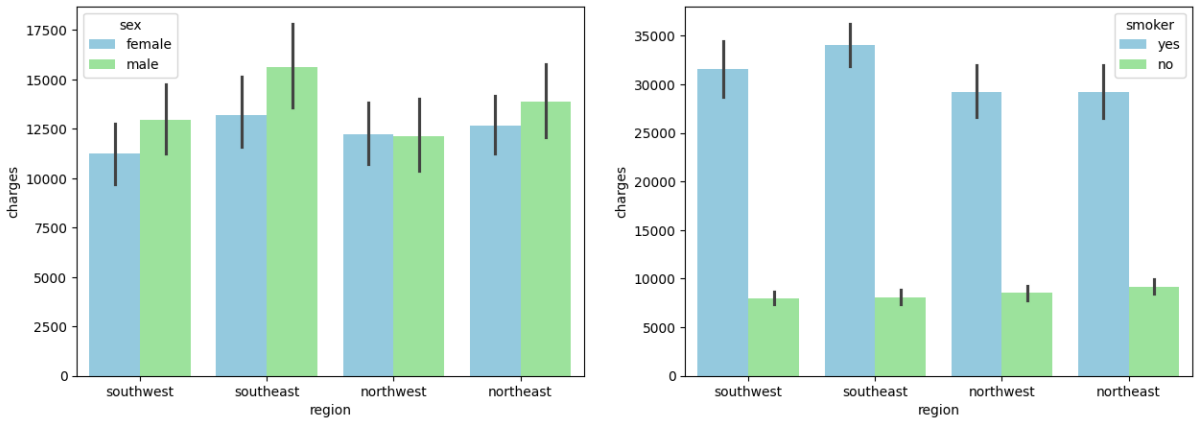

Figure 14. Bar chart for region, smoker and charges

From the analysis, it can be observed that, in general, males tend to have higher insurance charges compared to females, except for the cases in the Northwest and Southeast regions where charges are the highest for both genders. Additionally, it is evident that smokers have significantly higher charges compared to non-smokers, and the charges for non-smokers are relatively consistent across all regions.

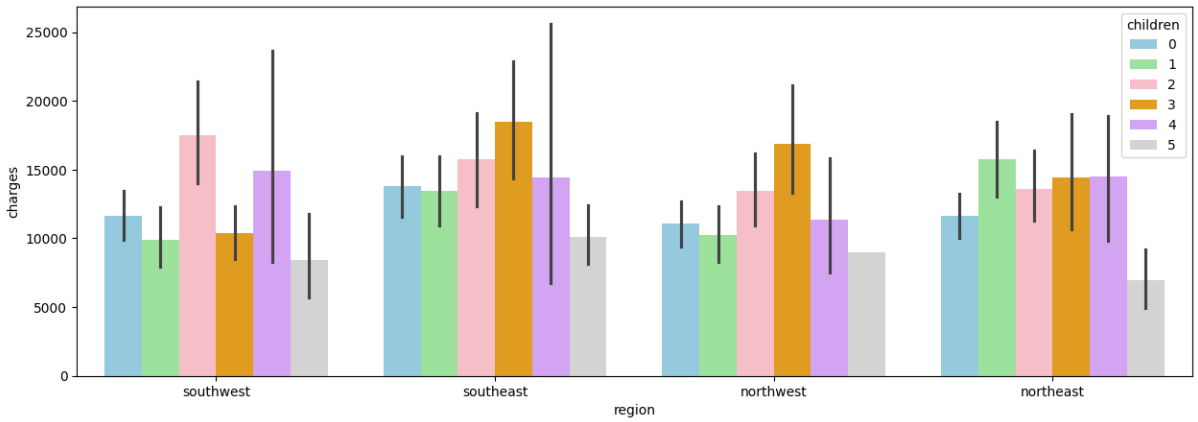

Figure 15. Bar chart for region, children and charges

It is evident that individuals with 5 children have lower charges than other groups across all regions. The Southeast region has higher charges compared to other regions, with the highest charges observed for individuals with 3 children. In contrast, in other regions, individuals with 2-3 children tend to have higher charges, while in the Northwest region, individuals with 1 child have the highest charges. These observations suggest that the number of children is a factor influencing insurance charges. In general,

having more children tends to be associated with lower charges, except in certain regions where different patterns emerge. The Southeast region stands out with higher charges overall, while in other regions, the specific number of children may have varying impacts on charges.

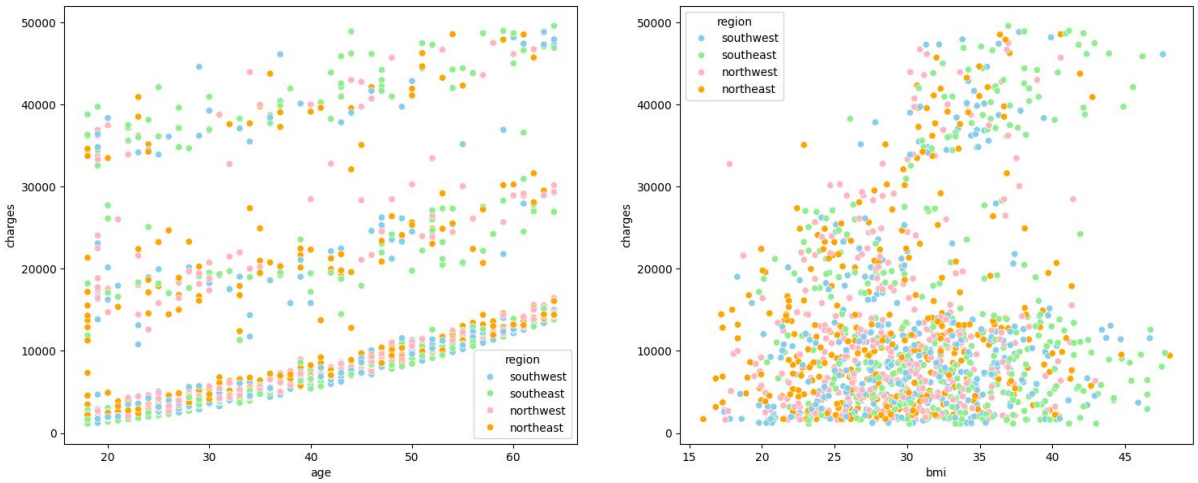

Figure 16. Scatter plot for age, bmi and charges

Regarding the age chart, it can be observed that the distribution of ages is relatively even across most regions. However, the Southeast region has a higher concentration of charges around 40,000 for certain age groups. Looking at the bmi chart, there is a diverse distribution of bmi and charges across regions. The Southeast region has a higher number of individuals with a bmi above 43, while the Northeast region has a higher number of individuals with a bmi below 25, resulting in lower charges compared to other regions.

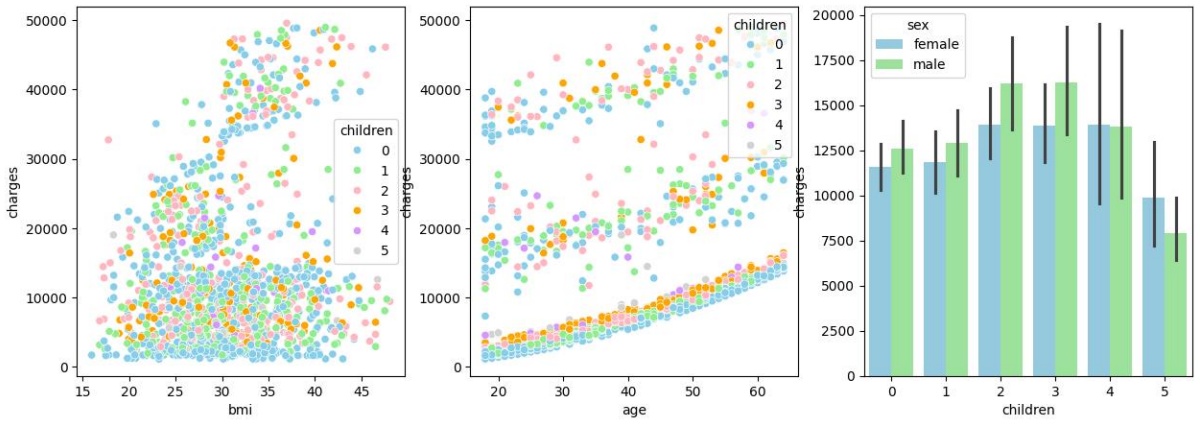

Figure 17. Scatter plot, bar chart for bmi, age, children, sex and charges

Regarding the BMI, it is obvious that there is diversity and an even distribution of BMI values across different levels of charges. This suggests that there is no clear

relationship between BMI, children, and charges. Looking at the age chart, for charges below 15,000, there is a positive relationship between charges, age, and children. As charges increase, the positive relationship still holds but with some attenuation. Regarding the bar chart for children, both males and females with 2 to 3 children have the highest charges, while individuals with 5 children have the lowest charges.

5. Suitable model

In the case of insurance data, where the goal is often to predict insurance charges or analyze factors affecting insurance costs, various types of models can be used. Here are a few commonly used models for insurance data:

Linear regression: Linear regression is a simple and interpretable model used in insurance data analysis. It is employed to predict insurance expenses based on input features such as age, BMI, number of children, and smoker status. The linear regression model assumes a linear relationship between these input features and the target variable. There are several reasons why linear regression should be used:

(i) Interpretability: Linear regression provides interpretable results. The coefficients estimated by the model represent the impact of each predictor variable on the insurance charges. For example, if the coefficient for the "smoker" variable is positive and significant, it indicates that being a smoker is associated with higher insurance charges. This interpretability allows individuals to understand the relationships between the variables and the target variable.

(ii) Simplicity: Linear regression is a straightforward and easy-to-understand model. It has a clear mathematical formulation, making it accessible to analysts and stakeholders who may not have a deep understanding of complex machine learning algorithms.

(iii) Efficiency: Linear regression is computationally efficient, especially for large datasets. The model can be trained quickly, making it suitable for scenarios where real-time or near-real-time predictions are required.

(iv) Feature importance: Linear regression can help identify the most important features in predicting insurance expenses. By examining the magnitude and significance of the coefficients, analysts can determine which variables have the most influence on

the target variable. This information can be valuable in decision-making and feature selection processes.

Ridge Regression*:* Ridge regression is a variant of linear regression that introduces a regularization term to the loss function. The regularization term, known as the L2 penalty, helps to prevent overfitting and improve the model's generalization performance. Ridge regression can be used to analyze the impact of variables such as BMI, age, sex, and children on insurance charges. It estimates the coefficients associated with each variable, revealing the direction and magnitude of their influence on the charges. For example, the BMI coefficient can show how charges change with BMI, while coefficients for age, sex, and children can provide insights into their respective effects. This model should be used because:

(i) Handling multicollinearity: Ridge regression is particularly useful when dealing with multicollinearity, which occurs when the predictor variables are highly correlated. In your dataset, it's possible that age, BMI, and smoker status, for example, could be correlated. Ridge Regression can handle this situation by reducing the impact of correlated variables and providing more stable and reliable estimates of their effects on charges.

(ii) Regularization and feature selection: Ridge regression incorporates a regularization term that helps prevent overfitting by shrinking the regression coefficients towards zero. This regularization can be beneficial when you have a limited amount of data or when dealing with high-dimensional datasets. It can also help with feature selection by automatically reducing the impact of less important variables. In your case, if some of the variables have a weaker relationship with charges, Ridge Regression can effectively downweight their influence on the predictions.

(iii) Robustness to outliers: Ridge regression is known to be more robust to outliers compared to ordinary least squares regression. Outliers in the dataset, such as unusually high or low insurance charges, might exist and can have a significant impact on the regression results. By penalizing large coefficient values, ridge regression can mitigate the influence of outliers and provide more robust predictions.

(iv) Interpretable results: Ridge Regression provides interpretable results by assigning weights (coefficients) to each predictor variable. These weights indicate the

strength and direction of the relationship between each variable and the target variable (charges). With Ridge Regression, analyst can easily interpret how each variable contributes to the predicted insurance charges, making it useful for understanding the factors that drive insurance costs.

Decision Trees: Decision trees are non-linear models that can capture complex relationships between input features and insurance expenses. They can handle both numerical and categorical variables and provide interpretable rules for making predictions. Decision trees can be further improved by using ensemble methods like Random Forests or Gradient Boosting. There are numerous reasons why decision trees should be utilized:

(i) Interpretability: Decision Trees provide easily interpretable rules that can be understood and explained. For instance, the model might split the data based on age, indicating specific age ranges where the likelihood of claims is higher or lower. This interpretability can help insurance professionals understand the factors influencing claims and make informed decisions.

(ii) Handling categorical and numerical variables: Decision Trees can handle both categorical and numerical variables effectively. In the insurance context, this is particularly useful since you may have a combination of variables like age (numerical) and smoking status (categorical). The model can automatically determine the optimal splitting points based on these variables, capturing their impact on claims.

(iii) Nonlinear relationships: Decision Trees can capture nonlinear relationships between variables. For example, the model can identify that for policyholders in certain occupations or with specific medical histories, the likelihood of claims may vary significantly. This flexibility allows Decision Trees to capture complex interactions that linear models might miss.

(iv) Feature importance: Decision Trees provide a measure of feature importance, indicating the variables that have the most significant influence on the predictions. This information can guide decision-making and risk assessment. For instance, people may find that age and medical history have the highest importance, allowing them to focus on these factors when determining insurance premiums or designing targeted interventions.

(v) Handling missing data: Decision Trees can handle missing values by utilizing available information for splitting the data. This means you don't have to discard incomplete records, preserving valuable data for analysis.

Gradient Boosting: Gradient boosting models (e.g., XGBoost, LightGBM) are also ensemble methods that combine weak learners (usually decision trees) in a sequential manner. They iteratively correct the mistakes of previous models and build a strong predictive model. Gradient boosting models often provide high predictive accuracy and can handle heterogeneous data types. Here are some specific applications and benefits of Gradient Boosting in the insurance domain:

(i) Risk assessment: Insurance companies rely on accurate risk assessment to determine premiums and make informed decisions. Gradient boosting models can effectively analyze diverse variables such as age, sex, BMI, number of children, smoking status, and geographic region to predict insurance charges. By leveraging the ensemble learning approach, these models can capture complex interactions and nonlinear relationships within the data, resulting in more precise risk assessments.

(ii) Claim prediction and fraud detection: Gradient boosting models can be employed to predict the likelihood of insurance claims and identify potential fraud cases. By training on historical claim data, these models can learn patterns and indicators that are associated with fraudulent activities. This helps insurance companies proactively flag suspicious claims and allocate resources efficiently to investigate them, reducing financial losses due to fraudulent activities.

(iii) Customer segmentation and personalized pricing: Gradient boosting models can segment customers into different risk profiles based on their characteristics and behaviors. This enables insurance companies to tailor their pricing and coverage offerings to specific customer groups. By accurately assessing risk levels for each segment, insurers can develop personalized pricing strategies that align with the individual needs of customers, improving customer satisfaction and retention.

(iv) Loss reserving and actuarial analysis: Gradient boosting models can assist in estimating loss reserves, a crucial aspect of insurance operations. By analyzing historical claim data and incorporating relevant predictors, these models can provide more accurate projections of future losses. This information helps insurers set aside

appropriate funds to cover potential claims, ensuring financial stability and effective risk management.

6. Model application

6.1. Choosing model

To perform a model for predicting insurance charges based on the provided dataset, we can use regression analysis. A multiple linear regression model is considered suitable for predicting the insurance charges in this context for the following reasons:

Numerical target variable: The target variable, "charges," is numerical, making it appropriate for regression analysis. Linear regression models are designed to predict numerical outcomes.

Multiple features: The dataset includes multiple independent variables such as "age," "bmi," "children," "smoker," and "region." Multiple linear regression can handle multiple predictors, capturing potential relationships between these variables and the target.

Linear assumption: The model assumes a linear relationship between the independent variables and the dependent variable. While this assumption may not hold perfectly in all cases, it is a reasonable starting point for exploration.

Interpretability: Linear regression models are interpretable and provide insights into the impact of each predictor on the target variable. This interpretability can be valuable for understanding the factors influencing insurance charges.

Ease of implementation: Multiple linear regression is relatively straightforward to implement and does not require complex configurations. It serves as a good baseline model before considering more advanced techniques.

6.2. Discuss the result

Table 7. Model result

	MSE	MAE	EVS	R2 score
Train Metrics	33665880.38401735	4026.880703453504	0.7535653882262375	0.7535653882262375
Test Metrics	34214506.06871899	4023.9335455060796	0.7494814875484863	0.7468367936503437

Source: Author

Mean Squared Error (MSE) and Mean Absolute Error (MAE): The high values of MSE and MAE on both the training and test sets, especially when the average value of the dependent variable is not too large, may indicate that the model is struggling to accurately predict the actual values and needs adjustment to minimize errors.

Explained Variance Score (EVS) and R2 Score: The high values of EVS and R2 Score (close to 0.75) on both the training and test sets indicate that the model explains about 75% of the variability in the data. While this is a good result, there is still around 25% of unexplained variability, suggesting potential room for improvement in the model.

Comparison between train and test metrics: When comparing the results between the training and test sets, it is observed that the metrics on the test set do not deviate significantly from those on the training set. This implies that the model does not exhibit overfitting or underfitting tendencies, in other words, the model has the ability to generalize well to new data.

In summary, the model shows some challenges in accurately predicting the actual values, and there is an opportunity for improvement to enhance its performance. The good generalization to new data, as indicated by the comparable metrics on the training and test sets, is a positive aspect that suggests the model has been appropriately fitted to the data.

6.3. Conclusions for the model

The linear regression model has been evaluated using performance metrics on both the training and test sets. While the model provides a significant level of explanation for the variability of the dependent variable, the prediction accuracy is still high as indicated by the Mean Squared Error (MSE) and Mean Absolute Error (MAE). The consistency between the datasets and the cumulative explanatory power suggests that the model has good generalization ability for new data. However, to improve performance, further considerations are needed regarding hyperparameter tuning, experimenting with new feature engineering techniques, and re-evaluating the impact of each variable. The results should also be placed in the context of the specific industry to determine whether the current level of accuracy meets the requirements or not.

6.4. Recommendations

While the model provides a significant level of explanation for the variability in the dependent variable, the accuracy, as reflected by Mean Squared Error (MSE) and Mean Absolute Error (MAE), remains high. Here are some suggestions to address this issue and improve the model:

Checking and handling outliers: Outliers can have a significant impact on the model. It is important to check and handle outlier data points to ensure the accuracy and reliability of the model.

Considering nonlinear relationships: Linear regression models assume a linear relationship between the input variables and the target variable. However, there may be nonlinear relationships between variables. Consider exploring nonlinear relationships by adding interaction terms or transforming existing variables.

Expanding the feature set: Sometimes, adding new variables can improve the model. Consider adding new variables that may have an impact on the target variable, such as health indices, income, or smoking information. However, careful consideration should be given to the potential cost of acquiring additional information.

Using variable selection methods: Apply variable selection methods such as t-statistic-based selection, F-statistic-based selection, or automated variable selection methods like LASSO to determine the most important variables for the model.

7. Conclusion

Through this project, individuals can observe the correlations between variables in the insurance dataset and how independent variables affect the dependent variable "charges". Specifically, there is a clear positive relationship between smoking status and age with charges, while the other variables have positive but less pronounced or almost no relationship with charges.

When working with this dataset, there are certain data points to consider, such as outliers and duplicate data. Resolving these issues can help improve the accuracy and effectiveness of the machine learning model. Furthermore, in terms of the machine learning model itself, it is evident that linear regression can explain a large portion of the dependent variable based on the independent variables. However, the accuracy of the model is not yet optimal and requires further improvements and it may not capture

the complexity of the data fully . Exploring alternative machine learning models such as decision trees, random forests, or neural networks could offer more sophisticated patterns and nonlinear relationships between variables, potentially leading to improved model performance.

In conclusion, this project highlights the correlations between variables in the insurance dataset and the impact of independent variables on the dependent variable "charges". By addressing data issues, exploring alternative models, and enhancing the feature set, analysts can strive for an accurate and effective machine learning model that better predicts insurance charges.

REFERENCES

Bonthu, H. (n.d.). *Detecting and Treating Outliers | Treating the odd one out!* From Analytics Vydhia: https://www.analyticsvidhya.com/blog/2021/05/detecting-and-treating-outliers-treating-the-odd-one-out/#2e43

Consulting, D. (2023). From https://bookdown.org/thomas_pernet/Tuto/methods-for-describing-a-set-of-data.html

Thomas, A. (2022). *How to Write a Problem Statement for an Open Data Project*. From data.world: https://data.world/blog/how-to-write-problem-statement-open-data-project/

UEL. (n.d.). *Khảo sát hình dạng phân phối của tập dữ liệu.* From https://maths.uel.edu.vn/Resources/Docs/SubDomain/maths/TaiLieuHocTap/ToanUngDung/kho_st_hnh_dng_phn_phi_ca_tp_d_liu.html

What are the best ways to handle duplicate records in your data? (n.d.). From Linkedin: https://www.linkedin.com/advice/0/what-best-ways-handle-duplicate-records-your-data#identify-duplicates