тестового проекта Занятие 1

ИТ-программное решение для бизнеса

Независимый дизайнер тестового проекта: Рамин Мохаммаддуст

Независимый валидатор тестового проекта: Афшин Дехгани

Введение

Belle Croissant Lyonnais, известная пекарня-кондитерская в Лионе, собрала огромное количество данных о своих продажах, заказчиках и продукции. Эти данные содержат ценную информацию, которая может помочь в принятии обоснованных решений и способствовать дальнейшему успеху пекарни. На этом занятии вы будете использовать доступные в вашей производственной среде инструменты для анализа этих данных, выявления значимых закономерностей и тенденций и представления своих выводов в ясной и сжатой форме.

Ваш анализ будет сосредоточен на нескольких ключевых областях:

- **Эффективность продаж:** Оцените общую эффективность продаж пекарни, определив наиболее эффективные продукты, периоды пиковых продаж и потенциальные области для улучшения.
- **Поведение заказчиков:** Изучите предпочтения заказчиков, покупательские привычки и демографические данные, чтобы адаптировать маркетинговые стратегии и повысить удовлетворенность заказчиков.
- **Продуктовые тренды:** Анализируйте популярность продукта, выявляйте сезонные колебания и изучайте потенциальные предложения новых продуктов.
- **Операционная эффективность:** Оцените операционную эффективность, изучив такие факторы, как время выполнения заказов, оборачиваемость запасов и численность персонала.

Полученные результаты будут представлены руководству Belle Croissant Lyonnais, что позволит им оптимизировать операционную деятельность, увеличить продажи и улучшить качество обслуживания заказчиков.

Содержание

Данный учебный пакет содержит следующие материалы:

1. **Session Instructions (PDF):** Подробные инструкции с описанием задач, которые необходимо выполнить, и ожидаемых результатов для этого занятия.
2. **Данные о продажах (CSV-файлы):**

- sales_transactions.csv: Содержит данные на уровне транзакции, включая идентификатор транзакции, идентификатор заказчика, дату, идентификатор продукта, количество и цену.
- products.csv: Содержит информацию о продукте, включая идентификатор продукта, название, категорию, ингредиенты, цену, себестоимость, сезонный показатель, активный статус и дату выпуска.
- customers.csv: Содержит информацию о заказчике, включая идентификатор заказчика, имя, возраст, пол, почтовый индекс, адрес электронной почты, номер телефона, статус участника, дату присоединения, дату последней покупки, общие расходы, среднюю стоимость заказа, частоту, предпочитаемую категорию и статус оттока.

3. **Data Dictionary (PDF):** Подробные описания полей данных и их значений в каждом CSV-файле приведены в документе с инструкциями к занятию.
4. **Common Folder:** Эта папка содержит дополнительные ресурсы, такие как логотип, значки, руководство по стилю и другие элементы дизайна Belle Croissant Lyonnais, которые могут быть использованы при разработке приложения.
5. **Модели ARIMA (PDF):** Справочное руководство, объясняющее модель ARIMA (AutoRegressive Integrated Moving Average), ее реализацию и оценку для прогнозирования временных рядов.

Эти материалы предоставляют участникам все необходимые ресурсы для успешного анализа данных и составления отчетов.

Описание проекта и задач

На этом занятии вы проанализируете данные компании Belle Croissant Lyonnais, чтобы получить представление об их деятельности, заказчиках и эффективности продукции.

Методические рекомендации

1. **Простота в использовании:** Представление данных и аналитических данных в ясном и понятном формате.
2. **Привлекательный вид:** Следуйте руководству по стилю Belle Croissant Lyonnais для всех визуализаций и отчетов.
3. **Корректная работа:** Убедитесь, что все анализы и расчеты точны и безошибочны.
4. **Безопасность:** Обрабатывайте данные заказчиков конфиденциально и соблюдайте правила конфиденциальности данных.
5. **Своевременность:** Выполните все задания в течение указанного срока.

Технические особенности

1. **Очистка данных:** Устраните пропущенные значения, несоответствия и проблемы с форматированием в предоставленных наборах данных.
2. **Анализ данных:** Применяйте соответствующие статистические методы для анализа тенденций, прогнозирования и сегментации.
3. **Визуализация данных:** Создавайте четкие и информативные диаграммы и таблицы для представления полученных результатов.
4. **Моделирование:** Реализуйте прогнозирование временных рядов, кластеризацию и другие соответствующие алгоритмы.

Дополнительные факторы

- Анализ должен быть воспроизводимым и хорошо документированным.
- Используйте понятные надписи и пояснения для всех визуализаций и таблиц.
- Логически организуйте информацию, чтобы заинтересованным сторонам было проще ее понимать.

Инструкции для участника

1.1 Загрузка и изучение данных

Цель

Продемонстрировать свою способность загружать, проверять и понимать предоставленные наборы данных, выявляя потенциальные проблемы с качеством данных и подготавливая их для дальнейшего анализа.

Задачи

1. Загрузка данных:

- Импортируйте предоставленные CSV-файлы (sales_transactions.csv, products.csv и customers.csv) в выбранную вами среду анализа данных.

2. Первоначальное исследование:

- Отобразите первые 5 строк каждого фрейма данных, чтобы продемонстрировать структуру и содержимое.
- Определите типы данных для каждого столбца и определите нечисловые столбцы.
- Проверьте, нет ли пропущенных значений и несоответствий в данных.

Результаты:

- **Имя файла:** Session1_DataExploration.txt
- Предоставьте следующую информацию для каждого из трех CSV-файлов:
  - Типы данных для каждого столбца
  - Несоответствия и аномалии:
    - **Недопустимые даты:** Количество строк с датами, выходящими за пределы ожидаемого диапазона (например, "2023-14-01").
    - **Отрицательные значения: Количество строк с отрицательными количествами или ценами.**
    - **Недопустимые идентификаторы: количество строк с идентификаторами продуктов или заказчиков, которых нет в соответствующих файлах.**
    - **Неожиданные значения: Количество строк с неожиданными значениями в категориальных столбцах по отношению к предоставленному словарю данных.**
    - **Проблемы с форматированием: количество строк с дополнительными пробелами или несогласованное форматирование в соответствующих столбцах в соответствии с** предоставленным словарем данных.

1.2 Очистка и преобразование данных

Цель

Продемонстрируйте свою способность очищать, преобразовывать и стандартизировать данные для обеспечения точности, согласованности и пригодности для анализа.

Задачи

1. Пропущенные значения:

- Заполните пропущенные значения в столбце age файла customers.csv, указав средний возраст.
- Заполните пропущенные значения в столбце phone_number файла customers.csv значением "0".
- Заполните пропущенные значения в столбце promotion_id файла sales_transactions.csv значением "0".

2. Преобразование типов данных:

- Преобразуйте столбцы даты как в sales_transactions.csv, так и в customers.csv в тип данных datetime. Что касается времени, укажите случайное время с 9 утра до 5 вечера.

3. Стандартизация данных:

- Стандартизируйте номера телефонов в файле customers.csv, удалив все нечисловые символы, кроме "+" (пробелы, тире, круглые скобки).

Результаты

1. **Имя файла:** customers_cleaned.csv

- Тип файла: CSV-файл (.csv)

2. **Имя файла:** sales_transactions_cleaned.csv

- **Тип файла:** CSV-файл (.csv)

1.3 Анализ тенденций продаж

Цель

Рассчитать и визуализировать динамику продаж Belle Croissant Lyonnais в динамике по времени.

Задачи

1. Рассчитайте общий доход от продаж, количество транзакций и среднюю стоимость заказа за месяц.
2. Создайте линейные диаграммы для каждого из трех показателей в динамике по времени (ежемесячно).
3. Определите 3 месяца с наибольшим доходом от продаж и сведите их в таблицу.

Результаты:

- **Имя файла:** Session1_SalesTrends.pdf
- **Тип файла:** Отчет в формате PDF, содержащий:
    - Линейный график: Общий доход от продаж за месяц
    - Линейный график: Количество транзакций в месяц
    - Линейный график: Средняя стоимость заказа в месяц
    - Таблица: 3 лучших месяца по выручке от продаж (месяц, Общая выручка)

1.4 Анализ эксплуатационных характеристик продукта

Цель

Проанализируйте и визуализируйте показатели продаж продукции Belle Croissant Lyonnais.

Задачи

1. Рассчитайте общее количество проданных товаров и общую выручку по каждому продукту.
2. Рассчитайте норму прибыли для каждого продукта (цена - себестоимость).
3. Создайте столбчатую диаграмму, показывающую общий доход по каждой товарной категории ("Выпечка", "Хлеб", "Тарты").
4. Создайте таблицу, в которой представлены 3 самых продаваемых продукта по количеству проданных товаров, включая их названия, общее количество и общий доход.

Результаты

- **Имя файла:** Session1_ProductPerformance.pdf
- **Тип файла:** Отчет в формате PDF, содержащий:
    - Гистограмма: Общий доход по категориям продуктов
    - Таблица: Топ-3 самых продаваемых продукта (Название продукта, Общее количество проданных продуктов, Общая выручка)

1.5 Анализ заказчиков

Цель

Анализируйте и визуализируйте демографию заказчиков и участие в программах лояльности.

Задачи

1. Рассчитайте и визуализируйте с помощью гистограммы распределение заказчиков по возрастным группам (18-24, 25-34, 35-44, 45+).

2. Рассчитайте и отобразите в таблице распределение заказчиков по полу ("M", "F") в процентах.
3. Рассчитайте и отобразите в таблице средние расходы на одного заказчика для каждого уровня лояльности ("Basic", "Silver", "Gold").

Результаты

- **Имя файла:** Session1_CustomerAnalysis.pdf
- **Тип файла:** Отчет в формате PDF, содержащий:
  - Гистограмма: Распределение заказчиков по возрастным группам
  - Таблица: Процентное распределение заказчиков по полу
  - Таблица: Средние расходы на каждый уровень лояльности

1.6 Прогнозирование временных рядов

Цель

Спрогнозировать общий объем ежедневных продаж Belle Croissant Lyonnais на следующие 30 дней, используя модель прогнозирования временных рядов.

Задачи

1. Выберите и внедрите модель ARIMA, используя данные об общих ежедневных продажах из файла sales_transactions_cleaned.csv.
2. Составьте прогнозы продаж на следующие 30 дней.
3. Вычислите среднюю абсолютную погрешность (MAE) модели.

Результаты

- **Имя файла:** Session1_SalesForecast.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:**Date (ГГГГ-ММ-ДД)
  - **Столбец 2:** Predicted_Sales (с плавающей точкой)

1.7 Сегментация заказчиков и рекомендации

Цель

Продемонстрировать свою способность сегментировать заказчиков на основе их покупательского поведения и разработать базовую систему рекомендаций по продуктам для Belle Croissant Lyonnais.

Задачи

1. Сегментация заказчиков:

- **Разработка функционала:** Создайте два новых столбца в файле customers.csv:
  - total_purchases: Подсчитайте общее количество транзакций для каждого заказчика.
  - avg_purchase_value: Вычислите среднюю стоимость транзакции для каждого заказчика.
- **Кластеризация:** Используя столбцы total_purchases и avg_purchase_value, примените кластеризацию K-средних значений с 3 кластерами для сегментирования заказчиков.

2. Механизм рекомендаций:

- **Соответствие продукта:** Для каждого продукта определите 3 других продукта, которые чаще всего приобретаются вместе в рамках одной транзакции.
- **Рекомендации:** Для каждого заказчика порекомендуйте 3 лучших продукта, которые он еще не приобрел, основываясь на продуктах, которые часто покупают другие заказчики в своем сегменте.

Результаты

- **Имя файла:** Session5_Segmentation_and_Recommendations.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:** customer_id
  - **Столбец 2:** cluster_label (1, 2 или 3)
  - **Столбец 3:** recommended_product_1 (product_id)
  - **Столбец 4:** recommended_product_2 (product_id)
  - **Столбец 5:** recommended_product_3 (product_id)

1.8 Анализ характеристик продукта и оптимизация цен

Цель

Продемонстрировать свою способность анализировать характеристики продукта, выявлять тенденции ценообразования и предлагать корректировки цен на продукцию Belle Croissant Lyonnais на основе полученных данных.

Задачи

1. Анализ эксплуатационных характеристик продукта:

- **Объем продаж:** Рассчитайте общее проданное количество и общий доход, полученный по каждому продукту. Отсортируйте товары по общему доходу в порядке убывания.
- **Рентабельность:** Рассчитайте норму прибыли (прибыль/выручка) для каждого продукта и отсортируйте продукты по величине прибыли в порядке убывания.
- **Тенденции продаж:** Проанализируйте тенденции продаж каждого продукта в динамике по времени (ежемесячно). Определите любую сезонность или закономерности в продажах.

2. Анализ цен:

- **Чувствительность к цене:** Для каждого товара рассчитайте ценовую эластичность спроса (PED). PED измеряет, насколько чувствителен объем спроса на продукт к изменениям его цены. Используйте следующую формулу для расчета PED:
- PED = (% изменения требуемого количества) / (% изменения цены)
  Вы можете использовать простой расчет процентного изменения или более сложный метод, такой как логарифмическая регрессия.
- **Оптимизация цен:** На основе рассчитанных значений PED и нормы прибыли предложите оптимальную корректировку цен для каждого продукта. Учитывайте следующие рекомендации:
  - Если продукт имеет высокий PED (эластичный спрос), небольшое снижение цены может привести к значительному увеличению объема продаж и потенциально более высокому общему доходу.
  - Если продукт имеет низкий PED (неэластичный спрос), небольшое повышение цены может незначительно повлиять на объем продаж и привести к увеличению выручки.
  - Предлагая скорректировать цену, учитывайте маржу прибыли от продукта. Стремитесь максимизировать прибыль при сохранении или увеличении объема продаж.

Результаты

1. **Имя файла:** Session5_Product_Performance.csv

- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:** product_id
  - **Столбец 2:** total_quantity_sold
  - **Столбец 3:** total_revenue
  - **Столбец 4: profit_margin (рассчитывается как (total_revenue - total_cost) / total_revenue)**

2. **Имя файла:** Session5_Price_Analysis.csv

- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:** product_id
  - **Столбец 2:**price_elasticity_of_demand
  - **Столбец 3:** suggested_price_change (в процентах, например, увеличение на 5% или уменьшение на 3%)

Дополнительные примечания

- Ценовую эластичность спроса можно рассчитать различными методами. Выберите метод, который, по вашему мнению, наиболее подходит для данных.
- Помните, что оптимизация цен — это сложный процесс, который включает в себя множество факторов. Ваши предложения должны основываться на имеющихся данных и трезвом расчете.

1.9 Расчет жизненного цикла заказчика (CLTV)

Цель

Рассчитайте CLTV для каждого заказчика.

Задачи

1. Рассчитайте среднюю стоимость покупки для каждого заказчика из файла sales_transactions_cleaned.csv.
2. Рассчитайте частоту покупок (количество транзакций в месяц) для каждого заказчика.
3. Рассчитайте CLTV по формуле:

   CLTV = (Средняя стоимость покупки) * (Частота покупок) * 36

Результаты:

- **Имя файла:** Session1_CLTV.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:** customer_id (int)
  - **Столбец 2:** cltv (с плавающей точкой, округленной до 2 знаков после запятой)

1.10 Анализ оттока

Цель

Проанализировать и сравнить CLTV для постоянных и активных заказчиков.

Задачи

1. Идентифицируйте отток заказчиков из файла customers_cleaned.csv.
2. Рассчитайте общий уровень оттока (процент оттока заказчиков).
3. Рассчитайте средний CLTV для постоянных и активных заказчиков по отдельности.

Результаты:

- **Имя файла:** Session1_Churn_Analysis.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:** churn_rate (с плавающей точкой, в процентах, округленных до 2 знаков после запятой)
  - **Столбец 2:** avg_cltv_churned (число с плавающей точкой, округленное до 2 знаков после запятой)
  - **Столбец 3:** avg_cltv_active (значение с плавающей точкой, округленное до 2 знаков после запятой)