Spotify Streaming Analysis: Pre and Post Pandemic

Your Name

2025-03-05

Introduction

This report explores the effect of the COVID-19 pandemic on Spotify music consumption. The dataset contains daily streaming numbers across multiple artists and countries from January 2019 to December 2021. The goal is to assess whether there was a noticeable impact on streaming trends after March 2020.

Load Data

```
# Load the dataset
spotify_data <- read_csv("spotify_data.csv")
```

```
## Rows: 219200 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (4): country, artist, song, pandemic_period
## dbl  (1): streams
## date (1): date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(spotify_data)
```

```
## # A tibble: 6 x 6
##   date       country artist   song   streams pandemic_period
##   <date>     <chr>   <chr>    <chr>    <dbl> <chr>
## 1 2019-01-01 USA     Artist 1 Song 1  457488 Before Pandemic
## 2 2019-01-01 UK      Artist 2 Song 2  468601 Before Pandemic
## 3 2019-01-01 USA     Artist 3 Song 3  143784 Before Pandemic
## 4 2019-01-01 UK      Artist 4 Song 4  415393 Before Pandemic
## 5 2019-01-01 USA     Artist 5 Song 5  321231 Before Pandemic
## 6 2019-01-01 UK      Artist 6 Song 1  260029 Before Pandemic
```

```
# View basic structure
str(spotify_data)
```

```
## spc_tbl_ [219,200 x 6] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ date           : Date[1:219200], format: "2019-01-01" "2019-01-01" ...
##  $ country        : chr [1:219200] "USA" "UK" "USA" "UK" ...
##  $ artist         : chr [1:219200] "Artist 1" "Artist 2" "Artist 3" "Artist 4" ...
##  $ song           : chr [1:219200] "Song 1" "Song 2" "Song 3" "Song 4" ...
##  $ streams        : num [1:219200] 457488 468601 143784 415393 321231 ...
##  $ pandemic_period: chr [1:219200] "Before Pandemic" "Before Pandemic" "Before Pandemic" "Before Pan
##  - attr(*, "spec")=
##   .. cols(
##   ..   date = col_date(format = ""),
##   ..   country = col_character(),
##   ..   artist = col_character(),
##   ..   song = col_character(),
##   ..   streams = col_double(),
##   ..   pandemic_period = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data Exploration

Summary Statistics

```
# Summary of streaming numbers
summary(spotify_data$streams)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    862.6 116896.6 233932.9 235138.4 350972.1 499996.0
```

```
# Check for missing values
colSums(is.na(spotify_data))
```

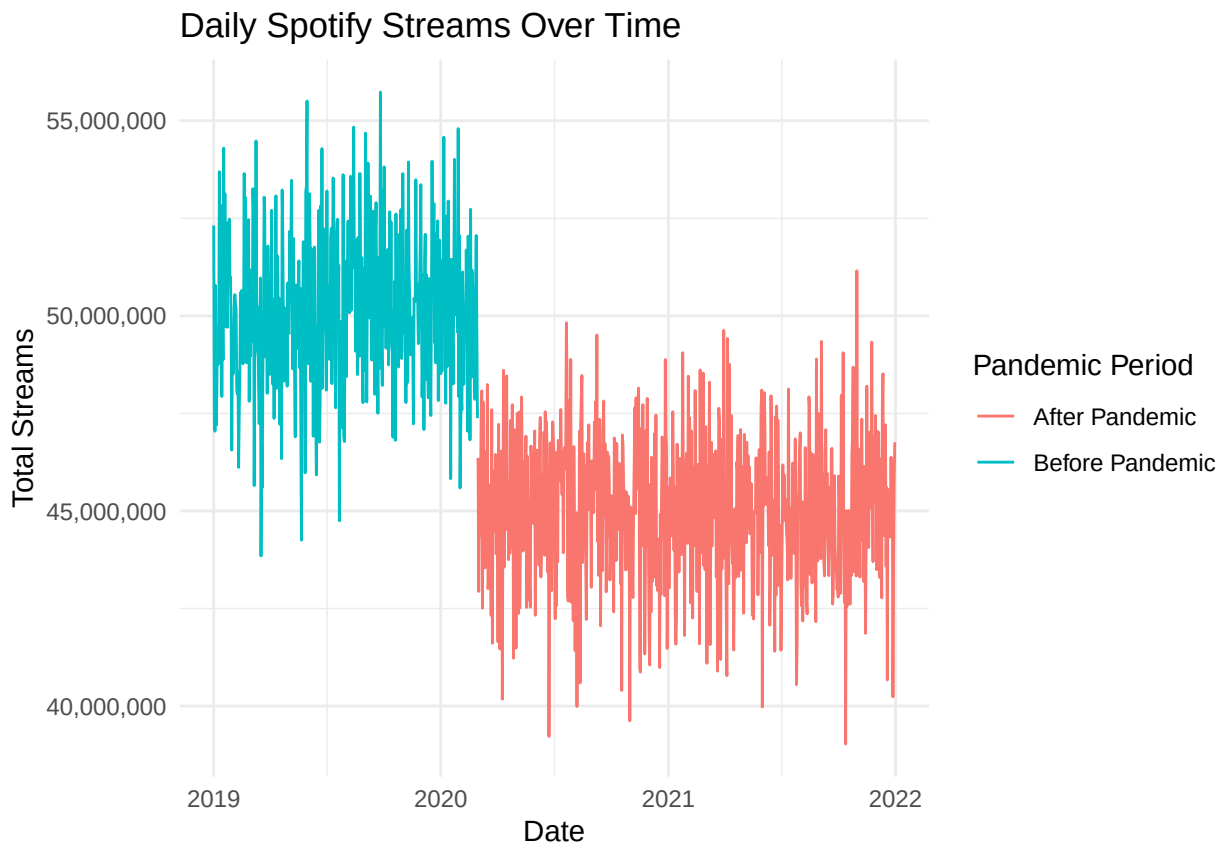
```
##            date         country          artist            song         streams
##               0               0               0               0               0
## pandemic_period
##               0
```

Streaming Trends Over Time

```
# Aggregate streams per day
daily_streams <- spotify_data %>%
  group_by(date, pandemic_period) %>%
  summarize(total_streams = sum(streams), .groups = 'drop')

# Plot daily streaming trends
ggplot(daily_streams, aes(x = date, y = total_streams, color = pandemic_period)) +
  geom_line() +
  labs(title = "Daily Spotify Streams Over Time",
       x = "Date",
```

```
       y = "Total Streams",
       color = "Pandemic Period") +
  theme_minimal() +
  scale_y_continuous(labels = comma)
```



Comparing Pre-Pandemic and Post-Pandemic Streaming

```
# Average streams before and after the pandemic
stream_comparison <- spotify_data %>%
  group_by(pandemic_period) %>%
  summarize(avg_streams = mean(streams), .groups = 'drop')

# Print comparison table
kable(stream_comparison, caption = "Comparison of Streaming Activity Before and After March 2020")
```

Table 1: Comparison of Streaming Activity Before and After March 2020

pandemic_period	avg_streams
After Pandemic	225155.6
Before Pandemic	250899.5

Country-wise Analysis

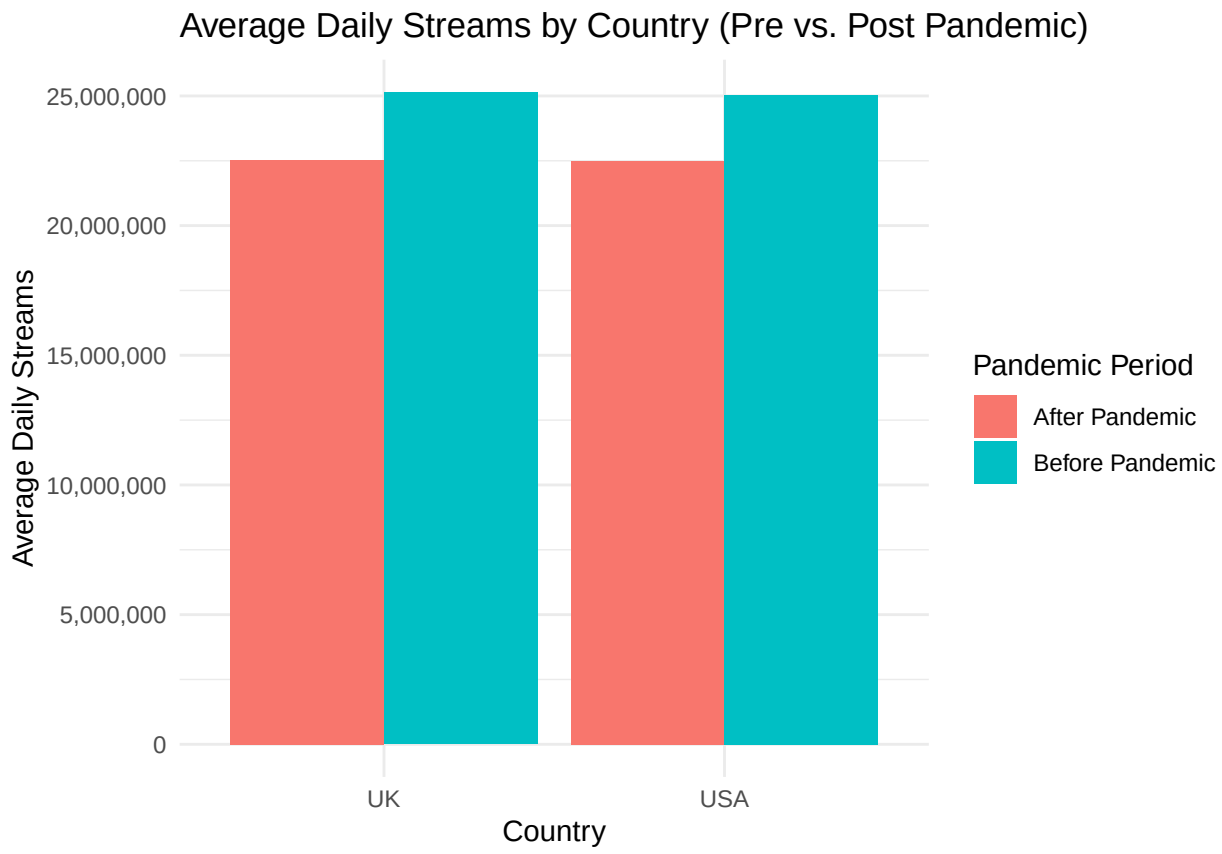
```
daily_country_streams <- spotify_data %>%
  group_by(date, country, pandemic_period) %>%
  summarize(daily_streams = sum(streams), .groups = "drop")

country_streams <- daily_country_streams %>%
  group_by(country, pandemic_period) %>%
  summarize(avg_daily_streams = mean(daily_streams), .groups = "drop")

# Print summary
summary(country_streams)
```

```
##    country          pandemic_period    avg_daily_streams
##  Length:4           Length:4           Min.   :22491547
##  Class :character   Class :character   1st Qu.:22527560
##  Mode  :character   Mode  :character   Median :23790193
##                                        Mean   :23802754
##                                        3rd Qu.:25065388
##                                        Max.   :25139082
```

```
# Plot streams by country
ggplot(country_streams, aes(x = country, y = avg_daily_streams, fill = pandemic_period)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Average Daily Streams by Country (Pre vs. Post Pandemic)",
       x = "Country",
       y = "Average Daily Streams",
       fill = "Pandemic Period") +
  theme_minimal() +
  scale_y_continuous(labels = comma)
```



Conclusion

We see a sharp decline in daily streams after pandemic. Also, the average streams decline post-pandemic for both US and UK.