Dataquest Guided Project: Answering Business Questions using SQL

Cindy Zhang

10/7/2020

Contents

Introduction

This is my solution to Dataquest's Guided Project from the Intermediate SQL in R course, which practices writing SQL queries to extract data from `chinook`, a database of sales for a fictional music store called Chinook.

More details such as the RMD and database files can be found in the repository in GitHub.

Findings

Step 1: Creating Helper Functions

After importing the RSQLite and DBI libraries, I created a function that takes in a SQL query as an argument and returns a result in a dataframe.I then created a function that calls the query function to return a list of all tables and views in the database so I could get a glimpse of the tables within the `chinook` database.

```
run_query <- function(q) {
  conn <- dbConnect(SQLite(), 'chinook.db')
  result <- dbGetQuery(conn, q)
  dbDisconnect(conn)
  return(result)
}
show_tables <- function() {
  q = "SELECT name, type FROM sqlite_master WHERE type IN ('table', 'view')"
  return(run_query(q))
}
show_tables()
```

```
##              name  type
## 1           album table
## 2          artist table
## 3        customer table
## 4        employee table
## 5           genre table
## 6         invoice table
## 7    invoice_line table
## 8      media_type table
## 9        playlist table
## 10 playlist_track table
## 11          track table
```

Step 2: Selecting Albums to Purchase

The first query answers the question for the following fictional scenario:

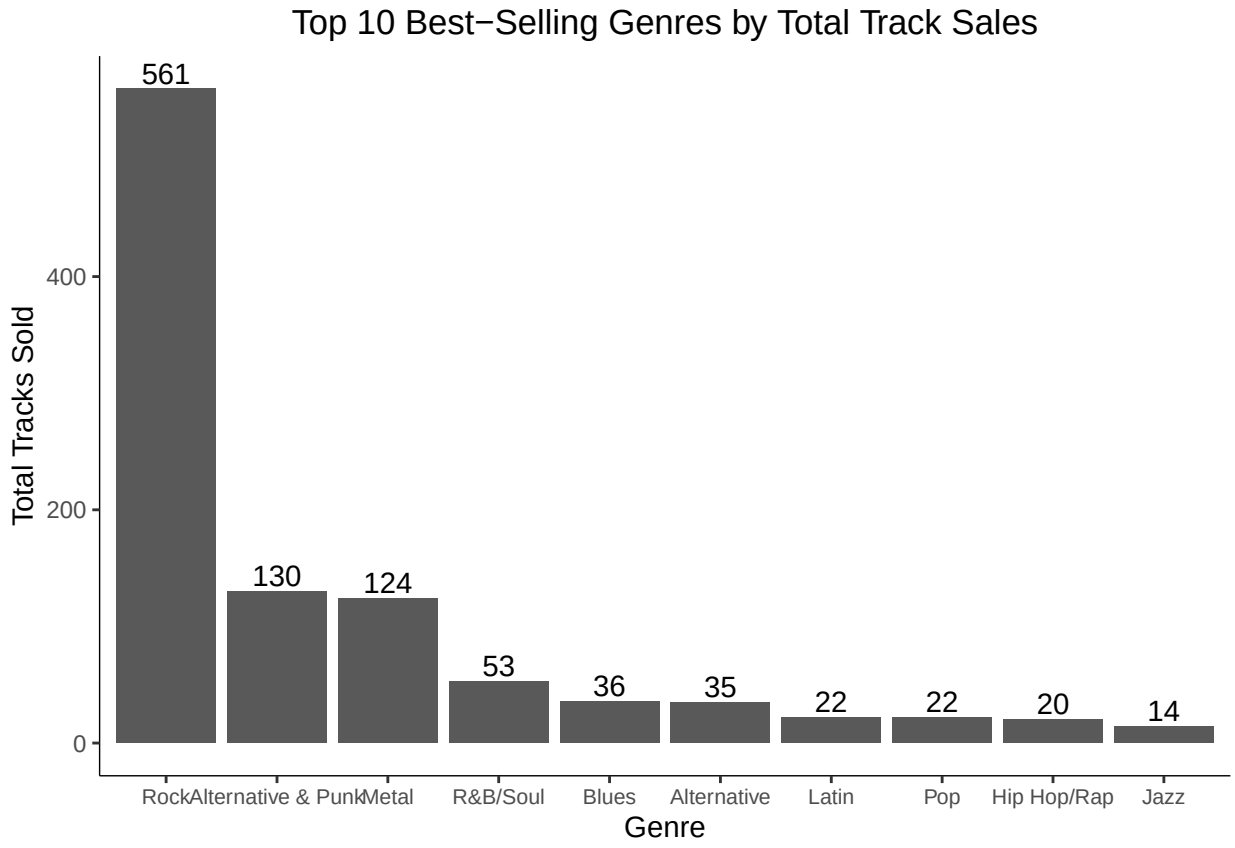
- **Fictional Scenario**: Chinook has just signed a deal with a new record label, and you're in charge of choosing the first three albums to be added to the store. There are four albums to choose from (see table below), and all four are by artists who don't have any tracks in the store right now.
- **Question 1**: Which of the three artists' albums should be added to the store?

Artist	Genre
Regal	Hip-Hop
Red Tone	Punk
Meteor and the Girls	Pop
Slim Jim Bites	Blues

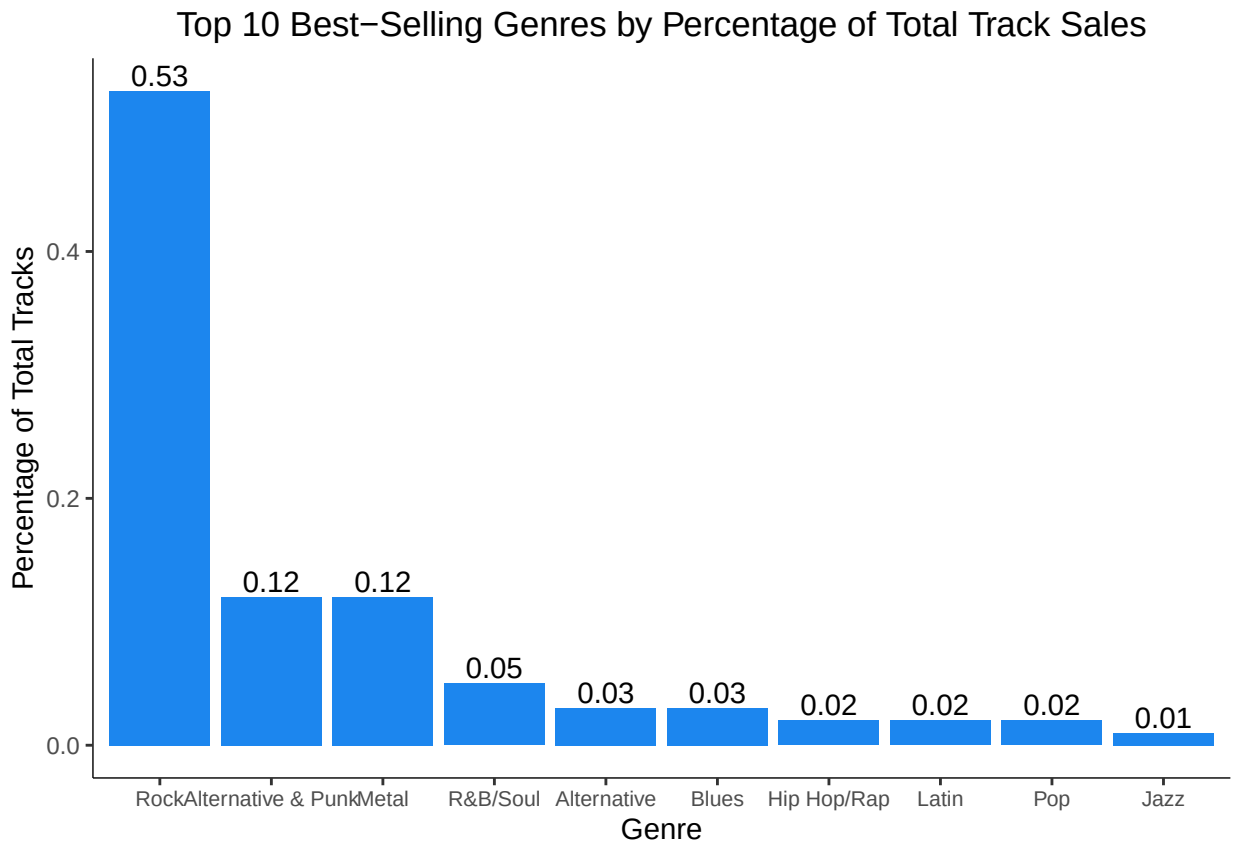
To answer this question, I calculated the number of tracks sold in the USA, in absolute numbers and percentages, by genre; logically, the genres that have sold the most tracks at Chinook would predict which of the four albums would sell well. After writing the query to obtain the data I want, I created two bar graph charts to display the total tracks sold and by percentage of total tracks sold for each genre, descending from largest to smallest.

```
album_query <- '
WITH usa_tracks_sold AS
   (
    SELECT il.* FROM invoice_line il
```

```
    INNER JOIN invoice i on il.invoice_id = i.invoice_id
    INNER JOIN customer c on i.customer_id = c.customer_id
    WHERE c.country = "USA"
   )
SELECT
    g.name genre,
    COUNT(uts.invoice_line_id) tracks_sold,
    ROUND(CAST(COUNT(uts.invoice_line_id) AS FLOAT) / (
        SELECT COUNT(*) from usa_tracks_sold
    ), 2) percentage_sold
FROM usa_tracks_sold uts
INNER JOIN track t on t.track_id = uts.track_id
INNER JOIN genre g on g.genre_id = t.genre_id
GROUP BY 1
ORDER BY 2 DESC
LIMIT 10;
'
album_df <- data.frame(run_query(album_query))

ggplot(data=album_df)+
  aes(x=reorder(genre,-tracks_sold), y=tracks_sold)+
  geom_bar(stat="identity")+
  geom_text(aes(label=tracks_sold), position=position_dodge(width=0.9), vjust=-0.25) +
  labs(y="Total Tracks Sold", x="Genre", title="Top 10 Best-Selling Genres by Total Track Sales")+
  theme(panel.background=element_rect(fill="white"), axis.line = element_line(size=0.25, colour = "black
```



```
ggplot(data=album_df)+
  aes(x=reorder(genre,-percentage_sold), y=percentage_sold)+
  geom_bar(stat="identity", fill="dodgerblue2")+
  geom_text(aes(label=percentage_sold), position=position_dodge(width=0.9), vjust=-0.25) +
  labs(y="Percentage of Total Tracks", x="Genre", title="Top 10 Best-Selling Genres by Percentage of To
  theme(panel.background=element_rect(fill="white"), axis.line = element_line(size=0.25, colour = "blac
```



Based on the query results and data visualization, Chinook should purchase the albums by Red Tone, Slim Jim Bites, and Meteor and the Girls since their respective genres sold the most tracks at Chinook, both in absolute numbers and percentages.

Step 3: Analyzing Employee Sales Performance

The second query answers the question for the following fictional scenario:

- **Fictional Scenario**: Each customer is assigned to a sales support agent within the company when they first make a purchase.
- **Question 2**: Are there sales support agents performing either better or worse than the others by total dollar amount of sales?

This question encourages experimenting with adding other variables that might be relevant to answering the question. I decided to include the following additional variables:

- name and title of the supervisor each employee reports to

- city, state, and country they work in
- birthday
- hire date
- total sales
- total invoices
- average dollar amount per sale

I then created another bar graph displaying total sales made by each sales support agent.

```
employee_query <- '
WITH employee_supervisor AS
  (
  SELECT
    e1.first_name || " " || e1.last_name employee_name,
    e1.title employee_title,
    e2.first_name || " " || e2.last_name supervisor_name,
    e2.title supervisor_title,
    e1.city,
    e1.state,
    e1.country,
    e1.birthdate,
    e1.hire_date,
    e1.employee_id
  FROM employee e1
  LEFT JOIN employee e2 ON e1.reports_to = e2.employee_id
  ORDER BY 1
  )
SELECT
  es.employee_name,
  es.employee_title,
  es.supervisor_name,
  es.supervisor_title,
  es.city,
  es.state,
  es.country,
  es.birthdate,
  es.hire_date,
  SUM(i.total) total_sales,
  COUNT(i.invoice_id) total_invoices,
  COUNT(DISTINCT(c.customer_id)) total_customers,
  AVG(i.total) avg_sales
FROM employee_supervisor es
LEFT JOIN customer c on c.support_rep_id = es.employee_id
LEFT JOIN invoice i on i.customer_id = c.customer_id
WHERE es.employee_title = "Sales Support Agent"
GROUP BY 1
ORDER BY 10 DESC;
'
employee_df <- data.frame(run_query(employee_query))
```

```
pandoc.table(employee_df, style = "rmarkdown", caption = "Sales Support Employee Information")
```

Table 2: Sales Support Employee Information (continued below)

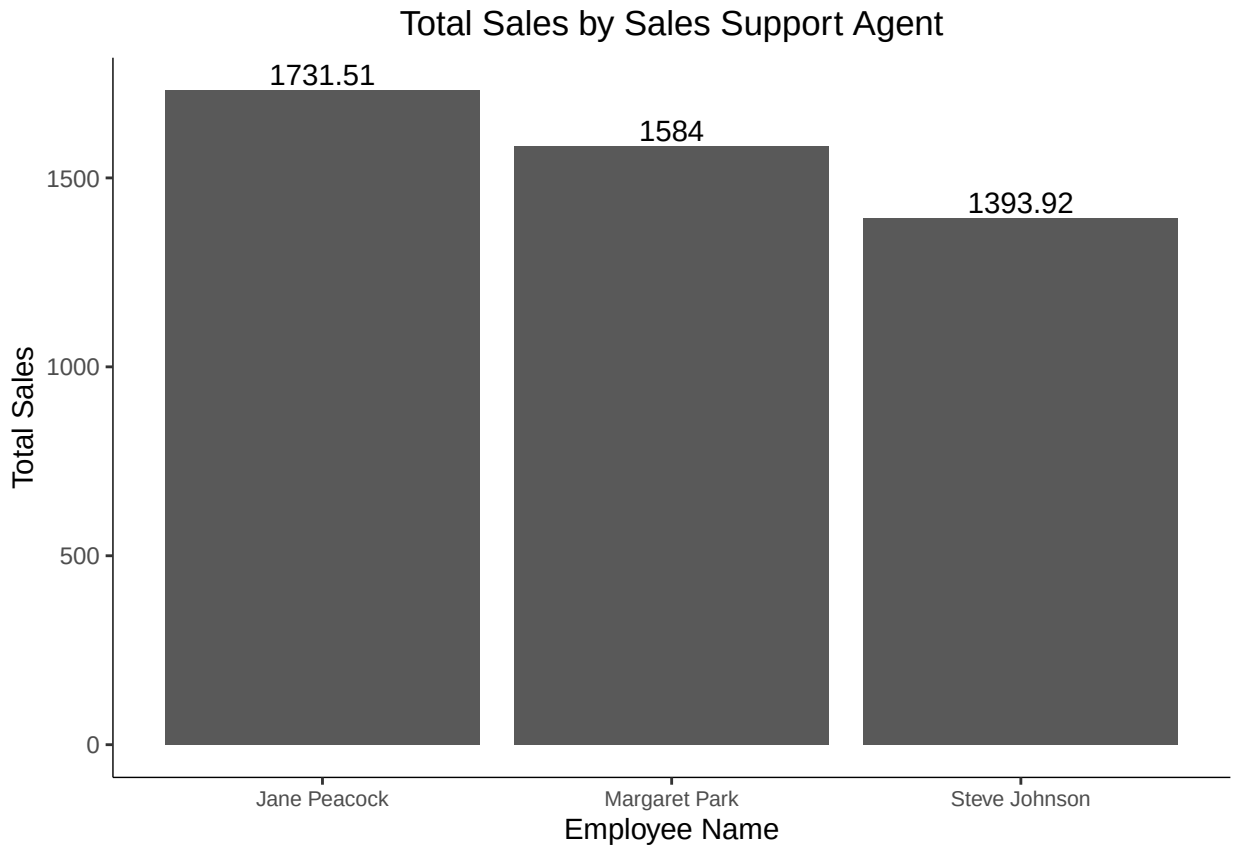
employee_name	employee_title	supervisor_name	supervisor_title
Jane Peacock	Sales Support Agent	Nancy Edwards	Sales Manager
Margaret Park	Sales Support Agent	Nancy Edwards	Sales Manager
Steve Johnson	Sales Support Agent	Nancy Edwards	Sales Manager

Table 3: Table continues below

city	state	country	birthdate	hire_date
Calgary	AB	Canada	1973-08-29 00:00:00	2017-04-01 00:00:00
Calgary	AB	Canada	1947-09-19 00:00:00	2017-05-03 00:00:00
Calgary	AB	Canada	1965-03-03 00:00:00	2017-10-17 00:00:00

total_sales	total_invoices	total_customers	avg_sales
1732	212	21	8.168
1584	214	20	7.402
1394	188	18	7.414

```
ggplot(data=employee_df)+
  aes(x=reorder(employee_name,-total_sales), y=total_sales)+
  geom_bar(stat="identity")+
  geom_text(aes(label=total_sales), position=position_dodge(width=0.9), vjust=-0.25) +
  labs(y="Total Sales", x="Employee Name", title="Total Sales by Sales Support Agent")+
  theme(panel.background=element_rect(fill="white"), axis.line = element_line(size=0.25, colour = "blac
```



Based on the query results, there are only three sales support agents who all work in the same location and report to the same supervisor, Sales Manager Nancy Edwards, making these variables controls. Jane Peacock has the highest total sales, most likely because she was hired earlier than her fellow employees, giving her more time to interact with more customers and accumulate more invoices and sales.

Step 4: Analyzing Sales by Country

The third query answers the question for the following fictional scenario:

- **Fictional Scenario**: Chinook services customers from several different countries- some with only one customer per country. Chinook management prefers that countries with only one customer be grouped together as "Other."
- **Question 3**: What is the sales data for each country and what does the data suggest in terms of which countries have potential for sales growth? Include the following variables:
  - Total number of customers
  - Total value of sales
  - Average value of sales per customer
  - Average order value

Since this was a tricky query using CASE WHEN and creating dummy variables to determine which countries should be included in the "Other" category, ("Other" = 1, "Else" = 0), writing the query is its own section.

```
country_sales_query <- '
WITH country_or_other AS
    (
```

```
     SELECT
       CASE
           WHEN (
                 SELECT count(*)
                 FROM customer
                 where country = c.country
                ) = 1 THEN "Other"
           ELSE c.country
       END AS country,
       c.customer_id,
       il.*
     FROM invoice_line il
     INNER JOIN invoice i ON i.invoice_id = il.invoice_id
     INNER JOIN customer c ON c.customer_id = i.customer_id
    )
SELECT
    country,
    customers,
    total_sales,
    average_order,
    customer_lifetime_value
FROM
    (
    SELECT
        country,
        count(distinct customer_id) customers,
        ROUND(SUM(unit_price), 2) total_sales,
        ROUND(SUM(unit_price) / count(distinct customer_id), 2) customer_lifetime_value,
        ROUND(SUM(unit_price) / count(distinct invoice_id), 2) average_order,
        CASE
            WHEN country = "Other" THEN 1
            ELSE 0
        END AS sort
    FROM country_or_other
    GROUP BY country
    ORDER BY sort ASC, total_sales DESC
    );
'
country_sales_df <- data.frame(run_query(country_sales_query))
```

Step 5: Visualizing Sales by Country

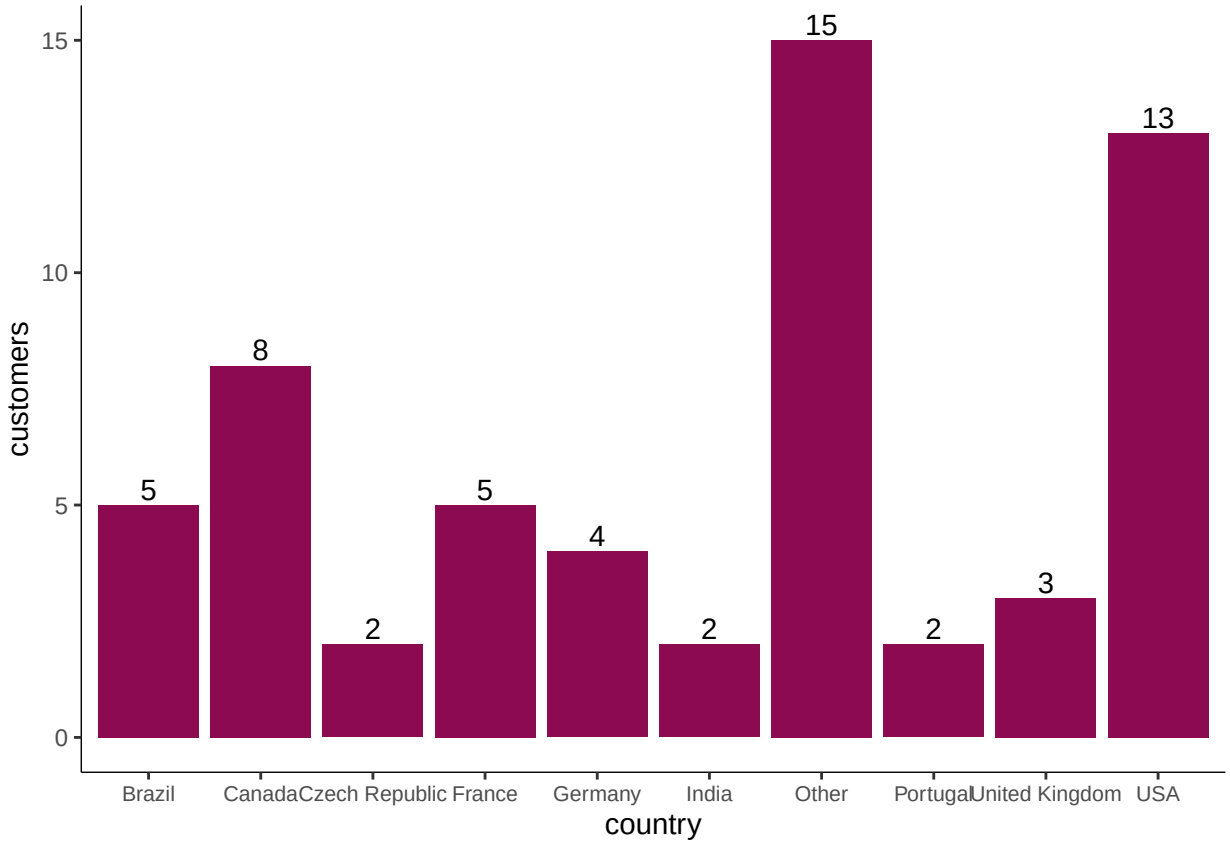
I wrote a function that generates a bar graph for any variables affiliated with the `country_sales_df` dataframe. To use the `map2` function, I created a character vector containing the country names and a character vector for the column titles of the other variables.

However, this function isn't perfect- I wasn't able to figure out how to sort the bar graphs in descending order (I am very open to any suggestions to improve this!).

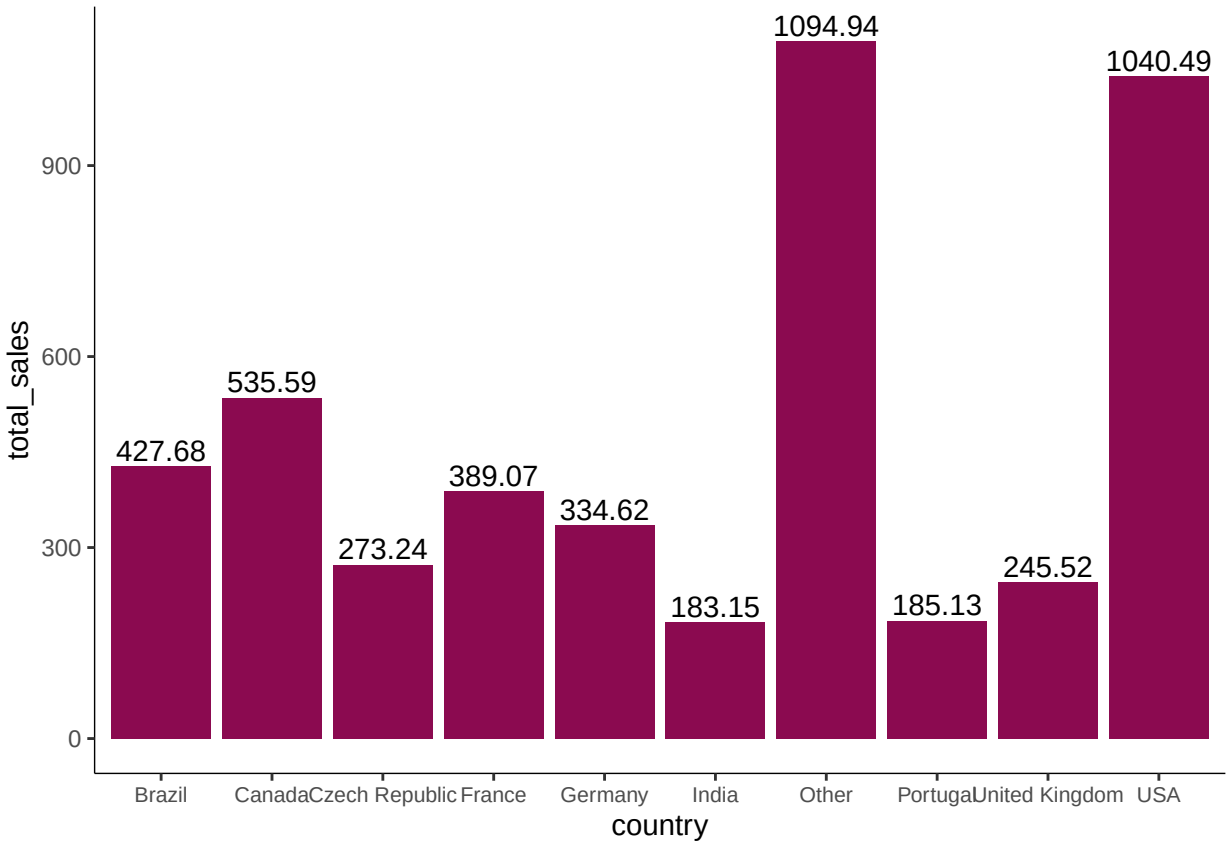
```
create_bar <- function(x,y){
    ggplot(data=country_sales_df)+
    (aes_string(x=x, y=y)) +
    geom_bar(stat="identity", fill= "deeppink4")+
```

```
    geom_text(aes(label=!!sym(y)), position=position_dodge(width=0.9), vjust=-0.25) +
    theme(panel.background=element_rect(fill="white"), axis.line = element_line(size=0.25, colour = "bla
}

x_countries <- names(country_sales_df)[1]
y_var <- names(country_sales_df)[2:5]

map2(x_countries, y_var, create_bar)
```

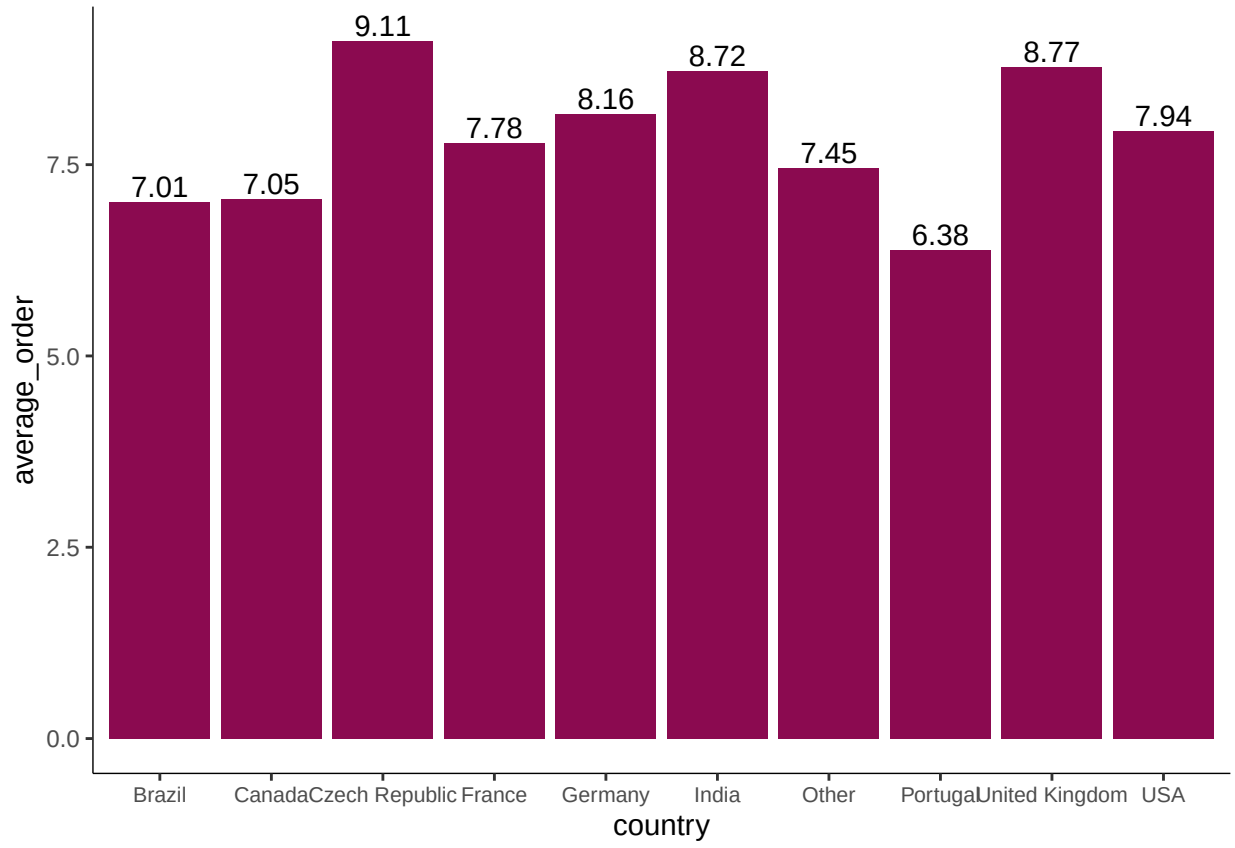
```
## [[1]]
```



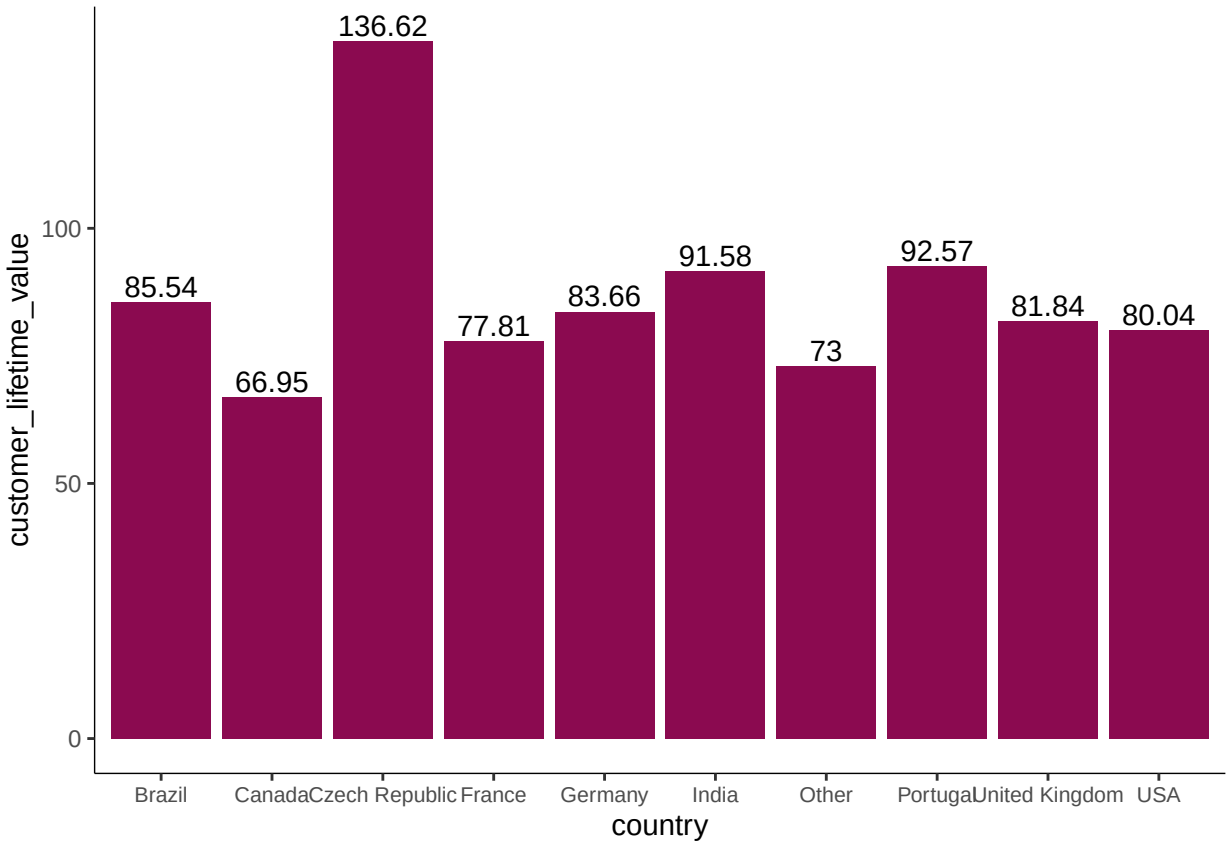

```
##
## [[2]]
```


1094.94
1040.49
900
600
total_sales
535.59
427.68
389.07
273.24
334.62
300
245.52
183.15
185.13
0
Brazil
Canada
Czech Republic
France
Germany
India
Other
Portugal
United Kingdom
USA
country

```
##
## [[3]]
```


9.11
8.72
8.77
8.16
7.94
7.78
7.45
7.01
7.05
6.38
7.5
5.0
2.5
0.0
average_order
Brazil
Canada
Czech Republic
France
Germany
India
Other
Portugal
United Kingdom
USA
country

```
##
## [[4]]
```



Despite there being more total customers in the "Other" category of countries, certain countries with multiple customers have potential for growth. While the USA accounts for the highest total sales for any one country, its average order per customer and customer lifetime value is middling compared to other countries. The Chinook team should consider the Czech Republic as a country with a lot of potential. Although there were only two customers from the Czech Republic, it had the highest average order per customer and customer lifetime value. India and Portugal also have potential for growth with high average orders and customer lifetime values.

Step 6: Album vs. Individual Tracks

The fourth and final query answers the question for the following fictional scenario:

- **Fictional Scenario**: Chinook customers can make purchases in two ways: 1) purchase a whole album or 2) purchase a collection of one or more individual tracks. Customers cannot purchase a whole album adn then add individual tracks to the same purchase, unless they do so by choosing each track manually. When customers purchase albums,, they are charged the same price as if they had purchased each of those tracks separately. Management is considering changing their purchasing strategy to save money by buying only the most popular tracks from each album from record companies, isntead of buying every track from an album.

- **Question 4**: What percentrage of purchases are individual tracks versus whole albums?

This query was the most challenging to write and involved writing a subquery that performs Boolean comparisons using the `EXCEPT` operator to determine whether a specific invoice was an album purchase or not.

```
album_track_query <- '
WITH album_boolean AS (
SELECT
  CASE
    WHEN (
      SELECT t.track_id FROM track t WHERE t.track_id = il.track_id GROUP BY t.album_id
      EXCEPT
      SELECT il.track_id FROM invoice_line il GROUP BY il.invoice_id
    ) IS NULL
    AND
        (
      SELECT il.track_id FROM invoice_line il WHERE il.track_id = t.track_id GROUP BY il.invoice_id
      EXCEPT
      SELECT t.track_id FROM track t GROUP BY t.album_id
        ) IS NULL
    THEN "Album Purchase"
    ELSE "Not Album Purchase"
  END AS album_or_not,
  il.track_id,
  il.invoice_id
  FROM track t
  INNER JOIN invoice_line il ON il.track_id = t. track_id
)
SELECT
  album_purchases,
  total_invoices,
  CAST(album_purchases AS FLOAT)/CAST(total_invoices AS FLOAT) album_purchase_percentage
FROM (
  SELECT (
    SELECT COUNT(album_or_not) FROM album_boolean WHERE album_or_not = "Album Purchase"
  ) album_purchases,
  COUNT(DISTINCT invoice_id) total_invoices
  FROM album_boolean
)
'
album_track_df <- data.frame(run_query(album_track_query))
pandoc.table(album_track_df, style = "rmarkdown", caption = "Album Purchase Breakdown")
```

Table 5: Album Purchase Breakdown

album_purchases	total_invoices	album_purchase_percentage
352	614	0.5733

Based on the query results, Chinook should continue to buy full albums as over half of all purchases are album purchases.

Conclusion

To recap the findings for each query:

1. Chinook should purchase the albums by Red Tone, Slim Jim Bites, and Meteor and the Girls since their respective genres sold the most tracks at Chinook, both in absolute numbers and percentages.
2. Jane Peacock has the highest total sales, most likely because she was hired earlier than her fellow employees, giving her more time to interact with more customers and accumulate more invoices and sales.
3. Chinook marketing should consider targeting the Czech Republic, India, and Portugal as countries with strong growth potential.
4. Chinook should continue to buy full albums as over half of all purchases are album purchases.

Thanks to everyone who's been reading so far and stay tuned for more projects!