Analysis of COVID-19 cases in Chicago Area

IBM Data Science Professional Certificate

Capstone Project

Ibrahim Benjelloun - November 11th, 2020

Summary

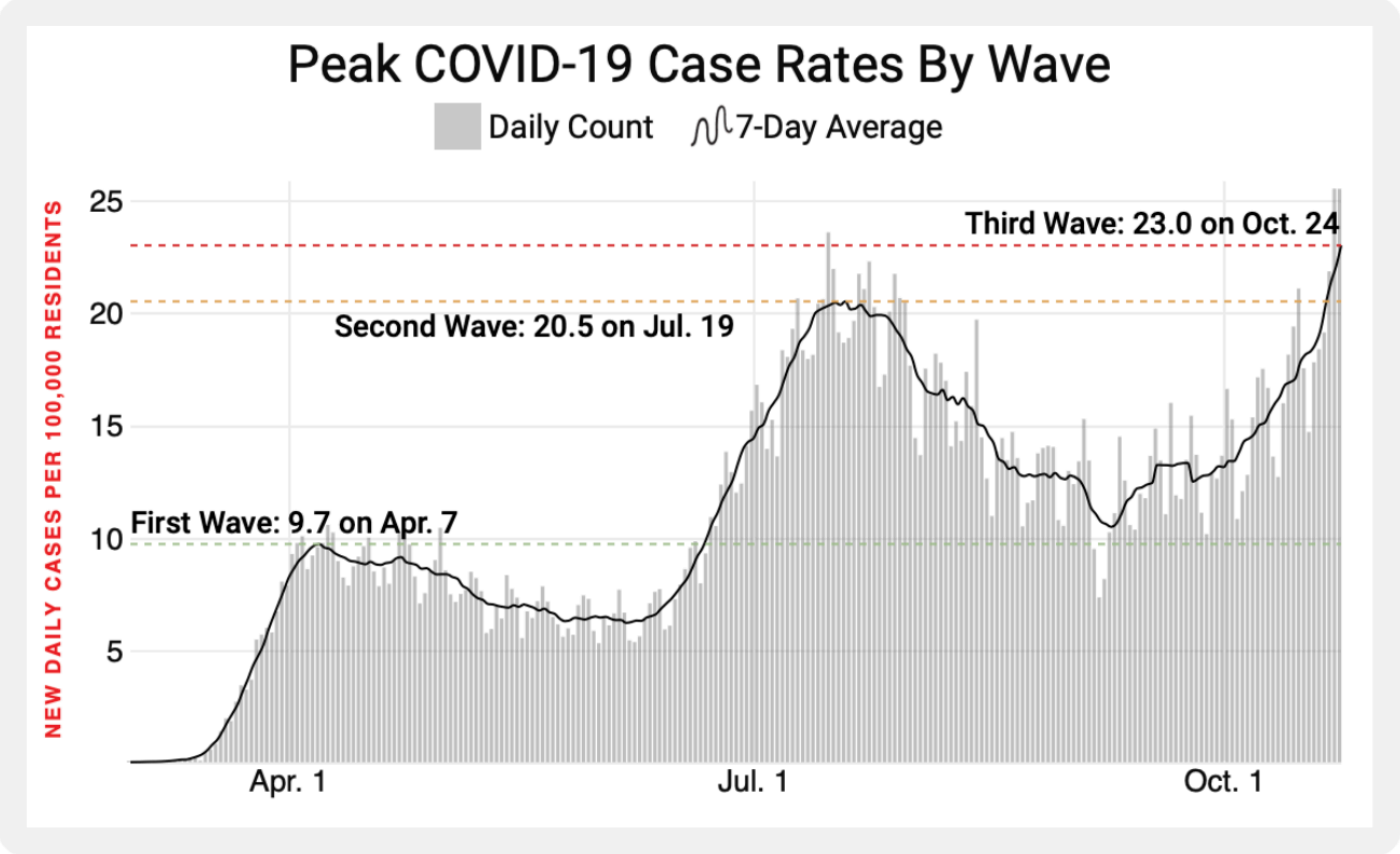
- Introduction
- Data
- Machine Learning: K-Means clustering
- Results & Discussion
- Conclusion

Introduction

Introduction

Key numbers

- The U.S. is the most impacted country from COVID-19
- 9 millions cases and 230,000 deaths
- Illinois is the 5th ranked state in terms of cases with 436,000 cases



Introduction

Objective

“ Study an eventual relation between the concentration of certain types of venues in the Chicago area with the amount of the registered COVID-19 cases ”

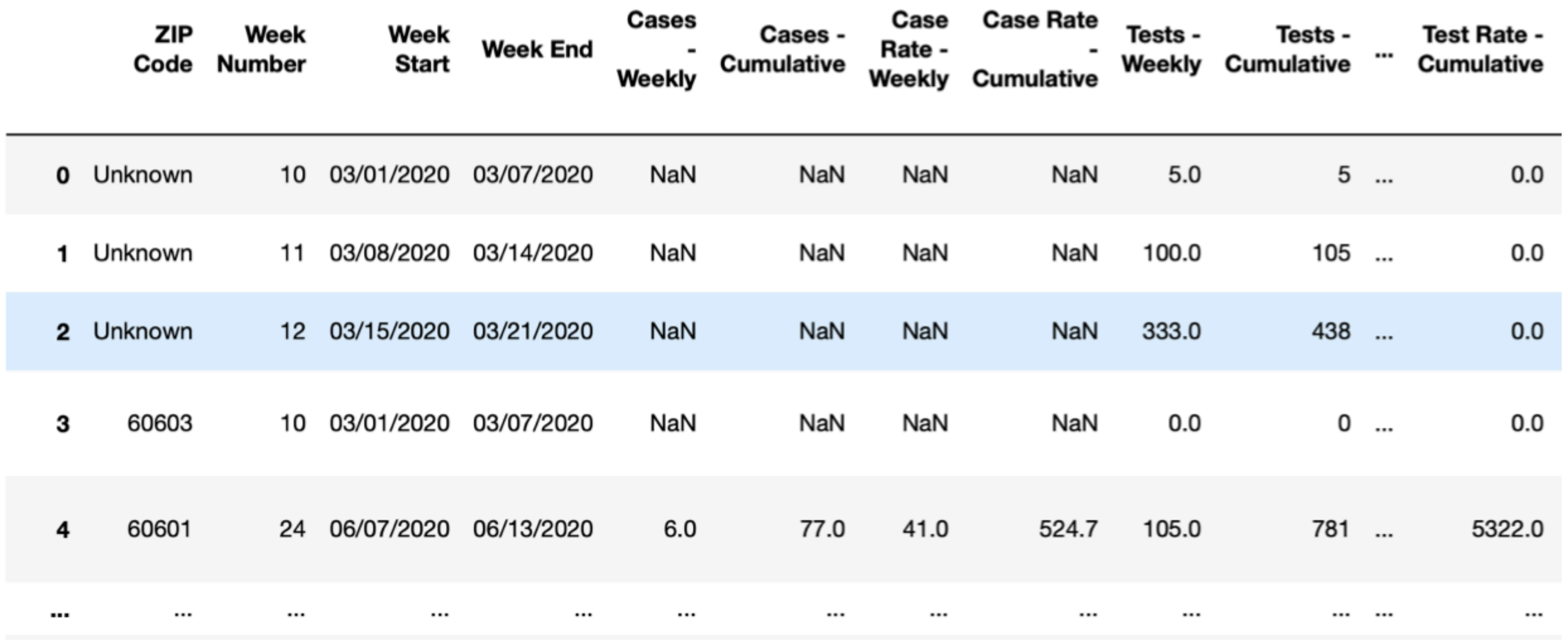
Data

Data (1/2)

COVID-19 data for the city of Chicago

Key information retrieved

- **ZIP code**: Home ZIP code of the cases and people tested
- **Case Rate Cumulative**: Cumulative case rate per 100,000 population
- **Coordinates**: geographic coordinates for the ZIP code



	ZIP Code	Week Number	Week Start	Week End	Cases - Weekly	Cases - Cumulative	Case Rate - Weekly	Case Rate - Cumulative	Tests - Weekly	Tests - Cumulative	...	Test Rate - Cumulative
0	Unknown	10	03/01/2020	03/07/2020	NaN	NaN	NaN	NaN	5.0	5	...	0.0
1	Unknown	11	03/08/2020	03/14/2020	NaN	NaN	NaN	NaN	100.0	105	...	0.0
2	Unknown	12	03/15/2020	03/21/2020	NaN	NaN	NaN	NaN	333.0	438	...	0.0
3	60603	10	03/01/2020	03/07/2020	NaN	NaN	NaN	NaN	0.0	0	...	0.0
4	60601	24	06/07/2020	06/13/2020	6.0	77.0	41.0	524.7	105.0	781	...	5322.0
...	...	...	...	...	...	...	...	...	...	...	...	...

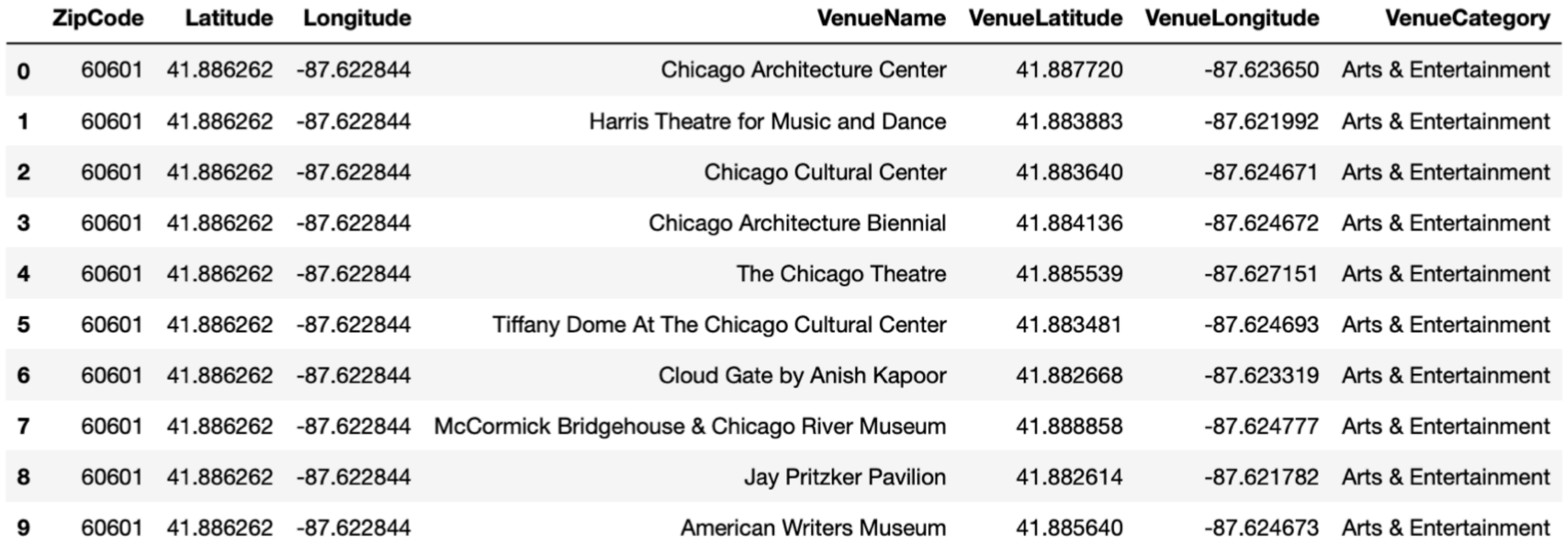
Source: Chicago Data Portal

Data (2/2)

Venues data (Foursquare API)

Key information retrieved

- **ZIP code**: ZIP code of the venue
- **Venue location**: geographic coordinates of each venue
- **Venue category**: Food, Arts & Entertainment, Shop & Service, Professional & Other Places, etc.



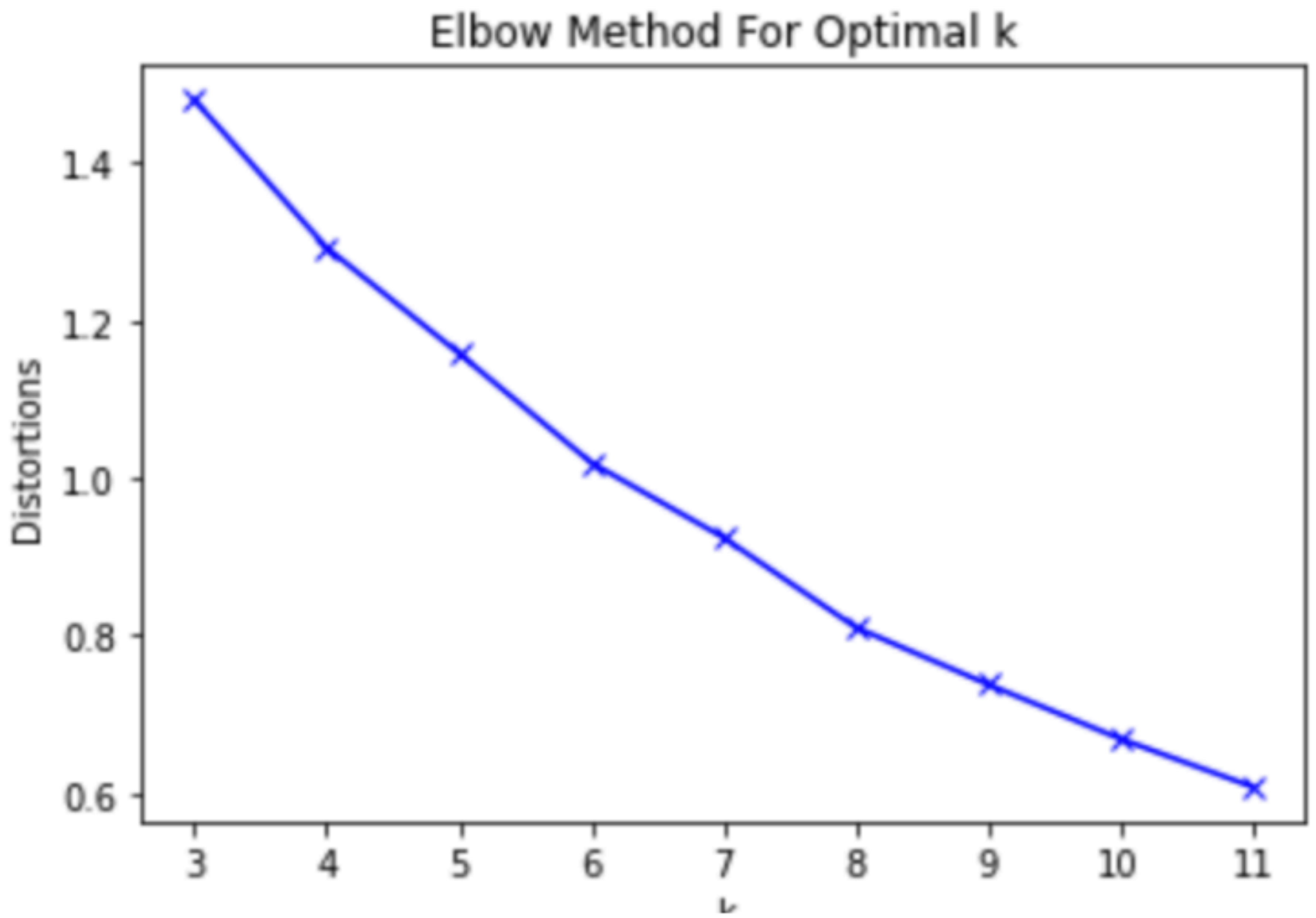
	ZipCode	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	60601	41.886262	-87.622844	Chicago Architecture Center	41.887720	-87.623650	Arts & Entertainment
1	60601	41.886262	-87.622844	Harris Theatre for Music and Dance	41.883883	-87.621992	Arts & Entertainment
2	60601	41.886262	-87.622844	Chicago Cultural Center	41.883640	-87.624671	Arts & Entertainment
3	60601	41.886262	-87.622844	Chicago Architecture Biennial	41.884136	-87.624672	Arts & Entertainment
4	60601	41.886262	-87.622844	The Chicago Theatre	41.885539	-87.627151	Arts & Entertainment
5	60601	41.886262	-87.622844	Tiffany Dome At The Chicago Cultural Center	41.883481	-87.624693	Arts & Entertainment
6	60601	41.886262	-87.622844	Cloud Gate by Anish Kapoor	41.882668	-87.623319	Arts & Entertainment
7	60601	41.886262	-87.622844	McCormick Bridgehouse & Chicago River Museum	41.888858	-87.624777	Arts & Entertainment
8	60601	41.886262	-87.622844	Jay Pritzker Pavilion	41.882614	-87.621782	Arts & Entertainment
9	60601	41.886262	-87.622844	American Writers Museum	41.885640	-87.624673	Arts & Entertainment

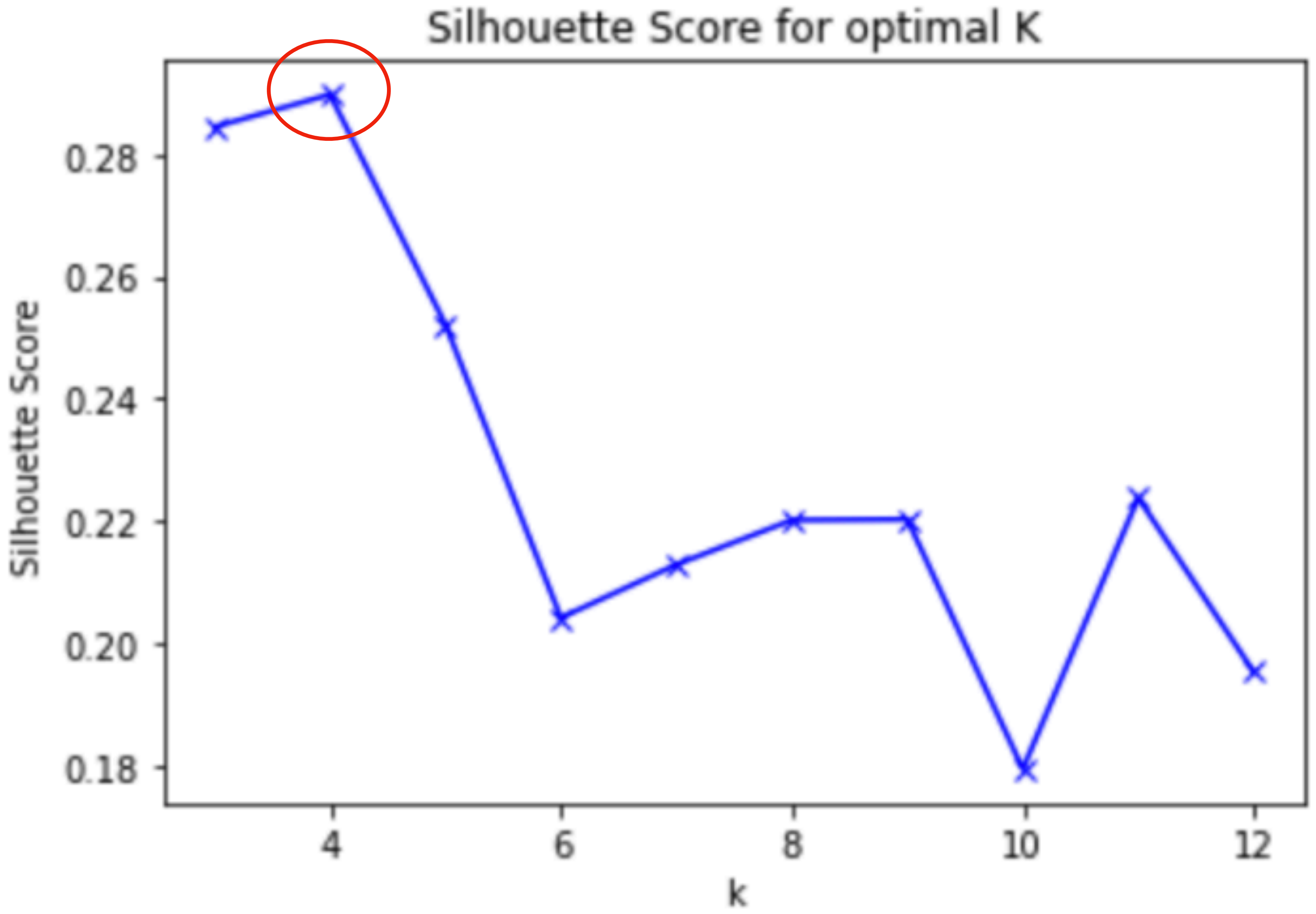
Source: Foursquare API

Machine Learning: K-Means clustering

Machine Learning: K-Means clustering

Selecting the optimal “K” (number of clusters)

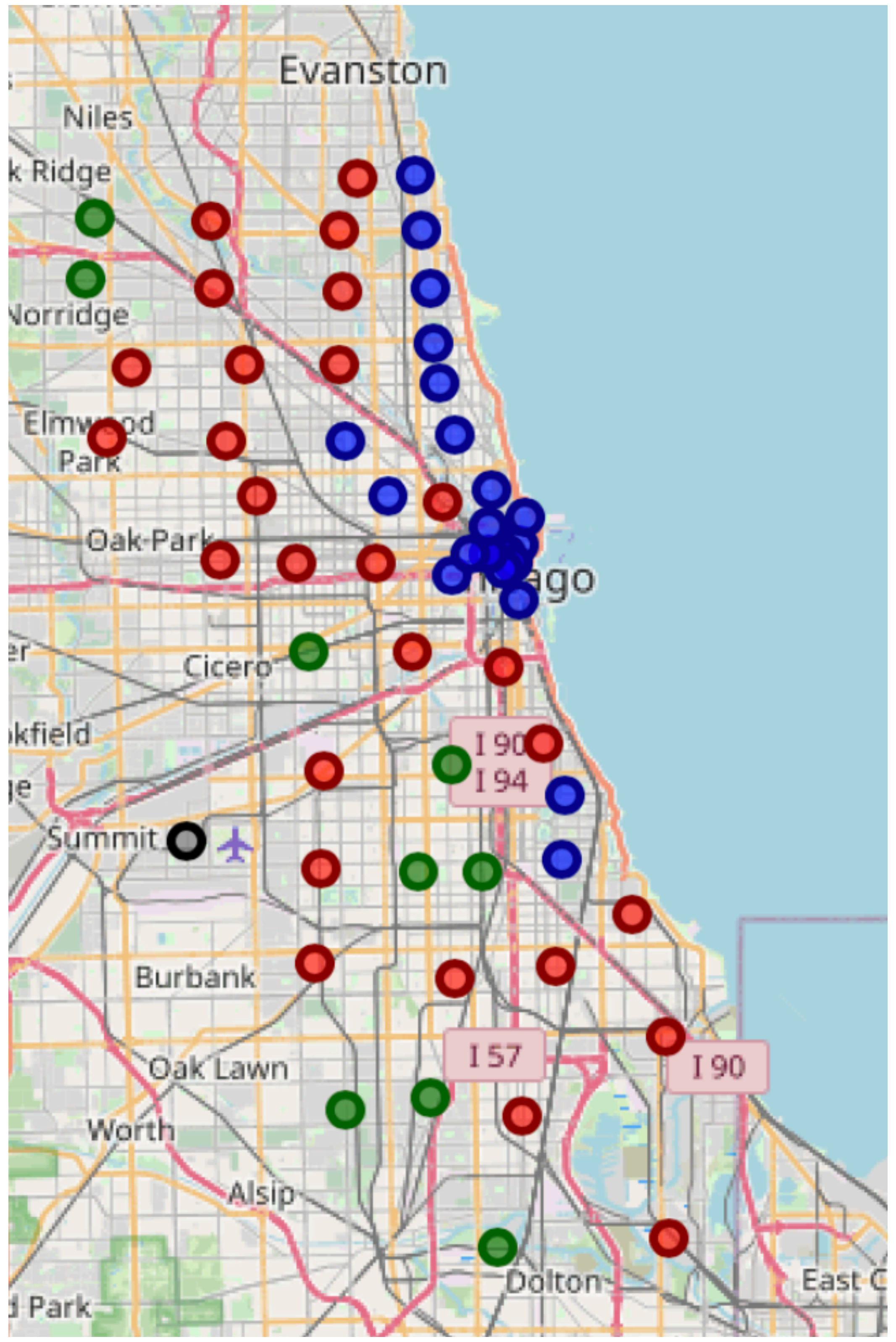




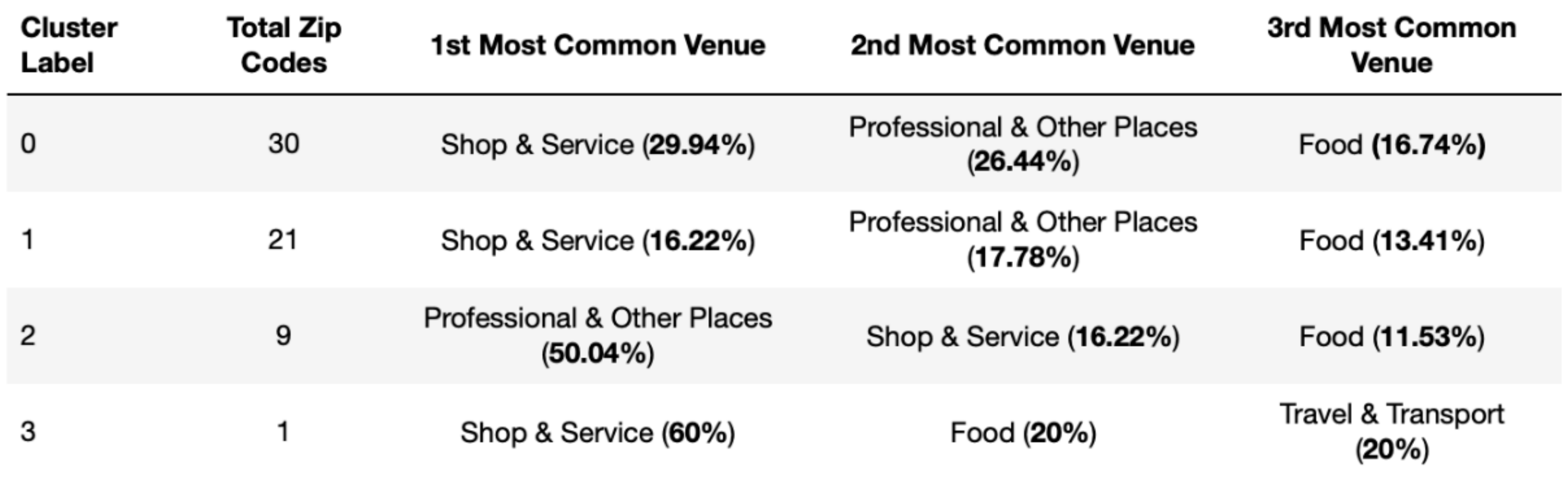
- **Elbow method**: difficult to distinguish a clear optimal “K”
- **Silhouette method**: presents a clear pic at **K=4**

Machine Learning: K-Means clustering

Resulting clusters



Red: “Cluster 0”, Blue: “Cluster 1”, Green: “Cluster 2”, Black: “Cluster 3”



Cluster Label	Total Zip Codes	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue
0	30	Shop & Service (**29.94%**)	Professional & Other Places (**26.44%**)	Food (**16.74%**)
1	21	Shop & Service (**16.22%**)	Professional & Other Places (**17.78%**)	Food (**13.41%**)
2	9	Professional & Other Places (**50.04%**)	Shop & Service (**16.22%**)	Food (**11.53%**)
3	1	Shop & Service (**60%**)	Food (**20%**)	Travel & Transport (**20%**)

Most common venues

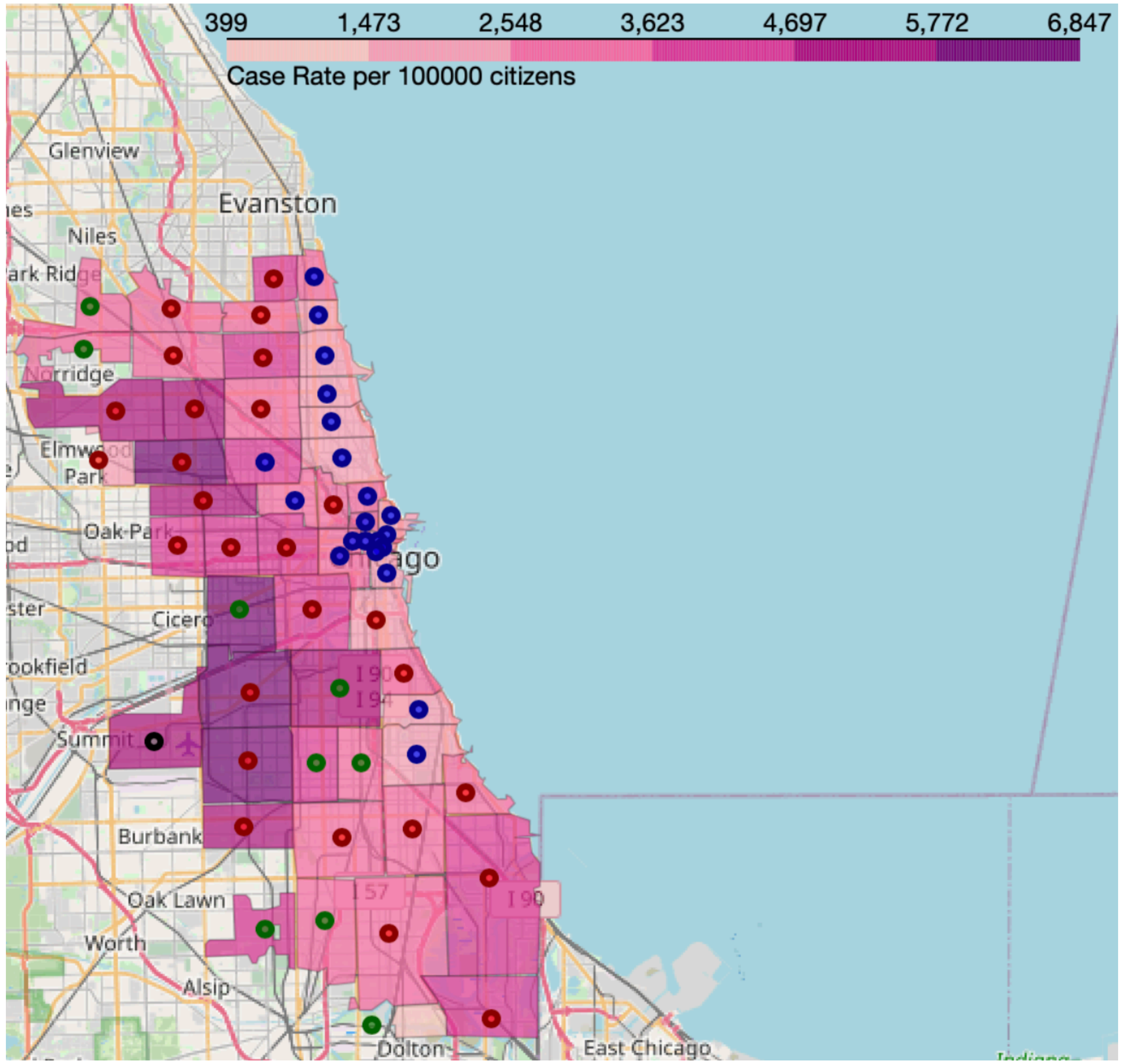
“Shop & Service”, “Professional & Other Places”, “Food” and “Travel & Transport”

Machine Learning: K-Means clustering

Chicago clusters vs COVID-19 data (1/2)

- Case rates go from 399 to 6,847 per 100,000 population
- Case rates are not homogeneous within clusters except for Clusters 1 and 3
- Cluster 1 appears to be the less impacted one

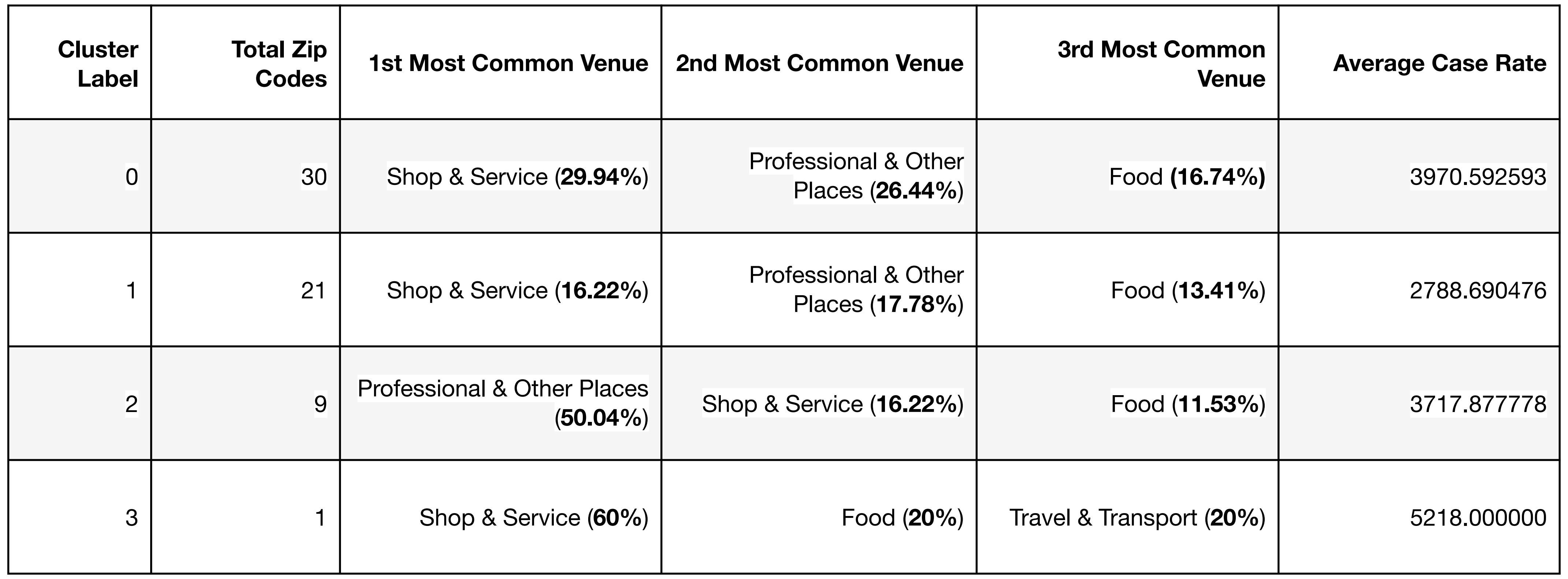
Chicago clusters and Covid case rates per ZIP code



Red: “Cluster 0”, Blue: “Cluster 1”, Green: “Cluster 2”, Black: “Cluster 3”

Machine Learning: K-Means clustering

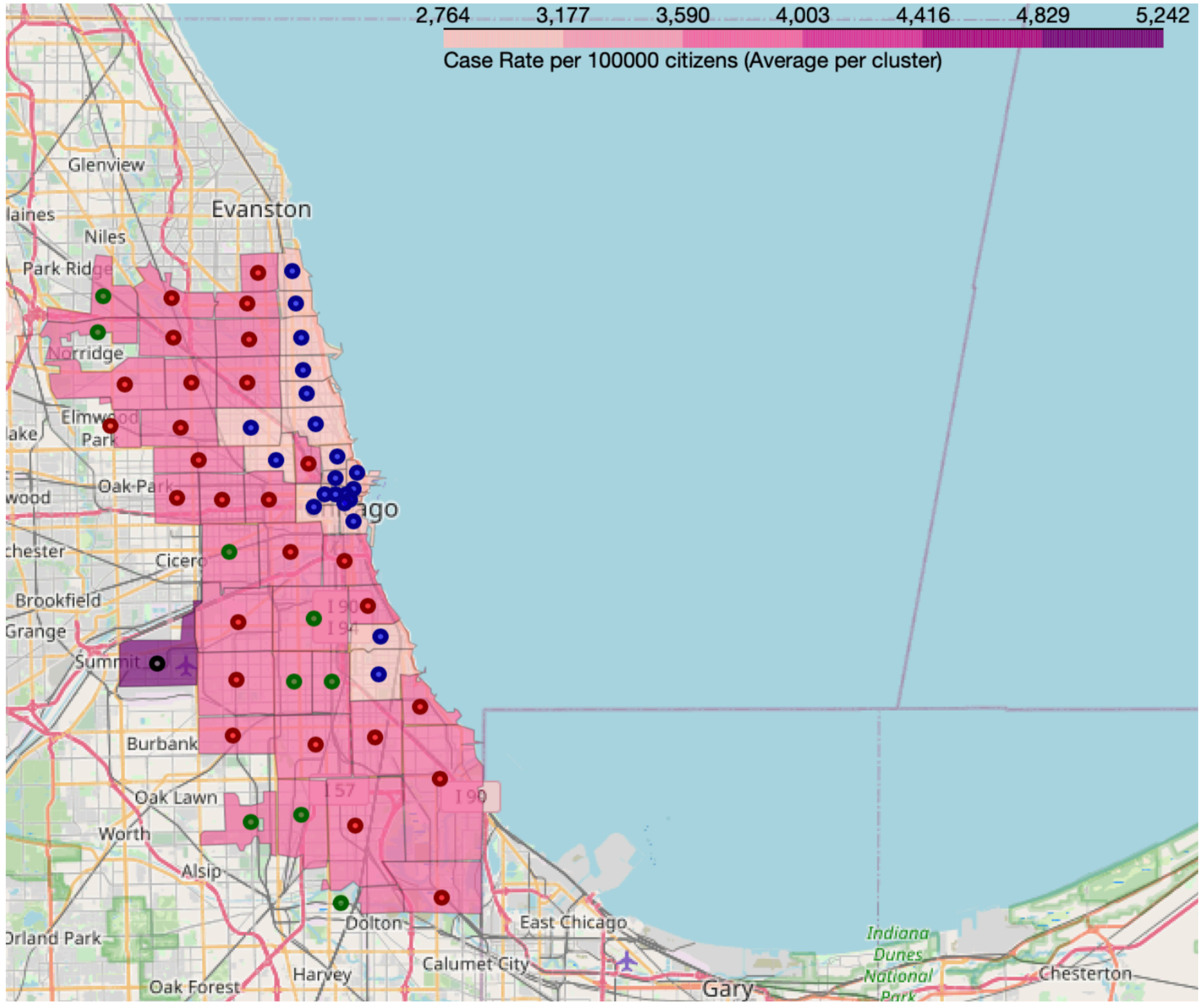
Chicago clusters vs COVID-19 data (2/2)



Cluster Label	Total Zip Codes	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	Average Case Rate
0	30	Shop & Service **(29.94%)**	Professional & Other Places **(26.44%)**	Food **(16.74%)**	3970.592593
1	21	Shop & Service **(16.22%)**	Professional & Other Places **(17.78%)**	Food **(13.41%)**	2788.690476
2	9	Professional & Other Places **(50.04%)**	Shop & Service **(16.22%)**	Food **(11.53%)**	3717.877778
3	1	Shop & Service **(60%)**	Food **(20%)**	Travel & Transport **(20%)**	5218.000000

Summary of clusters with average case rates

Chicago clusters and Covid average case rates per cluster



Red: “Cluster 0”, Blue: “Cluster 1”, Green: “Cluster 2”, Black: “Cluster 3”

Results and Discussion

Results & Discussion

Results

- **“Cluster 1”** is the least impacted cluster with an average of 2788 cases per 100,000 people. It is the most homogenous one as the frequencies of each venue type varies between 10% and 16% —> **Residential areas**

- Following is the **“Cluster 2”** with an average of 3717 case rate. The main difference with “Cluster 1” is the considerable proportion of “Professional & Other Places” venues with 50% —> **Business areas**

- “Cluster 0” is slightly more impacted than “Cluster 2” with 3970 average case rate. In comparison with “Cluster 1”, “Shop & Service” and “Professional & Other Places” venues proportions have shifted to nearly 30% each —> **Commercial areas**

- “Cluster 3” is the most impacted one with 5218 average case rate. This could be interpreted with the very important proportion of “Shop & Service” venues (60%) and also a relatively large proportion of “Travel & Transport” venues —> **Airport zones**

Results & Discussion

Discussion

- Although Foursquare data gave us some useful directions for our study, it lacks some insightful informations such as the size of a venue, its daily traffic, etc.
- A more interesting and complete approach would be to explore COVID data at individual level, with the type of venues visited during a specified time frame (7 days for examples) before the testing positive.
- Combining these data would be much more insightful although more complex.

Conclusion

Conclusion

- The results obtained in this study are high level and more qualitative than quantitative.
- They confirm some of our intuitions regarding the risk of getting infected depending on the type of area we are.
- An interesting thing to do would be to confirm these results by comparing the data of another area and see if our conclusions apply there.

https://github.com/i-benjelloun/Coursera_Capstone