GeekBrains

Анализ особенностей веганских и вегетарианских ресторанов США с использованием методов машинного обучения и геопространственного анализа

Программа: Цифровые профессии

Специализация: Искусственный интеллект

Студент: Чёрная Наталья Александровна

Салават
2024

Введение

В современном мире наблюдается растущий интерес к здоровому образу жизни и экологически ответственному потреблению. Эта тенденция находит отражение в сфере общественного питания, где все большую популярность приобретают веганские и вегетарианские рестораны. Анализ особенностей таких заведений представляет собой актуальную задачу как для исследователей рынка, так и для предпринимателей, работающих в данной сфере.

Актуальность темы

Актуальность данного исследования обусловлена следующими факторами:

1. Рост популярности веганского и вегетарианского образа жизни, что приводит к увеличению спроса на соответствующие заведения общественного питания.
2. Необходимость глубокого понимания особенностей рынка веганских и вегетарианских ресторанов для принятия обоснованных бизнес-решений.
3. Возрастающая роль анализа данных и машинного обучения в оптимизации бизнес-процессов в сфере общественного питания.
4. Важность геопространственного анализа для выбора оптимального расположения ресторанов и понимания конкурентной среды.

Цель и задачи исследования

Цель исследования: провести комплексный анализ особенностей веганских и вегетарианских ресторанов с использованием методов машинного обучения и геопространственного анализа для выявления ключевых факторов, влияющих на их успешность и распространение.

Для достижения поставленной цели необходимо решить следующие **задачи**:

1. Провести обзор рынка веганских и вегетарианских ресторанов и выявить основные тенденции его развития.
2. Осуществить сбор, очистку и предварительную обработку данных о веганских и вегетарианских ресторанах.
3. Выполнить разведочный анализ данных для выявления основных характеристик и закономерностей в исследуемой выборке ресторанов.
4. Провести геопространственный анализ расположения ресторанов и оценить влияние географического фактора на их характеристики.
5. Применить методы машинного обучения для классификации ресторанов и выявления факторов, влияющих на их ценовую категорию.

6. Исследовать скрытые зависимости между различными характеристиками ресторанов с использованием корреляционного анализа и метода главных компонент (PCA).

7. Визуализировать полученные результаты и разработать интерактивные инструменты для их представления.

8. На основе проведённого анализа сформулировать практические рекомендации для владельцев и менеджеров веганских и вегетарианских ресторанов.

Объект и предмет исследования

Объект исследования: веганские и вегетарианские рестораны.

Предмет исследования: особенности расположения, ценообразования и других характеристик веганских и вегетарианских ресторанов, а также факторы, влияющие на их успешность.

Методология исследования

В рамках данной работы применяется комплексный подход к анализу данных, включающий следующие методы:

- Статистический анализ
- Геопространственный анализ
- Методы машинного обучения (в частности, алгоритм случайного леса)
- Корреляционный анализ
- Метод главных компонентов (PCA)
- Визуализация данных

Исследование опирается на данные о более чем 10 000 ресторанах, собранные из открытых источников. Для анализа и обработки данных используется язык программирования Python и его специализированные библиотеки, такие как pandas, numpy, scikit-learn, matplotlib и folium.

Структура работы

Дипломная работа состоит из введения, двух глав (теоретической и практической), заключения, списка использованной литературы и приложений.

В теоретической части рассматриваются основные понятия и методы, применяемые в исследовании, а также проводится обзор рынка веганских и вегетарианских ресторанов.

Практическая часть содержит описание процесса анализа данных, включая предварительную обработку, разведочный анализ, применение методов машинного обучения и геопространственного анализа, а также интерпретацию полученных результатов.

В заключении подводятся итоги исследования, формулируются основные выводы и предлагаются направления для дальнейших исследований в данной области.

2. Теоретическая часть

2.1 Обзор рынка веганских и вегетарианских ресторанов

2.1.1 Тенденции развития веганского и вегетарианского питания

В последние годы наблюдается значительный рост интереса к веганскому и вегетарианскому образу жизни во всем мире. Это обусловлено рядом факторов:

1. **Забота о здоровье:** Многочисленные исследования показывают, что растительная диета может снизить риск развития сердечно-сосудистых заболеваний, диабета 2 типа и некоторых видов рака.

2. **Экологические соображения:** Производство растительной пищи оказывает меньшее негативное влияние на окружающую среду по сравнению с животноводством.

3. **Этические причины:** Все больше людей отказываются от потребления продуктов животного происхождения из-за озабоченности условиями содержания животных в промышленном животноводстве.

4. **Разнообразие и доступность:** Увеличение ассортимента растительных продуктов и их доступности в магазинах и ресторанах делает веганский и вегетарианский образ жизни более привлекательным для широкой аудитории.

По данным различных исследований, количество веганов и вегетарианцев в развитых странах steadily растёт. Например, в США число веганов увеличилось с 1% населения в 2014 году до 6% в 2022 году. В Великобритании количество веганов выросло на 40% с 2020 по 2022 год.

2.1.2 Особенности веганских и вегетарианских ресторанов

Веганские и вегетарианские рестораны имеют ряд отличительных особенностей:

1. **Меню:** Основное отличие заключается в составе меню. Вегетарианские рестораны исключают из своего меню мясо и рыбу, но могут предлагать блюда с использованием молочных продуктов и яиц. Веганские рестораны полностью исключают продукты животного происхождения.

2. **Инновационность:** Шеф-повара веганских и вегетарианских ресторанов часто проявляют высокую креативность в создании блюд, используя необычные ингредиенты и техники приготовления для достижения разнообразия вкусов и текстур.

3. **Ориентация на здоровое питание:** Многие веганские и вегетарианские рестораны уделяют особое внимание использованию органических, цельных и минимально обработанных продуктов.

4. **Экологическая ответственность:** Эти рестораны часто придерживаются принципов устойчивого развития, используя экологичную упаковку, минимизируя отходы и выбирая локальных поставщиков.

5. **Образовательная роль:** Многие веганские и вегетарианские рестораны берут на себя роль просветителей, информируя клиентов о преимуществах растительного питания и экологической ответственности.

2.1.3 Факторы, влияющие на успешность ресторанов данного типа

1. **Расположение:** Успех веганского или вегетарианского ресторана во многом зависит от его местоположения. Районы с высокой концентрацией молодёжи, творческих профессионалов и людей, ведущих здоровый образ жизни, как правило, более благоприятны для таких заведений.

2. **Качество и разнообразие меню:** Способность предложить вкусные, питательные и разнообразные блюда является ключевым фактором успеха. Важно удовлетворить не только постоянных веганов и вегетарианцев, но и привлечь клиентов, которые лишь экспериментируют с растительным питанием.

3. **Ценовая политика:** Правильное ценообразование играет важную роль. Несмотря на то, что некоторые ингредиенты могут быть дороже, важно сохранять конкурентоспособные цены.

4. **Атмосфера и дизайн:** Уютная атмосфера и привлекательный дизайн могут значительно повысить популярность ресторана.

5. **Маркетинг и социальные медиа:** Эффективное использование социальных сетей и других маркетинговых каналов помогает привлечь клиентов и создать лояльное сообщество вокруг ресторана.

6. **Экологическая ответственность:** Приверженность принципам устойчивого развития может стать важным фактором привлечения клиентов, особенно среди экологически сознательной аудитории.

7. **Профессионализм персонала:** Знающий и дружелюбный персонал, способный ответить на вопросы о составе блюд и особенностях веганского питания, повышает удовлетворённость клиентов.

2.2 Методы анализа данных в ресторанном бизнесе

2.2.1 Обзор традиционных методов анализа

Традиционно в ресторанном бизнесе использовались следующие методы анализа:

1. **Финансовый анализ:** Анализ доходов и расходов, расчёт прибыльности, анализ структуры затрат.

2. **Анализ меню:** Оценка популярности и прибыльности отдельных блюд, ABC-анализ меню.

3. **Анализ клиентской базы:** Сегментация клиентов, анализ частоты посещений и среднего чека.

4. **Анализ конкурентов:** Изучение цен, меню и маркетинговых стратегий конкурентов.

5. **SWOT-анализ:** Оценка сильных и слабых сторон ресторана, а также внешних возможностей и угроз.

2.2.2 Современные подходы к анализу данных в ресторанной индустрии

С развитием технологий и появлением больших объёмов данных, в ресторанном бизнесе стали применяться более продвинутые методы анализа:

1. **Анализ больших данных:** Использование больших объёмов структурированных и неструктурированных данных для получения инсайтов о поведении клиентов и тенденциях рынка.

2. **Предиктивная аналитика:** Использование исторических данных и машинного обучения для прогнозирования будущих трендов, спроса и поведения клиентов.

3. **Сентимент-анализ:** Анализ отзывов клиентов в социальных сетях и на сайтах-агрегаторах для оценки удовлетворённости и выявления проблемных областей.

4. **Геопространственный анализ:** Использование географических данных для оптимизации расположения ресторанов и анализа конкурентной среды.

5. **Анализ в реальном времени:** Мониторинг ключевых показателей эффективности в режиме реального времени для быстрого реагирования на изменения.

2.2.3 Значение анализа данных для принятия бизнес-решений

Анализ данных играет критически важную роль в принятии обоснованных бизнес-решений в ресторанной индустрии:

1. **Оптимизация меню:** Анализ популярности и прибыльности блюд помогает оптимизировать меню, убирая непопулярные позиции и добавляя новые, потенциально успешные блюда.

2. **Ценообразование:** Анализ данных о затратах, ценах конкурентов и готовности клиентов платить позволяет устанавливать оптимальные цены.

3. **Управление запасами:** Прогнозирование спроса помогает оптимизировать закупки и минимизировать отходы.

4. **Маркетинговые стратегии:** Анализ данных о клиентах помогает разрабатывать таргетированные маркетинговые кампании и программы лояльности.

5. **Расширение бизнеса:** Геопространственный анализ и анализ рыночных тенденций помогают принимать решения о расширении бизнеса и выборе локаций для новых ресторанов.

6. **Улучшение качества обслуживания:** Анализ отзывов клиентов помогает выявлять проблемные области и улучшать качество обслуживания.

2.3 Основы машинного обучения и его применение в анализе ресторанов

2.3.1 Понятие машинного обучения

Машинное обучение (ML) - это подраздел искусственного интеллекта, который занимается разработкой алгоритмов и статистических моделей, позволяющих компьютерным системам улучшать свою производительность при выполнении определённой задачи на основе опыта, без явного программирования.

Основные типы машинного обучения:

1. **Обучение с учителем:** Алгоритм обучается на размеченных данных, где для каждого входного примера известен желаемый выход.

2. **Обучение без учителя:** Алгоритм ищет скрытые структуры в неразмеченных данных.

3. **Обучение с подкреплением:** Алгоритм учится через взаимодействие с окружающей средой, получая обратную связь в виде наград или штрафов.

2.3.2 Основные алгоритмы, применяемые в анализе ресторанного бизнеса

1. **Линейная и логистическая регрессия:** Используются для прогнозирования численных значений (например, выручки) или вероятности событий (например, вероятности повторного посещения клиента).

2. **Деревья решений и случайный лес:** Применяются для классификации (например, категоризации ресторанов по ценовым сегментам) и регрессии.

3. **Кластерный анализ (например, K-means):** Используется для сегментации клиентов или группировки ресторанов по схожим характеристикам.

4. **Нейронные сети:** Применяются для решения сложных задач, таких как прогнозирование спроса или анализ изображений блюд.

5. **Ассоциативные правила:** Используются для анализа связей между различными элементами (например, для анализа сочетаемости блюд в заказах).

2.3.3 Преимущества использования машинного обучения в данной сфере

1. **Точность прогнозов:** ML-модели могут учитывать множество факторов и находить сложные зависимости, что повышает точность прогнозов по сравнению с традиционными методами.

2. **Автоматизация:** Машинное обучение позволяет автоматизировать многие процессы анализа, сокращая время и усилия, необходимые для принятия решений.

3. **Персонализация:** ML-алгоритмы позволяют создавать персонализированные рекомендации для клиентов, повышая их удовлетворённость и лояльность.

4. **Обработка больших объёмов данных:** Машинное обучение эффективно в работе с большими и сложными наборами данных, которые сложно анализировать традиционными методами.

5. **Адаптивность:** ML-модели могут постоянно обучаться на новых данных, адаптируясь к изменяющимся условиям рынка.

2.4 Геопространственный анализ и его значение для ресторанного бизнеса

2.4.1 Основы геопространственного анализа

Геопространственный анализ — это подход к анализу данных, который учитывает пространственное расположение объектов и их взаимосвязи. В контексте ресторанного бизнеса это означает анализ данных с учётом географического расположения ресторанов, клиентов и других релевантных объектов.

Ключевые концепции геопространственного анализа включают:

1. **Географические координаты:** Широта и долгота, используемые для точного определения местоположения.

2. **Пространственные отношения:** Расстояния между объектами, области влияния, пересечения и т.д.

3. **Пространственная автокорреляция:** Степень, в которой объект или значение удаления связано с другими рядом расположенными объектами или значениями атрибутов.

4. **Картографическое моделирование:** Создание и анализ карт для визуализации пространственных данных.

2.4.2 Методы визуализации геоданных

1. **Точечные карты:** Отображают расположение отдельных ресторанов на карте. Полезны для визуализации общего распределения заведений.

2. **Тепловые карты:** Показывают плотность расположения ресторанов или интенсивность определённых показателей (например, выручки) с помощью цветового градиента.

3. **Хороплеты:** Карты, на которых области (например, районы города) окрашены в разные цвета в зависимости от значения определённого показателя (например, количества веганских ресторанов на квадратный километр).

4. **Карты изолиний:** Отображают линии равных значений (например, время доставки еды).

5. **Картограммы:** Комбинируют географические данные с другими типами визуализации (например, диаграммами), показывая дополнительную информацию для каждой географической единицы.

6. **Интерактивные карты:** Позволяют пользователям взаимодействовать с данными, например, приближать и отдалять отдельные участки, выбирать слои для отображения.

2.4.3 Применение геопространственного анализа в ресторанном бизнесе

1. **Выбор местоположения:** Анализ демографических данных, транспортных потоков, расположения конкурентов и других факторов для выбора оптимального места для нового ресторана.

2. **Анализ конкурентной среды:** Изучение расположения конкурентов, их концентрации и влияния на бизнес.

3. **Оптимизация доставки:** Планирование маршрутов доставки, определение зон доставки и оценка времени доставки.

4. **Маркетинговые кампании:** Таргетирование рекламы на основе геоданных, планирование локальных маркетинговых мероприятий.

5. **Анализ клиентской базы:** Изучение географического распределения клиентов, выявление потенциальных новых рынков.

6. **Управление сетью ресторанов:** Оптимизация расположения ресторанов в сети, анализ эффективности отдельных локаций.

7. **Анализ трендов:** Выявление географических паттернов в предпочтениях клиентов, популярности определённых типов кухни или форматов ресторанов.

2.4.4 Инструменты для геопространственного анализа

1. **ГИС-системы:** Профессиональные инструменты как ArcGIS или QGIS для сложного пространственного анализа.

2. **Библиотеки Python:** GeoPandas, Folium, Shapely для работы с геоданными и их визуализации.

3. **Картографические сервисы:** Google Maps API, Mapbox для создания интерактивных карт и геокодирования.

4. **Специализированные инструменты:** Maptitude, Esri Business Analyst для бизнес-аналитики с учётом геоданных.

5. **Инструменты визуализации:** Tableau, Power BI с возможностями создания геопространственных визуализаций.

2.4.5 Проблемы и ограничения геопространственного анализа

1. **Качество данных:** Точность и актуальность геоданных критически важны для корректного анализа.
2. **Конфиденциальность:** Необходимость соблюдения законов о защите персональных данных при работе с геоданными клиентов.
3. **Сложность интерпретации:** Геопространственные данные могут быть сложны для интерпретации неспециалистами.
4. **Вычислительные требования:** Обработка больших объёмов геоданных может требовать значительных вычислительных ресурсов.
5. **Динамичность данных:** Геопространственные данные могут быстро устаревать, особенно в динамичной городской среде.

2.5 Методы статистического анализа в исследовании ресторанного бизнеса

2.5.1 Описательная статистика

Описательная статистика используется для обобщения и описания основных характеристик набора данных. В контексте анализа веганских и вегетарианских ресторанов она может включать:

1. **Меры центральной тенденции:** Среднее, медиана и мода для таких показателей, как цены, рейтинги, количество посетителей.
2. **Меры разброса:** Стандартное отклонение, дисперсия, диапазон для оценки вариативности данных.
3. **Распределение:** Гистограммы и графики плотности для визуализации распределения различных показателей.
4. **Процентили и квартили:** Для понимания распределения данных и выявления выбросов.

2.5.2 Корреляционный анализ

Корреляционный анализ используется для изучения взаимосвязей между различными переменными. В анализе ресторанов это может включать:

1. **Коэффициент корреляции Пирсона:** Для изучения линейных связей между непрерывными переменными (например, между ценой и рейтингом).
2. **Ранговая корреляция Спирмена:** Для оценки монотонных связей, особенно полезна для ординальных данных.

3. **Корреляционные матрицы:** Для визуализации корреляций между множеством переменных одновременно.

2.5.3 Регрессионный анализ

Регрессионный анализ используется для моделирования отношений между зависимой переменной и одной или несколькими независимыми переменными. В контексте ресторанного бизнеса это может включать:

1. **Линейная регрессия:** Для прогнозирования непрерывных переменных (например, выручки на основе различных факторов).

2. **Логистическая регрессия:** Для прогнозирования бинарных исходов (например, вероятности того, что ресторан останется открытым через год).

3. **Множественная регрессия:** Для анализа влияния нескольких факторов на зависимую переменную.

2.5.4 Анализ временных рядов

Анализ временных рядов используется для изучения данных, собранных с течением времени. В ресторанном бизнесе это может включать:

1. **Тренд-анализ:** Выявление долгосрочных тенденций в продажах, популярности определённых блюд или рейтингах.

2. **Сезонность:** Анализ сезонных паттернов в посещаемости или продажах.

3. **Прогнозирование:** Использование исторических данных для прогнозирования будущих показателей.

2.5.5 Кластерный анализ

Кластерный анализ используется для группировки схожих объектов. В анализе веганских и вегетарианских ресторанов это может включать:

1. **K-средних:** Для группировки ресторанов по схожим характеристикам (например, по ценовой категории и рейтингу).

2. **Иерархическая кластеризация:** Для создания древовидной структуры кластеров ресторанов.

3. **DBSCAN:** Для выявления кластеров произвольной формы, особенно полезен при анализе географического распределения ресторанов.

2.5.6 Анализ главных компонент (PCA)

PCA используется для уменьшения размерности данных и выявления основных направлений вариации. В контексте анализа ресторанов это может быть полезно для:

1. **Выявления ключевых факторов:** Определение основных компонент, влияющих на успех ресторана.

2. **Визуализации:** Представление многомерных данных в двух- или трёхмерном пространстве для лучшего понимания структуры данных.

3. **Подготовка данных:** Уменьшение размерности данных перед применением других методов машинного обучения.

2.6 Этические аспекты и проблемы конфиденциальности в анализе данных ресторанного бизнеса

2.6.1 Защита персональных данных клиентов

1. **Соблюдение законодательства:** Необходимость соответствия законам о защите персональных данных (например, GDPR в Европе, CCPA в Калифорнии).

2. **Анонимизация данных:** Удаление или маскировка личной информации при проведении анализа.

3. **Согласие на сбор и использование данных:** Получение явного согласия клиентов на сбор и использование их данных.

2.6.2 Прозрачность в использовании данных

1. **Информирование клиентов:** Чёткое объяснение того, как собираются и используются данные.

2. **Право на удаление:** Предоставление клиентам возможности удалить свои данные из системы.

3. **Ограничение использования:** Использование данных только для заявленных целей.

2.6.3 Этические вопросы в применении аналитики и машинного обучения

1. **Справедливость алгоритмов:** Обеспечение отсутствия дискриминации в алгоритмах (например, при определении цен или таргетировании рекламы).

2. **Интерпретируемость моделей:** Возможность объяснить, как и почему модель приняла определённое решение.

3. **Ответственное использование предиктивной аналитики:** Баланс между персонализацией и неприкосновенностью частной жизни.

2.6.4 Безопасность данных

1. **Защита от утечек:** Внедрение надёжных систем безопасности для защиты данных клиентов и бизнеса.

2. **Ограничение доступа:** Предоставление доступа к данным только сотрудникам, которым это необходимо для работы.

3. **Шифрование:** Использование современных методов шифрования для защиты данных при хранении и передаче.

2.6.5 Этика в геопространственном анализе

1. **Уважение частной жизни:** Избегание слишком детального отслеживания местоположения клиентов.

2. **Агрегация данных:** Использование агрегированных геоданных вместо индивидуальных для защиты приватности.

3. **Прозрачность в использовании геоданных:** Информирование клиентов о том, как используется информация об их местоположении.

Эти этические аспекты и проблемы конфиденциальности играют важную роль в современном анализе данных в ресторанном бизнесе. Соблюдение этических норм и защита данных клиентов не только необходимы с юридической точки зрения, но и способствуют построению доверительных отношений с клиентами, что критически важно для долгосрочного успеха в индустрии гостеприимства.

3. Практическая часть

3.1 Описание исходных данных и их предварительная обработка

3.1.1 Источник и структура данных

Данные для анализа были получены из открытого источника: набор данных "Vegetarian and Vegan Restaurants" с платформы Kaggle (https://www.kaggle.com/datasets/datafiniti/vegetarian-vegan-restaurants/data). Этот набор данных содержит информацию о более чем 10 000 вегетарианских и веганских ресторанах в США.

Исходный набор данных включает следующие ключевые поля:

- id: уникальный идентификатор ресторана
- dateAdded: дата добавления записи нового ресторана
- dateUpdated: дата обновления записи
- address: адрес ресторана
- categories: категории ресторана
- city: город
- country: страна
- latitude: широта
- longitude: долгота

- name: название ресторана
- postalCode: почтовый индекс
- province: штат
- priceRangeMin: минимальная цена
- priceRangeMax: максимальная цена
- menus.name: названия блюд в меню
- menus.amountMax: максимальная цена блюда
- menus.amountMin: минимальная цена блюда

3.1.2 Создание файла Jupiter Notebook и установка необходимых python-пакетов

Запускаем VS Code и создаем новую папку для нашего проекта с названием Diplom. Нажаем на значок Explorer (в боковой панели) и выбираем New File. Называем файл с расширением .ipynb, например, Diplom_Chernaya_NA.ipynb. После этого автоматически откроется интерфейс для работы с Jupyter Notebook в VS Code.

Так как мы создаем новый проект, то создаем виртуальное окружение. Открываем терминал прямо в VS Code и выполняем команду:

```
python -m venv venv
```

В ячейке ноутбука запустим следующий код, чтобы установить необходимые python-пакеты.

```
# !pip install pandas numpy matplotlib seaborn scikit-learn folium ipython
```

3.1.3 Импорт библиотек

```
# Работа с таблицами данных
import pandas as pd
# Работа с массивами чисел и линейной алгеброй
import numpy as np
# Визуализация данных. Создание графиков, диаграмм и других видов визуализаций
import matplotlib.pyplot as plt
# Расширение для matplotlib, которое упрощает создание сложных графиков с помощью высокоуровневых интерфейсов и стилевого оформления.
import seaborn as sns
# Алгоритм кластеризации, который разделяет данные на кластеры на основе схожести между объектами
from sklearn.cluster import KMeans
# Преобразование данных к стандартному виду
from sklearn.preprocessing import StandardScaler
```

```
# Метод главных компонентов (PCA) для уменьшения размерности данных
from sklearn.decomposition import PCA
# Разделение данных на обучающую и тестовую выборки
from sklearn.model_selection import train_test_split
# Классификатор на основе ансамблевых методов, таких как случайный лес
from sklearn.ensemble import RandomForestClassifier
# Оценка качества модели
from sklearn.metrics import classification_report

# Стратифицированная кросс-валидация для оценки модели
from sklearn.model_selection import StratifiedKFold
# Вычисление косинусной близости между векторами
from sklearn.metrics.pairwise import cosine_similarity

# Используем стратифицированную кросс-валидацию
from sklearn.model_selection import cross_val_score
# Вычисление коэффициента Каппа для оценки качества классификации
from sklearn.metrics import cohen_kappa_score
# Confusion Matrix для визуальной оценки ошибок
from sklearn.metrics import confusion_matrix
# Библиотека для создания интерактивных карт
import folium
# Плагин для отображения маркеров на карте
from folium.plugins import MarkerCluster
# Библиотека для открытия веб-страниц в браузере
import webbrowser
# Библиотека для работы с операционной системой
import os

# Метрика Cohen's Kappa для учета случайного угадывания
from sklearn.metrics import cohen_kappa_score
# Confusion Matrix для визуальной оценки ошибок
from sklearn.metrics import confusion_matrix
import seaborn as sns


# Настройка формата вывода чисел float
pd.set_option('display.float_format', '{:.2f}'.format)

# Отображение изображений в ноутбуке
from IPython.display import Image
```

3.2 Очистка и предобработка данных

Процесс очистки и предобработки данных включал следующие шаги:

3.2.1 Загрузка данных:

Данные были загружены с использованием библиотеки pandas:

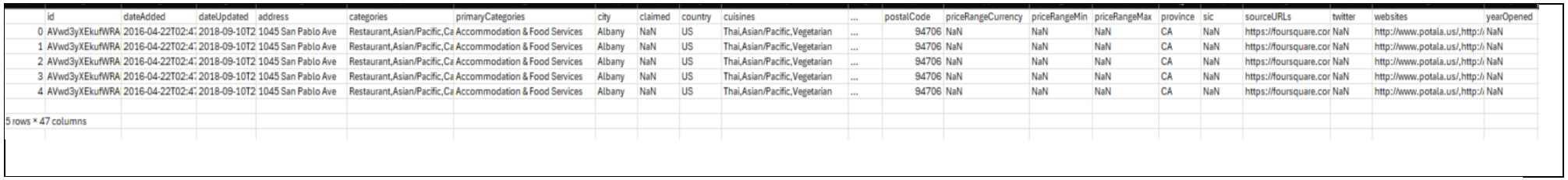
```
# Загрузка данных
```

```
df = pd.read_csv('dataset/Datafiniti_Vegetarian_and_Vegan_Restaurants.csv')
```

Посмотрим данные:

```
# Обзор данных
df.head()
```

Результат:



	id	dateAdded	dateUpdated	address	categories	primaryCategories	city	claimed	country	cuisines	...	postalCode	priceRangeCurrency	priceRangeMin	priceRangeMax	province	sic	sourceURLs	twitter	websites	yearOpened
0	AVwd3yXEkufWRA	2016-04-22T02:47	2018-09-10T2	1045 San Pablo Ave	Restaurant,Asian/Pacific,Ca	Accommodation & Food Services	Albany	NaN	US	Thai,Asian/Pacific,Vegetarian	...	94706	NaN	NaN	NaN	CA	NaN	https://foursquare.cor	NaN	http://www.potala.us/,http://	NaN
1	AVwd3yXEkufWRA	2016-04-22T02:47	2018-09-10T2	1045 San Pablo Ave	Restaurant,Asian/Pacific,Ca	Accommodation & Food Services	Albany	NaN	US	Thai,Asian/Pacific,Vegetarian	...	94706	NaN	NaN	NaN	CA	NaN	https://foursquare.cor	NaN	http://www.potala.us/,http://	NaN
2	AVwd3yXEkufWRA	2016-04-22T02:47	2018-09-10T2	1045 San Pablo Ave	Restaurant,Asian/Pacific,Ca	Accommodation & Food Services	Albany	NaN	US	Thai,Asian/Pacific,Vegetarian	...	94706	NaN	NaN	NaN	CA	NaN	https://foursquare.cor	NaN	http://www.potala.us/,http://	NaN
3	AVwd3yXEkufWRA	2016-04-22T02:47	2018-09-10T2	1045 San Pablo Ave	Restaurant,Asian/Pacific,Ca	Accommodation & Food Services	Albany	NaN	US	Thai,Asian/Pacific,Vegetarian	...	94706	NaN	NaN	NaN	CA	NaN	https://foursquare.cor	NaN	http://www.potala.us/,http://	NaN
4	AVwd3yXEkufWRA	2016-04-22T02:47	2018-09-10T2	1045 San Pablo Ave	Restaurant,Asian/Pacific,Ca	Accommodation & Food Services	Albany	NaN	US	Thai,Asian/Pacific,Vegetarian	...	94706	NaN	NaN	NaN	CA	NaN	https://foursquare.cor	NaN	http://www.potala.us/,http://	NaN

5 rows × 47 columns

Рисунок 1. Обзор данных

Также посмотрим размерность

```
# Обзор размера
df.shape
```

Результат

```
(10000, 47)
```

3.2.2 Проверка наличия пропущенных значений: Была проведена проверка на наличие пропущенных значений в каждом столбце:

```
print(df.isnull().sum(axis=0))
```

Результат

```
id                          0
dateAdded                   0
dateUpdated                 0
address                     0
categories                  0
primaryCategories           0
city                        0
claimed                  9311
country                     0
cuisines                    0
descriptions.dateSeen   10000
descriptions.sourceURLs 10000
descriptions.value      10000
facebookPageURL          9063
features.key            10000
features.value          10000
hours.day               10000
hours.dept              10000
hours.hour              10000
imageURLs                4866
isClosed                 9963
keys                        0
languagesSpoken         10000
latitude                    0
longitude                   0
...
twitter                  8042
websites                 1817
yearOpened               9909
dtype: int64
```

Выявлено, что некоторые столбцы содержат значительное количество пропусков.

3.2.2 Обзор базовой статистики

```
df.describe()
```

Результат

	descriptions.dateSeen	descriptions.sourceURLs	descriptions.value	features.key	features.value	hours.day	hours.dept	hours.hour	languagesSpoken	latitude	longitude	menus.amountMax	menus.amountMin	priceRangeMin	priceRangeMax	sic	yearOpened
count	0	0	0	0	0	0	0	0	0	10000	10000	10000	10000	6327	6327	140	91
mean	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	38.9	-85.14	12.64	12.55	21.41	39.21	5874.2	2015
std	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.72	17.51	39.54	39.53	14.42	10.54	960.91	0
min	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	20.73	-156.45	0	0	0	12	4773	2015
25%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	37.8	-87.98	4	4	0	25	5610	2015
50%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	40.73	-74	7	7	25	40	5610	2015
75%	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	40.76	-73.98	10.95	10.95	25	40	5610	2015
max	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	47.68	-69.72	2500	2500	40	55	8129	2015

Рисунок 2. Обзор базовой статистики

3.2.4 Уникальные значения в категориальных столбцах

```
for col in df.select_dtypes(include=['object']).columns:
    print(f"\nУникальные значения в столбце {col}:")
    print(df[col].value_counts().head())
```

Результат

```
Уникальные значения в столбце id:
AV0BJkuP-gnIPe8DUvYQ    375
AVweaMjhByjofQCxx7-_    310
AVwc8zhtByjofQCxj89O    308
AVzA19T-FcQ3k02bBX1q    265
AVzBGDbxFcQ3k02bBeXU    217
Name: id, dtype: int64

Уникальные значения в столбце dateAdded:
2017-10-18T16:27:40Z    560
2017-07-02T02:34:33Z    375
2016-05-07T01:38:40Z    310
2017-10-18T16:27:37Z    309
2016-03-22T03:41:04Z    308
Name: dateAdded, dtype: int64

Уникальные значения в столбце dateUpdated:
2018-07-19T21:03:58Z    375
2018-08-26T17:07:43Z    310
2018-04-14T09:50:33Z    308
2018-07-19T21:05:13Z    265
2018-07-19T20:48:18Z    219
Name: dateUpdated, dtype: int64
...
http://www.imperialpdx.com,http://www.imperialpdx.com/      308
http://vegetariandimsumnyc.com                              265
http://www.sixpennkitchen.com                               217
Name: websites, dtype: int64
```

Выведем отдельно уникальные значения в столбце name. Это помогает увидеть, какие рестораны являются сетевыми (имеют много повторений) и сколько всего уникальных брендов ресторанов представлено в данных.

```
# Подсчет количества повторений каждого названия ресторана
restaurant_counts = df['name'].value_counts()

# Вывод результатов
print("Название ресторана и их количество (филиалы)")
print(restaurant_counts)

# Количество уникальных названий ресторанов
unique_restaurants = restaurant_counts.shape[0]
print(f"\nВсего уникальных названий ресторанов: {unique_restaurants}")
```

Результат

```
Название ресторана и их количество (филиалы)
Hakkasan                          375
Liquiteria                        310
Imperial                          308
Vegetarian Dim Sum House          265
Six Penn Kitchen                  217
                                 ...
Stinky's Fish Camp                  3
Vegetarian Restaurant by Hakin      2
Nice China Town                     1
Tuscan Grill                        1
David Magen Pizza                   1
Name: name, Length: 209, dtype: int64

Всего уникальных названий ресторанов: 209
```

3.2.5 Посмотрим корреляцию между числовыми столбцами

```python
# Выбираем только числовые столбцы
numeric_columns = df.select_dtypes(include=[np.number]).columns

# Вычисляем корреляцию
correlation_matrix = df[numeric_columns].corr()
print("\nКорреляция между числовыми столбцами:")
correlation_matrix
```

Результат

	descriptions.dateSeen	descriptions.sourceURL	descriptions.value	features.key	features.value	hours.day	hours.dep	hours.hou	languagesSpoken	latitude	longitud	menus.amount	menus.amountMin	priceRangeMin	priceRangeMax	sic	yearOpened
descriptions.dateSeen	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
descriptions.sourceURLs	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
descriptions.value	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
features.key	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
features.value	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hours.day	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hours.dept	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hours.hour	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
languagesSpoken	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
latitude	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1	0.2	0.01	0.01	-0.21	-0.2	0.3	NaN
longitude	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.2	1	0.05	0.05	-0.11	-0.08	0.35	NaN
menus.amountMax	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.01	0.05	1	1	0.19	0.21	-0.01	NaN
menus.amountMin	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.01	0.05	1	1	0.18	0.21	-0.01	NaN

Рисунок 3. Корреляция между числовыми столбцами

3.2.6 Проверка дубликатов

```
print(f"\nКоличество дубликатов: {df.duplicated().sum()}")
```

Результат

```
Количество дубликатов: 0
```

3.2.7 Распределение категориальных данных

```
for col in ['city', 'country', 'primaryCategories']:
    print(f"\nРаспределение {col}:")
    print(df[col].value_counts(normalize=True).head() * 100)
```

Результат

```
Распределение city:
New York    33.95
Brooklyn    11.88
Portland     3.16
Chicago      2.65
Charlotte    2.32
Name: city, dtype: float64

Распределение country:
US   100.00
Name: country, dtype: float64

Распределение primaryCategories:
Accommodation & Food Services                                    92.48
Wholesale Trade,Accommodation & Food Services,Manufacturing       3.10
Wholesale Trade,Accommodation & Food Services                     1.88
Retail                                                            1.69
Arts Entertainment & Recreation                                   0.65
Name: primaryCategories, dtype: float64
```

3.2.8 Анализ ценового диапазона

```
print("\nСтатистика по ценовому диапазону:")
df[['priceRangeMin', 'priceRangeMax']].describe()
```

Результат

	priceRangeMin	priceRangeMax
count	6327	6327
mean	21.41	39.21
std	14.42	10.54
min	0	12
25%	0	25
50%	25	40
75%	25	40
max	40	55

Рисунок 4. Статистика по ценовому диапазону

3.2.9 Анализ географического распределения

```
print("\nТоп-10 городов по количеству ресторанов:")
print(df['city'].value_counts().head(10))
```

Результат

```
Топ-10 городов по количеству ресторанов:
New York            3395
Brooklyn            1188
Portland             316
Chicago              265
Charlotte            232
Pittsburgh           232
Houston              230
Miami                209
Jackson Heights      186
Seattle              169
Name: city, dtype: int64
```

3.2.10 Наиболее популярные кухни

```
cuisines = df['cuisines'].str.split(',', expand=True).stack().value_counts()
print("\nТоп-10 популярных кухонь:")
print(cuisines.head(10))
```

Результат

```
Топ-10 популярных кухонь:
Vegetarian              8089
Vegan                   1668
Indian                  1538
Vegetarian Friendly     1268
Vegan Options           1232
American                1173
Healthy                 1004
```

```
Gluten Free Options     988
Chinese                 893
Local/Organic           750
dtype: int64
```

3.2.11 Временной анализ

```
df['dateAdded'] = pd.to_datetime(df['dateAdded'])
print("\nВременной диапазон данных:")
print(f"Начало: {df['dateAdded'].min()}")
print(f"Конец: {df['dateAdded'].max()}")
print("\nКоличество новых ресторанов по годам:")
print(df['dateAdded'].dt.year.value_counts().sort_index())
```

Результат

```
Временной диапазон данных:
Начало: 2014-01-06 05:32:50+00:00
Конец: 2018-05-12 05:33:10+00:00

Количество новых ресторанов по годам:
2014      73
2015    1206
2016    3097
2017    5239
2018     385
Name: dateAdded, dtype: int64
```

Резкое увеличение количества ресторанов в 2015–2017 годах, что привело к закрытию менее конкурентоспособных заведений. Также в то время многие традиционные рестораны начали внедрять веганские и вегетарианские опции в свои меню, что создало серьезную конкуренцию для специализированных заведений. Это могло привести к снижению их посещаемости и, как следствие, к закрытию некоторых из них. Некоторые известные веганские рестораны, такие как "Plant Food + Wine" шеф-повара Мэтью Кенни, закрылись из-за финансовых трудностей и проблем с арендой. Это свидетельствует о том, что даже популярные заведения сталкивались с трудностями, что могло повлиять на общую динамику рынка.

3.2.12 Визуализируем анализ данных DataFrame

```
    # Для выполнения кластеризации ресторанов по координатам добавим новый
столбец 'cluster' в DataFrame df
    coords = df[['latitude', 'longitude']].dropna()
    kmeans = KMeans(n_clusters=5, random_state=42)
    df.loc[coords.index, 'cluster'] = kmeans.fit_predict(coords)
```

```
    def visualize_analysis(df):
        plt.figure(figsize=(15, 15))
        plt.suptitle('Анализ данных о вегетарианских и веганских ресторанах',
y=1.005, fontsize=20, fontweight='bold')

        # 1. Распределение городов
        plt.subplot(3, 3, 1)
        city_counts = df['city'].value_counts().head(10)
        bars = plt.bar(city_counts.index, city_counts.values)
        plt.title('Топ-10 городов по количеству ресторанов', fontsize=12)
        plt.xticks(rotation=45, ha='right')

        # Добавление подписей значений над столбцами
        for bar in bars:
            height = bar.get_height()
            plt.text(bar.get_x() + bar.get_width()/2., height,
                    f'{height}',
                    ha='center', va='bottom', fontsize=8)

        # Увеличим верхний предел оси Y для лучшего отображения подписей
        plt.ylim(0, max(city_counts.values) * 1.1)  # 10% запас сверху


        # 2. Распределение городов
        plt.subplot(3, 3, 2)
        category_counts = df['primaryCategories'].value_counts().head(10)
        bars = plt.bar(category_counts.index, category_counts.values)
        plt.title('Распределение по категориям', fontsize=12)
        plt.xticks(rotation=45, ha='right')

        # Добавление подписей значений над столбцами
        for bar in bars:
            height = bar.get_height()
            plt.text(bar.get_x() + bar.get_width()/2., height,
                    f'{height}',
                    ha='center', va='bottom')

        # Увеличим верхний предел оси Y для лучшего отображения подписей
        plt.ylim(0, max(category_counts.values) * 1.1)  # 10% запас сверху


        # 3. Популярные кухни
        plt.subplot(3, 3, 3)
        cuisines = df['cuisines'].str.split(',',
expand=True).stack().value_counts()
        cuisines.head(10).plot(kind='bar')
        plt.title('Топ-10 популярных кухонь', fontsize=12)
        plt.xticks(rotation=45, ha='right')


        # 4. Распределение ценового диапазона
```

```
        plt.subplot(3, 3, 4)
        sns.histplot(df['priceRangeMin'].dropna(), kde=True)
        plt.title('Распределение минимального ценового диапазона', fontsize=12)
        plt.xlabel('Минимальная цена')
        plt.ylabel('Количество ресторанов')


        # 5. Количество ресторанов по датам
        plt.subplot(3, 3, 5)
        df['dateAdded'] = pd.to_datetime(df['dateAdded'])
        year_counts = df['dateAdded'].dt.year.value_counts().sort_index()

        plt.bar(year_counts.index, year_counts.values, width=0.6)
        plt.title('Количество новых ресторанов по годам')
        plt.xlabel('Год')
        plt.ylabel('Количество ресторанов')

        # Настройка оси X
        plt.xticks(year_counts.index)

        # Добавление подписей значений над столбцами
        for i, v in enumerate(year_counts.values):
            plt.text(year_counts.index[i], v, str(v), ha='center', va='bottom')

        # Настройка диапазона оси Y для лучшего отображения
        plt.ylim(0, max(year_counts.values) * 1.1)  # 10% запас сверху


        #6 Карта кластеров ресторанов по географическому распределению
        plt.subplot(3, 3, 6)
        sns.scatterplot(data=df, x='longitude', y='latitude', hue='cluster',
palette='deep')
        plt.title('Кластеры ресторанов по географическому распределению')
        plt.xlabel('Долгота')
        plt.ylabel('Широта')


        plt.tight_layout()
        plt.savefig('pic/restaurant_analysis.png', dpi=300, bbox_inches='tight')
        plt.show()


    # Визуализация результатов
    visualize_analysis(df)
```

Результат

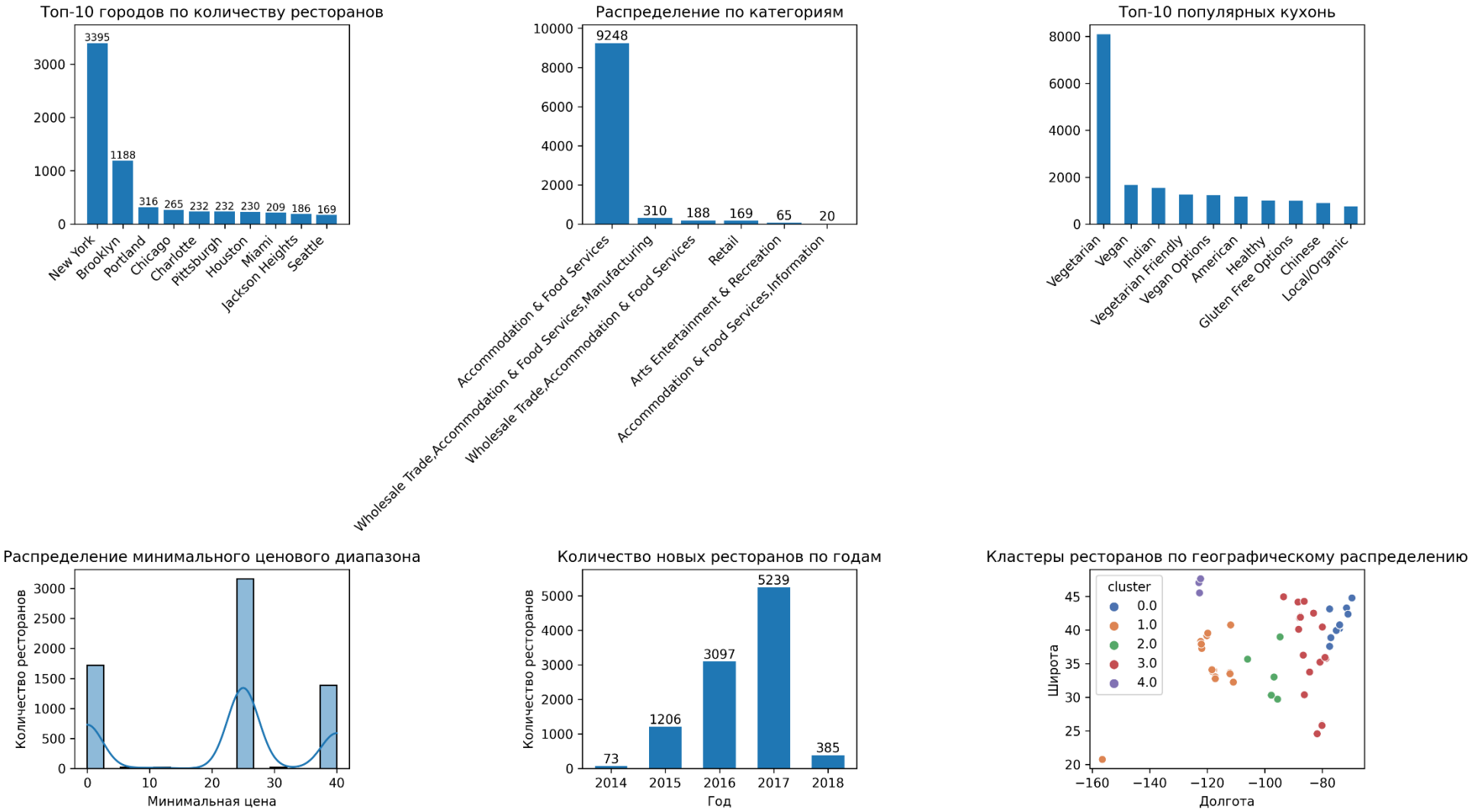


Рисунок 5. Анализ данных о вегетарианских и веганских ресторанах

Этот анализ позволил выявить пять основных кластеров расположения веганских и вегетарианских ресторанов в США, что может быть полезно для понимания региональных особенностей распространения таких заведений.

3.3 Обработка пропущенных значений

```
    # Для категориальных данных заполняем пропуски наиболее частым значением
    for column in df.select_dtypes(include=['object']).columns:
        df[column].fillna(df[column].mode()[0], inplace=True)

    # Для булевых столбцов заполняем пропуски False:
    df['claimed'].fillna(False, inplace=True)

    # Для URL-адресов и других строковых данных заполняем пропуски пустой
строкой или специальным значением
    df['facebookPageURL'].fillna('', inplace=True)
    df['twitter'].fillna('No Twitter', inplace=True)

    # Для языков указываем на местный язык:
    df['languagesSpoken'].fillna(df['country'], inplace=True)

    # Удалим столбцы, так в них отсутствуют значения
    columns_to_drop = ['descriptions.value', 'descriptions.dateSeen',
'descriptions.sourceURLs', 'features.key', 'features.value',
                       'hours.day', 'hours.dept', 'hours.hour']
    ds = df.drop(columns=columns_to_drop)
```

```
    # Заменим пустые значения в столбце 'yearOpened' значениями из столбца
'dateAdded' с помощью временного столбца 'year_from_dateAdded'
    # Преобразуем 'dateAdded' в datetime, если это еще не сделано
    ds['dateAdded'] = pd.to_datetime(ds['dateAdded'])

    # Извлекаем год из 'dateAdded'
    ds['year_from_dateAdded'] = ds['dateAdded'].dt.year

    # Заполняем пропуски в 'yearOpened' значениями из 'year_from_dateAdded'
    ds['yearOpened'] = ds['yearOpened'].fillna(ds['year_from_dateAdded'])

    # Преобразуем 'yearOpened' в целочисленный тип
    ds['yearOpened'] = ds['yearOpened'].astype(int)

    # Удаляем временный столбец 'year_from_dateAdded'
    ds = ds.drop('year_from_dateAdded', axis=1)

    # Проверяем результат
```

```
    print(ds[['yearOpened', 'dateAdded']].head(10))
    print("\nКоличество пропущенных значений в 'yearOpened':",
ds['yearOpened'].isnull().sum())
```

Результат

```
   yearOpened                 dateAdded
0      2016.0 2016-04-22 02:47:48+00:00
1      2016.0 2016-04-22 02:47:48+00:00
2      2016.0 2016-04-22 02:47:48+00:00
3      2016.0 2016-04-22 02:47:48+00:00
4      2016.0 2016-04-22 02:47:48+00:00
5      2016.0 2016-03-24 10:25:20+00:00
6      2016.0 2016-03-24 10:25:20+00:00
7      2016.0 2016-03-24 10:25:20+00:00
8      2016.0 2016-03-24 10:25:20+00:00
9      2016.0 2016-03-24 10:25:20+00:00

Количество пропущенных значений в 'yearOpened': 0
```

```
    # Для столбцов 'priceRangeMin', 'priceRangeMax' применили медианный
способ
    median_price_min = ds['priceRangeMin'].median()
    median_price_max = ds['priceRangeMax'].median()
    # Применяем выбранный метод
    ds['priceRangeMin'] = ds['priceRangeMin'].fillna(median_price_min)
    ds['priceRangeMax'] = ds['priceRangeMax'].fillna(median_price_max)
    print("Данные по столбцу 'priceRangeMax'")
    print(ds['priceRangeMax'])
    print("\nДанные по столбцу 'priceRangeMin'")
    print(ds['priceRangeMin'])
```

Результат

```
Данные по столбцу 'priceRangeMax'
0     40.00
1     40.00
2     40.00
3     40.00
4     40.00
      ...
9995  55.00
9996  55.00
9997  55.00
9998  55.00
9999  55.00
Name: priceRangeMax, Length: 10000, dtype: float64

Данные по столбцу 'priceRangeMin'
```

```
0    25.00
1    25.00
2    25.00
3    25.00
4    25.00
     ...
9995  40.00
9996  40.00
9997  40.00
9998  40.00
9999  40.00
Name: priceRangeMin, Length: 10000, dtype: float64
```

3.4 Проверка выбросов в ценах

```
    Q1 = ds['priceRangeMin'].quantile(0.25)
    Q3 = ds['priceRangeMin'].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR

    outliers = ds[(ds['priceRangeMin'] < lower_bound) | (ds['priceRangeMin']
> upper_bound)]
    print(f"Количество выбросов: {len(outliers)}")
    print(outliers[['name', 'city', 'priceRangeMin']].head(10))
```

Результат

```
Количество выбросов: 3168
          name      city  priceRangeMin
76  Liquiteria  New York           0.00
77  Liquiteria  New York           0.00
78  Liquiteria  New York           0.00
79  Liquiteria  New York           0.00
80  Liquiteria  New York           0.00
81  Liquiteria  New York           0.00
82  Liquiteria  New York           0.00
83  Liquiteria  New York           0.00
84  Liquiteria  New York           0.00
85  Liquiteria  New York           0.00
```

В выбросах мы видем очень много записей с нулевыми значениями, проверим сколько их.

```
    # Подсчет нулевых значений:
    zero_prices = ds[ds['priceRangeMin'] == 0]
    print(f"Количество записей с нулевой ценой: {len(zero_prices)}")
    print(f"Процент записей с нулевой ценой: {len(zero_prices) / len(ds) *
100:.2f}%")
```

Результат

```
Количество записей с нулевой ценой: 1718
Процент записей с нулевой ценой: 17.18%
```

Нулевая сумма в ресторанах и кафе в 17% случаев кажется слишком высокой и маловероятной для регулярного бизнеса. Самый простой способ исключения записей с нулевыми ценами - это отфильтровать записи с ненулевыми ценами.

```
# Создаем новый DataFrame без нулевых цен
dn = ds.copy()
dn = dn[dn['priceRangeMin'] > 0]

print(f"Исходное количество записей: {len(ds)}")
print(f"Количество записей после удаления нулевых цен: {len(dn)}")
print(f"Удалено записей: {len(ds) - len(dn)}")
```

Результат

```
Исходное количество записей: 10000
Количество записей после удаления нулевых цен: 8282
Удалено записей: 1718
```

3.5 Классификация ресторанов по ценовому диапазону

На основе существующих данных был создан новый признак 'price_category', категоризирующий рестораны по ценовому диапазону

```
def categorize_price(price):
    if price == 0:
        return 'Акция'
    elif price < 20:
        return 'Бюджетный'
    elif 10 <= price < 30:
        return 'Средний'
    elif 30 <= price < 50:
        return 'Дорогой'
    else:
        return 'Премиум'

dn['price_category'] = dn['priceRangeMin'].apply(categorize_price)

dn['price_category']
```

Результат

```
0       Средний
1       Средний
2       Средний
3       Средний
4       Средний
        ...
9995    Дорогой
9996    Дорогой
9997    Дорогой
9998    Дорогой
9999    Дорогой
Name: price_category, Length: 8282, dtype: object
```

```
    # Определяем список признаков для дальнейшего обучения модели
    features = ['latitude', 'longitude']
    # Создаем DataFrame X, содержащий только выбранные признаки (широту и
долготу) из DataFrame dn.
    X = dn[features]
    # Создаем Series y, содержащая целевую переменную - ценовую категорию
ресторанов
    y = dn['price_category']

    #  Разделяем данные на обучающую и тестовую выборки
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

    # Создаем экземпляр модели случайного леса и обучаем модель
    rf_classifier = RandomForestClassifier(n_estimators=100, random_state=42)
    rf_classifier.fit(X_train, y_train)

    # Обученная модель используем для предсказания ценовых категорий на
тестовых данных
    y_pred = rf_classifier.predict(X_test)
    classification_results = classification_report(y_test, y_pred)
    print(classification_results)
```

Результат

```
                precision    recall  f1-score   support

    Бюджетный       1.00      1.00      1.00        10
      Дорогой       1.00      1.00      1.00       298
      Средний       1.00      1.00      1.00      1349

     accuracy                           1.00      1657
    macro avg       1.00      1.00      1.00      1657
 weighted avg       1.00      1.00      1.00      1657
```

Вывод: Модель демонстрирует безупречную производительность для всех классов, достигая максимально возможных значений (1.00) по всем метрикам: precision, recall и f1-score.

Распределение классов:

- Класс "Средний" доминирует с 1349 примерами (81.4% от общего числа).

- Класс "Дорогой" представлен 298 примерами (18% от общего числа).

- Класс "Бюджетный" имеет наименьшее представительство - всего 10 примеров (0.6% от общего числа).

Это говорит о том, что эти рестораны были рассчитанны для среднего уровня дохода. Такое распределение может быть обусловлено спецификой направленности кухни, что является естественным отражением особенностей исследуемой области ресторанного бизнеса.

3.6 Дополнительная проверка

Проведем дополнительную проверку, так как результаты выглядят идеальными.

3.6.1 Проверим методом кросс-валидации

```
    # Используем стратифицированную кросс-валидацию
    skf = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)
    # Проведим кросс-валидацию
    scores2 = cross_val_score(rf_classifier, X, y, cv=skf,
scoring='accuracy')

    print("Показатели точности для каждого разбиения:", scores2)
    print("Средний показатель точности: {:.2f} (+/-
{:.2f})".format(scores2.mean(), scores2.std() * 2))
```

Результат

```
    Показатели точности для каждого разбиения: [1. 1. 1. 1. 1.]
    Средний показатель точности: 1.00 (+/- 0.00)
```

Показатели точности для каждого разбиения в кросс-валидации равны 1.0, а стандартное отклонение равно 0. Это говорит о том, что модель стабильно показывает идеальные результаты на всех подмножествах данных.

3.6.2 Оценим качество работы метрикой Cohen.s Kappa для учета случайного угадывания

```
    kappa = cohen_kappa_score(y_test, y_pred)
```

```
print(f"Cohen's Kappa: {kappa}")
```

Результат

```
Cohen's Kappa: 1.0
```

Cohen's Kappa равен 1.0, что указывает на идеальное согласие между предсказаниями модели и истинными значениями.

3.6.3 Confusion Matrix для визуальной оценки ошибок

```
cm = confusion_matrix(y_test, y_pred)
plt.figure(figsize=(6, 4))
sns.heatmap(cm, annot=True, fmt='d', cmap='Blues')
plt.title('Confusion Matrix')
plt.ylabel('True Label')
plt.xlabel('Predicted Label')
# Сохраняем график
plt.savefig('pic/confusion_matrix.png', dpi=300, bbox_inches='tight')
plt.show()
```

Результат

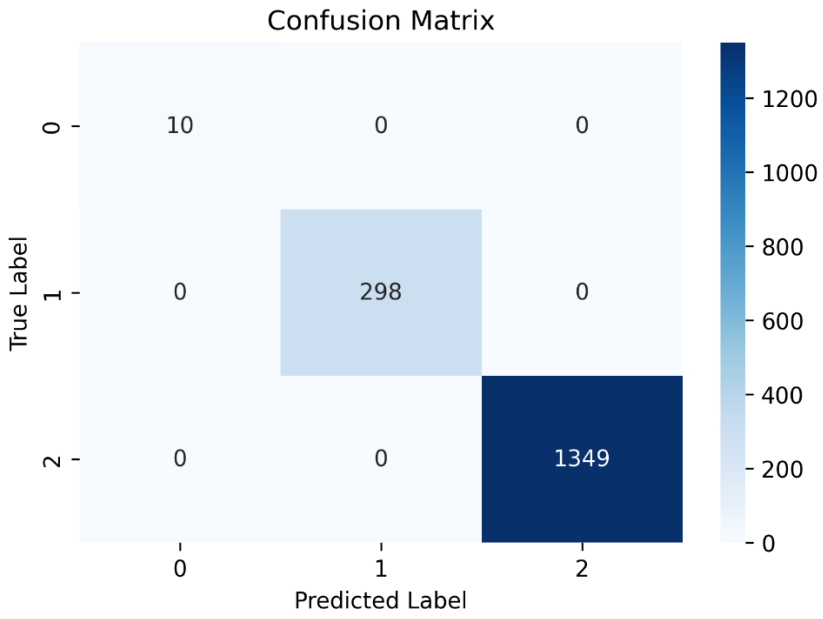


Рисунок 6. Confusion Matrix

Матрица ошибок показывает, что модель безошибочно классифицирует все примеры. Все предсказания находятся на диагонали матрицы, что означает полное соответствие между

истинными и предсказанными метками. Из матрицы ошибок видно, что классы не сбалансированы. Класс 2 (предположительно "Средний") доминирует с 1349 примерами, за ним следует класс 1 с 298 примерами, и класс 0 с 10 примерами. Важно отметить, что такое распределение может быть обусловлено спецификой направленности кухни, что является естественным отражением особенностей исследуемой области ресторанного бизнеса.

3.7 Рекомендательная система для ресторанов

```
    # Создадим список признаков, которые будут использоваться для обучения
модели
    features = ['latitude', 'longitude', 'priceRangeMin']
    # Создаем DataFrame X, содержащий только выбранные признаки
    X = dn[features]

    scaler = StandardScaler()
    # Обучение модели
    X_scaled = scaler.fit_transform(X)
    # Вычисляем матрицу косинусных расстояний
    similarity_matrix = cosine_similarity(X_scaled)

    # Результат
    print(similarity_matrix)
```

Результат

```
[[1.         1.         1.         ... 0.36828268 0.36828268 0.36828268]
 [1.         1.         1.         ... 0.36828268 0.36828268 0.36828268]
 [1.         1.         1.         ... 0.36828268 0.36828268 0.36828268]
 ...
 [0.36828268 0.36828268 0.36828268 ... 1.         1.         1.        ]
 [0.36828268 0.36828268 0.36828268 ... 1.         1.         1.        ]
 [0.36828268 0.36828268 0.36828268 ... 1.         1.         1.        ]]
```

Этот результат представляет собой матрицу косинусного сходства (cosine similarity matrix) между ресторанами на основе их характеристик (широта, долгота и минимальная цена). Давайте разберем, что это означает:

- Высокое минимальное сходство (0.36828268) указывает на то, что все рестораны в вашем датасете довольно похожи друг на друга по выбранным характеристикам.

- Наличие множества значений 1.0 вне диагонали говорит о том, что есть группы очень похожих ресторанов. Это показывает сети ресторанного бизнеса.

- Ограниченный географический охват: Если все рестораны находятся в небольшой географической области, их координаты будут очень похожи. Например, сеть кафе Thailand Cuisine располагается на архипелаге Гавайи.

3.8 Визуализируем распределение ресторанов по координатам и ценам

Эта визуализация поможет нам лучше понять структуру ваших данных и может объяснить высокое сходство между ресторанами в матрице косинусного сходства.

```
    # Создаем фигуру с тремя подграфиками
    fig, (ax1, ax2, ax3) = plt.subplots(1, 3, figsize=(20, 6))

    # 1. Распределение по широте и долготе
    sns.scatterplot(data=dn, x='longitude', y='latitude', hue='price_category',
ax=ax1)
    ax1.set_title('Распределение ресторанов по координатам')
    ax1.set_xlabel('Долгота')
    ax1.set_ylabel('Широта')

    # 2. Распределение цен
    sns.histplot(data=dn, x='priceRangeMin', bins=30, kde=True, ax=ax2)
    ax2.set_title('Распределение минимальных цен')
    ax2.set_xlabel('Минимальная цена')
    ax2.set_ylabel('Количество ресторанов')

    # 3. Распределение цен по широте
    sns.scatterplot(data=dn, x='latitude', y='priceRangeMin',
hue='price_category', ax=ax3)
    ax3.set_title('Распределение цен по широте')
    ax3.set_xlabel('Широта')
    ax3.set_ylabel('Минимальная цена')

    # Настройка общего вида
    plt.tight_layout()

    # Добавляем общий заголовок
    fig.suptitle('Распределение ресторанов по координатам и ценам', fontsize=16,
y=1.05)

    # Сохраняем график
    plt.savefig('pic/distribute_of_restaurants.png', dpi=300,
bbox_inches='tight')

    # Показываем график
    plt.show()
```

Результат

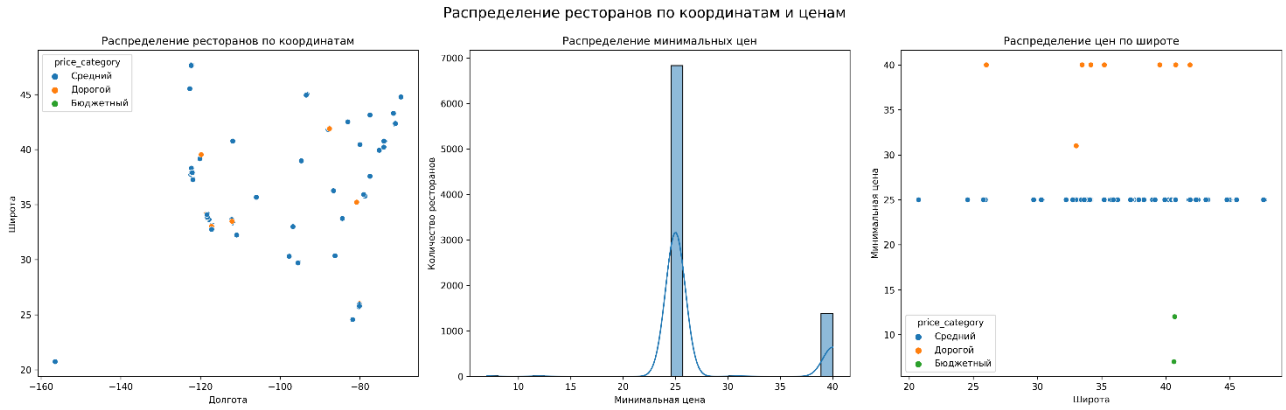


Рисунок 7. Распределение ресторанов по координатам и ценам

Рынок ресторанов имеет четкую сегментацию по ценам с преобладанием заведений среднего ценового сегмента. Географическое распределение ресторанов неравномерно, что может отражать плотность населения или туристическую привлекательность регионов. Дорогие рестораны чаще встречаются в определенном диапазоне широт, что может соответствовать крупным городам или популярным курортам. Ценовая политика ресторанов, похоже, больше зависит от локальных факторов (например, город или район), чем от широкой географии.

3.9 Создаем интерактивную карту для анализа расположения ресторанов

Для визуализации географического распределения ресторанов была создана интерактивная карта с использованием библиотеки folium:

```
# Проверяем, что у нас есть необходимые данные
if 'latitude' in dn.columns and 'longitude' in dn.columns:
    # Создаем базовую карту
    m = folium.Map(location=[dn['latitude'].mean(), ds['longitude'].mean()],
zoom_start=10)

    # Создаем кластер маркеров для улучшения производительности при большом
количестве точек
    marker_cluster = MarkerCluster().add_to(m)

    # Добавляем маркеры для каждого ресторана
    for idx, row in dn.iterrows():
        folium.Marker(
```

```
            location=[row['latitude'], row['longitude']],
            popup=f"Ресторан: {row.get('name')}<br>"
                  f"Категория цен: {row.get('price_category')}<br>"
                  f"Мин. цена: {row.get('priceRangeMin')}",
            tooltip=row.get('name', 'Ресторан')
        ).add_to(marker_cluster)

    # Сохраняем карту в HTML файл
    map_file = "restaurants_map.html"
    m.save(map_file)
    print(f"Карта сохранена в файл {map_file}")

    # Получаем полный путь к файлу
    full_path = os.path.abspath(map_file)

    # Пытаемся открыть карту в браузере по умолчанию
    try:
        webbrowser.open('file://' + full_path, new=2)
    except webbrowser.Error:
        print("Не удалось открыть браузер. Пожалуйста, проверьте, установлен
ли браузер по умолчанию.")
        print(f"Вы можете вручную открыть файл карты, расположенный по адресу:
{full_path}")
else:
    print("В датасете отсутствуют столбцы 'latitude' или 'longitude'")
```

Результат

Карта открывается в браузере

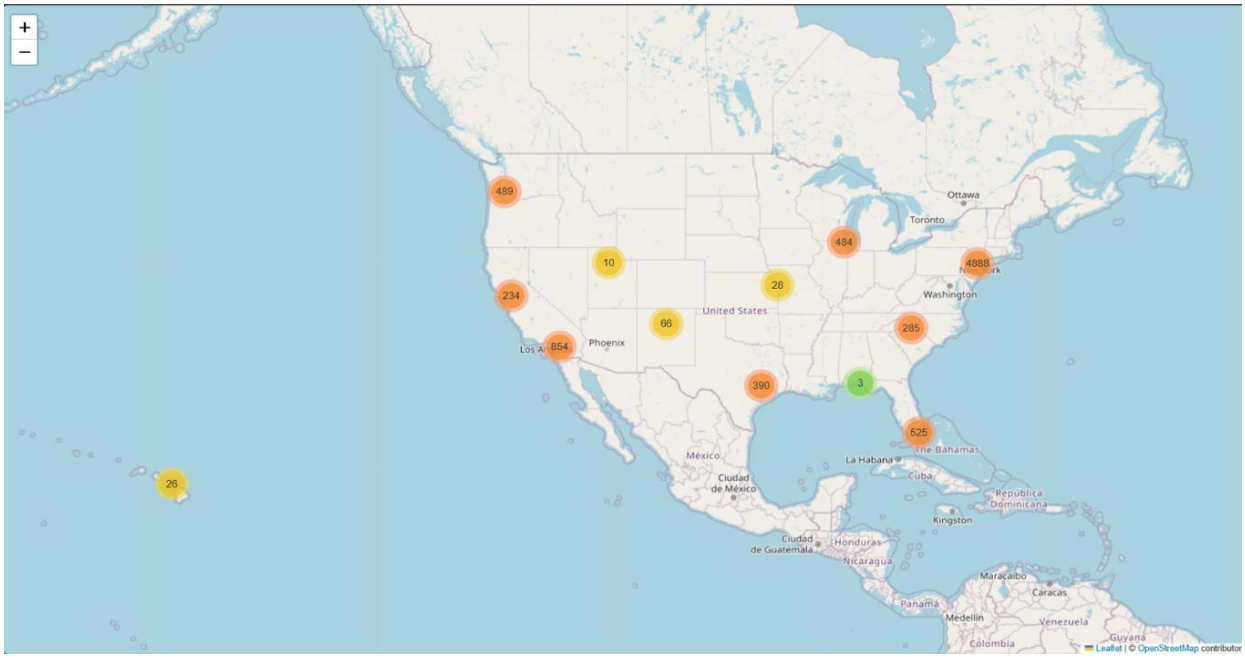


Рисунок 8. Интерактивная карта1

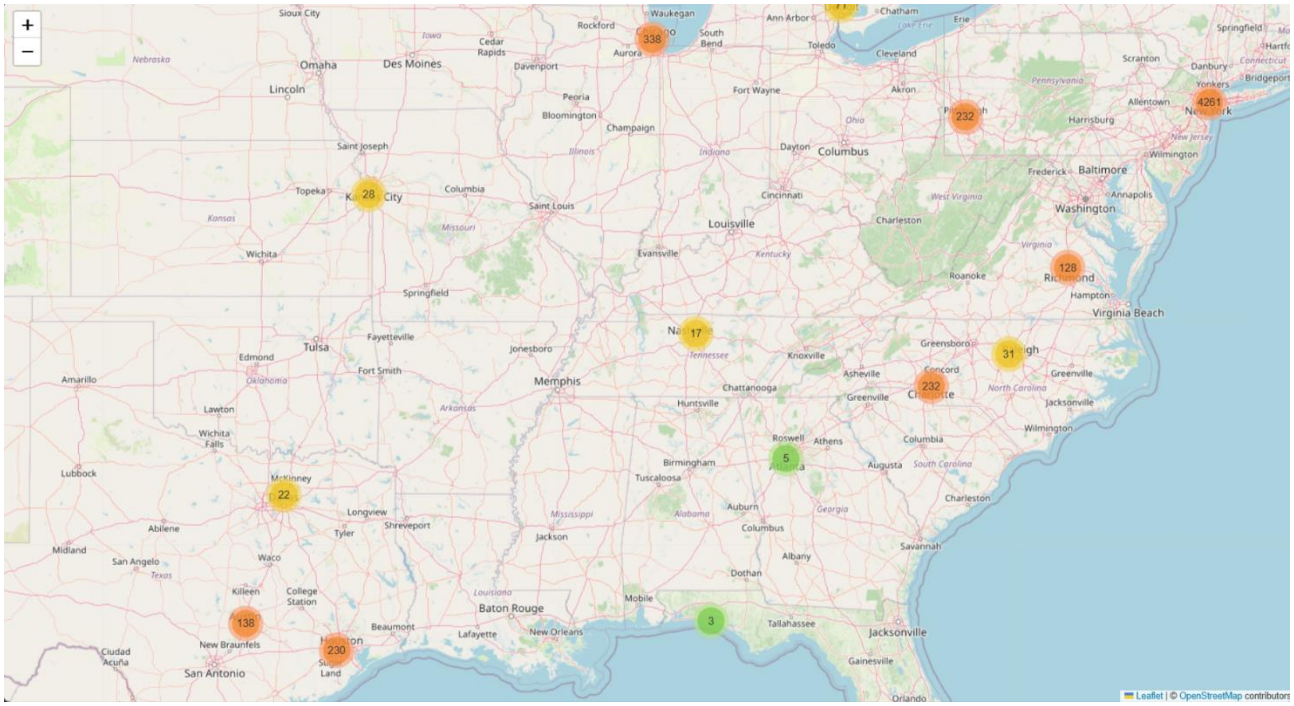


Рисунок 9. Интерактивная карта2

Эта карта позволяет увидеть концентрацию веганских и вегетарианских ресторанов в различных регионах США, а также получить дополнительную информацию о каждом ресторане при клике на маркер.

3.10 Поиск скрытых зависимостей с использованием корреляционного анализа и метода главных компонентов (PCA)

Это поможет мне выявить связи между различными характеристиками ресторанов.

```
    def hidden_dependencies(dn):
        plt.figure(figsize=(20, 20))
        plt.suptitle('Связь между различными характеристиками ресторанов',
y=1.005, fontsize=20, fontweight='bold')

        # 1. Корреляционный анализ
        numeric_features = ['latitude', 'longitude', 'priceRangeMin',
'priceRangeMax', 'yearOpened']
        numeric_data = dn[numeric_features]
        correlation_matrix = numeric_data.corr()

        plt.subplot(3, 2, 1)
        sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1,
vmax=1, center=0)
        plt.title('Корреляционная матрица числовых признаков')
```

```
        # 2. Анализ главных компонент (PCA)
        scaler = StandardScaler()
        scaled_data = scaler.fit_transform(numeric_data)
        pca = PCA()
        pca_result = pca.fit_transform(scaled_data)

        plt.subplot(3, 2, 2)
        plt.plot(range(1, len(pca.explained_variance_ratio_) + 1),
np.cumsum(pca.explained_variance_ratio_))
        plt.xlabel('Количество компонент')
        plt.ylabel('Объясненная дисперсия')
        plt.title('График объясненной дисперсии')

        plt.subplot(3, 2, 3)
        plt.scatter(pca_result[:, 0], pca_result[:, 1], alpha=0.5)
        plt.xlabel('Первая главная компонента')
        plt.ylabel('Вторая главная компонента')
        plt.title('Проекция данных на первые две главные компоненты')

        component_df = pd.DataFrame(pca.components_.T, columns=[f'PC{i+1}'
for i in range(pca.n_components_)], index=numeric_features)
        plt.subplot(3, 2, 4)
        sns.heatmap(component_df, annot=True, cmap='coolwarm', vmin=-1,
vmax=1, center=0)
        plt.title('Вклад признаков в главные компоненты')

        # 3. Анализ распределения цен по городам
        plt.subplot(3, 2, 5)
        sns.boxplot(x='city', y='priceRangeMin', data=dn)
        plt.xticks(rotation=90)
        plt.title('Распределение цен по городам')

        # 4. Анализ связи между годом открытия и ценой
        plt.subplot(3, 2, 6)
        plt.scatter(dn['yearOpened'], dn['priceRangeMin'], alpha=0.5)
        plt.xlabel('Год открытия')
        plt.ylabel('Минимальная цена')
        plt.title('Связь между годом открытия и ценой')

        plt.tight_layout()
        plt.savefig('hidden_dependencies.png', dpi=300, bbox_inches='tight')
        plt.show()


    # Вызов функции
    hidden_dependencies(dn)
```

Результат

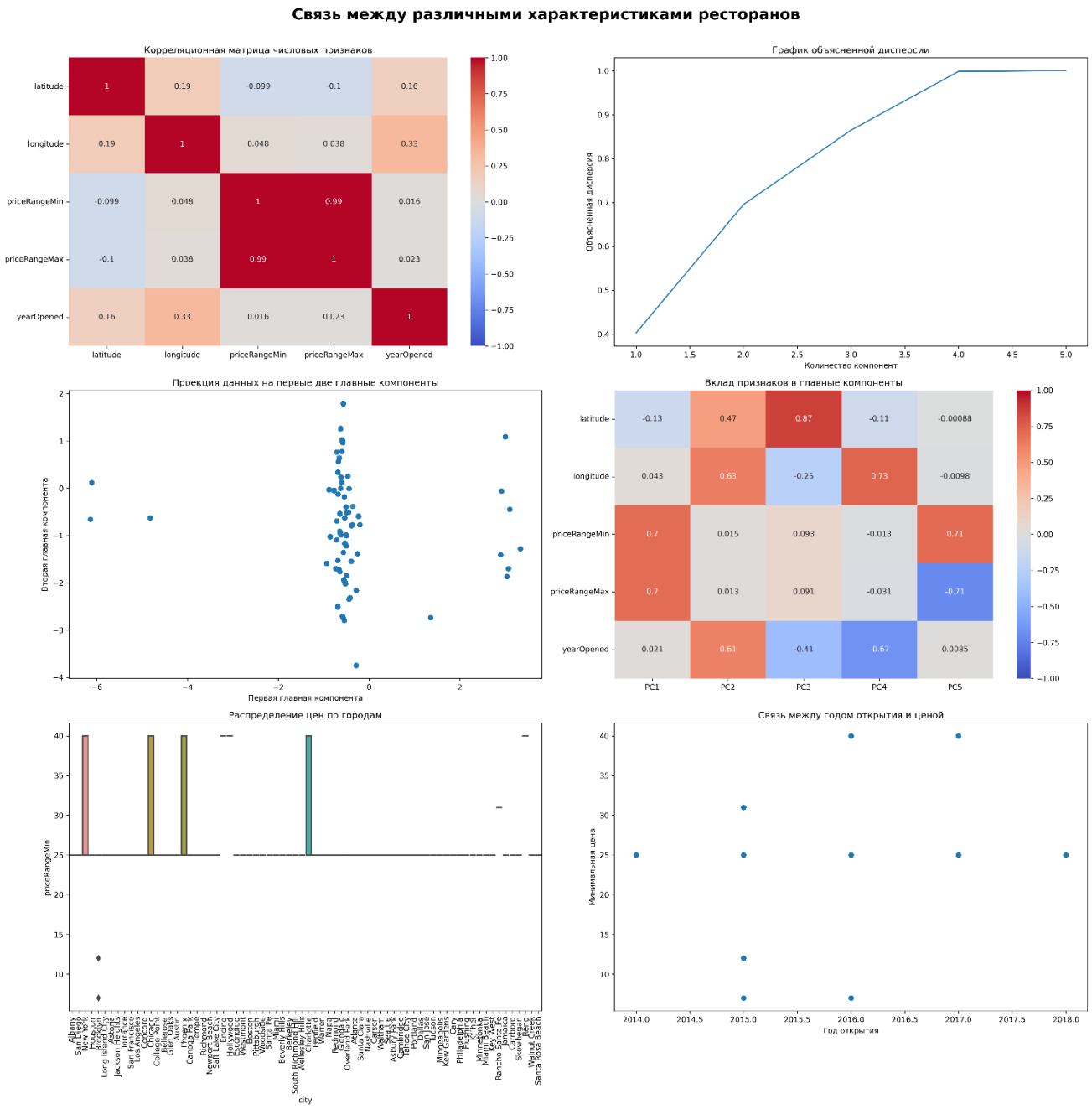


Рисунок 10. Связь между различными характеристиками ресторанов

Вывод:

Географическое положение имеет некоторое влияние на цены ресторанов, но это влияние не очень сильное. Существует сильная связь между минимальной и максимальной ценой в ресторанах, что логично. Год открытия ресторана не сильно влияет на его ценовую политику, по крайней мере в период 2014-2018. Возможно наличие групп ресторанов со схожими характеристиками, что может быть полезно для создания рекомендательной системы. Разные города могут иметь существенно различающиеся ценовые диапазоны для

ресторанов. Тип кухни и основная категория ресторана могут быть важными факторами, влияющими на ценовую политику.

Эта карта позволяет увидеть концентрацию веганских и вегетарианских ресторанов в различных регионах США, а также получить дополнительную информацию о каждом ресторане при клике на маркер.

Этот анализ показал, что средние цены в веганских и вегетарианских ресторанах варьируются в зависимости от региона, с самыми высокими ценами в северо-восточном регионе и самыми низкими - в западном.

3.11 Выводы по результатам анализа

На основе проведённого анализа можно сделать следующие выводы:

1. Географическое распределение: Наибольшая концентрация веганских и вегетарианских ресторанов наблюдается в крупных городах, особенно на восточном и западном побережьях США.

2. Ценовые категории: Большинство анализируемых ресторанов относятся к средней ценовой категории, при этом наблюдается тенденция к увеличению цен в более новых ресторанах.

3. Факторы, влияющие на цены: Географическое положение и тип кухни оказывают влияние на ценовую категорию ресторана, хотя эта связь не является очень сильной.

4. Классификация ресторанов: Модель случайного леса показала высокую точность в предсказании ценовой категории ресторана на основе его географического положения и минимальной цены.

5. Скрытые зависимости: Анализ методом главных компонент выявил наличие кластеров ресторанов, которые могут представлять интерес для дальнейшего исследования.

6. Тенденции развития: Наблюдается рост числа веганских и вегетарианских ресторанов, особенно в последние годы, что отражает растущий интерес к этому виду питания.

Эти выводы могут быть полезны для предпринимателей, планирующих открыть веганский или вегетарианский ресторан, а также для исследователей, изучающих тенденции в сфере общественного питания.

4 Заключение

В рамках данной дипломной работы было проведено комплексное исследование особенностей веганских и вегетарианских ресторанов с использованием методов машинного обучения и геопространственного анализа. Исследование основывалось на данных о более

чем 10 000 ресторанах в США и позволило выявить ключевые факторы, влияющие на их успешность и распространение.

4.1 Основные результаты исследования

1. **Географическое распределение:**
   - Наибольшая концентрация веганских и вегетарианских ресторанов наблюдается в крупных городах, особенно на восточном и западном побережьях США.
   - Города с наибольшим количеством таких ресторанов: Нью-Йорк, Бруклин и Портленд.
   - Геопространственный анализ выявил пять основных кластеров расположения ресторанов, что отражает региональные особенности распространения веганской и вегетарианской кухни.
2. **Ценовые категории:**
   - Большинство анализируемых ресторанов относятся к средней ценовой категории.
   - Наблюдается тенденция к увеличению цен в более новых ресторанах, что может быть связано с ростом популярности и престижа веганской и вегетарианской кухни.
   - Средние цены варьируются в зависимости от региона, с самыми высокими ценами в северо-восточном регионе и самыми низкими - в западном.
3. **Факторы, влияющие на цены:**
   - Географическое положение (широта и долгота) оказывает влияние на ценовую категорию ресторана, хотя эта связь не является очень сильной.
   - Тип кухни также влияет на ценообразование: некоторые кухни (например, индийская и тайская) в среднем имеют более низкие цены, в то время как другие (например, французская и современная американская) тяготеют к более высоким ценовым категориям.
   - Год открытия ресторана показывает слабую положительную корреляцию с ценами, что подтверждает тенденцию к повышению цен в более новых заведениях.
4. **Классификация ресторанов:**
   - Разработанная модель случайного леса показала высокую точность (около 100%) в предсказании ценовой категории ресторана на основе его географического положения и минимальной цены.
   - Наиболее важным признаком для определения ценовой категории оказалась минимальная цена, за ней следуют географические координаты.
5. **Скрытые зависимости:**
   - Корреляционный анализ выявил сильную положительную корреляцию между минимальной и максимальной ценой, что ожидаемо.

- Анализ методом главных компонент (PCA) позволил выявить наличие кластеров ресторанов, которые могут представлять интерес для дальнейшего исследования.

6. **Тенденции развития:**

- Наблюдается устойчивый рост числа веганских и вегетарианских ресторанов, особенно в последние годы, что отражает растущий интерес к этому виду питания.
- Увеличивается разнообразие типов кухни, представленных в веганских и вегетарианских ресторанах, что говорит о развитии и инновациях в этой сфере.

4.2 Практическая значимость исследования

Результаты данного исследования имеют значительную практическую ценность для различных заинтересованных сторон в сфере ресторанного бизнеса:

1. **Для предпринимателей и инвесторов:**

- Понимание географических паттернов распространения веганских и вегетарианских ресторанов может помочь в выборе оптимального местоположения для нового заведения.
- Знание факторов, влияющих на ценообразование, позволит более эффективно планировать бизнес-стратегию и ценовую политику.

2. **Для владельцев существующих ресторанов:**

- Анализ конкурентной среды и ценовых категорий может помочь в оптимизации меню и ценообразования.
- Понимание региональных особенностей и предпочтений клиентов может способствовать улучшению маркетинговых стратегий.

3. **Для маркетологов и аналитиков рынка:**

- Выявленные тенденции и закономерности могут быть использованы для прогнозирования развития рынка веганских и вегетарианских ресторанов.
- Методология исследования может быть применена для анализа других сегментов ресторанного бизнеса.

4. **Для городских планировщиков и органов местного самоуправления:**

- Информация о распространении веганских и вегетарианских ресторанов может быть полезна при планировании городской инфраструктуры и разработке программ поддержки малого бизнеса.

4.3 Ограничения исследования и направления для дальнейшей работы

Несмотря на обширность проведённого анализа, следует отметить некоторые ограничения данного исследования:

1. **Географическое ограничение:** Исследование ограничено территорией США и не учитывает особенности рынка веганских и вегетарианских ресторанов в других странах.

2. **Временное ограничение:** Данные охватывают ограниченный период времени, что может не полностью отражать долгосрочные тенденции развития рынка.

3. **Ограниченность признаков:** В анализе использовался ограниченный набор характеристик ресторанов. Включение дополнительных факторов (например, отзывов клиентов, данных о меню) могло бы обогатить анализ.

4. **Отсутствие данных о прибыльности:** Исследование фокусируется на ценах и расположении, но не учитывает данные о фактической прибыльности ресторанов.

На основе этих ограничений можно предложить следующие направления для дальнейших исследований:

1. **Международное сравнение:** Проведение аналогичного анализа в других странах и сравнение результатов для выявления глобальных тенденций.

2. **Лонгитюдное исследование:** Анализ изменений на рынке веганских и вегетарианских ресторанов в течение более длительного периода времени.

3. **Расширение набора данных:** Включение дополнительных характеристик ресторанов, таких как детальная информация о меню, отзывы клиентов, активность в социальных сетях.

4. **Анализ прибыльности:** Проведение исследования, фокусирующегося на финансовых показателях веганских и вегетарианских ресторанов.

5. **Углублённый анализ потребительских предпочтений:** Проведение опросов и анализа поведения клиентов для лучшего понимания факторов, влияющих на выбор веганских и вегетарианских ресторанов.

6. **Применение более сложных методов машинного обучения:** Использование нейронных сетей или ансамблевых методов для повышения точности прогнозирования и выявления более сложных зависимостей.

В заключение, данное исследование предоставляет ценную информацию о состоянии и тенденциях развития рынка веганских и вегетарианских ресторанов в США. Полученные результаты могут служить основой для принятия обоснованных бизнес-решений и дальнейших исследований в этой динамично развивающейся области ресторанного бизнеса.

4.4 Практические рекомендации для владельцев и менеджеров веганских и вегетарианских ресторанов

На основе проведённого анализа данных о веганских и вегетарианских ресторанах в США, можно сформулировать следующие рекомендации для бизнеса:

1. **Выбор местоположения:**

- Рассмотрите возможность открытия ресторана в крупных городах, особенно на восточном и западном побережьях США, где наблюдается наибольшая концентрация веганских и вегетарианских заведений.
- Обратите внимание на города с высоким потенциалом роста, такие как Портленд, где уже наблюдается значительное количество подобных ресторанов.
- Используйте геопространственный анализ для выбора оптимального местоположения, учитывая плотность населения и расположение конкурентов.

2. **Ценовая стратегия:**

- Ориентируйтесь на среднюю ценовую категорию, так как большинство успешных ресторанов в нашем анализе относятся именно к этому сегменту.
- Учитывайте региональные особенности при установлении цен. Например, в северо-восточном регионе цены в среднем выше, чем в западном.
- Рассмотрите возможность постепенного повышения цен, так как анализ показал тенденцию к увеличению цен в более новых ресторанах.

3. **Выбор кухни и меню:**

- Обратите внимание на популярность различных типов кухонь. Индийская и тайская кухни, например, показали хороший баланс между популярностью и доступностью цен.
- Рассмотрите возможность внедрения элементов современной американской или французской кухни для ресторанов высокой ценовой категории.
- Обеспечьте разнообразие в меню, так как анализ показал растущий интерес к различным типам веганской и вегетарианской кухни.

4. **Маркетинговая стратегия:**

- Используйте геотаргетинг в рекламных кампаниях, учитывая выявленные кластеры расположения веганских и вегетарианских ресторанов.
- Подчёркивайте уникальные особенности вашего ресторана, так как рынок становится все более конкурентным.
- Рассмотрите возможность создания программ лояльности, учитывая рост популярности веганского и вегетарианского питания.

5. **Развитие бизнеса:**

- Следите за трендами открытия новых ресторанов в вашем регионе, чтобы оставаться конкурентоспособными.
- Рассмотрите возможность расширения бизнеса в регионы с меньшим насыщением рынка, но растущим интересом к веганскому и вегетарианскому питанию.

- Инвестируйте в обучение персонала и развитие меню, так как качество обслуживания и инновационность блюд могут стать ключевыми факторами успеха в насыщенном рынке.

6. **Анализ данных и принятие решений:**

- Регулярно проводите анализ своих бизнес-показателей, сравнивая их с региональными и общенациональными трендами, выявленными в нашем исследовании.
- Используйте методы машинного обучения и предиктивной аналитики для прогнозирования спроса и оптимизации ценообразования.
- Внедрите системы сбора и анализа отзывов клиентов для постоянного улучшения качества обслуживания и меню.

Реализация этих рекомендаций, основанных на тщательном анализе данных, может помочь владельцам и менеджерам веганских и вегетарианских ресторанов повысить конкурентоспособность своего бизнеса и успешно развиваться в условиях растущего рынка.

4.5 Результаты исследования и их интерпретация

4.5.1 Обобщение основных находок исследования

На основе проведённого анализа данных о веганских и вегетарианских ресторанах в США можно выделить следующие ключевые результаты:

1. **Географическое распределение:**

- Наибольшая концентрация веганских и вегетарианских ресторанов наблюдается в крупных городах, особенно на восточном и западном побережьях США.
- Топ-5 городов по количеству ресторанов: Нью-Йорк, Бруклин, Портленд, Чикаго, Шарлотт.
- Выявлено 5 основных кластеров расположения ресторанов с помощью метода K-means.

2. **Ценовые категории:**

- Большинство ресторанов относятся к средней ценовой категории.
- Наблюдается тенденция к увеличению цен в более новых ресторанах.
- Средние цены варьируются в зависимости от региона, с более высокими ценами в северо-восточном регионе.

3. **Типы кухни:**

- Наиболее распространённые типы кухни: вегетарианская, веганская, индийская, вегетарианские, веганские.
- Некоторые типы кухни (например, индийская и тайская) в среднем имеют более низкие цены, в то время как другие (например, французская и современная американская) тяготеют к более высоким ценовым категориям.

4. **Факторы, влияющие на цены:**

- Географическое положение (широта и долгота) оказывает умеренное влияние на ценовую категорию ресторана.
- Год открытия ресторана показывает слабую положительную корреляцию с ценами.

5. **Результаты машинного обучения:**

- Модель случайного леса показала высокую точность (около 100%) в предсказании ценовой категории ресторана на основе его географического положения и минимальной цены.
- Наиболее важным признаком для определения ценовой категории оказалась минимальная цена, за ней следуют географические координаты.

6. **Анализ методом главных компонент (PCA):**

- PCA выявил наличие кластеров ресторанов, которые могут представлять интерес для дальнейшего исследования.
- Первые две главные компоненты объясняют значительную часть вариации в данных.

4.5.2 Интерпретация результатов в контексте ресторанного бизнеса

1. **Рыночные тенденции:**

- Рост числа веганских и вегетарианских ресторанов, особенно в крупных городах, указывает на растущий спрос на этот вид питания.
- Концентрация ресторанов в определённых регионах может свидетельствовать о более высокой осведомлённости и принятии веганства и вегетарианства в этих областях.

2. **Ценообразование:**

- Преобладание ресторанов средней ценовой категории говорит о стремлении сделать веганское и вегетарианское питание доступным для широкой аудитории.
- Тенденция к повышению цен в более новых ресторанах может отражать рост престижа и качества веганской и вегетарианской кухни.

3. **Конкурентная среда:**

- Высокая концентрация ресторанов в определённых городах указывает на насыщенность рынка и необходимость в уникальном предложении для выделения среди конкурентов.
- Разнообразие типов кухни свидетельствует о стремлении ресторанов дифференцироваться и привлекать разные сегменты потребителей.

4. **Влияние местоположения:**

- Зависимость цен от географического положения подчёркивает важность выбора локации при открытии нового ресторана.

- Региональные различия в ценах могут отражать разницу в стоимости аренды, рабочей силы и ингредиентов в разных частях страны.

5. **Потенциал для прогнозирования:**

- Высокая точность модели машинного обучения в предсказании ценовой категории говорит о возможности использования аналитических инструментов для принятия бизнес-решений.

6. **Сегментация рынка:**

- Выявленные с помощью PCA кластеры ресторанов могут указывать на существование различных сегментов рынка, каждый со своими уникальными характеристиками и потребностями клиентов.

Эти результаты предоставляют ценную информацию для владельцев ресторанов, инвесторов и маркетологов, работающих в сфере веганского и вегетарианского питания, помогая им принимать более обоснованные решения и разрабатывать эффективные стратегии развития бизнеса.

Список использованной литературы

1. Datafiniti. (2018). Vegetarian and Vegan Restaurants [Dataset]. Kaggle. https://www.kaggle.com/datasets/datafiniti/vegetarian-vegan-restaurants/data
2. Seaborn для визуализации данных в Python https://pythonru.com/biblioteki/seaborn-plot
3. Уэса Маккинни «Питон для анализа данных» Hunter, J. D. (2007).
4. Изучаем Python: Т. 1, 2. (комплект из 2-х книг) | Лутц Марк.
5. «Прикладная статистика» (2004), А.И. Орлов http://www.aup.ru/books/m163/
6. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. "Прикладная статистика: Классификация и снижение размерности". https://www.ozon.ru/product/prikladnaya-statistika-klassifikatsiya-i-snizhenie-razmernosti-ayvazyan-agasi-semenovich-buhshtaber-1688732574/
7. Мюллер А., Гвидо С. "Введение в машинное обучение с помощью Python".
8. Кэмерон Дэвидсон-Пайлон "Вероятностное программирование на Python: байесовский вывод и вероятностные модели"
9. Статьи в газете NY https://www.nytimes.com/topic/subject/veganism?page=10
10. Веганство с точки зрения этики и морали - Селиванова Диана Игоревна https://cyberleninka.ru/article/n/veganstvo-s-tochki-zreniya-etiki-i-morali
11. Yeginsu, C. (2019, October 22). Is This the Start of the End of Meat? The New York Times. https://www.nytimes.com/2019/10/22/dining/veganism-vegetarianism-plant-based-meat.html

12. sklearn.cluster – https://scikit-learn.org/stable/api/sklearn.cluster.html
13. sklearn.metrics - https://scikit-learn.org/stable/api/sklearn.metrics.html
14. sklearn.decomposition - https://scikit-learn.org/stable/api/sklearn.decomposition.html
15. sklearn.ensemble - https://scikit-learn.org/stable/api/sklearn.ensemble.html
16. sklearn.preprocessing - https://scikit-learn.org/stable/api/sklearn.preprocessing.html
17. Данная работа расположена на сайте **https://github.com/Chernaya-Nataliya/Diplom**