Data analytics using R

Digvijay Singh
digvijay.singh@iitb.ac.in

R Team, FOSSEE,
Indian Institute of Technology, Bombay

05-03-2022

Topics to cover

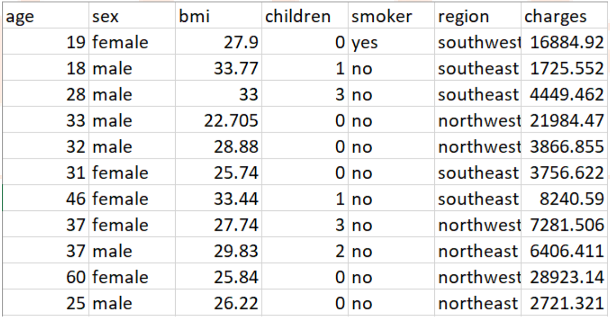
1. Insurance dataset
2. Objective
3. Statistical data analysis - I
4. Statistical data analysis - II

Insurance dataset

1.1 Data extracted from the chapter 6 of the book **Machine Learning with R** by *Brett Lantz*.

1.2 Data was originally created by the **U.S. Census Bureau**.

1.3 Dataset includes 1,338 examples of beneficiaries currently enrolled in an insurance plan.



age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552
28	male	33	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.47
32	male	28.88	0	no	northwest	3866.855
31	female	25.74	0	no	southeast	3756.622
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.506
37	male	29.83	2	no	northeast	6406.411
60	female	25.84	0	no	northwest	28923.14
25	male	26.22	0	no	northeast	2721.321

Figure 1: Glimpse of the Insurance dataset.

Insurance dataset

1.4 Variables contained in the dataset (3 continuous & 4 discrete)

- **age** *(continuous)*: Age of the primary beneficiary (excluding those above 64 years).
- **sex** *(discrete)*: Policy holder's gender.
- **bmi** *(continuous)*: The body mass index (BMI) of policy holder.
- **children** *(discrete)*:The number of children / dependents covered by the insurance plan.
- **smoker** *(discrete)*: Indicates whether the insured regularly smokes tobacco or not.
- **region** *(discrete)*: The beneficiary's place of residence in the U.S. in terms of geographic region.
- **charges** *(continuous)*: Yearly medical expenses of each individual.

Objective

2. To find meaningful patterns in the data, especially those associated with the `charges` variable.

Statistical data analysis - I

1. Let's switch to RStudio.

Statistical data analysis - II

1. Let's switch to RStudio.

Thank You!