customers

Luke Yee

8/9/2020

I will be observing customer data from a (fictional) digital media store. - There are 11 tables in the chinook sample database.

- `Employee` table stores employees data such as employee id, last name, first name, etc.
- `Customer` table stores customers data.
- `Invoice` & `InvoiceLine` tables: these two tables store invoice data.
- `Artist` table stores artists data. It is a simple table that contains only the artist id and name.
- `Albums` table stores data about a list of tracks. Each album belongs to one artist. However, one artist may have multiple albums.
- `MediaType` table stores media types such as MPEG audio and AAC audio file.
- `Genre` table stores music types such as rock, jazz, metal, etc.
- `Track` table store the data of songs. Each track belongs to one album.
- `Playlist` & `PlaylistTrack` tables: playlists table store data about playlists.

A basic query for the title of the album with an album id = 31:

```
select * from "Album" Where "AlbumId" = 31;
```

Table 1: 1 records

AlbumId	Title	ArtistId
31	Bongo Fury	23

A query that returns all albums whose artist have the word "black" in their name, using tidyverse

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ----------------------------------------------------------------------- tidyve
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
## Warning: package 'stringr' was built under R version 3.6.3

## Warning: package 'forcats' was built under R version 3.6.3

## -- Conflicts -------------------------------------------------------------------------- tidyverse_co
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Artist1 <- chinook %>% tbl("Artist") %>% collect()
Album1 <- chinook %>% tbl("Album") %>% collect()
Artist1 %>% filter(str_detect(Name,"Black")) %>% inner_join(Album1, by = "ArtistId") %>% collect()
```

```
## # A tibble: 6 x 4
##   ArtistId Name                AlbumId Title
##      <int> <chr>                 <int> <chr>
## 1       11 Black Label Society      14 Alcohol Fueled Brewtality Live! [Disc 1]
## 2       11 Black Label Society      15 Alcohol Fueled Brewtality Live! [Disc 2]
## 3       12 Black Sabbath            16 Black Sabbath
## 4       12 Black Sabbath            17 Black Sabbath Vol. 4 (Remaster)
## 5      137 The Black Crowes        209 Live [Disc 1]
## 6      137 The Black Crowes        210 Live [Disc 2]
```

And the same query in SQL:

```
select a."Title", b."Name"
from "Album" a join (
select * from "Artist" where "Name" like '%Black%') b
on b."ArtistId" = a."ArtistId"
```

Table 2: 6 records

Title	Name
Alcohol Fueled Brewtality Live! [Disc 1]	Black Label Society
Alcohol Fueled Brewtality Live! [Disc 2]	Black Label Society
Black Sabbath	Black Sabbath
Black Sabbath Vol. 4 (Remaster)	Black Sabbath
Live [Disc 1]	The Black Crowes
Live [Disc 2]	The Black Crowes

A query for the length and name of all tracks that are between 30 and 40 seconds, and of the genre Latin:

```
select ("GenreId", "Name", "Milliseconds") from "Track" where "GenreId" = 7 and "Milliseconds" > 300000
```

Table 3: Displaying records 1 - 10

row
(7,“Vai Passar”,369763)
(7,“Geni E O Zepelim”,317570)
(7,“O Estrangeiro”,374700)
(7,“Fora Da Ordem”,354011)
(7,Imperatriz,339173)
(7,Beija-Flor,327000)
(7,Viradouro,344320)
(7,“Unidos Da Tijuca”,338834)
(7,Salgueiro,305920)
(7,Portela,319608)

A query that list each country and the number of customers in that country:

```
select "Country", count("Country") as "customer_amount" from "Customer" group by "Country"
```

Table 4: Displaying records 1 - 10

Country	customer_amount
France	5
Netherlands	1
Australia	1
Chile	1
USA	13
Ireland	1
Canada	7
United Kingdom	3
Italy	1
Sweden	1
And its equivalen	t in tidyverse:

```
chinook %>% tbl("Customer") %>% group_by(Country) %>% count() %>% collect()
```

```
## # A tibble: 11 x 2
##    Country              n
##    <chr>          <int64>
##  1 France               5
##  2 Netherlands          1
##  3 Australia            1
##  4 Chile                1
##  5 USA                 13
##  6 Ireland              1
##  7 Canada               7
##  8 United Kingdom       3
##  9 Italy                1
## 10 Sweden               1
## 11 India                2
```

Finally, a query that returns the artists whose listeners span the most number of countries, ie the artist listened to by the most countries

```
select e."Name", count(distinct(a."BillingCountry")) as "countries_reached"
from "Invoice" a
join "InvoiceLine" b on a."InvoiceId" = b."InvoiceId"
join "Track" c on b."TrackId" = c."TrackId"
join "Album" d on c."AlbumId" = d."AlbumId"
join "Artist" e on d."ArtistId" = e."ArtistId"
group by e."Name" order by "countries_reached" desc
```

Table 5: Displaying records 1 - 10

Name	countries_reached
Iron Maiden	8
Led Zeppelin	7
U2	7
Creedence Clearwater Revival	7

Name	countries_reached
Metallica	6
Faith No More	5
Miles Davis	5
Lost	5
Santana	5
R.E.M.	5