Call notes

Use the playlist generation as a test of your model

1. Use NLP on lyrics
2. PCA on audio

The distribution of the number of followers is like an extreme power law or exponential distribution

A lot of concentration in playlists that are not super followed and then mega-follower playlists are a few

The median is skewed

Going to have to make decisions as to how do you define a followed playlist

Maybe define what a reasonably followed playlist would be

There’s a lot of times you can run models that give you feature importance, that is useful, take those features and fit them to models that fit your data better and the procedure better

Do tree based models!

That gives you feature importance

Then kind of do PCA thing and take from here up

Foolproof substantiated good method of doing that

Don’t run trees really deep

I fyou have a categorically rich dataset, the dimensionality will expand very quickly with dummy vars

A way around this is that you can factorize

Look up python factorize categorical variables

Think about noise

Think about variance bias tradeoff

Think about error, how you define your error, maybe don’t do sum of squared errors

Build a github repo

Make an account with github pages

Please remember for EDA to label everything, put titles, make it easy

Remember that presentations should be agnostic of the time that your audience is taking, make good captions, put an arrow when you need them