initial_analysis

Jonathan Kogan

2022-07-15

```
library(dplyr)
```

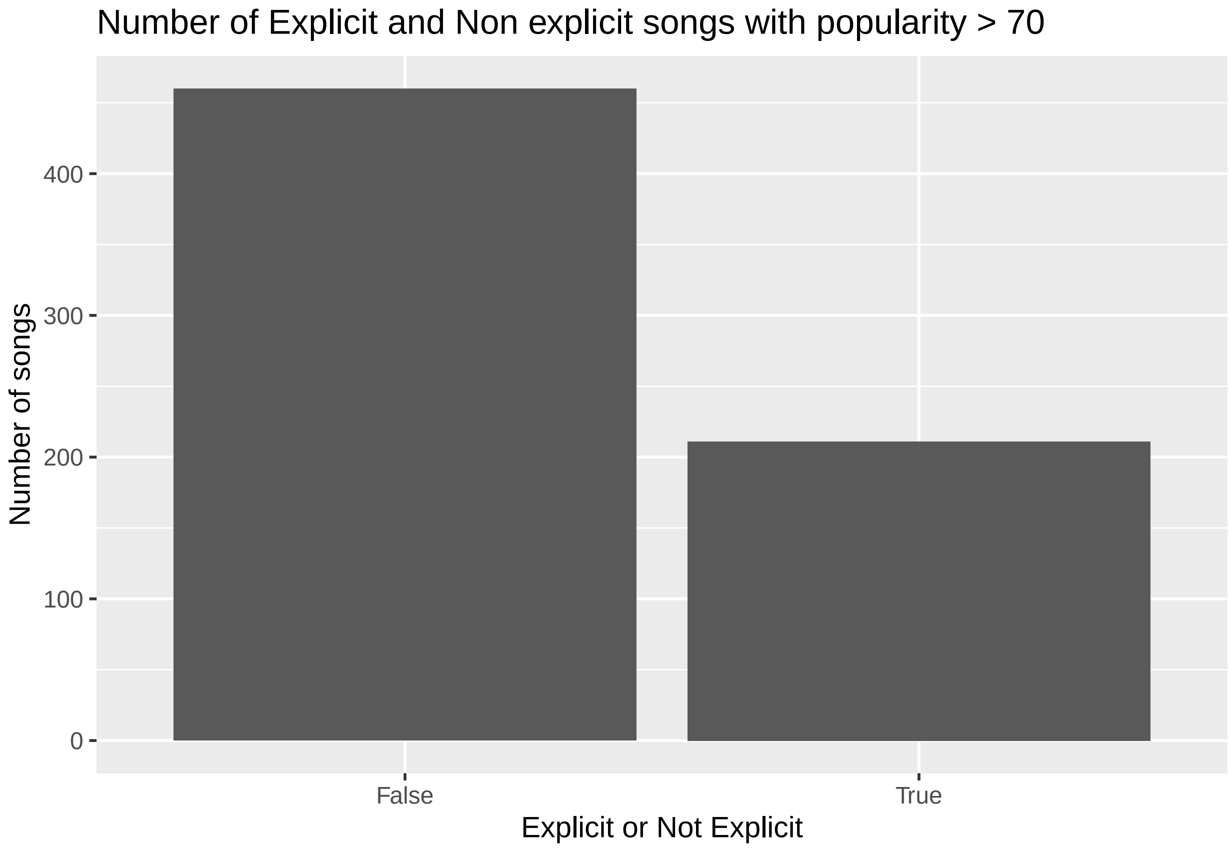
```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(stringr)
songs <- read.csv("data/songs_normalize.csv")
```

```
over70 <- filter(songs, popularity > 70)
#over70

Drake <- filter(songs, artist == "Drake")
head(arrange(Drake, desc(popularity)))
```

```
##   artist                                     song duration_ms explicit year
## 1  Drake                                One Dance      173986    False 2016
## 2  Drake                               God’s Plan      198973     True 2018
## 3  Drake                            Hotline Bling      267066    False 2016
## 4  Drake                                  Nonstop      238614     True 2018
## 5  Drake                            Nice For What      210746     True 2018
## 6  Drake Money In The Grave (Drake ft. Rick Ross)      205426     True 2019
##   popularity danceability energy key loudness mode speechiness acousticness
## 1         84        0.792  0.625   1   -5.609    1      0.0536      0.00776
## 2         81        0.754  0.449   7   -9.211    1      0.1090      0.03320
## 3         77        0.891  0.628   2   -7.863    1      0.0551      0.00258
## 4         77        0.912  0.412   7   -8.074    1      0.1230      0.01650
## 5         77        0.585  0.909   8   -6.474    1      0.0707      0.08910
## 6         76        0.831  0.502  10   -4.045    0      0.0460      0.10100
##   instrumentalness liveness valence   tempo             genre
## 1         1.80e-03   0.3290   0.370 103.967 hip hop, pop, R&B
## 2         8.29e-05   0.5520   0.357  77.169 hip hop, pop, R&B
## 3         1.90e-04   0.0504   0.552 134.966 hip hop, pop, R&B
## 4         1.26e-02   0.1040   0.423 154.983 hip hop, pop, R&B
## 5         9.70e-05   0.1190   0.758  93.372 hip hop, pop, R&B
## 6         0.00e+00   0.1220   0.101 100.541 hip hop, pop, R&B
```

```
ggplot(data = over70, aes(x =  explicit)) + geom_bar() + labs(x = "Explicit or Not Explicit", y = "Numbe
```



```
notExplicit <- filter(songs, explicit== "False")
#notExplicit

top20 <- slice_max(songs, order_by = popularity, n = 20)
#top20

selected <- select(songs, popularity, song, danceability, energy, explicit, acousticness, liveness)
#selected

biggerSongs <- mutate(songs,
       duration_s = duration_ms/1000,
       duration_minutes = duration_s/60)
head(biggerSongs)
```

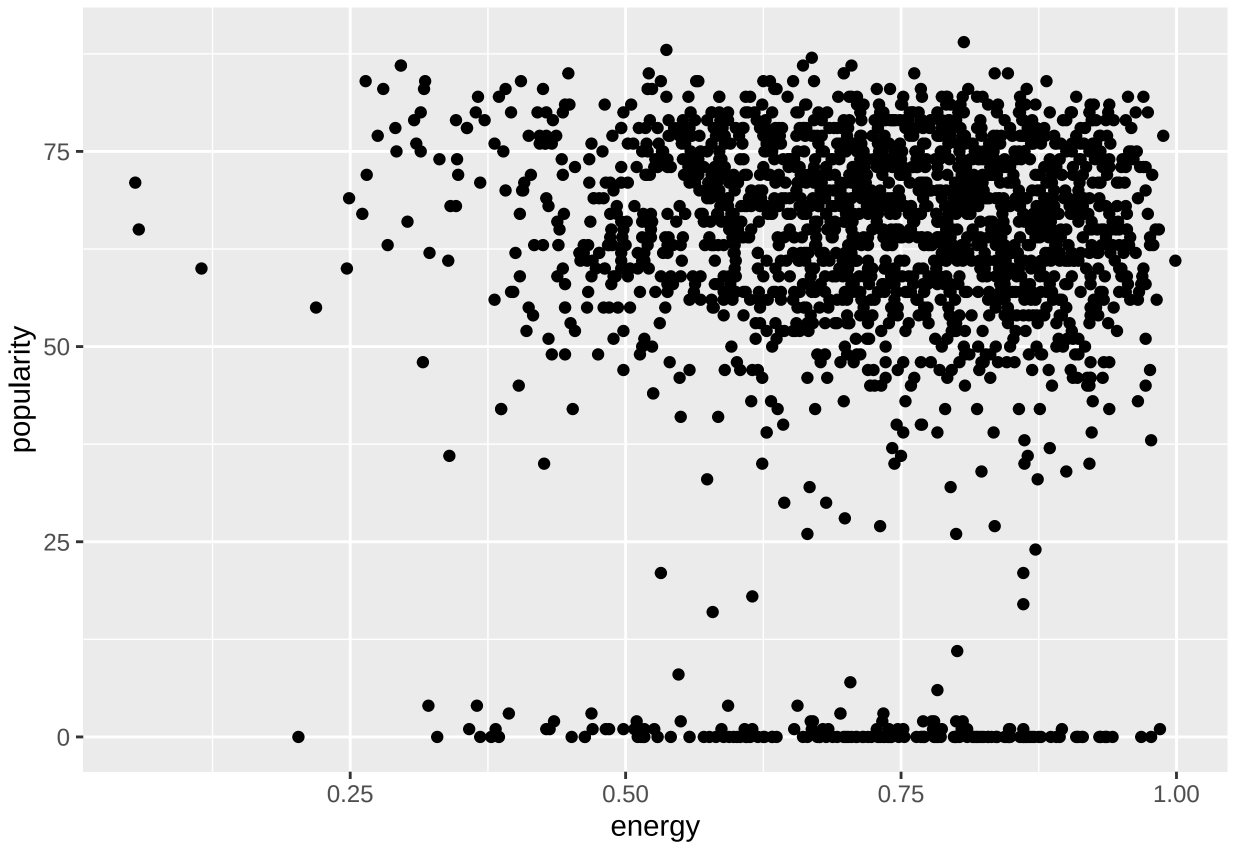
```
##            artist                   song duration_ms explicit year popularity
## 1 Britney Spears Oops!...I Did It Again      211160    False 2000         77
## 2      blink-182    All The Small Things      167066    False 1999         79
## 3     Faith Hill                Breathe      250546    False 1999         66
## 4       Bon Jovi           It’s My Life      224493    False 2000         78
## 5         *NSYNC            Bye Bye Bye      200560    False 2000         65
## 6          Sisqo             Thong Song      253733     True 1999         69
##   danceability energy key loudness mode speechiness acousticness
## 1        0.751  0.834   1   -5.444    0      0.0437       0.3000
## 2        0.434  0.897   0   -4.918    1      0.0488       0.0103
## 3        0.529  0.496   7   -9.007    1      0.0290       0.1730
## 4        0.551  0.913   0   -4.063    0      0.0466       0.0263
## 5        0.614  0.928   8   -4.806    0      0.0516       0.0408
## 6        0.706  0.888   2   -6.959    1      0.0654       0.1190
##   instrumentalness liveness valence   tempo             genre duration_s
## 1         1.77e-05   0.3550   0.894  95.053               pop    211.160
## 2         0.00e+00   0.6120   0.684 148.726         rock, pop    167.066
## 3         0.00e+00   0.2510   0.278 136.859      pop, country    250.546
## 4         1.35e-05   0.3470   0.544 119.992       rock, metal    224.493
## 5         1.04e-03   0.0845   0.879 172.656               pop    200.560
## 6         9.64e-05   0.0700   0.714 121.549 hip hop, pop, R&B    253.733
##   duration_minutes
## 1         3.519333
## 2         2.784433
## 3         4.175767
## 4         3.741550
## 5         3.342667
## 6         4.228883
```

```
shortestSongs <- arrange(biggerSongs, duration_minutes)
longestSongs <- arrange(biggerSongs, desc(duration_minutes))
shortestModified <- select(longestSongs, popularity, song, duration_minutes)
#shortestModified
longestModified <- select(longestSongs, popularity, song, duration_minutes)
#longestModified

summarized <- summarise(songs, mean_popularity = mean(popularity, na.rm = T))
#summarized
numOfPopularity <- summarise(group_by(songs, popularity), count = n()) # n()
#numOfPopularity
```

```
## Practice subsetting data
# use a combination of filter, select, mutate, arrange, summarise, group_by, sample, and/or slice
# create a visulaization using your new subset of data
mostEnergetic <- filter(songs, energy > 0.50)
#mostEnergetic

arrangedEnergetic <- arrange(mostEnergetic, desc(popularity))
head(select(arrangedEnergetic, artist, song, popularity, energy))
```

```
##              artist                song popularity energy
## 1 The Neighbourhood     Sweater Weather         89  0.807
## 2         Tom Odell        Another Love         88  0.537
## 3            Eminem          Without Me         87  0.669
## 4            Eminem The Real Slim Shady         86  0.661
## 5            WILLOW      Wait a Minute!         86  0.705
## 6            Eminem    ’Till I Collapse         85  0.847
```

```
ggplot(data = songs, aes(x = energy, y = popularity)) + geom_point()
```



```
head(songs %>%
  group_by(popularity) %>%
  sample_n(1))
```

```
## # A tibble: 6 x 18
## # Groups:   popularity [6]
##   artist   song  duration_ms explicit  year popularity danceability energy   key
##   <chr>    <chr>       <int> <chr>    <int>      <int>        <dbl>  <dbl> <int>
## 1 Astrid S Hurt~      208728 False     2016          0        0.672  0.589     7
## 2 Lil Way~ 6 Fo~      248586 True      2011          1        0.364  0.752     2
## 3 Adele    Set ~      242973 False     2011          2        0.603  0.67      2
## 4 The Whi~ Seve~      231920 False     2003          3        0.741  0.469     4
## 5 Baby Ba~ Baby~      219920 True      2005          4        0.899  0.365     9
## 6 Avicii   Wake~      247426 False     2013          6        0.532  0.783     2
## # ... with 9 more variables: loudness <dbl>, mode <int>, speechiness <dbl>,
## #   acousticness <dbl>, instrumentalness <dbl>, liveness <dbl>, valence <dbl>,
## #   tempo <dbl>, genre <chr>
```

```
head(songs %>%
  group_by(year) %>%
  sample_n(1))
```

```
## # A tibble: 6 x 18
## # Groups:   year [6]
##   artist   song  duration_ms explicit  year popularity danceability energy   key
##   <chr>    <chr>       <int> <chr>    <int>      <int>        <dbl>  <dbl> <int>
## 1 Missy E~ Hot ~      215466 True      1998         49        0.727  0.445     1
## 2 Dido     Than~      218360 False     1999         73        0.725  0.583     1
## 3 Eminem   Stan       404106 True      2000         83        0.78   0.768     6
## 4 S Club 7 Don’~      233626 False     2001         63        0.822  0.672     7
## 5 JAY-Z    Excu~      281240 True      2002         56        0.714  0.862     6
## 6 Three D~ I Ha~      231480 False     2003         72        0.498  0.83      6
## # ... with 9 more variables: loudness <dbl>, mode <int>, speechiness <dbl>,
## #   acousticness <dbl>, instrumentalness <dbl>, liveness <dbl>, valence <dbl>,
## #   tempo <dbl>, genre <chr>
```

```r
mu <- mean(songs$loudness)
sig <- sd(songs$loudness)
iqr <- IQR(songs$loudness)
q1 <- as.numeric(quantile(songs$loudness, 0.25))
q3 <- as.numeric(quantile(songs$loudness, 0.75))
mu - 3*sig #min1
```

```
## [1] -11.31288
```

```r
mu + 3*sig #max1
```

```
## [1] 0.2880115
```

```r
q1 - iqr *1.5
```

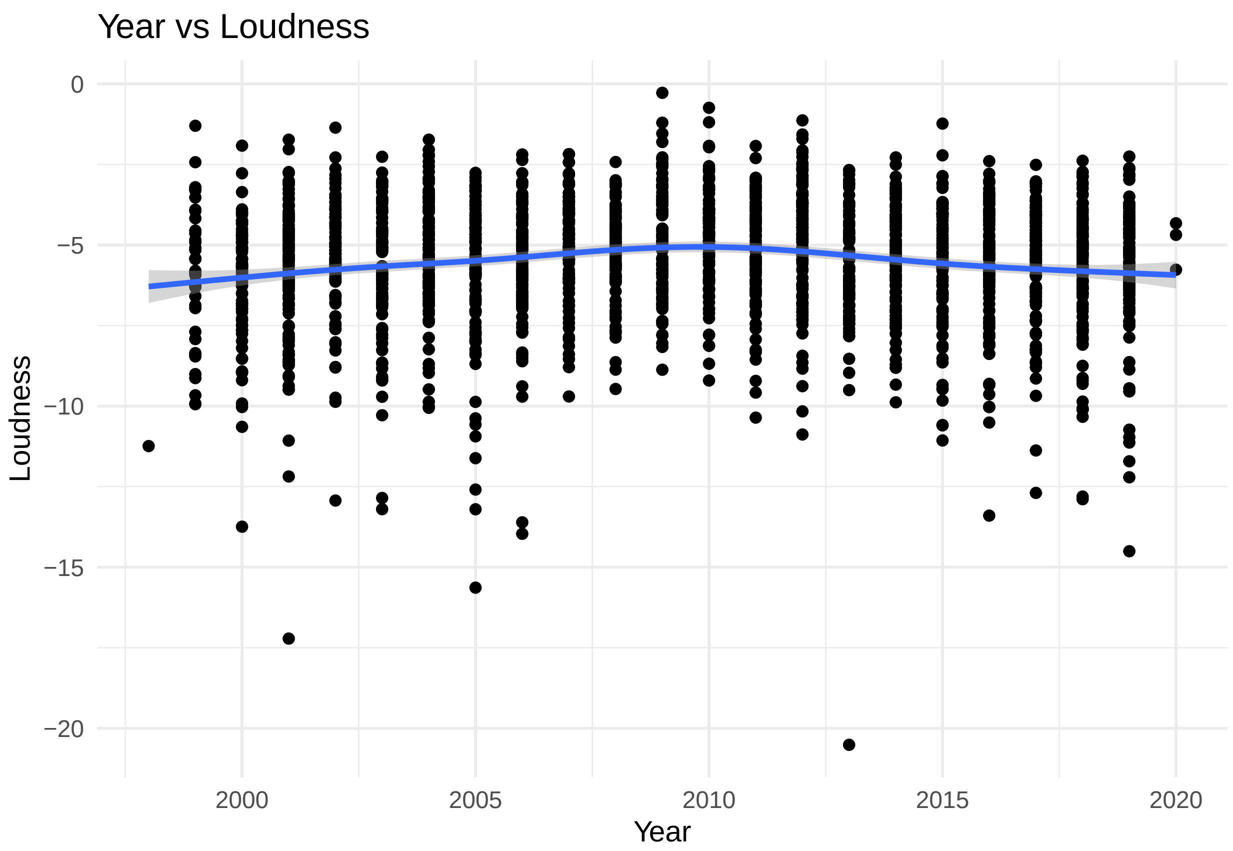
```
## [1] -9.974
```

```r
q3 + iqr *1.5
```

```
## [1] -0.684
```

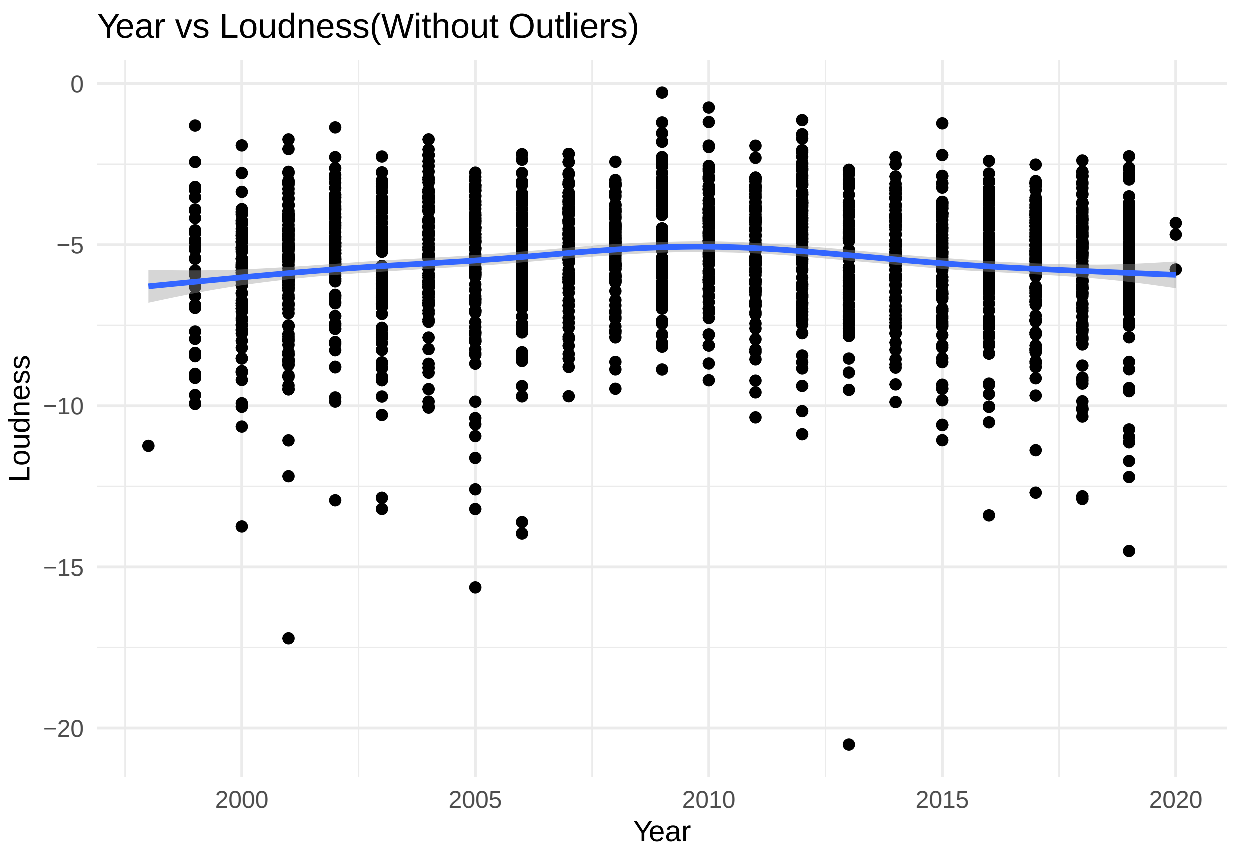
```r
withoutOutliers <- filter(songs, loudness > 9.974, loudness < -0.684)
ggplot(data = songs, aes(x = year, y = loudness)) + geom_point()  + theme_minimal() + labs(x = "Year", 
```

```
## ‘geom_smooth()‘ using method = ’gam’ and formula ’y ~ s(x, bs = "cs")’
```



```
ggplot(data = songs, aes(x = year, y = loudness)) + geom_point()  + theme_minimal() + labs(x = "Year", y
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



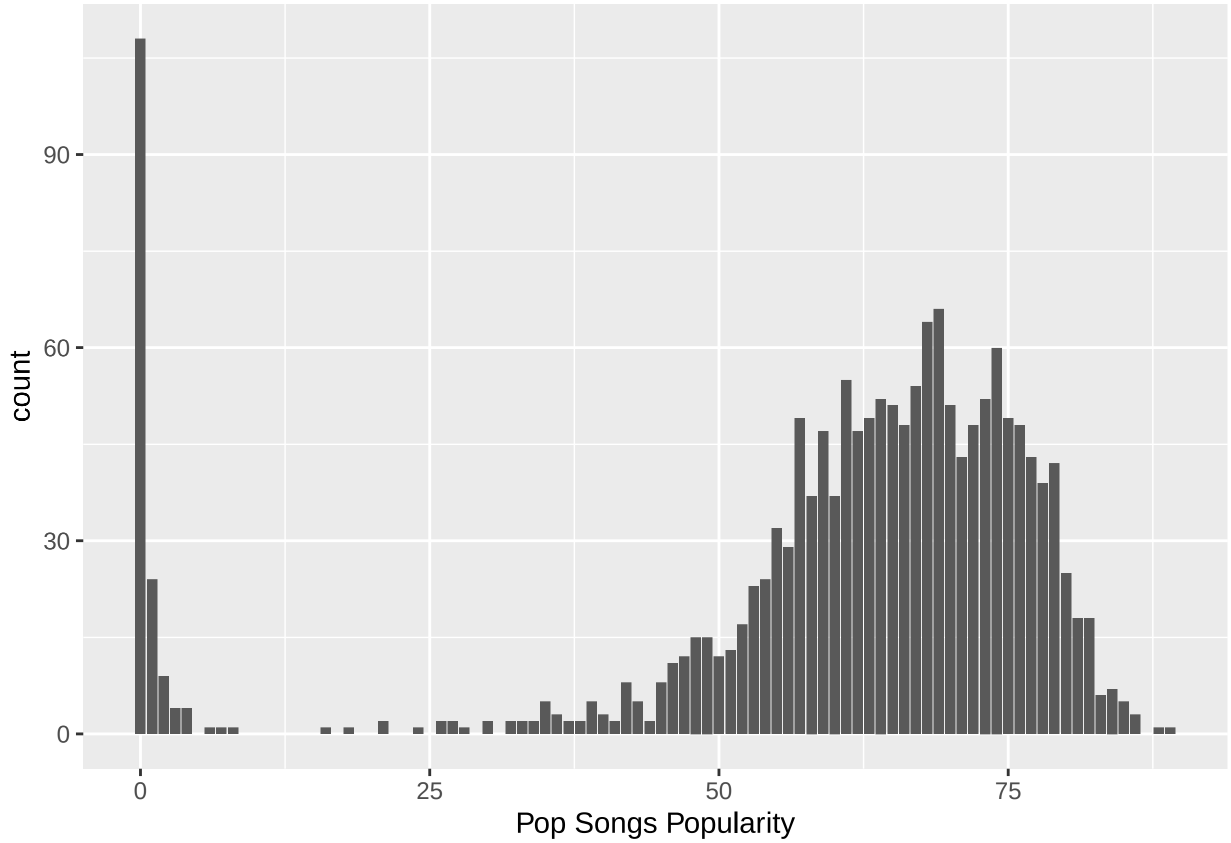
```
unique(songs$genre)
```

```
##  [1] "pop"
##  [2] "rock, pop"
##  [3] "pop, country"
##  [4] "rock, metal"
##  [5] "hip hop, pop, R&B"
##  [6] "hip hop"
##  [7] "pop, rock"
##  [8] "pop, R&B"
##  [9] "Dance/Electronic"
## [10] "pop, Dance/Electronic"
## [11] "rock, Folk/Acoustic, easy listening"
## [12] "metal"
## [13] "hip hop, pop"
## [14] "R&B"
## [15] "pop, latin"
## [16] "Folk/Acoustic, rock"
## [17] "pop, easy listening, Dance/Electronic"
## [18] "rock"
## [19] "rock, blues, latin"
## [20] "pop, rock, metal"
## [21] "rock, pop, metal"
## [22] "hip hop, R&B"
## [23] "pop, Folk/Acoustic"
## [24] "set()"
## [25] "hip hop, pop, latin"
## [26] "hip hop, Dance/Electronic"
## [27] "hip hop, pop, rock"
## [28] "World/Traditional, Folk/Acoustic"
## [29] "Folk/Acoustic, pop"
## [30] "rock, easy listening"
## [31] "World/Traditional, hip hop"
## [32] "hip hop, pop, R&B, latin"
## [33] "rock, blues"
## [34] "rock, R&B, Folk/Acoustic, pop"
## [35] "latin"
## [36] "pop, R&B, Dance/Electronic"
## [37] "World/Traditional, rock"
## [38] "pop, rock, Dance/Electronic"
## [39] "pop, easy listening, jazz"
## [40] "rock, Dance/Electronic"
## [41] "World/Traditional, pop, Folk/Acoustic"
## [42] "country"
## [43] "hip hop, pop, Dance/Electronic"
## [44] "hip hop, pop, country"
## [45] "World/Traditional, rock, pop"
## [46] "World/Traditional, pop"
## [47] "hip hop, pop, R&B, Dance/Electronic"
## [48] "pop, R&B, easy listening"
## [49] "rock, pop, Dance/Electronic"
## [50] "Folk/Acoustic, rock, pop"
## [51] "rock, pop, metal, Dance/Electronic"
## [52] "pop, rock, Folk/Acoustic"
## [53] "country, latin"
## [54] "rock, classical"
## [55] "rock, Folk/Acoustic, pop"
## [56] "hip hop, rock, pop"
## [57] "easy listening"
## [58] "hip hop, latin, Dance/Electronic"
## [59] "hip hop, country"
```

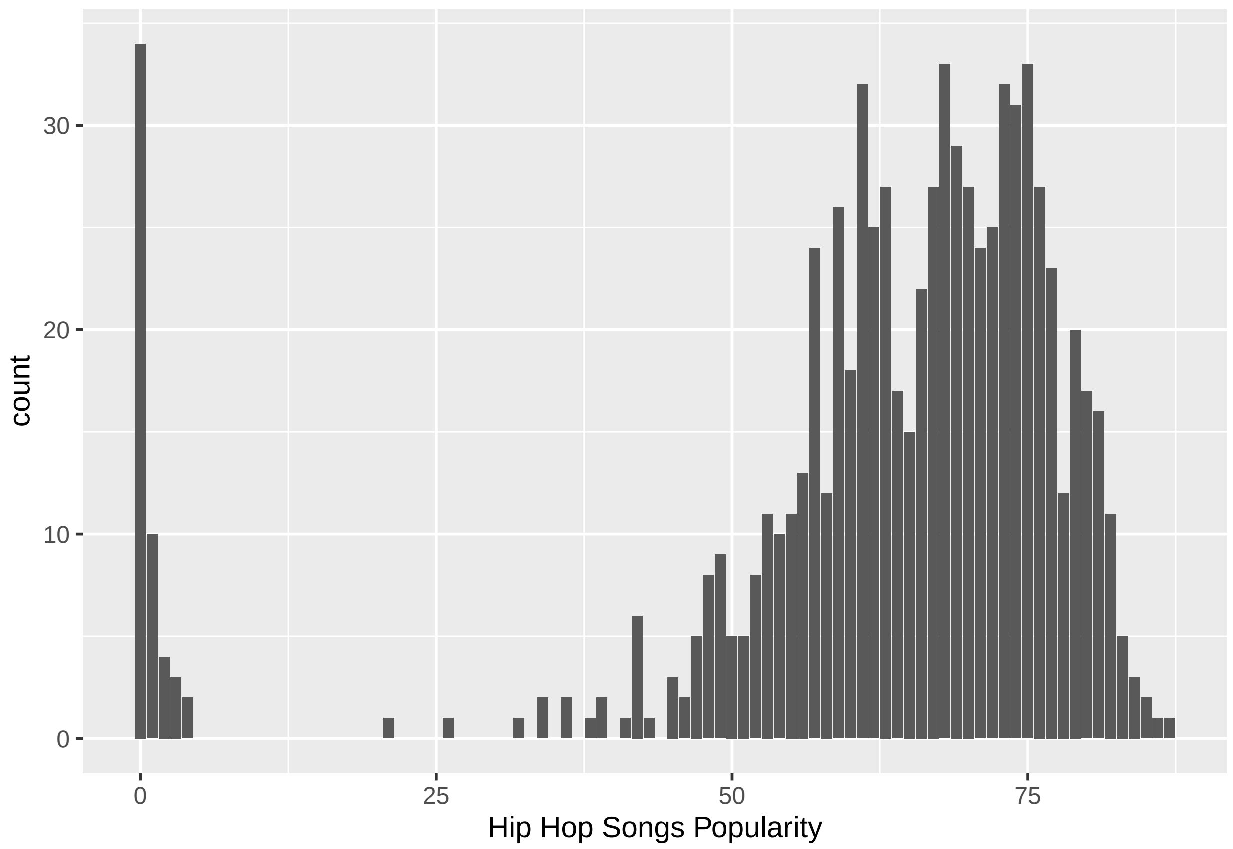
```r
popSongs <- filter(songs, str_detect(genre, "pop"))
#popSongs
hipHopSongs <- filter(songs, str_detect(genre, "hip hop"))
#hipHopSongs
rockSongs <- filter(songs, str_detect(genre, "rock"))
#hipHopSongs
metalSongs <- filter(songs, str_detect(genre, "metal"))
#metalSongs
bluesSongs <- filter(songs, str_detect(genre, "blues"))
#bluesSongs
edmSongs <- filter(songs, str_detect(genre, "Dance/Electronic"))
#edmSongs
countrySongs <- filter(songs, str_detect(genre, "country"))
#countrySongs
folkSongs <- filter(songs, str_detect(genre, "Folk/Acoustic"))
#folkSongs
latinSongs <- filter(songs, str_detect(genre, "latin"))
```

```
#latinSongs
RandBSongs <- filter(songs, str_detect(genre, "R&B"))
#RandBSongs
```

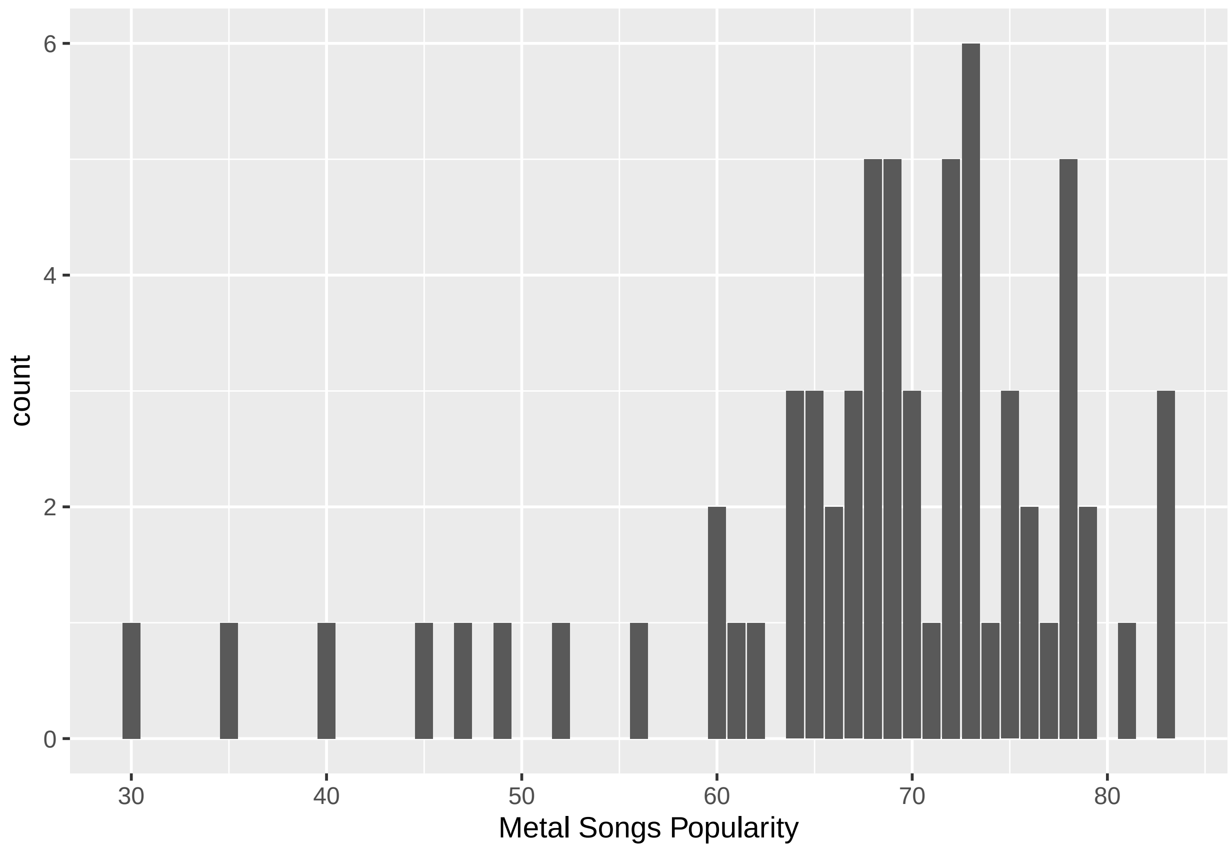
```
genres <- c(popSongs, hipHopSongs,hipHopSongs,metalSongs,bluesSongs,edmSongs ,countrySongs,folkSongs,la
ggplot(data = popSongs, aes(x = popularity)) + geom_bar() + labs(x = "Pop Songs Popularity")
```



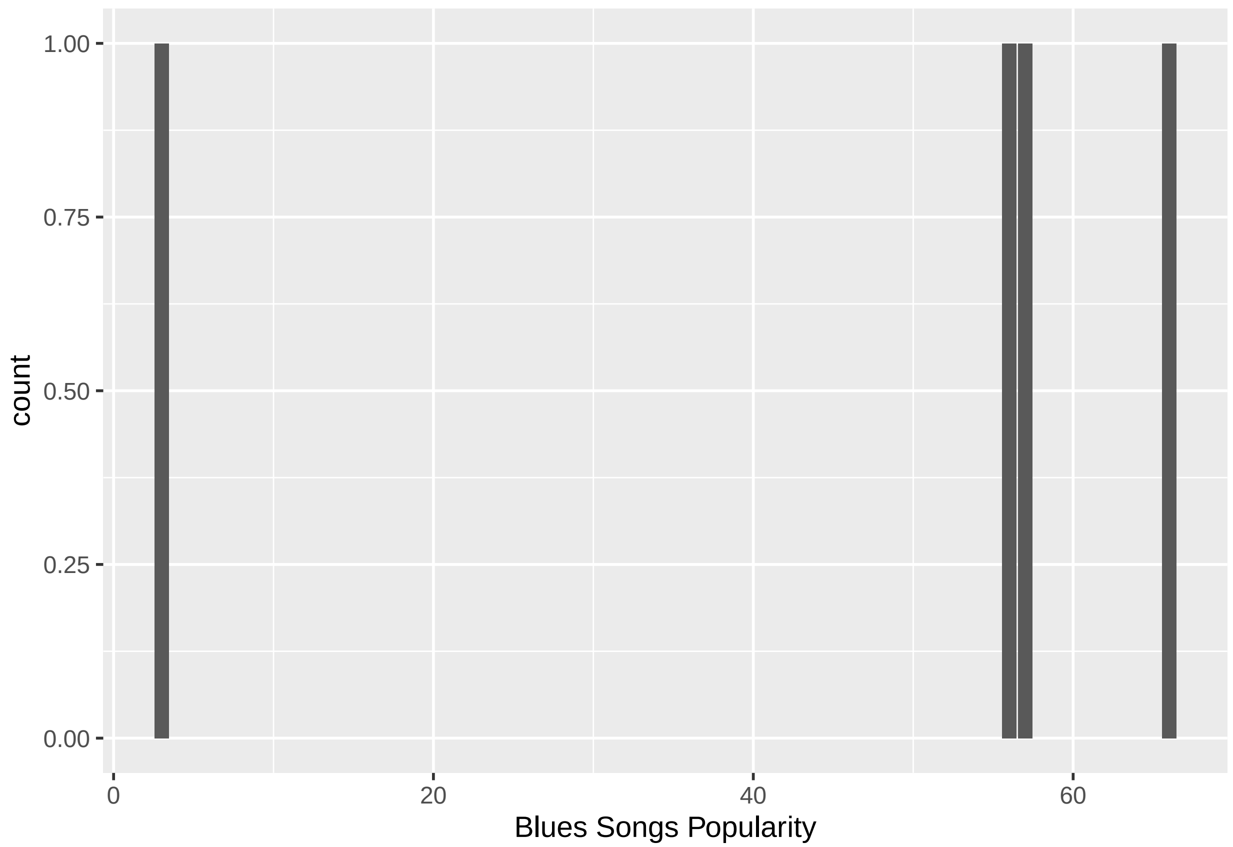
```
ggplot(data = hipHopSongs, aes(x = popularity)) + geom_bar() + labs(x = "Hip Hop Songs Popularity")
```



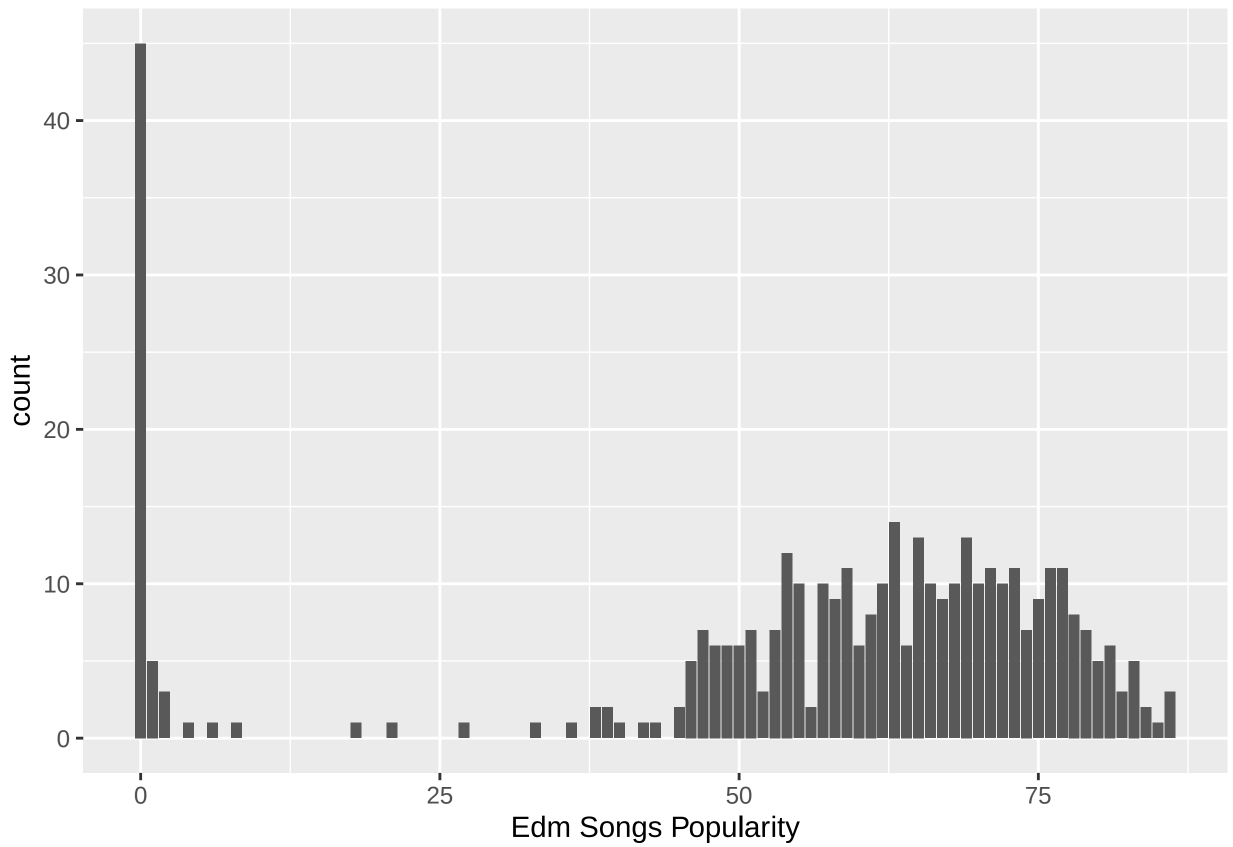
```
ggplot(data = metalSongs, aes(x = popularity)) + geom_bar() + labs(x = "Metal Songs Popularity")
```



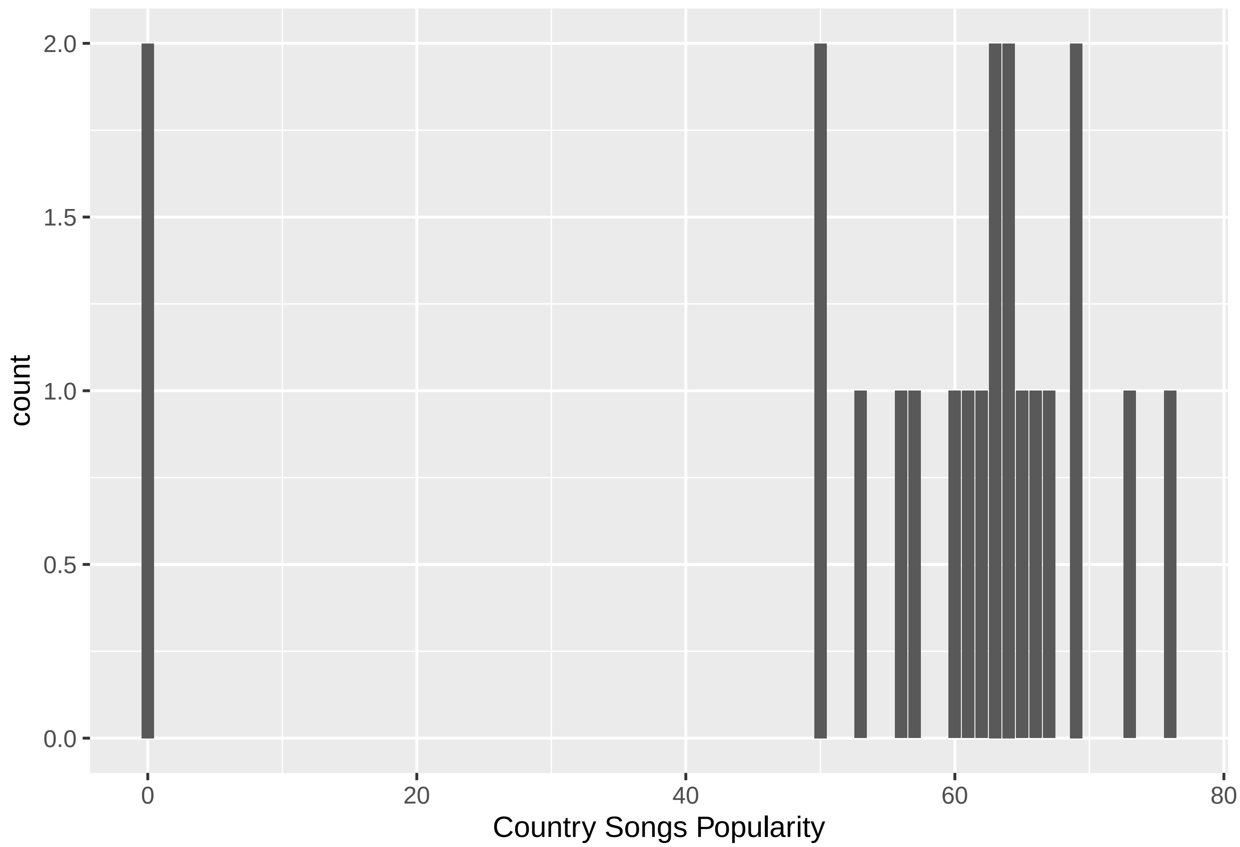
```
ggplot(data = bluesSongs, aes(x = popularity)) + geom_bar() + labs(x = "Blues Songs Popularity")
```



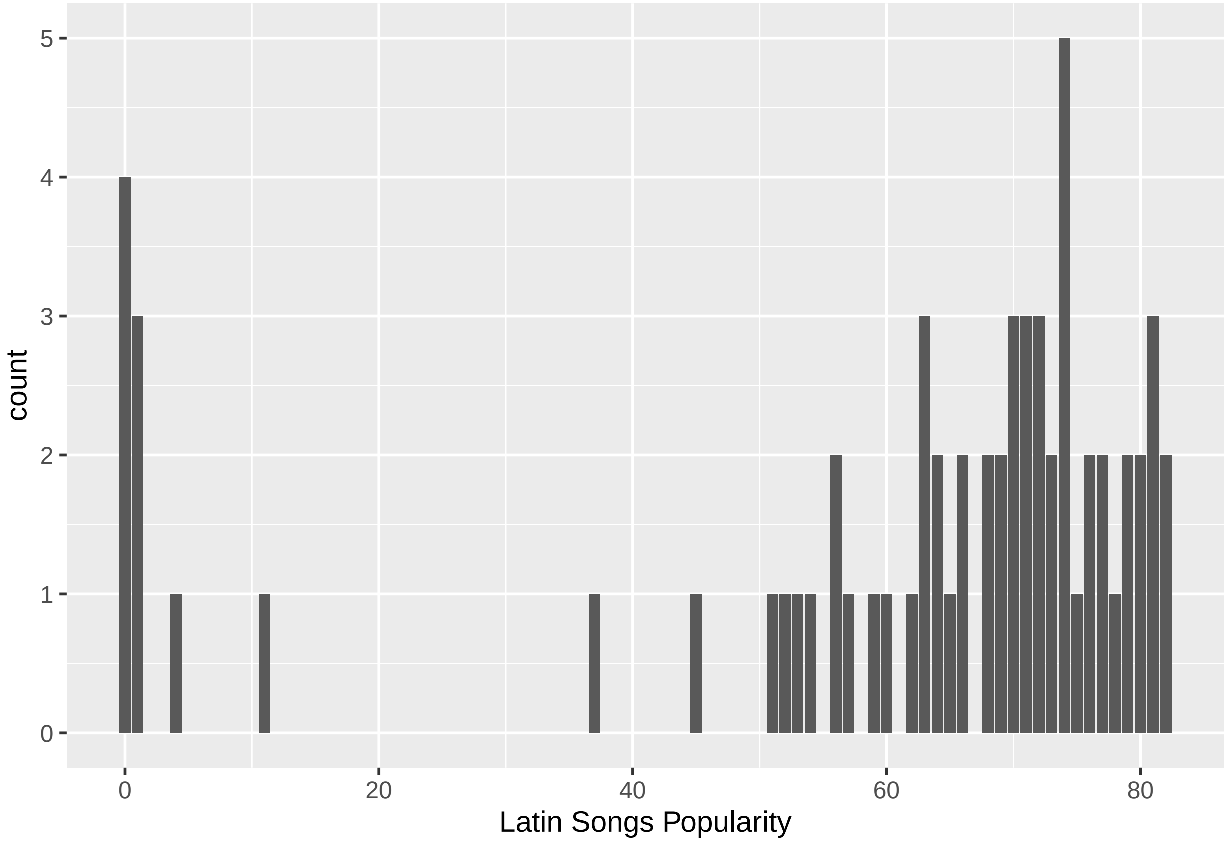
```
ggplot(data = edmSongs, aes(x = popularity)) + geom_bar() + labs(x = "Edm Songs Popularity")
```



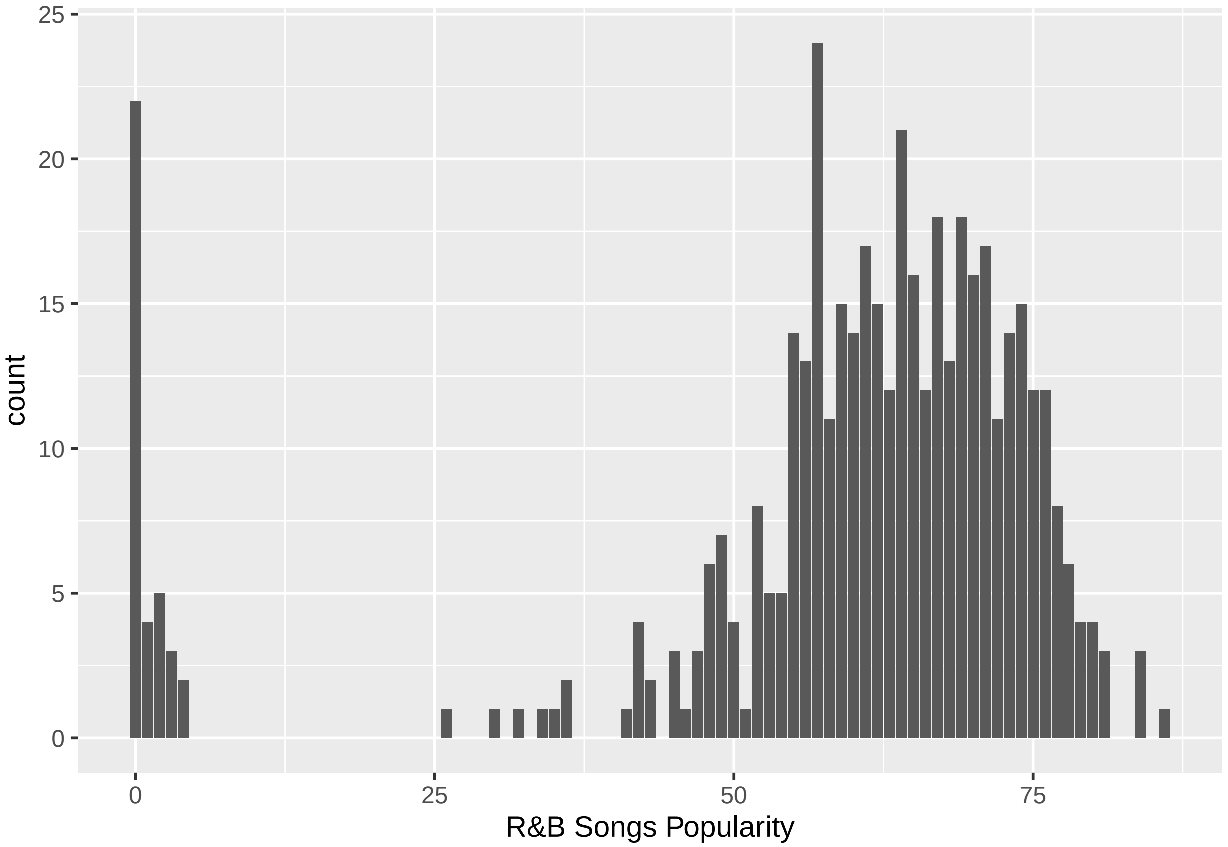
```r
ggplot(data = countrySongs, aes(x = popularity)) + geom_bar() + labs(x = "Country Songs Popularity")
```



```
ggplot(data = latinSongs, aes(x = popularity)) + geom_bar() + labs(x = "Latin Songs Popularity")
```



```
ggplot(data = RandBSongs, aes(x = popularity)) + geom_bar() + labs(x = "R&B Songs Popularity")
```



25
20
15
10
5
0
count
0
25
50
75
R&B Songs Popularity