Lab 1 - SQL

Purpose of this lab is to use the knowledge of SQL and relational database learnt from this course to find insights from DVD rental data. You are a data analyst in a DVD rental company. There is some legacy data stored in sqlite database, your manager "The Manager" asks you to migrate it to duckdb, analyze them and present the insights to The Manager. Create a public github repo to save your work over the entire project.

Task 0 - data ingestion

We will be using the Sakila database which we have gone through in lecture 11-12 and 17-20. Follow the instruction in the lecture materials in the course repo to create a duckdb database file using the Sakila database data from kaggle.

You can follow 18_pandas_duckdb to load the sakila sqlite database into a duckdb database file.

NOTE

> If you aim for VG and want to do task 3 (BI report), you can follow the instruction in lecture 19 to load sakila sqlite into duckdb using data load tool (dlt).

Task 1 - EDA in python

The Manager would like you to explore the data and analyze the questions below. He wants you to make it in jupyter notebook so that the team can directly see the results of your EDA. Combine duckdb and pandas to do the EDA.

It's important that you show the answers, i.e. run your cells with outputs directly in the notebook.

a) Which movies are longer than 3 hours (180 minutes), show the title and its length?

b) Which movies have the word "love" in its title? Show the following columns

- title
- rating
- length
- description

c) Calculate descriptive statistics on the length column, The Manager wants, shortest, average, median and longest movie length

d) The rental rate is the cost to rent a movie and the rental duration is the number of days a customer can keep the movie. The Manager wants to know the 10 most expensive movies to rent per day.

e) Which actors have played in most movies? Show the top 10 actors with the number of movies they have played in.

f) Now it's time for you to choose your own question to explore the sakila database! Write down 3-5 questions you want to answer and then answer them using pandas and duckdb.

Task 2 - graphs

Also, on the same `.ipynb` file, the The Manager wants to see some meaningful graphs from each of the dataframe.

a) Who are our top 5 customers by total spend? The Manager wants to know so that they can reward them with special offers. Create a bar chart showing the top 5 customers by total spend.

b) How much money does each film category bring in? Make a bar chart showing total revenue per film category.

```
Hint: join category -> film_category -> film -> inventory -> rental ->
payment
```

Task 3 (Bonus) - BI report

In lecture 20_evidence_dashboard, we have built a simple dashboard using evidence and duckdb on the Sakila database. The Manager wants some of your findings from task 1 and 2 to be included in the dashboard. Make sure to combine relevant text, tables and graphs. Also feel free to add more analysis and graphs to the dashboard.

Task 4 - video for stakeholder presentation

Create a video 5-10 min long where you record the screen and you go through your repository and solutions to the taks. Your target audience should be The Manager and its data team. Explain with correct computer science terminology but keep it simple.

The Manager and the team have limited amount of time, so keep the time limit stated above. If you have done task 3, you should present it in the video as well both a demo and short explanation of the code structure.

For simplicity use microsoft teams and open up a meeting with yourself, then share screen and record. Afterwards download the video file and upload to your learning platform. If the file size is too big then you could upload it to youtube but keep it unlisted. Or use google drive, onedrive, dropbox or similar, but make sure that anyone with the link can view the file.

LLM usage

LLMs are allowed, for smaller parts, for generating ideas, but not for solving entire parts. Very important is that you understand the code. If there are parts where you use LLMs, it's important that you make a comment that it is LLM generated.

Submission

Hand in the following to your learning platform:

- a link to your public github repository for this lab
- a video file or link to the video

Grading

If you have taken ideas of codes from someone or found them online, it is **important** that you state the source and understand how the codes work. Write a comment next to these codes.

Criteria for G:

- solved tasks 0, 1, 2 and 4 correctly
- the video is clear and easy to follow
- several relevant commits with descriptive commit messages

Criteria for VG:

- also solved task 3 correctly
- the coding is well structured and holds higher quality
- also your video, you use correct computer science and SQL terminology