Data Science

Business Intelligence (BI)
Primary Topic: Data Processing and Visualization
Course: 2020-2A – Group: 32 – Submission Date: 2021-04-16

Hemanthkumar Sureshkumar
University of Twente
h.sureshkumar@student.utwente.nl

Marzieh Adineh
University of Twente
m.adineh@student.utwente.nl

ABSTRACT

The report focuses on Business Intelligence using the dataset "Sakila-db". The data set was originally created by a former team member from MySQL. In this dataset, we have used creative methods such as XAMPP tool for SQL service and DBeaver tool to run and connect SQL scripts with the PostgreSQL server. The reliance on the R programming tool has been the traditional method for statistical analysis of the datasets and connections with SQL scripts with the server. However, a different approach has been implemented to utilize different tools and understand their productivity on cross application effectiveness.

KEYWORDS

Data Preparation and Visualization, Tableau, SQL scripts, ETL, schema, dashboard.

1 INTRODUCTION

Data Preparation and Visualization (DPV) is a method that is widely used to carry out multiple processes with the use of data in a series of iterations. The data is prepared with the extraction of relevant data, cleaning the data and storing it in a usable format based on the principles of the ETL (extract, transform and clean). Using this method is the steppingstone for the field of Business Intelligence (BI). Here, the BI includes the data analysis of the business information. Sakila dataset can be interpreted as the data of a film DVD rental store where multiple dimensions revolve around the unique fact. The data derived from the store conveys the sales of the film DVDs and income generated by the store over rentals and sales respectively. In order to do so, we had to first perform the data warehousing using an SQL tool. This required the analysis of appropriate tools that would be the best use case.

The chosen MySQL Workbench tool that we worked on has been developed by the former member of the MySQL team. It was logical to adhere with the standards as this eased into the processing of data and feeding inputs to the respective tables. In MySQL, we can perform several functions depending on the requirements. They are:

1. Visualize the sakila schema along with the appropriate connectors.
2. Visualize the tables and commands associated.
3. Create tables and values from the given dataset.
4. Map with servers or link with external tools like R, XAMPP, etc.

On completion of the data preparation, there would be the visualization using the Tableau tool to interpret meaningful information required for answering the framed business questions based on the balanced scorecard perspectives.

The following sections would describe our background knowledge, approach, results and conclusion.

2 BACKGROUND KNOWLEDGE

Business Intelligence is the understanding of different aspects discovering problem solving, continuous improvement and sustainability in context of any business [1]. MySQL paved the way for SQL scripting which is an open-source relational database management system. The application is widely used in data warehousing, e-commerce and interlinking with other applications. Furthermore, the tools like XAMPP and DBeaver are also free and open-source data operational tools. XAMPP helps in transitioning the data from a local environment to a live running server easily. Originally, the tool was used by developers to test cases before deploying to a live environment. DBeaver behaves more as a plugin to connect the environments with less hassles. We have also utilized the R programming tool for data cleansing and retrieval of the required meaningful data. Though R and Python are experienced tools in the field of data science with similar syntaxes, R is specialized with data manipulations whereas Python is dedicated towards the machine learning algorithms of data science.

3 APPROACH

The business questions were framed based on the scorecard perspective. They are as follows:

Q1. Financial: What were the top five film categories rented from the store?

Q2. Customer: Which special features DVD rented makes the most amount?

Q3. Internal Business Process: At which hour is there more crowding that we must manage at the store?

Q4. Employee and Organization Innovation and Learning: How could we improvise income for our stores in different cities?

While these questions are framed, they answer different stakeholders about the performance of the DVD store. The Q1 provides information about the sales performance aspect of the store. Using this, the store can understand which film categories are most preferred by their customers and plan to stock more on those respective categories. The Q2 gives detailed information for the customers as the question answers about the special features set of the DVDs sold for renting to the customers. The customers will be able to judge the features most liked based on the sale and take conscious decisions on the current trends. Q3 on an interesting note facilitates the details to plan the business process that the store can invest for development. The busiest hour identification can answer the store to plan the human resource allocation to address customers and help in increased renting. The Q4 focuses on the learning and development for the employee and the store that can add value.

3.1 Data Processing

The "sakila" dataset is an SQL dataset that requires MySQL to process and work. It includes a pre-processing stage where the values for the corresponding tables were built. The XAMPP, as shown in the Figure 10 of Appendix section, application ran the server in the local host. On completion of the process, the data is then imported to our university PostgreSQL platform. During the pre-processing stage, we used MySQL to view the entire dataset schema and it gave a broad understanding of the primary key and foreign keys respectively. To run the script, we then used the DBeaver tool as you can see in Figure 1. You can also find the related images on script executions using the tool in Figure 11 and Figure 12 of Appendix. The connected database platform is then used for retrieving tables in the data processing tool R where the commands for tables are run as shown in Figure 9 of Appendix. The required datasets were selected as part of the data cleaning process which would be required for the visualization on Tableau [2].

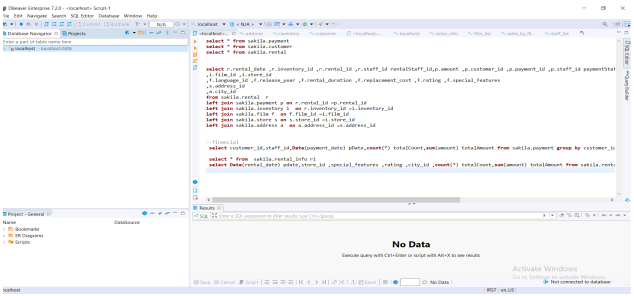

Figure 1: Initial scripting on the DBeaver tool.

3.2 Data Visualization

The data has been visualized using the Tableau. The data source must be first connected with our PostgreSQL where the tables are imported into Tableau. Once the connection is live on the data source section of Tableau, the tables are interconnected like a schema representation based on the referenced foreign and primary keys as shown in Figure 2.

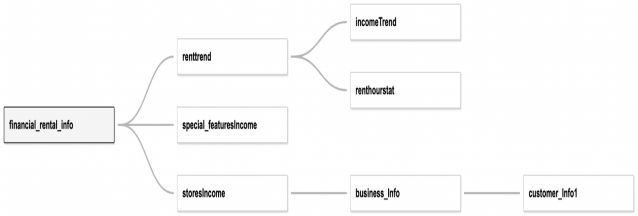

Figure 2: Data source representation on Tableau.

4 RESULTS AND DISCUSSION

The results are presented in multiple deliverables as Star schema, Dashboard and the answers to the framed business questions of the balanced scorecard perspectives.

4.1 Schema

The schema has been divided into the following layers: **Customer Data, Business, Inventory and the Views**. They represent the customer related information, data required to run the business, movie database and special view on certain data used for appraisals respectively. Using the given schema, on processing the data as per our business requirement, we framed a fundamental star schema.

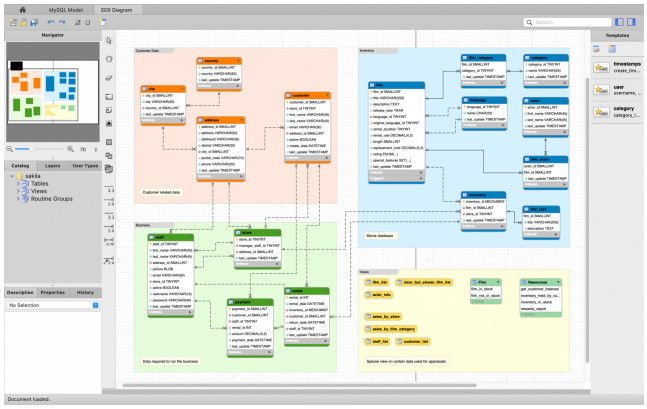

Figure 3: The star schema representation using the MySQL workbench.

The star schema will represent the fact as "Revenue" with the dimensions as "Customer", "Time", "Movie", "Stores" and "Staff".

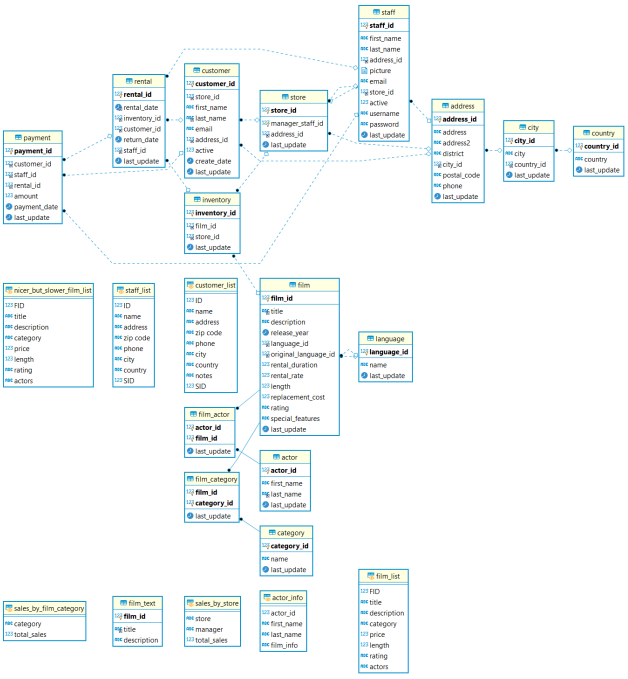

Figure 4: Star schema as per the chosen business questions of balanced scorecard perspective.

The schema also recommends a detailed note on the foreign keys for the corresponding fact and the dimension tables. Here, we were able to understand what our Dashboard connection on Tableau should be to visualize the results for our business questions as well.

4.2 Dashboard

The Dashboard data is visualized on connection to PostgreSQL with the login credentials over the Tableau. Once the connection is complete, the table is updated to load the data on the tool. This operation is now explained under the graphical representation on four different scenarios.

4.3 Graphical Representation

To illustrate the top five film categories, we used the sales margin for the corresponding categories as shown below in Figure 4. The data gives a broad idea on which film categories are sold or rented the most from the DVD store. On calculating, we observed that the top five were Sports, Sci-Fi, Animation, Drama and Comedy.

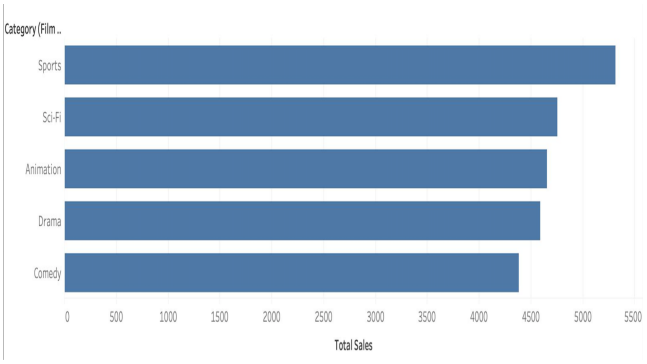

Figure 5: Revenue generation for the top 5 film categories.

To understand how the customer would view us, our approach focused on the special features rented most from the store. The customers would like to view the maximum rented income. These sales figures are generally posted as banners outside stores to promote their businesses. From this information, customers can stay on top of the trend. The maximum rented special feature was found to be Trailers, Commentaries, Behind the Scenes.

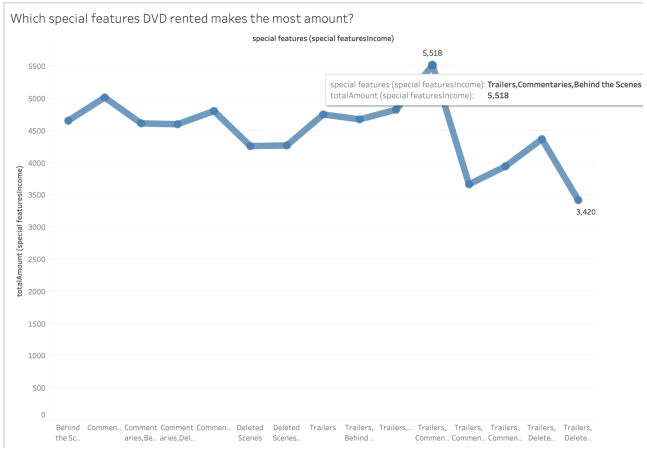

Figure 6: The maximum special features trend.

The busy hour calculation helps in planning the business process for the store. Figure 6 depicts the hourly calculation for the count of customers in the store. It was found that the maximum rentals were at the 15:00 hours of the day.

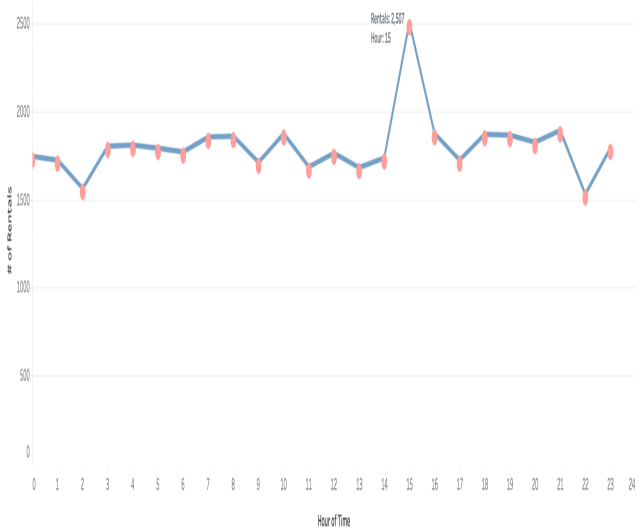

Figure 7: Hourly rent trend analysis of the days.

The learning curve for the organization and employee can be interpreted through what they can continue to improve and add value. The result visualized in Figure 7 gives an idea for the organization for planning their inventory based on the stores at respective cities for a weekly schedule.

<What rating type of DVDs should we have more on stock at which city and store respectively?>

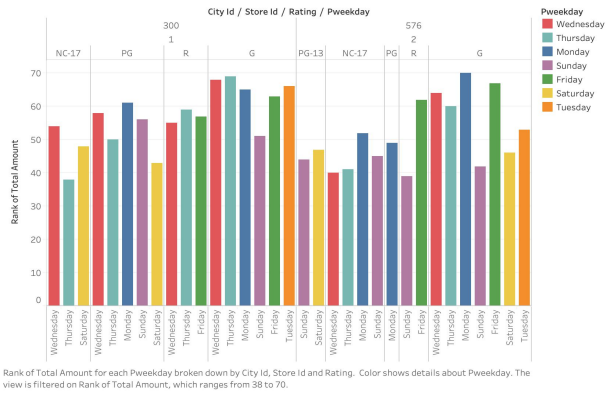

Figure 8: Busy stores at respective cities with the corresponding days of week.

This elucidates that the busy days are generally on Wednesdays where the most rented DVD rating was found to be G in the city with ID 300. Due to many stores available in different cities, the store ID can help in tracing the location of the store from the database of the company. Using this information, the inventory control is made easier.

5 CONCLUSION AND FUTURE RESEARCH

The study on Business Intelligence helped us to learn about the scope of every firm to analyze their business goals along with alignment of mission and vision. This explains broadly the real-life context of a DVD store in the "sakila-db". To achieve the visualization, we were able to incorporate the skills for using Tableau. The independent research groups and organizations that worked on this excellent dataset have also showered an immense amount of knowledge base to our project. We were able to think on the business questions aligning with the "why" component for every question.

However, we believe that there can be an additional set of methods that we could utilize to analyze the datasets in future. The research on such qualitative dataset could incorporate the linear regression models. A linear regression would be useful in contributing for more explanatory power in analysis as the technique focuses on the accuracy of each predicted value.

ACKNOWLEDGMENTS

We would like to thank our professor, Elena Mocanu, for her constructive feedback on our project. We also would like to extend our sincere thanks to every other source for their immense support and contributions on our learning curve.

REFERENCES

[1] Liang, T. P., & Liu, Y. H. (2018). Research Landscape of Business Intelligence and Big Data analytics: A bibliometrics study. Expert Systems with Applications, 111, 2–10. https://doi.org/10.1016/j.eswa.2018.05.018

[2] Nestorov, S., Jukić, B., Jukić, N., Sharma, A., & Rossi, S. (2019). Generating insights through data preparation, visualization, and analysis: Framework for combining clustering and data visualization techniques for low-cardinality sequential data. Decision Support Systems, 125, 113119. https://doi.org/10.1016/j.dss.2019.113119

APPENDIX

Figure 9: Pre-processing and data cleaning over R.

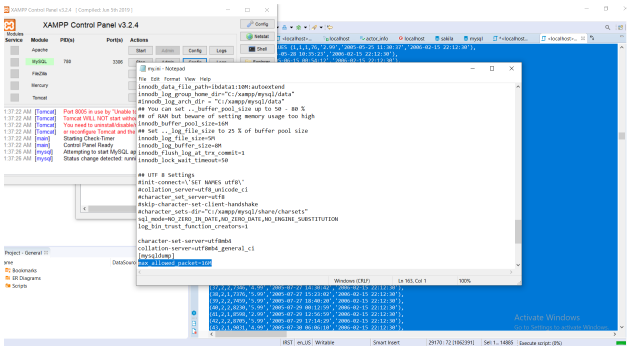

Figure 10: Connection using the XAMPP tool.

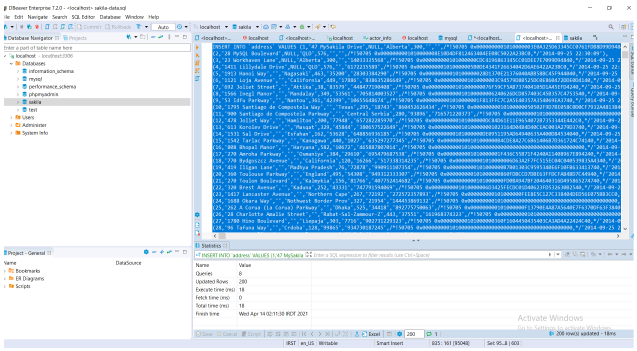

Figure 11: Running scripts over the DBeaver tool.

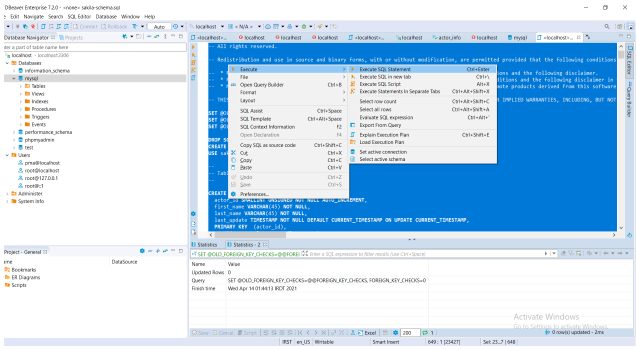

Figure 12: SQL script execution stage on the DBeaver tool.