Sports Store Analysis

Sebastian Ortuno

2025-06-13

# **Sports Store project**

### **Author: Sebastian Ortuno**

### **Tools used: RStudio.**

## **Project Overview:**

* In this project, I use SQL to analyze and clean data from a fictional sports store. The goal is to answer key business questions and extract insights on revenue, profit, customer ratings, and geographic trends.

### **Libraries:**

Before we start with the business requirements, I load the libraries needed for this project.

library(openxlsx)  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(dplyr)  
library(lubridate)  
library(ggplot2)

## **Load and view the dataset**

orders<-read.csv("C:/Users/sebas/OneDrive/Documents/NEC MASTERS/Projects Portfolio/Projects Portfolio/R Projects/Sports Analytics/Orders.csv")  
  
customer<-read.csv("C:/Users/sebas/OneDrive/Documents/NEC MASTERS/Projects Portfolio/Projects Portfolio/R Projects/Sports Analytics/Customers.csv")  
  
head(orders)

## date order\_id customer\_id sport revenue profit shipping\_cost rating  
## 1 1/1/2022 10001 102278 Baseball 183.60 97.29 0 NA  
## 2 1/1/2022 10002 102279 Basketball 185.76 103.40 0 NA  
## 3 1/1/2022 10003 102280 Basketball 128.16 66.27 0 NA  
## 4 1/1/2022 10004 102281 Hockey 45.62 15.46 7 NA  
## 5 1/1/2022 10005 102282 Football 106.30 21.75 0 NA  
## 6 1/1/2022 10006 102283 Football 58.11 12.08 0 3

head(customer)

## customer\_id full\_name email State  
## 1 102278 Alica Reary areary0@sciencedaily.com Florida  
## 2 102279 Delmor Rubin drubin1@yahoo.co.jp Indiana  
## 3 102280 Joanie Hoyt jhoyt2@bloglovin.com Pennsylvania  
## 4 102281 Madelena Boat mboat3@surveymonkey.com Nevada  
## 5 102282 Sayers Patkin spatkin4@sogou.com New York  
## 6 102283 Merwyn Stout mstout5@sfgate.com Michigan

## **– Data Cleaning and Business Requirements:**

### – 1) Convert ‘date’ column (in text format) to a proper DATE type and store in ‘Date\_New’.

### – 2) KPIs: total revenue, profit, number of orders, profit margin.

### – 3) KPIs by sport: revenue, profit, orders, profit margin.

### – 4) Customer ratings: number, the percentage of ratings the company got from all the orders, average rating.

### – 5) Ratings distribution: number of orders by rating, revenue by rating, profit by rating, and profit margin by rating.

### – 6) Revenue, profit, and profit margin by State.

### – 7) Monthly profit trends and comparison with previous month.

### – 8) Monthly profit trends and comparison with previous month.

### **1) Convert ‘date’ column (in text format) to a proper DATE type and store in ‘Date\_New’.**

orders<-orders%>%mutate(New\_date=as\_date(date, format="%m/%d/%Y"))  
orders<-orders%>%select("order\_id","customer\_id","sport","revenue","profit","shipping\_cost",  
 "rating","New\_date")

### **2) KPIs: total revenue, profit, number of orders, profit margin.**

KPI<-orders%>% # Start with the orders   
 summarize(Total\_Revenue=sum(revenue,na.rm=TRUE),#Summarize revenue   
 Total\_Profit=sum(profit), # Summarize orders  
 N\_orders=n\_distinct(order\_id),# Count orders using summarize  
 Profit\_Margin=round((Total\_Profit/Total\_Revenue)\*100,2)) #Profit Margin   
  
KPI

## Total\_Revenue Total\_Profit N\_orders Profit\_Margin  
## 1 459418.4 284821.9 2847 62

* Total Revenue: $459,418.40

The store generated nearly half a million in total sales — strong revenue.

* Total Profit: $284,821.90

Profit makes up a significant portion of revenue, indicating healthy operations.

* Number of Orders: 2,847

On average, each order generates about $161.35 in revenue (459,418.4 ÷ 2,847)

* Profit Margin: 62%

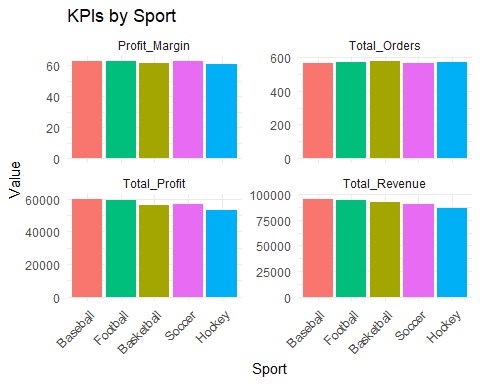
Very high margin — over half of every dollar earned is profit, which is excellent for retail.

### **3) KPIs by sport: revenue, profit, orders, profit margin.**

KPI\_sports<-orders%>%  
 group\_by(sport)%>%  
 summarise(Total\_Revenue=sum(revenue),  
 Total\_Profit=sum(profit),  
 Total\_Orders=n\_distinct(order\_id),  
 Profit\_Margin=round((Total\_Profit/Total\_Revenue)\*100,2))%>%  
 arrange(desc(Total\_Profit), desc(Total\_Orders))  
   
KPI\_sports

## # A tibble: 5 × 5  
## sport Total\_Revenue Total\_Profit Total\_Orders Profit\_Margin  
## <chr> <dbl> <dbl> <int> <dbl>  
## 1 Baseball 95364. 59699. 565 62.6  
## 2 Football 94768. 59329. 572 62.6  
## 3 Soccer 90158. 56641. 561 62.8  
## 4 Basketball 92116. 56275. 577 61.1  
## 5 Hockey 87012. 52878. 572 60.8

KPI\_sports\_long <- KPI\_sports %>%  
 pivot\_longer(cols = c(Total\_Revenue, Total\_Profit, Total\_Orders, Profit\_Margin),  
 names\_to = "KPI",  
 values\_to = "Value")  
ggplot(KPI\_sports\_long, aes(x = reorder(sport, -Value), y = Value, fill = sport)) +  
 geom\_bar(stat = "identity") +  
 facet\_wrap(~KPI, scales = "free\_y") + # Create one chart per KPI  
 theme\_minimal() +  
 theme(axis.text.x = element\_text(angle = 45, hjust = 1),  
 legend.position = "none") +  
 labs(title = "KPIs by Sport", x = "Sport", y = "Value")



* All sports have profit margins above 60%, which is a strong indicator of overall profitability.
* Soccer has the highest margin, while Basketball has the highest order volume.
* Football and Baseball show excellent balance between high revenue and strong margins.
* Hockey, although slightly behind in margin, still performs well and could improve further with cost optimization.

### **4) Customer ratings: number, percentage of ratings from all orders, average rating.**

# We need to work with NA values for rating column.  
  
Customer\_Ratings<- orders%>% summarise(Average\_Rating=round(mean(rating, na.rm=TRUE),2),   
 Total\_Rating=sum(!is.na(rating)),  
 Percentage\_rating= round((Total\_Rating/2847)\*100,2))  
  
  
  
  
Customer\_Ratings

## Average\_Rating Total\_Rating Percentage\_rating  
## 1 3.13 1193 41.9

### **5) Ratings distribution: number of orders by rating, revenue by rating, profit by rating, and profit margin by rating.**

rating\_distribution<-orders %>%  
 group\_by(rating) %>% summarise(orders\_by\_ratings=n(),  
 revenue\_by\_rating=sum(revenue),  
 profit\_by\_rating=sum(profit),  
 profit\_margin=round((profit\_by\_rating/revenue\_by\_rating)\*100,2))%>%  
 arrange(desc(rating))  
  
  
rating\_distribution

## # A tibble: 6 × 5  
## rating orders\_by\_ratings revenue\_by\_rating profit\_by\_rating profit\_margin  
## <int> <int> <dbl> <dbl> <dbl>  
## 1 5 297 40566. 23958. 59.1  
## 2 4 216 29468. 17304. 58.7  
## 3 3 240 38663. 24209. 62.6  
## 4 2 225 31839. 19251. 60.5  
## 5 1 215 28597. 16340. 57.1  
## 6 NA 1654 290285. 183761. 63.3

* Rating 3 shows the highest profit margin (62.62%) despite not having the most orders.
* Rating 5 leads in revenue and profit, but with a lower margin (59.06%).
* Lower ratings (1–2) have the lowest margins, indicating potential customer dissatisfaction.
* No direct correlation between higher rating and better profitability.

### **6) Analyze revenue, profit, and profit margin by state.**

## Best profiability efficiency (Top 3)  
  
inner\_join(orders, customer, by="customer\_id") %>%  
 group\_by(State) %>%  
 summarise(Revenue\_by\_state= sum(revenue),  
 profit\_by\_state = sum(profit),  
 profit\_margin= round((profit\_by\_state/Revenue\_by\_state)\*100,2)) %>%  
 mutate(rank\_by\_margin = as.integer(dense\_rank(desc(profit\_margin)))) %>%   
 arrange(rank\_by\_margin) %>%  
 filter(rank\_by\_margin<=3)

## # A tibble: 3 × 5  
## State Revenue\_by\_state profit\_by\_state profit\_margin rank\_by\_margin  
## <chr> <dbl> <dbl> <dbl> <int>  
## 1 Utah 5257. 3657. 69.6 1  
## 2 Massachusetts 8665. 6023. 69.5 2  
## 3 New Mexico 2997. 2044. 68.2 3

# Highest revenue and profit (top 3).  
  
inner\_join(orders, customer, by= "customer\_id") %>%   
 group\_by(State) %>%  
 summarise(profit\_state= sum(profit),  
 revenue\_state= sum(revenue),  
 profit\_margin= round( (profit\_state/revenue\_state)\*100,2))%>%  
 arrange(desc(profit\_state), desc(revenue\_state)) %>%  
 mutate(rank=row\_number())%>%  
 filter(rank<=3)

## # A tibble: 3 × 5  
## State profit\_state revenue\_state profit\_margin rank  
## <chr> <dbl> <dbl> <dbl> <int>  
## 1 California 34554. 55470. 62.3 1  
## 2 Texas 32235. 52306. 61.6 2  
## 3 Florida 22398. 36251. 61.8 3

# 3 least profitable and least revenue.  
inner\_join(orders, customer, by="customer\_id")%>%  
 group\_by(State) %>%  
 summarise(profit\_state= sum(profit),  
 revenue\_state= sum(revenue),  
 profit\_margin= round( (profit\_state/revenue\_state)\*100,2)) %>%  
 arrange(profit\_state, revenue\_state)%>%  
 mutate(rank= row\_number()) %>%  
 filter(rank<=3)

## # A tibble: 3 × 5  
## State profit\_state revenue\_state profit\_margin rank  
## <chr> <dbl> <dbl> <dbl> <int>  
## 1 Maine 16.7 91.1 18.4 1  
## 2 Rhode Island 299. 560. 53.4 2  
## 3 North Dakota 454. 706. 64.2 3

# 4 -- Smaller states like Delaware and New Hampshire show high margins despite lower total revenue.  
  
inner\_join(orders, customer, by="customer\_id")%>%  
 group\_by(State) %>%  
 summarise(profit\_state= sum(profit),  
 revenue\_state= sum(revenue),  
 profit\_margin= round( (profit\_state/revenue\_state)\*100,2)) %>%  
 arrange( desc(profit\_margin))

## # A tibble: 48 × 4  
## State profit\_state revenue\_state profit\_margin  
## <chr> <dbl> <dbl> <dbl>  
## 1 Utah 3657. 5257. 69.6  
## 2 Massachusetts 6023. 8665. 69.5  
## 3 New Mexico 2044. 2997. 68.2  
## 4 Delaware 1659. 2447. 67.8  
## 5 New Hampshire 1012. 1497. 67.6  
## 6 Iowa 3368. 5030. 67.0  
## 7 Kentucky 4598. 6973. 65.9  
## 8 Nebraska 2608. 3956. 65.9  
## 9 Illinois 5625. 8568. 65.6  
## 10 South Dakota 679. 1045. 64.9  
## # ℹ 38 more rows

* Utah, Massachusetts, and New Mexico have the best profit efficiency.
* California, Texas, and Florida rank highest in both Revenue and Profit, but not in margin.
* Maine and Rhode Island are at the bottom in all three metrics: least profitable and least revenue.
* Smaller states like Delaware and New Hampshire show high margins despite lower total revenue.

### **7) Monthly profit trends and month-over-month comparisons.**

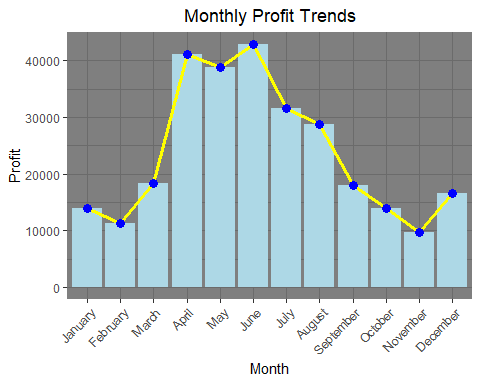
I first create a new column called Month\_trend by extracting the month number from the New\_date column. Then, I replace those numeric month values with their full month names using recode(). This way, my Month\_trend column is easier to understand because it shows names like “January” instead of just numbers.

orders<-orders %>%   
 mutate(Month\_trend= month(New\_date))  
  
  
  
Monthly\_Trend<-orders %>%  
 mutate(Month\_trend=recode(Month\_trend, "1"="January",  
 "2"="February",  
 "3"="March",  
 "4"="April",  
 "5"="May",  
 "6"="June",  
 "7"="July",  
 "8"="August",  
 "9"="September",  
 "10"="October",  
 "11"="November",  
 "12"="December"),  
 Month\_trend=factor(Month\_trend,  
 levels = c("January", "February", "March", "April", "May", "June", "July", "August", "September", "October", "November", "December"))) %>%  
 group\_by(Month\_trend) %>%   
 summarise(Monthly\_Profit= sum(profit))  
  
Monthly\_Trend

## # A tibble: 12 × 2  
## Month\_trend Monthly\_Profit  
## <fct> <dbl>  
## 1 January 14014.  
## 2 February 11244.  
## 3 March 18336.  
## 4 April 41131   
## 5 May 38847.  
## 6 June 42802.  
## 7 July 31550.  
## 8 August 28681.  
## 9 September 17992   
## 10 October 13895.  
## 11 November 9761.  
## 12 December 16568.

ggplot(Monthly\_Trend, aes(x = Month\_trend, y = Monthly\_Profit)) +  
 geom\_bar(stat = "identity", fill = "lightblue") +  
 geom\_line(aes(group = 1), color = "yellow", size = 1.2) +  
 geom\_point(color = "blue", size = 3) +  
 theme\_dark() +  
 theme(  
 axis.text.x = element\_text(angle = 45, hjust = 1),  
 plot.title = element\_text(hjust = 0.5)   
 ) +  
 labs(  
 title="Monthly Profit Trends",  
 x= "Month",  
 y="Profit"  
 )

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



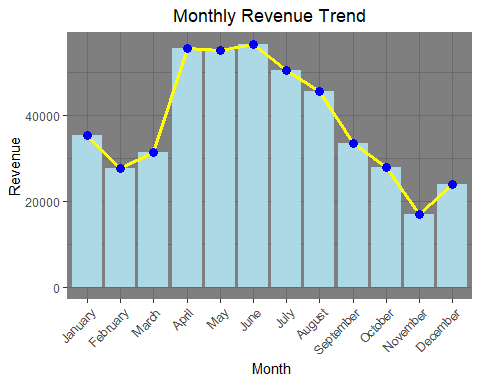
* June has the highest profit ($42,802.26), indicating a peak in sales or business activity mid-year.
* April ($41,131.00) and May ($38,847.24) also show strong profits, suggesting a strong spring season.
* January ($14,013.52) and February ($11,244.50) have relatively low profits, possibly due to post-holiday slowdowns or seasonal effects.
* November has the lowest profit ($9,760.52), which might be surprising since it’s close to the holiday season; this could suggest inventory issues, lower sales, or external factors affecting business.
* Profit fluctuates noticeably month-to-month, with some sharp increases from March to April and decreases after July.
* Summer months (June and July) show solid performance, but July’s profit ($31,550.40) dips compared to June, maybe reflecting some mid-summer slowdowns.
* October and September are on the lower side ($13,895.44 and $17,992.00 respectively), which might indicate seasonal variation or operational challenges.
* December ($16,567.86) rebounds from November’s low, possibly due to holiday shopping but doesn’t reach spring/summer highs.
* Overall, the data suggests strong seasonality, with peak sales in late spring and early summer, and dips in late fall and early year.

### **8) Monthly Revenue trends and month-over-month comparisons.**

Revenue\_Trend<-orders %>% mutate(Month\_trend =recode( Month\_trend, "1"="January",  
 "2"="February",  
 "3"="March",  
 "4"="April",  
 "5"="May",  
 "6"="June",  
 "7"="July",  
 "8"="August",  
 "9"="September",  
 "10"="October",  
 "11"="November",  
 "12"="December"),  
 Month\_trend=factor(Month\_trend, levels= c("January", "February", "March",  
 "April", "May","June", "July",  
 "August","September","October","November","December"))) %>%  
 group\_by(Month\_trend) %>%   
 summarise(Revenue= sum(revenue))  
  
Revenue\_Trend

## # A tibble: 12 × 2  
## Month\_trend Revenue  
## <fct> <dbl>  
## 1 January 35283.  
## 2 February 27718.  
## 3 March 31311.  
## 4 April 55438.  
## 5 May 55082.  
## 6 June 56407.  
## 7 July 50390.  
## 8 August 45469.  
## 9 September 33367.  
## 10 October 27995.  
## 11 November 17088.  
## 12 December 23870.

ggplot(Revenue\_Trend, aes(x=Month\_trend,y=Revenue))+  
 geom\_bar(stat = "identity", fill="lightblue") +  
 geom\_line(aes(group=1), color="yellow", size=1.2) +  
 geom\_point(color= "blue", size=3) +theme\_dark()+  
 theme( axis.text.x = element\_text(angle= 45, hjust=1),  
 plot.title = element\_text (hjust = 0.5))+  
 labs( title= "Monthly Revenue Trend",  
 x= " Month",  
 y="Revenue")

 - June had the highest revenue ($56,406.87), indicating peak business performance in early summer.

* April ($55,437.76) and May ($55,082.04) also showed strong revenue, suggesting a highly profitable spring season.
* July ($50,390.34) and August ($45,468.72) maintained solid performance, continuing the strong trend into summer.
* November recorded the lowest revenue ($17,088.32), which is unusually low considering seasonal events like Black Friday.
* December revenue ($23,869.79) improved from November but remained below the yearly average.
* January to March showed modest revenues ($27K–$35K), reflecting a slow start to the year.
* A sharp increase from March to April suggests seasonal growth or successful marketing campaigns.
* Revenue declined gradually from September ($33,366.54) to October ($27,995.24), showing post-summer slowdown.
* The data indicates a clear seasonal trend, with strong performance in late spring and early summer, followed by a steady decline into the last quarter of the year.

### **Executive Summary**

The business demonstrates strong profitability with an average profit margin of 62%, highlighting sports such as soccer and basketball for their volume and profit margins.

Customer ratings do not show a direct correlation with profitability; however, low ratings indicate potential areas for improvement.

Geographically, states like Utah and Massachusetts stand out for margin efficiency, while high-volume states such as California, Texas, and Florida dominate in total revenue and profit.

Seasonality is evident, with peaks in spring and summer (April to June) and notable declines in winter and fall, especially in November.

It is recommended to focus marketing campaigns during peak months and investigate the causes of low profitability in underperforming states and months.