Производственная практика
на тему «Предсказание доходов ресторана»

Выполнил:
Сенкевич Эдвард

1. Исследование данных

Основные поля в наборе данных

Можем выделить следующие поля:

id — идентификационный номер ресторана.

Open Date — дата открытия ресторана.

Type — тип ресторана, делится на IL, FC, DT.

City — город в котором находится ресторан.

City Group — тип города к которому принадлежит ресторан (Big cities or Other).

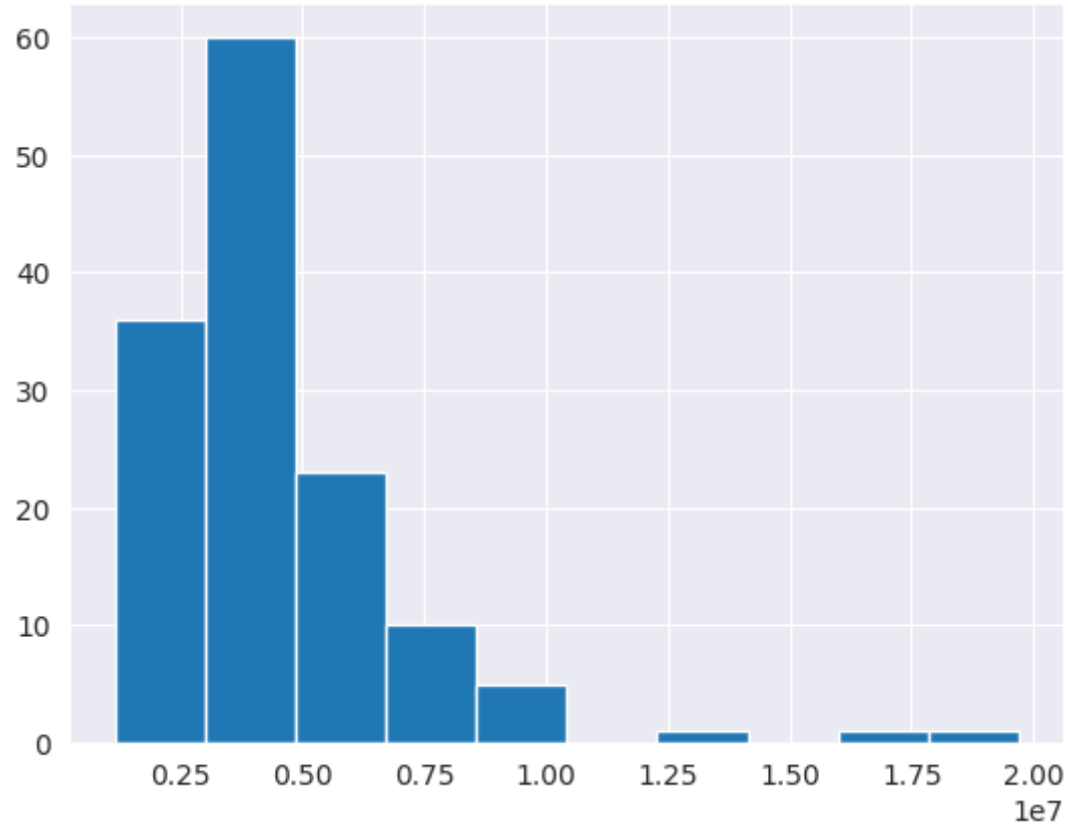
P1..P37 — специфические параметры ресторана.

Revenue — Доход ресторана.

Исследование набора данных

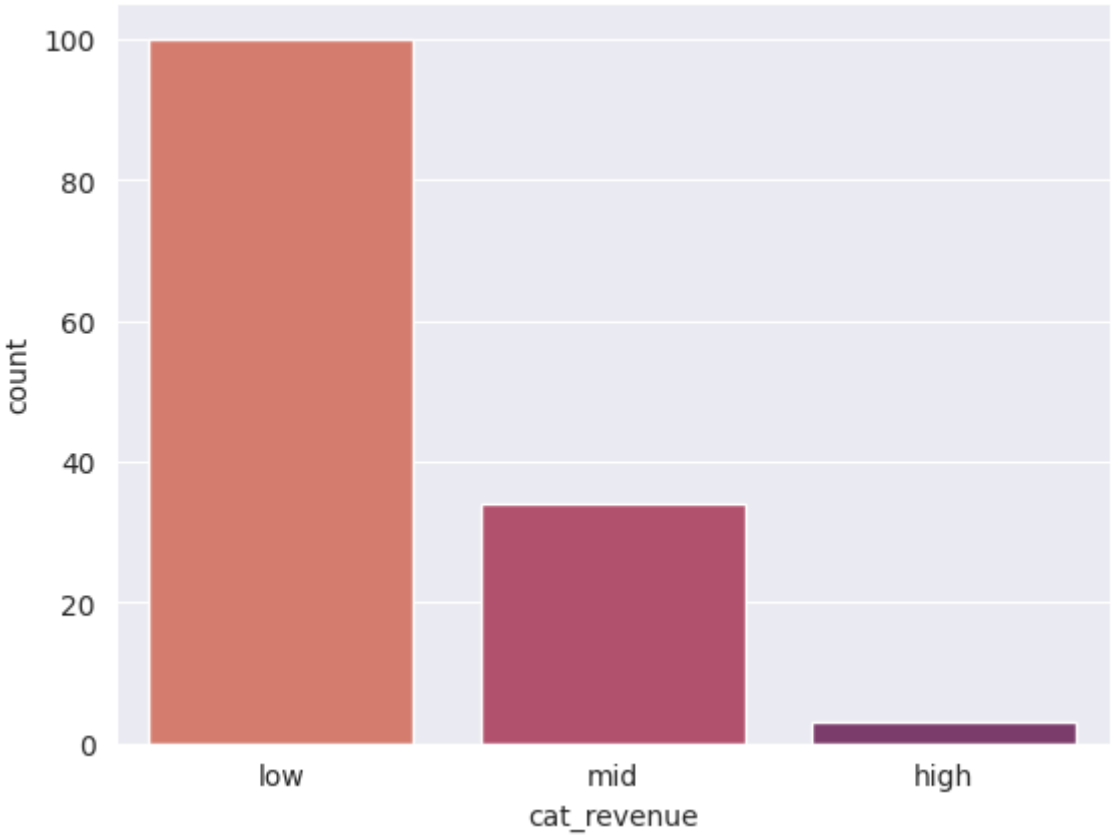
Построим гистограмму доходов всех ресторанов в тренировочном наборе данных:

Здесь мы видим что наибольший доход ресторанов около 60 единиц, минимальный 5 единиц.

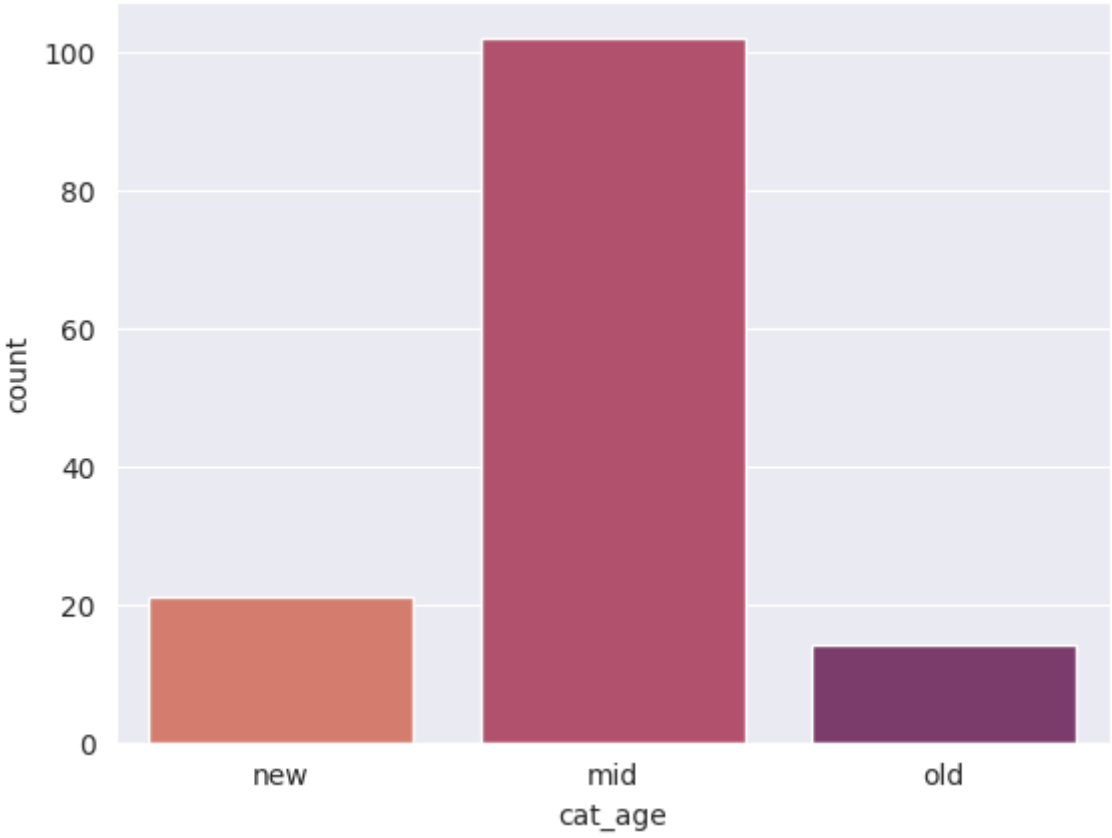


Можем для простоты восприятия разделить данные на «высокий», «средний» и «низкий» доход, разделение происходило по доходам ниже 0, 5000000, 10000000 соответственно:

Можно заметить, что низкий доход преобладает в наборе данных.

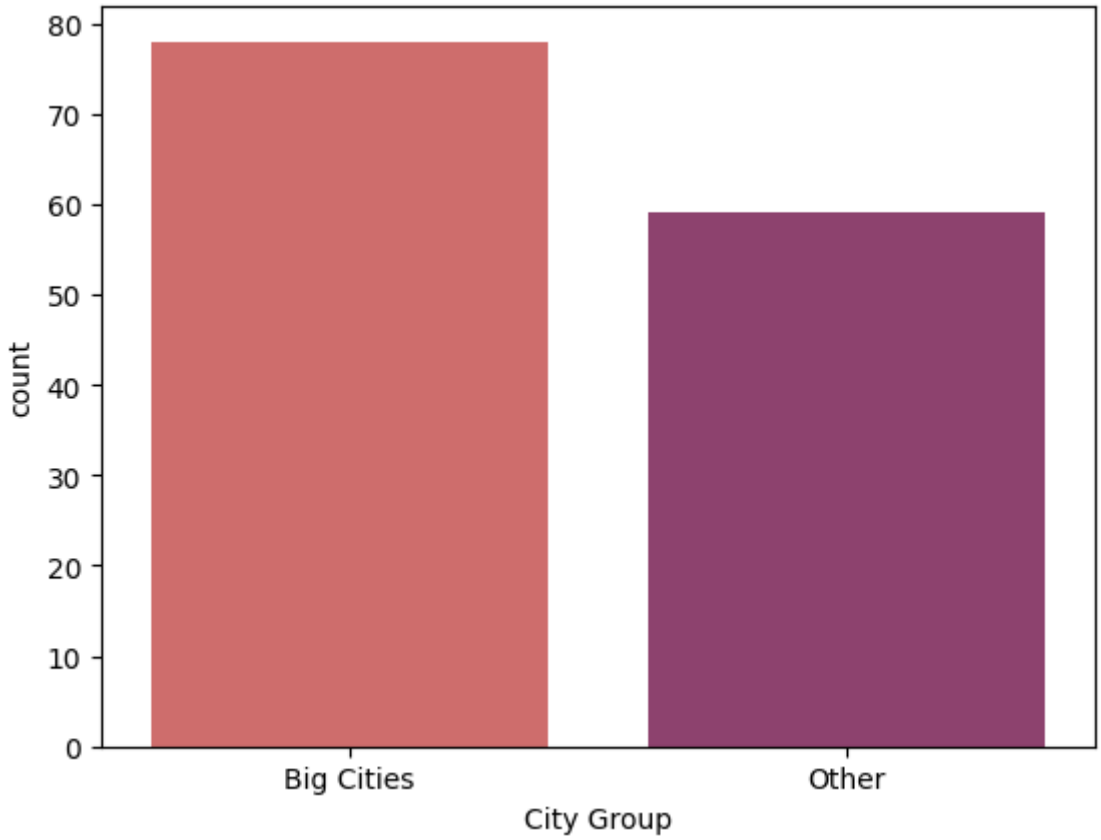


Проанализируем возраст ресторанов:



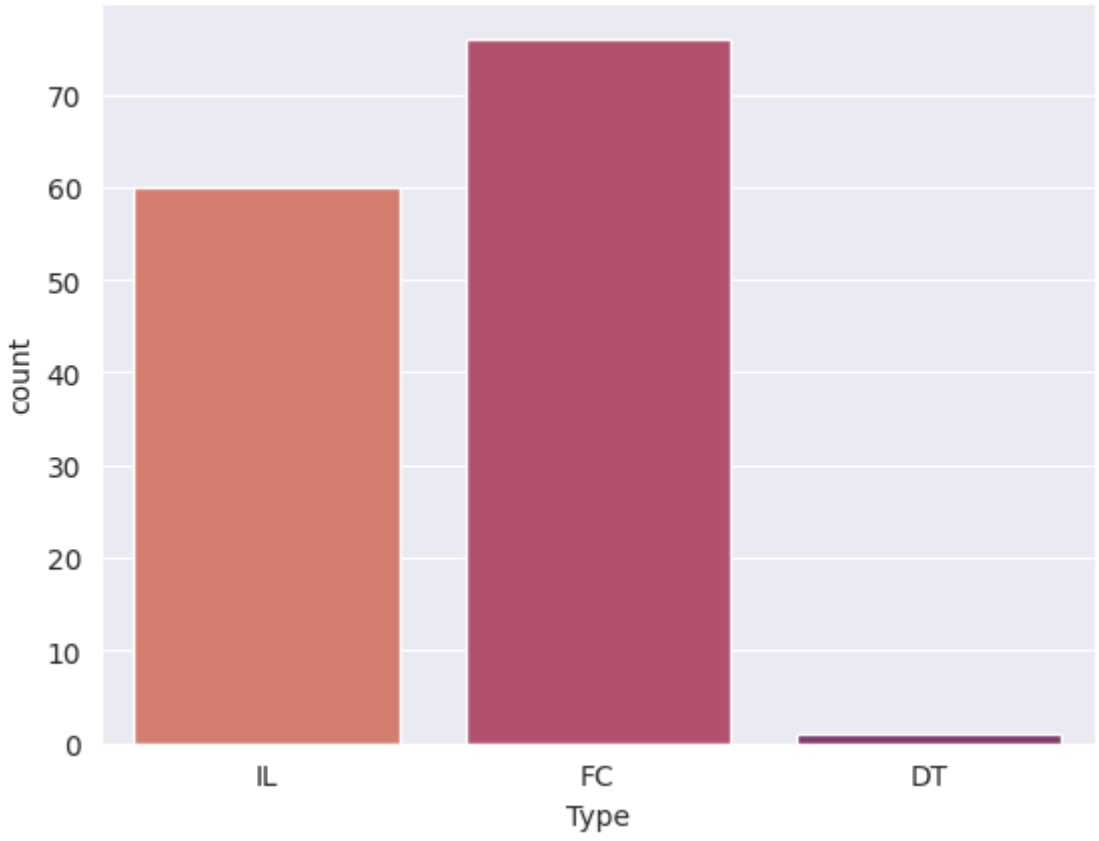
Среди них преобладают рестораны со средним возрастом.

Теперь посмотрим распределением по группам городов (City Groups):



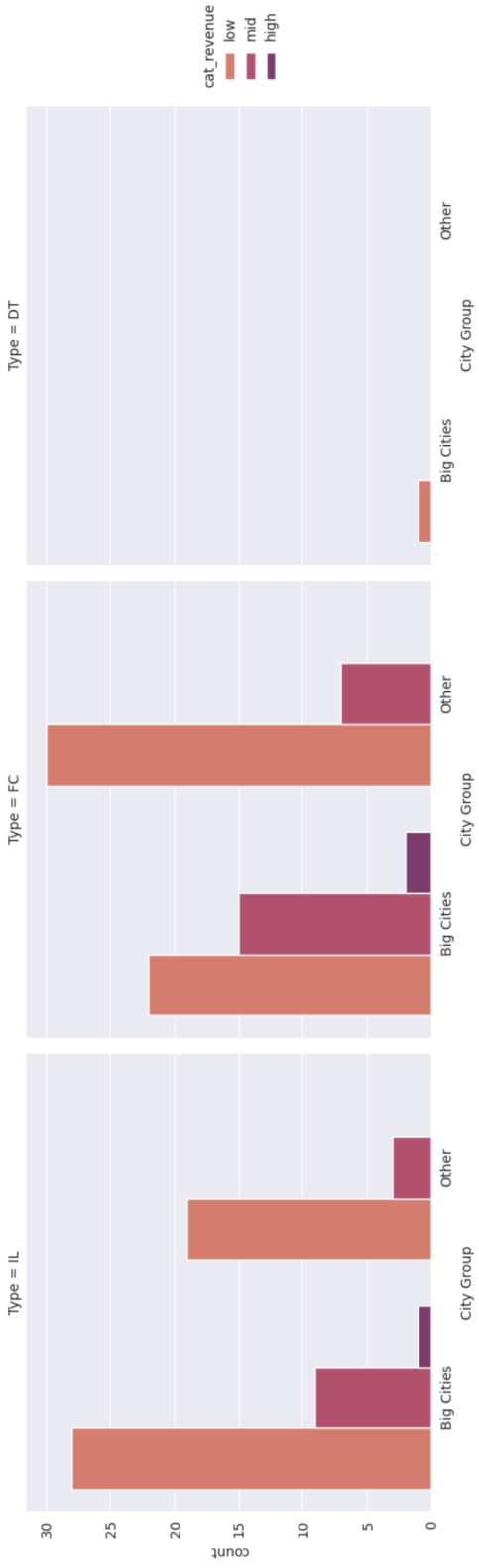
Количество ресторанов в больших городах **не сильно** преобладает чем в других.

Распределение ресторанов по типам:



Количество ресторанов с типом FC **не сильно** отличается от типа IL, ресторанов DT типа меньше всего.

Посмотрим на доходы ресторанов в разных группах городов для конкретного типа:

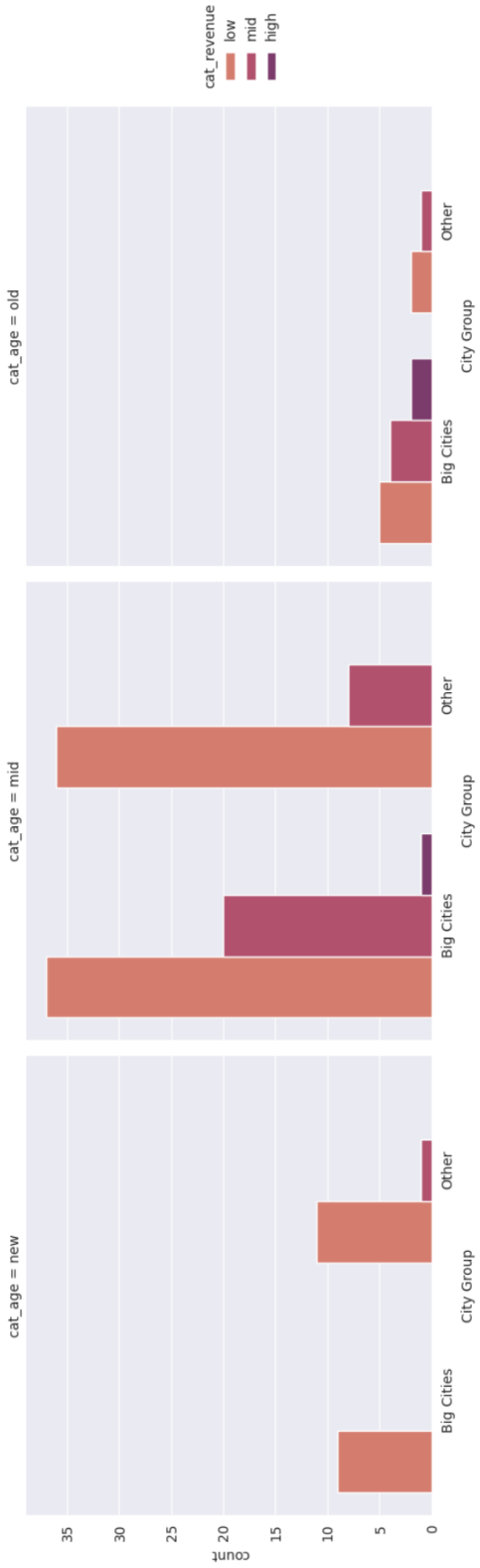


Можем заметить, что:

- рестораны с большим доходом находятся в больших городах.

- рестораны типа DT по показателям сильно отстают от других типов, можно предположить, что такие рестораны пользуются малой популярностью.

Теперь посмотрим на доход ресторанов в разных городах, разделенных по возрасту:



Можем заметить, что:

- Рестораны с большим доходом находятся в больших городах

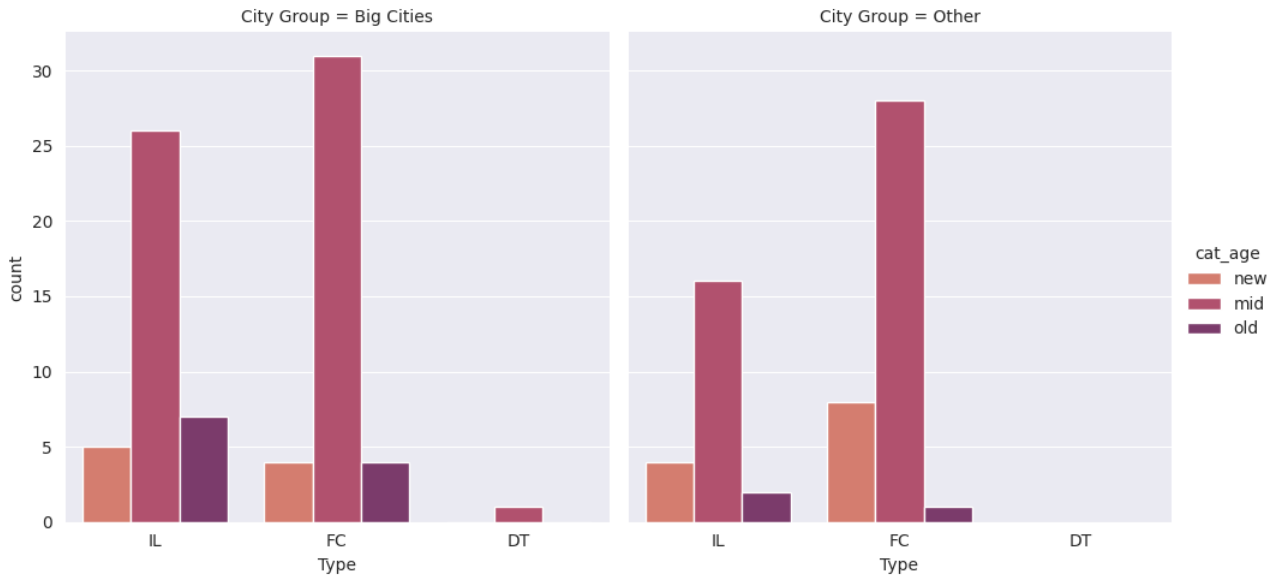
- Можем сказать, что количество ресторанов со средним доходом в среднем и молодом возрасте почти одинаково.
- количество новых ресторанов в крупных городах невелико по сравнению с другими городами. возможно, из-за высокой стоимости строительства в больших городах.

Разделим рестораны на две группы: «большие города» и «другие», и посмотрим на доход ресторанов по их типам в этих группах:

Можно заметить, что:

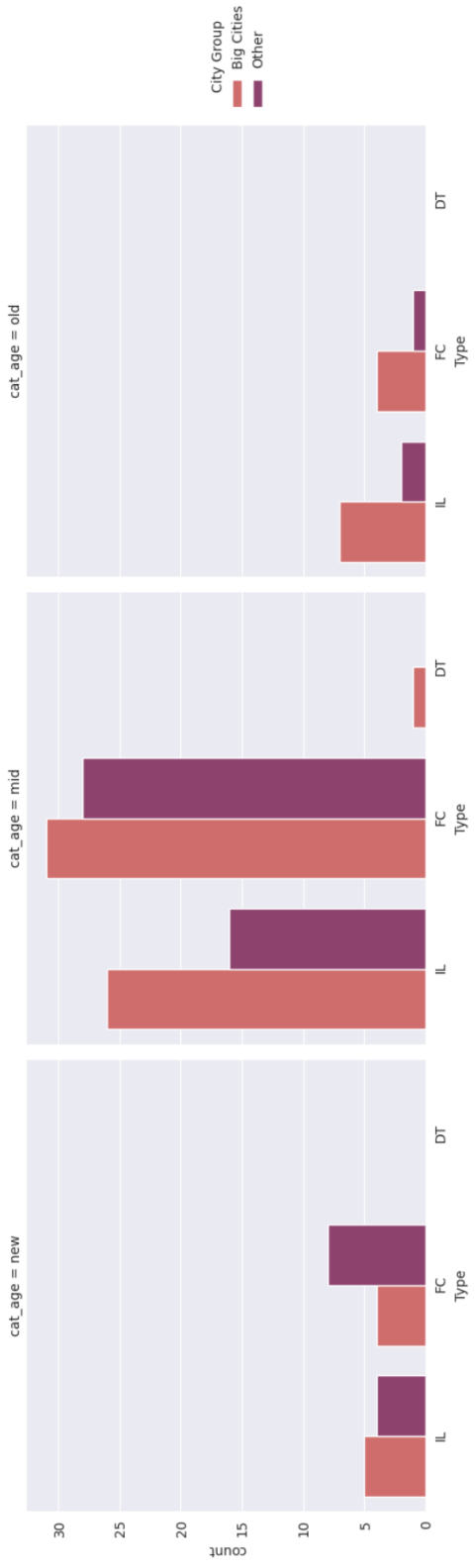
- Общее количество ресторанов типа FC больше, чем других.
- В обоих типах городов FC имеет наибольший доход.
- В других городах слабый доход, возможно это связано с количеством населения.

Далее рассмотрим такое же разделение городов по группам как в оценке выше, только теперь будем смотреть на возраст ресторанов:



Можно заметить, что:

- Большинство ресторанов среднего возрастаю .
- Количество новых ресторанов FC ресторанов в другой группе городов, больше, возможно из-за дешевизны строительства таких ресторанов в этой группе городов.
- Если детально рассмотреть группу городов по каждому типу, и разделенных по возрасту:



Можем заметить, что:

- FC ресторанов в других городах больше, чем в больших городах, что противоположно другим типам.

Выводы:

- Рестораны с большим доходом находятся в больших городах.
- Низкий доход преобладает в наборе данных.
- В наборе преобладают рестораны со средним возрастом.
- Количество ресторанов в больших городах не сильно преобладает чем в других.
- Количество ресторанов с типом FC не сильно отличается от типа IL, ресторанов DT типа меньше всего.
- Рестораны типа DT по показателям сильно отстают от других типов, можно предположить, что такие рестораны пользуются малой популярностью.
- Можем сказать, что количество ресторанов со средним доходом в среднем и молодом возрасте почти одинаково.
- Количество новых ресторанов в крупных городах невелико по сравнению с другими городами. возможно, из-за высокой стоимости строительства в больших городах.
- Общее количество ресторанов типа FC больше, чем других.
- В обоих типах городов FC имеет наибольший доход.
- В других городах слабый доход, возможно это связано с количеством населения.
- Большинство ресторанов среднего возрастаю .
- Количество новых ресторанов FC ресторанов в другой группе городов, больше, возможно из-за дешевизны строительства таких ресторанов в этой группе городов.
- FC ресторанов в других городах больше, чем в больших городах, что противоположно другим типам.
- **Тип FC лучше чем другие по всем показателям.**

Машинное обучение

Для предсказания доходов идеально подходят алгоритмы машинного обучения, по сути они помогают найти корреляцию между входным и выходным набором данных, и на основе этого давать предсказания.

Были выбраны следующие модели:

- K-nearest neighbors algorithm
- Random Forest Regression
- Light Gradient Boosting Machine
- Linear regression

Первые три модели используются для предсказания промежуточных данных. То есть своих локальных предсказаний на основе одного входного набора данных.

Вычислим их абсолютные ошибки, то есть абсолютная сумма разницы между валидными данными и предсказанными, деленная на общее количество данных.

KNN mean_absolute_error: 1608334.7431

RandomForest mean_absolute_error: 1670680.2016

LightGBM mean_absolute_error: 1560284.4391

Можно заметить что Random Forest дает большую ошибку.

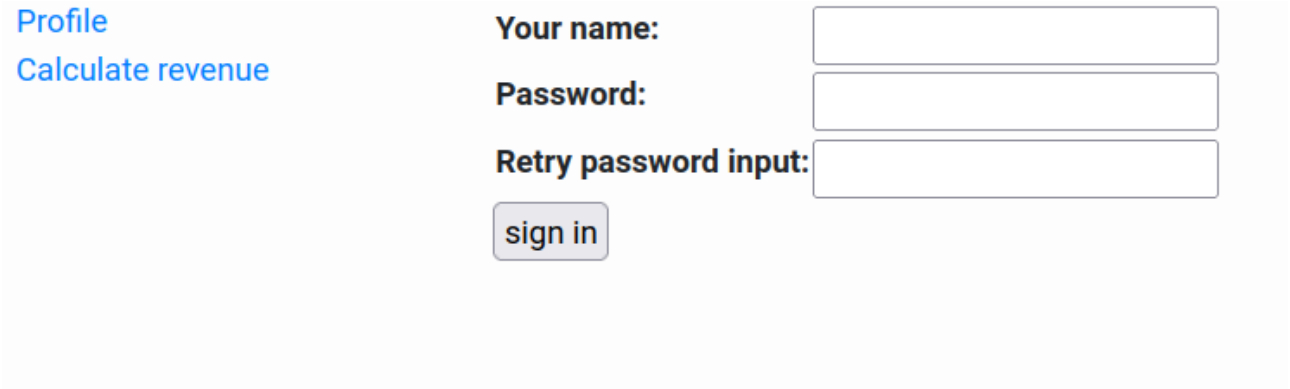
Далее предсказания каждой модели мы подаем на вход линейного регрессора (не изменяя веса каждого входа, то есть в нашем случае вес равен 0,33). При этом абсолютная ошибка равна: 1623166.3113, что примерно является средним значением ошибок входных данных.

Веб-приложение

Для демонстрации приложения была выбрана веб версия. В качестве фреймворка использовался Django, в качестве ORM встроенная DjangoORM, в качестве бд — SQLite.

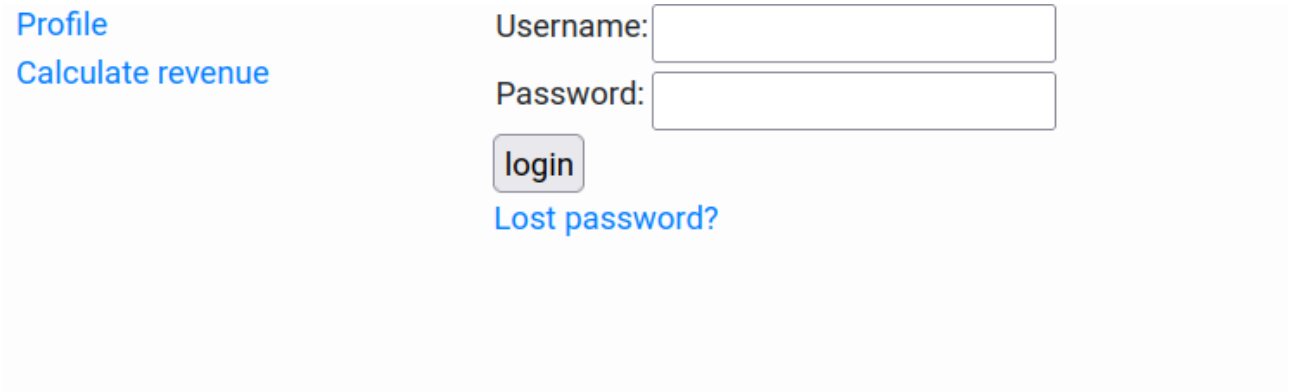
Основной сценарий описан в задании.

Регистрация пользователя



Пользователь вводит данные и подтверждает их нажимая «sign up», далее он переходит на следующий экран, по прежде чем рассматривать это, возможна ситуация когда пользователь уже зарегистрирован, поэтому рассмотрим вход пользователя.

Вход пользователя

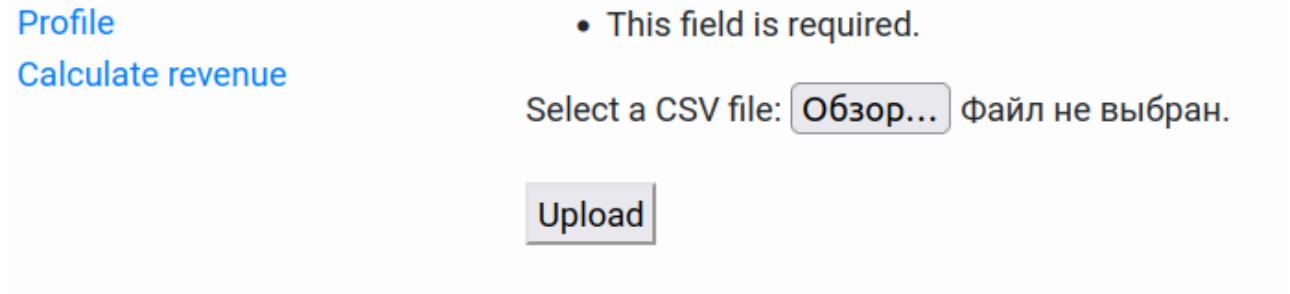


В данном случае пользователь вводит уже существующие данные подтверждает их кнопкой login и входит в систему. Попадая на главный экран, где он может поменять свои данные и посмотреть уже вычисленные им данные

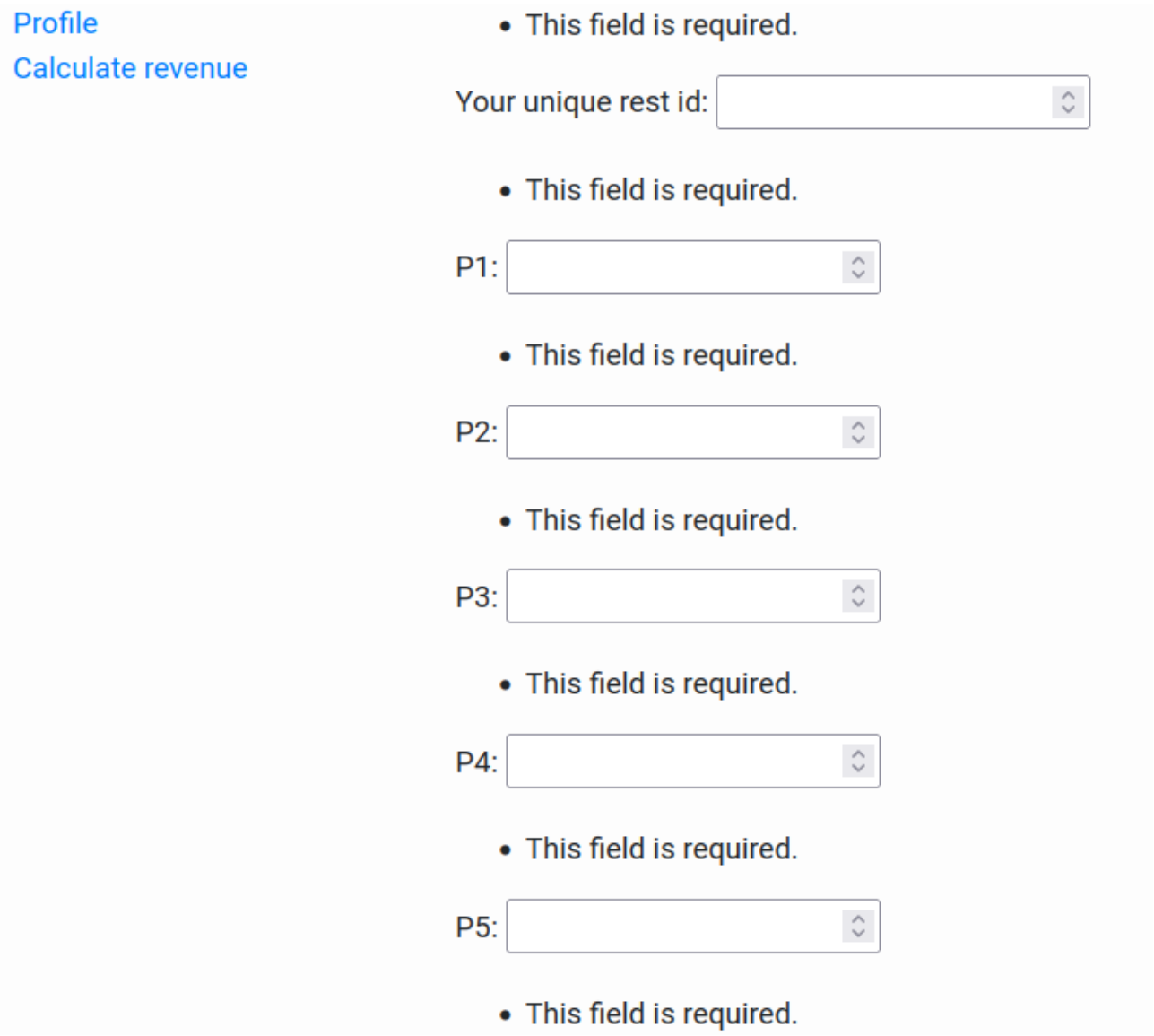
Главный экран

Страницы с вычислениями

1 страница позволяет загрузить таблицу с данными и получить в ответе таблицу с результатами, полезно для аналитики большого количества данных.



2 страница позволяет загрузить едичиные данные и получить сразу резальтат.



Результат:

City: Istambul

City Group: Big Cities

Type: FC

Upload

[4350193.44864116]