Maven Music Challenge

Koundi(Cody)

Maven Music Challenge

About The Data Set

Spotify user's complete music streaming history data, including timestamps, track, artist, and album names, and reasons for playing and ending each track.

Objective:

To identify patterns in user listening history from the 12 years data

Methodology

1. First let us install the packages tidyverse and lubridate.

   Tidyverse assists in data importing, tidying up, maipulating and Visualizing

   lubridate assists in working with dates and times

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```
library(lubridate)
library(hrbrthemes)
```

2. Read the csv file

   Read the csv file using read.csv command

   ```
   df<- read.csv('Spotify.csv')
   head(df)
   ```

   ```
          spotify_track_uri             ts   platform ms_played
   1 2J3n32GeLmMjwuAzyhcSNe 8/07/2013 2:44 web player      3185
   2 1oHxIPqJyvAYHy0PVrDU98 8/07/2013 2:45 web player     61865
   3 487OPlneJNni3NWC8SYqhW 8/07/2013 2:50 web player    285386
   4 5IyblF777jLZj1vGHG2UD3 8/07/2013 2:52 web player    134022
   5 0GgAAB0ZMllFhbNc3mAod0 8/07/2013 3:17 web player         0
   6 50VNvhzyaSplJCKWchN7a8 8/07/2013 3:17 web player     63485
                                        track_name       artist_name
   1                           Say It, Just Say It      The Mowgli's
   2 Drinking from the Bottle (feat. Tinie Tempah)     Calvin Harris
   3                                   Born To Die      Lana Del Rey
   4                              Off To The Races      Lana Del Rey
   5                                     Half Mast Empire Of The Sun
   6                                    Impossible      James Arthur
                             album_name reason_start reason_end shuffle skipped
   1               Waiting For The Dawn     autoplay   clickrow   FALSE   FALSE
   2                          18 Months     clickrow   clickrow   FALSE   FALSE
   3 Born To Die - The Paradise Edition     clickrow    unknown   FALSE   FALSE
   4 Born To Die - The Paradise Edition    trackdone   clickrow   FALSE   FALSE
   5                 Walking On A Dream     clickrow    nextbtn   FALSE   FALSE
   6                         Impossible     clickrow   clickrow   FALSE   FALSE
   ```

3. Lets convert ts column to a proper date time format

   ```
   df$ts<- as.POSIXct(df$ts,format="%d/%m/%Y %H:%M")
   ```

4. Lets check and remove any missing values

   ```
   sum(is.na(df))
   ```

   ```
   [1] 0
   ```

   ```
   which(is.na(df))
   ```

   ```
   integer(0)
   ```

```
df<-df%>% drop_na()
```

5. Lets convert the shuffle and skipped columns to integers true for 1 and false for 0

```
df$shuffle <- as.integer(as.logical(df$shuffle))
df$skipped <- as.integer(as.logical(df$skipped))
head(df)
```

```
       spotify_track_uri                  ts   platform ms_played
1 2J3n32GeLmMjwuAzyhcSNe 2013-07-08 02:44:00 web player      3185
2 1oHxIPqJyvAYHy0PVrDU98 2013-07-08 02:45:00 web player     61865
3 487OPlneJNni3NWC8SYqhW 2013-07-08 02:50:00 web player    285386
4 5IyblF777jLZj1vGHG2UD3 2013-07-08 02:52:00 web player    134022
5 0GgAAB0ZMllFhbNc3mAod0 2013-07-08 03:17:00 web player         0
6 50VNvhzyaSplJCKWchN7a8 2013-07-08 03:17:00 web player     63485
                                     track_name       artist_name
1                           Say It, Just Say It      The Mowgli's
2 Drinking from the Bottle (feat. Tinie Tempah)     Calvin Harris
3                                   Born To Die      Lana Del Rey
4                              Off To The Races      Lana Del Rey
5                                    Half Mast Empire Of The Sun
6                                    Impossible      James Arthur
                          album_name reason_start reason_end shuffle skipped
1               Waiting For The Dawn     autoplay   clickrow       0       0
2                          18 Months     clickrow   clickrow       0       0
3 Born To Die - The Paradise Edition     clickrow    unknown       0       0
4 Born To Die - The Paradise Edition    trackdone   clickrow       0       0
5                 Walking On A Dream     clickrow    nextbtn       0       0
6                         Impossible     clickrow   clickrow       0       0
```

2. Create a new column for the hour of the day

```
df$hour<- hour(df$ts)
```

Exploratory Data Anlysis (EDA)

6. Most Played Tracks

```
most_played_tracs<-df%>%
  group_by(track_name,artist_name)%>%
  summarise(total_plays =n(), total_time_played=sum(ms_played))%>%
  arrange(desc(total_plays))
```

```
`summarise()` has grouped output by 'track_name'. You can override using the
`.groups` argument.
```

```
print(most_played_tracs)
```

```
# A tibble: 14,639 x 4
# Groups:   track_name [13,839]
   track_name                         artist_name   total_plays total_time_played
   <chr>                              <chr>               <int>             <int>
 1 Ode To The Mets                    The Strokes           207          67431580
 2 In the Blood                       John Mayer            181          38427087
 3 Dying Breed                        The Killers           166          36182653
 4 Caution                            The Killers           164          35619945
 5 19 Dias y 500 Noches - En Directo  Joaquín Sabi~         148          42914042
 6 All These Things That I've Done    The Killers           142          35754915
 7 Concerning Hobbits                 Howard Shore          142          19239222
 8 Come Together - Remastered 2009    The Beatles           137          22682658
 9 Yesterday - Remastered 2009        The Beatles           134          14934173
10 Crucify Your Mind                  Rodríguez             131          19842588
# i 14,629 more rows
```

7. Most Played artists

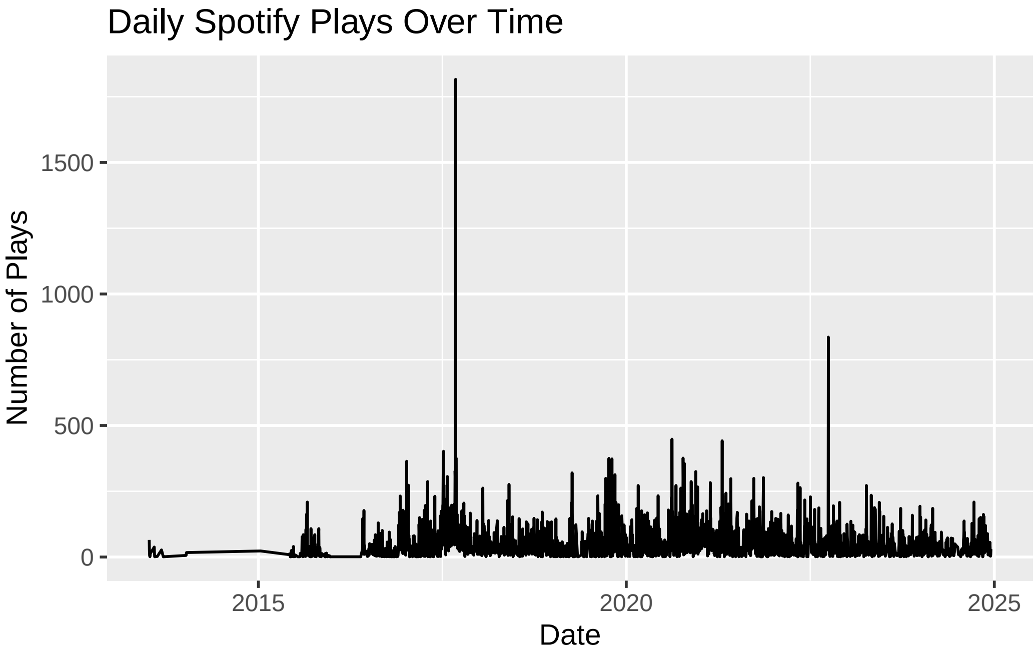
```
most_played_artists<-df%>%
  group_by(artist_name)%>%
  summarise(total_plays =n(), total_time_played=sum(ms_played))%>%
  arrange(desc(total_plays))

print(most_played_artists)
```

```
# A tibble: 4,113 x 3
   artist_name        total_plays total_time_played
   <chr>                    <int>             <int>
 1 The Beatles              13621        1210184552
 2 The Killers               6878        1059556516
 3 John Mayer                4855         725219443
 4 Bob Dylan                 3814         569456396
 5 Paul McCartney            2697         357354370
 6 Led Zeppelin              2482         248338279
 7 Johnny Cash               2478         239690064
 8 The Rolling Stones        2390         307917009
 9 Radiohead                 2305         216657418
10 The Black Keys            2231         192035798
# i 4,103 more rows
```

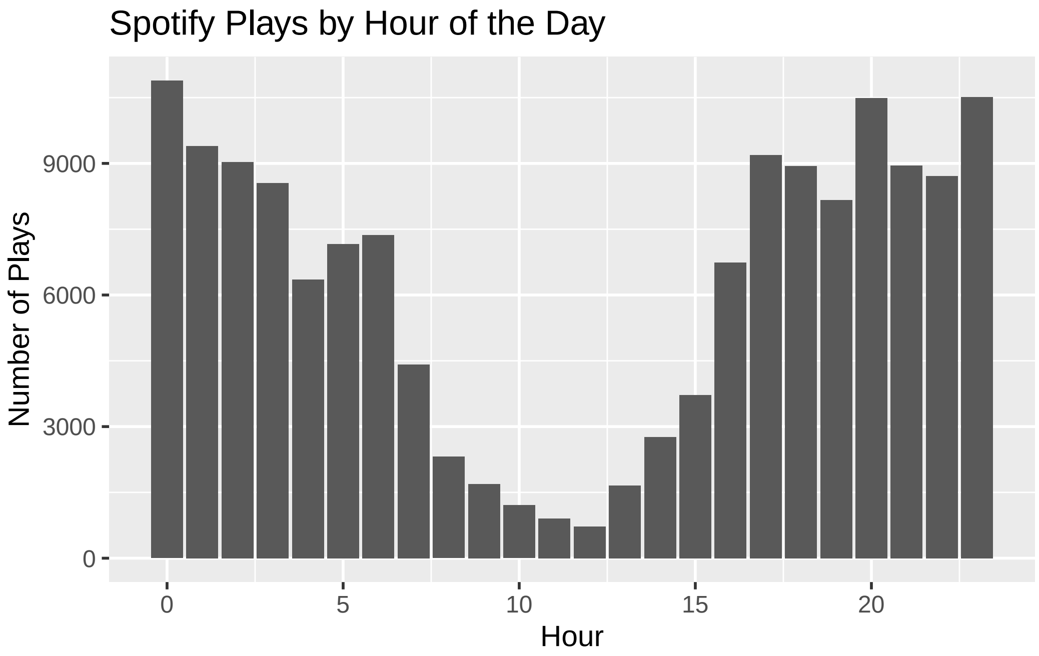
8. Listening Patterns over time

```
df%>%
  group_by(date = as.Date(ts))%>%
  summarise(daily_plays=n())%>%
  ggplot(aes(x=date,y=daily_plays))+
  geom_line()+
  labs(title = "Daily Spotify Plays Over Time", x = "Date", y = "Number of Plays")
```



9. Listening patterns by hour of the day

```
df%>%
  group_by(hour)%>%
  summarise(hourly_plays =n())%>%
  ggplot(aes(x=hour,y=hourly_plays))+
  geom_bar(stat ="identity")+
  labs(title = "Spotify Plays by Hour of the Day", x = "Hour", y = "Number of Plays")
```



10. Skipped vs Shuffle tracks

```
skip_tracks<- df%>%
  group_by(skipped)%>%
  summarise(count=n())%>%
  mutate(percentage = count/sum(count)*100)
skip_tracks
```

```
# A tibble: 2 x 3
  skipped  count percentage
    <int>  <int>      <dbl>
1       0 141991      94.7
2       1   7869       5.25
```

11. Shuffle usage

```
shuffle_tracks<-df%>%
  group_by(shuffle)%>%
  summarise(count=n())%>%
  mutate(percentage = count/sum(count)*100)
shuffle_tracks
```

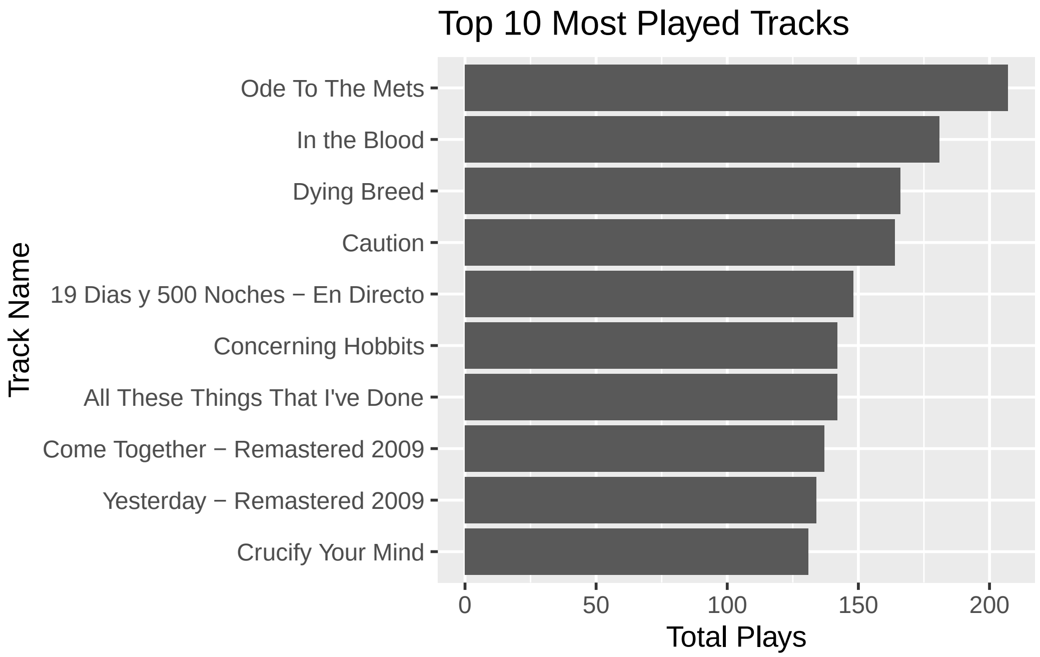
```
# A tibble: 2 x 3
  shuffle  count percentage
    <int>  <int>      <dbl>
1       0  38277       25.5
```

```
2        1 111583       74.5
```

Visualization

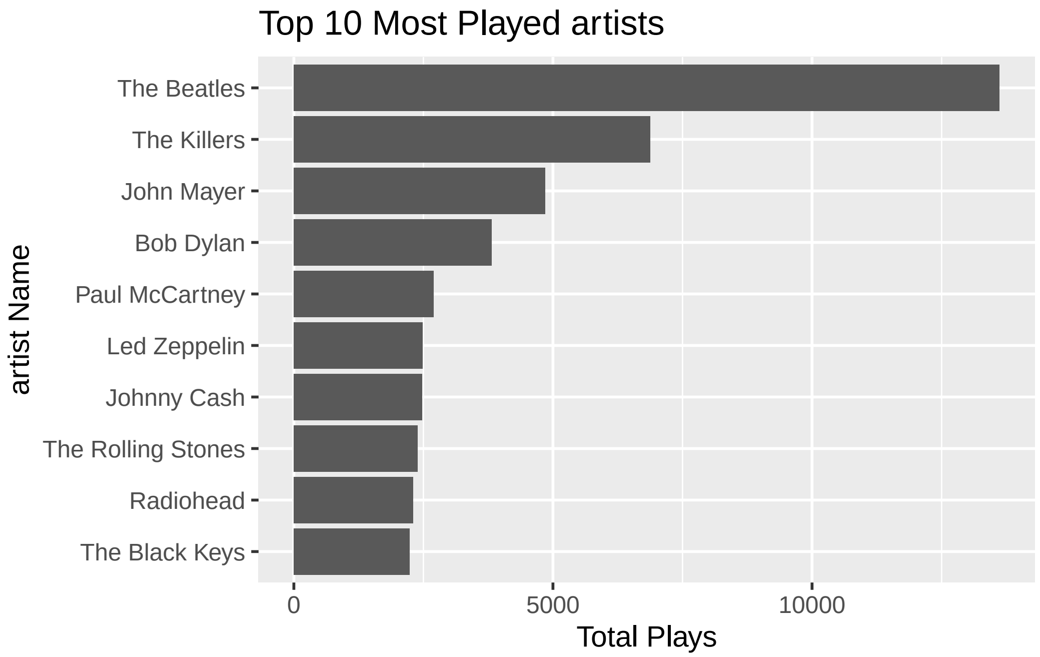
12. Top 10 most played tracks

```r
most_played_tracs%>%
  head(10)%>%
  ggplot(aes(x=reorder(track_name,total_plays), y=total_plays))+
  geom_bar(stat = "identity")+
  coord_flip()+
  labs(title="Top 10 Most Played Tracks", x = "Track Name", y = "Total Plays")
```



13. Top 10 most played artists

```r
most_played_artists%>%
  head(10)%>%
  ggplot(aes(x=reorder(artist_name,total_plays), y=total_plays))+
  geom_bar(stat = "identity")+
  coord_flip()+
  labs(title="Top 10 Most Played artists", x = "artist Name", y = "Total Plays")
```

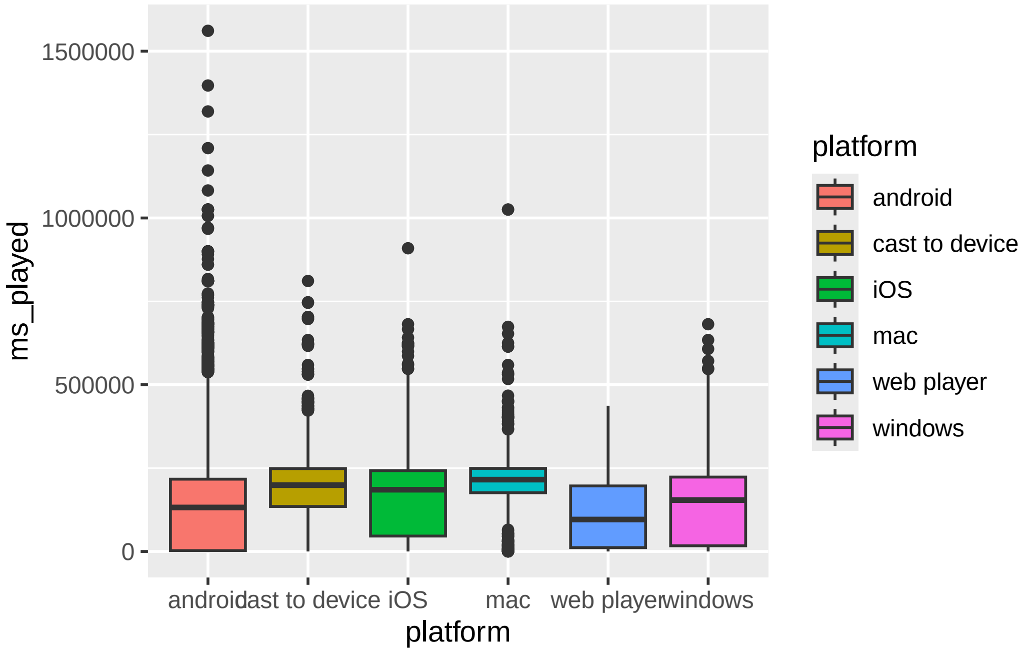


14. Lets make a boxplot to compare distributions across platforms

    - Android has the highest number of plays (close to 1,500,000).
    - Cast to Device and iOS have significantly fewer plays compared to Android.
    - Mac, Web Player, and Windows have the lowest number of plays.

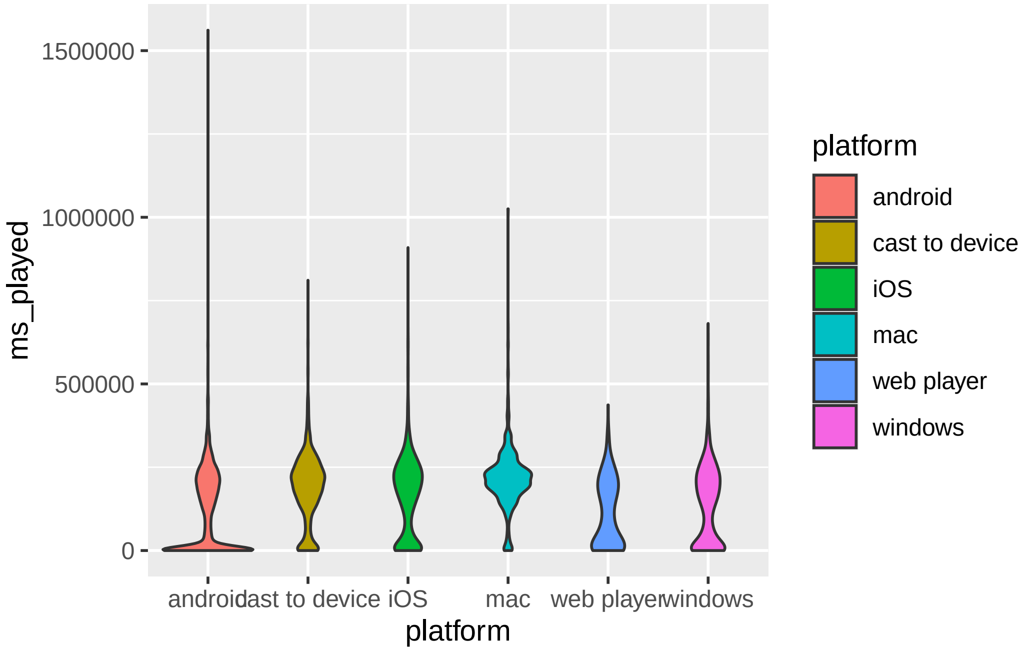
```
library(tidyverse)
library(hrbrthemes)
library(viridisLite)

df %>%
  ggplot(aes(x=platform, y=ms_played, fill = platform))+
  geom_boxplot()
```



15. Lets look at the Violin plot

```
df%>%
  ggplot(aes(x=platform, y=ms_played, fill =platform))+
  geom_violin()
```



16. Based on the analysis, we can draw several insights:

    1. **Most Played Tracks and Artists**: The most played tracks and artists can help identify user preferences.

    2. **Listening Patterns**: Users tend to listen more during certain hours of the day, which could be useful for targeted marketing.

    3. **Skipped Tracks**: A significant percentage of tracks are skipped, which might indicate user dissatisfaction with certain tracks or playlists.

    4. **Shuffle Usage**: The shuffle feature is used frequently, suggesting that users enjoy a randomized listening experience.