Использование SQL-запросов для исследования датасета на примере данных с сайта OZON

Работа выполнена Коротковой Л.С.

Подготавливаем среду для работы с датасетом и SQL-запросами

```
import sqlite3
import pandas as pd
! pip install pandas_profiling
import pandas_profiling

# подключение к бд
con = sqlite3.connect('my_ozon_2020-12-18_2021-01-18.db', timeout=10)
cur = con.cursor()
```

Знакомимся с датасетом

```
# готовим таблицу с данными, создадим датафрейм из csv-файла
df = pd.read_csv('data\ozon_2020-12-18_2021-01-18.csv')
df.shape
df.info()
df.profile_report()
```

Предобработка данных

Выведем информацию о пропусках в процентном соотношении
df.isnull().mean() * 100

Далее:

- Удалим столбцы с категориальными данными, в которых более 30 % пропусков
- Заполним оставшиеся ячейки константами
- Переименуем столбцы

Постановка задачи

```
# загрузим теперь наши данные в бд, указав название таблицы
df.to_sql(con=con, name='ozon_table', index=False)
```

считывание данных из таблицы

```
#считывание данных из таблицы
data_test = cur.execute('select * from ozon_table')
con.commit()
cur.fetchall()
или
sql = """select * from ozon_table"""
df = pd.read_sql_query(sql, con=con)
df
```

выведем названия всех фирм-продавцов, представленных в датасете

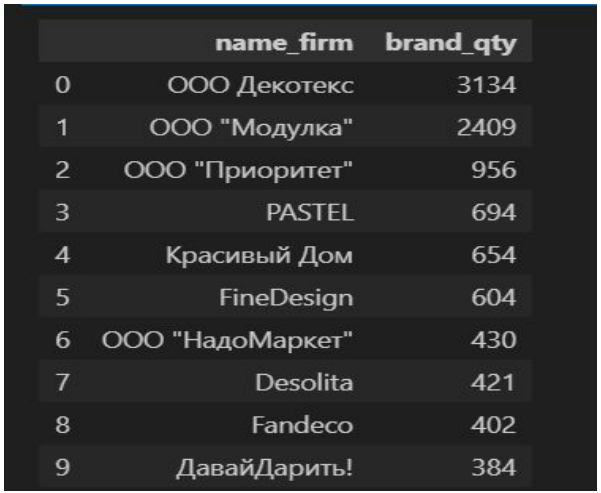
```
sql = """SELECT SUBSTR(seller, 1, INSTR(seller, ',') - 1) AS name_firm
        FROM ozon_table
        GROUP BY 1"""
df = pd.read_sql_query(sql, con=con)
df
```

выведем бренд, цену и отсортированный по категории скидка товара столбец

```
sql = """SELECT brand,
            price,
            MAX(sales) AS max_s
        FROM ozon_table
        GROUP BY 1
        ORDER BY 3 DESC"""
df = pd.read_sql_query(sql, con=con)
df
```

выведем пот-10 названий фирм-продавцов, с наибольшим числом представленных товаров одного бренда

```
sql = """SELECT SUBSTR(seller, 1, INSTR(seller, ',') - 1) AS name_firm,
            COUNT(brand) AS brand_qty
        FROM ozon_table
        GROUP BY 1
        ORDER by rating DESC, 2 DESC
        LIMIT 10"""
df = pd.read_sql_query(sql, con=con)
df
```



	name_firm	brand_qty
0	ООО Декотекс	3134
1	ООО "Модулка"	2409
2	ООО "Приоритет"	956
3	PASTEL	694
4	Красивый Дом	654
5	FineDesign	604
6	ООО "НадоМаркет"	430
7	Desolita	421
8	Fandeco	402
9	ДавайДарить!	384

выведем категорию схема доставки, отсортированную по общему количеству товаров в доставке

```
sql = """SELECT deliveryscheme,
            COUNT(deliveryscheme) AS deliveryscheme_qty
        FROM ozon_table ot
        GROUP BY 1
        ORDER BY 2 DESC"""
df = pd.read_sql_query(sql, con=con)
df
```

	Deliveryscheme	deliveryscheme_qty
0	FBS	264189
1	FBO	55627
2	Retail	13494
3	Cross	5224
4	0,0000	1

выведем товары, условно сгруппированные по количеству упоминаний в датасете

```
sql = """SELECT
        CASE
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Искусственная'
            THEN 'Искусственная елка'
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Ёлочный'
            THEN 'Ёлочный шарик'
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Уличное'
            THEN 'Уличное освещение'
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Подвесной'
            THEN 'Подвесной светильник'
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Анатомическая'
            THEN 'Анатомическая подушка'
          WHEN SUBSTR(Name, 1, INSTR(Name, ' ') - 1) = 'Циркуляционный'
            THEN 'Циркуляционный насос'
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Настольная'
            THEN 'Настольная лампа'
          WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Кухонный'
            THEN 'Кухонный нож'
          ELSE SUBSTR(name, 1, INSTR(name, ' ') - 1)
        END AS object,
        COUNT(name) AS qty
        FROM ozon_table
        GROUP BY object
        ORDER BY 2 DESC"""
```

выведем название, количество и рейтинг товаров в категории "Хозяйственные товары", отсортированные по количеству и рейтингу товара

```
sql = """SELECT
        CASE
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Искусственная'
                THEN 'Искусственная елка'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Ёлочный'
                THEN 'Ёлочный шарик'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Уличное'
                THEN 'Уличное освещение'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Подвесной'
                THEN 'Подвесной светильник'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Анатомическая'
                THEN 'Анатомическая подушка'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Циркуляционный'
                THEN 'Циркуляционный насос'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Настольная'
                THEN 'Настольная лампа'
            WHEN SUBSTR(name, 1, INSTR(name, ' ') - 1) = 'Кухонный'
                THEN 'Кухонный нож'
            ELSE SUBSTR(name, 1, INSTR(name, ' ') - 1)
        END AS object,
        COUNT(name) AS qty,
        rating
    FROM ozon_table
    WHERE SUBSTR(SUBSTR(full_category, INSTR(full_category, '/') + 1), 1, INSTR(SUBSTR(full_category,
INSTR(full_category, '/')+2),'/')) == 'Хозяйственные товары'
    GROUP BY object
    ORDER BY rating DESC, qty DESC"""
```

	object	qty	Rating
0	Штора	3127	5,0000
1	Полка	721	5,0000
2	Дозатор	697	5,0000
3	Сумка	192	5,0000
4	Ёршик	150	5,0000
...	...	...	...
1237	зарядное	1	0,0000
1238	микрофибры	1	0,0000
1239	набор	1	0,0000
1240	распылительная	1	0,0000
1241	сушилка	1	0,0000

1242 rows × 3 columns

Выводы

1. В наших данных 5524 фирмы-продавца
2. Наибольшая скидка в денежных единицах составляет 4237 рублей, при этом стоимость товара после скидки составляет 732 рубля
3. Наибольшее количество брендов представлено ООО Декотекс - 3134, закрывает десятку по представленным брендам ДавайДарить! - 384
4. В категории схема доставки всего 4 вида, при этом на первом месте FBS - 264189, - который в 5 раз превосходит следующий за ним FBO (55627)
5. Наиболее широко представлены разного рода наборы (15529 единиц) и комплекты (10905 единиц)
6. В категории "Хозяйственные товары" с рейтингом 5 баллов наиболее широко представлены шторы - 3127, полки - 721, дозаторы - 697, сумки - 192 и ёршики - 150. При этом, общее число разных товаров в категории - 1243.

Благодарю за внимание!