CLUSTERING THE FIFTY LARGEST METROPOLITAN AREAS IN THE WORLD

INTRODUCTION

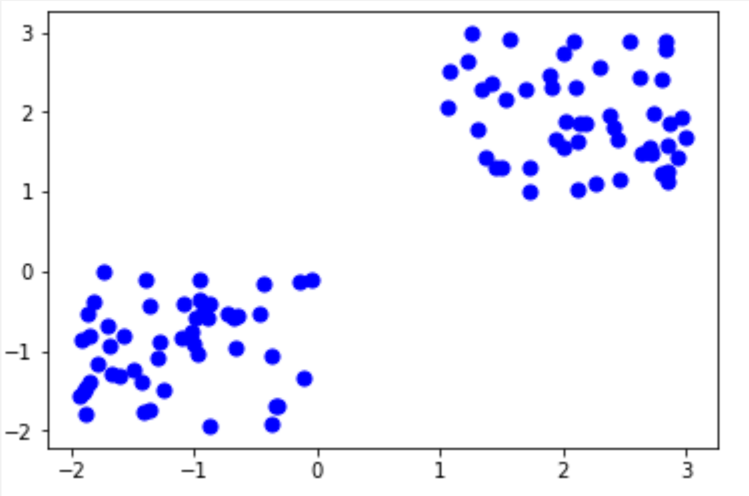
- I am running an immigration agency in Paris, France.
- Many of my clients who want to start a new life ask me which cities are most culturally similar to Paris.
- Clustering is used to find cities most similar to Paris

DATA

- Data is scraped from Wikipedia using the BeautifulSoup Library
- Latitudes and Longitudes obtained via the Geocoders Library

METHODOLOGY

- K-means Clustering is used
  - 10 groups
  - Frequency of each category is used to calculate Euclidean Distance



RESULTS

Paris was clustered with the following cities

Tokyo	Japan	3
São Paulo	Brazil	3
Istanbul	Turkey	3
Rio de Janeiro	Brazil	3
Los Angeles	United States	3
Moscow	Russia	3
Paris	France	3
Seoul	South Korea	3
Nagoya	Japan	3
Tehran	Iran	3
Chicago	United States	3

DISCUSSION/CONCLUSION

- Found the clustering pretty arbitrary
- Next Steps: Try different factors (per-capita GDP, tourism numbers, urban density), and try different number of cluster groups
- Overall, great experience with learning the entire data science pipeline