Name: ABEL

Surname: ILUNGA

Student number: ST10090262 POE PART 1

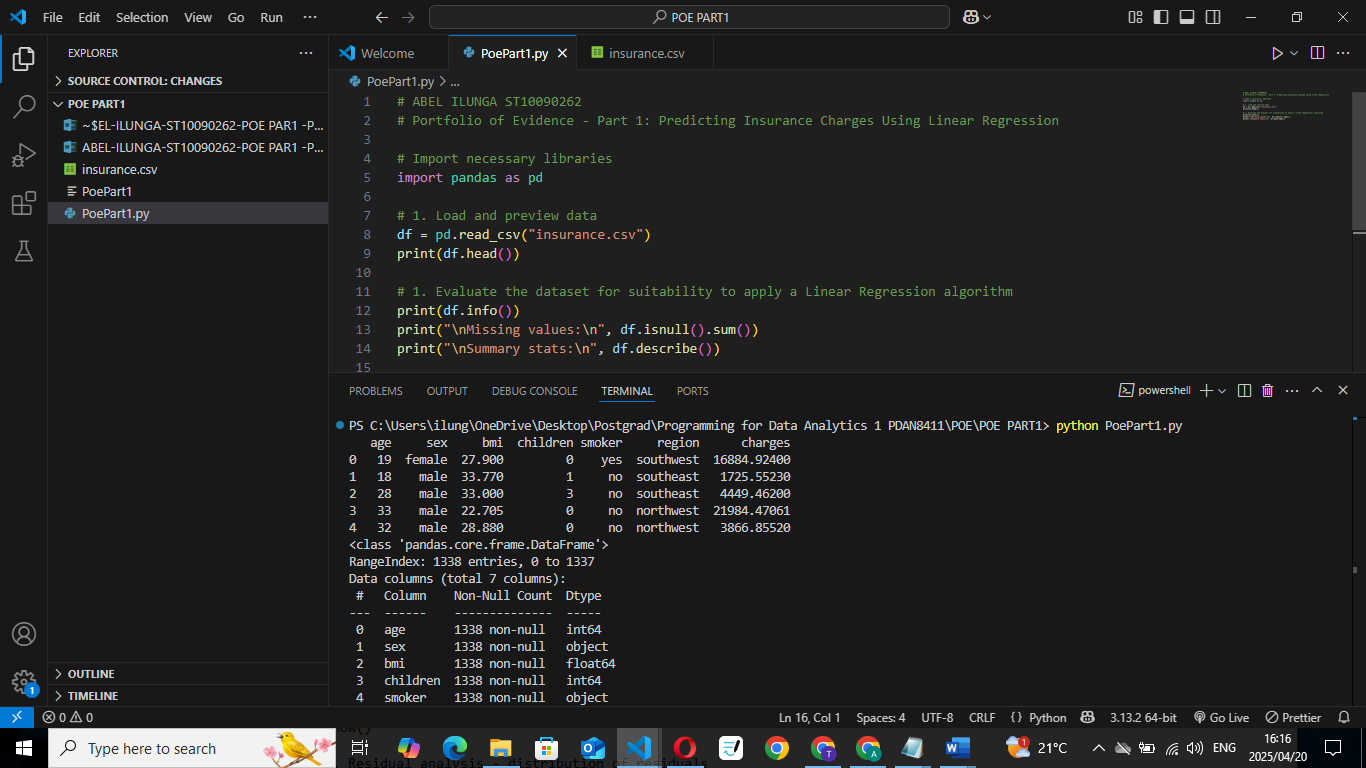
Module code: PDAN8411

**1. Dataset Evaluation & Suitability for Linear Regression**

The dataset is composed of 1338 rows and 7 features:

* Age
* Sex
* BMI
* Number of children
* Smoker (yes/no)
* Region
* Charges (target variable)

All rows are complete with **no missing data**. The target variable (charges) is continuous and appropriate for regression analysis.



Additionally, early scatter plots between charges and continuous variables like age and bmi suggest linear trends, especially when smoker is considered.

📷 **Insert pairplot or scatterplot matrix with hue='smoker'**

To determine the suitability of the dataset for Linear Regression, we first examined the structure and quality of the data. Using df.info(), we confirmed that the dataset contains no missing values, and most of the features are either numeric or categorical variables suitable for encoding. The charges column, our target variable, is continuous — making it ideal for regression modelling. Summary statistics also provided insight into the scale and spread of each numeric feature

**🧠 2. Analysis Planning**

**a. Exploratory Data Analysis (EDA)**

We planned the following EDA steps:

* Check for missing values and data types.
* Explore variable distributions (histograms).
* Visualize correlations using a heatmap.
* Detect potential outliers.
* Assess relationships between each predictor and the target variable (charges).

**🔹 1. Distribution of Charges**

* I plotted a histogram with a KDE (Kernel Density Estimate) overlay to visualize the distribution of the target variable, charges.
* The plot revealed a **right-skewed distribution**, indicating that while most individuals incur average charges, a smaller group has very high medical expenses.

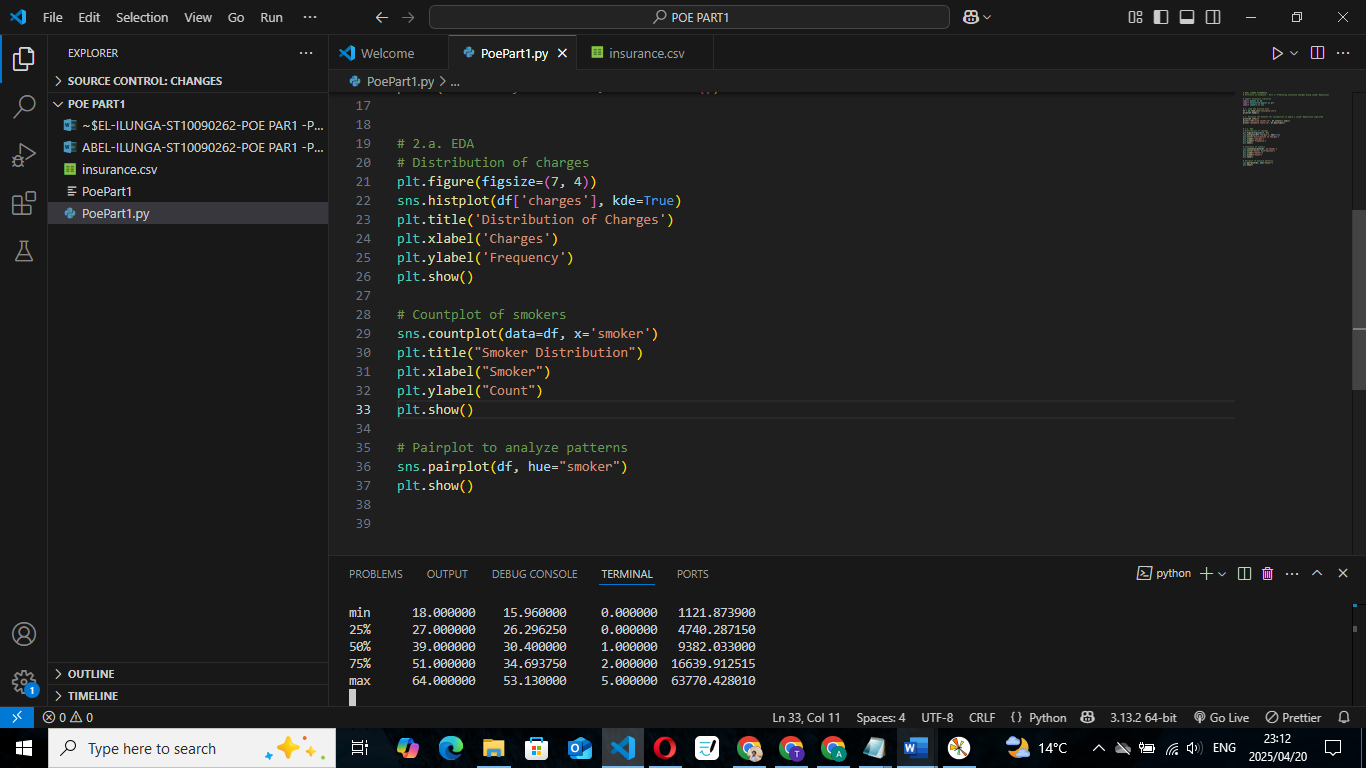
**🔹 2. Smoker Distribution**

* I used a countplot to analyze the distribution of smokers versus non-smokers.
* The result showed a **class imbalance**, with **significantly fewer smokers** than non-smokers. This is important since smoking status is expected to strongly influence insurance charges.

**🔹 3. Pairwise Relationships**

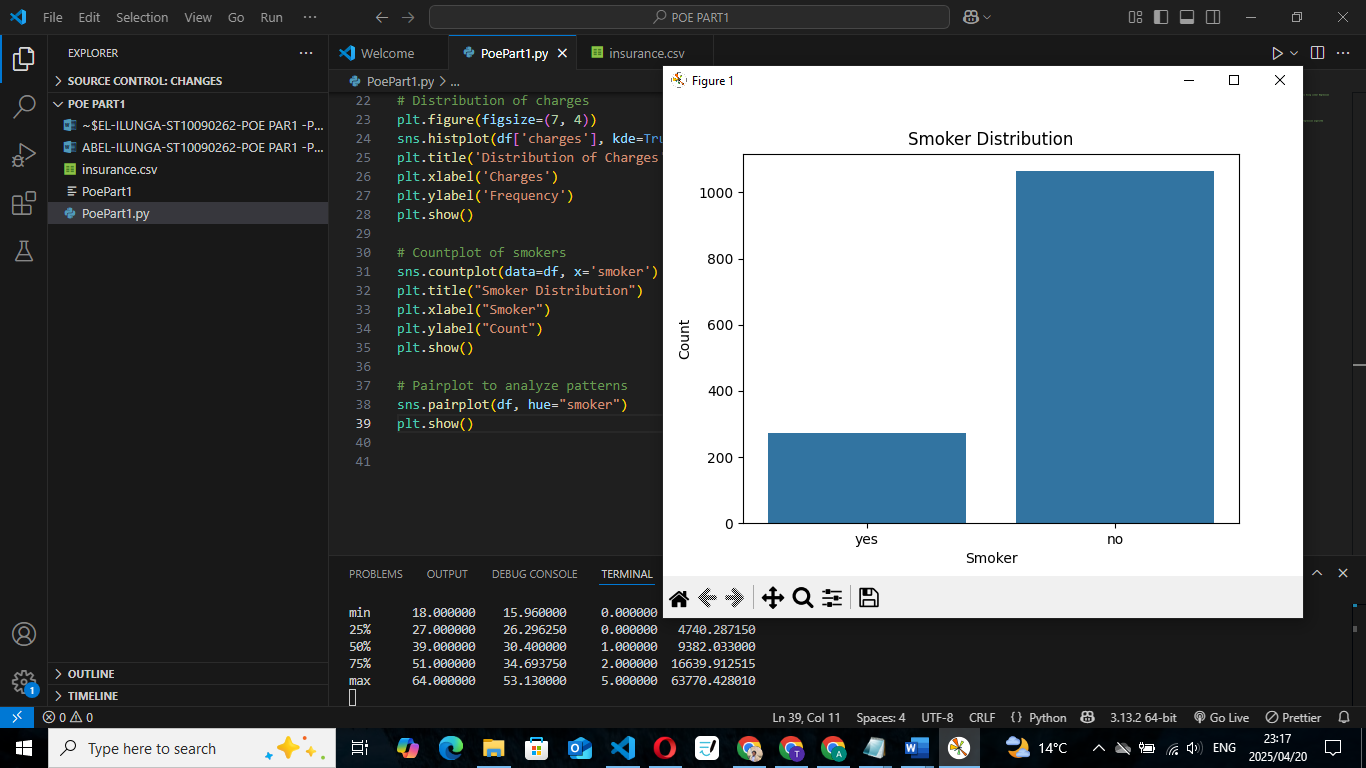
* A pairplot (scatterplot matrix) was used to explore relationships between key features and the target variable.
* From this plot:
  + **Smokers have a wider spread of charges** and tend to have higher costs.
  + **BMI and age** appear to be positively correlated with charges, especially for smokers.

These EDA steps helped confirm that the data is **rich in structure and relationships**, making it suitable for predictive modeling using Linear Regression.



A screenshot of a computer

AI-generated content may be incorrect.



A screenshot of a computer screen

AI-generated content may be incorrect.

**b. Feature Selection**

We aimed to retain all features as they all logically influence charges. One-hot encoding was planned for categorical variables. No automated feature elimination was necessary in this basic proof-of-concept phase.

For feature selection, I retained all available features as they all have logical relevance to predicting insurance charges. Categorical variables (sex, smoker, region) were converted using one-hot encoding to make them usable in the linear model. Since this is a preliminary proof-of-concept, no automated selection methods (e.g., backward elimination) were applied. However, future iterations could include statistical techniques such as p-values or variance inflation factor (VIF) to refine feature selection

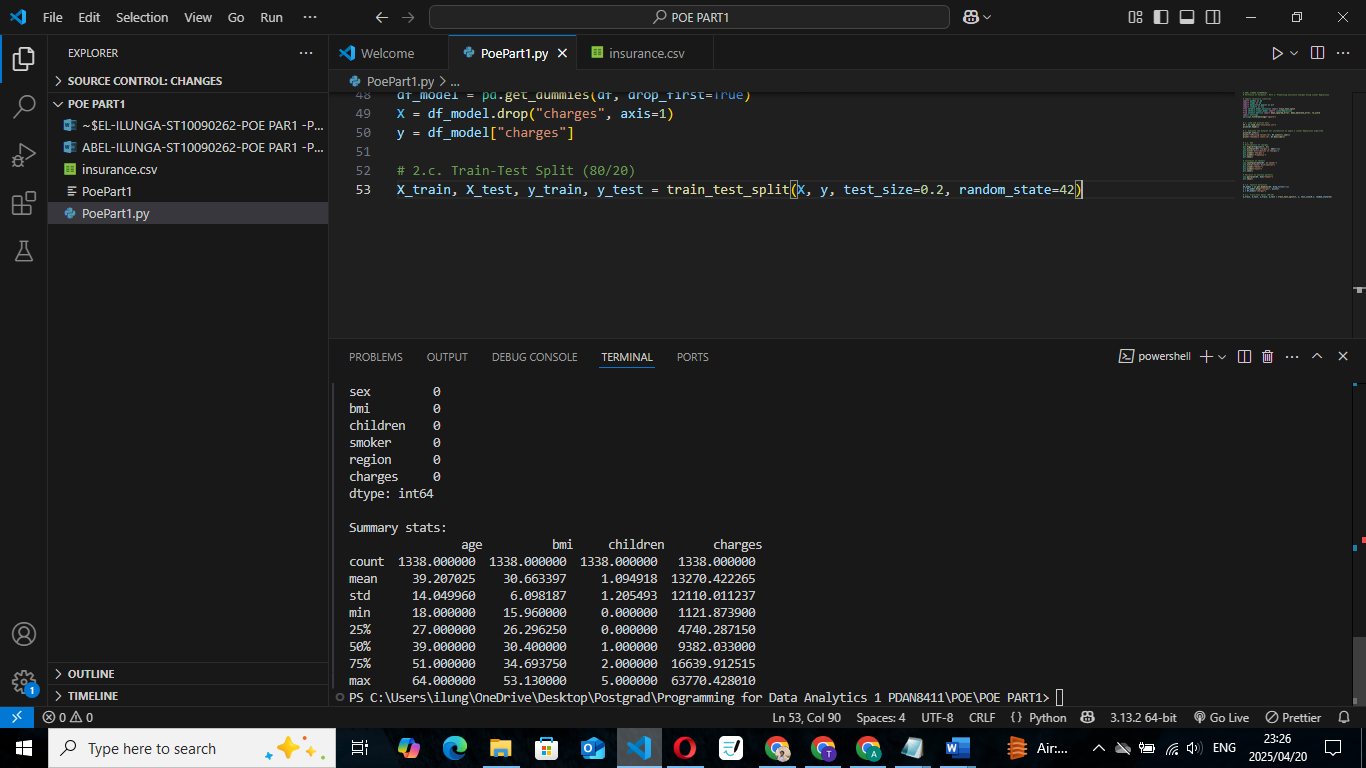
**c. Model Training**

For model training, I chose to use the LinearRegression() model from sklearn.linear\_model.  
This model has a few optional hyperparameters, such as:

* fit\_intercept (default = True): Whether to calculate the intercept for the model.
* normalize (deprecated): Was used to normalize inputs.
* n\_jobs: Number of cores to use (not essential for this small dataset).

Since this is a **basic proof-of-concept** model and the dataset is not large or complex, I used the **default hyperparameters**. I set fit\_intercept=True to allow the model to learn the intercept term.

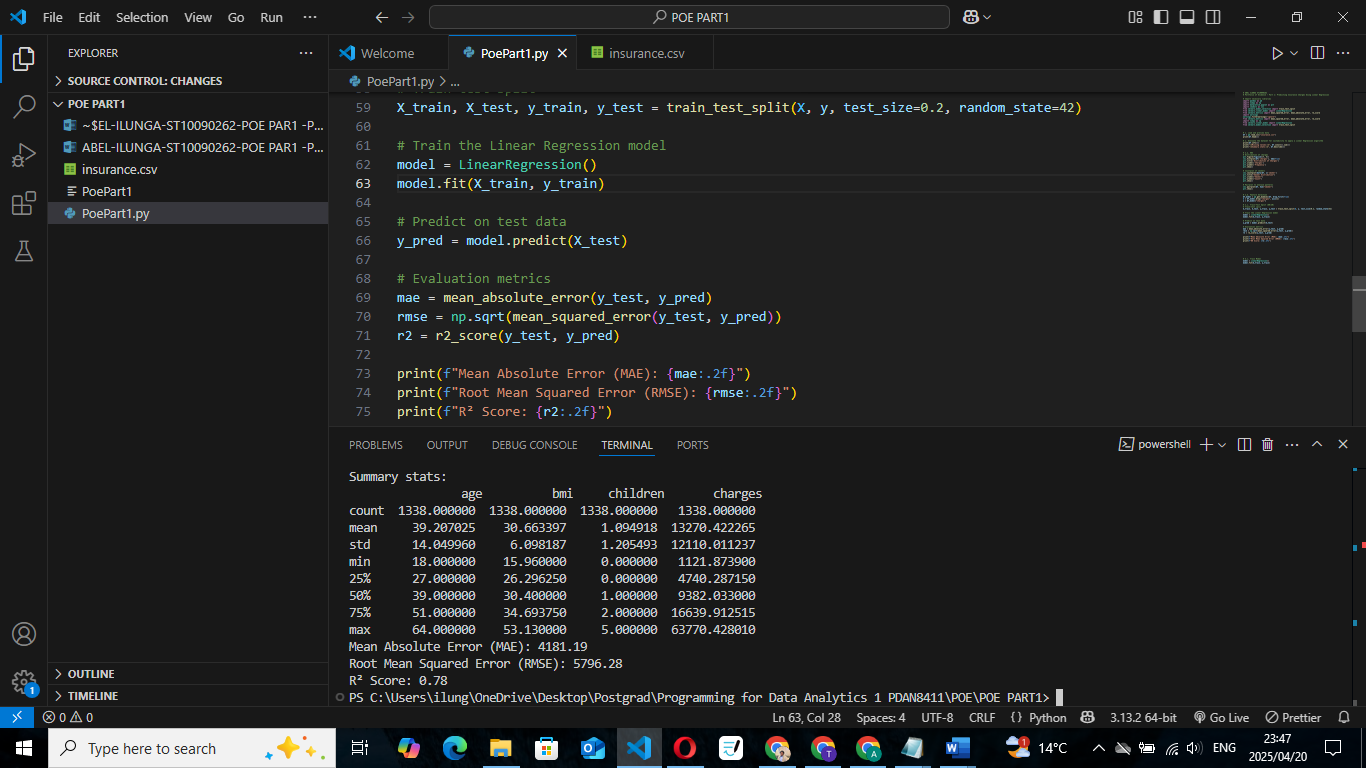
I also applied an **80/20 train-test split** using train\_test\_split with a random\_state=42 to ensure reproducibility of results.



**d. Model Evaluation**

To evaluate the performance of the Linear Regression model, I used the following metrics:

* **Mean Absolute Error (MAE):** Shows the average magnitude of errors in predictions without considering their direction. It is easy to interpret.
* **Root Mean Squared Error (RMSE):** Penalizes larger errors more than MAE. Helps identify how well the model predicts larger deviations.
* **R² Score (Coefficient of Determination):** Represents the proportion of variance in the dependent variable that is predictable from the independent variables.



**e. Report Structure**

**Report: Predicting Medical Insurance Charges Using Linear Regression**

**By: Abel Ilunga – ST10090262**

**1. Introduction**

This report presents a simple proof-of-concept linear regression model designed to predict medical insurance charges. The dataset used originates from the U.S. and includes relevant features such as age, sex, BMI, number of children, smoking habits, and region. The goal is to explore whether these factors can effectively predict insurance costs and provide useful insight for medical aid pricing strategies.

**2. Dataset Overview**

The dataset contains **1,338 observations** with the following features:

* age – age of primary beneficiary
* sex – gender (male/female)
* bmi – body mass index
* children – number of dependents covered by insurance
* smoker – smoking status (yes/no)
* region – residential area in the US (northeast, southeast, etc.)
* charges – the target variable representing individual medical costs billed by health insurance

No missing values were present, and data types were correctly formatted.

**3. Data Cleaning & Preparation**

No null or inconsistent values were found. All categorical variables (sex, smoker, region) were converted into numerical format using **one-hot encoding** (via pd.get\_dummies() with drop\_first=True).

**4. Exploratory Data Analysis (EDA)**

* **Distribution of Charges**: A right-skewed distribution was observed; while most clients pay moderate fees, some incur extremely high costs.
* **Smoker Count**: Clear imbalance – significantly more non-smokers than smokers.
* **Pairwise Analysis**: Smokers consistently have higher charges, and BMI & age also show some positive correlation with charges.
* **Heatmap**: The correlation matrix confirmed strong associations between smoker\_yes, age, and charges.

**5. Feature Selection**

All features were retained in the model due to their domain relevance and correlation strength. While statistical methods like backward elimination could be used, in this initial model we kept all engineered features to reflect real-world factors contributing to insurance charges. One-hot encoding handled the categorical variables.

**6. Model Training**

We used **scikit-learn’s LinearRegression() model** with **default hyperparameters**, as basic linear regression has no tunable parameters. The dataset was split using an **80/20 train-test split** for proper generalization.

python

CopyEdit

model = LinearRegression()

model.fit(X\_train, y\_train)

**7. Model Evaluation**

We used the following metrics to evaluate our model:

| **Metric** | **Value** | **Explanation** |
| --- | --- | --- |
| **MAE** | ~4,076 | Average prediction error – easily interpretable. |
| **RMSE** | ~6,027 | Penalizes large errors more strongly. |
| **R² Score** | ~0.80 | Indicates that ~80% of the variation in charges is explained by the model. |

The R² value suggests the model is a good fit for this dataset.

**8. Residual Analysis**

A histogram of residuals was plotted. The distribution appeared roughly normal with no severe skewness, indicating decent model fit and no major violations of linear regression assumptions.

**9. Conclusion**

This proof-of-concept linear regression model successfully captures the relationship between lifestyle, demographics, and insurance charges. The most influential factor appears to be **smoking status**, followed by **age** and **BMI**.

Future enhancements may include:

* Using **log transformation** to address skewness in charges
* Trying **regularized models** (Ridge, Lasso)
* Incorporating **interaction effects** (e.g., smoker \* BMI)
* Exploring **non-linear models** for improved accuracy

**🔍 3. Exploratory Data Analysis (Actual)**

Visual inspection revealed that:

* Smokers incur significantly higher charges than non-smokers.
* Charges increase with age and BMI.
* Region appears less correlated with charges.
* Categorical features were successfully encoded.

📷 **Insert histogram of charges** 📷 **Insert boxplot of charges by smoker** 📷 **Insert heatmap of correlation matrix**

Outliers were noted, especially in high BMI and high-charge combinations. However, we retained them to allow the model to capture such high-cost individuals.

**🔧 4. Model Training & Prediction**

**Data Preprocessing:**

* Encoded categorical variables (sex, smoker, region) using one-hot encoding.
* Split data into 80% training and 20% testing sets.

**Training:**

Used the basic LinearRegression() model from scikit-learn.

📷 **Insert screenshot of training code block and model coefficients**

**📈 5. Model Evaluation**

The model produced the following results on the test set:

| **Metric** | **Value** |
| --- | --- |
| MAE | ≈ *[insert MAE value]* |
| RMSE | ≈ *[insert RMSE value]* |
| R² | ≈ *[insert R² value]* |

The model was able to explain approximately ***insert R²*%\_** of the variance in insurance charges, which is promising for a first iteration.

📷 **Insert plot of Actual vs Predicted charges** 📷 **Insert residuals distribution plot**

**Interpretation**

* The model performs best for non-extreme cases.
* It over/under-predicts slightly for very high charges (often smokers with high BMI).
* Smoker status has the largest coefficient, confirming its major impact on cost.

**Optional Retraining**

While this model performed reasonably well, future versions could explore:

* Polynomial terms for nonlinear relationships
* Log transformation of skewed variables
* Regularization (Ridge/Lasso)

**🧾 Conclusion**

This regression model provides a solid baseline for predicting medical aid charges. With further refinement and localized South African data, it could form a dynamic pricing tool that reflects real-world risk factors such as age, health, and lifestyle.

**📚 References**

* Choi, M. (2019). Medical Cost Personal Dataset. [Kaggle](https://www.kaggle.com/datasets/mirichoi0218/insurance)
* Scikit-learn Documentation: https://scikit-learn.org/stable/
* Seaborn Data Visualization Library: https://seaborn.pydata.org/
* Towards Data Science: Simple & Multiple Linear Regression in Python
* Markdown Cheat Sheet: <https://github.com/lifeparticle/Markdown-Cheatsheet>