Original Insurance data dictionary

Source: https://www.kaggle.com/datasets/fibonamew/insurance-data

This data was found originally on the web from Kaggle. Copyright provisions are not well defined as this is publicly available data once you create a Kaggle account.

Insurance data for my final project in biostatistics 1 with Dr. Gaddis. The data will be used in hypothesis testing.

This dataset is useful in understanding variables important in determining insurance premium(s) for prospective insurance purchasers.

File: original_insurance.csv

About this file: The insurance.csv dataset contains 1033 tuples (rows) and 7 attributes (columns). The dataset contains 4 numerical attributes (age, bmi, children and expenses) and 3 nominal attributes (sex, smoker and region) that were converted into factors with numerical value designated for each level. The data is collected from 4-part regions in the United States. It has 1033 tuples and 7 attributes originally but I added an 8th field unique_id for ease of use in a table joining if needed. Attributes are named:

1. unique_id (unique identifier),
2. age (in years),
3. sex (male vs female),
4. bmi = body mass index (kg/cm^2),
5. children = number of children (no unit),
6. smoker (smoking status reported as yes or no),
7. region (region of origin as in southeast, northwest etc),
8. expenses (patient health expenses in dollars)

Any extra attributes result from my data analysis and interpretation for setup in hypothesis testing and exploratory data analysis. This version of the data does not contain my modifications but modifications will be provided when my analysis is done.

Proposed Tasks:

1. hypothesis testing
2. Statistical Modeling
3. Exploratory Data Analytics