SPOTIFY ANALYTICS 2010-2019 (R STUDIO)

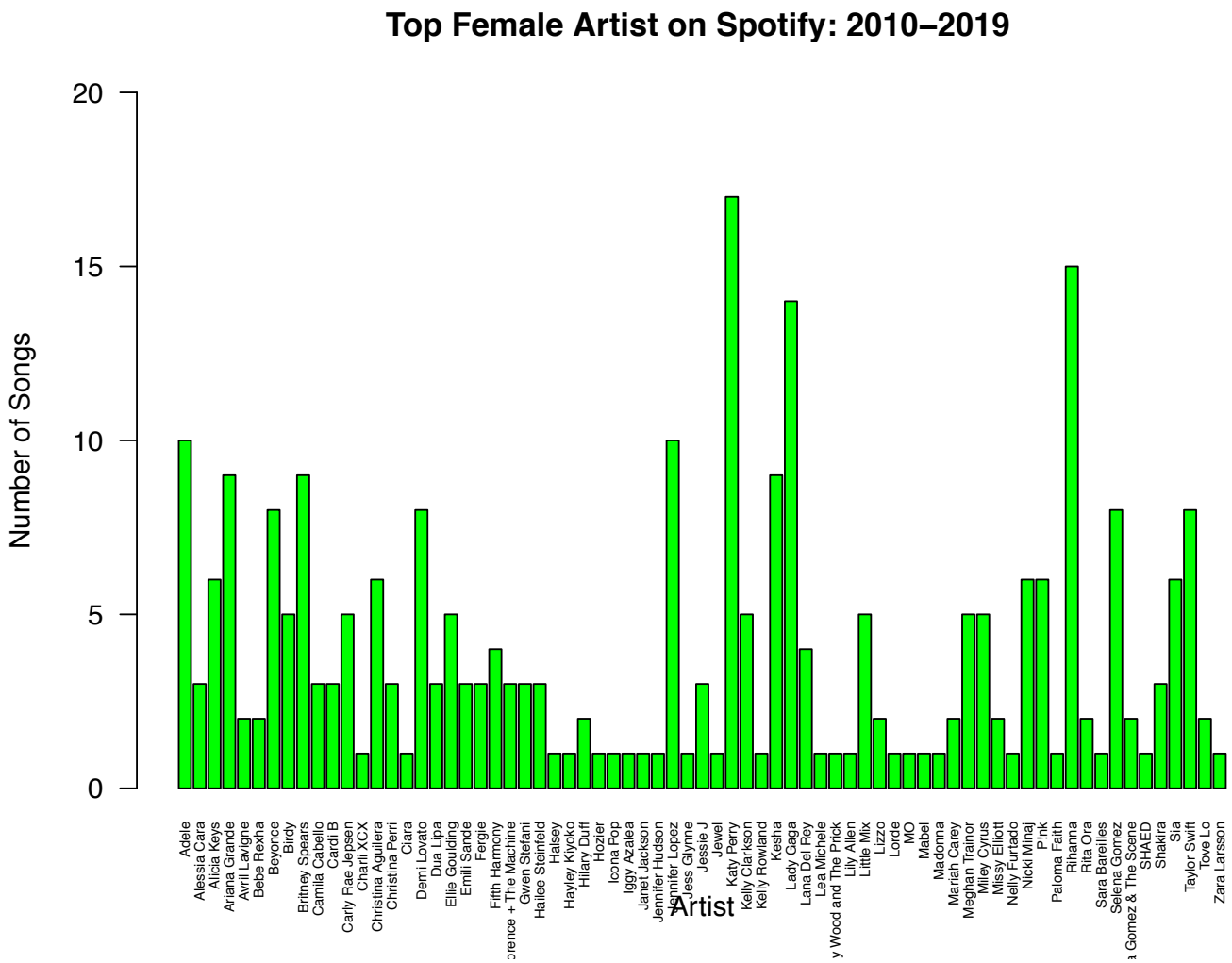
In this project I cleaned, visualized, and used linear regression model to see if there is any correlation between a songs popularity and its Spotify's song metrics to have a better understanding of this dataset. Its safe to assume that since amount of users on Spotify has seen exponential growth since the start of the decade (2010), the platforms users has grown from a niche market to a more mainstream market. Thus we see more accurate models in the later half of the decade rather than the beginning.

Listed below are all of 184 artists within this dataset, after cleaning levels with special characters:

```
  [1] "3OH!3"                  "5 Seconds of Summer"      "A Great Big World"
  [4] "Adam Lambert"           "Adele"                    "Alan Walker"
  [7] "Alessia Cara"           "Alesso"                   "Alicia Keys"
 [10] "Ansel Elgort"           "Ariana Grande"            "Austin Mahone"
 [13] "Avicii"                 "Avril Lavigne"            "BORNS"
 [16] "Bastille"               "Bebe Rexha"               "Beyonce"
 [19] "Birdy"                  "Britney Spears"           "Bruno Mars"
 [22] "Calvin Harris"          "Camila Cabello"           "Cardi B"
 [25] "Carly Rae Jepsen"       "Cashmere Cat"             "Charli XCX"
 [28] "Charlie Puth"           "Chris Brown"              "Christina Aguilera"
 [31] "Christina Perri"        "Ciara"                    "Clean Bandit"
 [34] "CNCO"                   "Coldplay"                 "Daddy Yankee"
 [37] "Daft Punk"              "Dan + Shay"               "David Guetta"
 [40] "Demi Lovato"            "Disclosure"               "DJ Khaled"
 [43] "DJ Snake"               "DNCE"                     "Drake"
 [46] "Dua Lipa"               "Ed Sheeran"               "Ellie Goulding"
 [49] "Emili Sande"            "Eminem"                   "Enrique Iglesias"
 [52] "Far East Movement"      "Fergie"                   "Fifth Harmony"
 [55] "Flo Rida"               "Florence + The Machine"   "fun."
 [58] "G-Eazy"                 "Galantis"                 "Gwen Stefani"
 [61] "Gym Class Heroes"       "Hailee Steinfeld"         "Halsey"
 [64] "Harry Styles"           "Hayley Kiyoko"            "Hilary Duff"
 [67] "Hot Chelle Rae"         "Hozier"                   "Icona Pop"
 [70] "Iggy Azalea"            "J Balvin"                 "James Arthur"
 [73] "Janet Jackson"          "Jason Derulo"             "Jennifer Hudson"
 [76] "Jennifer Lopez"         "Jess Glynne"              "Jessie J"
 [79] "Jewel"                  "Joey Montana"             "John Legend"
 [82] "John Newman"            "Jonas Blue"               "Jonas Brothers"
 [85] "Justin Bieber"          "Justin Timberlake"        "Kanye West"
 [88] "Katy Perry"             "Kelly Clarkson"           "Kelly Rowland"
 [91] "Kesha"                  "Khalid"                   "Kygo"
 [94] "Labrinth"               "Lady Gaga"                "Lana Del Rey"
 [97] "Lea Michele"            "Lewis Capaldi"            "Liam Payne"
[100] "Lilly Wood and The Prick" "Lily Allen"            "Little Mix"
[103] "Lizzo"                  "LMFAO"                    "Lorde"
[106] "Lost Frequencies"       "Luis Fonsi"               "Lukas Graham"
[109] "MO"                     "Mabel"                    "Macklemore & Ryan Lewis"
[112] "Madonna"                "MAGIC!"                   "Major Lazer"
[115] "Mariah Carey"           "Mark Ronson"              "Maroon 5"
[118] "Marshmello"             "Martin Garrix"            "Martin Solveig"
[121] "Meghan Trainor"         "Michael Jackson"          "Mike Posner"
[124] "Miley Cyrus"            "Missy Elliott"            "Mr. Probz"
[127] "N.E.R.D"                "Naughty Boy"              "Ne-Yo"
[130] "Nelly Furtado"          "Neon Trees"               "Niall Horan"
[133] "Nick Jonas"             "Nicki Minaj"              "Olly Murs"
[136] "One Direction"          "OneRepublic"              "Owl City"
[139] "P!nk"                   "Paloma Faith"             "Passenger"
[142] "Pharrell Williams"      "Pitbull"                  "R3HAB"
[145] "RedOne"                 "Ricky Martin"             "Rihanna"
[148] "Rita Ora"               "Robin Schulz"             "Robin Thicke"
```

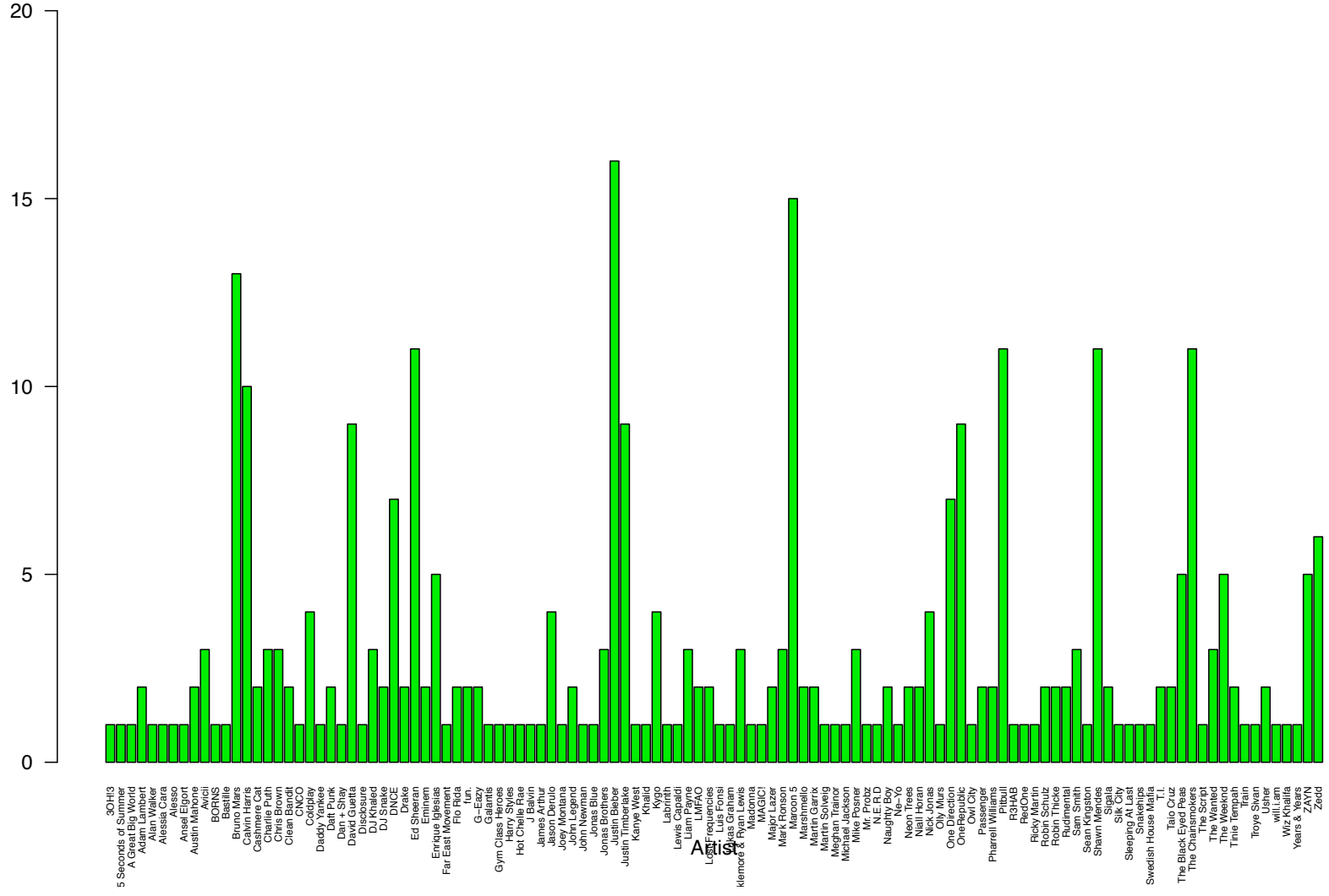
```
[151] "Rudimental"              "Sam Smith"              "Sara Bareilles"
[154] "Sean Kingston"           "Selena Gomez"           "Selena Gomez & The Scene"
[157] "SHAED"                   "Shakira"                "Shawn Mendes"
[160] "Sia"                     "Sigala"                 "Silk City"
[163] "Sleeping At Last"        "Snakehips"              "Swedish House Mafia"
[166] "T.I."                    "Taio Cruz"              "Taylor Swift"
[169] "The Black Eyed Peas"     "The Chainsmokers"       "The Script"
[172] "The Wanted"              "The Weeknd"             "Tinie Tempah"
[175] "Tove Lo"                 "Train"                  "Troye Sivan"
[178] "Usher"                   "will.i.am"              "Wiz Khalifa"
[181] "Years & Years"           "Zara Larsson"           "ZAYN"
[184] "Zedd"
```

I wanted to see which artist had the most Top 50 songs and the top genre throughout the decade. To organize the data I subsetted the artist by genders, used the plot function to visualize, included the summaries of each gender, and followed with a graph of the top genres.

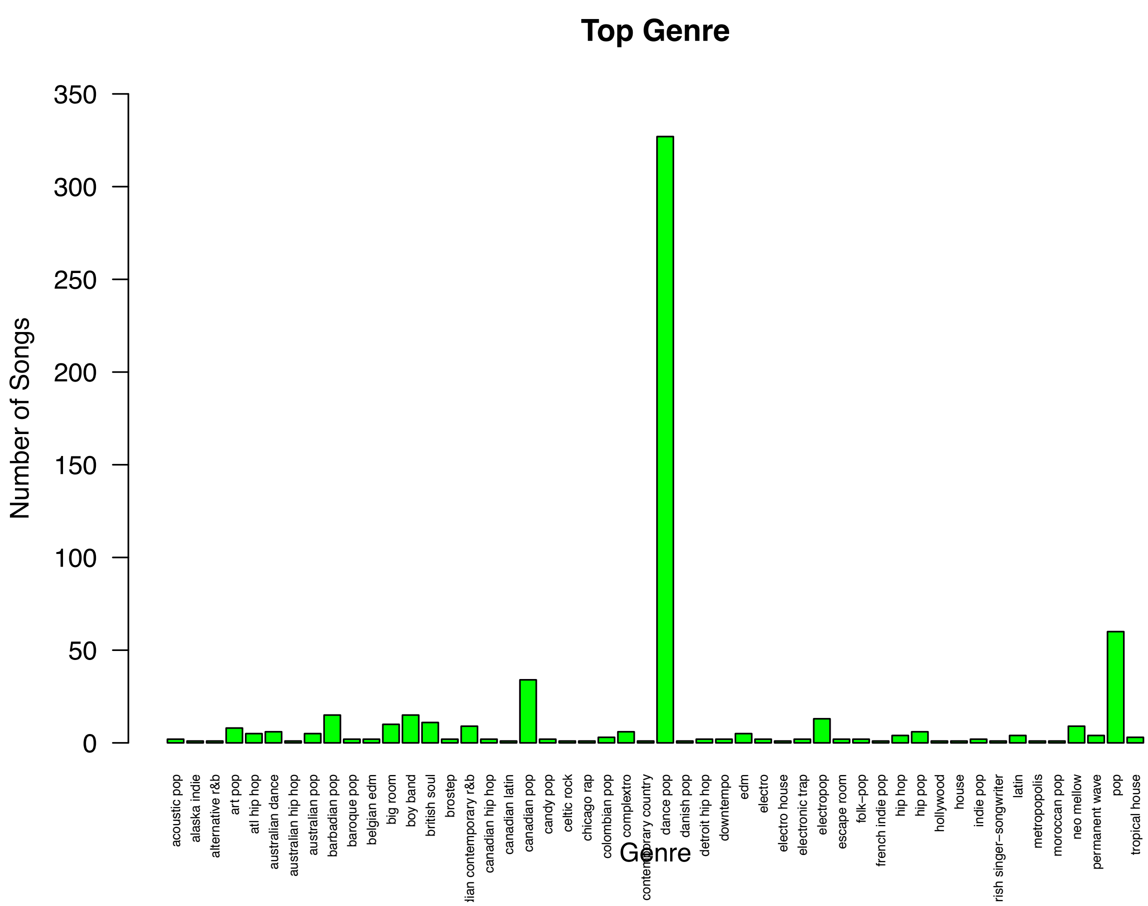


```
       X                                                  title          artist          top.genre
 Min.   :  3.0   A Little Party Never Killed Nobody (All We Got):  2   Katy Perry    : 17   dance pop    :185
 1st Qu.:128.8   All I Ask                                      :  2   Rihanna       : 15   barbadian pop: 15
 Median :270.5   Here                                           :  2   Lady Gaga     : 14   british soul : 11
 Mean   :280.9   We Are Never Ever Getting Back Together         :  2   Adele         : 10   pop          : 11
 3rd Qu.:418.2   ...Ready For It? - BloodPop\xa8 Remix          :  1   Jennifer Lopez: 10   art pop      :  8
 Max.   :587.0   #Beautiful                                     :  1   Ariana Grande :  9   canadian pop :  7
                 (Other)                                        :266   (Other)       :201   (Other)      : 39
      year           bpm             nrgy            dnce              dB              live            val
 Min.   :2010   Min.   :  0.0   Min.   : 0.00   Min.   : 0.00   Min.   :-60.000   Min.   : 0.00   Min.   : 0.00
 1st Qu.:2012   1st Qu.:100.0   1st Qu.:60.00   1st Qu.:56.00   1st Qu.: -6.000   1st Qu.: 9.00   1st Qu.:33.00
 Median :2014   Median :120.0   Median :73.00   Median :65.00   Median : -5.000   Median :13.00   Median :51.00
 Mean   :2014   Mean   :120.4   Mean   :69.23   Mean   :63.53   Mean   : -5.641   Mean   :17.92   Mean   :50.16
 3rd Qu.:2016   3rd Qu.:130.0   3rd Qu.:81.00   3rd Qu.:73.00   3rd Qu.: -4.000   3rd Qu.:24.00   3rd Qu.:68.25
 Max.   :2019   Max.   :206.0   Max.   :96.00   Max.   :97.00   Max.   : -2.000   Max.   :74.00   Max.   :96.00

      dur             acous            spch             pop          gender
 Min.   :157.0   Min.   : 0.0   Min.   : 0.000   Min.   : 0.00   F:276
 1st Qu.:203.8   1st Qu.: 1.0   1st Qu.: 4.000   1st Qu.:56.75
 Median :221.0   Median : 5.0   Median : 5.000   Median :66.00
 Mean   :225.8   Mean   :14.2   Mean   : 8.362   Mean   :63.06
 3rd Qu.:243.0   3rd Qu.:16.0   3rd Qu.:10.000   3rd Qu.:73.25
 Max.   :403.0   Max.   :97.0   Max.   :48.000   Max.   :97.00
```

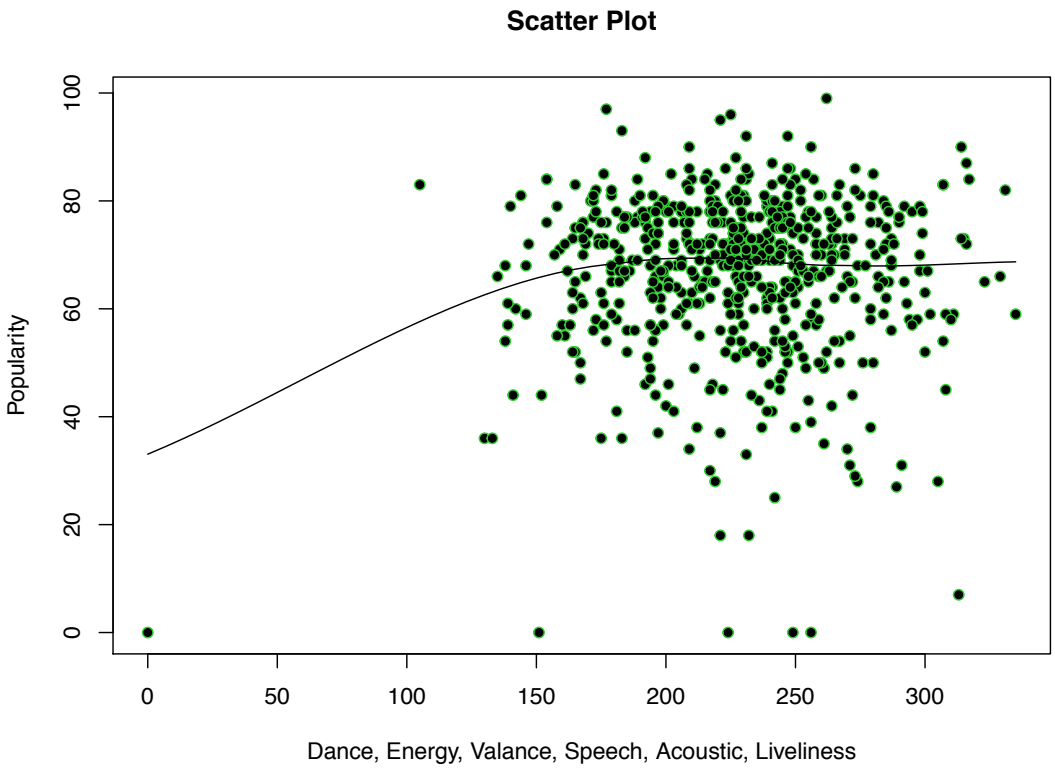
Top Male Artist on Spotify: 2010–2019



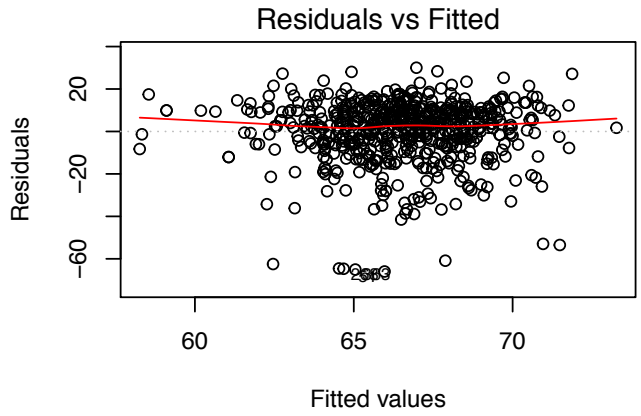
```
       X                                               title                artist              top.genre          year     
 Min.   :  1.0   Castle Walls (feat. Christina Aguilera):  2   Justin Bieber: 16   dance pop    :142   Min.   :2010  
 1st Qu.:168.5   Company                                :  2   Maroon 5     : 15   pop          : 49   1st Qu.:2013  
 Median :323.0   First Time                             :  2   Bruno Mars   : 13   canadian pop: 27   Median :2015  
 Mean   :319.8   Just the Way You Are                   :  2   Ed Sheeran   : 11   boy band     : 15   Mean   :2015  
 3rd Qu.:480.5   Kissing Strangers                      :  2   Pitbull      : 11   electropop   : 12   3rd Qu.:2017  
 Max.   :603.0   Love Yourself                          :  2   Shawn Mendes : 11   big room     : 10   Max.   :2019  
                 (Other)                                :315   (Other)      :250   (Other)      : 72                 
      bpm          nrgy            dnce            dB               live            val              dur       
 Min.   : 65   Min.   : 4.00   Min.   :26.0   Min.   :-15.000   Min.   : 2.00   Min.   : 4.00   Min.   :134.0  
 1st Qu.:100   1st Qu.:61.00   1st Qu.:59.0   1st Qu.: -7.000   1st Qu.: 9.00   1st Qu.:37.50   1st Qu.:202.0  
 Median :120   Median :75.00   Median :67.0   Median : -5.000   Median :12.00   Median :54.00   Median :220.0  
 Mean   :117   Mean   :71.58   Mean   :65.1   Mean   : -5.526   Mean   :17.65   Mean   :53.97   Mean   :223.8  
 3rd Qu.:128   3rd Qu.:83.50   3rd Qu.:74.0   3rd Qu.: -4.000   3rd Qu.:24.50   3rd Qu.:70.50   3rd Qu.:236.0  
 Max.   :192   Max.   :98.00   Max.   :93.0   Max.   : -2.000   Max.   :70.00   Max.   :98.00   Max.   :424.0  

     acous            spch             pop          gender 
 Min.   : 0.00   Min.   : 3.000   Min.   : 0.00   M:327  
 1st Qu.: 2.00   1st Qu.: 4.000   1st Qu.:64.00          
 Median : 7.00   Median : 6.000   Median :72.00          
 Mean   :14.43   Mean   : 8.355   Mean   :69.44          
 3rd Qu.:17.50   3rd Qu.: 9.000   3rd Qu.:78.00          
 Max.   :99.00   Max.   :45.000   Max.   :99.00          
```

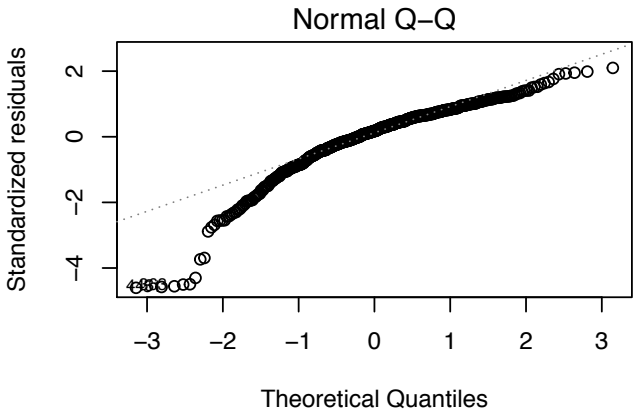


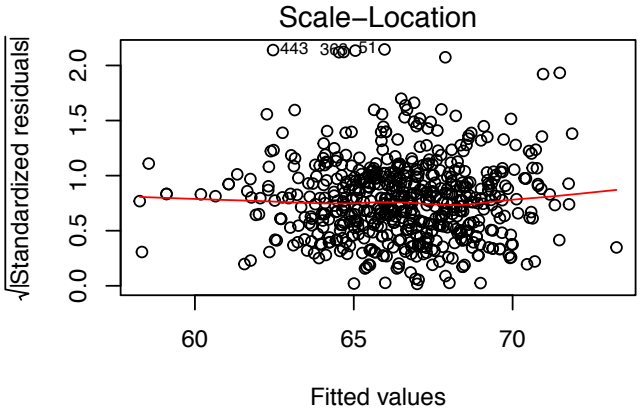
In my R code I go in depth with the analysis' and visualization by doing the same graphing process each year in the decade. But to avoid redundancy, we can skip to the the multiple linear regression models. The first model was built comparing popularity to Spotify's song metric variables which are listed as Danceability, Energy, Valance, Spechiness, Acousticness, and Liveliness. I used a scatter plot to observe the initial data, and I can see that there are some outliers. I created our first linear model, and summary to identify those outliers and remove them.

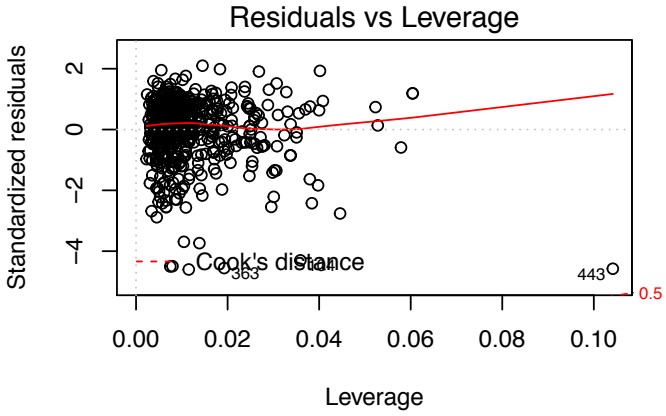



```
Call:
lm(formula = spotify$pop ~ spotify$dnce + spotify$nrgy + spotify$val +
    spotify$spch + spotify$acous + spotify$live, data = spotify)

Residuals:
    Min      1Q  Median      3Q     Max
-65.968  -5.972   2.518   9.457  30.029

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    62.459456   4.652240  13.426   <2e-16 ***
spotify$dnce    0.133895   0.051995   2.575   0.0103 *
spotify$nrgy   -0.051024   0.047467  -1.075   0.2828
spotify$val     0.006441   0.033010   0.195   0.8454
spotify$spch   -0.048667   0.080512  -0.604   0.5458
spotify$acous   0.014787   0.034932   0.423   0.6722
spotify$live   -0.062050   0.046159  -1.344   0.1794
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.42 on 596 degrees of freedom
Multiple R-squared:  0.02386,      Adjusted R-squared:  0.01403
F-statistic: 2.428 on 6 and 596 DF,  p-value: 0.02507
```





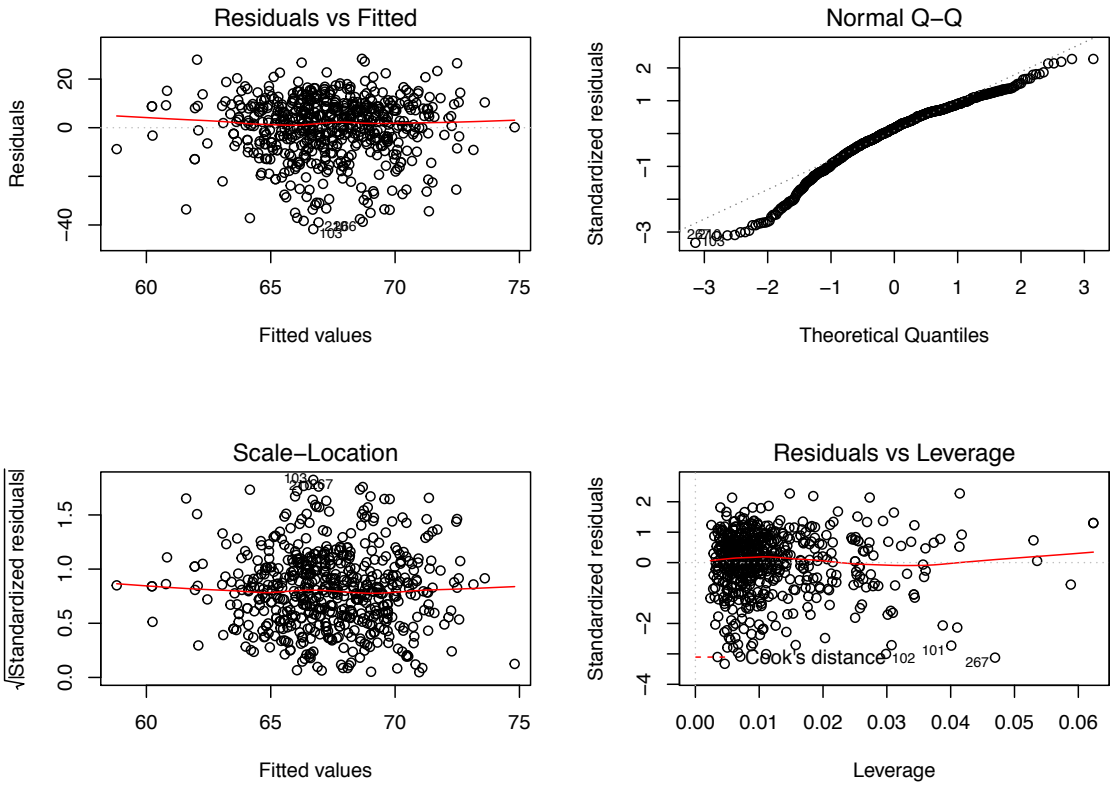




From this initial linear model I could clearly see which values are outliers (139,51,363,443,104,268,362,442). All of these have some kind of error in the data, but the majority of them have all values set at zero. These outliers could cause errors in the model, I reduce the dataset down by removing these values and building a new model.

```
Call:
lm(formula = spotify2$pop ~ spotify2$dnce + spotify2$nrgy + spotify2$val +
    spotify2$spch + spotify2$acous + spotify2$live, data = spotify2)

Residuals:
    Min      1Q  Median      3Q     Max
-41.709  -6.458   1.981   8.665  28.331

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     69.66831    4.35515  15.997  < 2e-16 ***
spotify2$dnce    0.09859    0.04692   2.101  0.03603 *
spotify2$nrgy   -0.11207    0.04332  -2.587  0.00992 **
spotify2$val     0.01772    0.02901   0.611  0.54147
spotify2$spch   -0.04911    0.07100  -0.692  0.48935
spotify2$acous  -0.01642    0.03148  -0.522  0.60205
```

```
spotify2$live  -0.05844    0.04041  -1.446  0.14870
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.57 on 588 degrees of freedom
Multiple R-squared:  0.03374,      Adjusted R-squared:  0.02388
F-statistic: 3.422 on 6 and 588 DF,  p-value: 0.0025
```



When looking at the summary and models it is safe to reject the null hypotheses, thus there seems to be no correlations in a songs attributes and its popularity. This is more than likely inferred by the fact that the popularity of a lot of these songs are based on the how popular the artist is. So in conclusion this model would not be efficient to use. However moving forward I will reduce the dataset down to the songs by artist that only appear once. From there I can see if there is any correlations since there will not be any artist popularity bias.