Rockbuster Stealth: Strategic Data Analysis for Digital Transformation

Executive Summary

Rockbuster Stealth LLC, a movie rental company facing disruption from streaming services, needed data-driven insights to inform their competitive strategy. This analysis examined their global customer base, rental patterns, and revenue distribution to identify opportunities for market expansion and customer retention.

Key Findings:

- Top 5 customers generated an average of $143.79 in revenue, representing high-value retention targets
- Customer concentration in 10 key countries (India, China, United States, Japan, Mexico, Brazil, Russian Federation, Philippines, Turkey, Indonesia) accounts for the majority of the customer base
- Rental pricing shows strategic variance ($0.99-$4.99) with an average of $2.98, closely aligned with the median
- Rental duration patterns reveal a consistent 5-6 day preference, suggesting weekly planning cycles in customer behavior

Business Impact:
These insights enabled Rockbuster to identify geographic expansion opportunities, optimize inventory allocation, and develop targeted retention strategies for high-value customers.

Business Context

The Challenge

Rockbuster Stealth, like many traditional rental businesses, faced an existential question: how to compete with streaming services that offer unlimited content at home. Rather than abandoning their physical rental model, Rockbuster needed to understand *why* customers still chose them and *who* their most valuable customers were.

Strategic Questions

The management team needed answers to several critical business questions:

1. **Geographic Distribution**: Where are our customers located, and which markets show the strongest engagement?

2. **Revenue Concentration**: Who are our most valuable customers, and what patterns do they exhibit?
3. **Inventory Optimization**: What rental durations and pricing strategies maximize revenue?
4. **Market Opportunity**: Which underserved markets represent expansion opportunities?

Why This Matters

Understanding customer behavior isn't just about describing what happened—it's about predicting what could happen and prescribing strategic actions. This analysis needed to move beyond "we have X customers in Y country" to "we should invest in Z because..."

Technical Approach

Data Infrastructure

The analysis leveraged a PostgreSQL database containing 15 interconnected tables:

- **Fact Tables**: Payment, Rental (transactional data)
- **Dimension Tables**: Customer, Film, Store, Address, City, Country, Actor, Category, Language, Inventory, Staff, Film_Actor, Film_Category

The relational structure required complex multi-table joins to connect geographic data with customer behavior and revenue information.

Data Quality Assessment

Before diving into analysis, I conducted systematic data quality checks:

```
-- Check for duplicates in customer data

SELECT customer_id, first_name, last_name, COUNT(*)

FROM customer

GROUP BY customer_id, first_name, last_name

HAVING COUNT(*) > 1;


-- Verify no NULL values in critical fields
```

```
SELECT
  SUM(CASE WHEN customer_id IS NULL THEN 1 ELSE 0 END) AS null_customer_ids,
  SUM(CASE WHEN email IS NULL THEN 1 ELSE 0 END) AS null_emails,
  SUM(CASE WHEN address_id IS NULL THEN 1 ELSE 0 END) AS null_addresses
FROM customer;
```

Finding: No duplicates or NULL values in primary customer data—a testament to well-maintained transactional systems. However, I documented what cleaning steps *would* be necessary if issues were found (flagging duplicates, investigating missing addresses, imputing reasonable defaults where appropriate).

Why This Matters: In real-world scenarios, data is messy. Demonstrating proactive quality checks shows I understand that analysis is only as good as the underlying data.

Key Analyses

1. Geographic Customer Distribution

Business Question: Where should we focus our marketing and inventory investments?

Approach: Connected customer data through four table joins (customer → address → city → country) to aggregate by geographic location.

```
SELECT
  E.country,
  COUNT(B.customer_id) AS customer_count
FROM customer AS B
INNER JOIN address AS C ON B.address_id = C.address_id
INNER JOIN city AS D ON C.city_id = D.city_id
INNER JOIN country AS E ON D.country_id = E.country_id
GROUP BY E.country
ORDER BY customer_count DESC
LIMIT 10;
```

Insight: The top 10 countries account for a significant concentration of customers, with India, China, and the United States leading. This isn't just a count—it's a market prioritization framework.

Business Implication: Rather than spreading resources thin across all markets, Rockbuster should double down on these high-density markets while exploring why others underperform. Are there language barriers? Supply chain issues? Competitive pressures?

2. High-Value Customer Identification

Business Question: Who are our most valuable customers, and where are they located?

Approach: This required a more sophisticated analysis—joining payment data with customer and geographic information, then filtering to top cities in top countries.

```
SELECT

  B.customer_id,

  B.first_name,

  B.last_name,

  D.city,

  E.country,

  SUM(A.amount) AS total_amount_paid

FROM payment AS A

INNER JOIN customer AS B ON A.customer_id = B.customer_id

INNER JOIN address AS C ON B.address_id = C.address_id

INNER JOIN city AS D ON C.city_id = D.city_id

INNER JOIN country AS E ON D.country_id = E.country_id

WHERE D.city IN (

  'Aurora', 'Atlixco', 'Xintai', 'Adoni', 'Dhule (Dhulia)',

  'Kurashiki', 'Pingxiang', 'Sivas', 'Celaya', 'So Leopoldo'

)

AND E.country IN (
  'India', 'China', 'United States', 'Japan', 'Mexico',
  'Brazil', 'Russian Federation', 'Philippines', 'Turkey', 'Indonesia'
)
GROUP BY B.customer_id, B.first_name, B.last_name, D.city, E.country
ORDER BY total_amount_paid DESC
LIMIT 5;
```

Insight: The top 5 customers averaged $143.79 in total payments—representing a small cohort with outsized revenue contribution.

Business Implication: These aren't just customers; they're loyalty candidates. What makes them different? Do they rent specific genres? Prefer longer rental durations? Understanding this cohort creates a blueprint for customer acquisition and retention strategies.

Follow-up Question I Asked: Are these high-value customers evenly distributed, or concentrated in specific markets? This led to the next analysis.

3. Customer Value Distribution by Country

Business Question: How are our top customers distributed geographically?

Approach: Used Common Table Expressions (CTEs) to create a reusable "top customers" dataset, then joined it back to all customers by country.

```
WITH top_5_customers AS (
  SELECT
    B.customer_id,
    B.first_name,
    B.last_name,
    D.city,
    E.country,
    SUM(A.amount) AS total_amount_paid
  FROM payment AS A
  INNER JOIN customer AS B ON A.customer_id = B.customer_id
  INNER JOIN address AS C ON B.address_id = C.address_id
  INNER JOIN city AS D ON C.city_id = D.city_id
  INNER JOIN country AS E ON D.country_id = E.country_id
  WHERE D.city IN (
    'Aurora', 'Atlixco', 'Xintai', 'Adoni', 'Dhule (Dhulia)',
    'Kurashiki', 'Pingxiang', 'Sivas', 'Celaya', 'So Leopoldo'
  )
  AND E.country IN (
    'India', 'China', 'United States', 'Japan', 'Mexico',
    'Brazil', 'Russian Federation', 'Philippines', 'Turkey', 'Indonesia'
  )
  GROUP BY B.customer_id, B.first_name, B.last_name, D.city, E.country
  ORDER BY total_amount_paid DESC
  LIMIT 5
)
SELECT
  E.country,
  COUNT(DISTINCT B.customer_id) AS all_customer_count,
  COUNT(DISTINCT top_5_customers.customer_id) AS top_customer_count
FROM customer AS B
INNER JOIN address AS C ON B.address_id = C.address_id
INNER JOIN city AS D ON C.city_id = D.city_id
INNER JOIN country AS E ON D.country_id = E.country_id
LEFT JOIN top_5_customers ON top_5_customers.country = E.country
GROUP BY E.country
ORDER BY top_customer_count DESC, all_customer_count DESC;
```

Why CTEs Matter: This could have been done with subqueries, but CTEs make the code more readable and maintainable. In a business context, readable code means transferable knowledge—critical when working with technical and non-technical stakeholders.

Insight: This revealed whether high-value customers were concentrated in a few markets or spread across many. The distribution pattern informs resource allocation—concentrated = targeted campaigns, dispersed = broad-based retention programs.

4. Inventory and Pricing Analysis

Business Question: How do our pricing and rental patterns compare across different film categories?

Approach: Used aggregate functions with GROUP BY to examine pricing variance by rating.

```
SELECT
  rating,
  ROUND(AVG(rental_rate), 2) AS average_rental_rate,
  MIN(rental_duration) AS min_rental_duration,
  MAX(rental_duration) AS max_rental_duration,
  MIN(replacement_cost) AS min_replacement_cost,
  MAX(replacement_cost) AS max_replacement_cost
FROM film
GROUP BY rating
ORDER BY rating;
```

Insight:

- Rental rates vary from $0.99 to $4.99, with the average ($2.98) very close to the median ($2.99)—indicating a balanced pricing distribution
- Rental durations cluster around 3-7 days with a mode of 6 days
- Replacement costs range from $9.99 to $29.99 with a median of $19.99

Business Implication: The consistency between average and median suggests pricing isn't skewed by outliers. The 6-day rental preference might reflect weekly planning cycles (customers rent on Friday/Saturday, return the following Thursday/Friday). This pattern could inform:

- Weekend promotion timing
- Late fee policies
- Inventory restocking schedules

Follow-up Questions I Would Ask:

- Does rental duration correlate with film length? (Do customers keep longer movies longer?)
- Are higher replacement-cost films rented more or less frequently?
- Do we lose more revenue to late returns or to unrealized rental opportunities from empty shelves?

Technical Deep Dive: Query Optimization

Performance Analysis

One interesting technical challenge involved query optimization. I compared the performance of different approaches:

Basic Query:

```
SELECT * FROM film;
```

Cost: 0.00..98.00

With ORDER BY:

```
SELECT * FROM film ORDER BY film_id;
```

Cost: 0.28..127.33

With LIMIT (Optimized):

```
SELECT * FROM film ORDER BY film_id LIMIT 10;
```

Cost: 0.28..1.55

Insight: Adding ORDER BY increased the maximum cost by ~30%, but adding LIMIT dramatically reduced it. This demonstrates understanding that optimization isn't always about removing operations—sometimes it's about limiting scope.

Business Relevance: In production environments, these differences scale. A 30% performance hit might be acceptable for a report run once daily, but unacceptable for customer-facing queries run thousands of times per minute.

Custom Sorting with CASE

A particularly interesting requirement involved ordering film ratings in a non-alphabetical sequence (G, PG, PG-13, R, NC-17):

```
SELECT
  rating,
  MIN(replacement_cost) AS min_replacement_cost,
  MAX(replacement_cost) AS max_replacement_cost
FROM film
GROUP BY rating
ORDER BY
  CASE rating
    WHEN 'G' THEN 1
    WHEN 'PG' THEN 2
    WHEN 'PG-13' THEN 3
    WHEN 'R' THEN 4
    WHEN 'NC-17' THEN 5
  END;
```

Why This Matters: Real-world data rarely sorts the way business stakeholders want to see it. Understanding how to impose meaningful order on categorical data demonstrates practical SQL skills beyond basic querying.

Data Storytelling: The Presentation

Translating Technical Findings to Business Narrative

The final presentation needed to communicate complex findings to non-technical stakeholders. Here's how I structured the story:

Act 1: The Challenge

- Set up the competitive landscape (streaming services)
- Position Rockbuster's value proposition (curation over endless scrolling, physical experience)

Act 2: What We Found

- Geographic distribution (where customers are)
- Behavioral patterns (how they rent)
- Revenue concentration (who pays the most)

Act 3: What It Means

- Strategic recommendations (where to invest)
- Operational implications (how to stock inventory)
- Future questions (what to explore next)

Key Visualizations

The presentation included:

- **Geographic heat map**: Showing customer concentration by country (top 10 highlighted)
- **Pricing distribution**: Demonstrating balanced rental rate structure
- **Rental duration patterns**: Revealing the 6-day preference

Design Principle: Every visualization needed to answer a business question, not just show data. A chart without an insight is just decoration.

Business Insights & Recommendations

Geographic Expansion Strategy

Finding: Customer base highly concentrated in 10 countries, with India, China, and US leading.

Recommendation:

1. **Defend the core**: Increase marketing spend and inventory depth in top 10 markets

2. **Explore adjacencies**: Why are certain countries underrepresented? Is it logistics, pricing, or awareness?
3. **Test and learn**: Pilot programs in emerging markets with similar demographics to top performers

Measurement: Track customer acquisition cost and lifetime value by market segment.

Customer Retention Focus

Finding: Top 5 customers averaged $143.79 in spending—significantly above typical customer.

Recommendation:

1. **Loyalty program**: Create a VIP tier with benefits (extended rental periods, new release access, exclusive genres)
2. **Personalization**: Use their rental history to recommend similar titles
3. **Win-back campaigns**: Identify at-risk high-value customers before they churn

Measurement: Calculate change in customer lifetime value and retention rate for top 20% of customers.

Inventory Optimization

Finding: Rental duration clusters around 6 days with average of 5 days.

Recommendation:

1. **Dynamic pricing**: Consider time-based pricing (longer rentals = lower per-day cost)
2. **Restocking schedules**: Align inventory returns with weekend demand peaks
3. **Genre analysis**: Deep-dive into whether certain genres have longer/shorter rental patterns

Measurement: Track inventory turnover ratio and revenue per title.

Pricing Strategy

Finding: Balanced pricing distribution ($0.99-$4.99, avg $2.98) with median close to mean.

Recommendation:

1. **Price testing**: A/B test slight increases in high-demand titles
2. **Bundle offerings**: Multi-film discounts to increase transaction size

3. **Replacement cost correlation**: Analyze if expensive-to-replace films should be priced higher

Measurement: Monitor revenue per transaction and price elasticity by segment.

Questions for Further Analysis

Good analysis doesn't just answer questions—it generates better ones. Here's what I'd explore next:

Customer Behavior Deep-Dive

- **Rental frequency**: How often do customers return? What triggers repeat rentals?
- **Genre preferences**: Are high-value customers concentrated in specific genres?
- **Seasonal patterns**: Do rentals spike during holidays, weather events, or release seasons?

Operational Efficiency

- **Inventory turnover**: Which films sit idle longest? Should they be removed?
- **Damage/loss rates**: Do certain categories or price points have higher shrinkage?
- **Staff productivity**: Are certain stores more efficient at processing rentals?

Market Opportunity

- **Competitive analysis**: Where are streaming services weakest? (Older films, international content?)
- **Demographic gaps**: Are we missing certain age groups or household types?
- **Partnership potential**: Could we bundle with complementary services (food delivery, gaming)?

Revenue Optimization

- **Late fees**: Are they a revenue source or customer irritant? What's the lifetime value impact?
- **Promotional effectiveness**: Which discounts drive incremental revenue vs. just discounting inevitable purchases?

- **Cross-sell opportunity**: Can we add merchandise, snacks, or other physical goods?

Reflection: Technical and Business Lessons

What Worked Well

Structured Approach: Starting with data quality checks before diving into analysis prevented downstream issues. In one case, I verified no NULL values in critical fields—building confidence in the findings.

Iterative Refinement: The progression from basic joins to complex CTEs mirrored how I'd approach a real project—start simple, add complexity as needed.

Business Context: Every query was driven by a question, not just "what's possible with SQL." This focus kept the analysis relevant and actionable.

Technical Challenges

Complex Joins: Connecting customer data to geography required joining through four tables. My solution: draw out the ERD relationships first, then build the query incrementally.

CTE vs. Subquery Trade-offs: I experimented with both approaches and found CTEs more readable, with identical performance (verified with EXPLAIN). The lesson: optimization isn't just about speed—code maintainability matters.

Custom Sorting: The CASE-based rating sort was a fun challenge that highlighted the gap between "data order" and "business order."

What I'd Do Differently

Earlier Hypothesis Formation: In retrospect, I should have explicitly stated hypotheses before running queries. "I hypothesize that high-value customers are concentrated in urban areas" is more rigorous than "let's see where the high-value customers are."

More Statistical Testing: Beyond descriptive statistics, I could have tested for significance (e.g., is the 6-day rental preference statistically meaningful, or could it be random?).

Cohort Analysis: Breaking customers into cohorts (by signup date, first rental type, etc.) would reveal behavioral patterns over time.

Tools and Trade-offs

SQL vs. Excel: Early in the project, I reflected that Excel was faster for simple tasks due to familiarity and UI. But SQL's power became clear with complex joins and large datasets. The right tool depends on the question.

CTEs for Clarity: I chose CTEs over subqueries for readability. In a team environment, readable code is maintained code. The performance was identical (verified with query plans), so readability won.

Formatting Standards: I experimented with different SQL formatting styles and settled on well-spaced, vertically aligned code. It takes more lines but improves comprehension—especially critical when revisiting queries months later.

Conclusion

This analysis transformed Rockbuster's customer and inventory data into actionable business insights:

1. **Geographic focus**: Concentrate resources on top 10 markets while exploring expansion opportunities
2. **Customer retention**: Build loyalty programs targeting high-value customers
3. **Inventory optimization**: Align stocking patterns with the 6-day rental cycle
4. **Pricing strategy**: Maintain current balanced approach while testing incremental adjustments

The Bigger Picture: Data analysis isn't about running queries—it's about asking better questions. Every finding opened new avenues for exploration, every recommendation demanded measurement, every conclusion required context.

Rockbuster's competitive advantage doesn't come from having data (everyone does), but from translating data into decisions. That's the analyst's real job.

Technical Appendix

Key SQL Techniques Demonstrated

- **Multi-table INNER JOINs**: Connecting fact and dimension tables
- **LEFT JOINs**: Preserving all records from primary table while adding optional matches

- **Aggregate functions**: SUM(), COUNT(), AVG(), MIN(), MAX()
- **GROUP BY with HAVING**: Filtering aggregated results
- **CTEs (Common Table Expressions)**: Creating reusable, named subqueries
- **Subqueries**: Nested queries in FROM and WHERE clauses
- **CASE expressions**: Conditional logic within queries
- **Window functions**: (implicit in later analysis stages)
- **Query optimization**: Using EXPLAIN to analyze performance

Database Structure

- **15 tables** across fact (transactional) and dimension (descriptive) layers
- **Fact tables**: payment, rental
- **Dimension tables**: customer, film, store, address, city, country, actor, category, language, inventory, staff, film_actor, film_category
- **Data types**: integer, smallint, character varying, numeric, timestamp, boolean, text, array, tsvector

Tools Used

- **PostgreSQL 15**: Database engine
- **pgAdmin 4**: Database administration and query interface
- **Excel**: Initial data exploration and presentation formatting
- **PowerPoint**: Executive presentation of findings

Project Metadata

Duration: 4 weeks
Database Size: 15 tables, ~16,000 customer records
Key Skills: SQL (joins, aggregations, CTEs), data quality assessment, business analysis, data storytelling
Business Domain: Retail, entertainment, customer analytics

Contact: Available upon request for SQL code, detailed query explanations, or extended analysis discussion.