Quality of Life - A Comprehensive Analysis of Regional Living Standards Focusing on Tri-States

Goergen Institute for Data Science Graduate '24 Wonha Shin (wshin7@ur.rochester.edu)

1. Introduction

"Exploring Economic and Environmental Indicators in the Tri-State Area: A Data-Driven Approach to Quality of Life Enhancement"

In our data-centric world, the intersection of economic and environmental indicators offers invaluable insights into the quality of life. This project delves into the heart of this nexus, with a concentrated focus on the Tri-State area of New York, New Jersey, and Connecticut. Leveraging a rich dataset from Kaggle, I would like to take insights to understand how these indicators not only reflect living standards but also influence them in this specific region.

I've found this dataset at Kaggle which is a rich compilation of data that reflects the multifaceted aspects of living standards across various regions. It's a treasure trove of data, ripe for analysis, offering a glimpse into how economic prosperity, environmental sustainability, and quality of life intertwine.

Through this project, I aim to decipher the meanings and implications of various indicators. What aspects of our economic environment most significantly affect our lives? How does our natural environment contribute to or detract from our sense of well-being? These are some of the questions I seek to address. The dataset is not just a collection of numbers and statistics; it's a mirror reflecting the multifaceted aspects of our lives, and a guide that can lead us to make more informed decisions for our collective future.

2. Dataset Overview

The "City, ZIP, County FIPS - Quality of Life" dataset from Kaggle offers a detailed overview of various quality of life indicators across different geographic areas. It could provide four different kinds of informations as follows :

- 1. Demographic Information: population statistics, age distribution, and other demographic characteristics of different regions.
- 2. Economic Indicators: detailed data on income levels, employment rates, and other economic factors provide insights into the financial well-being of communities.
- 3. Environmental Metrics: aspects such as air quality, green spaces, and other environmental conditions that affect the quality of life.
- 4. Geographical Diversity: from a diverse range of locations, offering a comparative view across different cities, ZIP codes, and counties.

By Analyzing this dataset, I would like to yield just a bit of insights that could be crucial for policy-making, urban planning, and community development, enhancing our understanding of the factors that influence quality of life in diverse areas.

(1) Initial Examination of the Dataset

Before diving into a detailed analysis, we will conduct an initial examination of the dataset, focusing on several key aspects. This preliminary review is as follows.

df.	head()												
	countyhelper	LSTATE	NMCNTY	FIPS	LZIP	ULOCALE	Overall Rank	2022 Population	2016 Crime Rate	Unemployment	2020PopulrVoteParty	2020 PopulrMajor%	AQI%Good
0	VACharles City County	VA	Charles City County	51036	23030	42-Rural: Distant	NaN	6,605	8/1000	3.21%	D	54.11%	93.76%
1	TXMcmullen County	тх	McMullen County	48311	78072	43-Rural: Remote	NaN	576	47/1000	1.81%	R	52.06%	75.33%
2	TXTerrell County	ΤХ	Terrell County	48443	79848	43-Rural: Remote	NaN	693	20/1000	3.54%	R	52.06%	75.33%
3	AKSkagway Municipality	AK	Skagway Municipality	2230	99840	43-Rural: Remote	NaN	1,081	13/1000	7.19%	R	52.83%	87.86%
4	GABaker County	GA	Baker County	13007	39870	42-Rural: Distant	NaN	2,788	0	4.19%	D	49.50%	83.30%

df.columns

Index(['countyhelper', 'LSTATE', 'NMCNTY', 'FIPS', 'LZIP', 'ULOCALE', 'Overall Rank', '2022 Population', '2016 Crim e Rate', 'Unemployment', '2020PopulrVoteParty', '2020 PopulrMajor%', 'AQI%Good', 'WaterQualityVPV', 'ParkScore2023 Rank', '%CvgCityPark', 'NtnlPrkCnt', '%CvgStatePark', 'Cost of Living', '2022 Median Income', 'AVG C2I', '1p0c', '1 p1c', '1p2c', '1p3c', '1p4c', '2p0c', '2p1c', '2p2c', '2p3c', '2p4c', 'Stu:Tea Rank', 'Diversity Rank (Race)', 'Div ersity Rank (Gender)'], dtype='object')

df.isnull().sum()

countyhelper	0
LSTATE	0
NMCNTY	0
FIPS	0
LZIP	0
ULOCALE	0
Overall Rank	3134
2022 Population	0
2016 Crime Rate	0
Unemployment	0
2020PopulrVoteParty	0
2020 PopulrMajor%	0
AQI%Good	0
WaterQualityVPV	0
ParkScore2023 Rank	0
%CvgCityPark	0
NtnlPrkCnt	0
%CvgStatePark	0
Cost of Living	0

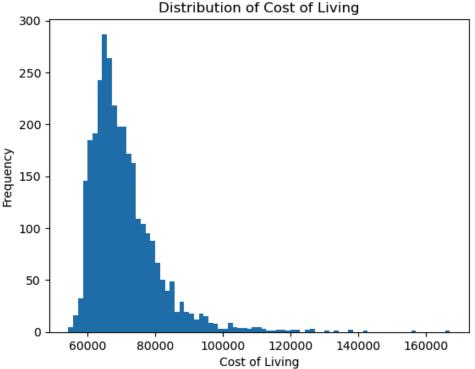
df.describe().round(1)

	FIPS	LZIP	Overall Rank	WaterQualityVPV	ParkScore2023 Rank	NtnlPrkCnt	Stu:Tea Rank	Diversity Rank (Race)	Diversity Rank (Gender)
count	3134.0	3134.0	0.0	3134.0	3134.0	3134.0	3134.0	3134.0	3134.0
mean	30393.0	53173.5	NaN	2.7	-0.9	1.2	1482.6	1567.5	1567.5
std	15162.1	23271.7	NaN	10.4	2.0	1.6	902.9	904.9	904.9
min	1001.0	1098.0	NaN	-1.0	-1.0	0.0	-1.0	1.0	1.0
25%	18179.5	34239.8	NaN	0.0	-1.0	0.0	698.2	784.2	784.2
50%	29178.0	54431.5	NaN	1.0	-1.0	1.0	1481.5	1567.5	1567.5
75%	45080.5	71727.5	NaN	3.0	-1.0	1.0	2264.8	2350.8	2350.8
max	56045.0	99929.0	NaN	456.0	52.0	9.0	3048.0	3134.0	3134.0

3. Geometrical Analysis

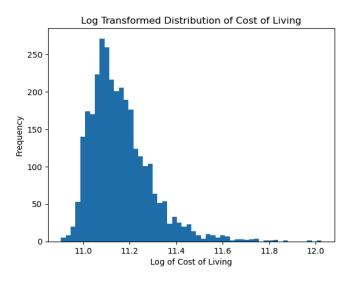
- (1) Cost of Living
 - (a) Data Preparation

For the analysis for cost of living distribution, we will take a look from seeing the distribution for the datasets as below.



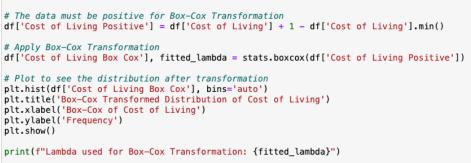
The dataset is left-skewed, which means that the bulk of the data is concentrated to the right. Since normal distribution is often a key assumption for parametric statistical tests, regression models, and other analyses, we will try to transform this with log transformation and box cox transformation respectively.

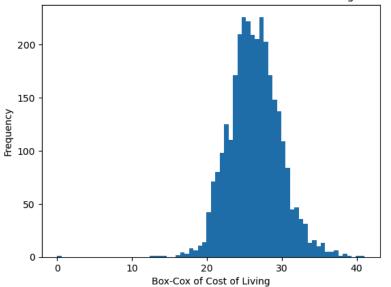
```
import numpy as np
# Apply log transformation
df['Cost of Living Log'] = np.log(df['Cost of Living'])
# Plot to see the distribution after transformation
plt.hist(df['Cost of Living Log'], bins='auto')
plt.title('Log Transformed Distribution of Cost of Living')
plt.xlabel('Log of Cost of Living')
plt.ylabel('Frequency')
plt.show()
```



Still, the log transferred distribution is left skewed. We will try with Box-Cox Transformation in the Scipy library. Code and the result as below.

from scipy import stats





Box-Cox Transformed Distribution of Cost of Living

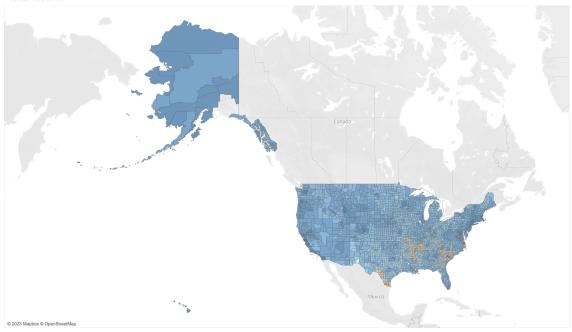
* Lambda used for Box-Cox Transformation: 0.1849793087308169

So after box-cox transformation, the datasets look like a normal distribution. We will use the box-cox transformed datasets for further analysis.

d	df.head()															
21	1p0c	1p1c	1p2c	1p3c	1p4c	2p0c	2p1c	2p2c	2p3c	2p4c	Stu:Tea Rank	Diversity Rank (Race)	Diversity Rank (Gender)	Cost of Living Log	Cost of Living Positive	Cost of Living Box Cox
%	51.54%	74.45%	91.51%	111.16%	119.90%	67.97%	90.07%	105.57%	123.82%	131.87%	135	1	25	11.232303	21101.35	28.696376
%	54.47%	73.72%	86.94%	105.46%	111.95%	72.03%	90.73%	104.21%	120.05%	127.11%	3	2	87	11.065282	9483.26	24.006475
%	67.29%	90.58%	106.09%	127.10%	135.84%	87.96%	110.73%	125.11%	145.91%	153.79%	12	3	47	11.072263	9931.00	24.258546
%	50.91%	79.00%	99.19%	121.39%	128.32%	69.18%	94.01%	113.02%	132.18%	139.30%	15	4	9	11.381783	33279.30	31.694980
%	63.61%	85.92%	103.34%	122.57%	131.91%	85.54%	108.73%	124.45%	141.99%	153.63%	26	5	60	10.991869	4959.27	20.682704

(b) Map Visualisation

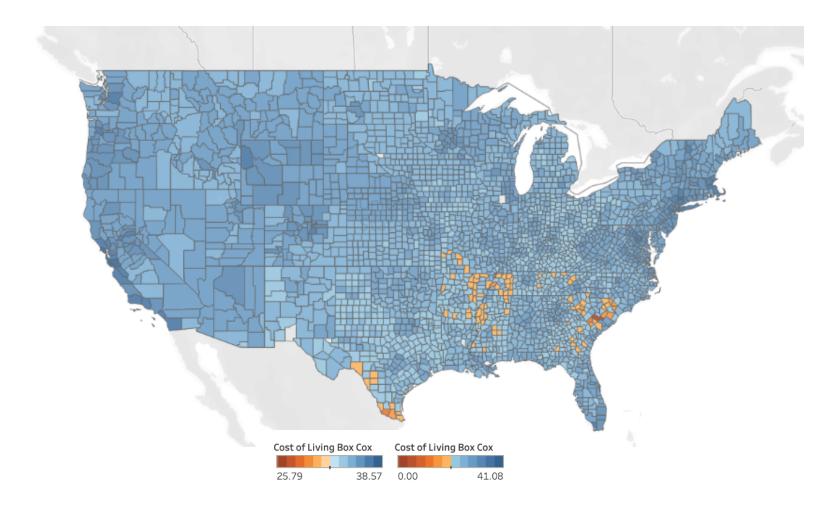
The map below illustrates the Box-Cox transformed "Cost of Living" data across different locations in the United States. Areas with darker or more intense colors represent higher transformed cost of living values, while lighter colors indicate lower values. It seems that there is a concentration of higher cost of living scores in certain regions, potentially urban areas.



<Cost of Living Analysis- Whole Country>

Map based on Longitude (generated) and Latitude (generated). For marks layer Tri_state_datasets.csv. Nmcnty: Color shows sum of Cost of Living Box Cox. Details are shown for Lzip, Lstate and Nmcnty. For marks layer United States: Color shows sum of Cost of Living Box Cox. Details are shown for Lzip, Lstate and Nmcnty, Lzip and Lstate.

Cost of Living Bo>	Cox	Cost of Living	Box Cox
25.79	38.57	0.00	41.08



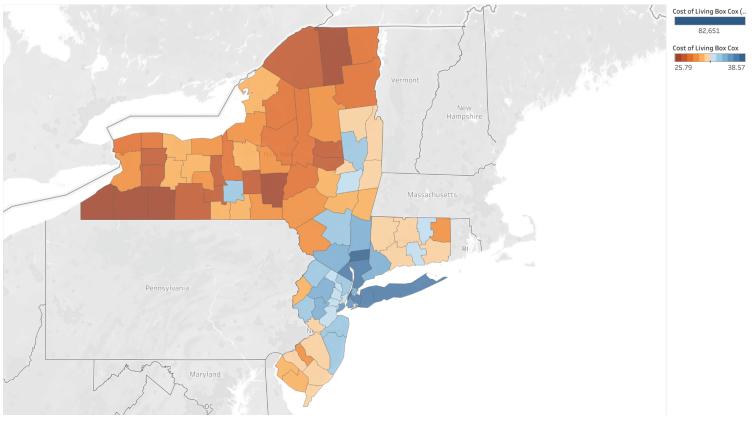
The geographic regions are divided into counties, with each county's color indicating its relative cost of living according to the transformed data. Darker shades would indicate a higher cost of living, while lighter shades represent a lower cost of living after the Box-Cox transformation.

Darker shades, indicating a higher transformed cost of living value, seem to be concentrated around certain areas.Typically, urban areas (like New York City) have a higher cost of living due to factors such as housing demand and the price of services. The map likely shows this, with darker shades potentially corresponding to metropolitan areas.

Also, areas with similar economic activities or characteristics might show similar colors on the map, indicating comparable costs of living. These clusters can be indicative of shared economic drivers, such as industry presence or economic policies - like New York City, Long Island, and Hudson County.

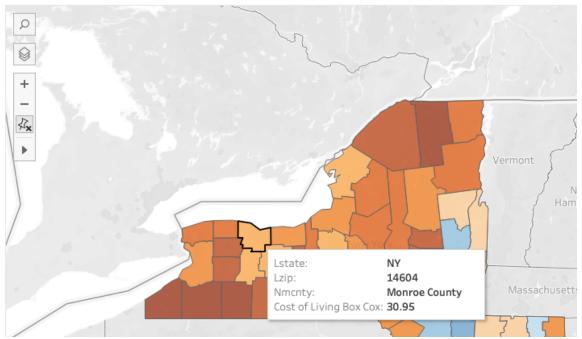
<Cost of Living Analysis- Tri States>

--Box Cox Transformed-

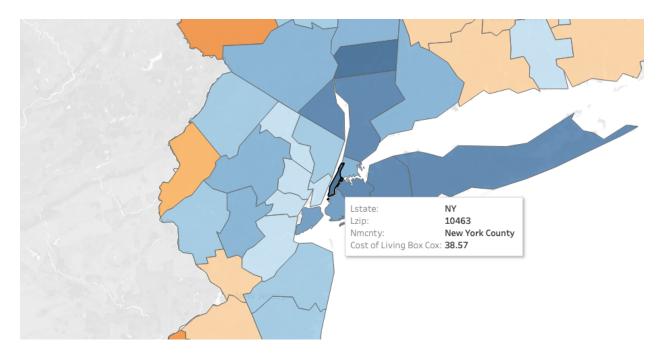


<Cost of Living Analysis- Tri States>

--Box Cox Transformed-



Monroe County is a relatively low living cost region amongst other counties in Tri-States regions. Meanwhile, New York City has the highest living cost.



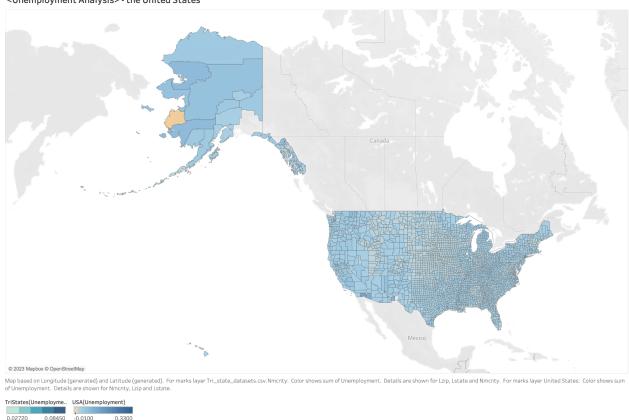
(c) Analysis : Monroe County vs. New York City

- 1. **Monroe County Affordability**: Monroe County is depicted with lighter shades on the map, indicating it has a lower cost of living relative to other counties in the Tri-State area. This could be due to a variety of factors such as lower housing costs, a more affordable price for goods and services, or lower taxes.
- 2. **New York City's Premium**: Marked by darker tones, New York City stands out for its high living expenses. The city's premium cost of living is driven by intense economic activity, with a dense concentration of high-paying industries, corporate offices, and cultural hubs, which inflate expenses across the board.
- 3. **Underlying Factors**: After conducting research for the reason why this difference occurred, I have arranged the reasons as below.
 - <u>Economic Dynamics</u>: New York City's economy thrives on high-stakes finance and corporate sectors, necessitating expensive infrastructure and services. Conversely, Monroe County's diversified economic base leans on industries that sustain a more cost-effective living environment.
 - <u>Transportation Infrastructure</u>: The comprehensive public transit system in New York City, while efficient, demands significant investment, influencing the city's

living costs. In Monroe County, personal vehicle use prevails, sparing the region from comparable public infrastructure expenses.

- <u>Real Estate Market</u>: NYC's real estate scene is fiercely competitive, inflating housing costs notably in central boroughs. Monroe County's more relaxed housing market benefits from greater land availability and reduced demand, offering residents more affordable living spaces.

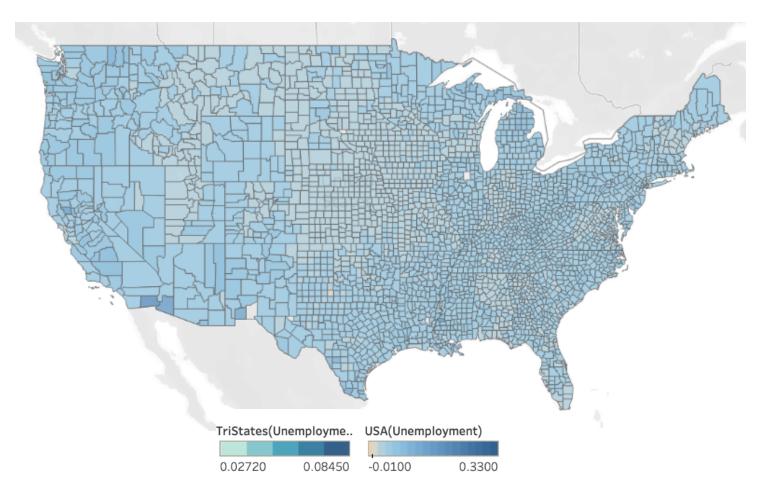
This analysis offers a snapshot of how economic structure, transportation, and housing dynamics shape living costs in these distinct regions, illuminating the stark contrast between urban and suburban living in the Tri-State area.



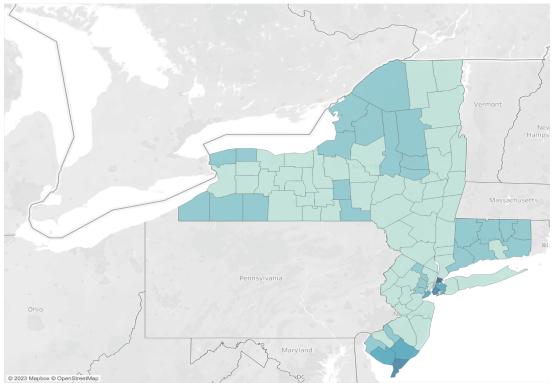
(2) Unemployment

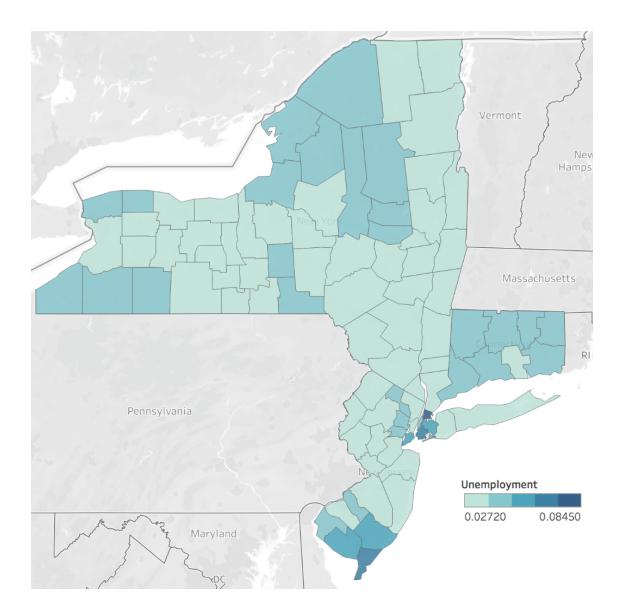
<Unemployment Analysis> - the United States

There are noticeable differences in unemployment rates across the country. Some areas are colored much darker than others, suggesting higher unemployment in those regions. Certain regions, particularly those in darker blue, might indicate economic distress or a lack of job opportunities. In contrast, lighter areas could point to more robust job markets.



<Unemployment Analysis> - Tri States





The overall ranking charts grouped by States are also next pages.

- New York State: In New York, the Bronx has the highest unemployment rates, with Kings (Brooklyn), Richmond (Staten Island), and Queens also reporting high numbers. This pattern suggests that urban counties, particularly within New York City, are experiencing the brunt of unemployment issues. Factors could include a high cost of living driving out businesses, automation, shifts in industry demands, or a mismatch between job seekers' skills and job availability.

- **New Jersey:** Cape May and Atlantic City are highlighted for their high unemployment rates. These areas have economies heavily reliant on seasonal tourism, which can lead to

significant employment fluctuations. Economic diversification in these regions may be limited, and job losses in key sectors like hospitality can disproportionately affect the local job market.

- **Connecticut**: Windham, New Haven, and Hartford in Connecticut show milder yet significant unemployment rates. These areas might be experiencing structural changes, such as the decline of manufacturing jobs, shifts to a service-based economy, or educational disparities that affect job opportunities.

<Unemployment Analysis> - Tri States

NY 10456 Bronx County NY 10456 Bronx County	NY 14715 Allegany County NY 14715	NY 14779	NY 133		NY 136 Jeff Cou NY 136	ferson inty	NY 1348 Lewis Coun NY 1348	s ity	NY 12032 Fulton County NY 12032	: 	NY 13475 Herkimer County NY 13475	NY 14775	NJ 08210 Cape May County NJ 08210 Cape May County	NJ 082 Atla Cou NJ 082 Atla		Cou NJ 083		and	NJ 08069 Salem Count NJ 08069 Salem	n :y Ə
NY 11208 Kings County NY 11208	NY 14028 Niagara Cour NY 13158	nty	NY 12093 Schoharie County NY 12093	NY 14224 Erie County NY 14224	y	12852 Essex County NY		NY 12986 Franklin County NY 12986	NY 13471 Oneida County NY 13471		NY 12485 Greene County NY 12485	NY 14591 Wyoming County NY 14591	NJ 07456 Passaic County NJ NJ	S	7871 ussex ounty	Gl	071 ouces unty	ster	NJ 07094 Hudso Count NJ	n
NY 10305 Richmond County NY	Cortland Cou		NY 13776 Otsego Co	ounty	Or	3210 nondaga		ngston	NY 13780 Chenango		NY 12758 Sullivan	NY 12861	07079 Essex County NJ	0	Warren Berg		NJ NJ 07652 0700 Bergen County		NJ 01 0806	
10305 NY 11364	13904 Broome Cour	me County NY 128)1 ren County		Y Y	Cour NY	NY	County NY NY		County NY NY	NY	NJ 08037 Camden County	N O		NJ 07652 Berger				
Queens County NY	14098 Orleans Cour	nty	NY 14604 Monroe Co	-		1 3148 eneca Coun	ity	10706	5 11746 Suffolk County NY		14020 Genesee County NY	12571 Dutchess County NY	NJ 07202 Union County	-	J 7712 onmouth		NJ 08807 Somerset			ΝJ
NY 10463 New York County NY	NY 14865 Schuyler Cou	inty	NY 13021			Y 1883 oga County	y	NY 12180	11746	N	14020	NY 12571 NY 11572	NJ 08087 Ocean County	0	NJ 08625 Mercer County		NJ 07960 nty Morris Cou		ty	
NY 12847 Hamilton County	NY 13731 Delaware Co		Cayuga Co NY 13052 Madison C	-		Y)924 range Cour	ity	Renssel NY 14548		C	utnam ounty IY NY	Nassau County NY	CT 06226 Windham County		CT 06351 New Lo	ondon Co	ounty	C	T)6076 Tolland	
NY 13669 St. Lawrence County	NY 14901 Chemung Co		NY 12901 Clinton Co			Y 1489 'ayne Coun	ty	Ontario NY 12211			0962 148		CT 06418		CT 06851 Fairfield Coun			(County T	
NY 13145 Oswego County	NY 14826 Steuben Cou		NY 12302 Schenecta			Y 2528 Ister Count	.y	Albany NY 14837 Yates C	ŕ	N	IY		New Haven County CT 06489 Hartford County		CT 06752		·	C M C	i 6498 Aiddle County T	sex

<Unemployment Analysis> - Tri States

10456 Bronx County NY	10463 New York County NY	13145 Oswego County NY	06752 Litchfield County CT	13731 Delaware County NY	14901 Chemung County NY	06498 Middlesex County CT	14826 Steube County NY	n Sc	093 hoharie unty	14224 Erie Co NY		12852 Essex County NY	l	12986 Franklin County NY
	07079 Essex County	14779 Cattaraugus County NY	14028 Niagara County NY											
11208 Kings County	LN			13471 Oneida Coun	tv	13052 Madison	07652 Bergen	0700 Midd	1 llesex	12901 Clinton		07712 Monmout		12302
	06226	13317 Montgomery County	13158 Cortland County	NY	~	County NY	County NJ	Coun NJ		County NY		County NJ		
	Windham County CT	NY	NY	12485										
10305 Richmond County NY		08037 Camden County	13904 Broome County	Greene Coun NY		13210	14517	1378	0	12758		12861	1	13148
	06418 New Haven County	NJ	NY	Or 14591 Cc Wyoming County NY NY		Onondaga County NY	Livingstor County NY		ango	Sullivan County NY	livan Inty	Washingt County NY	ton S	Seneca County NY
11364 Queens County NY	СТ	13618 Jefferson County NY	06076 Tolland County CT											
	12847 Hamilton County	N T	CI	07863 Warren Cour NJ	ity	14883		10706	11746 Suffo		1020 enesee	1218	0	14548 Ontario
08215 Atlantic County	NY	13489 Lewis County	14098 Orleans County			Tioga County NY			Count		ounty			County
LИ	06489	NY	NY	13776 Otsego Coun NY		10924 Orange Count	V							
08361	Hartford County CT	12032	08087			NY	,							
Cumberland County NJ		Fulton County NY	Ocean County NJ	12801 Warren Cour NY		14489 Wayne County NY		12211 Albany Co NY	unty	08809 Hunte NJ) Irdon C		14817	12513
08069 Salem County NJ	06351 New London County CT	07202 Union County NJ	14865 Schuyler County NY	14604 C Monroe County M		08625 Mercer Count NJ		14837 Yates Cou NY	nty	11572 Nassa NY	2 Iu Cour	ıty		
07456 Passaic County NJ	13669 St. Lawrence County NY	14775 Chautauqua County NY	08071 Gloucester County NJ	13021 Cayuga Cour NY		12528 Ulster County NY		10541 Putnam Co NY	ounty	10962 Rockli NY	2 and Co	unty	12134 Sarato NY	oga County

Lzip, Nmcnty and Lstate. Color shows sum of Unemployment. Size shows sum of Unemployment. The marks are labeled by Lzip, Nmcnty and Lstate. The data is filtered on sum of Water Quality VPV (filtered_water+), which keeps non-Null values only.

Unemployment

0.02720 0.08450

(3) Crime Rate of 2016

(1) Data Preparation

For effective analysis and visualization, the format of the crime rate data is crucial. The original datasets have format as follows, I added a column "Crime Rate Per Capita", a converted number, which is a more insightful method, the crime rates into a per capita basis, such as a crime rate per 1000 people.

In	[6]:	tri_s	state_data[[' <mark>20</mark>)16 Crime	Rate']
Out	[6]:		2016 Crime Rate		
		320	11/1000		
		542	19/1000		
		726	6/1000		
		757	13/1000		
		993	15/1000		
		1054	19/1000		
		1182	11/1000		
		1454	13/1000		
		1550	25/1000		
		1614	13/1000		

1

```
# Replace non-numeric values with NaN and convert to float
tri_state_data['2016 Crime Rate'] = pd.to_numeric(tri_state_data['2016 Crime Rate'].str.rstrip('/1000'),
```

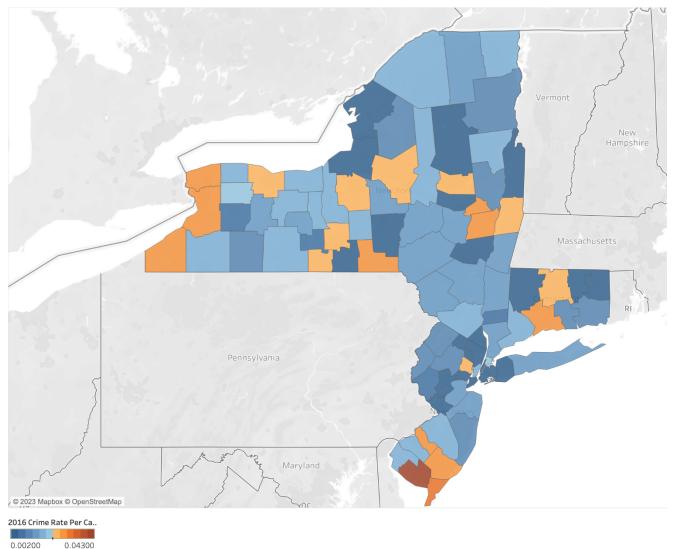
```
# Calculate crime rate per capita
tri_state_data['2016 Crime Rate Per Capita'] = tri_state_data['2016 Crime Rate'] / 1000
```

From the code above, the result is as follows.

In [8]:	tri_stat	te_data[['2	2016 Crime	Rate']]
Out[8]:	201	6 Crime Rate		
	201	o onine nate		
	320	NaN		
	542	19.0		
	726	6.0		
	757	13.0		
	993	15.0		
	1054	19.0		
	1182	NaN		
	1454	13.0		
	1550	25.0		
	1614	13.0		

<2016 Crime Rate Analysis> - Tri States

- 2016 Crime Rate Per Capita



From the provided visualization, it seems that Cumberland, Cape May, Atlantic, and Camden Counties in New Jersey have higher crime rates. In New York, Schenectady, Niagara, and Albany Counties stand out, while in Connecticut, New Haven and Hartford Counties appear to have higher crime rates.

This suggests that certain urban or densely populated areas may experience higher crime rates, which could be due to a variety of socio-economic factors that often correlate with crime, such as poverty levels, unemployment rates, and population density. For a deeper analysis, one would typically look at the root causes, the types of crimes contributing to these

rates, and how these rates have changed over time to develop a comprehensive understanding of the crime dynamics within these regions.

The overall ranking grouped by States is as follows.

<2016 Crime Rate Analysis> - Tri States

- 2016 Crime Rate Per Capita per each States

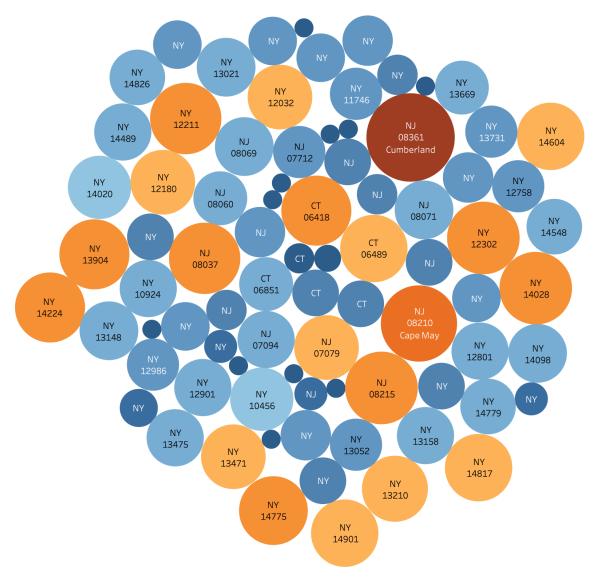
NY 12302 Schenectady County	NY 14817 Tompkins County	NY 14901 Chemung County	NY 13210 Onond County	laga	NY 12032 Fulton Co	unty	NY 12180 Rensse County		NY 134 One		County		NJ 08361 Cumberland County	NJ 0821 Cape	0 May Coun		NJ 08215 Atlantio	c County
NY 14028 Niagara County																		
	NY 10456 Bronx County	NY 12801 Warren County	NY 12901 Clinto Count	n Co	, 158 rtland unty	NY 13475 Herkin Count	5 mer	NY 14489 Wayn Count	e	NY 1454 Onta Cour	ario		NJ 08037	NJ 070		NJ 08060		
NY 12211 Albany County	NY												Camden County	Hud Cou				
	14020 Genesee County	NY 14779		NY 12986	NY 13052		NY 13731				571		NJ 07079					
NY 13904 Broome County	NY 13021	Cattaraugus	County	Franklin County			Delaware County						Essex County	NJ 07001 Middlesex County		IJ		
NY	Cayuga County	NY 14826 Steuben Cou	nty	NY		NY		NY		NY		_	NJ 08069 Salem County	NJ 078	63			
14224 Erie County	NY 13148 Seneca County	NY		14517 - Livingstor	n County	13776 Otseg Count				1213	atoga		NJ 08071	LИ	rren Count		NJ 0880	ци е
NY		13669 St. Lawrence	County	NY 10706		NY			NY	_	NY	_	Gloucester County	080 Oce	87 an County			NJ
14775 Chautauqua County	NY 14098 Orleans County	NY 11746		Westches NY		12852	2 County		14591 Wyoming		10541		CT 06418 New Haven County		CT 06851 Fairfield	l Count	y Mi	498 iddlesex
NY		Suffolk Coun	ty		County	NY 14715		57	NY 14865	NY	' NY	,					Co	
14604 Monroe County	NY 10924 Orange County	NY 12758 Sullivan Cour		NY 12513		Allegany Count		y County		⊨	NY NY		CT 06489 Hartford County		CT 06351		ст	
				Columbia	County	13489)			NY	/ NY	NY		New Lor	laon	СТ		

2016 Crime Rate Per Ca..

0.00200 0.04300

<2016 Crime Rate Analysis> - Tri States

- 2016 Crime Rate Per Capita all states



2016 Crime Rate Per Ca.. 0.00200 0.04300

4. Water Quality Analysis

(1) Data Preparation

Prior to the analysis, I assessed the distribution of the Quality of Life dataset and identified the presence of outliers. To address this, I will utilize the Interquartile Range (IQR) method to detect and trim these outliers, ensuring a robust dataset that will yield more accurate insights and facilitate a clearer understanding of the underlying trends and patterns. This process will help in mitigating the impact of extreme values that could skew the results and potentially lead to misleading conclusions.

df['WaterQualityVPV'].describe()								
count	3134.000000							
mean	2.689215							
std	10.376495							
min	-1.000000							
25%	0.00000							
50%	1.000000							
75%	3.000000							
max	456.000000							
Name:	WaterQualityVPV,	dtype:	float64					

```
# Calculate Q1 (25th percentile) and Q3 (75th percentile)
Q1 = df['WaterQualityVPV'].quantile(0.25)
Q3 = df['WaterQualityVPV'].quantile(0.75)
IQR = Q3 - Q1
# Define bounds for what is considered an outlier
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
# Filter out outliers
filtered_df = df[(df['WaterQualityVPV'] >= lower_bound) & (df['WaterQualityVPV'] <= upper_bound)]</pre>
```

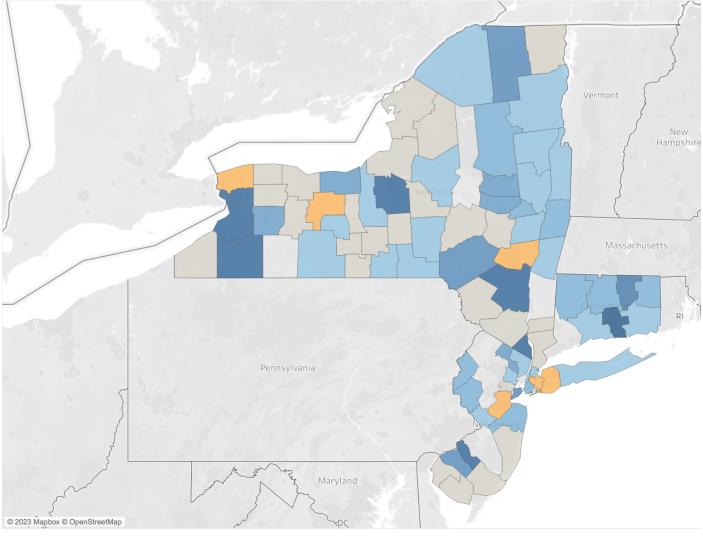
(2) WaterQualityVPV

The map represents the water quality in the Tri-State area after addressing outliers through IQR trimming, with varying shades indicating different levels of quality.

But due to numerous outliers and missing values, I plan to examine the 2022 datasets for more accurate and current insights.

<Water Quality> - Tri States

- After trimming with IQR method



Water Quality VPV

-1

7

<2022 Water Quality Analysis> - New York

NEWBURGH 12550 Orange	WARWICK 10990 Orange SLATE HILL 10973 Orange	LAKE GEORGE 12845 Warren	KINGSTON 12401 Ulster	NEW PALTZ 12561 Ulster
MONROE 10950 Orange	SAUGERTIES 12477 Ulster	MIDDLETOWN 10940 Orange PLATTSBURGH 12901 Clinton		MONTGOMERY 12549 Orange
GOSHEN 10924 Orange	WALLKILL 12589 Uister	POUGHKEEPSIE 12603 Dutchess		CAMPBELL HALL 10916 Orange

City Name, Zip Code and Counties Served. Color shows sum of # of Violations. Size shows sum of # of Violations. The marks are labeled by City Name, Zip Code and Counties Served. The view is filtered on sum of # of Violations, which ranges from 3,500 to 7,053.

of Violations

3,513 6,819

(3) Water Quality 2022 Analysis in NYS

<2022 Water Quality Analysis> - Cities in NYS

NEWBURGH Orange 12550	SLATE HILL Orange 10973 SAUGERTIES	MIDDLETOWN Orange 10940	Greene SP/	atoga 12	GHLAND ster 528	OLD FORGE Herkimer 13420		DSON umbia 34	ROCK TAVERN Orange 12575	PORT JERVIS Orange 12771
MONROE Orange 10950	Ulster 12477 WALLKILL	PLATTSBURGH Clinton 12901	GLOVERSVILLE Fulton 12078	PINE ISLAND Orange 10969		ROSEN Ulster 12472	(DNEONTA Dtsego .3820		POTSDAM St. Lawrence 13676
GOSHEN Orange 10924	Ulster 12589	POUGHKEEPSIE Dutchess 12603	ELLENVILLE Ulster 12428	VALATIE Columbia 12184	R	IDGE (IORTH REEK Varren 2853	NEW HAMPTON Orange 10958		FORT PLAIN Herkimer 13339
	AMSTERDAM Montgomery 12010	MONTGOMERY Orange 12549	COOPERSTOWN Otsego 13326	DOVER PLAINS Dutchess 12522	P	POUGHKEEPSIE Orange 12601 PULASKI Osweqo		GREA		
WARWICK Orange 10990	LAKE GEORGE Warren 12845 KINGSTON Ulster 12401	CAMPBELL HALL Orange 10916	CLIFTON PARK Saratoga 12065	CHESTERTOWN Warren 12817	1 P			Nass		
WAPPINGERS FALLS Dutchess 12590 FALLS Orange 12590			PLEASANT VALLEY Dutchess 12569	KEESEVILLE Clinton 12944	1 S	ARANAC	ONTA St. La 14519	wrence	BLOOMING GROVE Orange 10914	GRANVILLE
		PERU Clinton 12972	MAHOPAC Putnam 10541	MILLBROOK Dutchess 12545 PARISH	1 B	BREWSTER Putnam		ARD PARK 7 KLYN		
HOPEWELL JUNCTION HOPEWELL Dutchess JCTN 12533 Dutchess 12533	NEW PALTZ Ulster 12561	SARATOGA SPRINGS Saratoga 12866	ELLENBURG DEPOT Clinton 12935	Oswego 13131 WEST WINFIELD Herkimer 13491	D P Y	ENN YAN ates 4527	BROO Sulliv BROO Sulliv 11213	van L KLYN van	PREBLE Cortland 13141	SELKIRK Fulton 12158

Zip Code, Counties Served and City Name. Color shows sum of # of Violations. Size shows sum of # of Violations. The marks are labeled by Zip Code, Counties Served and City Name. The view is filtered on sum of # of Violations, which ranges from 1.400 to 7.053.

of Violations

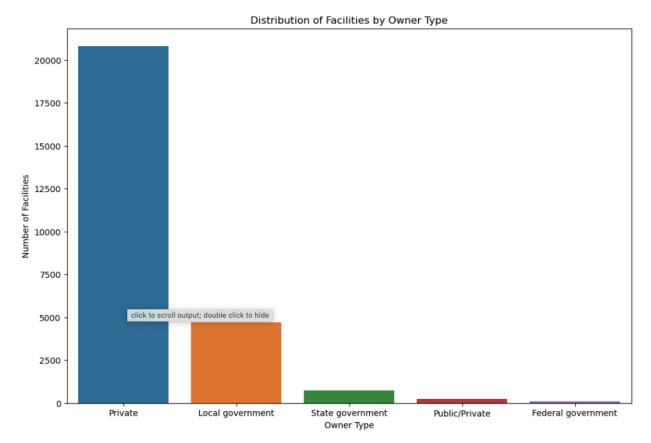
1,400 6,819

I've compiled the 2022 water system maintenance violation data over 22,000 from the EPA's Water System Summary for New York State cities, correlating each with its respective ZIP code and additional details. The dataset underwent rigorous scrutiny and preprocessing, including data trimming, to ensure accuracy and relevance for analysis.

The treemap based on EPA's Water System Summary data for 2022 indicates that within New York State, certain cities in Orange County, such as Newburgh, Monroe, Goshen, and Warwick, have reported higher water quality violations. This is followed by notable numbers in Wappingers Falls, Hopewell Junction, and Slate Hill.

(4) Exploratory Data Analysis with 2022 Water System report



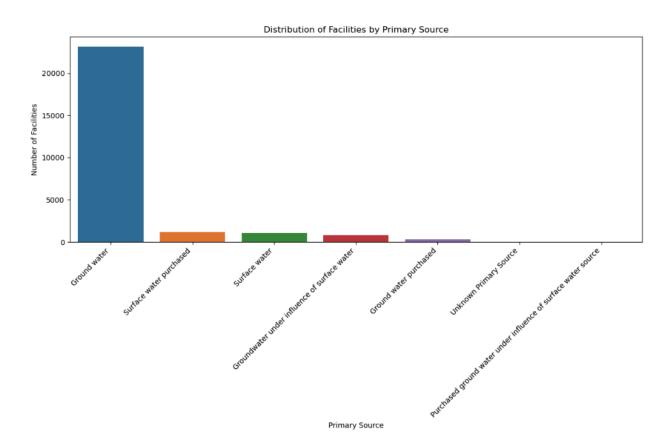


The bar chart provided appears to illustrate the distribution of facilities by ownership type. Here is a breakdown of the information presented in the chart:

- (1) **Private**: This category has the highest number of facilities by a significant margin, as indicated by the tallest bar on the chart. This suggests that the private sector owns the majority of facilities in this dataset.
- (2) **Local Government**: Represented by the second bar, local government ownership is markedly less than private but still substantial. This shows that a considerable number of facilities are managed at the municipal or local level.
- (3) **State Government**: The state government owns the fewest facilities among the categories shown, as indicated by the small size of the bar. This suggests that state-level ownership or operation of facilities is relatively uncommon in comparison to the other types of ownership.

- (4) **Public/Private**: This category indicates facilities that have a mixed ownership or partnership between public and private entities. The bar here is very small, suggesting that this type of ownership is relatively rare.
- (5) **Federal Government**: The federal government ownership is represented by the last bar, which is also quite small. This indicates that, like the state government, the federal government does not own a large number of facilities compared to the private sector.

The chart suggests that local governments play a more significant role than state or federal governments in facility ownership, which could be due to the local nature of many services (such as water, schools, and parks).



(b) Facilities by Owner Type

The bar chart illustrates the distribution of facilities according to their primary source of water. Each bar represents a different water source category, with the height of the bar indicating the number of facilities using that particular source.

- Ground Water: The tallest bar represents facilities that primarily use groundwater. This
 category has by far the largest number of facilities, indicating that groundwater is the
 most common source among the listed options.
- Surface Water Purchased: The second bar, much shorter than the first, signifies facilities that purchase surface water. This indicates that while some facilities rely on surface water, it is less common than groundwater.
- Surface Water: The third bar represents facilities that source water directly from surface water. This category has even fewer facilities compared to purchased surface water, suggesting that direct utilization of surface water is less common.

(c) Future work

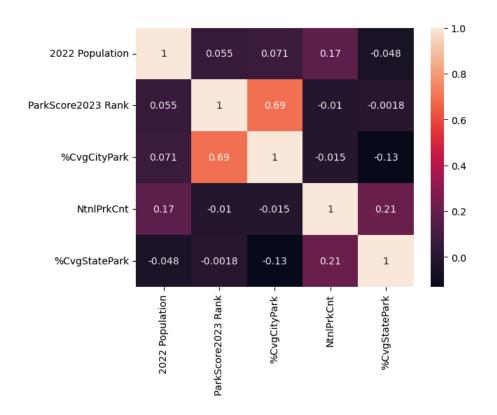
Upon reviewing the dataset, I would like to conduct a comprehensive study on water contamination and various factors on it, which could exert influence on everyday life.

- Trend Analysis: Analyzing the data for trends over time would be a key area of focus.
 This could involve looking at changes in water source usage, shifts in ownership patterns, or the emergence of new sources.
- Comparative Studies: Comparing the data from this dataset with other related datasets, such as population growth, climate change data, or industrial growth figures, could provide insights into how external factors influence water sourcing and facility distribution.
- Policy Impact Analysis: Examining the impact of current policies on the distribution and operation of water facilities and using this data to inform policy revisions or the development of new regulations.

5. Conclusion

My initial foray into this dataset analysis has revealed potential metrics that might significantly influence daily living standards within the Tri-State area. Regrettably, certain states are characterized by high unemployment rates, low median incomes, and high crime rates, all of which could adversely affect the quality of life.

The dataset also encompasses a diverse array of indicators that warrant a more multifaceted analytical approach. These include Park Scores Ranking, which evaluates the accessibility and quality of local parks; City Parking availability, which could affect urban mobility; and the proportions of land dedicated to National and State Parks, which reflect on a community's commitment to conservation and public recreation spaces.



In pursuit of a detailed understanding, forthcoming analyses could be expected to investigate the interplay between these environmental and infrastructural indicators and broader socioeconomic conditions. For instance, the availability and quality of parklands within a community may be indicative of higher real estate values, serving not only as a gauge for fiscal health but also for the well-being of its residents. The scarcity of parking provisions in city centers, on the other hand, could be symptomatic of a transition towards greener transportation methods or highlight potential shortcomings in urban planning.

Further, the proportion of land designated for National and State Parks could be emblematic of a region's dedication to preserving natural landscapes, which, in turn, might shed light on local investment in recreational spaces and the tourism sector.

By broadening the analytical framework to encompass these variables, it is conceivable to unearth subtle correlations between the availability of environmental resources and the

economic and social fabric of a community. This holistic approach promises to enrich our comprehension of the myriad elements that contribute to the overall quality of life, facilitating informed decision-making aimed at enhancing communal living standards.

[1] Z. Vaughan, "City, ZIP, County, FIPS - Quality of Life," Kaggle, 2023. [Online]. Available: <u>https://www.kaggle.com/datasets/zacvaughan/cityzipcountyfips-quality-of-life</u>. [Accessed: Dec. 22, 2023].

[2] U.S. Environmental Protection Agency, "Safe Drinking Water Information System Federal Reports," U.S. Environmental Protection Agency. [Online]. Available: https://ordspub.epa.gov/ords/sfdw_rest/r/sfdw/sdwis_fed_reports_public/21?clear=RIR. [Accessed: Dec. 22, 2023].