ISQA 522 Final Project

Skye Gilbreth

1/31/2021

Introduction

I love music. Since March 2020 (when the pandemic started), I have been home a lot more than ever before! To stay focused and motivated, I find a multitude of ways to occupy myself. One of the ways I am able to do this is through music. So, for my final project, I decided to analyze music from Spotify. This project will take what I've learned these past four weeks and apply it to data collected from Spotify. What I will be doing with this data is looking for patterns from the year 2020 to now in the trending music.

About the Data

This data was pulled off of Kaggle's website:

https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks

The raw data contains 174,390 observations and 19 variables. The original poster of this data from Kaggle pulled the data by using the Spotify web API.

Load Data

```
# read in the data from a CSV file
Spotifydata <- read.csv("/cloud/project/Final/Data/data.csv")

# display the data in the newly created data frame
head(Spotifydata)
```

```
##   acousticness                          artists danceability duration_ms
energy explicit                    id
## 1     0.991000                    ['Mamie Smith']        0.598      168333
```

```
0.224        0 0cS0A1fUEUd1EW3FcF8AEI
## 2     0.643000         ["Screamin' Jay Hawkins"]        0.852      150200
0.517        0 0hbkKFIJm7Z05H8Zl9w30f
## 3     0.993000                  ['Mamie Smith']        0.647      163827
0.186        0 11m7laMUgmOKqI3oYzuhne
## 4     0.000173               ['Oscar Velazquez']        0.730      422087
0.798        0 19Lc5SfJJ5O1oaxY0fpwfh
## 5     0.295000                         ['Mixe']        0.704      165224
0.707        1 2hJjbsLCytGsnAHfdsLejp
## 6     0.996000 ['Mamie Smith & Her Jazz Hounds']        0.424      198627
0.245        0 3HnrHGLE9u2MjHtdobfWl9
##   instrumentalness key liveness loudness mode
name popularity
## 1        0.0005220   5   0.3790  -12.628    0
Keep A Song In Your Soul         12
## 2        0.0264000   5   0.0809   -7.261    0
I Put A Spell On You           7
## 3        0.0000176   0   0.5190  -12.098    1
Golfing Papa          4
## 4        0.8010000   2   0.1280   -7.311    1 True House Music - Xavier Sa
ntos & Carlos Gomix Remix         17
## 5        0.0002460  10   0.4020   -6.036    0
Xuniverxe          2
## 6        0.7990000   5   0.2350  -11.470    1                           Cr
azy Blues - 78rpm Version          9
##   release_date speechiness   tempo valence year
## 1         1920      0.0936 149.976  0.6340 1920
## 2   1920-01-05      0.0534  86.889  0.9500 1920
## 3         1920      0.1740  97.600  0.6890 1920
## 4   1920-01-01      0.0425 127.997  0.0422 1920
## 5   1920-10-01      0.0768 122.076  0.2990 1920
## 6         1920      0.0397 103.870  0.4770 1920
```

Code book

A code book is a dictionary of the variables in a data set that provides their names, labels and any relevant value labels. I am now going to document the variables by reading in the xlsx codebook I created into a data frame. Doing this, will help during the data cleanup and analysis phases to help deepen my understanding on the data I am working with.

```r
# read in the data from an XLSX file
codebook <- Read("/cloud/project/Final/Data/codebook.xlsx", var_labels=FALSE,
quiet=TRUE)

# display the codebook
head(codebook)
```

```
##   variable_Name
## 1  acousticness
## 2       artists
## 3  danceability
## 4   duration_ms
## 5        energy
## 6      explicit
##
varaible_defintion
## 1
The relative metric of the track being acoustic
## 2
The list of artists credited for production of the track
## 3 Danceability describes how suitable a track is for dancing based on a co
mbination of musical elements including tempo, rhythm stability, beat strengt
h, and overall regularity.
## 4
The length of the track in milliseconds (ms)
## 5
A  perceptual measure of intensity and activity of the track
## 6
The binary value whether the track contains explicit content or not
##                                                    variable_value
## 1                                                             <NA>
## 2                                                             <NA>
## 3 A value of 0.0 is least danceable and 1.0 is most danceable.
## 4                                                             <NA>
## 5                                  Is a measure from 0.0 to 1.0
## 6                0 = No explicit content, 1 = Explicit content
```

Data Prep & Clean Up

Now that I know what the data looks like and have a code book to understand the data, my next step is to clean up the data before I conduct any sort of analysis.

First, I will look for null values in the dataset.

```
#find null values
Spotifydata[!complete.cases(Spotifydata), ]

##  [1] acousticness     artists          danceability     duration_ms      e
nergy           explicit
##  [7] id               instrumentalness key              liveness         l
oudness         mode
## [13] name             popularity       release_date     speechiness      t
empo            valence
## [19] year
## <0 rows> (or 0-length row.names)
```

Since there weren't any null values returned, I will now create a new data frame (d) by subsetting the data with logical criterion. This criterion will extract data from the year 2020 and 2021. I will also remove any variables I found unnecessary for this analysis.

```
#Create a data frame that only grabs data from 2020 and 2021 using lessR.
d <- Spotifydata[.(year > "2019"),]

#update data frame d to only include the specified columns
d <- d[,.(-7)]
```

Along with removing variables, I will now look for duplicated data and remove this from the data set.

```
#Find unique values in the data set and remove duplicate using the dplyr dist
inct() command.
d <- d %>% distinct()
```

This new data set now has 4,730 observations with 18 variables. By cleaning up and filtering out the data, I have reduced the observation by 169,659 and one variable.

Factor

Next I will factor by adding value labels to integer variables "mode" and "explicit". Within the code book, these two fields are defined as a 0 or 1 and what these integer values mean. To have more meaningful data, I will update the values to what the code book defines 0 and 1 to represent.

```
# update the mode variable to display a categorical value instead of an int
d$mode <- factor(d$mode, levels=0:1, labels=c("Minor", "Major"))

# update the explicit variable to display a categorical value instead of an i
nt
d$explicit <- factor(d$explicit, levels=0:1, labels=c("No explicit content",
"Explicit content"))
```

I will also update the key column to display the corresponding tonal counterpart with it's pitch class. These values were pulled from the listed wiki page on Spotify's API.

```
#create a vector to label the pitch class tonal counterparts
pitchclass <- c("C","C♯,D♭","D","D♯,E♭","E","F","F♯,G","G","G♯,A♭","A","A♯,B♭"
,"B")

# update the key column to display the corresponding tonal counterpart with i
t's pitch class
d$key <- factor(d$key, levels=0:11, labels=pitchclass)
```

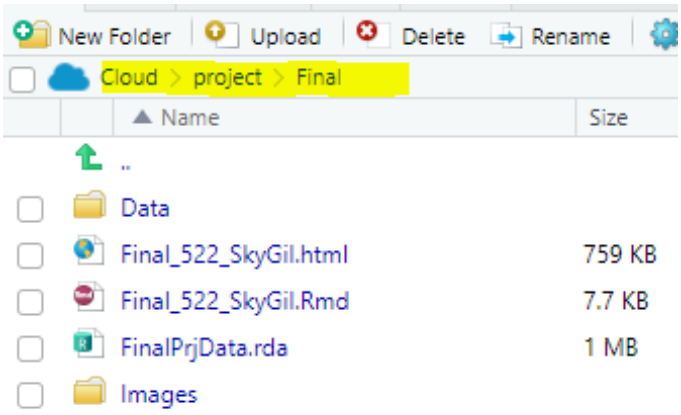
Export data

Now that I am done manipulating the original data, I will export this data into an R binary file labeled, 'FinalPrjData'. When working with large data files, it is best to write these files as an R binary format. This will be handy for when I want to share this data file with someone else using R or Python.

```
# Writing the manipulated data into an R binary file
Write("FinalPrjData", format="R")

## The d data frame contents written at the current working directory
##         FinalPrjData.rda  in:   /cloud/project/Final
```

As you can see, the d data frame contents were written into the current working directory /cloud/project/Final.



Data Analysis

Data Frame Manipulation

Now I am going into the analysis phase of the project. First, I will split up the d data frame into two data frames (d1 and d2), then I will create a new column to act as a unique index (primary key) and finally I will use the sqldf package to merge the two data frames into a new data frame named "d3".

Create two new data frames

Below, I am splitting up the d data frame into two separate data frames called d1 and d2. Both of these newly created data frames are only grabbing the first 20 rows of data. In d1, there will only be 10 variables ranging from artists to mode. In d2, there will be only two variables called artists and name.

```
#create a data frame named d1 with columns ID to mode
d1 = d[1:20, .(artists:mode)]

#create a data frame named d2 with columns ID, artists and name
d2 = d[1:20, .(artists, name)]

#remove all rows with at least one missing data value
d1 <- na.omit(d1)
d2 <- na.omit(d2)

#display data frame d1 and d2
d1

##
artists danceability duration_ms energy
## 1                                                                    ['Joni M
itchell']       0.6440      313093 0.2120
## 2                                                                    ['Joni M
itchell']       0.6270      295093 0.1840
## 3                                                                    ['Joni M
itchell']       0.5810      183440 0.3310
## 4                                                                    ['Joni M
itchell']       0.4420      147907 0.3990
## 5                                                                    ['Joni M
itchell']       0.5700       64173 0.1760
## 6                                                                    ['Joni M
itchell']       0.5650      232640 0.1530
## 7                                                                    ['Joni M
itchell']       0.5980      233520 0.2120
## 8                                                                    ['Joni M
itchell']       0.3670      213840 0.3070
## 9                                                                    ['Joni M
itchell']       0.6380       92560 0.1900
## 10                                                                   ['Joni M
itchell']       0.7000       61400 0.1640
## 11 ['Richard Wagner', 'Birgit Nilsson', 'Bayreuth Festival Orchestra', 'Ka
rl Böhm']       0.1080      375280 0.2540
## 12                                                                      ['Ri
chaadEB']       0.2770      243873 0.9570
## 13                                                           ['The Rolling
Stones']       0.1580      444160 0.8550
## 14                                                                       [
'Busted']       0.5700      196387 0.7700
```

```
## 15                                                                    ['Zol
a Bryon']       0.6330      170305 0.6940
## 16                                                                  ['Ema
Spatula']       0.6990       83853 0.5580
## 17                     ['not applicable', 'Riccardo Muti', 'Wiener Philhar
moniker']       0.7110      217360 0.0983
## 18                                                                      ['
Hammock']       0.0692      533707 0.2850
## 19                   ['Johann Strauss II', 'Riccardo Muti', 'Wiener Philhar
moniker']       0.2690      654107 0.0833
## 20                                                          ['Schoolgirl
Byebye']       0.3140       74302 0.0855
##                explicit instrumentalness   key liveness loudness  mode
## 1  No explicit content        0.0000222     B   0.7980  -14.118 Major
## 2  No explicit content        0.0001620 C♯,D♭   0.0986  -15.533 Major
## 3  No explicit content        0.0000150  F♯,G   0.1470  -14.087 Major
## 4  No explicit content        0.0004990  F♯,G   0.9120  -12.661 Major
## 5  No explicit content        0.0000000  F♯,G   0.1470  -22.676 Minor
## 6  No explicit content        0.0000000 A♯,B♭   0.3580  -21.606 Minor
## 7  No explicit content        0.0000232  F♯,G   0.6920  -15.078 Minor
## 8  No explicit content        0.0000000     B   0.7300  -12.420 Minor
## 9  No explicit content        0.0000000     D   0.2730  -22.268 Major
## 10 No explicit content        0.0000000 A♯,B♭   0.2370  -19.645 Minor
## 11 No explicit content        0.0149000     B   0.1390  -14.490 Major
## 12 No explicit content        0.6430000 G♯,A♭   0.2930   -5.656 Major
## 13 No explicit content        0.0219000     D   0.7080   -7.001 Major
## 14 No explicit content        0.0000000     B   0.0831   -8.431 Major
## 15 No explicit content        0.0000000 G♯,A♭   0.1770   -9.187 Minor
## 16 No explicit content        0.8830000     A   0.0932  -14.713 Major
## 17 No explicit content        0.0000000 C♯,D♭   0.6200  -28.235 Major
## 18 No explicit content        0.6730000     D   0.1990  -14.854 Major
## 19 No explicit content        0.9080000     D   0.0673  -18.492 Major
## 20 No explicit content        0.7950000     A   0.1600  -15.775 Major

d2

##
artists
## 1                                                                   ['Joni M
itchell']
## 2                                                                   ['Joni M
itchell']
## 3                                                                   ['Joni M
itchell']
## 4                                                                   ['Joni M
itchell']
## 5                                                                   ['Joni M
itchell']
## 6                                                                   ['Joni M
```

```
itchell']
## 7                                                                    ['Joni M
itchell']
## 8                                                                    ['Joni M
itchell']
## 9                                                                    ['Joni M
itchell']
## 10                                                                   ['Joni M
itchell']
## 11 ['Richard Wagner', 'Birgit Nilsson', 'Bayreuth Festival Orchestra', 'Ka
rl Böhm']
## 12                                                                     ['Ri
chaadEB']
## 13                                                               ['The Rolling
Stones']
## 14                                                                       [
'Busted']
## 15                                                                    ['Zol
a Bryon']
## 16                                                                   ['Ema
Spatula']
## 17                      ['not applicable', 'Riccardo Muti', 'Wiener Philhar
moniker']
## 18                                                                       ['
Hammock']
## 19                   ['Johann Strauss II', 'Riccardo Muti', 'Wiener Philhar
moniker']
## 20                                                                ['Schoolgirl
Byebye']
##
name
## 1                                          The Circle Game - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 2                                           Urge For Going - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 3                                  What's The Story Mr. Blue - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 4                                             Brandy Eyes - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 5                                    Intro To Urge For Going - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 6                                   Intro To The Circle Game - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 7                                             Eastern Rain - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 8                                      Night In The City - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 9                                 Intro To Night In The City - Live at The 2n
d Fret, Philadelphia, PA, 11/1966
## 10                       Intro To What's The Story Mr. Blue - Live at The 2n
```

```
d Fret, Philadelphia, PA, 11/1966
## 11 Tristan und Isolde, WWV 90 / Act 3: Mild und leise wie er lächelt - Liv
e at Bayreuther Festspiele / 1966
## 12
To You, In 2000 Years
## 13                                          2000 Light Years From Home / She’s
A Rainbow / Keyboard Duet - Live
## 14
Year 3000
## 15
Year 2000 Flow
## 16
Year 2020
## 17                                                 Neujahrsgruß / New Year's
Address / Allocution du Nouvel An
## 18
Longest Year - 2020
## 19                                                                 An der schö
nen blauen Donau, Walzer, Op. 314
## 20
Year,2015
```

Manipulate Data Frames

Next,I will create a new variable called “ID” in both d1 and d2. This is a unique index which could also be referred to as the primary key.

```
#create a common ID field named 'ID' in data frame d1 and d2
d1$ID = row.names(d1)
d2$ID = row.names(d2)

#display data frame d1 and d2
d1[1:5,]
```

```
##             artists danceability duration_ms energy           explicit in
strumentalness   key liveness loudness  mode
## 1 ['Joni Mitchell']        0.644      313093  0.212 No explicit content
0.0000222     B   0.7980  -14.118 Major
## 2 ['Joni Mitchell']        0.627      295093  0.184 No explicit content
0.0001620 C♯,D♭   0.0986  -15.533 Major
## 3 ['Joni Mitchell']        0.581      183440  0.331 No explicit content
0.0000150  F♯,G   0.1470  -14.087 Major
## 4 ['Joni Mitchell']        0.442      147907  0.399 No explicit content
0.0004990  F♯,G   0.9120  -12.661 Major
## 5 ['Joni Mitchell']        0.570       64173  0.176 No explicit content
0.0000000  F♯,G   0.1470  -22.676 Minor
##   ID
## 1  1
## 2  2
## 3  3
```

```
## 4  4
## 5  5

d2[1:5,]

##             artists
name ID
## 1 ['Joni Mitchell']          The Circle Game - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  1
## 2 ['Joni Mitchell']           Urge For Going - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  2
## 3 ['Joni Mitchell'] What's The Story Mr. Blue - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  3
## 4 ['Joni Mitchell']              Brandy Eyes - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  4
## 5 ['Joni Mitchell']   Intro To Urge For Going - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  5
```

After creating a unique ID in both data frames, I will remove some rows from both data frames. Doing this, there will be some observations in d1 that d2 does not have and vice versa.

```
#Delete some rows from both data frames to show more discernible differences
when doing the merging
row.names(d1) <- NULL
row.names(d2) <- NULL

#delete some rows from both data frames
d1 = d1[-c(1,3,6,8), ]
d2 = d2[-c(2,4), ]

#display data frame d1 and d2
d1[1:5,]

##             artists danceability duration_ms energy            explicit in
strumentalness   key liveness loudness  mode
## 2 ['Joni Mitchell']        0.627      295093  0.184 No explicit content
0.0001620 C♯,D♭   0.0986  -15.533 Major
## 4 ['Joni Mitchell']        0.442      147907  0.399 No explicit content
0.0004990  F♯,G   0.9120  -12.661 Major
## 5 ['Joni Mitchell']        0.570       64173  0.176 No explicit content
0.0000000  F♯,G   0.1470  -22.676 Minor
## 7 ['Joni Mitchell']        0.598      233520  0.212 No explicit content
0.0000232  F♯,G   0.6920  -15.078 Minor
## 9 ['Joni Mitchell']        0.638       92560  0.190 No explicit content
0.0000000     D   0.2730  -22.268 Major
##   ID
## 2  2
## 4  4
## 5  5
```

```
## 7  7
## 9  9

d2[1:5,]

##             artists
name ID
## 1 ['Joni Mitchell']          The Circle Game - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  1
## 3 ['Joni Mitchell'] What's The Story Mr. Blue - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  3
## 5 ['Joni Mitchell']   Intro To Urge For Going - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  5
## 6 ['Joni Mitchell']  Intro To The Circle Game - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  6
## 7 ['Joni Mitchell']             Eastern Rain - Live at The 2nd Fret, Phil
adelphia, PA, 11/1966  7
```

Now some observations in d1 are not in d2. For example,observations with ID 1 and 3 are missing from the d1 data frame and observations with ID 2 and 4 are missing from the d2 data frame.

Next, I will use the SQLDF package, to do an inner join of these two data frames and merge them into a new data frame called d3.

```
#create the new data frame where d1 and d2 inner join
d3 <- sqldf("SELECT distinct *
             FROM d1
             JOIN d2 on d1.ID = d2.ID"
            )

#display the results
d3

##
artists danceability duration_ms energy
## 1                                                                    ['Joni M
itchell']       0.5700       64173 0.1760
## 2                                                                    ['Joni M
itchell']       0.5980      233520 0.2120
## 3                                                                    ['Joni M
itchell']       0.6380       92560 0.1900
## 4                                                                    ['Joni M
itchell']       0.7000       61400 0.1640
## 5  ['Richard Wagner', 'Birgit Nilsson', 'Bayreuth Festival Orchestra', 'Ka
rl Böhm']       0.1080      375280 0.2540
## 6                                                                       ['Ri
chaadEB']       0.2770      243873 0.9570
## 7                                                             ['The Rolling
Stones']       0.1580      444160 0.8550
## 8                                                                        [
```

```
'Busted']       0.5700      196387 0.7700
## 9                                                                    ['Zol
a Bryon']       0.6330      170305 0.6940
## 10                                                                  ['Ema
Spatula']       0.6990       83853 0.5580
## 11                     ['not applicable', 'Riccardo Muti', 'Wiener Philhar
moniker']       0.7110      217360 0.0983
## 12                                                                     ['
Hammock']       0.0692      533707 0.2850
## 13                   ['Johann Strauss II', 'Riccardo Muti', 'Wiener Philhar
moniker']       0.2690      654107 0.0833
## 14                                                               ['Schoolgirl
Byebye']       0.3140       74302 0.0855
##                explicit instrumentalness   key liveness loudness  mode ID
## 1  No explicit content        0.0000000  F♯,G   0.1470  -22.676 Minor  5
## 2  No explicit content        0.0000232  F♯,G   0.6920  -15.078 Minor  7
## 3  No explicit content        0.0000000     D   0.2730  -22.268 Major  9
## 4  No explicit content        0.0000000 A♯,B♭   0.2370  -19.645 Minor 10
## 5  No explicit content        0.0149000     B   0.1390  -14.490 Major 11
## 6  No explicit content        0.6430000 G♯,A♭   0.2930   -5.656 Major 12
## 7  No explicit content        0.0219000     D   0.7080   -7.001 Major 13
## 8  No explicit content        0.0000000     B   0.0831   -8.431 Major 14
## 9  No explicit content        0.0000000 G♯,A♭   0.1770   -9.187 Minor 15
## 10 No explicit content        0.8830000     A   0.0932  -14.713 Major 16
## 11 No explicit content        0.0000000 C♯,D♭   0.6200  -28.235 Major 17
## 12 No explicit content        0.6730000     D   0.1990  -14.854 Major 18
## 13 No explicit content        0.9080000     D   0.0673  -18.492 Major 19
## 14 No explicit content        0.7950000     A   0.1600  -15.775 Major 20
##
artists
## 1                                                                  ['Joni M
itchell']
## 2                                                                  ['Joni M
itchell']
## 3                                                                  ['Joni M
itchell']
## 4                                                                  ['Joni M
itchell']
## 5  ['Richard Wagner', 'Birgit Nilsson', 'Bayreuth Festival Orchestra', 'Ka
rl Böhm']
## 6                                                                    ['Ri
chaadEB']
## 7                                                               ['The Rolling
Stones']
## 8                                                                      [
'Busted']
## 9                                                                    ['Zol
a Bryon']
## 10                                                                 ['Ema
```

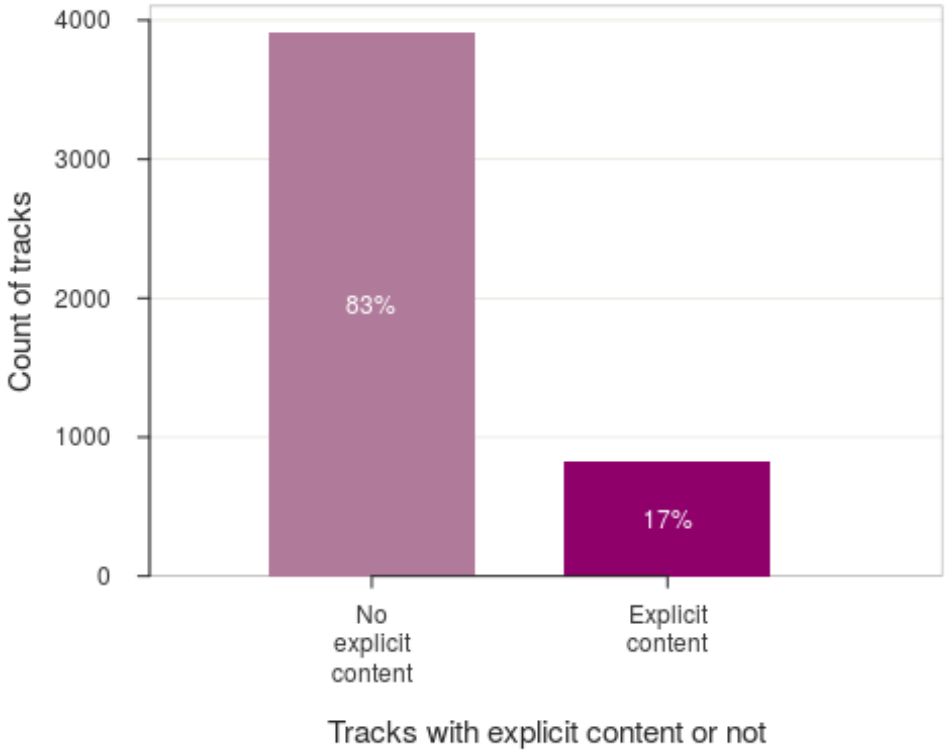
```
Spatula']
## 11                    ['not applicable', 'Riccardo Muti', 'Wiener Philhar
moniker']
## 12                                                                    ['
Hammock']
## 13                 ['Johann Strauss II', 'Riccardo Muti', 'Wiener Philhar
moniker']
## 14                                                          ['Schoolgirl
Byebye']
##
name ID
## 1                                  Intro To Urge For Going - Live at The 2n
d Fret, Philadelphia, PA, 11/1966  5
## 2                                           Eastern Rain - Live at The 2n
d Fret, Philadelphia, PA, 11/1966  7
## 3                               Intro To Night In The City - Live at The 2n
d Fret, Philadelphia, PA, 11/1966  9
## 4                      Intro To What's The Story Mr. Blue - Live at The 2n
d Fret, Philadelphia, PA, 11/1966 10
## 5  Tristan und Isolde, WWV 90 / Act 3: Mild und leise wie er lächelt - Liv
e at Bayreuther Festspiele / 1966 11
## 6
To You, In 2000 Years 12
## 7                                      2000 Light Years From Home / She’s
A Rainbow / Keyboard Duet - Live 13
## 8
Year 3000 14
## 9
Year 2000 Flow 15
## 10
Year 2020 16
## 11                                            Neujahrsgruß / New Year's
Address / Allocution du Nouvel An 17
## 12
Longest Year - 2020 18
## 13                                                          An der schö
nen blauen Donau, Walzer, Op. 314 19
## 14
Year,2015 20
```

The mentioned observations with ID’s 1-4 are now excluded from the d3 data frame. This is because they do not have matching records to link with one another. Data frame, d3, only shows observations that are in both d1 and d2. As you can see from the data above, there are now only 14 observations these two data frames have in common.

Exploratory Analysis

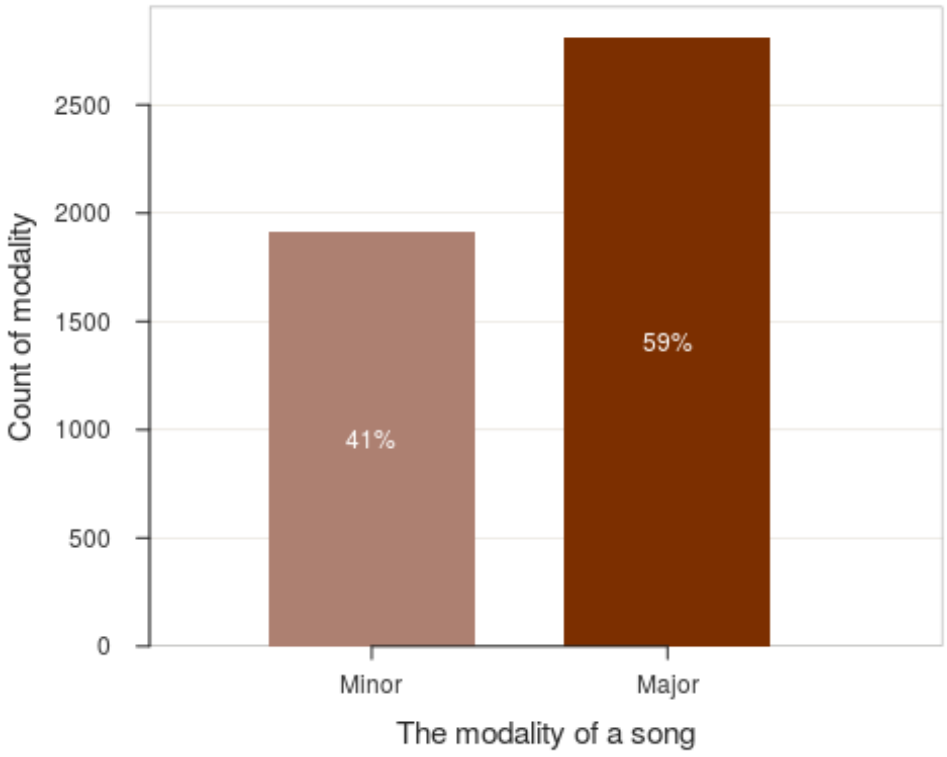
In this next section, I am going to play around with the data and see what can be uncovered through the variety of visualizations.

The below barchart is pulling data from the d data frame. What this bar chart tells me is, under 20% of songs from 2020 and 2021 had explicit content in an individual song. Of course, this only takes into an account what users listen to through Spotify.



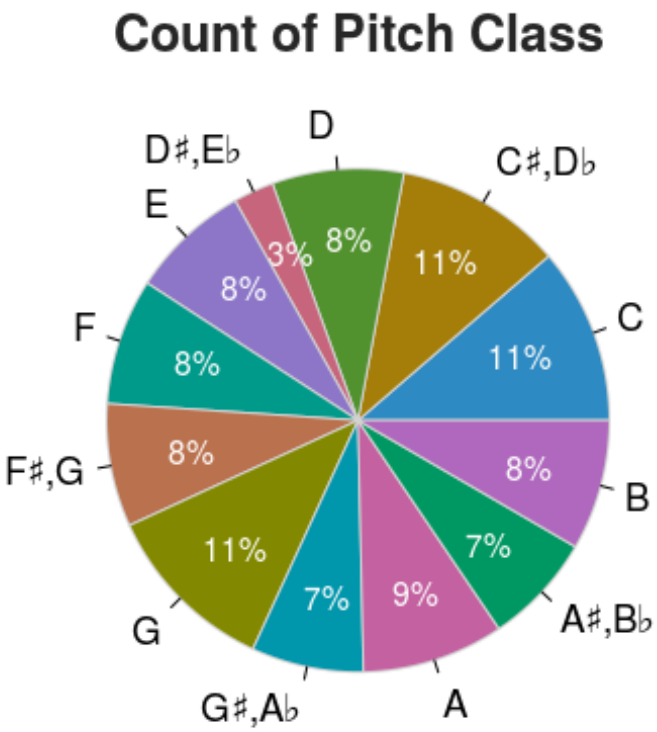
In the below barchart, I will see if the majority of music was in the major or minor modality for 2020-2021. This is displaying a count of the variable "mode", from the d data frame. Since I factored out the integer values to a more descriptive meaning, this barchart is much easier to understand. I also included descriptive labels on the y and x axis.

```
BarChart(mode,fill = "rusts", ylab="Count of modality", xlab="The modality of a song",quiet=TRUE)
```



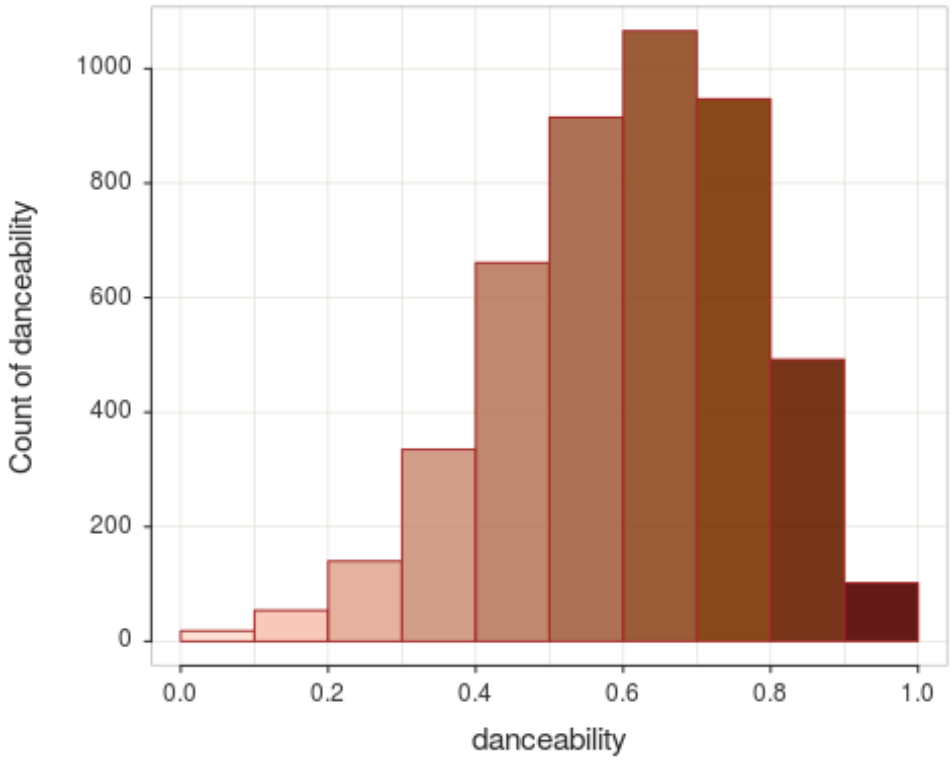
In this next graphic, I am displaying a pie chart with the key variable using the “hues” color scheme. What this pie chart is telling me is, the top tones used in this data set are in “G”, “C” and “C♯,D♭”. Which is interesting because the “G” and “C” tones can be said to have a similar quality of octave equivalence. This could be used to do a more in-depth analysis on quality of pitch and a correlation between the other variables, such as valence, instrumentalness and etc.

```
PieChart(key, hole=0, fill = "hues",main="Count of Pitch Class",quiet=TRUE)
```



Next, I will use a histogram to display how the tracks are distributed based off of their danceability rating. This is one of the better variables to use in this data set since it is a continuous numerical value ranging from 0 to 1.The bin width is auto set to 0.1 with 10 bins displayed, 11 outliers and a peak of a danceability score at 0.7. What this tells me is, a majority of the songs in this data set are more suitable for dancing.

```
#Create a histogram with the variable danceability
hs(danceability,fill="rusts", color="brown", trans=.1)
```




```
##
##
## --- danceability ---
##
##        n   miss         mean           sd          min          mdn
max
##      4730      0      0.60989      0.16877      0.00000      0.62700
0.98700
##
##
## (Box plot) Outliers: 29
##
## Small      Large
## -----      -----
##  0.0
##  0.0
##  0.0
##  0.0
##  0.0
##  0.1
##  0.1
##  0.1
##  0.1
##  0.1
##  0.1
##  0.1
##  0.1
```

```
##  0.1
##  0.1
##  0.1
##  0.1
##  0.1
##
## + 11 more outliers
##
##
## Bin Width: 0.1
## Number of Bins: 10
##
##        Bin  Midpnt  Count    Prop  Cumul.c  Cumul.p
## ------------------------------------------------------
##  0.0 > 0.1    0.05     18    0.00       18     0.00
##  0.1 > 0.2    0.15     54    0.01       72     0.02
##  0.2 > 0.3    0.25    140    0.03      212     0.04
##  0.3 > 0.4    0.35    335    0.07      547     0.12
##  0.4 > 0.5    0.45    661    0.14     1208     0.26
##  0.5 > 0.6    0.55    915    0.19     2123     0.45
##  0.6 > 0.7    0.65   1066    0.23     3189     0.67
##  0.7 > 0.8    0.75    947    0.20     4136     0.87
##  0.8 > 0.9    0.85    492    0.10     4628     0.98
##  0.9 > 1.0    0.95    102    0.02     4730     1.00
```

Now I am going to look into aggregating some data to see if I can find out any information regarding the popularity of a song. First, I want to subset the data to pull out songs with a popularity score of 50 or higher. Then I will arrange the data to display most popular to least. After doing this, I will put this is a data frame and only keep the top 10 observations.

```
# Aggregate the data by first creating a subset of data that is filtered by i
ndices and has a popularity rating of 50 or higher
agg <- d[, .(2:3, 6,11:13)]
agg <- filter(agg, popularity > 49)

#sort the data from most popular to least
agg <- Sort(agg, by=popularity, direction="-", quiet = TRUE)

#Grab the top 10 popular songs
agg <- agg[1:10,]

#display the results
agg
```

```
##                          artists danceability            explicit  mode
name popularity
## 101           ['Olivia Rodrigo']        0.585    Explicit content Major
drivers license        100
## 1      ['24kGoldn', 'iann dior']        0.700    Explicit content Minor Mo
od (feat. iann dior)         96
```

```
## 3              ['Ariana Grande']        0.737    Explicit content Major
positions         96
## 13  ['Bad Bunny', 'Jhay Cortez']        0.731    Explicit content Minor
DÁKITI         95
## 58                  ['KAROL G']        0.863    Explicit content Minor
BICHOTA         95
## 4              ['Ariana Grande']        0.830    Explicit content Major
34+35         94
## 5                         ['CJ']        0.711    Explicit content Minor
Whoopty         94
## 6              ['The Kid LAROI']        0.662    Explicit content Major
WITHOUT YOU         94
## 8              ['Billie Eilish']        0.889 No explicit content Minor
Therefore I Am         94
## 34      ['Bad Bunny', 'ROSALÍA']        0.856 No explicit content Major
LA NOCHE DE ANOCHE         94
```

Conclusion

In conclusion, I have discovered, the song “drivers license” was the most popular song listened to on Spotify in 2020-2021 so far. Ariana Grande was the most popular artist, the top 10 songs averaged closer to 1.0 showing the top listened songs all were songs listeners could dance to, listener’s prefer explicit content and the mode wasn’t a huge factor in songs the listener chose to listen to. I found this data set very interesting and there is a variety of different ways the exploratory analysis could go with this data.