**Link to Github**

[**https://github.com/Mbali-CMD/POE-Part-1?tab=readme-ov-file#poe-part-1**](https://github.com/Mbali-CMD/POE-Part-1?tab=readme-ov-file#poe-part-1)

1. **Analysis Planning**
2. Asserting our goal: **The goal is to create a linear regression model that provides a sliding scale of charges according to the different features**
3. EDA - Exploratory Data Analysis:

* View the columns.
* Check for missing data.
* Understand data types (numeric, categorical).
* Look at basic stats (mean, median, min, max, standard deviation).
* Plot simple visualizations: histograms (for distribution of features), boxplots (to check for outliers)

1. Data Cleaning

* Check for and handle missing data (included in EDA phase)
* Encode categorical variables (Sex, Smoker, Region using One-Hot Encoding)
* Remove or correct obvious outliers if they are errors.

1. Feature Selection,

* Figuring out: Which variables actually matter for predicting charges. Use methods like correlation matrix (for numerical variables), SelectKBest and visualizations (boxplots comparing smoker/non-smoker charge

1. Model Building

* Since charges are continuous and the question explicitly asked for Linear Regression, that is what I’ll do.

1. Model Evaluation

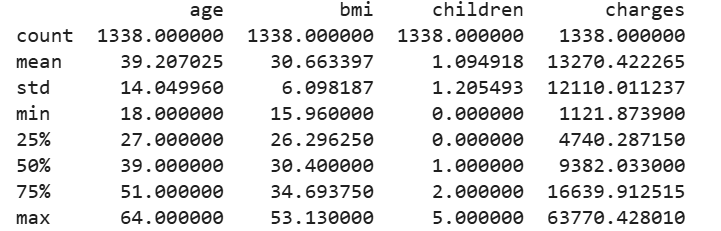
* I'd split the data: Train/Test split 80/20.
* Measure model performance: R² score (how much variation is explained), RMSE / MAE (error metrics).
* Predicted vs Actual charges (scatterplot).

1. Final Insights

* Write a summary
* Suggest real-world actions based on findings

**EDA**

From the output we learn that the dataset contains 1338 rows and 7 columns, no missing values, 3 categorical variables which are smoker, region and sex. The numeric variables are age, bmi, children and charges. All other variables are independent variables and charges is the target variable. Here is screenshot of the statistics:

****I also created histograms for numerical variables to visualise the type of distribution of numerical values. BMI is the closest to normal distribution out of all numerical variables. Pie charts to compare the proportions of smoker vs non smokers, residents in each region, male and female.

**Data Cleaning**

I cleaned the data set by removing outliers, checking for missing data encoding categorical variables

**Feature Selection**

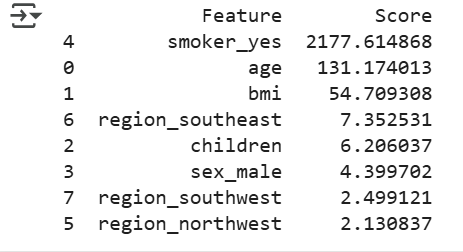
Belcic and Stryker(n.d) state that feature is the process of identifying and choosing the most important variables (or "features") in your dataset that contribute the most to predicting your target variable, in our case medical charges.

For feature selection we used SelectKbest(with a scoring function) and Correlation Matrix(exclusive of categorical features)

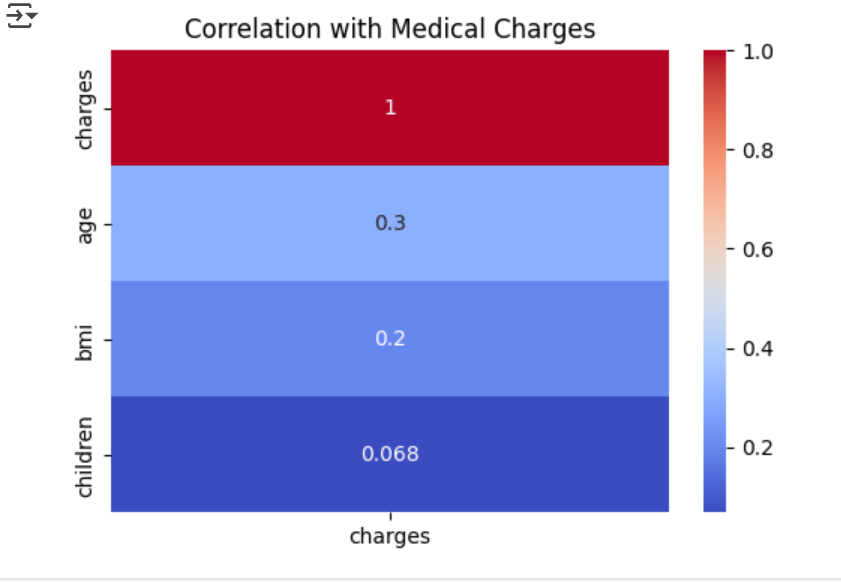
Output- SelectKbest

The scores are laid out in ascending order.

* Smoker is a very strong predictor therefore the most important feature. Being a smoker massively affects medical charges.
* Age and BMI are moderate predictors
* lastly children and region are weak predictors.



Correlation Matrix Output



* **charges vs charges = 1**:
  + Perfect correlation. A variable is always perfectly correlated with itself.
* **age has a correlation of 0.3 with charges age is somewhat important for predicting charges:**
  + This is a moderatepositive correlation.
  + As age increases, medical charges tend to increase.
* **bmi has a correlation of 0.2 with charges BMI is a little important for predicting charges**:
  + A weak positive correlation.
  + Higher BMI (Body Mass Index) slightlytends to mean higher medical charges.

In the end I ended up using all features because some features might have a high correlation but not necessarily the best predictor and some features may have non-linear effects, interaction effects, or become important only when combined with others

* **children has a correlation of 0.068 with charges, number of children is not really important for predicting charges:**
  + Very weak (almost no) correlation.
  + The number of children a person has doesn't strongly affect their medical charges.

**Model Building**

Created a linear regression

**Model Evaluation**

I have split the model in 80/20 proportions

A number and equation on a white background

AI-generated content may be incorrect.

Intercept: -11,931.22

This is the baseline cost when all features are 0. Since it's hard to interpret in a real-world sense, we will not focus on although it's part of the prediction formula.

Coefficients

These represent how much the charges(we’ll assume it’s USD) change with a one-unit increase in each corresponding feature, holding others constant.

* age: 256.98 Each additional year of age increases charges by about $257.
* bmi: 337.09 Each BMI unit adds about $337.
* children: 425.28 Each child adds around $425 to charges.
* sex\_male: -18.59 Males pay slightly less, but the difference is negligible.
* smoker\_yes: 23,651.13 Smoker status adds a massive charge increase .
* Regions (negative coefficients) Compared to the base category (likely northeast, which was dropped due to drop\_first=True), these other regions slightly reduce charges.

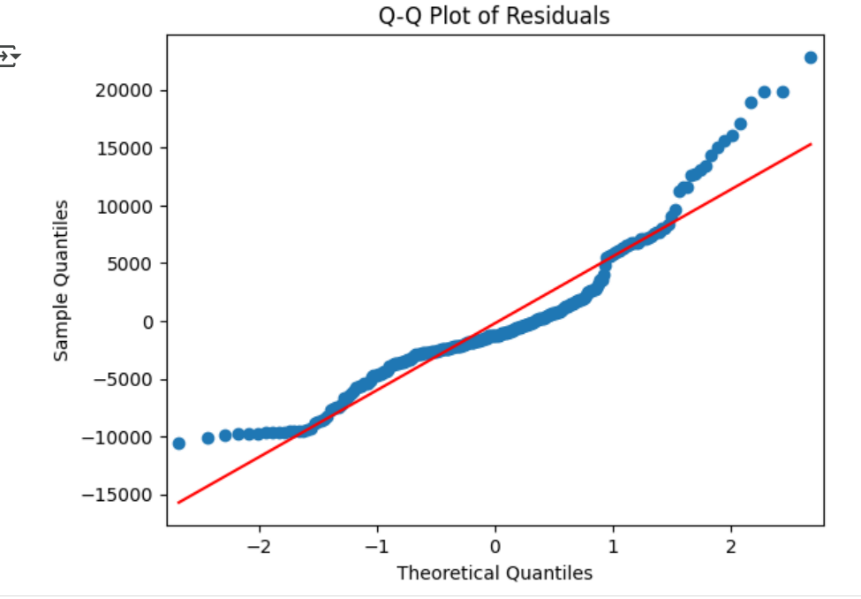
Homoscedasticity

Homoscedasticity (constant variance of residuals): The Residual vs Predicted plot exhibits heteroscedasticity, meaning that the variance of the residuals is not constant. The residuals increase in magnitude as the predicted values increase, suggesting that the model makes larger errors at higher values. This violates the assumption that the variance of the residuals should remain constant across all levels of the independent variable.

Multicollinearity

Multicollinearity: The Variance Inflation Factor (VIF) value is less than 5, indicating that there is no significant multicollinearity in the model. This suggests that the independent variables are not highly correlated with each other and each variable has an independent effect on the outcome.

Quantile-Quantile(Q-Q) plot

Ford(2015) states that a Q-Q plot is a scatter plot used to help us assess what theoretical distribution a set of data came from. Ideally all point would align along the 45 degree line as for a perfect linear regression residuals(errors) are normally distributed.  
  


What this Q-Q plot shows:

* Early left side (lower quantiles):  
  The points are below the line — suggesting negative skew (residuals more negative than expected at the low end).
* Middle section (around 0):  
  The points are reasonably close to the line, but with slight deviation — not terrible, but not perfect either.
* Right tail (higher quantiles):  
  Points bend sharply above the line — strong positive skew in the residuals at the higher end (some extreme positive errors).

Interpretation of Q-Q:

* Residuals are not perfectly normally distributed.
* There's some skewness — especially at the higher charge predictions.
* The model underpredicts for very high medical charges.
* Potential reasons could be:
  + Presence of outliers (very expensive medical cases).
  + Non-linear relationships not captured well by a simple linear model.
  + Some missing variables that explain the high charges

Our model is good but not perfect.

Retraining model with different parameters



Plotting Actual Against Predicted Values

**Summary and Final Thoughts**

Based on the metrics we used we can conclude that our model is good but not perfect. Among the features in the dataset smoking status seems to have the highest impact on charges and region, so the insurance company should charge smokers more as medical expenses increase if one is a smoker. BMI and Children seem to have a positive moderate effect on medical expenses. The Pie charts also show that smokers while they are make up less than 21% of the dataset they contribute about 49.5% of medical charges. did not use hyperparameters as linear Regression e.g. Ridge, Lasso. My focus was on understanding the data, building a clean model, and evaluating its performance.

References

[1] A. Kruger, “Statistics and Data Visualization | Q-Q plots,” *YouTube*, 2022. [Online]. Available: <https://www.youtube.com/watch?v=Ek4fdRwBejc>

[2] University of Virginia Library, “Understanding Q-Q Plots,” *UVA Library*. [Online]. Available: <https://library.virginia.edu/data/articles/understanding-q-q-plots>

[3] Alireza151, “AI Course - Session 4 - Visualization,” *Kaggle Notebooks*, 2023. [Online]. Available: <https://www.kaggle.com/code/alireza151/ai-course-session-4-visualization>

[4] AnalyticaObscura, “Medical Cost Analysis,” *Kaggle Notebooks*, 2023. [Online]. Available: <https://www.kaggle.com/code/analyticaobscura/medical-cost-analysis>

[5] A. Singh, “Hyperparameter Tuning: Beyond the Basics,” *Medium*, 2023. [Online]. Available: <https://medium.com/@abhaysingh71711/hyperparameter-tuning-beyond-the-basics-34d36b014482>

[6] “Linear Regression | What is Regression | Statistics & Probability | Don't Memorise,” *YouTube*, 2021. [Online]. Available: <https://www.youtube.com/watch?v=ntBa7YKc9XM>

[7] “Linear Regression in Python | Machine Learning Tutorial,” *YouTube*, 2021. [Online]. Available: <https://www.youtube.com/watch?v=iJ5c-XoHPFo>

[8] “Understanding Feature Selection | Python ML Tutorial,” *YouTube*, 2023. [Online]. Available: <https://www.youtube.com/watch?v=VBOoiR0uOck>