DATA 607: Week 6.3

Cindy Lin and William Forero

2025-03-05

INTRODUCTION

Week 6 project's goal is to transform three untidy data set and tidy/transform them and generate the discussed analysis.

Data Set#3 - Sales data

Loading library

```
library (tidyverse)
```

I am loading the tidyverse library because there are functions that can help tidy the loaded data.

Loading the data and removing rows

```
get_data <- read.csv("Sales.csv")
# Skip the first row to get the header

glimpse(get_data)
```

```
## Rows: 9
## Columns: 8
## $ Product.Name <chr> "Product A", "Product A", "Product A", "Product B", "Prod~
## $ Region       <chr> "North", "South", "East", "North", "South", "East", "Nort~
## $ Jan.Sales    <int> 100, 200, 300, 150, 250, 350, 50, 100, 150
## $ Feb.Sales    <int> 110, 210, 310, 160, 260, 360, 55, 105, 155
## $ Mar.Sales    <int> 120, 220, 320, 170, 270, 370, 60, 110, 160
## $ Apr.Sales    <int> 130, 230, 330, 180, 280, 380, 65, 115, 165
## $ May.Sales    <int> 140, 240, 340, 190, 290, 390, 70, 120, 170
## $ Jun.Sales    <int> 150, 250, 350, 200, 300, 400, 75, 125, 175
```

Loading the data

Long format

```
df_long <- get_data %>%
  pivot_longer(
    cols = starts_with("Jan") | starts_with("Feb") | starts_with("Mar") |
           starts_with("Apr") | starts_with("May") | starts_with("Jun"),
    names_to = "Month",
    values_to = "Sales"
  ) %>%
  mutate(Month = gsub(".Sales", "", Month))  #

df_long
```

```
## # A tibble: 54 x 4
##    Product.Name Region Month Sales
##    <chr>        <chr>  <chr> <int>
##  1 Product A    North  Jan     100
##  2 Product A    North  Feb     110
##  3 Product A    North  Mar     120
##  4 Product A    North  Apr     130
##  5 Product A    North  May     140
##  6 Product A    North  Jun     150
##  7 Product A    South  Jan     200
##  8 Product A    South  Feb     210
##  9 Product A    South  Mar     220
## 10 Product A    South  Apr     230
## # i 44 more rows
```

Shaping to long format for analysis

Year to Year Trend

```
df_long$Month <- factor(df_long$Month, levels =
                        c("Jan", "Feb", "Mar", "Apr", "May", "Jun"))

sales_trends <- df_long %>%
  group_by(Product.Name, Month) %>%
  summarise(Total_Sales = sum(Sales))
```

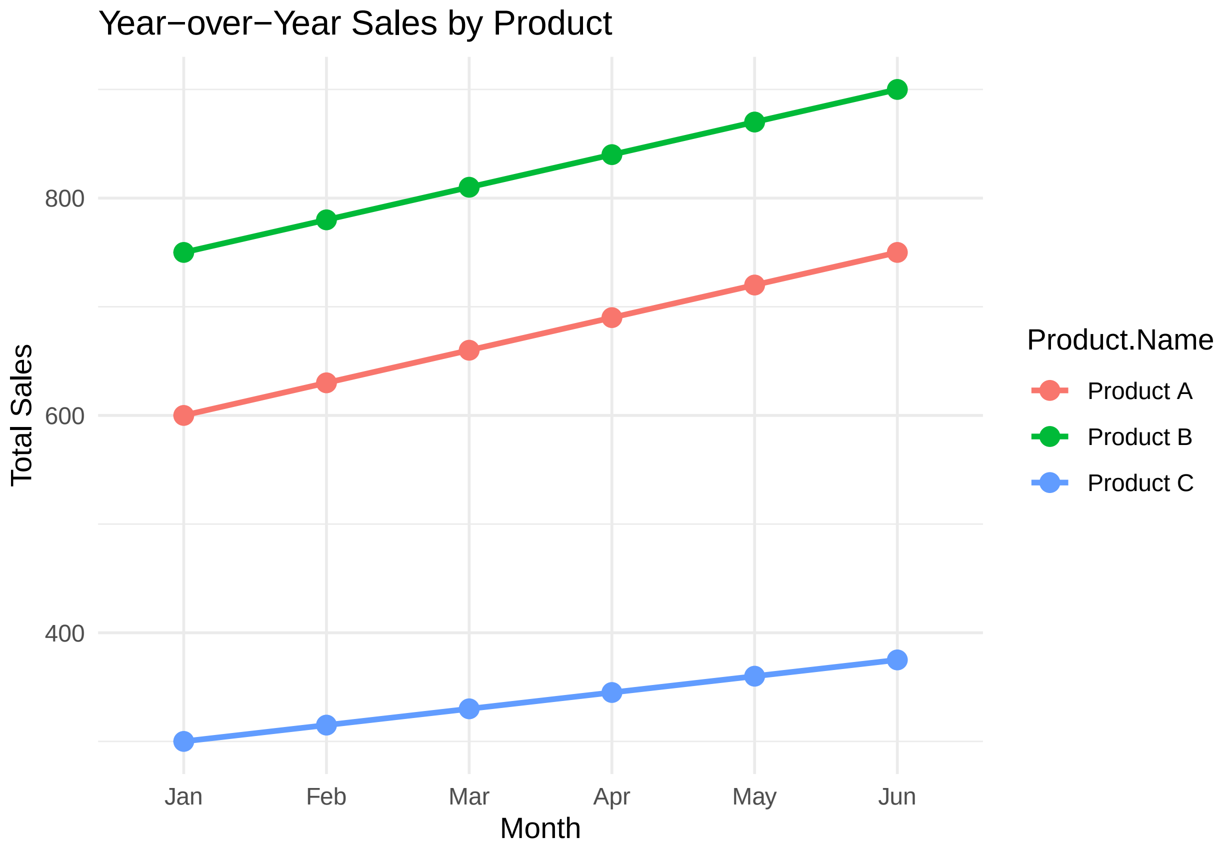
```
## ‘summarise()‘ has grouped output by ’Product.Name’. You can override using the
## ‘.groups‘ argument.
```

```
#sum of sales by product

sales_trends_region <- df_long %>%
  group_by(Region, Month) %>%
  summarise(Total_Sales = sum(Sales))
```

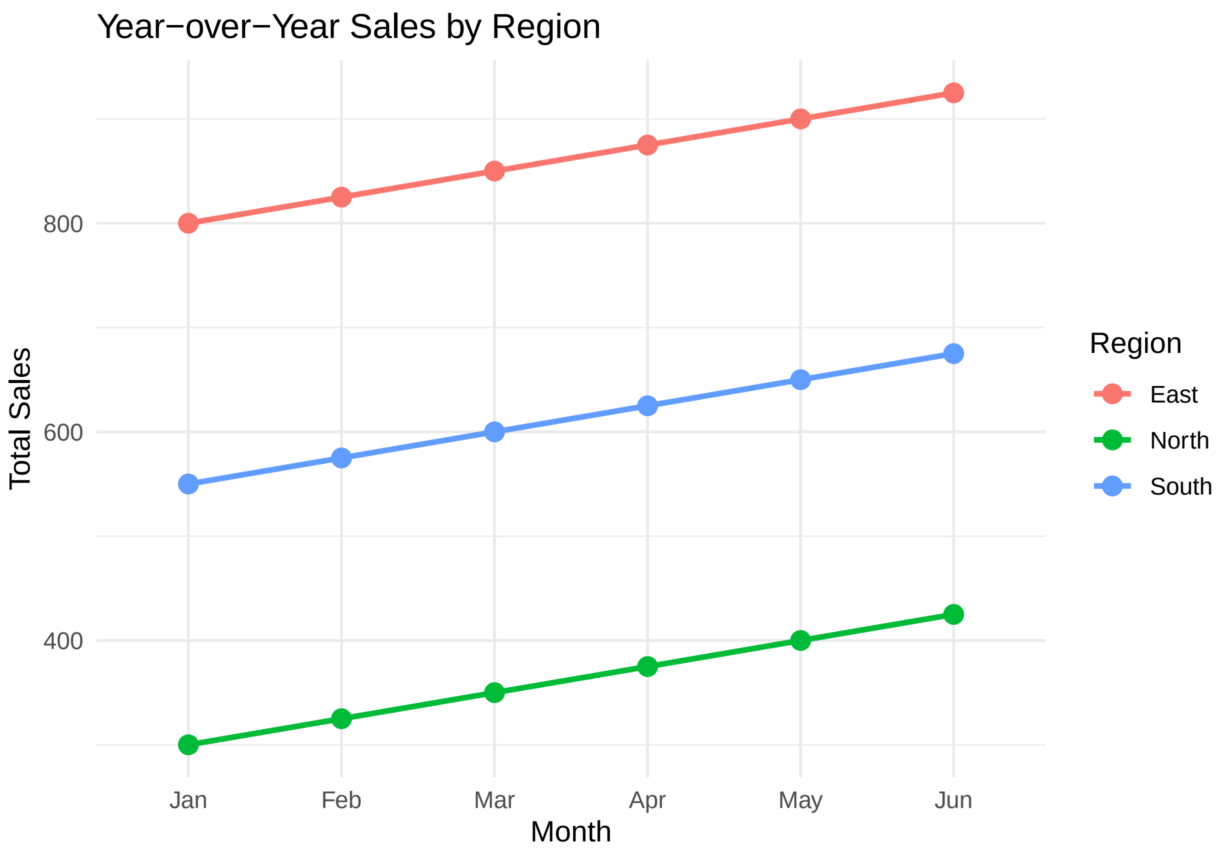
```
## ‘summarise()‘ has grouped output by ’Region’. You can override using the
## ‘.groups‘ argument.
```

```r
#sum of sales by product

ggplot(sales_trends, aes(x = Month,
                         y = Total_Sales,
                         group = Product.Name ,
                         color = Product.Name)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(title = "Year-over-Year Sales by Product", x = "Month", y = "Total Sales") +
  theme_minimal()
```

```
## 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.
```



```r
# plot trends

ggplot(sales_trends_region, aes(x = Month, y = Total_Sales, group = Region , color = Region)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  labs(title = "Year-over-Year Sales by Region", x = "Month", y = "Total Sales") +
  theme_minimal()
```



```
# plot trends
```

June sales tends to have a higher for all products. It seems that for the most part, sales increase throughout the year. With Product B having the most sales and Prodcut C with the less amount of the three.

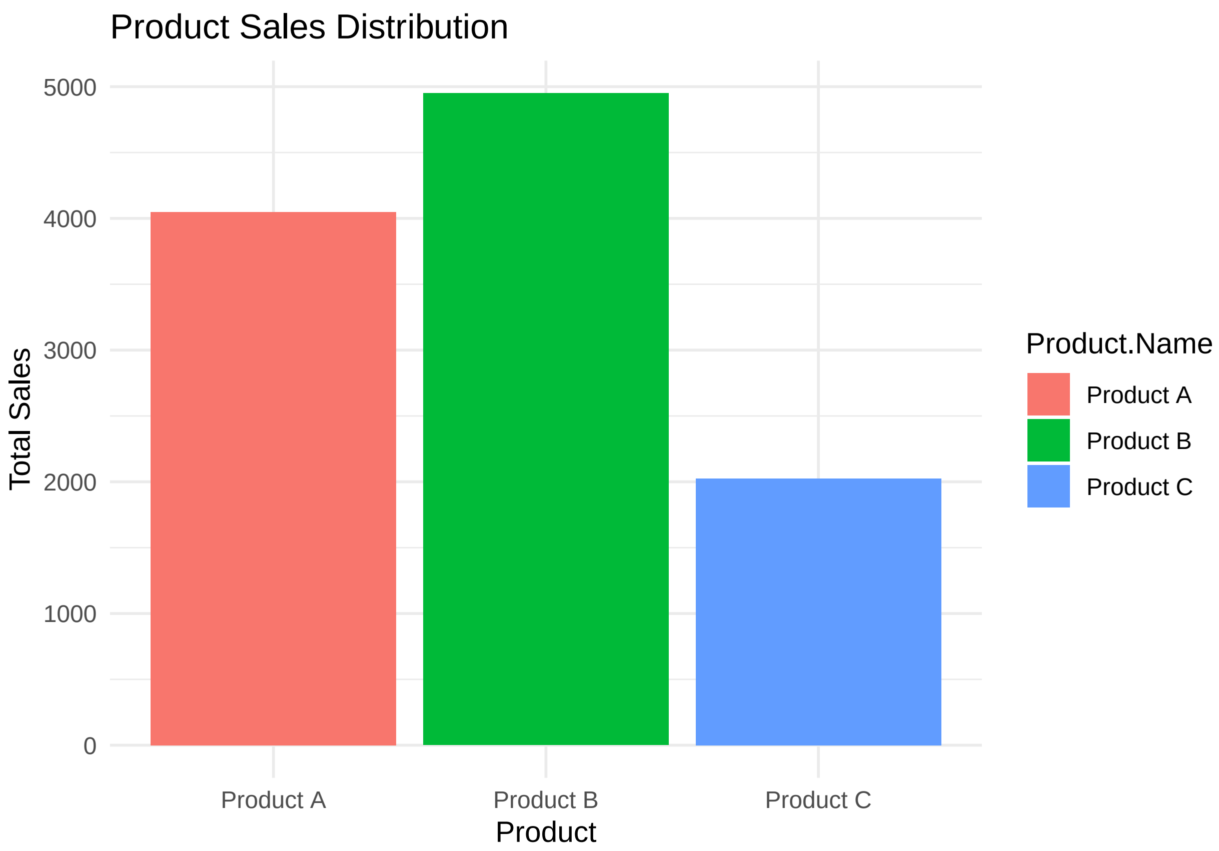
The East region also have the higher sales with north having less than the three.

Product Sales Distribution

```
sales_trends2 <- df_long %>%
  group_by(Product.Name) %>%
  summarise(Total_Sales = sum(Sales))

df_long |>
  group_by(Product.Name) |>
  summarise(Total_Sales = sum(Sales, na.rm = TRUE),
            Average_Sales = mean(Sales),
            Max_Sales = max(Sales),
            Min_Sales = min(Sales),
            .groups = "drop")
```

```
## # A tibble: 3 x 5
##   Product.Name Total_Sales Average_Sales Max_Sales Min_Sales
##   <chr>              <int>         <dbl>     <int>     <int>
## 1 Product A           4050          225       350       100
## 2 Product B           4950          275       400       150
## 3 Product C           2025          112.      175        50
```

```
ggplot(sales_trends2, aes(x = Product.Name,
                          y = Total_Sales,
                          fill = Product.Name)) +
  geom_bar(stat = "identity") +
  labs(title = "Product Sales Distribution", x = "Product", y = "Total Sales") +
  theme_minimal()
```



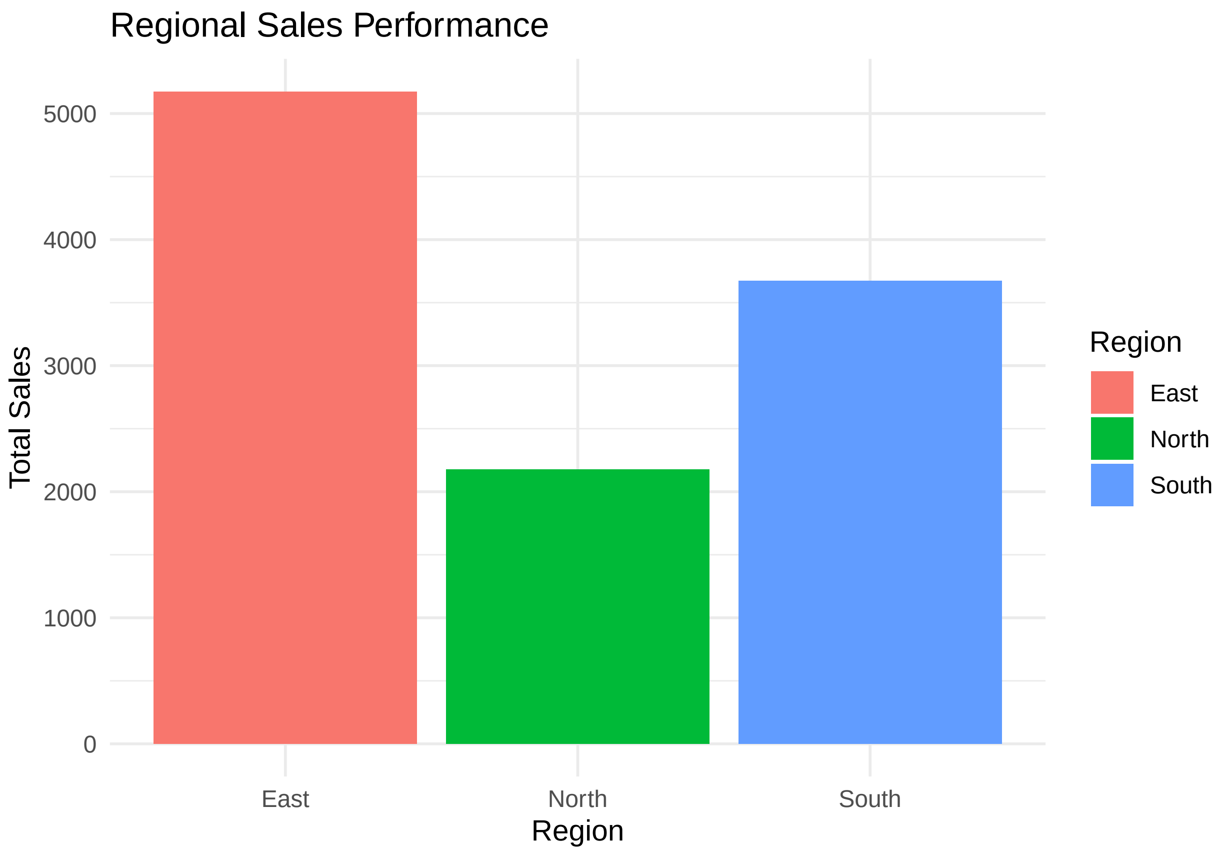
Product B had roughly 5000 total sales, Product A has a little over 4000 sales, and Product C has 2000 total sales.

Regional Performance Comparison

```
sales_trends_region2 <- df_long %>%
  group_by(Region) %>%
  summarise(Total_Sales = sum(Sales))
#sum of sales by product

df_long |>
  group_by(Region) |>
  summarise(Total_Sales = sum(Sales, na.rm = TRUE),
            Average_Sales = mean(Sales),
            Max_Sales = max(Sales),
            Min_Sales = min(Sales),
            .groups = "drop")
```

```
## # A tibble: 3 x 5
##   Region Total_Sales Average_Sales Max_Sales Min_Sales
##   <chr>        <int>         <dbl>     <int>     <int>
## 1 East          5175          288.       400       150
## 2 North         2175          121.       200        50
## 3 South         3675          204.       300       100
```

```
ggplot(sales_trends_region2, aes(x = Region,
                                 y = Total_Sales,
                                 fill = Region)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Regional Sales Performance", x = "Region", y = "Total Sales") +
  theme_minimal()
```



East region has roughly over 5000 sales total, north has a little over 2000 sales, and South has around 3500 total sales.

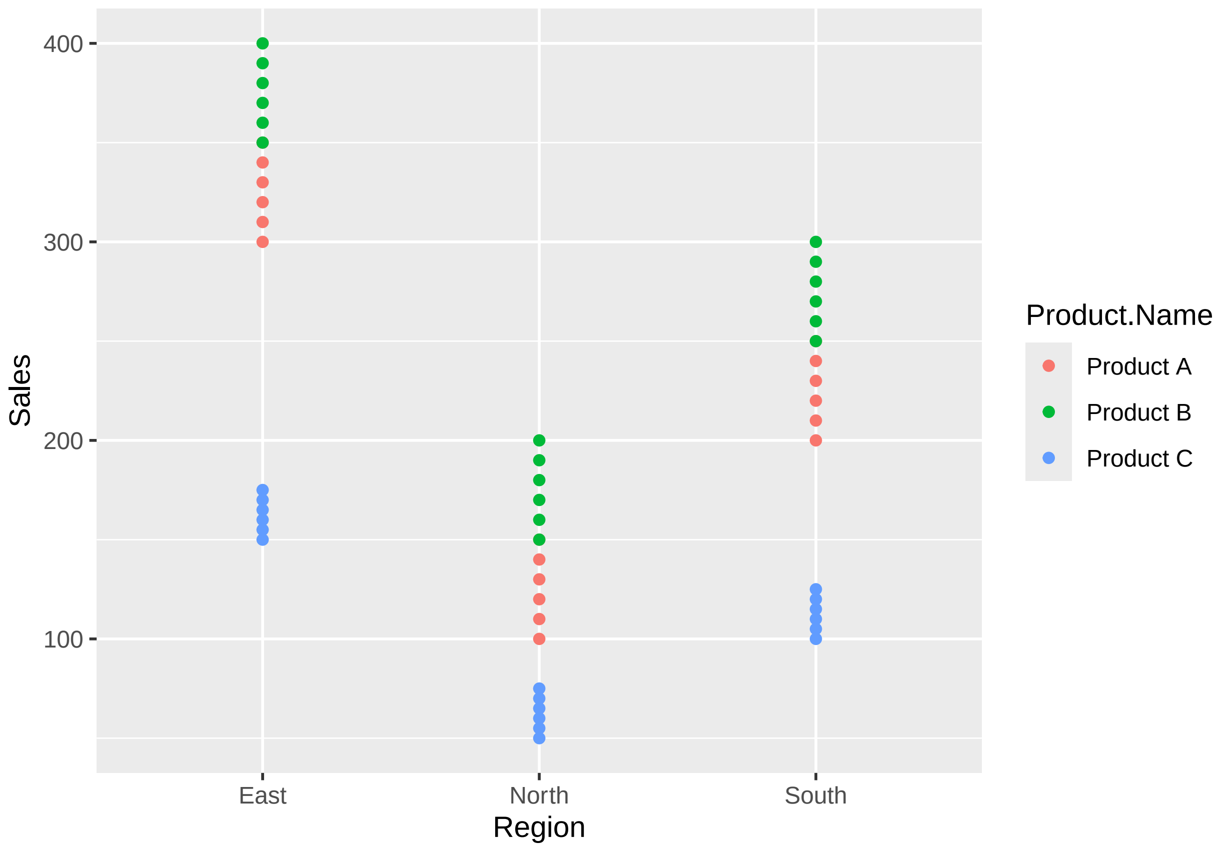
By Product and Region

```
df_long |>
  group_by(Region, Product.Name) |>
  summarise(Total_Sales = sum(Sales, na.rm = TRUE),
            Average_Sales = mean(Sales),
            Max_Sales = max(Sales),
            Min_Sales = min(Sales),
            .groups = "drop")
```

```
## # A tibble: 9 x 6
##   Region Product.Name Total_Sales Average_Sales Max_Sales Min_Sales
##   <chr>  <chr>              <int>         <dbl>     <int>     <int>
## 1 East   Product A           1950         325         350       300
## 2 East   Product B           2250         375         400       350
## 3 East   Product C            975         162.        175       150
## 4 North  Product A            750         125         150       100
## 5 North  Product B           1050         175         200       150
## 6 North  Product C            375          62.5        75        50
## 7 South  Product A           1350         225         250       200
## 8 South  Product B           1650         275         300       250
## 9 South  Product C            675         112.        125       100
```

```
ggplot(df_long, aes(x=Region, y= Sales, color = Product.Name)) + geom_point(postion = "jitter")
```

```
## Warning in geom_point(postion = "jitter"): Ignoring unknown parameters:
## ‘postion‘
```



Here we see that product A and B are relately close in sales for all 3 regions but for product C, the difference between the other two is greater. We can see the largest difference in East region.

Conclusion

Visualizations provide us quick insights on the data we are looking at, and in this scenario, it was helpfully in concluding the results of the sales and which regions are performing better than the rest. In this case, we see the the East region and product B performed the best versus the north and product C performed the worse.