Nirali Bandaru, Rohan Gangisetty f20-spotify2 CPSC6300

Group Assignment 2 - Exploratory Data Analysis - Spotify 2

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

QUESTION 1

Basic Information about Dataset:

Unit of Analysis	Song/Track
Total Observations	169,909
Unique Observations	154417
Time Period Covered	1921-2020

Predictor Details:

Predictor Name	**Data Type**	**Range before cleaning**	**Range after cleaning**
Acousticness	Numeric	0 - 1	0-0.996
Danceability	Numeric	0 - 1	0-0.988
Popularity	Numeric	0 - 100	0-100
Mode	Binary	0, 1	0-1
Key	Categorical	0 - 11	0-11
Loudness	Float	-60.00 - 3.855	-60-0
Liveness	Numeric	0 - 1	0-1
Instrumentalness	Numeric	0 - 1	0-1
Artist	Categorical	NA	removed
Year	Categorical	1921-2020	1921-2020
Name	Categorical	NA	removed
Duration (ms)	Integer	5k-500k	60000-360000
Explicit	Binary	0, 1	removed
Speechiness	Numeric	0 - 1	0-0.967
Tempo	Numeric	0-250	0-244.09
Release Date (yyyy-mm-dd)	Categorical	NA	removed

Note: Precision of date varies

Question 2

Summary of data cleaning

- Removed unwanted columns
    - Artists
    - Release Date
    - Explicit
    - ID
    - Name
- Checked for missing values
    - No missing values
- Checked for duplicate observations
    - Checked and deleted
- Removed unwanted observations
    - Non-unique values
- Double checked data types
    - No discrepancies
- Checked for illegal values (most columns have values from 0-1)
    - Loudness had a max value of 3.8, out of bounds from range -60 to 0. Set to 0.
    - Minimum value is given as “5108ms” in the data summary, which is a mere 5 seconds. We have decided to not consider this a song. A threshold value at 1 min is chosen at which we determine whether a track is a “song” or not.
- Check for typos
    - None
- Check for mislabeled classes
    - None
- Do we have outliers that we don’t want?
    - Examined in data visualization

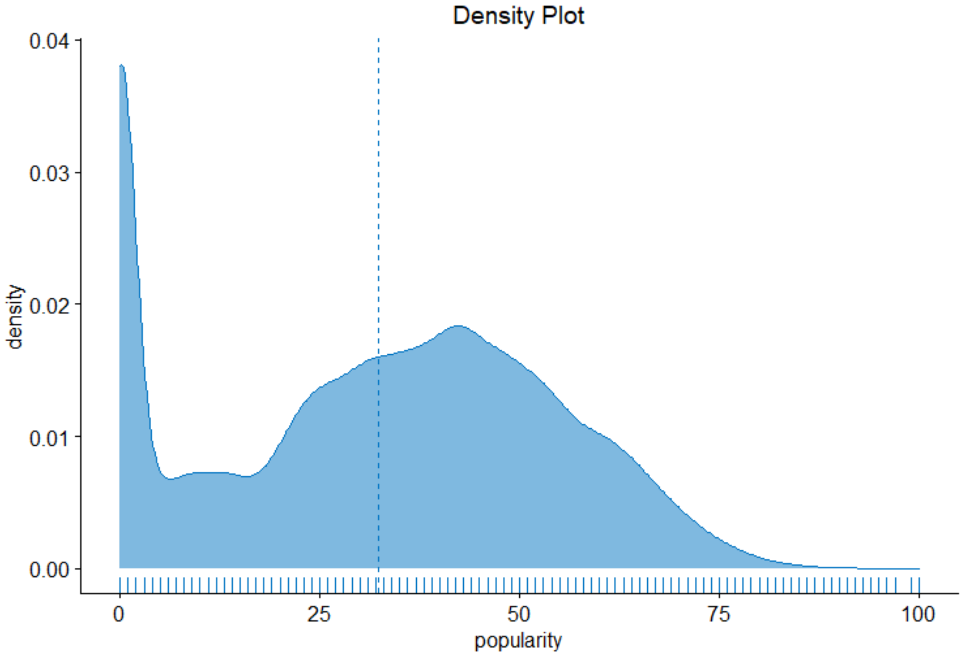
Number of observations before cleaning - 169909

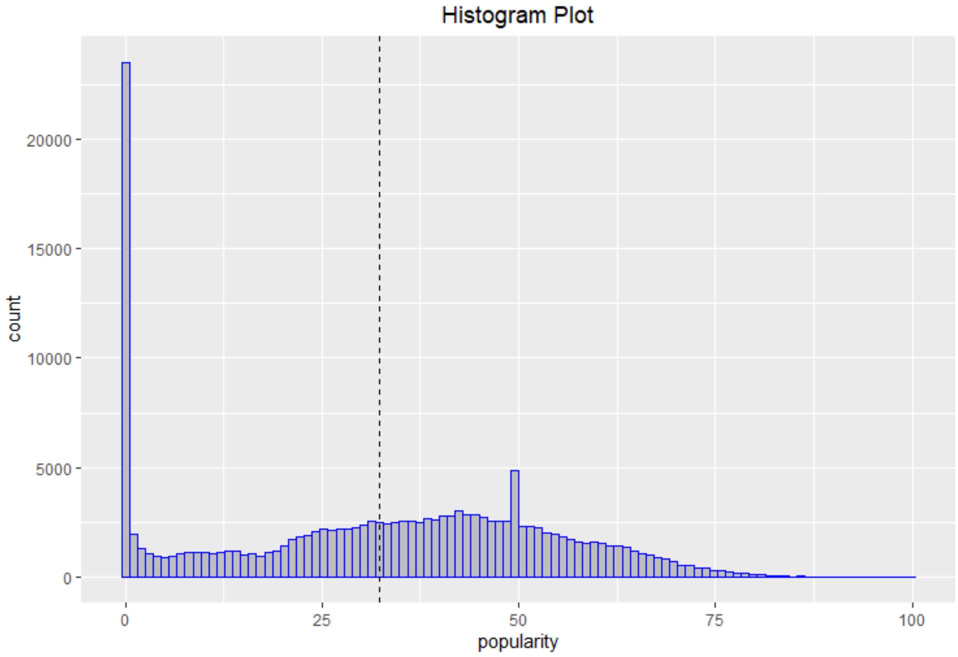
Number of observations after cleaning - 154417

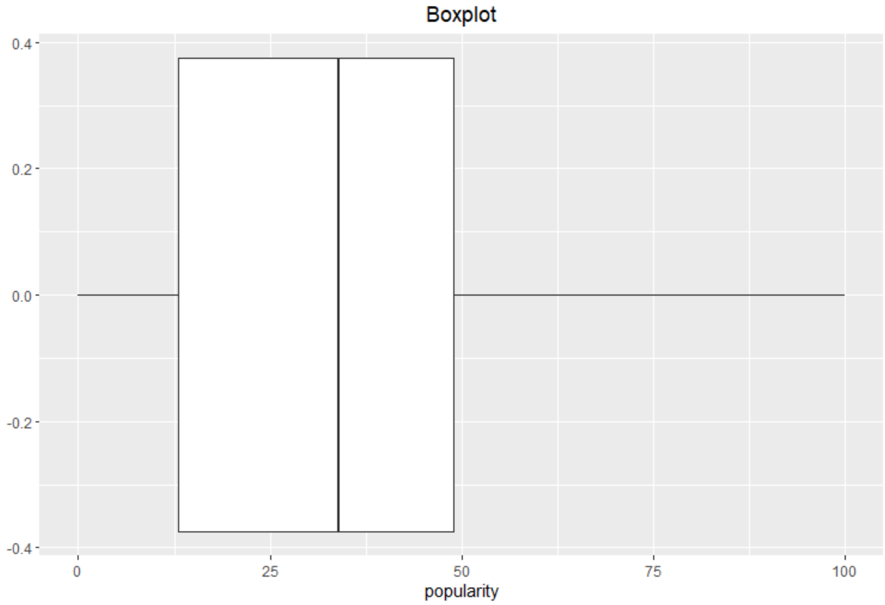
Question 3

Summary of response with appropriate visualization technique

Response variable : Popularity


Density Plot
density
0.04
0.03
0.02
0.01
0.00
0
25
50
75
100
popularity


Histogram Plot
count
20000
15000
10000
5000
0
0
25
50
75
100
popularity

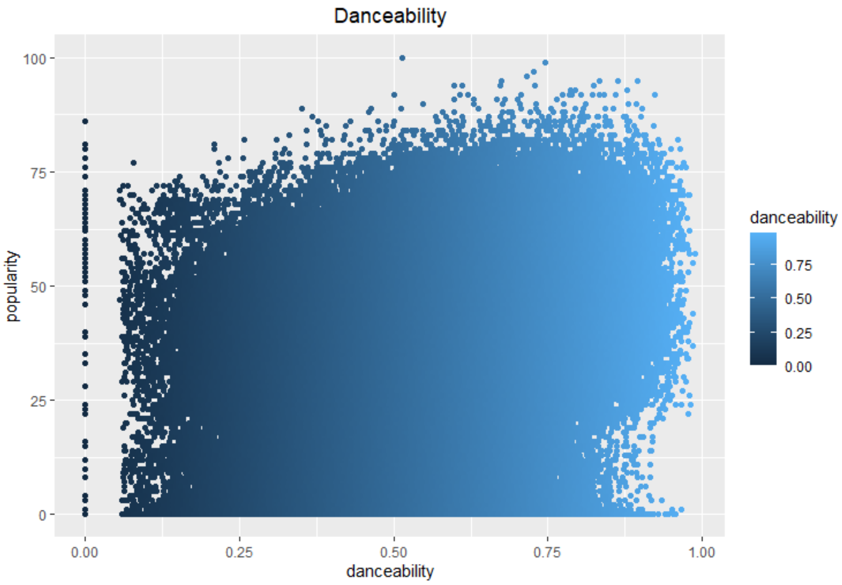


Question 4

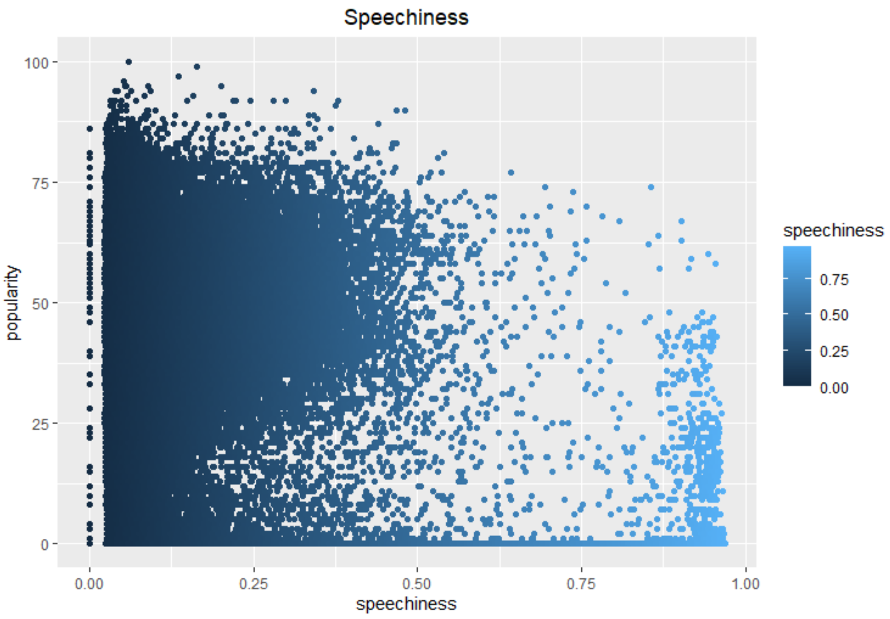
Summary of key predictors with appropriate visualization techniques that compare predictors to the response

Key predictors

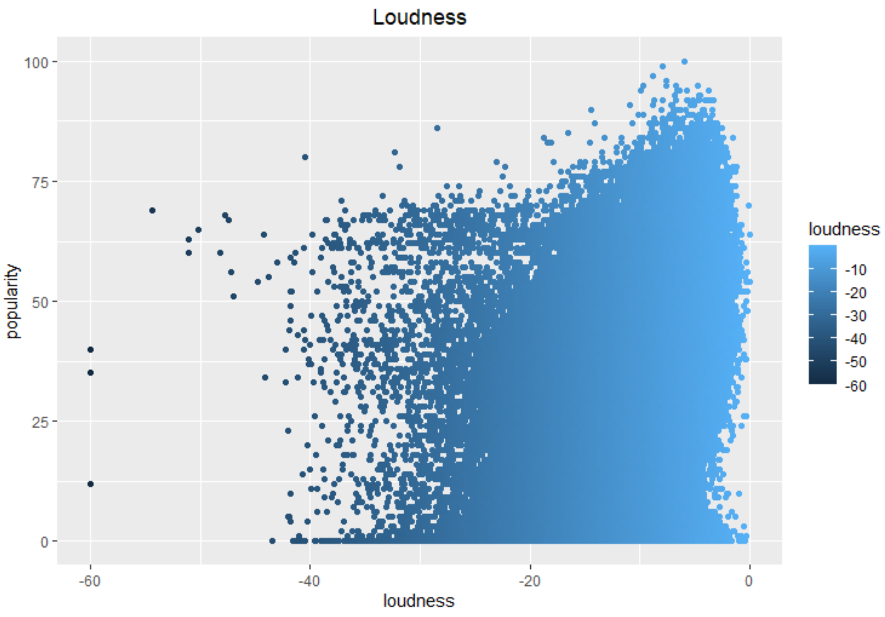
- Danceability



- Speechiness



- Loudness



RESOURCES:

Handling missing data: https://www.statmethods.net/input/missingdata.html

https://towardsdatascience.com/data-cleaning-with-r-and-the-tidyverse-detecting-missing-values-ea23c519bc62

Data Cleaning : https://elitedatascience.com/data-cleaning

Audio Features on Spotify: https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

Sorting data: https://www.displayr.com/how-to-sort-data-in-r/

Plots:

https://cran.r-project.org/web/packages/ggpubr/readme/README.html

www.sthda.com/english/articles/32-r-graphics-essentials/133-plot-one-variable-frequency-graph-density-distribution-and-more/

Ggplot2 aesthetics: https://stackoverflow.com/questions/40675778/center-plot-title-in-ggplot2

https://towardsdatascience.com/explore-two-variables-using-r-68ece9cbcd81