Final Project Specification: SQL for Data Analytics

This project mirrors real-world analytics workflows, giving you the opportunity to showcase both your **SQL expertise** and **data storytelling skills**. You will transform raw transactional data into a structured data warehouse, apply advanced SQL techniques, and extract meaningful insights for a stakeholder.

Treat this as a **consulting challenge**—your stakeholder relies on your analysis to make informed business decisions. **Go beyond the obvious, uncover unique insights, and tell a compelling data-driven story.** Most importantly, have fun and think like a data professional!

Outline

Your report should have at least five sections:

1. Abstract (Placed at the Beginning)

- In **under 10 sentences**, summarize:
  - Who your stakeholder is (be specific—e.g., Head of Marketing, Supply Chain Manager).
  - Your **analytics objective**—what problem or opportunity you are addressing.
  - A high-level overview of your **key insights and findings**.
  - The main **recommendations** for your stakeholder.
- Keep this concise but compelling—this is your project's executive summary.

2. Background and Analytics Objective

- Provide a clear **description** of the data science project you are conducting.
- Answer the following:
  - **What is the project about?** What business problem or opportunity are you addressing?
  - **Who is your primary stakeholder?** (Avoid broad groups like "Marketing"—choose a specific role, such as "Head of Customer Retention" or "Operations Lead for Logistics.")
  - **How will your stakeholder benefit from this analysis?** Explain the business impact.

Example:
Instead of: *"My stakeholder is the Marketing Team, and I will analyze sales trends."*
Write: *"My stakeholder is the Head of Digital Marketing. They are responsible for optimizing paid advertising spend. My analysis will focus on customer acquisition costs and conversion rates by campaign type, helping them allocate budget more efficiently."*

- **Define at least three analytical questions** that drive your analysis.
  - These should be **specific** and tied to business decisions. Avoid overly broad questions.
  - Each question should explain its business relevance to the stakeholder.

Examples:
Too broad: "What are the sales trends?"

Better: "Which product categories have the highest sales growth in the past 6 months, and how does seasonality impact them?"
Too generic: "How can we improve customer retention?"
Better: "What are the top three factors driving customer churn based on purchase history, support interactions, and engagement levels?"

3. Data Definition and Dimensional Model

- **Dataset:** You will be working with the **fillians_toy_shop** or **fillians_record_shop** dataset, which represents a transactional database. See
- **Entity-Relationship Diagrams (ERD):**
    - **Step 1:** Draw an **ERD** of the existing **transactional database** (Fillian's Toy Shop or Fillian's Record Shop). This should include:
        - **Entities** with attributes (fields)
        - **Primary keys and foreign keys**
        - **Relationships across entities**, including their cardinalities (one-to-many, many-to-many, etc.)
    - **Step 2:** Design a **dimensional model** based on your analysis needs.
        - **Draw an ERD** for your dimensional model, showing:
            - Fact and dimension tables
            - Granularity of the fact table
            - Relationships between fact and dimension tables

G*uide Questions:*

- What transactional tables should be transformed into dimensions and facts?
- How do the relationships change from a normalized transactional structure to a denormalized star schema?
- **Building the Dimensional Model:**
    - Define at least **one fact table** and **two or more dimension tables**.
    - **Follow proper naming conventions:**
        - Fact tables: fact_<subject> (e.g., fact_sales)
        - Dimension tables: dim_<subject> (e.g., dim_customers)
    - **Define the granularity** of each table—what does a single row represent?

Example:
A **fact_sales** table might have:

- **Granularity:** One row per product sale per transaction.
- **Key fields:** sale_id, customer_id, product_id, sale_date, quantity, total_price.

A **dim_customers** table might have:

- **Granularity:** One row per unique customer.
- **Key fields:** customer_id, first_name, last_name, signup_date, segment.
- **Write SQL queries** to create each table based on the raw dataset.
- **Be specific** about how you filter and structure your data to align with your analysis.
    - **Avoid loading the entire dataset**—apply filters to keep only relevant data (e.g., limit by date range, region, category).
- Present a table summarizing:

- **Table Name** | **Column Names** | **Granularity** | **SQL Query**

4. Data Transformation (Python for ETL, SQL for Analysis)

- You will use **Python for ETL (Extract, Transform, Load) operations** to create your dimensional model:
  - Select data from the **fillians_record_shop transactional database**.
  - Transform it according to your **dimensional model schema**.
  - Insert the transformed data into the dimensional model tables.
- **However, the bulk of your analysis must be performed using SQL.**
- Think of Python as your tool to **"build the warehouse"**, but SQL as your **"analytical engine."**
- This mirrors real-world data warehousing:
  - **ETL processes (Extract, Transform, Load) are handled with Python.**
  - **SQL is used to query, analyze, and generate insights.**

5. Analysis and Insights

- Use **tables and graphs** to present your findings. You may choose any tool to generate your graphs.
- Clearly explain the **statistical and analytical methods** used in your project.
  - Why did you choose them? How do they support your stakeholder's decision-making?
- **Each analytical question from Section 2 should have a corresponding analysis and visualization.**

Example:

If you analyze customer retention, you might:

- Show a **churn rate graph** over time.
- Use a **cohort analysis** to compare new vs. repeat customers.
- Apply **predictive modeling** (if applicable) to identify churn risk factors.
- Ensure your insights **"tell a story."**
  - Don't just present numbers—explain what they mean for your stakeholder.
  - Structure it logically: What you found → What it means → What should be done next.

6. Recommendations and Business Impact

- **For each analytical question, provide a clear recommendation** based on your findings.
- Support recommendations with data—use graphs, charts, and tables.
- Be **actionable**—what should your stakeholder do next based on the insights?

Example:

Instead of: *"Customer churn is high in the first 3 months."*

Write: *"Customers with fewer than 2 purchases in their first month have a 60% churn rate. We recommend a targeted onboarding campaign offering discounts or personalized product recommendations to improve retention."*

Grading Rubric

- **60% Dimensional Model Design, Correctness and Execution**
  - **ERD of Transactional Database (10%)**
    - Clearly depicts all relevant entities, attributes, relationships, and cardinalities.
    - Accurately represents the existing database structure.
    - Correctly identifies primary and foreign keys.
  - **Dimensional Model Diagram (15%)**
    - Shows correct fact and dimension tables, with appropriate granularity.
    - Clearly defines relationships and keys.
    - Demonstrates an improvement from the transactional model for analytics.
  - **Dimensional Model Specifications (20%)**
    - Includes clear documentation of fact and dimension tables.
    - Defines appropriate primary keys, foreign keys, and data types.
    - Justifies design choices (e.g., why certain fields were moved to dimensions).
  - **SQL Queries for Data Warehouse Table Creation (15%)**
    - Correctly implements CREATE TABLE statements.
    - Uses proper data types.
    - Aligns SQL implementation with the ERD and model specifications.
    - Includes well-structured INSERT, UPDATE, DELETE statements if relevant.
- **20% Analysis & Insights (SQL Execution + Data Storytelling)**
  - This section evaluates the depth, originality, and business relevance of your findings. Your analysis should go beyond surface-level observations and provide meaningful insights that directly address your stakeholder's needs.
  - **Depth & Relevance (10%)** – Are the insights well-supported by SQL queries, tables, and graphs? Do they directly answer the business questions and provide value to the stakeholder?
  - **Originality & Uniqueness (10%)** – Strive to uncover insights that are not obvious or widely shared across multiple groups. If many groups identify the same trend or pattern, its value diminishes. The most valuable insights are those that highlight unique findings, challenge assumptions, or offer new perspectives.
  - Tip: Go beyond the most apparent trends. Think critically about why certain patterns emerge, what hidden relationships exist in the data, and how your stakeholder can take action based on your findings.
- **10% Business Case and Relevance**
  - The analysis is well-motivated and relevant to the stakeholder.
  - Clearly defines analytical questions that are specific and actionable.
  - Findings connect back to business impact (e.g., "This insight helps optimize marketing spend").
- **10% Formatting, Design, and Adherence to convention**
  - Uses consistent formatting for SQL queries (proper indentation, casing).
  - Report is well-structured with headings, sections, and clear narratives.
  - Adheres to naming conventions (e.g., fact_sales, dim_customers).
  - Graphs and tables are well-labeled and easy to interpret.

Bonus Points Opportunity

Challenge yourself with a new dataset!

Originally, the plan was to use fillians_record_shop as the primary transactional database. However, to make the project more accessible and build on our past work, we have shifted the default dataset to fillians_toy_shop. You can leverage your previous queries and data exploration from our earlier sessions to inform your analysis on this dataset.

If you're ready to push your limits and embrace a more challenging data exploration experience, you have the option to earn up to 15 bonus points by working with fillians_record_shop instead. This dataset will require you to perform deeper analysis, as you haven't had as much exposure to its structure and nuances.

Why Take on the Challenge?

- **Limited Prior Exploration:** The record shop dataset is new territory, demanding that you explore its structure, relationships, and business implications from scratch.
- **Higher Effort, Bigger Reward:** Overcoming these challenges will demonstrate advanced SQL and data exploration skills, setting your work apart and earning you valuable bonus points.

Bonus Points Breakdown

Your ability to leverage SQL effectively in delivering a robust, in-depth analysis and insights for Fillians_Record_Shop will be assessed based on the following criteria:

- **Depth of SQL Utilization:** Effective use of **advanced joins** to combine data from multiple tables in a meaningful way. Use of **window functions** (e.g., ranking, running totals, moving averages) to uncover trends or key metrics. Proper application of **CTEs (Common Table Expressions)** or **subqueries** to break down complex queries into manageable parts.
- **Quality of Insights & Storytelling:** Insights are **unique** (not generic) and **specific to Fillians_Record_Shop**, considering its industry and dataset structure. Analysis follows a **logical, well-structured narrative** that makes sense for a business stakeholder. Use of **SQL-driven analysis** that support conclusions.
- **Business Impact & Relevance:** Findings are **actionable**, with clear recommendations based on the data. Insights tie back to a real-world business decision that a record shop would need to make (e.g., inventory management, customer preferences, sales trends).

Go beyond basic sales analysis—think of insights related to customer purchasing behavior, seasonality trends, artist/genre performance, or pricing strategies.

Formatting Guidelines

Document format: PDF
Follow the APA 6th guidelines for citations.
Headings: Use bold or heading styles for clarity (e.g., **H1 for section titles, H2 for subsections**).
Spacing: 1.5 line spacing, standard margins (1" on all sides).
Code Formatting:

- SQL and Python code should be in **monospace font** (e.g., Consolas, Courier New).
- Format queries properly with indentation for readability.

Tables & Graphs:

- Clearly labeled, with appropriate titles and captions.
- Include units, legends, and source notes where applicable.

Word Limit

The word limit for the final report is 2,000-4,000 words. References at the end of the report (consisting of a list of URLs and/or cited reports) are not included in the word count. Note that staying within the word limit demonstrates your ability to write concisely. For this reason, a penalty may be applied to reports exceeding the limit, or the marker may ignore the excess of the report. Focus on clarity and conciseness rather than hitting a word count. Well-structured tables, properly formatted SQL queries, and insightful analysis are more valuable than excessive text.

Presentation

Each group will have **7 minutes** to present their project. Given the time limit, focus on the most important aspects of your analysis and avoid unnecessary details. Your presentation should be **clear, concise, and impactful**—think of it as a business pitch to your stakeholder.

Tip: Think of this as a pitch to a business executive—what do they need to know in 7 minutes to make a decision?

Suggested Presentation Structure

1. **Introduction**
    - Briefly introduce your **stakeholder** (specific role, not just a department).
    - Clearly state the **business problem(s)** your analysis aims to solve.
2. **Dimensional Model**
    - Show and **explain your dimensional model ERD** (fact and dimension tables, relationships).
    - Justify your design choices—how does this model support your analysis?
3. **Key Insights & Analysis**
    - Present your **most valuable and unique insights** from the data.
    - Use **graphs, tables, or SQL results** to support your findings.
    - Highlight **any advanced SQL techniques** used (e.g., CTEs, window functions).
4. **Recommendations & Business Impact (1.5 min)**
    - Provide actionable **recommendations** for your stakeholder based on your insights.

- Explain **why these recommendations matter** and how they can drive business decisions.

Additional Guidelines

Be selective—focus on highlights, not every detail.
Make it engaging—tell a compelling data story, not just a technical report.
Use visuals effectively—ERD, charts, and key SQL outputs should be **clear and readable**.
Practice timing—stay within the 7-minute limit to ensure fairness for all groups.
Be prepared for Q&A—after your presentation, the instructor may ask follow-up questions.