Cities Recommender

Mark Valentino

The goal of this project is to recommend cities and towns to individuals based on a city or town's "safety", and "economy". Many people often relocate to new cities that they think are better than the ones they already live in. Determining which city to move to is challenging. Cities that are both safe and economical are ideal but finding an ideal city to move to is hard to determine.

This AI can solve the issue by showing the best cities it finds via Pareto optimization, where cities shown are optimized between "safety" and "economy". First, to keep terminology simple, a "city" can also be a town, "safety" is referred to as a city's population divided by total numbers of crime committed, and "economy" is referred to as a city's median income divided by its cost of living. Higher values for safety and economy are better. Pareto optimization in this case involves making a virtual plot of cities, where cities with the lowest distance from the plot origin are the best cities.

In order to determine where cities are plotted, the values of their "safety" and "economy" are ranked in descending order, with the value '0' being the top possible rank. Plotted cities have their distance from the origin calculated, which is referred to as "magnitude", and lower magnitudes are better. Magnitudes then indicate how good a city is overall with its safety and economy values. The lower the magnitude, the more Pareto efficient a city is. Magnitudes can then be sorted in descending order, leaving the best cities at the top and the worst at the bottom.

Naturally as this sorted list is iterated through, many cities will start to increasingly skew more and more either to being more safe or more economical. Users most likely would have a preference of having a safer city or a more economical city, but users probably would not like to live in a city that is very safe with a bad economy or vice versa. This magnitude sorted list of cities can then be divided into two categories, one that skews safer, one that skews economical. These categories will still be sorted in terms of balance between safety and economy. Two cities from both categories can be shown to the user.

Datasets used are from the sources as follows:

Cost of Living of Cities: https://advisorsmith.com/data/coli/

Median Income of Cities: https://www.census.gov/data/datasets.html

Crime and Population of Cities: https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/tables/10tbl08.xls/view

Cost of Living of States: https://meric.mo.gov/data/cost-living-data-series

In order to determine how good or bad cities are, data from the datasets on the previous slide were first combined into one dataset. To help with combining data, inconsistent naming practices were dealt with, such as the state of Alabama being listed as both "AL" and "ALABAMA" in different datasets. Another example is that Honolulu was listed as both "Urban Honolulu" and just "Honolulu". The selection of cities in the census dataset were chosen as the main selection of cities since the census dataset was in the middle in terms of how many cities it covered. There were cities from the FBI crime dataset and the Advisorsmith dataset that were left over in that they could not be applied to cities from the census dataset. The dataset of cities with combined data had nearly 4,000 cities.

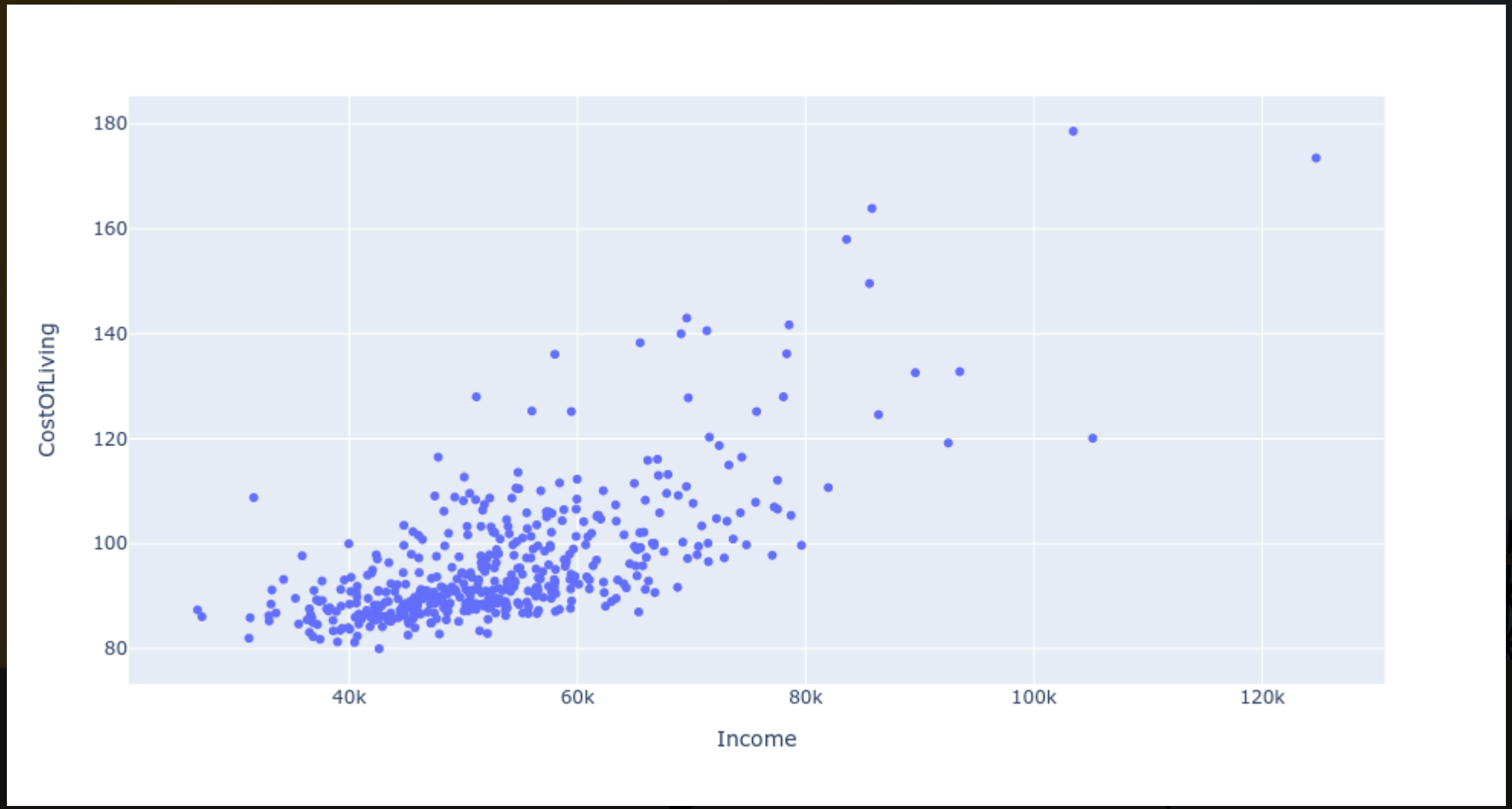
Tail of combined results into one dataframe:

	Name	State	EconomyRanking	Economy	SafetyRanking	Safety	Magnitude	Income	CostOfLiving	StateCOL	Crime	Population	RelocationNumber
3875	Willmar	MN	-1	-1	-1	-1	-1	48536	-1.0	99.6	1203	17854	-1
3876	Willows	CA	-1	-1	-1	-1	-1	45192	-1.0	146.9	394	6234	-1
3877	Wills Point	TX	-1	-1	-1	-1	-1	34727	-1.0	92.6	64	3910	-1
3878	Wilmington	NC	-1	-1	-1	-1	-1	56113	99.0	96.4	12873	102649	-1
3879	Wilmington	OH	-1	-1	-1	-1	-1	36080	-1.0	92.9	1277	12578	-1
3880	Wilmore	KY	-1	-1	-1	-1	-1	49479	-1.0	93.9	171	6028	-1
3881	Wilson	NC	-1	-1	-1	-1	-1	42007	86.3	96.4	4870	49134	-1
3882	Wilton	IA	-1	-1	-1	-1	-1	52381	-1.0	90.3	94	2867	-1
3883	Winamac	IN	-1	-1	-1	-1	-1	39798	-1.0	91.1	-1	-1	-1
3884	Winchester	IN	-1	-1	-1	-1	-1	44464	-1.0	91.1	549	4554	-1

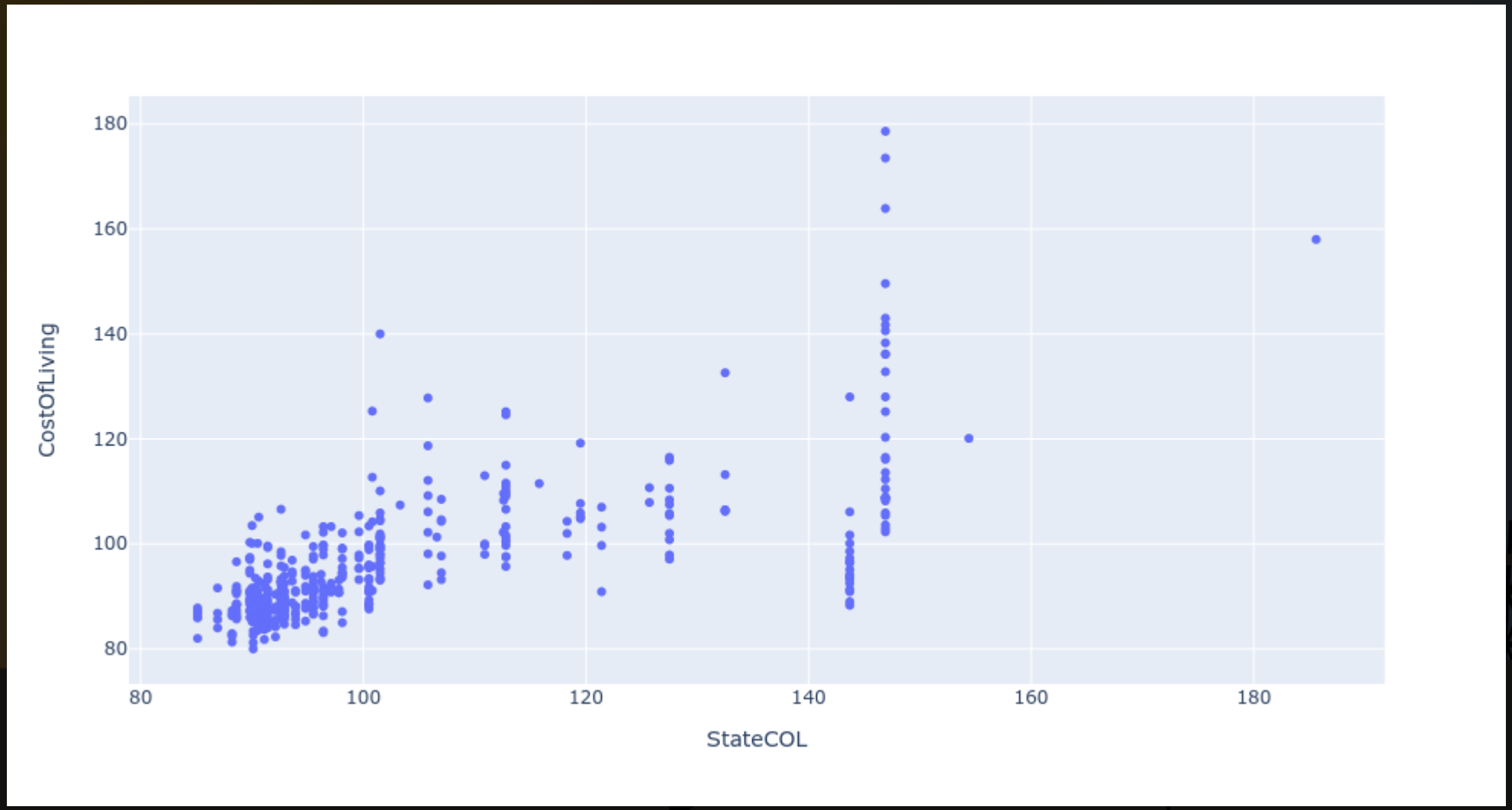
Most missing values were for the city COL. Some were for crime and population. If a city had no crime value, it had no population value, since one dataset had values on crime and population.

The combined results only had about 500 cities with no missing data. Better recommendations could be made if there were even more cities with no missing data, so datatypes on the cities were analyzed to see if there were correlations between them in hopes of being able to accurately impute data.

It was found that there was a correlation between COL and income:

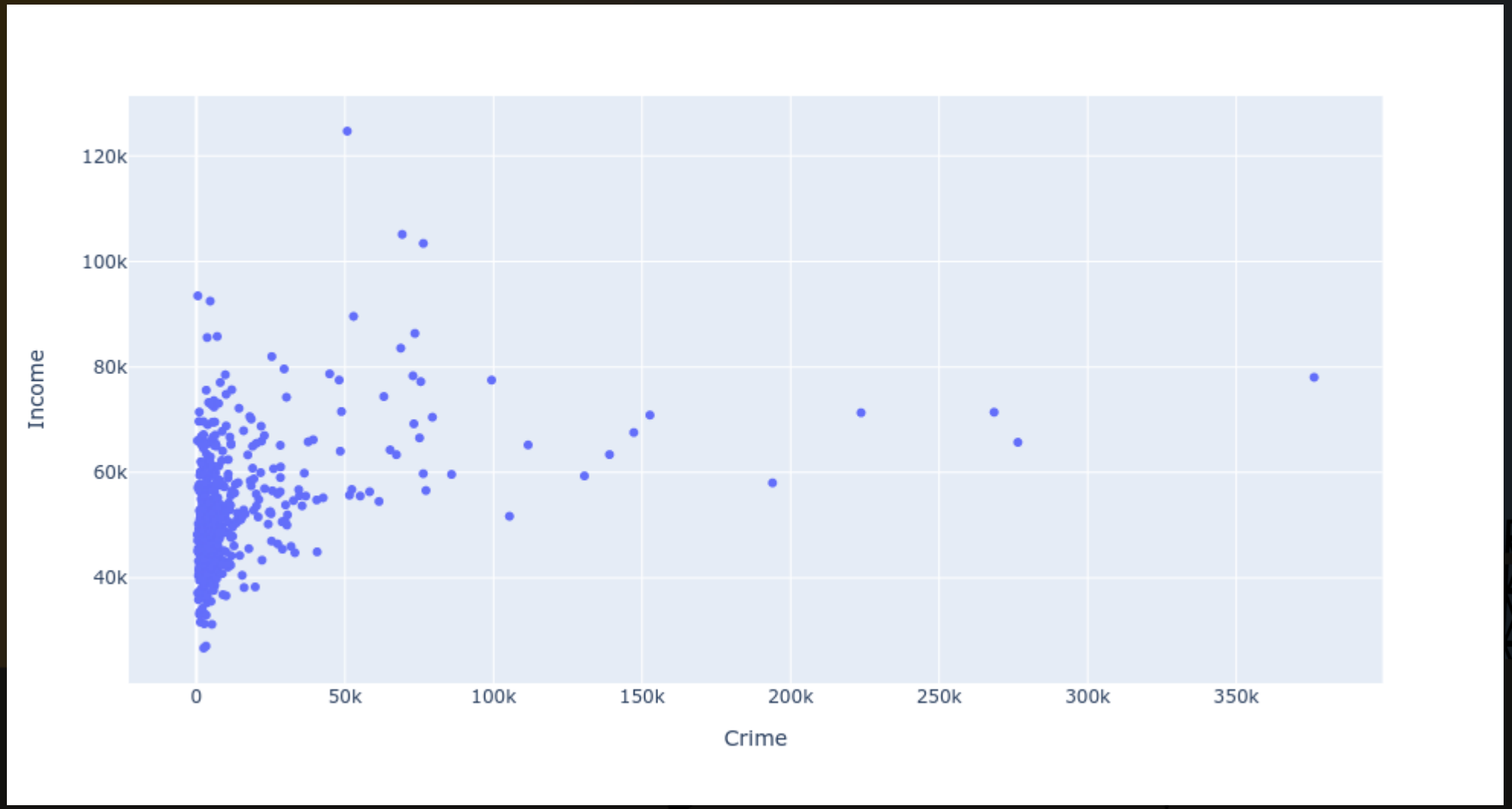


State COL and city COL were found to be related:

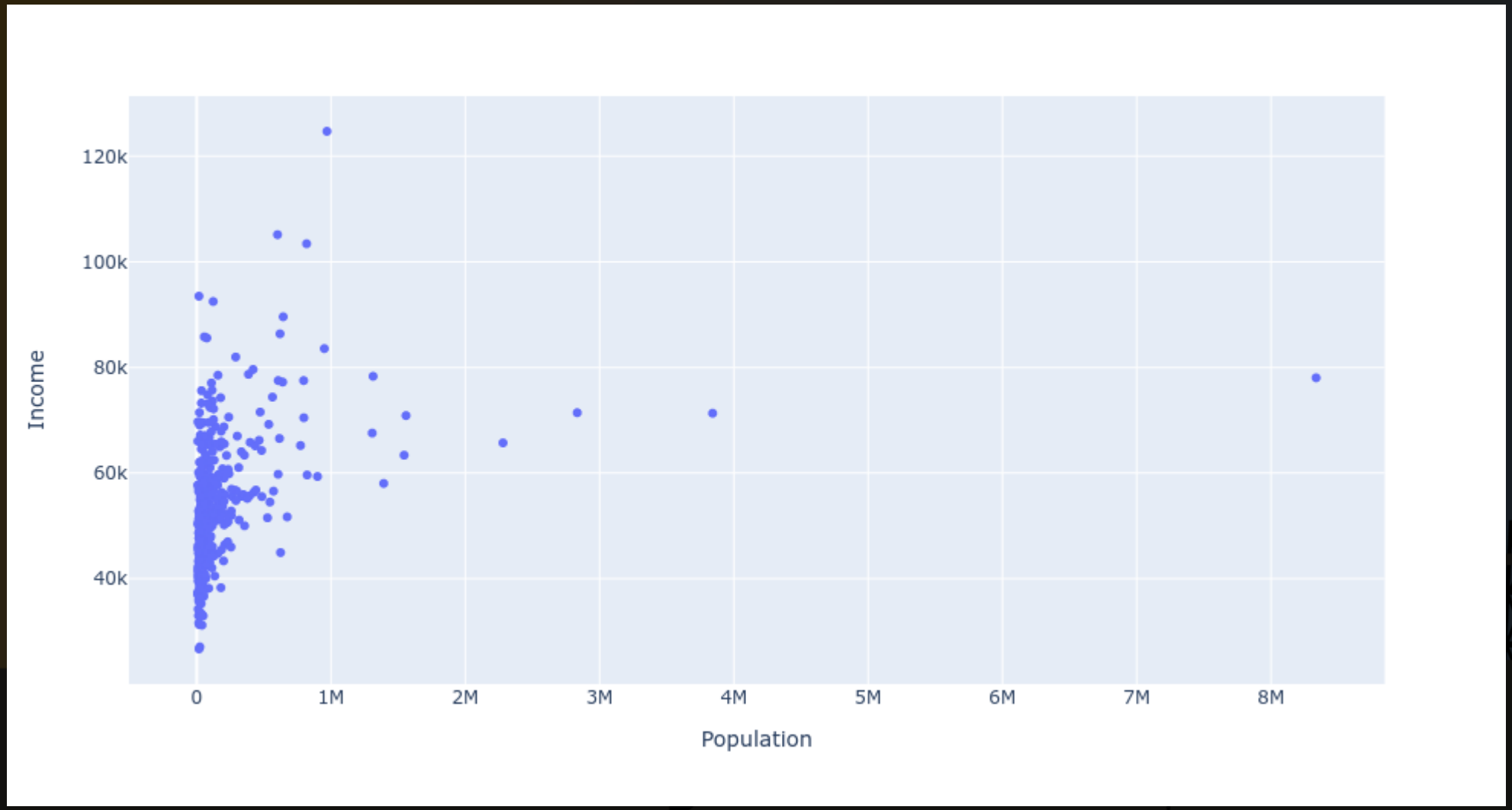


Population, crime, and safety did not show strong enough correlations with COL.

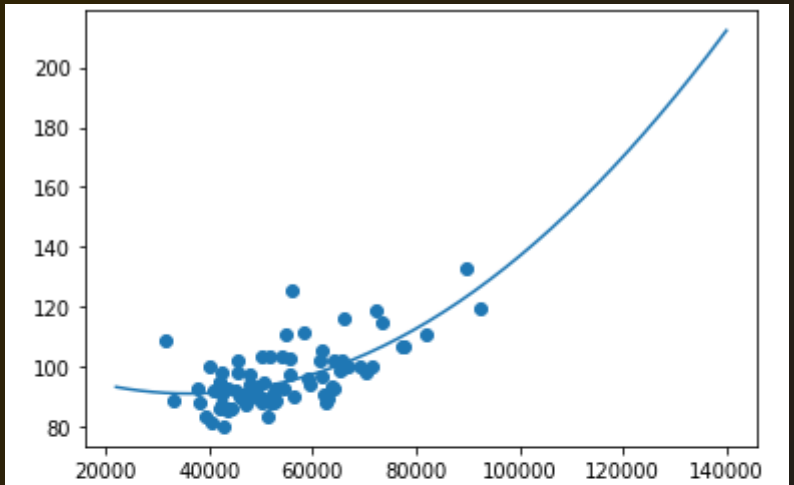
There was not a good correlation for crime and income:

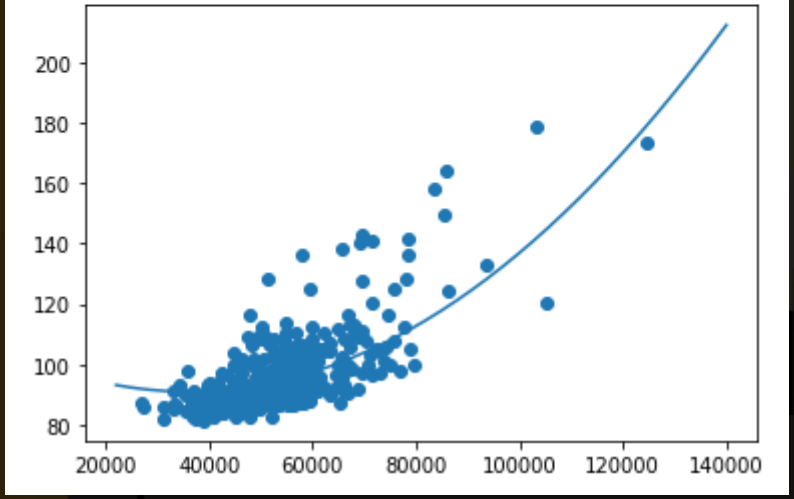


There was not a good correlation for income and population:



Crime and population would not have been able to have been imputed accurately enough.





A machine learning function was generated which was trained on both income and COL for cities with no missing data in those categories.

The machine learning function was found to be off by 6.44 points on average.

A test was done on how accurate just the values of a city's state COL compare to actual COL and was found to be off by 8.86 points on average.

It was found that if the ML predicted function and state COL were combined in a 70/30 ratio, accuracy increased to being 5.28 points off on average.
ML predicted * 0.70 + state COL * 0.30

It was then found that accuracy could be increased further. Cities were divided up into two categories. One with median income <= the US median income (62843) and cities >= the US median income.

It was found that the 70/30 ratio was still the best for lower income cities, and accuracy was off 4.78 points off on average.

For higher income cities, a 40/60 ratio was the best with accuracy at 6.67 points off on average. ML predicted * 0.40 + state COL * 0.60
Higher income cities were contributing to lower accuracy before the cities were dived up.

Accuracies of higher income cities:

	34	56	86	98	151	176	193	278	319	337	373
Name	Albany	Allentown	Anchorage	Ann Arbor	Atlanta	Austin	Baltimore	Bend	Bismarck	Bloomington	Boston
State	NY	PA	AK	MI	GA	TX	MD	OR	ND	IL	MA
Income	66722	65272	81974	65491	69230	77522	77238	66155	63516	62981	89606
CostOfLiving	100.1	98.9	110.7	99.3	100.3	106.6	107.0	115.9	93.1	89.0	132.6
StateCOL	143.7	100.5	125.7	91.4	89.8	92.6	121.4	127.5	97.8	90.5	132.5
MLPCOL	101.521808	100.54549	114.649383	100.689923	103.321873	110.278039	110.014275	101.134417	99.426299	99.099067	123.177842
4060COL	126.828723	100.518196	121.279753	95.115969	95.208749	99.671216	116.84571	116.953767	98.45052	93.939627	128.771137

Accuracies of lower income cities:

	5	11	27	29	31	33	38	43	48	69	74
Name	Aberdeen	Abilene	Akron	Alamogordo	Albany	Albany	Albertville	Albuquerque	Alexandria	Altoona	Amarillo
State	WA	TX	OH	NM	GA	OR	AL	NM	LA	PA	TX
Income	47642	50504	52179	41956	42428	61758	42552	54514	43263	46207	52448
CostOfLiving	97.6	89.1	89.4	85.8	87.3	105.4	90.9	92.9	86.2	90.8	88.2
StateCOL	112.8	92.6	92.9	90.6	89.8	127.5	88.6	90.6	92.8	100.5	92.6
MLPCOL	92.448861	93.289037	93.866016	91.324982	91.390665	98.375144	91.40875	94.775396	91.519107	92.096782	93.964544
7030COL	98.554203	93.082326	93.576211	91.107487	90.913465	107.112601	90.566125	93.522777	91.903375	94.617747	93.555181

It was decided that these predicted COLs were not accurate enough to rank cities for recommendation but would be useful in giving users the option to compare their own city with a top ranked city to see if it is worth getting a recommendation for a city.

Safety and economy were calculated for all cities. Then, rankings in economy and safety were calculated. How many people could relocate to the cities without overloading a city’s economy was estimated based off of how many open apartments Manhattan had and Manhattan's total population. That figure was then cut in half to be safe. Magnitude (distance from origin) was finally calculated. Cities were then sorted by magnitude with the best cities on top. If COL was estimated for a city, it was flagged.

To see if the results were sensible, some googling of the cities were done.

It was interesting to note that cities that were lowest on the list had a bad reputation on sites like Reddit. Cities highest on the list appeared to have favorable reputations, except for complaints like there being “no night life”.

Name	State	EconomyRanking	Economy	SafetyRanking	Safety	Magnitude	Income	CostOfLiving	StateCOL	Crime	Population	RelocationNumber	COLIsEstimated
Erie	CO	18.0	861.590158	17.0	218.279070	24.758837	103820	120.498127	105.8	86	18772	29.0	T
Lincoln	ND	22.0	854.831714	20.0	174.444444	29.732137	94292	110.304752	97.8	18	3140	5.0	T
Becker	MN	28.0	838.191895	74.0	73.116667	79.120162	92653	110.539127	99.6	60	4387	7.0	T
Smithton	IL	43.0	809.397396	80.0	71.000000	90.824006	80625	99.611143	90.5	50	3550	6.0	T
Brandon	SD	47.0	804.695568	81.0	70.593220	93.648278	84058	104.459380	96.2	118	8330	13.0	T
Los Alamos	NM	4.0	951.816023	152.0	48.441799	152.052622	109341	114.876192	90.6	378	18311	29.0	T
Williamston	MI	116.0	750.778524	107.0	59.078125	157.813181	73125	97.398897	91.4	64	3781	6.0	T
Plain City	OH	69.0	785.355486	143.0	50.416667	158.776573	78798	100.334182	92.9	72	3630	6.0	T
Hampshire	IL	3.0	954.655421	167.0	45.500000	167.026944	110071	115.299193	90.5	132	6006	9.0	T
Whispering Pines	NC	104.0	755.704608	136.0	51.904762	171.207476	76860	101.706406	96.4	42	2180	3.0	T
Purcellville	VA	1.0	973.743460	184.0	42.891473	184.002717	140049	143.825356	98.1	129	5533	9.0	T
Peotone	IL	30.0	825.530720	185.0	42.882353	187.416648	83059	100.612852	90.5	102	4374	7.0	T
Winneconne	WI	163.0	723.416406	94.0	63.000000	188.162164	71964	99.477976	95.5	40	2520	4.0	T
Heber	UT	89.0	766.509264	181.0	43.166667	201.697794	81767	106.674510	101.5	234	10101	16.0	T
St. Charles	MN	196.0	703.065096	55.0	82.930233	203.570627	71583	101.815608	99.6	43	3566	6.0	T
Kasson	MN	134.0	739.336190	156.0	47.132231	205.650188	76524	103.503658	99.6	121	5703	9.0	T
Shallowater	TX	172.0	719.205264	114.0	56.761905	206.349219	69808	97.062693	92.6	42	2384	4.0	T
Mahomet	IL	11.0	883.241105	214.0	39.905882	214.282524	92939	105.224949	90.5	170	6784	11.0	T

Erie, CO | Official Website
erieco.gov

News Flash • Erie, CO • Ci...
erieco.gov

Live Free or Die if You Must, Say ...
khn.org

A curious fruit obsession led me to the ...
9news.com

Bastrop, Louisiana - Wikipedia
en.wikipedia.org

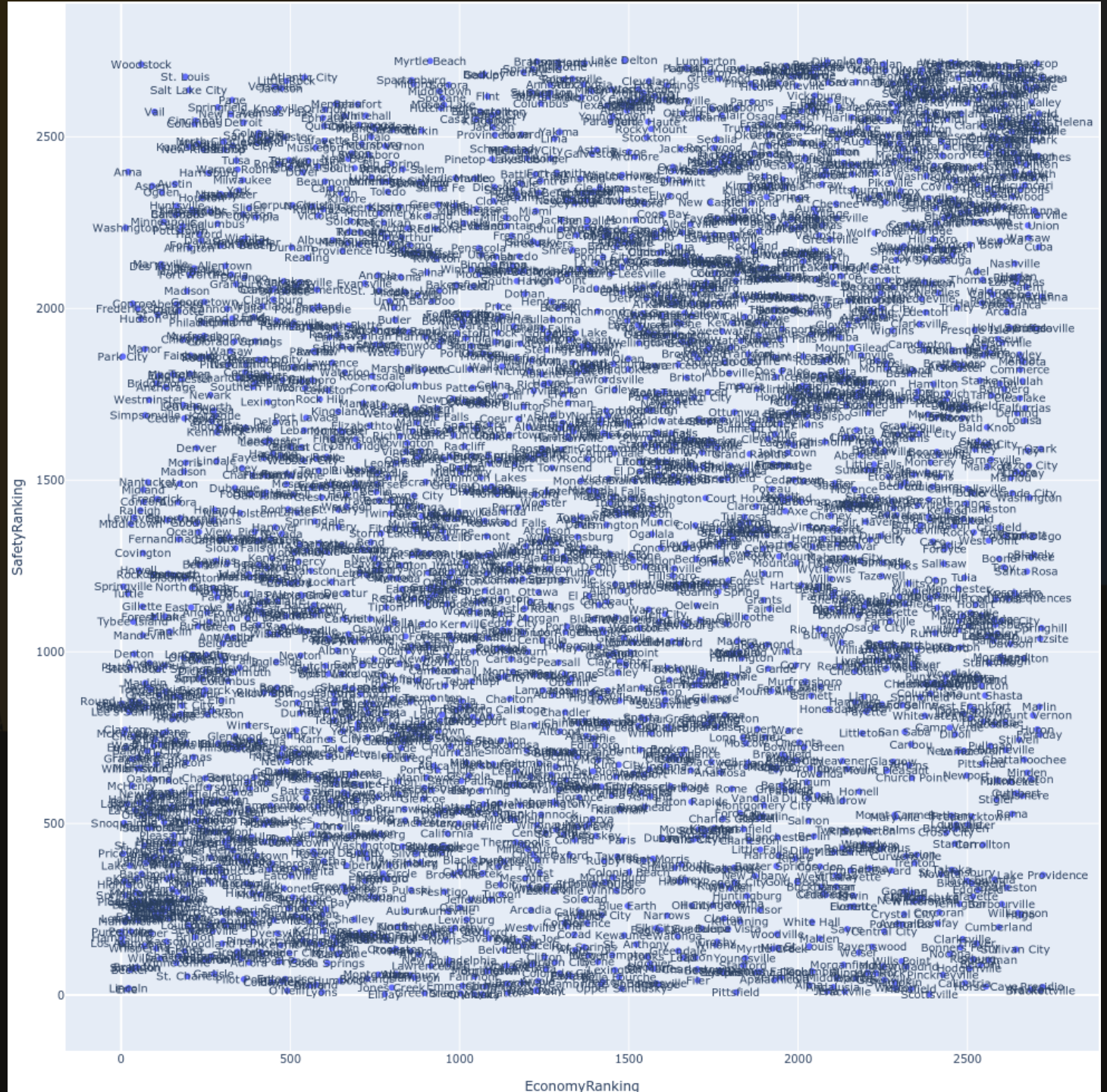
Bastrop, Louisiana - Wikipedia
en.wikipedia.org

Bastrop, Louisiana - Wikipedia
en.wikipedia.org

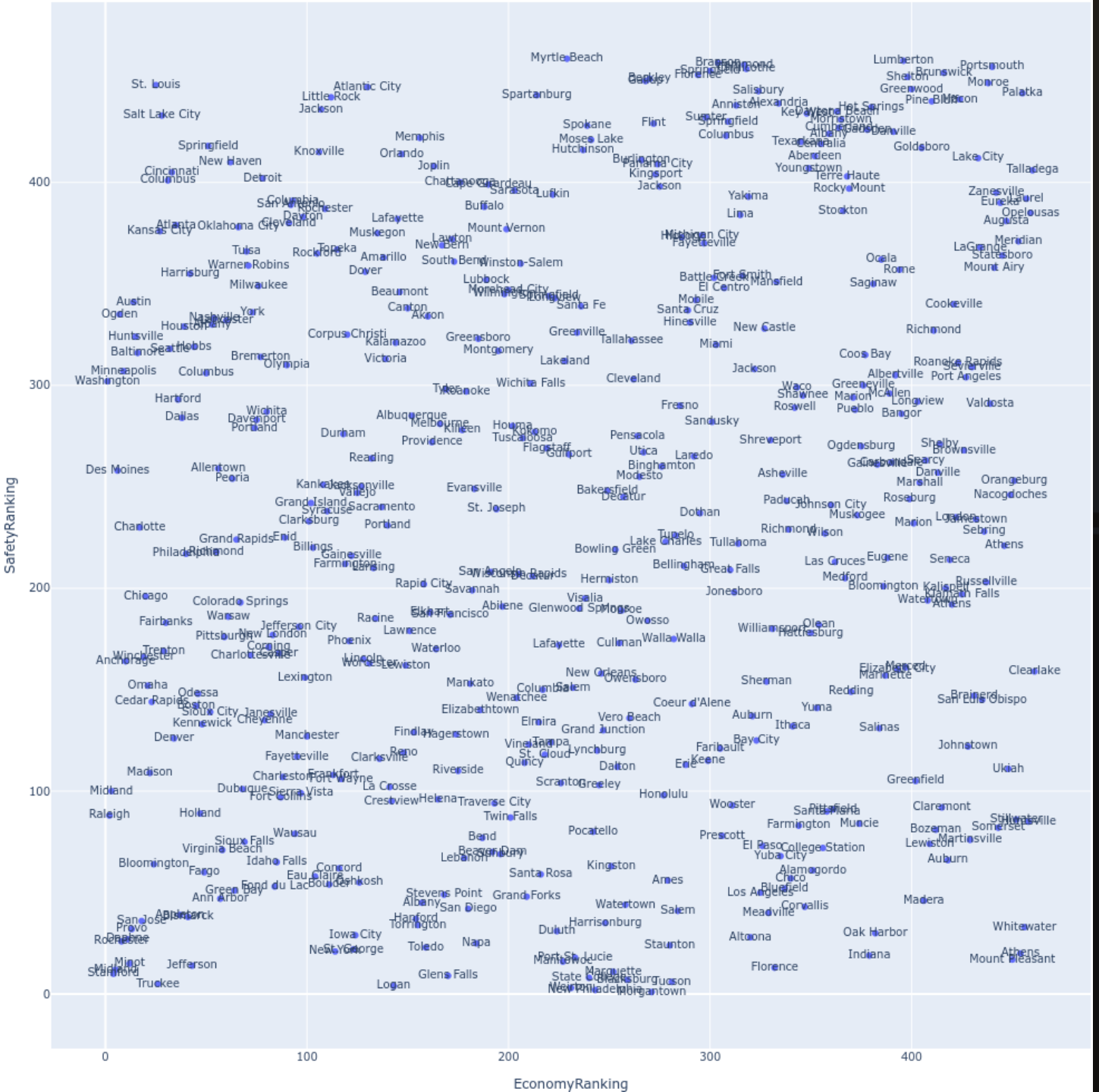
Best Places to Live in Bastrop, Lou...
bestplaces.net

Beautiful Downtown Bastrop, Louisiana ...
pinterest.com

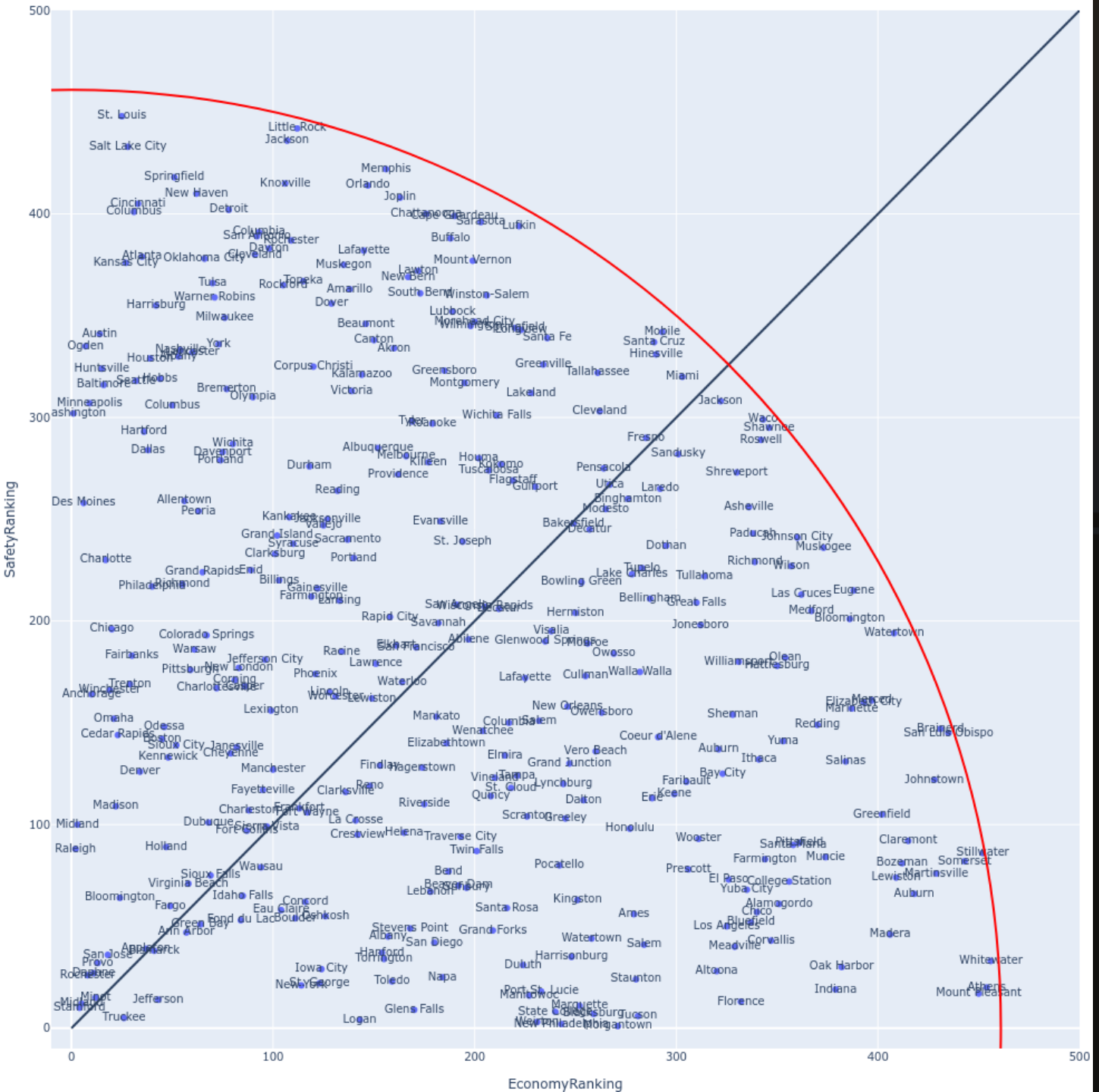
Comparing the two image search results, you get the impression that Erie is a nicer city than Bastrop.
Some quick research on Bastrop showed that Bastrop's economy has been severely affected by a paper mill closing.


SafetyRanking
EconomyRanking
0
500
1000
1500
2000
2500

These are all cities that can be recommended. Rankings needed to be recalculated free from the influence of cities with an estimated COL.

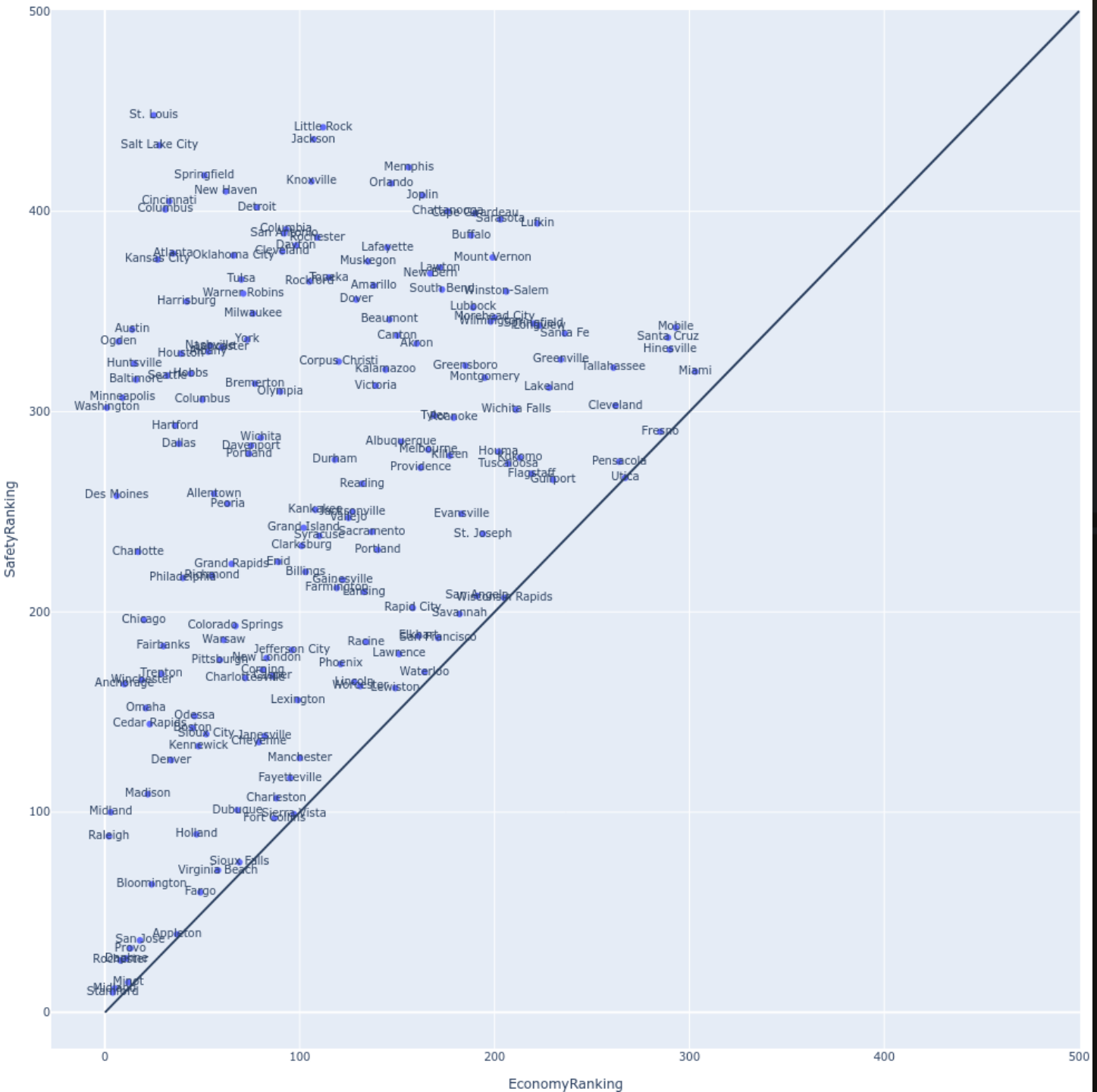


With rankings recalculated the top recommended cities change order slightly.

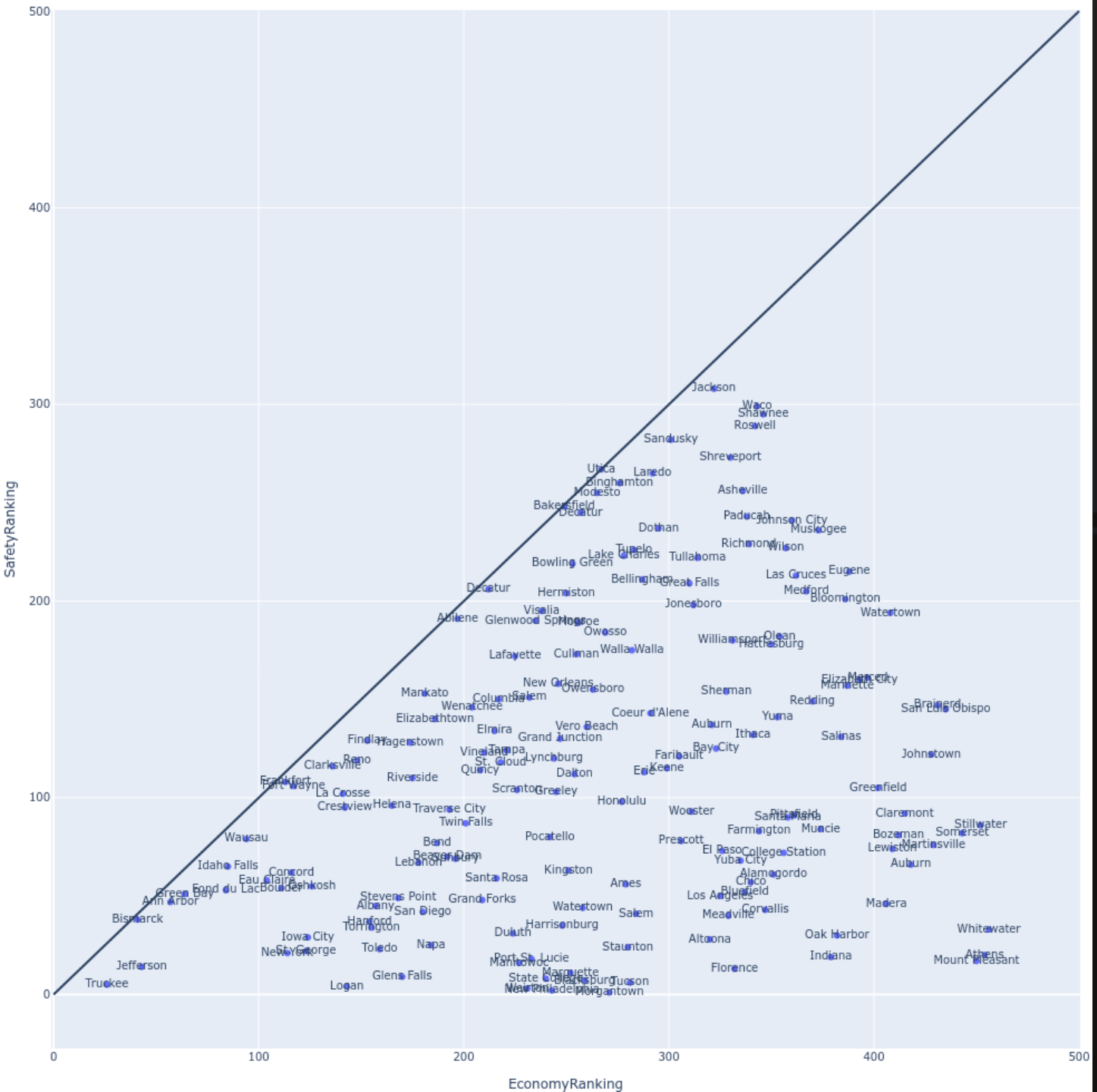


With Cities are divided into two categories, one skewing more economical and one skewing more to safer.

Cities that were beyond this red quarter circle were removed. The x and y bounds of the arc were the max economy and safety ranking. Cities have little semblance of balance between safety and economy beyond this point and are bad cities to recommend anyway. Cities beyond this arc have no chance of being the safest or the most economical.



These are cities that skew more economical. Notice that economy rankings of cities don’t go much beyond 300.



These are cities that skew more to safety. Notice that safety rankings of cities don’t go much beyond 300.

Bastrop has an estimated safety and economy below average compared to all cities.

Disclaimer: Cities are ranked from a pool of 2723 cities. The cost of living for most cities are estimated.
The Economy and Low Cost of Living Rankings can be misleading if two cities have Rankings very close to each other.
Crime Ranking and High Population Ranking are determined from an FBI dataset from 2010.

Napa has an estimated safety and economy above average compared to all cities.

Disclaimer: Cities are ranked from a pool of 2723 cities. The cost of living for most cities are estimated.
The Economy and Low Cost of Living Rankings can be misleading if two cities have Rankings very close to each other.
Crime Ranking and High Population Ranking are determined from an FBI dataset from 2010.

Stamford's economy is the best out of two.
Stamford is the most balanced between safety and economy out of two.
Truckee's safety is the best out of two.

Disclaimer: Cities are ranked from a pool of 461 cities. Cost Of Living is NOT estimated in any of these cities.
Crime Ranking and High Population Ranking are determined from an FBI dataset from 2010.

	index	Name	State	EconomyRanking	Economy	SafetyRanking	Safety	Magnitude
0	111	Stamford	CT	4.0	775.981544	10.0	26.437204	10.770330
1	124	Midland	MI	5.0	751.310345	12.0	25.665394	13.000000
2	187	Minot	ND	12.0	736.317530	15.0	23.256030	19.209373
3	247	Rochester	MN	8.0	748.828366	26.0	20.739929	27.202941
4	249	Daphne	AL	11.0	739.896480	27.0	20.679752	29.154759

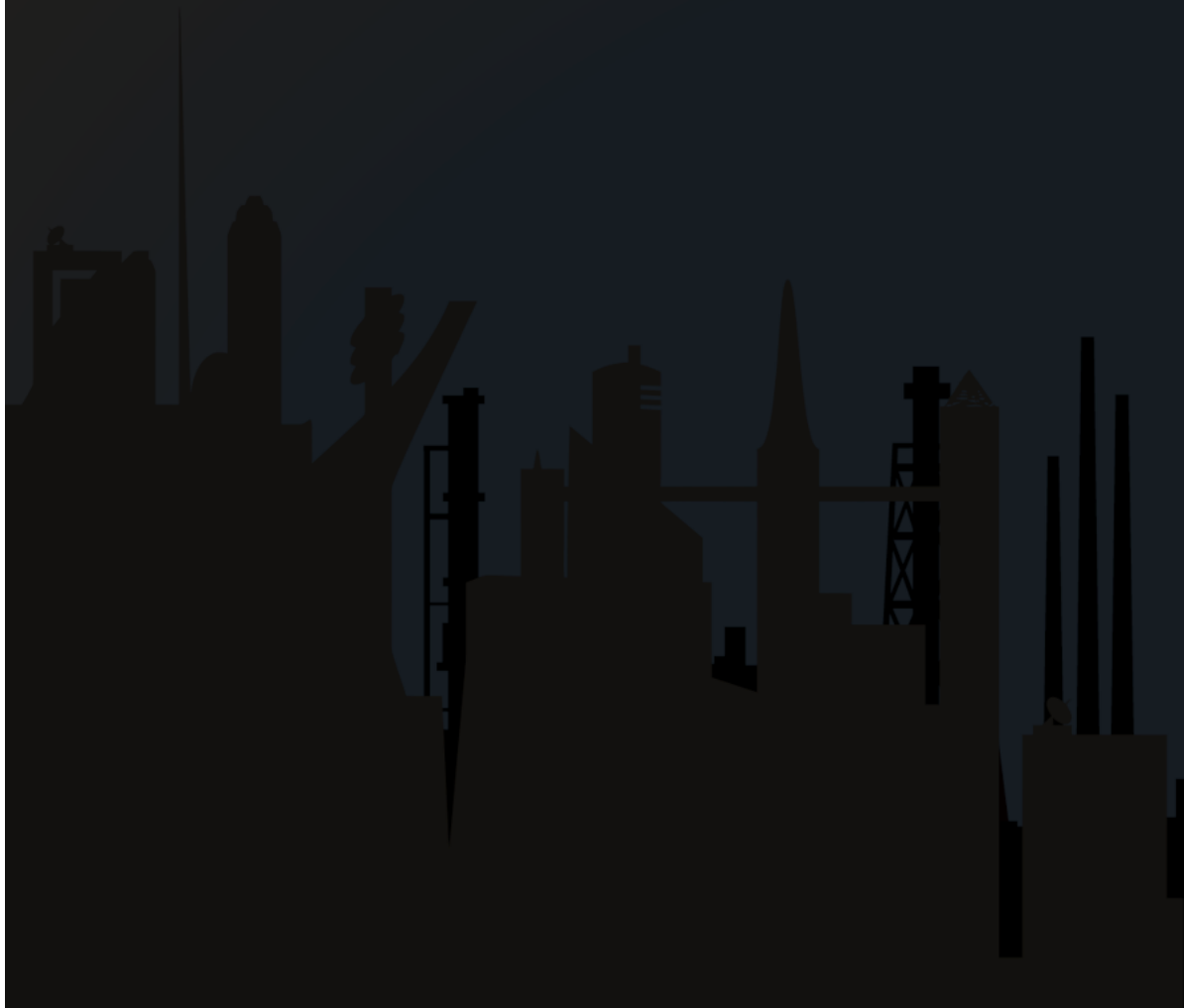
More economical cities

	index	Name	State	EconomyRanking	Economy	SafetyRanking	Safety	Magnitude
0	57	Truckee	CA	26.0	704.066265	5.0	35.162047	26.476405
1	198	Jefferson	GA	43.0	677.977413	14.0	23.723343	45.221676
2	342	Bismarck	ND	41.0	682.234157	38.0	18.782430	55.901699
3	433	Ann Arbor	MI	57.0	659.526687	47.0	17.287283	73.878278
4	457	Green Bay	WI	64.0	649.923414	51.0	17.068733	81.835200

Safer cities

	NameOfPerson	CityPreference	Name	State	Magnitude
0	Matthew Hill	2	Murray	KY	2922.593540
1	Tiffany Sanchez	2	Tyrone	PA	1933.364166
2	Jennifer Barber	2	Nevada	IA	1425.468695
3	Melissa Bell	1	Hartwell	GA	3652.420567
4	Mrs. Christina Haney MD	1	Raymond	WA	2164.883831
5	Courtney Gonzalez	2	Smithville	TX	1320.193925
6	Ebony Clarke	1	Colorado City	AZ	1297.554719
7	Amy Rasmussen	1	Moultrie	GA	3745.018024
8	Cynthia Jones	1	Oakridge	OR	2801.441950
9	Jesse Burch	2	Bridgeville	DE	2626.795957
10	Katie Mclaughlin	1	Archbold	OH	492.468273
11	Daniel Walker	1	Sumter	SC	2895.832177
12	Cheryl Hill	2	Mount Pleasant	TN	2744.452587
13	Charles Davis	2	New Hampton	IA	1455.709449
14	Parker Fox	2	Muncy	PA	1067.002929
15	Karen Howard	2	Holbrook	AZ	2947.511662
16	Kristen Hicks	2	Parker	AZ	2333.163089
17	Scott Rodriguez	2	Johnson City	TN	2559.240512

A fake list of 1000 people were randomly generated. They “lived” in cities that were mostly considered worse. These people also had a preference for either a safer or economical city.



	NameOfPerson	CityPreference	Name	State	CityChosen	StateOfCityChosen
950	April Moore	1	St. Anthony	ID	skipped	skipped
951	Stacey Robbins	1	Colusa	CA	Green Bay	WI
952	Stacey Jackson	1	Schulenburg	TX	Green Bay	WI
953	Michelle Cohen	2	Delano	CA	Daphne	AL
954	Jenna Adams	2	Blackstone	VA	Daphne	AL
955	Laura Martin	2	Brainerd	MN	Daphne	AL
956	Dawn Walker	1	Fayette	MO	skipped	skipped
957	Ariel Salazar	1	Lake City	MN	skipped	skipped
958	Andrew Avila	2	Oxford	AL	Daphne	AL
959	John Dunn	1	Salem	OR	Green Bay	WI
960	Jennifer Ward	2	Washington	NJ	skipped	skipped
961	Stephanie Thompson	2	Fairmount	IN	Daphne	AL
962	Gregory Martin	1	Salem	OH	skipped	skipped
963	Theresa Anderson	2	Ashland City	TN	Daphne	AL
964	Carolyn Harrison	2	Texarkana	TX	Daphne	AL
965	Kendra Russell	2	Manistique	MI	Daphne	AL
966	Jesse Anderson	1	Estherville	IA	skipped	skipped

A simulation was run where people decided to relocate if the recommendation was significantly better than the city they already live in. The simulation was “first come first serve”, which arguably isn’t the fairest.

This image is part of the tail end of the list of people, where the top best cities were already filled up. As you can see, people decided to skip relocating.

To make the simulation fairer, people could be sorted where they live in estimated worse cities first, and people who live in worse cities would be given more priority.