Task6

rm

2025-04-10

```
#install and load imp library
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(readr)
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

Load dataset

```
data <- read_csv("data[1].csv")
```

```
## Rows: 541909 Columns: 8
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (5): InvoiceNo, StockCode, Description, InvoiceDate, Country
## dbl (3): Quantity, UnitPrice, CustomerID
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Structure and info

```
glimpse(data)
```

```
## Rows: 541,909
```

```
## Columns: 8
## $ InvoiceNo   <chr> "536365", "536365", "536365", "536365", "536365", "536365"~
## $ StockCode   <chr> "85123A", "71053", "84406B", "84029G", "84029E", "22752", ~
## $ Description <chr> "WHITE HANGING HEART T-LIGHT HOLDER", "WHITE METAL LANTERN~
## $ Quantity    <dbl> 6, 6, 8, 6, 6, 2, 6, 6, 6, 32, 6, 6, 8, 6, 6, 3, 2, 3, 3, ~
## $ InvoiceDate <chr> "12/1/2010 8:26", "12/1/2010 8:26", "12/1/2010 8:26", "12/~
## $ UnitPrice   <dbl> 2.55, 3.39, 2.75, 3.39, 3.39, 7.65, 4.25, 1.85, 1.85, 1.69~
## $ CustomerID  <dbl> 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17850, 17~
## $ Country     <chr> "United Kingdom", "United Kingdom", "United Kingdom", "Uni~
```

Summary statistics

```
summary(data)
```

```
##   InvoiceNo          StockCode         Description          Quantity
##  Length:541909      Length:541909      Length:541909      Min.   :-80995.00
##  Class :character   Class :character   Class :character   1st Qu.:     1.00
##  Mode  :character   Mode  :character   Mode  :character   Median :     3.00
##                                                           Mean   :     9.55
##                                                           3rd Qu.:    10.00
##                                                           Max.   : 80995.00
##
##  InvoiceDate          UnitPrice           CustomerID       Country
##  Length:541909      Min.   :-11062.06   Min.   :12346    Length:541909
##  Class :character   1st Qu.:     1.25   1st Qu.:13953    Class :character
##  Mode  :character   Median :     2.08   Median :15152    Mode  :character
##                     Mean   :     4.61   Mean   :15288
##                     3rd Qu.:     4.13   3rd Qu.:16791
##                     Max.   : 38970.00   Max.   :18287
##                                         NA's   :135080
```

reemove missing values

```
orders_clean <- data %>%
  filter(!is.na(InvoiceNo), !is.na(CustomerID), !is.na(UnitPrice), !is.na(Quantity))
```

Convert InvoiceDate to Date type

```
orders_clean <- orders_clean %>%
  mutate(InvoiceDate = as.POSIXct(InvoiceDate, format="%m/%d/%Y %H:%M"))
```

Calculate 'amount' = UnitPrice * Quantity

```
orders_clean <- orders_clean %>%
  mutate(amount = UnitPrice * Quantity)
```

Monthly Sales Trend Analysis

```
monthly_summary <- orders_clean %>%
  mutate(
    year = year(InvoiceDate),
    month = month(InvoiceDate)
  ) %>%
  group_by(year, month) %>%
  summarise(
    total_revenue = sum(amount, na.rm = TRUE),
    order_volume = n_distinct(InvoiceNo),
    .groups = "drop"
  ) %>%
  arrange(year, month)
```

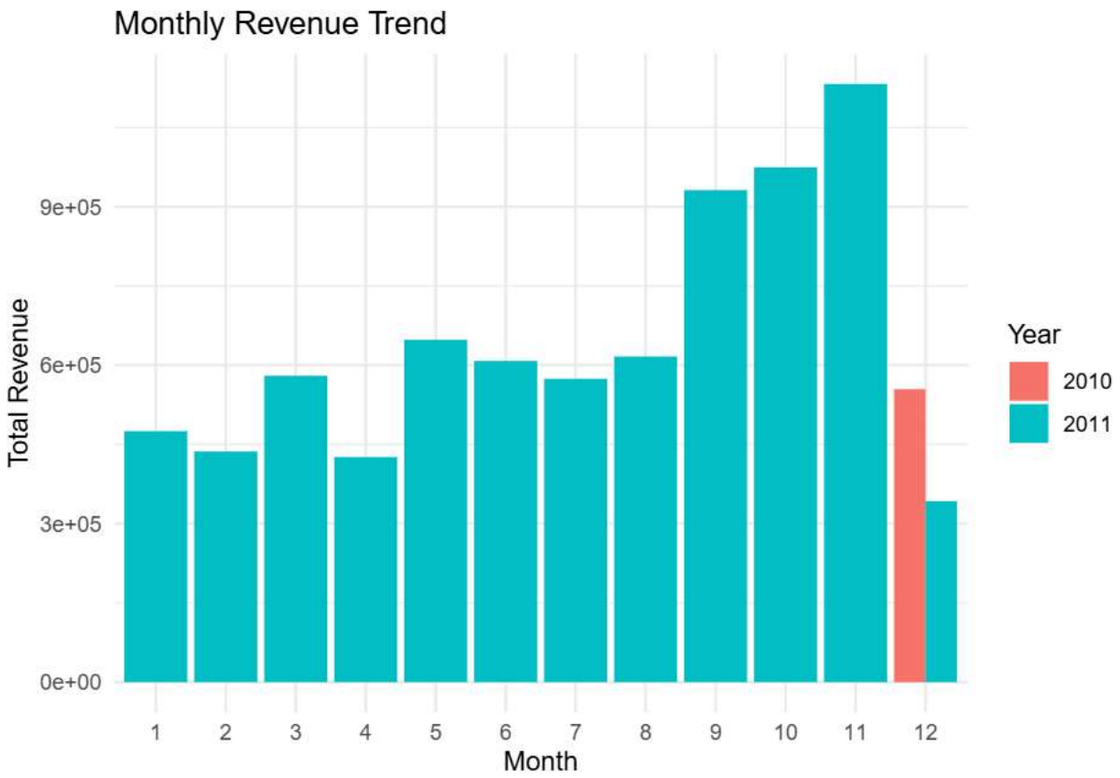
View the result

```
print(monthly_summary)
```

```
## # A tibble: 13 x 4
##     year month total_revenue order_volume
##    <dbl> <dbl>         <dbl>        <int>
##  1  2010    12       554604.         1708
##  2  2011     1       475074.         1236
##  3  2011     2       436546.         1202
##  4  2011     3       579965.         1619
##  5  2011     4       426048.         1384
##  6  2011     5       648251.         1849
##  7  2011     6       608013.         1707
##  8  2011     7       574238.         1593
##  9  2011     8       616368          1544
## 10  2011     9       931440.         2078
## 11  2011    10       974604.         2263
## 12  2011    11      1132408.         3086
## 13  2011    12       342506.          921
```

Plot the Monthly Revenue Trend

```
ggplot(monthly_summary, aes(x = factor(month), y = total_revenue, fill = factor(year))) +
  geom_col(position = "dodge") +
  labs(
    title = "Monthly Revenue Trend",
    x = "Month",
    y = "Total Revenue",
    fill = "Year"
  ) +
  theme_minimal()
```


Monthly Revenue Trend
Total Revenue
9e+05
6e+05
3e+05
0e+00
1
2
3
4
5
6
7
8
9
10
11
12
Month
Year
2010
2011