Data Analytics for Immersive Environments - Individual Project

SQL, Quarto & Linear Regression

Niall McGuinness

2023-12-14

Learning Outcomes

To practice the following:

- Writing and executing complex SQL queries to retrieve and manipulate data.
- Understanding and applying SQL concepts in a database context.
- Analysing and interpreting data from a SQL database effectively.
- Developing and interpreting linear regression models in the context of game development data.
- Understanding the relationship between different variables and how they impact each other.
- Evaluating the performance of a linear regression model and understanding its limitations and applicability.
- Utilising Quarto to create interactive Shiny web applications.
- Designing user-friendly interfaces for data interaction and visualisation.
- Implementing data visualisation techniques to effectively communicate data insights.
- Applying data analytics techniques to real-world scenarios in game development.
- Analysing and interpreting data specific to game development projects or products.
- Demonstrating the ability to solve problems by applying analytical skills in a structured manner.

Functional Requirements

Part A - SQL (20 Marks)

You will be provided with an SQLite database file (via email before **5PM 18th December 2023**) containing information about various game development projects and their associated assets.

This database includes tables for projects, assets, developers, timelines, and customers. The structure of each table is shown.

Projects

Column Name	Data Type	Description
ProjectID	Integer	Unique identifier for each project (Primary Key)
ProjectName	String	Name of the game development project
StartDate	Date	Start date of the project
EndDate	Date	End date of the project
Budget	Decimal	Total budget allocated for the project
Status	String	Current status of the project (e.g., In Progress, Completed, Cancelled)
CustomerID	Integer	Identifier linking to the Customers table (Foreign Key)

Assets

Column Name	Data Type	Description
AssetID	Integer	Unique identifier for each asset (Primary Key)
ProjectID	Integer	Identifier linking to the Projects table (Foreign Key)
AssetName	String	Name of the asset
Type	String	Type of the asset (e.g., 3D Model, Animation, Texture)
CreationDate	Date	Date when the asset was created

Developers

Column Name	Data Type	Description
DeveloperID	Integer	Unique identifier for each developer (Primary Key)
Name	String	Name of the developer
Specialisation	String	Area of specialisation (e.g., Programmer, Animator, Artist)
ExperienceYears	Integer	Number of years of experience in the field

ProjectDevelopers

Column Name	Data Type	Description
ProjectID	Integer	Identifier linking to the Projects table (Foreign Key)

Column Name	Data Type	Description
DeveloperID	Integer	Identifier linking to the Developers table (Foreign Key)
Role	String	Role of the developer in the project (e.g., Lead, Contributor)

Timelines

Column Name	Data Type	Description
TimelineID	Integer	Unique identifier for each timeline entry (Primary Key)
ProjectID	Integer	Identifier linking to the Projects table (Foreign Key)
Milestone	String	Description of the milestone
ExpectedCompletionDate	Date	Expected date of completion for the milestone
ActualCompletionDate	Date	Actual date of completion for the milestone

Customers

Column Name	Data Type	Description
CustomerID	Integer	Unique identifier for each customer (Primary Key)
CustomerName	String	Name of the customer
CustomerCity	String	City of the customer
CustomerCountry	String	Country of the customer

You are required to write **three** core SQL queries to perform the following tasks:

1. List the total budget allocated for projects in each country, along with the count of projects per country. Display sorted by the total budget in descending order.
2. List the average development time for projects, categorized by the number of assets used.
3. List the top three developers based on the number of successful projects they've been involved in. Display the results.

You are also required to demonstrate the following **three** SQL concepts using no fewer than **three** distinct SQL statements:

1. SELECT with LIKE and OR
2. SELECT with DISTINCT and ORDER BY
3. Subquery with SELECT

Display the results in formatted table(s) in a Quarto Notebook file inside a **Quarto Project**. You should add pagination to the tables if the number of rows is larger than 8-10 (see Recommended Reading on creating tables)

Part B - Linear Regression (20 Marks)

Utilise the provided dataset, which contains details about project budgets, timelines, team sizes, and success rates and apply linear regression analysis to understand trends or patterns in the game development lifecycle.

You are required to perform the following tasks:

1. Perform linear regression to predict the success rate of a project based on its budget **and** team size and present the data in an appropriate plot.
2. Interpret the model coefficients and discuss what insights they provide about game development.
3. Comment on the reliability of the the linear regression model and any outliers present in the data.

Output the results in a Quarto Notebook file inside a **Quarto Project**.

Part C - Interactive Shiny Web Application (30 Marks)

Create an interactive web application using Quarto Dashboards and Shiny to display and interact with data from the database. You are required develop an interactive web web application containing **two** pages/tabs which allow the following:

1. The first page/tab should allow the user to view the contents of the five tables in the database with no filtering or queries executed.
2. The second page/tab should allow users to visualize data through 2-3 appropriate plots and interact with at least 1 plot by hovering over the data presented. Typically, plots could include data related to project timelines, budget utilization, success rates, etc.
3. The second page/tab should allow for user interaction, such as selecting specific projects or developers to view more detailed data. You must include a minimum of 2 fields for interaction (e.g., textfield, dropdown, slider, checkbox) for each plot.

Output the results in a properly a **Quarto Project**. The web application does not need to be published to a remote server and may be run from your machine.

Part D - Quality & Conclusions (20 Marks)

You are required to present Assess the student's ability to present a Quarto project containing three well-formatted documents. You will be assessed in terms of adherence to best practices in coding and data visualization and the quality of your conclusion. The assessment criteria for this component are listed below:

Document Formatting (4 Marks)

- Overall structure and organization of the report.
- Effective use of headings, subheadings, and bullet points.
- Consistency in font, spacing, and margins.
- Proper pagination and table of contents (if applicable).

Suitability of Tables in Part A (4 Marks)

- Clarity and readability of tables.
- Appropriate labeling for tables.
- Consistent and legible formatting in tables and figures.

Suitability of Plots in Part B (4 Marks)

- Relevance and effectiveness of chosen plots to represent the data.
- Ability to convey the intended message or findings clearly.
- Innovative and insightful use of data visualization techniques.

Adherence to Best Practices in Coding and Data Visualization (5 Marks)

- Code clarity, including commenting and organization.
- Use of efficient and appropriate coding methods.
- Adherence to data visualization best practices, including color choice, scale, and avoiding misleading representations.

Conclusion and Reflection (4 Marks)

- Inclusion of a 1-2 paragraph Conclusion section.
- Depth of reflection on what was learned from each part of the assignment.
- Insightful discussion on how the skills and knowledge gained apply to game development and computer animation.

Part E - Technology Exploration (10 Marks)

Finally, you will be awarded a maximum of 10 marks for creative use of R packages which have **not** been covered in class. This could be a package for rendering tables (i.e. reactable), or effective use of OPML fields, or rendering the geospatial data from the database (i.e. customer city and customer country) using additional R packages (i.e. ggplot2 and ggmap).

Version Control Requirements

You must use a recognised online code repository (e.g., GitHub) and make regular well-named commits to your private repository. A link to your code repo must be included in a README as part of the final submission and you must add your lecturer as a developer to the repository. The repository must be named *2023_DAIE_ICA_StudentInitials*.

Your grade for this component will depend directly upon the **regularity** of your commits. A development project of this size should consist of a minimum of **5+ distinct commit messages** spread over the lifetime of the development. Committing all your code in one commit, before the deadline, will be interpreted negatively.

Any submission made which does not include a repository link will not be graded.

Submission Requirements

1. A link to your R source code repository. Ensure that **no** changes are made to the repository following the submission deadline. You should create a separate fork for this submission and leave it unchanged after the deadline. Ask your lecturer for details on fork creation.
2. The assignment must be entirely the work of each student. Students are **not** permitted to share any pseudocode or source code from their solution with any other group in the class.
3. Students may **not** distribute the source code of their solution to any other student, in any format (i.e., electronic, verbal, or hardcopy transmission).
4. Plagiarised assignments will receive a mark of **zero**. This also applies to the individual/group allowing their work to be plagiarised. Any plagiarism will be reported to the Head of Department and a report will be added to your permanent academic record.
5. Late assignments will **only** be accepted if accompanied by the appropriate medical note. This documentation must be received within 10 working days of the project deadline. The Institute standard penalties for late submission will apply.
6. Each student **must** complete and **sign** a single assignment cover sheet. Please submit the signed cover sheet before 5 pm on the Friday of the week of the deadline.

Recommended Reading

- Recommended Text - OpenIntro Statistics Videos & Slides
- SQL Tutorial
- A Comprehensive Introduction to Working with Databases using R
- How to Make Beautiful Tables in R
- Simple Linear Regression | An Easy Introduction & Examples

- Linear regression using R programming
- Add Regression Line to ggplot2 Plot in R
- Fitting and visualizing linear regression models with the ggplot2 R package (CC237)
- ggplot for plots and graphs. An introduction to data visualization using R programming
- Quarto Dashboards | Charles Teague | Posit
- Dashboards with Shiny for R

Additional Resources

The resources below are provided as a mixture of **optional background reading**, technical reference, and useful tutorials.

- YouTube - Hello, Quarto: A World of Possibilities (for Reproducible Publishing)
- YouTube - Beautiful reports and presentations with Quarto
- OPML Options