Exercise 6.1 Project Information Brief Insurance

Data

1 Datasource

Data source and collection method. Kaggle is a collaoration platorm for data scientists and analysts. The chosen dataset is the insurance.csv obtained on Kaggle via https://www.kaggle.com/datasets/mirichoi0218/insurance/data and links bank to a learning data set provided by GitHub. The information on Kaggle states that the dataset is a synthesized dataset that was updated around 7 years ago. This informs us that this datasource is not **trustworthy** for real-life health policy implications. In addition, no timeframe was given for the dataset.

Content. The dataset has a total of 1338 Rows with 7 variable columns and gives information on how following health and location variables might impact health insurance expense claims.

- **age**: age of primary beneficiary
- **sex**: insurance contractor gender, female, male
- **bmi**: body mass index, providing an understanding of body, weights that are relatively high or low relative to height
- **objective index of body weight** (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9
- **children**: number of children covered by health insurance / number of dependents
- **smoker**: smoking, yes or no.
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges**: Individual medical costs billed by health insurance

Limitations: The **synthetic** nature of the dataset and the absence of a defined time period limit its applicability for real-world policy recommendations. Any insights derived should be interpreted cautiously, as they may not fully reflect actual health insurance trends or patient behaviors. Additionally, biases in data generation may skew results, leading to misleading conclusions if applied without validation against real-world data.

It is noteworthy, that **BMI** alone is not a sufficient indicator of health, physical fitness or body composition. It is known to be oversimplified and only gives a

The binary variable for **smoking** is also simplified. For health it also matters how much is smoked, though this is admittedly hard to measure, since the most common data on this would be self-reported and underly self-reporting bias. For simplification of analysis a binary variable is sufficient

Ethical considerations.

- While the dataset does not contain personally identifiable information (PII), there remains a risk of bias, which could contribute to **discrimination in insurance pricing or policy decisions.** Health insurers must ensure that predictive models do not disproportionately impact vulnerable populations or reinforce existing disparities.

Furthermore, transparency in data use and adherence to ethical AI practices are critical to maintaining fairness and public trust in insurance decision-making.
- Considering the **limitations of BMI and smoking** variables mentioned before, health insurance providers need to treat this information with caution and acknowledge that both variables are strongly oversimplified.

Choice of dataset. I acknowledge that the dataset is both older and fictional; however, I chose to work with it due to its relevant content and potential for gaining initial insights into the health insurance domain. Given its limitations, I consider this a training dataset to familiarize myself with analyzing health-related data.

Research context. I chose this dataset due to the ongoing discussions surrounding rising healthcare costs and insurance affordability. As healthcare expenditures continue to increase globally, analyzing health insurance data provides valuable insights into cost drivers, risk assessment, and potential areas for policy improvement. This relevance makes the dataset particularly useful for exploring broader industry challenges.

There are many common preconceptions about individuals with health conditions, particularly the assumption that factors like high age or BMI directly correlate with increased health risks, such as obesity or cardiovascular disease. While these factors can contribute to overall health risk, I want to examine whether there is supporting evidence in data. Understanding these nuances is crucial for developing fair and accurate insurance models that do not rely on oversimplified risk assumptions.

2 Data profile

Data cleaning

Data was checked for missing values, duplicates, mixed-type data, and correct data type. No deviations were found.

Both categorical and numerical data were checked for inconsisteny, no inconsistencies present.

The columns have correct column names and datatypes.

Understanding data - Descriptive analysis

[24]:

	age	bmi	children	charges
count	1338.000000	1338.000000	1338.000000	1338.000000
mean	39.207025	30.663397	1.094918	13270.422265
std	14.049960	6.098187	1.205493	12110.011237
min	18.000000	15.960000	0.000000	1121.873900
25%	27.000000	26.296250	0.000000	4740.287150
50%	39.000000	30.400000	1.000000	9382.033000
75%	51.000000	34.693750	2.000000	16639.912515
max	64.000000	53.130000	5.000000	63770.428010

Age. Youngest patients are 18 years old (old enough to have their own health insurance), older are 64 years old. Mean= Median, points to a normal distribution.

BMI. Mean=median at 30 meaning that the average patient is to be considered obese (with a normal distribution).

The Body Mass Index (BMI) categories are defined as follows by CDC:

- **Underweight:** BMI less than 18.5
- **Normal weight:** BMI 18.5–24.9
- **Overweight:** BMI 25–29.9
- **Obesity:** BMI 30 or greater

https://www.cdc.gov/bmi/adult-calculator/bmi-categories.html

Children. Normal distribution with mean=median, average patient has 1 child, the max. is 5 children.

Charges. Range between $1121-63770 which is awide range, there are large values but the are reasonable hence we keep them. The distribustion is right-skewed. The distribution is reasonable because:

Health insurance claims often have:

- Many small/medium claims (doctor visits, minor treatments).
- A few very expensive claims (surgeries, chronic diseases).
- No extreme negative values (claims can't be negative).

Questions to explore

Is there evidence for the common assumption for positive linear relationship of bmi, age, smoking to health costs?

How does do the health variables (smoking, BMI) differ between gender? Do these differnces also translate to different health charges in gender?

How is obesity (BMI) connected to smoking? Is there a positve linear relationship?

What types of regional differences can be detected?

Task Description

1. If you haven't done so already, download your Achievement 6 project brief (.pdf).
2. The data you use for your project will need to meet certain criteria as defined in the brief. Read through the data requirements now to be sure the data you choose is appropriate.
3. **Source your data.** Use the requirements (and your own interests) to source an open data set from the web. We introduced you to several sources in this Exercise but feel free to look elsewhere, as well.
4. Create a new document to detail your project information.
5. Create a "Data Source" section in your project document and provide the following information:
    - A summary of your data source. We recommend you revisit Exercise 1.4: Sourcing the Right Data for a recap on what to include in your summary.
    - An explanation for why you've chosen this data set.

6. **Clean your data.** Conduct some basic data cleaning and consistency checks in Jupyter to ensure your data is ready for further analysis.
7. **Understand your data.** Develop a basic understanding of your data set by reviewing the variables and performing basic descriptive statistical analysis. You might want to make a data profile similar to what you did in Achievement 1.
8. **Consider limitations and ethics.** Outline any limitations and ethical considerations presented by the content of your data, its source, and/or how it was collected.
9. Include the results of steps 6 to 8 in a second section of your project document. This second section can be titled something like “Data Profile.”

10. **Define questions to explore.** In a third section of your project document, define a list of questions to explore with your analysis. As mentioned in the Exercise, you may want to revisit Exercise 1.2: Starting with Requirements for a recap on writing good questions.
11. Submit your project document and Jupyter notebook to your tutor for review.

Achievement 6 milestones

1. Data sourcing, cleaning
2. Exploratory analysis
3. Geographical visualisations
4. Regression with supervised machine learning
5. Unsupervised machine learning Clustering
6. Time series data
7. Data dashboards

8. GitHub Repository:
    - Clean scripts