Programming for Data Analytics 1

Case study : A Prominent Medical Aid Scheme in South Africa

POE PART 1

Mukeba Mbunda Eliezer
ST10486340
PDDA0801 VCWCCR Term1 GRO1
Varsity College Cape Town Newlands

DUE: April 25, 2025

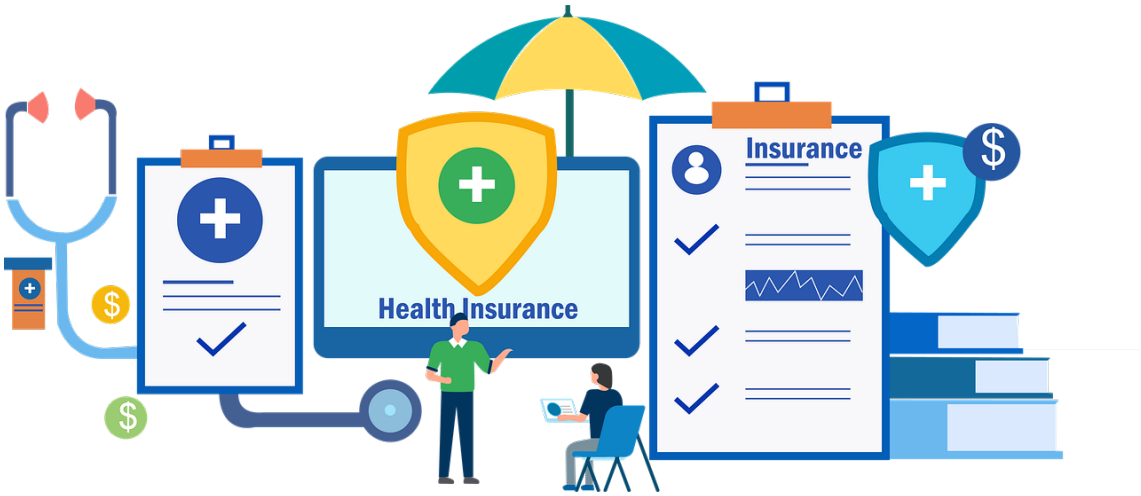
LECTURER NAME: Dr Rudolf Holzhausen

Listings

List of Figures

Contents

1 Introduction



We have been contacted by a prominent medical aid scheme in South Africa to create a Linear Regression model that can be useful to tailor their medical services according to the lifestyles and geographic regions of their customers. Before jumping into the process of building the model, we should remember that nowadays, medical insurance is in a stage of continuous market growth, and patients tend to pay more and more attention to their health. According to Kodiyan and Francis (2019), personal expenditure on medical aid has been increasing faster than the overall economy, leading to more to more pressure on poeple's budgets. High medical costs lead to the need to anticipate financial risks, both for individuals and insurance providers. Therefore, medical cost data analysis is necessary to predict future medical expenses.

In this context, using a linear regression model as requested would definitely be a good idea because it offers a strong foundation for making reliable predictions. Its simplicity and effectiveness make it one of the most widely used techniques in data analysis, especially when trying to understand how certain factors influence an outcome(dataquest 2026). By modeling the relationship between variables, linear regression can help us generate accurate cost predictions, which is incredibly valuable in the healthcare industry. These predictions allow decision-makers to plan ahead, manage resources more efficiently, and ultimately provide better

services to patients.

This report aims to provide detailed findings on how we applied the linear regression model. We begin by explaining how we evaluated the dataset and conducted exploratory data analysis (EDA) to understand the patterns within the data. We also cover other important aspects necessary for training the model. Finally, we explain how the model was trained, interpreted, and evaluated.

2 Evaluate Dataset

In this section, we used the `pandas` library to explore and evaluate the dataset. `Pandas` is a powerful Python library commonly used for data manipulation and analysis. It provides data structures and functions that make it easy to clean, analyze, and visualize structured data(Wschool 2019).

2.1 Exploring dataset

We use here the medical cost personal dataset from Kaggle.

Column	Description
age	Age of the primary beneficiary
sex	Gender of the insurance contractor (female or male)
bmi	Body Mass Index, providing an estimate of body fat based on height and weight
children	Number of children covered by health insurance (number of dependents)
smoker	Indicates whether the person is a smoker
region	Residential area of the beneficiary in the US (northeast, southeast, southwest, northwest)
charges	Annual medical insurance charges billed to the individual

Table 1: Description of the Medical Cost Personal Dataset

To gain an overview of the dataset, we printed the first 20 rows to examine its structure and contents as shown in Figure 1.

[426]:

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.900	0	yes	southwest	16,884.924
1	18	male	33.770	1	no	southeast	1,725.552
2	28	male	33.000	3	no	southeast	4,449.462
3	33	male	22.705	0	no	northwest	21,984.471
4	32	male	28.880	0	no	northwest	3,866.855
5	31	female	25.740	0	no	southeast	3,756.622
6	46	female	33.440	1	no	southeast	8,240.590
7	37	female	27.740	3	no	northwest	7,281.506
8	37	male	29.830	2	no	northeast	6,406.411
9	60	female	25.840	0	no	northwest	28,923.137
10	25	male	26.220	0	no	northeast	2,721.321
11	62	female	26.290	0	yes	southeast	27,808.725
12	23	male	34.400	0	no	southwest	1,826.843
13	56	female	39.820	0	no	southeast	11,090.718
14	27	male	42.130	0	yes	southeast	39,611.758
15	19	male	24.600	1	no	southwest	1,837.237
16	52	female	30.780	1	no	northeast	10,797.336
17	23	male	23.845	0	no	northeast	2,395.172
18	56	male	40.300	0	no	southwest	10,602.385
19	30	male	35.300	0	yes	southwest	36,837.467

Figure 1: First 20 rows in from the dataset

We got the number of rows and columns as shown in Figure 2 , which are the dimensions, and we also obtained the summary information of the dataset((data types, non-null counts, column names, and memory usage), as shown in Figure 3.

```
[436]: (1338, 7)
```

Figure 2: Dimensions

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1338 non-null   int64
 1   sex       1338 non-null   object
 2   bmi       1338 non-null   float64
 3   children  1338 non-null   int64
 4   smoker    1338 non-null   object
 5   region    1338 non-null   object
 6   charges   1338 non-null   float64
dtypes: float64(2), int64(2), object(3)
memory usage: 73.3+ KB
```

Figure 3: Summary Information of the Dataset

We got as well a statistical summury of the dataset, The statistical summary provides key descriptive statistics for the numerical columns in the dataset. These statistics help us understand the distribution and spread of the data, including the total number of entries (count), the average value (mean), the variation in the data (standard deviation), the lowest and highest values (min and max), and the quartiles (25th, 50th, and 75th percentiles), which give insight into the data's spread and central tendency(Cuemath 2025).

[446]:

	age	bmi	children	charges
count	1,338.000	1,338.000	1,338.000	1,338.000
mean	39.207	30.663	1.095	13,270.422
std	14.050	6.098	1.205	12,110.011
min	18.000	15.960	0.000	1,121.874
25%	27.000	26.296	0.000	4,740.287
50%	39.000	30.400	1.000	9,382.033
75%	51.000	34.694	2.000	16,639.913
max	64.000	53.130	5.000	63,770.428

Figure 4: Statistical summary

2.2 Cleaning dataset

We started the cleaning process by checking for missing values as shown in Figure 5 and identifying any duplicated rows in the data sett as shown in Figure 6,we finally obtained the new dimension as shown in Figure 7.

[450]:
```
age         0
sex         0
bmi         0
children    0
smoker      0
region      0
charges     0
dtype: int64
```

Figure 5: Missing values

[459]:

	age	sex	bmi	children	smoker	region	charges
195	19	male	30.590	0	no	northwest	1,639.563
581	19	male	30.590	0	no	northwest	1,639.563

Figure 6: Duplicate values

```
[470]: (1337, 7)
```

Figure 7: New dimensions

3 Exploratory Data Analysis

According to Geeksforgeeks (2025) , Exploratory Data Analysis (EDA) is an important first step in data science projects. It involves looking at and visualizing data to understand its main features, find patterns, and discover how different parts of the data are connected. In this section, we analyzed the dataset by exploring three types of relationships: univariate, bivariate, and multivariate. We used the `matplotlib`, `seaborn`, and `plotly` libraries for data visualization.

3.1 Univariate analysis

We performed a univariate analysis, which involves exploring one variable at a time to understand its effect within the dataset.(Rzayev 2024).

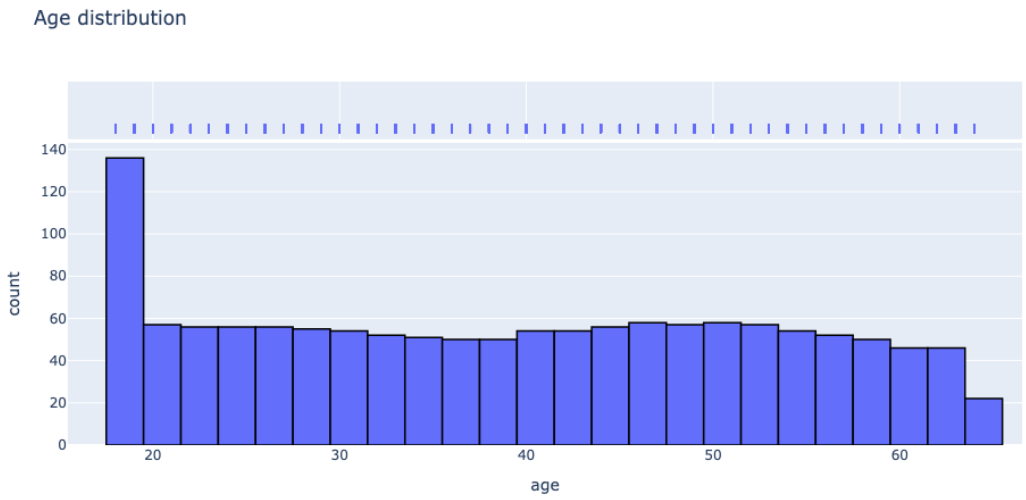


Figure 8: Age distribution

According to Figure 8, most patients in the dataset are between 18 and 19 years old (136 people), with 22 patients between 64 and 65 years old.

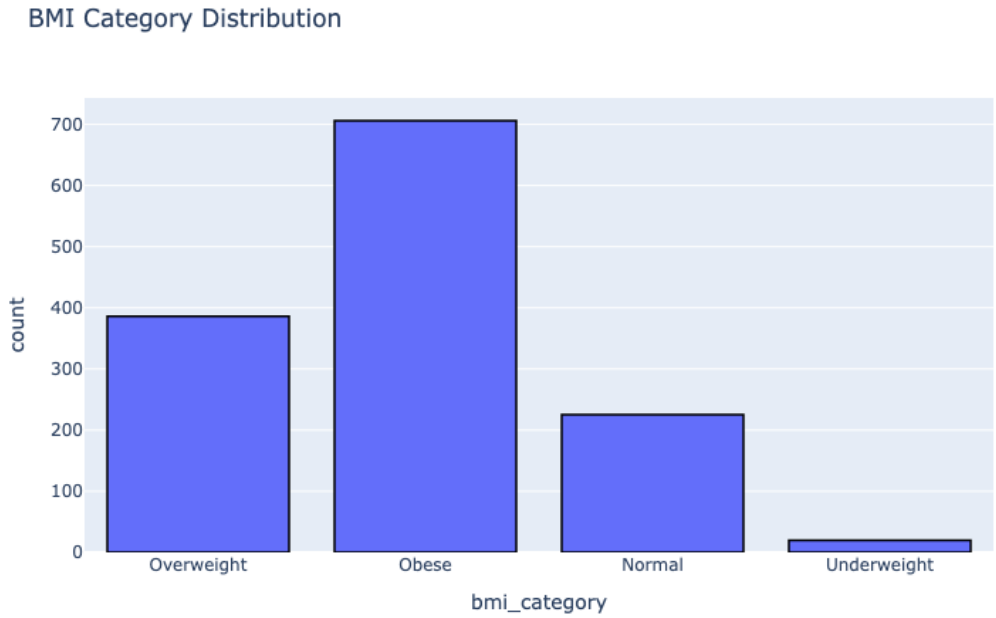


Figure 9: BMI distribution by category

According to Figure 9, most patients in the dataset are obese, while fewer are underweight.

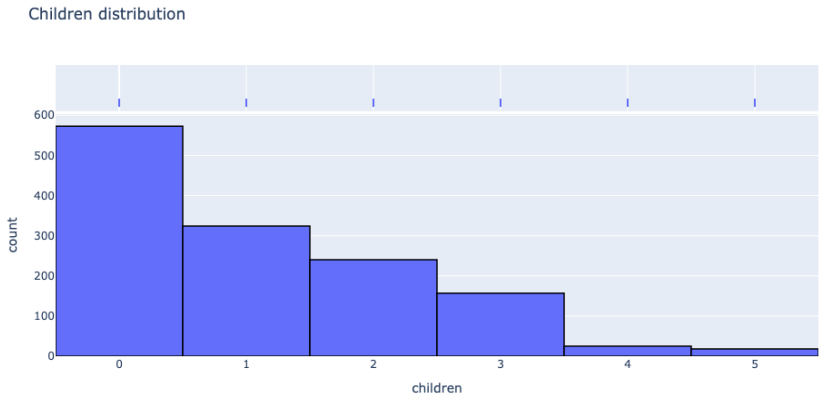


Figure 10: Children distribution

According to Figure 10, most patients in the dataset do not have any children (573 people), and few have more children (18 people).

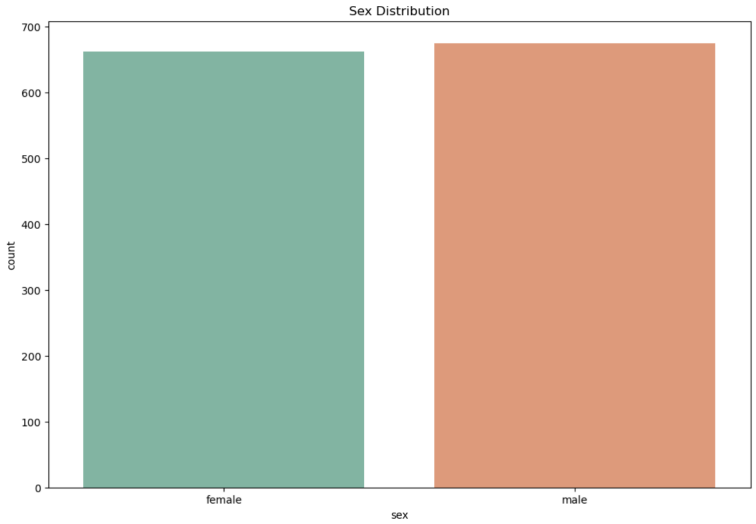


Figure 11: Sex distribution

According to Figure 11, the number of males and females in the dataset is quite balanced, but males are more than females

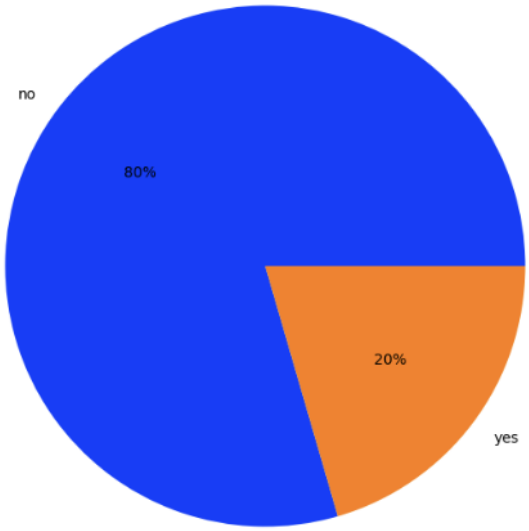


Figure 12: Smoker distribution

According to Figure 12, 80%/ of patients are not smokers, while 20%/ are smokers.

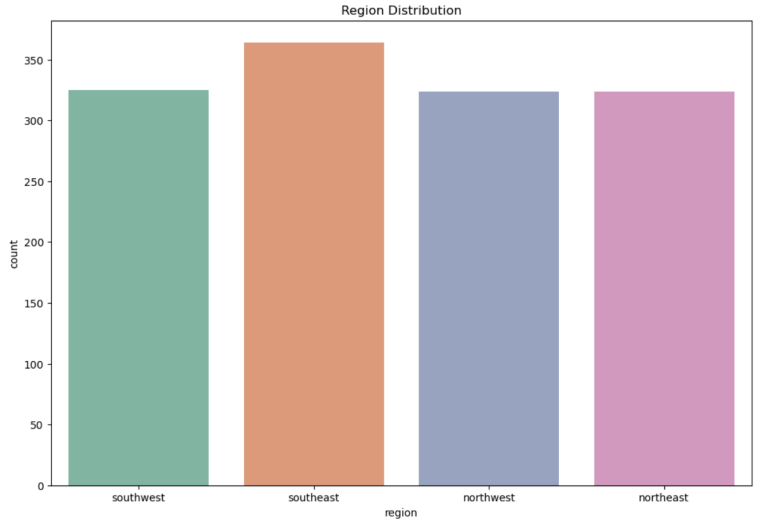


Figure 13: Region distribution

According to Figure 13, the number of patients from the northeast is slightly higher than in other regions, but the counts from the other regions are nearly equal.

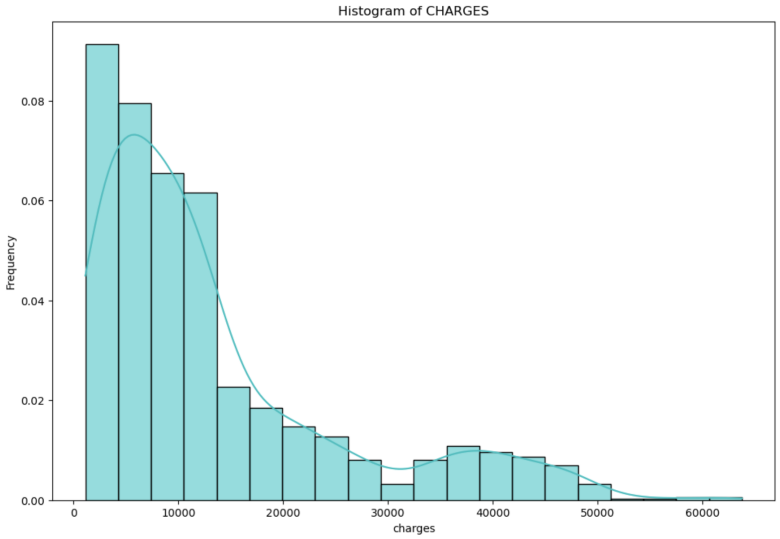


Figure 14: Region distribution

According to Figure 14, We can see that most medical expenses are below 10,000, and only a small number of patients have expenses exceeding 50,000. This is the target variable, and we can observe that the distribution is skewed to the right, meaning that more patients have higher charges in medical insurance.

3.2 Bivariate analysis

Bivariate analysis examines the relationship between two variables and determines if they are dependent on each other(Rzayev 2024). we wanted to answer this question : Is there a relationship between two variables, and if so, what is its nature? , We analyzed each feature in relation to the target variable, using the smoker column to color-code for better understanding.

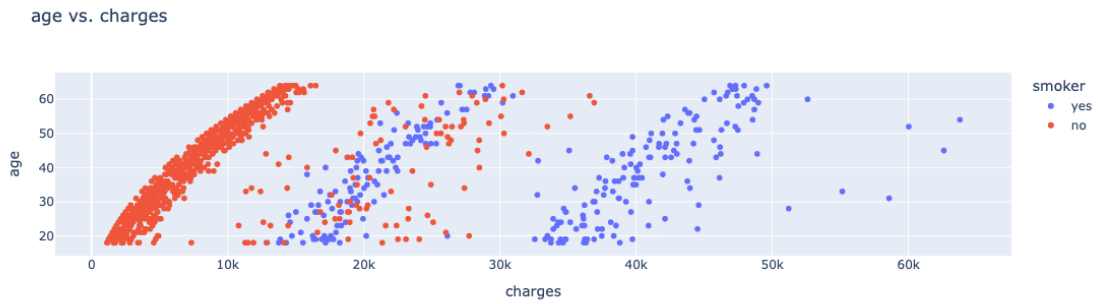


Figure 15: Age vs. Charges distribution

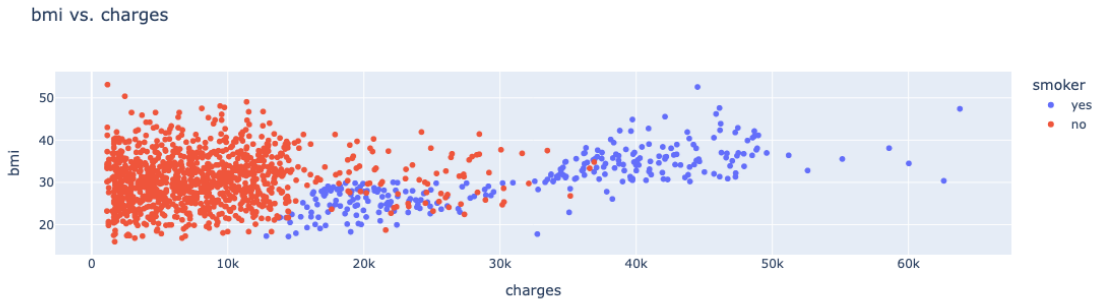


Figure 16: BMI vs. Charges distribution

According to Figure 15, there is some moderate positive relationship between age and charges, we can see that as age increases the charges also tend to increase. However, this relationship is not linear. We observed that some younger patients have higher charges due to other influencing factors, such as smoking status in this case. Additionally, Figure 16 shows that there is no strong relationship between bmi and charges.

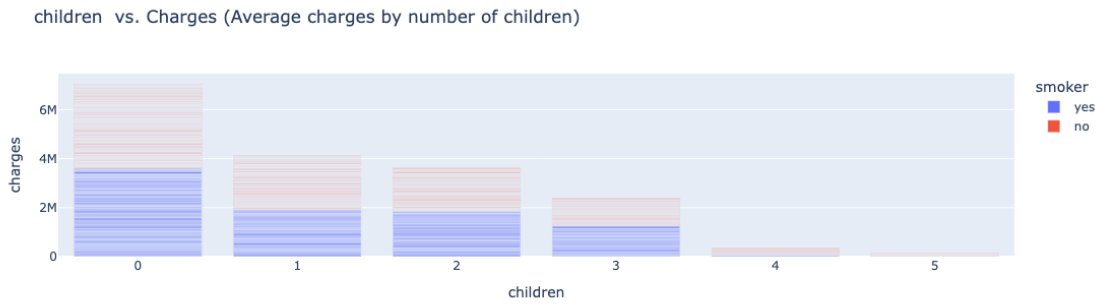


Figure 17: Children vs. Charges distribution

According to Figure 17, Patients with fewer children tend to have higher charges compared to those with more children. However, we should not hastily conclude that there is a strong correlation.

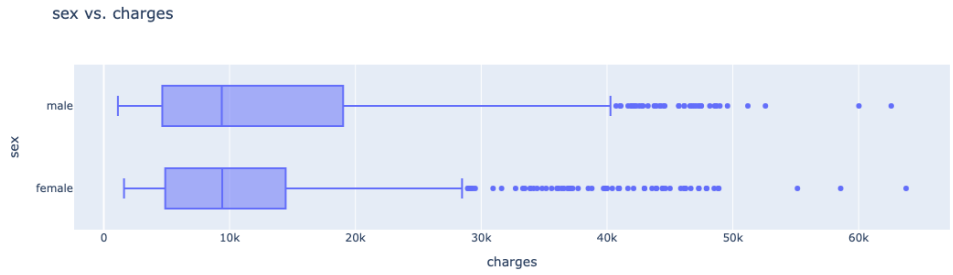


Figure 18: Sex vs. Charges distribution

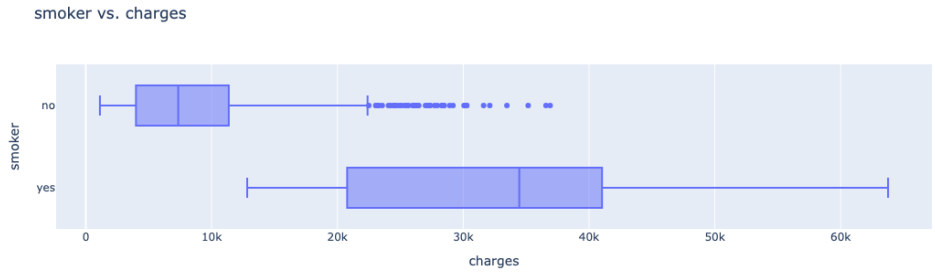


Figure 19: Smoker vs. Charges distribution

According to Figure 18 on average, females have higher charges than males. However, males tend to have a greater variability than females due to the wider box reflecting a range of lower to higher changes. This is not definitive as the result might be infulenced by others factors . However We notice that here smokers tend to have more charges than non-smokers, the average and the variability of smokers are higher. This allows us to say there is a strong relationship between smokers and charges as shown in Figure 17.

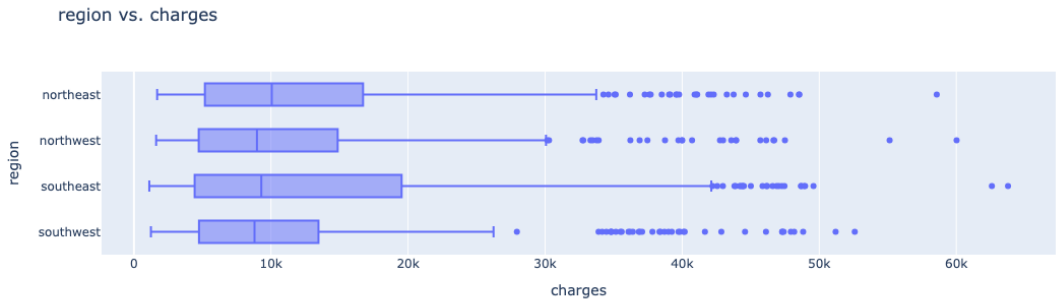


Figure 20: region vs. Charges distribution

According to Figure 20 , We can see that the Southeast region tends to have a stronger relationship with charges compared to the other regions.

3.3 Multivariate analysis

Multivariate analysis is similar to bivariate analysis, but instead of two, more than two variables are analyzed at once. For three variables, we can create a 3D model to study the relationship (also known as Trivariate Analysis)(Rzayev 2024).

We started by plotting the pairplot to explore the relationships between the numeric variables: age, bmi, children, and charges. The pairplot displays scatter plots and distributions, allowing us to visually assess potential correlations, patterns, and outliers within the dataset, and better understand how these variables interact with one another.

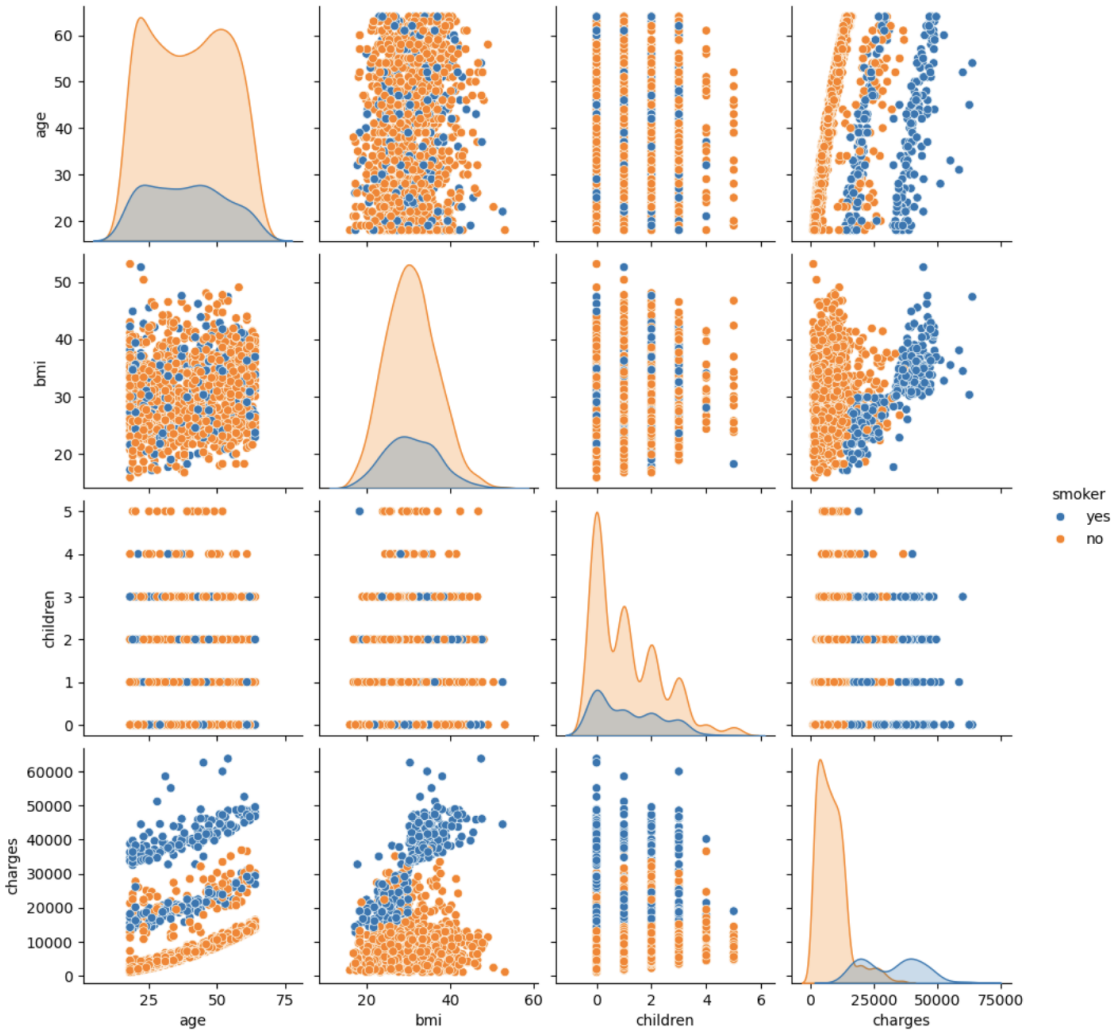


Figure 21: Pairplot distribution

According to Figure 22 shown above , it is clear that in the southeast region, more male patients tend to spend on charges than in other regions, However the 3D distribution helps us to we visualizes the relationship between charges, age, BMI, and number of children. Charges tend to increase with age and BMI, especially for older patients with higher BMI. The color dimension, representing the number of children, shows that patients with more children are scattered across different charge levels but are more frequent among middle-aged. However, there doesn't appear to be a strong pattern suggesting that the number of children alone significantly influences charges, as similar charge levels occur across different child counts. Instead, age and BMI remain the dominant factors influencing higher insurance costs, with the number of children adding some variation but not a

definitive trend.

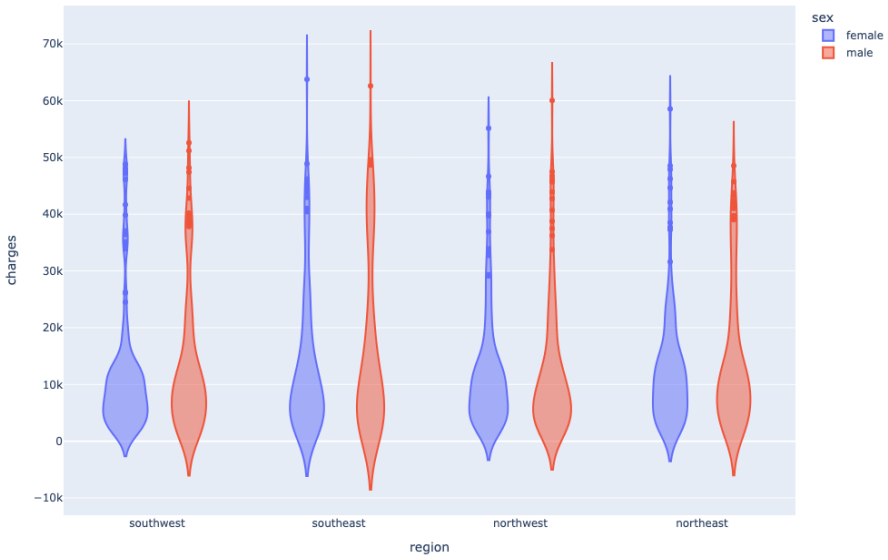


Figure 22: Violin distribution

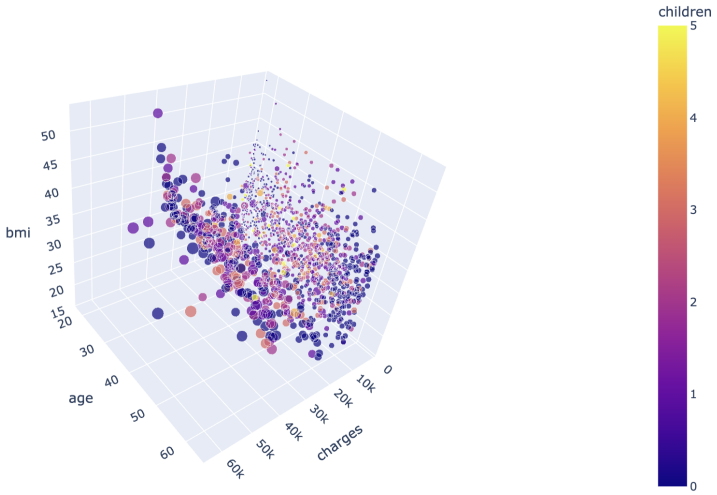


Figure 23: 3D Visualisation

Before proceeding with the correlation matrix, we start by converting categorical features such as sex, smoker, and region.

The figure above shows the relationship between each column in the dataset. However, we are going to stay focused on the relationship between features and the target variable, charges. We notice that there is a strong correlation between smoker and charges.

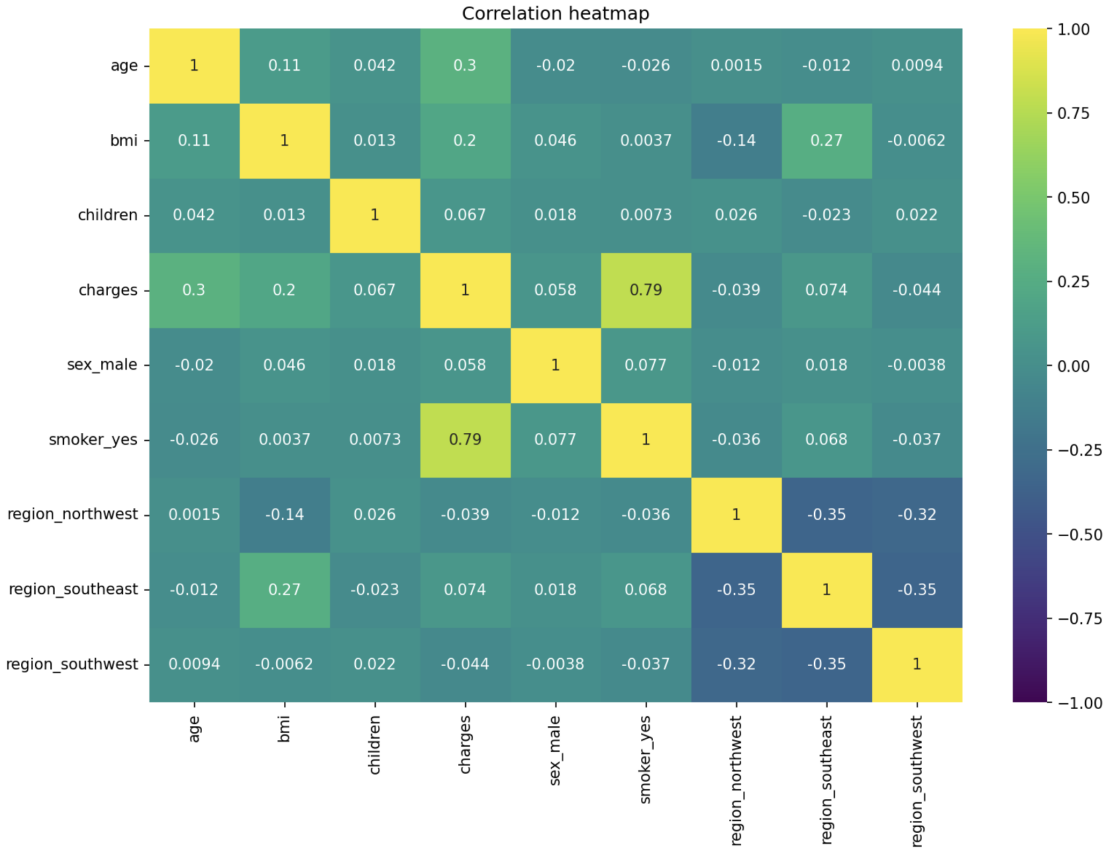


Figure 24: Correlation heatmap

4 Model training : Linear Regression

4.1 Feature selection

Before creating the model, we needed to select the best features that would be useful for training the model. We used two methods to do this:

1. Correlation Matrix Heatmap:

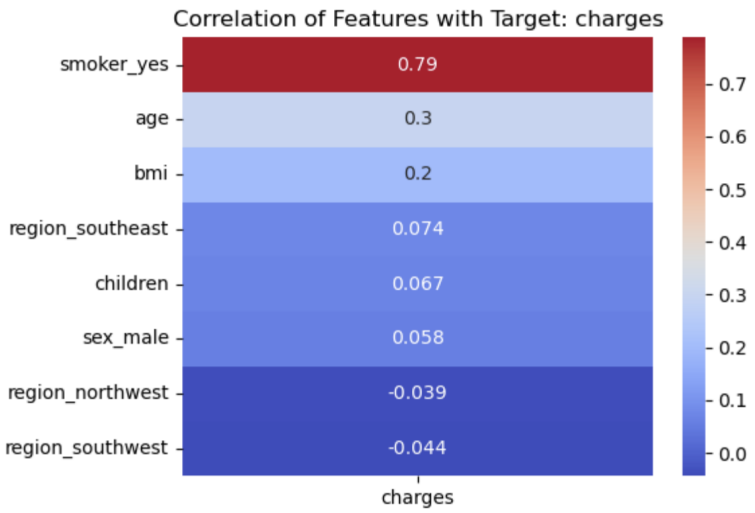
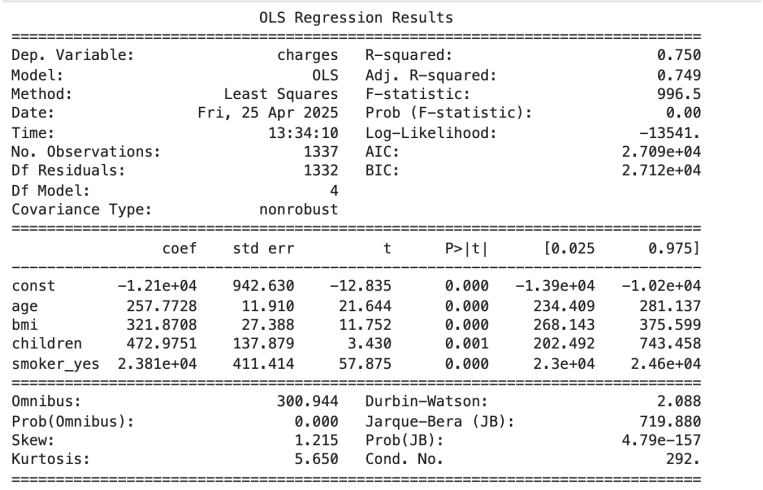


Figure 25: Correlation Matrix Heatmap

As shown Figure 25 , This method in this case helped us to identify how strongly each feature was linearly related to the target variable charges. Features with higher absolute correlation values were likely to be better predictors (Simplilearn 2023). Based on the results obtained, we selected the following features:

- smoker_yes
- age
- bmi
- region_southeast

2 Backward Elimination (via OLS process):



OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	996.5
Date:	Fri, 25 Apr 2025	Prob (F-statistic):	0.00
Time:	13:34:10	Log-Likelihood:	-13541.
No. Observations:	1337	AIC:	2.709e+04
Df Residuals:	1332	BIC:	2.712e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.21e+04	942.630	-12.835	0.000	-1.39e+04	-1.02e+04
age	257.7728	11.910	21.644	0.000	234.409	281.137
bmi	321.8708	27.388	11.752	0.000	268.143	375.599
children	472.9751	137.879	3.430	0.001	202.492	743.458
smoker_yes	2.381e+04	411.414	57.875	0.000	2.3e+04	2.46e+04

Omnibus:	300.944	Durbin-Watson:	2.088
Prob(Omnibus):	0.000	Jarque-Bera (JB):	719.880
Skew:	1.215	Prob(JB):	4.79e-157
Kurtosis:	5.650	Cond. No.	292.

Figure 26: Backward Elimination (via OLS process)

This technique is used to iteratively remove features with the highest p-values, which indicate weaker statistical significance (daython 20243). In this context, we were not focusing on building an OLS model itself, but rather using the OLS process as a tool to help eliminate features that are not statistically significant for predicting charges. we selected the following features As shown Figure 26:

- smoker_yes
- age
- bmi
- children

4.2 Training Model

We trained a linear regression model using all the features, as well as using the selected features from both methods we applied (Correlation Matrix Heatmap,Backward Elimination). This helped us how the model's performance was affected. After splitting the dataset into training data and testing data , we trained the model.

We started with a linear regression model using all the features, then we trained

a linear regression model using the features selected by the Correlation Matrix Heatmap method, and finally, we trained a linear regression model using the features selected by the Backward Elimination method.

```
# Split the dataset into training and testing data
X_train, X_test,y_train, y_test = train_test_split(X, y,test_size
    =0.2, random_state=42)
# create the linearRegresson model
reg=LinearRegression()
# fit the model
reg.fit(X_train,y_train)
# predict new values
y_pred=reg.predict(X_test)
# create a list with y_test and y_pred
features_linear=[y_test, y_pred]
# coefficient of determination
r2_linear=r2_score(y_test, y_pred)
# mean absolute error
mae_linear=mean_absolute_error(y_test, y_pred)
# mean squared error
mse_linear=mean_squared_error(y_test, y_pred)
rmse_linear= np.sqrt(mean_squared_error(y_test, y_pred))
# root mean squared error
```

Listing 1: Linear Regression model

5 Model Evaluation and Interpretation

Before interpreting the results, we demonstrated that removing unnecessary columns, as in feature selection(using Correlation Matrix Heatmap and Backward Elimination method , did not have a significant impact on the model's perfomance.

We rounded the values and saw that the results were almost the same the same.

	Linear Regression(All features)	Linear Regression(selected features method1)	Linear Regression(selected features method2)
R^2	0.807	0.806	0.805
MAE	4,177.046	4,183.602	4,198.593
MSE	35,478,020.675	35,683,805.517	35,914,551.480
RMSE	5,956.343	5,973.592	5,992.875

Figure 27: Models performances

We stayed focused on the Linear Regression model trained with all the features for evaluation and interpretation.

```
R²              0.81
MAE         4,177.05
MSE    35,478,020.68
RMSE        5,956.34
Name: Linear Regression(All features), dtype: float64
```

Figure 28: Linear Regression performance

We used four metrics to evaluate and interpret the model's performance:

R-Squared (R^2): 81% of the variance in the charges (target variable) is explained by the features in the model (*age, sex, bmi, smoker, and region*).

Mean Absolute Error (MAE): Higher average error between the predicted charges and the actual charges.

Mean Squared Error (MSE): Higher average squared error, which increases the risk of outliers influencing the results more strongly.

Root Mean Squared Error (RMSE): Higher average deviation between the predicted charges and the actual charges.

- **R-Squared (R^2):** This metric shows the proportion of variance in the target variable explained by the model. An R^2 closer to 1 indicates a better fit (Durga 2024).
- **Mean Absolute Error (MAE):** This metric measures the average magnitude of errors in predictions, without considering their direction. It is calculated as the average of the absolute differences between predicted and actual

values (Farshadk 2024).

- **Mean Squared Error (MSE):** This metric measures the average of the squared differences between predicted and actual values. It penalizes larger errors more severely, making it sensitive to outliers (Durga 2024).
- **Root Mean Squared Error (RMSE):** This metric is the square root of the MSE, bringing it back to the same units as the target variable. It provides an easily interpretable measure of average error size (Farshadk 2024).

Our model has performed well, achieving an R^2 score of 0.81. However, further improvements may still be possible.

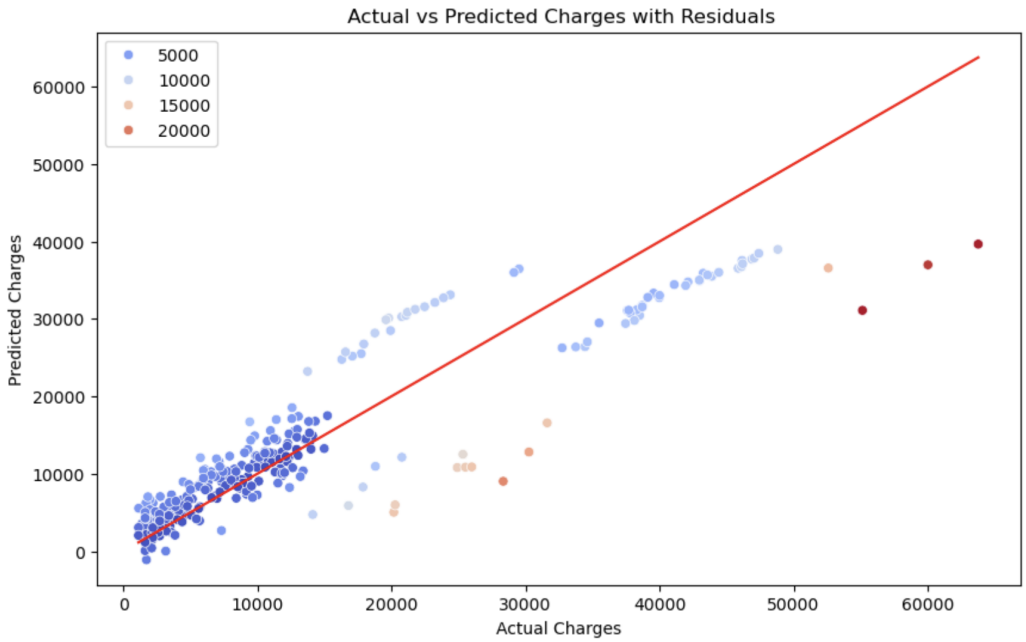


Figure 29: Actual VS predicted charges with residuals

we can clearly see how well our model fits the data, the plot shows that most of the points are close to the line, indicating a good fit.

6 Retrain model

As the target in the dataset was skewed to the right, this could affect the model's performance because the distribution was not normalized. To address this, we normalized the distribution and retrained the model. This may help improve the evaluation metrics; however, in real-life scenarios, such preprocessing might not always reflect the true nature of the data.

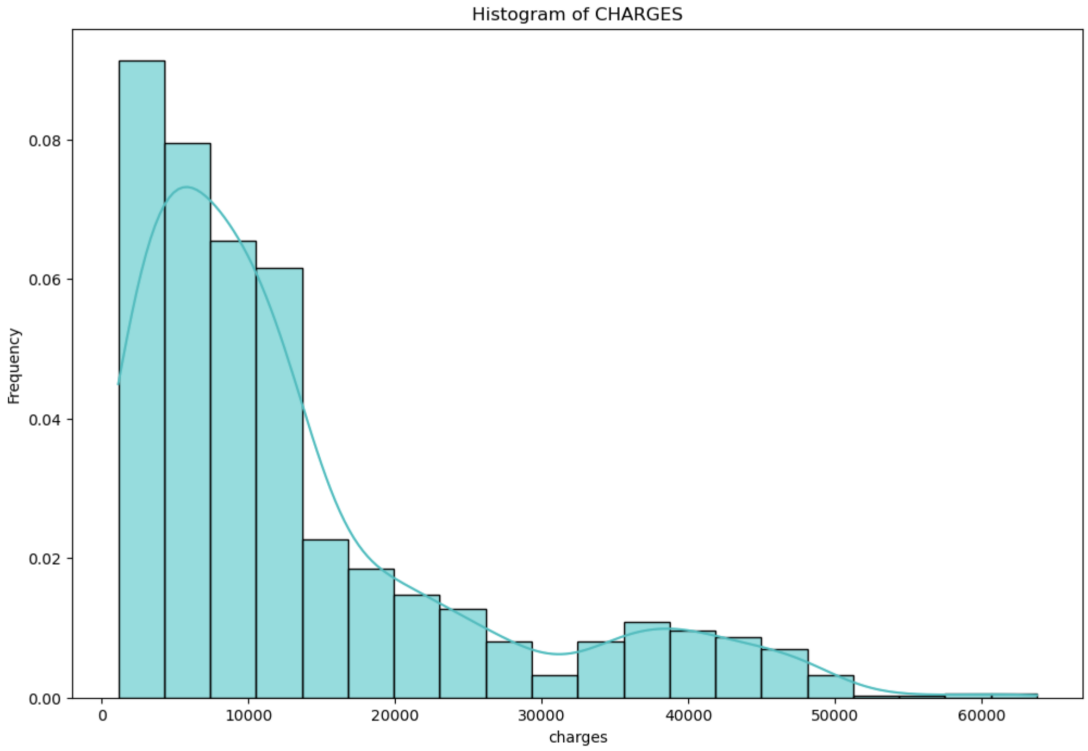


Figure 30: Skewed-right distribution

Both figures show the distribution of charges. The first one (Figure 30) is skewed to the right, just like the original dataset. The second one, however, is more normally distributed after applying a log transformation.

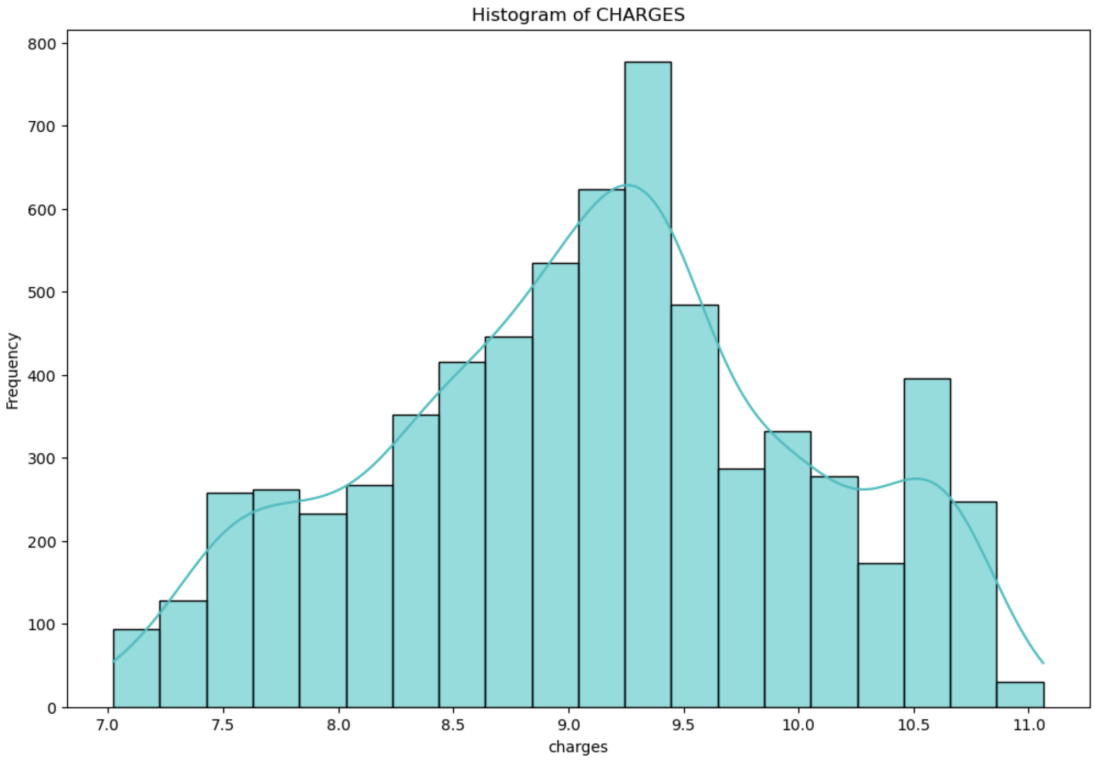


Figure 31: Normal distribution

We then retrained the model using the transformed data and achieved the following score:

626]:

	Linear Regression retrained
R^2	0.83
MAE	0.26
MSE	0.16
RMSE	0.40

Figure 32: model's performances

The model has performed very well, achieving an R^2 score of 83%. The average MAE is very low, indicating minimal errors, along with similarly low MASE and RMSE values, confirming the model's strong performance

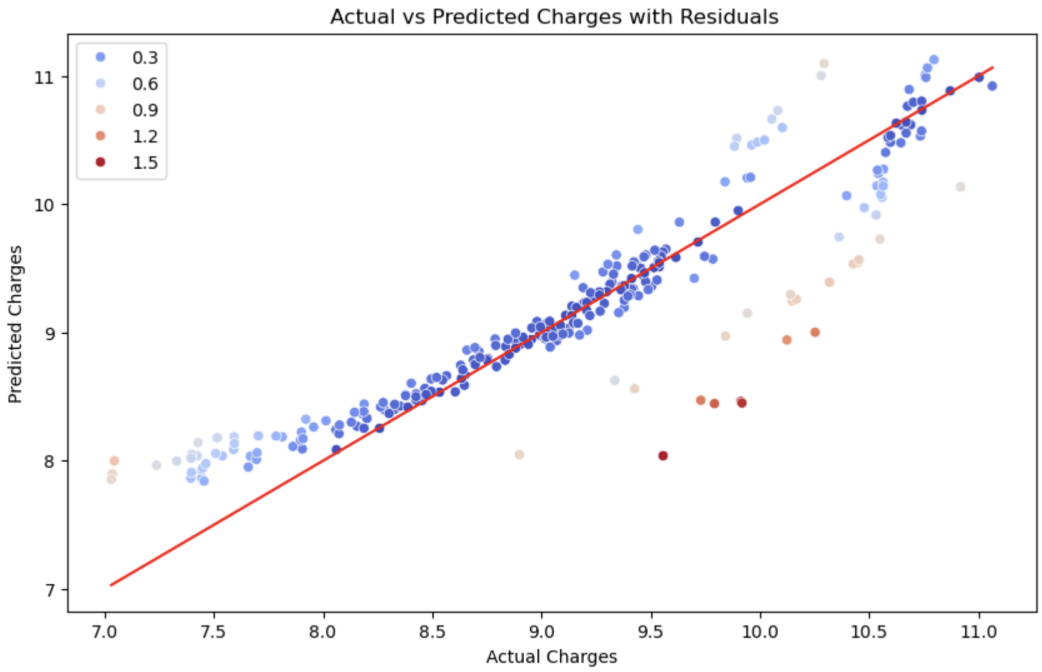


Figure 33: Actual vs. Predicted charges with Residals

7 Bonus

We trained several regression models and obtained the following results

[684]:

	Decision Tree	Random Forest	Linear Regression	Lasso Regression	Ridge Regression	Elastic Net
R²	0.90	0.88	0.81	0.81	0.81	0.80
MAE	2,621.31	2,630.54	4,177.05	4,178.75	4,178.75	4,224.74
MSE	18,886,631.25	22,703,695.13	35,478,020.68	35,495,705.71	35,495,705.71	36,182,293.89
RMSE	4,345.88	4,764.84	5,956.34	5,957.83	5,957.83	6,015.17

Figure 34: Regression Model Performances

References

Cuemath (2025). *Summary statistics*. Accessed: 22 April 2025. URL: https://www.cuemath.com/data/summary-statistics/.

dataquest (2026). *Using Linear Regression for Predictive Modeling in R*. Accessed: 21 April 2025. URL: https://www.dataquest.io/blog/statistical-learning-for-predictive-modeling-r/.

daython (20243). *Mastering the Art of Feature Selection: Python Techniques for Visualizing Feature Importance*. Accessed: 21 April 2025. URL: https://medium.com/@nihat.rzayev1357/univariate-bivariate-and-multivariate-analysis-f01bd339c825.

Durga (2024). *Model Evaluation and Interpretation for Linear Regression*. Accessed: 21 April 2025. URL: https://medium.com/itversity/model-evaluation-and-interpretation-for-linear-regression-fba70e66fe53.

Farshadk (2024). *Essential Regression Evaluation Metrics: MSE, RMSE, MAE, R^2, and Adjusted R^2*. Accessed: 21 April 2025. URL: https://farshadabdulazeez.medium.com/essential-regression-evaluation-metrics-mse-rmse-mae-r%C2%B2-and-adjusted-r%C2%B2-0600daa1c03a.

Geeksforgeeks (2025). *What is Exploratory Data Analysis?* Accessed: 22 April 2025. URL: https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/.

Kodiyan, Akhil Alfons and Kirthy Francis (Dec. 2019). *Linear regression model for predicting medical expenses based on insurance data*.

Rzayev, Nihat (2024). *Univariate, Bivariate, and Multivariate Analysis*. Accessed: 23 April 2025. URL: https://medium.com/@nihat.rzayev1357/univariate-bivariate-and-multivariate-analysis-f01bd339c825.

Simplilearn (2023). *What is backward elimination technique in machine learning?* Accessed: 21 April 2025. URL: ttps://www.simplilearn.com/what-is-backward-elimination-technique-in-machine-learning-article.

Wschool (2019). *Pandas Introduction*. Accessed: 22 April 2025. URL: https://www.w3schools.com/python/pandas/pandas_intro.asp.