Juan Ignacio Galvalisi

Exercise 6.1 – Sourcing Open Data

Medical Cost Personal

Executive Summary

This data is a pratical is used in the book Machine Learning with R by Brett Lantz; which is a book that provides an introduction to machine learning using R. This dataset is in the public domain and explains the cost of a small sample of USA population Medical Insurance Cost Personal based on different attributes.

Data Sources

The data for this project is an open-source data downloaded from Kaggle according to the following resource: https://www.kaggle.com/datasets/mirichoi0218/insurance.

Limitation and Data Ethics

- There is no data about regions, cities, and postal codes, useful to identify more precisely where the USA population is settled.
- There is no information regarding previous years to compare whether there are variations over time.
- The dataset does not contain any name related to the data of each person, so it accomplishes with the General Data Protection Regulation (GDPR).

Data Cleaning and Data Consistency Checks

- Change data types
- Check numerical variables
- Looking for missing data
- Looking for duplicate data.

Data Profile

The dataset has 7 columns and 1338 rows. After the data wrangling and consistency check, the dataset contains 7 columns and 1337 rows.

Column Details

Column	Description	Quantitative/ Qualitative	Type	Time
age	Age of the primary beneficiary	Qualitative	Ordinal	Variant
sex	Insurance contractor gender	Qualitative	Nominal	Invariant
bmi	Body mass index	Quantitative	Continuous	Invariant
children	Number of dependents	Quantitative	Discrete	Variant
smoker	If he/she is a smoker or not	Qualitative	Nominal	Invariant
region	Residential area in the US	Qualitative	Nominal	Invariant
charges	Individual medical costs	Quantitative	Discrete	Invariant

Questions to Inquire

- What is the worst age range for insurance charges?
- Being younger implies a lower or higher risk translated into costs? And the elderly people?
- Is having a high degree of body mass index an indicator of risk?
- Being a woman or a man implies a greater or lesser risk?
- Which region has the highest costs? Are any relevant correlations between one region and another concerning a more significant number of the elderly population?
- Is being a smoker an indicator of an increase in charges?
- Does having children make the number of charges go up or down?