**I intro**

. my project was an effort ot try and predict resale prices for vintage vinyl records – specifally jazz records.

. My goal today is to walk you though my approach intrying to understand vintage vinyl rec prices better and see if we can find some important factors related to resale prices

**slide 2**

. more specifically i'm going to walk through:

. my motivation for this project and why i'ts important

. details around obtaining data to support the analyis

. general exploratory findings and then our model selection

. finally results of the analyiss

**slide 3**

. what's the motivation behind this analysis

Business :slide

. emphasize market growth - units sold and price (mostly new)

. this only captures new albums

. some research has shown used album sales are much greater than new. It’s just not reported and recorded sas often given record companies don’t benefit from them

. just to throw out a few more numbers

**Slide 4**

Personsal

. i started ollecting records wihtin the past 2-3 years so i'm still a newbie.

. aside from the enjoyment of analog sound or the enjoyment of discoverying a hidden gem from the past

. i noteiced discripancy in prices

. some intuiitons tells us why theres a differenc but i sought to find some answers

**Slide 5**

GOAL:

help buyers and sellers figure out a good target resale price for the ablum considered

what are some important features that impact the resale price

**Data collection Feature**

. to obtain data I went to the website discogs.com which in addition to other things is mostly an online marketplace for buyers and sellers of records – used and new

. Discogs lists 5.7 million used vinyl items in its marketplace from U.S. sellers (Discogs doesn't distinguish between new and used; it only lists items by condition. The 5.7 million figure doesn't count a million items rated as "Mint"

. page – list of album with a number of features

. price, label, seller ratings, user ratings, have, wants, et

. dig further you’ll find notes

. keywords such as first pressing, original or rare or deep groove can be extracted futher

. lay out a few conditions I used upfront

. jazz recoreds only

. between 1950-1970 (prime years of jazz)

. price $20+ (based on domain knowl peope in this space are looking for records on the higher end

**Slide 6 (key feature)**

. I list the majority of the features pulled and highlight a few that I thought were important from the start

. media gread – basically is a scale of 1-8 Mint is the best, and P is the worst

. Notes extract

. roughly 8000 records in total

**Slide 7 (EDA)**

. we can move on to a few desctiptives to get an idea of what the data looks like

. our outcome, price is heavily skewed to the right

. unsurprising and surprising at the same time

. majority of recs jazz are on the lower end – ad you can see how the avg is pulled up by many higher priced albums (mean vs. median)

**Slide 8 (EDA)**

**.** pair plots. Noticed some possible collinearity between variables.

. makes sense that the number. Of people that have may give ratings for the records so I had to think of a way. To deal.

**Slide 9**

**.** despite some of what I saw in EDA I decided to run an OLS regression

. highglighted a few variables that seemed to be more noise than anything

. x notes – that was my attempt ot create a binary flag on any notes that contained keywrods

. recoyear – intueitvly makes sense the trend

. haves – collinear

. seller rating – almost always near 100%

**Slied 10-11**

. shows the change in features prior to modeling

**Slide 12**

. so once we had that I went through a number of models to test and compare for model selection

**Slide 13**

. I tried a good number of models but I’m only showing 4 that were most impactful

. the baseline mode was simlpl OLS R^@

.

**Slide 14**

. modeling methods / outcomes

. tried a few bt without a ton of domain knowledge I decided to jump in and throw the kitchen sink into the model

. I applied interaction terms , polynomials with degree of 2, standardization and ran through Linear reg, Lasso and Ridge

Too complex

Overfitting