BILLBOARD HOT 100 ANALYSIS

WHAT MAKES A SONG POPULAR?

Laurie Cagney | Regis University | Summer 2021

INTRODUCTION

- Context
- The Dataset
- Data Prep
- Data Model
- Conclusion
- Appendix

CONTEXT

THE QUESTION

Can characteristics of a song be used to determine **when** the song appeared on the Billboard Hot 100 Chart?

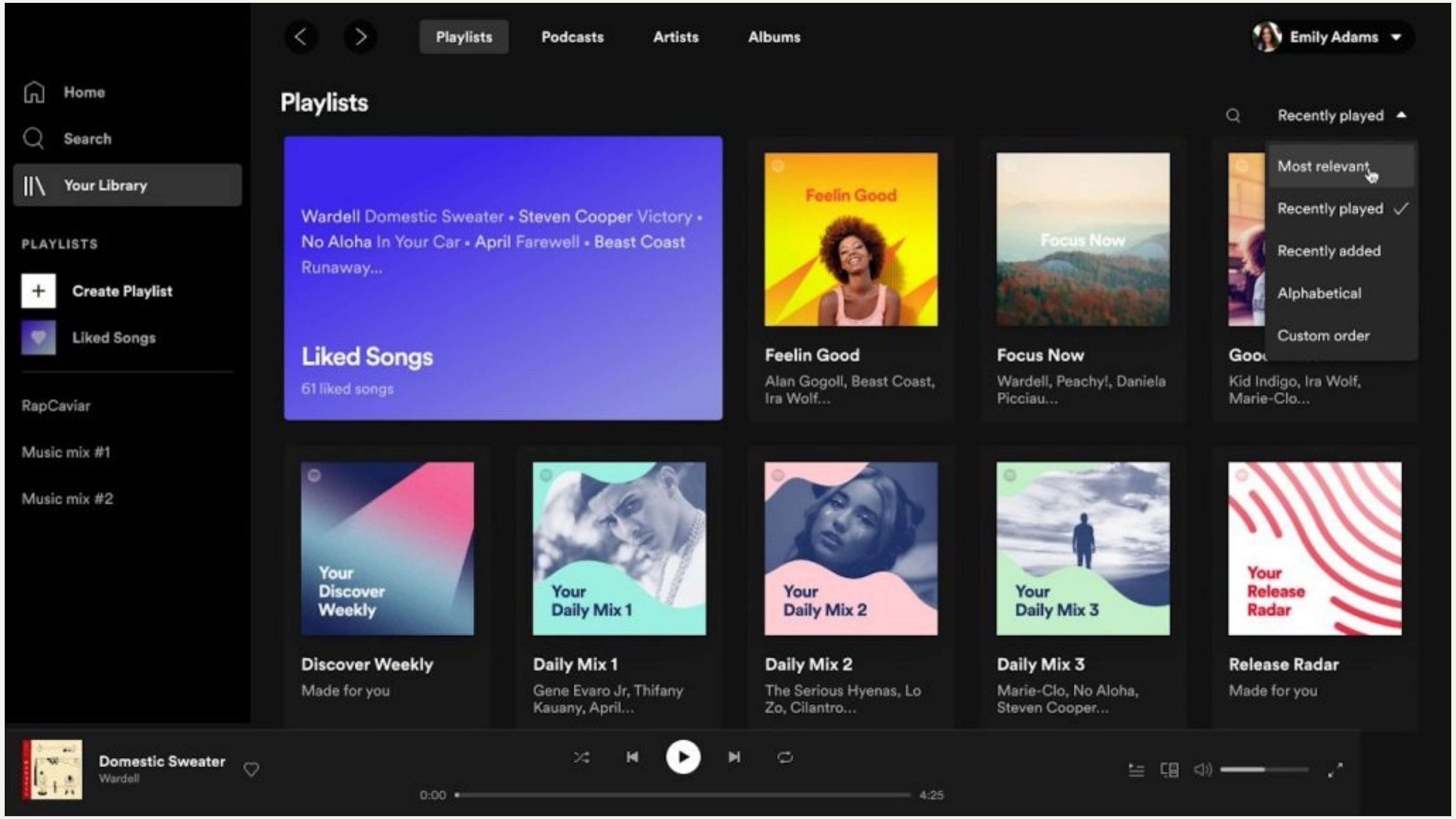
THE BILLBOARD HOT 100

billboard
HOT 100

	SONG	ARTIST
1	**Permission To Dance**	BTS
2	**good 4 u**	Olivia Rodrigo
3	**Stay**	The Kid LAROI & Justin Bieber
4	**Levitating**	Dua Lipa ft. DaBaby
5	**Kiss Me More**	Doja Cat ft. SZA
6	**Bad Habits**	Ed Sheeran
7	**Butter**	BTS
8	**Montero (Call Me By Your Name)**	Lil Nas X
9	**Save Your Tears**	The Weeknd & Ariana Grande
10	**deja vu**	Olivia Rodrigo

chart dated July 24, 2021

SONG CHARACTERISTICS FROM SPOTIFY



INTRODUCING THE DATASET

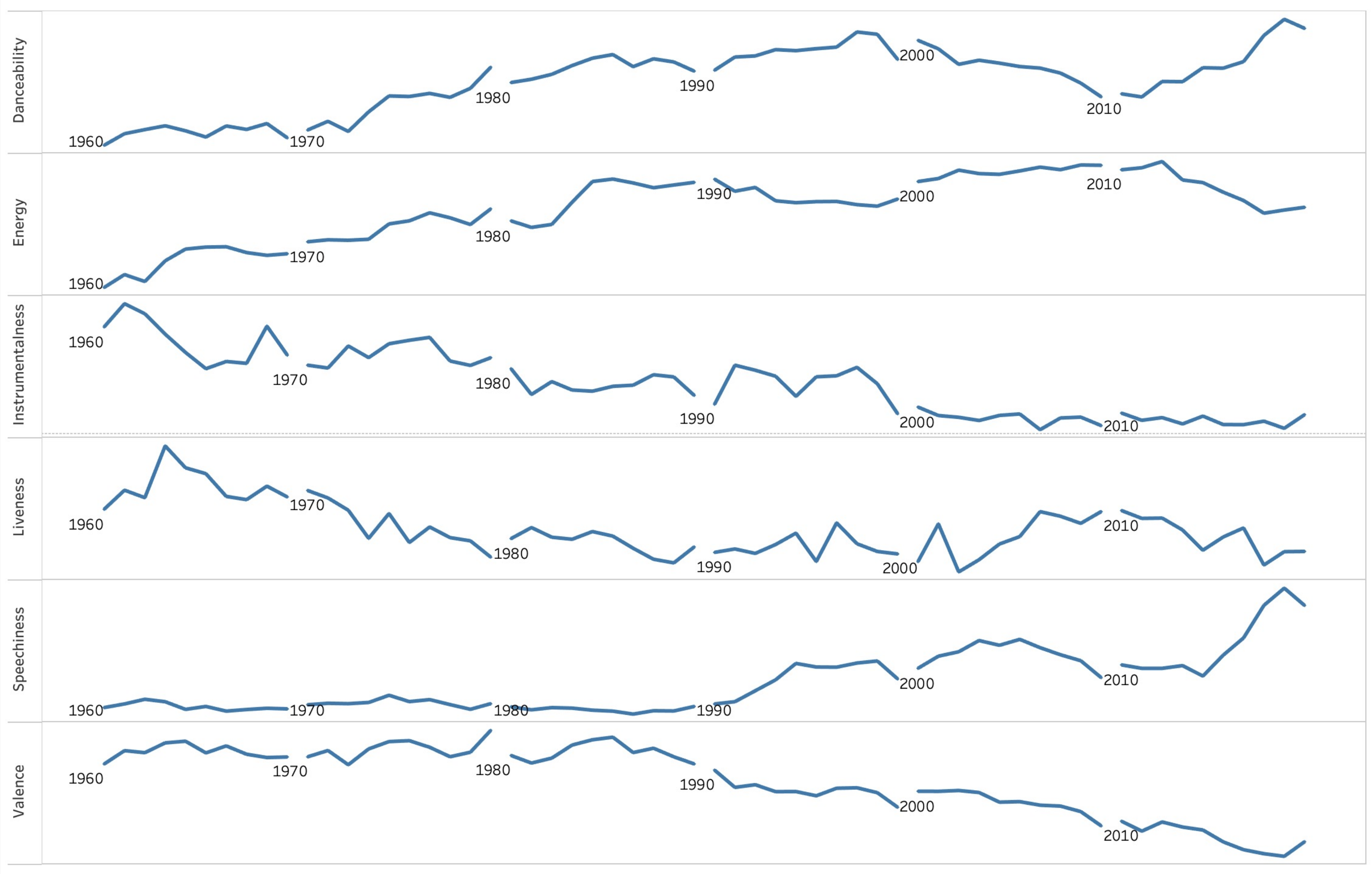
DATASET ORIGIN

Characteristics of how the public received it

+

Characteristics of it’s sound

MUSIC CHARACTERISTICS OVER TIME



DATA PREP

CLEANING THE DATA

Removed

- Highly correlated columns
- Empty Rows (no metadata)
- Columns where all the values were the same

Added

- Season (Fall, Spring, etc.)
- Isolated Year from Week value

Modified

- Min Max scaling rows
- Dummy Variables

THE MODEL

LGBM REGRESSOR

FINAL MODEL STATS

INPUTS

Learning Rate	0.1
Max_Depth	-1
N_Estimators	300
Num_Leaves	56

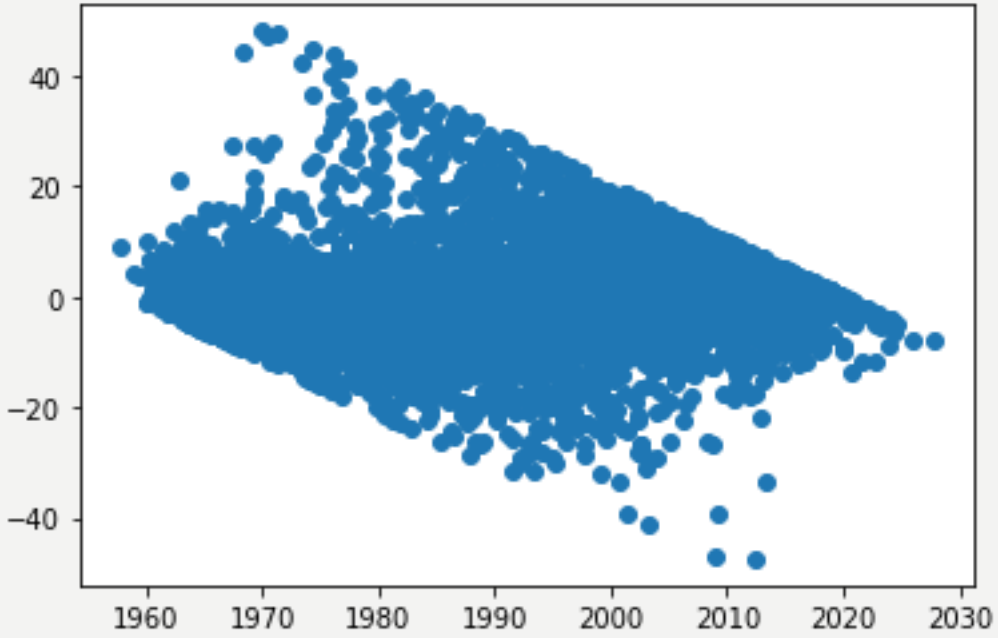
OUTPUTS

RMSE	9.0
MAE	6.4
r2	77.7

IMPORTANT VARIABLES

The song's **duration**, **speechiness**, and **loudness** were among the most important characteristics along with **how long it charted** and **how high.**

RESIDUAL PLOT



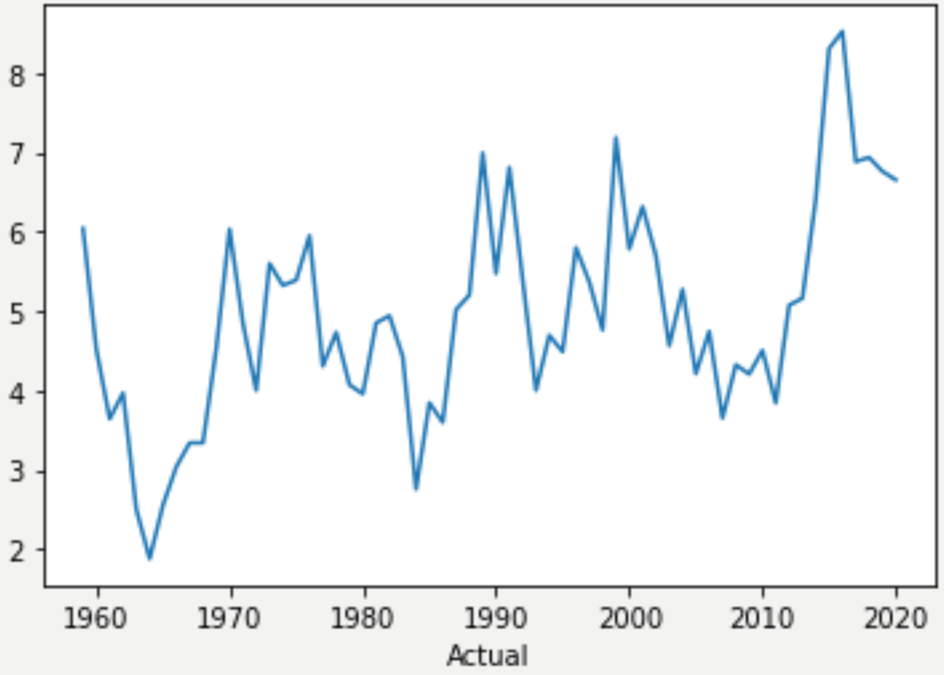
The linear pattern may be due to a missing variable

CONCLUSION

SO CAN I PREDICT IT?

- To a certain extent, if the song doesn't resemble songs from previous decades
  - Lots of music is derivative and sound is cyclical

Median Actual v. Predicted Delta Over Time



NEXT STEPS…

- Include genre once reducing cardinality, see if that addresses the residual plot
- Tag and remove song covers
- Look at it from a derivation standpoint – who sounds like who?

THANK YOU!

Laurie Cagney
Regis University
LCagney@regis.edu

APPENDIX

ALL THE EXTRA STUFF, ANNOTATED

LINKS

- Data.World Datasets: https://data.world/kcmillersean/billboard-hot-100-1958-2017
- GitHub: https://github.com/lcagney/MSDS-Practicum-2
- YouTube: https://youtu.be/Q3zYRNRokrc

SPOTIFY DEFINITIONS

Term	Definition
Acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
Instrumentalness	Predicts whether a track contains no vocals.
Key	The key the track is in.
Liveness	Detects the presence of an audience in the recording.
Loudness	The overall loudness of a track in decibels (dB).
Mode	Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.
Speechiness	Speechiness detects the presence of spoken words in a track.
Tempo	The overall estimated tempo of a track in beats per minute (BPM).
Valence	A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.

Obtained from https://developer.spotify.com/documentation/web-api/reference

FINAL DATASET VALUES

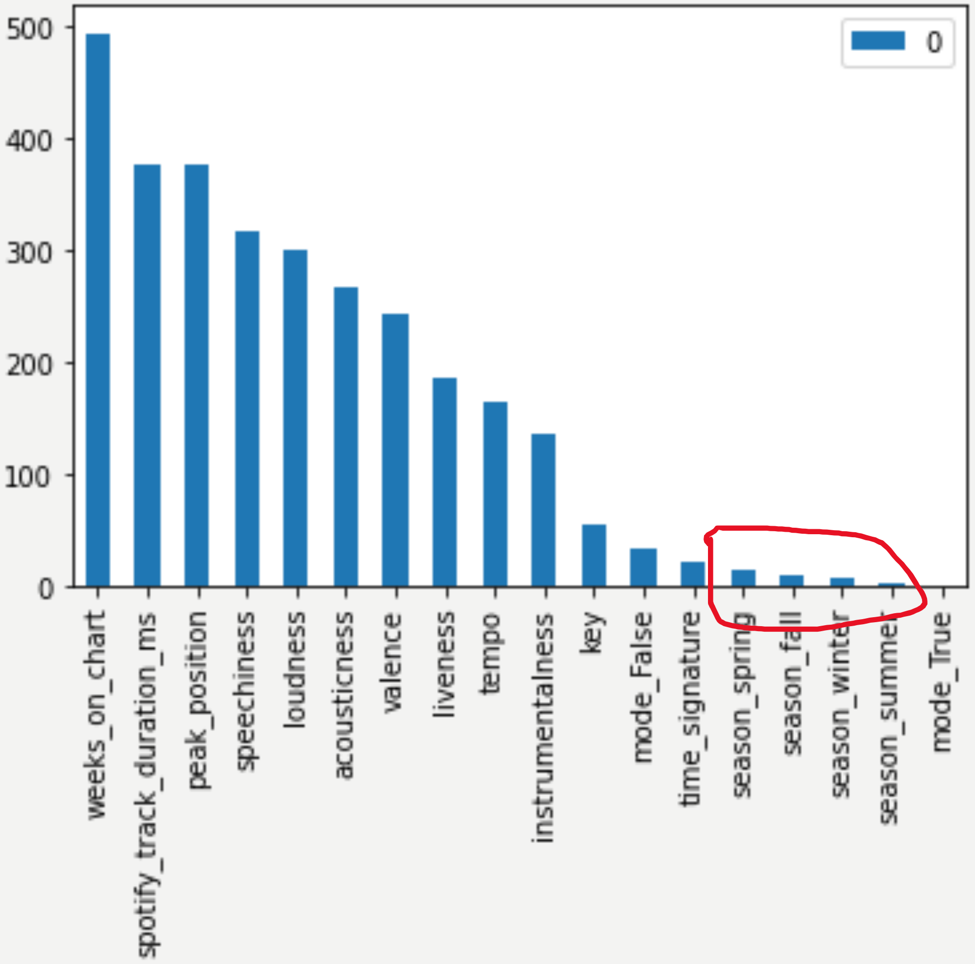
TARGET VARIABLE

- Year

PREDICTOR VARIABLES

- weeks_on_chart
- peak_position
- spotify_track_duration_ms
- acousticness
- loudness
- tempo
- time_signature
- key
- speechiness
- instrumentalness
- liveness
- valence
- mode_False
- mode_True
- season_fall
- season_spring
- season_summer
- season_winter

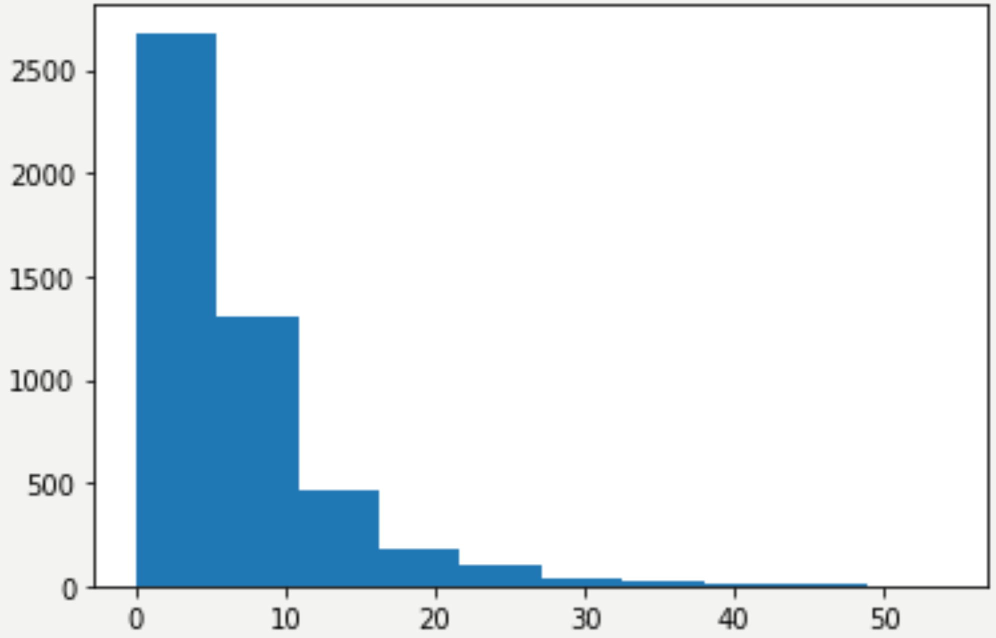
FULL IMPORTANT VARIABLES



The variables I created for the dataset added virtually no predictive value

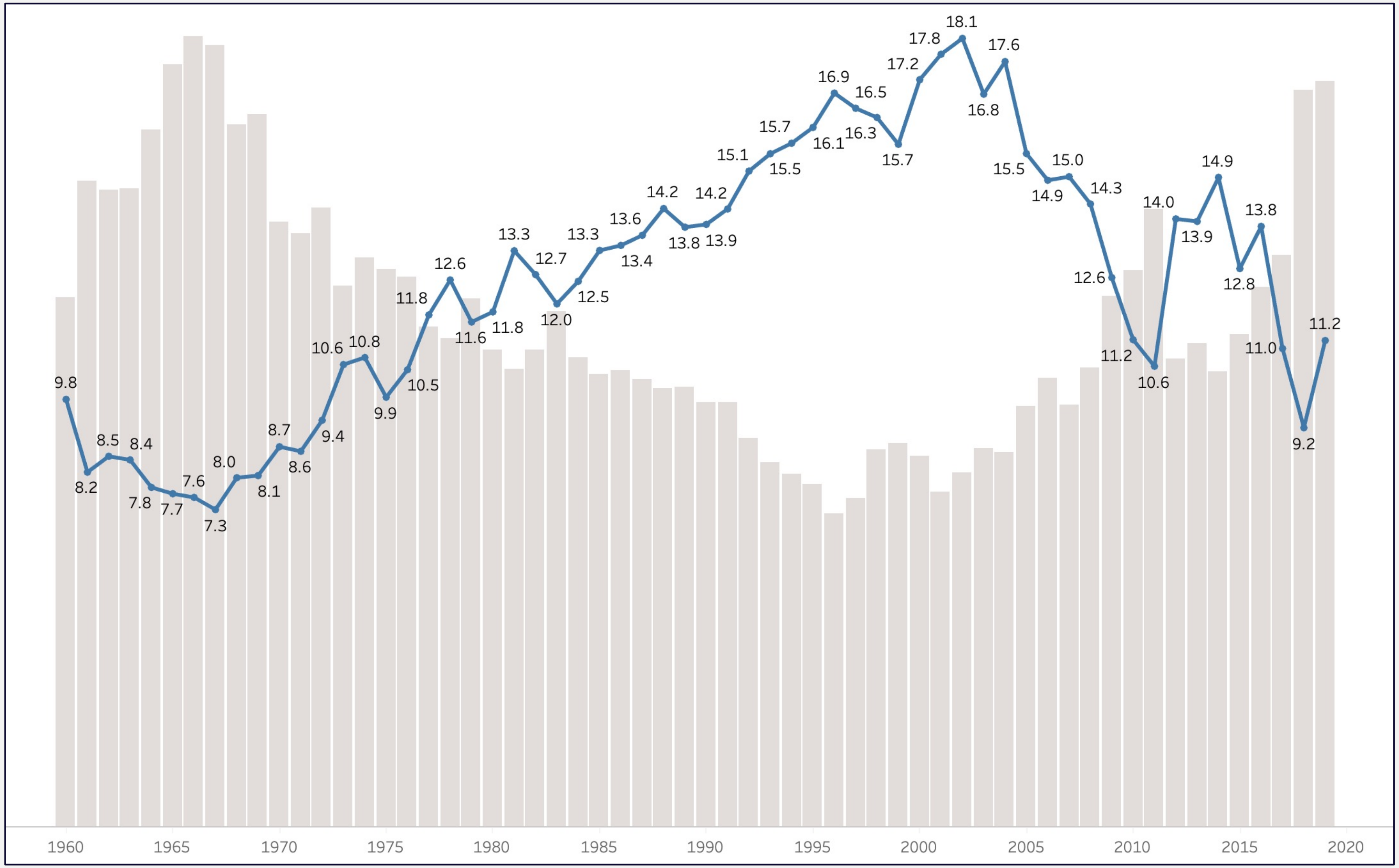
COMPARING ACTUAL VS. PREDICTED

Delta Between Actual & Predicted



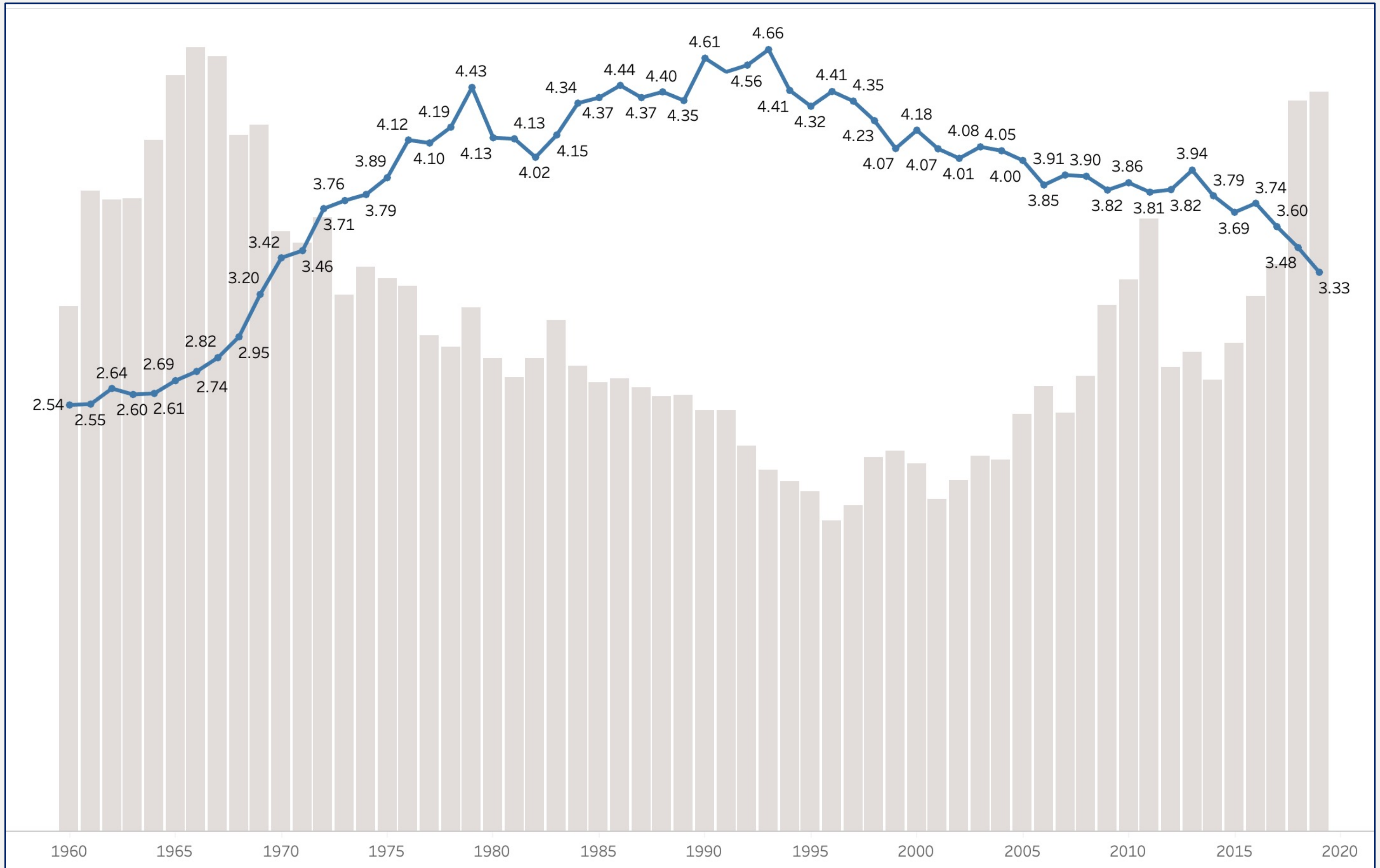
Comparing the differences between actual and predicted. More than half of the test dataset got the year correctly within a small variance; however, when the prediction was off – it was *really* off. My hypothesis is because of either covers of songs or because of throwback sounds.

INVERSE RELATIONSHIP BETWEEN LENGTH ON CHART AND NUMBER OF SONGS



Average length on chart on line chart and number of songs by year on bar plot

SIMILAR PATTERN WITH LENGTH OF SONG AND YEAR



Length of song on line chart and count of year on bar plot

LAZY PREDICT REGRESSION RESULTS

Model	Adjusted R-Squared	R-Squared	RMSE	\
LGBMRegressor	0.77	0.77	9.09	
HistGradientBoostingRegressor	0.77	0.77	9.10	
XGBRegressor	0.77	0.77	9.14	
ExtraTreesRegressor	0.75	0.75	9.49	
RandomForestRegressor	0.75	0.75	9.55	
GradientBoostingRegressor	0.74	0.74	9.72	
BaggingRegressor	0.73	0.73	9.97	
SVR	0.65	0.65	11.22	
NuSVR	0.65	0.65	11.23	
KNeighborsRegressor	0.62	0.62	11.73	
LinearRegression	0.56	0.56	12.65	
TransformedTargetRegressor	0.56	0.56	12.65	
PoissonRegressor	0.56	0.56	12.66	
LassoLarsIC	0.56	0.56	12.66	
LassoCV	0.56	0.56	12.66	
ElasticNetCV	0.56	0.56	12.66	
BayesianRidge	0.56	0.56	12.66	
RidgeCV	0.56	0.56	12.66	
Ridge	0.56	0.56	12.66	
LarsCV	0.56	0.56	12.66	
LassoLarsCV	0.56	0.56	12.66	
Lars	0.56	0.56	12.66	
SGDRegressor	0.56	0.56	12.66	
HuberRegressor	0.55	0.56	12.71	
Lasso	0.53	0.53	13.06	
LinearSVR	0.53	0.53	13.09	
OrthogonalMatchingPursuitCV	0.51	0.51	13.35	
ElasticNet	0.50	0.50	13.54	

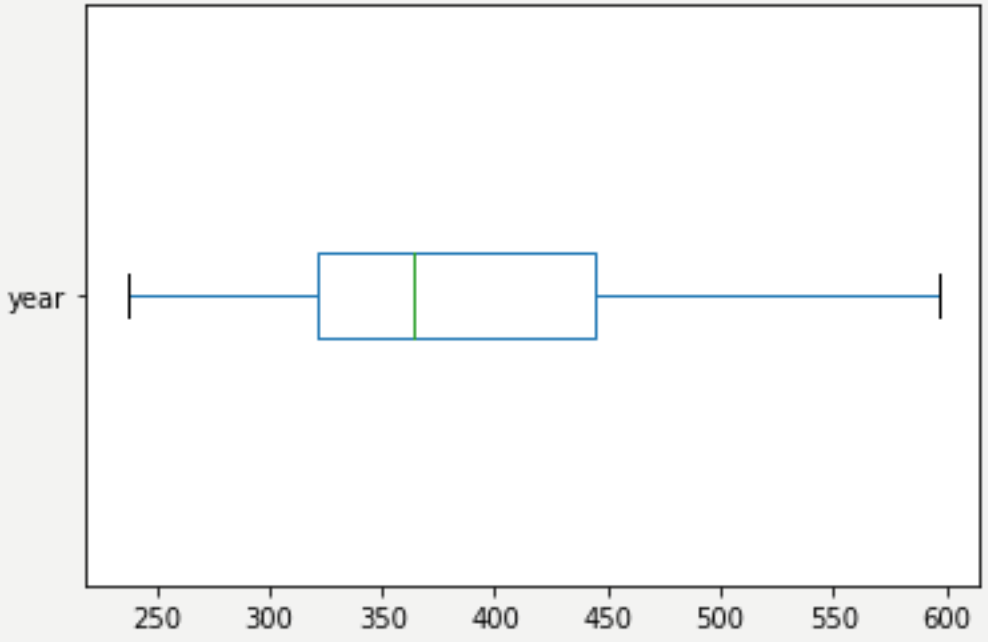
Could have really gone with any of the top 3

WORD CLOUD- TOP ARTISTS SINCE 1960

Glee ruled the charts when it was releasing their covers in the early 2010s

SONG COUNT DISTRIBUTION BY YEAR

MEDIAN IS AROUND 360 SONGS



Slightly right skewed, with some years charting an abnormally large amount of songs.

SQL STATEMENT TO COMBINE DATASET FROM DATA.WORLD

```
SELECT a.weekid as weekid,week_position,a.song as song,a.performer as performer,weeks_on_chart,peak_position
,spotify_track_duration_ms,spotify_track_explicit,danceability,energy,c.key,loudness,mode,speechiness
,acousticness,instrumentalness,liveness,valence,tempo,time_signature
FROM hot_stuff_2 as a
join(
        SELECT max(weekid) as weekid, songid
        FROM hot_stuff_2
        GROUP BY songid) as b
on b.weekid = a.weekid and b.songid = a.songid
join hot_100_audio_features as c
on c.songid = a.songid
```