Assignment 2: Data Wrangling - Metropolitan Segregation

Kai Yoshino

Info 370

Introduction

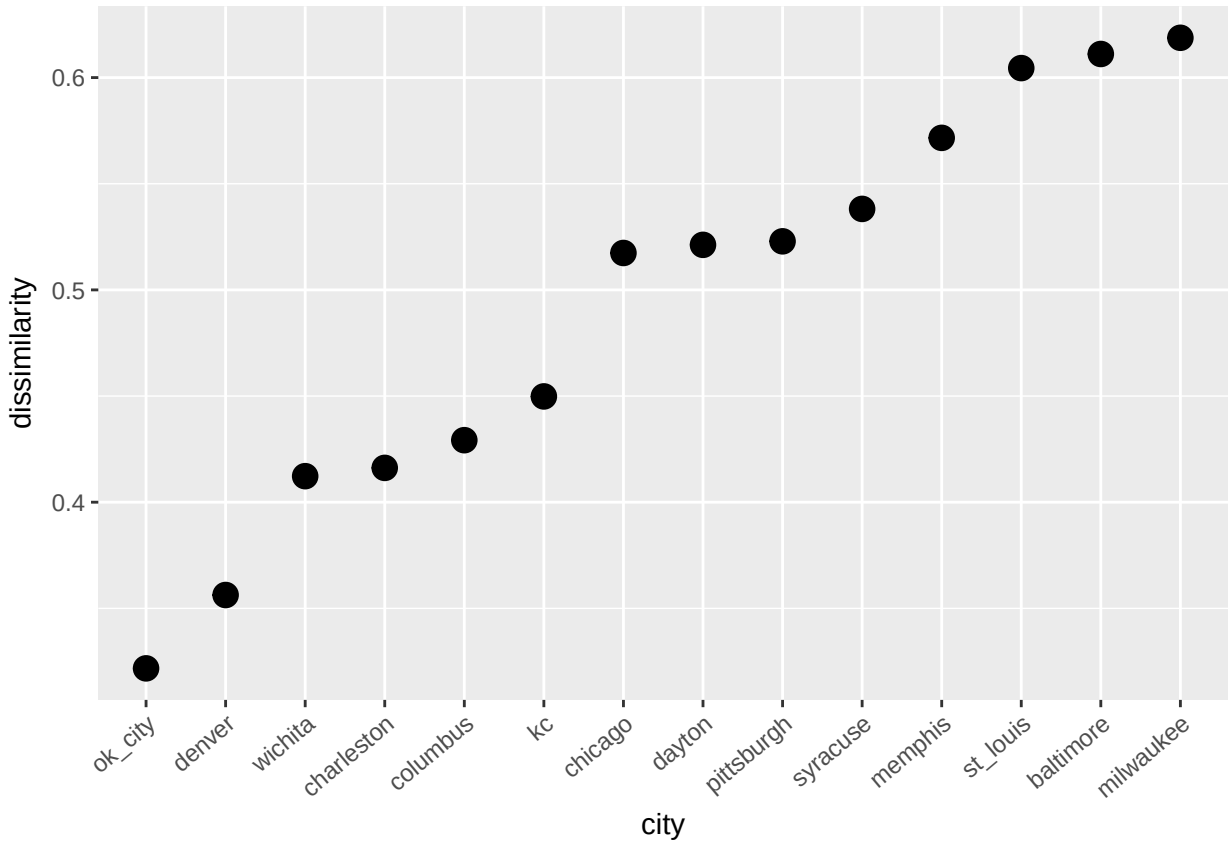
The purpose of this assignment was to aggregate and interpret an metropolitan demographics data set in order to measure segregation of each city within it. The dataset included 14 different cities. For each city, I calculated four different metrics that measure different aspects of segregation. I will go over my calulations, show how they measure segregation, and how the results for each city compare to the rest.

Dissimilarity

First I measured the Dissimilarity index. This measures what percentage of the population of a city would have to move for each neighborhood to have an equal ratio of white to non white residents. Having high dissimilairty is a common indicator of segregation. The formula is as shown.

$$\frac{\sum_{i=1}^{n}[t_i|p_i - P|]}{[2TP(1-P)]}$$

To calculate you take the sum of the total population (t_i), multiplied by the absolute value of the ratio of minority population to the total population (p_i) for each area, minus that ratio for the whole city(P), then divide that sum by two, multiplied by the total population, multiplied by the P, and $1 - P$



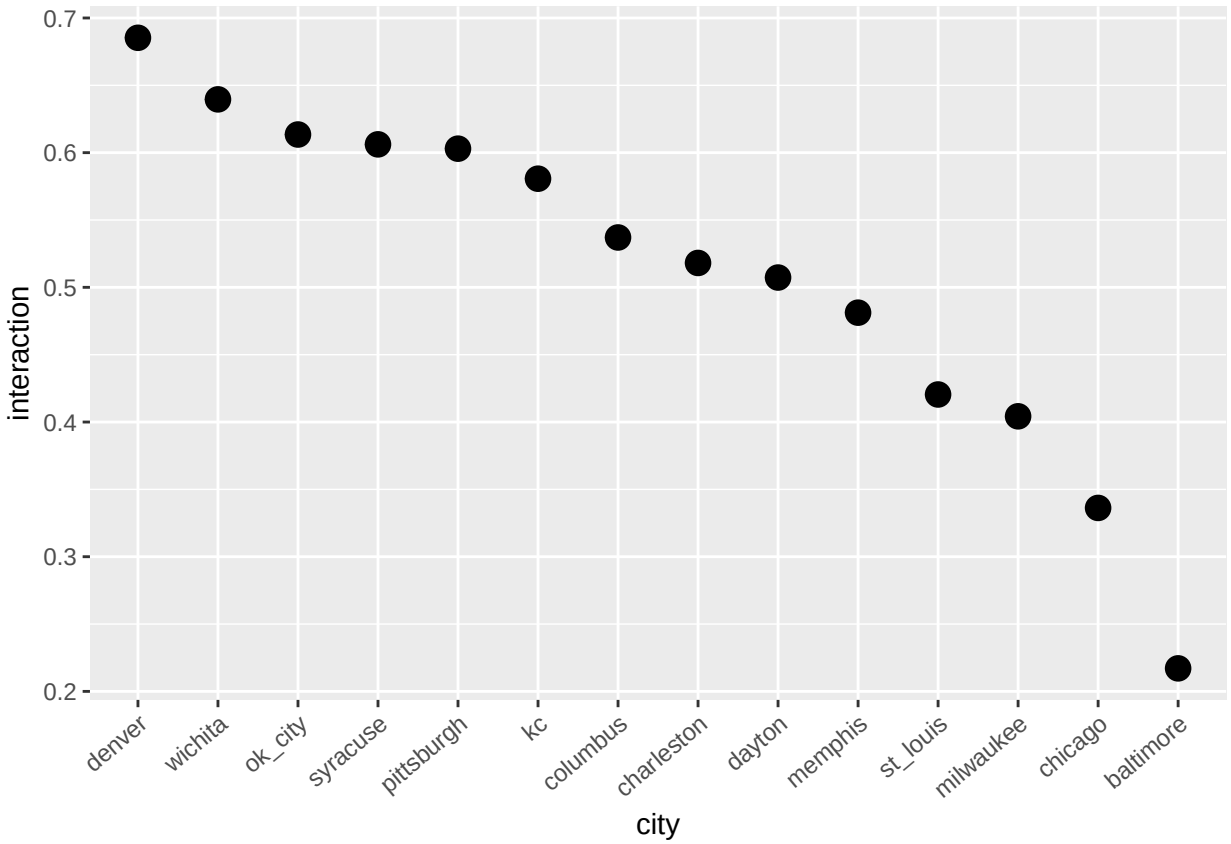
We see from the plot that Oklohoma City has the least dissimilarity, and Milwaukee has the most.

Interaction

Next I looked at interaction. This measures the probability that a minority person shares a neighborhood with a majority group person. Having a low measure of interaction is a sign of segregation.

$$\sum_{i=1}^{n}[(\frac{x_i}{X})(\frac{y_i}{t_i})]$$

To calculate I took the sum of the minority population of each area (x_i), divided by the minority population of the city(X), multiplied by the majority population of each area (y_i), divided by t_i.



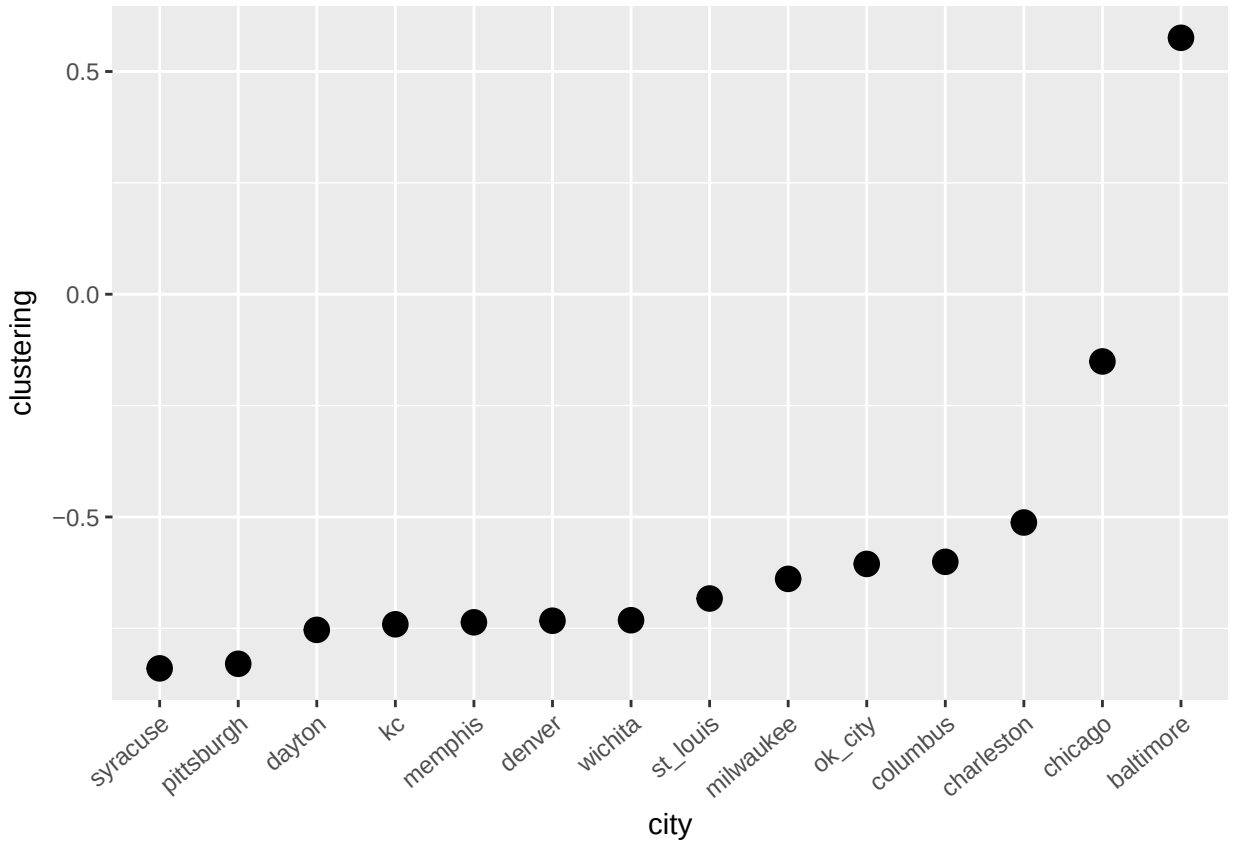
This measure shows Baltimore having the least interaction and Denver having the most.

Clustering

Next I looked at Relative Clustering, which measures the average distance between minority members with the average distance between majority members. Having a positive clustering measure means that minorities display greater clustering than the majority and having negative means the majority clusters more, clustering of 0 means both gourps cluster equally. This metric is a bit more interesting because it gives context to how cities might be segregated, weighting segregation of the minority or majority groups. It doesn't as easily compare to the other metrics I have looked at here, but gives more context.

$$(\frac{P_{xx}}{P_{yy}}) - 1$$

This metric is calculated by taking the ratio of ratio of the entire city that is in the minority (P_{xx}), then divide it by the ratio of the city's population that is in the majority (P_{yy}), and subtract one from it.



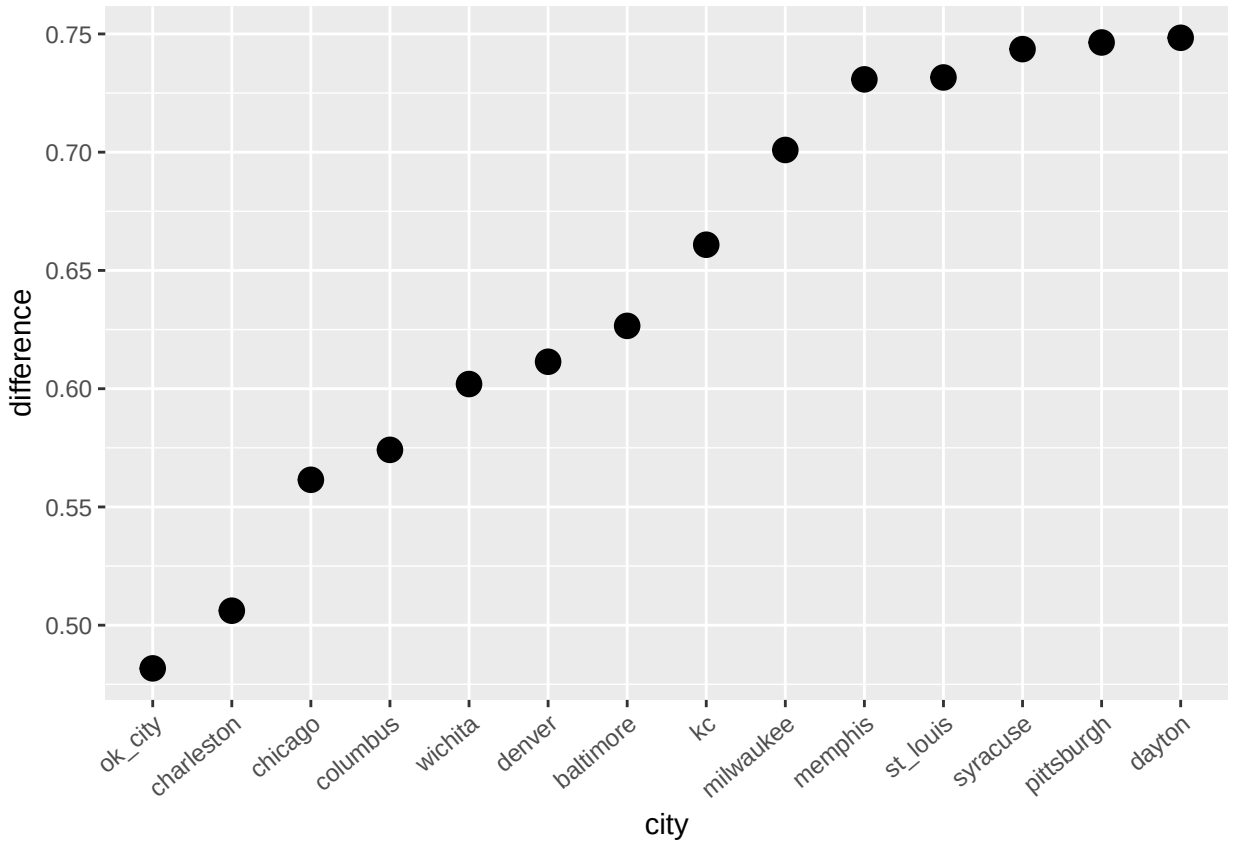
This measure Syracuse having the most majority clustering, Baltimore with the most minority clustering, and Chicago with the most unifrom amount of any clustering.

Metric Proposal

Finally, I implemented my own metric for segregation: average difference in group population percentages. This measures the average difference between percentage populations between minority and majority groups in a given neighborhood. Having a high difference means that neighborhoods of that city are not diverse.

$$\frac{\sum_{i=1}^{n} |p_i - z_i|}{n}$$

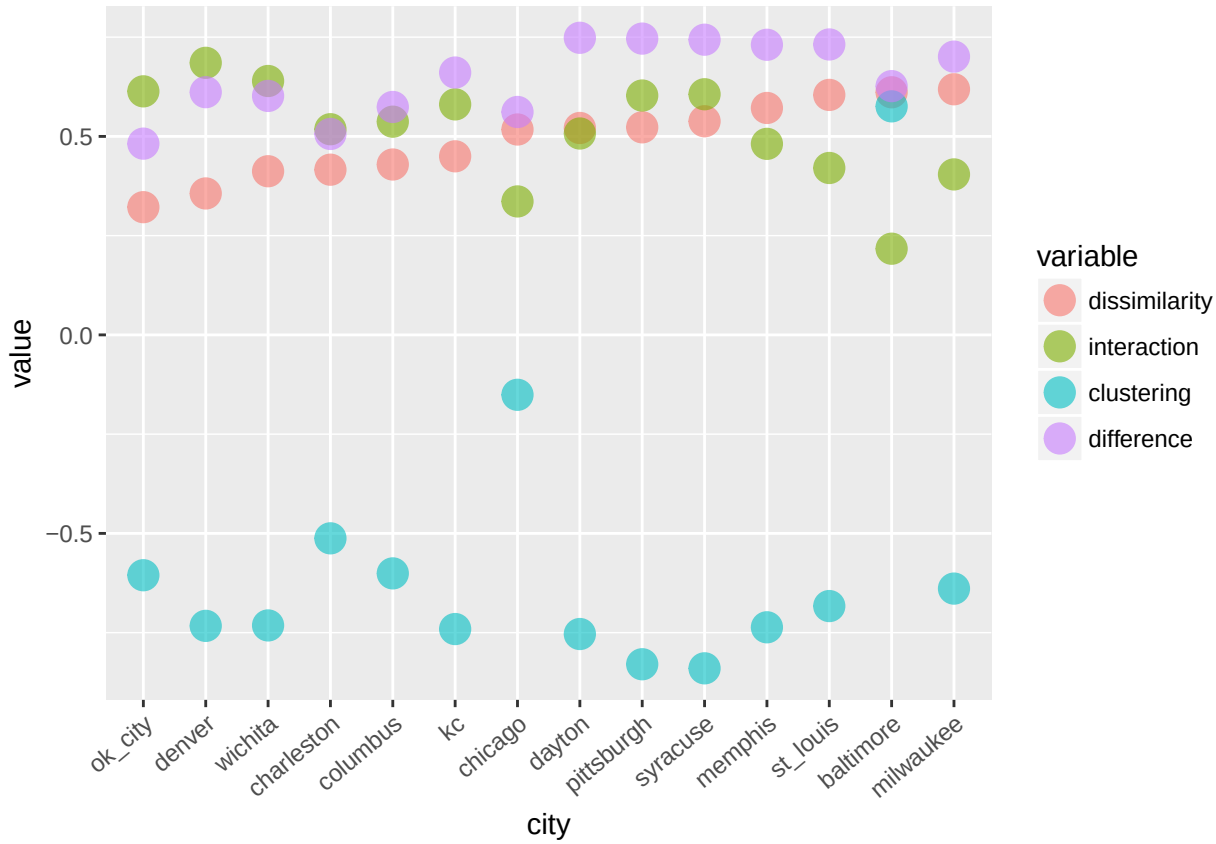
To calculate this metric you simply take the sum of the differences between the percentage of minority population (p_i) and percentage of majority population for each neighborhood (z_i), then divide that sum by the total number of neighborhoods.



Above is how each city measured up with my metric. Some of the ordering is similar to dissimilarity and inversely to interaction. You can see this below where I plot all the metrics together. While it does give another dimension to the data, it is also reinforcing how the other measuremnts ranked a city's segregation.

Analysis of Results

Lets take a look at how all of these metrics stack up with each other.




```
kable(results)
```

city	dissimilarity	interaction	clustering	difference
baltimore	0.6111194	0.2170937	0.5756543	0.6265630
charleston	0.4161646	0.5181057	-0.5124482	0.5061323
chicago	0.5173891	0.3362153	-0.1509064	0.5614761
columbus	0.4292138	0.5370268	-0.6007188	0.5741342
dayton	0.5212164	0.5072700	-0.7534688	0.7483584
denver	0.3562831	0.6852802	-0.7329294	0.6113847
kc	0.4498272	0.5806695	-0.7409620	0.6608686
memphis	0.5716063	0.4811954	-0.7364952	0.7307809
milwaukee	0.6187397	0.4042414	-0.6390241	0.7009274
ok_city	0.3217329	0.6134721	-0.6054068	0.4817471
pittsburgh	0.5228730	0.6029657	-0.8296152	0.7463824
st_louis	0.6044759	0.4204528	-0.6830677	0.7315889
syracuse	0.5381421	0.6061708	-0.8398871	0.7435688
wichita	0.4122590	0.6395624	-0.7318369	0.6019636

All of these metrics provide us with another way to look at segregation. And even still, the amount of ways people choose or discover how to measure segregation changes and grows each year. This is good, and important, for data scientists and statisticians to expand how we measure something can be somewhat subjective.