Code ▾

# CS5200 Practicum II, Part III

Victoria Klimkowski and Maik Katko

Fall 2023

## Setup and Connection

### We first connect to the database and import the relevant libraries.

Hide

```
# 1. Library
library(RMySQL, quietly=T)
library(ggplot2, quietly=T)

# 2. Settings freemysqlhosting.net (max 5MB)
db_name_fh <- "sql9665320"
db_user_fh <- "sql9665320"
db_host_fh <- "sql9.freemysqlhosting.net"
db_pwd_fh <- "ITbDar1jGA"
db_port_fh <- 3306

# 3. Connect to remote server database
mydb.fh <-  dbConnect(RMySQL::MySQL(), user = db_user_fh, password = db_pwd_fh,
                      dbname = db_name_fh, host = db_host_fh, port = db_port_fh)

mydb <- mydb.fh
```

### Verify that the db has the correct tables

Hide

```
SHOW TABLES
```

| **Tables_in_sql9665320**<br><chr> |
|---|
| rep_facts |
| sales_facts |

2 rows

### Verify that the tables are populated correctly.

Hide

```
SELECT * FROM rep_facts
LIMIT 20
```

| total_sold<br><dbl> | total_qty_sold<br><int> | total_transactions<br><int> | sales_rep<br><chr> | y…<br><int> | quarter<br><int> | product<br><chr> |
|---:|---:|---:|---|---|---:|---|
| 8924 | 9700 | 9 | Helmut Schwab | 2020 | 1 | Alaraphosol |
| 50310 | 13000 | 11 | Helmut Schwab | 2020 | 1 | Bhiktarvizem |
| 14976 | 10400 | 11 | Helmut Schwab | 2020 | 1 | Clobromizen |
| 13038 | 10600 | 9 | Helmut Schwab | 2020 | 1 | Colophrazen |
| 420 | 10500 | 8 | Helmut Schwab | 2020 | 1 | Diaprogenix |
| 19822 | 10600 | 13 | Helmut Schwab | 2020 | 1 | Gerantrazeophe |
| 10200 | 10200 | 11 | Helmut Schwab | 2020 | 1 | Presterone |
| 20293 | 9100 | 8 | Helmut Schwab | 2020 | 1 | Proxinostat |
| 6204 | 2200 | 5 | Helmut Schwab | 2020 | 1 | Xinoprozen |
| 4680 | 6500 | 6 | Helmut Schwab | 2020 | 1 | Xipramin |

1-10 of 20 rows Previous **1** 2 Next

# Question 2: Analytical Queries

## Analytical Query I

### We first query for all the data on sales reps, the year, their total sold, total qty, and total transactions. We then use r to filter for the top 5 sales reps.

Hide

```
sql <- "SELECT
```

```
Warning message:
In .local(conn, statement, ...) :
  Decimal MySQL column 0 imported as numeric
```

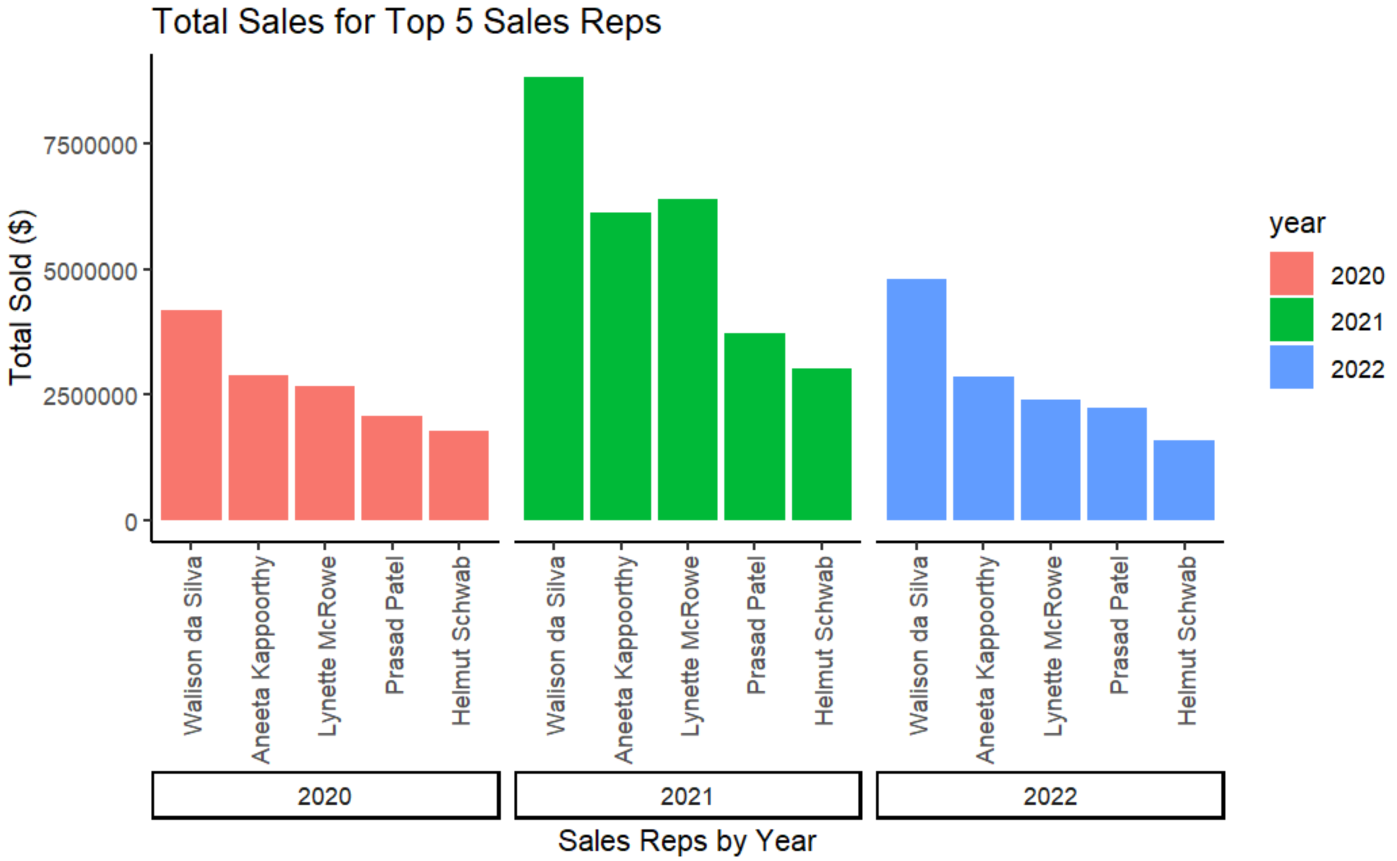
Hide

```
            sales_rep,
            year,
            SUM(total_sold) AS total_sold_per_year,
            SUM(total_qty_sold) AS total_qty_per_year,
            SUM(total_transactions) AS total_transactions_per_year
        FROM
            rep_facts
        GROUP BY
            sales_rep,
            year
        ORDER BY
            sales_rep,
            total_sold_per_year DESC;"

top_reps <- dbGetQuery(mydb, sql)

# Takes the data and splits it by year to filter for only the top 5 sales reps
filter_top_5 <- function(df) {
  df_list <- split(df, df$year)
  top_5_list <- lapply(df_list, function(x) x[order(-x$total_sold_per_year), ][1:5, ])
  do.call(rbind, top_5_list)
}

top_reps_filtered <- filter_top_5(top_reps)
```

To visualize the top 5 sales reps in a way that’s useful, we create a bar chart that organizes the sales reps by year and orders it in decreasing order of total_sold. Here we use ggplot to help with the visualization. We offer 2 versions of this data visualization.

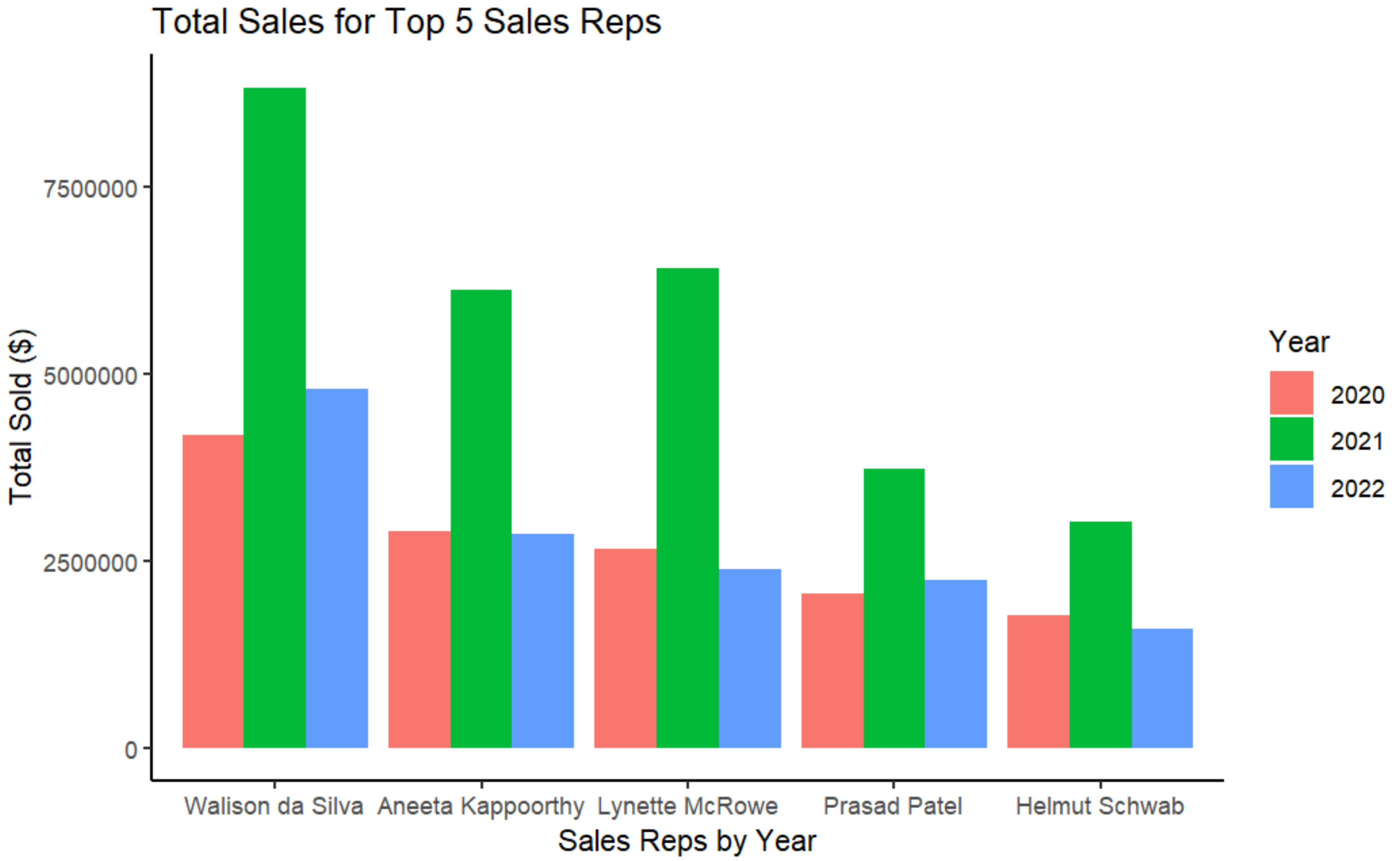
Hide

```
# Version 1: Grouping by year, with a bar for each sales rep.
# This version helps to show trends of relative performance between the reps
# .. for each year.
ggplot(top_reps_filtered, aes(x = reorder(sales_rep, -total_sold_per_year), y = total_sold_per_y
ear, group = factor(year), fill = factor(year)))+
  geom_col(position = position_dodge())+
  facet_wrap(~year, strip.position = "bottom")+
  theme_classic()+
  theme(strip.placement = "outside")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(x = "Sales Reps by Year", y = "Total Sold ($)")+
  ggtitle("Total Sales for Top 5 Sales Reps")+
  guides(fill=guide_legend(title="year"))
```



Hide

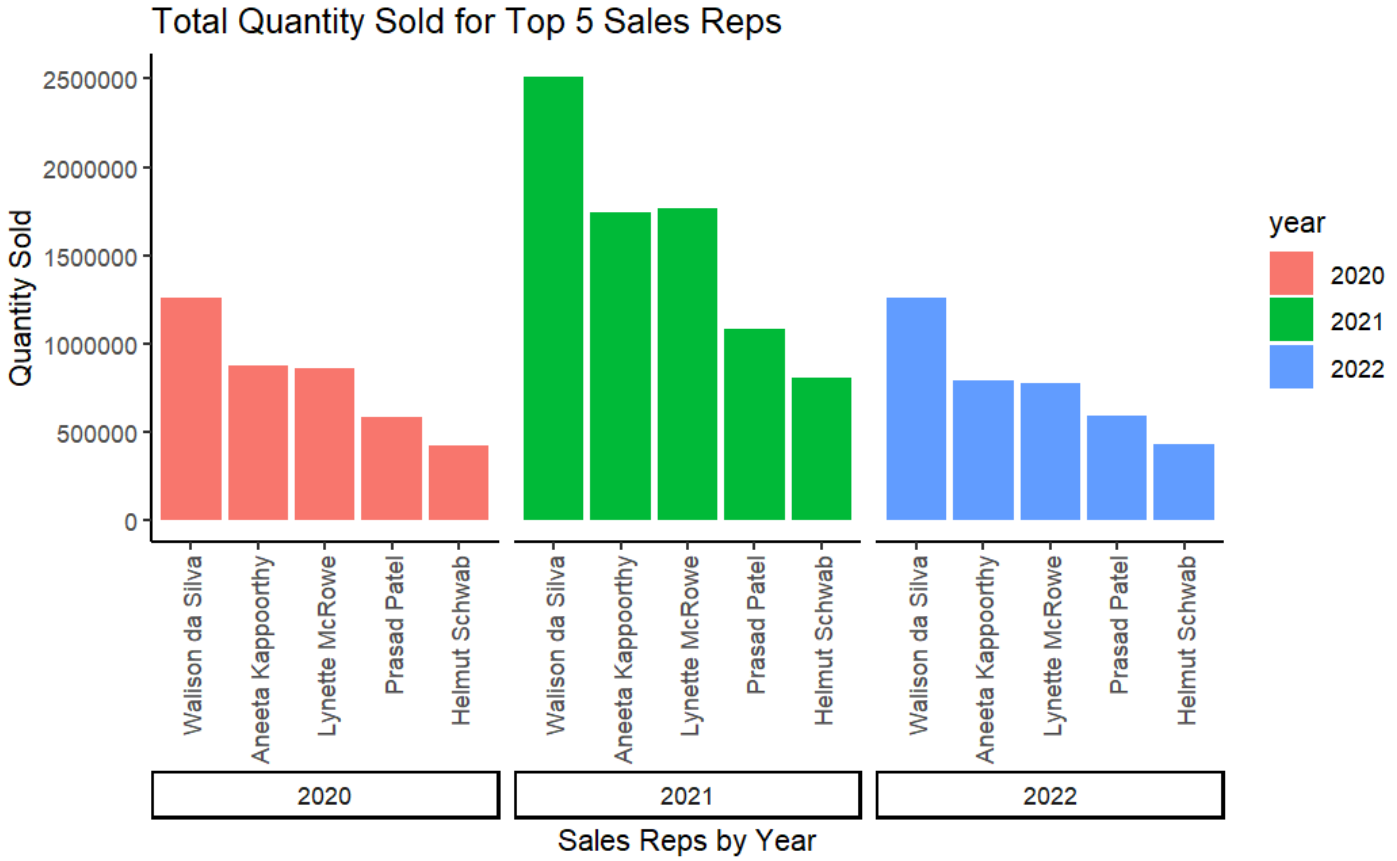
```
# Version 2: Grouping by Sales Rep, with a bar for each year.
# This version helps to show an individual rep's contribution over the time span.
ggplot(top_reps_filtered, aes(x = reorder(sales_rep, -total_sold_per_year), y = total_sold_per_year, fill = factor(year))) +
  geom_col(position = position_dodge()) +
  theme_classic() +
  labs(x = "Sales Reps by Year", y = "Total Sold ($)") +
  ggtitle("Total Sales for Top 5 Sales Reps") +
  guides(fill = guide_legend(title = "Year")) +
  theme(strip.placement = "outside")
```



Based on these graphs, It’s clear that Walison da Silva is the top sales rep by a wider margin than the other reps. Aneeta typically outperformed Lynette, but in 2021 was outpaced by her by a slight margin. Also, it looks like 2021 was an unusually good year for sales overall.

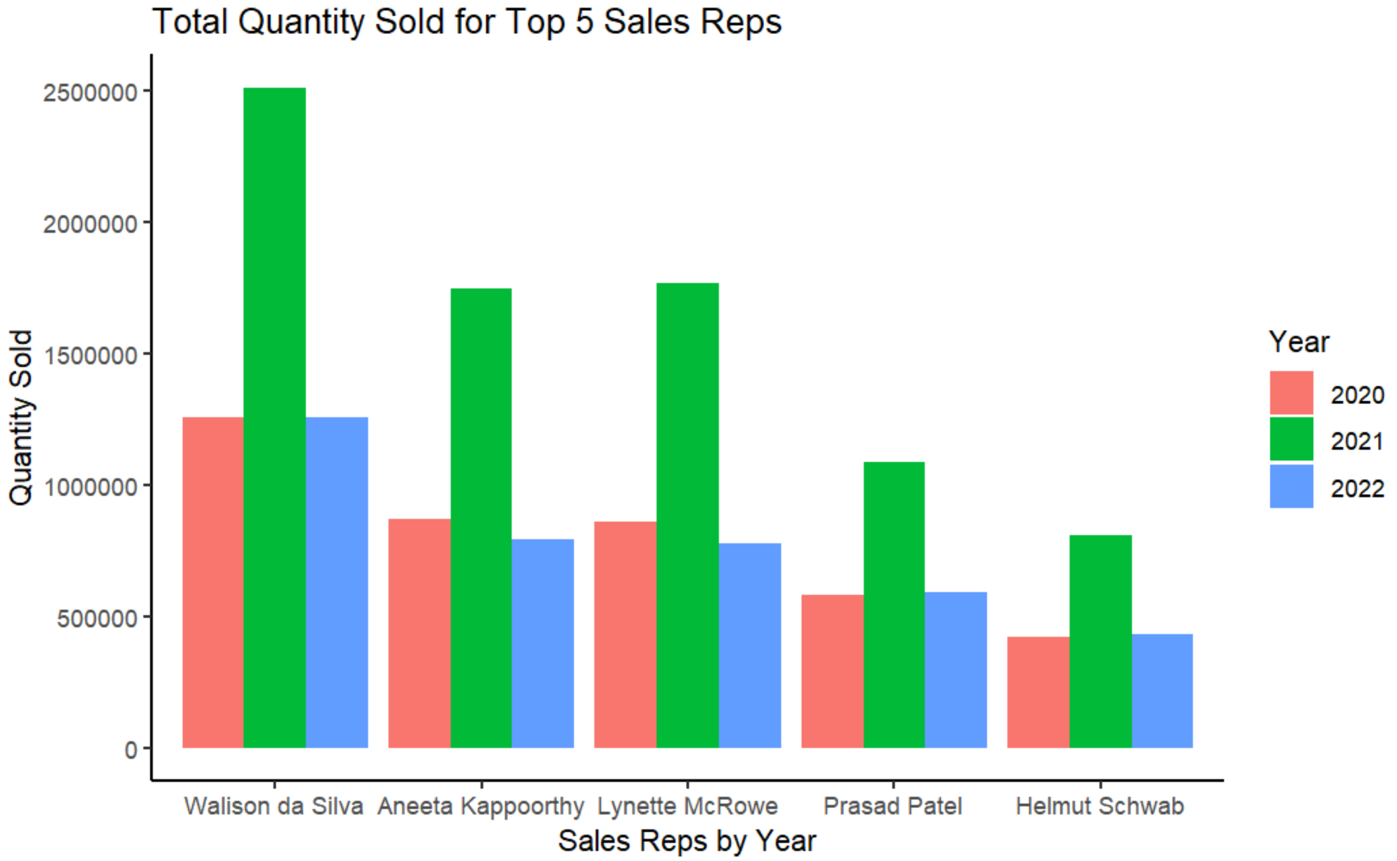
Hide

```
# Version 1: Grouping by year, with a bar for each sales rep.
# This version helps to show trends of relative performance between the reps
# .. for each year.
ggplot(top_reps_filtered, aes(x = reorder(sales_rep, -total_qty_per_year), y = total_qty_per_yea
r, group = factor(year), fill = factor(year)))+
  geom_col(position = position_dodge())+
  facet_wrap(~year, strip.position = "bottom")+
  theme_classic()+
  theme(strip.placement = "outside")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(x = "Sales Reps by Year", y = "Quantity Sold")+
  ggtitle("Total Quantity Sold for Top 5 Sales Reps")+
  guides(fill=guide_legend(title="year"))
```



Hide

```
# Version 2: Grouping by Sales Rep, with a bar for each year.
# This version helps to show an individual rep's contribution over the time span.
ggplot(top_reps_filtered, aes(x = reorder(sales_rep, -total_qty_per_year), y = total_qty_per_yea
r, fill = factor(year))) +
  geom_col(position = position_dodge()) +
  theme_classic() +
  labs(x = "Sales Reps by Year", y = "Quantity Sold") +
  ggtitle("Total Quantity Sold for Top 5 Sales Reps") +
  guides(fill = guide_legend(title = "Year")) +
  theme(strip.placement = "outside")
```



The quantity sold matches closely with the total sold, which makes sense assuming the prices for individual products aren’t changing drastically across the years.

## Analytical Query 2

We first query for the regional total by year and save it in a data frame.

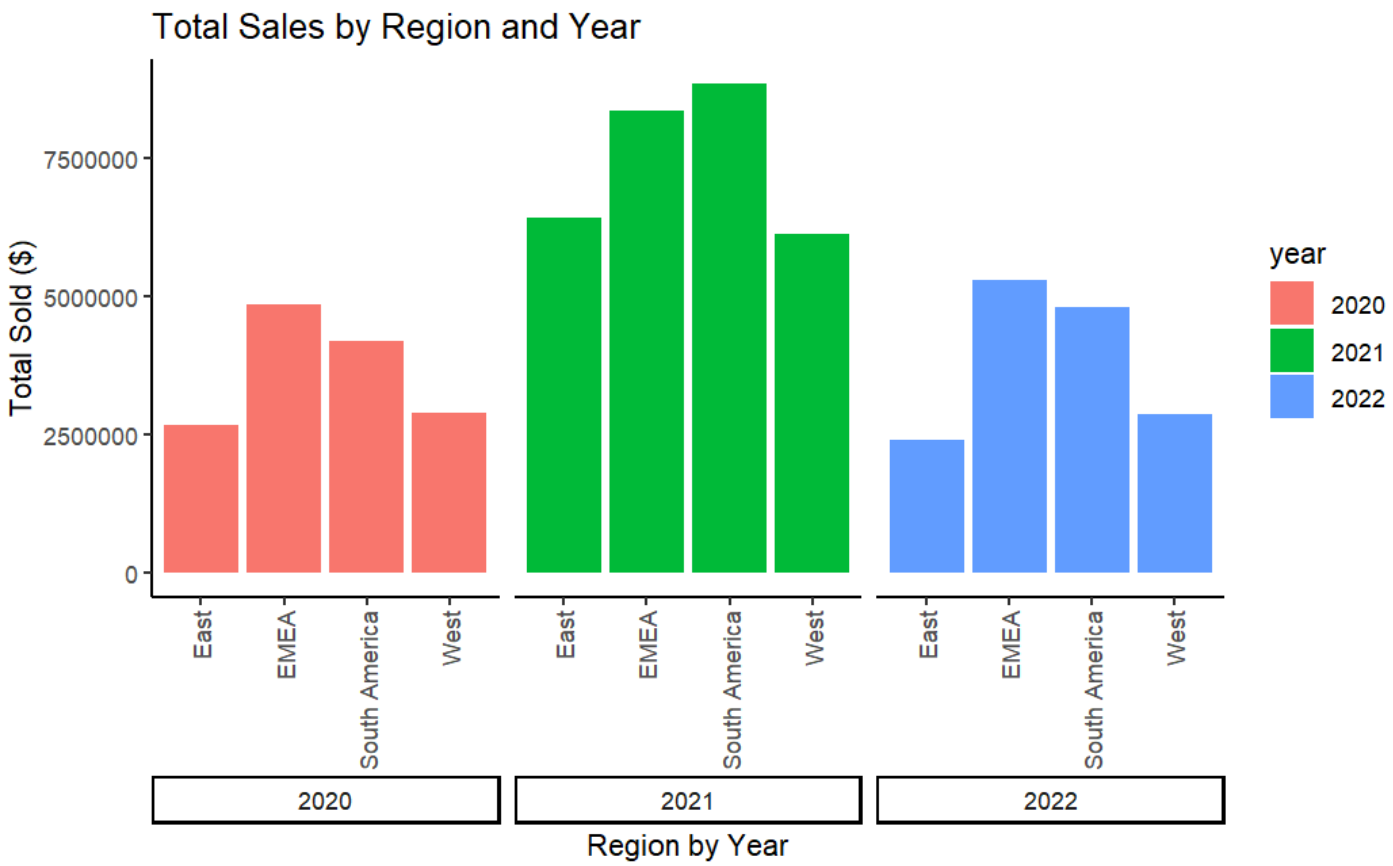
Hide

```
sql <- "SELECT
          year,
          region,
          SUM(total_sold) AS regional_total
        FROM sales_facts
        GROUP BY year, region"

regional_totals <- dbGetQuery(mydb, sql)
```

We use ggplot to visualize the data in an easy-to-read way. We offer 2 versions of this data visualization.

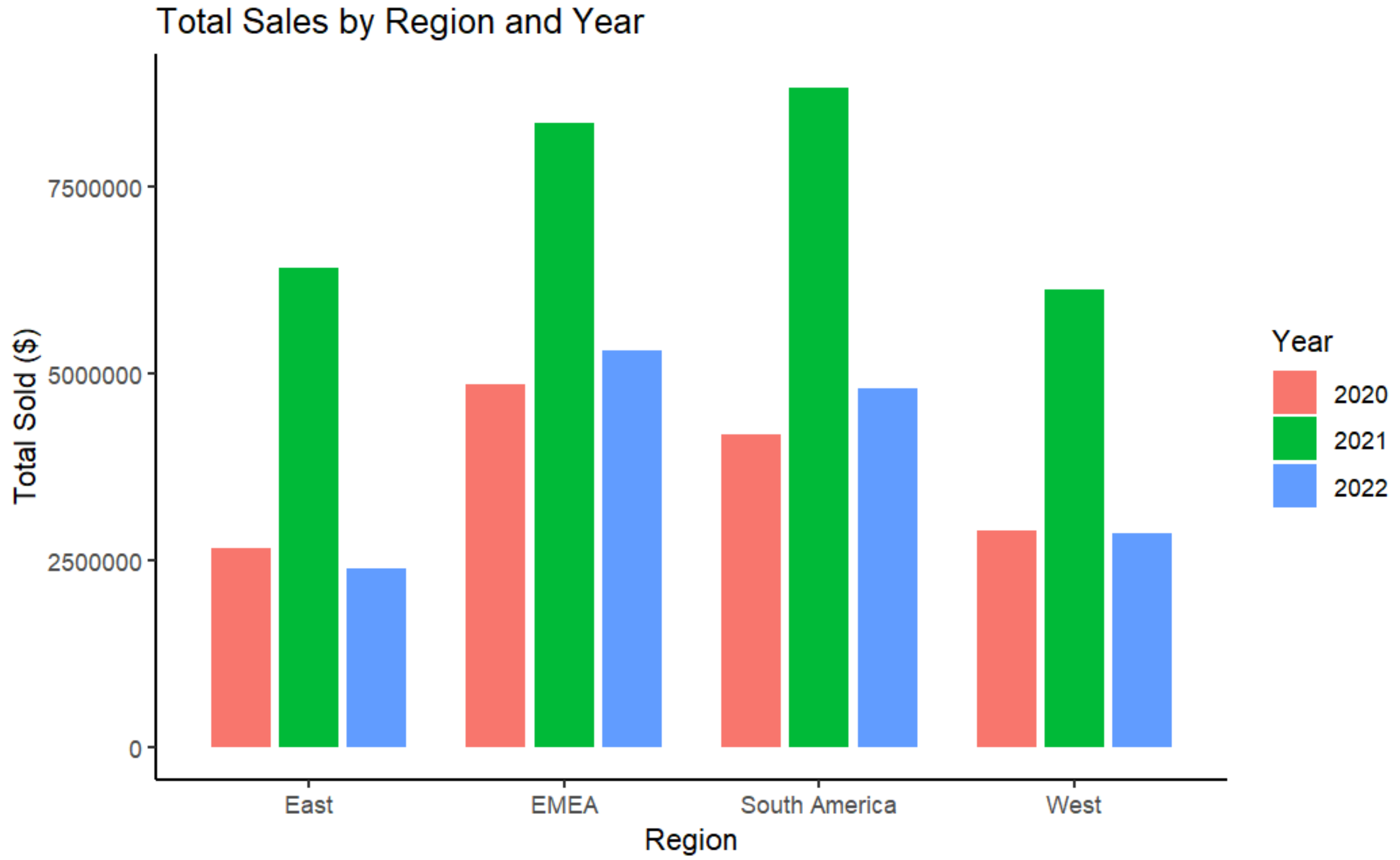
Hide

```
# Version 1: Group by year then by region.
# Helps to visualize how sales changed over the year relative to the other regions.
ggplot(regional_totals, aes(x = region, y = regional_total, group = factor(year), fill = factor
(year)))+
  geom_col(position = position_dodge())+
  facet_wrap(~year, strip.position = "bottom")+
  theme_classic()+
  theme(strip.placement = "outside")+
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  labs(x = "Region by Year", y = "Total Sold ($)")+
  ggtitle("Total Sales by Region and Year")+
  guides(fill=guide_legend(title="year"))
```



Hide

```
# Version 2: Group by region with a bar for each year.
# Helps to visualize how a region performed over the course of the time period.
ggplot(regional_totals, aes(x = region, y = regional_total, fill = factor(year))) +
  geom_col(position = position_dodge(width = 0.8), width = 0.7) +
  theme_classic() +
  labs(x = "Region", y = "Total Sold ($)") +
  ggtitle("Total Sales by Region and Year") +
  guides(fill = guide_legend(title = "Year")) +
  theme(strip.placement = "outside")
```



Based on these graphs, its again clear that 2021 was a particularly strong year for sales with each region showing a large amount of growth compared to 2020. It also seems like South America outperformed EMEA in 2021, while trailing in the other years.

# Analytical Query III

We first query for total by year and quarter and save it to a data frame.

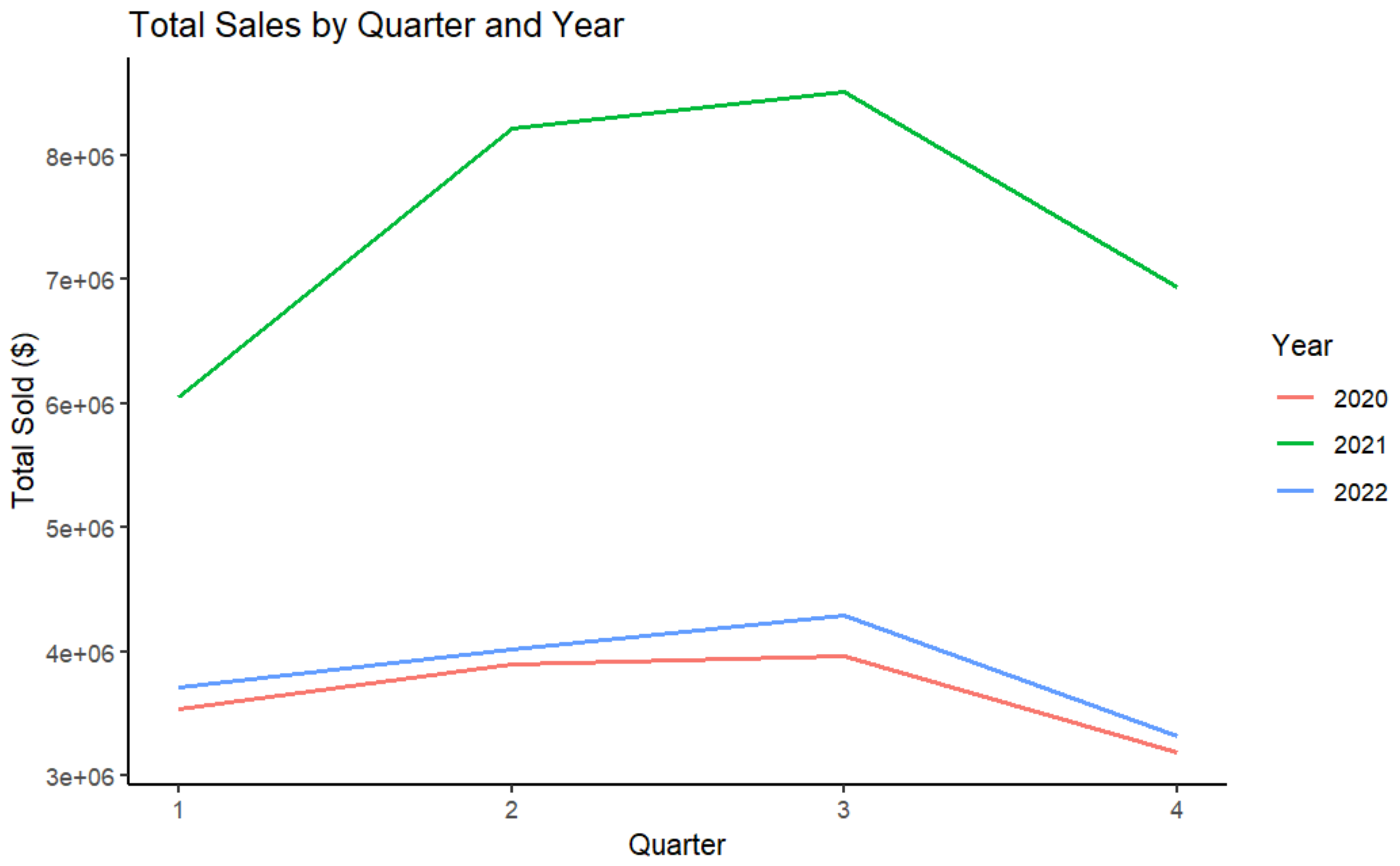
Hide

```
sql <- "SELECT
          year,
          quarter,
          SUM(total_sold) AS total
        FROM sales_facts
        GROUP BY year, quarter"

totals.df <- dbGetQuery(mydb, sql)
```

To visualize the data, we use a line plot in ggplot.

Hide

```
ggplot(totals.df, aes(x = quarter, y = total, group = factor(year), color = factor(year))) +
  geom_line(lwd = 0.75) +
  theme_classic() +
  labs(x = "Quarter", y = "Total Sold ($)") +
  ggtitle("Total Sales by Quarter and Year") +
  guides(color = guide_legend(title = "Year"))
```



This graph also confirms 2021 as a particularly strong year. It seems like the 2nd and 3rd quarters are typically better than the 1st and 4th quarters.

## Disconnect from database

Hide

```
dbDisconnect(mydb)
```

```
[1] TRUE
```