Задачи

1. Загрузите [датасет](https://yadi.sk/d/20P4PUpwkigr2g) transaction\_data.csv. Проверьте размер таблицы, типы переменных, число пропущенных значений и описательную статистику.
2. Какие значения находятся в колонке transaction? Сколько наблюдений относятся к тому или иному уровню? Визуализируйте результат с помощью барплота. Подумайте, как можно улучшить график.
3. Сколько транзакций завершились ошибкой?
4. Сколько успешных транзакций осуществил каждый из пользователей? Постройте гистограмму распределения числа успешных транзакций.
5. Коллега прислал Вам обновленные [данные](https://yadi.sk/d/NtNP5zY0qU4ZAw). Постройте сводную таблицу user\_vs\_minute\_pivot, где в качестве столбцов будут использованы имена пользователей, строк – минуты, значений – число совершенных операций в указанную минуту. Пропущенные значения заполните нулями.
6. Изучите представленный график. Помогите коллеге разобраться,  есть ли в данных ошибка, или же всё хорошо.
7. Если в данные закралась ошибка, исправьте её и сохраните правильное число минут, прошедших с начала дня, в колонку true\_minute.

Описание данных

Имеются следующие данные:

* date – дата (2020-03-02 00:00:00 – 2020-03-02 01:29:05)
* name – имя пользователя
* transaction – статус транзакции
* minute\* – сколько минут прошло с начала дня

\*Доступна в данных с 5 шага.

Загрузите датасет transaction\_data.csv. Проверьте размер таблицы, типы переменных, число пропущенных значений и описательную статистику.

**Вопрос:** сколько в датасете пропущенных значений?

[Подсказки](https://www.notion.so/5-1-python-44ee3f473c6a4c94a1d7e74673ad767e)

Проверьте, какие значения принимает колонка transaction . Сколько наблюдений относятся к тому или иному уровню? Визуализируйте результат с помощью барплота.

**Вопрос:** расположите уровни переменной в порядке убывания числа относящихся наблюдений (от большего к меньшему)

Подумайте, хорош ли полученный график?

[Подсказки](https://www.notion.so/5-3-python-ae76c9af79884ce6a4136476d9a4e92b)

Да, это просто график мечты!

Плохо видно число наблюдений у некоторых значений

У графика неправильно подписаны оси

На барплоте не может быть квалитативных переменных

Сколько произошло ошибок?

[Подсказки](https://www.notion.so/5-4-python-c1924c7997a14ac383b59132c940fd83)

Сколько транзакций осуществил каждый из пользователей? Осуществлёнными считаются транзакции со значением successfull

Посмотрите на характеристики распределения, а затем постройте гистограмму числа успешных транзакций, совершённых каждым пользователем, где

* по оси x – число транзакций
* по оси y – доля пользователей с таким числом транзакций

Выберите верные утверждения.

[Подсказки](https://www.notion.so/5-5-python-9b0119cb6cfb4740b625bdc489dce4ce)

Пользователи в среднем совершили примерно по 66 транзакций

Максимальное число транзакций – 82

Пользователи в среднем совершили по 64 транзакции

Примерно 25% пользователей совершили менее 61 транзакции

Примерно 25% пользователей совершили более 61 транзакции

Пользователи в среднем совершили примерно по 56 транзакций

Максимальное число транзакций – 112

Пользователи в среднем совершили примерно по 94 транзакции

💀 Сложное задание! 💀

К вам пришёл коллега с [немного обновленными данными](https://stepik.org/media/attachments/lesson/361624/transaction_data_updated.csv). Постройте сводную таблицу user\_vs\_minute\_pivot, где:

* столбцы – имена пользователей
* строки – минута времени осуществления операции (minute)
* значения – число совершённых операций

Пропущенные значения заполните нулями.

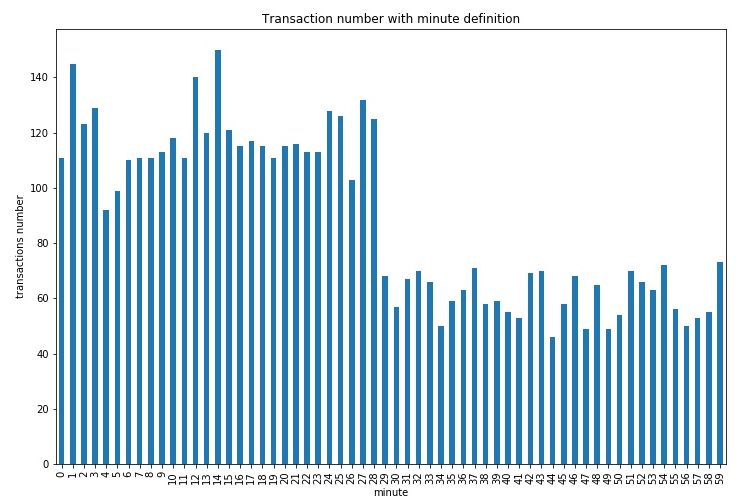
Note: фильтровать успешные транзакции не нужно - постройте табличку по всем.

Новые данные сохранены в df.

[Подсказки](https://www.notion.so/5-6-python-220ecbb5fccc4ef6aa1345ad579e1d5c)

Коллега построил следующий график по сводной таблице из предыдущего шага, и сомневается, что в данных нет ошибки.

* по оси x – минута транзакции (сколько минут прошло с начала дня)
* по оси y – число транзакций



Как вы думаете, из-за чего мы видим такую картину? Произошла ошибка, или с данными всё хорошо?

[Подсказки](https://www.notion.so/5-7-python-c203c7b7d49e4668afa08dbdbe156df9)

Исправьте ошибку коллеги (если она есть 😉), и посчитайте число минут, прошедших с начала дня. Результат сохраните в колонку true\_minute. Постройте график, как в предыдущем шаге, и посмотрите, произошли ли изменения (график в код включать не нужно).

Например:

date name transaction minute true\_minute

2020-03-02 00:37:20 Ekaterina\_Petrova cancelled 37 37

2020-03-02 01:06:14 Ekaterina\_Smirnova successfull 6 66

Данные записаны в df , колонка date прочитана как дата.

[Подсказки](https://www.notion.so/5-8-python-72d4cd63da384ba0a7e4e212d8cb9d48)

Дополнительный проект

В папке subsid (shared/homeworks/python\_ds\_miniprojects/5\_subsid) находятся файлы (tm\_sales\_1, tm\_sales\_2, ...) с продажами продуктов через телемаркетинг. Каждый файл содержит, как минимум, 4 колонки (поля): FILIAL\_ID, SUBS\_ID, PROD\_ID, ACT\_DTTM.

Суть задачи в том, чтобы проверить подключения продуктов определенным пользователям, соединив файлы о продажах с логами по подключениям в системе.

Особенности данных:

1. сотрудники телемаркетинга не всегда указывают полный id, если 'id' нет в начале SUBS\_ID, то нужно его добавить
2. поля в файлах могут быть расположены абсолютно случайным образом, но названия полей статичны
3. продажа не засчитывается, если отключение (END\_DTTM) произошло меньше чем через 5 минут после подключения (START\_DTTM)
4. если в файле с продажами встречается строка без указанного SUBS\_ID, она пропускается

Сохраните результат в csv файл с разделителем ;, содержащий корректные подключения.

*Note:* обратите внимание на то, как pandas переводит дату из строки, возможно вам понадобится параметр format

Пример содержания итогового файла (колонка difference это разница между START\_DTTM и END\_DTTM):

💀 Сложное задание! 💀

Для проверки результатов введите SUBS\_ID из полученного датасета в порядке возрастания, через запятую с пробелом. Например:

id1, id2

[Подсказки](https://www.notion.so/5-1-python-2f6d9a22d3764e4bafa398de3b3f409c)