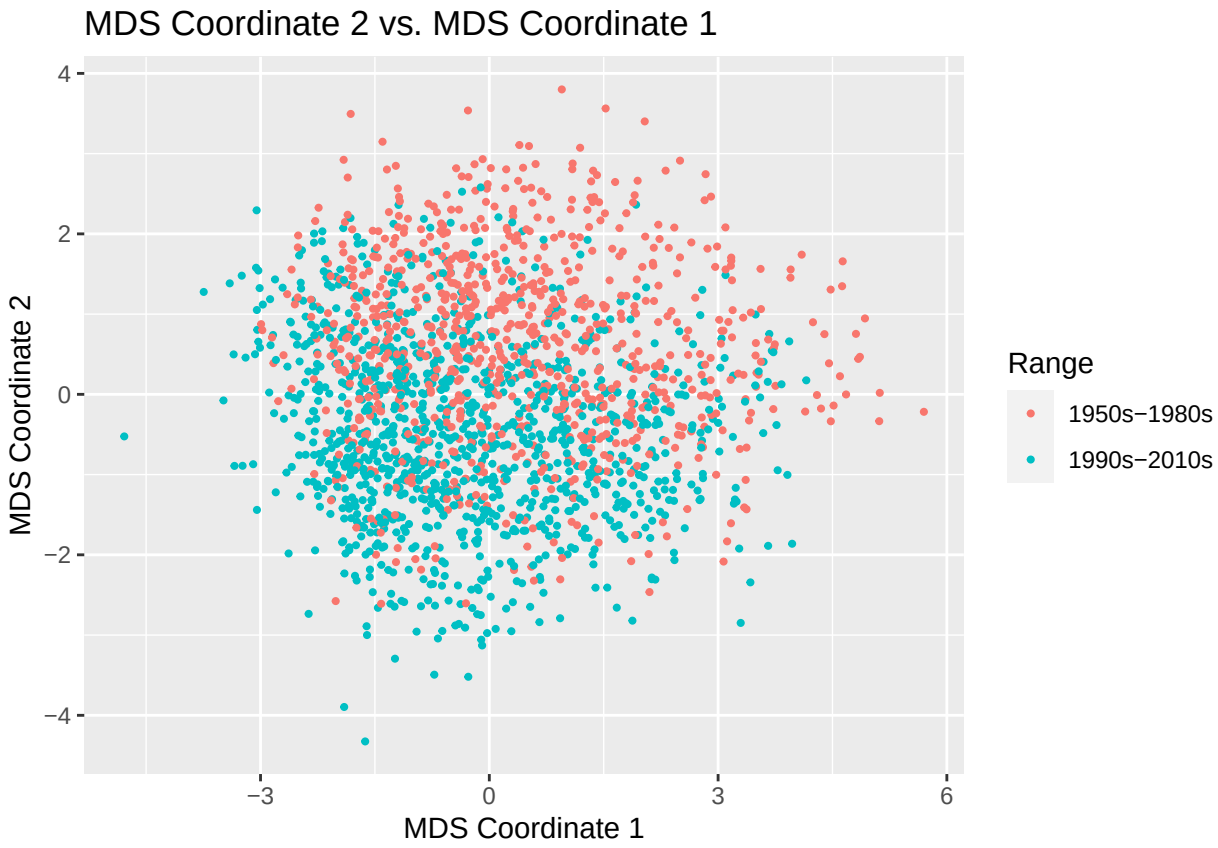
35-315 Final Project Report

Kyra Balenzano, Evan Feder, David Yuan

5/2/2020

Question 3: Time



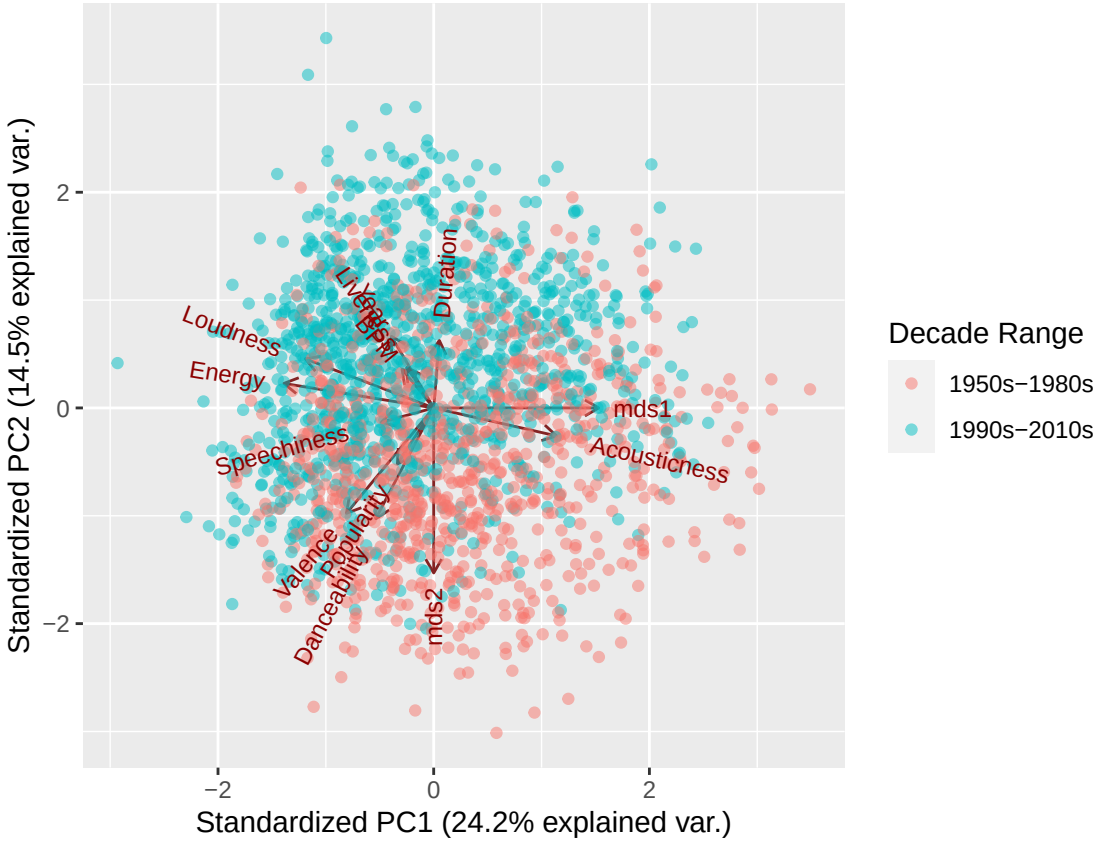
(Probably the one I would vote to go first due to the nuance of analysis needed)

Since we have so many quantitative variables, we first conducted principal component

analysis (PCA). We then made a graph plotting the first two components, and colored our datapoints by the Decade Range variable so that we could make some comparisons regarding time without clouding the graph with too many overlapping colors.

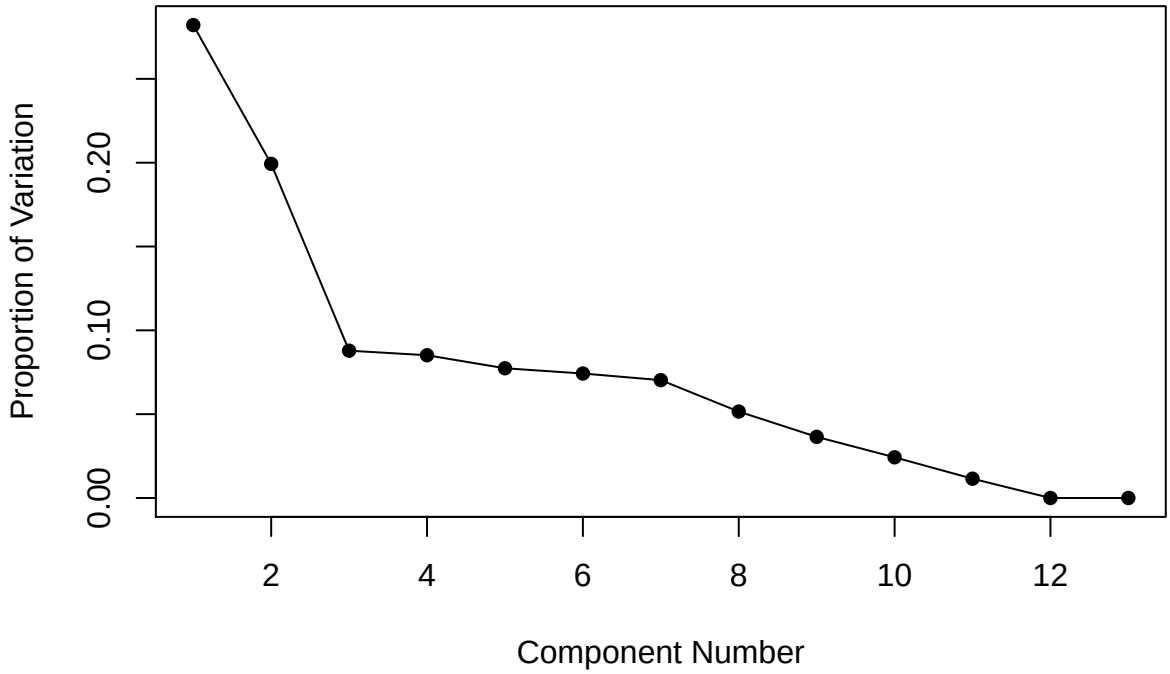
```
## [1] 1994   13
```

```
##          PC1          PC2          PC3          PC4          PC5          PC6
## 1.914899e+00 1.609519e+00 1.068740e+00 1.052050e+00 1.002986e+00 9.823782e-01
##          PC7          PC8          PC9         PC10         PC11         PC12
## 9.558305e-01 8.185955e-01 6.884235e-01 5.617186e-01 3.864984e-01 2.202649e-15
##         PC13
## 1.276801e-15
```

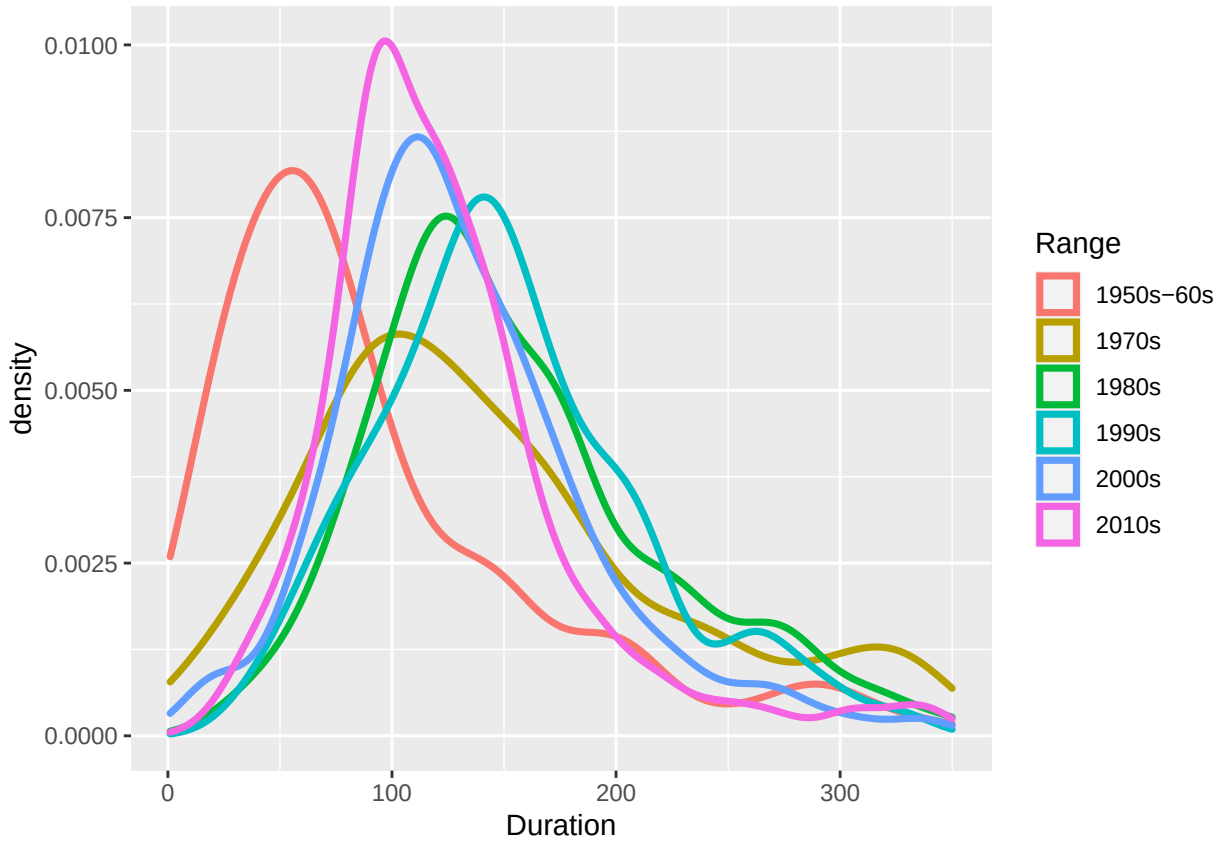


We can see that Decade Range slightly clusters by the first two components, since there are mostly blue datapoints on the top and mostly red datapoints on the bottom. One

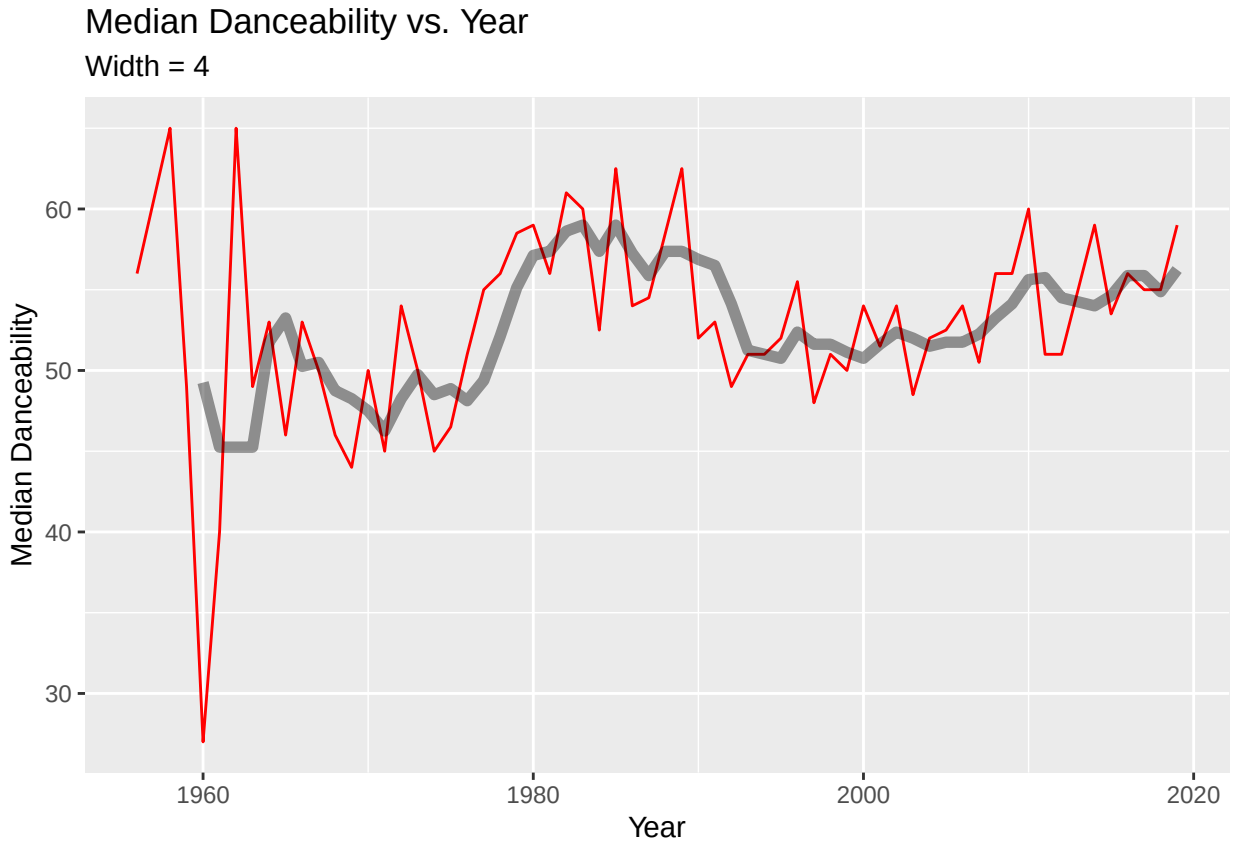
example of a conclusion that can be drawn from the graph is that as BPM increases, both PC1 decreases and PC2 increases, which further allows us to conclude that songs from the 1990s-2010s tend to have a greater number of beats per minute than songs from the 1950s-1980s. That being said, since the Normal distribution ellipses overlap quite a bit, there is not enough evidence to conclude that the two groups are significantly different with respect to their principal components.

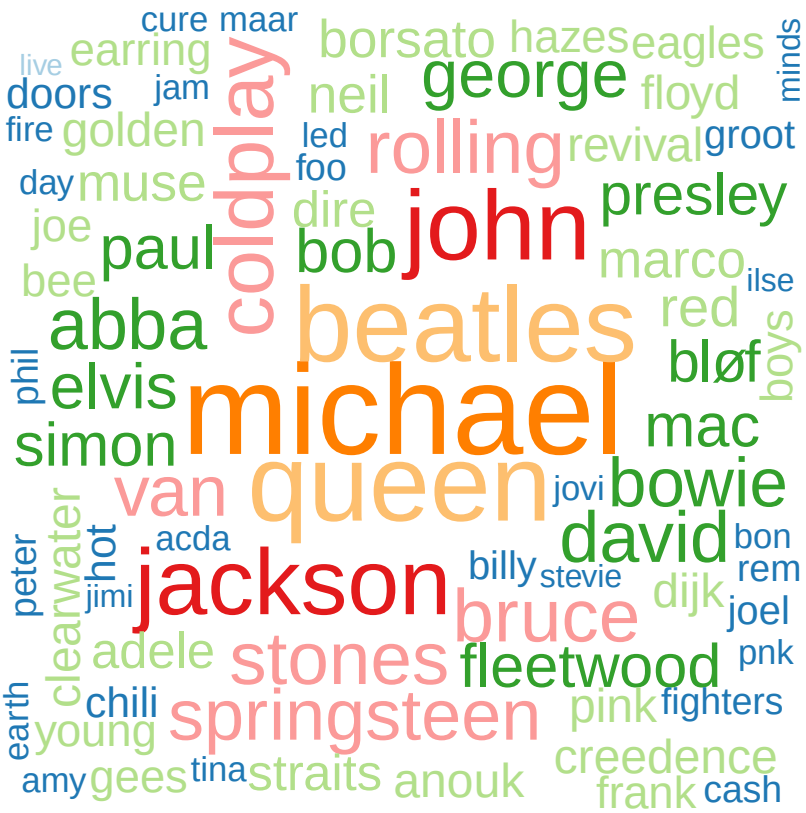


Notes: Lecture 18 R Demo – arrow plot thing, making ellipses around the clusters. PCA – arrows. Note that scree plot suggests use of third dimension but we did not use. Use statistical analyses – t tests in terms of x variable.



Interpretation: song duration has changed throughout the decades. All are right skewed but 1950s-1960s seemed to have the smallest duration mode, followed by 1970s, 2010s, 2000s, 1980s, and then 1990s. Seemed to have cycled around.


Median Danceability vs. Year
Width = 4
Median Danceability
60
50
40
30
1960
1980
2000
2020
Year


cure
maar
earring
live
borsato
hazes
eagles
minds
doors
jam
neil
george
floyd
fire
golden
led
rolling
revival
groot
day
muse
foo
presley
joe
dire
john
paul
bob
marco
ilse
bee
red
abba
coldplay
beatles
bløf
boys
phil
elvis
michael
mac
simon
queen
van
jovi
bowie
david
bon
peter
clearwater
hot
acda
jackson
billy
stevie
rem
jimi
dijk
bruce
joel
adele
stones
fleetwood
pnk
earth
chili
springsteen
pink
fighters
young
creedence
amy
gees
tina
straits
anouk
frank
cash

Note limitations!!

Conclusions

Future work: adding and interpreting a third dimension to the PCA plot. Adding the third dimension has been learned, but not yet interpreting. Explore every quantitative variable in time analysis individually.