*Analyzing borrowers’ risk of defaulting*

**Project Description**

This is a project to prepare a report for a bank’s loan division. The answer that I am looking for is if a customer’s marital status and number of children has an impact on whether they will default on a loan. The data has some record of a bank customers’ credit worthiness.

**Describing data**

*children*: the number of children in the family

*days\_employed*: how long the customer has worked

*dob\_years*: the customer’s age

*education*: the customer’s education level

*education\_id*: identifier for the customer’s education

*family\_status*: the customer’s marital status

*family\_status\_id*: identifier for the customer’s marital status

*gender*: the customer’s gender

*income\_type*: the customer’s income type

*debt*: whether the client has ever defaulted on a loan

*total\_income*: annual income

*purpose*: reason for taking out a loan

**Instructions for completing the project**

**Task 1. Conduct research**

**Step 1**. Open the data file /datasets/credit\_scoring\_eng.csv and have a look at the general information. [Download dataset](https://code.s3.yandex.net/datasets/credit_scoring_eng.csv).

**Step 2**. Preprocess the data:

—Identify and fill in missing values.

—Replace the real number data type with the integer type.

—Delete duplicate data.

—Find lemmas for values in the column listing reasons for taking out a loan.

—Categorize the data.

As you complete **task 1**, describe what you’re doing in the *markdown* cells.

Explain:

—which missing values you identified;

—possible reasons these missing values were present;

—which method you used to fill in missing values;

—which method you used to find and delete duplicate data and why;

—possible reasons why duplicate data was present;

—which method you used to change the data type and why;

—how you used lemmatization on the reasons for getting a loan;

—which dictionaries you've selected for this data set and why.

Format:

Complete **task 1** in the *Jupyter Notebook.* It opens when you click Next.

**Task 2. Present your conclusions**

**Step 1.** Answer these questions:

—Is there a relation between having kids and repaying a loan on time?

—Is there a relation between marital status and repaying a loan on time?

—Is there a relation between income level and repaying a loan on time?

—How do different loan purposes effect on-time repayment of the loan?

Interpret your answers. Explain what the results you obtained mean.

**Step 2.** Write an overall conclusion.

Format:

Present **task 2** as a coherent text in the *markdown* cell at the end of your *Jupyter Notebook* project. Use formatting and headings.

**How will my project be assessed?**

Your mentor will use assessment criteria to evaluate your project. Before getting started, read these criteria carefully. They will help you ensure you’re doing a good job.

Here’s what mentors look at when assessing your project:

—How do you describe the problems identified in the data?

—What methods do you use to replace data types and to process missing values and duplicate data?

—Have you used lemmatization effectively?

—Did you categorize the data? Why did you do it the way you did?

—Did you export the final data into pivot tables?

—Did you use the *try–except* clauses in your code to handle any unexpected errors?

—Did you follow the correct project structure and write clean code?

—Did you draw accurate and useful conclusions?

—Did you leave comments on each step you took?

The knowledge and skills you acquired in previous lessons will help you successfully complete your project.