The Regression Dataset

The dataset I'm using for this regression task is: Medical Cost (https://www.kaggle.com/datasets/mirichoi0218/insurance). The goal of this dataset is to predict the dependent variable charges (how much treatment cost) based on the independent variables: age, sex, bmi, children, smoker and region.
In this notebook I try to conclude with which variables have a strong correlation with charges and if they should be used in making a model.

The Final Dataset

Data cleaning:

This is where I remove columns I know will not be a part of the model, fix null-values, encode categorical values, and normalize attributes.
At first glance I couldn't see any attributes I know I should remove and therefor I wait with this step until after the Feature Correlation Analysis. The dataset didn't contain any null-values and this step was then skipped.

Encoding

I changed the datatypes of the columns Region, BMIGroup and AgeGroup to category, so I could then one-hot encode them.

Normalization

I chose to crate bins containing Age and BMI intervals and then one-hot encoded these intervals.
I also created two function; min-max-normalizing and z-score-normalizing just to have the opportunity to check the possible effects of this on a model

EDA

Feature Correlation Analysis

I see great correlations between the dependent variable charges and the independent variables: smoker, bmi and age. There is also a good correlation between the region southeast and charges, but this is due to the more unhealthy population (more smokers and overweight people). However, the variables sex and children didn’t seem to effect the charges by a significant amount.

Feature Selection

Normally these variables (sex and children) would most likely be removed in bigger datasets due to the computational cost, but since this is such a small dataset, I see no reason to do this.