Exploratory Data Analysis of Gentrification Factors in Riverwest, Milwaukee

Emily Bender[*]

*University of Minnesota- Twin Cities, Department of Geography, Environment & Society, Minneapolis, MN, 55455, USA

Received: May 5, 2025

1 Introduction

For my final project, I was inspired to do an exploratory analysis into key gentrification factors in the Riverwest neighborhood of Milwaukee, Wisconsin. I grew up on the west side of Milwaukee, and while, unfortunately, we weren't taught about Milwaukee's history in school, it is something you become aware of as a local. Unlike Milwaukee's segregation, which you can largely observe on your own by exploring the city, the factors of gentrification require more digging into the data. Goetz et al. [1], with the Center for Urban and Regional Affairs (CURA) at the University of Minnesota studied gentrification in Minneapolis and St. Paul between the years of 2000 and 2015, and found evidence of gentrification in both cities. It is their study that informed my approach, although gentrification can look different across different neighborhoods. In the summary of their quantitative results, they found that gentrifying neighborhoods experienced increases of population with bachelor's degrees at rates that far exceeded citywide trends, median household income for the top 10% of households increased by about 15%, and housing costs for renters and owners increased at higher rates in gentrifying neighborhoods. They also noted that racial change was an inconsistent pattern across these neighborhoods.

2 Database Description

Most of the data I used for this project was loaded into the database as a csv, and did not include a "geometry" column for spatial querying. All of my datasets could be connected by zip code, because I had three zip codes of interest for my project: 53202, 53211, and 53212, alongside some city-wide and country-wide data for comparison. I wanted to get granular with my analysis, but I will say that it was difficult to find data at the zip code level. However, I was able to find a United States Census Bureau [2] shapefile with data at the zip code level for spatial joins and visualizations. I did not need much else from that shapefile besides the actual polygon geometry that it provided.

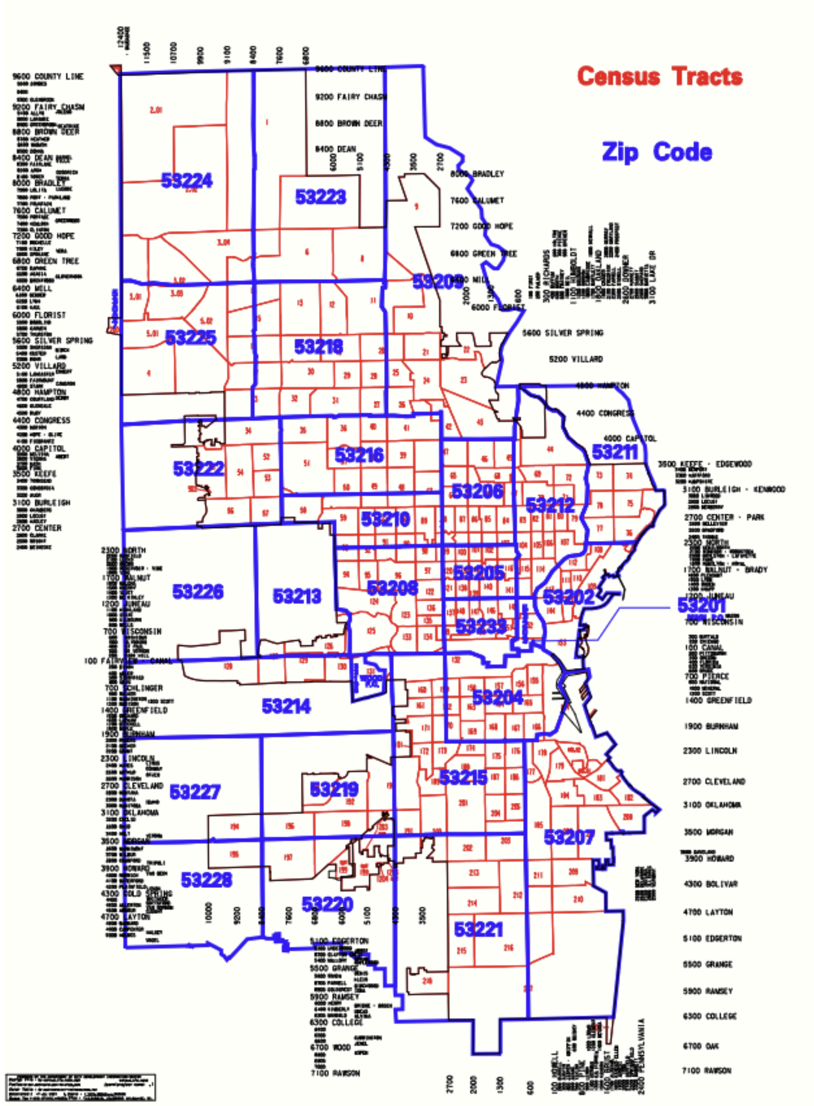
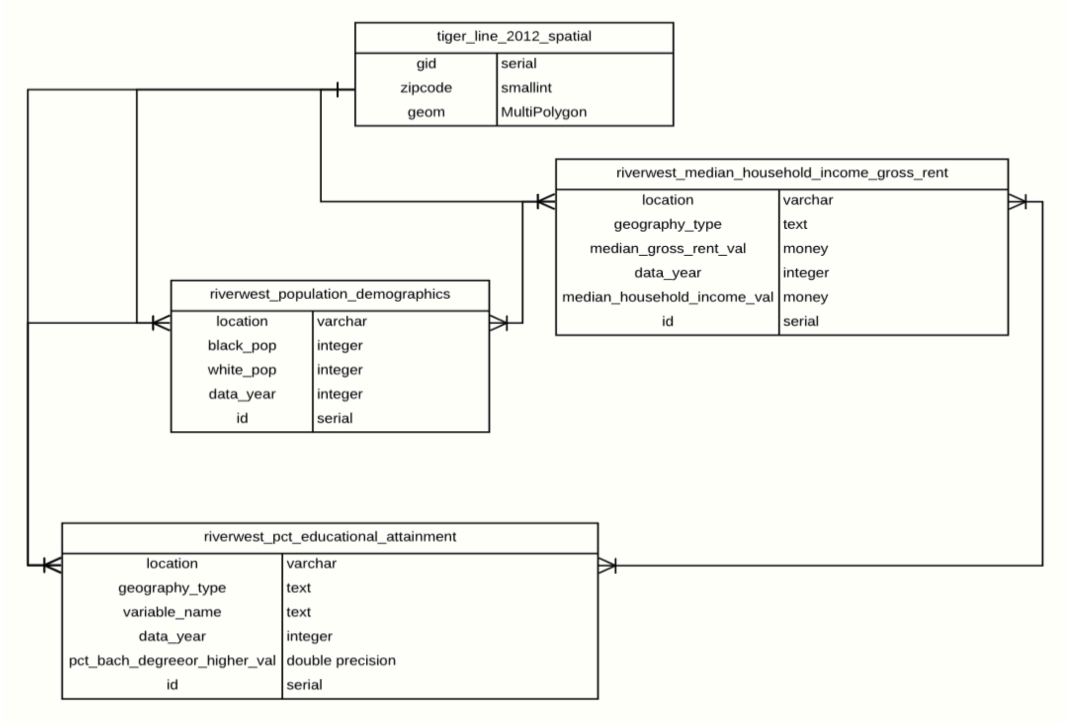


Figure 1 Milwaukee census tracts overlaid by zip code polygons for reference. Photo taken from the official website of the City of Milwaukee [3]

3 Data (types & sources)

SimplyAnalytics [4] is where I generated most of my data. I was able to search for my variables of interest at the zip code level and gather time series data, which was really important for this kind of analysis. The data available on SimplyAnalytics went back to the year 2010, but many of the columns at the zip code level were N/A values for 2010, so my analysis goes back as far as 2011, and stops at the year 2023. SimplyAnalytics allows for exporting as a csv, but after that there was some data cleaning and other reshaping steps I took before loading it into the database to reduce redundancy and make querying more seamless.

My data types are a mix between text, integers, and character varying. There were a few columns that I thought should be “money” types (median household income value and median gross rent value), but for the sake of running certain queries, it worked better when those values were specified as integers. Figure 2 illustrates the tables and fields used for the analysis and how they can be joined together via a common field.



tiger_line_2012_spatial	
gid	serial
zipcode	smallint
geom	MultiPolygon

riverwest_median_household_income_gross_rent	
location	varchar
geography_type	text
median_gross_rent_val	money
data_year	integer
median_household_income_val	money
id	serial

riverwest_population_demographics	
location	varchar
black_pop	integer
white_pop	integer
data_year	integer
id	serial

riverwest_pct_educational_attainment	
location	varchar
geography_type	text
variable_name	text
data_year	integer
pct_bach_degreeor_higher_val	double precision
id	serial

Figure 2 Database schema, with indicators for connections between tables on a common field and specified data types. In all tables but the tiger_line_2012_spatial table, there are many rows of a similar zipcode (1 per year of data)

4 Methods

I used one of two methods for bringing my various datasets into the database. First, for csv's there is the aspect of creating tables in the database via a CREATE TABLE statement, then there is work to be done in the command prompt, after a database connection has been established for loading the tables. For my shapefile, all that I needed to do was call the shp2pgsql command to insert the shapefile into my database. However, I was able to extract the DDL from the shapefile once it was placed in the database.

5 Analysis

My first query looks at the median household income and median gross rent values by year, by zipcode/location. It calculates percent change from a "baseline" year that I set to be 2011, since that is where our data begins. When you run the query, you can see that there is some gradual change in both the rising cost of rent and the median income, but in some cases, rent is increasing more rapidly than income. I think this query gets at the question of affordability. We expect the rent to rise year after year, but is income increasing to match that?

My second query looks at pearson's correlation coefficient between percent of population with a bachelor's degree or higher and median household income values at each location in my data -

53202, 53212, 53211, Milwaukee, and USA. We expect this to be a strong positive correlation (very close to 1), and largely that is what I observed across all locations, but 53202 is quite low in comparison to the others at 0.4. This means that people with bachelor's degrees either aren't getting paid to reflect their educational attainment, or they are getting paid more than would be expected for someone without that higher level degree. This kind of analysis can not tell us which is the case in these areas.

My third query looks at rent by year, by zip code side-by-side the city of Milwaukee rent values, compares them to the previous year's rent and calculates the percent change. This is an interesting query because you can watch how the rent raises and lowers, and the rent in these three zip codes is not always higher than the citywide trend.

For my spatial query, I used the census data to be able to join and plot rent percent change onto each of the three zip codes. This just looks at the rent in 2011 and 2023 and tracks that jump. You can see that zip code 53202 experienced a larger change in rent from 2011 to 2023 at 63.88%, than the other two zip codes included in my study, which were both situated at about 40% change.

I also looked at black and white populations at these locations and ran the pearson's correlation coefficient to see how the populations are moving in relation to each other. I noticed that for the most part black and white populations in the city of Milwaukee as a whole have been decreasing, which is indicated by the positive correlation. Zip code 53202 has a strong negative correlation which makes sense because the black population fluctuates quite a bit while the white population trends upward.

6 Challenges

I learned that it takes a lot of thoughtfulness to get your data into a state that will be usable. Due to some trial and error, I also know that you need to invest time into forming the conceptual model of your database. I ended up loading the data into the database early on and just having to backtrack and rework the tables and reload them a few times. Another important step is coming up with your hypothesis, and writing queries that will test that. Having that hypothesis and thinking about ways to test it is another step that should be taken early on in the project process.

7 Solution

Truthfully, I think I need to add more data than I included in this study to gain a better understanding of this complex phenomena. What I did was an initial analysis, to see if I could find the data variables that can act as gentrification factors and see what I could glean from them. I also think it is important to involve subject matter experts in this kind of research to help with the analyses, interpretation and conclusions. What I came up with is the beginning of a solution, but not exactly a solution. If I had more time, and resources, I think I could definitely draw some more definite conclusions. In the study that inspired this one, they also surveyed the community for what they call "qualitative results", AKA what do the people who live in these areas perceive is happening around them, which I think is meaningful when you have people who aren't from these neighborhoods attempting to study some sensitive phenomenon.

What I can say based on my analysis is that rent has increased since 2011 in these zip codes, at a higher rate than it is increasing for the city of Milwaukee overall, but not always higher than the country-wide trend. I can also say that for zip code 53202, the pearson's correlation coefficient between population with at least a bachelor's degree and median household income is closer to 0 than it is to 1, as opposed to the other locations included in my study that are all very close to 1. As far as rent change year after year, 53202 experienced a steeper increase in the cost of rent from previous years than the citywide trend. The other two zip codes did not always surpass the citywide trend, there seemed to be more fluctuation there. As far as the black and white populations of these areas, while both populations seem to be slightly declining for the city overall, we can see this in 53212 as well. The relationship between these populations is less clear in my other locations that share a negative correlation. This indicates that one population may be rising while the other is falling.

Appendix

1. Steps taken to load csv data into database

```
Unset
--in SQL
DROP TABLE IF EXISTS riverwest_med_household_income_gross_rent;

CREATE TABLE riverwest_med_household_income_gross_rent
(
  location varchar(20),
  geography-type text,
  year integer,
  median_household_income_val integer,
  median_gross_rent_val integer
);

--in command prompt
```

```
\COPY riverwest_med_household_income_gross_rent FROM
'C:\git\class-project-ekbender\datasets\riverwest_med_household_income_gross_re
nt.csv' WITH CSV HEADER;
```

2. Steps taken to load shapefile into database

```
Unset
--in command prompt
Shp2pgsql -I -s 4326
C:\git\class-project-ekbender\datasets\tl_2012_us_zcta510\tl_2012_us_zcta510.sh
p tiger_line_2012 | psql -h spatial.healthgeog.org -d bende287 -p 5432 -U
bende287

*no need to write a create table statement
```

3. Queries

```
Unset
WITH baseline_values AS (  SELECT
       location,
       median_household_income_val::text::numeric AS baseline_income,
       median_gross_rent_val::text::numeric AS baseline_rent
       FROM
       riverwest_med_household_income_gross_rent2
       WHERE
       year = 2011 )  SELECT
       r.location,
       r.year,
       r.median_household_income_val::text::numeric AS income_val,
       ROUND(       (r.median_household_income_val::text::numeric -
b.baseline_income)         / b.baseline_income * 100,
       2      ) AS income_pct_change_from_2011,
       r.median_gross_rent_val::text::numeric AS rent_val,
       ROUND(       (r.median_gross_rent_val::text::numeric - b.baseline_rent)
       / b.baseline_rent * 100,
       2      ) AS rent_pct_change_from_2011
       FROM
       riverwest_med_household_income_gross_rent2 r
       JOIN
```

```
    baseline_values b
        ON r.location = b.location
    WHERE
    r.year >= 2011
    ORDER BY
    r.location,
    r.year;
```

Unset

```
SELECT
    i.location,
    CORR(i.median_household_income_val::numeric,
    e.pct_bachelor_degree_or_higher_val::numeric) AS pearson_corr
    FROM
    riverwest_med_household_income i
    JOIN
    riverwest_pct_educational_attainment e
        ON i.location = e.location
        AND i.year = e.year
    GROUP BY
    i.location;
```

Unset

```
WITH zip_rent AS (  SELECT
    location,
    year,
    median_gross_rent_val::text::numeric AS rent
    FROM
    riverwest_med_household_income_gross_rent2
    WHERE
    location IN ('53212', '53202', '53211') ), city_rent AS (   SELECT
    year,
    median_gross_rent_val::text::numeric AS city_rent
    FROM
    riverwest_med_household_income_gross_rent2
    WHERE
    location = 'Milwaukee' ), zip_yoy_change AS (  SELECT
    location,
    year,
    rent,
    LAG(rent) OVER (PARTITION
    BY
    location
    ORDER BY
    year) AS prev_year_rent,
    ROUND(              (rent - LAG(rent) OVER (PARTITION
    BY
    location
    ORDER BY
    year))              / NULLIF(LAG(rent) OVER (PARTITION
    BY
    location
    ORDER BY
    year),
    0) * 100,
    2      ) AS rent_pct_change
    FROM
    zip_rent ), city_yoy_change AS (  SELECT
    year,
    city_rent,
    LAG(city_rent) OVER (
    ORDER BY
    year) AS prev_year_rent,
    ROUND(              (city_rent - LAG(city_rent) OVER (
    ORDER BY
    year))              / NULLIF(LAG(city_rent) OVER (
    ORDER BY
    year),
    0) * 100,
    2      ) AS city_pct_change
    FROM
    city_rent )  SELECT
    z.location,
    z.year,
    z.rent,
    z.prev_year_rent,
    z.rent_pct_change,
    c.city_rent,
    c.prev_year_rent AS city_prev_year_rent,
    c.city_pct_change
    FROM
    zip_yoy_change z
    JOIN
```

```
        city_yoy_change c
                ON z.year = c.year
        ORDER BY
        z.location,
        z.year;
```

Unset

```
WITH rent_by_zip AS (      SELECT
        location,
        MAX(CASE
                WHEN year = 2011 THEN median_gross_rent_val::text::numeric
        END) AS rent_2011,
        MAX(CASE
                WHEN year = 2023 THEN median_gross_rent_val::text::numeric
        END) AS rent_2023
        FROM
        riverwest_med_household_income_gross_rent2
        WHERE
        location ~ '^\d{5}$'  -- restrict to ZIPs only
        GROUP BY
        location ), rent_change AS (      SELECT
        location,
        rent_2011,
        rent_2023,
        ROUND(((rent_2023 - rent_2011) / NULLIF(rent_2011,
        0)) * 100,
        2) AS rent_pct_change
        FROM
        rent_by_zip ), final_map AS (     SELECT
        r.location,
        r.rent_pct_change,
        s.geom
        FROM
        rent_change r
        JOIN
        tiger_line_2012 s
                ON r.location = s.zcta5ce10   )  SELECT
                *
        FROM
        final_map;
```

Unset

```
SELECT
       location,
       CORR(black_pop::numeric,
       white_pop::numeric) AS pearson_corr
   FROM
       riverwest_population_demographics
   GROUP BY
       location;
```

References

[1] Goetz, E. G., Lewis, B., Damiano, A., & Calhoun, M. (n.d.). *THE DIVERSITY OF GENTRIFICATION: Multiple Forms of Gentrification in Minneapolis and St. Paul | CURA Twin Cities Gentrification Project.*

[2] *Index of /geo/tiger*. (n.d.). Retrieved April 27, 2025, from https://www2.census.gov/geo/tiger/

[3] *City of Milwaukee*. (n.d.). Retrieved April 27, 2025, from https://city.milwaukee.gov/home

[4] *Data | SimplyAnalytics*. (n.d.). Retrieved April 27, 2025, from https://simplyanalytics.com/data