The Sure Tomorrow insurance company wants to solve several tasks with the help of Machine Learning, and you are asked to evaluate that possibility.

* Task 1: Find customers who are similar to a given customer. This will help the company's agents with marketing.
* Task 2: Predict whether a new customer is likely to receive an insurance benefit. Can a prediction model do better than a dummy model?
* Task 3: Predict the number of insurance benefits a new customer is likely to receive using a linear regression model.
* Task 4: Protect clients' personal data without breaking the model from the previous task.

It's necessary to develop a data transformation algorithm that would make it hard to recover personal information if the data fell into the wrong hands. This is called *data masking*, or *data obfuscation*. But the data should be protected in such a way that the quality of machine learning models doesn't suffer. You don't need to pick the best model, just prove that the algorithm works correctly.

**Project Instructions**

1. Load the data.
2. Check that the data is free of issues — there is no missing data, extreme values, and so on.
3. Work on each task and answer the questions posed in the project template.
4. Draw conclusions based on your experience working on the project.

There is some precode in the project template, feel free to use it, some precode needs to be finished first. Also, there are two appendices in the project template with useful information.

**Data Description**

The dataset is stored in file /datasets/insurance\_us.csv. [You can download the dataset here.](https://practicum-content.s3.us-west-1.amazonaws.com/datasets/insurance_us.csv)

* **Features:** insured person's gender, age, salary, and number of family members.
* **Target:** number of insurance benefits received by an insured person over the last five years.