Chinook Business Analysis

Robert Hazell

3/24/2019

```
library(RSQLite)
library(DBI)
# the remaining packages can be found in the tidyverse package
# this is an explicit listing
library(dplyr)
library(dbplyr)
library(ggplot2)
library(magrittr)
library(purrr)
library(tidyr)
```

Make a connection between the database and R.

```
working_dir<- "/Users/roberthazell/Desktop/Dataquest/Chinook"
setwd(working_dir)
con <- DBI::dbConnect(RSQLite::SQLite(),
                      dbname = paste(c(working_dir, "/chinook.db"), collapse = ''))
```

Take a look at the tables.

```
dbListTables(con)
```

```
 [1] "album"          "artist"         "customer"       "employee"
 [5] "genre"          "invoice"        "invoice_line"   "media_type"
 [9] "playlist"       "playlist_track" "track"
```

Now we need to create a `tbl` of each table and then convert each to a `data.frame`. This can be done by defining functions that create names for each table and assigns each of them to the respective table-turned-`data.frame`.

```
# create names for each table by iterating through the table list
table_names <- map_chr(dbListTables(con),
                       function(t) {paste(c(t, "_db"), collapse = "")})
# create data frames from each table
chinook_tables <- map(dbListTables(con),
                      function(t) {tbl(con,t) %>% as.data.frame()})
# link the table names to the tables, respectively
names(chinook_tables) <- table_names
# attach to reference each table without calling chinook_tables
attach(chinook_tables)
```

Selecting Albums to Purchase

The Chinook record store has just signed a deal with a new record label, and the task is to select the first three albums that will be added to the store from a list of four. All four albums are by artists that don't have any tracks in the store right now - we have the artist names, and the genre of music they produce.

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

The record label specializes in artists from the USA, and they have given Chinook some money to advertise the new albums in the USA, so we're interested in finding out which genres sell the best in the USA. The following query answers this question.

```
genres_USA <- genre_db %>%
  inner_join(track_db, by = 'genre_id', suffix = c("_genre", "_track")) %>%
  inner_join(invoice_line_db, by = 'track_id', suffix = c("_track", "_invoice_line")) %>%
  inner_join(invoice_db, by = 'invoice_id')

# make sure no column names are duplicated
any(duplicated(colnames(genres_USA)))
```

```
[1] FALSE
```

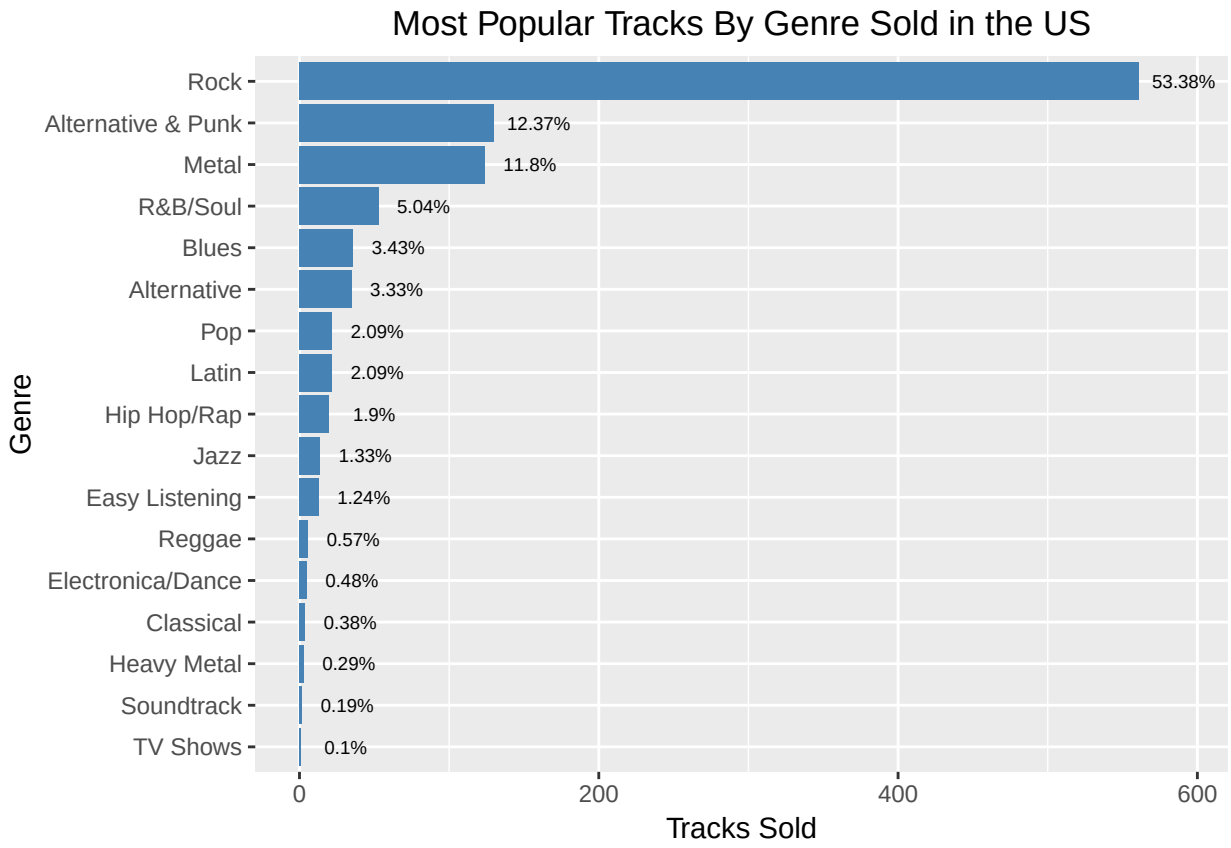
```
# create the summary
genres_USA_summary <- genres_USA %>%
  group_by('Genre'= name_genre) %>%
  filter(billing_country == "USA") %>%
  summarise(Total = length(Genre)) %>%
  arrange(desc(Total)) %>%
  mutate('Percent Sold' = round(Total/sum(Total) * 100, 2))

# create formatted table of results
genres_USA_summary %>%
  kable(align = rep('c',3)) %>%
  kable_styling(bootstrap_options = "striped", full_width = F)
```

Genre	Total	Percent Sold
Rock	561	53.38
Alternative & Punk	130	12.37
Metal	124	11.80
R&B/Soul	53	5.04
Blues	36	3.43
Alternative	35	3.33
Latin	22	2.09
Pop	22	2.09
Hip Hop/Rap	20	1.90
Jazz	14	1.33
Easy Listening	13	1.24
Reggae	6	0.57
Electronica/Dance	5	0.48
Classical	4	0.38
Heavy Metal	3	0.29
Soundtrack	2	0.19
TV Shows	1	0.10

Here's a bar plot summarizing this information.

```
ggplot(genres_USA_summary, aes(x = reorder(Genre, Total), y = Total)) +
  geom_bar(fill = 'steel blue', stat = "identity") +
  coord_flip() +
  geom_text(aes(label = paste0(`Percent Sold`, '%')),
                size = 2.5, nudge_y = 30) +
  ggtitle("Most Popular Tracks By Genre Sold in the US") +
  ylab("Tracks Sold") +
  xlab("Genre") +
  theme(plot.title = element_text(hjust = 0.5))
```



Based on the sales of tracks across different genres in the USA, we should purchase the new albums by the following artists:

- Red Tone (Punk)
- Slim Jim Bites (Blues)
- Meteor and the Girls (Pop)

It's worth noting that combined, these three genres only make up only 17% of total sales, so we should be on the lookout for artists and albums from the 'Rock' genre, which accounts for over 53% of sales.

Analyzing Employee Sales Performance

Each customer of the Chinook store gets assigned to a sales support agent within the company when they first make a purchase. Management requests an analysis of customer purchases belonging to each employee to see if any sales support agent is performing either better or worse than the others.

One method is to determine how many sales an employee generates, and this requires joining the `employee_db`,

customer_db, and invoice_db tables together.

```
# append "emp" to avoid duplicate names when joining employee and customer tables
# retrieve the original column names - keep the first column the same
emp_cols <- colnames(employee_db)[2:dim(employee_db)[2]]
# create function to append "emp"
new_emp_cols <- map_chr(emp_cols, function(t) {paste0(t, "_emp")})
# rename columns 2-15
names(employee_db)[2:15] <- new_emp_cols
# now join and summarise
employee_db %>%
  inner_join(customer_db, by = c('employee_id' = 'support_rep_id')) %>%
  inner_join(invoice_db, by = 'customer_id') %>%
  select(employee_id:first_name_emp, total, hire_date_emp) %>%
  group_by(employee_id, last_name_emp, first_name_emp, hire_date_emp) %>%
  summarise(`Total Sales ($)` = sum(total)) %>%
  unite("Employee Name", c("first_name_emp", "last_name_emp"), sep=" ") %>%
  arrange(desc(`Total Sales ($)`))
```

```
# A tibble: 3 x 4
# Groups:   employee_id [3]
  employee_id `Employee Name` hire_date_emp       `Total Sales ($)`
        <int> <chr>           <chr>                           <dbl>
1           3 Jane Peacock    2017-04-01 00:00:00             1732.
2           4 Margaret Park   2017-05-03 00:00:00             1584
3           5 Steve Johnson   2017-10-17 00:00:00             1394.
```

We can see that only three employes work at Chinook. Jane, who has the most sales, is also the longest hired amongst the three.