# Modified 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)
9. BMI_codeverbose is a transformation of BMI numeric data into categories using the CDC guidelines for adult BMI interpretation,
10. BMI_codenum is my conversion of BMI_codeverbose into numeric identifiers for the BMI_codeverbose,
11. Family_size is my interpretation of the numeric variable 5 above (children) into categories for purposes of analysis in SPSS. Family_size encodes the numeric variable children into 3 levels of “small family size” for patients with 0 kids, “normal family size” for patients with 1<=kids<2,”large family size” for patients with >3 kids,
12. Famsize_num is the conversion of Family_size into numeric identifiers for analysis in SPSS. 1 = small family size, 2= normal family size and 3=large family size,

13. Region_num is conversion of the string region into numeric identifiers for analysis in SPSS. 1 = southwest,2=northwest,3=southeast,4=northeast,
14. Family_size_recode is my interpretation of the numeric variable 5 above (children) into categories for purposes of analysis in SPSS. Family_size_recode encodes the numeric variable children into 2 levels of “normal family size” for patients with 0<=kids<2, and” large family size” for patients with >3 kids,
15. Family_size_recode_sex is very much identical to Family_size (variable 11) but this time there are 4 levels resulting from separating the small family size into male and female subgroups. The aim of this variable was to get to the root of the bimodal distribution of the small family size,
16. Family_size_recode_sex_num is simply the assignment of numeric identifiers to the levels encoded in Family_size_recode_sex,
17. Family_size_recode_smoker is very much identical to Family_size (variable 11) but this time there are 4 levels resulting from separating the small family size into smoker and non-smoker subgroups. The aim of this variable was to get to the root of the bimodal distribution of the small family size,
18. Family_size_recode_smoker_num is simply the assignment of numeric identifiers to the levels encoded in Family_size_recode_smoker

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