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.

```r
library(openxlsx)
library(tidyverse)
```

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

```r
library(dplyr)
library(lubridate)
library(ggplot2)
```

Load and view the dataset

```r
orders<-read.csv("C:/Users/sebas/OneDrive/Documents/NEC MASTERS/Projects Portfolio/Projects Portfolio/R

customer<-read.csv("C:/Users/sebas/OneDrive/Documents/NEC MASTERS/Projects Portfolio/Projects Portfolio/

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
```

```r
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%

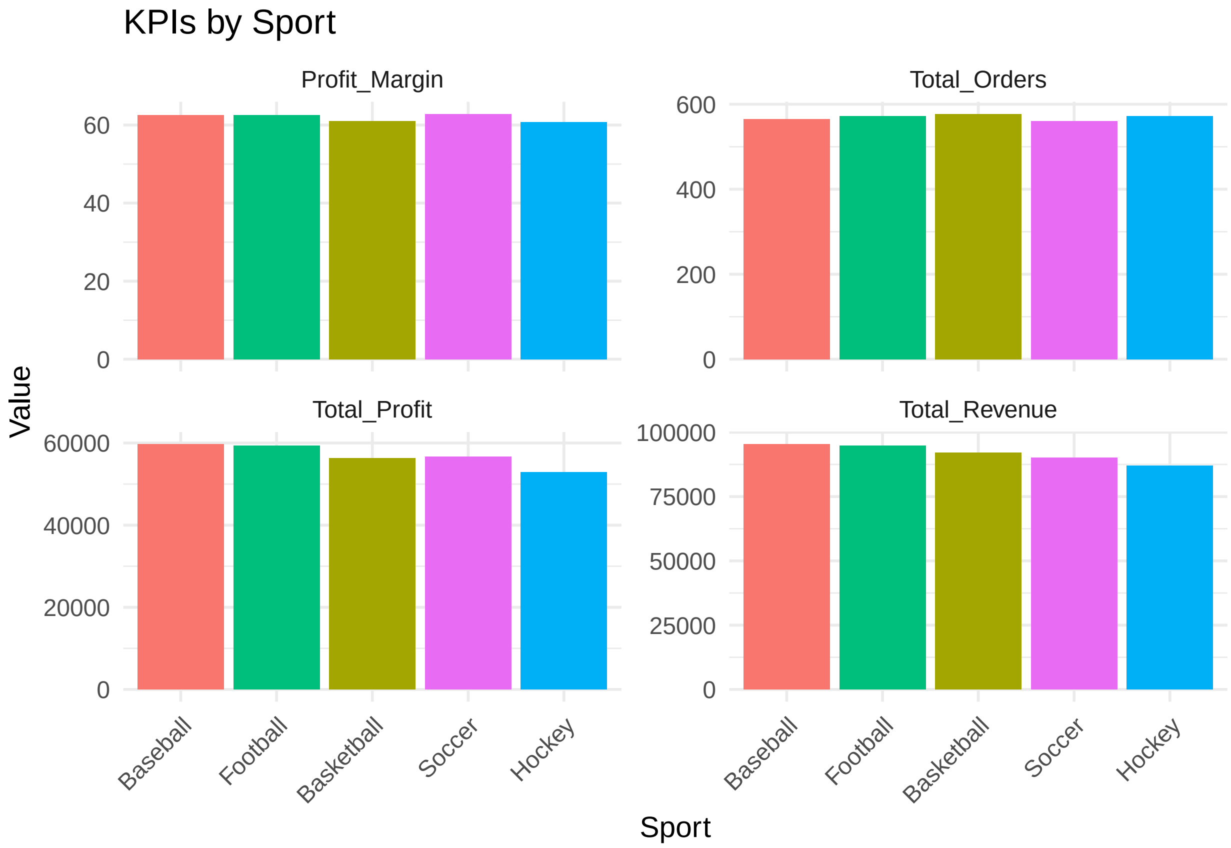
Very high margin — over half of every dollar earned is profit, which is excellent for retail.

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 x 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 x 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 x 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 x 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 x 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 x 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.

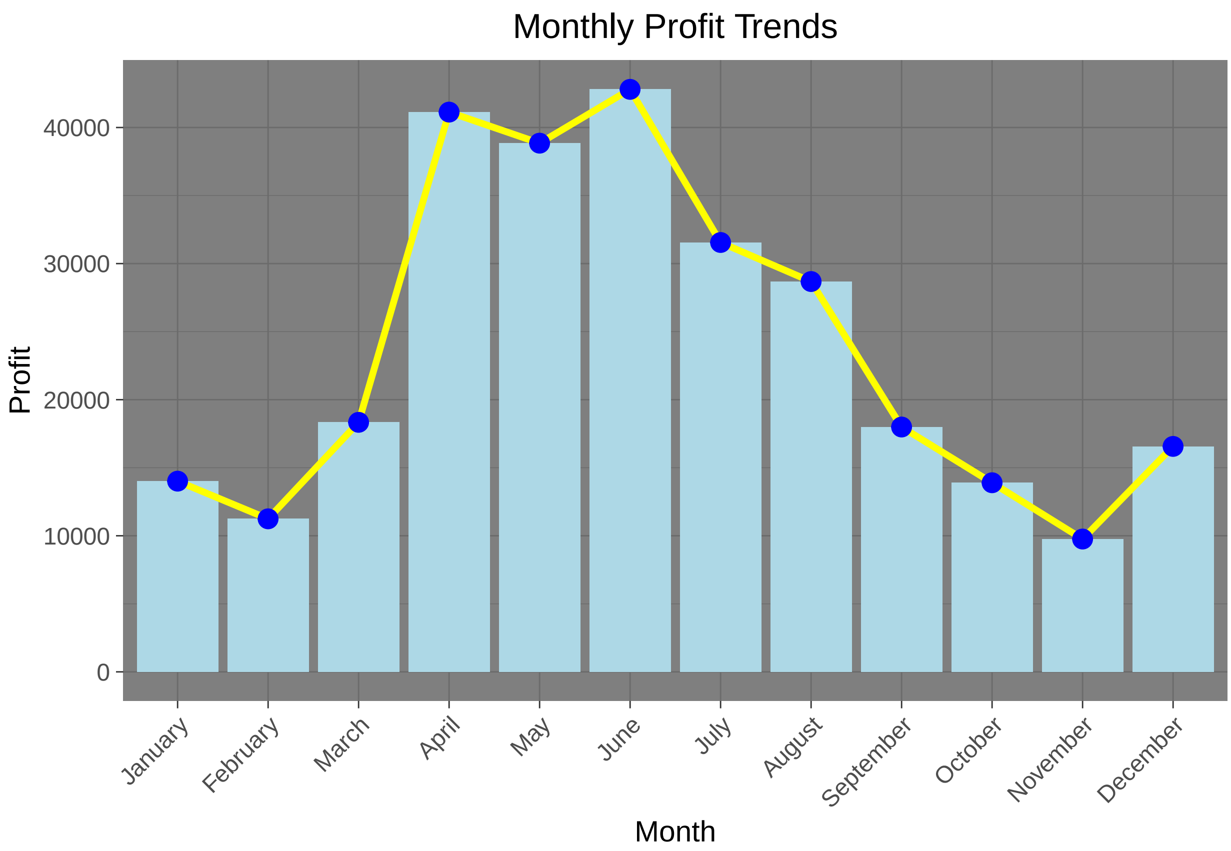
I first create a new column called Month_trend by extracting the month number from the New_date column. Then, I replace those numeric month values with their full month names using recode(). This way, my Month_trend column is easier to understand because it shows names like “January” instead of just 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",
          group_by(Month_trend) %>%
          summarise(Monthly_Profit= sum(profit))

Monthly_Trend
```

```
## # A tibble: 12 x 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.
```



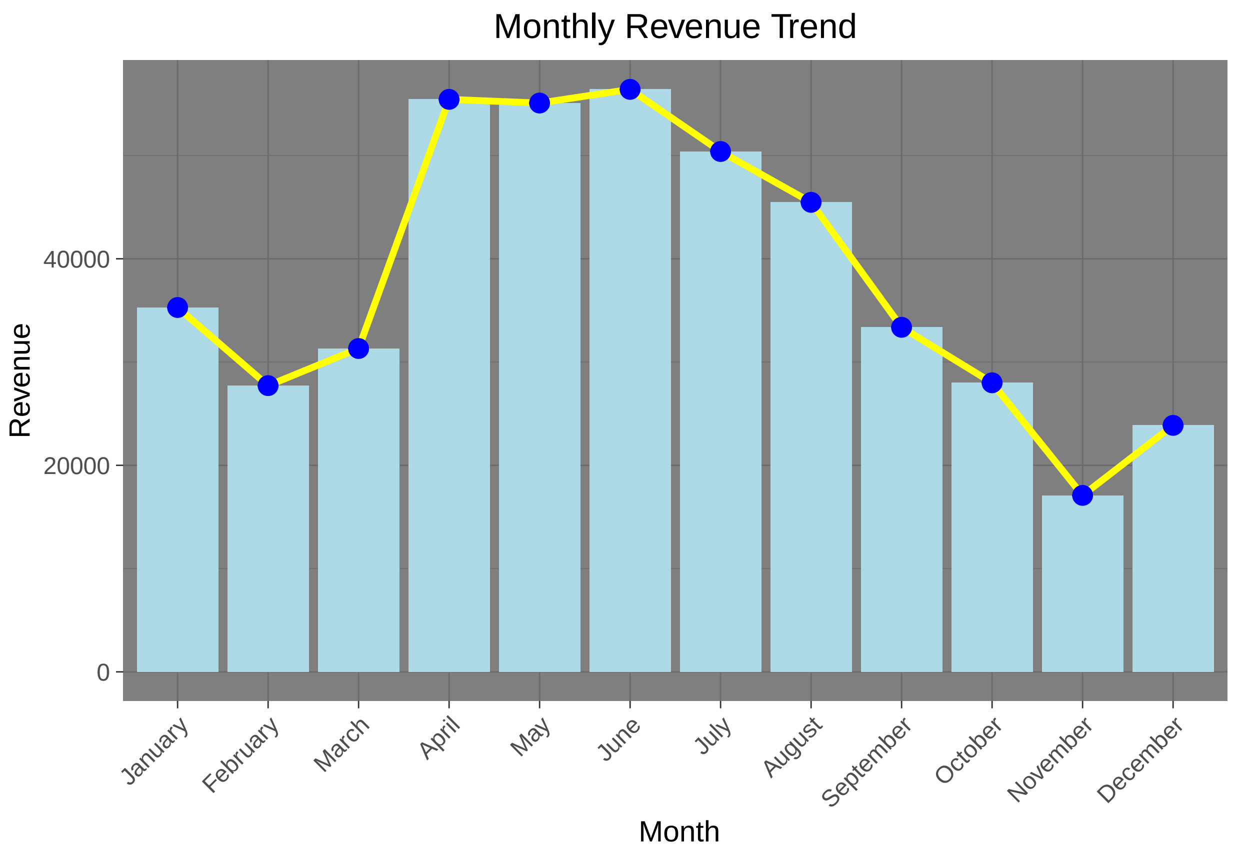
- June has the highest profit ($42,802.26), indicating a peak in sales or business activity mid-year.
- April ($41,131.00) and May ($38,847.24) also show strong profits, suggesting a strong spring season.
- January ($14,013.52) and February ($11,244.50) have relatively low profits, possibly due to post-holiday 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.

```r
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 x 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.
```

```r
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.