Databases and Data Visualization Group Project Report
Group 10

Jiahao Tang - 12106208, Kevin Veenboer - 12604275,
Nitai Nijholt - 12709018, Kelvin Gilian Mulder - 12253286

Contents

1. Introduction, database and case description

In a modern business, competitive pressures firms to either grow or perish (Volberda, Morgan Reinmoller, Hitt, Ireland Hoskisson, 2011). Given this fact, businesses are increasingly focussed on making optimal strategic decisions in order to achieve growth. To optimize these decisions and their corresponding outcomes, management needs to turn the information they have access to into relevant knowledge which can support the decision process to ensure their decisions have enough traction to affect their circumstances in a positive way. An important starting place is the information that is gained in the business from its day-to-day operations. From this, insights can be gained, which are uniquely relevant not only to the operations of the business, but also the local business environment.

This collection of information needs to be stored in a database, in order to be exchanged, searched, corrected or supplemented and thus be useful to the business. Also it must get corrupted or lost, as the information stored in it is a key asset to the business (Amrit, 2021a).

For a database to be useful to the business, it must be designed towards the specific requirements of this business. This means providing the right information to actors and in turn delivering the right output, as well as acting upon the data/changing it as required (Domdouzis, Lake and Crowther, 2021, 71 - 72).

Also, businesses have different requirements they place on databases in terms of consistency, availability, and partitioning, and navigate the trade-off between these characteristics database design according to their use case (Amrit, 2021c).

The Sakila dataset is a database modelling a DVD rental business with 2 stores.

There are four main parts of the Sakila database according to the EER diagram.

1. Customer data. The personal information of customers are stored in this session, including the addresses, names, and e-mail addresses of customers. The tables in this session provide insights over KPIs from customer perspectives.
2. Inventory. Since Sakila models a DVD rental business, the only inventories in the dataset are films. The information of movies are stored in this session, such as film names, categories, actors, languages, rental rates, release years, and so on.
3. Business. The tables in the business session store the data of the stores and employees and the daily business transactions. The ‘rental table’ is a junction table, which describes a many-to-many relationship between customers and films. The ‘store’ table

is also a table connecting the customer database and the film database, and we can see the performance of different stores by looking at this connection. The information in this session is helpful when analyzing the KPIs over all four aspects.

4. Views. The Views session gives some insights over certain data, such as the sales by store, sales by film category, etc.

During this project, we aim to extract insights from this database by designing a data warehouse. This can be used to acquire business intelligence through the designing of certain KPIs which help answer relevant business questions. This information can then be communicated to management in order to help them support their decision process and in turn optimize business outcomes. The design of these KPI's are based on the balanced scorecard model which is explained next.

2. KPIs selection and explanation

Figure 1

Financial	**Customer**	**Internal Business Processes**	**Employee and Organisation Innovation and Learning**
Revenue per region	Customer per country	Revenue per category per day rented out	Workload per staff
Revenue per month	Active customers per store	Rental Duration per Film	Revenue per staff

A 'balanced scorecard' is a tool used by firms to balance the interests and requirements of different parts of a business in order to achieve optimal performance of the whole. It is used to determine and control strategy, finances and optimally satisfy all the stakeholders (Customers, shareholders, employees). It is also used to strike an appropriate balance between short term exploitation (such as profit taking, share buybacks, etc.) and investment for the longer term (expansion, employee training programs, customer service capability investments) (Volberda, Morgan, Robert, Reinmoeller, Hitt, Ireland & Hoskisson, 2011).

1. Financial perspective. Business owners always want to know the profitability of this business, such that they can ensure the business keeps operating. Revenues and costs

are two essential elements of the profit. Revenue growth is also important as it determines strategic decisions and shareholder ROI. As many cost categories are unknown, we choose revenue per region (indicated by the proxy payments) and revenue per month (indicated by the proxy payments per month) (Kaplan and Norton, 2005, 72.)

2. Customer perspective - concerned with the value customers perceive by the firm's products or services (ibid.).
3. Internal Business Process - focus on the business processes which enable the firm to satisfy customers, employees and shareholders (ibid.).
4. Employee and Organization Innovation and Learning - concerned with the firm's ability to realise innovation, change and growth (ibid.).

3. Data Warehouse design.

Figure 2

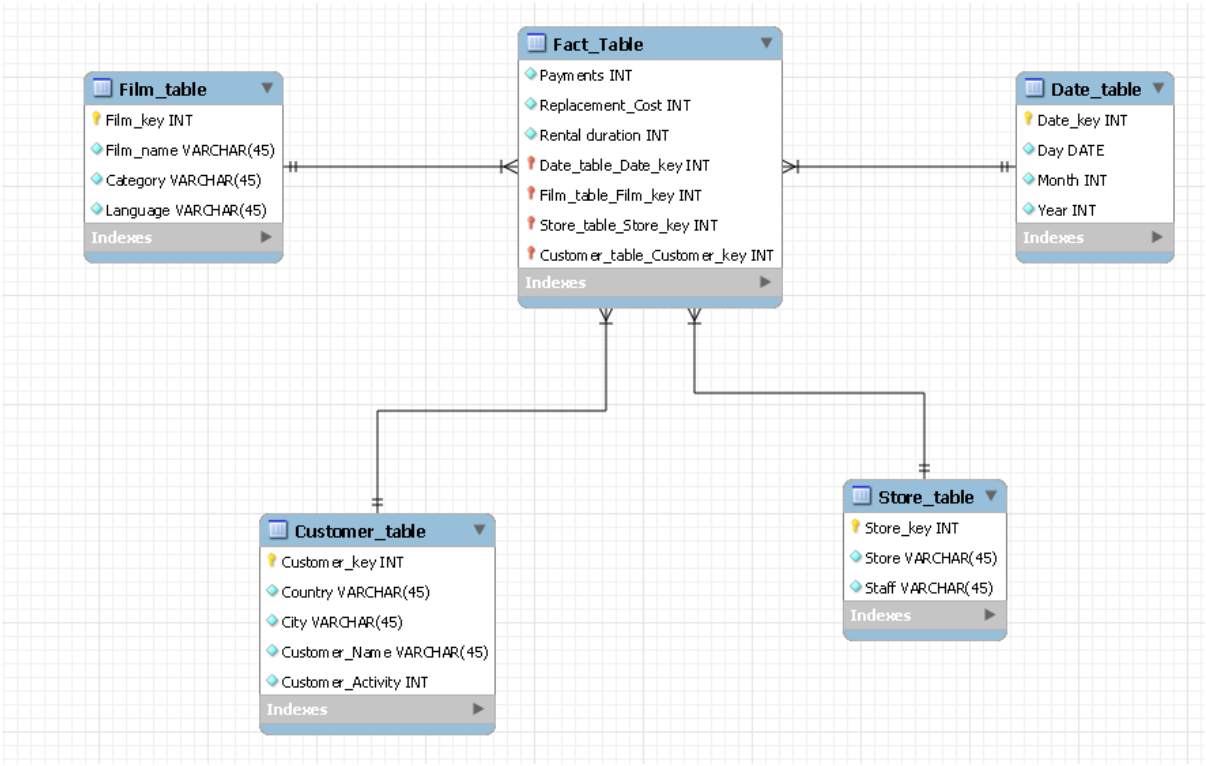


Figure 2 shows our data warehouse design. This section outlines the arguments for the design choices we made in coming up with this design based on the requirements.

Firstly we dive into some theoretical background.

A data warehouse is used to support business intelligence processes. It is used for data analysis and decision making, and therefore it has different requirements than transactional databases.

Data warehouses do not contain all the data of an organisation, but only the data that is necessary for KPIs. In addition, the data in a data warehouse does not need to be updated in real-time. Updating the data when the period between updates has become meaningful for review, e.g. a month for most business intelligence purposes, is fine. Transactional databases have different requirements as they are used for different means. These are used for everyday operations like customer orders, checking inventory status, producing new products, management, administration, etc. Therefore, transactional databases need to contain all the data of an organization. This data has to be updated in real-time to ensure that business processes are continued, as any disturbance might result in a loss of revenue (Domdouzis, Lake and Crowther, 2021, 190-191).

We are transferring the data of the transactional Sakila database within a non-transactional data warehouse, better suited to conduct complex queries. These queries can then be used to obtain and visualize the data, gather insights (business intelligence) and communicate these with management to inform strategy. We base our data warehouse design choices on these requirements. We prefer optimized business queries over minimized data redundancy as we desire fast access to relevant business data so that the time to visualize and draw strategic conclusions can be minimized. Therefore we limited the normalization of the data warehouse to merely second normal form by using a star schema instead of a snowflake schema. Our choice of schema is based on our desire for increased query efficiency, as less normalization will require less joins and result in faster queries at the cost of increased data redundancy (Domdouzis, Lake and Crowther, 2021 , 33-34 ; Amrit, 2021b). Some of the dimension tables could be further normalized to achieve third normal form but this decrease in data redundancy could lead to longer query time, which we wish to minimize.

Within this star schema, we design our data warehouse based on our KPI selection. As listed in the balanced scorecard above, we measure eight KPIs. Based on the data needed for these eight KPI's, we created a central fact table containing Payments, Replacement_costs, and Rental_Duration and the foreign keys linking it to four tables for each dimension, which are film, date, customer and store as can be seen in figure 2. We then created tables in MySQLWorkbench in accordance with our data warehouse design, and loaded the data into these tables. All the SQL scripts used to create and populate the warehouse can be found in the zip file 'All sql stuff relating to the warehouse'. This folder also contains a ReadMe.txt file explaining how to use the scripts and a dump file we used to ease collaboration since populating the warehouse using the populating script took quite some time. To minimize the time lost on

this we created a dump file after the warehouse was successfully populated by one of the team members. Besides the forelder, the query code can also be found in the Appendix under appendix 1 and 2. Screenshots of the filled fact- and dimension tables can be found under appendix 3.

Finally, we used Tableau to visualize these eight KPI's and made a workbook with the name 'BI Appliction.tbwx'. Visualization is added because it can help to understand different interrelationships between different sets of data which can reveal important business insights (Domdouzis, Lake and Crowther, 2021 , 205-210 ; Jesen, Pedersen, and Tomsen, 2010). The results and insights gained are discussed in the next section.

4. Data visualization and Analysis

We will now discuss the results of visualizing our eight KPIs using Tableau. We use barcharts, tables, a line chart, and a map in our analysis, based on what is most suited for the data on display. Our motivation for using these specific visualizations relies on the 'Data Visualization' chapter in the book 'Essentials of Business Analytics' written by John F. Tripp. In his work, Tripp elaborates on human perception of interrelationships between sets of data and the do's and don't of visualizing data most effectively (Tripp, 2019).

Tripp argues that the visualization used should be above all clear (directly representing the relation that is meant to be shown and non obscured), in service of to the data (high data to ink ratio), representative and fair (consistency between tick sizes, no misleading scales) consistent with known human perceptual intuitions/limitations (not combining to many dimensions in 1 graph, not asking users to compare multiple dimensions at once etc.). We strive to adhere to these principles when choosing our visualizations in order to communicate the facts most clearly (Tripp, 2019).

Figure 3 is for the first KPI, which is 'Revenue per region'. Since there are only two stores, there are only two regions to interpret and thus a table is chosen, for a high data to ink ratio. Also a table is chosen over a bar/line chart because it allows for clearer comparison of 2 values when there is only a slight difference between them (Tripp, 2019). It is clear that 'Australia' has received a total of 33.924,1 summed payment amounts while 'Canada' has received a total of 33.482,5 summed payment amounts. This means that 'Australia' performs better than 'Canada' but the difference is very small. Therefore we can conclude that both regions perform about equal.

Figure 3

Figure 4 is for the second KPI, which is 'Revenue per month' which is important as the business can use this as an indication for growth. We chose a line chart here and added specific data points in order to make the revenue totals easily comparable while showing the month-on-month trend from which the growth can be inferred. Analyzing this growth, we can observe that from May to July, monthly revenue has increased from 4.823 to 28.369. This rise of 488% in monthly revenue shows that the company has shown an increasing trend in performance. However, from July to August, monthly revenue has decreased from 28.369 to 24.070. This drop of 15% in monthly revenue could be worth investigating. When combined with managerial day-to-day knowledge of strategic decisions and operations, this information could be used to inform a reaction to this drop. For example, for example, a drop in efficiency of promotion efforts, which might signal the need for new marketing campaigns. Interesting to note is that this drop in revenue occurred during summer, so this might indicate a seasonality effect.

Finally, the single data point on the right for February is data for only one single day. Either February is an extreme outlier, or its main relevance is to point out that data is missing from August to February.

Figure 4

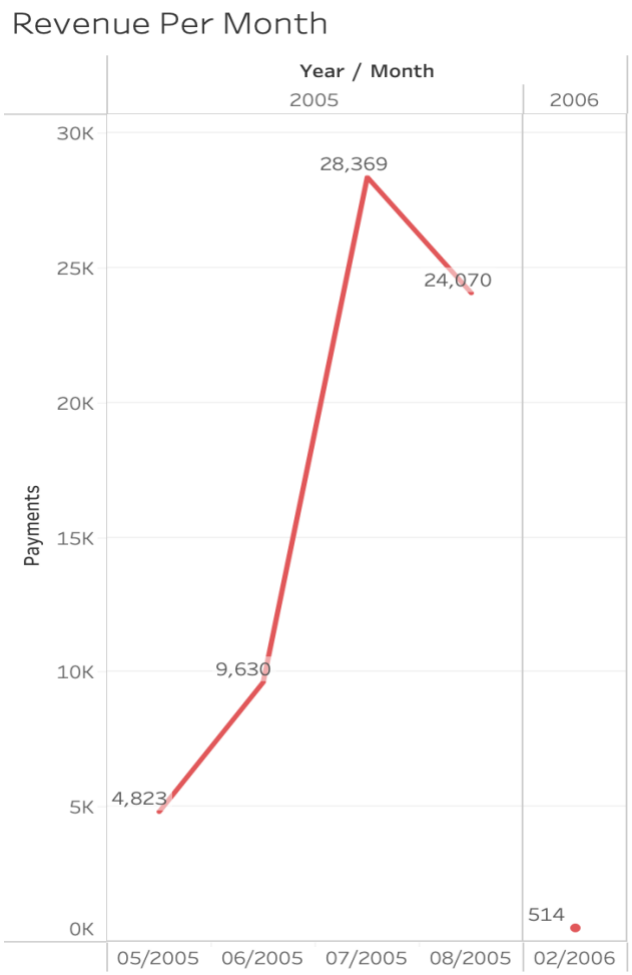


Figure 5 and 6 are for the third KPI, which is 'Customer per country'. In figure 5, dots with a greater size represent more customers from that specific country. Judging from the size of these dots, we can see that most customers are from China and India. The table (figure 6) with corresponding specific customer figures per country confirms this. We chose these visualizations to make the comparison of the top 10 countries easy (table) and we chose the world map in order to give a rough overview of countries which have less customers in them (instead of a detailed overview e.g. in a table, because these countries are less relevant to answer the business question). Customer numbers in other countries shown in descending order from most to least are the U.S., Mexico, Brazil, Russia, and Japan. When looking at figure 6, which is a table containing the exact number of customers per country, we observe that the top countries align with the previous observations in graph three, but specify the rest of the top 10 (the countries Philippines, Turkey, and Indonesia) more clearly. This information can be used to inform strategic recommendations where the business can best expand to (as it already has customers over there which could be better serviced and might be an indication for future

demand). Another strategic recommendation could be to focus marketing strategies on these ten countries.

Figure 5

Figure 6

Customer Per Country

Country	
India	60
China	53
United States	36
Japan	31
Mexico	30
Russian Federation	28
Brazil	28
Philippines	20
Turkey	15
Indonesia	14

Figure 7 is for the fourth KPI, which is the amount of 'active customers' versus the 'total customers'. A table is again chosen for a high data to ink ratio and easy comparison of the values (Tripp, 2019). This graph shows that both stores contain a total number of 599 customers in total, of which, 584 customers are active. This means that there are about 15 inactive customers in total, of which we have address information. This could be an opportunity to increase the total amount of active customers per store which will increase the revenue per customer. This company can utilize address information of inactive customers to try and convert them into active customers by sending them promotional material or advertisements by mail.

Figure 7

Active Customer VS Total Customer

Count of customer_table	599.00
Customer Activity	584.00

For the sixth KPI, we chose Revenue per category per day rented out (Figure 8). A barchart is chosen to give a visual intuition of the magnitude of the differences between categories in a single dimension, with values added to allow direct comparison between categories (Tripp, 2019). The sakilla dataset contains little direct data on business processes, so in order to add perspective to the business inventory management we use revenue per category per day rented out as a proxy for inventory efficiency. It shows which categories are contributing the most to the bottom line per day rented out and thus which categories most efficiently contribute to gross

margin. The top five best performing film categories in this sense are ‘New’, ‘Comedy’, ‘Sports’, ‘Sci-Fi’ and ‘Horror’. The worst performing film categories are ‘Travel’, ‘Children’ and ‘Family’. These results can provide guidance for the film (category) purchasing policy which is also a key internal business process for a DVD rental store. Categories which have higher revenue per day rented out give an indication for revenue generation potential, either by soliciting more revenue per rental (because customers are willing to pay more) or having a lower number of days rented out (meaning they can be rented out again soon and generate a rental payment quicker). Conversely, categories which have low revenues per day rented out, either solicit lower payments per rental (because customers are unwilling to pay more) or have a higher number of days rented out (meaning they are not using the inventory space reserved for them as efficiently because they are in the customers home). So if the inventory is to be managed most efficiently, it could make sense to purchase more of the high revenue generating categories. These categories give the highest return per day rented out as they both use storage space more efficiently, reduce storage costs, are more in demand (customers willing to pay more), and as a consequence generate more revenues.

Figure 8

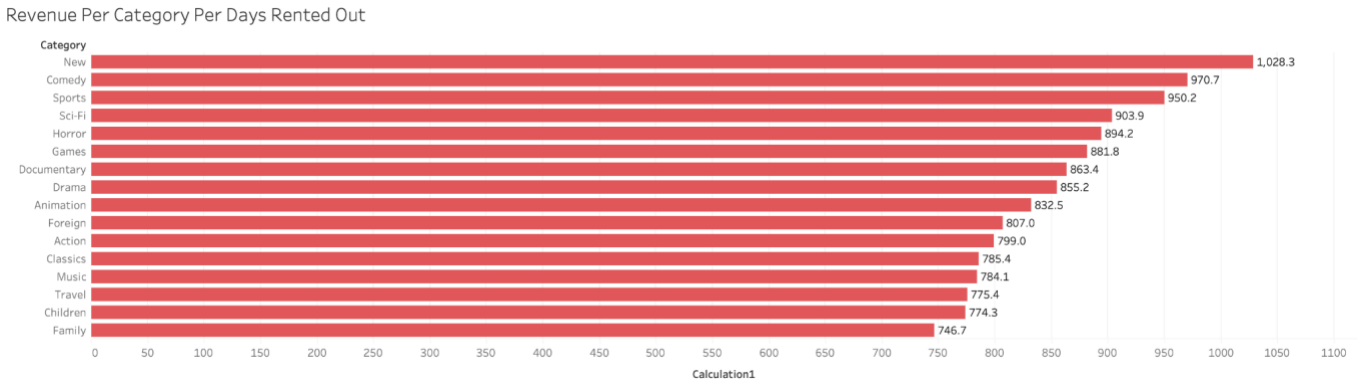


Figure 9 is for the sixth KPI, Rental Duration per Film. This is a proxy for the most popular films. A table is again chosen for a high data to ink ratio and easy comparison of the values (Tripp, 2019). The reason we chose days rented out instead of simply listing the films with the most transactions, is in order to use this measure to inspect inventory efficiency as a business process. This can be done by recognizing that a copy of a film which is rented out, cannot be rented out again in the meantime. This means that in order to exploit the popularity of this film in order to generate maximum revenue, more copies should be purchased. This indicator can in that way guide the purchasing policy & process of the business. This graph shows the total

rental duration per specific film for the top films in this KPI. These are, from best to worst, ‘Bucket Brotherhood’, ‘Ace Goldfinger’, ‘Robbers Joon’, ‘Shock Cabin’, ‘Muscle Bright’, ‘Married Go’, ‘Moon Bunch’, ‘Deep Virginian’, ‘Trip Newton’, ‘Story Side’, ‘Scalawag Duck’, ‘Forward Temple’, ‘Primary Glass’, ‘Operation Operation’, ‘Lies Treatment’, ‘Forrester Comancheros’, ‘Curtain Videotape’, ‘Blackout Private’, ‘Westward Seabiscuit’, ‘Island Exorcist’ and finally, ‘Contact anonymous’. These twenty films have been rented for either almost or over two hundred days, and this is indicative for movie popularity. These results can then guide the internal business of promotion by informing the company which movies are likely to be popular with consumers and thus good candidates for promoting, for example to inactive customers.

Figure 9

Rental Duration Per Film

Film name	
BUCKET BROTHERHOOD	238
ROBBERS JOON	217
SHOCK CABIN	210
MUSCLE BRIGHT	210
MARRIED GO	210
MOON BUNCH	203
DEER VIRGINIAN	203
TRIP NEWTON	196
STORY SIDE	196
SCALAWAG DUCK	192
FORWARD TEMPLE	192
PRIMARY GLASS	189
OPERATION OPERATION	189
LIES TREATMENT	189
FORRESTER COMANCHER..	189
CURTAIN VIDEOTAPE	189
BLACKOUT PRIVATE	189
WESTWARD SEABISCUIT	182
ISLAND EXORCIST	182
CONTACT ANONYMOUS	182

Figure 10 is for the seventh KPI, which is ‘Workload per Staff’. This KPI is measured by summing the total amount of rentals per staff member. As the difference in workload per staff member is close, and there are just 2 employees, we visualized it using a table. Here, we can see that the employee ‘Mike Hillyer’ has a workload of 8.054 rentals while employee ‘Jon Stephens’ has a workload of 7.990 rentals. This difference is small and based on the data.

We observe in the data that the staff members handle an average of 8.022 rentals per staff member over 3 months which comes down to an average of 87,7 per day or 11 per hour (assuming an 8 hour workday) and certainly seems quite busy. A personal interview with both employees might reveal that they personally experience their workload as being too high, which might have negative effects on revenue. If this is the case, then more employees should be hired. Based on just these 2 data points however, we exercise caution in saying the workload is too high or too low as we are not familiar with the in-store handling of purchasing processes, which might be partially automated. Based purely on the data, the only conclusion we can draw is that the relative workload does not differ significantly. Although we would recommend an employee interview to assess the perceived weight of the workload.

Figure 10

Workload Per Staff

Staff	
Jon Stephens	7,990
Mike Hillyer	8,054

Figure 11 is for our final eight KPI, which is 'Revenue per Staff'. This KPI is measured by summing the total amount of summed payments per staff member. A table is again chosen for a high data to ink ratio and easy comparison of the values (Tripp, 2019). It is clear that employee 'Jon Stephens' is responsible for generating a revenue of '33.924,1' while employee 'Mike Hillyer' is responsible for generating a revenue of '33.482,5'. This difference is small. The dataset gives no indication of variable costs (only replacement costs for lost movies) so it is hard to draw conclusions about profitability per employee as a proxy for employee productivity, which is what would be nice as based on that we can make an argument if investing in additional employee training makes sense financially. If we however use the revenues per staff as a proxy for profit, and consider the fact that each store has only 1 employee which is responsible for bringing in all the revenues of that specific store, investing in employee training might make sense. Both from an standpoint of improving efficiency, but also because investing in employee training increases employee retention and losing the only employee in the store could mean that store becomes non-operational (Sandhya, Kumar, 2011).

Figure 11

Revenue Per Staff

Staff	
Jon Stephens	33,924.1
Mike Hillyer	33,482.5

5. Conclusion

In summary, we have analyzed the performance of the Sakila DVD rental business. We did so by visualizing eight KPIs for data stored in a data warehouse created within a relational database for the Sakila dataset. These eight KPIs are 'Revenue per Region', 'Revenue per month', 'Customer per Country', 'Active customers vs inactive customers per Store', 'Revenue per category per day rented out', 'Rental Duration per Film', 'Workload per staff', and 'Revenue per staff'. Our results revealed the following important business insights.

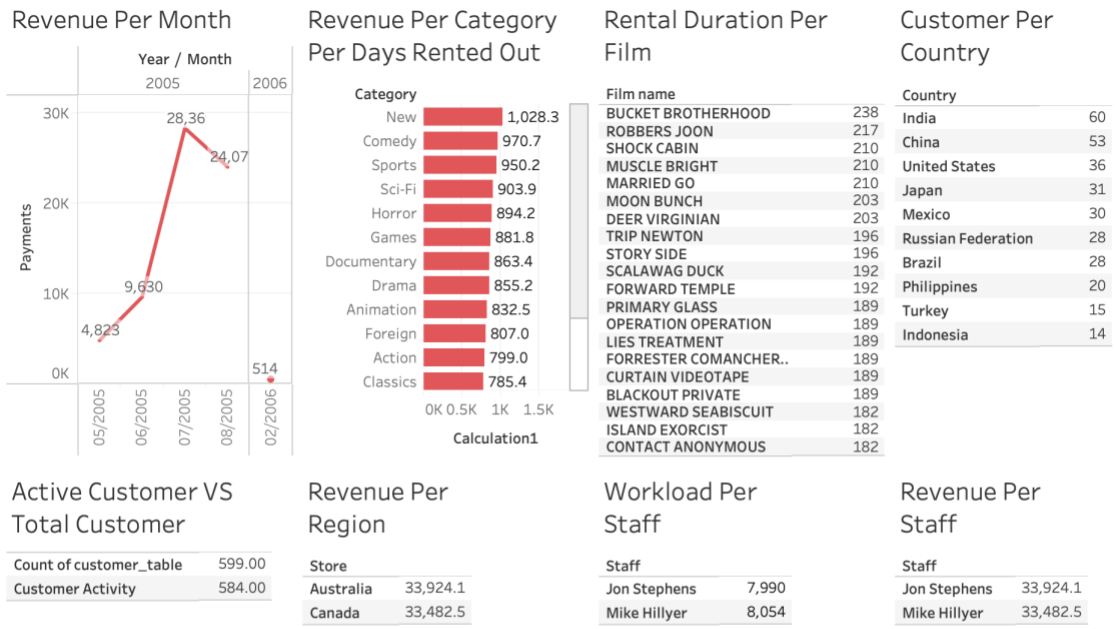
The monthly revenue of this company has increased 488% from May to July, but has decreased 15% from July to August. This drop can be attributed to different potential explanations such as the seasonality effect of the summer, or an unsuccessful marketing strategy. There is no significant difference in revenue per region, and therefore this is not a potential explanation for the drop in monthly revenue. However, when analyzing the amount of customers per country, we conclude that there are opportunities for expanding to other countries. With most to least customers in a country listed in descending order India, China, U.S, Japan, Mexico, Brazil, Russia, Philippines, Turkey, and Indonesia.

We also see opportunities for trying to increase monthly revenue, by directly targeting inactive customers. This could be done by using promotion to convert them into active customers, as the database contains both the addresses of these customers and their activity status. We could promote the most revenue generating categories per day rented out to them. Our analysis reveals that the top five most rented film categories are, from best to worst, 'New', 'Comedy', 'Sports', 'Sci-Fi' and 'Horror'. Of the entire film catalog, the ten best performing films are, from best to worst, 'Bucket Brotherhood', 'Robbers Joon', 'Shock Cabin', 'Muscle Bright', 'Married Go', 'Moon Bunch', 'Deer Virginian', 'Trip Newton', 'Story Side', and either 'Scalawag Duck' or 'Forward Temple' which both tie at 192, these could be promoted to the inactive customers as well.

We did not find any meaningful difference in workload per staff member. Since the workload seems high, the company might conduct personal interviews with these employees

to evaluate their perceived workload, which might gain more meaningful insights such as for example if they need training or if hiring more employees is needed. We show the relevant visualizations in our Tableau dashboard (figure 12) below:

Figure 12



6. Bibliography

Amrit, Chintan. (2021a). *Lecture 1, Intro to Databases*, lecture notes, Databases and Data Visualization 6012B0415Y, University of Amsterdam, delivered 26 October 2021.

Amrit, Chintan. (2021b). *Lecture 4, Data Warehousing*, lecture notes, Databases and Data Visualization 6012B0415Y, University of Amsterdam, delivered 16 November 2021.

Amrit, Chintan. (2021c). *Lecture 6, NO SQL Databases*, lecture notes, Databases and Data Visualization 6012B0415Y, University of Amsterdam, delivered 12 December 2021.

Domdouzis, Konstantinos, Peter Lake, and Paul Crowther. (2021). *Concise Guide To*

Databases: A Practical Introduction. Cham: Springer.

Jensen, Christian S. Pedersen, Torben B. Thomsen, Christian. (2010). "Multidimensional databases and data warehousing". Synthesis Lectures on Data Management, 2(1), 1-111.

Kaplan, Robert. S. Norton, David. P. (2005). "The balanced scorecard: measures that drive performance". Harvard business review, 83(7), 70-80.

Sandhya, Kethavath. Kumar, D. Pradeep. (2011). Employee retention by motivation. Indian Journal of science and technology, 4(12), 1778-1782.

Tripp, John F. (2019). *Data Visualization*. In: Pochiraju B., Seshadri S. (eds) Essentials of Business Analytics. International Series in Operations Research & Management Science, vol 264. Springer, Cham. https://doi.org/10.1007/978-3-319-68837-4_5

Volberda, Henk W. Morgan, Robert E. Reinmoeller, Patrick. Hitt, Michael A. R. Ireland, Duane. Hoskisson, Robert E. (2011). *Strategic Management Concepts and Cases Competitiveness*
and Globalization. Boston: Cengage Learning.

7. Appendix

1. **Query code for making data warehouse**

```
-- MySQL Workbench Forward Engineering

SET @OLD_UNIQUE_CHECKS=@@UNIQUE_CHECKS, UNIQUE_CHECKS=0;
SET @OLD_FOREIGN_KEY_CHECKS=@@FOREIGN_KEY_CHECKS,
FOREIGN_KEY_CHECKS=0;
SET @OLD_SQL_MODE=@@SQL_MODE,
SQL_MODE='ONLY_FULL_GROUP_BY,STRICT_TRANS_TABLES,NO_ZERO_IN_D
ATE,NO_ZERO_DATE,ERROR_FOR_DIVISION_BY_ZERO,NO_ENGINE_SUBSTITU
TION';

-- -----------------------------------------------------
-- Schema sakila_warehouse
-- -----------------------------------------------------
DROP SCHEMA IF EXISTS `sakila_warehouse` ;

-- -----------------------------------------------------
-- Schema sakila_warehouse
-- -----------------------------------------------------
CREATE SCHEMA IF NOT EXISTS `sakila_warehouse` DEFAULT CHARACTER SET
utf8 ;
USE `sakila_warehouse` ;

-- -----------------------------------------------------
-- Table `sakila_warehouse`.`Date_table`
-- -----------------------------------------------------
DROP TABLE IF EXISTS `sakila_warehouse`.`Date_table` ;
```

```
CREATE TABLE IF NOT EXISTS `sakila_warehouse`.`Date_table` (
  `Date_key` INT NOT NULL AUTO_INCREMENT,
  `Day` DATE NOT NULL,
  `Month` VARCHAR(45) NOT NULL,
  `Year` VARCHAR(45) NOT NULL,
  PRIMARY KEY (`Date_key`))
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `sakila_warehouse`.`Film_table`
-- -----------------------------------------------------
DROP TABLE IF EXISTS `sakila_warehouse`.`Film_table` ;

CREATE TABLE IF NOT EXISTS `sakila_warehouse`.`Film_table` (
  `Film_key` INT NOT NULL AUTO_INCREMENT,
  `Film_name` VARCHAR(45) NOT NULL,
  `Category` VARCHAR(45) NOT NULL,
  `Language` VARCHAR(45) NOT NULL,
  PRIMARY KEY (`Film_key`))
ENGINE = InnoDB;


-- -----------------------------------------------------
-- Table `sakila_warehouse`.`Store_table`
-- -----------------------------------------------------
DROP TABLE IF EXISTS `sakila_warehouse`.`Store_table` ;

CREATE TABLE IF NOT EXISTS `sakila_warehouse`.`Store_table` (
  `Store_key` INT NOT NULL AUTO_INCREMENT,
  `Store` VARCHAR(45) NOT NULL,
  `Staff` VARCHAR(45) NOT NULL,
  PRIMARY KEY (`Store_key`))
ENGINE = InnoDB;
```

```
-- -----------------------------------------------------
-- Table `sakila_warehouse`.`Customer_table`
-- -----------------------------------------------------
DROP TABLE IF EXISTS `sakila_warehouse`.`Customer_table` ;

CREATE TABLE IF NOT EXISTS `sakila_warehouse`.`Customer_table` (
  `Customer_key` INT NOT NULL AUTO_INCREMENT,
  `Country` VARCHAR(45) NOT NULL,
  `City` VARCHAR(45) NOT NULL,
  `Customer_Name` VARCHAR(45) NOT NULL,
  `Customer_Activity` INT NOT NULL,
  PRIMARY KEY (`Customer_key`))
ENGINE = InnoDB;



-- -----------------------------------------------------
-- Table `sakila_warehouse`.`Fact_Table`
-- -----------------------------------------------------
DROP TABLE IF EXISTS `sakila_warehouse`.`Fact_Table` ;

CREATE TABLE IF NOT EXISTS `sakila_warehouse`.`Fact_Table` (
  `Payments` DECIMAL(5,2) NOT NULL,
  `Replacement_Cost` DECIMAL(5,2) NOT NULL,
  `Rental_duration` INT NOT NULL,
  `Date_key` INT NOT NULL,
  `Film_key` INT NOT NULL,
  `Store_key` INT NOT NULL,
  `Customer_key` INT NOT NULL,
  INDEX `fk_Fact_Table_Date_table_idx` (`Date_key` ASC) VISIBLE,
  INDEX `fk_Fact_Table_Film_table1_idx` (`Film_key` ASC) VISIBLE,
  INDEX `fk_Fact_Table_Store_table1_idx` (`Store_key` ASC) VISIBLE,
  INDEX `fk_Fact_Table_Customer_table1_idx` (`Customer_key` ASC) VISIBLE,
```

```
  PRIMARY KEY (`Film_key`, `Store_key`, `Customer_key`, `Date_key`),
  CONSTRAINT `fk_Fact_Table_Date_table`
    FOREIGN KEY (`Date_key`)
    REFERENCES `sakila_warehouse`.`Date_table` (`Date_key`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_Fact_Table_Film_table1`
    FOREIGN KEY (`Film_key`)
    REFERENCES `sakila_warehouse`.`Film_table` (`Film_key`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_Fact_Table_Store_table1`
    FOREIGN KEY (`Store_key`)
    REFERENCES `sakila_warehouse`.`Store_table` (`Store_key`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION,
  CONSTRAINT `fk_Fact_Table_Customer_table1`
    FOREIGN KEY (`Customer_key`)
    REFERENCES `sakila_warehouse`.`Customer_table` (`Customer_key`)
    ON DELETE NO ACTION
    ON UPDATE NO ACTION)
ENGINE = InnoDB;


SET SQL_MODE=@OLD_SQL_MODE;
SET FOREIGN_KEY_CHECKS=@OLD_FOREIGN_KEY_CHECKS;
SET UNIQUE_CHECKS=@OLD_UNIQUE_CHECKS;
```

2. Query code for populating fact table and dimension tables

```
-- Making sure it does not fill full tables
DELETE FROM sakila_warehouse.date_table;
DELETE FROM sakila_warehouse.film_table;
DELETE FROM sakila_warehouse.store_table;
```

```
DELETE FROM sakila_warehouse.customer_table;
DELETE FROM sakila_warehouse.fact_table;

-- Filling Date_Table
insert into sakila_warehouse.date_table(Day,Month,Year)
select CAST(payment_date as DATE), month(CAST(payment_date as DATE)),
year(CAST(payment_date as DATE))
from sakila.payment group by CAST(payment_date as DATE);

-- Filling Film_Table
insert into sakila_warehouse.film_table(Film_name, Category, Language)
select f.title , c.name, l.name from sakila.film f
inner join sakila.language l on f.language_id = l.language_id
inner join sakila.film_category fc on f.film_id = fc.film_id
inner join sakila.category c on fc.category_id = c.category_id
order by f.title;

-- Filling Store_table
insert into sakila_warehouse.store_table(Store,Staff)
SELECT co.country, concat(sf.first_name,' ',sf.last_name) FROM sakila.store st
inner join sakila.staff sf on st.store_id = sf.store_id
inner join sakila.address a on st.address_id = a.address_id
inner join sakila.city ci on a.city_id = ci.city_id
inner join sakila.country co on ci.country_id = co.country_id;

-- Filling Customer_location_table
insert into
sakila_warehouse.customer_table(country,City,Customer_Name,Customer_Activity)
SELECT co.country, ci.city, concat(cu.first_name, ' ', cu.last_name), cu.active
FROM sakila.customer cu
inner join sakila.address a on cu.address_id = a.address_id
inner join sakila.city ci on a.city_id = ci.city_id
inner join sakila.country co on ci.country_id = co.country_id;
```

```
-- Filling fact table

insert into sakila_warehouse.fact_table(Payments,Replacement_cost, Rental_duration,
Date_key, Film_key, Store_key, Customer_key)
Select X.payment, X.replacement, X.duration,
D.Date_key,F.Film_key,S.Store_key,C.Customer_key
FROM (
        select p.amount as payment, f.replacement_cost as replacement, f.rental_duration as
duration,
   CAST(payment_date as DATE) as Day_filter, concat(cu.first_name, ' ', cu.last_name) as
C_filter,
   f.title as F_filter, concat(sf.first_name,' ',sf.last_name) as S_filter
        from sakila.payment p
        inner join sakila.rental r on p.rental_id = r.rental_id
        inner join sakila.inventory i on r.inventory_id = i.inventory_id
        inner join sakila.film f on i.film_id = f.film_id
   inner join sakila.customer cu on p.customer_id = cu.customer_id
   inner join sakila.staff sf on p.staff_id = sf.staff_id

        ) X

   inner join
   (
   select Date_key, Day from sakila_warehouse.date_table
   ) D ON D.Day = X.Day_filter

   inner join
   (
   select Customer_key, Customer_Name from sakila_warehouse.customer_table
   ) C on C.Customer_Name = X.C_filter

   inner join
   (
   select Film_key,Film_name from sakila_warehouse.film_table
```

```
) F ON F.Film_name = X.F_filter

inner join
(
select Store_key, Staff from sakila_warehouse.store_table
) S ON S.staff = X.S_filter
```

3. Screenshots of tables in data warehouse

Fact table

Payments	Replacement_Cost	Rental_duration	Date_key	Film_key	Store_key	Customer_key
0.99	20.99	6	80	1024	4	1057
0.99	20.99	6	76	1024	4	1062
0.99	20.99	6	83	1024	4	1115
3.99	20.99	6	82	1024	4	1302
0.99	20.99	6	78	1024	4	1324
1.99	20.99	6	88	1024	4	1382
0.99	20.99	6	86	1024	4	1429
0.99	20.99	6	70	1024	4	1454
0.99	20.99	6	83	1024	4	1564
3.99	20.99	6	82	1024	4	1577
1.99	20.99	6	87	1024	4	1604
0.99	20.99	6	84	1024	4	1610
1.99	20.99	6	69	1024	4	1620
0.99	20.99	6	76	1024	5	1031
0.99	20.99	6	91	1024	5	1067
0.99	20.99	6	91	1024	5	1184
0.99	20.99	6	85	1024	5	1193
0.99	20.99	6	73	1024	5	1275
0.99	20.99	6	82	1024	5	1367
0.99	20.99	6	66	1024	5	1368
1.99	20.99	6	98	1024	5	1434
1.99	20.99	6	75	1024	5	1510
3.99	20.99	6	78	1024	5	1541
9.99	12.99	3	83	1025	4	1162
4.99	12.99	3	91	1025	4	1168
8.99	12.99	3	78	1025	4	1203
4.99	12.99	3	97	1025	4	1389
8.99	12.99	3	79	1025	4	1556
9.99	12.99	3	74	1025	5	1105
4.99	12.99	3	77	1025	5	1294
2.99	18.99	7	66	1026	4	1123
4.99	18.99	7	77	1026	5	1024
2.99	18.99	7	93	1026	5	1048

Customer dimension table

	Customer_key	Country	City	Customer_Name	Customer_Activ...
▶	1024	Japan	Sasebo	MARY SMITH	1
	1025	United States	San Bernardino	PATRICIA JOHNSON	1
	1026	Greece	Athenai	LINDA WILLIAMS	1
	1027	Myanmar	Myingyan	BARBARA JONES	1
	1028	Taiwan	Nantou	ELIZABETH BROWN	1
	1029	United States	Laredo	JENNIFER DAVIS	1
	1030	Yugoslavia	Kragujevac	MARIA MILLER	1
	1031	New Zealand	Hamilton	SUSAN WILSON	1
	1032	Oman	Masqat	MARGARET MOORE	1
	1033	Iran	Esfahan	DOROTHY TAYLOR	1
	1034	Japan	Sagamihara	LISA ANDERSON	1
	1035	India	Yamuna Nagar	NANCY THOMAS	1
	1036	Turkey	Osmaniye	KAREN JACKSON	1
	1037	United States	Citrus Heights	BETTY WHITE	1
	1038	India	Bhopal	HELEN HARRIS	1
	1039	United King…	Southend-on-…	SANDRA MARTIN	0
	1040	Russian Fe…	Elista	DONNA THOMPSON	1
	1041	Nigeria	Kaduna	CAROL GARCIA	1
	1042	South Africa	Kimberley	RUTH MARTINEZ	1
	1043	Pakistan	Mardan	SHARON ROBINSON	1
	1044	Bangladesh	Tangail	MICHELLE CLARK	1
	1045	Morocco	Sal	LAURA RODRIGUEZ	1
	1046	Latvia	Liepaja	SARAH LEWIS	1
	1047	Argentina	Crdoba	KIMBERLY LEE	1
	1048	Pakistan	Shikarpur	DEBORAH WALKER	1
	1049	Holy See (V…	Citt del Vaticano	JESSICA HALL	1
	1050	Philippines	Davao	SHIRLEY ALLEN	1
	1051	India	Munger (Mon…	CYNTHIA YOUNG	1
	1052	Japan	Shimonoseki	ANGELA HERNAN…	1
	1053	Taiwan	Lungtan	MELISSA KING	1
	1054	India	Kamarhati	BRENDA WRIGHT	1
	1055	India	Jhansi	AMY LOPEZ	1
	1056	Italy	Alessandria	ANNA HILL	1

Date dimension table

Date_key	Day	Month	Year
64	2005-05-25	5	2005
65	2005-05-28	5	2005
66	2005-06-15	6	2005
67	2005-06-16	6	2005
68	2005-06-18	6	2005
69	2005-06-21	6	2005
70	2005-07-08	7	2005
71	2005-07-09	7	2005
72	2005-07-11	7	2005
73	2005-07-27	7	2005
74	2005-07-28	7	2005
75	2005-07-29	7	2005
76	2005-07-31	7	2005
77	2005-08-01	8	2005
78	2005-08-02	8	2005
79	2005-08-17	8	2005
80	2005-08-18	8	2005
81	2005-08-19	8	2005
82	2005-08-21	8	2005
83	2005-08-22	8	2005
84	2005-05-27	5	2005
85	2005-06-17	6	2005
86	2005-07-10	7	2005

Film dimension table

Film_key	Film_name	Category	Language
1024	ACADEMY DINOSAUR	Documentary	English
1025	ACE GOLDFINGER	Horror	English
1026	ADAPTATION HOLES	Documentary	English
1027	AFFAIR PREJUDICE	Horror	English
1028	AFRICAN EGG	Family	English
1029	AGENT TRUMAN	Foreign	English
1030	AIRPLANE SIERRA	Comedy	English
1031	AIRPORT POLLOCK	Horror	English
1032	ALABAMA DEVIL	Horror	English
1033	ALADDIN CALENDAR	Sports	English
1034	ALAMO VIDEOTAPE	Foreign	English
1035	ALASKA PHANTOM	Music	English
1036	ALI FOREVER	Horror	English
1037	ALICE FANTASIA	Classics	English
1038	ALIEN CENTER	Foreign	English
1039	ALLEY EVOLUTION	Foreign	English
1040	ALONE TRIP	Music	English
1041	ALTER VICTORY	Animation	English
1042	AMADEUS HOLY	Action	English
1043	AMELIE HELLFIGHT…	Music	English
1044	AMERICAN CIRCUS	Action	English
1045	AMISTAD MIDSUMMER	New	English
1046	ANACONDA CONFE…	Animation	English
1047	ANALYZE HOOSIERS	Horror	English
1048	ANGELS LIFE	New	English
1049	ANNIE IDENTITY	Sci-Fi	English
1050	ANONYMOUS HUMAN	Sports	English
1051	ANTHEM LUKE	Comedy	English
1052	ANTITRUST TOMAT…	Action	English
1053	ANYTHING SAVANNAH	Horror	English
1054	APACHE DIVINE	Family	English
1055	APOCALYPSE FLAMI…	New	English
1056	APOLLO TEEN	Drama	English

Store dimension table

Store_key	Store	Staff
4	Canada	Mike Hillyer
5	Australia	Jon Stephens