Team N.14

Medical Cost Predictor For Insurance Providers.

Team Members:

Name	Section	Bench Number
Mohamed Gamal	1	48
Aya Salah	1	12
Reem Adel	2	1
Yassmen Sayed	2	46
Sara Mahmoud Hussein	1	28

Table of contents

1 Introduction

1.1 The importance of prediction algorithms in healthcare 4

1.2 How our prediction algorithm works 4

2 Methods

2.1 Importing data and libraries 5

2.2 Data handling 5

2.3 Describing data 5

2.4 Standardizing the features 6

2.5 Normality for each feature 6

2.6 Calculating linear regression 7

2.6.1 Simple linear regression 7

2.6.2 Multiple linear regression 7

3 Results and Discussion

3.1 Data handling 8

3.1.1 Duplicated data 8

3.1.2 Outliers 8

3.2 Describing data 9

3.3 Standardizing the features 9

3.4 Feature distributions 10

3.5 Normality test 13

3.6 Correlation coefficient 14

3.7 Linear regression 14

3.8 Multivariable linear regression 14

4 Conclusion 15

5 Member Contribution 16

Table of figures

3.1.2.1 Data before removing outliers 8

3.1.2.2 Data after removing outliers 8

3.3.1 Part of data after standardization 9

3.4.1 Smoker distribution 10

3.4.2 Region distribution 10

3.4.3 Number of children distribution 11

3.4.4 Gender distribution 11

3.4.5 BMI distribution 12

3.4.6 Age distribution 12

3.4.7 Charges distribution 13

3.6.1 Correlation coefficient results 14

3.7.1 Linear regression from scratch 14

1 Introduction

1.1 The importance of prediction algorithms in healthcare

The global health insurance market size was valued at USD 1,966.6 billion in 2020 and will grow from USD 2,088.5 billion in 2021 to USD 3,038.6 billion in 2028, it's one of the biggest markets to exist.

Predicting the medical expenses of potential clients can aid insurance providers, policymakers, and individuals in making knowledgeable decisions. Medical costs can have a large impact on individuals, families, and healthcare systems. Understanding the factors driving these costs is important for effective financial planning and resource planning.

Prediction algorithms like ours can greatly help insurance providers predict the probable medical cost for a client that they will have to oblige, using a set of factors like the number of children in the household, the age of the client, or their bmi (body mass index). The insights gained from this analysis can help with providing personalized insurance plans, informed risk assessment, and efficient resource allocation in the healthcare sector.

1.2 How our prediction algorithm works

Our prediction algorithm uses linear regression, to figure out the correlation between each factor like (gender, region, age) to figure out the probable medical cost for the client.

We adjusted our prediction algorithm model to provide high accuracy by the usage of a medical cost forecast dataset with over 1,300 different inputs of the age, gender, bmi, smoking status, the number of children, the region and of course the medical charges to reach an accurate prediction.

The findings from this study have the potential to contribute to personalized healthcare planning, knowledgeable decision-making by insurance providers, and improved cost-control strategies within the healthcare industry.

2 Methods

2.1 Importing data & libraries

```
import warnings

warnings.filterwarnings('ignore')

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from sklearn.metrics import mean_squared_error, r2_score

from math import sqrt

from scipy.stats import probplot

from scipy import stats

from sklearn.linear_model import LinearRegression

from scipy.stats import shapiro
```

2.2 Data handling

After importing data we made sure that there are no missing values then we divided the features into categorical [sex-smoker-region] and numerical [age-bmi-charges-children].

we found that there were duplicated rows of data so we removed it. We also found outliers in bmi and charges so we handled this using the **z-score**.

After removing outliers we made sure there are no missing values.

2.3 Describing data

We used a method called describe() to calculate a set of statistics to describe the data [count-mean-std-min-25%-50%-75%-max-mode-variance].

2.4 Standardizing the features

Using our calculations for the descriptive statistics we standardized the features (bmi , charges , age) that are continuous using the z score method .

$$z = \frac{x - \mu}{\sigma}$$

$$\mu = \text{Mean}$$
$$\sigma = \text{Standard Deviation}$$

2.5 Normality for each feature

For each feature we plotted our data after standardizing the feature then we observed the type of each one of them [Gaussian, exponential, uniform, etc].

then we test the normality of of each feature using a Shapiro-Wilk test that return the p value

and we compared it with the alpha value = .05

Ho : null hypothesis (our feature is normally distributed)

Ha :alternative hypothesis (our feature is not normally distributed).

p_value < alpha

Reject the null hypothesis: The data does not follow a normal distribution

p_value > alpha

Fail to reject the null hypothesis: The data may follow a normal distribution

we mapped the categorical data :

"female":0, "male":1

"yes":1, "no":0

"southwest":0, "southeast":1, "northwest":2, "northeast":3

Then we calculated the correlation coefficient for each feature with our response feature [charges].

2.6 Calculating linear regression

2.6.1 Simple linear regression

we implemented our linear regression between our target value [charges] and each feature one by one . We calculate rmse(root mean square error) and r2_score (also known as the coefficient of determination, is a statistical measure used to evaluate the goodness-of-fit of a regression model).

$$Y = mx+b \qquad (m : \text{coefficient} , b : \text{bias})$$

Our linear regression model calculates the coefficient of each feature and its bias.

Simple linear regression is implemented from scratch and using python package each is implemented in 2 methods.

2.6.2 Multiple linear regression

We build our multivariable linear regression using LinearRegression() from the sklearn.linear_model module.

The model is trained on the predictor variables x and the response variable y using the fit() method of the linear regression model.

The trained model is used to make predictions on the predictor variables x and assign them to the variable Y_pred.

then to assess the multivariable regression quality we calculated

- The root mean squared error (RMSE) is calculated using the mean_squared_error() function from sklearn.metrics by comparing the actual response variable y with the predicted values Y_pred.
- The R-squared (R2) score is calculated using the score() method of the linear regression model, which returns the coefficient of determination indicating the quality of the model's fit to the data.
- The intercept and coefficients of the linear regression model are printed using model.intercept_ and model.coef_, respectively.

The code outputs the RMSE, R2 score, intercept, and coefficients of the linear regression model. These metrics provide information about the accuracy and quality of the model in predicting the "charges" variable based on the given predictor variables.

3 Results and Discussion

3.1 Data handling

3.1.1 Duplicated data

we found 1 duplicated row of data (row 581) and we removed it .

3.1.2 Outliers

We plotted our data and found outliers in bmi and charges.

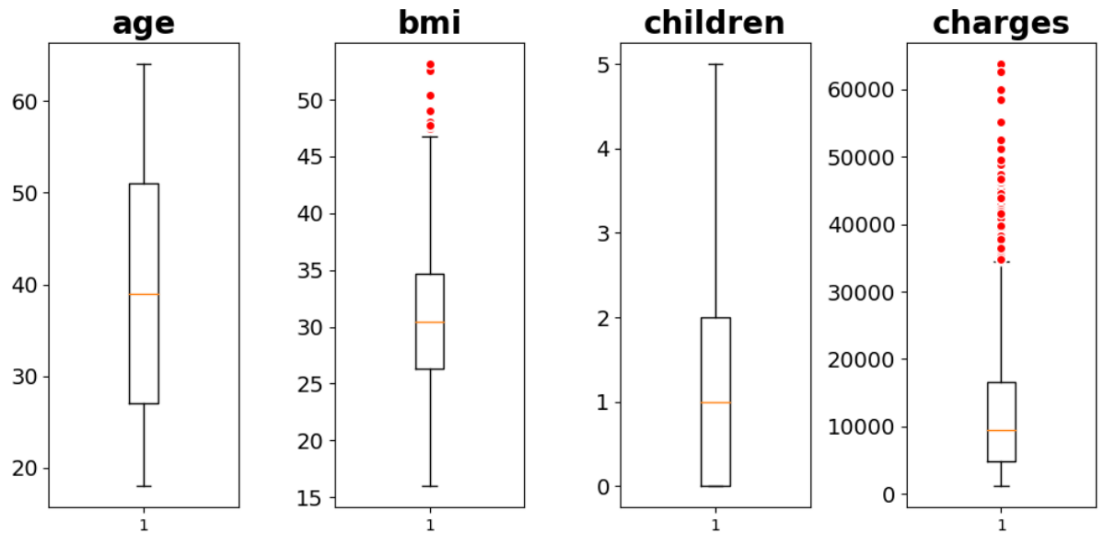


fig 3.1.2.1 data before removing outliers

We used z-score to remove the outliers . we choose the z threshold to be 2 because this fits our data.

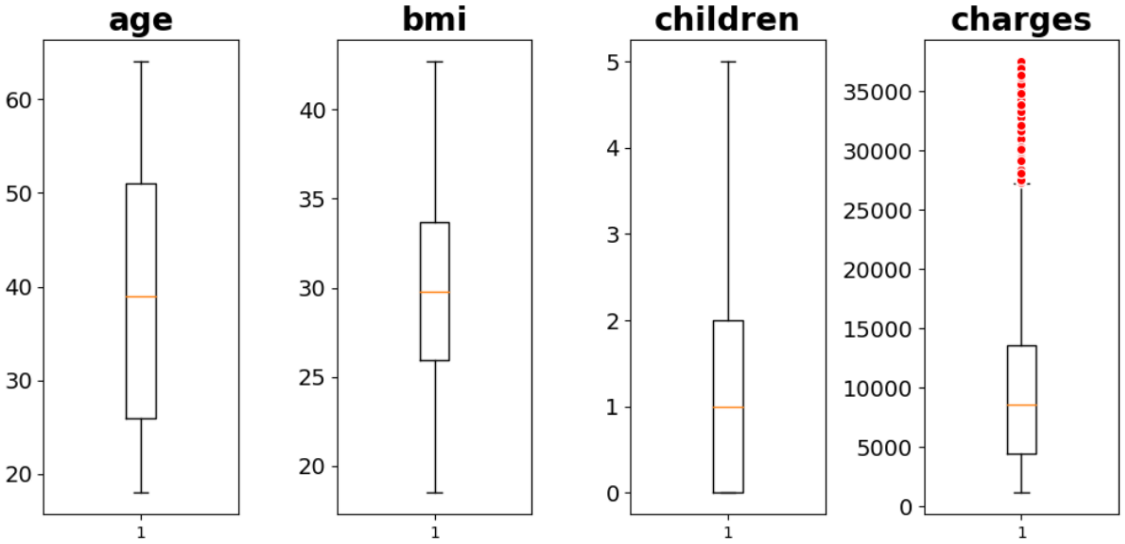


fig 3.1.2.2 data after removing outliers

3.2 Describing data

Mean and standard deviation for each feature were found to be :

mean : bmi (30.012358) - age(38.768840) - children (1.071126) - charges (1.072234e+04)

std : bmi (5.334893) - age(14.127983) - children (1.212934) - charges (8.345511e+03)

We will use these values in Standardization using z-score.

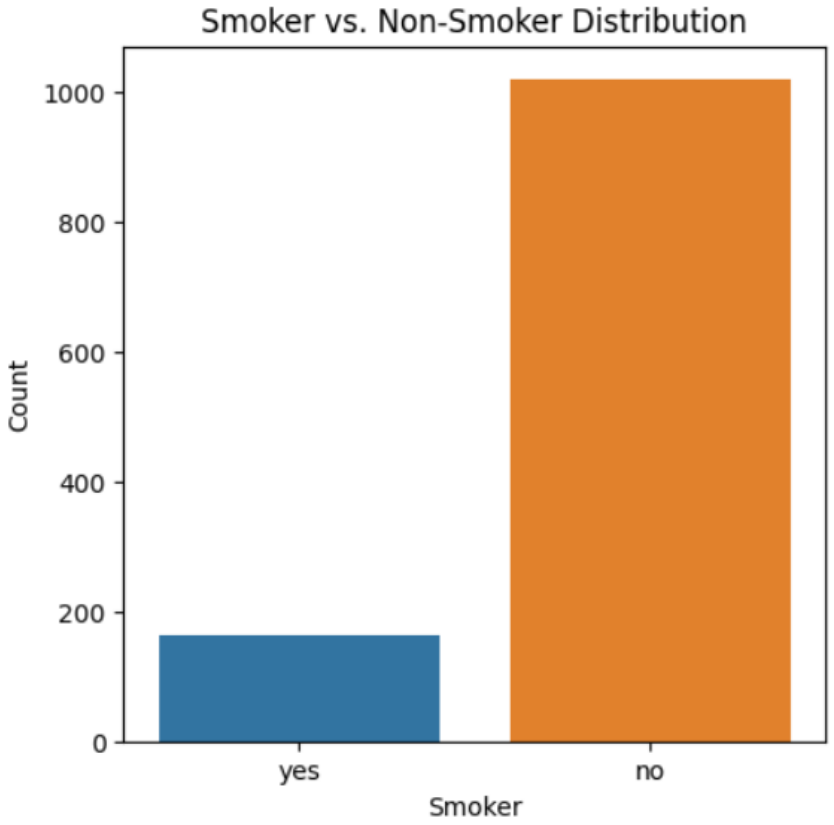
3.3 Standardizing the features

After using the z-score these are the first 5 row of our dataset:

	age	sex	bmi	children	smoker	region	charges
0	-1.399268	female	-0.395951	0	yes	southwest	0.738432
1	-1.470050	male	0.704352	1	no	southeast	-1.078039
2	-0.762235	male	0.560019	3	no	southeast	-0.751646
3	-0.408327	male	-1.369729	0	no	northwest	1.349484
4	-0.479109	male	-0.212255	0	no	northwest	-0.821457

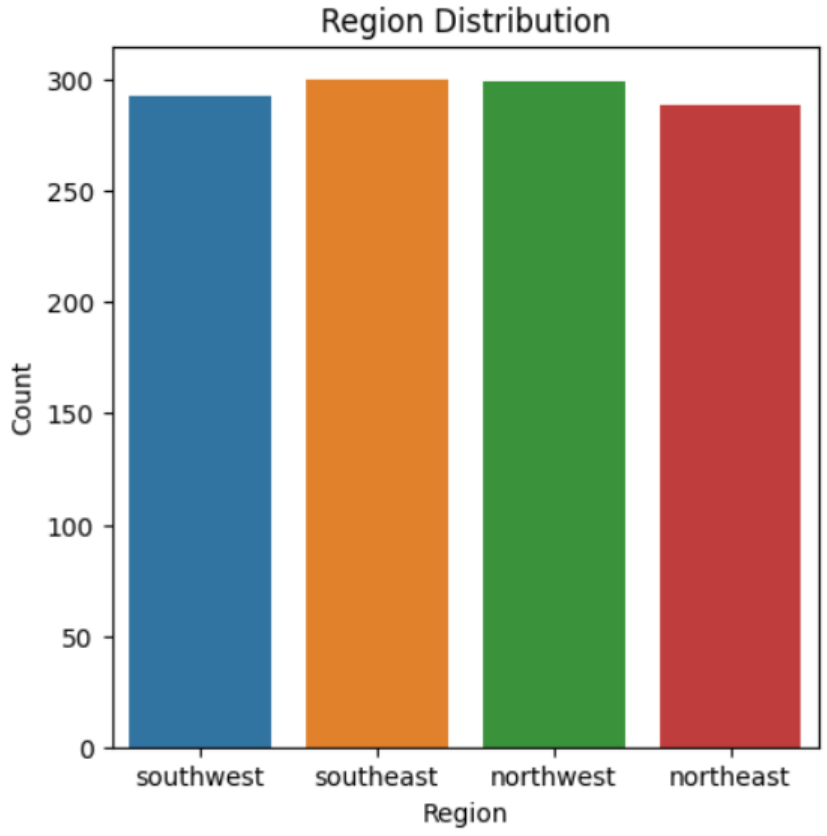
fig 3.3.1 part of data after standardization

3.4 Features distributions



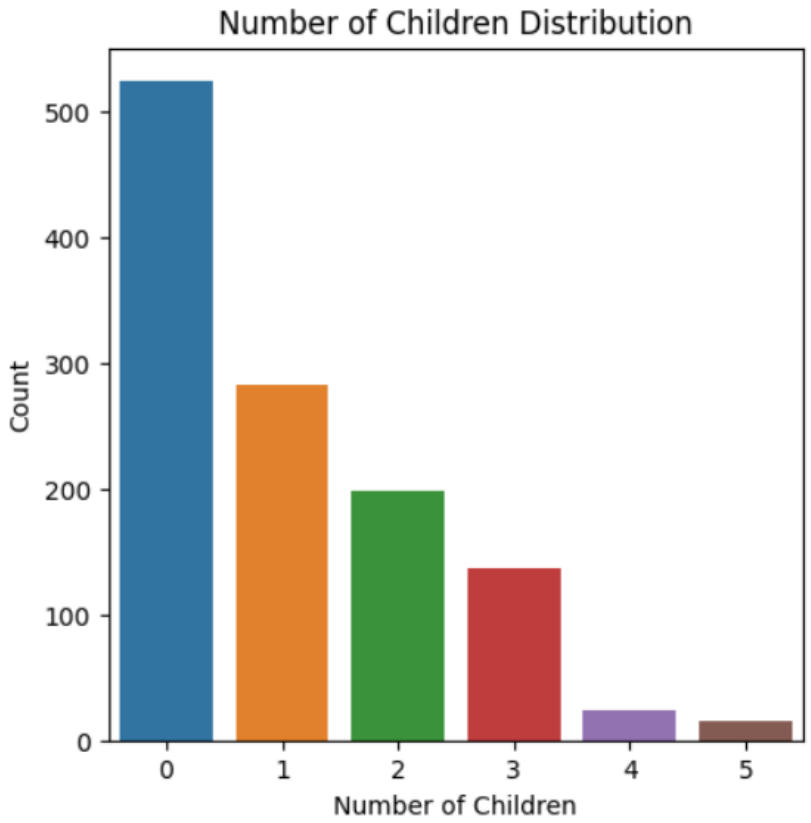
type of distribution: Bernoulli distribution

fig 3.4.1 smoker distribution



type of distribution: Uniform distribution

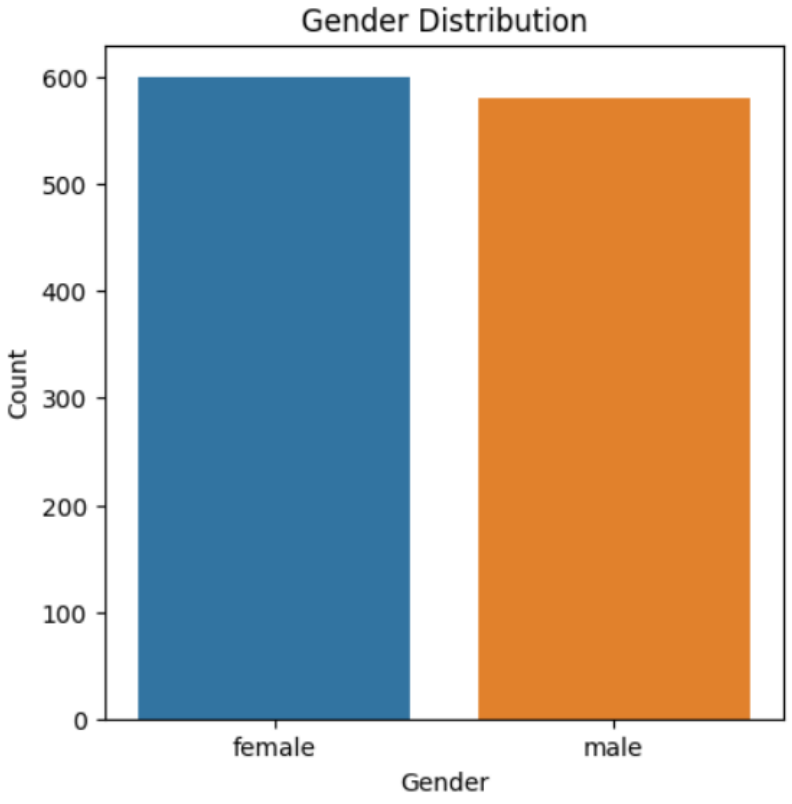
fig 3.4.2 region distribution



type of distribution:

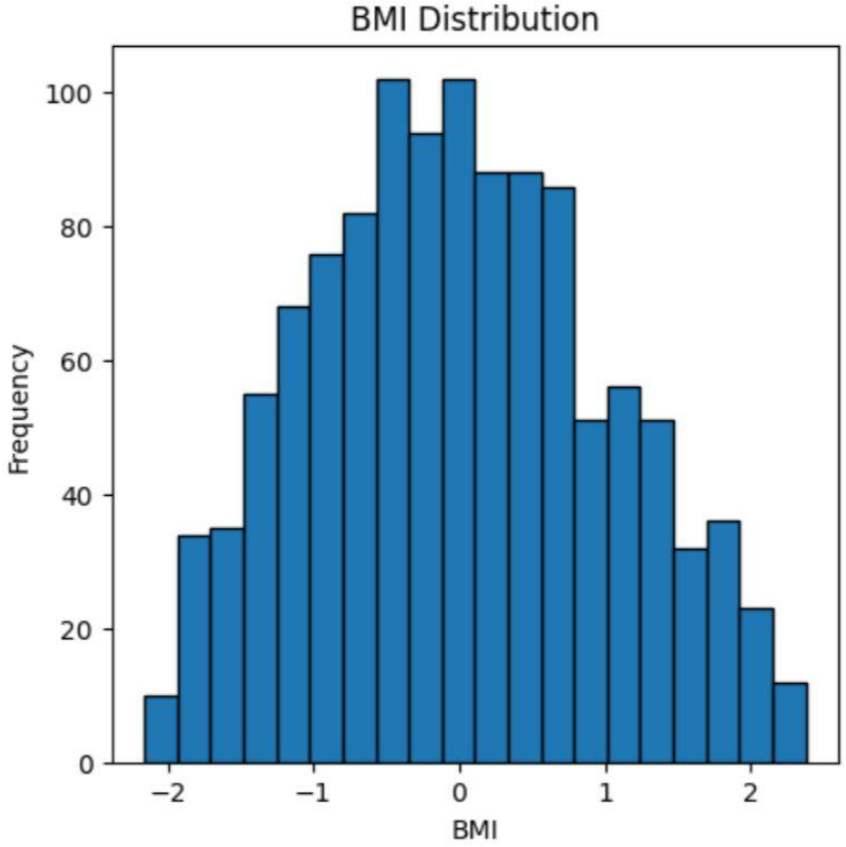
Geometric distribution

fig 3.4.3 number of children distribution



type of distribution: Bernoulli distribution

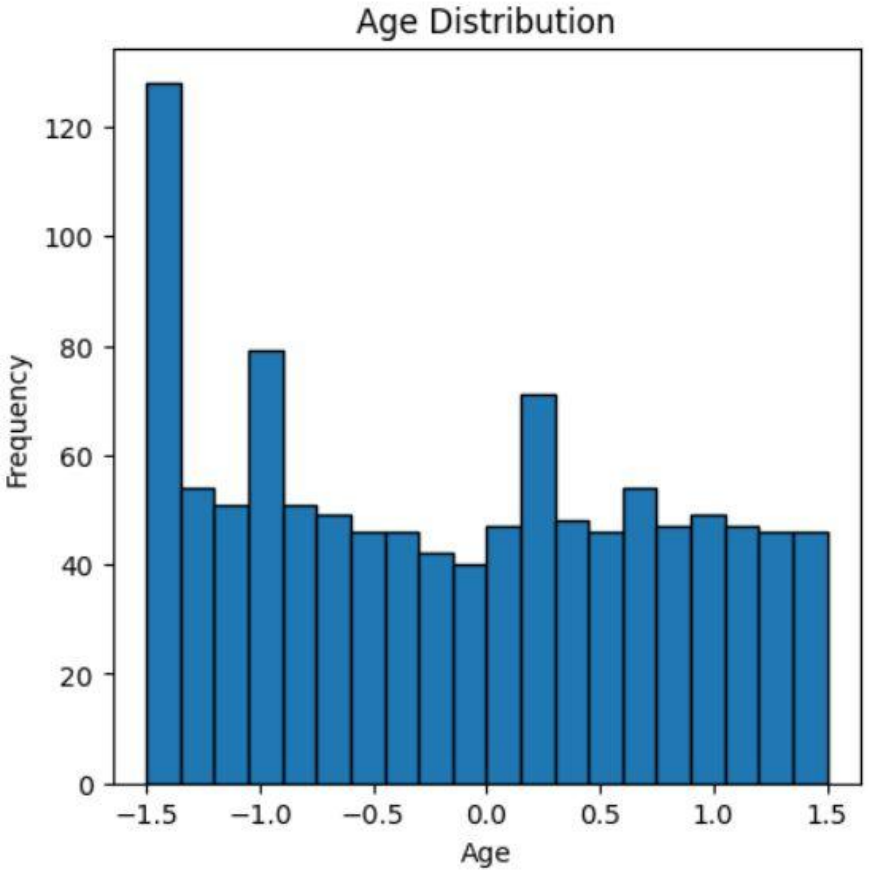
fig 3.4.4 gender distribution



type of distribution: Normal distribution

Which most data points cluster toward the middle of the range, while the rest taper off symmetrically toward either extreme.

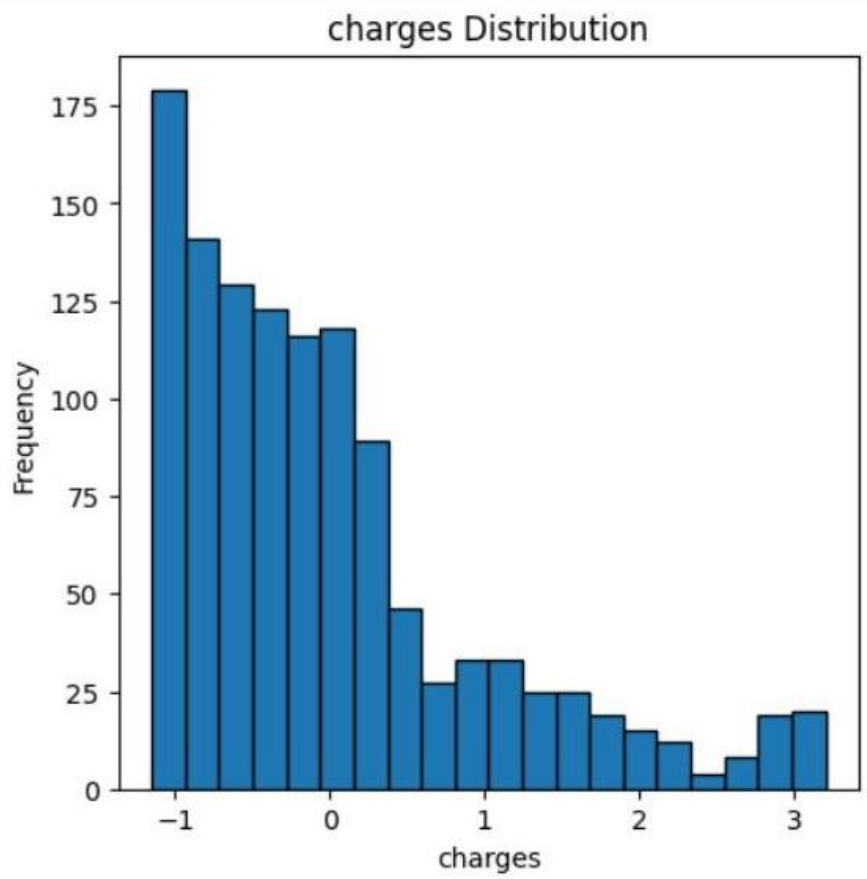
fig 3.4.5 BMI distribution



type of distribution : Uniform distribution

it tends to be uniform although there are some values have more frequency than the rest

fig 3.4.6 age distribution



type of distribution: log-normal distribution

fig 3.4.7 charges distribution

3.5 Normality test

- **age**
  p-value: 3.2019030600933976e-21
  Reject the null hypothesis: The data does not follow a normal distribution.
- **bmi**
  p-value: 3.72329331810306e-08
  Reject the null hypothesis: The data does not follow a normal distribution.
- **charges**
  p-value: 2.2347923075799168e-30
  Reject the null hypothesis: The data does not follow a normal distribution.

3.6 Correlation coefficient

we calculated the correlation coefficient between each continuous feature and the response feature [charges] and this is the results :

bmi	0.004397
charges	1.000000
age	0.305213

fig 3.6.1 correlation coefficient results

3.7 linear regression

We calculated linear regression from scratch and plotted the results

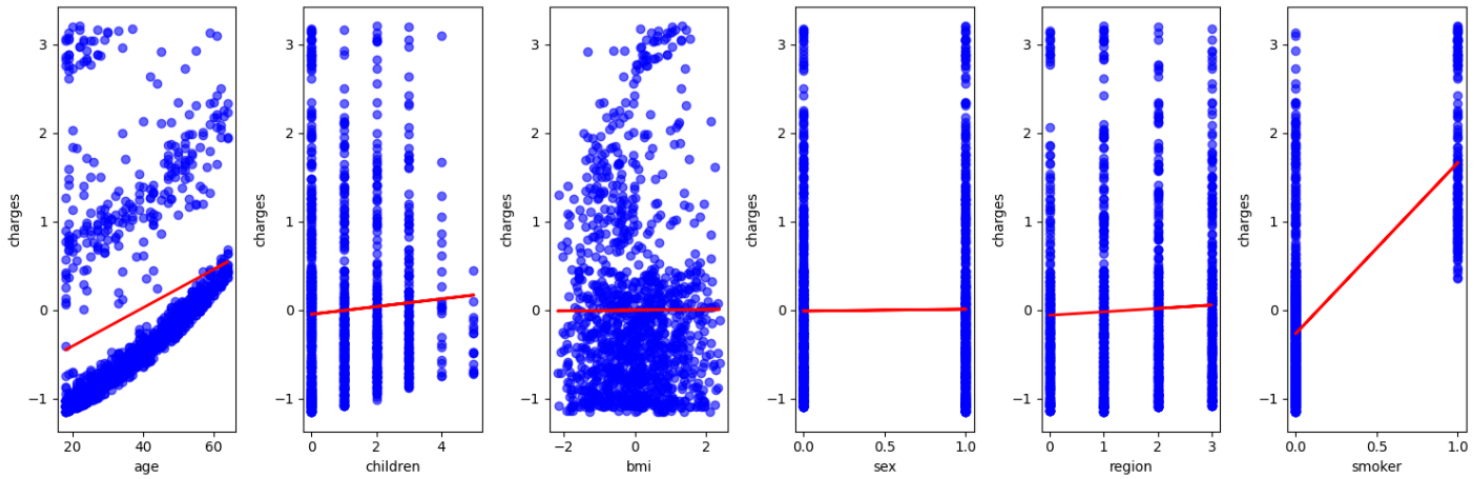


fig 3.7.1 linear regression from scratch

3.8 multivariable linear regression

we calculated multivariable linear regression with method called linearregression() to get the coefficients and bias:

RMSE : 0.6250645069864977

R2 Score : 0.6089632556329347

Intercept: -0.3937709504864599

Coefficients: age [0.37788738]sex [-0.03080995]bmi [0.11831168]number of children [0.05378538]smoker [2.13357776] region [0.03924418]

4 Conclusion

In this study, we have applied multivariable linear regression to predict the charges for healthcare services based on various demographic and health-related factors. The model yielded promising results, with a root mean squared error (RMSE) of 0.6250645069864977 and an R2 score of 0.6089632556329347.

Our analysis revealed that several factors significantly influence healthcare charges. The coefficients of the independent variables provided insights into the direction and magnitude of their impact on the charges. Notably, variables such as age, BMI, number of children, smoking status, and region showed statistically significant associations with healthcare charges.

The findings suggest that as age and BMI increase, healthcare charges tend to rise. Furthermore, individuals who smoke demonstrated significantly higher charges compared to non-smokers. Additionally, the number of children was found to have a positive effect on charges. The influence of region was also observed, with certain regions exhibiting higher charges compared to others.

The predictive performance of the model, as indicated by the R2 score, suggests that approximately 60.9% of the variance in healthcare charges can be explained by the included independent variables. However, it is important to note that there may be other unaccounted factors or complexities that could further enhance the predictive accuracy of the model.

This study demonstrates the potential of multivariable linear regression as a valuable tool for understanding and predicting healthcare charges. The insights gained from this research can assist healthcare providers, insurers, and policy-makers in better understanding the factors driving charges and making informed decisions regarding resource allocation and healthcare cost management.

Although the results obtained from this study are encouraging, it is essential to acknowledge the limitations. The dataset used for analysis may not capture all possible variables influencing healthcare charges, and there may be other non-linear relationships or interactions between variables that were not explored. Additionally, the generalizability of the findings may be limited to the specific population and timeframe of the dataset.

Future research could focus on incorporating additional variables, exploring alternative regression techniques, and assessing the robustness of the model on different datasets. Furthermore, investigating the potential impact of interventions or policy changes on healthcare charges could provide valuable insights for healthcare cost containment strategies.

5 Member Contribution

Reem Adel	report(methods-results)-choosing dataset-linear regression model
Sara Mahmoud Hussein	report(methods-results)-outliers visualization-multivariable linear regression -normality test-notebook organizing
Aya Salah	report(conclusion)- mapping data-standardization - descriptive statistics
Yassmen Sayed	report(methods-results)-handling data(outliers-duplicated - categorical or numerical) - correlation coefficient
Mohamed Gamal	report(introduction)-linear regression model - distribution plotting