Data Source:

External. Obtained from the website Kaggle

https://www.kaggle.com/datasets/mirichoi0218/insurance

Data Collection:

Administrative Data and Survey: Collected primarily through insurance companies.

Data Contents:

Summary**:** The dataset contains information about individuals' medical costs and related personal characteristics.

Variables (Columns):

- age: Age of the primary beneficiary.
- sex: Gender of the insurance contractor (female, male).
- bmi: Body mass index.
- children: Number of children/dependents covered by insurance.
- smoker: Smoking status (yes/no).
- region: Beneficiary's residential region in the US (northeast, southeast, southwest, northwest).
- charges: Individual medical costs billed by health insurance.

Limitations:

Limitations might include how the data was collected (e.g., self-reported data might have biases), the representativeness of the sample, and potential missing values.

Relevance:

Scenario 1: Project Objective: Predict medical costs.

- Hypothesis: Smoking status is a strong predictor of medical costs.
- Relevance:
    - age, sex, bmi, children, smoker, region: *All are potentially relevant*. These variables could influence medical costs. smokers are particularly relevant to the hypothesis.
    - Charges: This is the dependent variable

Scenario 2: Project Objective: Investigate regional differences in medical costs.

- Hypothesis: Medical costs are higher in the Northeast region compared to other regions.
- Relevance:
    - Region: This is the key variable for this objective.
    - charges: This is what you're comparing across regions.
    - age, sex, bmi, children, smoker: *Potentially relevant*.

Scenario 3: Project Objective: Analyze the impact of family size on medical costs.

- Hypothesis: The number of children/dependents is positively correlated with medical costs.
- Relevance:
    - Children: This is the independent variable of interest.
    - Charges: This is the dependent variable.
    - age, sex, bmi, smoker, region:

Project Goals:

Identify the key factors that contribute to high medical costs.

Understand regional variations in medical costs.

Data Profile:

Data Profile for df_clean

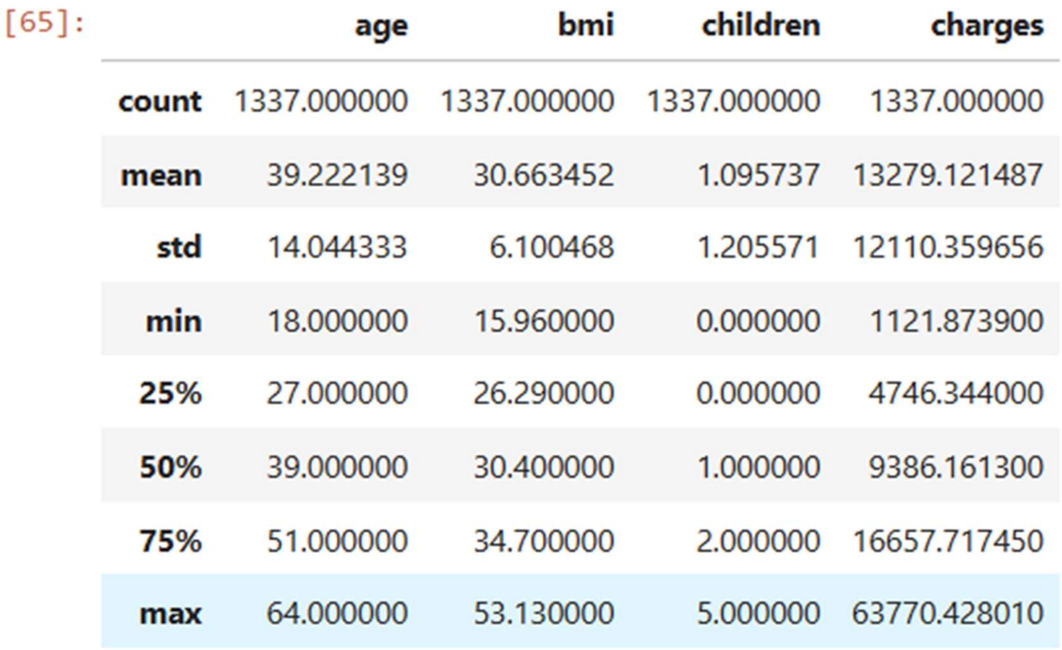
1. Variables and Data Types:

- age: Quantitative, time-invariant, discrete
- sex: Qualitative, time-invariant, nominal
- bmi: Quantitative, time-invariant, continuous
- children: Quantitative, time-invariant, discrete
- smoker: Qualitative, time-invariant, binary
- region: Qualitative, time-invariant, nominal
- charges: Quantitative, time-variant, continuous

2. Data Integrity Issues:

- age:
  - Check for negative ages (impossible).
  - Check for unreasonably high ages (e.g., > 120).
- sex:
  - Check for consistent categories (e.g., "male," "female," or other specified categories). Look for typos or inconsistencies (e.g., "Male", "MALE").
- bmi:
  - Check for values less than or equal to 0 (impossible).
  - Check for extremely high BMI values (e.g., > 60), which might be errors or represent extreme cases that need special handling.
- children:
  - Check for negative numbers of children
  - Consider very high numbers of children; while possible, they might be outliers.

- smoker:
    - Check for consistent categories. Look for typos or inconsistencies.
- region:
    - Check for the correct US regions (northeast, southeast, southwest, northwest). Look for typos or inconsistencies.
- charges:
    - Check for negative charges.
    - Check for unusually high charges, which might be errors or represent extreme cases.



[65]:

	age	bmi	children	charges
count	1337.000000	1337.000000	1337.000000	1337.000000
mean	39.222139	30.663452	1.095737	13279.121487
std	14.044333	6.100468	1.205571	12110.359656
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.290000	0.000000	4746.344000
50%	39.000000	30.400000	1.000000	9386.161300
75%	51.000000	34.700000	2.000000	16657.717450
max	64.000000	53.130000	5.000000	63770.428010

Data cleaning:

Duplicate row found and removed.

Questions:

Hypothesis: Smoking status is a strong predictor of medical costs.

- How does smoking prevalence vary across different demographic groups?
- Within the smoker group, are there other factors (like bmi or children) that further influence the charges, and do these factors interact with sex?
-

Hypothesis: Medical costs are higher in the Northeast region compared to other regions.

- Are we comparing *average* charges across regions, or total charges for each region?

Hypothesis: The number of children/dependents is positively correlated with medical costs.

- Are there specific age groups where the correlation between children and charges is stronger?
- Does the relationship between children and charges vary by sex or smoking status?