CSI-300-02
Python SQL Integration
Kimberly Benson, Michael Kerwood, Harrison Labrecque,

Abstract

The goal of this paper is to make recommendations for the DVD rental stores represented by the Sakila database. Within this paper exists multiple sections that analyze the behavior of customers, sales of genres, the effect of actors on film sales, and total store revenue. Based on these analyses, this paper makes numerous recommendations about how the stores could improve business performance. These recommendations include focusing on high rental volume months such as July and August and focusing on high-selling genres like Sports, Animation, and Action. The paper also notes the profit differences between stores, focusing on how both stores have similar overall revenue, and as a result, should have equal focus on growth.

1. Introduction

In this project, we will be acting as data analysts for a fictitious DVD rental company represented by the Sakila sample database that comes with MySQL. The Sakila database is a large collection of data with information detailing transactions between a DVD rental company and its customers. It contains information for its customers, staff, inventory, and the specifics of the content on the DVDs. Using this extensive database, we will extract actionable insights to understand and interpret trends in the data. The objectives of our analysis are as follows:

- Analyze customer behavior by extracting and visualizing data
- Evaluate film performance
- Assess actor performance
- Analyze revenue and payment trends

Using this analysis, we can better understand trends in the business, which will lead to insights on how to improve business practices for better performance.

2. Methodology

For this project, everyone used a combination of Python and SQL. SQL queries were used to extract the data from the Sakila database. The Sakila-based was connected to using the MySQL Connector library, and individual queries were executed using the Pandas library. The data was stored in a Pandas DataFrame object. Once inside the DataFrame object, the data was plotted using the Matplotlib and Numpy libraries to create bar plots and scatter plots.

Analysis of the data is split into multiple sections to focus on the previously mentioned objectives. The analysis is broken into four broader themes (customer and rental, film and category, actor and film, and revenue and store). Within each of these themes are 3 individual data analysis sections that include three different pieces: first, a relevant sub-question for the theme, second, a visualization of the sub-question's data, and third, a discussion of the sub-question's data. These discussions focus on the key insights that can be taken away from the data and its visualization.

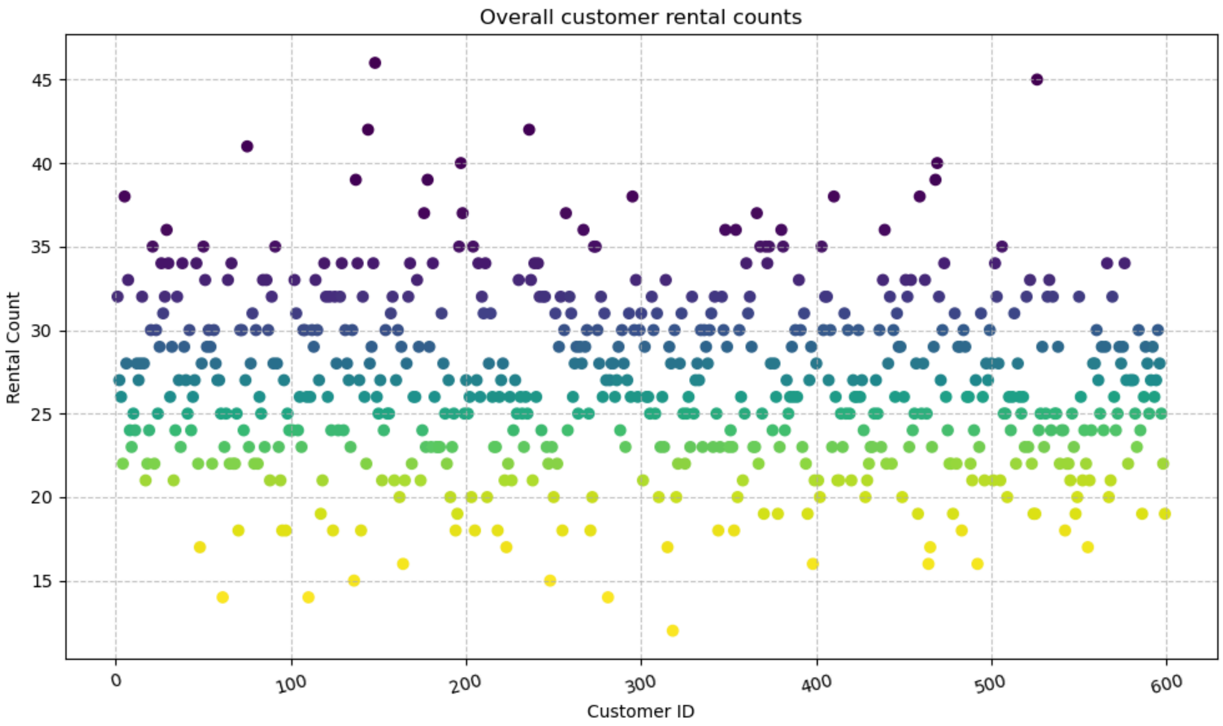
3. Analysis

3.1: Customer and Rental Insights

Analyze customer behavior by extracting and visualizing data related to:

- Total rentals per customer

Visualizations:



(Figure 1 - This graph shows the amount of rentals done per customer, organized by Customer ID)

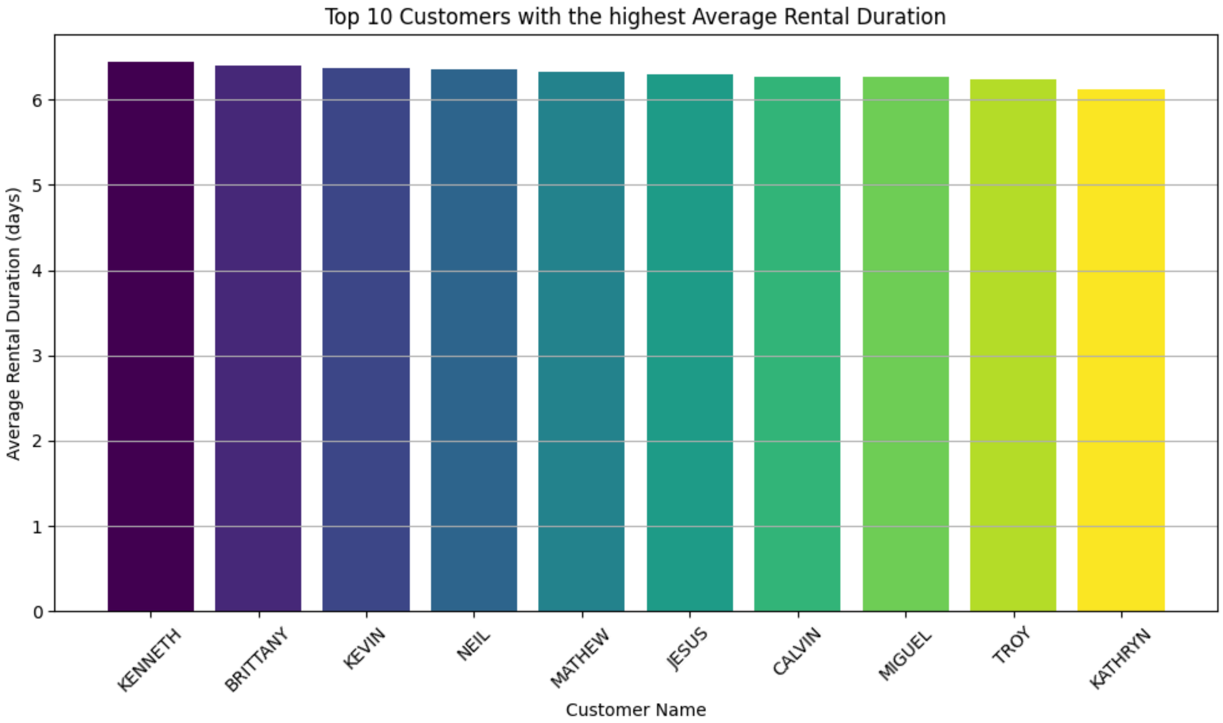
Discussion:

Based on Figure 1, we can understand the general range of rentals from customers. Generally, most of the data seems to fall within the range of 20 to 35 rentals per customer. So, pricing schemes should seek to target these customers. Prices per rental should be set at amounts that generate appropriate revenue from individual rentals because it is unlikely that individual customers will exceed 40 rentals.

Another conclusion that can be drawn from this data is that while there are outliers from the previously established 20 to 35 range, only 4 customers fall below 15 rentals. So, stores could try to offer bundle deals for rentals that encourage slightly more spending to push these customers above the 15 rental range, and as a result, push these outlier customers to rent more while simultaneously encouraging existing customers to push into the 30-35 range for rentals.

- **Average rental duration**

Visualizations:



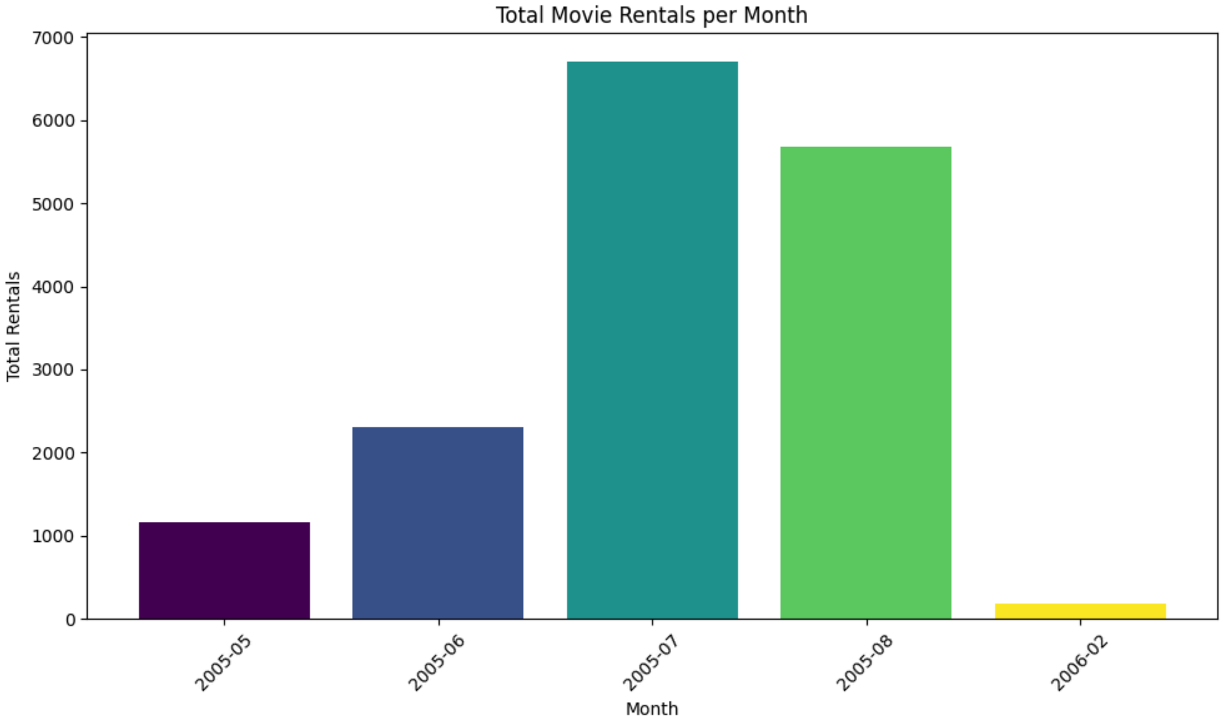
(Figure 2 - The top 10 customers with the highest average rental duration)

Discussion:

Figure 2 shows the top 10 customers with the highest average rental duration. Meaning, these are the customers who rent movies the longest. As rental duration is tied to the amount it costs to rent the movie, these customers are also the ones likely to pay the most amount of money. Everyone in the top 10 rents movies on an average of more than 6 days. Because of the length that these people are renting the movies they rent, they're likely very engaged with the kinds of movies they rent. So looking into what kinds of movies they rent and finding similarities between them would allow the business to know what movies to stock up on for these customers to enjoy. As they seem committed to renting movies, they'll likely notice any new additions that may pique their interest.

- Distribution of rental counts

Visualizations:



(Figure 3 - The total number of rentals per month)

Discussion:

Figure 3 shows the total number of rentals from the DVD shop per month in the year 2005. July was by far the most popular month to rent movies from, nearing the 7000 total rentals mark, with August following closely behind. No data was available for the months between August 2005 and February 2006, giving the graph a discrepancy in the data able to be provided.

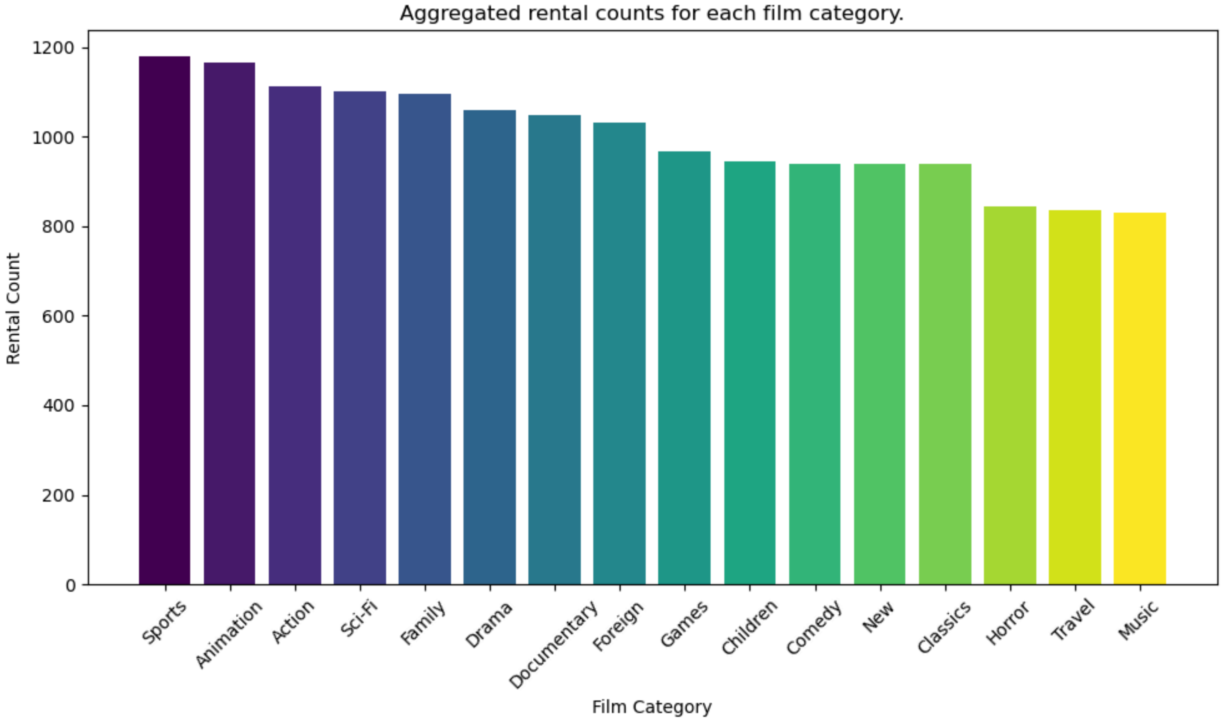
For the store, it would be beneficial to stock up around the July-August time period as that's when most of their rentals occur in the store. The May-June time frame starts to ramp up how many films get released, yet they come nowhere close to July or August, so it wouldn't make much sense to have a larger category during those months when the rentals are at a lower point.

3.2: Film and Category Analysis

Evaluate film performance by analyzing:

- Total rentals per film category

Visualizations:



(Figure 4 - This plot shows the rental count for each film category)

Discussion:

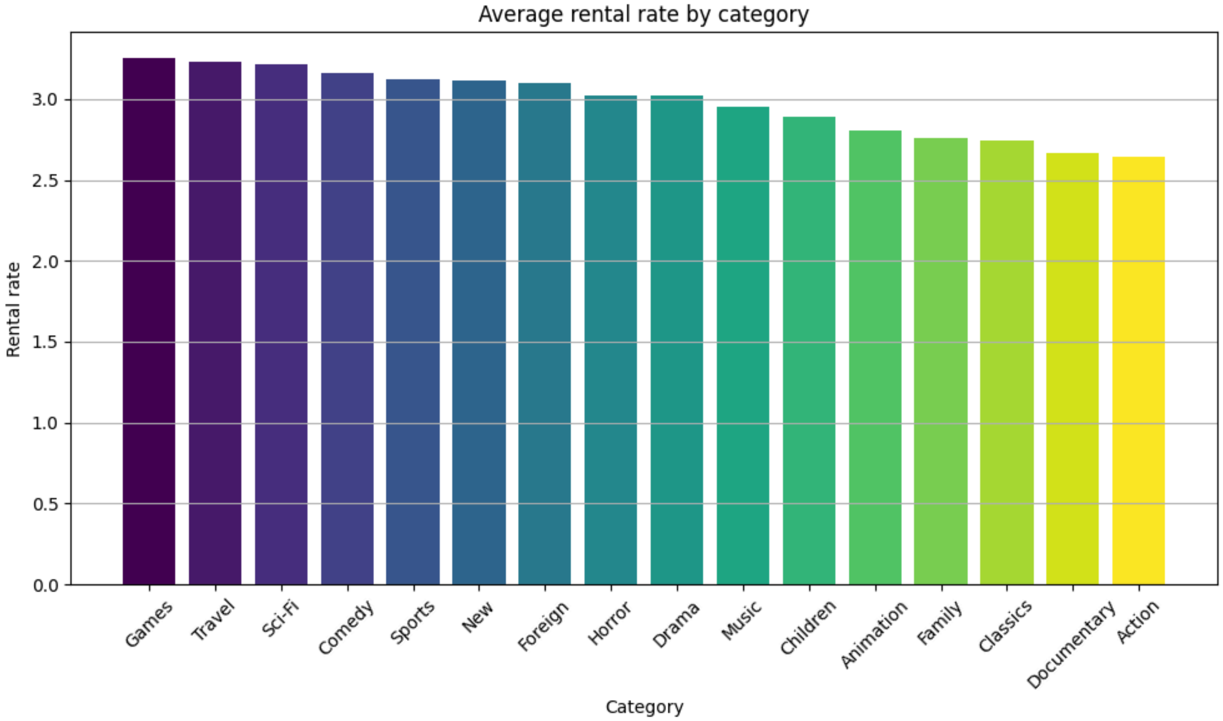
Figure 4 displays the rental count for each individual film category. The data allows for an understanding of what type of film generates the most rentals. As seen within the data, 7 out of the 15 total categories have over 1000 rentals, and none eclipse 1200 rentals. Based on this number, we can conclude that around half of the genres sell above 1000, and use 1000 as a metric for average sales. If a genre's rental count falls below 1000, this is a sign of it being a genre that generates low consumer investment. This includes all of the genres from Games to Music in Figure 4.

Figure 4 also displays that the bottom 3 genres (Horror, Travel, Music) all have a drop in sales compared to the bottom 50% of the dataset, with all 3 of them not reaching 900 sales, making them 200 away from the previously established metric for average genre sales. Therefore, these genres should have the least investment put into them, and have minimal inventory within the store.

Meanwhile, the top 3 genres (Sports, Animation, and Action) should all get more investment from the store and continue to be stocked, because they get the most rentals. However, the rest of the top 50% falls above the previously established 1000 rentals margin, so they should continue to be stocked.

- Average rental rate for each category

Visualizations:



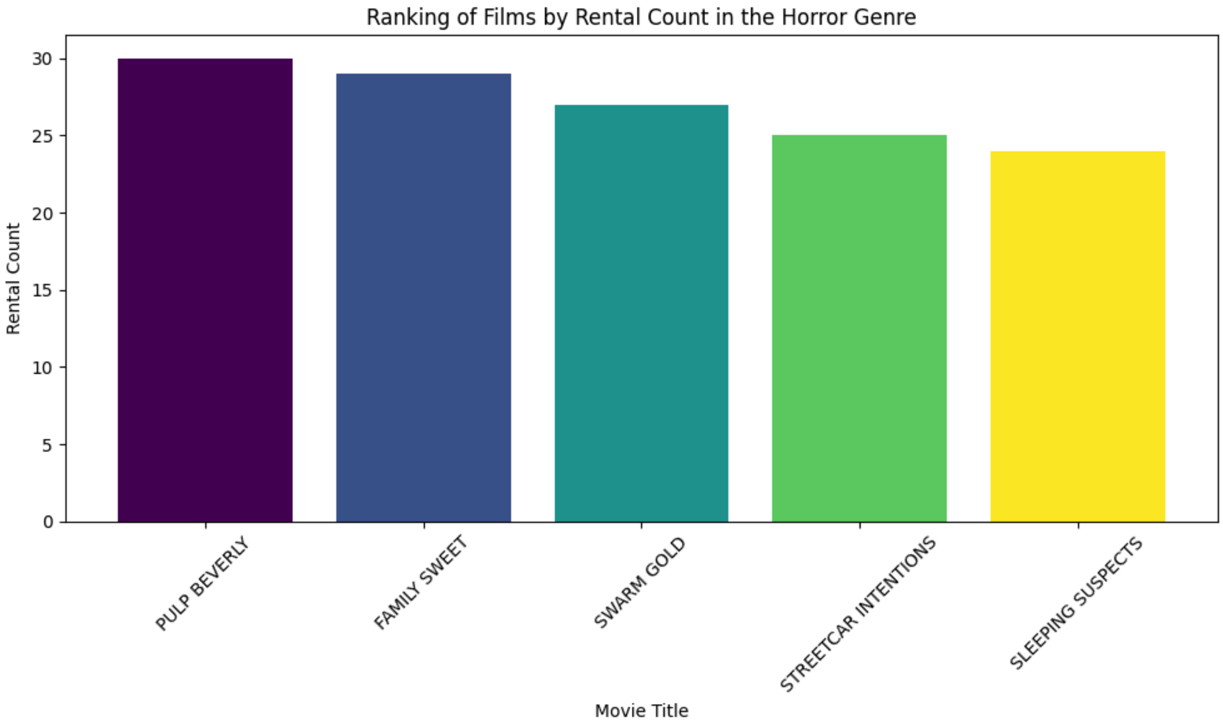
(Figure 5 - This plot is the relationship between categories and their average rental rate)

Discussion:

Figure 5 displays the relationship between a movie category and its rental rate. Rental rate is the cost to rent the film for its rental period. The higher the rate, the higher the cost to return it. The category with the highest average rental rate is ‘Games’, with a rental rate of 3.25. ‘Games’ having the highest average rental rate means that this category is the most expensive to rent for its given rental period. As seen in Figure 4, Sports is the category with the most amount of total rentals, while Games is the 9th most rented category. The rental rates of Sports and Games are both above average, however, Sports is far ahead of Games in terms of the number of rentals. Travel is also in a similar situation, with it being the 2nd lowest rented category as seen in Figure 4. Given the data here, it might be an advantageous idea to lower the rental rate for underperforming categories, as this may result in more overall rentals.

- The top 5 most rented films within one or more specific categories

Visualizations:



(Figure 6 - This plot shows the relationship between movies in the Horror Genre and rental count)

Discussion:

In Figure 6, there are the top five rented movies in the Horror genre. The rental count is the total number of times that each movie was rented out of the store. Of the movies seen in Figure 6, Pulp Beverly has the most rentals of all of the horror movies at 30 rentals total, and Sleeping Suspects has the least rentals of the top five horror movies at 24.

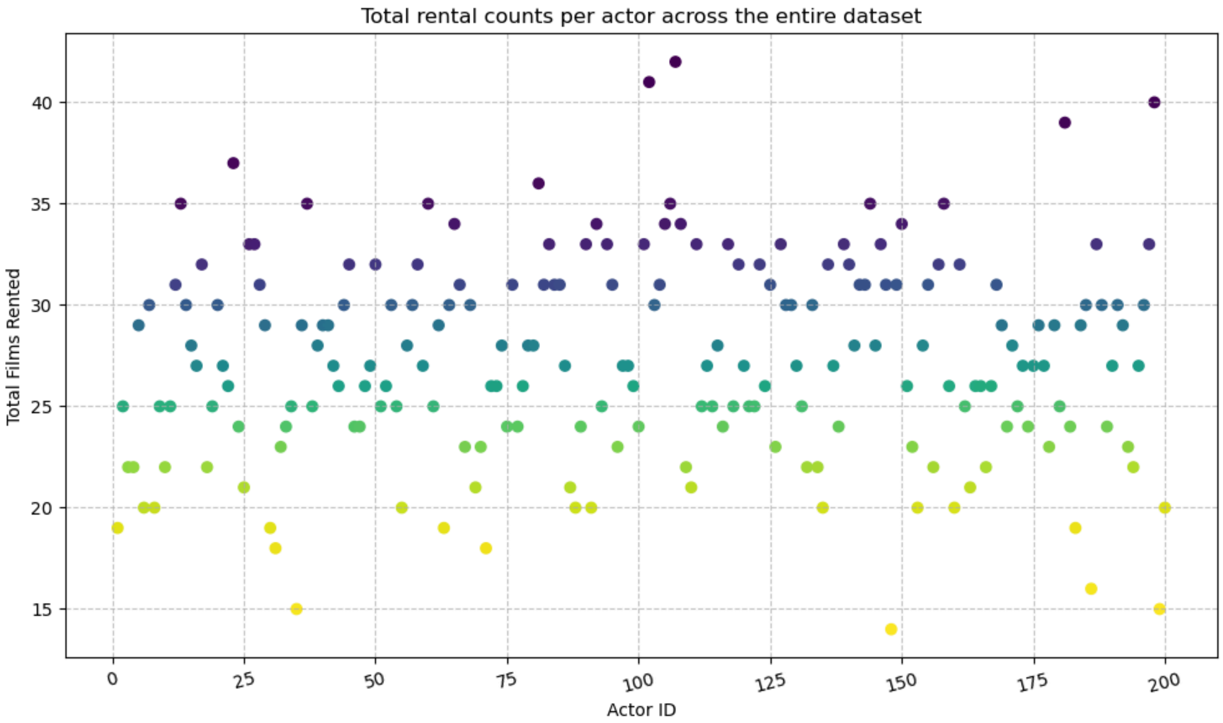
As seen in Figure 4, horror is the third least popular genre in the DVD rental store. As Horror has less than 900 total rentals in the store, it would be better for the store to only stock the most popular horror movies. Given their already small traction in the business, a smaller focus group on the most popular movies in this genre could assist in sales due to these few movies taking up the majority of the total rentals.

3.3: Actor and Film Performance

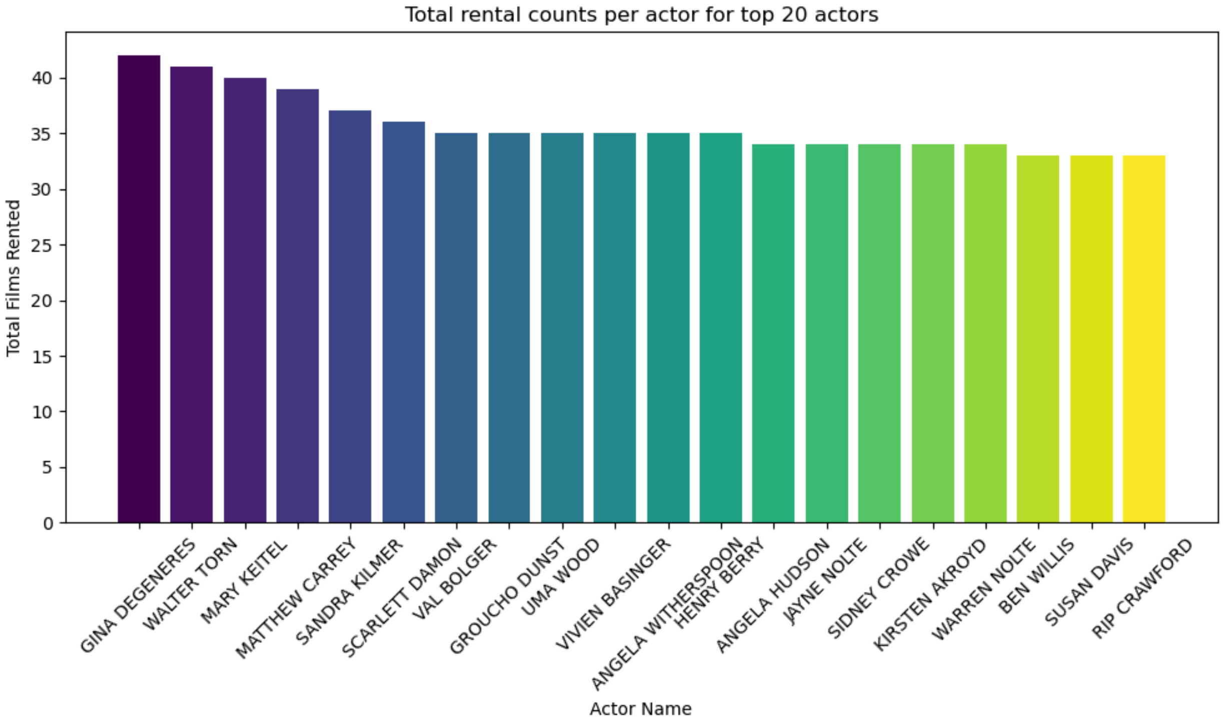
Assess actor performance by linking the rental performance of films to the actors featured in them. Your analysis should address:

- Overall rental counts associated with each actor.

Visualizations:



(Figure 7.1 - The total rental counts for every actor, represented with Actor ID, across the entire dataset)



(Figure 7.2 - Total rental counts for the top 20 actors within the dataset)

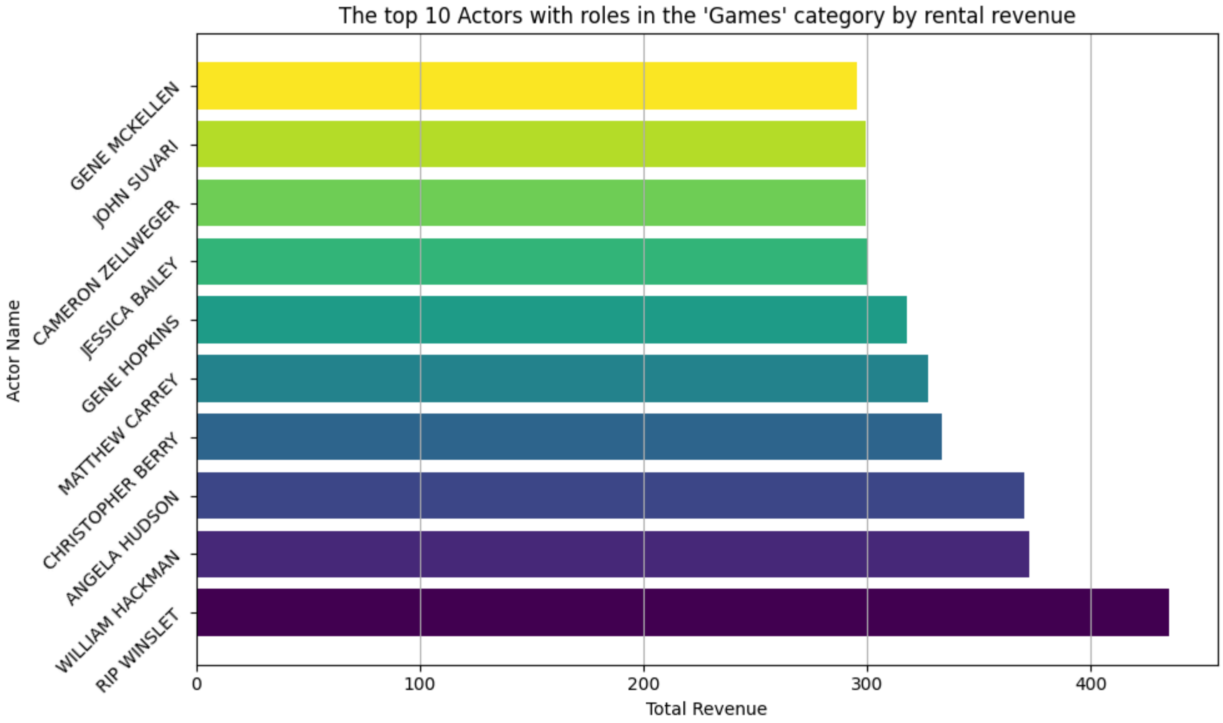
Discussion:

As showcased by Figure 7.1, there is a large range of total rental counts per actor, however, most actors fall within the 20 to 35 range. So, based on this distribution, it can be assumed that while certain actors may be more or less appealing than others, as a whole, the choice of actor doesn't massively affect the rentals for movies. There is a wide range of actors to choose from, so, the stores should stock movies with a wide range of actors to choose from to appeal to a wide market.

Figure 7.2 further proves this point, because it shows that within the top 20 actors, there is a less than 10 rental difference between the top actor and 20th actor. So, this means that all of the top 20 actors should be stocked, with a focus on the top actor, Gina Degeneres, but also the rest of the top 20.

- Comparative performance of actors within a specific film category.

Visualizations:



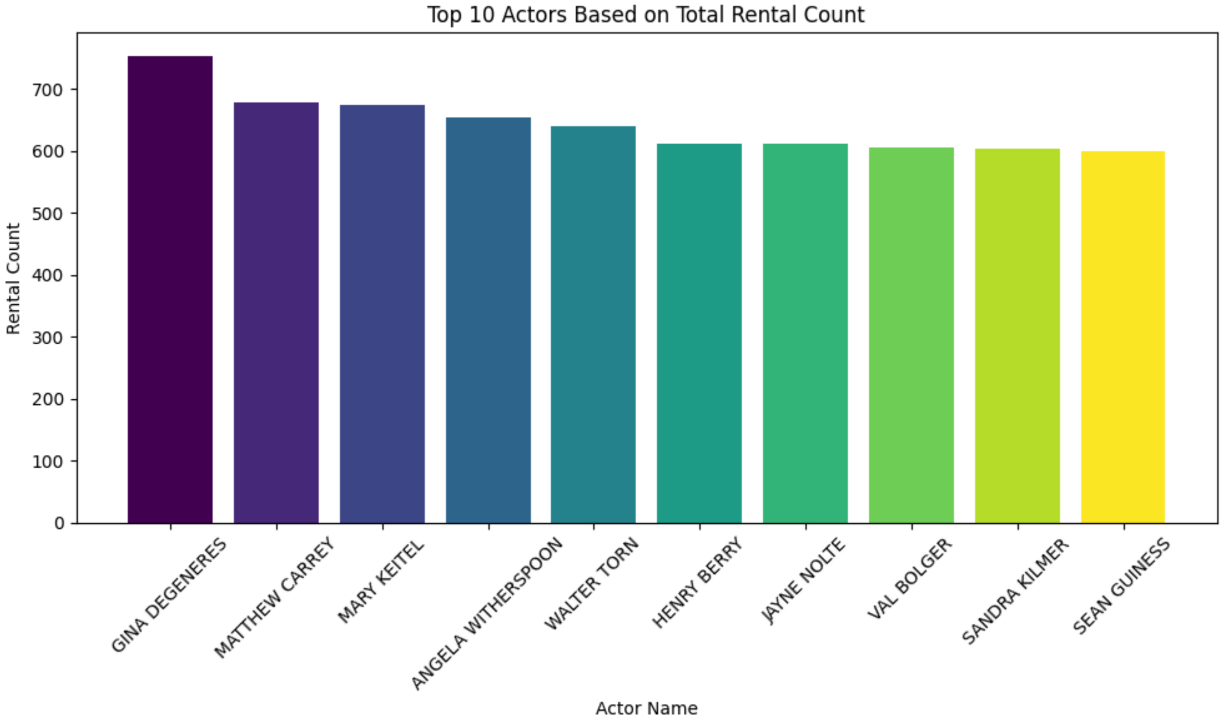
(Figure 8 - The top 10 actors in the 'Games' category by total rental revenue)

Discussion:

Figure 8 shows actors with roles in the 'Games' category and the collective total rental revenue those movies have. Using this figure we can see the popularity of specific actors in specific genres. Basically, this chart shows how well specific actors perform in specific genres. So when assessing what movies to add to the store's inventory, we could take a look if, for example, Rip Winslet is an actor in it. If they are, then that would mean that movie would likely be one that makes more money than one with another actor in it. This would also work for advertising, as marketing popular actors would lead to people renting more of those movies.

• Identification of the top 10 actors based on aggregated rental counts.

Visualizations:



(Figure 9 - This graph displays the relationship between the actors and how many times their films have been rented out and displays the 10 that have been rented out the most)

Discussion:

Figure 9 shows a relationship between movie actors and how often their movies have been rented. The graph counts the total number of rentals that each actor has, represented by ‘Rental Count’, and limits it to the top 10 actors with the most rentals. The more often the actors' movies are rented, the further left they are in the bar graph. As seen in Figure 9, Gina Degeneres has the most number of rented movies with all but one cracking the 600 rented movies milestone.

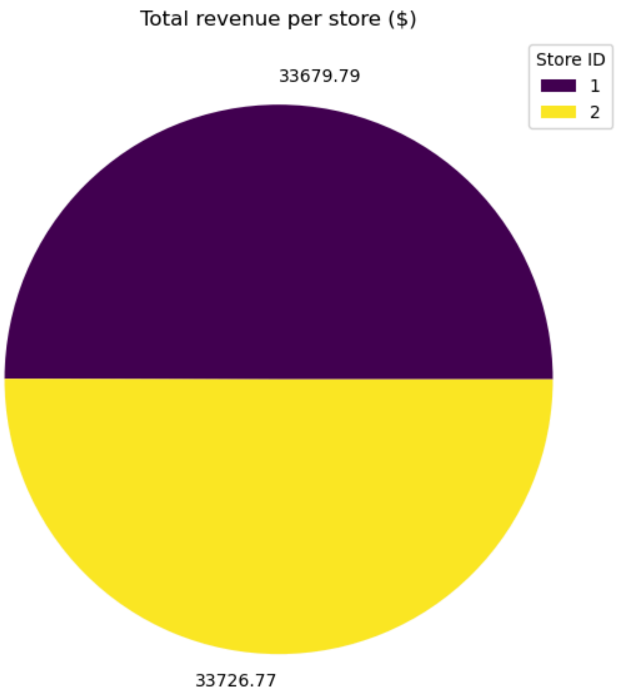
Based on the graph, it would be ideal for the DVD rental store to stock more movies that contain these 10 actors. In addition, they could create promotional materials to better promote the films that contain these actors in an effort to boost sales even further. As these actors are the most popular, focusing on getting their movies containing these actors could boost the store’s sales if the relationship between the actors and the total number of rentals of movies with the actor in it share a relationship not influenced by outside factors.

3.4: Revenue and Store Analysis

Perform an advanced analysis of revenue and payment trends by:

- Calculating total revenue per store.

Visualizations:



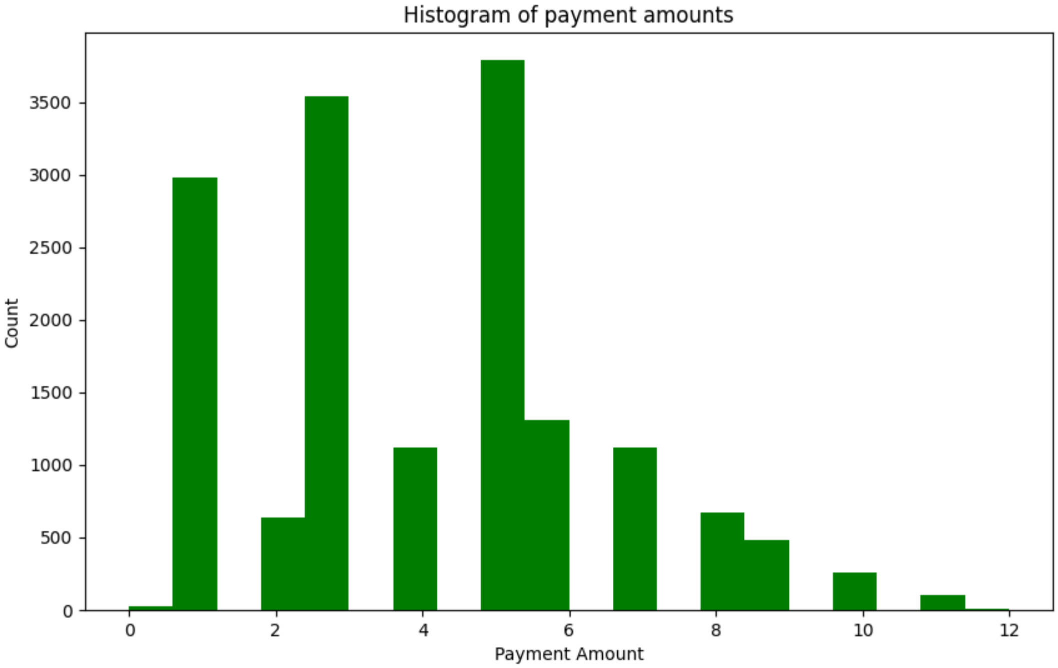
(Figure 10 - This graph shows the overall revenue generated per store in terms of dollars. Store 1 generated $33,679.79 and Store 2 generated $33,726.77)

Discussion:

Based on Figure 10, it can be observed that both stores make similar amounts. Both stores make around $33,000 each. This means that as a whole, the business has made around $66,000. So, it is worth equally investing in both stores to try to continue to keep profits up. If both stores keep up their current trends with equal investment, they will grow at similar rates. So, based on this existing relationship, it can be concluded that if one store doubles in revenue, it is likely that a similar effect may happen to the revenue of the other store.

- Analyzing the average payment amount per rental.

Visualizations:



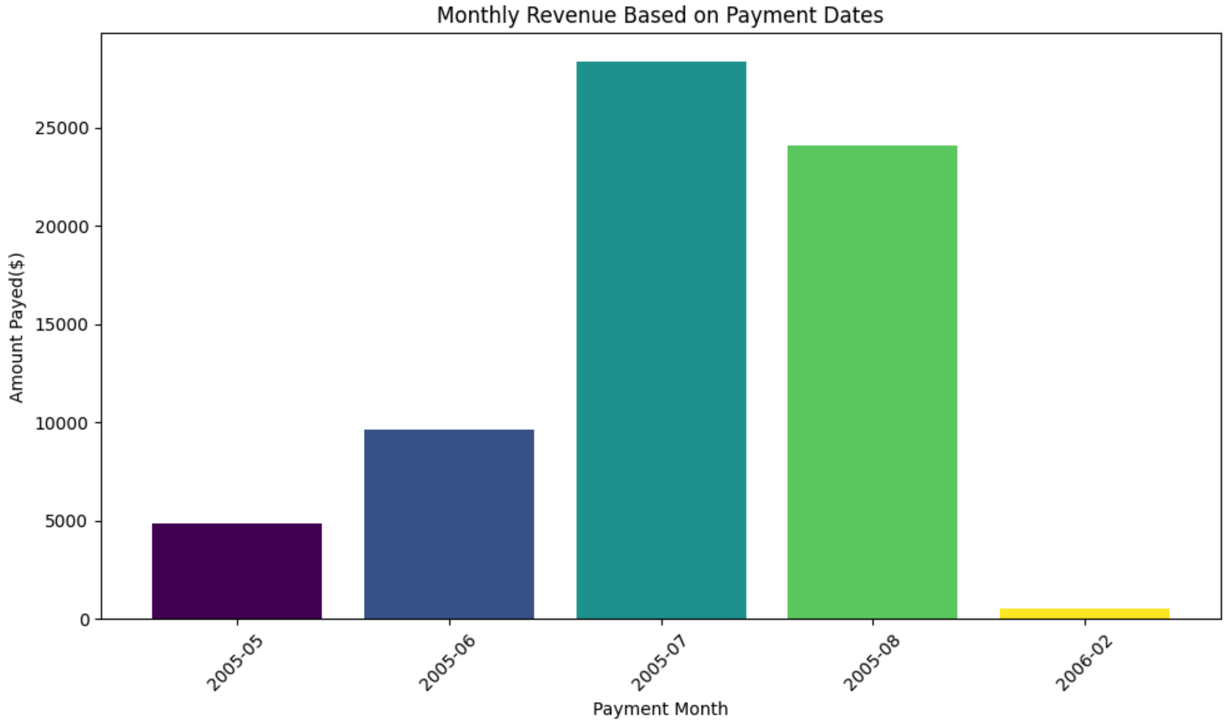
(Figure 11)

Discussion:

Figure 11 is the distribution of the payment amount for every rental transaction. As you can see, the payment amount generally follows a normal distribution. However, three different price points have sold the absolute best. That being 4.99 with 3789 transactions, 2.99 with 3542 transactions, and 0.99 with 2977 transactions. 64% of all transactions had an amount of these prices. More than half of the customers who have rented anything have paid this price. Meaning, that these are the prices that customers are the most willing to pay. The other prices follow a nice normal distribution. This tells us that customers are less willing to rent something the more the renting costs. A strategy to increase rental sales would be to reduce some of the higher prices to a more recognizable number, like the three largest listed above. This way, more customers would be willing to rent more movies, leading to a higher profit overall.

- Tracking monthly revenue trends.

Visualizations:



(Figure 12 - The total monthly revenue gained per month)

Discussion:

Figure 12 displays the monthly revenue based on when the payment came into the DVD business. The July-August timeframe had the highest revenue gained with July being over 25,000 dollars and August falling just shy of that same milestone. The data here lines up nearly identically to the data in Figure 3 (Total Movie Rentals Per Month), as both cover the same time frame for a similar statistic. While Figure 12 covers the payment that the shop receives, Figure 5 shows how many rentals the shop was able to sell per month.

Using these two figures, the shop can see that even though July wasn’t too far above August in rentals, the revenue gained in July surpassed the amount gained in August by a large amount. The other months in Figure 12 also follow a similar trend to the months in Figure 3, so what could be said about one can be said about both. For the shop, it would be ideal to promote more movies or bring in some with more popular actors like the ones seen in Figure 9 (Top 10 Actors based on rental count) to help further drive up the rentals to bring in a larger revenue stream to the business.

4. Individual Contributions (Harry, Kim, Michael)

4.1 Harry

I was responsible for 4 queries total, the second one in each section to be exact. Along with the queries I also made corresponding figures for each one. I also wrote the introduction and did some final formatting of the document.

My overall experience with this project was good. My only major problem was trying to figure out how to join all of the necessary tables I needed for a given query. Other than that, the process of figuring out what features I needed wand here to get the query result I wanted came quite easily to me. The fact that the query questions were very open-ended at first made this project a bit more difficult to begin, as starting an analysis without much bearing on a dataset is quite hard. But once I got through a query or two and ran some statistics tests to better understand the data things became more clear. One thing I really appreciated about this project was the UML diagram that was made with MySQL. Seeing the database in a connected, and visual manner was extremely helpful when trying to figure out what tables were related to each other. Without it, joins would have been 10x harder. Additionally, the documentation for the Sakila database was a very important tool. As some of the variable names, I think, can be interpreted differently than what they actually represent, it was a very valuable resource to have in understanding the database. I’ve learned a lot after doing this project. Getting actual practice with SQL and joins specifically has cleared up many of the misunderstandings I had when going into this project. Actually being able to apply things we've been doing in class has been great for my understanding of SQL.

4.2 Kim

I was responsible for 5 queries in the entire project, specifically the first one of each section. Alongside the queries, I also created data visualizations for each of my queries. I handled 4 of the subsections within the analysis section, but for 3.3, I did two queries to create two different visualizations. Outside of my analysis, I also did the abstract and methodology sections.

Overall, I had a great experience with this project. I have experience making data visualizations and presenting them, but they’re more focused on a research angle rather than a business one. So, it was interesting to see how to make appealing visualizations for business decisions. It was particularly exciting making visualizations for the total rental counts for actors because I struggled to figure out how to represent such a large dataset, which led to me discovering how I could use 2 visualizations at once for the same set in different ways to prove one argument. Getting to learn how to use the queries for such a big database was a challenge at first, but by the end of the project, I felt like I had overcome it. The hardest query for me ended up being the 2nd one, because I had to use so many SQL joins in one query to connect the data. I also struggled with the revenue one at first, because there were so many options for how I wanted to string together my joins, so it took a while to figure out the most optimal way, and even then, there might be a way to do it with less joins that I didn’t realize. So, I really enjoyed that challenge, because the solutions to my problems didn’t feel obvious, and sometimes I had to consult the diagram created by MySQL and draw out how I was going to solve it.

4.3 Michael

I was responsible for four of the queries done on the project, the third one in each section. I also created the graphs associated with each query, and I also wrote the conclusion for the paper.

I enjoyed working on this project. It made me think differently when trying to come up with the queries for each of the questions I was trying to answer for each one. It was a struggle to get to this point, as some of the queries required a lot While not a huge chunk of my time put into the project, I did spend a good amount just looking through the Sakila database before making the queries to have an idea of which tables I would need to use for each one so I wouldn't be scrambling around looking through each one for every query. I think that helped the most as it also led to me understanding what was in the database available for me to use and what ways I could go about joining these tables before really putting in the effort to write the queries themselves. Outside of this, a couple of my queries had me stumped for a bit due to needing to join several tables together to get them to work. It took me a decent amount of time to figure out that I needed to join several tables together and then getting somewhat comfortable with joining all of these tables together was another thing in itself. At the end of this project, I do feel more confident in creating queries in MySQL based on a database and somewhat comfortable with using joins, but I do feel like I will need more practice going forward with some more complex queries to understand parts of joining together better than I currently do.

5. Conclusion and Recommendations (Michael)

Our analysis of the DVD rental business through the use of the Sakila database has given us insight into customer behavior, rental trends, film and actor performance, and revenue. Through SQL-based queries and charts created through the use of Python with matplotlib, key patterns were able to be better extracted and put into a visual format to help better drive decision-making for the DVD rental business.

Key Findings

1. Customer and Rental

Most customers who rented movies fell within the 20-35-year-old demographic, suggesting a core group of customers who rent out movies more often. There are some outliers in this demographic who fall below the average amount of rentals. There is also a group of individuals who rent out films more often than others based on their average rental duration. The best time of year for the rental company seems to be in July or August, as those two months have the highest number of rentals happening within their months.

2. Film and Category

Some genres, such as Sports, Animation, and Action, consistently perform well with rental counts constantly exceeding 1000, while other genres such as Horror, Travel, and Music struggle to hit even 900 rentals. In addition, the Games genre has the longest average rental rate out of all of the genres the store stocks, while Action has the shortest average rental rate from the store. Within the categories, such as Horror, some films stand out from the rest with a much larger set of rentals when compared to other films even in the same genre.

3. Actor and Film

While the popularity of the actors may have a limited impact on the rentals, certain actors, such as Gina Degeneres show a greater rental frequency than some other actors. While other factors may have a greater influence over the choice of a rental, certain actors may attract a loyal audience who want to watch all of their movies. The appearance of a well-known celebrity may lead to greater revenue, but it doesn't appear to be a primary factor.

4. Revenue and Store

Both store locations generate around the same amount of money, indicating that the customers are split rather evenly between the two locations. From the rentals at the stores, three price points, those being $0.99, $2.99, and $4.99 are the most common prices for the DVDs to be rented out. In addition, the revenue gained per month seems to peak for the stores around July and August, while being at its worst during February.

Recommendations

1. Targeted Promotions

The store could implement bundle deals and loyalty programs to encourage customers to rent more often, incentivizing those who rent less frequently to rent more. These rewards could offer more frequent engagement among those who visit the store, leading to the store being more likely to be recommended to those people who live near one of the stores. Personalized promotions based on previous purchases could also help keep customers coming back for more.

2. Optimize Inventory Allocation

Focusing on keeping an inventory of the high-performing genres while also minimizing the investment in those that underperform could help maximize profits. Sports, Animation, and Action movies should have their inventory increased due to the popularity of the genres, while also keeping some inventory free in case one of the other genres becomes popular enough to warrant a larger inventory.

3. Leverage Actor Popularity

Promoting films featuring high-performing actors through things such as marketing campaigns and promotions would help boost customer interactions. Highlighting the popular actors can boost the number of rentals the movies that they starred in receive, as well as recommending them to other customers who have watched previous films that include an actor in them as well.

4. Seasonal Pricing and Planning

Keeping the pricing in the more common ranges for the payment amounts would help keep customers comfortable with the pricing and have them keep coming back to the stores. Also, having special promotions or additional options for rentals during the peak months of rental activity can further increase the profits gained by the stores if more people come to rent movies more often.