ЛР 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.

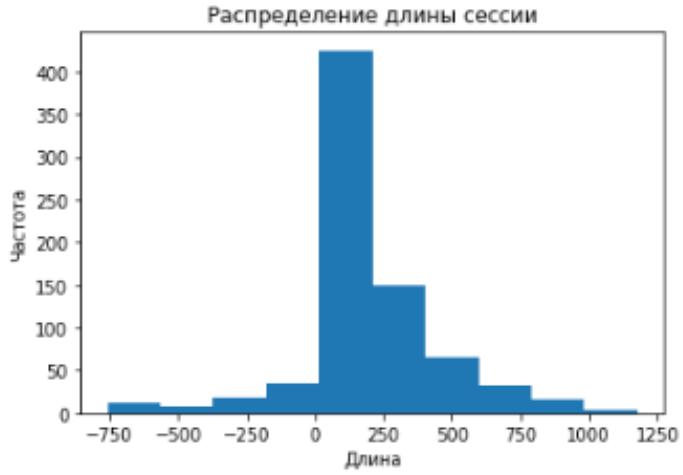
Запрос:

```
with cte_first_action as (
   select *
   from very_good_table1
   where page ='rooms.homework-showcase'
),

cte_second_action as (
   select *
   from very_good_table1
   where page = 'rooms.view.step.content'
),

cte_third_action as (
   select *
   from very_good_table1
   where page = 'rooms.lesson.rev.step.content'
),

cte_all_sessions as (
   select f.id,
          DATETIME(f.time) as curr_session_start_time,
          DATETIME(s.time) as curr_session_interim_time,
          DATETIME(t.time) as curr_session_end_time,
         lag(DATETIME(f.time), 1, 0) over(partition by f.id order by DATETIME(f.time))
                                                                        as prev_session_start_time,
         lag(DATETIME(s.time), 1, 0) over(partition by f.id order by DATETIME(s.time))
                                                                        as prev_session_interim_time,
         lag(DATETIME(t.time), 1, 0) over(partition by f.id order by DATETIME(t.time))
                                                                        as prev_session_end_time
   from cte_first_action f
   join cte_second_action s
   on s.id = f.id
   join cte_third_action t
   on t.id = f.id

   where Cast ((
            JulianDay(curr_session_interim_time) - JulianDay(curr_session_start_time)
            ) *24*60 As Integer) between 0 and 60

   and Cast ((
            JulianDay(curr_session_end_time) - JulianDay(curr_session_interim_time)
            ) *24*60 As Integer) between 0 and 60
)

select id,  DATETIME(curr_session_start_time, '-60 minutes') as start,
           DATETIME(curr_session_end_time, '+60 minutes') as end
from cte_all_sessions
where curr_session_start_time != prev_session_start_time
and curr_session_interim_time != prev_session_interim_time
and curr_session_end_time != prev_session_end_time
order by id, start, end
'''
```

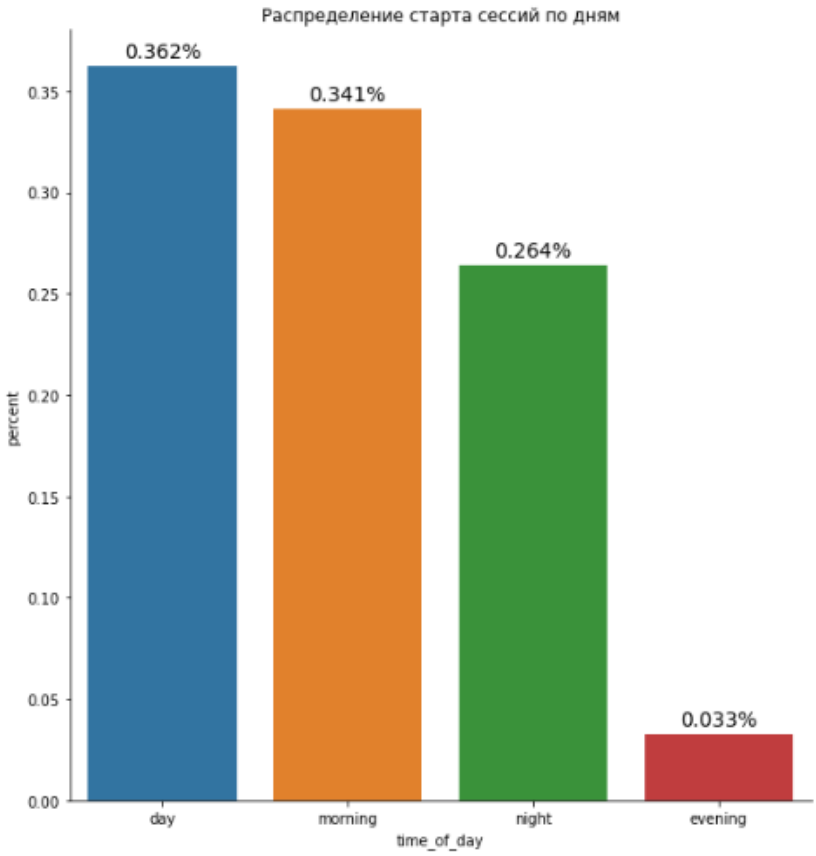
	id	start	end
0	2231	2017-03-01 04:38:50	2017-03-01 07:00:43
1	3301	2017-03-01 14:28:55	2017-03-01 18:00:35
2	3339	2017-03-01 15:01:40	2017-03-01 18:07:05
3	5487	2017-03-01 14:14:10	2017-03-01 17:04:38
4	6127	2017-03-01 17:34:39	2017-03-01 20:03:21
...	...	...	...
371	226755	2017-03-01 10:51:24	2017-03-01 13:03:45
372	226970	2017-03-01 06:59:15	2017-03-01 09:01:18
373	229522	2017-03-01 16:52:47	2017-03-01 19:00:01
374	233349	2017-03-01 14:47:21	2017-03-01 17:55:44
375	236412	2017-03-01 04:53:47	2017-03-01 07:00:41

376 rows × 3 columns

Распределения длины сессии:



Распределения стартов сессий по времени дня:



ЛР 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:

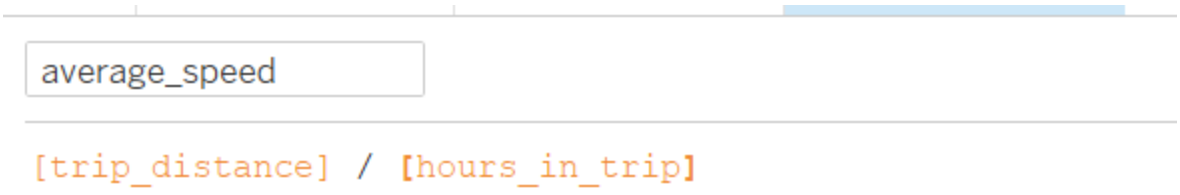
- расход топлива во время поездки можно вычислить по формуле FE = 0.2 * AS + 20 миль на галлон, где AS — средняя скорость во время поездки, миль в час;
- стоимость бензина в 2015 году $2.43 за галлон;
- оплата труда водителя в среднем составляла $15 в час;
- утилизация труда водителя (доля рабочего времени, 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).

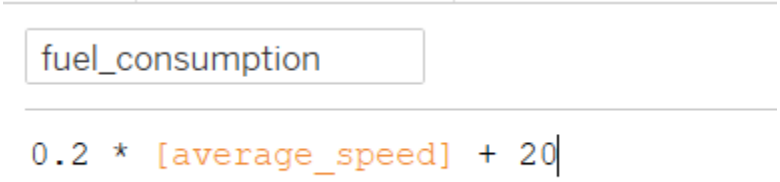
Инструкция, как подключаться к BigQuery https://help.tableau.com/current/pro/desktop/en-us/examples_googlebigquery.htm

Добавим к датасету ряд вычисляемых полей, которые потребуются для дашборда:

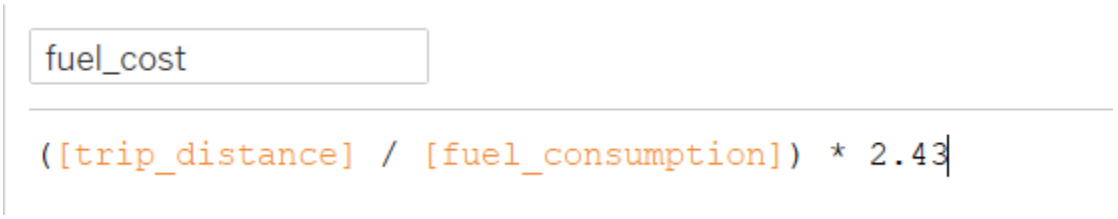
Средняя скорость, рассчитанная как отношение пройденного расстояния (в милях) к количеству часов



Расход топлива по формуле - число миль, которое можно пройти на одном галлоне бензина с данной скоростью:



Стоимость топлива - отношение пройденного расстояния к числу миль, которые можно пройти на одном галлоне (по сути, число галлонов), умноженное на стоимость одного галлона



Оплата труда водителя:

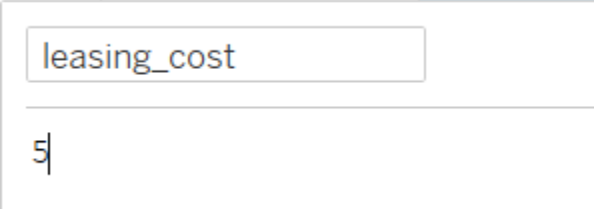
Так как время работы водителя составляет в среднем 35% от всего времени, то, если мы будем просто умножать зарплату в час на количество часов поездки, то суммарно за все поездки мы учтем примерно 35% от всего расхода на зарплату. Это можно скомпенсировать, добавляя к каждой поездке расход так, как будто поездка длилась на 65% времени дольше. Можно решить простое уравнение:
расход - 0.35

x - 1
—--
x = 1*расход/0.35

Расход это, как мы сказали зарплата в час, умноженная на количество часов в пути

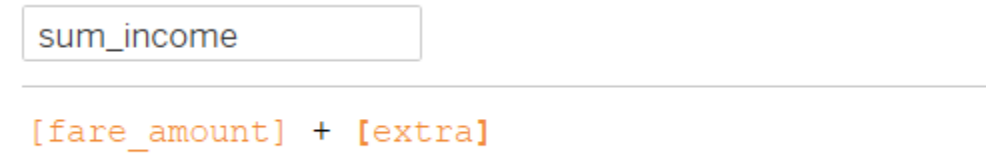
Количество часов в пути - время конца минус время начала (т.к. результат по умолчанию дается в сутках, то нужно умножить на 24, чтобы представить его в часах)

Плата за лизинг - константа, равная 5 для всех поездок



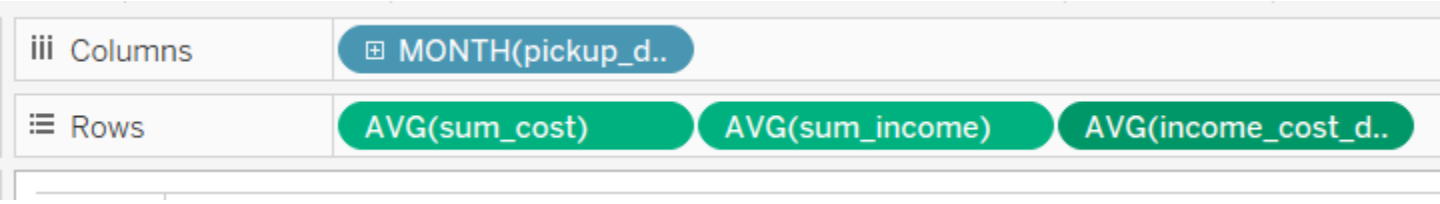
Суммарный расход за поездку

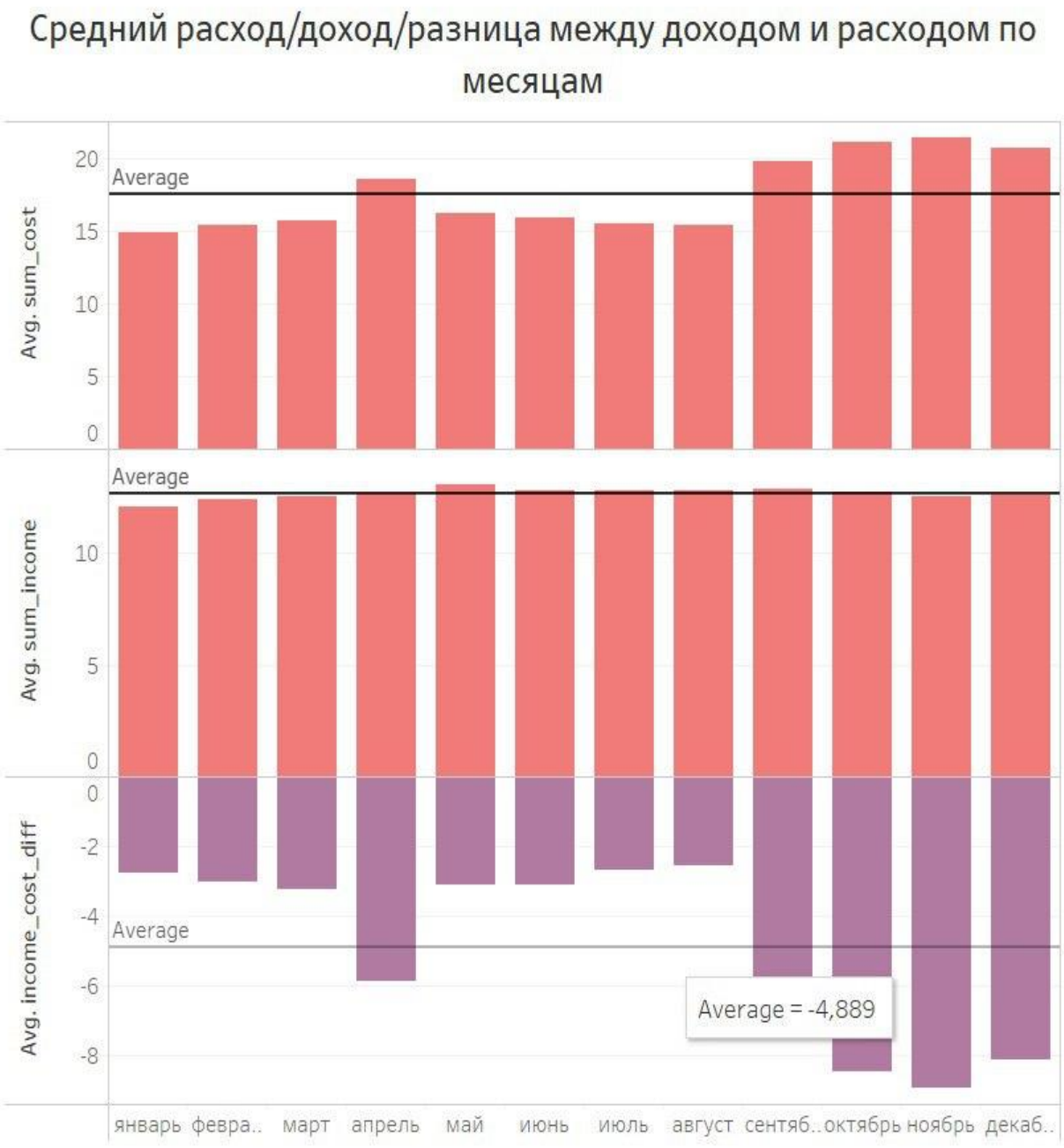
Суммарный доход с поездки:



Разница между доходом и расходом:

По последним трем полям можно построить дашборд, группируя значения по месяцам, а также отобразив линию среднего по всему году





Таким образом, на одну поездку мы в среднем тратим больше, чем с неё получаем