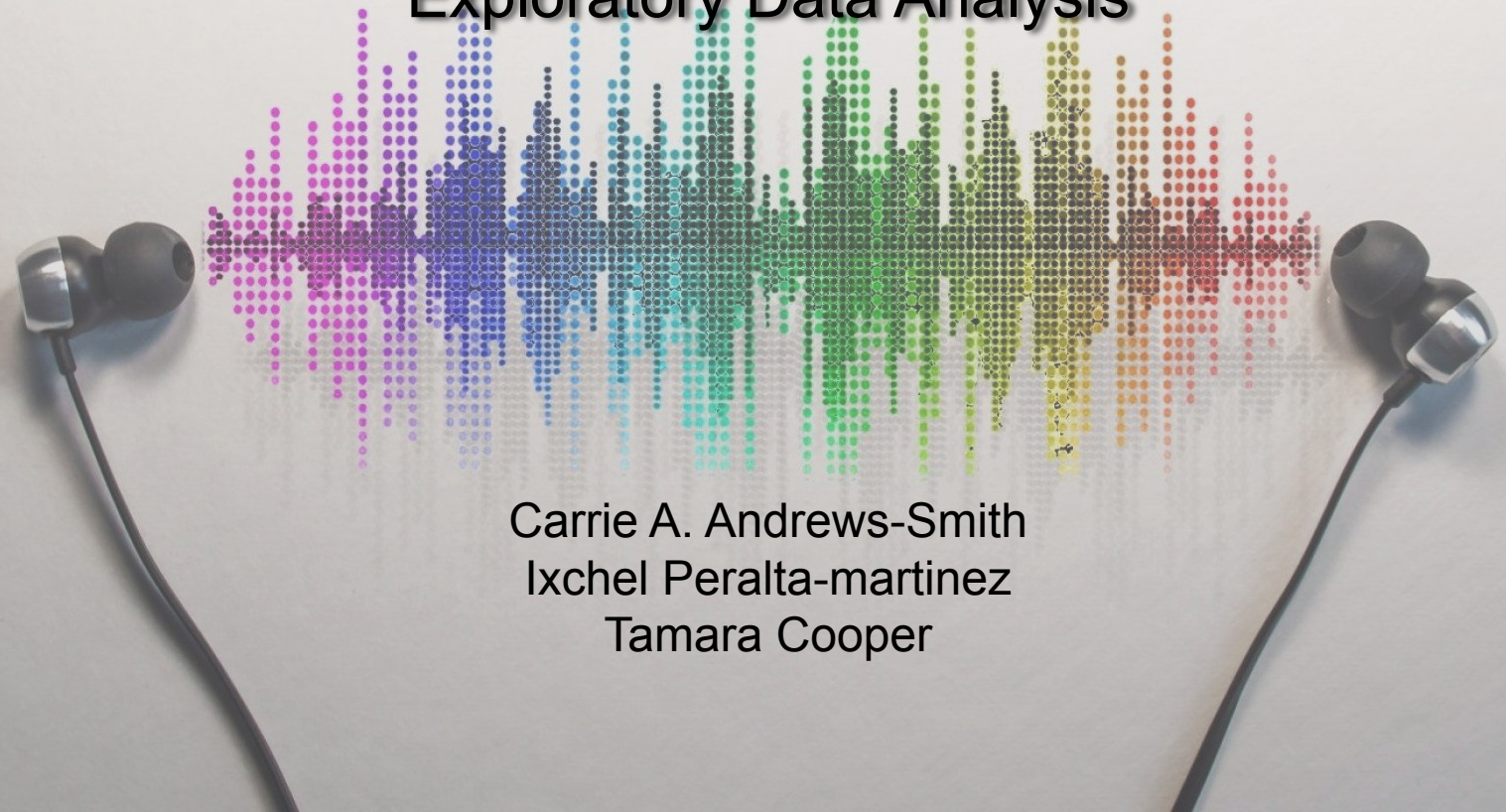
Spotify - All Time Top 2000s Mega Dataset

Exploratory Data Analysis



Carrie A. Andrews-Smith
Ixchel Peralta-martinez
Tamara Cooper

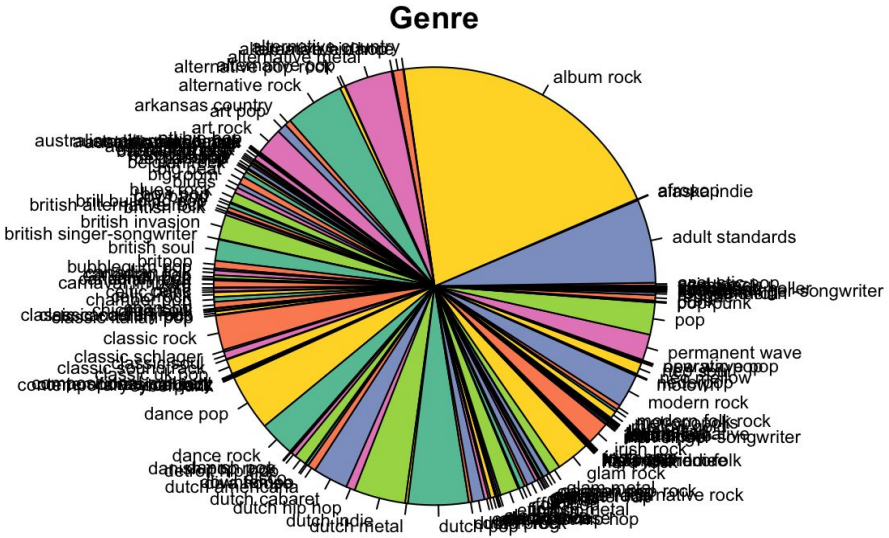
Dataset Description *"Spotify - All Time Top 2000s Mega Dataset"*

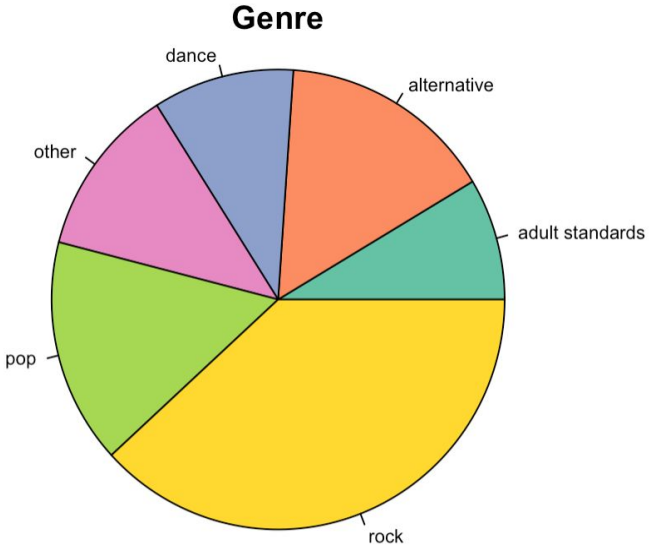
- The data set Spotify All Time Top 2000 tracks contains 1994 values
- The categorical variables are Title, Artist, and Genre
- All remaining variables are numerical

Index	Title	Artist	Top Genre	Year	BPM	Energy	Danceability	Loudness (dB)	Liveness	Valence	Length	Acousticness	Speechiness	Popularity
1	Sunrise	Norah Jones	adult standards	2004	157	30	53	-14	11	68	201	94	3	71
2	Black Night	Deep Purple	album rock	2000	135	79	50	-11	17	81	207	17	7	39
3	Clint Eastwood	Gorillaz	alternative hip hop	2001	168	69	66	-9	7	52	341	2	17	69
4	The Pretender	Foo Fighters	alternative metal	2007	173	96	43	-4	3	37	269	0	4	76
5	Waitin' On A Sunny D	Bruce Springsteen	classic rock	2002	106	82	58	-5	10	87	256	1	3	59
6	The Road Ahead (Mile	City To City	alternative pop rock	2004	99	46	54	-9	14	14	247	0	2	45
7	She Will Be Loved	Maroon 5	pop	2002	102	71	71	-6	13	54	257	6	3	74
8	Knights of Cydonia	Muse	modern rock	2006	137	96	37	-5	12	21	366	0	14	69
9	Mr. Brightside	The Killers	modern rock	2004	148	92	36	-4	10	23	223	0	8	77
10	Without Me	Eminem	detroit hip hop	2002	112	67	91	-3	24	66	290	0	7	82

Data and Categorical variables

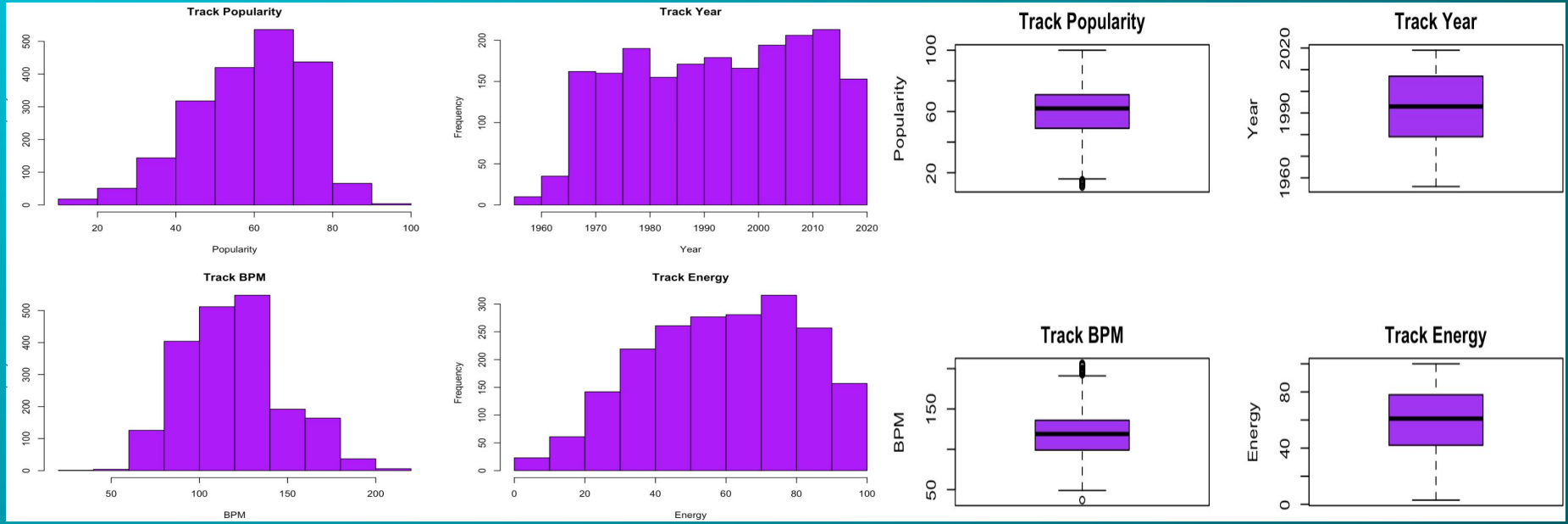
- Original data set contains 149 unique genres





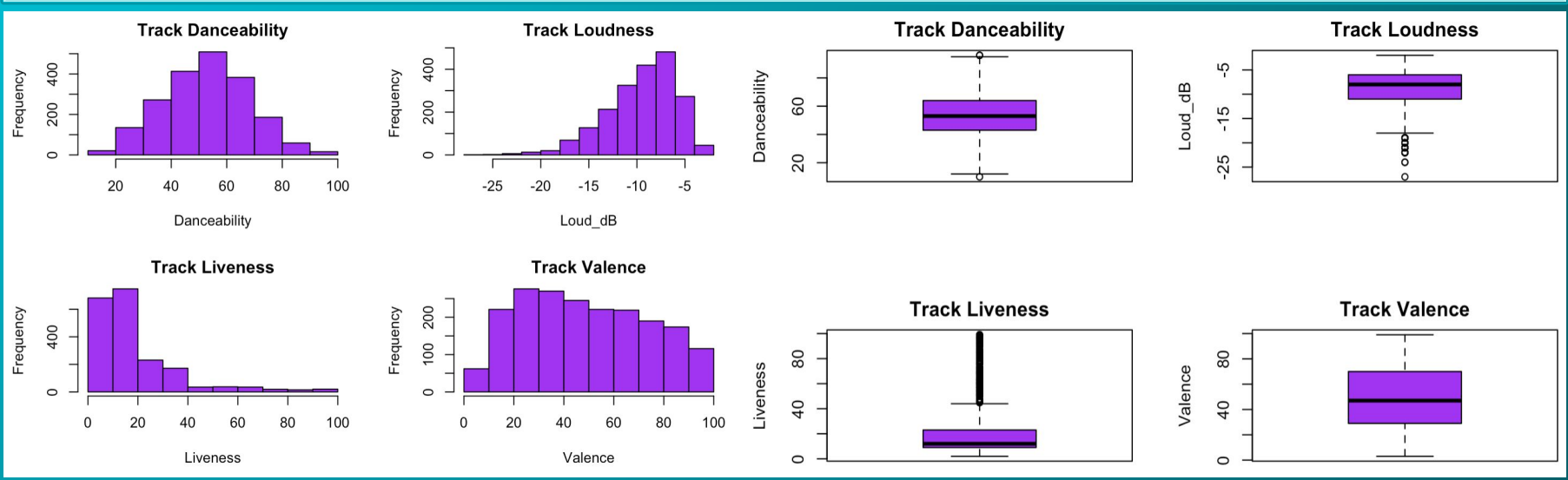
Numerical variables

- Approximately normally distributed, except year which is approximately uniformly distributed



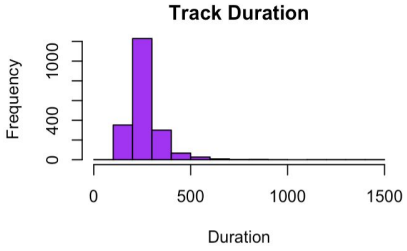
Numerical variables

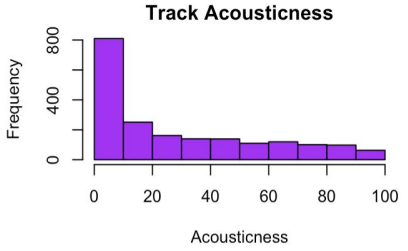
- Danceability and Valence approximately normally distributed
- Loudness and Liveness are skewed left and right respectively
- Large amount of outliers in Liveness

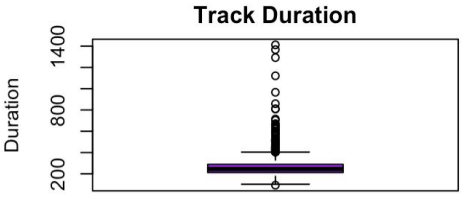


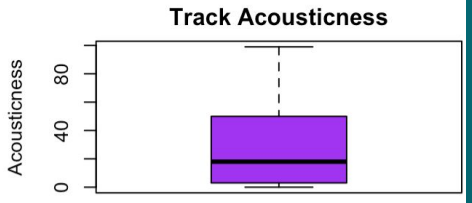
Numerical variables

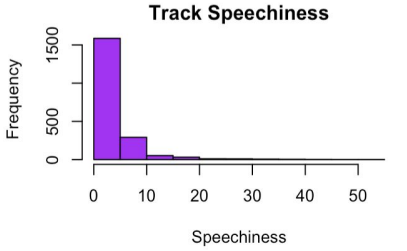
- Right skewed
- Duration and Speechiness have a lot of outliers

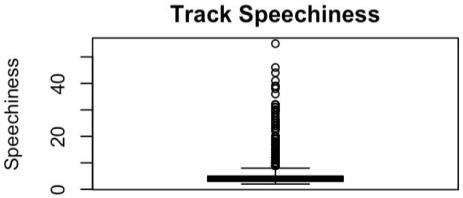






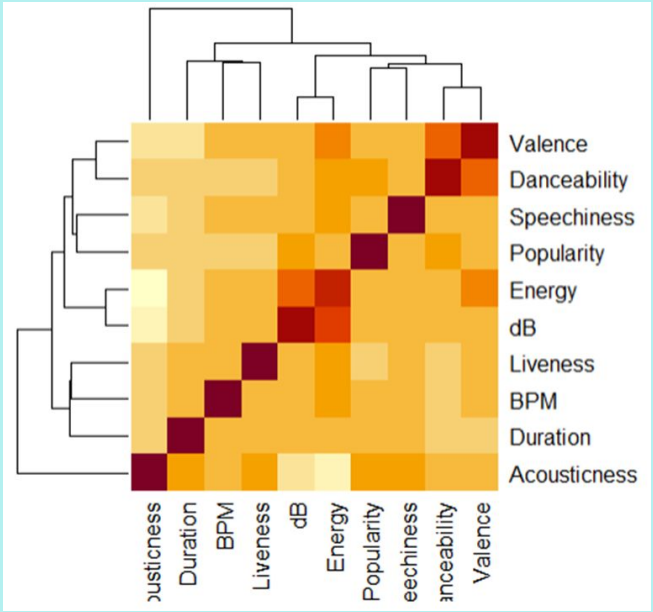






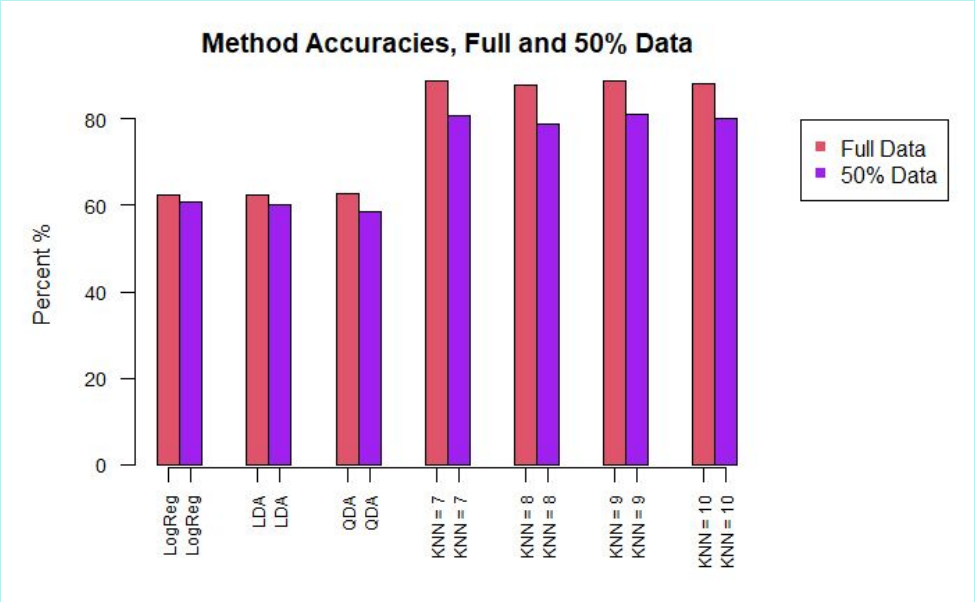
Heat Map

- Mostly weak correlation between the variables, especially when we look at Popularity as a response variable.

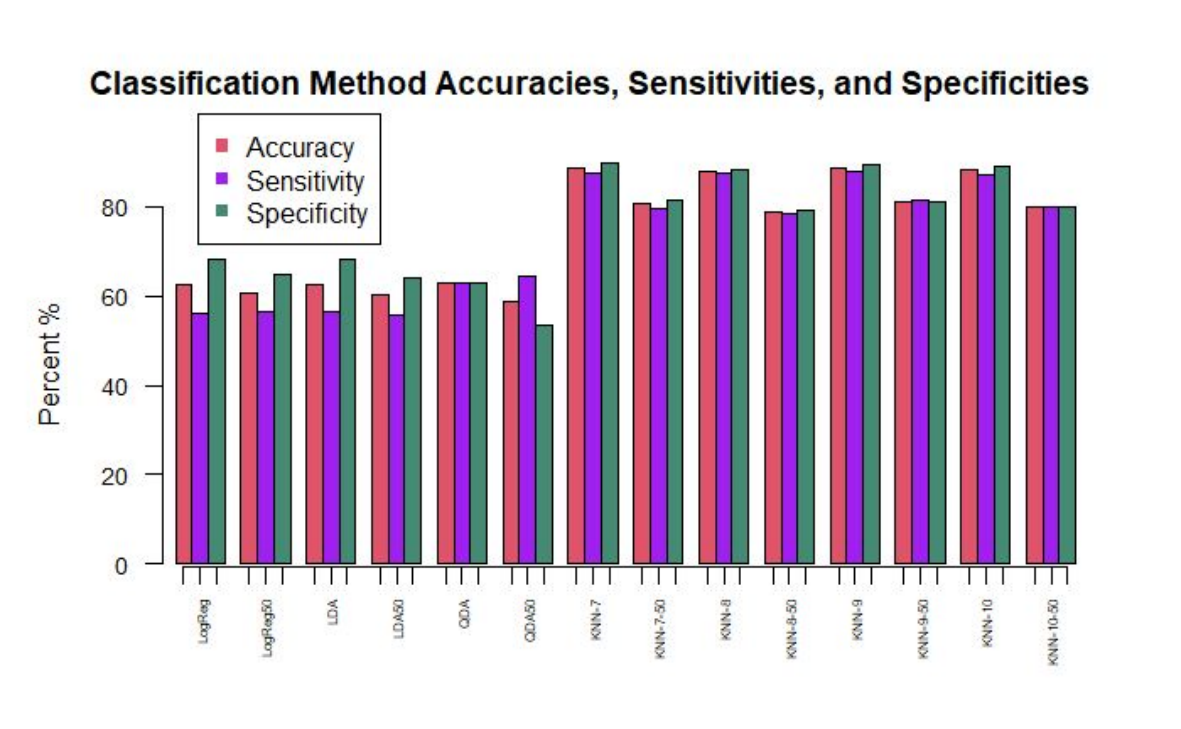


Classification: Full Data vs 50% Data

- Categorical popularity ~ All numerical variables
- Predictors: Year, BPM, Energy, Danceability, dB, Liveness, Valence, Duration, Acousticness, Speechiness
- 50% data accuracies:
  - Log Regression, LDA, & QDA similar (~60%)
  - KNN (k=7-10) similar (~80%)



Classification statistics



Classification Methods: Test Error & Run Time

	Full Data				50% Data			
Method	**% Sensitivity**	**% Specificity**	**% Test Error**	**Run Time (s)**	**% Sensitivity**	**% Specificity**	**% Test Error**	**Run Time (s)**
LogReg	56.23	68.24	37.51	0.033	56.39	64.62	39.32	0.027
LDA	56.34	68.24	37.46	0.028	55.77	64.23	39.82	0.026
QDA	62.83	62.75	37.21	0.026	64.36	53.27	41.42	0.025
KNN-7	87.43	89.99	11.23	0.150	79.66	81.54	19.36	0.037
KNN-8	87.64	88.26	12.04	0.069	78.62	79.04	21.16	0.040
KNN-9	88.06	89.32	11.28	0.072	81.34	80.96	18.86	0.041
KNN-10	87.33	89.03	11.79	0.079	80.08	80.00	19.96	0.047

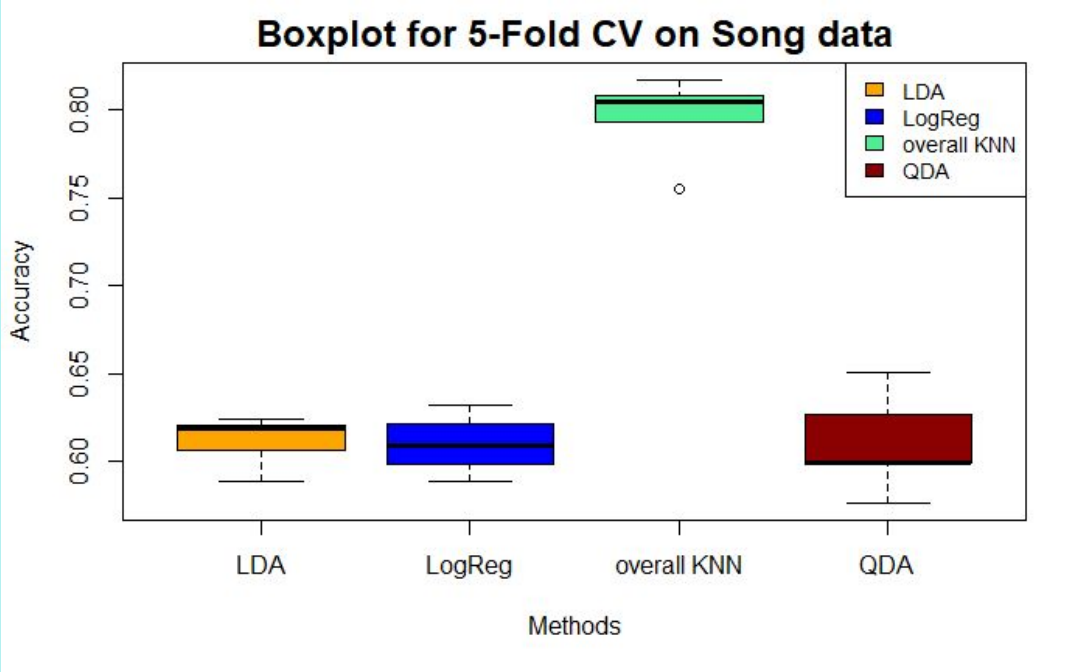
Best Overall Classification Model

- As expected, a decrease in accuracy was observed for the 50% data model in comparison to the full data model
- 50% data KNN classification models outperformed Log Regression, LDA, & QDA methods
- k = 9 yielded best accuracy, sensitivity, & specificity (81% for each)

	Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN-7	80.64	81.54	79.66
KNN-8	78.84	79.04	78.62
KNN-9	81.14	80.96	81.34
KNN-10	80.04	80.00	80.08

k-folds Cross Validation

- The overall KNN 5-fold CV outperformed Log Regression, LDA, & QDA CV models
- Accuracy comparisons:
  - Log Reg = 60.98%
  - overall KNN = 79.55%
  - LDA = 60.98%
  - QDA = 61.03%

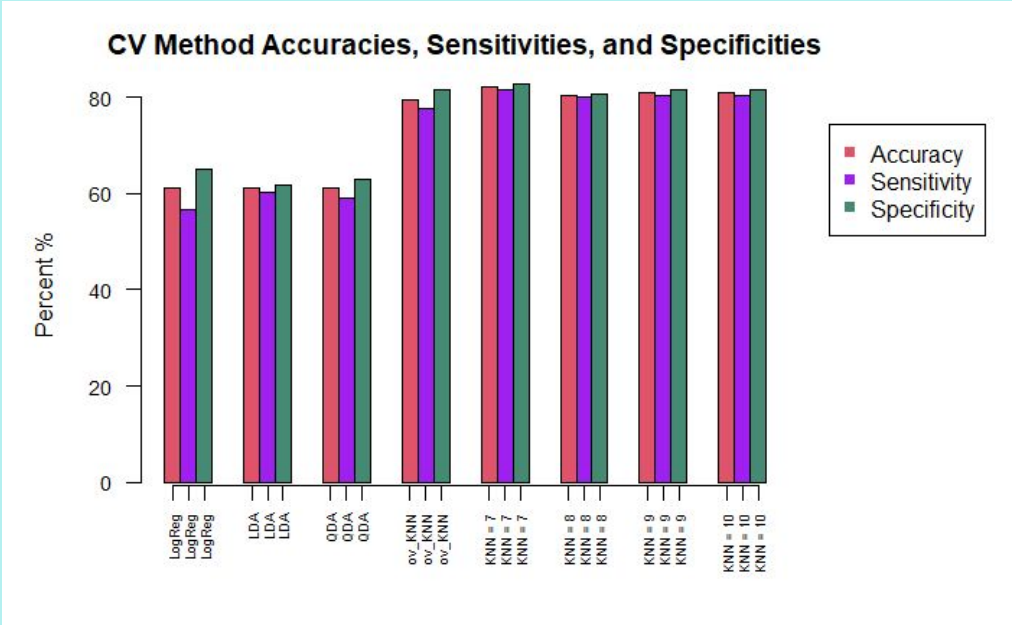


k-folds Cross Validation Statistics Data

k = 5 CV	%Accuracy	%Sensitivity	%Specificity	Run Time (s)
LogReg	60.98	56.54	65.06	0.083
LDA	60.98	60.32	61.86	0.085
QDA	61.03	59.14	62.84	0.077
overall KNN	79.55	77.57	81.37	0.545
KNN-7	82.15	81.56	82.68	0.092
KNN-8	80.24	79.87	80.57	0.073
KNN-9	80.89	80.27	81.45	0.079
KNN-10	80.94	80.42	81.41	0.113

k-folds Cross Validation Statistics

- Statistical analysis supports findings of KNN dominating CV models
- KNN CV models with specified k-values:
  - Higher accuracy
  - Higher sensitivity
  - Higher specificity
    - Similar results for k-values 7 - 10.
    - CV for KNN, k = 7, shows best fit

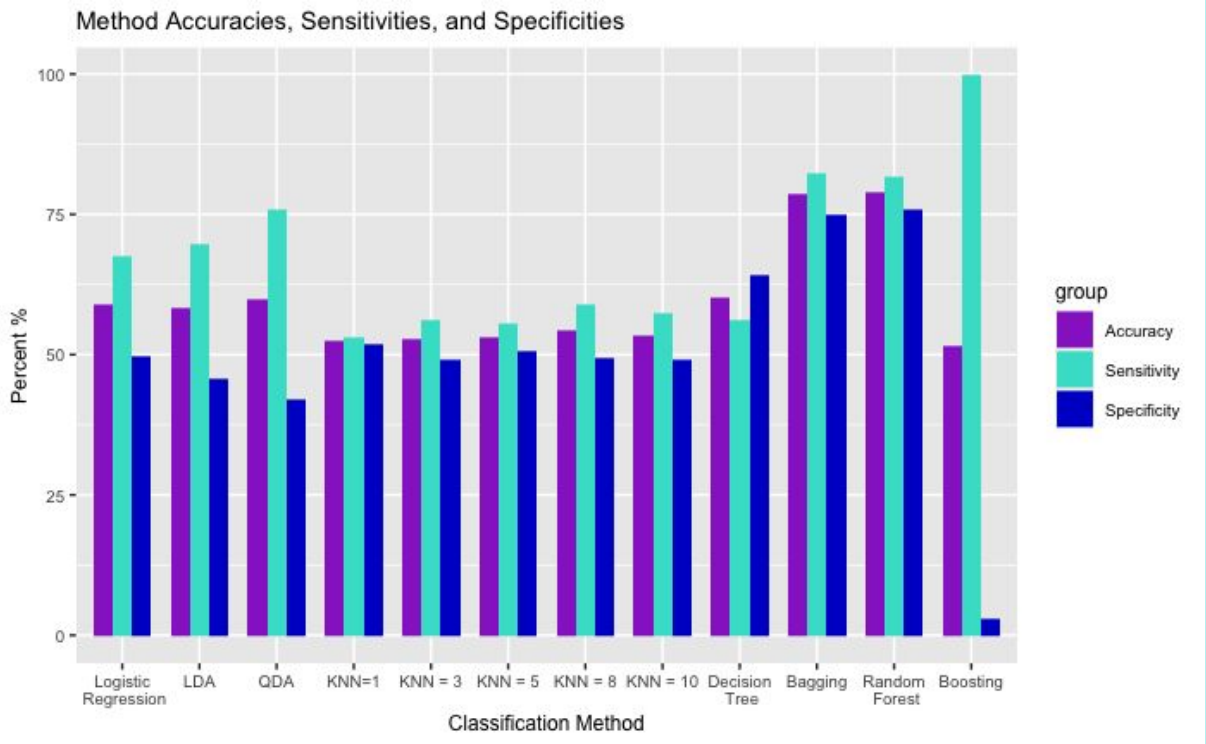


All Eight Classification Methods

- Popularity = β_0 + β_1Year + β_2BPM + β_3Energy + β_4Danceability + β_5Loudness + β_6Liveness + β_7Valence + β_8Duration + β_9Acousticness + β_{10}Speechiness + β_{11}Genre
- Model using all numeric predictor variables, including genre numbered 1-6
- Bagging and Random Forest have the highest Accuracy, Sensitivity, and Specificity, but the longest run times

	Accuracy	Error Rate	Sensitivity	Specificity	Running Time (in seconds)
Logistic Regression	0.591	0.409	0.677	0.497	0.028
LDA	0.583	0.417	0.698	0.457	0.053
QDA	0.598	0.402	0.760	0.421	0.026
KNN = 1	0.525	0.475	0.531	0.518	0.020
KNN = 3	0.528	0.472	0.562	0.491	0.022
KNN = 5	0.532	0.468	0.556	0.505	0.022
KNN = 8	0.545	0.455	0.590	0.495	0.022
KNN = 10	0.534	0.466	0.573	0.491	0.024
Decision Tree	0.602	0.398	0.563	0.641	0.065
Bagging	0.787	0.213	0.824	0.749	0.284
Random Forest	0.789	0.211	0.818	0.759	1.586
Boosting	0.516	0.484	0.998	0.030	0.285

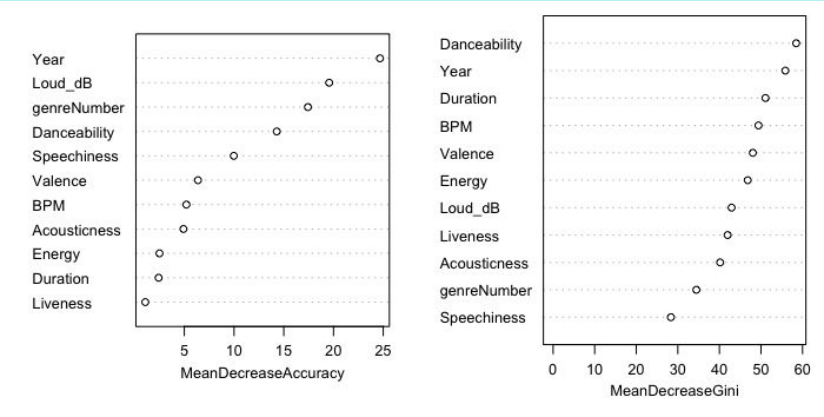
All Eight Classification Methods

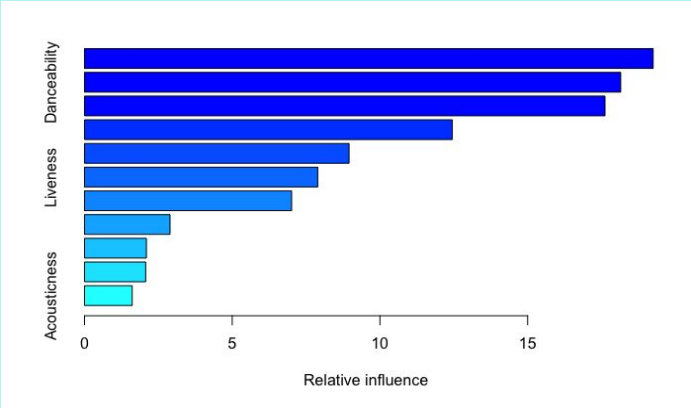


Most Important Variables

- Random Forest and Boosting methods show that Year, Danceability, and Loudness are important variables
- We may also want to consider Genre

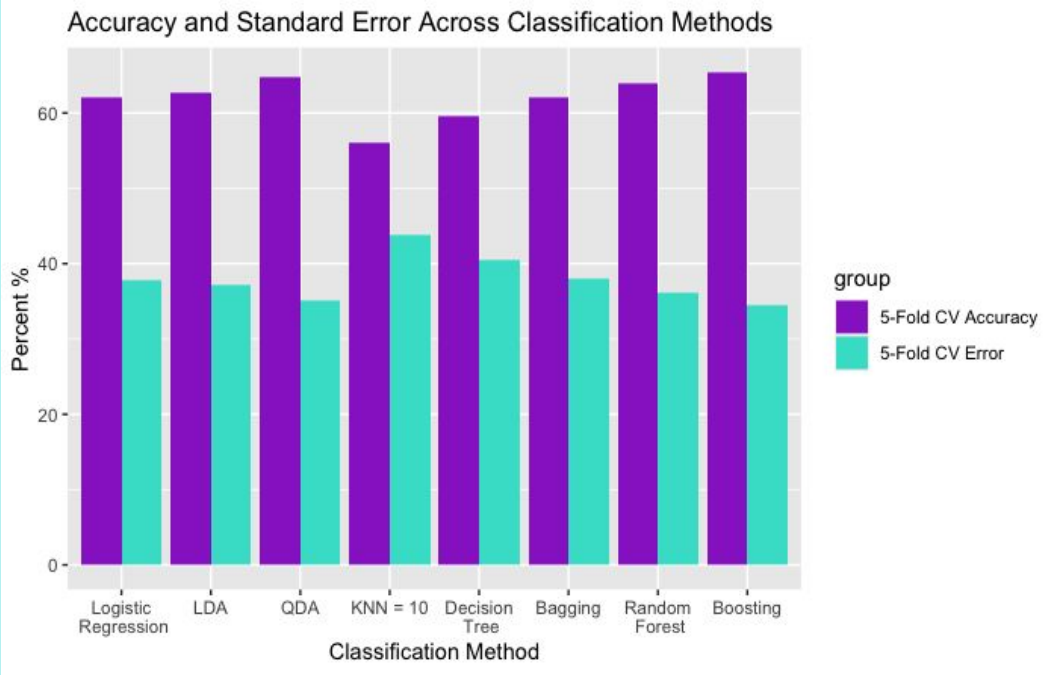
	var <chr>	rel.inf <dbl>
Year	Year	19.247870
Danceability	Danceability	18.150101
Loud_dB	Loud_dB	17.618374
genreNumber	genreNumber	12.449698
Duration	Duration	8.954744
Liveness	Liveness	7.894001
Speechiness	Speechiness	7.010718
BPM	BPM	2.894884
Valence	Valence	2.094760
Energy	Energy	2.071149
Acousticness	Acousticness	1.613701



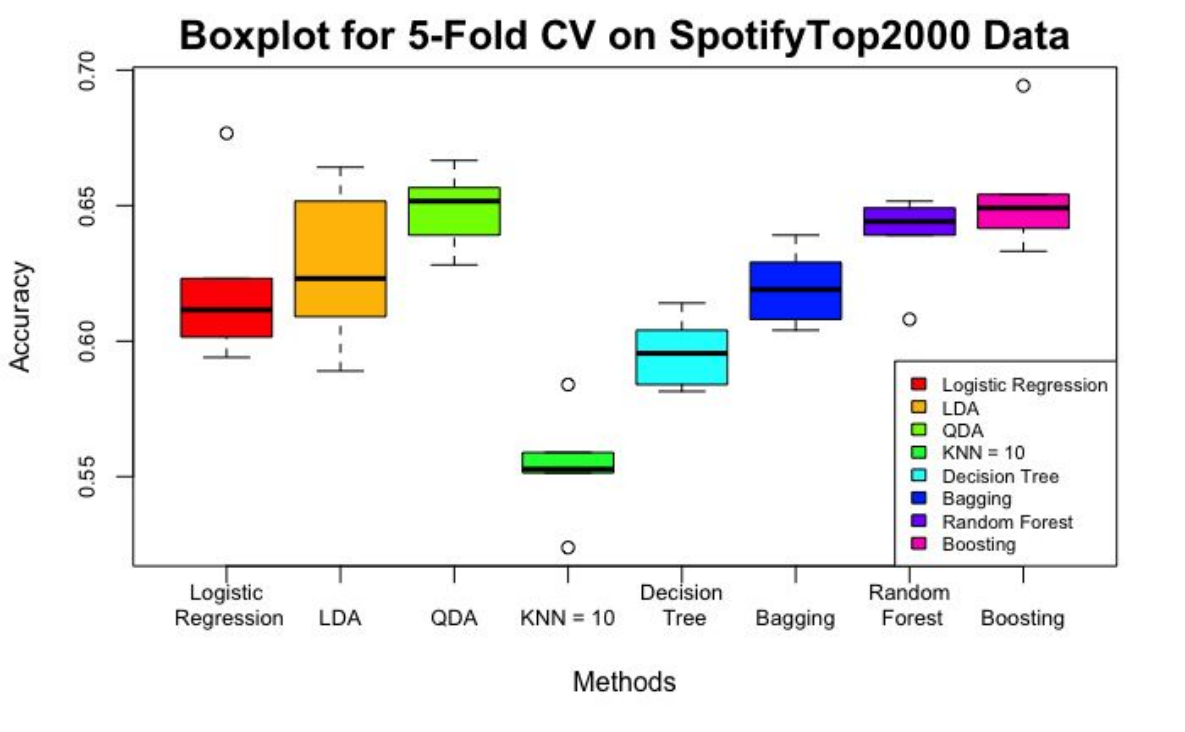


5-fold CV on All Eight Classifications - Accuracy and Test Error

- Same model with all numeric predictors
- QDA and Boosting methods produce the best accuracies
- Since Boosting is more time intensive we should choose the QDA method

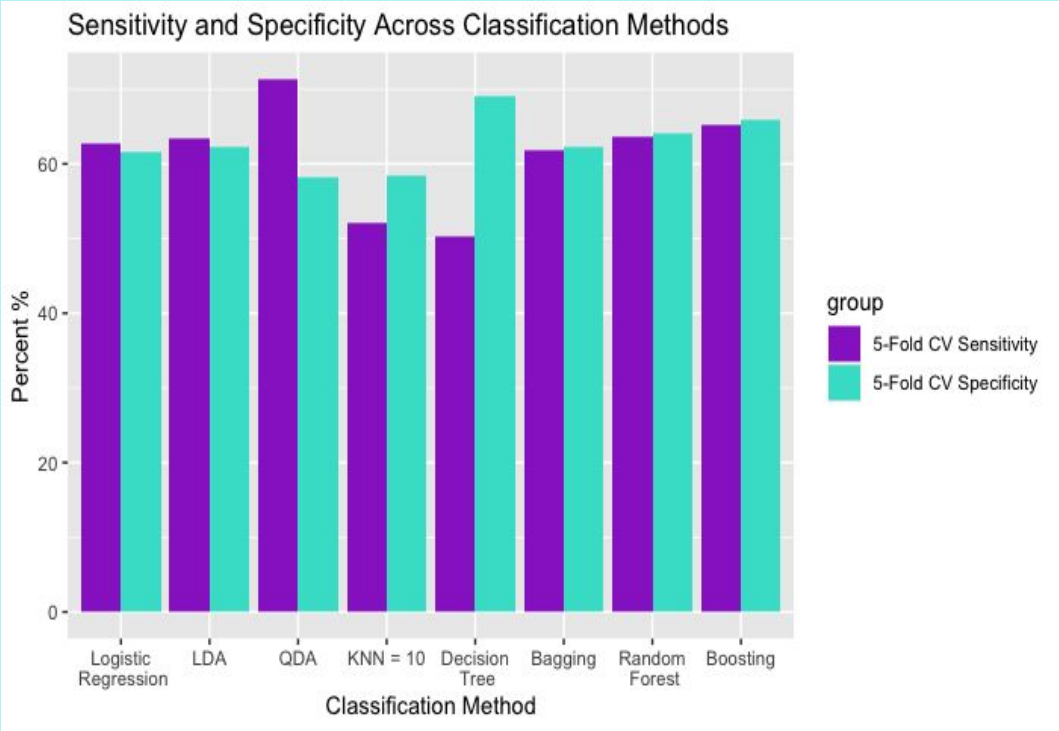


5-fold CV on All Eight Classifications - Boxplots of Accuracies



5-fold CV on All Eight Classifications - Sensitivity and Specificity

- Same model with all numeric predictors
- QDA method produces the best Sensitivity
- Decision Tree produces the best Specificity
- Boosting appears the best since Sensitivity and Specificity is about the same, but at increased run time
- May be better to choose LDA



Source

- https://www.kaggle.com/datasets/iamsumat/spotify-top-2000s-mega-dataset
- “This data is extracted from the Spotify playlist - Top 2000s on PlaylistMachinery(@plamere) using Selenium with Python. More specifically, it was scraped from http://sortyourmusic.playlistmachinery.com/.”