## 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 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
   * 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

**Примечание**

«Мегалайн» всегда округляет секунды до минут, а мегабайты — до гигабайт. Каждый звонок округляется отдельно: даже если он длился всего 1 секунду, будет засчитан как 1 минута.

Для веб-трафика отдельные сессии не считаются. Вместо этого общая сумма за месяц округляется в бо́льшую сторону. Если абонент использует 1025 мегабайт в этом месяце, с него возьмут плату за 2 гигабайта. Примите, что неиспользованные за предыдущий месяц звонки, смс, интернет на следующий месяц не переносятся.

### Инструкция по выполнению проекта

**Шаг 1. Откройте файл с данными и изучите общую информацию**

Путь к файлам:

* /datasets/calls.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/calls.csv)
* /datasets/internet.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/internet.csv)
* /datasets/messages.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/messages.csv)
* /datasets/tariffs.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/tariffs.csv)
* /datasets/users.csv. [Скачать датасет](https://code.s3.yandex.net/datasets/users.csv)

**Шаг 2. Подготовьте данные**

* Приведите данные к нужным типам;
* Найдите и исправьте ошибки в данных, если они есть.

Поясните, какие ошибки вы нашли и как их исправили. В данных вы найдёте звонки с нулевой продолжительностью. Это не ошибка: нулями обозначены пропущенные звонки, поэтому их не нужно удалять.

Посчитайте для каждого пользователя:

* количество сделанных звонков и израсходованных минут разговора по месяцам;
* количество отправленных сообщений по месяцам;
* объем израсходованного интернет-трафика по месяцам;
* помесячную выручку с каждого пользователя (вычтите бесплатный лимит из суммарного количества звонков, сообщений и интернет-трафика; остаток умножьте на значение из тарифного плана; прибавьте абонентскую плату, соответствующую тарифному плану).

**Шаг 3. Проанализируйте данные**

Опишите поведение клиентов оператора, исходя из выборки. Сколько минут разговора, сколько сообщений и какой объём интернет-трафика требуется пользователям каждого тарифа в месяц? Посчитайте среднее количество, дисперсию и стандартное отклонение. Постройте гистограммы. Опишите распределения.

**Шаг 4. Проверьте гипотезы**

* средняя выручка пользователей тарифов «Ультра» и «Смарт» различаются;
* средняя выручка пользователей из Москвы отличается от выручки пользователей из других регионов.

Пороговое значение alpha задайте самостоятельно.

Поясните:

* как вы формулировали нулевую и альтернативную гипотезы;
* какой критерий использовали для проверки гипотез и почему.

**Шаг 5. Напишите общий вывод**

**Оформление:** Задание выполните в Jupyter Notebook. Программный код заполните 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, то тариф ещё действовал на момент выгрузки данных)
* city — город проживания пользователя
* tariff — 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 сообщения
* 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)

**Примечание.** Если объединение таблиц командой merge приводит к ошибке dead kernell, примените join.

## ****Как будут проверять мой проект?****

Прежде чем приступить к решению кейса, изучите критерии оценивания.

На что обращают внимание ревьюеры:

* Как вы описываете выявленные в данных проблемы?
* Как готовите данные к анализу?
* Какие графики строите для распределений?
* Как интерпретируете полученные графики?
* Как рассчитываете стандартное отклонение и дисперсию?
* Формулируете ли альтернативную и нулевую гипотезы?
* Какие методы применяете для проверки гипотез?
* Интерпретируете ли результат проверки гипотезы?
* Соблюдаете структуру проекта и поддерживаете аккуратность кода?
* Какие выводы делаете?
* Оставляете ли комментарии к шагам?