Unsupervised Clustering for Tour Route Planning

Using data from Billboard, Spotify, and Songkick

Kevin Giroux

May 2021

Problem statement

For any developing musical artist, building an engaged following is the key to growing your career. Today, in the age of social media, it is easier than ever for a musician to build a global fanbase online.

Social media fans can be transient, though, and the reality is it is less clear than ever where to start when it comes to organically building a local and regional fanbase from the ground-up, especially for artists on a limited budget.

“If I can only afford to hit 3 cities on my next tour, where should I perform between Memphis, Albuquerque, Grand Rapids, Cleveland, Tampa, or Milwaukee?”

Goals

1. Classify a set of 3rd- and 4th-tier US cities by their taste for various musical genres

2. Use the results to define genre “archetypes" for each set of cities that artists can reference when planning their live tour routes

3. Derive insights about which genres are well-received in a given city AND which genres are not well-received

4. Come up with as equal-sized clusters as possible, so that each cluster will contain a set of multiple recommended cities for touring through.

Data sources

Billboard (web scraping):

- Scraped the weekly Top Emerging Artists list for the names of the top developing artists across all genres between January 2018 and March 2020

Spotify API:

- Enriched artist list with the associated genre(s) of each artist

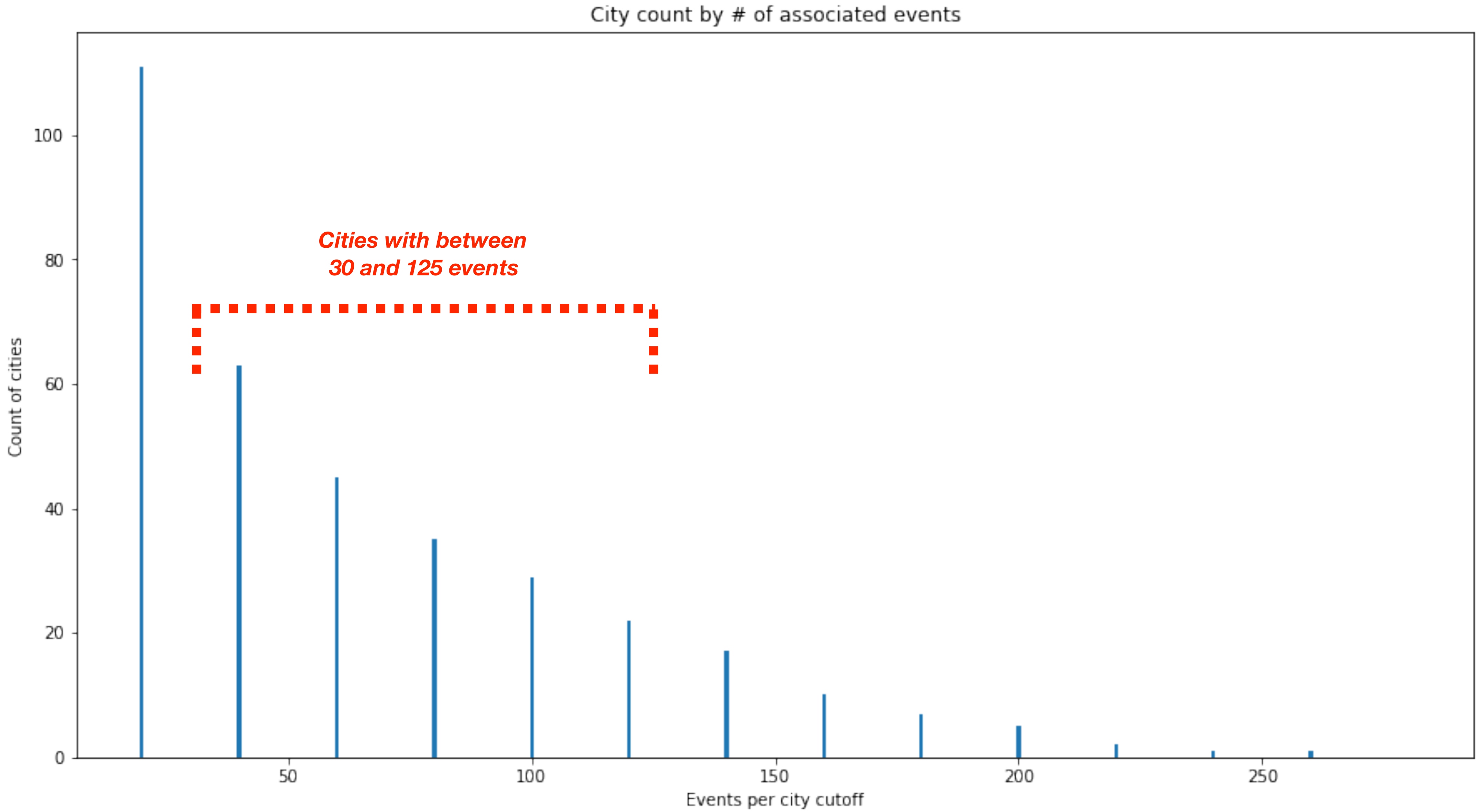
Songkick API:

- Pulled each artist’s full historical “gigography" (list of live performance events) between January 2018 and March 2020 using Songkick’s API

- Each event object included the city and state of the event as well as a popularity score for the event

Methodology

Cutting cities with fewer than 30 associated events and more than 125 associated events resulted in a set of 52 regional cities for clustering.



Pre-scrubbing:

- 14,659 events
- 925 cities

Post-scrubbing:

- 3,111 events
- 52 cities

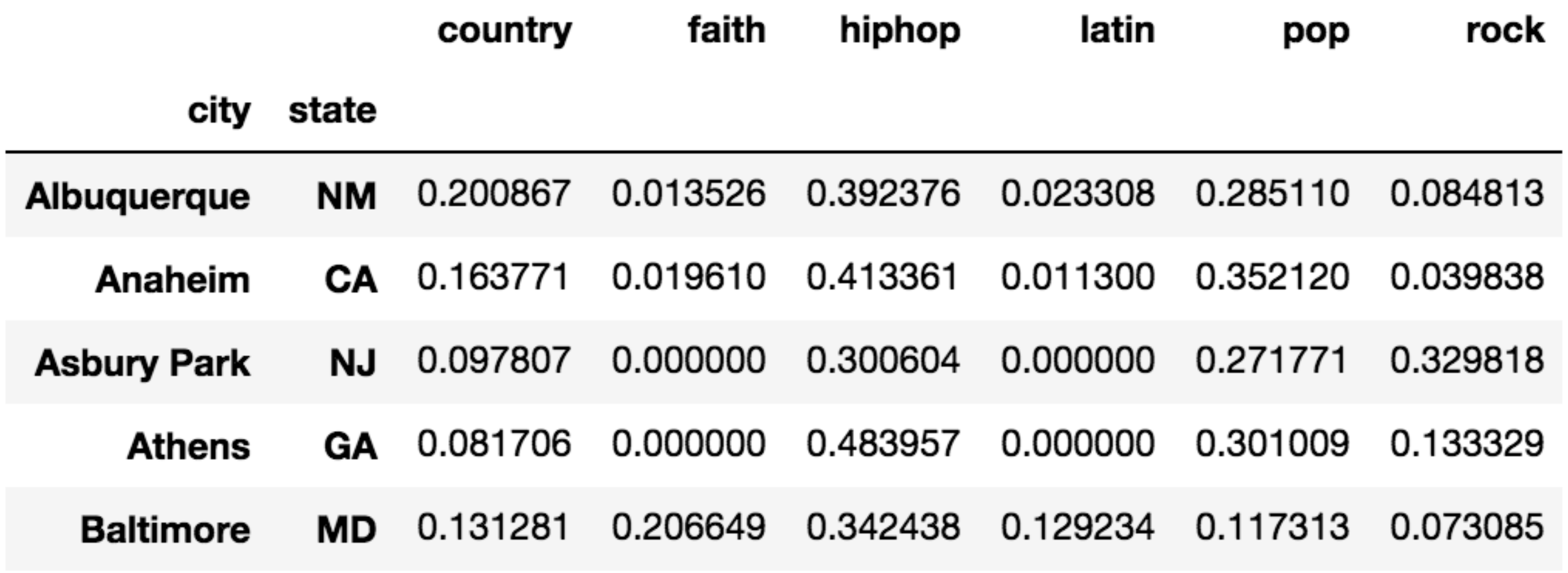
Methodology

For the purposes of this analysis, I bucketed each artist under one of the follow genres:

- Pop
- Hip hop
- Country
- Rock
- Latin
- Faith

Then I calculated an average popularity score for each genre in each city, based on each city’s associated set of events.

Final dataset



		country	faith	hiphop	latin	pop	rock
city	**state**						
Albuquerque	**NM**	0.200867	0.013526	0.392376	0.023308	0.285110	0.084813
Anaheim	**CA**	0.163771	0.019610	0.413361	0.011300	0.352120	0.039838
Asbury Park	**NJ**	0.097807	0.000000	0.300604	0.000000	0.271771	0.329818
Athens	**GA**	0.081706	0.000000	0.483957	0.000000	0.301009	0.133329
Baltimore	**MD**	0.131281	0.206649	0.342438	0.129234	0.117313	0.073085

Finally, I normalized the data by the sum of each row’s individual genre score values, to remove bias in event popularity arising from the varying size of each included city.

(i.e. events in a 1st tier city will always have a higher popularity score than events in 4th tier city)

RESULTS

Unsupervised clustering

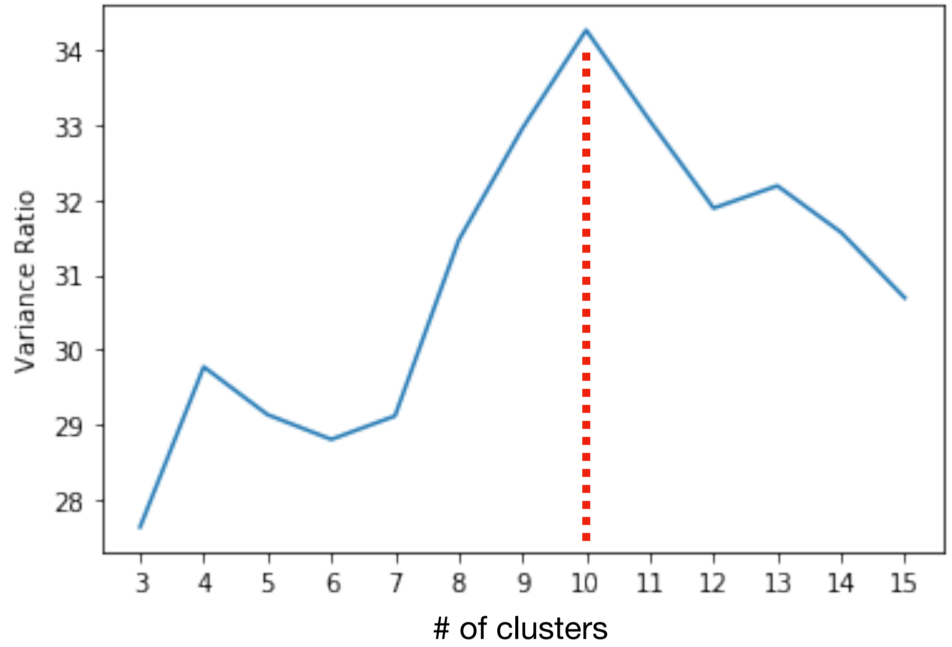
- Employed 2 different clustering techniques:
    - K-means clustering
    - Agglomerative hierarchical clustering
- Compared results of clustering algorithms across 4 different versions of the final dataset, which varied in normalization technique employed

Optimal model: Hierarchical agglomerative clustering, with 9 clusters

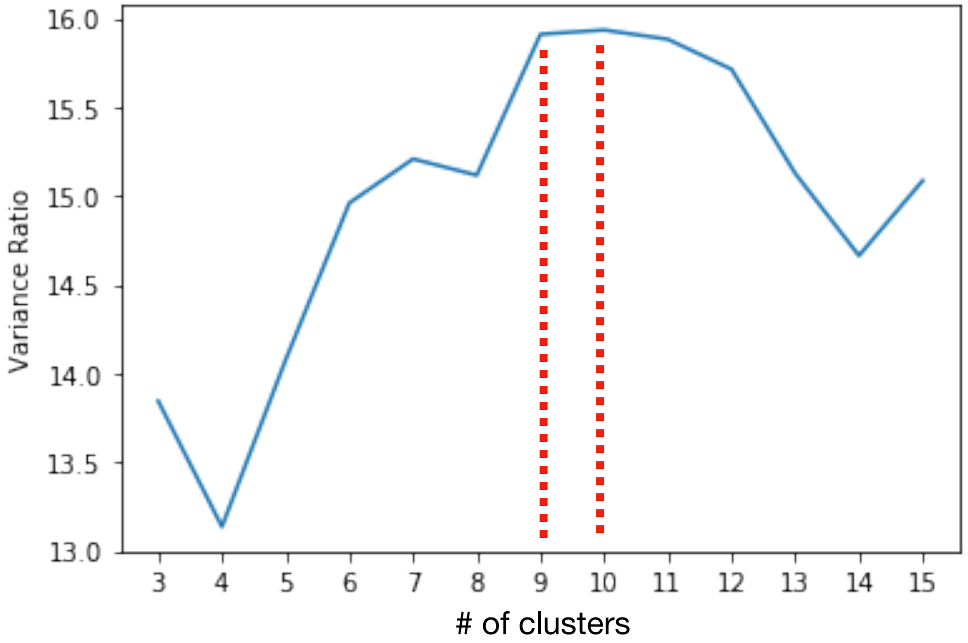
CH & WCSS Scores

The following tests were employed to determine the optimal number of clusters for classifying the dataset.

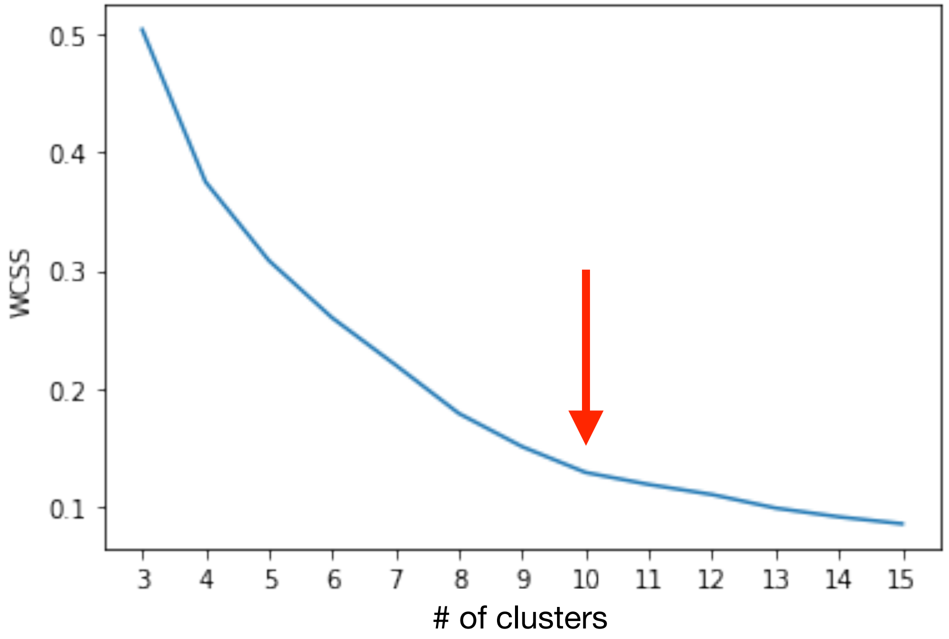
Calinski Harabasz Scores for Different Values of K (Unnormalized data)



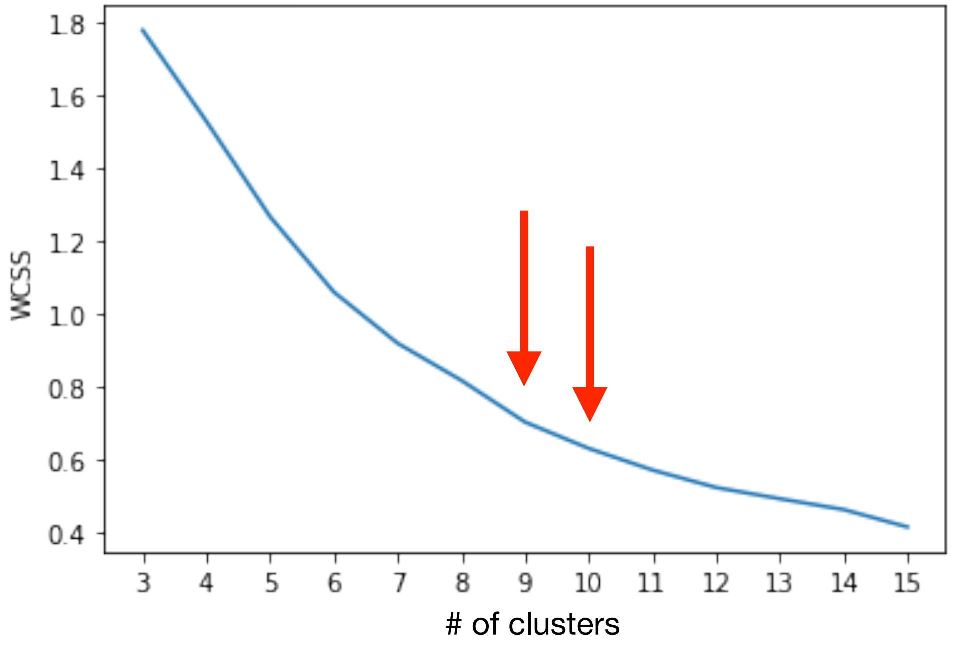
Calinski Harabasz Scores for Different Values of K (Normalized data)



Within Cluster Sum of Squares Score for Different Values of K (Unnormalized data)

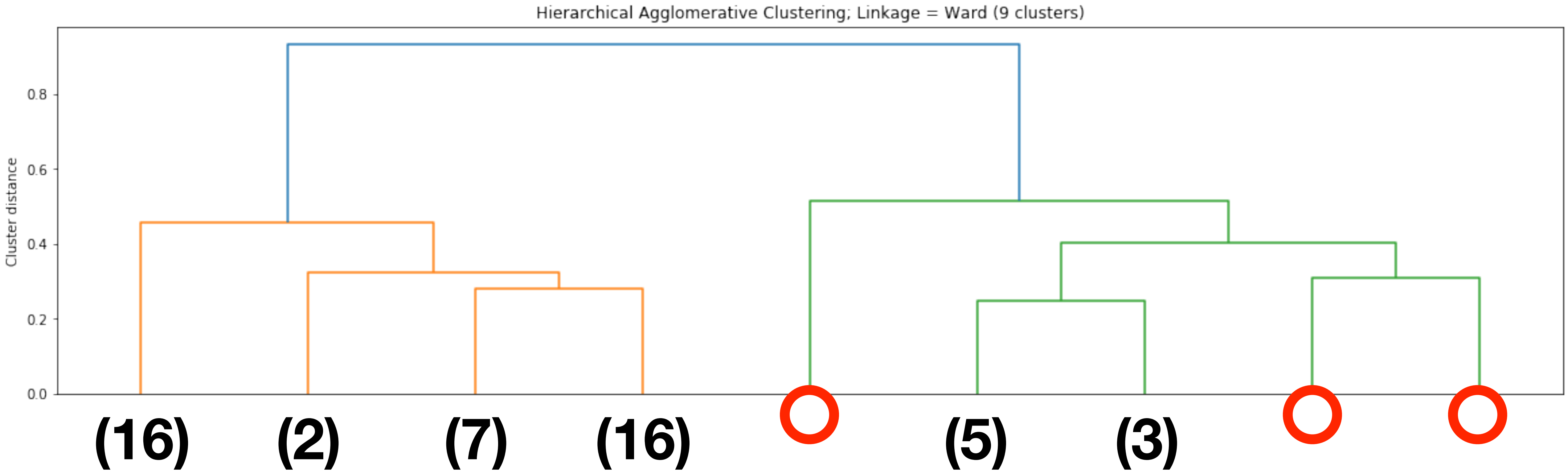


Within Cluster Sum of Squares Score for Different Values of K (Normalized data)



Conclusion: Optimal number of clusters is 9 or 10

Dendrogram (unnormalized data)

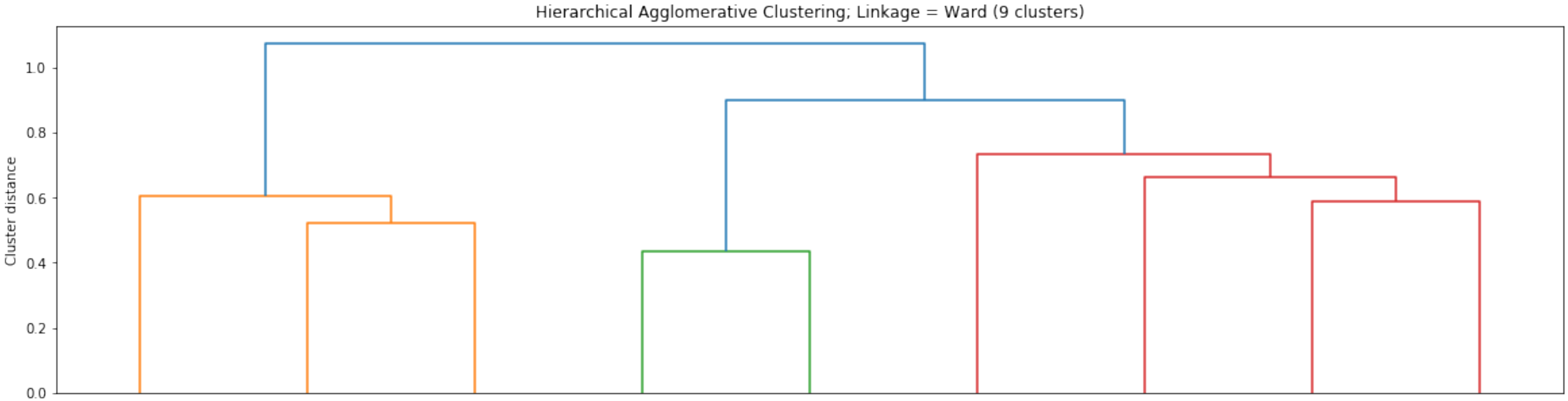


(# of cities in each cluster)

○ : *Denotes a cluster that contains only one city*

Conclusion: Unnormalized data does not result in evenly-sized clusters

2. Dendrogram (normalized data)

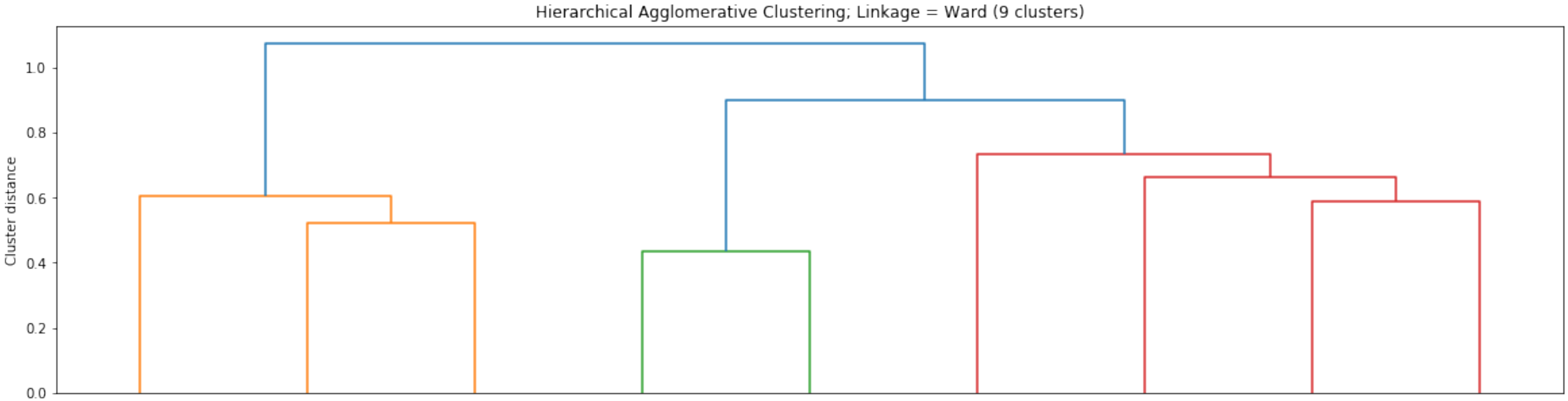


(2) (3) (2) (2) (10) (8) (7) (5) (13)

(# of cities in each cluster)

Conclusion: Normalized data results in relatively evenly-sized clusters (no loners)

2. Dendrogram (normalized data)



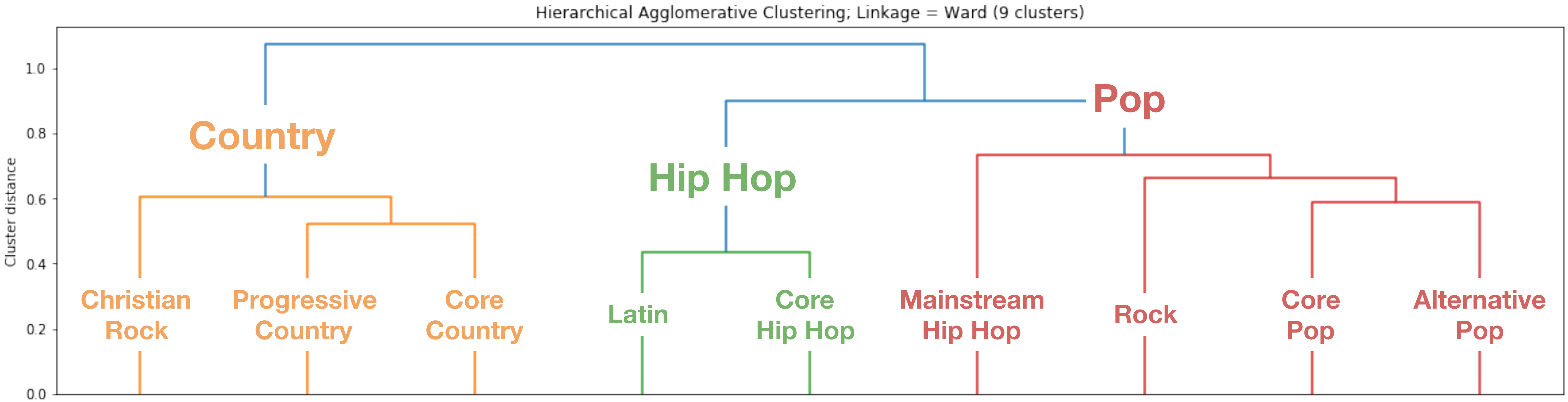
(2) (3) (2) (2) (10) (8) (7) (5) (13)

Country Hip Hop Pop

(# of cities in each cluster)

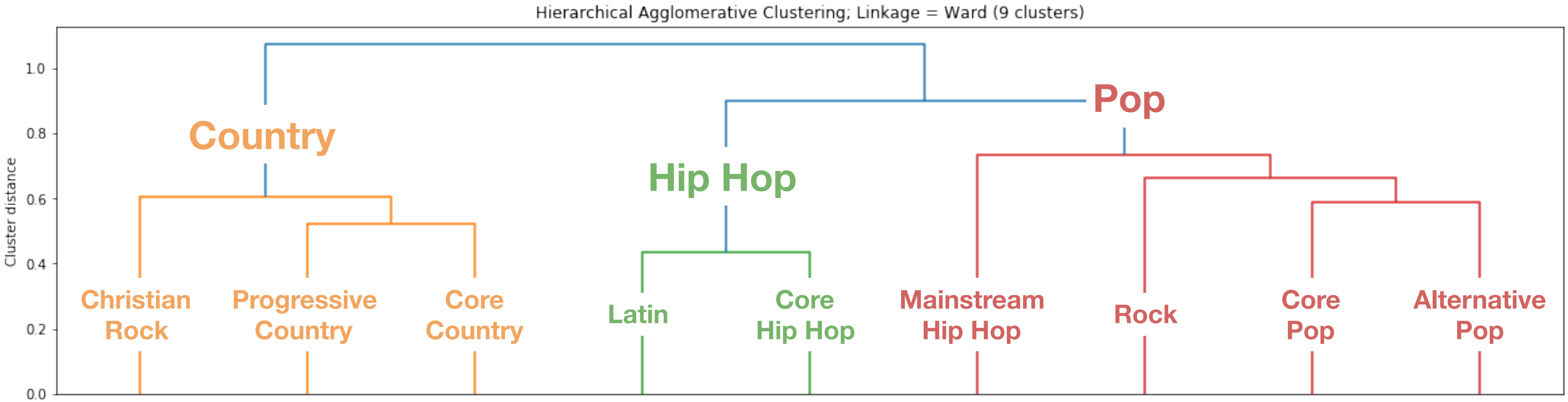
Conclusion: Normalized data results in clear higher-order genre clustering.

Cluster archetypes



5:	7:	2:	8:	1:	3:	4:	6:	0:
Faith *(Rock / Country)*	**Hip hop** *(Country / Rock)*	**Country** *(Hip hop)*	**Latin** *(Hip hop)*	**Hip hop / Pop** *(Latin)*	**Hip hop** *(Pop)*	**Rock / Pop / Hip hop**	**Pop** *(Hip hop)*	**Pop / Hip hop** (Country)
Greenville, SC Spokane, WA	Jacksonville, FL Knoxville, TN Raleigh, NC	Oakland, CA Springfield, MO	Columbus, OH Santa Ana, CA	Baltimore, MD Charlotte, NC Cleveland, OH Grand Rapids, MI Orlando, FL Phoenix, AZ Pittsburgh, PA Sacramento, CA San Antonio, TX Tampa, FL	Albuquerque, NM Anaheim, CA Athens, GA Cincinnati, OH Morrison, CO Richmond, VA Rosemont, IL San Jose, CA	Asbury Park, NJ Cambridge, MA Charleston, SC Columbia, SC Madison, WI Silver Spring, MD Tucson, AZ	Kansas City, MO Memphis, TN Milwaukee, WI New Orleans, LA Tulsa, OK	Birmingham, AL Boise, ID Buffalo, NY Des Moines, IA Greensboro, NC Indianapolis, IN Louisville, KY Mesa, AZ Norfolk, VA Oklahoma City, OK Omaha, NE St Louis, MO St. Paul, MN

Cluster archetypes



5:	7:	2:	8:	1:	3:	4:	6:	0:
Faith *(Rock / Country)*	**Hip hop** *(Country / Rock)*	**Country** *(Hip hop)*	**Latin** *(Hip hop)*	**Hip hop / Pop** *(Latin)*	**Hip hop** *(Pop)*	**Rock / Pop / Hip hop**	**Pop** *(Hip hop)*	**Pop / Hip hop** (Country)
Greenville, SC Spokane, WA	Jacksonville, FL Knoxville, TN Raleigh, NC	Oakland, CA Springfield, MO	Columbus, OH Santa Ana, CA	Baltimore, MD Charlotte, NC Cleveland, OH Grand Rapids, MI Orlando, FL Phoenix, AZ Pittsburgh, PA Sacramento, CA San Antonio, TX Tampa, FL	Albuquerque, NM Anaheim, CA Athens, GA Cincinnati, OH Morrison, CO Richmond, VA Rosemont, IL San Jose, CA	Asbury Park, NJ Cambridge, MA Charleston, SC Columbia, SC Madison, WI Silver Spring, MD Tucson, AZ	Kansas City, MO Memphis, TN Milwaukee, WI New Orleans, LA Tulsa, OK	Birmingham, AL Boise, ID Buffalo, NY Des Moines, IA Greensboro, NC Indianapolis, IN Louisville, KY Mesa, AZ Norfolk, VA Oklahoma City, OK Omaha, NE St Louis, MO St. Paul, MN

Other thoughts

- Analysis was based on 6 genres, but was optimized at 9 to 10 clusters
    - Speaks to the mix of tastes that exists in every market
    - Also speaks to the fact that genre is a spectrum, and that many genres are sonically related to one another

- Hip hop was universally the most popular genre
    - Indication that hip hop is more widely enjoyed than pop across the cities in question

- Context and subject matter expertise are required
    - How to structure your dataset (e.g. how to normalize)
    - Cluster parameters (e.g. avoiding solutions that result in clusters of 1 city)

- The output of this analysis could be used for many tasks other than tour planning - e.g. partnering with local media outlets, like radio or podcast

Future work

1. Automate the higher-order genre mapping step of the analysis

2. Perform the same analysis with a bigger sample size - some surprising results may have been due to outliers or class imbalance

3. Perform the same analysis within a single macro, such as Hip Hop, and evaluate US cities for their taste in various sub-genres of Hip Hop
    - This would mitigate the manual genre mapping required in my original analysis
    - It could result in more distinct clusters which would be easier to derive actionable insights from

4. Learn from Songkick what the technical definition of the "popularity" attribute is

5. Create a front-end where users can run their own analyses based on their own lists of comparable artists

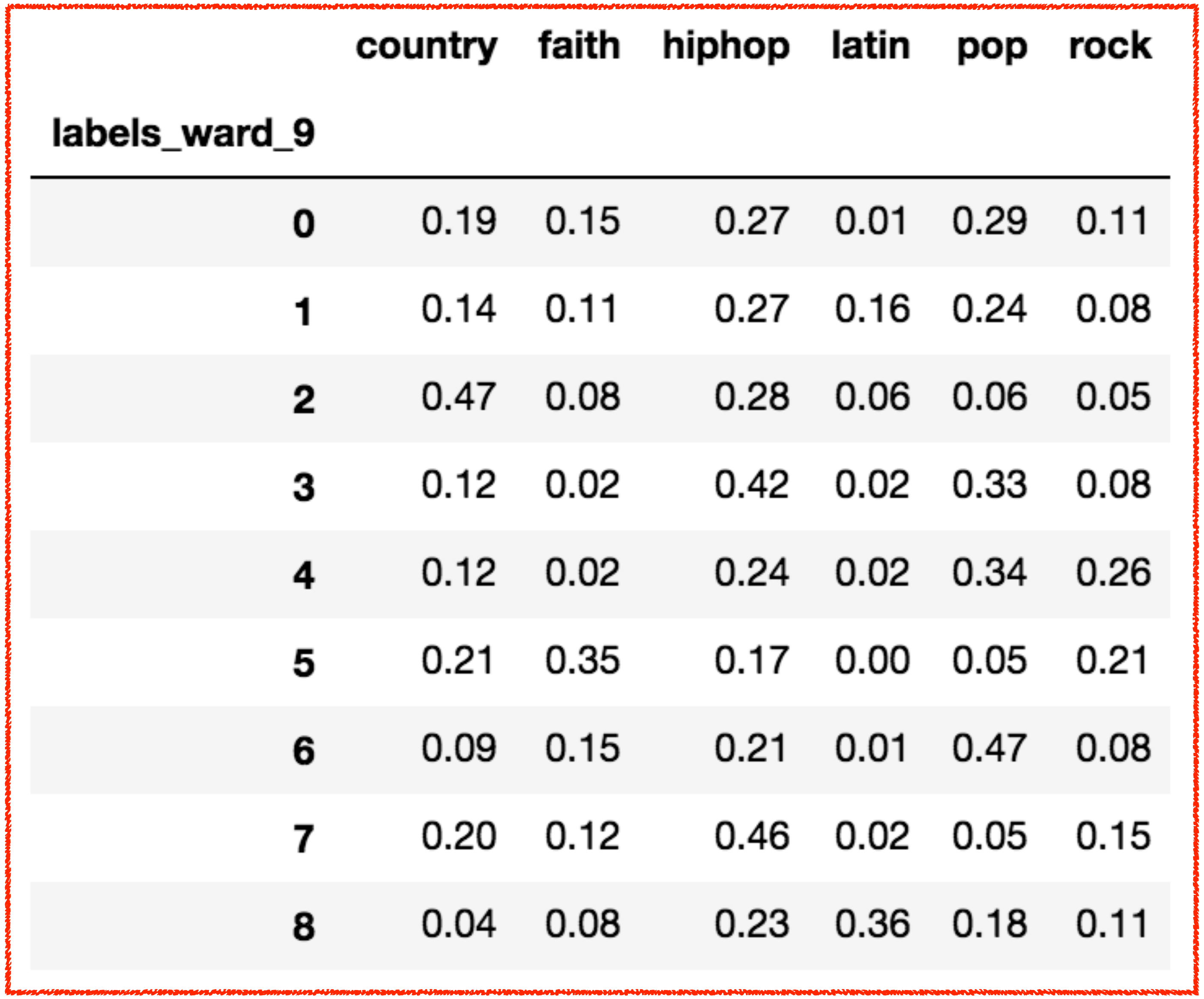
6. Perform the same analysis using international countries rather than US cities

Thank you!

Kevin Giroux
kevinsgiroux@gmail.com

Appendix

Final data table of popularity scores



		country	faith	hiphop	latin	pop	rock
	labels_ward_9						
Alternative Pop	**0**	0.19	0.15	0.27	0.01	0.29	0.11
Core Hip Hop	**1**	0.14	0.11	0.27	0.16	0.24	0.08
Core Country	**2**	0.47	0.08	0.28	0.06	0.06	0.05
Mainstream Hip Hop	**3**	0.12	0.02	0.42	0.02	0.33	0.08
Rock	**4**	0.12	0.02	0.24	0.02	0.34	0.26
Christian Rock	**5**	0.21	0.35	0.17	0.00	0.05	0.21
Core Pop	**6**	0.09	0.15	0.21	0.01	0.47	0.08
Progressive Country	**7**	0.20	0.12	0.46	0.02	0.05	0.15
Latin	**8**	0.04	0.08	0.23	0.36	0.18	0.11

Final city groupings by cluster ID

Alternative Pop

0: Pop / Hip hop (Country)

Birmingham, AL
Boise, ID
Buffalo, NY
Des Moines, IA
Greensboro, NC
Indianapolis, IN
Louisville, KY
Mesa, AZ
Norfolk, VA
Oklahoma City, OK
Omaha, NE
St Louis, MO
St. Paul, MN

Core Hip Hop

1: Hip hop / Pop (Latin)

Baltimore, MD
Charlotte, NC
Cleveland, OH
Grand Rapids, MI
Orlando, FL
Phoenix, AZ
Pittsburgh, PA
Sacramento, CA
San Antonio, TX
Tampa, FL

Core Country

2: Country (Hip hop)

Oakland, CA
Springfield, MO

Mainstream Hip Hop

3: Hip hop (Pop)

Albuquerque, NM
Anaheim, CA
Athens, GA
Cincinnati, OH
Morrison, CO
Richmond, VA
Rosemont, IL
San Jose, CA

Rock

4: Rock / Pop / Hip hop

Asbury Park, NJ
Cambridge, MA
Charleston, SC
Columbia, SC
Madison, WI
Silver Spring, MD
Tucson, AZ

Christian Rock

5: Faith (Rock / Country)

Greenville, SC
Spokane, WA

Core Pop

6: Pop (Hip hop)

Kansas City, MO
Memphis, TN
Milwaukee, WI
New Orleans, LA
Tulsa, OK

Progressive Country

7: Hip hop (Country / Rock)

Jacksonville, FL
Knoxville, TN
Raleigh, NC

Latin

8: Latin (Hip hop)

Columbus, OH
Santa Ana, CA