



SPA
DRP

Cluster Analysis

Renee Chien

Mentors: Daniel Suen, David Marciano

Fall 2021

Table of contents

01

Clustering
Overview,
K-means

03

Results Analysis +
Discussion

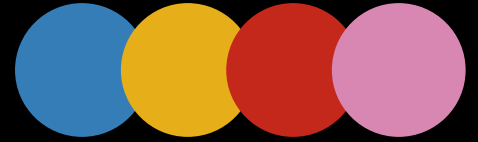
02

Data Set +
Features

04

Conclusions +
Takeaways

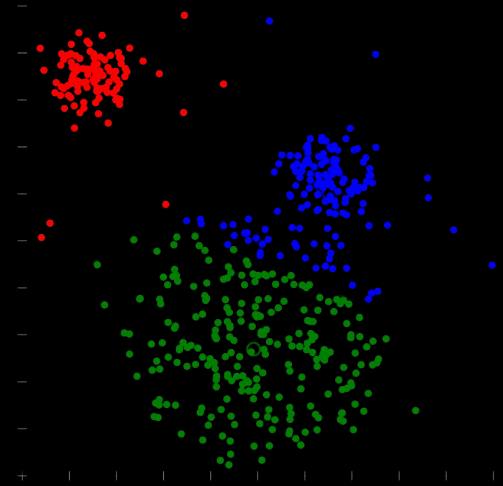
Clustering Overview

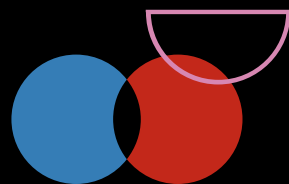
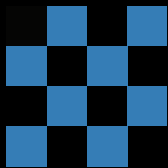


- **A Way to Find Subgroups within a Data Set**

- **Two Methods Discussed in ISL**

K-Means and Hierarchical

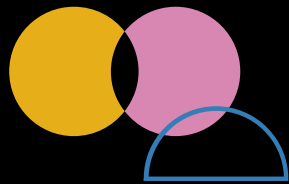


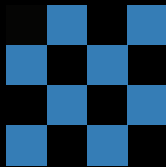


K-Means Clustering

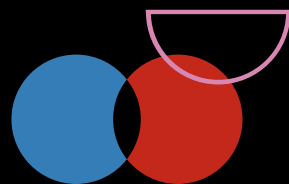
Dissimilarity measure: Euclidean distance

$$W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2,$$
$$\text{minimize}_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}.$$



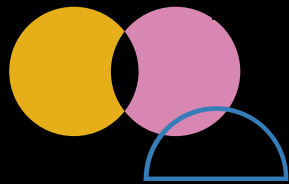


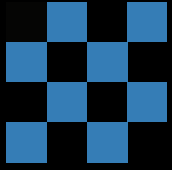
K-Means Clustering



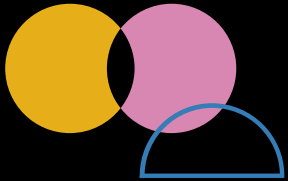
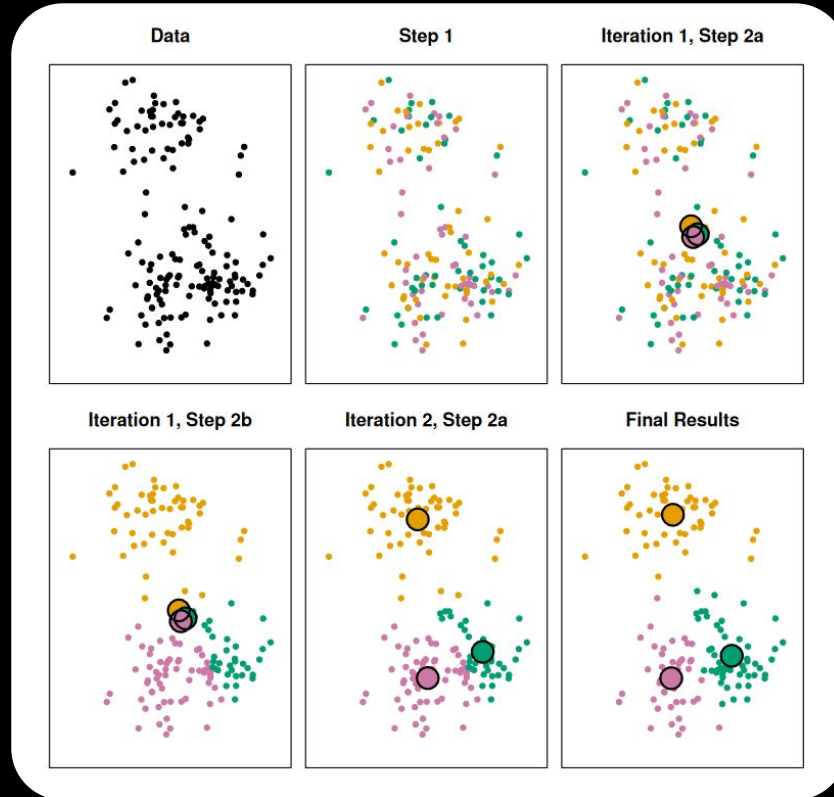
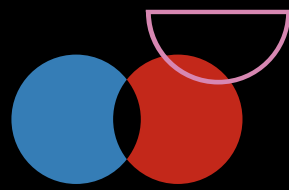
Algorithm 12.2 *K*-Means Clustering

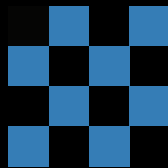
1. Randomly assign a number, from 1 to K , to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
 - (a) For each of the K clusters, compute the cluster *centroid*. The k th cluster centroid is the vector of the p feature means for the observations in the k th cluster.
 - (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



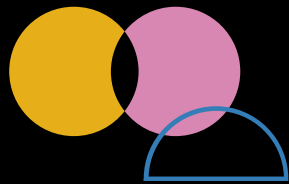
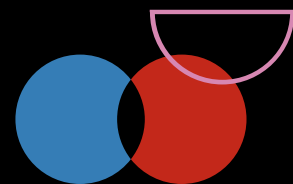


Visualization of K-Means





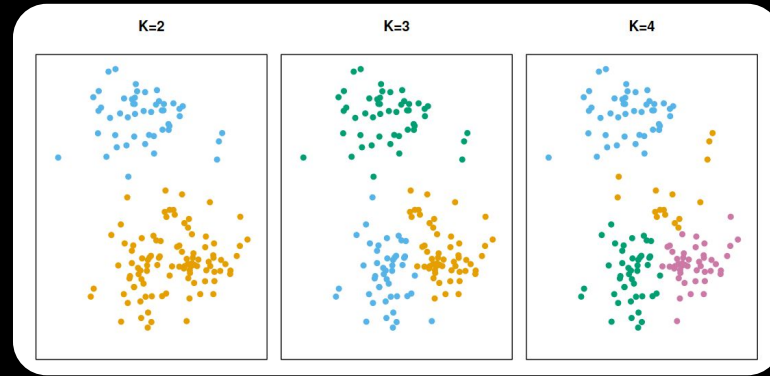
Different Runs of K-Means



Challenges with K-Means/Clustering

K-means

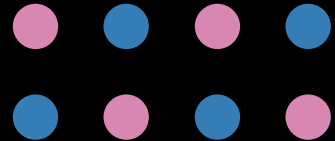
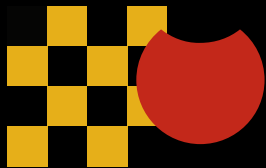
“How many clusters?”



Clustering in General

Potential “outliers” that don’t truly belong in any cluster

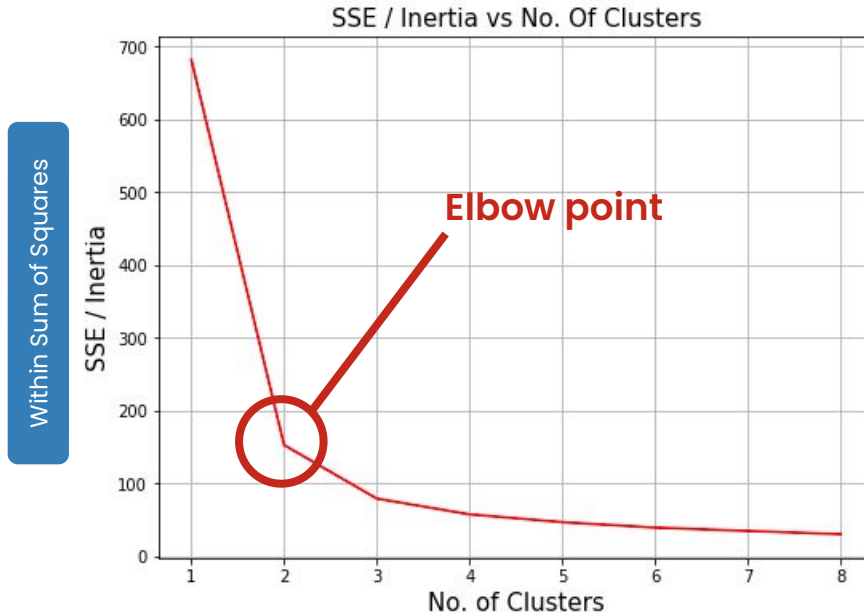
Perturbations in Data – changes in the set effect clusters drastically



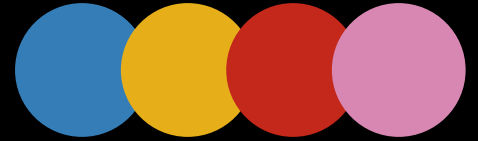
Elbow Method




For deciding the number of clusters to use for k-means clustering

```
In [10]: drawSSEPlotForKMeans(df, [0, 1, 2, 3])
```



K-means Simulation

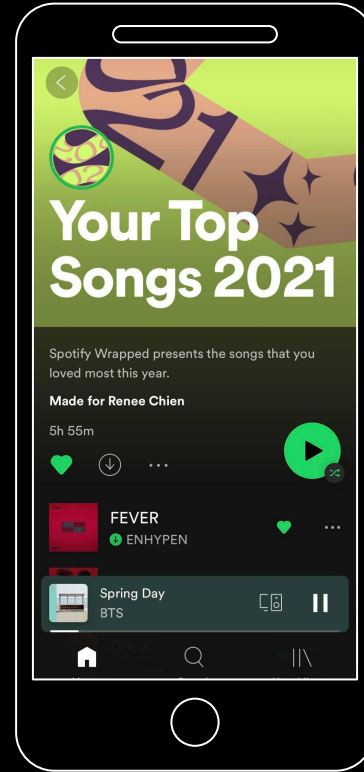


```
12 ▾ ##### K-means Clustering with K=2
13 We begin with a simple simulated example in which there truly are two clusters in the
    data: the first 25 observations have a mean shift relative to the next 25 observations.
14
15 ▾ {r Two Clusters}   
16 set.seed(2)
17 x <- matrix(rnorm(50 * 2), ncol = 2)
18 x[1:25, 1] <- x[1:25, 1] + 3
19 x[1:25, 2] <- x[1:25, 2] - 4
20
21 # We now perform K-means clustering with K = 2.
22 km.out <- kmeans(x, 2, nstart = 20)
23
24 # The cluster assignments of the 50 observations are contained in km.out$cluster
25 km.out$cluster
26
27
28 #The K-means clustering perfectly separated the observations into two clusters even
    though we did not supply any group information to kmeans(). We can plot the data, with
    each observation colored according to its cluster assignment.
29 #par(mfrow = c(1, 2))
30 plot(x, col = (km.out$cluster + 1),
31 main = "K-Means Clustering Results with K = 2",
32 xlab = "", ylab = "", pch = 20, cex = 2)
33 ▸
```



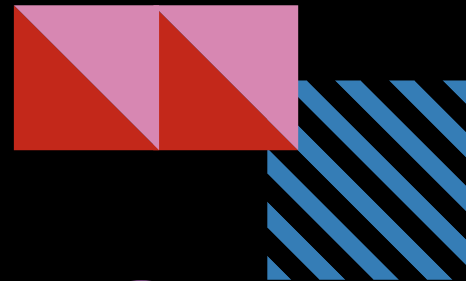
My Data Set

My Spotify Wrapped playlist:
My 100 most-played songs of 2021



Track Features

As described by the Spotify API



danceability

Based on tempo, rhythm stability, beat strength, & overall regularity

energy

perceptual measure of intensity and activity— *fast, loud, and noisy?*

valence

musical “positiveness” conveyed by a track

speechiness

Presence of spoken words?

acousticness

Is it acoustic?

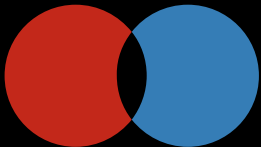
loudness

overall loudness of a track

tempo

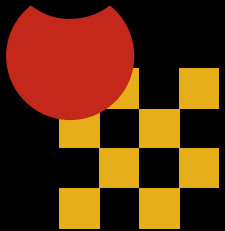
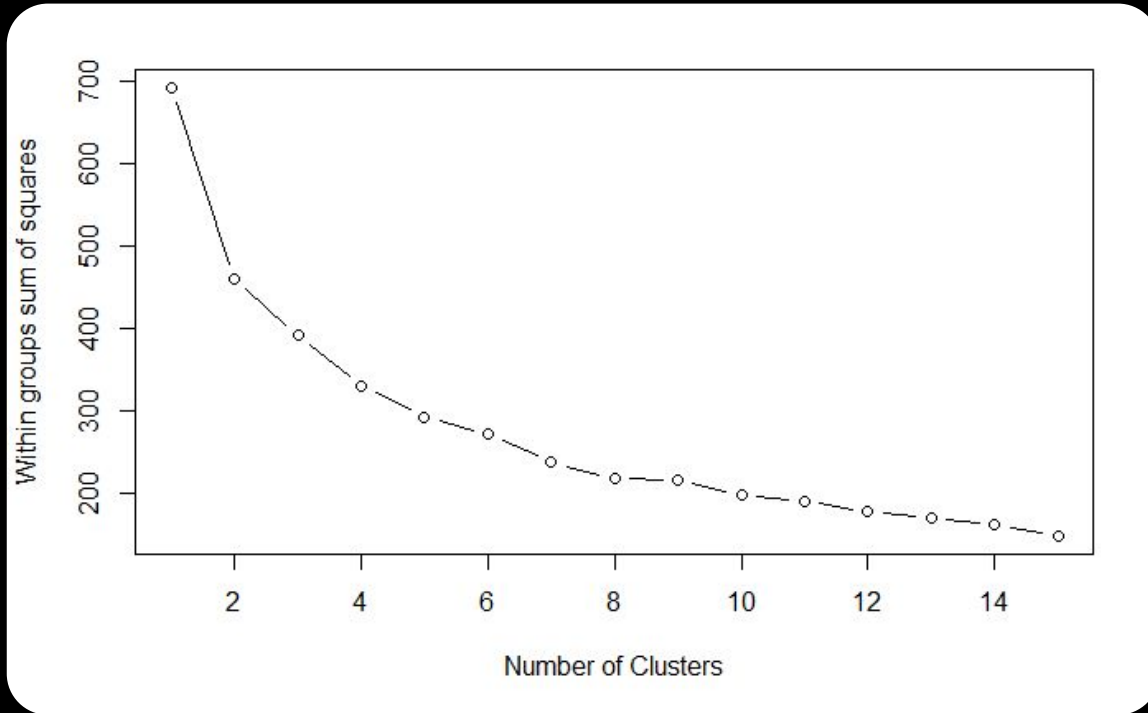
BPM

**All features were normalized through the scale function*



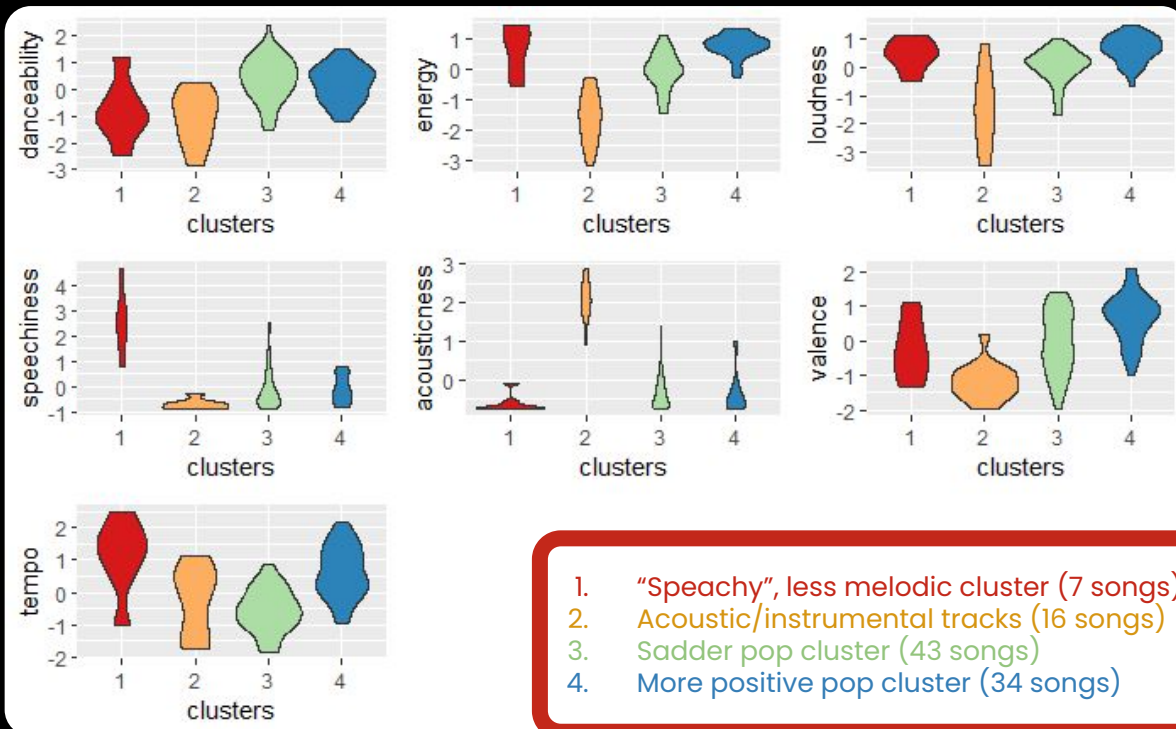
My Choice of K

(How many clusters to use)



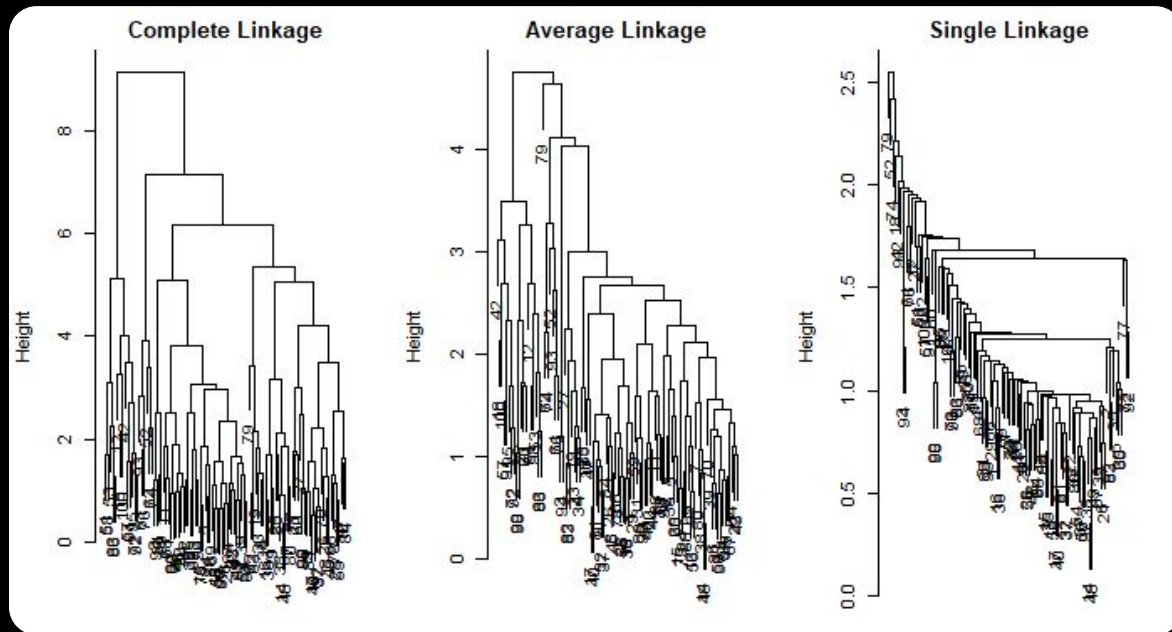
Applying K-Means Clustering

On my data set

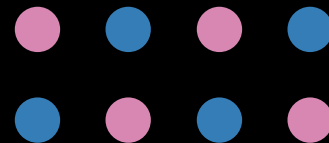
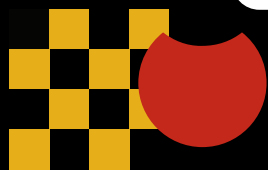


1. "Speechy", less melodic cluster (7 songs)
2. Acoustic/instrumental tracks (16 songs)
3. Sadder pop cluster (43 songs)
4. More positive pop cluster (34 songs)

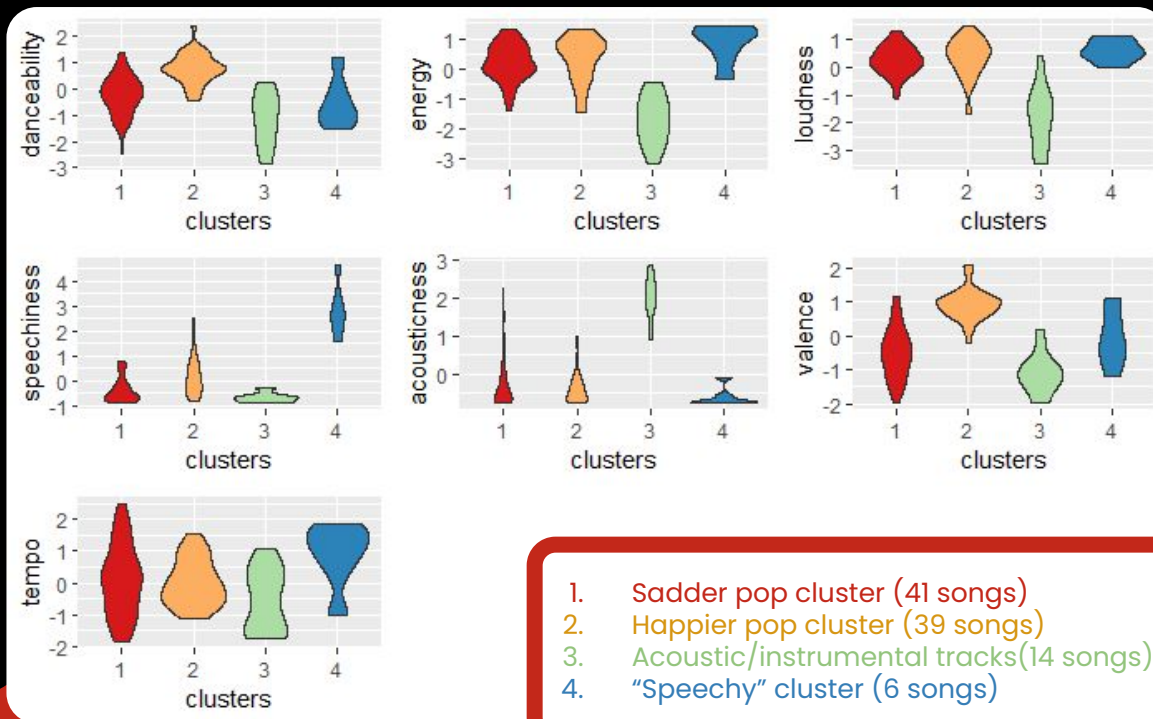
What about Hierarchical?



Dendrograms of Three Modes of Linkage



Results from Hierarchical (Complete)

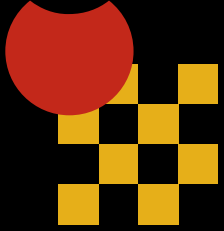


1. Sadder pop cluster (41 songs)
2. Happier pop cluster (39 songs)
3. Acoustic/instrumental tracks (14 songs)
4. "Speechy" cluster (6 songs)



Sushi
Burrito!

Takeaways from the DRP





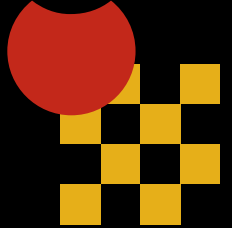
Citations

Information

An Introduction to Statistical Learning
Daniel Suen and David Marcano

Images

K-Means Clustering Plot - [wikimedia.org](https://www.wikimedia.org)
Elbow Plot - vitalflux.com
Book Cover - statlearning.com



Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Second Edition

 Springer





Thank You

for your audience!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**

