**TEAM 1J:**

Bangshi Zeng

Harini Sarankapani

Jyotiraditya Guleria

Manchen Ding

Yishi Qian

**CIS 467 final group project (due by Wednesday, March 5 at 11:59 PM).**

**This is a group project (total 300 points). The groups have been created on Blackboard. Please make only one submission per group and put all your team members’ full names into this Word document and submit this Word document. Please also submit a Tableau Workbook file .twb into the Final Project folder on Blackboard together with the Excel file of your Data Warehouse which you uploaded to Tableau and used for visualizations.**

The script files sakila-schema.sql and sakila-data.sql create a database which contains tables (the database schema is below), with transactional data related to some company operations.

**Please check early that you can create the sakila database on your machine. First, run this code - sakila-schema.sql, and second, run this code - sakila-data.sql.**

**Very Important! All three parts of the final project should be on One topic/subject of the data warehouse. For example, if you decide to track customers as your topic/subject, part 1 (Data Warehouse), part 2 (Queries) and part 3 (visualizations) should only be related to customers and should NOT include any other topics.**

**If you use Chat GPT, please use the “Share” button (looks like ‘upward arrow’) in the right corner of ChatGPT chat, and ‘copy link’ and share the link to that chat in this Word document and briefly explain how you used it for your Final Project (for each question if you used it). No points will be taken off for using ChatGPT (it is allowed to use it for Final Project) but you are required to share the link to a chat if you used it.**

**More information on how to share a chat here:** <https://help.openai.com/en/articles/7925741-chatgpt-shared-links-faq>

A computer screen shot of a diagram

AI-generated content may be incorrect.

Please put all your work into **this** **single Word doc and also submit a Tableau Workbook file .twb and the Excel file of your Data Warehouse that you used for Tableau visualizations**. Please see instructions for Tableau below in question 3.

1. **(121 points)** Design and create a data warehouse for the provided database. The decisions about which fields to include and how to aggregate the data are left to you. You do not need to include every single data point from the tables given. Use your judgement as to what will be interesting/useful for the organization. But please make sure that you pull (combine) data from **at least six tables** and compute relevant aggregate statistics. Please compute relevant aggregate statistics for each table that you join. **In your queries later in part 2, you may join your Data Warehouse with other tables to answer useful questions**. Please see many examples from class lectures and you may adapt those codes for your purpose (for this dataset).

**Submit a screenshot of the first 25 rows of your data warehouse (paste into this Word document) and the SQL code that you used to create it. Please copy and paste your SQL code into this Word document. If your PC does not show 25 rows of data, please submit what you have (i.e., rows you can see on a screenshot) with a comment that you cannot show 25 rows of data. Please add a full description of what your Data Warehouse will be tracking for a company. Please treat this assignment as a business case. So, the more you describe the better. Please also create an Excel file (Export from MySQL) of your data warehouse and use it for part 3 – Tableau visualizations.**

**Description:**

The **Customer Data Warehouse** is designed to centralize and analyze customer interactions, spending behavior, and rental preferences for a movie rental business. This warehouse aggregates data from multiple operational tables in the **Sakila** database to provide valuable insights into customer engagement, store performance, and revenue generation.

By structuring the data warehouse around customers, the business can make **data-driven decisions** regarding **loyalty programs, personalized recommendations, inventory management, and marketing strategies.**

**Query:**

-- Creating a Customer-Centric Data Warehouse in MySQL

-- This script builds a customer data warehouse by aggregating rentals, payments, and customer preferences.

-- Drop and recreate the data warehouse database

DROP DATABASE IF EXISTS sakila\_dw;

CREATE DATABASE sakila\_dw;

USE sakila\_dw;

DROP TABLE IF EXISTS customer\_data\_warehouse;

CREATE TABLE customer\_data\_warehouse AS

SELECT

c.customer\_id,

CONCAT(c.first\_name, ' ', c.last\_name) AS customer\_name,

c.email AS customer\_email,

c.create\_date AS customer\_since,

-- Calculating Total Rentals and Total Payments per Customer

(SELECT COUNT(\*) FROM sakila.rental r WHERE r.customer\_id = c.customer\_id) AS total\_rentals,

(SELECT COUNT(\*) FROM sakila.payment p WHERE p.customer\_id = c.customer\_id) AS total\_payments,

(SELECT COALESCE(SUM(p.amount), 0) FROM sakila.payment p WHERE p.customer\_id = c.customer\_id) AS total\_spent,

-- Determining First and Last Rental Dates

(SELECT MAX(r.rental\_date) FROM sakila.rental r WHERE r.customer\_id = c.customer\_id) AS last\_rental\_date,

(SELECT MIN(r.rental\_date) FROM sakila.rental r WHERE r.customer\_id = c.customer\_id) AS first\_rental\_date,

-- Identifying the Most Rented Category per Customer

(SELECT cat.name

FROM sakila.rental r2

JOIN sakila.inventory i ON r2.inventory\_id = i.inventory\_id

JOIN sakila.film f ON i.film\_id = f.film\_id

JOIN sakila.film\_category fc ON f.film\_id = fc.film\_id

JOIN sakila.category cat ON fc.category\_id = cat.category\_id

WHERE r2.customer\_id = c.customer\_id

GROUP BY cat.category\_id, cat.name

ORDER BY COUNT(fc.film\_id) DESC, cat.name ASC

LIMIT 1) AS favorite\_category,

-- Extracting Customer's Home Store and Assigned Staff

st.store\_id AS home\_store,

s.staff\_id AS primary\_staff,

CONCAT(s.first\_name, ' ', s.last\_name) AS staff\_name

FROM sakila.customer c

LEFT JOIN sakila.store st ON c.store\_id = st.store\_id

LEFT JOIN sakila.staff s ON st.manager\_staff\_id = s.staff\_id

-- Sorting Customers by Total Spending

ORDER BY total\_spent DESC;

-- Verify Data

SELECT \* FROM customer\_data\_warehouse;

**Snips of the first 25 rows of the Warehouse:**

**A black screen with white text

Description automatically generatedA table with numbers and names

Description automatically generated**

2. **(104 points)** Create **eight** SQL queries **on your data warehouse** (not on the original dataset) that answer interesting questions. At least **6** queries should be more complex queries. For example, more complex queries could include Joins, a Group By, UNION elements or a subquery or use some aggregate functions and summary calculations and conditional logic codes (see examples in the class lectures’ slides). **If needed, you may join your Data Warehouse with other tables (which are not a part of Data Warehouse) to answer useful questions.**

**Submit a copy of each query SQL code (paste into this Word document), and the screenshot of each query results (or the first 25 rows if it is longer or how many rows you can get on your PC) and full description of the question your SQL code was addressing and what you found in the results. The question that each query answers should be useful for a company to make decisions and act upon.**

1. **Purpose:** Identify Customers with the Highest Total Spending

**SQL Query:**

SELECT

customer\_id,

customer\_name,

total\_spent,

total\_rentals

FROM customer\_data\_warehouse

ORDER BY total\_spent DESC

LIMIT 10;

**Output:**

**A screenshot of a black screen

Description automatically generated**

**Insight:** This query identifies the **top 10 highest-spending customers**, helping the company focus on **loyalty programs and retention strategies**. The highest spender has **spent $221.55**, indicating strong engagement. The company can **offer personalized promotions** to retain these high-value customers and maximize revenue.

**2. Purpose:** ToAnalyze the Most Popular Movie Categories by Revenue

**SQL Query:**

SELECT

cat.name AS category\_name,

COUNT(DISTINCT r.rental\_id) AS total\_rentals,

SUM(p.amount) AS total\_revenue

FROM customer\_data\_warehouse cdw

JOIN sakila.rental r ON cdw.customer\_id = r.customer\_id

JOIN sakila.payment p ON r.rental\_id = p.rental\_id

JOIN sakila.inventory i ON r.inventory\_id = i.inventory\_id

JOIN sakila.film f ON i.film\_id = f.film\_id

JOIN sakila.film\_category fc ON f.film\_id = fc.film\_id

JOIN sakila.category cat ON fc.category\_id = cat.category\_id

GROUP BY cat.name

ORDER BY total\_revenue DESC

LIMIT 5;

**Output:**

**A screenshot of a data

Description automatically generated**

**Insight:** This query identifies the most popular movie categories by revenue, helping the company optimize inventory and marketing strategies. **Sports** is the highest-grossing category with **$5314.21** in revenue, followed by **Sci-Fi and Animation**. The high rental volume in these genres indicates strong customer demand, suggesting the company should **prioritize these categories, offer targeted promotions, and adjust inventory** to maximize profitability.

**3. Purpose:** Evaluate store performance by total customer spending

**SQL Query:**

SELECT

st.store\_id,

COUNT(DISTINCT cd.customer\_id) AS total\_customers,

SUM(cd.total\_spent) AS total\_revenue

FROM customer\_data\_warehouse cd

JOIN sakila.store st ON cd.home\_store = st.store\_id

GROUP BY st.store\_id

ORDER BY total\_revenue DESC;

**Output:**

**A screenshot of a black and white screen

Description automatically generated**

**Insight:** This query evaluates **store performance based on total customer spending**, helping the company optimize resource allocation. **Store 1 generates the highest revenue at $36,997.53**, serving **326 customers**, while **Store 2 follows with $30,409.03** from **273 customers.** The company can **invest more in high-performing stores and implement strategies to boost revenue in lower-performing locations.**

**4. Purpose:** Identifying Inactive Customers Who Haven’t Rented in the Last Three Months

**SQL Query:**

SELECT

customer\_id,

customer\_name,

total\_rentals,

last\_rental\_date

FROM customer\_data\_warehouse

WHERE DATEDIFF(NOW(), last\_rental\_date) > 90

ORDER BY last\_rental\_date ASC;

**Output:A table with numbers and names

Description automatically generated**

**Insight:** This query identifies **inactive customers who haven’t rented in over three months**, helping the company detect potential churn. Customers like **Herbert Kruger and Jose Andrew** last rented in August 2005, indicating a drop in engagement. The company can **offer personalized discounts or promotional campaigns** to re-engage these inactive customers and boost retention.

## 5. Purpose: Determine the average rental frequency per customer

## SQL Query:

SELECT

    ROUND(AVG(total\_rentals), 2) AS avg\_rental\_frequency\_per\_customer

FROM customer\_data\_warehouse;

**Output:**

A screen shot of a computer

Description automatically generated

**Insight:** On average, customers rent movies every 10-15 days. High-frequency renters could be targeted for VIP memberships, while low-frequency renters might benefit from promotions.

## 6. Purpose: Find out which employees handle the most customers

## SQL Query:

SELECT

    staff\_name,

    COUNT(customer\_id) AS managed\_customers

FROM customer\_data\_warehouse

GROUP BY staff\_name

ORDER BY managed\_customers DESC;

**Output:**

A screenshot of a computer

Description automatically generated

**Insight:** Certain employees manage significantly more customers, indicating possible workload imbalance. Redistributing customer management duties might improve service efficiency.

1. **Purpose:** Identify the top 5 countries with the highest number of customers

**SQL Query:**

SELECT

    co.country AS country\_name,

    COUNT(cd.customer\_id) AS total\_customers

FROM customer\_data\_warehouse cd

JOIN sakila.customer c ON cd.customer\_id = c.customer\_id

JOIN sakila.address a ON c.address\_id = a.address\_id

JOIN sakila.city ci ON a.city\_id = ci.city\_id

JOIN sakila.country co ON ci.country\_id = co.country\_id

GROUP BY country\_name

ORDER BY total\_customers DESC

LIMIT 5;

**Output:**

A screenshot of a computer

Description automatically generated

**Insight:** Countries with the highest customer counts should be prioritized for regional marketing strategies and potential store expansions.

1. **Purpose:** Calculate customer spending as a percentage of total revenue

**SQL Query:**

SELECT

    customer\_id,

    customer\_name,

    total\_spent,

    ROUND((total\_spent / (SELECT SUM(total\_spent) FROM customer\_data\_warehouse)) \* 100, 2) AS percentage\_of\_total\_revenue

FROM customer\_data\_warehouse

ORDER BY percentage\_of\_total\_revenue DESC;

**Output:**

A screenshot of a data

Description automatically generated

**Insights:** A small group of customers contributes a significant portion of the total revenue, reinforcing the importance of loyalty programs and personalized marketing.

3. (**75 points**) Create **five** Tableau individual visualizations (graphs) **on your data warehouse** (**plus one dashboard** as discussed below) with valuable information to present findings to senior management of the company. Save each visualization as a png file (as I show in class, and we will also practice in the lab 5) and paste each individual visualization png file **into this Word** document with the full explanation of what the visualizations show, how they are useful to a company and how company management could make decisions based on what you show. Finally, combine those **five** visualizations into one **Dashboard** (as I show in class, and we will also practice in the lab 5), and save this Dashboard as a png file and **paste the Dashboard into this Word** document.

**Important! You might visualize some queries from part 2 or create completely new visualizations. If for queries (in part 2), you used data warehouse together with (joining) some additional tables, you need to create a new source Excel file for Tableau, and submit it as described below in the next paragraph (please explain in your submission what source you used). Alternatively, you may do visualizations only on Data Warehouse Excel file (from part 1) and submit that file as a source file for Tableau (see below). In general, we need one Excel source file for Tableau to check your visualizations and the Tableau workbook (see below).**

**Please also save the whole Tableau project as a Tableau Workbook file .twb (In Tableau use File - Save as) and submit to the Final Project folder on Blackboard together with this Word document and together with the Excel file of your Data Warehouse which you uploaded to Tableau and used for visualizations.**

Q3

**1. Trend of Inactive Customers Over the Past 90 Days**

**Description:**

This line chart visualizes the trend of customers who have not rented in the past 90 days. The x-axis represents time, and the y-axis shows the number of inactive customers.

**Business Value:**

This visualization helps **track customer churn trends**, identifying periods with high churn rates to take timely action. By analyzing churn patterns, the company can improve customer retention, enhance marketing strategies, and assess the impact of churn on overall revenue.

* **Monitor churn trends** to develop targeted recovery strategies.
* **Improve retention efforts** through personalized marketing.
* **Evaluate revenue impact** by quantifying customer churn.

**Decision Making:**

To reduce customer churn, the company can implement:

·**Targeted Marketing** – Send **personalized recommendations and discounts** to re-engage customers.  
 **·Data-Driven Analysis** – Identify churn reasons (e.g., content gaps, competition) and optimize services.  
 ·**Promotional Incentives** – Offer **limited-time deals** or **loyalty rewards** to retain users.

These strategies help minimize churn, boost customer loyalty, and improve long-term revenue growth.

图表, 折线图

AI 生成的内容可能不正确。

**2. Most Popular Movie Categories by Spending**

**Description:**

This bar chart showcases the most popular movie categories based on total spending. The y-axis represents total spending, and the x-axis lists different movie categories.

**Business Value:**

This visualization helps **identify the best-selling movie categories**, allowing the company to optimize inventory and purchasing strategies. By analyzing customer preferences, the company can refine its marketing efforts and improve personalized recommendations. Additionally, data insights can guide **content investment decisions**, ensuring better allocation of resources and maximizing return on investment (ROI).

* **Optimize inventory and purchasing** by identifying high-demand categories and adjusting stock levels.
* **Enhance marketing and recommendations** by understanding customer preferences and targeting promotions more effectively.
* **Support content investment decisions** to allocate resources strategically and increase overall profitability.

**Decision Making:**

To improve its efficiency, the company can implement:

**·Inventory Optimization** – Increase the supply of popular movie categories while reducing stock for less popular ones.  
 ·**Pricing Strategy Adjustments** – Raise rental prices for high-demand genres to maximize profitability.  
 ·**Personalized Recommendations** – Use customer preference data to improve recommendation algorithms, ensuring users receive content that matches their interests, boosting engagement and conversion rates.

图表, 条形图

AI 生成的内容可能不正确。

**3. Customer Geographic Distribution**

**Description:**

This map visualization displays customer distribution across different countries, with shading intensity representing customer count.

**Business Value:**

This visualization helps **analyze the geographical distribution of customers**, allowing the company to focus on high-potential regions and improve market expansion strategies. It also **identifies new market opportunities**, supporting international growth by adapting marketing strategies to regional consumption patterns.

* **Optimize market expansion** by identifying key high-density areas and adjusting resource allocation.
* **Discover growth opportunities** to support international market expansion.
* **Analyze regional behavior** to implement localized marketing strategies and enhance engagement.

**Decision Making:**

To align business strategy with regional demand, the company can implement:

**·Increase Regional Marketing Efforts** – Invest in high-density areas through social media campaigns and local promotions.  
 · **Consider International Expansion** – If a country has a large customer base, evaluate the feasibility of establishing localized operations.  
 · **Enhance Localized Recommendations** – Use regional data and customer preferences to improve recommendations and boost engagement.

By applying these strategies, the company can **expand market reach, increase customer engagement, and drive global business growth**.

地图

AI 生成的内容可能不正确。

**4. Total Spending Based On Top 5 Country**

**Description:**

This chart shows the total spending distribution of the top 5 countries, with India (28.43%) and China (24.87%) contributing the most, followed by the United States (17.63%), Japan (14.89%), and Mexico (14.18%), reflecting the company's key revenue markets.

**Business Value:**

This visualization displays **the global distribution of customers**, combined with **the top 5 highest-spending countries**, helping the company identify key markets. By analyzing customer concentration and revenue contribution, the company can better allocate resources and strengthen its market position.

* **Enhanced Market Focus** – The geographic distribution map shows overall customer locations, while the spending chart pinpoints the top 5 revenue-generating countries, allowing for precise market targeting.
* **Optimized Marketing & Promotions** – Understanding customer spending behavior in these key markets enables the company to design more effective, data-driven marketing campaigns.
* **Strategic Global Expansion** – The data helps determine whether to reinforce operations in existing high-revenue regions or explore new market opportunities for future growth.

**Decision Making:**

Based on regional customer distribution and top 5 spending markets, the company can implement the following strategies:

·Targeted Market Investment – Increase ad spending and localized promotions in the top 5 highest-revenue countries to boost brand presence.  
·Operational Expansion – Strengthen physical and digital operations in these key markets, such as adding retail locations or expanding streaming services.  
· Content & Product Optimization – Tailor content recommendations and movie availability based on regional preferences, ensuring higher engagement and customer retention.

By leveraging insights from both geographic distribution and revenue analysis, the company can make more informed decisions, optimize marketing and operational strategies, and drive sustainable long-term growth.

图表, 饼图

AI 生成的内容可能不正确。

**5. Top 10 Highest-Spending Customers**

**Description:**

This bar chart displays the top 10 customers based on total spending. The y-axis represents the total amount spent. The x-axis lists the names of the highest-spending customers.

**Business Value:**

This visualization helps identify **customers who contribute the highest revenue** to the company. Understanding the spending patterns of these high-value customers allows the business to develop targeted strategies to improve customer retention and engagement.

* **Identify Loyal Customers**: Recognizing customers with high spending habits enables the company to focus on customer satisfaction and long-term retention.
* **Optimize Customer Relationship Management (CRM)**: By analyzing spending trends, the company can create personalized offers to encourage repeat rentals.
* **Discover Growth Opportunities**: Understanding high-value customer behavior helps the company refine its marketing strategies and improve product offerings.

**Decision Making:**

Based on the analysis, the company can take the following actions to enhance customer engagement and **increase repeat transactions**:

**Offer Exclusive Discounts**

* Introduce **special promotions or loyalty rewards** for top-spending customers.
* Implement a **points-based system**, where frequent renters accumulate rewards for discounts or free rentals.

**Personalized Recommendations**

* Use **historical rental data** to suggest movies tailored to customer preferences.
* Adjust **inventory planning** to ensure that frequently rented movie categories remain well-stocked.

**Develop a VIP Membership Program**

* Introduce **tiered membership levels**, providing exclusive benefits to high-spending customers.
* Offer perks such as **early access to new movies** or **discounted rental rates** to encourage long-term engagement.

图表, 条形图

AI 生成的内容可能不正确。

**6.**  Top 10 Customers by Total Rentals

**Description:**

This visualization highlights the top 10 customers based on the number of rentals. The y-axis shows the total number of rentals, and the x-axis lists customer names.

**Business Value:**

1. **Identifies highly engaged customers** who frequently rent movies, highlighting a key customer segment for retention efforts.
2. **Frequent renters contribute significantly** to revenue and engagement, making them ideal candidates for loyalty programs.
3. **Understanding rental behavior** helps optimize movie inventory by ensuring availability of popular titles.
4. **Targeting high-frequency renters** with exclusive promotions can further increase rental activity and overall spending.

**Decision Making:**

1. **Introduce a VIP Membership Program** with exclusive benefits such as unlimited rentals, early access to new releases, or loyalty rewards.
2. **Personalized Engagement Strategies:** Offer tailored promotions and rental discounts to high-frequency renters to encourage repeat business.
3. **Optimize Inventory Management** based on rental trends, ensuring the most-rented categories remain well-stocked.
4. **Develop Subscription-Based Offers** for frequent renters like Eleanor Hunt and Karl Seal, such as fixed monthly rental plans to increase long-term engagement.

.

**图表, 条形图

AI 生成的内容可能不正确。**

**7. Top 10 Highest Spendings and Rentals Customers**

图表, 条形图

AI 生成的内容可能不正确。

**Description:**

**Top Spending Customers:**

* The highest-spending customer, **Karl Seal**, has spent **$221.55**.
* Following closely behind are **Eleanor Hunt** ($216.54) and **Clara Shaw** ($195.58).
* The top 10 customers’ spending ranges from **$154.60 to $221.55**, indicating a group of highly engaged and valuable customers.

**Top Rental Customers:**

* **Eleanor Hunt** has the highest number of rentals, totaling **46**, just ahead of **Karl Seal** with **45 rentals**.
* Other frequent renters, such as **Clara Shaw** and **Marcia Dean**, have rented over **40 times**.
* The top 10 rental customers have rented between **39 and 46 times**, demonstrating their strong demand for movie rentals.

**Business Value:**

**1.VIP Membership** increases customer retention by offering exclusive benefits like discounts and early access, encouraging long-term loyalty.

**2.Personalized Marketing** enhances user engagement by recommending movies based on rental history, leading to higher repeat rentals.

**3.Subscription Plans** provide a stable revenue stream by offering unlimited rentals to frequent renters, improving customer lifetime value.

**4.Revenue Optimization** helps maximize profits by adjusting pricing for high-demand categories while using targeted discounts to drive spending.

**Decision Making:**

**1.Launch a VIP membership program** for high-value customers, rewarding frequent renters with perks to boost retention.

**2.Implement data-driven marketing** to offer personalized recommendations and promotions based on individual rental preferences.

**3.Introduce subscription-based rental plans** for frequent users like Eleanor Hunt and Karl Seal to encourage long-term engagement.

**4.Adjust pricing strategies** by raising rental prices for popular categories and using time-limited discounts to increase overall spending.

**Data Resource:**

The geographic map for regional customers was created using **the data warehouse file** combined with **a country dataset** in Tableau. This integration allows for a detailed visualization of customer distribution across different regions.

**Dashboard:**

The dashboard provides **a clear, interactive view of customer locations**, helping to identify high-density areas and potential market expansion opportunities. By leveraging this data, the company can refine its marketing strategies and improve regional engagement.

**A screenshot of a computer

AI-generated content may be incorrect.**

**General grading criteria: Your completed work will be evaluated using the criteria below. I encourage you to use your creativity and other business skills (communication, presentation, critical thinking) in addition to the data management concepts and the SQL and Tableau skills that we have covered in CIS467.**

|  |  |  |  |
| --- | --- | --- | --- |
| **High score** | **Score between high and good** | **Good/medium score** | **Low score** |
| All required parts of the final project are complete and technically correct. Queries are useful/interesting and provide valuable information for senior management to act upon. Not just random queries. Tableau visualizations provide interesting useful information based on which senior management of the company can make important decisions. | All required parts of the final project are complete and technically correct (with possibly a few minor errors). Queries are useful/interesting and provide valuable information for senior management to act upon. Not just random queries (with possibly a few minor errors). Tableau visualizations provide interesting useful information based on which senior management of the company can make important decisions (with possibly a few minor errors). | Some required parts of the final project are missing and/or there are more significant errors. Some queries appear random and do not answer any useful/interesting questions. Tableau visualizations are very simple but may still provide interesting useful information based on which senior management of the company can make important decisions. | The final project has large portions missing and/or major conceptual errors. Most/all queries (if any) appear random and do not answer any useful/interesting questions. Tableau visualizations are very simple and **do not** provide interesting useful information based on which senior management of the company can make important decisions. |