Dishan Wickremasinghe

Sumati Bhala

Muqtasid Alam

**Week 13 - ETL Project Report**

The ETL project comprised of Toronto Public Library data, which combined multiple data sources to perform data extraction, normalization and finally transformed them to depict how different branches performed along with their in-house programs offered in the past years.

The primary web source we used was “<https://opendata.tpl.ca/>.” This link lead us to some equally interesting data to work with. The data was in different formats like JSON, CSV, KML, and XML.

1. **Data Sets Extracted**

First step for the project was to extract the data. We also used the URL for the list of library branches. Some of the datasets we considered for our project are about active-cardholders, branch information, hours of operation, workstations and new registrations.

Following is the list of the csv files used to build the database.

1. Active\_Cardholders *= pd.read\_csv("resources/Active\_Cardholders.csv")*
2. Active\_Cardholders\_by\_Cardholder\_Type = *pd.read\_csv("resources/Active\_Cardholders\_by\_Cardholder\_Type.csv")*
3. Annual\_Visits = *pd.read\_csv("resources/Annual\_Visits.csv")*
4. Catchment\_Population = *pd.read\_csv("resources/Catchment\_Population.csv")*
5. Collection\_Size = *pd.read\_csv("resources/Collection\_Size.csv")*
6. Hours\_of\_Operation = *pd.read\_csv("resources/Hours\_of\_Operation.csv")*
7. New\_Registrations = *pd.read\_csv("resources/New\_Registrations.csv")*
8. Registered\_Cardholders = *pd.read\_csv("resources/Registered\_Cardholders.csv")*
9. Registered\_Cardholders\_by\_Cardholder\_Type = *pd.read\_csv("resources/Registered\_Cardholders\_by\_Cardholder\_Type.csv")*
10. Circulation\_by\_Cardholder\_Type = *pd.read\_csv("resources/Circulation\_by\_Cardholder\_Type.csv")*
11. Workstations = *pd.read\_csv("resources/Workstations.csv")*
12. *Workstation\_Users = pd.read\_csv("resources/Workstation\_Users.csv")*

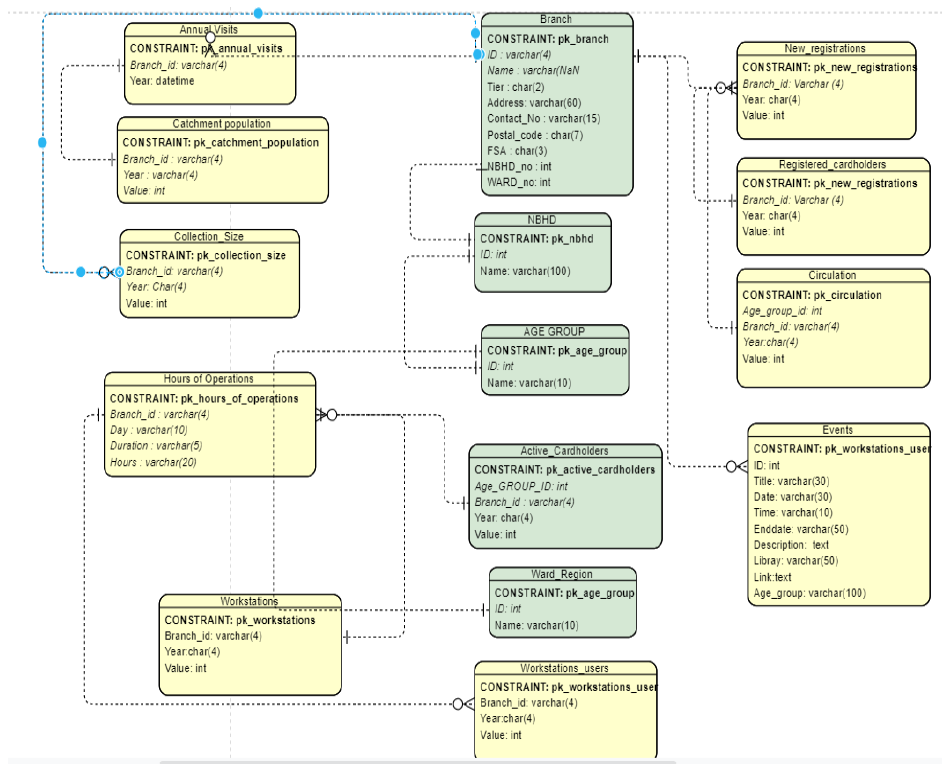
JSON

1. url = "https://opendata.tpl.ca/resources/events"
2. **Transformation**

The data sets had multiple values so the first step was to normalize the data. Followed by data cleaning with python scripts. There were different tables that need to be connected through primary key. It took us a lot of time to decide the schema. We used ERD to finalize the schema.

**ERD Diagram**

Database is normalized upto 3rd normal fom, and the relationships between parent and child tables are shown in the underneath entity relationship diagram.

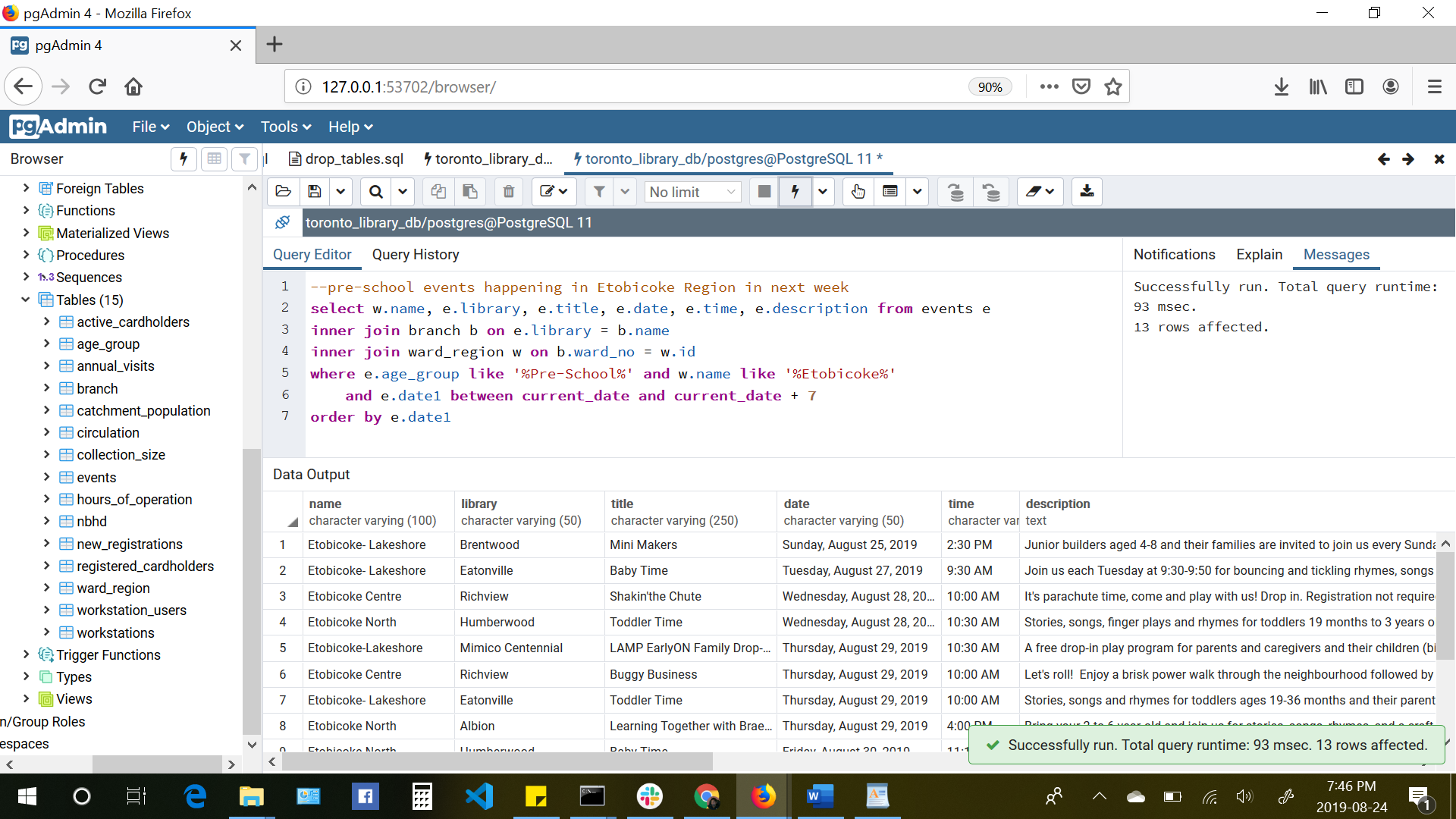
****

1. **Loading to the database**

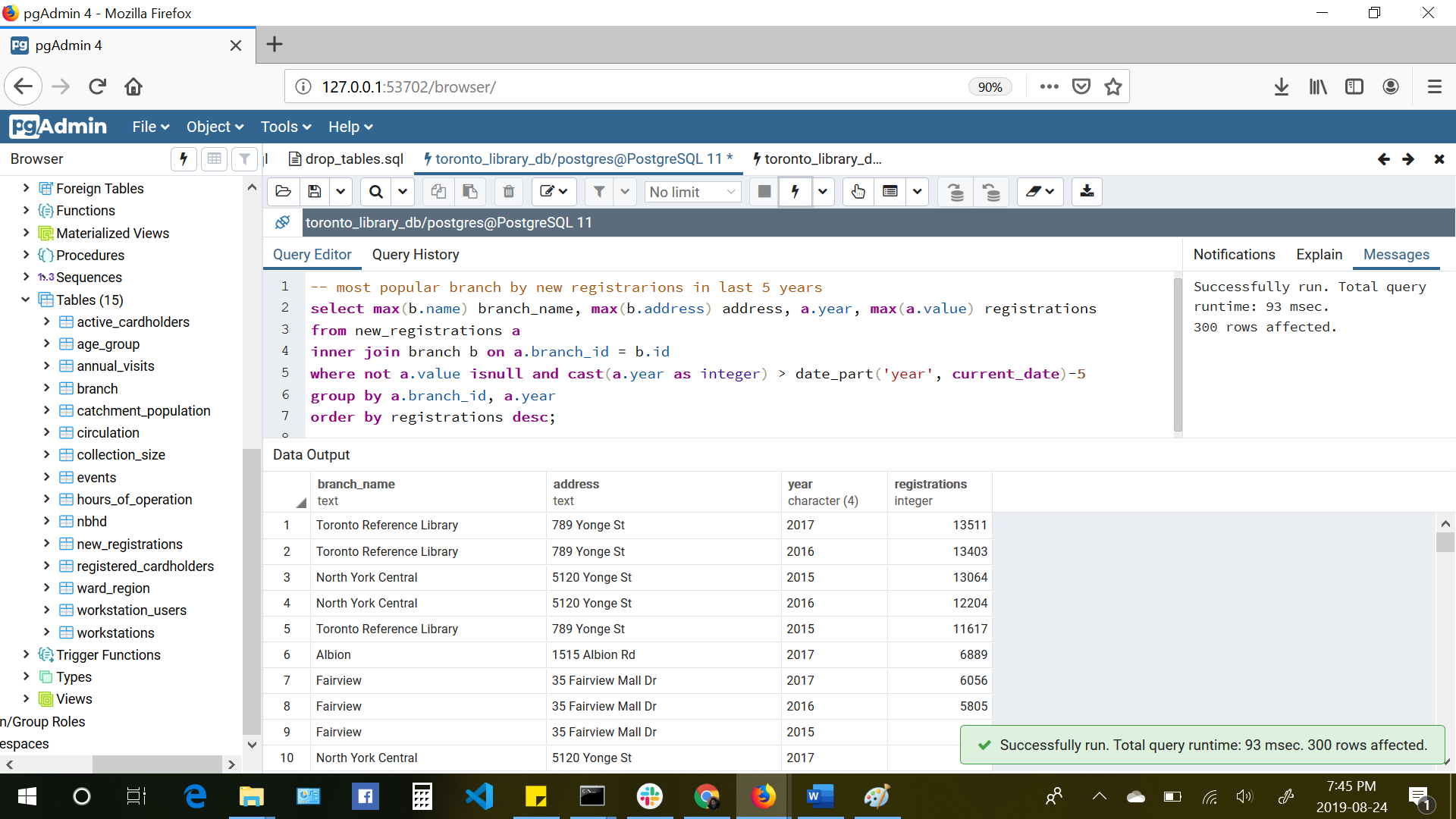
The final step was to write a script to create the database, tables and relationships. We used SQL alchemy to handle all data input into the PostgreSQL database. This data can be further used to run queries and find out more information about Toronto Public library.

Few queries run against the database are shown below.

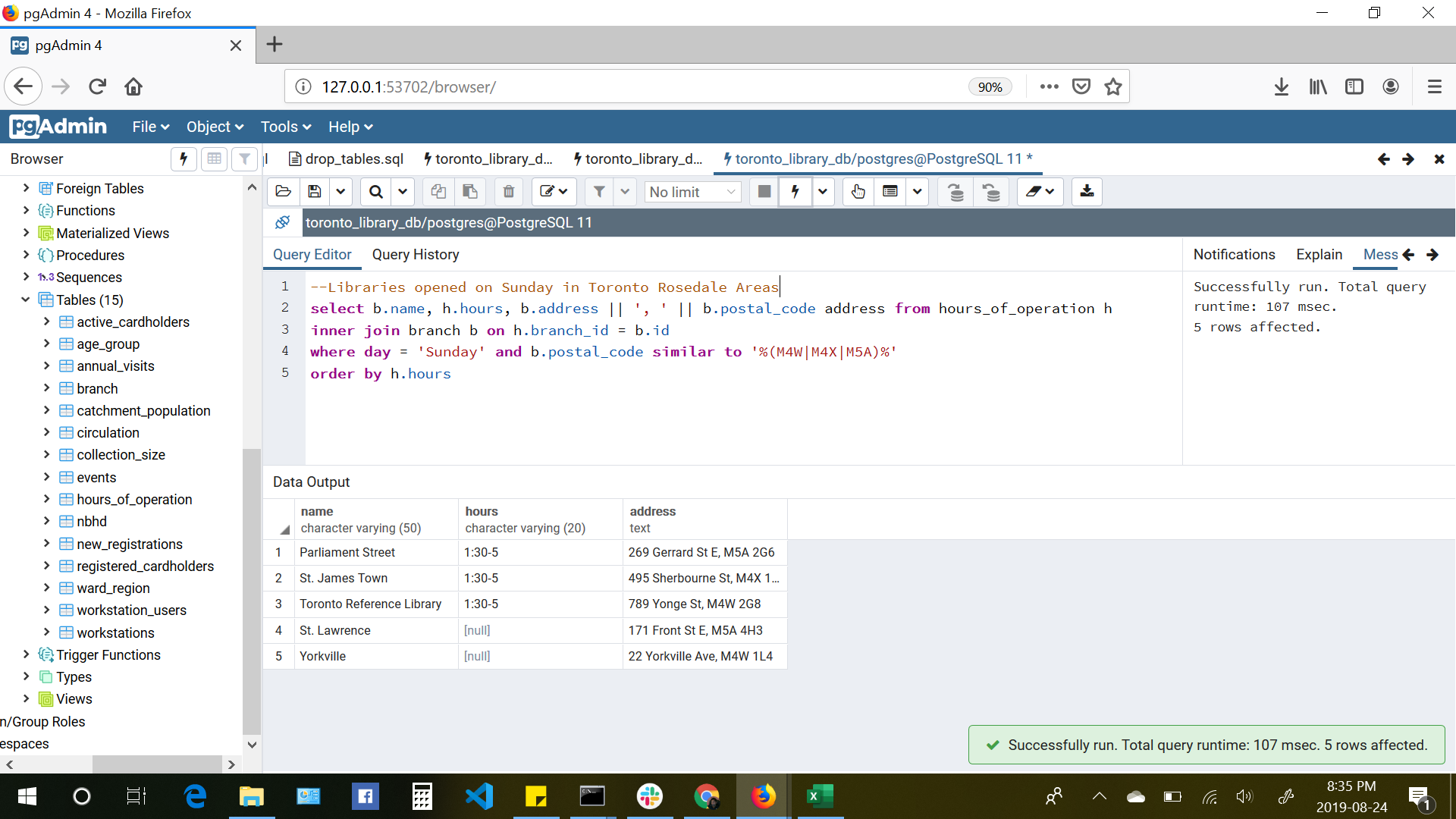
1. Pre-school events happening in “Etobicoke Region” in next week (Aug-24 to Aug-31)

****

1. Branches drew high demand of registration in last 5 years

****

1. Libraries open on Sundays in Toronto Rosedale area



**Why we chose this?**

Majority of the data was in CSV format so it seemed like a perfect choice to use relational database to store the final database.

**Additional learning:**

Web scraping with beautiful soup. We tried to use some of the web scraping skills we learned this to get a list of books from the library web page.

<https://www.torontopubliclibrary.ca/books-video-music/new-items.jsp?category=Adult+Fiction>