Академическая гимназия имени Д.К. Фаддеева
Санкт-Петербургского государственного университета

Анализ корреляции дополнительных услуг и рейтинга заведений Horeca (кафе)

Научный руководитель:
Мултанен Татьяна Генриховна
Выполнили:
Рыжкова Валерия Игоревна
Карташова Екатерина Алексеевна

Санкт-Петербург
2023 г.

Содержание

1. Общие сведения

1.1. Контекст

Сначала стоит упомянуть, что проект выполнялся в рамках хакатона по анализу данных «DANO» от НИУ ВШЭ. На этом хакатоне нам необходимо было исследовать предложенный датасет, презентовав свою работу жюри. Никаких ограничений указано не было, поэтому тему, способ, цель, формат исследования мы определяли самостоятельно.

1.2. Введение

Посещение кафе и ресторанов стало неотъемлемой частью жизни жителей России. Для удобства выбора существует система рейтингов и оценок, которые пользователи могут оставлять посещенным местам. При выборе кафе или ресторана люди часто ориентируются на рейтинги, составленные предыдущими посетителями, что делает рейтинги важными и для самих организаций. В исследовании мы рассмотрели отзывы, оставленные жителями Москвы и Санкт-Петербурга на предприятия общественного питания (далее — организации), в частности кафе, работающие в соответствующих городах.

1.3. База данных

Данные получены с сервиса “Яндекс Карты” за 3 года, начиная с февраля 2018, в обоих городах жителями было оставлено свыше 3 млн. отзывов, из которых было случайно отобрано 104,278. Пользователи могут как посещать организации в своем городе (жители Москвы — в Москве, жители Санкт-Петербурга — в Санкт-Петербурге), так и путешествовать и оставлять отзывы в чужом (москвичи в организациях Санкт-Петербурга и наоборот). Данные об отзывах (оценка, город и дата) были объединены с данными об организациях (средний рейтинг, средний чек, город и направления/особенности деятельности).

Описание переменных:

user_id — идентификатор пользователя (93,786 уникальных значений).

org_id — идентификатор организации (22,925 уникальных значений).

rating — поставленная оценка в отзыве (от 1 до 5).

ts — дата отзыва (последовательный номер дня, начиная с 01.02.2018 — это 0 день).

user_city — информация о городе проживания пользователя (msk — Москва, spb — Санкт-Петербург).

org_city — информация о городе организации (msk — Москва, spb — Санкт-Петербург).

average_bill — средний чек в рублях (округленный с точностью до 500 рублей).
rating_org — средний рейтинг организации (от 1 до 5).
rubrics — рубрика, указывающая на основной вид деятельности организации (например, «Ресторан», «Кафе», «Быстрое питание» и т. п.).
food_delivery, handmade_goods, …— 63 последних столбца в таблице представляют собой дамми-переменные (или бинарные переменные), связанные с особенностями организаций. Каждая из них принимает значение 1, если данная особенность относится к организации, и 0, если нет. Например, если переменная breakfast равна 1, это означает, что в организации подают завтраки. Как правило, множество особенностей организации определяется ее владельцем и может быть неточным.

2. Гипотеза

2.1. Актуальность

Почему это важно и актуально:

- В России предпринимательской деятельностью занимаются примерно 3% населения, 4,2 млн человек (по данным из исследования компании “Сбербанк”, которые были получены анализом big data о движении средств по счетам и учитывают только реально действующие юрлица и ИП, по данным РБК)
- Многие предприниматели неграмотно оценивают ситуацию на рынке, вследствие чего имеют убыточный бизнес, подверженный банкротству.
- Для того, чтобы грамотно выстроить свой бизнес, необходимо провести ряд исследований, которые покажут, насколько бизнес прибылен, какие технологии стоит внедрять и на какие не стоить тратить средства
- Одним из факторов, оказывающих влияние на прибыльность заведения является наличие определенных дополнительных услуг. Согласно последним исследованиям, посетители кафе чаще обращают внимание на качество обслуживания, внешнюю обстановку, чем на вкус еды.
- Отзывы - неотъемлемая часть любого продукта. Согласно исследованиям, Более 50% потребителей предварительно читают онлайн-отзывы о продукте, и 95% покупателей читают отзывы, размещенные на сайтах-отзовиках. Исходя из того, что посетители кафе прежде чем выбрать место для приема пищи обращают внимание на отзывы и рейтинги кафе изменяются в зависимости от наличия или отсутствия той или иной услуги, мы пришли к выводу о том что существуют наборы оптимальных дополнительных услуг, которые смогли вы внедрить предприниматели в свои кафе для повышения отзывов и прибыли как следствие.

2.2. Гипотезы и цель исследования

Гипотезы:

- Наш продукт будет актуален для целевой аудитории
- Непостоянные посетители, которые являются целевой аудиторией, выбирают кафе с помощью отзывов
- Дополнительные услуги влияют на рейтинг заведения

Цели:

- Выявить список тех дополнительных услуг, которые могут повысить рейтинг заведения.
- Создать список рекомендованных дополнительных услуг для кафе

3. Исследование

3.1. Задачи исследования

- Провести анализ рынка и выявить потребности целевой аудитории
- Провести анализ базы данных при помощи Excel, Python, библиотек Pandas, matplotlib и т.д.
- Выявить корреляционную зависимость рейтинга кафе от наличия/отсутствия дополнительных услуг
- Выдвинуть рекомендации для предпринимателей

3.2. Как формируются отзывы?

Оставленные отзывы формируются следующим образом:
Если человек оставил отзыв и при этом написал к нему какой-либо комментарий, то за этот отзыв ему начисляется 20 баллов, если пользователь оставил отзыв без комментария, то тогда он получает 5 баллов. В зависимости от количества баллов, отзывы оставленные пользователями, имеют разный вес. С ростом количества баллов пользователя, вес его отзывов растет. То есть на рейтинг заведения влияют следующие факторы:

- Достоверность — ставил ли пользователь оценки раньше и сколько их было. Если ставил, его оценка имеет больший вес. Если нет — оценка будет учитываться с меньшим весом.
- Влиятельность — как человек оценивал компании раньше. Если он всегда ставил пятерки, а новому заведению поставил четыре, алгоритмы понимают, что этого посетителя довольно сложно расстроить. Поэтому его четверка повлияет на рейтинг компании больше, чем если бы он снова поставил пятерку.

3.3. Целевая аудитория

1 категория: группа людей, предприниматели имеющие свой бизнес (кафе), который является убыточным, желающие выйти на точку безубыточности и предотвратить банкротство, боятся потерять репутацию среди знакомых и друзей

2 категория: неопытные, начинающие предприниматели, которые только планируют открыть свой бизнес и не имеют достаточно знаний, боятся потерять деньги, не хотят тратить время на самостоятельный анализ и средства на заработную плату аналитикам

3 категория: предприниматели сетевики, которые планируют открыть еще точки

4 категория: аналитики, маркетологи, не желающие тратить время на работу, которые хотят выделить время под саморазвитие или выполнение других действий

5 категория: аналитики, маркетологи которые желают возвыситься в глазах управляющих, боятся неодобрений, ошибок

3.4. SWOT-анализ

Мы провели внешний и внутренний анализ.

Сильные стороны проекта:

- Инновационность в использовании данных с отзывом посетителей (данный вид анализа с выявлением определенного набора дополнительных услуг еще не был применен на рынке, есть только подобные исследования, которые не включают в себя получение набора дополнительных услуг, которые будут гарантировать успешный бизнес)
- Легко контролировать процесс и управлять изменениями (небольшое количество людей, маленький объем выполняемых параллельно процессов)

Слабые стороны проекта:

- Отсутствие опыта в работе
- Возможность наличия погрешностей (проводится анализ данных за определенный период, поэтому нет гарантий, что полученные результаты все еще актуальны)
- Универсальность сервиса (нет индивидуального подхода к каждому клиенту, есть только общий набор данных)

Возможности:

- Нет стартапа на рынке, выполняющего такие же функции
- Коллаборации с известными аналитическими компаниями (Озон, Яндекс)

Риски:

- Незаинтересованность ЦА

- Отсутствие доверия клиентов (из-за непопулярности на рынке, небольшая узнаваемость, низкое доверие из-за отсутствия рекламы)
- Появление новых игроков на рынке (появление конкурентов, использующих аналогичную технологию, которые быстро стали пользоваться спросом у идентичной целевой аудитории)

3.5. Конкуренты

1) Исследовательская компания «Restteam»

- Аудит ресторана, бара и кафе с проработкой рекомендаций
- Предлагают полный аудит, операционный аудит, маркетинговый аудит, аудит кухни ресторана, кафе или бара.
- Анализ соответствия выбранной концепции месту расположения предприятия. Трафик пешеходного потока.
- Анализ соответствия выбранной концепции месту расположения предприятия. Трафик пешеходного потока.
- Анализ наличия диссонансов в процессе нахождения гостя в ресторане (интерьер, декор, функц. зонирование, музыка, температура, посторонние запахи/звуки, мебель, доп. услуги, посуда, сервировка столов, туалет и т.п.)
- Анализ меню (отчёты по продажам - себестоимость/кол-во, наполняемость категорий, презентации подачи блюд/напитков, скорость приготовления)
- Анализ организации системы обслуживания (униформа и внешний вид сотрудников, общение с гостями, скорость обслуживания, работа в конфликтных ситуациях, использования инструментов продаж, чтение гостей, расстановка приоритетов)
- Анализ работы менеджера/администратора на смене с гостями (визиты к столу, аккумуляция обратной связи гостей, работа в конфликтной ситуации, «восьмёрка»)
- Анализ маркетинговых зон ресторана (мерчендайзинг, оформление торговых мест)
- Анализ предложения конкурентов (ассортимент, цены, нюансы концепции)
- Анализ маркетинг-плана и проводимых в заведении акций локального маркетинга

- Стоимость 88000 рублей
- Собирают необходимые данные 1-3 дня, т.е. многие факторы могут не учесть
- За выезд нужно отдельно оплачивать проживание и все расходы

2) Исследовательская компания «Mozg.rest»

- Отчет показывает динамику по наполняемости блюд и напитков в разрезе дня и вечера.
- Помогают быстро определить тенденцию и работать на исправление ошибок при снижении наполняемости.
- Сравнительный анализ выручки, посещаемости, среднего чека по дням в сравнении с планом.
- Показывают по какой причине не выполнен/выполнен план: гости или средний чек.
- Стоимость 8000-10000
- Не предоставляют конкретный анализ дополнительных услуг
- Предоставляют бесплатную 30 минутную консультацию
- Имеют крупных сетевых ресторанов и кафе в числе клиентов

3) Исследовательская компания «trade-drive»

- Убрать невыгодные позиции и добавить нужные,
- Ликвидировать ценовую и смысловую конкуренцию блюд,
- Правильно отрегулировать выход блюд в граммах,
- Отрегулировать наценки на блюда и пр.
- Ввести стимулирующие акции и предложения, точно соответствующие ожиданиям разных групп гостей.
- Разумно сократить издержки, списания сырья, ликвидировать «недовложения» в тарелку гостя.
- Четко определить концепцию вашего ресторана глазами гостя.
- Составить портрет вашей ключевой аудитории.
- Предложить каждой целевой группе оптимальное соотношение цены и качества.
- Стоимость 17000-25000 рублей
- Не предоставляют конкретные фичи, которые можно добавить; не анализируют конкурентов

- Работают удаленно

Исходя из анализа конкурентов, мы сделали следующие выводы:

Достоинства конкурентов: индивидуальный анализ с предварительным сбором данных, выдвижение персональных рекомендаций каждому клиенту, полные анализы, предоставление бесплатных консультаций

Недостатки конкурентов: дороговизна, длительный период ожидания результатов исследования, возможность ошибок и неучет некоторых факторов

3.6. Какой рейтинг считается хорошим?

Чтобы вывести наборы дополнительных услуг, которые, положительно повлияли бы на рейтинг заведений, нам нужно было определить, какой рейтинг является хорошим. Согласно данным исследования компании Aliexpress: отрицательная оценка — 1-2 звезды, нейтральная — 3 звезды, положительная — 4-5 звезд. Из чего можно сделать вывод, что “хорошим” считается рейтинг в интервале от 4 до 5, то есть, заведения с таким рейтингом содержат дополнительные услуги, которые следует внедрить предпринимателю, желающему открыть свое новое заведение. Далее на сайте компании “Совкомбанк” мы нашли информацию о том, что максимально возможный рейтинг не вызывает доверие у людей, рейтинг “5,0” мы уже исключили из наших данных. В исследованиях компании также говорилось, что если максимальный рейтинг = “5”, то лучше всего поднимать продажи будет рейтинг в диапазоне 4,5 – 4,7 баллов.

3.7. Сбор недостающих данных и сокращение выборки

3.7.1. Заполнение пропусков

Несмотря на то, что выбранная категория заведения(кафе) содержала наибольшее количество информации по оценкам пользователя и дополнительным услугам, данные по некоторым значениям рейтинга кафе отсутствовали, сокращать выборку на данном этапе не представлялось приемлемым. По этой причине мы при помощи сводных таблиц взяли среднее значение всех оставленных оценок пользователей каждого из заведений и заполнили этими данными рейтинг заведения, там где он отсутствовал.

3.7.2. Сокращение выборки

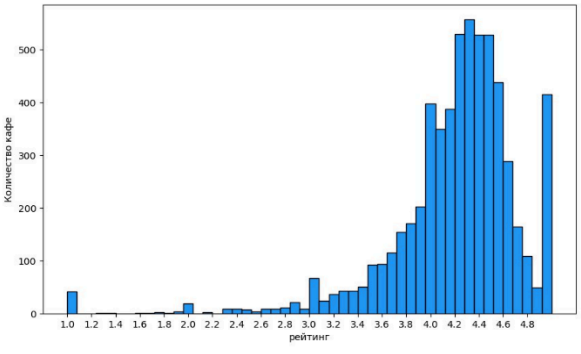
Так как мы решили искать наборы дополнительных услуг, которые положительно могут повлиять на рейтинг, нам необходимо сократить выборку и оставить только те заведения, у которых рейтинг выше определенного значения (мы приняли, что хороший рейтинг равен от 4.0 и выше), также, проанализировав данные на нормальность, отбросили заведения, у которых рейтинг 4.9 и выше, так как количество данных значений выше нормы. Это связано с тем, что мы при заполнении

пропущенных данных рейтинга заведения брали среднее значение всех оценок пользователей, которые они оставили конкретному заведению, но так как во многих заведениях пользователь оставил только одну оценку (во многих случаях это была оценка 5), то кафе со средним рейтингом 5 оказалось большое количество, поэтому мы приняли решение не рассматривать их при анализе.

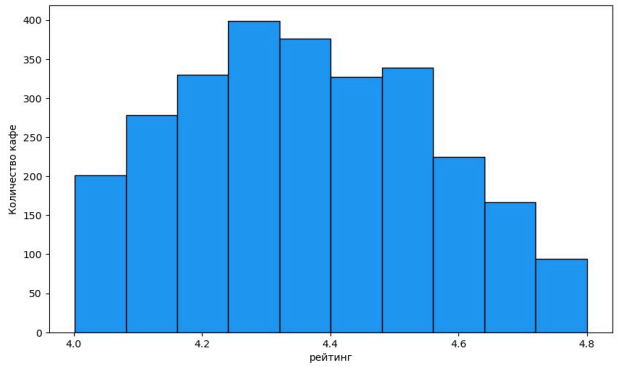
3.7.3. Проверка на нормальность

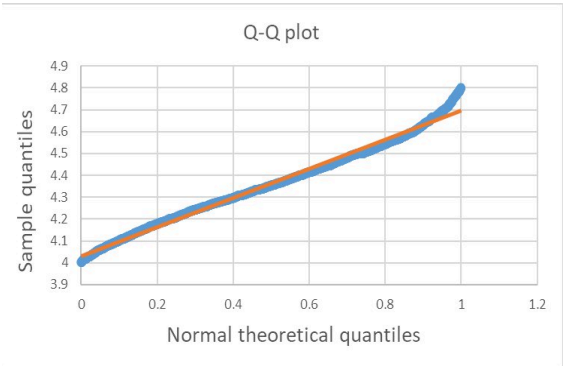
Нормальное распределение служит основной упрощенной моделью моделью для реальных процессов. Для дальнейшего анализа нам необходимо было проверить данные на нормальность. Мы использовали метод QQ (сокращение от графика “квантиль-квантиль”) в Excel. Данный график показывает, насколько выборочные значения данных соответствуют предсказанным значениям данных, если бы распределение было бы нормальным.Если два сравниваемых распределения похожи, точки на графике Q – Q будут приблизительно лежать на линии y = x. Для построения графика мы отсортировали данные по возрастанию и поставили ранговое значение. Далее мы рассчитали процентили, требуемы для подстановки по формуле $\frac{(R_k-0.5)}{n}$, где R_1 - это ранг k-го элемента в столбце, а n=const - количество рассматриваемых данных. Далее мы применили эту формулу для всех ячеек. Для расчета нормальных теоретических квантилей мы использовали формулу =НОРМ.СТ.ОБР, которая возвращает обратное значение стандартного нормального интегрального распределения. Для рассчета квантилей данных (Z-score) мы использовали формулу=НОРМАЛИЗАЦИЯ(A1,AVERAGE($A$2:$A$5995),СТАНДОТКЛОН($A$2:$A$5995)), которая возвращает нормализованное значение, использовав при этом данные находящиеся в ячейке A2 и скопировали формулу на все ячейки. Далее мы выделили ячейки с нормальным отклонением и полученным и построили точечный график QQ.

Сначала мы проверили на нормальность исходные данные, стандартное отклонение = 0.54420721.



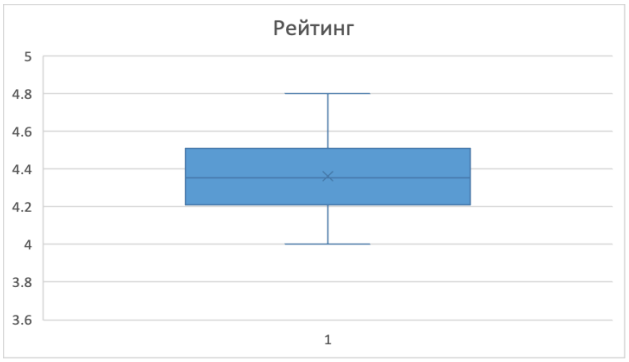
По графикам видно, что данное распределение не является нормальным. Тогда мы предположили, что необходимо сделать, для того, чтобы получить нормальное распределение. Мы проанализировали данные разных городов: Санкт-Петербург и Москва и проанализировали их этим же методом. В Москве отклонение оказалось наименьшим, поэтому для того, чтобы было проще привести к нормальному виду данные, мы выбрали именно их. По формулам, описанным выше мы доказали, что исследуемые нами данные имеют нормальное распределение.

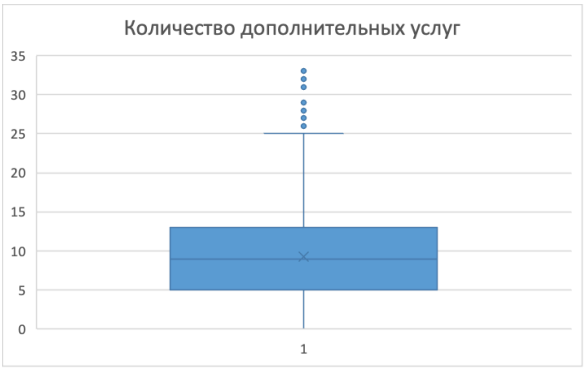




3.8. Описательная статистика

Построение графика размаха “Ящик с усами”





Медианное значение = 4,35

Размах = 0,9

Медианное значение = 9

Размах = 25

Для того, чтобы проанализировать данные, мы провели проверку на нормальность, которая показала нормальное распределение в г. Москва у заведений с рейтингом от 4,0 до 4,9. Исходя из этих данных можно сделать вывод о том, что для того, чтобы предприниматель повысил рейтинг у уже существующего своего заведения, ему следует внедрить наборы дополнительных услуг, содержащиеся в заведениях с рейтингом 4,5-4,7

3.9. Кластерный анализ

3.9.1. Определение

Кластерный анализ - анализ, предназначенный для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу (кластер) были максимально “схожи”, а элементы из разных групп были максимально “отличными” друг от друга.

3.9.2. Наша модель

Мы решили разбивать на группы те дополнительные услуги, которые могут повлиять на рейтинг. Мы преобразовали наши данные в другой формат, каждой доп.услуге присвоили номер (начиная с первой и далее по возрастанию). Для этого нам необходимо было посчитать средний рейтинг заведений, где есть та или иная доп.услуга и количество раз, сколько она встречается у всех организаций. Метод позволяет выявить наборы тех доп.услуг, которые хорошо взаимодействуют между собой и положительно влияют на рейтинг. Вот по как мы это рассчитывали:
Программа начинает с K случайно выбранных кластеров, а затем изменяет принадлежность объектов к ним, чтобы: минимизировать изменчивость внутри кластеров, и максимизировать изменчивость между кластерами. В кластеризации методом K средних программа перемещает объекты из одних групп (кластеров) в другие для того, чтобы получить наиболее значимый результат. этих кластеров. Число кластеров k задается исследователем заранее.
Метод k-средних – это метод кластерного анализа, цель которого является разделение m наблюдений (из пространства R^n) на k кластеров, при этом каждое наблюдение относится к тому кластеру, к центру (центроиду) которого оно ближе всего.

В качестве меры близости используется Евклидово расстояние:

$$\rho(x, y) = \|x - y\| = \sqrt{\sum_{p=1}^{n}(x_p - y_p)^2}$$, где $x, y \in R^n$

Итак, рассмотрим ряд наблюдений $(\mathbb{x}^{(1)}, \mathbb{x}^{(2)}, \ldots, \mathbb{x}^{(m)})$, $x^{(j)} \in R^n$.

Метод k-средних разделяет m наблюдений на k групп (или кластеров) ($k \leq m$) $\mathbf{S} = \{S_1, S_2, \ldots, S_k\}$, чтобы минимизировать суммарное квадратичное отклонение точек кластеров от центроидов этих кластеров:

$$\min\left[\sum_{i=1}^{k}\sum_{x^{(j)} \in S_i}\left\|x^{(j)} - \mu_i\right\|^2\right]$$, где $x^{(j)} \in R^n, \mu_i \in R^n$

μ_i - центроид для кластера S_i.

Рассмотрим первоначальный набор k средних (центроидов) $\mu_1, \ldots, \mu_k$ в кластерах $S_1, S_2, \ldots, S_k$. На первом этапе центроиды кластеров выбираются случайно или по определенному правилу (например, выбрать центроиды, максимизирующие начальные расстояния между кластерами).

Относим наблюдения к тем кластерам, чье среднее (центроид) к ним ближе всего. Каждое наблюдение принадлежит только к одному кластеру, даже если его можно отнести к двум и более кластерам.

Затем центроид каждого i-го кластера вычисляется по следующему правилу:

$$\mu_i = \frac{1}{s_i} \sum_{x^{(j)} \in S_i} x^{(j)}$$

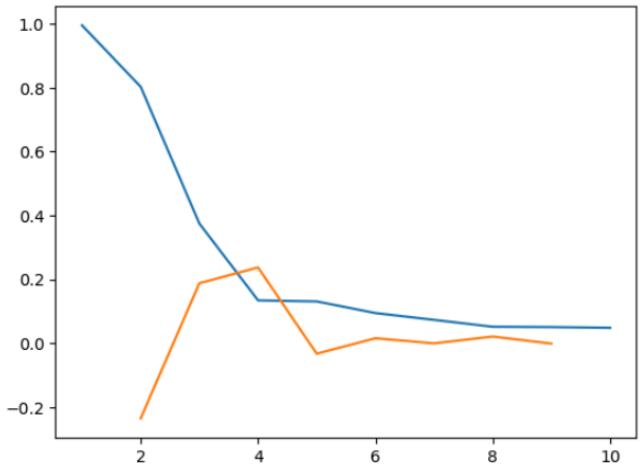
Таким образом, алгоритм k-средних заключается в перечислении на каждом шаге центроида для каждого кластера, полученного на предыдущем шаге.

Алгоритм останавливается, когда значения μ_i не меняются: $\mu_i^{\text{mar } t} = \mu_i^{\text{mar } t+1}$

3.9.3. Реализация исследования на Python

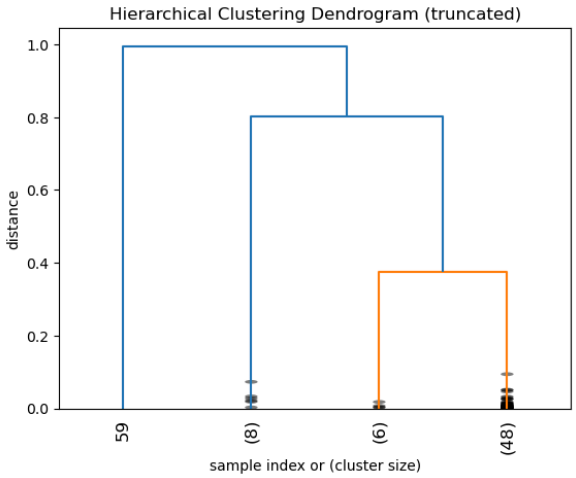
Общий алгоритм для исследования был такой (стоит учесть, что собрали дополнительные данные и очистили выборку мы до этого):

1) Сделать нормирование данных, чтобы привести их к единой единице измерения
   - для этого мы используем библиотеку sklearn
2) Вычислить евклидово расстояние (расстояние между каждым набором данных)
   - для этого используем библиотеку Pandas
3) Определить оптимальное количество кластеров при помощи метода “локтя”
   - для этого используем библиотеки matplotlib и numpy

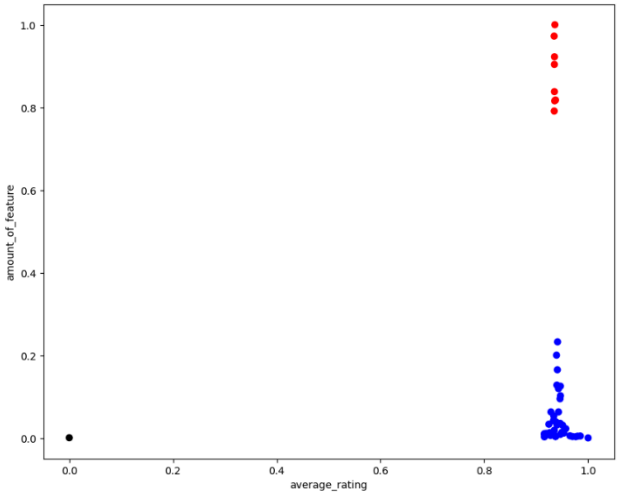



```
Рекомендованное количество кластеров: 4
```

4) Построить дендограмму



5) Построить диаграмму распределения кластеров



6) Интерпретировать результаты

3.9.4. Вывод

Мы получили 4 кластера, которые содержат в себе наборы дополнительных услуг, содержащиеся в кафе с хорошим рейтингом.

KMeans	average_rating	amount_of_feature	amount_of_features_in_klaster
1	4.358876	2049.625000	8
2	4.360673	957.833333	6
3	4.387619	89.062500	48
4	0.000000	0.000000	0

В первом кластере нами были получены номера (1-8) дополнительных услуг в наборе , количество заведений, среднее количество заведений, содержащих эти дополнительные услуги и средним рейтингом заведений.

	number_of_feature	average_rating	amount_of_feature	KMeans
0	1	4.354821	2258	1
1	2	4.359597	1894	1
2	3	4.367357	1899	1
3	4	4.355655	1837	1
4	5	4.357912	2142	1
5	6	4.357373	1946	1
6	7	4.361262	2322	1
7	8	4.357033	2099	1

Во втором кластере были такие дополнительные услуги как 9, 10, 11, 12, 13, 18

	number_of_feature	average_rating	amount_of_feature	KMeans
8	9	4.374491	1023	2
9	10	4.356770	915	2
10	11	4.360961	1006	2
11	12	4.340062	916	2
12	13	4.371965	876	2
17	18	4.359788	1011	2

В третьем кластере количество дополнительных услуг = 48, что говорит о том, что между услугами третьего кластера нет корреляции ни в рейтинге, ни в количестве заведений, содержащих данные услуги, эти услуги не вошли ни в одну из первых двух групп, поэтому их количество большое.

	number_of_feature	average_rating	amount_of_feature	KMeans
13	14	4.376640	465	3
14	15	4.384061	383	3
15	16	4.385872	540	3
16	17	4.394268	277	3
18	19	4.411545	291	3
19	20	4.379644	297	3
20	21	4.326439	146	3
21	22	4.407855	220	3
22	23	4.411804	237	3
23	24	4.394154	146	3
24	25	4.356010	43	3
25	26	4.307879	77	3
26	27	4.351907	123	3
27	28	4.371954	90	3
28	29	4.315560	30	3
29	30	4.367496	9	3
30	31	4.367496	9	3

В четвертом кластере находится только 1 дополнительная услуга, которая не содержится ни в одном из кафе и является выбросом.

	number_of_feature	average_rating	amount_of_feature	KMeans
59	60	0.0	0	4

Исходя из кластерного анализа можно сделать вывод о том, что для того, чтобы кафе имело высокий рейтинг необходимо внедрить такие дополнительные услуги как:

1. food_delivery - доставка еды
2. breakfast - завтрак
3. takeaway - еда на вынос
4. summer_terrace - летняя терраса
5. wi_fi - вай-фай доступ
6. business_lunch - комплексный обед
7. payment_by_credit_card - возможность оплаты картой
8. coffee_to_go - кофе на вынос

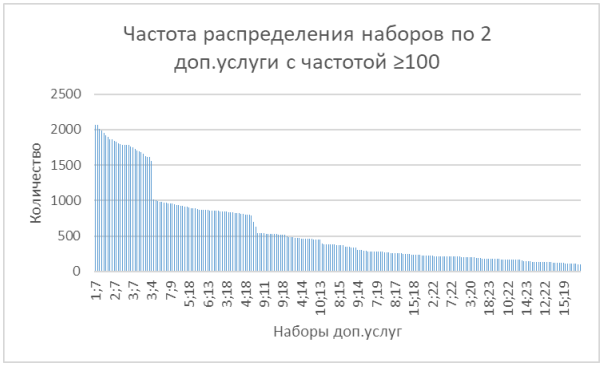
Для того чтобы улучшить рейтинг, предпринимателю необходимо добавлять следующие дополнительные услуги:

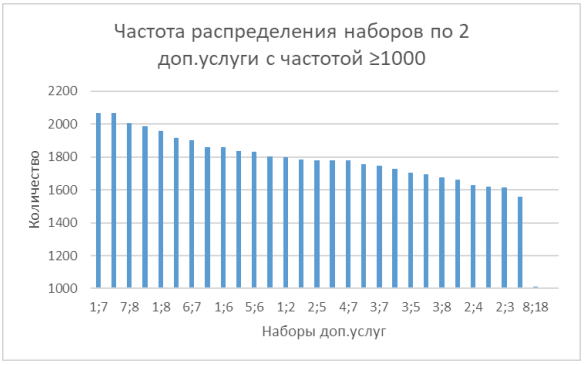
9. closed_for_quarantine - меры безопасности в случае карантина
10. online_takeaway - возможность заказа на вынос онлайн
11. karaoke - караоке
12. special_menu - специальное меню
13. sports_broadcasts - спортивные передачи
18. wheelchair_access - доступ для людей с ограниченными возможностями

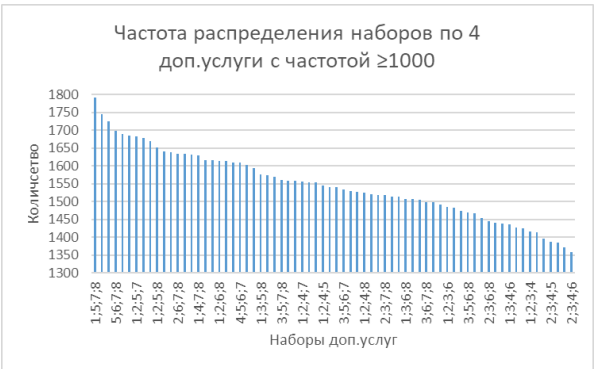
3.10. Анализ другим методом

3.10.1. Описание и реализация исследования на Python

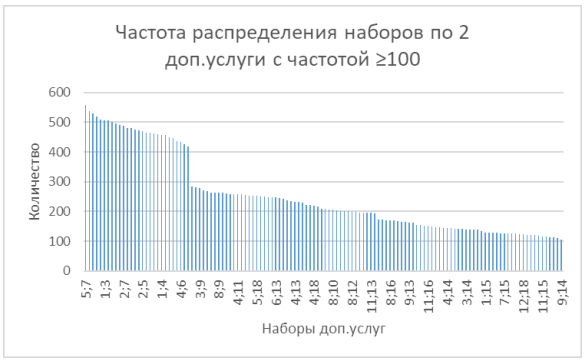
Мы решили рассмотреть какие наборы дополнительных услуг встречаются чаще всего, для этого мы пользовались Python и библиотекой combinations. Для этого мы преобразовали наши данные в другой формат, каждой доп.услуге присвоили номер (начиная с первой и далее по возрастанию). После этого загрузили базу данных в программу и смотрели, начиная с 1 доп. услуги в наборе, сколько раз тот или иной набор встречается. Выводы представлены на диаграммах ниже. На левых графиках изображены те наборы, которые встречаются больше 100 раз, а справа те наборы, которые встречаются 1000 и более раз. Мы это сделали для того, чтобы лучше посмотреть, какие наборы доп.услуг встречаются наибольшее количество раз.

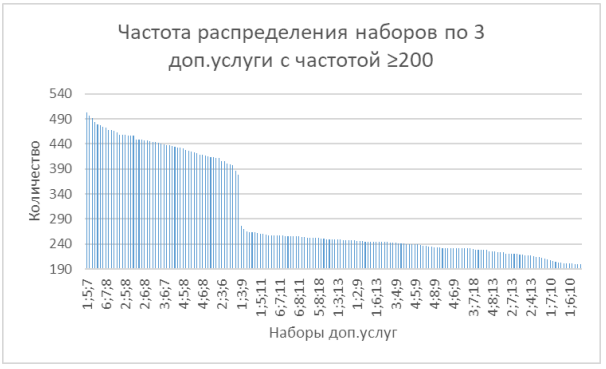


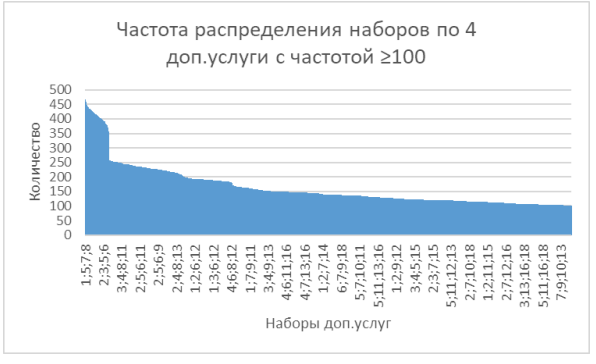


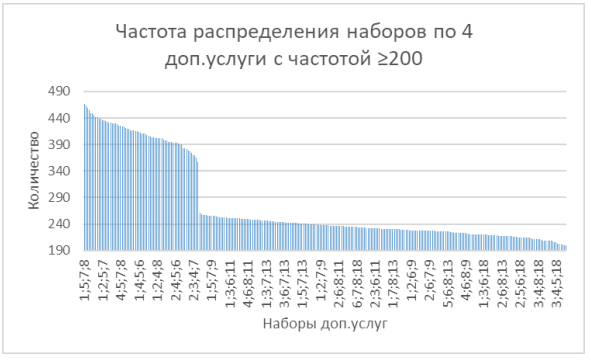


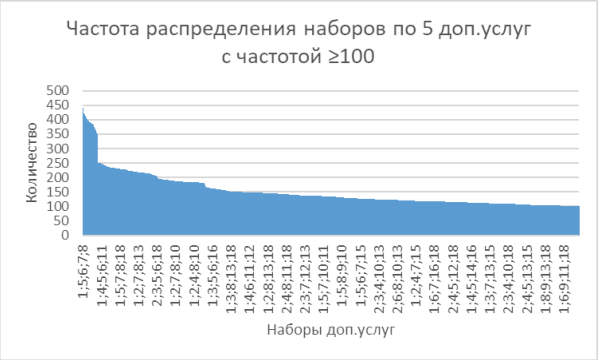
Далее мы решили рассмотреть кафе с отличным рейтингом (4.5 и выше), для этого мы убрали все лишние значения, преобразовали базу данных и выполнили все те же действия, которые описаны выше. На левых графиках изображены те наборы, которые встречаются больше 100 раз, а справа те наборы, которые встречаются 200 и более раз. Мы это сделали для того, чтобы лучше посмотреть, какие наборы доп.услуг встречаются наибольшее количество раз.

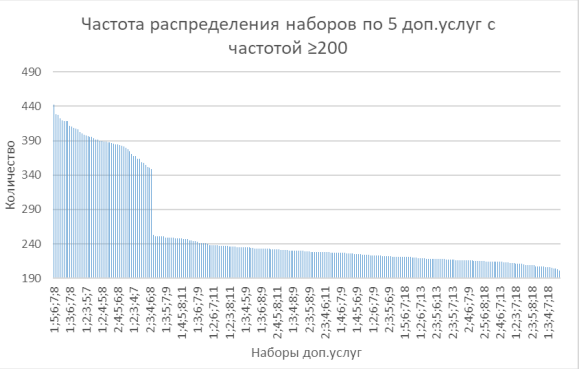












3.10.2. Вывод

Нам удалось подсчитать максимально только для 5 доп.услуг в наборе, но даже до 5 мы можем увидеть, что чаще всего встречаются доп. услуги 1,2,3,4,5,6,7,8: food_delivery, breakfast, takeaway, summer_terrace, wi_fi, business_lunch, payment_by_credit_card, coffee_to_go. Далее, мы увеличивали кол-во доп.услуг в наборе. Из этого можно сделать вывод, что перечисленные выше доп.услуги встречаются в хорошем рейтинге. Далее мы сократили базу данных и посмотрели, какие наборы доп.услуг встречаются в кафе с отличным рейтингом (т.е. с рейтингом 4.5 и выше). Получили что в кафе с отличным рейтингом встречаются те же доп.услуги, что и в хорошем и отличном вместе взятом рейтингах. Но при этом в отличном рейтинге также встречаются такие доп.услуги, как 9, 10, 11, 12, 13, 18: closed_for_quarantine, online_takeaway, karaoke, special_menu, sports_broadcasts, wheelchair_access.

3.11. Результаты

Мы провели анализ двумя разными способами и в обоих случаях нам выдал одни и те же результаты. Мы вывели список дополнительных услуг, которые встречаются в заведениях с хорошим рейтингом и предпринимателям, которые хотят открыть свое кафе или уже имеют, необходимо добавить их в свои заведения. Также кластерным анализом мы вывели еще одну группу дополнительных услуг, которые встречаются в заведениях с хорошим рейтингом, но не так часто. Проверив это другим методом анализа, мы пришли к выводу, что данный набор встречается в заведениях с отличным рейтингом(выше 4.5). Значит, эти дополнительные услуги(табл.) следует внедрять в заведения, чтобы иметь отличный рейтинг.

3.12. Ограничения и перспективы

В ходе проделанной работы нами были выявлены недостатки нашего анализа и разработаны методы его улучшения, а также планируемые действия для дальнейшей

работы с проектом. Так так нам было предоставлено ограниченные данные, содержащие не все заведений Москвы, то анализ нельзя считать в полной мере объективной оценкой. В некоторых ячейках с рейтингом кафе были представлены пропуски, что повлекло за собой необходимость их самостоятельного заполнения, которое также может отличаться от действительного. Еще одним важным ограничением является факт того, что данные были представлены за прошедший период, на них можно опираться говоря об общей ситуации в сфере бизнеса кафе, но нельзя применять к конкретному году. В работе, с целью читабельности данных были проведены сокращения, оценки, что может повлечь за собой неточности в расчетах и выводимых результатах.

В перспективах планируется улучшать проект, дорабатывая слабые стороны и используя новые алгоритмы, которые будут изучены позже, например будет проводится регрессионных анализ. Также планируется создать на первых стадиях лендинг-пейдж, отражающую результаты нашего анализа, далее эта доработка будет переделана в сайт, где пользователь сможет вводить данные или выбирать предложенные, и по полученным данным пользователю будет предложен его персональный набор для того чтобы улучшить свое кафе, а для пользователей, планирующих впервые открыть свой бизнес будет представлен набор дополнительных услуг, составленный с учетом деталей, таких как расположение, город, размер планируемого заведения.

4. Использованная литература

1) https://pntr.io/27facts - сайт с исследованиями о значимости отзывов
2) https://thebell.io/amp/v-rossii-3-predprinimatelej-eto-namnogo-menshe-chem-v-mire- статистика предпринимателей в России
3) https://www.rbc.ru/spb_sz/25/11/2020/5fbe22cb9a794707aa1d09c1 - информация о упадочном бизнесе ресторанов и причинах превышения закрытий над открытиями
4) https://iom.anketolog.ru/2021/02/15/kafe-restorany - на что обращат внимание посетители при выборе заведения приема пищи
5) https://restteam.ru/consulting/audit-restorana?ysclid=ld0qibp98c561647791 - один из игроков на рынке
6) https://info.mozg.rest - один из игроков на рынке
7) https://trade-drive.ru/automatization/rest/?ysclid=ldf39ja5jk85641997 - один из игроков на рынке
8) https://yandex.ru/support/reviews/review.html

9) https://yandex.ru/support/business-priority/manage/rating.html - как формируются отзывы

10) https://sovcombank.ru/blog/biznesu/kak-rabotayut-otzivi-trendi-i-rekomendatsii/amp?utm_referrer=https%3A%2F%2Fwww.google.com%2F - анализ понятия “хороший рейтинг”

11) https://business.aliexpress.ru/help/article/otzivi-pokupatelei - анализ понятия “хороший рейтинг”

12) https://en.wikipedia.org/wiki/Q–Q_plot - нормальнольность

13) https://habr.com/ru/post/578754/ - нормальность

14) http://statistica.ru/theory/klasterizatsiya-metod-k-srednikh/ - кластерный анализ