Data Visualization in R

Kittirot Pannukul

Libraries

```
library(tidyverse)
library(lubridate)
library(patchwork)
library(scales)
library(glue)
```

Data Overview

Use data queried from the Chinook database, which represents a digital media store consisting of tables for artists, albums, media tracks, invoices, and customers.

```
tracks <- read_csv("small_chinook_tracks.csv")
invoices <- read_csv("small_chinook_invoices.csv")
```

```
glimpse(tracks)
```

```
## Rows: 3,503
## Columns: 7
## $ artist <chr> "AC/DC", "AC/DC", "AC/DC", "AC/DC", "AC/DC", "AC/DC", "AC/DC", ~
## $ album  <chr> "For Those About To Rock We Salute You", "For Those About To Ro~
## $ track  <chr> "For Those About To Rock (We Salute You)", "Put The Finger On Y~
## $ genre  <chr> "Rock", "Rock", "Rock", "Rock", "Rock", "Rock", "Rock", "Rock",~
## $ min    <dbl> 5.73, 3.43, 3.90, 3.51, 3.39, 4.39, 3.33, 4.39, 3.43, 4.51, 5.7~
## $ mb     <dbl> 10.65, 6.40, 7.28, 6.54, 6.29, 8.21, 6.26, 8.20, 6.40, 8.41, 5.~
## $ price  <dbl> 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.9~
```

```
glimpse(invoices)
```

```
## Rows: 2,240
## Columns: 10
## $ name         <chr> "Luís Gonçalves", "Luís Gonçalves", "Luís Gonçalves", "Lu~
## $ company      <chr> "Embraer - Empresa Brasileira de Aeronáutica S.A.", "Embr~
## $ date         <dttm> 2010-03-11, 2010-03-11, 2010-06-13, 2010-06-13, 2010-06-~
## $ bill_city    <chr> "São José dos Campos", "São José dos Campos", "São José d~
## $ bill_country <chr> "Brazil", "Brazil", "Brazil", "Brazil", "Brazil", "Brazil~
## $ total_price  <dbl> 3.98, 3.98, 3.96, 3.96, 3.96, 3.96, 5.94, 5.94, 5.94, 5.9~
```

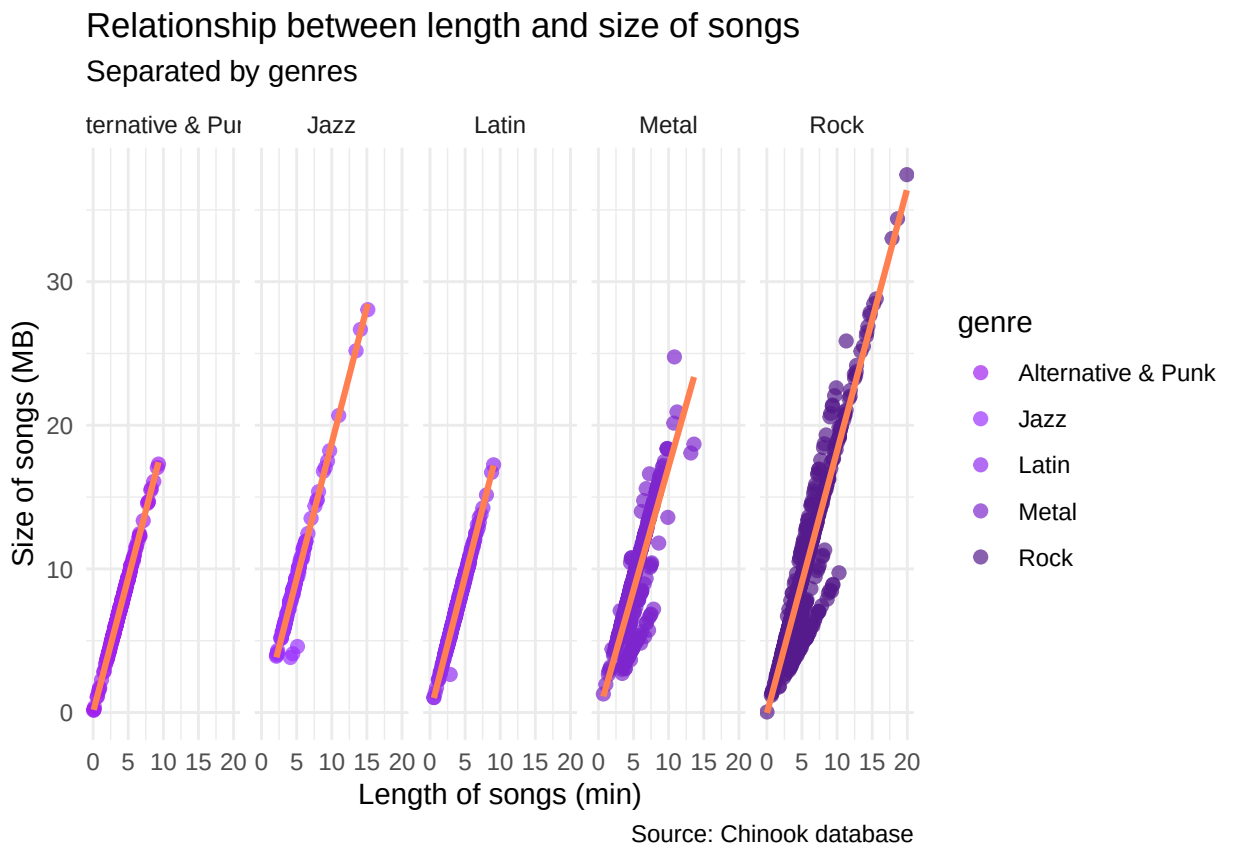
```
## $ unit_price   <dbl> 1.99, 1.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.99, 0.9~
## $ quantity     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ track        <chr> "Experiment In Terra", "Take the Celestra", "Shout It Out~
## $ genre        <chr> "Sci Fi & Fantasy", "Sci Fi & Fantasy", "Rock", "Rock", "~
```

Data Visualization

Chart 1: Relationship between length and size of songs

```
five_genre <- tracks %>%
  count(genre) %>%
  arrange(desc(n)) %>%
  head(5)

tracks %>%
  filter(genre %in% five_genre$genre,
         min < 25) %>%
  ggplot(aes(min, mb, color = genre)) +
    geom_point(size = 2,
               alpha = 0.7) +
    geom_smooth(method = 'lm',
                se = F,
                color = "coral",
                size = 1.1) +
    scale_color_manual(values = c("purple",
                                  "purple1",
                                  "purple2",
                                  "purple3",
                                  "purple4")) +
    facet_wrap(~ genre,
               ncol = 5) +
    theme_minimal() +
    labs(title= "Relationship between length and size of songs",
         subtitle = "Separated by genres",
         x = "Length of songs (min)",
         y = "Size of songs (MB)",
         caption = "Source: Chinook database")
```

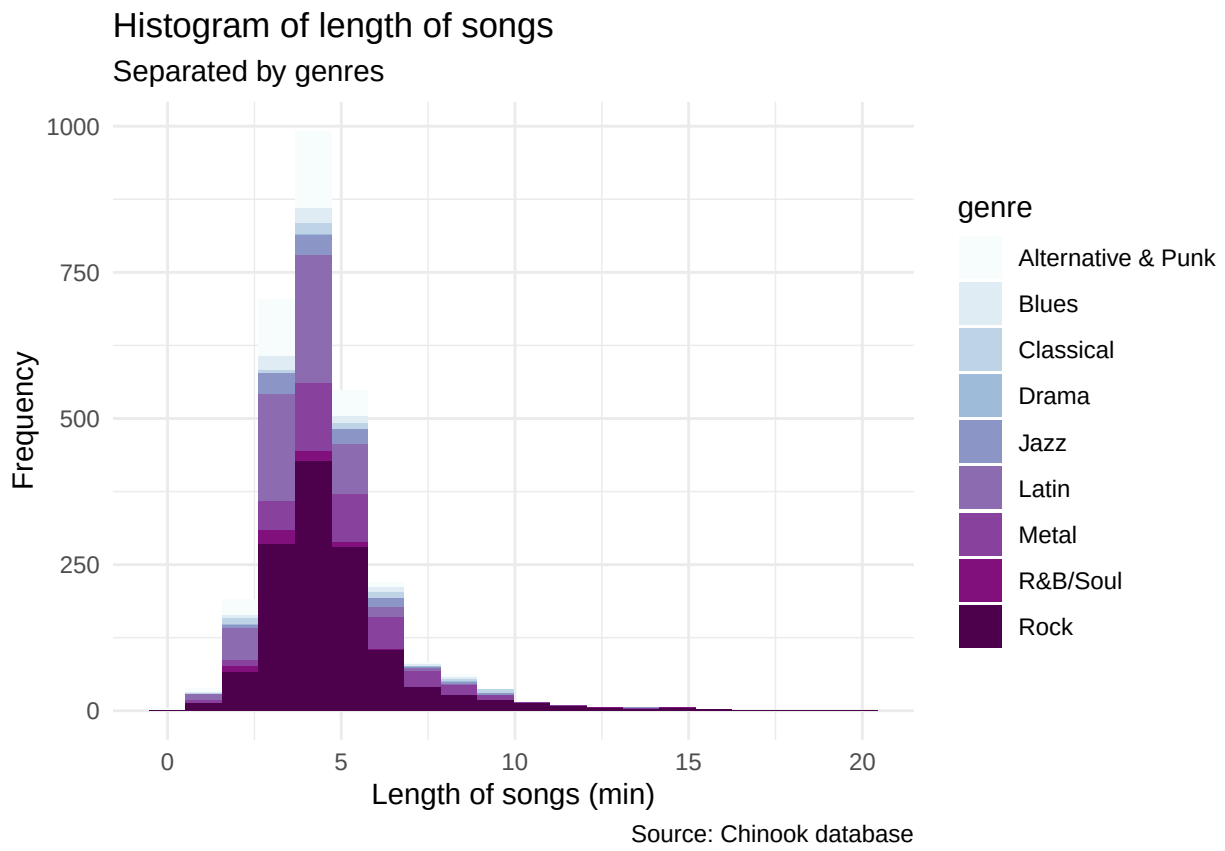


The scatter plot shows the relationship between length and size of songs separated by genres. The longer the song is, the bigger the size will be.

Chart 2: Histogram of length of songs

```
ten_genre <- tracks %>%
  count(genre) %>%
  arrange(desc(n)) %>%
  head(10)

tracks %>%
  filter(min < 20,
         genre %in% ten_genre$genre) %>%
  ggplot(aes(min, fill = genre)) +
    geom_histogram(bins = 20) +
    scale_fill_brewer(type = "seq", palette = "BuPu") +
    theme_minimal() +
    labs(title = "Histogram of length of songs",
         subtitle = "Separated by genres",
         x = "Length of songs (min)",
         y = "Frequency",
         caption = "Source: Chinook database")
```

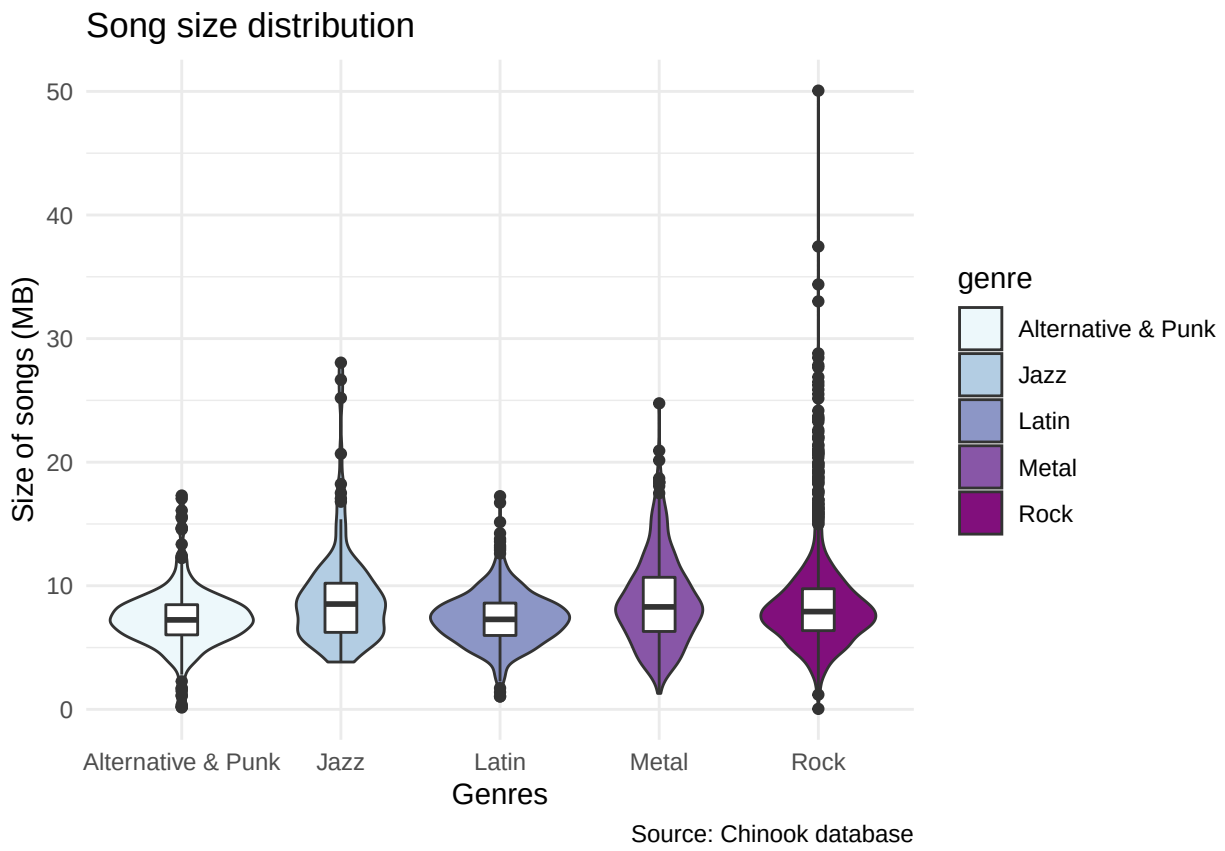


This histogram shows the frequency of the length of songs (min) by genre.

Chart 3: Size of songs by genre

```
five_genre <- tracks %>%
  count(genre) %>%
  arrange(desc(n)) %>%
  head(5)

tracks %>%
  filter(genre %in% five_genre$genre) %>%
  ggplot(aes(genre, mb, fill = genre)) +
  geom_violin() +
  geom_boxplot(width=0.2, fill = "white") +
  scale_fill_brewer(palette = "BuPu") +
  theme_minimal() +
  labs(title = "Song size distribution",
       x = "Genres",
       y = "Size of songs (MB)",
       caption = "Source: Chinook database")
```



The violin plot and box plot show the distribution of song size by genre. There are too many outliers in Rock music.

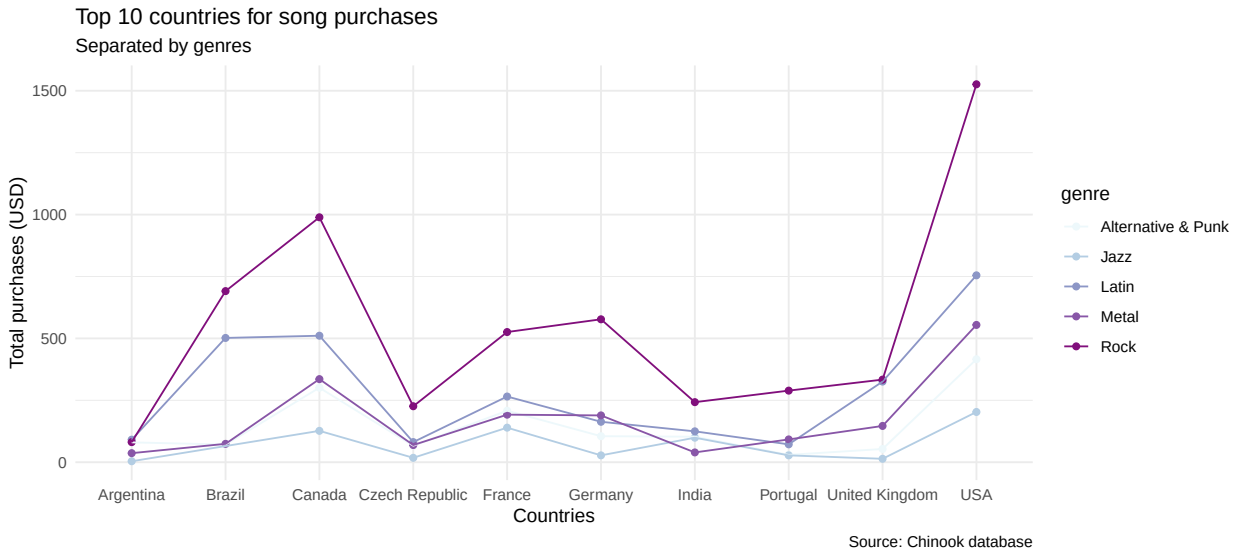
Chart 4: Top 10 countries for song purchases

```
five_genre_invoices <- invoices %>%
  count(genre) %>%
  arrange(desc(n)) %>%
  head(5)

top_spending_countries <- invoices %>%
  group_by(country = bill_country) %>%
  summarise(sum_quantity = sum(quantity),
            sum_price = sum(total_price)) %>%
  arrange(desc(sum_quantity)) %>%
  head(10)

invoices %>%
  filter(genre %in% five_genre_invoices$genre,
         bill_country %in% top_spending_countries$country) %>%
  group_by(country = bill_country, genre) %>%
  summarise(total = sum(total_price)) %>%
  ggplot(aes(country, total, group = genre, color = genre)) +
    geom_point() +
```

```
    geom_line() +
    scale_color_brewer(palette = "BuPu") +
    theme_minimal() +
    labs(title = "Top 10 countries for song purchases",
         subtitle = "Separated by genres",
         x = "Countries",
         y = "Total purchases (USD)",
         caption = "Source: Chinook database")
```

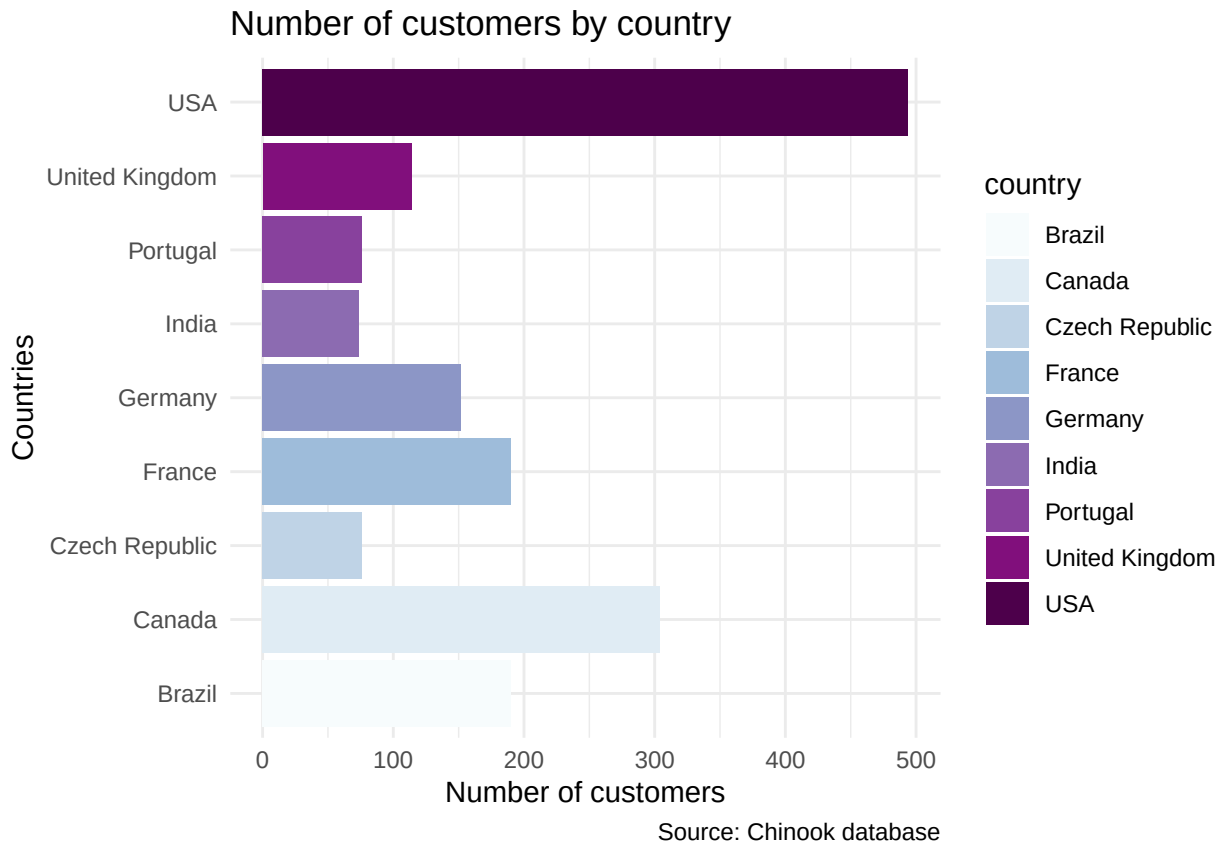


It can be seen that Rock music was purchased the most by all the top 10 spending countries on music purchases.

Chart 5: Number of customers by country

```
country <- invoices %>%
  group_by(bill_country) %>%
  summarise(n = n()) %>%
  filter(n > 50)

invoices %>%
  filter(bill_country %in% country$bill_country) %>%
  select(country = bill_country) %>%
  ggplot(aes(y = country, fill = country, )) +
    geom_bar() +
    scale_fill_brewer(palette = "BuPu") +
    theme_minimal() +
    labs(title = "Number of customers by country",
         x = "Number of customers",
         y = "Countries",
         caption = "Source: Chinook database")
```



The bar plot shows that the number of customers purchasing songs in the USA was the highest.