A CLUSTERING ANALYSIS

CITY IN NORTH AMERICA

INTRODUCTION

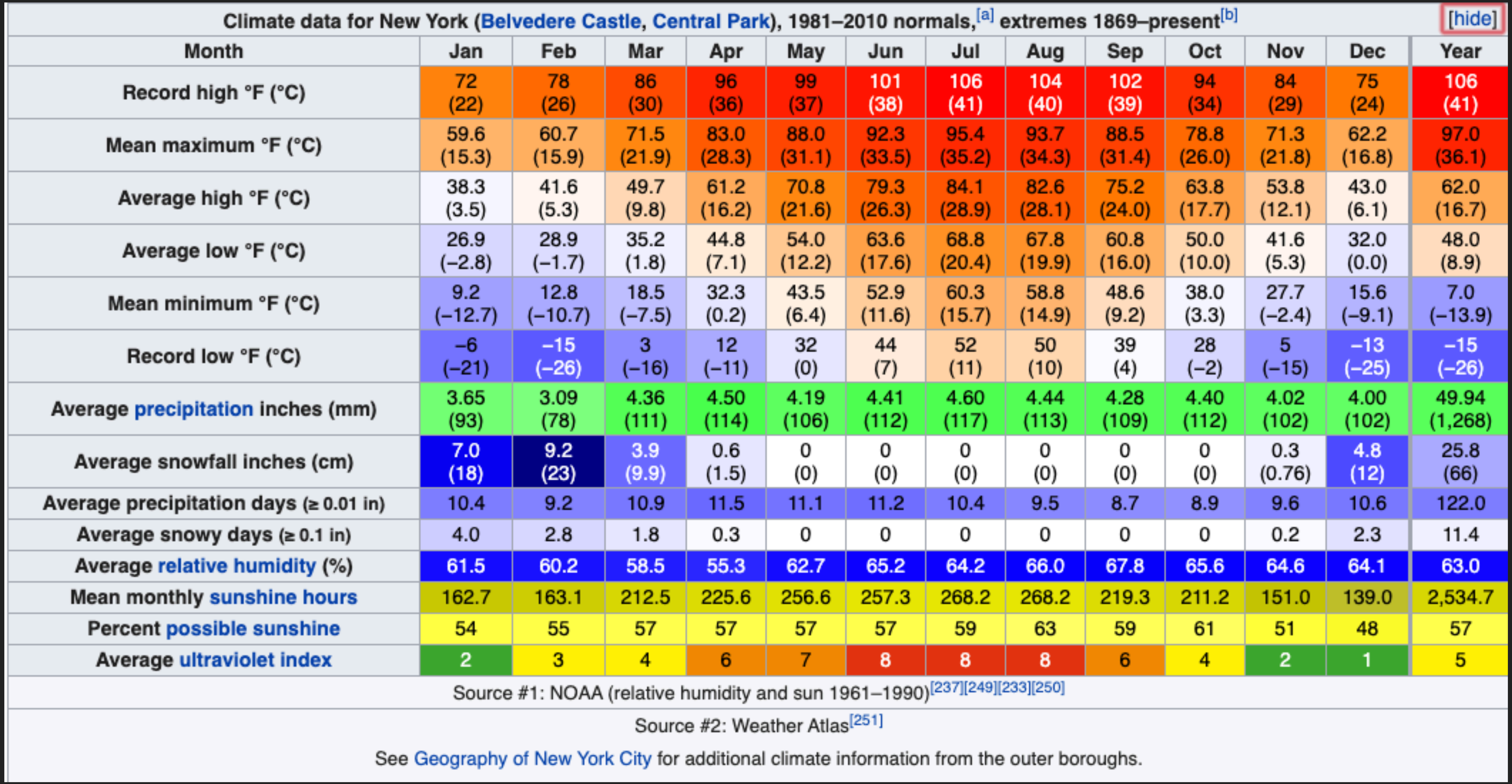
- As it becomes more common for people to relocate between cities within or even across countries, it is useful to provide a comparison between cities in different scales such as climate, demographic, living cost etc.
- In this project, we are going to explore the different cities in North America, and compare them in different aspects.

DATA ACQUISITION AND CLEANING

- For this project, we need three types of data:
  - City Climate Data From Wikipedia
  - Demographic Data From Wikipedia
  - City Venues Data From Foursquare

CLIMATE DATA

- For climate data, We Scrape from Wikipedia Dat



Climate data for New York (Belvedere Castle, Central Park), 1981–2010 normals,[a] extremes 1869–present[b] [hide]

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Year
Record high °F (°C)	72 (22)	78 (26)	86 (30)	96 (36)	99 (37)	101 (38)	106 (41)	104 (40)	102 (39)	94 (34)	84 (29)	75 (24)	106 (41)
Mean maximum °F (°C)	59.6 (15.3)	60.7 (15.9)	71.5 (21.9)	83.0 (28.3)	88.0 (31.1)	92.3 (33.5)	95.4 (35.2)	93.7 (34.3)	88.5 (31.4)	78.8 (26.0)	71.3 (21.8)	62.2 (16.8)	97.0 (36.1)
Average high °F (°C)	38.3 (3.5)	41.6 (5.3)	49.7 (9.8)	61.2 (16.2)	70.8 (21.6)	79.3 (26.3)	84.1 (28.9)	82.6 (28.1)	75.2 (24.0)	63.8 (17.7)	53.8 (12.1)	43.0 (6.1)	62.0 (16.7)
Average low °F (°C)	26.9 (−2.8)	28.9 (−1.7)	35.2 (1.8)	44.8 (7.1)	54.0 (12.2)	63.6 (17.6)	68.8 (20.4)	67.8 (19.9)	60.8 (16.0)	50.0 (10.0)	41.6 (5.3)	32.0 (0.0)	48.0 (8.9)
Mean minimum °F (°C)	9.2 (−12.7)	12.8 (−10.7)	18.5 (−7.5)	32.3 (0.2)	43.5 (6.4)	52.9 (11.6)	60.3 (15.7)	58.8 (14.9)	48.6 (9.2)	38.0 (3.3)	27.7 (−2.4)	15.6 (−9.1)	7.0 (−13.9)
Record low °F (°C)	−6 (−21)	−15 (−26)	3 (−16)	12 (−11)	32 (0)	44 (7)	52 (11)	50 (10)	39 (4)	28 (−2)	5 (−15)	−13 (−25)	−15 (−26)
Average precipitation inches (mm)	3.65 (93)	3.09 (78)	4.36 (111)	4.50 (114)	4.19 (106)	4.41 (112)	4.60 (117)	4.44 (113)	4.28 (109)	4.40 (112)	4.02 (102)	4.00 (102)	49.94 (1,268)
Average snowfall inches (cm)	7.0 (18)	9.2 (23)	3.9 (9.9)	0.6 (1.5)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.3 (0.76)	4.8 (12)	25.8 (66)
Average precipitation days (≥ 0.01 in)	10.4	9.2	10.9	11.5	11.1	11.2	10.4	9.5	8.7	8.9	9.6	10.6	122.0
Average snowy days (≥ 0.1 in)	4.0	2.8	1.8	0.3	0	0	0	0	0	0	0.2	2.3	11.4
Average relative humidity (%)	61.5	60.2	58.5	55.3	62.7	65.2	64.2	66.0	67.8	65.6	64.6	64.1	63.0
Mean monthly sunshine hours	162.7	163.1	212.5	225.6	256.6	257.3	268.2	268.2	219.3	211.2	151.0	139.0	2,534.7
Percent possible sunshine	54	55	57	57	57	57	59	63	59	61	51	48	57
Average ultraviolet index	2	3	4	6	7	8	8	8	6	4	2	1	5

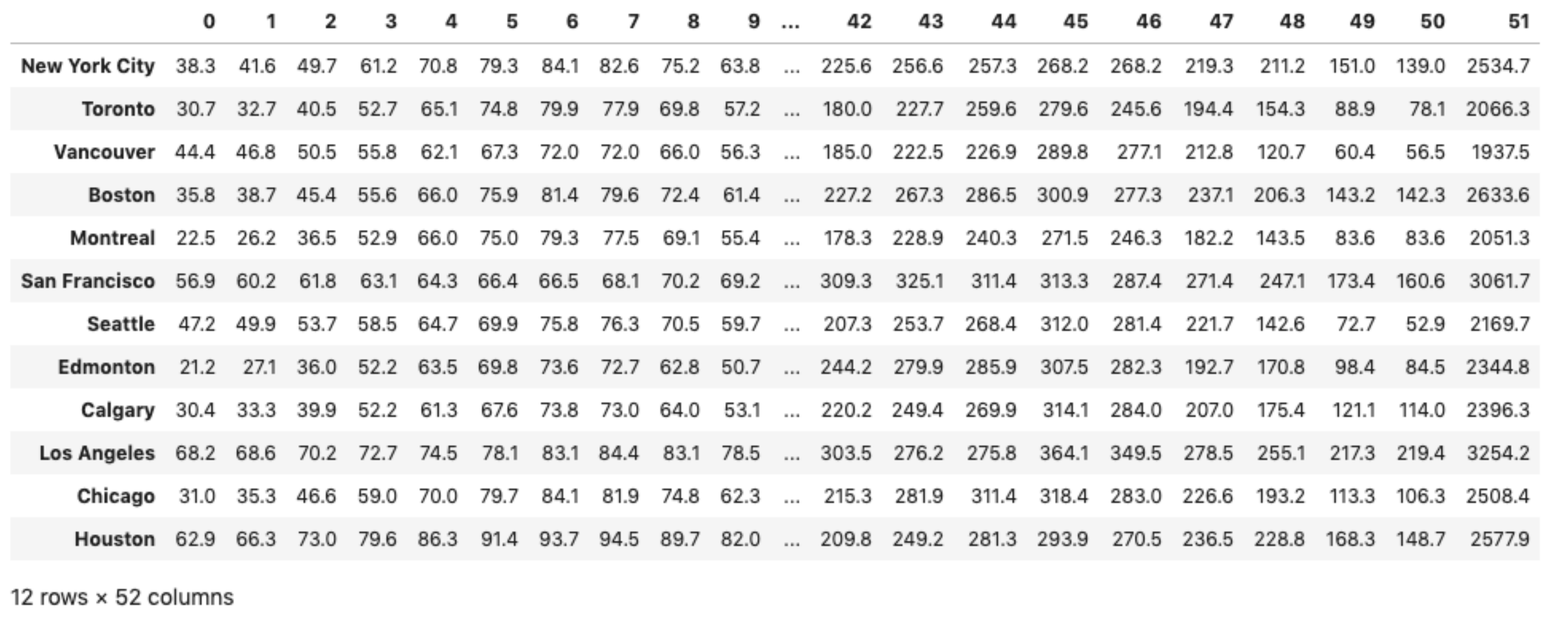
Source #1: NOAA (relative humidity and sun 1961–1990)[237][249][233][250]

Source #2: Weather Atlas[251]

See Geography of New York City for additional climate information from the outer boroughs.

CLIMATE DATA

- We use 5 data sets from climate data, “Average High, Average Low, Precipitation Days, Rainy Days, Monthly Sunshine Hours”, and vectorize into one dimensional data for each city.

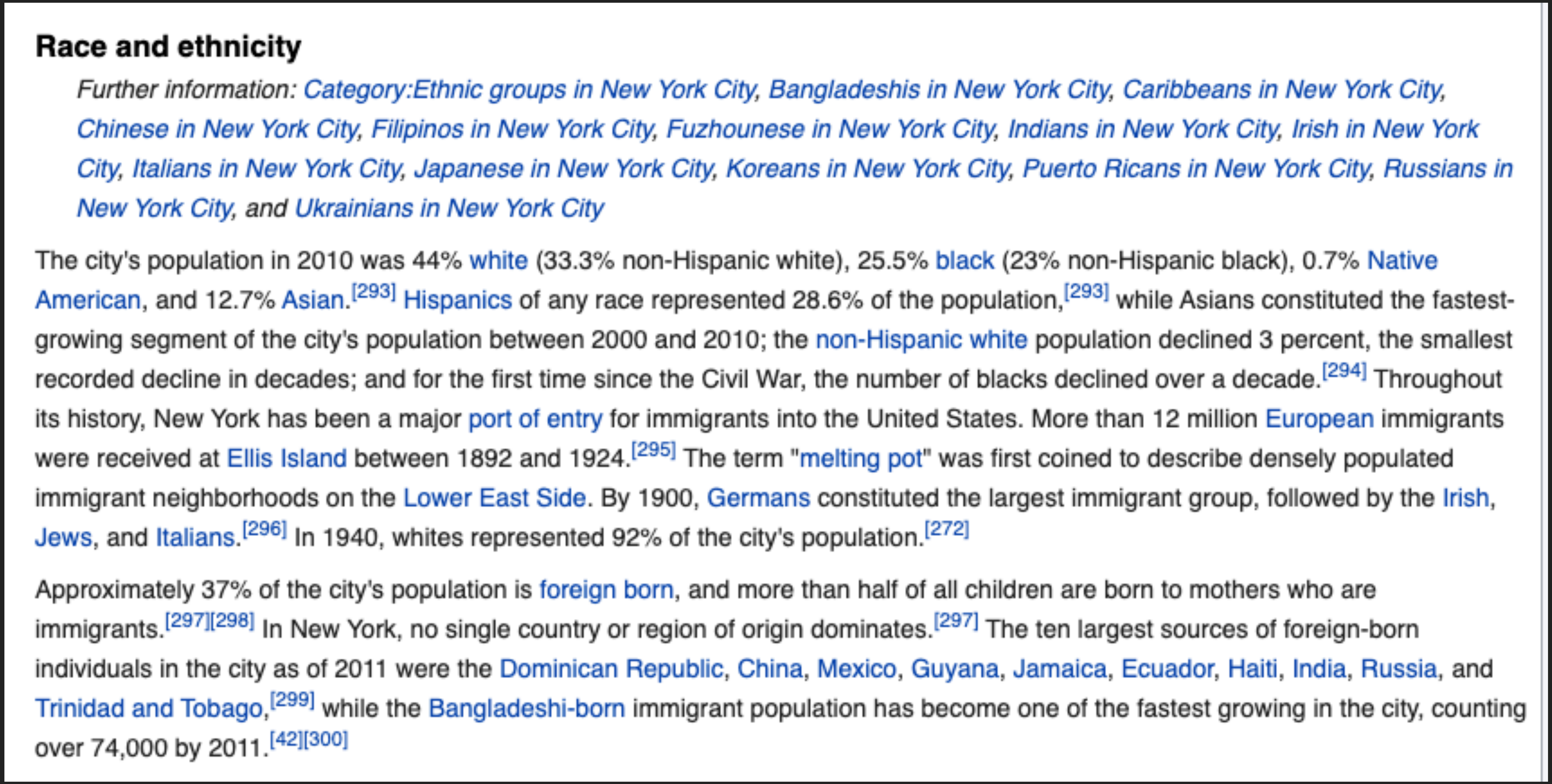


	0	1	2	3	4	5	6	7	8	9	...	42	43	44	45	46	47	48	49	50	51
New York City	38.3	41.6	49.7	61.2	70.8	79.3	84.1	82.6	75.2	63.8	...	225.6	256.6	257.3	268.2	268.2	219.3	211.2	151.0	139.0	2534.7
Toronto	30.7	32.7	40.5	52.7	65.1	74.8	79.9	77.9	69.8	57.2	...	180.0	227.7	259.6	279.6	245.6	194.4	154.3	88.9	78.1	2066.3
Vancouver	44.4	46.8	50.5	55.8	62.1	67.3	72.0	72.0	66.0	56.3	...	185.0	222.5	226.9	289.8	277.1	212.8	120.7	60.4	56.5	1937.5
Boston	35.8	38.7	45.4	55.6	66.0	75.9	81.4	79.6	72.4	61.4	...	227.2	267.3	286.5	300.9	277.3	237.1	206.3	143.2	142.3	2633.6
Montreal	22.5	26.2	36.5	52.9	66.0	75.0	79.3	77.5	69.1	55.4	...	178.3	228.9	240.3	271.5	246.3	182.2	143.5	83.6	83.6	2051.3
San Francisco	56.9	60.2	61.8	63.1	64.3	66.4	66.5	68.1	70.2	69.2	...	309.3	325.1	311.4	313.3	287.4	271.4	247.1	173.4	160.6	3061.7
Seattle	47.2	49.9	53.7	58.5	64.7	69.9	75.8	76.3	70.5	59.7	...	207.3	253.7	268.4	312.0	281.4	221.7	142.6	72.7	52.9	2169.7
Edmonton	21.2	27.1	36.0	52.2	63.5	69.8	73.6	72.7	62.8	50.7	...	244.2	279.9	285.9	307.5	282.3	192.7	170.8	98.4	84.5	2344.8
Calgary	30.4	33.3	39.9	52.2	61.3	67.6	73.8	73.0	64.0	53.1	...	220.2	249.4	269.9	314.1	284.0	207.0	175.4	121.1	114.0	2396.3
Los Angeles	68.2	68.6	70.2	72.7	74.5	78.1	83.1	84.4	83.1	78.5	...	303.5	276.2	275.8	364.1	349.5	278.5	255.1	217.3	219.4	3254.2
Chicago	31.0	35.3	46.6	59.0	70.0	79.7	84.1	81.9	74.8	62.3	...	215.3	281.9	311.4	318.4	283.0	226.6	193.2	113.3	106.3	2508.4
Houston	62.9	66.3	73.0	79.6	86.3	91.4	93.7	94.5	89.7	82.0	...	209.8	249.2	281.3	293.9	270.5	236.5	228.8	168.3	148.7	2577.9

12 rows × 52 columns

DEMOGRAPHIC RACE DATA

- For Race data, because it’s unstructured, we have to grad it manually from wikipedia pages.



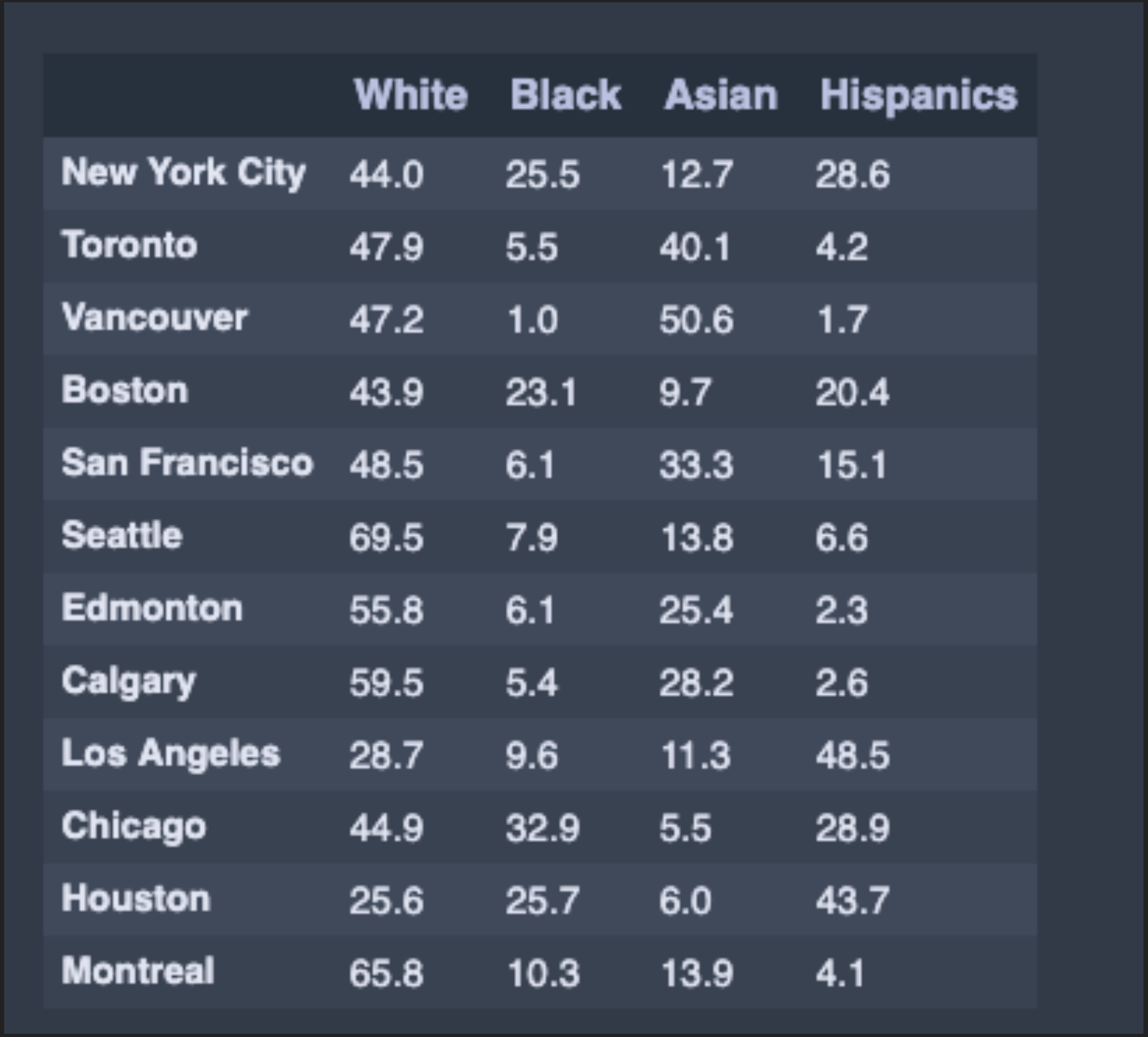
Race and ethnicity

Further information: Category:Ethnic groups in New York City, Bangladeshis in New York City, Caribbeans in New York City, Chinese in New York City, Filipinos in New York City, Fuzhounese in New York City, Indians in New York City, Irish in New York City, Italians in New York City, Japanese in New York City, Koreans in New York City, Puerto Ricans in New York City, Russians in New York City, and Ukrainians in New York City

The city's population in 2010 was 44% white (33.3% non-Hispanic white), 25.5% black (23% non-Hispanic black), 0.7% Native American, and 12.7% Asian.[293] Hispanics of any race represented 28.6% of the population,[293] while Asians constituted the fastest-growing segment of the city's population between 2000 and 2010; the non-Hispanic white population declined 3 percent, the smallest recorded decline in decades; and for the first time since the Civil War, the number of blacks declined over a decade.[294] Throughout its history, New York has been a major port of entry for immigrants into the United States. More than 12 million European immigrants were received at Ellis Island between 1892 and 1924.[295] The term "melting pot" was first coined to describe densely populated immigrant neighborhoods on the Lower East Side. By 1900, Germans constituted the largest immigrant group, followed by the Irish, Jews, and Italians.[296] In 1940, whites represented 92% of the city's population.[272]

Approximately 37% of the city's population is foreign born, and more than half of all children are born to mothers who are immigrants.[297][298] In New York, no single country or region of origin dominates.[297] The ten largest sources of foreign-born individuals in the city as of 2011 were the Dominican Republic, China, Mexico, Guyana, Jamaica, Ecuador, Haiti, India, Russia, and Trinidad and Tobago,[299] while the Bangladeshi-born immigrant population has become one of the fastest growing in the city, counting over 74,000 by 2011.[42][300]

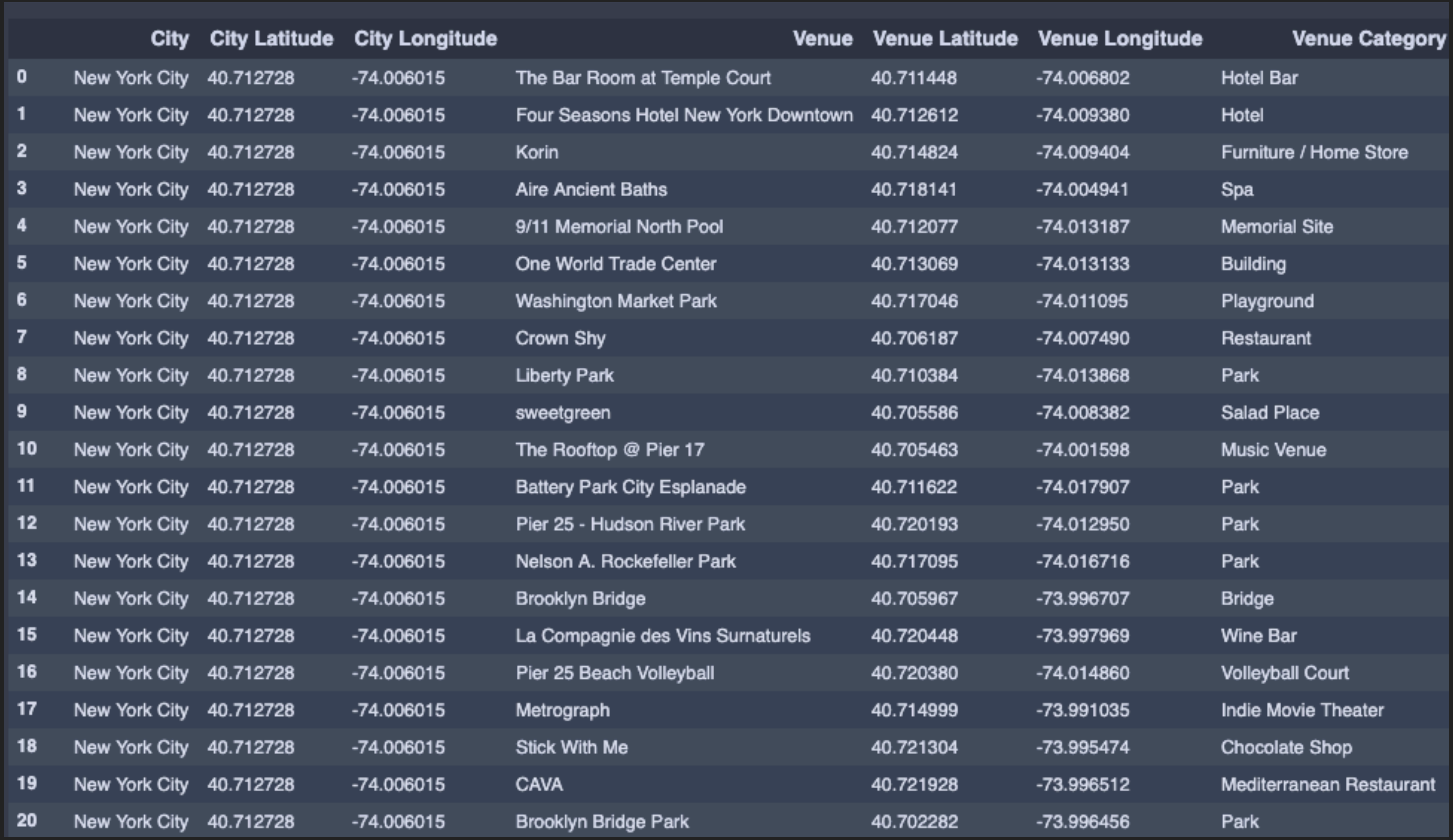
DEMOGRAPHIC RACE DATA



	White	Black	Asian	Hispanics
New York City	44.0	25.5	12.7	28.6
Toronto	47.9	5.5	40.1	4.2
Vancouver	47.2	1.0	50.6	1.7
Boston	43.9	23.1	9.7	20.4
San Francisco	48.5	6.1	33.3	15.1
Seattle	69.5	7.9	13.8	6.6
Edmonton	55.8	6.1	25.4	2.3
Calgary	59.5	5.4	28.2	2.6
Los Angeles	28.7	9.6	11.3	48.5
Chicago	44.9	32.9	5.5	28.9
Houston	25.6	25.7	6.0	43.7
Montreal	65.8	10.3	13.9	4.1

FOURSQUARE

- We just use Foursquare Search API to get venues data for specific city.



	City	City Latitude	City Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	New York City	40.712728	-74.006015	The Bar Room at Temple Court	40.711448	-74.006802	Hotel Bar
1	New York City	40.712728	-74.006015	Four Seasons Hotel New York Downtown	40.712612	-74.009380	Hotel
2	New York City	40.712728	-74.006015	Korin	40.714824	-74.009404	Furniture / Home Store
3	New York City	40.712728	-74.006015	Aire Ancient Baths	40.718141	-74.004941	Spa
4	New York City	40.712728	-74.006015	9/11 Memorial North Pool	40.712077	-74.013187	Memorial Site
5	New York City	40.712728	-74.006015	One World Trade Center	40.713069	-74.013133	Building
6	New York City	40.712728	-74.006015	Washington Market Park	40.717046	-74.011095	Playground
7	New York City	40.712728	-74.006015	Crown Shy	40.706187	-74.007490	Restaurant
8	New York City	40.712728	-74.006015	Liberty Park	40.710384	-74.013868	Park
9	New York City	40.712728	-74.006015	sweetgreen	40.705586	-74.008382	Salad Place
10	New York City	40.712728	-74.006015	The Rooftop @ Pier 17	40.705463	-74.001598	Music Venue
11	New York City	40.712728	-74.006015	Battery Park City Esplanade	40.711622	-74.017907	Park
12	New York City	40.712728	-74.006015	Pier 25 - Hudson River Park	40.720193	-74.012950	Park
13	New York City	40.712728	-74.006015	Nelson A. Rockefeller Park	40.717095	-74.016716	Park
14	New York City	40.712728	-74.006015	Brooklyn Bridge	40.705967	-73.996707	Bridge
15	New York City	40.712728	-74.006015	La Compagnie des Vins Surnaturels	40.720448	-73.997969	Wine Bar
16	New York City	40.712728	-74.006015	Pier 25 Beach Volleyball	40.720380	-74.014860	Volleyball Court
17	New York City	40.712728	-74.006015	Metrograph	40.714999	-73.991035	Indie Movie Theater
18	New York City	40.712728	-74.006015	Stick With Me	40.721304	-73.995474	Chocolate Shop
19	New York City	40.712728	-74.006015	CAVA	40.721928	-73.996512	Mediterranean Restaurant
20	New York City	40.712728	-74.006015	Brooklyn Bridge Park	40.702282	-73.996456	Park

ANALYZE METHODOLOGY

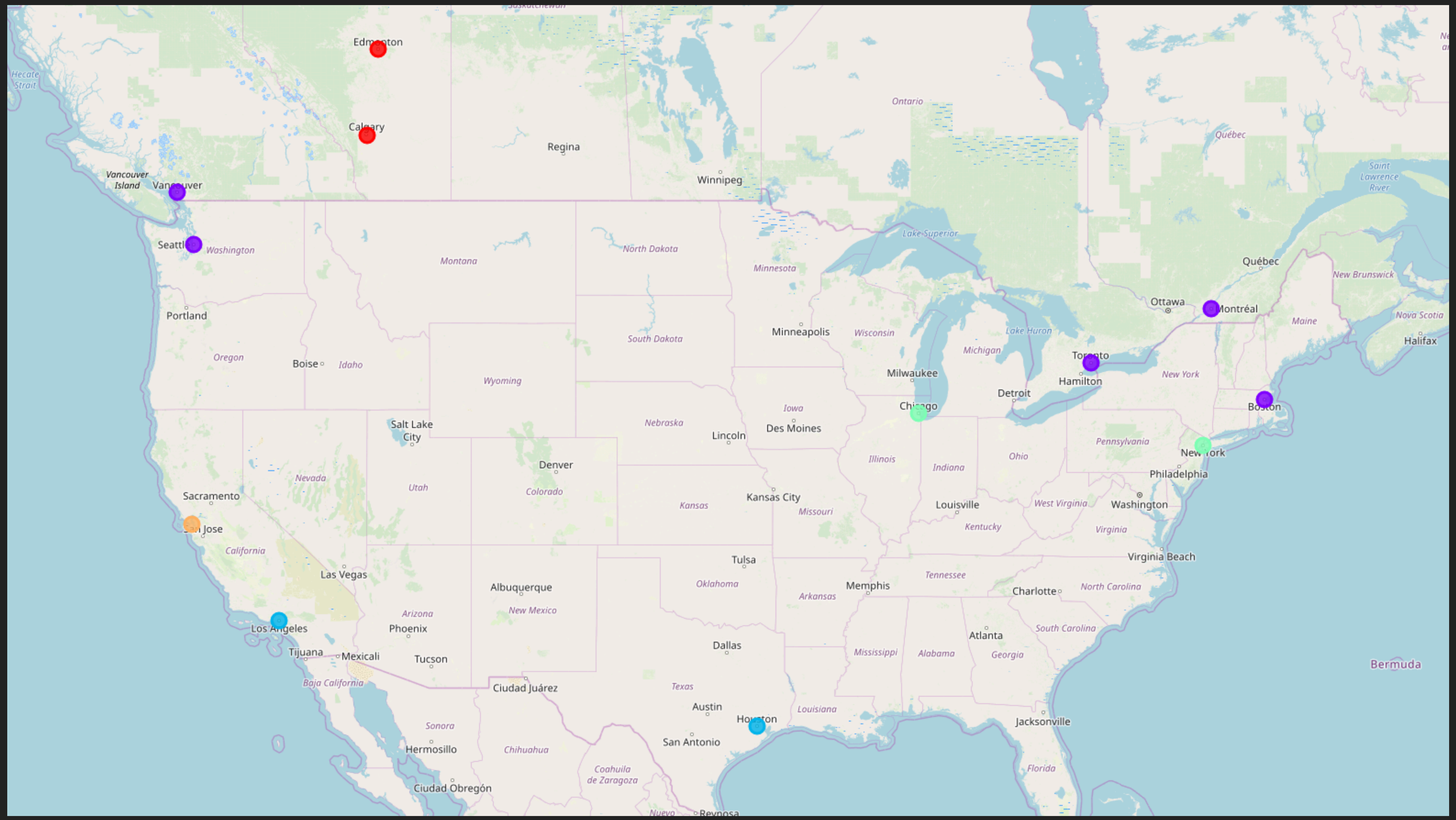
- Hierarchical Clustering
- Dendrogram
- Choose Number of Clustering makes most sense

ANALYZE METHODOLOGY

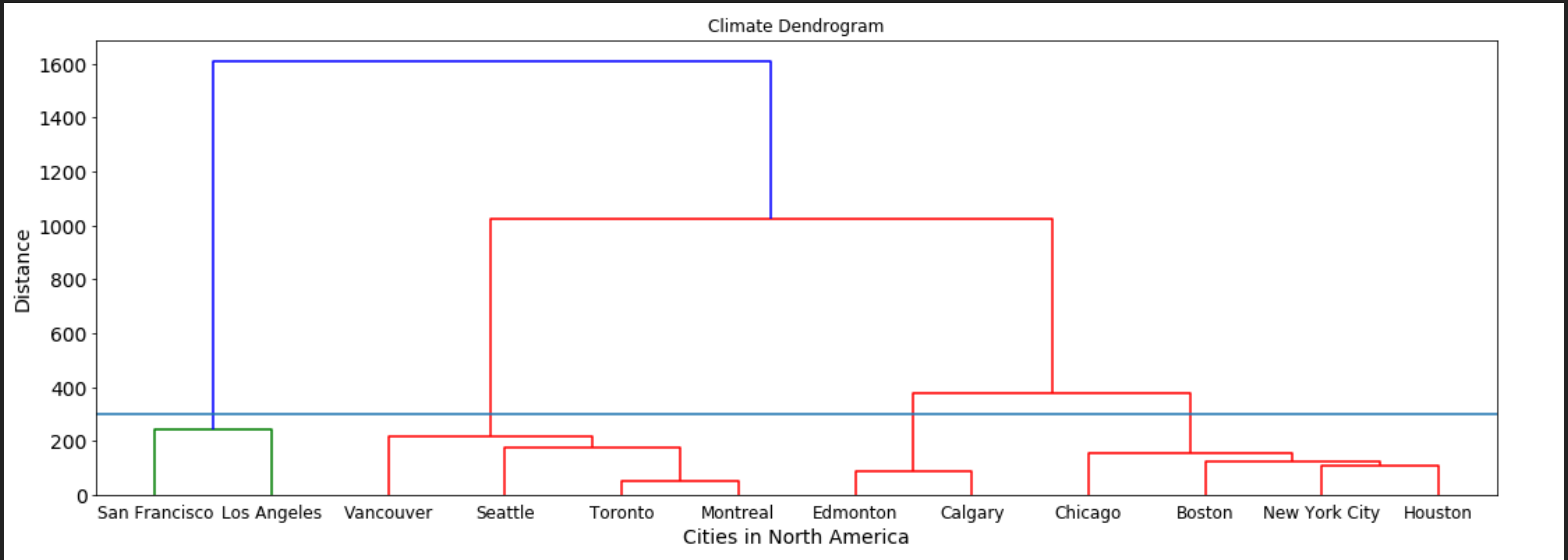
- We use 5 data sets from climate data, "Average High, Average Low, Precipitation Days, Rainy Days, Monthly Sunshine Hours", and vectorize into one dimensional data for each city.
- For Demographic Data, since the original data is just 5 dimensions, We don't have to do preprocessing.
- For Venus data, we count the number of venues for each Category.

- Then We do hierarchical clustering to see the best number to do clustering for all three data set.
- Finally, we combine all the data with Venus and Climate data preprocessed by PCA.

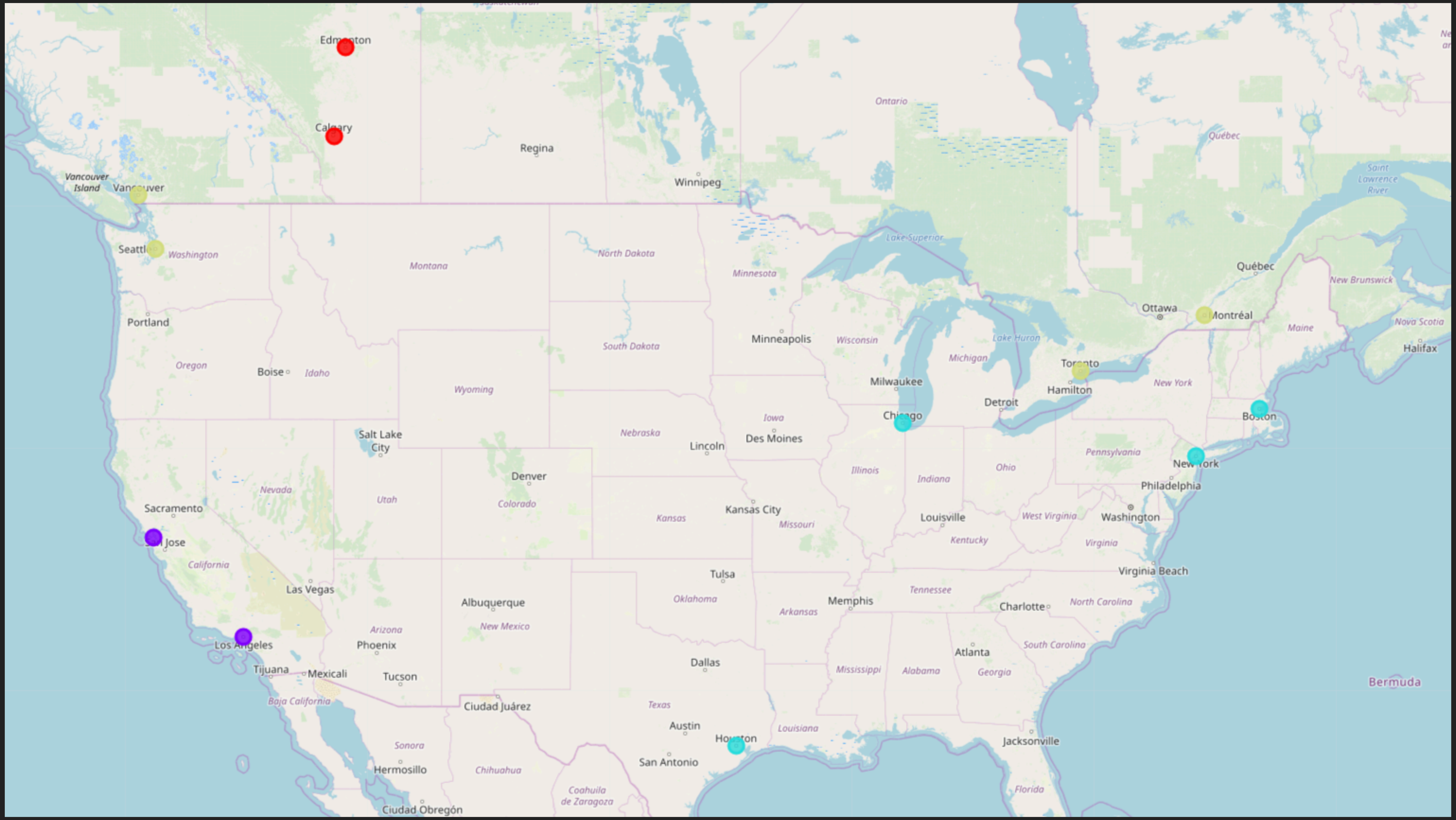
ANALYSIS RESULTS



CLIMATE DENDROGRAM

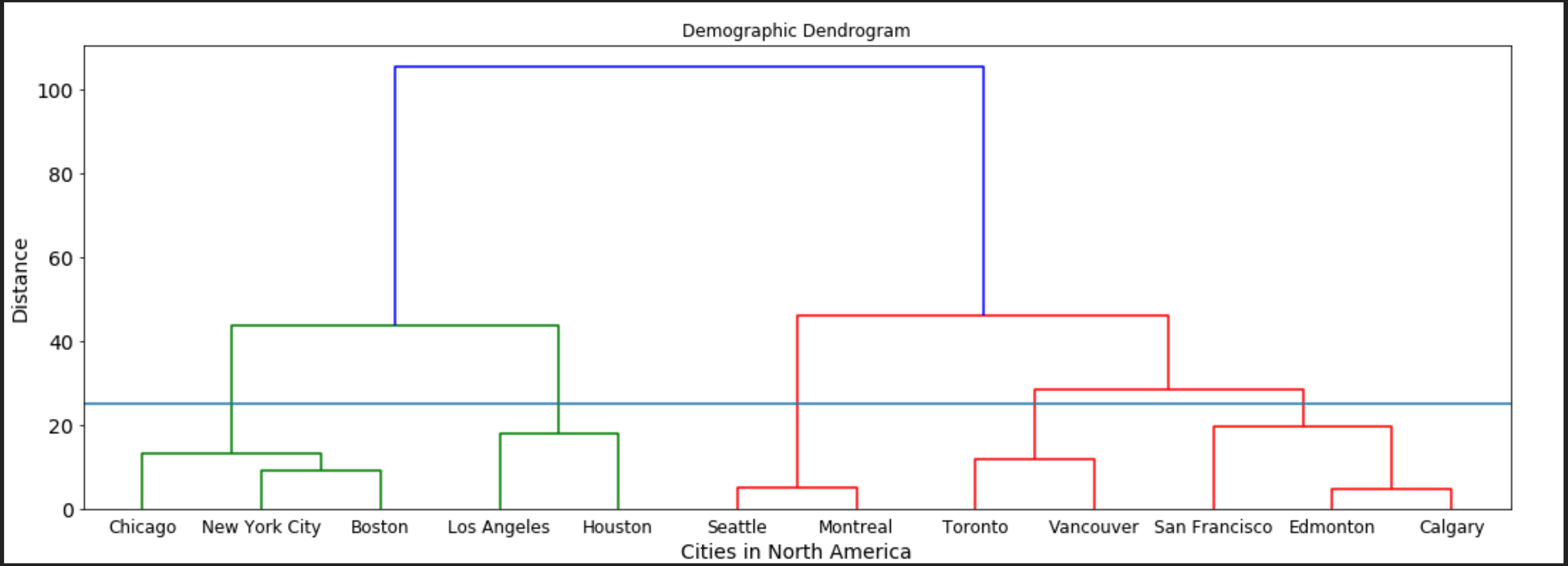


CLIMATE CLUSTER

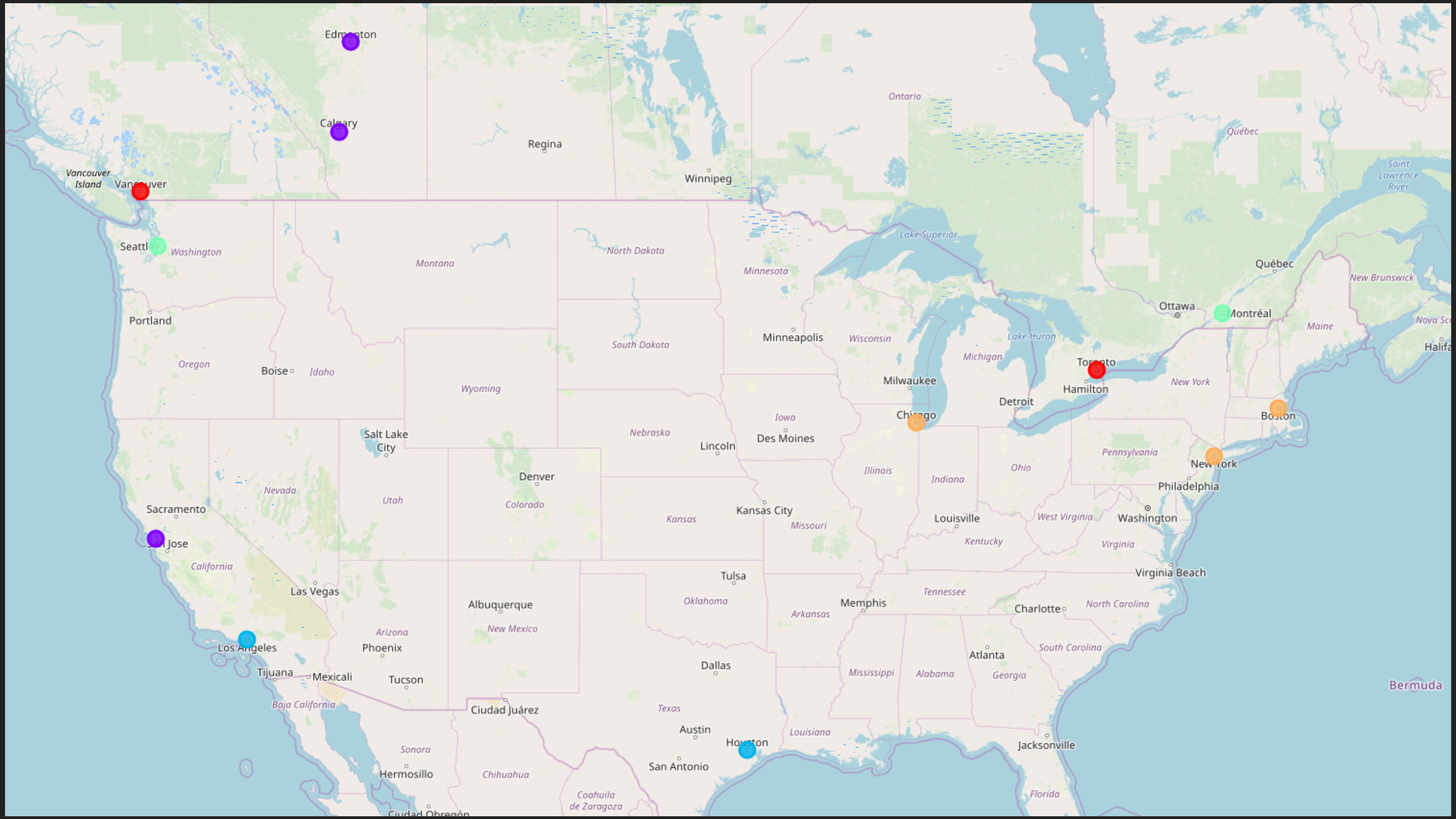


- climate clustering has significant geolocation influence.

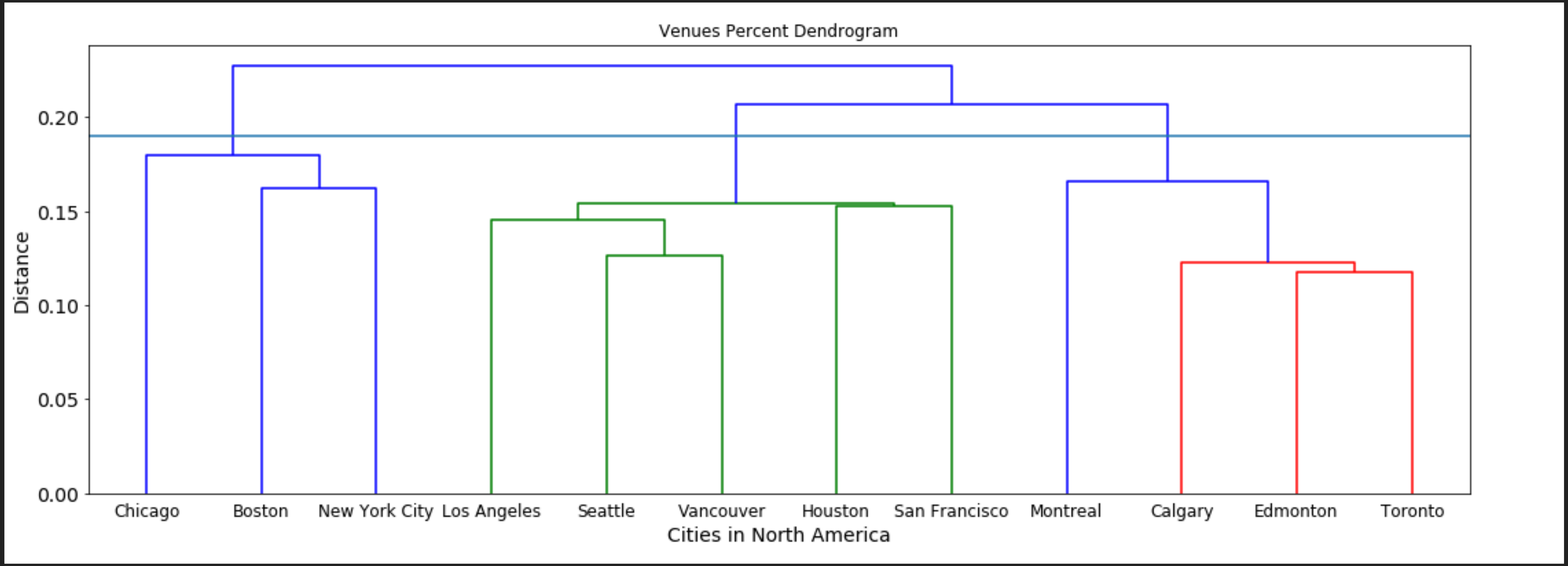
DEMOGRAPHIC DENDROGRAM



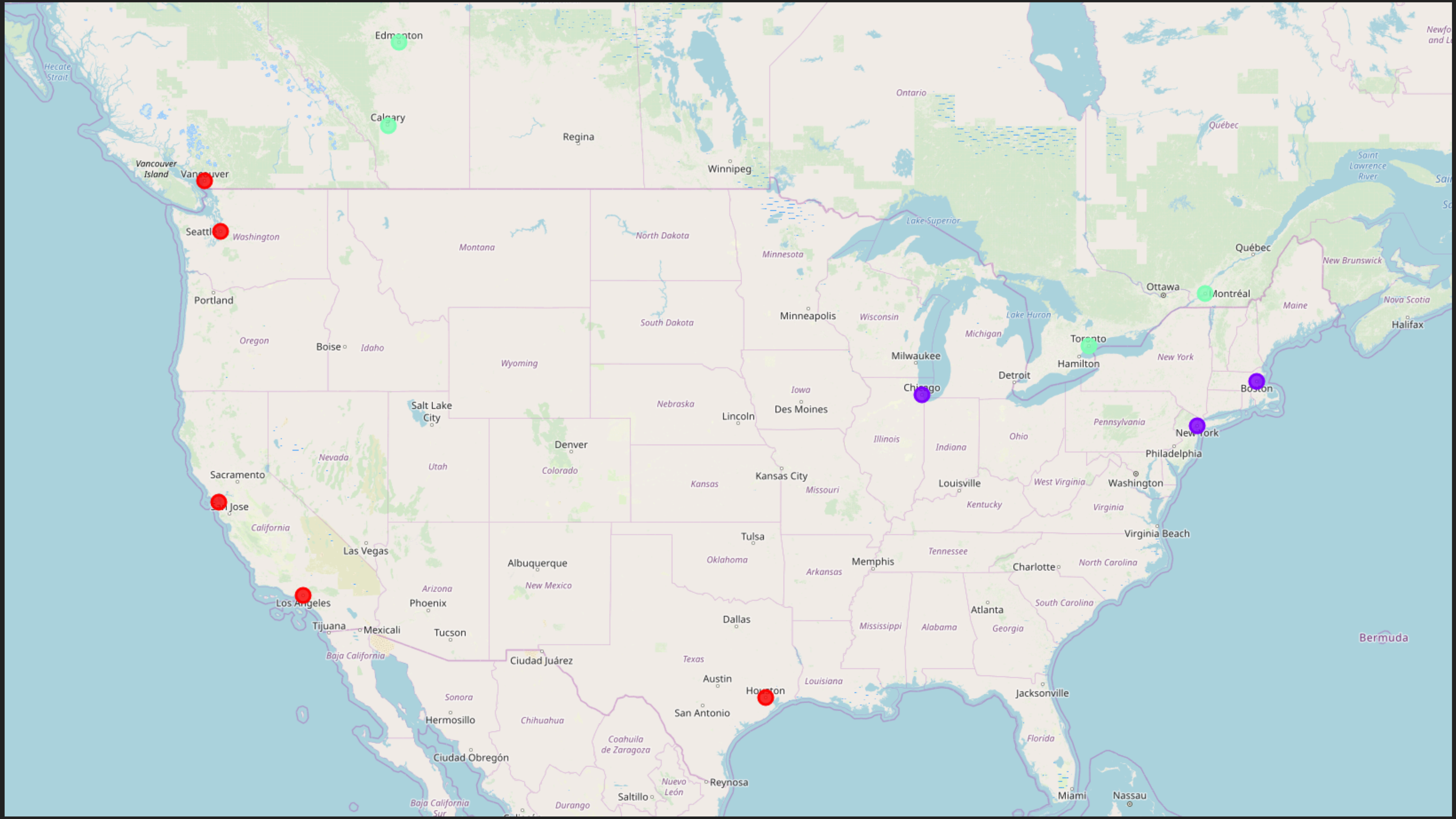
DEMOGRAPHIC CLUSTER



VENUES DENDROGRAM



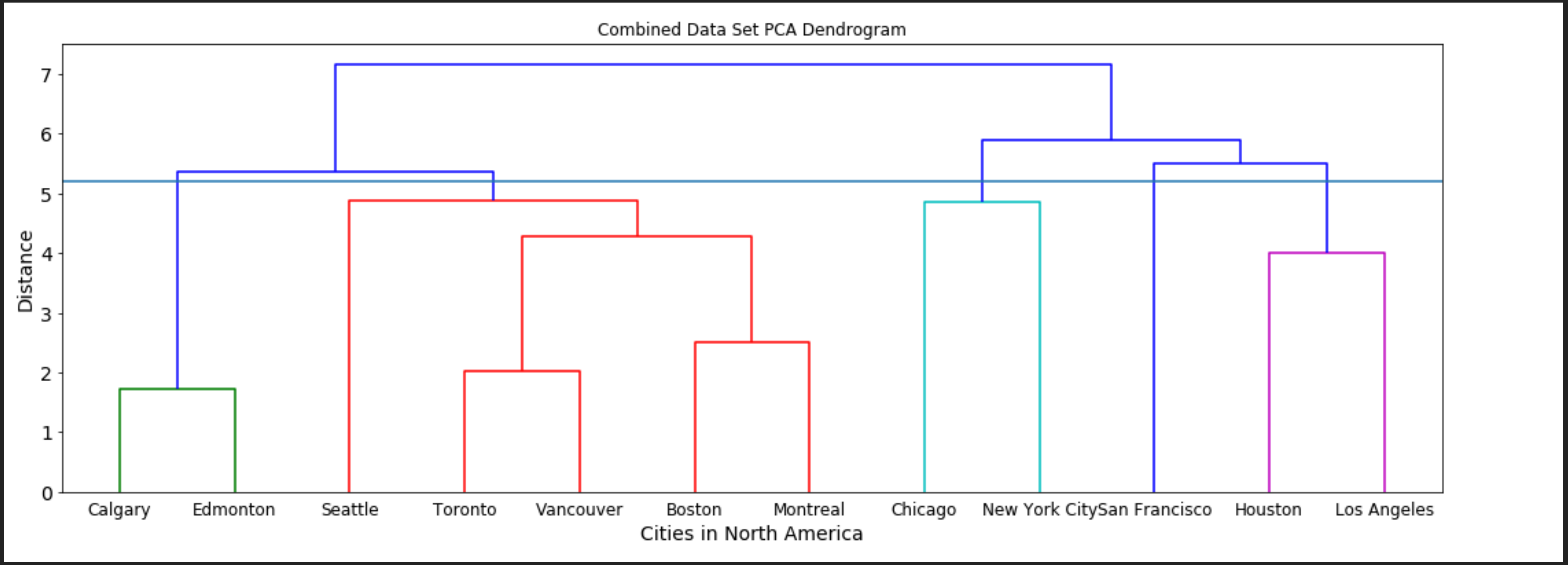
VENUES CLUSTERING



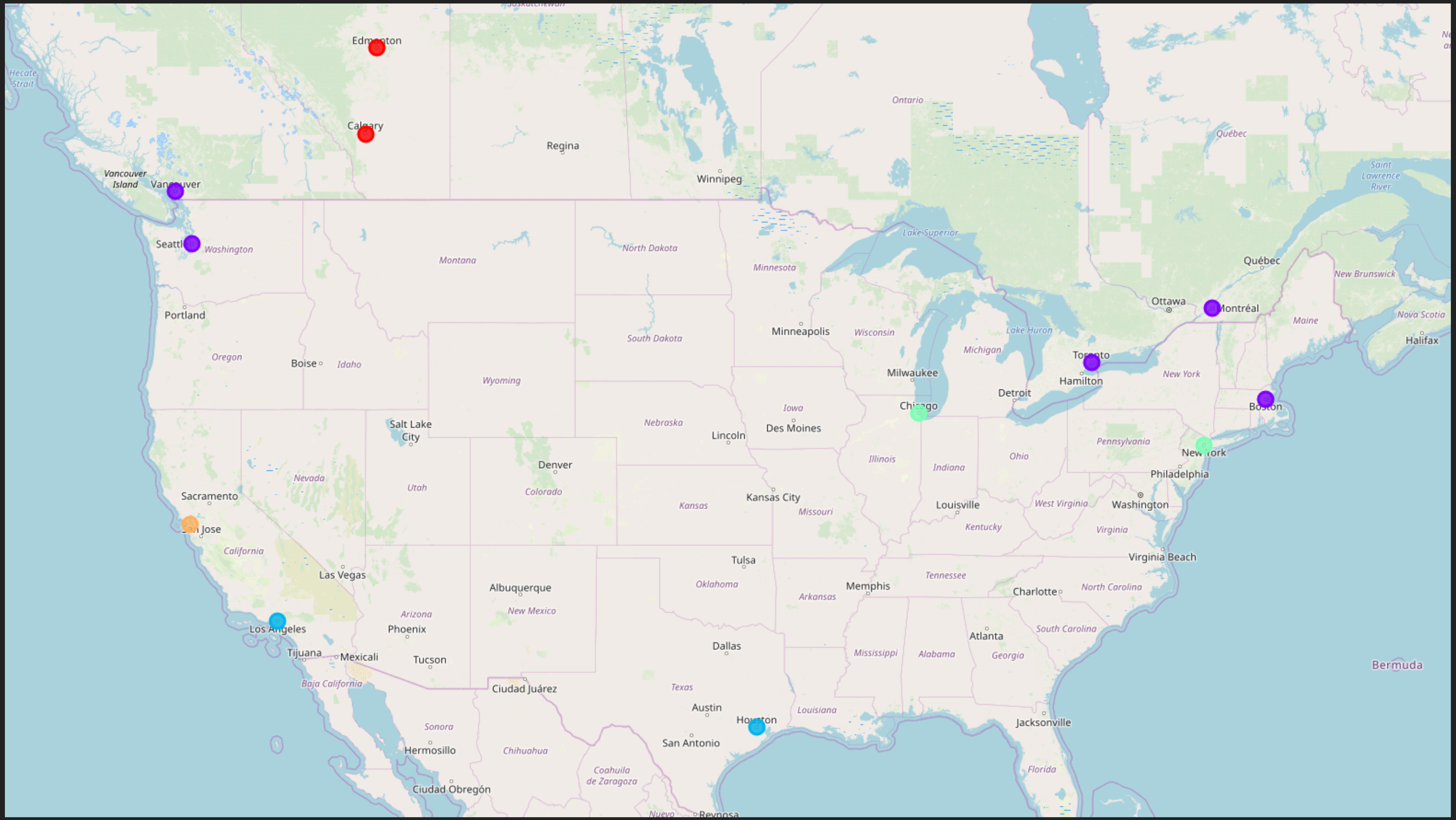
COMBINE ALL THREE DATA SET

- Use Principle Component Analysis(PCA) to reduce Dimension
- Analyze Combined data with Hierarchical Clustering
- Cluster Into Groups

COMBINE ALL THREE DATA DENDROGRAM



COMBINE DATA CLUSTERING



CONCLUSION

- We can conclude that venue category together with climate and demographic all have geolocation influence on it.

THANK YOU FOR REVIEWING