ETL. Семинар №2

1. Викторина
2. Блок 1

Задание 1

Скачайте датасет. Проанализируйте его на наличие пропусков используя pandas.

Напишите в чат какие пропущенные значения вы обнаружили и причины их появления.

Решение:

```
import pandas as pd

data = pd.read_csv('/datasets/data.csv')

print(data.head(20))
print(data.info())

data.isna().sum()
```

```
Output:
children               0
days_employed       2174
dob_years              0
education              0
education_id           0
family_status          0
family_status_id       0
gender                 0
income_type            0
debt                   0
total_income        2174
purpose                0
dtype: int64
```

Здесь как раз важно описание проблемы, должно быть что-то такое:
В двух столбцах есть пропущенные значения. Один из них — `days_employed`. Другой столбец с пр опущенными значениями — `total_income` — хранит данные о доходах. На сумму дохода сильнее в

сего влияет тип занятости, поэтому заполнить пропуски в этом столбце нужно медианным значением по каждому типу из столбца `income_type`. Например, у человека с типом занятости `сотрудник` пр опуск в столбце `total_income` должен быть заполнен медианным доходом среди всех записей с тем же типом.

Задание 2

Найдите в датафрейме дубликаты. И удалите их. Значения могут быть одинаковыми но написаны по разному. Например может отличаться размер регистра(заглавные и строчные буквы)

Напишите в чат возможные причины появления дубликатов.

Решение:

```
# посчитайте дубликаты
data.duplicated().sum()

# удалите дубликаты
data = data.drop_duplicates().reset_index(drop=True)

# удаление неявных дубликатов
data['education'] = data['education'].str.lower()
```

3. Блок 2

Задание 3

Сделайте колонку purpose_category в которую войдут следующие категории:

- операции с автомобилем,
- операции с недвижимостью,
- проведение свадьбы,
- получение образования

В чат напишите какое количество строк у вас получилось в каждой категории.

Решение:

```
data['purpose'].unique()

# создайте функцию categorize_purpose()
def categorize_purpose(purpose):
    if 'авто' in purpose:
        return 'операции с автомобилем'
    if 'недвиж' in purpose or 'жиль' in purpose:
        return 'операции с недвижимостью'
    if 'свад' in purpose:
        return 'проведение свадьбы'
    if 'образов' in purpose:
        return 'получение образования'
print(categorize_purpose('покупка жилья'))

# примените функцию методом apply()
data['purpose_category'] = data['purpose'].apply(categorize_purpose)
data
```

Задание 4

Постройте иерархическую и реляционную модели описывающие структуру предприятия состоящие из объектов Отдел, Начальник, Сотрудник

Нарисуйте схему моделей используя app.diagrams.net и поделитесь картинкой в чате.

Решение: