**Проектная работа**

**Что такое проект**

Поздравляем! Вы прошли курс в тренажёре. Самое время проверить полученные знания на практике и решить аналитический кейс. Выполнять работу вы будете самостоятельно.

Это проект с полностью автоматической проверкой. Он состоит из нескольких заданий.

В середине проекта есть ячейки с уже готовым кодом — просмотрите их, запустите, проанализируйте результат и переходите к следующим заданиям. Пока тренажёр не зачтёт все задания, пожалуйста, не добавляйте новый код к краткому исследовательскому анализу в этой части проекта.

Проект завершён, когда тренажёр примет все задания. После выполнения проекта вы по желанию можете вернуться к тетрадке и расширить приведённый в шаблоне исследовательский анализ данных по своему усмотрению. Проверяться вручную специалистом эта работа не будет.

**Описание проекта**

Вы аналитик компании «Мегалайн» — федерального оператора сотовой связи. Клиентам предлагают два тарифных плана: «Смарт» и «Ультра». Чтобы скорректировать рекламный бюджет, коммерческий департамент хочет понять, какой тариф приносит больше денег.

Вам предстоит сделать предварительный анализ тарифов на небольшой выборке клиентов. В вашем распоряжении данные 500 пользователей «Мегалайна»: кто они, откуда, каким тарифом пользуются, сколько звонков и сообщений каждый отправил за 2018-й год. Нужно проанализировать поведение клиентов и сделать вывод — какой тариф лучше.

**Описание тарифов**

**Тариф «Смарт»**

1. Ежемесячная плата: 550 рублей
2. Включено 500 минут разговора, 50 сообщений и 15 Гб интернет-трафика
3. Стоимость услуг сверх тарифного пакета:
   * Минута разговора — 3 рубля. Количество использованных минут и мегабайтов «Мегалайн» всегда округляет вверх. Если пользователь проговорил всего 1 секунду, в тарифе засчитывается целая минута.
   * Сообщение — 3 рубля.
   * 1 Гб интернет-трафика — 200 рублей.

**Тариф «Ультра»**

1. Ежемесячная плата: 1950 рублей
2. Включено 3000 минут разговора, 1000 сообщений и 30 Гб интернет-трафика
3. Стоимость услуг сверх тарифного пакета:
   * Минута разговора — 1 рубль;
   * Сообщение — 1 рубль;
   * 1 Гб интернет-трафика: 150 рублей.

**Пути к файлам:**

[/datasets/calls.csv](https://code.s3.yandex.net/datasets/calls.csv)

[/datasets/internet.csv](https://code.s3.yandex.net/datasets/internet.csv)

[/datasets/messages.csv](https://code.s3.yandex.net/datasets/messages.csv)

[/datasets/tariffs.csv](https://code.s3.yandex.net/datasets/tariffs.csv)

[/datasets/users.csv](https://code.s3.yandex.net/datasets/users.csv)

**Инструкция по выполнению проекта**

**Шаг 1. Изучите общую информацию о данных из файла**

**Задание 1.** Откройте файл /datasets/calls.csv, сохраните датафрейм в переменную calls.

**Задание 2.** Выведите первые 5 строк датафрейма calls.

**Задание 3.** Выведите основную информацию для датафрейма calls с помощью метода info().

**Задание 4.** С помощью метода hist() выведите гистограмму для столбца с продолжительностью звонков. Изучите распределение данных.

**Задание 5.** Откройте файл /datasets/internet.csv, сохраните датафрейм в переменную sessions.

**Задание 6.** Выведите первые 5 строк датафрейма sessions.

**Задание 7.** Выведите основную информацию для датафрейма sessions с помощью метода info().

**Задание 8.** С помощью метода hist() выведите гистограмму для столбца с количеством потраченных мегабайт.

**Задание 9.** Откройте файл /datasets/messages.csv, сохраните датафрейм в переменную messages.

**Задание 10.** Выведите первые 5 строк датафрейма messages.

**Задание 11.** Выведите основную информацию для датафрейма messages() с помощью метода info().

**Задание 12.** Откройте файл /datasets/tariffs.csv, сохраните датафрейм в переменную tariffs.

**Задание 13.** Выведите весь датафрейм tariffs.

**Задание 14.** Выведите основную информацию для датафрейма tariffs с помощью метода info().

**Задание 15.** Откройте файл /datasets/users.csv, сохраните датафрейм в переменную users.

**Задание 16.** Выведите первые 5 строк датафрейма users.

**Задание 17.** Выведите основную информацию для датафрейма users с помощью метода info().

**Шаг 2. Подготовьте данные**

**Задание 18.** Приведите столбцы

* reg\_date из таблицы users
* churn\_date из таблицы users
* call\_date из таблицы calls
* message\_date из таблицы messages
* session\_date из таблицы sessions

к новому типу с помощью метода to\_datetime().

**Задание 19.** В данных вы найдёте звонки с нулевой продолжительностью. Это не ошибка: нулями обозначены пропущенные звонки, поэтому их не нужно удалять.

Однако в столбце duration датафрейма calls значения дробные. Округлите значения столбца duration вверх с помощью метода numpy.ceil() и приведите столбец duration к типу int.

**Задание 20.** Удалите столбец Unnamed: 0 из датафрейма sessions. Столбец с таким названием возникает, когда данные сохраняют с указанием индекса df.to\_csv(..., index=column). Этот столбец сейчас не понадобится.

**Задание 21.** Создайте столбец month в датафрейме calls с номером месяца из столбца call\_date.

**Задание 22.** Создайте столбец month в датафрейме messages с номером месяца из столбца message\_date.

**Задание 23.** Создайте столбец month в датафрейме sessions с номером месяца из столбца session\_date.

**Задание 24.** Посчитайте количество сделанных звонков разговора для каждого пользователя по месяцам и сохраните в переменную calls\_per\_month. Для этого нужно:

* сгруппировать датафрейм с информацией о звонках по двум столбцам — с идентификаторами пользователей и номерами месяцев;
* затем применить метод для подсчёта количества: .agg({'duration': 'count'}).

Выведите первые 30 строчек calls\_per\_month.

**Задание 25.** Посчитайте количество израсходованных минут разговора для каждого пользователя по месяцам и сохраните в переменную minutes\_per\_month. Для этого нужно:

* сгруппировать датафрейм с информацией о звонках по двум столбцам — с идентификаторами пользователей и номерами месяцев;
* затем применить метод для подсчёта суммы: .agg(minutes=('duration', 'sum')); так вы переименуете столбец в minutes после вычисления суммы.

Выведите первые 30 строчек minutes\_per\_month.

**Задание 26.** Посчитайте количество отправленных сообщений по месяцам для каждого пользователя и сохраните в переменную messages\_per\_month. Для этого нужно:

* сгруппировать датафрейм с информацией о сообщениях по двум столбцам — с идентификаторами пользователей и номерами месяцев;
* затем применить метод для подсчёта количества: .agg(messages=('message\_date', 'count')), так вы переименуете столбец в messages после вычисления количества.

Выведите первые 30 строчек messages\_per\_month.

**Задание 27.** Посчитайте количество потраченных мегабайт по месяцам для каждого пользователя и сохраните в переменную sessions\_per\_month. Для этого нужно:

* сгруппировать датафрейм с информацией о потраченных мегабайтах по двум столбцам — с идентификаторами пользователей и номерами месяцев;
* затем применить метод для подсчёта суммы: .agg({'mb\_used': 'sum'}).

**Шаг 3. Анализ данных**

Исследовательский анализ данных и подсчёт помесячной выручки уже есть в проекте.

Помесячная выручка вычисляется так:

* отнимите бесплатный лимит от суммарного количества звонков, сообщений и интернет-трафика;
* остаток умножьте на значение из тарифного плана;
* к результату прибавьте абонентскую плату, соответствующую тарифному плану.

Запустите ячейки этого шага, изучите код и результат. Какие выводы можно сделать из результатов этого этапа?

**Шаг 4. Проверьте гипотезы**

Проверьте две гипотезы о выручке пользователей. В качестве уровня значимости *alpha* возьмите 0.05 (5%).

**Задание 28**

Проверьте гипотезу, что средняя выручка с пользователей тарифов «Ультра» и «Смарт» различается. Воспользуйтесь методом из библиотеки scipy и выведите на экран значение *p-value:* только значение, без округления и пояснений.

Затем выведите на экран фразу «Отвергаем нулевую гипотезу», если есть основания отвергнуть нулевую гипотезу или фразу «Не получилось отвергнуть нулевую гипотезу», если оснований отвергнуть нулевую гипотезу нет. Здесь вам поможет условный оператор и сравнение получившегося *p-value* с заданным уровнем значимости.

Формат ответа должен быть таким:

Скопировать код

...

Отвергаем нулевую гипотезу

или

Скопировать кодPYTHON

...

Не получилось отвергнуть нулевую гипотезу

Где … — это значение *p-value*.

**Задание 29**

Проверьте гипотезу, что средняя выручка с пользователей из Москвы отличается от выручки с пользователей других регионов. Воспользуйтесь методом из библиотеки scipy и выведите на экран значение *p-value*. Формат вывода полностью аналогичен предыдущему заданию.

**Оформление:** Выполните задание в *Jupyter Notebook*. Заполните программный код в ячейках типа *code,* текстовые пояснения — в ячейках типа *markdown*. Примените форматирование и заголовки.

**Описание данных**

Таблица users — информация о пользователях:

* user\_id — уникальный идентификатор пользователя
* first\_name — имя пользователя
* last\_name — фамилия пользователя
* age — возраст пользователя (годы)
* reg\_date — дата подключения тарифа (день, месяц, год)
* churn\_date — дата прекращения пользования тарифом (если значение пропущено, значит, тариф ещё действовал на момент выгрузки данных)
* city — город проживания пользователя
* tarif — название тарифного плана

Таблица calls — информация о звонках:

* id — уникальный номер звонка
* call\_date — дата звонка
* duration — длительность звонка в минутах
* user\_id — идентификатор пользователя, сделавшего звонок

Таблица messages — информация о сообщениях:

* id — уникальный номер звонка
* message\_date — дата сообщения
* user\_id — идентификатор пользователя, отправившего сообщение

Таблица internet — информация об интернет-сессиях:

* id — уникальный номер сессии
* mb\_used — объём потраченного за сессию интернет-трафика (в мегабайтах)
* session\_date — дата интернет-сессии
* user\_id — идентификатор пользователя

Таблица tariffs — информация о тарифах:

* tariff\_name — название тарифа
* rub\_monthly\_fee — ежемесячная абонентская плата в рублях
* minutes\_included — количество минут разговора в месяц, включённых в абонентскую плату
* messages\_included — количество сообщений в месяц, включённых в абонентскую плату
* mb\_per\_month\_included — объём интернет-трафика, включённого в абонентскую плату (в мегабайтах)
* rub\_per\_minute — стоимость минуты разговора сверх тарифного пакета (например, если в тарифе 100 минут разговора в месяц, то со 101 минуты будет взиматься плата)
* rub\_per\_message — стоимость отправки сообщения сверх тарифного пакета
* rub\_per\_gb — стоимость дополнительного гигабайта интернет-трафика сверх тарифного пакета (1 гигабайт = 1024 мегабайта)