### **Тестовое задание**

#### **Введение**

Цель этого задания — показать Ваши навыки работы с данными, использования библиотек и умение выстраивать пайплайн решения задачи. Важно понимать, что задание не направлено на создание идеальной модели или решение сложных бизнес-проблем. Оно должно занять 4–8 часов и быть интересным исследованием, а не марафоном. Наслаждайтесь процессом!

#### **Данные**

Для работы потребуется следующий датасет он находится в архиве data\_for\_test\_arch.zip основан он на [Dominick’s dataset](https://www.chicagobooth.edu/research/kilts/research-data/dominicks), в архиве ты найдешь таблицы:

beer\_sales\_data.parquet - продажи пива

cig\_sales\_data.parquet - продажи сигарет

ana\_sales\_data.parquet - продажи анальгетиков

demographic\_data.parquet - данные демографии и магазина

beer\_upc.parquet - таблица с описанием товара пиво

cig\_upc.parquet - таблица с описанием товара сигареты

ana\_upc.parquet - таблица с описанием товара анальгетики

описание таблиц sales:

|  |  |
| --- | --- |
| Variable | Description |
| upc | Номер UPC |
| store | Номер магазина |
| week | Номер недели |
| move | Количество проданных единиц |
| price | Розничная цена |
| qty | Количество товаров в комплекте |
| sale | Код распродажи (B,C,S) Эта переменная указывает, продавался ли продукт по акции на этой неделе. Код «B» указывает на бонусную покупку, «C» указывает на купон, «S» указывает на простое снижение цены. |

описание таблицы upc

|  |  |
| --- | --- |
| Variable | Description |
| 'com\_code' | не важно |
| 'upc' | номер UPC |
| 'descrip' | Название товара |
| 'size' | Размер товара |
| 'case' | не важно |
| 'nitem' | Количество товаров в ящике |

описание таблиц demographic

|  |  |
| --- | --- |
| Variable | Description |
| 'store' | номер магазина |
| 'age60' | % населения старше 60 лет |
| 'age9' | % населения моложе 9 лет |
| 'educ' | % выпускников колледжей |
| 'ethnic' | % чернокожих и латиноамериканцев |
| 'income' | логарифм медианного дохода |
| 'hhlarge' | % домохозяйств с 5 или более людьми |
| 'workwom' | % работающих женщин с полной занятостью |
| 'hval150' | % домохозяйств со стоимостью более 150 000 долларов США |
| 'sstrdist' | расстояние до ближайшего склада-магазина |
| 'sstrvol' | соотношение продаж этого магазина к ближайшему складу-магазину |
| 'cpdist5' | среднее расстояние в милях до ближайших 5 супермаркетов |
| 'cpwvol5' | соотношение продаж этого магазина к среднему показателю ближайших пяти магазинов |

Рекомендации:

* Вы можете использовать весь набор данных, если считаешь, что он необходим для выполнения задачи, но это не обязательно. Выбирайте только те данные, которые считаете важными.
* Постарайтесь сосредоточиться на решении задачи, а не на бесконечной обработке данных.
* Таблицу с номерами недель можно восстановить, используя данные с 21 страницы [мануала](https://www.chicagobooth.edu/research/kilts/research-data/-/media/enterprise/centers/kilts/datasets/dominicks-dataset/dominicks-manual-and-codebook_kiltscenter). Там же есть информация о праздниках, которую можно использовать для анализа сезонности.

#### **Задача**

Цель: Прогнозировать спрос на товар или категорию товаров.

Ожидаемые шаги:

1. Провести исследовательский анализ данных (EDA), чтобы:
   * Понять структуру данных.
   * Найти возможные зависимости спроса от параметров, включая сезонность, тренды и другие факторы.
2. Построить модель(и) прогноза спроса:
   * Хорошо, если Вы попробуете несколько подходов (например, бустинг, AutoML, статистические модели) и сравнишь их.
   * Желательно, чтобы выбор моделей был обоснован.
3. Выделите зависимость спроса от цены и построй кривые эластичности. Это необязательно, но будет преимуществом.

#### **Ожидаемый результат**

* Описательный EDA: основные выводы и визуализации.
* Модели с их сравнением и интерпретацией результатов.
* (Необязательно, но инетересно ;-)) Анализ эластичности и графики зависимости спроса от цены.