Test Project Session 6

IT Software Solution for Business

Введение

Национальная компания «КазМунайГаз» (КМГ), ведущее нефтегазовое предприятие в Казахстане, управляет обширными наборами данных, описывающими объемы производства, ассортимент продукции, экспортные операции и деятельность подрядчиков. Эти наборы данных содержат ценные сведения, которые поддерживают принятие решений на основе данных для повышения эффективности, доходов и управления партнерскими отношениями.

В этой сессии участники анализируют данные о производстве и продажах КМГ, чтобы выявить операционные закономерности, рыночные тенденции и факторы эффективности. Участники должны подготовить четкие, воспроизводимые отчеты и визуальные материалы в соответствии с аналитическими стандартами WorldSkills.

Ваш анализ будет сосредоточен на нескольких ключевых областях:

- **Эффективность производства:** Оценка общей эффективности производства и экспорта КМГ, выявление групп продуктов с высоким доходом и пиковых месяцев.
- **Поведение партнеров/подрядчиков:** Изучение типов партнеров, объемов контрактов и уровней активности для поддержки сегментации и анализа оттока.
- **Тенденции по продуктам:** Анализ динамики на уровне категорий (Сырая нефть, Продукты переработки, Газ/СУГ, Нефтехимия).
- **Операционная эффективность:** Оценка эффективности цепочки поставок и логистики посредством частоты транзакций и средней стоимости заказа.

Ваши выводы будут представлены руководству КМГ, предоставляя им действенные сведения для оптимизации операций, увеличения продаж и улучшения опыта партнеров.

Содержание

Пакет материалов для этой сессии содержит следующее:

1. **Инструкции к сессии (PDF):** Подробные инструкции, описывающие аналитические задачи, которые необходимо выполнить, и ожидаемые результаты для этой сессии.
2. **Операционные данные (файлы CSV):**
    - **sales_transactions.csv** – Содержит ежедневные операционные данные, включая ID транзакции, ID подрядчика, дату, ID продукта, количество, цену за единицу, метод оплаты и канал продаж.
    - **products_KMG.csv** – Содержит информацию о продуктах KMG, включая ID продукта, название продукта, категорию (Crude Oil, Refined Products, Gas/LPG, Petrochemicals), спецификации, цену, себестоимость, индикатор сезонности, статус активности и дату выпуска.
    - **partners_KMG.csv** – Содержит информацию о подрядчиках и партнерах, включая ID подрядчика, имя, регион, почтовый индекс, адрес электронной почты, номер телефона, уровень членства (Basic/Silver/Gold), дату присоединения, дату последней транзакции, общие расходы, среднюю стоимость заказа, частоту транзакций, предпочтительную категорию продукта и статус оттока.
3. **Словарь данных (PDF):** Подробные описания полей данных и их значений в каждом файле CSV содержатся в документе с инструкциями к сессии.
4. **Общая папка (Руководство):** Эта папка содержит дополнительные ресурсы, такие как логотип KMG, иконки, руководство по стилю и другие дизайнерские активы, которые можно использовать при разработке приложения.
5. **Модели ARIMA (PDF):** Справочное руководство, объясняющее модель ARIMA (авторегрессионная интегрированная модель скользящего среднего), ее реализацию и оценку для прогнозирования временных рядов.

Эти материалы предоставляют все необходимые ресурсы для успешного завершения сессии по анализу данных и составлению отчетов.

Описание Проекта и Задач

В рамках этой сессии вы будете анализировать данные Национальной компании «КазМунайГаз» (КМГ) , чтобы получить представление о ее производственных операциях, сети подрядчиков и эффективности продуктов.

Руководство (Guidelines)

1. **Простота использования (Easy to Use):** Представляйте данные и выводы в ясном, понятном формате.
2. **Эстетика (Looks Good):** Следуйте Корпоративному руководству по стилю KMG для всех визуализаций и отчетов.
3. **Качество работы (Works Well):** Обеспечьте точность и безошибочность всех анализов и расчетов.
4. **Безопасность (Secure):** Обрабатывайте данные подрядчиков конфиденциально и соблюдайте правила конфиденциальности данных.
5. **Своевременность (On Time):** Выполните все задачи в указанные сроки.

Технические Требования (Technical Considerations)

1. **Очистка данных (Data Cleaning):** Устраните пропущенные значения, несоответствия и проблемы форматирования в предоставленных наборах данных.
2. **Анализ данных (Data Analysis):** Применяйте соответствующие статистические методы и методы бизнес-анализа для определения производственных тенденций, сегментации подрядчиков и прогнозирования доходов.
3. **Визуализация данных (Data Visualization):** Создайте четкие, информативные профессиональные диаграммы и таблицы для представления выводов.
4. **Моделирование (Modeling):** Реализуйте прогнозирование временных рядов, кластеризацию и другие соответствующие алгоритмы.

Дополнительные Требования (Additional Considerations)

- Анализ должен быть воспроизводимым и хорошо задокументированным.
- Используйте четкие подписи и пояснения для всех визуализаций и таблиц.
- Организуйте информацию логически, чтобы облегчить ее понимание заинтересованными сторонами.
- Используйте ссылки на определения и правила данных из официального Словаря данных KMG (WSC2025_TP09_data_dictionary_en_KMG).

Инструкции для Участника

1.1 Загрузка и Исследование Данных (Data Loading and Exploration)

Цель

Продемонстрировать вашу способность загружать, проверять и понимать предоставленные наборы данных о производстве и продажах КМГ, выявляя потенциальные проблемы с качеством данных и подготавливая данные для дальнейшего анализа.

Задачи (Tasks)

1. **Загрузка данных (Load Data):** 2. Импортируйте предоставленные файлы CSV (sales_transactions.csv, products_KMG.csv и partners_KMG.csv) в выбранную вами среду для анализа данных.

2. **Первичное исследование (Initial Exploration):**

- Отобразите первые 5 строк каждого DataFrame, чтобы показать структуру и содержание.
- Определите типы данных каждого столбца и выявите нечисловые столбцы.
- Проверьте данные на наличие пропущенных значений и несоответствий.

Результаты (Deliverables)

- **Имя файла:** Session6_DataExploration.txt
- **Содержание:** Предоставьте следующую информацию для каждого из трех файлов CSV:
  - Типы данных каждого столбца.
  - **Несоответствия и Аномалии (Inconsistencies and Anomalies):**
    - **Неверные даты (Invalid Dates):** Количество строк с датами вне ожидаемого диапазона (например, "2023-14-01").
    - **Отрицательные значения (Negative Values):** Количество строк с отрицательным количеством (quantity) или ценами (price).
    - **Неверные ID (Invalid IDs):** Количество строк с ID продуктов или ID партнеров, которые не существуют в соответствующих файлах.
    - **Неожиданные значения (Unexpected Values):** Количество строк с неожиданными значениями в категориальных столбцах по отношению к предоставленному словарю данных.
    - **Проблемы форматирования (Formatting Issues):** Количество строк с лишними пробелами или несогласованным форматированием в соответствующих столбцах по отношению к предоставленному словарю данных.

1.2 Очистка и Преобразование Данных

Цель

Продемонстрировать вашу способность очищать, преобразовывать и стандартизировать данные для обеспечения точности, согласованности и пригодности для анализа.

Задачи

1. **Пропущенные значения (Missing Values):**
   - Заполнить пропущенные значения в столбце age файла partners_KMG.csv медианным возрастом всех контактных лиц партнеров.
   - Заполнить пропущенные значения в столбце phone_number файла partners_KMG.csv значением '0'.
   - Заполнить пропущенные значения в столбце promotion_id файла sales_transactions.csv значением '0'.
2. **Преобразование типов данных (Data Type Conversion):**
   - Преобразовать столбцы с датами в файлах sales_transactions.csv и partners_KMG.csv в тип данных datetime. Для времени добавьте случайное время между 9:00 и 17:00.
3. **Стандартизация данных (Data Standardization):**
   - Стандартизировать номера телефонов в файле partners_KMG.csv, удалив все нечисловые символы, кроме + (пробелы, тире, скобки).

Результаты (Deliverables)

1. **Имя файла:** partners_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 тенденции производства и продаж Национальной компании «КазМунайГаз» (КМГ) с течением времени.

Задачи

1. Рассчитать общую выручку от продаж, количество транзакций и среднюю стоимость заказа за месяц.
2. Создать линейные графики для каждой из трех метрик с течением времени (по месяцам).
3. Определите топ-3 месяца с самой высокой общей выручкой от продаж и отобразите их в таблице.

Результаты:

- **Имя файла:** Session6_SalesTrends.pdf
- **Тип файла:** Отчет в формате PDF, содержащий:
    - Линейный график: Общая выручка от продаж за месяц
    - Линейный график: Количество транзакций за месяц
    - Линейный график: Средняя стоимость заказа за месяц
    - Таблица: Топ-3 месяца по общей выручке от продаж (Месяц, Общая выручка)

1.4 Анализ Эффективности Продуктов (Product Performance Analysis)

Цель

Проанализировать и визуализировать эффективность продаж нефти, газа и нефтехимической продукции Национальной компании «КазМунайГаз» (КМГ).

Задачи

1. Рассчитать общее количество проданного и общую выручку для каждого продукта КМГ.
2. Рассчитать маржу прибыли для каждого продукта (Цена - Себестоимость).
3. Создать столбчатую диаграмму, показывающую общую выручку для каждой категории продуктов («Crude Oil» (Сырая нефть), «Refined Products» (Продукты переработки), «Gas/LPG» (Газ/СУГ), «Petrochemicals» (Нефтехимия)).
4. Создать таблицу, показывающую топ-3 самых продаваемых продуктов по количеству, включая их названия, общее количество и общую выручку.

Результаты

- **Имя файла:** Session6_ProductPerformance.pdf
- **Тип файла:** Отчет в формате PDF, содержащий:
    - Столбчатая диаграмма: Общая выручка по категории продуктов.
    - Таблица: Топ-3 самых продаваемых продуктов (Название продукта, Общее количество проданного, Общая выручка).

1.5 Анализ Партнеров и Подрядчиков (Partner and Contractor Analysis)

Цель

Проанализировать и визуализировать демографические данные и уровни вовлеченности партнеров и подрядчиков Национальной компании «КазМунайГаз» (КМГ).

Tasks

1. Рассчитать и визуализировать распределение **возрастных групп** партнеров (18-24, 25-34, 35-44, 45+) с помощью **столбчатой диаграммы**, используя поле **age** из файла partners_KMG.csv.
2. Рассчитать и отобразить распределение **пола** партнеров («M», «F») в **процентах** в таблице.
3. Рассчитать и отобразить **средние общие расходы** по уровню партнерства («Basic» (Базовый), «Silver» (Серебряный), «Gold» (Золотой)) в таблице, используя поля **membership_status** и **total_spending** из файла partners_KMG.csv.

Результаты:

- **Имя файла:** Session6_PartnerAnalysis.pdf
- **Тип файла:**Отчет в формате PDF, содержащий:
  - Столбчатая диаграмма: Распределение возрастных групп партнеров
  - Таблица: Процентное распределение пола партнеров
  - Таблица: Средние общие расходы по уровню партнерства

1.6 Прогнозирование Объемов Производства (Временные ряды)

Цель

Спрогнозировать ежедневный общий объем производства или продаж Национальной компании «КазМунайГаз» (КМГ) на следующие 30 дней, используя модель прогнозирования временных рядов.

Задачи

1. Выбрать и реализовать модель **ARIMA** (Авторегрессия — интегрированное скользящее среднее), используя ежедневные данные об общих продажах из файла sales_transactions_cleaned.csv.
2. Сгенерировать прогнозы объема производства или продаж на следующие **30 дней**.
3. Рассчитать **Среднюю абсолютную ошибку (MAE)** модели для оценки точности прогноза.

Результаты

- **Имя файла:** Session6_SalesForecast.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - Столбец 1: Дата (ГГГГ-ММ-ДД)
  - Столбец 2: Прогнозируемые_Продажи (Predicted_Sales) (число с плавающей точкой)

1.7 Сегментация Партнеров и Рекомендации

Цель

Продемонстрировать вашу способность сегментировать партнеров и подрядчиков Национальной компании «КазМунайГаз» (КМГ) на основе их покупательского и контрактного поведения, а также разработать базовую систему рекомендаций продуктов.

Задачи

1. **Сегментация партнеров (Partner Segmentation):**
   - **Инженерия признаков (Feature Engineering):** Создать два новых столбца в файле partners_KMG.csv:
     - total_purchases (общее количество покупок): Рассчитать общее число транзакций для каждого партнера на основе файла sales_transactions.csv.

- avg_purchase_value (средняя стоимость покупки): Рассчитать среднюю стоимость транзакции для каждого партнера.
- **Кластеризация (Clustering):** Используя столбцы total_purchases и avg_purchase_value, применить кластеризацию K-means с 3 кластерами для сегментации партнеров на поведенческие группы (например, подрядчики с малым, средним и большим объемом).

2. **Движок рекомендаций (Recommendation Engine):**
   - **Аффинность продуктов (Product Affinity):** Для каждого продукта в products_KMG.csv определить топ-3 других продукта, наиболее часто покупаемых совместно в рамках одной транзакции из sales_transactions.csv.
   - **Рекомендации (Recommendations):** Для каждого партнера рекомендовать топ-3 продукта, которые он еще не приобрел, на основе продуктов, часто покупаемых другими партнерами в том же кластере.

Результаты

- **Имя файла:** Session6_Segmentation_and_Recommendations.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
  - **Столбец 1:** partner_id
  - **Столбец 2:** cluster_label (1, 2, или 3)
  - **Столбец 3:** recommended_product_1 (ID продукта)
  - **Столбец 4:** recommended_product_2 (ID продукта)
  - **Столбец 5:** recommended_product_3 (ID продукта)

1.8 Анализ Эффективности Продуктов и Оптимизация Цен

Цель

Продемонстрировать вашу способность анализировать эффективность продуктов Национальной компании «КазМунайГаз» (КМГ), определять тенденции ценообразования и предлагать основанные на данных корректировки цен для нефти, газа и нефтехимической продукции.

Задачи

1. **Анализ эффективности продуктов (Product Performance Analysis):**
   - **Объем продаж (Sales Volume):** Рассчитать общее количество проданного и общую выручку, сгенерированную для каждого продукта. Отсортировать продукты по общей выручке в порядке убывания.
   - **Прибыльность (Profitability):** Рассчитать маржу прибыли (прибыль/выручка) для каждого продукта и отсортировать продукты по марже прибыли в порядке убывания.
   - **Тенденции продаж (Sales Trends):** Проанализировать ежемесячные тенденции производства и продаж для каждой категории продуктов (Сырая нефть, Продукты переработки, Газ/СУГ, Нефтехимия). Определить любую сезонность или повторяющиеся закономерности спроса.
2. **Анализ цен (Price Analysis):**
   - **Чувствительность к цене (Price Sensitivity):** Для каждого продукта рассчитать ценовую эластичность спроса (ЦЭС). ЦЭС измеряет, насколько чувствительно требуемое количество продукта реагирует на изменения его цены. Используйте следующую формулу для расчета ЦЭС:
     - ЦЭС = (изменение спроса в %) / (изменение цены в %). Вы можете использовать простой расчет процентного изменения или более сложный метод, например, лог-лог регрессию.
   - **Оптимизация цен (Price Optimization):** На основе рассчитанных значений ЦЭС и маржи прибыли предложите оптимальные корректировки цен для каждого продукта. Учитывайте следующие рекомендации:
     - Если продукт имеет высокую ЦЭС (эластичный спрос) — небольшое снижение цены может значительно увеличить объем продаж и общую выручку.
     - Если продукт имеет низкую ЦЭС (неэластичный спрос) — небольшое повышение цены может мало повлиять на объем, но увеличить прибыль.

- Всегда учитывайте маржу прибыли продукта при предложении корректировок цен, чтобы максимизировать прибыльность при сохранении или росте продаж.

Результаты

1. **Имя файла:** Session6_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:** Session6_Price_Analysis.csv
    - **Тип файла:** CSV-файл (.csv)
    - **Формат:**
        - Столбец 1: product_id
        - Столбец 2: price_elasticity_of_demand (ценовая эластичность спроса)
        - Столбец 3: suggested_price_change (предлагаемое изменение цены) (в процентах, например, 5% увеличение или -3% снижение)

Дополнительные Примечания (Additional Notes)

- Ценовая эластичность спроса может быть рассчитана с использованием различных методов. Выберите метод, который вы считаете наиболее подходящим для данных.
- Помните, что оптимизация цен — это сложный процесс, включающий множество факторов. Ваши предложения должны основываться на доступных данных и вашем наилучшем суждении.

1.9 Расчет Пожизненной Ценности Партнера (PLTV)

Цель

Рассчитать **Пожизненную Ценность Партнера (PLTV)** для каждого партнера или подрядчика Национальной компании **«КазМунайГаз» (КМГ)** на основе их истории транзакций и данных о вовлеченности.

Задачи

1. Рассчитать среднюю стоимость покупки на одного партнера, используя данные из файла sales_transactions_cleaned.csv.

2. Рассчитать частоту покупок (количество транзакций в месяц) для каждого партнера.

3. Рассчитать PLTV с использованием следующей формулы:

$$\text{PLTV} = (\text{Средняя стоимость покупки}) \times (\text{Частота покупок}) \times 36$$

Результаты:

- **Имя файла:** Session6_CLTV.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
    - Столбец 1: partner_id (целое число)
    - Столбец 2: pltv (число с плавающей точкой, округленное до 2 десятичных знаков)

1.10 Анализ Оттока Партнеров

Цель

Проанализировать и сравнить Пожизненную Ценность Партнера (PLTV) для отточных и активных

партнеров Национальной компании «КазМунайГаз» (КМГ).

Задачи

1. Идентифицировать отточных партнеров из файла partners_cleaned.csv (где churned = TRUE).
2. Рассчитать общий коэффициент оттока (процент партнеров, помеченных как отточные).
3. Рассчитать средний PLTV для отточных и активных партнеров отдельно.

Результат:

- **Имя файла:** Session6_Churn_Analysis.csv
- **Тип файла:** CSV-файл (.csv)
- **Формат:**
    - Столбец 1: churn_rate (число с плавающей точкой, процент, округленный до 2 десятичных знаков)
    - Столбец 2: avg_pltv_churned (число с плавающей точкой, округленное до 2 десятичных знаков)
    - Столбец 3: avg_pltv_active (число с плавающей точкой, округленное до 2 десятичных знаков)