итоговый проект по

EDA

развертоточному анализу данных

группа №3

Olga Pavlyuck || Polina Basko || Igor Ageev || Gulnaz Nursultanova || Raman Kamisarau

2023

Этапы

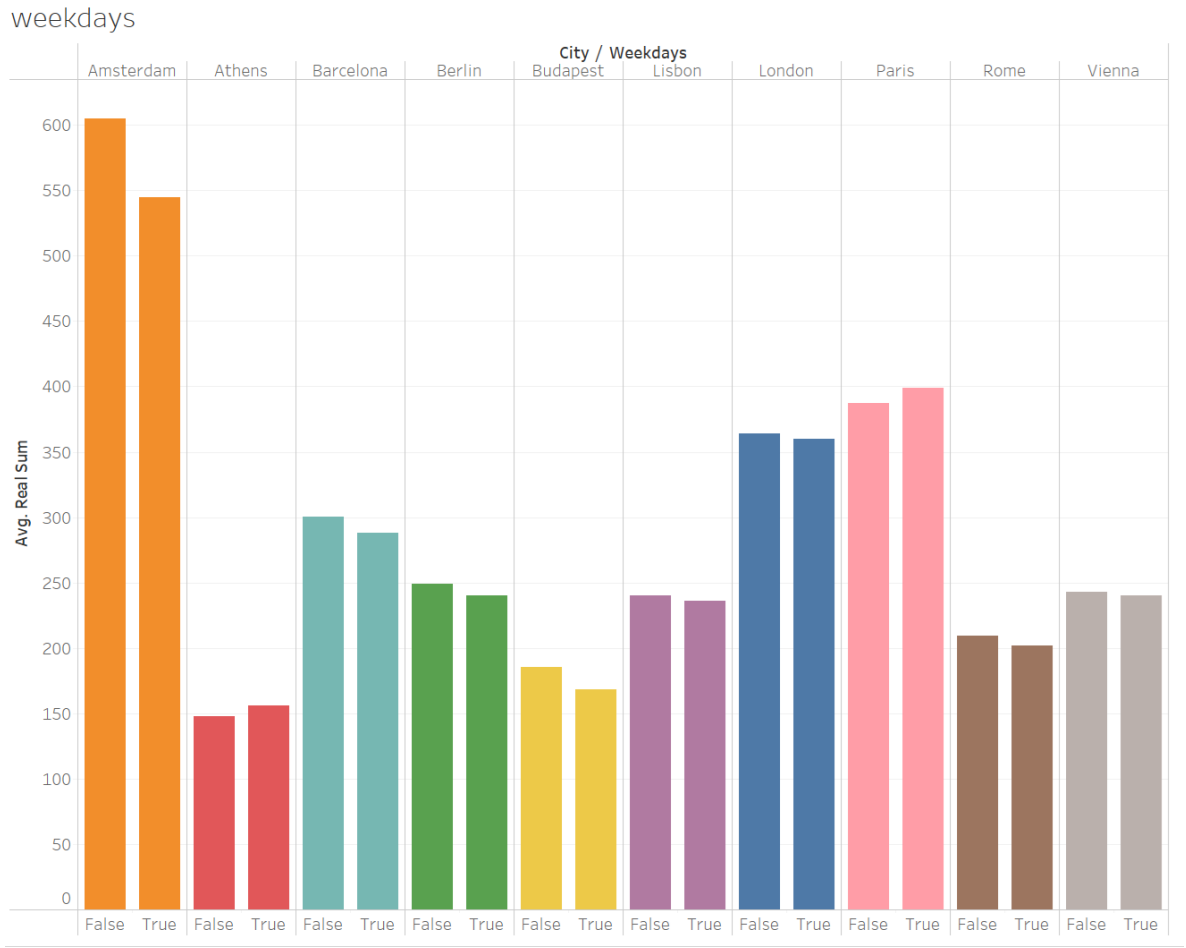
- **предобработка**
- **постановка задачи**
- **предварительный анализ**
- **алгоритм выбора размещения**
- **результат**
- что еще можно исследовать:
  профиль инвестора

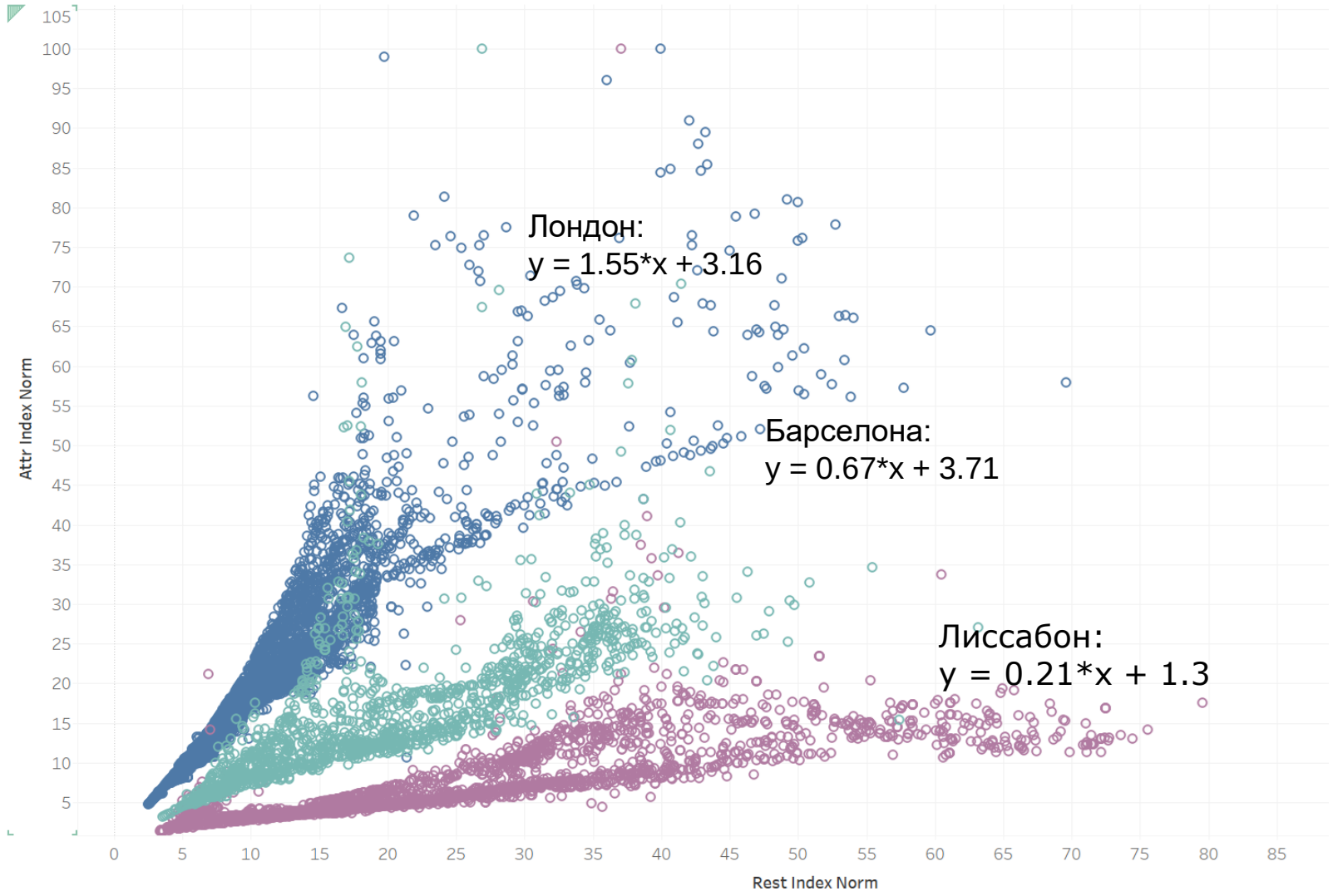
Постановка задачи

Студенты хотят поехать на каникулах в одну из европейских столиц

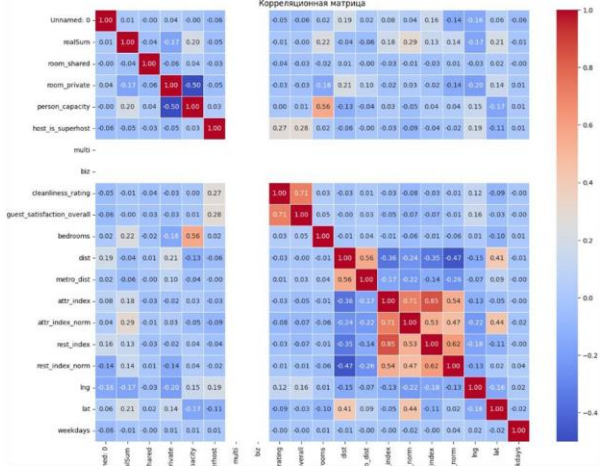
- ограниченный бюджет
- большая компания
- быть как можно ближе ко всем достопримечательностям и друг к другу
- получить максимум комфорта за минимум денег

Город

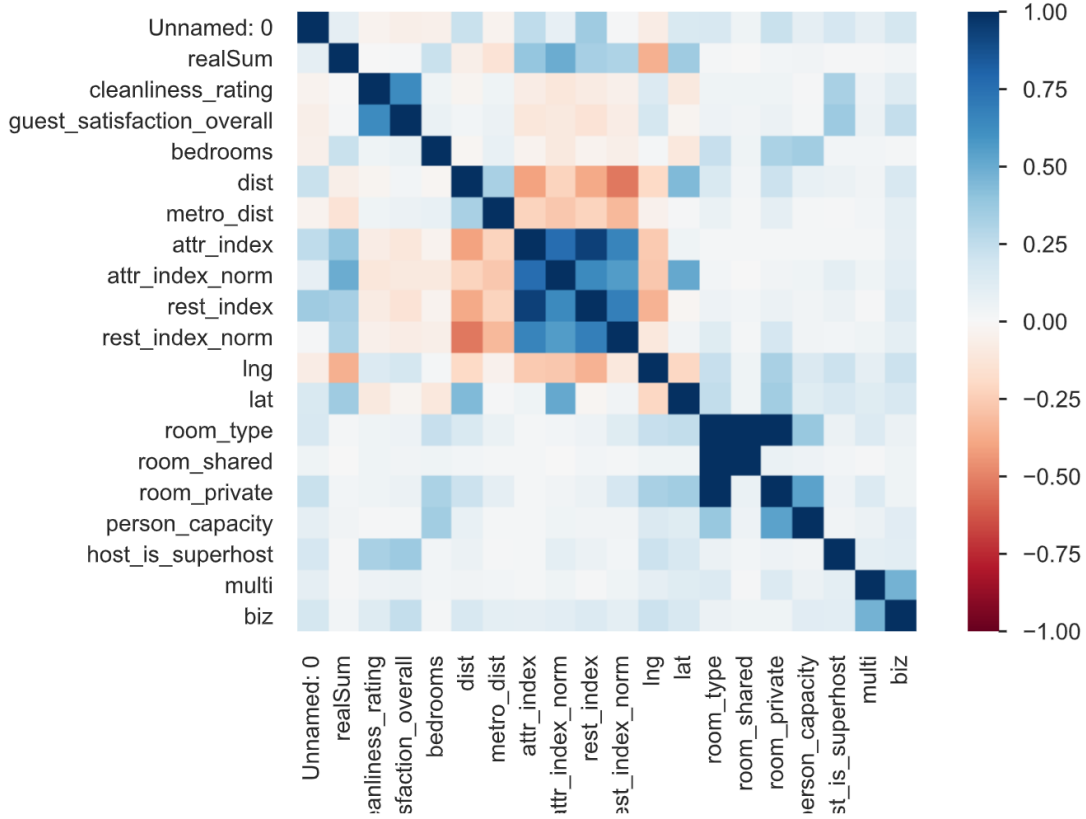

weekdays
City / Weekdays
Amsterdam
Athens
Barcelona
Berlin
Budapest
Lisbon
London
Paris
Rome
Vienna
Avg. Real Sum
600
550
500
450
400
350
300
250
200
150
100
50
0
False
True


Attr Index Norm
Rest Index Norm
Лондон:
y = 1.55*x + 3.16
Барселона:
y = 0.67*x + 3.71
Лиссабон:
y = 0.21*x + 1.3

Коэффициенты корреляции Phik (ϕk)



guest_satisfaction_overall = 0.53 * cleanliness_rating + 0.36 * host_is_superhost + 0.06 * rest_index + 0.05 * dist



Unnamed: 0	0.094108
realSum	0.000000
room_type	0.081310
room_shared	0.042905
room_private	0.075127
person_capacity	0.038964
host_is_superhost	0.477318
cleanliness_rating	0.698187
guest_satisfaction_overall	1.000000
bedrooms	0.058794
dist	0.072594
metro_dist	0.050088
attr_index	0.028521
attr_index_norm	0.113290
rest_index	0.076426
rest_index_norm	0.094259
lng	0.172962
lat	0.144924
city	0.235649
weekdays	0.000000

Итоговый коэффициент

ВАЖНО

- ПО ЛУЧШЕЙ ЦЕНЕ
- В ЛУЧШИХ УСЛОВИЯХ
- БЛИЖЕ К ЦЕНТРУ
- БЛИЖЕ К МЕТРО
- В ХОРОШЕМ РАЙОНЕ
- С ХОРОШИМИ РЕСТОРАНАМИ

```
budget_airbnb['cost_quality_index'] = (
      - 0.4 * (budget_airbnb['cost_per_person'] / max_cost)
      + 0.4 * (budget_airbnb['guest_satisfaction_overall'] / 100)
      - 0.4 * (budget_airbnb['dist'] / max_dist)
      - 0.2 * (budget_airbnb['metro_dist'] / max_metro_dist)
      + 0.4 * (budget_airbnb['attr_index_norm'] / 100)
      + 0.2 * (budget_airbnb['rest_index_norm'] / 100)
   )
```

[-1; 1]

Алгоритм

#1 Итоговый индекс

Просчитываем для всех точек, которые можем себе позволить

#3 Точки в радиусе

Получаем все точки в радиусе и селим по лучшим

#2 Лучшее место

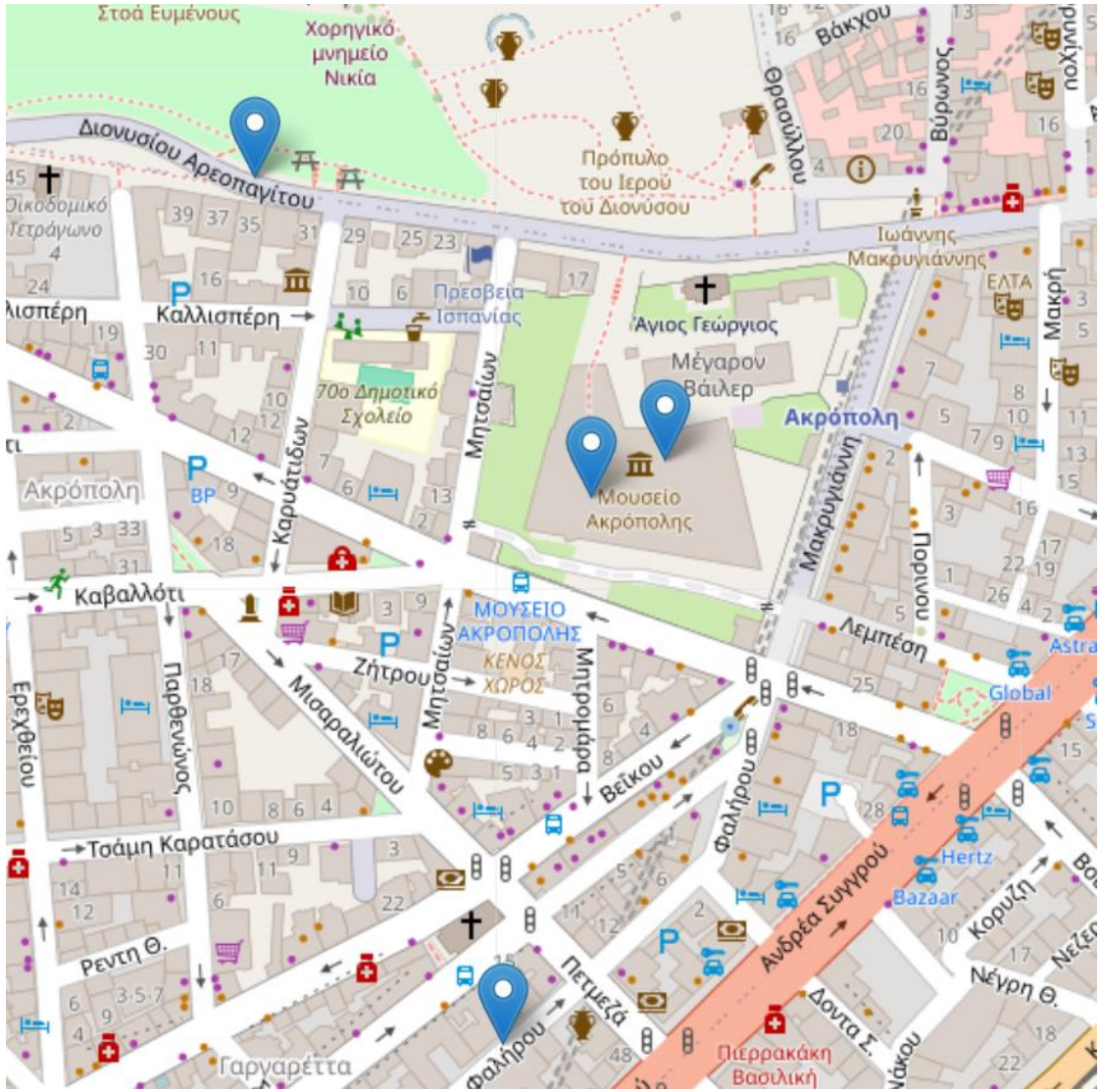
Выбираем точку с лучшим итоговым индексом

#4 Не получилось

Если не получилось разместить всех в радиусе, то выбираем новую точку и все заново.

```python
def rec_placement(data, count_persons, count_days, budget, weekdays, dist):
    global the_best
    global best_city
    global budget_airbnb
    global budget_airbnb_city

    # Рассчитываем максимальную стоимость размещения на человека в день
    max_cost_per_person = budget / count_persons / count_days

    # Рассчитываем стоимость размещения на человека в день по каждому предложению
    data['cost_per_person'] = data['realSum'] / data['person_capacity']

    # Отбрасываем предложения выходного дня и то, что не можем позволить по бюджету
    budget_airbnb = data.query(f'weekdays == {weekdays} and cost_per_person <= {max_cost_per_person}')

    # Рассчитываем максимальные значения для нормализации
    max_cost = budget_airbnb['cost_per_person'].max()
    max_dist = budget_airbnb['dist'].max()
    max_metro_dist = budget_airbnb['metro_dist'].max()

    # Высчитываем индекс идеального соотношения цена-качество (от -1 до 1)
    budget_airbnb['cost_quality_index'] = (
        - 0.4 * (budget_airbnb['cost_per_person'] / max_cost)
        + 0.4 * (budget_airbnb['guest_satisfaction_overall'] / 100)
        - 0.4 * (budget_airbnb['dist'] / max_dist)
        - 0.2 * (budget_airbnb['metro_dist'] / max_metro_dist)
        + 0.4 * (budget_airbnb['attr_index_norm'] / 100)
        + 0.2 * (budget_airbnb['rest_index_norm'] / 100)
    )

    budget_airbnb = budget_airbnb.sort_values(by='cost_quality_index', ascending=False)

    # Находим лучшее предложение и город лучшего предложения
    the_best = budget_airbnb.iloc[0]
    best_city = the_best.loc["city"]
    print(f'Рекомендуем вам размещение в городе {best_city}')

    # Координаты лучшей точки
    start_lat = the_best.loc['lat']
    start_lng = the_best.loc['lng']
    start_point = (start_lat, start_lng)

    # Функция для расчета расстояния между двумя точками
    def calculate_distance(row):
        point = (row['lat'], row['lng'])
        return geodesic(start_point, point).meters

    # Применение функции к каждой строке датафрейма и создание нового столбца 'distance'
    budget_airbnb['distance'] = budget_airbnb.apply(calculate_distance, axis=1)

    # Формируем список точек в "лучшем" городе в пределах {distance} метров от точки с наивысшим рейтингом
    budget_airbnb_city = budget_airbnb.query(f'city == "{best_city}" and distance <= {dist}')
```

Полученное решение



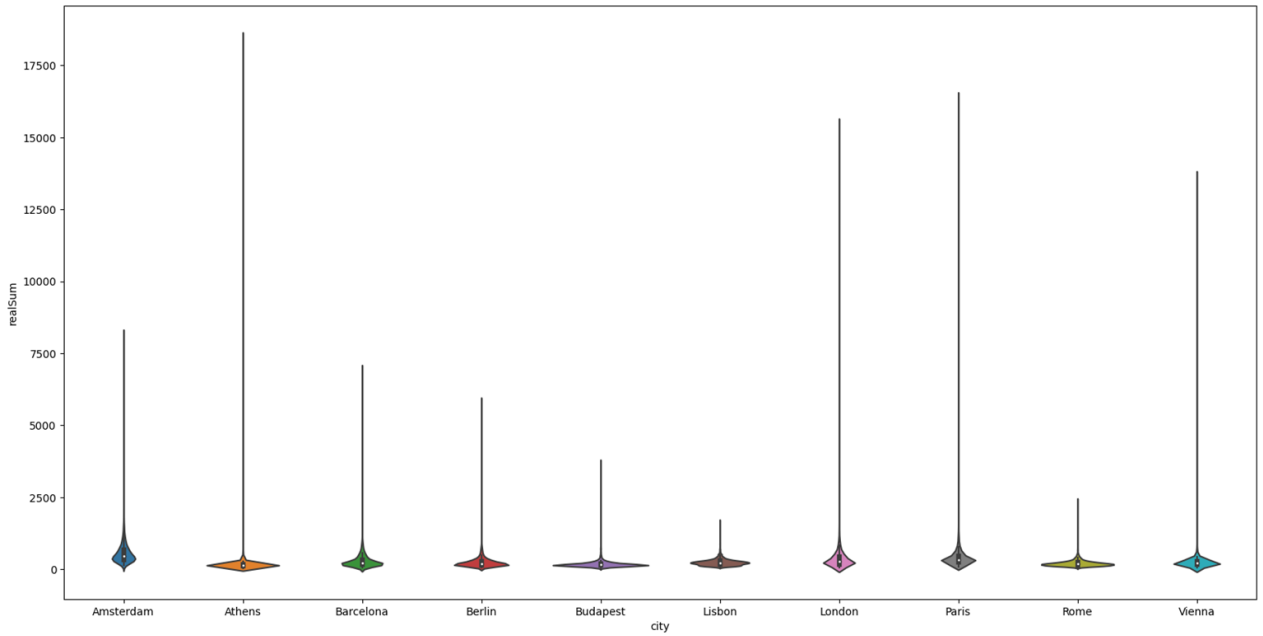
Количество студентов: 15
Бюджет: 10 000
Количество дней: 5
Радиус расселения: 300 м (по прямой)
Будни

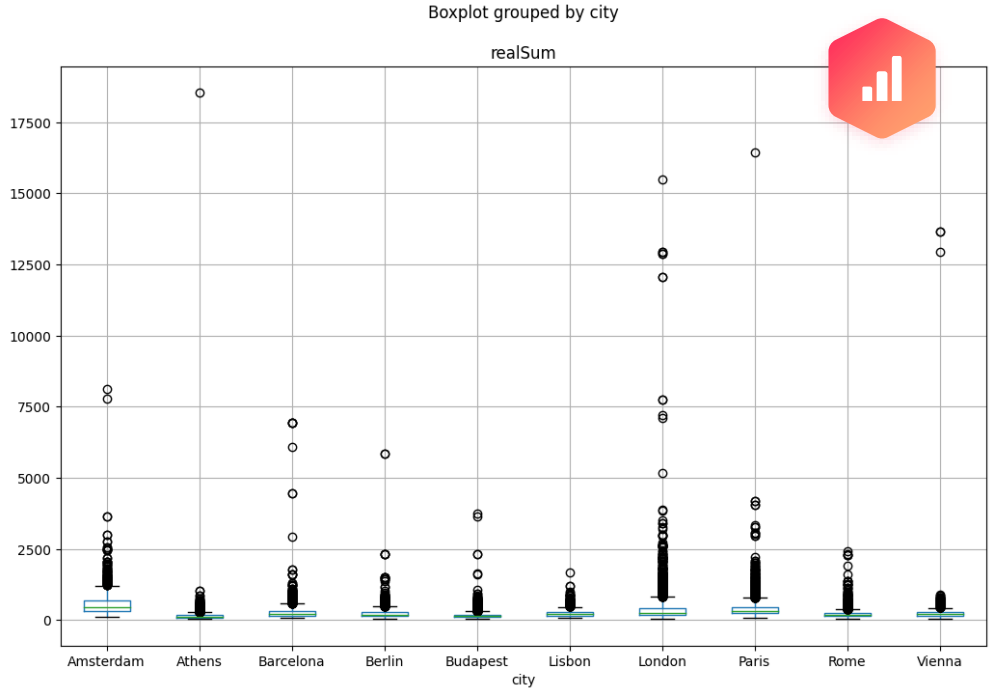
```
[ ]  rec_placement(opti_air, 15, 5, 10_000, True, 300)
```

Почему Афины?

Наименьшая медиана:

261.2949504927209	103.366
460.24418250415954	184.3085
127.71541724275303	**37.0941**
208.29939255707868	94.258385
191.17509582125828	78.7604
152.98209334022656	45.8241
225.375234521576	68.4803
317.5971665579271	121.6329
182.59182194374955	57.79347
208.49402800177643	66.61524



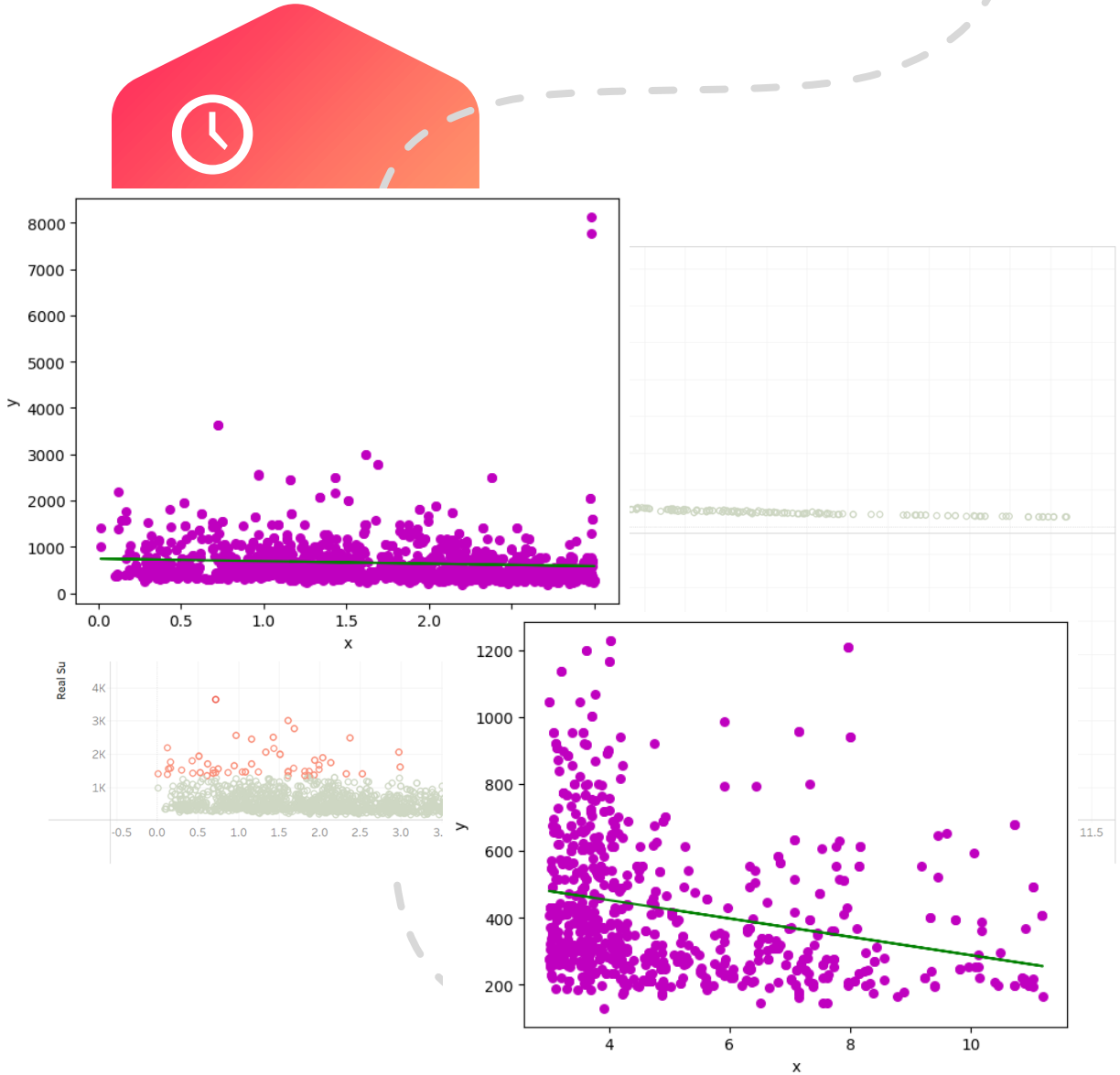


Что еще?

Профиль инвестора

Было предположение, что более драматично падает цена за съем, начиная с определенного расстояния от центра. Возможно, пока рейтинг ресторанов остается на приемлемом уровне, цена так сильно не опускается.
Так как земля в центре, вероятно, стоит дороже, то для инвестора было бы выгодным размещать объекты на границе в 3 км от центра.

Предположение не подтвердилось:
До 3 км: y = -53.92*x + 743.68
После 3 км: y = -27.41*x + 562.06



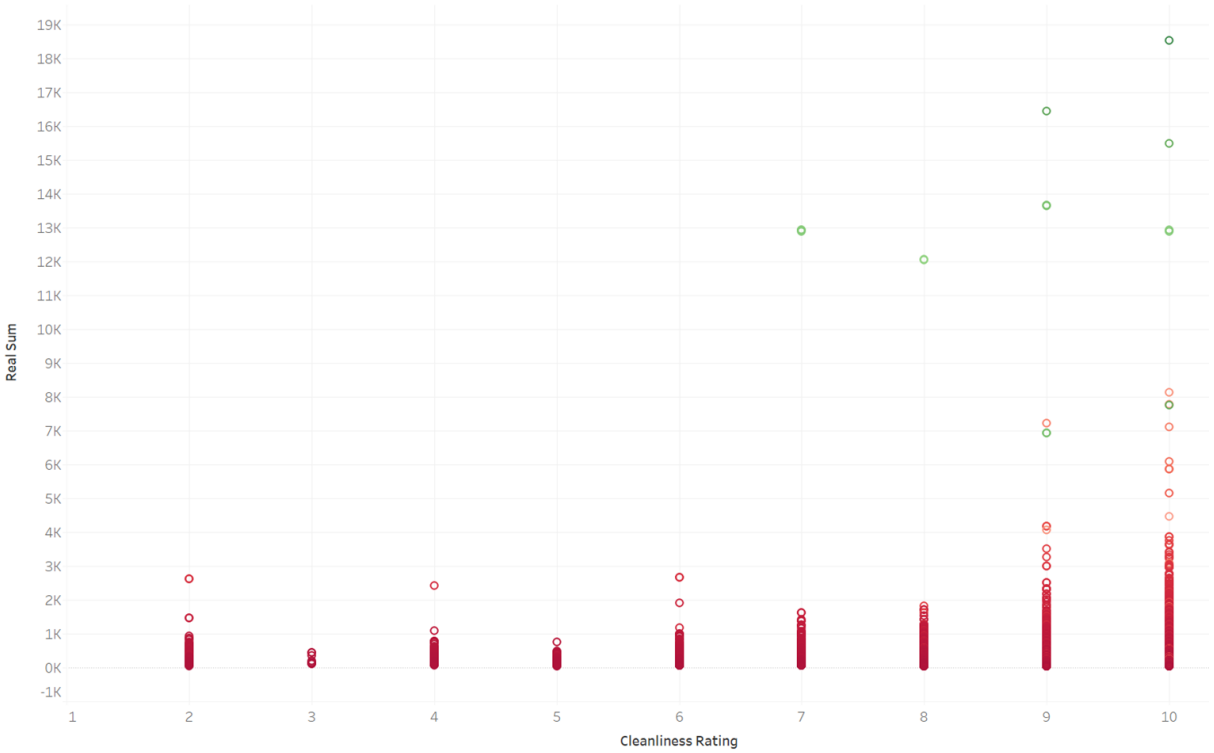
Что еще?

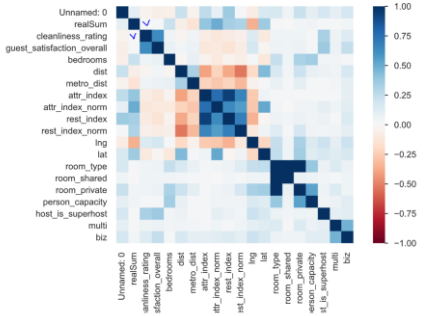
Профиль инвестора

Аналогичное предположение с привлекательностью района – не подтверждается

Чтобы найти наиболее инвестиционно привлекательный объект для вложений.
Не хватает данных:

- О расходах на обслуживание
- Площади помещений
- О ценах за квадратный метр
- О средних зарплатах в городе (для понимания расходов на сотрудников)





Что еще?

Профиль инвестора

Казалось, что начиная только с уровня оценки качества клининга в 7, стоимость сдачи начинает расти. Тогда логичней либо вообще не вкладываться в клининг, либо делать это на должном уровне.
Однако, видим, что реальной корреляции между качеством клининга и стоимости сдачи вообще нет:

THANK YOU!