## ****Описание проекта****

Заказчик — кредитный отдел банка. Нужно разобраться, влияет ли семейное положение и количество детей клиента на факт погашения кредита в срок. Входные данные от банка — статистика о платёжеспособности клиентов.

Результаты исследования будут учтены при построении модели кредитного скоринга — специальной системы, которая оценивает способность потенциального заёмщика вернуть кредит банку.

## ****Инструкция по выполнению****

**Шаг 1. Откройте таблицу и изучите общую информацию о данных**

Путь к файлу: /datasets/data.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/data.csv)

**Шаг 2. Предобработка данных**

1. В двух столбцах есть пропущенные значения. Один из них — days\_employed. Пропуски в этом столбце вы обработаете на следующем этапе. Найдите другой столбец и заполните пропущенные значения в нём медианным значением по столбцу:
   * опишите, какие пропущенные значения вы обнаружили;
   * проверьте, какую долю составляют пропущенные значения в каждом из столбцов с пропусками;
   * приведите возможные причины появления пропусков в данных;
   * объясните, почему заполнить пропуски медианным значением — лучшее решение для количественных переменных.
2. В данных могут встречаться артефакты (аномалии) — значения, которые не отражают действительность и появились по какой-то ошибке. Например, отрицательное количество дней трудового стажа в столбце days\_employed. Для реальных данных это нормально. Обработайте значения в столбцах с аномалиями и опишите возможные причины появления таких данных. После обработки аномалий заполните пропуски в days\_employed медианными значениями по этому столбцу.
3. Замените вещественный тип данных в столбце total\_income на целочисленный, например, с помощью метода astype().
4. Если в данных присутствуют строки-дубликаты, удалите их. Также обработайте неявные дубликаты. Например, в столбце education есть одни и те же значения, но записанные по-разному: с использованием заглавных и строчных букв. Приведите их к одному регистру. Проверьте остальные столбцы.

После удаления дубликатов сделайте следующее:

* + поясните, как выбирали метод для поиска и удаления дубликатов в данных;
  + приведите возможные причины появления дубликатов.

1. Создайте два новых датафрейма, в которых:
   * каждому уникальному значению из education соответствует уникальное значение education\_id — в первом;
   * каждому уникальному значению из family\_status соответствует уникальное значение family\_status\_id — во втором.

Удалите из исходного датафрейма столбцы education и family\_status, оставив только их идентификаторы: education\_id и family\_status\_id. Новые датафреймы — это те самые «словари» (не путайте с одноимённой структурой данных в Python), к которым вы сможете обращаться по идентификатору.

1. На основании диапазонов, указанных ниже, создайте столбец total\_income\_category с категориями:
   * 0–30000 — 'E';
   * 30001–50000 — 'D';
   * 50001–200000 — 'C';
   * 200001–1000000 — 'B';
   * 1000001 и выше — 'A'.

Например, кредитополучателю с доходом 25000 нужно назначить категорию 'E', а клиенту, получающему 235000, — 'B'.

1. Создайте функцию, которая на основании данных из столбца purpose сформирует новый столбец purpose\_category, в который войдут следующие категории:
   * 'операции с автомобилем',
   * 'операции с недвижимостью',
   * 'проведение свадьбы',
   * 'получение образования'.

Например, если в столбце purpose находится подстрока 'на покупку автомобиля', то в столбце purpose\_category должна появиться строка 'операции с автомобилем'.

Вы можете использовать собственную функцию и метод apply(). Изучите данные в столбце purpose и определите, какие подстроки помогут вам правильно определить категорию.

**Шаг 3. Ответьте на вопросы**

Ответы на вопросы можно разместить в ячейках тетрадок Jupyter Notebook с типом markdown.

* Есть ли зависимость между количеством детей и возвратом кредита в срок?
* Есть ли зависимость между семейным положением и возвратом кредита в срок?
* Есть ли зависимость между уровнем дохода и возвратом кредита в срок?
* Как разные цели кредита влияют на его возврат в срок?

Ответы сопроводите интерпретацией — поясните, о чём именно говорит полученный вами результат.

**Шаг 4. Напишите общий вывод**

**Оформление:** Задание выполните в Jupyter Notebook. Программный код заполните в ячейках типа code, текстовые пояснения — в ячейках типа markdown. Примените форматирование и заголовки.

## ****Описание данных****

* children — количество детей в семье
* days\_employed — общий трудовой стаж в днях
* dob\_years — возраст клиента в годах
* education — уровень образования клиента
* education\_id — идентификатор уровня образования
* family\_status — семейное положение
* family\_status\_id — идентификатор семейного положения
* gender — пол клиента
* income\_type — тип занятости
* debt — имел ли задолженность по возврату кредитов
* total\_income — ежемесячный доход
* purpose — цель получения кредита

## ****Как будут проверять мой проект?****

На что обращают внимание при проверке проектов:

* Как вы описываете найденные в данных проблемы?
* Какие методы замены типов данных, обработки пропусков и дубликатов применяете?
* Категоризируете ли данные? Почему именно таким образом?
* Выводите ли финальные данные в сводных таблицах с помощью метода pivot\_table()?
* Применяете ли конструкцию try...except... для обработки потенциальных ошибок?
* Соблюдаете ли структуру проекта и поддерживаете аккуратность кода?
* Какие выводы делаете?
* Оставляете ли комментарии к шагам?