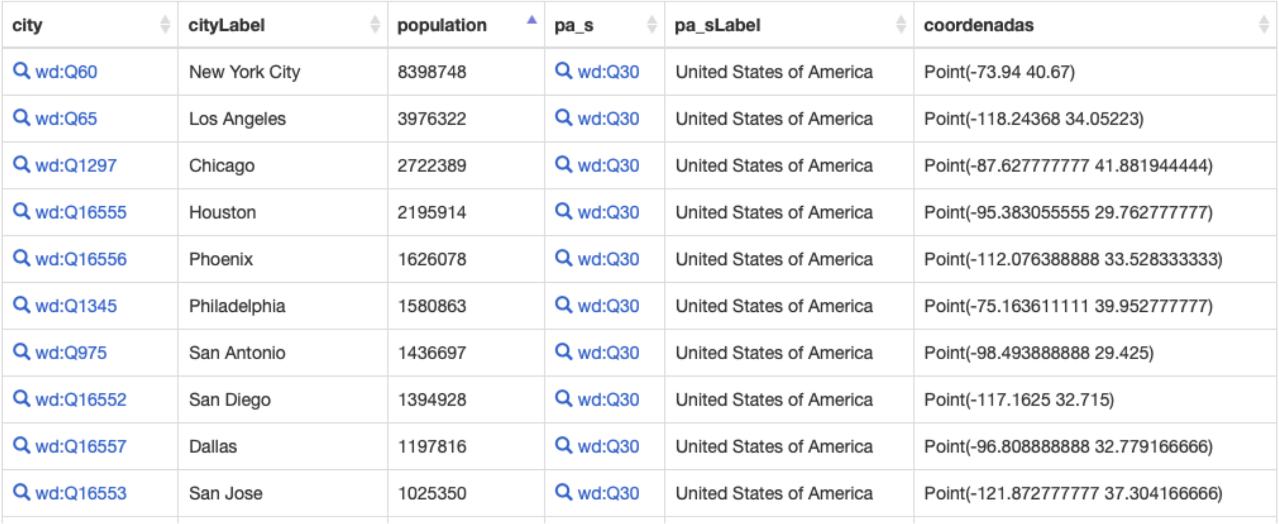
3. Data

A list of the cities and they coordinates in Wikidata is used. The data is filtered with country name of United State of America and sorted with population. The results is download and saved as .csv file (see pic 3.1).



city	cityLabel	population	pa_s	pa_sLabel	coordenadas
wd:Q60	New York City	8398748	wd:Q30	United States of America	Point(-73.94 40.67)
wd:Q65	Los Angeles	3976322	wd:Q30	United States of America	Point(-118.24368 34.05223)
wd:Q1297	Chicago	2722389	wd:Q30	United States of America	Point(-87.627777777 41.881944444)
wd:Q16555	Houston	2195914	wd:Q30	United States of America	Point(-95.383055555 29.762777777)
wd:Q16556	Phoenix	1626078	wd:Q30	United States of America	Point(-112.076388888 33.528333333)
wd:Q1345	Philadelphia	1580863	wd:Q30	United States of America	Point(-75.163611111 39.952777777)
wd:Q975	San Antonio	1436697	wd:Q30	United States of America	Point(-98.493888888 29.425)
wd:Q16552	San Diego	1394928	wd:Q30	United States of America	Point(-117.1625 32.715)
wd:Q16557	Dallas	1197816	wd:Q30	United States of America	Point(-96.808888888 32.779166666)
wd:Q16553	San Jose	1025350	wd:Q30	United States of America	Point(-121.872777777 37.304166666)

Pic 3.1 Top Ten Population Cities in USA

It had to be cleaned by splitting the coordinates column into Longitude and Latitude.

The venues for each cities can be queried by using the Foursquare API.

The two set of data can be merged into pandas data frame.

After creating the new dataset with cities, they are plotted on a map using Folium to check if the coordinates were ok.

A new dataframe with the most common venues for each city will be created.

K-Means and Hierarchical Cluster will be used to analyse. The results will be plotted on a Folium Map with colors for each cluster.