Final Data Scientist Challenge

Task: We need to predict the medical charges for an insurance company using Linear regression. A dataset called insurance is provided.

a. Find the summary statistics of the variable charges in the dataset
b. Display a table that contains the number of people in each region
c. Visualize the relationship among all features using a scatterplot matrix
d. Train a model on the data
e. Evaluate the model performance
f. Improve the model performance by Adding nonlinear relationship (Hint consider only age as input)

Since smoking and obesity may have a harmful impact, we assume the combination of the two may be worse. Build a model with the interaction effects of smokers and obesity (Obesity is considered for BMI >30)

- Improve your regression model by putting a, and b together

Evaluation criteria

- Data Understanding and Exploration
- Code Quality and Documentation
- Visualisation and Feature Relationships
- Model Performance and Evaluation