Graph Structured Data: DVD Rental Analyst

Sajed Karimy, December 12, 2022 *Algorithms and Data models, Dr. Sean Chester*

Sakila Databese[1]

I will perform exploratory data analysis on top of the sakila database that is distributed as a training example by MySQL, which you can find it here (direct download link). The version of the files that I currently use is sakila-data.sql Version 1.3 and sakila-schema.sql Version 1.4. I will design and implement a labeled property graph-structured logical data model for the relevant data in this dataset. In addition, I use another database called world.sql distributed by MySQL, which you can find it here (direct download link), to add the population of each country to sakila database.

The Sakila sample data is a relational database designed to represent a DVD rental store. The ERD represented in Figure 1, depict the entities and relationships between them. Costumers, Staffs, and Actors are the people involved in this database alongside their properties such as, First Name, Last Name, Email, Address, Picture, the Store they work in, Username, and Password. Country, City, and Address tables depict the exact address used for stores, customers, and staffs. In addition, The Film table includes title, description, release year, rental and replacement cost, etc. Film_Text table seems to be redundant, because all of its information is included in the Film table. However, This table is provided to allow for full text searching of the titles and descriptions of the films listed in the film table and is not supposed to get changed directly. At last, Payment, Rental, and Inventory tables contain information about the business.

Before moving on to the main analysis, I extract some information about this data.

- The release year, language id, and original language id of all films in this database are constant are equal to 2006, 1, and NULL respectively. Consequently, this data is redundant and constant, and we can not do any exploratory data analysis on them. [q1,q2,q3][2]
- There are two stores, and each one manages a store. [q4]

Summary

- **defining problems and requirements:** After making sense of the data, I made a list of problems that could be of interest to a data analyst, and picked the best and most complex ones to analyze. Then, I made a list of requirements based on the problems that my conceptual schema need to meet.
- **MySQL:** Via MySQL Workbench, I installed two databases and combined them to add the desired information to sakila database [q5]. After removing redundant attributes, I exported them to CSV files. At the current and previous steps, the following commands where helpful: SELECT, DISTINCT, FROM, WHERE, ALTER TABLE, ADD COLUMN, UPDATE, INNER JOIN, SET, DROP COLUMN, DROP CONSTRAINT, and DROP TABLE.
- **Neo4j:** I loaded CSV files [q7] and created the desired nodes and relationships, and finally, I analyzed the data. At this step, the following commands where helpful: LOAD CSV, MATCH, MERGE, CREATE, REMOVE, RETURN, WHERE, LIMIT, DISTINCT, ORDER BY, WITH, and DESC.

Data Exploration

As I am going to use labeled property graph as our logical data model, those kinds of analysis are of interest to us which are related to relationships. Because of that, I came up with the following problems.

1. How well do the staffs provide service to customers?
   (a) **Number of rentals:** From which staff costumers have rented more?

[1] Copyright © 2007, 2022, Oracle and/or its affiliates.

[2] All queries are kept in the queries.txt file, and it is referred to the query whenever an information is extracted and shown.

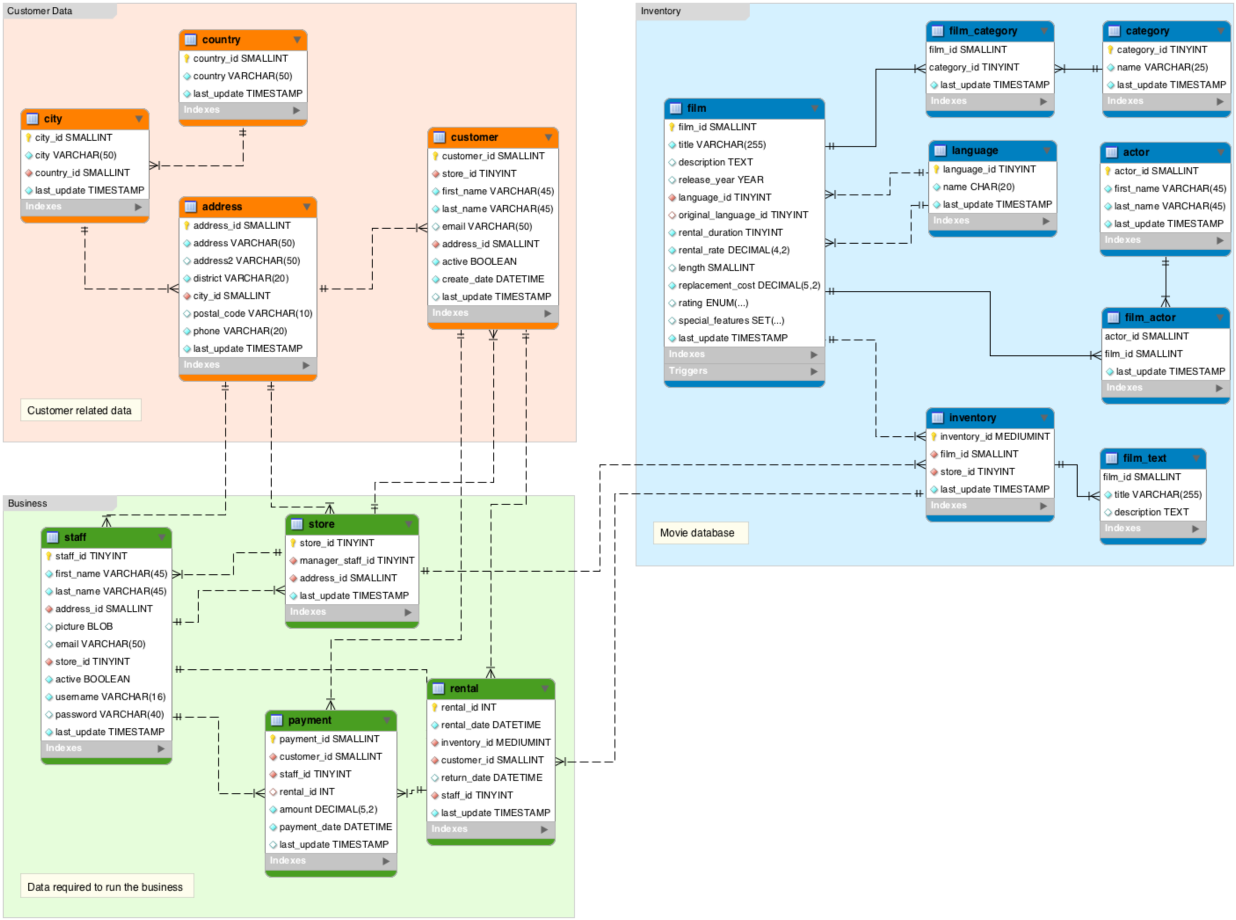

Figure 1: A database schema diagram of the Sakila Sample Database provided by Database.Guide. Note that *address.location* is missing from this diagram.

(b) **Customer loyalty:** If a customer has had rented films from both staffs, from which one they have rented more?

(c) Which categories/actors the staffs have rented more films from.

2. What is the relation between the time in a day that a customer rents a film and the category of that film?

3. Do customers search for movies with the same actors? How many films of the same actor they have rented?

4. Which countries have the most number of film renters relative to their population?

The problem 1c, determines that how much the staffs know about the movies in those kinds of films, so that they have been able to advertise them to the customers. As an example of an analysis of the problem 2, consider the following result: Most of the customers rent horror films at night and rent family films before noon. For the problem 3, it usually happens that you decide to watch a film based on the actors that have played in it. You may search for the movies that a special actor has played in.

Conceptual Design

Based on the problems that we are going to analyze, we need to have the data of country, customer, staff, film, category, actor, rental, and inventory and their relationships. You can see that not all the attributes of these tables are needed. Moreover, we can can join some tables and discard redundant information. Based on these requirements, Figure 2 depict our final ERD. Note that this ERD contains a ternary relationship as a challenge to draw our logical schema based on it. So in our logical schema that we will present, we keep the staff entity as an attribute of the rental relationship, and keep the staffs information separately to use them in our final visualizations.

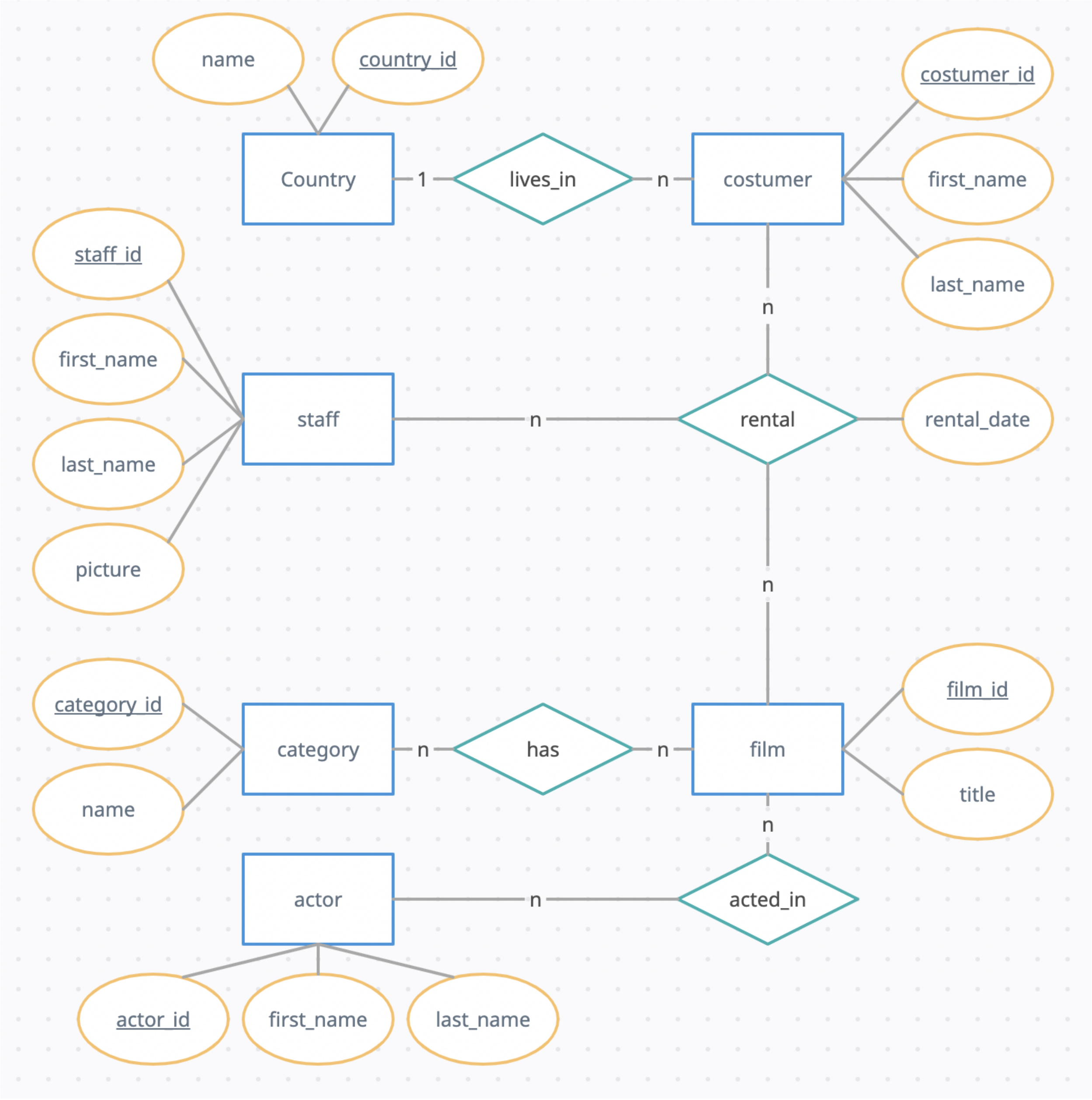

Figure 2: The final schema of our analysis. [q6]

Logical Design

Figure 3 depicts our labeled property graph schema. We have five kinds of nodes labeled by COUNTRY, COSTUMER, ACTOR, FILM, and CATEGORY, where they have directed edges as relationships between them labeled by LIVES_IN, RENTED, ACTED_IN, and HAS.

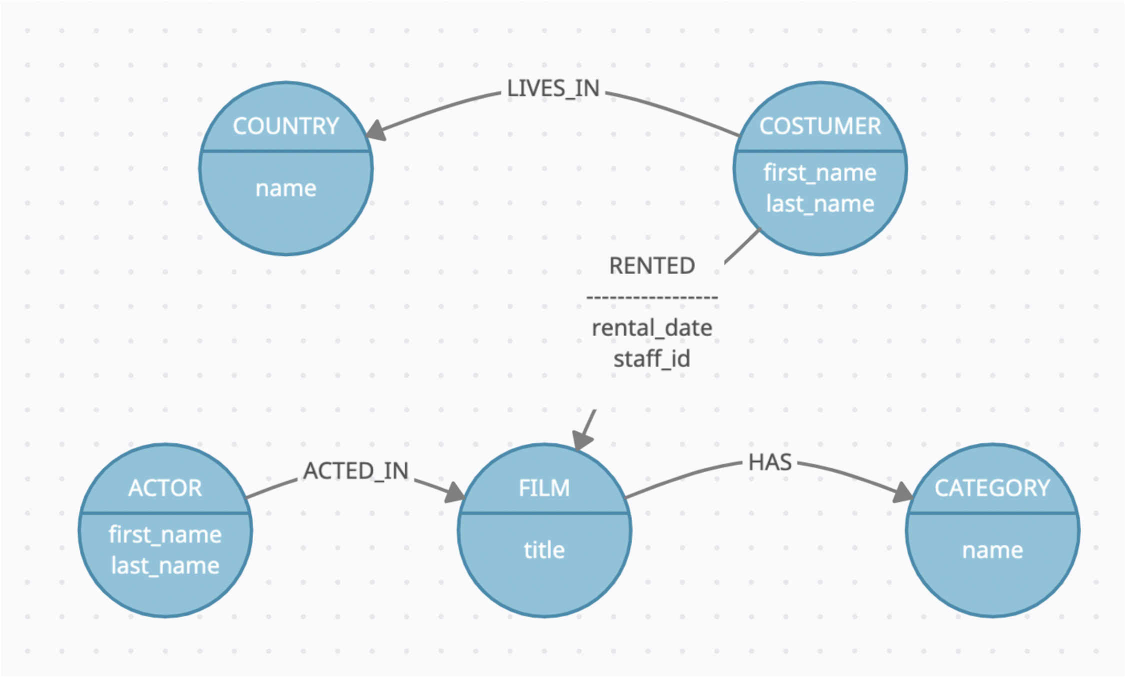

Figure 3: The final labeled property graph schema of our analysis.

This labeled property graph is a "good" schema in the following sense.

- It is **complete**: We show that all the data needed for analyzing the aforementioned problems are provided in this schema.

  1. (a) The number of rentals can be found by counting RENTED edges between COSTUMER and FILM nodes with staff_id of 1 and 2.

     (b) The costumer loyalty can be found by querying RENTED edges by counting the number of COSTUMER nodes based on the conditions on their RENTED edges.

     (c) The ACTOR and CATEGORY nodes and the number of RENTED edges with the satisfying conditions contain the data needed for this problem

  2. The CATEGORY nodes and rental_date properties in RENTED edges contain the needed data.
  3. The ACTOR nodes and number of RENTED edges with the specified conditions are enough to analyze this problem.
  4. The COUNTRY nodes and RENTED edges contain the needed data.

- It is **queryable**: We show that this schema is queryable based on the problems and the requirements by providing some queries. Furthermore, Note that COUNTRY and COSTUMER got related directly to extract the information more efficiently. In general, the denormalization has helped to query and traverse easier and faster.

  1. (a) MATCH (:CUSTOMER)-[r1:RENTED staff_id:’1’]->(:FILM)

     (b) MATCH (c:CUSTOMER)-[r1:RENTED staff_id:’1’]->(:FILM),
     (c)-[r2:RENTED staff_id:’2’]->(:FILM)

     (c) MATCH (c:CATEGORY)<–(f:FILM)<-[r:RENTED staff_id:’1’]-(:CUSTOMER)

  2. MATCH (c:CATEGORY)<-[:HAS]-(:FILM)<-[r:RENTED]-(:CUSTOMER)
  3. MATCH (c:CUSTOMER)–>(f1:FILM)<–(a:ACTOR)–>(f2:FILM)<–(c)
     WHERE f1<>f2

4. MATCH (co:COUNTRY)<–(cu:CUSTOMER)-[r]->(:FILM)

- It is **minimal**: All the redundant data is removed, and no duplicate data is kept.

In conclusion, this logical design adheres to all of the requirements minimally. To implement it as a graph in Neo4j, we just need to apply the create command on the CSV files.

Results of analysis

1. (a) Figure 4 depicts the results. It seems that both staffs have provided almost the same number of services, and they are in sort of competition. [q8]

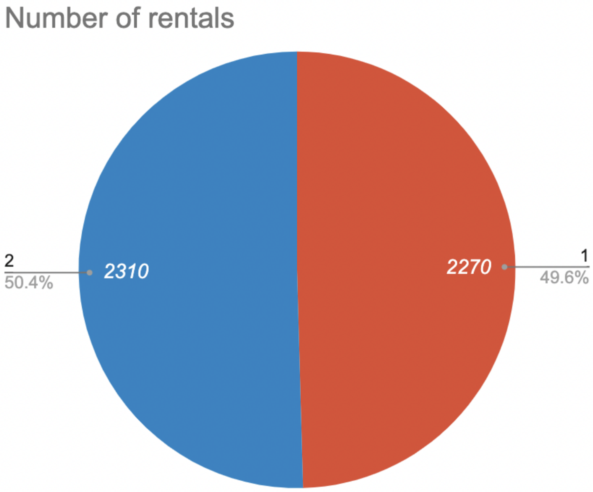

Figure 4: It seems that both staffs have provided almost the same number of services.

(b) Figure 5 depicts the results. Like the previous case, they are really close at this parameter. [q9]

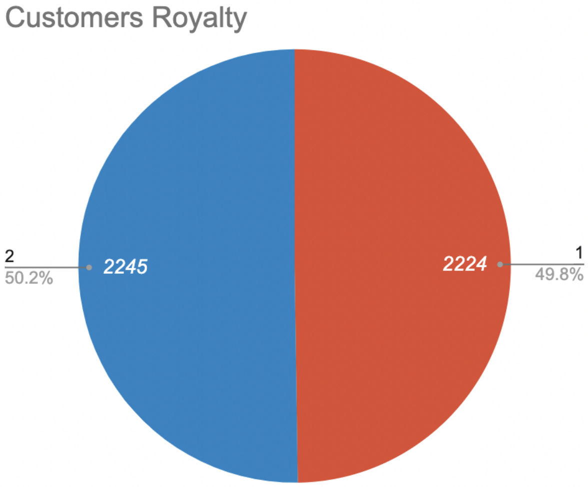

Figure 5: It seems that customers are almost royal to both stores.

(c) Figure 6 and 7 depict the results for Categories. Based on this observation, Staff number 1 mostly know about Action movies. However, Staff number 2 mostly know about Sports Films. [q10]

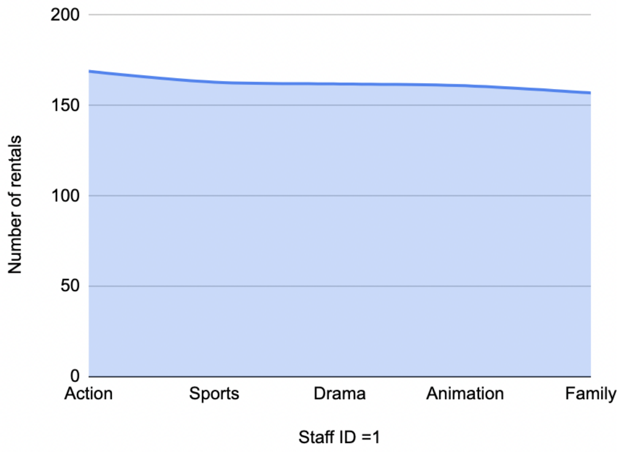

Figure 6: Categories: Staff number 1 has mostly advertised about Action Movies.

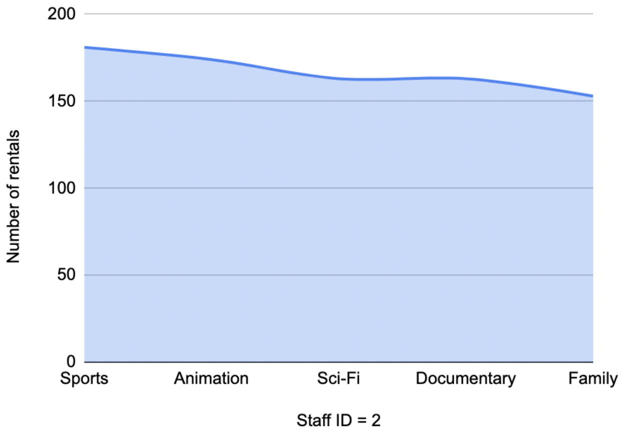

Figure 7: Categories: Staff number 2 has mostly advertised about Sports Movies.

(d) Figure 8 and 9 depict the results for Actors. It is interesting that 3 out of 5 of the first 5 result are common. Maybe this is not a good idea to measure their knowledge about actors in this way. [q10]

2. Figure 10 depicts the result. It is interesting that customers have rented DVDs between 6 am and 12 pm more than other times. Also Note that the two categories with most number of rentals between each period is shown below. [q11]

   - 00-06: Documentary and Sports
   - 06-12: Animation and Sports
   - 12-18: Sci-Fi and Action
   - 18-00: Sports and Animation

3. The number of customers that have seen two different movies having at least one actor in common, is 531. Note that there are only 599 customers, i.e., that is 88%. In addition, the number of customers that have seen three different movies which an actor has played in all of them is 112, i.e., 18%. [q12]

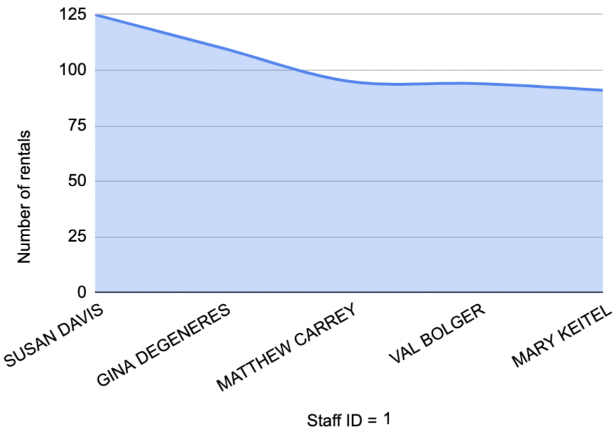

Figure 8: Actors: Staff number 1 has mostly advertised about SUSAN DAVIS.

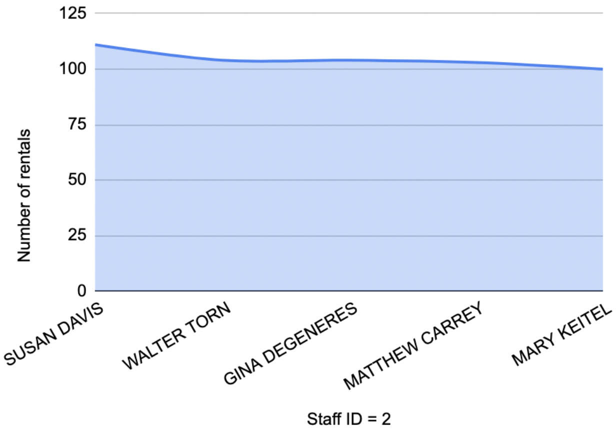

Figure 9: Actors: Staff number 2 has mostly advertised about SUSAN DAVIS.

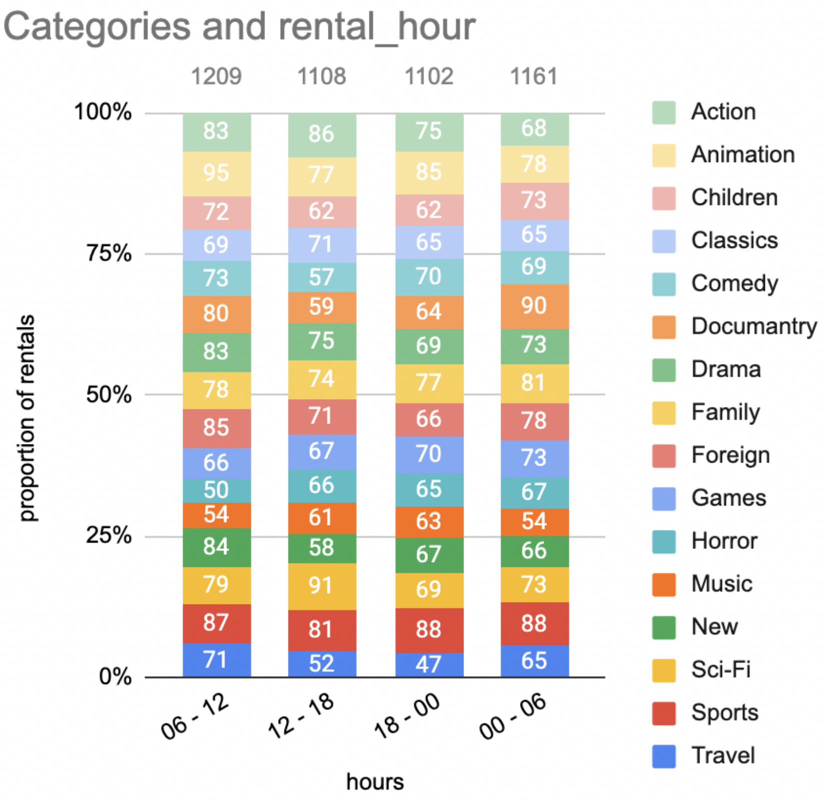

Figure 10: Note that customers have rented DVDs between 6 am and 12 pm more than other times.