NAME: ABEL

SURENAME: ILUNGA

STUDENT NUMBER: ST10090262 POE PART 1

MODULE: PDAN8411

**LINEAR REGRESSION REPORT: MEDICAL CHARGES A PREDICTIONS**

**INTRODUCTION**

The goal of medical assistance programs is to provide customers with reasonable, data-driven rates. This project offers a data-driven proof of concept that predicts medical insurance costs based on lifestyle and demographic parameters using linear regression. By identifying the main factors influencing medical expenses, the model will assist the business in creating more precise pricing plans that are suited to the needs of individual clients.

1. **DATA OVERVIEW**

The datasets consist of **1338** records with **7** features including demographics and lifestyles variables such

- age: Age of the individual

- sex: Gender (binary)

- bmi: Body Mass Index

- children: Number of children covered

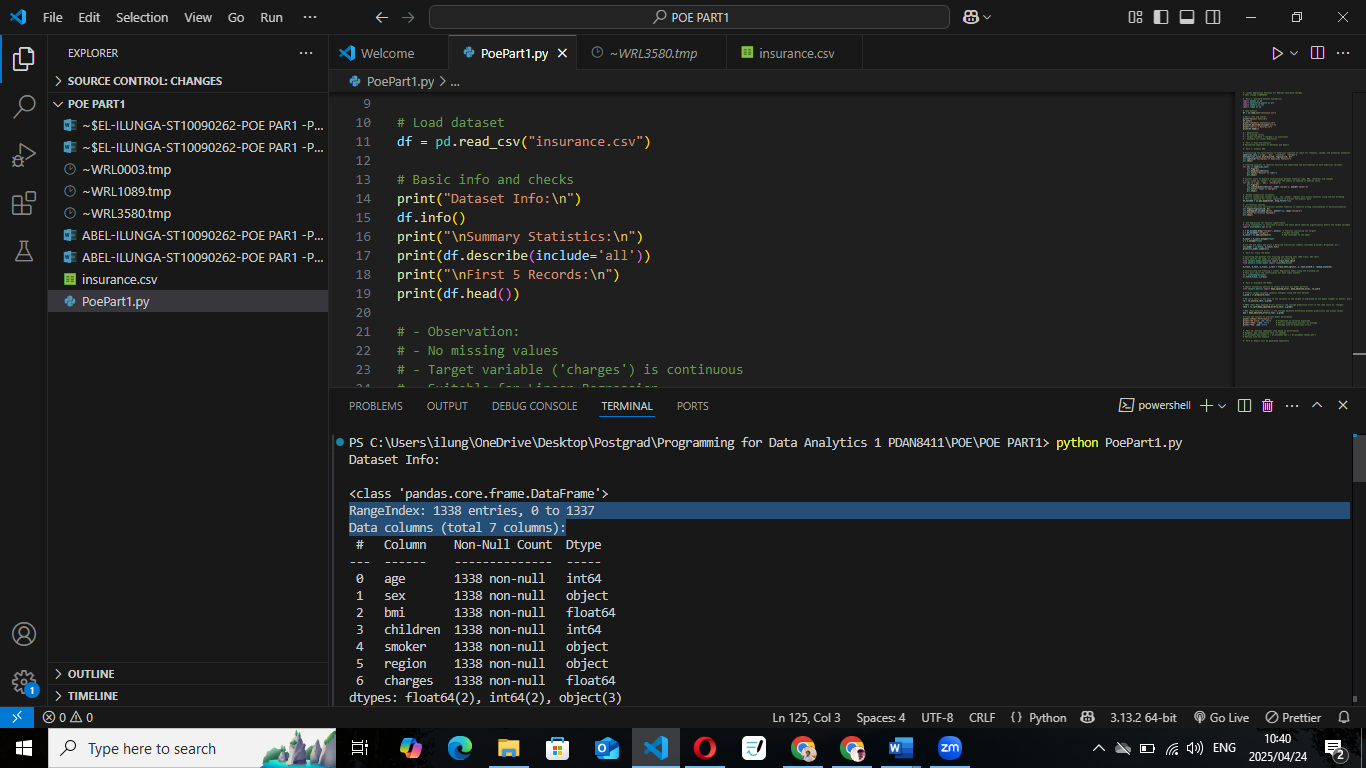
- smoker: Smoking status

- region: Geographic region in the US

- charges: Annual medical charges (target variable)

* 1. There are no mussing values
  2. The target variables (charges) is continuous, ideal for linear regression.
  3. Categorical feature like sex, smoker, and region will be encouraged.
  4. Suitability confirmed based on the variable types and structure.

*Code: to be found in the Jupyter file*

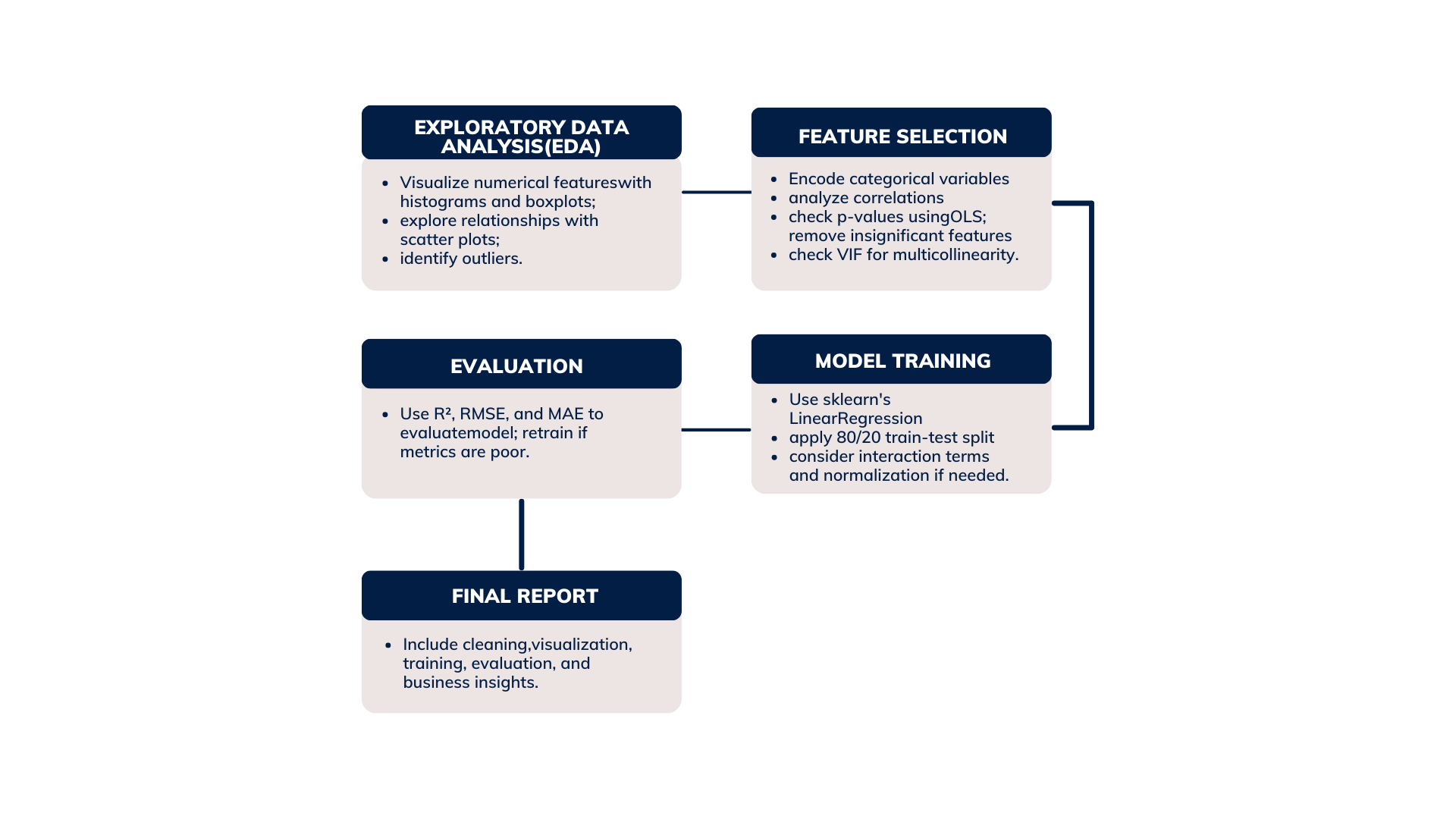


1. **PLAN OF ANALYSIS**

I created a precise analysis plan of creating a strong linear regression model to forecast medical insurance costs, making sure that every stage is rationally organised and in line with the modelling goals. This plan is visually summarised in the flowchart below, which is divided into five main stages:

**a. Exploratory Data Analysis (EDA)**: This phase includes visualizing distributions of numeric variables (e.g., age, BMI, charges), identifying outliers, and exploring relationships between features (e.g., smoking vs. charges). This step ensures data quality and highlights patterns.

1. **Feature Selection**: After encoding categorical features (e.g., sex, region), I will assess correlation, use OLS p-values to test significance, and check multicollinearity with VIF. Irrelevant or redundant features will be removed to improve model efficiency and interpretability.
2. **Model Training**: The model will be trained using a train-test split (typically 80/20). Scikit-lean’s Linear Regression will be used initially, with potential tuning or transformation (e.g., normalization or interaction terms) if needed.
3. **Model Evaluation**: Key evaluation metrics will include R-squared (for variance explanation), RMSE (for error magnitude), and MAE (for average error). These will guide any retraining or adjustments.
4. **Report and Insights**: Finally, the results will be compiled into a business-focused report that includes EDA findings, model interpretation, and insights relevant to the medical aid client’s goals.



1. **CONDUCTING THE ANALYSIS**

In this section, I use Python to illustrate the entire analysis pipeline. To guarantee openness, reproducibility, and conformity to regression modelling best practices, each step is explicably explained and supported.

**Exploratory Data Analysis (EDA)**

* Histograms are useful for visualising continuous variable distributions and identifying uniformity or skewness. (Kenton, 2022).
* To identify outliers and comprehend how data is distributed across quartiles, boxplots are utilised. (McKinney, 2018).
* The hue='smoker' property in scatter plots draws attention to possible correlations between factors and insurance premiums:
  + Smokers generally incur higher charges.
  + Age correlates positively with medical charges.
* A correlation heatmap aids in the detection of multicollinearity by visualising the direction and intensity of linear correlations between variables. (Seaborn Docs, 2023).

**Feature Selection**

* For model compatibility, one-hot encoding is used to transform categorical variables (sex, region, and smoker) into binary features. (Scikit-learn, 2023).
* OLS Regression is used to evaluate the statistical significance of features through p-values, identifying which features significantly impact the target (charges) (Brooks et al., 2019).

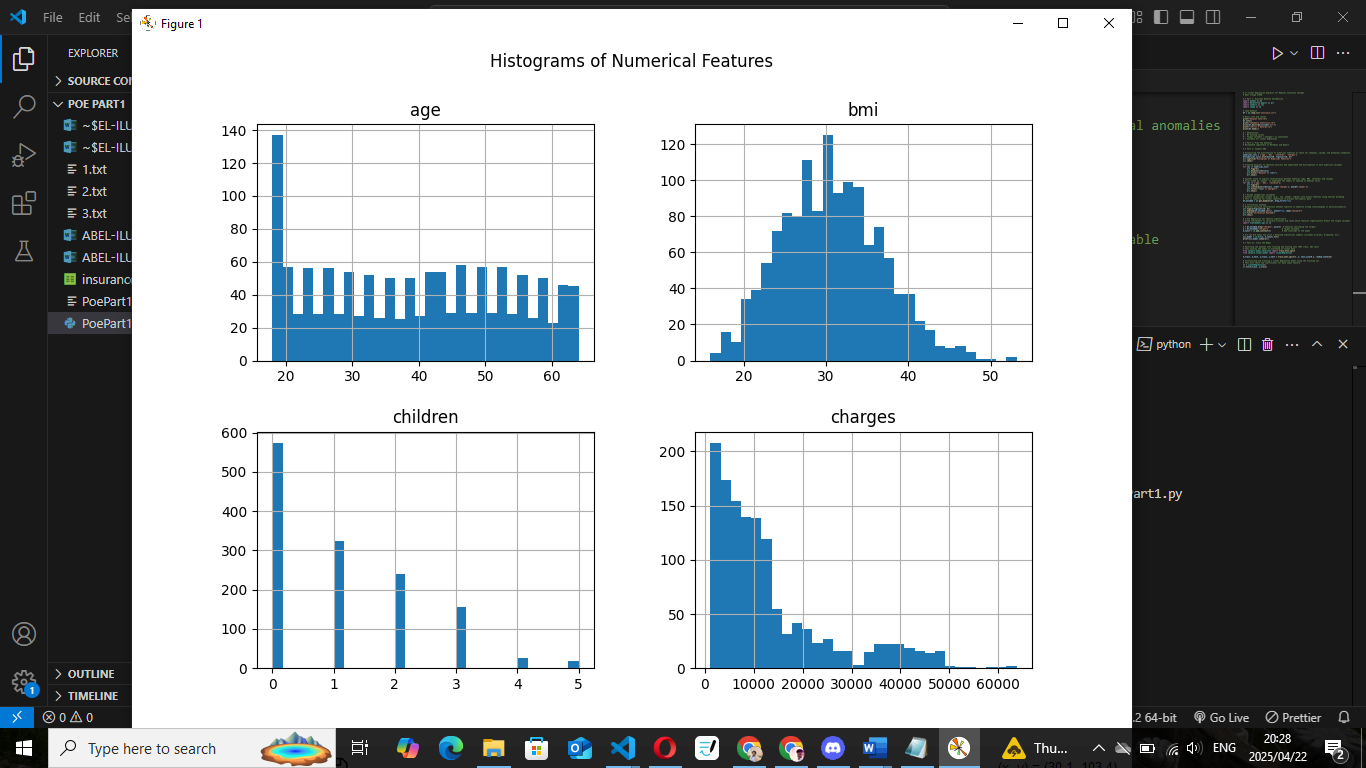
**Model Training**

* By conserving unseen data for testing, an 80/20 train-test split guarantees equitable model evaluation. (Scikit-learn, 2023).
* Scikit-learn is used to construct linear regression, which minimises the residual sum of squares between observed and expected values to estimate coefficients. (James et al., 2013).

This model acts as a baseline, even though basic linear regression lacks hyperparameters that can be adjusted. If assessment measures indicate underperformance, normalisation or interaction terms can be investigated later.

Each of these steps contributes to building a transparent, accurate, and explainable model tailored for predicting medical insurance charges, the following images demonstrates all details explained above, and more steps on the achievements are in the code *Jupyter* submitted along with this report

**Histograms**



**Boxplots (age)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Boxplots (bmi)A screenshot of a computer

AI-generated content may be incorrect.**

**Boxplots (children)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Boxplots(charges)**

**A screenshot of a computer

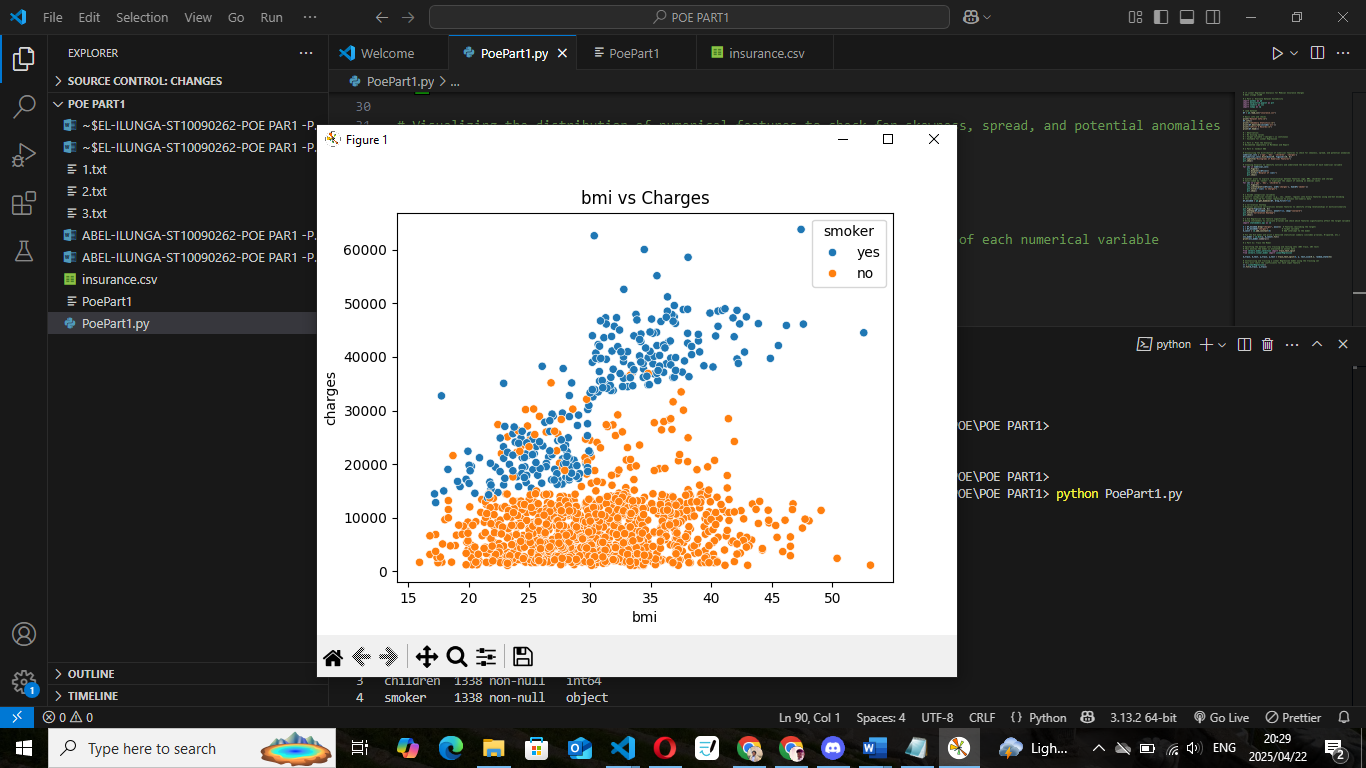
AI-generated content may be incorrect.**

**Scatter Plots (age vs. charges)**

**A screenshot of a computer

AI-generated content may be incorrect.**

**Scatter Plots(bmi vs charges)**

****

**Scatter Plots (children vs charges)**

**A screenshot of a computer screen

AI-generated content may be incorrect.**

**Correlation Heatmap A screenshot of a computer screen

AI-generated content may be incorrect.**

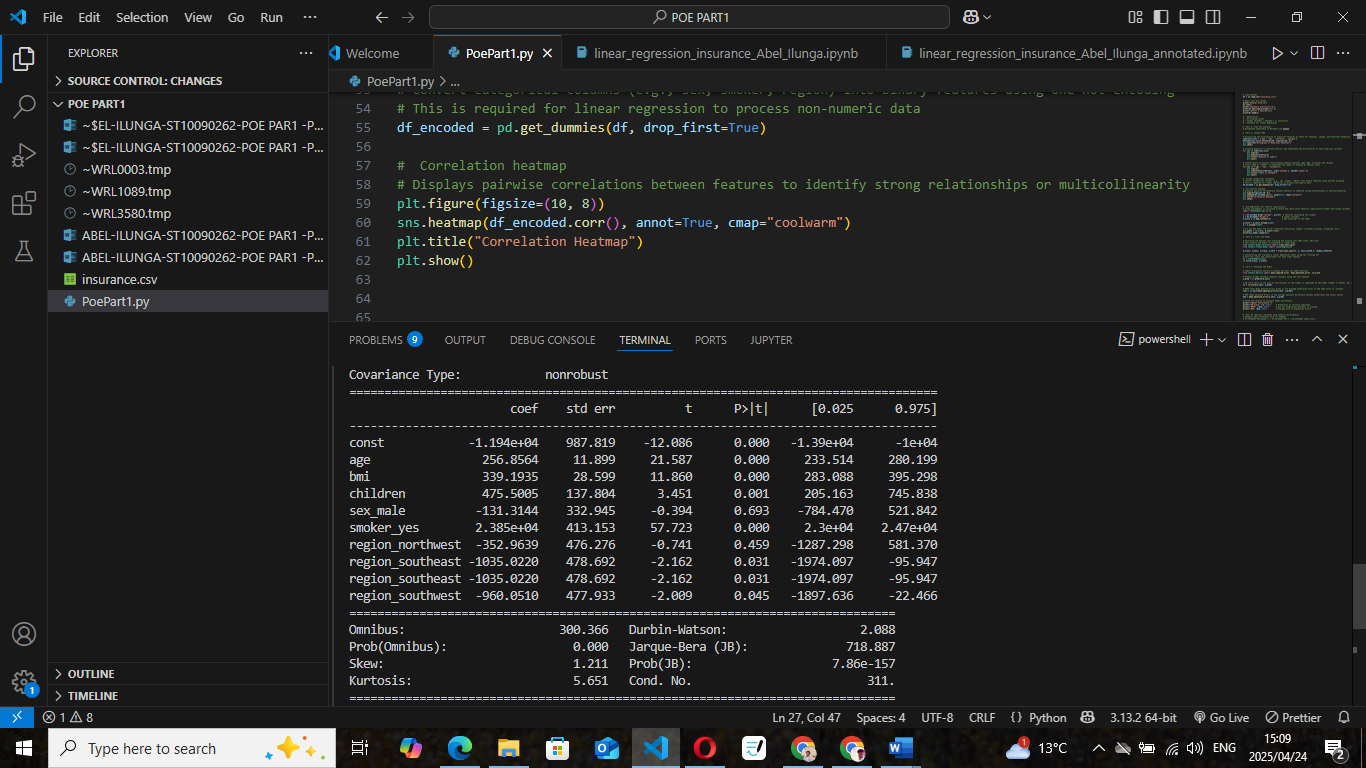
**4. EVALUATING THE MULTIPLE LINEAR REGRESSION MODEL**

Model Summary

I used Ordinary Least Squares (OLS) to train a Multiple Linear Regression model after preprocessing the data, which included one-hot encoding categorical variables and dividing the dataset into training and testing sets. The model overview is as follows:

* Dependent Variable: charges
* R-squared: 0.751
* Adjusted R-squared: 0.749
* F-statistic: 500.8 (p < 0.001)
* Number of Observations: 1338
* Number of Predictors (Df Model): 8

Model Coefficients:



Interpretation:

According to the regression results, the factors that have the biggest favourable effects on medical charges are age, BMI, and smoking status. For instance, when all other factors are held constant, smoking raises predicted expenses by about R23,850. About 75.1% of the variance in the target variable can be explained by the model, which is regarded as strong for data from the actual world.

Model Evaluation Metrics

To assess the model’s predictive performance, I used the following metrics on the test dataset:

* R² Score: 0.784  
  This means that 78.4% of the variation in medical charges is explained by the model. This suggests the model has good explanatory power.
* Root Mean Squared Error (RMSE): R5,796.28  
  This metric shows the standard deviation of prediction errors. A lower RMSE indicates better model accuracy. Given the wide range of medical expenses, this is a reasonable error level.
* Mean Absolute Error (MAE): R4,181.19  
  MAE reflects the average absolute difference between the predicted and actual values. It is easier to interpret and not as sensitive to outliers as RMSE.

Why these metrics were chosen:

* R² evaluates the model's overall fit.
* RMSE emphasizes larger prediction errors and is good for assessing variance.
* MAE is robust and provides a straightforward measure of average error.

According to *James et al. (2013)* and *Scikit-learn (2023)*, combining these metrics provides a comprehensive view of model performance. Given the results, the model performs well and can be considered reliable for predicting medical charges.

**A screenshot of a computer

AI-generated content may be incorrect.**

Conclusion on Retraining:  
Retraining the model is not necessary currently, according on the existing performance metrics (R2 = 0.784, RMSE = R5,796.28, MAE = R4,181.19). For predicting actual medical costs, the model already offers a high explanatory power and manageable error rates. Future enhancements like regularisation or feature engineering, however, might increase its performance even further.

Reference

Brooks, C., Burke, L., & Persand, G. (2019) The econometrics of financial markets. 3rd edn. Cambridge: Cambridge University Press.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013) An introduction to statistical learning: with applications in R. New York: Springer.

Kenton, W. (2022) Histogram Definition. Investopedia. Available at: https://www.investopedia.com/terms/h/histogram.asp (Accessed: 24 April 2025).

McKinney, W. (2018) Python for data analysis: data wrangling with Pandas, NumPy, and IPython. 2nd edn. Sebastopol, CA: O'Reilly Media.

Scikit-learn (2023) Scikit-learn: Machine Learning in Python. Available at: https://scikit-learn.org/stable/ (Accessed: 24 April 2025).

Seaborn Docs (2023) Seaborn: Statistical data visualization. Available at: https://seaborn.pydata.org/ (Accessed: 24 April 2025).

Choi, M. (2018) *Insurance Dataset*. Kaggle. Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance (Accessed: 24 April 2025).