Лабораторная работа 1

У вас есть датасет — таблица test.vimbox_pages со сведениями об активности пользователей следующего вида: <идентификатор-пользователя>,<страница>,<дата-время посещения>: https://www.dropbox.com/s/dhcltzfybwvnrqb/vimbox_pages.csv?dl=1
Нужно импортировать его в любую доступную базу данных и там написать SQL-запрос, который найдёт все сессии, в течение которых пользователь делал домашнее задание перед уроком, то есть совершил последовательность действий:

- 0 или более раз заходил на любые страницы;
- зашел на rooms.homework-showcase (раздел со списком домашних заданий);
- 0 или более раз заходил на любые страницы;
- зашел на rooms.view.step.content (страница домашнего задания);
- 0 или более раз заходил на любые страницы;
- зашел на rooms.lesson.rev.step.content (страница урока с преподавателем);
- 0 или более раз заходил на любые страницы.

Сессией называется активность пользователя, в которой между последовательными действиями проходит менее одного часа. Сессия начинается в момент первого из этих действий и заканчивается через час после последнего из них.
Результатом должна стать выгрузка сессий вида: <идентификатор-пользователя>,<дата-время начала сессии>,<дата-время окончания сессии>.
Нужно построить в любом инструменте такой график к полученной выгрузке, чтобы на нём можно было увидеть, какая обычно длина сессии, как зависит число сессий от времени суток, какие бывают выбросы (необычные сессии).
Разрешается добавлять в выгрузку дополнительные поля или использовать отдельные запросы для построения графика.

Лабораторная работа 2

Вам нужно построить в Tableau дэшборд (набор графиков), который будет показывать юнит-экономику сервиса такси (совокупность средних значений основных бизнес-метрик, которые определяют доходы и расходы в пересчете на одну поездку)

1. за весь год;
2. по месяцам;

по данным из таблицы tlc_green_trips_2015 датасета new_york_taxi_trips https://console.cloud.google.com/bigquery?project=maps-test-338719 (доступ выдаю на gmail-почту) при следующих допущениях:

- расход топлива во время поездки можно вычислить по формуле FE = 0.2 * AS + 20 миль на галлон, где AS — средняя скорость во время поездки, миль в час;
- стоимость бензина в 2015 году $2.43 за галлон;
- оплата труда водителя в среднем составляла $15 в час;
- утилизация труда водителя (доля рабочего времени, когда водитель везет пассажира) составляла в среднем 35%;
- плата за лизинг одного автомобиля составляет в среднем $5 в пересчете на одну поездку.

Чтобы выполнить эту задачу, сперва нужно построить иерархию метрик, из чего состоят доходы (поля fare_amount и extra) и расходы в пересчете на одну поездку (лизинг, оплата труда водителя, стоимость бензина).

Инструкция, как подключаться к BigQuery https://help.tableau.com/current/pro/desktop/en-us/examples_googlebigquery.htm

* Доп.задание: отобразить на карте самые прибыльные pickup-локации.

Лабораторная работа 3

Заказчик — кредитный отдел банка. Нужно разобраться, влияет ли семейное положение, уровень дохода, цель кредита 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

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

Шаг 2. Предобработка данных

1. определите и заполните пропущенные значения:
    - опишите, какие пропущенные значения вы обнаружили;
    - приведите возможные причины появления пропусков в данных;
    - объясните, по какому принципу заполнены пропуски;
2. замените вещественный тип данных на целочисленный:

    - поясните, как выбирали метод для изменения типа данных;
3. удалите дубликаты:
    - поясните, как выбирали метод для поиска и удаления дубликатов в данных;
    - приведите возможные причины появления дубликатов;
4. выделите леммы в значениях столбца с целями получения кредита:
    - опишите, как вы проводили лемматизацию целей кредита;
5. категоризируйте данные:
    - перечислите, какие «словари» вы выделили для этого набора данных, и объясните, почему.

В данных могут встречаться артефакты — значения, которые не отражают действительность. Например, отрицательное количество дней трудового стажа. Для реальных данных — это нормально. Нужно описать возможные причины появления таких данных и обработать их.

Шаг 3. Ответьте на вопросы

- Есть ли зависимость между наличием 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

Допзадание: построить модель кредитного скоринга

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

Лабораторная работа 4

В вашем распоряжении данные сервиса Яндекс.Недвижимость — архив объявлений о продаже квартир в Санкт-Петербурге и соседних населённых пунктах за несколько лет. Нужно научиться определять рыночную стоимость объектов недвижимости. Ваша задача — установить параметры. Это позволит построить автоматизированную систему: она отследит аномалии и мошенническую деятельность.
По каждой квартире на продажу доступны два вида данных. Первые вписаны пользователем, вторые — получены автоматически на основе картографических данных. Например, расстояние до центра, аэропорта, ближайшего парка и водоёма.

Инструкция по выполнению

Шаг 1. Откройте файл с данными и изучите общую информацию

Описание данных

- *airports_nearest* — расстояние до ближайшего аэропорта в метрах (м)
- *balcony* — число балконов
- *ceiling_height* — высота потолков (м)
- *cityCenters_nearest* — расстояние до центра города (м)
- *days_exposition* — сколько дней было размещено объявление (от публикации до снятия)
- *first_day_exposition* — дата публикации
- *floor* — этаж
- *floors_total* — всего этажей в доме
- *is_apartment* — апартаменты (булев тип)
- *kitchen_area* — площадь кухни в квадратных метрах (м²)
- *last_price* — цена на момент снятия с публикации
- *living_area* — жилая площадь в квадратных метрах(м²)
- *locality_name* — название населённого пункта
- *open_plan* — свободная планировка (булев тип)
- *parks_around3000* — число парков в радиусе 3 км
- *parks_nearest* — расстояние до ближайшего парка (м)

- *ponds_around3000* — число водоёмов в радиусе 3 км
- *ponds_nearest* — расстояние до ближайшего водоёма (м)
- *rooms* — число комнат
- *studio* — квартира-студия (булев тип)
- *total_area* — площадь квартиры в квадратных метрах (м²)
- *total_images* — число фотографий квартиры в объявлении

Пояснение: апартаменты — это нежилые помещения, не относящиеся к жилому фонду, но имеющие необходимые условия для проживания.

Шаг 2. Предобработка данных

- определите и изучите пропущенные значения:
    - для некоторых пропущенных значений можно предположить логичную замену. Например, если человек не указал число балконов — скорее всего, их нет. Такие пропуски правильно заменить на 0. Для других типов данных нет подходящего значения на замену. В этом случае правильно оставить эти значения пустыми. Отсутствие значения — тоже важный сигнал, который не нужно прятать;
    - заполните пропуски, где это уместно. Опишите, почему вы решили заполнить пропуски именно в этих столбцах и как выбрали значения;
    - укажите причины, которые могли привести к пропускам в данных.
- приведите данные к нужным типам:
    - поясните, в каких столбцах нужно изменить тип данных и почему.

Шаг 3. Посчитайте и добавьте в таблицу:

- цену квадратного метра;
- день недели, месяц и год публикации объявления;
- этаж квартиры; варианты — первый, последний, другой;
- соотношение жилой и общей площади, а также отношение площади кухни к общей.

Шаг 4. Проведите исследовательский анализ данных и выполните инструкции:

- Изучите следующие параметры: площадь, цена, число комнат, высота потолков. Постройте гистограммы для каждого параметра.
- Изучите время продажи квартиры. Постройте гистограмму. Посчитайте среднее и медиану. Опишите, сколько обычно занимает продажа. Когда можно считать, что продажи прошли очень быстро, а когда необычно долго?
- Уберите редкие и выбивающиеся значения. Опишите, какие особенности обнаружили.
- Какие факторы больше всего влияют на стоимость квартиры? Изучите, зависит ли цена от площади, числа комнат, удалённости от центра. Изучите зависимость цены от того, на каком этаже расположена квартира: первом, последнем или другом. Также изучите зависимость от даты размещения: дня недели, месяца и года.
- Выберите 10 населённых пунктов с наибольшим числом объявлений. Посчитайте среднюю цену квадратного метра в этих населённых пунктах. Выделите среди них населённые пункты с самой высокой и низкой стоимостью жилья. Эти данные можно найти по имени в столбце '*locality_name'*.
- Изучите предложения квартир: для каждой квартиры есть информация о расстоянии до центра. Выделите квартиры в Санкт-Петербурге (*'locality_name'*). Ваша задача — выяснить, какая область входит в центр. Создайте столбец с расстоянием до центра в километрах: округлите до целых значений. После этого посчитайте среднюю цену для каждого километра. Постройте график: он должен показывать, как цена зависит от удалённости от центра. Определите границу, где график сильно меняется — это и будет центральная зона.
- Выделите сегмент квартир в центре. Проанализируйте эту территорию и изучите следующие параметры: площадь, цена, число комнат, высота потолков. Также выделите факторы, которые влияют на стоимость квартиры (число комнат, этаж, удалённость от центра, дата размещения объявления). Сделайте выводы. Отличаются ли они от общих выводов по всему городу?

Шаг 5. Напишите общий вывод

Допзадание: построить модель прогноза цены продажи
Оформление: Выполните задание в *Jupyter Notebook* либо на SQL в BigQuery.

Лабораторная работа 5

Вы работаете в интернет-магазине «Стримчик», который продаёт по всему миру компьютерные игры. Из открытых источников доступны исторические данные о продажах игр, оценки пользователей и экспертов, жанры и платформы (например, *Xbox* или *PlayStation*). Вам нужно выявить определяющие успешность игры закономерности. Это позволит сделать ставку на потенциально популярный продукт и спланировать рекламные кампании.
Перед вами данные до 2016 года. Представим, что сейчас декабрь 2016 г., и вы планируете кампанию на 2017-й. Нужно отработать принцип работы с данными. Неважно, прогнозируете ли вы продажи на 2017 год по данным 2016-го или же 2027-й — по данным 2026 года.
В наборе данных попадается аббревиатура *ESRB (Entertainment Software Rating Board)* — это ассоциация, определяющая возрастной рейтинг компьютерных игр. *ESRB* оценивает игровой контент и присваивает ему подходящую возрастную категорию, например, «Для взрослых», «Для детей младшего возраста» или «Для подростков».

Инструкция по выполнению

Шаг 1. Откройте файл с данными и изучите общую информацию

Скачать датасет

Описание данных

- *Name* — название игры
- *Platform* — платформа
- *Year_of_Release* — год выпуска
- *Genre* — жанр игры
- *NA_sales* — продажи в Северной Америке (миллионы проданных копий)

- *EU_sales* — продажи в Европе (миллионы проданных копий)
- *JP_sales* — продажи в Японии (миллионы проданных копий)
- *Other_sales* — продажи в других странах (миллионы проданных копий)
- *Critic_Score* — оценка критиков (максимум 100)
- *User_Score* — оценка пользователей (максимум 10)
- *Rating* — рейтинг от организации *ESRB* (англ. *Entertainment Software Rating Board).* Эта ассоциация определяет рейтинг компьютерных игр и присваивает им подходящую возрастную категорию.

Данные за 2016 год могут быть неполными.

Шаг 2. Подготовьте данные

- Замените названия столбцов (приведите к нижнему регистру);
- Преобразуйте данные в нужные типы. Опишите, в каких столбцах заменили тип данных и почему;
- Обработайте пропуски при необходимости:
    - Объясните, почему заполнили пропуски определённым образом или почему не стали это делать;
    - Опишите причины, которые могли привести к пропускам;
    - Обратите внимание на аббревиатуру 'tbd' в столбцах с рейтингом. Отдельно разберите это значение и опишите, как его обработать;
- Посчитайте суммарные продажи во всех регионах и запишите их в отдельный столбец.

Шаг 3. Проведите исследовательский анализ данных

- Посмотрите, сколько игр выпускалось в разные годы. Важны ли данные за все периоды?
- Посмотрите, как менялись продажи по платформам. Выберите платформы с наибольшими суммарными продажами и постройте распределение по годам. За какой характерный срок появляются новые и исчезают старые платформы?

- Возьмите данные за соответствующий **актуальный период.** Актуальный период определите самостоятельно в результате исследования предыдущих вопросов. Основной фактор — эти данные помогут построить прогноз на 2017 год.
- Не учитывайте в работе данные за **предыдущие годы**.
- Какие платформы лидируют по продажам, растут или падают? Выберите несколько потенциально прибыльных платформ.
- Постройте график «ящик с усами» по глобальным продажам игр в разбивке по платформам. Опишите результат.
- Посмотрите, как влияют на продажи внутри одной популярной платформы отзывы пользователей и критиков. Постройте диаграмму рассеяния и посчитайте корреляцию между отзывами и продажами. Сформулируйте выводы.
- Соотнесите выводы с продажами игр на других платформах.
- Посмотрите на общее распределение игр по жанрам. Что можно сказать о самых прибыльных жанрах? Выделяются ли жанры с высокими и низкими продажами?

Шаг 4. Составьте портрет пользователя каждого региона
Определите для пользователя каждого региона (*NA, EU, JP*):

- Самые популярные платформы (топ-5). Опишите различия в долях продаж.
- Самые популярные жанры (топ-5). Поясните разницу.
- Влияет ли рейтинг ESRB на продажи в отдельном регионе?

Шаг 5. Проверьте гипотезы

- Средние пользовательские рейтинги платформ *Xbox One* и *PC* одинаковые;
- Средние пользовательские рейтинги жанров *Action* (англ. «действие», экшен-игры) и *Sports* (англ. «спортивные соревнования») разные.

Задайте самостоятельно пороговое значение *alpha*.
Поясните:

- Как вы сформулировали нулевую и альтернативную гипотезы;

- Какой критерий применили для проверки гипотез и почему.

Шаг 6. Напишите общий вывод
Оформление: Выполните задание в *Jupyter Notebook,* либо на SQL в BigQuery, либо в Tableau, и вставьте скриншоты в google doc.

Лабораторная работа 6

Допустим, вы работаете в добывающей компании «ГлавРосГосНефть». Нужно решить, где бурить новую скважину.
Шаги для выбора локации обычно такие:

- В избранном регионе собирают характеристики для скважин: качество нефти и объём её запасов;
- Строят модель для предсказания объёма запасов в новых скважинах;
- Выбирают скважины с самыми высокими оценками значений;
- Определяют регион с максимальной суммарной прибылью отобранных скважин.

Вам предоставлены пробы нефти в трёх регионах. Характеристики для каждой скважины в регионе уже известны. Постройте модель для определения региона, где добыча принесёт наибольшую прибыль. Проанализируйте возможную прибыль и риски техникой *Bootstrap.*

Инструкция по выполнению проекта

1. Загрузите и подготовьте данные. Поясните порядок действий.
2. Обучите и проверьте модель для каждого региона:
   2.1. Разбейте данные на обучающую и валидационную выборки в соотношении 75:25.
   2.2. Обучите модель и сделайте предсказания на валидационной выборке.
   2.3. Сохраните предсказания и правильные ответы на валидационной выборке.
   2.4. Напечатайте на экране средний запас предсказанного сырья и

RMSE модели.
2.5. Проанализируйте результаты.

3. Подготовьтесь к расчёту прибыли:
   3.1. Все ключевые значения для расчётов сохраните в отдельных переменных.
   3.2. Рассчитайте достаточный объём сырья для безубыточной разработки новой скважины. Сравните полученный объём сырья со средним запасом в каждом регионе.
   3.3. Напишите выводы по этапу подготовки расчёта прибыли.
4. Напишите функцию для расчёта прибыли по выбранным скважинам и предсказаниям модели:
   4.1. Выберите скважины с максимальными значениями предсказаний.
   4.2. Просуммируйте целевое значение объёма сырья, соответствующее этим предсказаниям.
   4.3. Рассчитайте прибыль для полученного объёма сырья.
5. Посчитайте риски и прибыль для каждого региона:
   5.1. Примените технику *Bootstrap* с 1000 выборок, чтобы найти распределение прибыли.
   5.2. Найдите среднюю прибыль, 95%-й доверительный интервал и риск убытков. Убыток — это отрицательная прибыль.
   5.3. Напишите выводы: предложите регион для разработки скважин и обоснуйте выбор.

Описание данных

Данные геологоразведки трёх регионов находятся в файлах:

- `geo_data_0.csv`. **Скачать датасет**
- `geo_data_1.csv`. **Скачать датасет**
- `geo_data_2.csv`. **Скачать датасет**
- *id* — уникальный идентификатор скважины;
- *f0, f1, f2* — три признака точек (неважно, что они означают, но сами признаки значимы);
- *product* — объём запасов в скважине (тыс. баррелей).

Условия задачи:

- Для обучения модели подходит только линейная регрессия (остальные — недостаточно предсказуемые).
- При разведке региона исследуют 500 точек, из которых с помощью машинного обучения выбирают 200 лучших для разработки.
- Бюджет на разработку скважин в регионе — 10 млрд рублей.
- При нынешних ценах один баррель сырья приносит 450 рублей дохода. Доход с каждой единицы продукта составляет 450 тыс. рублей, поскольку объём указан в тысячах баррелей.
- После оценки рисков нужно оставить лишь те регионы, в которых вероятность убытков меньше 2.5%. Среди них выбирают регион с наибольшей средней прибылью.

Данные синтетические: детали контрактов и характеристики месторождений не разглашаются.

Bootstrap
Чтобы получить нужную величину, например, среднее, из исходного набора данных формируют подвыборки (псевдовыборки). На каждой из них и вычисляют среднее. Теоретически формировать подвыборки и рассчитывать по ним нужную величину можно многократно. Так мы получим несколько значений интересующего показателя и оценим распределение.

Бутстреп применим для любых выборок. Это полезно, когда:

- Наблюдения не описываются нормальным законом;
- Для искомых величин нет статистических тестов.

В действительности не стоит всегда рассчитывать на нормальное распределение.

Пример бутстрепа для расчета 90%-го доверительного интервала 99%-квантили:

```
state = np.random.RandomState(12345)
values = []
for i in range(1000):
    subsample = data.sample(frac=1, replace=True, random_state=state)
    values.append(subsample.quantile(0.99))
```

```
values = pd.Series(values)

lower = values.quantile(0.05)
upper = values.quantile(0.95)

print(lower)
print(upper)
```