

Overview:

Extensive national research has demonstrated widespread vertical and racial inequities in property assessment and taxation across the country. Primarily, these studies show that low-value properties tend to be over-assessed compared to high value properties, often leading to a higher property tax burden for low-income homeowners and communities of color. These studies tend to raise possible hypotheses for structural issues that might contribute to the problem. One common hypothesis centers on the role of neighborhood delineations: the heterogeneity within neighborhood delineations of low-income communities, as well as the impact of means pulling disparate values towards the middle.

Our analysis plan aims to explore patterns in sales ratios and horizontal and vertical equity measures within and between property tax neighborhood delineations. It focuses on three counties, Durham, Orange, and Wake. This analysis will examine the last several years of qualified sales, but also look at the last assessment year's new assessment values compared to the previous year's qualified sales (a retrospective sales ratio analysis) to see if discernible neighborhood patterns are present. Using lessons learned from our grant partner's previous research, we will also examine factors or neighborhood characteristics which may be predictive of communities with inequity in property valuation. We will explore whether additional desired variables may be accessible for inclusion in our analysis through internet research, APIs, and web scraping.

Questions of Interest:

1. Where are there examples or evidence of improperly or unfairly assessed properties in Durham, Orange, and Wake counties? ("horizontal" dispersion)
2. Do these examples point to a systemic problem that may indicate racial inequities either within or between neighborhoods? ("vertical" inequities)
3. Are there features that are not available in the datasets used for appraisal that, if made available, could improve the appraisal process?
4. Are there neighborhood characteristics which may be associated with instances of vertical inequities?
5. Are these predictive neighborhood characteristics (if identified) generalizable to different communities (ie from one county to another)?

Equity Metrics:

In addition to general exploratory data analysis (EDA), the following measures will be assessed for neighborhoods in the three included counties:

- **Coefficient of dispersion (COD).** The average deviation of a group of numbers from the median expressed as a percentage of the median. In ratio studies, the average percentage deviation from the median ratio. This measure is the most generally useful measure of variability or uniformity, and is used to examine “horizontal”, or random, dispersion among the ratios in a stratum, regardless of the value of individual parcels.
- **Assessment progressivity (regressivity).** An appraisal bias such that high-value properties are appraised higher (or lower) than low-value properties in relation to market values. Another form of inequity exists when there are systematic differences in the appraisal of low- and high-value properties, termed “vertical” inequities. When low-value properties are appraised at greater percentages of market value than high-value properties, assessment regressivity is indicated. When low-value properties are appraised at smaller percentages of market value than high-value properties, assessment progressivity is the result. Appraisals made for tax purposes should be neither regressive nor progressive. An index statistic for measuring vertical equity is the PRD,
- **Price-related differential (PRD).** . The mean divided by the weighted mean. The statistic has a slight bias upward. Price-related differentials above 1.03 tend to indicate assessment regressivity; price-related differentials below 0.98 tend to indicate assessment progressivity.
- **PRD to COD / Ratio Scatter Plots:** The PRD may not be a sufficiently reliable measure of vertical inequities. A scatter plot of ratios versus appraised values or sale prices is a useful diagnostic tool

Important analysis variables/dimensions to explore:

1. **% minority population:** Prior research performed by our grant partners illustrated correlations between the % minority population for a neighborhood and the price related differential (PRD) and coefficient of dispersion (COD).
2. **Land value as a contributor to property valuation:** Land valuations may contribute to overvaluations in especially low-income neighborhoods, as illustrated by neighborhoods with land values that are higher than improved property sales.
3. **Age of homes:** Initial research may indicate a direct connection between low-income communities and the average age of the homes.
4. **Previously identified features:** Adding additional features to the dataset such as tenure and length of ownership would be a helpful addition to future analysis. This assessment

will explore whether additional resources may be available to aid in the analysis and build on prior work.

5. **Additional predictive features:** Using a labeled dataset of neighborhoods that are deemed potentially problematic from prior work in Wake County, we can develop a predictive model that identifies neighborhood characteristics that may help classify neighborhoods of interest for future analyses (ie. stage of development, changing demographics, neighborhood property sales rate, etc). These can be used as a starting point when identifying neighborhoods to look at more closely as we begin new county assessments.
6. **Neighborhood Delineations and comparison property sets :** Comparison properties may be derived from within set neighborhood boundaries set by the appraisal process. These boundaries do not take into account all properties that may be useful for comparison, as the area may not have many sales within such a limited geographic range. Appraisals may be improved by looking at a broader or differently defined geographic boundary, or other method of sales query for comparison.

Initial Timeline:

Spring 2024:

- Study previous work and methodologies
- Set up data repository
- Identify, obtain and clean analysis datasets
- Obtain labeled “potentially problematic” neighborhood list for Wake County
- Define analysis benchmarks, tables, and charts
- Clarify scope of additional variable inclusion/ assess if other variables are obtainable

Summer 2024:

- Flesh out statistical analysis plan with discrete analysis steps
- Begin analyses
- Complete analysis tables
- Review results internally
- Iterate on analysis plan as needed

Fall 2024

- Finalize materials for community presentations
- Incorporate feedback and needs assessed from community work

Winter 2024

- Identify avenues for incorporating results into advocacy work at a systemic level
- Document methodology and findings that may be generalizable to other communities across the state

Reference:

International Association for Assessing Officers. *Standard on Ratio Studies* [PDF].

https://www.iaao.org/media/standards/Standard_on_Ratio_Studies.pdf

August Update

Executive Summary

In summary, the past several months have involved exploratory data analyses and data cleaning to create data useful for modeling and analysis. Qualified sales data were joined to geocoded parcel data and spatially joined with demographic data from the US Census.

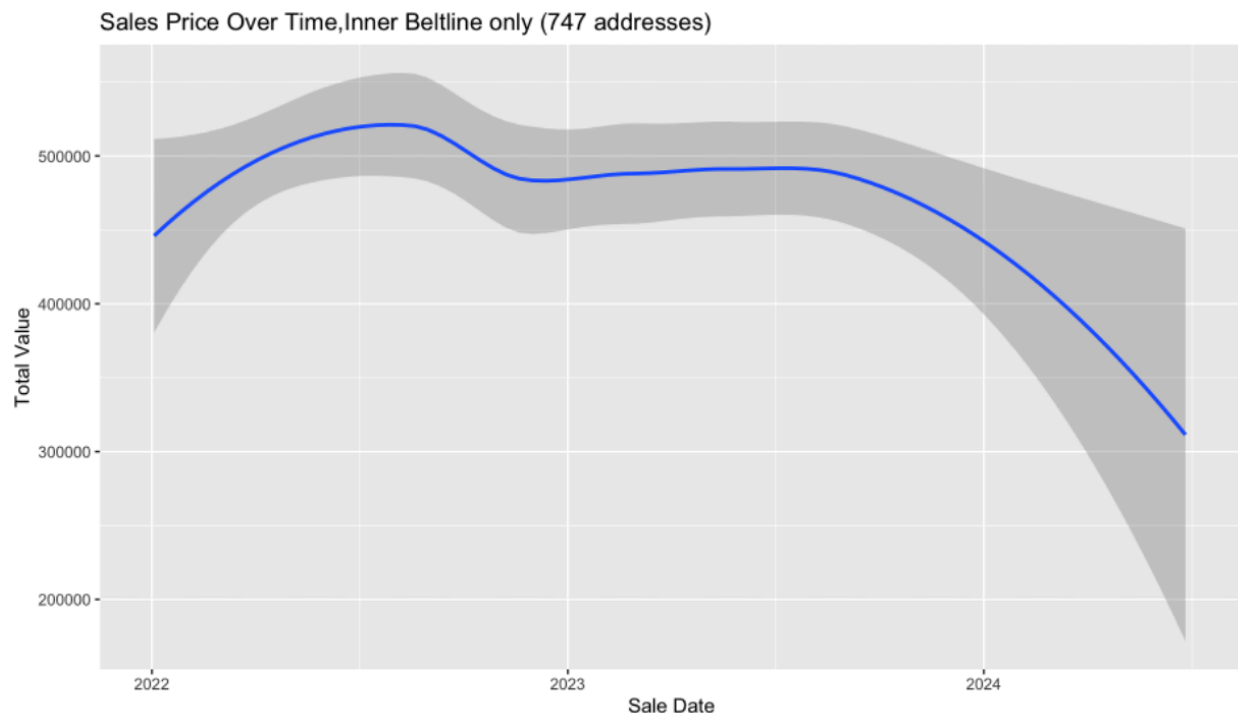
We learned that VCS neighborhoods are extremely homogenous and therefore the typical equity measures, such as PRD may not be the best for understanding inequity that may be occurring. However, additional analyses with “ground truthed” neighborhood boundaries also did not show significant differences for equity measures (**Appendix A**). This indicates that more focus may need to be made on identifying influential outliers that impact neighboring properties. We are now exploring different types of analyses to understand differences between VCS labeled neighborhoods and the way citizens may see or define Wake County neighborhoods (**Table 1**).

Table 1: Example of Labeling Hand-Drawn “Ground Truthed” Neighborhood Boundaries

Drawn Neighborhood Label	Description and Block groups /VCS Included
East Southpark	West side of Tarboro St of 507 and all 508
Battery Heights	E side Tarboro St of 507
Chavis Park	All west of Tarboro St (507,509, w side of 507)
Current Southpark	All of 509
Stratford Park	VCS of 01RA506 within 52002
Greater Stratford	All west of Dagian St within 52002
Worthdale	VCS of 01RA506 within 52002

Initial modeling of features to predict property sale ratios and sale price indicated that the timing of sale is highly influential and may need to be standardized as part of analyses and assessments (Graph 1). The percent of the population in the property's block group also was found to reduce the sale price of a property. (**Appendix B**)

Chart 1: Sale Price Greatly Influenced by Real Estate Market



A few new variables were added to the analysis in an attempt to measure gentrification in a neighborhood including a) properties in block groups that have seen a 25% or greater reduction in black homeowners between 2013 and 2019, and b) “investment” property designations (LLC or out of state investor owned properties). Being in a block group of declining black homeowners may impact sale ratios/sale price, but the sample sizes are too small (compared to the full county dataset) to statistically show significance.

If we re-examine Raleigh at the next re-evaluation period, we would like to do more work using the investment property designation to look at investment sales as a neighborhood characteristic of development. Future analyses would benefit from looking at ways to identify influential sales that influence properties around them. This kind of methodology may be beneficial and scalable to other county assessments. Additionally, we hope to explore the role land value may play as a proportion of the total sale value.

Updated 2024 Timeline

August

1. Iterate on and models with additional demographic and investment property variables
2. Add additional neighborhood comparison metrics (such as quartiles/deciles of sales price by neighborhood)

September

1. Interpret and write up model findings
2. Examine individual “influential” properties and asses impact on neighborhoods
3. Create initial materials to share with neighborhoods, get feedback from stakeholders

October

1. Add-in any additional results from characteristic-based group analyses with additional demographic features
2. Share results with community

November

1. Draft scalable methodology for use in Durham and Orange counties
2. Write final report and documentation for project

December

1. Final report to NNIP

Internal Progress Reports

May 31, 2024	Wake County EDA
July 17, 2024	Gentrification Proxy
July 19, 2024	Initial Sale Ratio Modeling
July 26, 2024	Neighborhood Equity Measures
Aug 5, 2024	Initial Sales Price Modeling
Aug 8, 2024	LLC EDA
Sep 2024	Merrick Moore Master Deck
Sep 17, 2024	Site Visit Overview
Oct 7, 2024	Durham County EDA
Nov 12, 2024	Neighborhood Deep Dives
Nov 27, 2024	Overvaluation Modeling
Dec 4, 2014	Full Review

Data Files

Durham Input

All Parcels
2020-2022 Qualified Sales Data
Neighborhood Shapefile
Neighborhood Properties

R File

2024_12_Durham_Tax_Reevaluation_Analysis.R

Folders: NNIP_DATA_DIR
NNIP_SCRIPT_DIR

Inputs:

(sales data) 2020_Qualified_Sales_Durham.csv
2021_Qualified_Sales_Durham.csv
2022_Qualified_Sales_Durham.csv
(Parcel data) 2024_10_All_Durham_BlockGroups_with_race_neighborhoods.csv

Outputs:

(Qualified res sales) 2020_Durham_Qualified_Residential_Sales.csv
2021_Durham_Qualified_Residential_Sales.csv
2022_Durham_Qualified_Residential_Sales.csv
(prop val stats) 2020_Durham_Qualified_Residential_with_Valuation_Stats.csv
2021_Durham_Qualified_Residential_with_Valuation_Stats.csv
2022_Durham_Qualified_Residential_with_Valuation_Stats.csv

Steps:

1. Loads Data
2. Renames columns and reduces number of columns
3. Merges parcels to sales data for each year of qualitative sales data
4. Filters just to residential properties
 - a. Filters to ones with "RES/" but excluding "VAC RES/" and "CONDOMINIUM"
5. Calculate Neighborhood Metrics
 - a. Neighborhood Median Sales Ratio
 - b. Neighborhood Mean Sales Ratio
 - c. Neighborhood Sum Sales Price
 - d. Neighborhood Sum Total Assessed Value
 - e. Neighborhood Weighted Ratio
 - f. Neighborhood PRD
 - g. Neighborhood Avg Abs Diff
 - h. Neighborhood Standard Deviation of Sales Ratio
 - i. Neighborhood Coefficient of Dispersion of Sales Ratio
 - j. Neighborhood Number of Comparison Properties

- k. Neighborhood Median Sales Price
 - l. Neighborhood Mean Sales Price
- 6. Save Residential datasets to output data folder
- 7. EDA of distribution of select statistics
 - a. Qual Sales Ratio
 - b. Qual Sale Price
 - c. Neighborhood PRD
 - d. Neighborhood Mean Sales Ratio
 - e. Neighborhood Coefficient of Dispersion of Sales Ratio
- 8. Identify VCS of Interest that meet following criteria:
 - a. PRD greater than 1.03 (literature)
 - b. 90th percentile of COD
 - c. At least 20 qualified sales that year
- 9. Create Sale Ratio by Sale Price Scatterplots for VCS of Interest
- 10. Create charts for Neighborhoods selected by community input
- 11. Add Valuation Stats for properties based on:
 - a. Sales Ratio
 - i. Overvalued when greater than 1
 - ii. Undervalued when less than 1
 - iii. Appropriately Valued when equal to 1
 - b. Difference between Property Sales Ratio and Neighborhood Median Sales Ratio
 - i. Overvalued when greater than 0
 - ii. Undervalued when less than 0
 - iii. Appropriately valued when equal to zero
- 12. Save Valuation statistics
- 13. Create Valuation Quartiles Dataset
 - a. 25% Upper Bound
 - b. 50% Upper Bound
 - c. 75% Upper Bound
 - d. 100% Upper Bound
 - e. Label for category falls into (.25,.50,.75, Top .25)
- 14. Create Quartile Valuation Charts (NEED TO UPDATE)

Full Write-Up (Live Document)

Initial EDA

Highlights:

- 1. Our Initial EDA (exploratory data analysis) brought mixed results.**
- 2. VCS neighborhoods scored well on equity quality metrics, unsurprisingly**
- 3. However, the VCS labels do not seem to give a full picture.**
- 4. We decided different analyses were needed to understand the VCS neighborhoods**

Our goal during our initial EDA process was to get a sense of how neighborhoods looked compared to standard quality control metrics (ie Median Sales Ratio, PRD, and COD for a VCS neighborhood). For this analysis duplicate records (properties sold more than once in the 2022-2024 time period) were removed by including only the most recent sale value of properties for the analysis dataset.

We learned there are more than 4,000 VCS neighborhoods in the Wake county data. We calculated sales ratio, median and mean sales ratio, weighted sales ratio, PRD, and COD for each VCS (neighborhood) and found most if not all of them met quality control metrics. This is unsurprising as these are the primary standard metrics for equity evaluation. Additionally the neighborhoods are intentionally extremely homogeneous, selecting a narrow band of properties close to each other with extremely similar characteristics. However, our anecdotal knowledge and experience working with homeowners has illustrated examples DO exist where individual or groups of sales can greatly impact property values within the same and nearby VCS neighborhoods. Therefore, we moved next to looking at neighborhoods that were on the high end of these metrics overall compared to the rest of the dataset..

We identified 734 properties from 22 neighborhoods that met the criteria for neighborhoods with “High PRD, High Dispersion (COD), and many comparison properties.” These were defined with a PRD > 1.03 (based on literature), a COD > 15.02 (the 90th percentile for the dataset), and at least 20 comparison properties (top 50% of the dataset). However the VCS still appeared homogeneous (as designed), and therefore don't give a full picture of what is actually located in the neighborhood. Therefore, PRD may not be the best metric to identify interesting neighborhoods (in the absence of better spatially oriented/labeled neighborhood data).

Instead we decided to look at characteristics of VCS groupings with highest and lowest median/mean sales ratios and looked at metrics for VCS groupings with the highest and lowest median sale values. Ultimately we also decided to geocode all of the data in order to do more in depth spatial analysis.

Additional VCS Analyses

Highlights:

- 1. Our results did not follow our hypothesis that neighborhoods with high sales ratios would identify neighborhoods with more disparities and lower sales ratios would be in more affluent communities**

2. **Instead, we discovered there were a lot data challenges that required further data cleaning**
3. **A new theory may be that newly built properties may obtain higher sales values as a way to account for/in anticipation of future market change to sale prices.**

We identified neighborhoods with high and low median sales ratios. The hypothesis behind this analysis is that neighborhoods with high sales ratios would likely identify neighborhoods with more disparities and lower sales ratios may indicate more affluent neighborhoods. However, our results did not follow our hypothesis.

Neighborhoods considered “High median sales ratio neighborhoods” have a median sales ratio greater than 110 (based on literature), and 20 or more comparison properties (top 50% of the dataset). Six neighborhoods met these criteria. Many of these neighborhoods were new developments in growth areas such as Garner, Wake Forest, Apex, Holly Springs, Morrisville. These neighborhoods tend to be in the upper mid range of cost (~ median sale values of approx. \$500,000- \$1 mill). Therefore this was different than our hypothesis. A new theory may be that newly built properties may obtain higher sales values, in hopes of anticipating a future market change.

The EDA analysis predominantly showed our team that the data required further data cleaning as issues were revealed during the analysis process. “Low median sales ratio neighborhoods” were defined as neighborhoods with a median sales ratio greater than 90 (based on literature) with more than 20 comparison properties. Five neighborhoods met these criteria. These neighborhoods tend to be in the mid range of cost (~ median sale values of approx. \$300- \$500k. However, many of these may be picked up due to a data comparison issue (ie -Townhouses being compared to single family homes). We determined we would need to link the qualified data set to a parcel data set from the county to pull out more information about property types and further filter the dataset to create a more homogeneous analysis and remove such confounders.

Additionally, looking specifically at properties with high sales ratios brought up examples such as sales of entire apartment complexes, a car dealership, and sales of commercial buildings. These were additional filters included in the following data cleaning process.

Data Cleaning

Highlights:

1. We filtered data only to single family home properties.
2. We geocoded parcel data from the county and linked to qualified sales data.
3. We downloaded Census data related to population demographics.
4. We performed a spatial join to add demographic data to the qualified sales data

Exploring Neighborhood Boundaries

Highlights:

1. In order to get better neighborhood boundaries for analysis, we ground truthed neighborhood boundaries with partners
2. We also explored a potential gentrification proxy measure from Census data

- a. Qualified parcels that fall within 2020 block groups where there was a 25% or greater reduction in black homeowners from 2013-2019 (2010 census boundaries)
3. We re-calculated equity measures for “groundtruthed” neighborhood boundaries
4. Most neighborhoods did not show major differences between the VCS boundaries and some of the initial ground truth neighborhoods we selected.

See Appendix A

Equity Measures Compared (calculated by VCS, labeled names, characteristic-based groups):

- a. Price-Related Differential (should be between .98 and 1.03)
- b. Coefficient of Dispersion (COD)
 - i. Below 10 for newer and homogenous residential neighborhoods
 - ii. Under 15 for older, heterogeneous neighborhoods
 - iii. Under 20 or 25 for vacant land in urban or rural areas
 - iv. Under 20 for rural residential and commercial
- c. Median Sales Ratio (should be between 90% and 110%)
- d. Median Sales Value
- e. No. of Comparison Properties

Modeling Sale Ratio and Sale Price

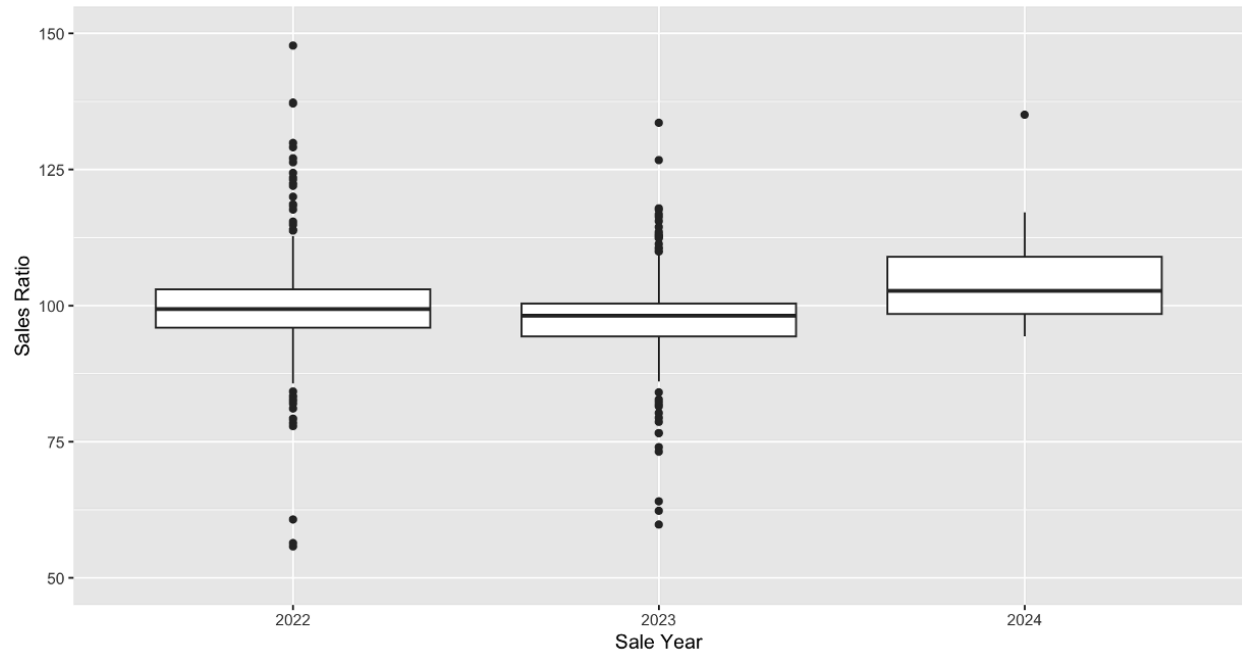
Highlights:

1. **Timing of sale** does seem to matter due to market conditions
 - a. Lower sales ratios occurred in **2023**, and higher sales ratios in **2024**
2. Being in a **block group of declining black homeowners** may matter, but the sample sizes are too small to statistically show significance
3. The **percent black population in a block group** does seem to be significant for both sales ratio and sale price, with lower sales ratios (maybe indicating flipped properties) and lower sale prices.
4. **Age of home** also does seem to be significant, with lower sales ratios (maybe indicating flipped properties), however higher sale values, which may indicate the value of more historic neighborhoods as compared to newer developments.
5. Overall, **properties in the I-540 beltline** appear to have lower sales ratios than those in the county as a whole (could indicate flipping)

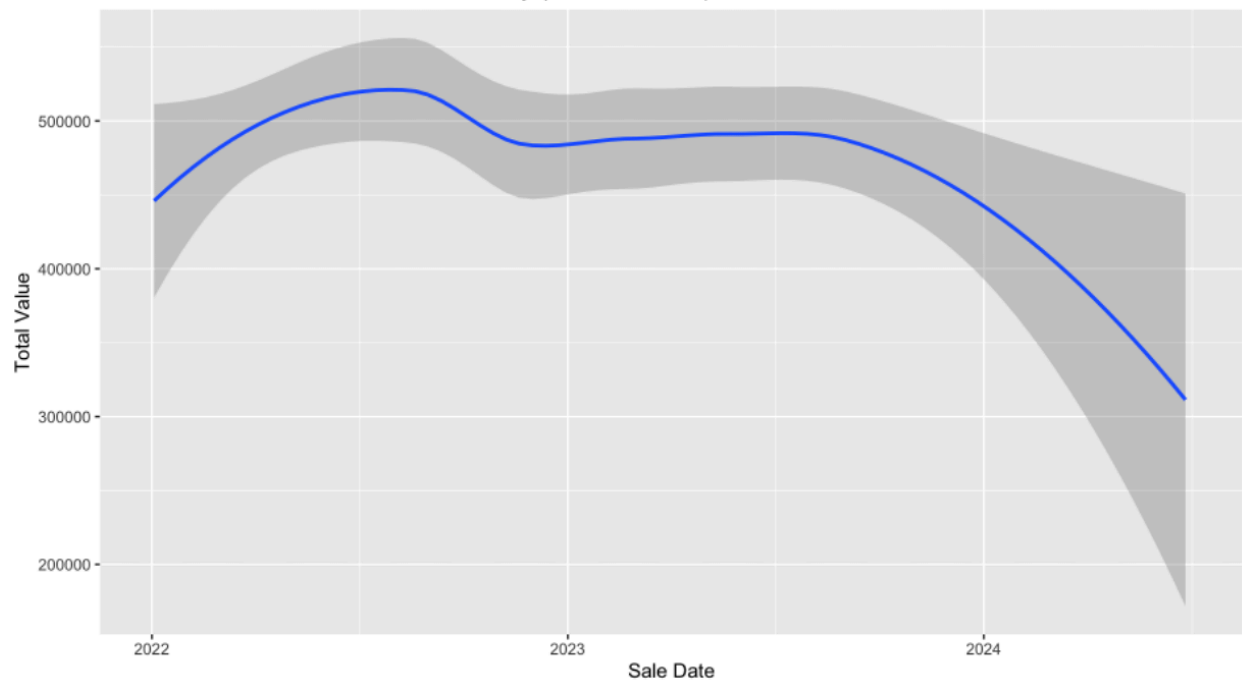
See Appendix B

Sales Ratio by Sale Year, Inner Beltline only (747 addresses)

note: removed outliers from 2022 over 150



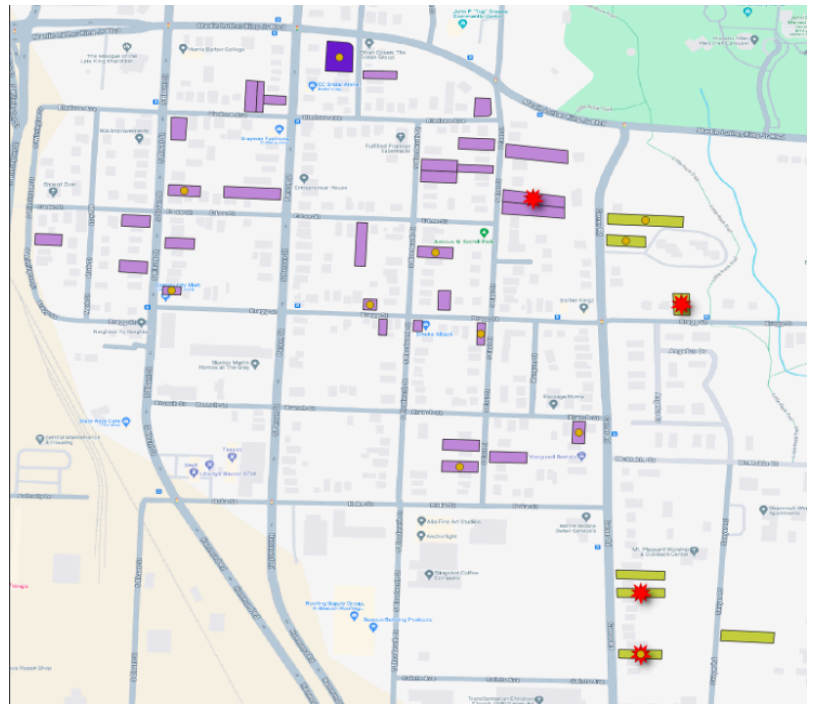
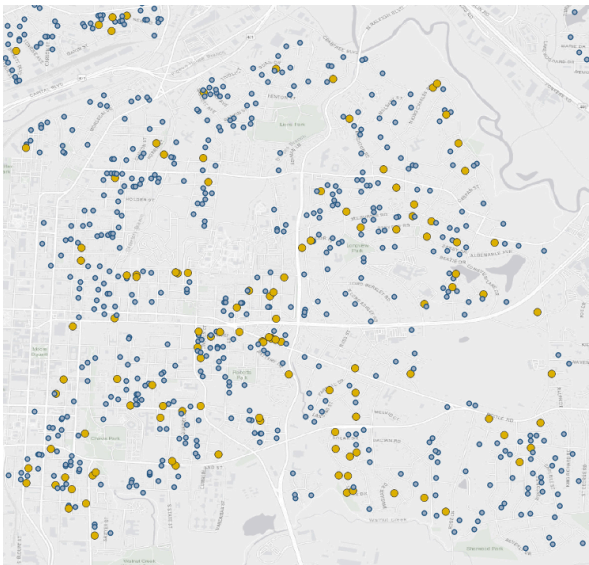
Sales Price Over Time, Inner Beltline only (747 addresses)



Part F: Next Steps

Highlights:

1. Our analysis would benefit from looking at ways to identify influential sales that influence properties around them. This kind of methodology may be beneficial and scalable to other county assessments.
2. Additionally, we hope to explore the role land value may play as a proportion of the total sale value.
3. We would like to do more work using the investment property designation to look at investment sales as a neighborhood characteristic in future analyses
 - a. Total of 2725 “Investment” Properties
 - i. 1562 “LLC” properties
 - ii. 1936 non-NC address owners (out of state investments)



LEFT CHART: The chart on the left shows a sample of “investment properties vs non investment properties mapped in the I-540 beltline region.

RIGHT CHART: The chart on the right is an example of one way to locate influential properties, and assess how they are included.

- Parcels colored by VCS
- Yellow points designate investment properties
- Red splotches are properties with sales prices more than twice the median of the neighborhood

Part G: Learnings for Durham and Orange County Analyses

1. Compare Neighborhoods
2. Identify influential Properties
3. Look at distribution of high/low sale prices
4. "Back Into" Problem Areas

Appendix A: Analysis Tables for Neighborhood Boundaries

VCS Analysis

VCS	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties
01RA545	1.03	20.14	100.29	475,000	37
01RA506	1.03	24.5	97.73	282,000	24
01RA530	1.027	10.00	98.96	632,500	10
01RA541	0.99	13.62	93.93	369,500	12

Drawn Neighborhood Approach

Neighborhood	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties
East Southpark	0.999	6.59	99.07	532,500	72
Battery Heights	0.997	6.24	97.67	410,000	61
Chavis Park	1.009	7.24	99.13	534,250	122
Current Southpark	1.02	8.02	99.71	534,250	46
Stratford Park	1.00	3.96	96.73	286,000	19
Greater Stratford Park	1.00	3.33	97.47	278,000	32
Worthdale	1.00	5.086	97.64	305,000	39

Characteristic based neighbors Approach

Grouping	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties
Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692

Property Impact Analysis:

Select several properties and fill out this table/set:

Address	Approach	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties
1107 S Wilmington St REID: 0048196 Sales Ratio: 1.03 Sale Price: 455,000	01RA545	1.03	20.14	100.29	475,000	37
	Chavis Park	1.009	7.24	99.13	534,250	122
	Current Southpark	1.02	8.02	99.71	534,250	46
	Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692
18 Washington St REID: 0102321 Sales Ratio: 95.55 Sale Price: 281,000	01RA506	1.03	24.5	97.73	282,000	24
	Stratford Park	1.00	3.96	96.73	286,000	19
	Greater Stratford Park	1.00	3.33	97.47	278,000	32
806 Postell St REID: 0002845 Sales Ratio: 88.53 Sale Price: 752,500	01RA530	1.027	10.00	98.96	632,500	10
	East Southpark	0.999	6.59	99.07	532,500	72
	Chavis Park	1.009	7.24	99.13	534,250	122
1611 Joe Louis Ave	01RA541	0.99	13.62	93.93	369,500	12

REID: 0024424 Sales Ratio: 62.32 Sale Price: 394000	Battery Heights	0.997	6.24	97.67	410,000	61
	Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692
618 E Cabarrus St REID: 0076365 Sales Ratio: 89.52 Sale Price: 1,240,000	01RA549	1.017	8.47	101.49	581,000	12
	Chavis Park	1.009	7.24	99.13	534,250	122
	Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692

Appendix B: Linear Regression Models

Model A (Sale Quarter) Output:

```
Call:
lm(formula = sale_price ~ DEED_ACRES + HEATEDAREA + bg_black_p +
    AGE + YR_QTR, data = model_final3)

Residuals:
    Min       1Q   Median       3Q      Max
-23805238  -84373    -5924    68528   37371758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -111728.8538    7479.9773  -14.937 < 0.0000000000000002 ***
DEED_ACRES   -9983.4488    3447.3063   -2.896    0.00378 **
HEATEDAREA    260.5282      0.4984  522.754 < 0.0000000000000002 ***
bg_black_p   -1826.8207     139.8992  -13.058 < 0.0000000000000002 ***
AGE           2099.2378     123.7394   16.965 < 0.0000000000000002 ***
YR_QTR22_02   38818.5619     9024.0896    4.302    0.0000170 ***
YR_QTR22_03   27630.6973     9335.6507    2.960    0.00308 **
YR_QTR22_04   10441.8481     9903.1443    1.054    0.29171
YR_QTR23_01   15959.8859    10327.9454    1.545    0.12228
YR_QTR23_02   42530.6372     9693.1240    4.388    0.0000115 ***
YR_QTR23_03   46528.3985    10065.3441    4.623    0.0000038 ***
YR_QTR23_04   46756.7946    10740.6798    4.353    0.0000135 ***
YR_QTR24_01   28147.7867    22724.6029    1.239    0.21549
YR_QTR24_02   30315.4139    20707.8569    1.464    0.14322
YR_QTR24_03   28829.9637    59974.0009    0.481    0.63073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 434000 on 30233 degrees of freedom
Multiple R-squared:  0.93,    Adjusted R-squared:  0.9299
F-statistic: 2.868e+04 on 14 and 30233 DF,  p-value: < 0.00000000000000022
```

Model A Parameters and Coefficients:

Intercept (22_01 Baseline)	Sale_YR_QTR	Deeded Acres	Heated Area	Age	% Black Population
\$-111,729	22_02: +38463 22_03: +27186 22_04: +10241 23_01: +15939 23_02: +42349 23_03: +46596 23_04: +46325 24_01: +28379 24_02: +30188 24_03: +28303	-9983 /Acre	+260 / Sq Ft	+2099 /Year	-1827 /% Black in Blockgroup

Model B (Sale Year, Simplified) Output:

```
Call:
lm(formula = sale_price ~ HEATEDAREA + bg_black_p + AGE + SALE_YEAR,
    data = model_final3)

Residuals:
    Min       1Q   Median       3Q      Max
-23821306  -84671   -5715   69361  37261333

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept)  -92231.5688    4822.2252  -19.126 < 0.0000000000000002 ***
HEATEDAREA     259.6765      0.4108  632.047 < 0.0000000000000002 ***
bg_black_p    -1829.3728    139.7554  -13.090 < 0.0000000000000002 ***
AGE             2038.4109    121.1322   16.828 < 0.0000000000000002 ***
SALE_YEAR2023  18120.0413    5136.0047    3.528    0.000419 ***
SALE_YEAR2024   9376.7163   14575.6961    0.643    0.520027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 434200 on 30242 degrees of freedom
Multiple R-squared:  0.9299,    Adjusted R-squared:  0.9299
F-statistic: 8.021e+04 on 5 and 30242 DF,  p-value: < 0.00000000000000022
```

Model B Parameters and Coefficients:

Intercept (2022 Baseline)	Sale_Year	Heated Area	Age	% Black Population
\$-92,232	2023: +18,120 2024: + 9,377	+260 /Sq Ft	+2038 /Year	-1829 /% Black Population

Model Evaluation/Comparison:

Model Name	R^2	RSE	F-statistic
Model A (Sale Quarter)	0.93	434000 on 30233 DOF	2.868e+004, p-value: < 0.00000000000000022
Model B (Sale Year, Simplified)	0.93	434200 on 30242 DOF	8.021e+004, p-value: < 0.00000000000000022

Appendix C: Full Analysis Plan (Live Working Document)

1. Introduction

- **Objective:** Our analysis plan aims to explore patterns in sales ratios and horizontal and vertical equity measures within and between property tax neighborhood delineations. It focuses on three counties, Durham, Orange, and Wake. This analysis will examine the last several years of qualified sales, but also look at the last assessment year's new assessment values compared to the previous year's qualified sales (a retrospective sales ratio analysis) to see if discernible neighborhood patterns are present. Using lessons learned from our grant partner's previous research, we will also examine factors or neighborhood characteristics which may be predictive of communities with inequity in property valuation. We will explore whether additional desired variables may be accessible for inclusion in our analysis through internet research, APIs, and web scraping.
- **Study Design:** This analysis plan has two approaches
 - A) modeling variables associated with a higher Sales Ratio, and
 - B) analyzing the impact of neighborhood boundaries and characteristic descriptions on key variables such as Price Related Differential (PRD).

2. Study Population

- **Inclusion/Exclusion Criteria:** The study population includes sales parcels from the qualified sales records for Wake County from 2022-2024.
- **Baseline Characteristics:** These parcels have been assessed and checked by local assessors to eliminate extreme outliers that may indicate a non-representative market value.

3. Variables

Modeling Approach

- **Dependent Variable(s):** Sales Ratio (SR)
- **Independent Variable(s):**
 - Sale QR+YR (categorical) [SALE_DATE]-> factored as [YR_QTR]
 - Heated square feet (continuous, standardize?) [HEATEDAREA]
 - Acreage (continuous)- [DEED_ACRES]
 - Year built (continuous, standardize?) [YEAR_BUILT]

- ~~Effective year built (continuous) --- check for collinearity with year built pick one~~
- % Black population (continuous)--- run with and without [bg_black_p]
- Location - [PLANNING_J]
- Gentrification proxy (decrease in black home owners) - [25decbhown]
- ~~Sale price (put in groups?) (categorical?)~~
- ~~Land value? (put in groups?) (binary flag?)~~

Neighborhood Boundary Approach

- Heated square footage
- Acreage
- Year built (effective year built)
- Proximity value/distance
- *Other variables as informed by modeling approach....*

4. Statistical Methods (Raleigh)

Modeling Approach

- **Descriptive Statistics/Pre-modeling work:**
 - Assess for correlations and collinearity of variables
 - Determine if any transformations are needed (log, standardization, etc)
- **Inferential Statistics:**
 - Linear Regression model
 - 80% Train , 20% Test data split
 - Cross validation
 - Model Tuning
 - Model Evaluation

Neighborhood Boundary Approach

- **Equity Measures (calculate by VCS, labeled names, characteristic-based groups):**
 - PRD
 - COD
 - Median Sales Ratio
 - Median Sales Value
 - No. of Comparison Properties
 - Visualization of scatterplot for boundary

5. Data Management (Raleigh)

- **Data Cleaning:**

Data still to explore?

Sales rate within a neighborhood? How to calculate?

Datasets:

Wake County Blockgroups (Census)
Wake County Parcels (Wake County GIS)
Race Data Tables (created in R using Census API)
Qualified Sales Dataset with Calculations (Wake County, with R calculations)

Pre- Analysis: (completed - 7/16)

- Create demographic data tables by blockgroup using Census API (R)
- Calculate race percentages (in R), export and load table into QGIS
- Download Qualified sales data and calculate neighborhood variable sin R by VCS (R)
- Merge qualified sales data with larger real estate data file from wake county, remove duplicates and select most recent, single home, non-townhouse/condo properties (R)
- Export and load qualified sales addresses (with REID and calculations) into QGIS
- Download and map the real estate parcel and census block group shapefiles (in QGIS)
- Merge qualified sales addresses with parcels layer on REID (QGIS)
- Map the centroids of the qualified parcel data to create a point layer (QGIS)
- Use a spatial join to merge the race data on to qualified parcel points dataset (QGIS)
- Export qualified parcel dataset as table (with race data, and calculations, etc) (QGIS)

Create Final Modeling Dataset (in R) (in progress- goal completion 8/9)

- Create a sale quarter + year categorical variable
- Create a location category or use a zoning one?--- Planning_Jurisdictionproxy
- Create a “active gentrification” proxy – Percent decrease in black home owners
- Trim to required modeling fields:
 - Sale QR+YR (categorical)
 - Heated square feet (continuous)
 - Agerage (continuous)
 - Year built (continuous)
 - Planning Jurisdiction (categorical)
 - Gentrification Proxy- 25% or greater reduction in black home own (binary)
 - Others:

- *Effective year built (continuous) == check for collinearity with year built-pick one*
 - *% Black population (continuous)--- run with and without*
 - *Sale price (put in groups?) (categorical?)*
- OUTCOME VARIABLE: Sales Ratio (continuous- but need a transformation?)
- Export to Modeling data folder

Create sample Neighborhood Boundary Dataset(s): (complete for some neighborhoods- iterative development, will move forward with some, learn, move to other neighborhoods)

- Create three datasets from the exported qualified parcel dataset
 - All qualified sales
 - Qualified sales within beltline (manually drawn)
 - Neighborhood(s) of interest dataset(s) (manually drawn)
 - East Southpark - W side of Tarboro of 507 and all 508
 - Battery Heights - E side Tarboro of 507
 - Chavis Park - All west of Tarboro (507,509, w side of 507)
 - Current Southpark – All of 509
 - Stratford Park - VCS of 01RA506 within 52002
 - Greater Stratford - All west of Dagian St within 52002
 - Worthdale - VCS of 01RA506 within 52002
 - *Still to determine (52101 and 52102)*
- Export tables to Wake Data Folder
- **Data Transformation:** Specify any transformations (e.g., log transformation) planned.

6. Statistical Software

- **Software:** Specify the statistical software package(s) to be used (e.g., R, SAS, SPSS).

7. Reporting and Interpretation

- **Results Presentation:**

Analysis Table Shells for Linear Regression Models

Model A (Sale Quarter) Output:

```
Call:
lm(formula = sale_price ~ DEED_ACRES + HEATEDAREA + bg_black_p +
    AGE + YR_QTR, data = model_final3)

Residuals:
    Min       1Q   Median       3Q      Max
-23805238  -84373    -5924    68528   37371758

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -111728.8538    7479.9773  -14.937 < 0.0000000000000002 ***
DEED_ACRES   -9983.4488    3447.3063   -2.896    0.00378 **
HEATEDAREA    260.5282     0.4984  522.754 < 0.0000000000000002 ***
bg_black_p   -1826.8207    139.8992  -13.058 < 0.0000000000000002 ***
AGE           2099.2378    123.7394   16.965 < 0.0000000000000002 ***
YR_QTR22_02   38818.5619    9024.0896    4.302    0.0000170 ***
YR_QTR22_03   27630.6973    9335.6507    2.960    0.00308 **
YR_QTR22_04   10441.8481    9903.1443    1.054    0.29171
YR_QTR23_01   15959.8859   10327.9454    1.545    0.12228
YR_QTR23_02   42530.6372    9693.1240    4.388    0.0000115 ***
YR_QTR23_03   46528.3985   10065.3441    4.623    0.0000038 ***
YR_QTR23_04   46756.7946   10740.6798    4.353    0.0000135 ***
YR_QTR24_01   28147.7867   22724.6029    1.239    0.21549
YR_QTR24_02   30315.4139   20707.8569    1.464    0.14322
YR_QTR24_03   28829.9637   59974.0009    0.481    0.63073
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 434000 on 30233 degrees of freedom
Multiple R-squared:  0.93,    Adjusted R-squared:  0.9299
F-statistic: 2.868e+04 on 14 and 30233 DF,  p-value: < 0.00000000000000022
```

Model A Parameters and Coefficients:

Intercept (22_01 Baseline)	Sale_YR_QTR	Deeded Acres	Heated Area	Age	% Black Population
\$-111,729	22_02: +38463 22_03: +27186 22_04: +10241 23_01: +15939 23_02: +42349 23_03: +46596 23_04: +46325 24_01: +28379 24_02: +30188 24_03: +28303	-9983 /Acre	+260 / Sq Ft	+2099 /Year	-1827 /% Black in Blockgroup

Model B (Sale Year, Simplified) Output:

```
Call:
lm(formula = sale_price ~ HEATEDAREA + bg_black_p + AGE + SALE_YEAR,
    data = model_final3)

Residuals:
    Min       1Q   Median       3Q      Max
-23821306  -84671   -5715   69361  37261333

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -92231.5688    4822.2252  -19.126 < 0.0000000000000002 ***
HEATEDAREA     259.6765      0.4108  632.047 < 0.0000000000000002 ***
bg_black_p    -1829.3728    139.7554  -13.090 < 0.0000000000000002 ***
AGE           2038.4109    121.1322   16.828 < 0.0000000000000002 ***
SALE_YEAR2023  18120.0413    5136.0047    3.528    0.000419 ***
SALE_YEAR2024   9376.7163   14575.6961    0.643    0.520027
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 434200 on 30242 degrees of freedom
Multiple R-squared:  0.9299,    Adjusted R-squared:  0.9299
F-statistic: 8.021e+04 on 5 and 30242 DF,  p-value: < 0.00000000000000022
```

Model B Parameters and Coefficients:

Intercept (2022 Baseline)	Sale_Year	Heated Area	Age	% Black Population
\$-92,232	2023: +18,120 2024: + 9,377	+260 /Sq Ft	+2038 /Year	-1829 /% Black Population

Model Evaluation/Comparison:

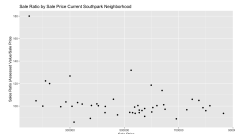
Model Name	R^2	RSE	F-statistic
Model A (Sale Quarter)	0.93	434000 on 30233 DOF	2.868e+004, p-value: < 0.000000000000000022
Model B (Sale Year, Simplified)	0.93	434200 on 30242 DOF	8.021e+004, p-value: < 0.000000000000000022

Analysis Table Shells for Neighborhood Boundaries Analysis:

VCS Analysis

VCS	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties	Scatterplot
01RA545	1.03	20.14	100.29	475,000	37	
01RA506	1.03	24.5	97.73	282,000	24	
01RA530	1.027	10.00	98.96	632,500	10	
01RA541	0.99	13.62	93.93	369,500	12	

Drawn Neighborhood Approach

Neighborhood	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties	Scatterplot
East Southpark	0.999	6.59	99.07	532,500	72	
Battery Heights	0.997	6.24	97.67	410,000	61	
Chavis Park	1.009	7.24	99.13	534,250	122	
Current Southpark	1.02	8.02	99.71	534,250	46	
Stratford Park	1.00	3.96	96.73	286,000	19	
Greater Stratford Park	1.00	3.33	97.47	278,000	32	
Worthdale	1.00	5.086	97.64	305,000	39	

Characteristic based neighbors Approach

Grouping	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties	Scatterplot
Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692	

Property Impact Analysis:

Select several properties and fill out this table/set:

Address	Approach	PRD	COD	Median Sales Ratio	Median Sales Value	No. of Comparison Properties	Scatterplot
1107 S Wilmington St REID: 0048196 Sales Ratio: 1.03 Sale Price: 455,000	01RA545	1.03	20.14	100.29	475,000	37	
	Chavis Park	1.009	7.24	99.13	534,250	122	
	Current Southpark	1.02	8.02	99.71	534,250	46	
	Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692	
18 Washington St REID: 0102321 Sales Ratio: 95.55 Sale Price: 281,000	01RA506	1.03	24.5	97.73	282,000	24	
	Stratford Park	1.00	3.96	96.73	286,000	19	
	Greater Stratford Park	1.00	3.33	97.47	278,000	32	

806 Postell St REID: 0002845 Sales Ratio: 88.53 Sale Price: 752,500	01RA530	1.027	10.00	98.96	632,500	10	
	East Southpark	0.999	6.59	99.07	532,500	72	
	Chavis Park	1.009	7.24	99.13	534,250	122	
1611 Joe Louis Ave REID: 0024424 Sales Ratio: 62.32 Sale Price: 394000	01RA541	0.99	13.62	93.93	369,500	12	
	Battery Heights	0.997	6.24	97.67	410,000	61	
	Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692	
618 E Cabarrus St REID: 0076365 Sales Ratio: 89.52 Sale Price: 1,240,000	01RA549	1.017	8.47	101.49	581,000	12	
	Chavis Park	1.009	7.24	99.13	534,250	122	
	Black Home ownership decreased by 25% +	1.00	7.10	98.29	417,750	692	

- **Interpretation:** Describe how the results will be interpreted in relation to the study objectives.
- **Subgroup Analyses (if applicable):** Specify any planned subgroup analyses.

Data Management (Durham)

Datasets:

Durham County Blockgroups (Census)
Durham County Parcels (Durham County GIS)
Race Data Tables (created in R using Census API)

*Qualified Sales Datasets with Calculations (Durham County, with R calculations)
(2020,2021,2022)*

Pre- Analysis:

- Create demographic data tables by blockgroup using Census API (R)
- Calculate race percentages (in R), export and load table into QGIS
- Download Qualified sales data and calculate neighborhood variable in R by VCS (R)
- Download and map the real estate parcel and census block group shapefiles (in QGIS)
- Map the centroids of the qualified parcel data to create a point layer (QGIS)
- Use a spatial join to merge the race data on to qualified parcel points dataset (QGIS)
- Export qualified parcel dataset as table (with race data, and calculations, etc) (QGIS)
- Merge qualified sales data with larger real estate data file from wake county, remove duplicates and select most recent, single home, non-townhouse/condo properties (R)
- *Export and load qualified sales addresses (with REID and calculations) into QGIS*
- *Merge qualified sales addresses with parcels layer on REID (QGIS)*

Neighborhood Deep Dives:

- Create a Crosswalk between Dataworks NC Compass neighborhoods and VCS / Tax labeled “neighborhoods”
- Use Neighborhood shapefile from Dataworks NC Compass to create subsets of the parcel points by neighborhood for focus neighborhoods, including:
 - Unity Village
 - Merrick Moore
 - Bragtown
 - Crest St
 - Cleveland-Holloway
 - College View
 - (Walltown)

VCS Charts:

1. Identify VCS neighborhoods with High Regressivity for each year of qualified Data (2020-2022)
 - Residential Properties Only (no vacant land)
 - 90th percentile of COD SR (>31)
 - PRD > 1.03

- 50th percentile of comparison properties (>28)
- 2. Create scatterplots for Sale price x Sale Ratio
- 3. Calculate Qualified Sale Quartiles for each VCS
- 4. Create Chart for “Undervalued” vs “Overvalued” for High Regressivity neighborhoods
 - a. Overvalued if Sale Ratio > 1, Undervalued if Sale Ratio < 1, Appropriate if = 1
 - b. Show percent overvalued vs undervalued by Qualified Sale Quartile
- 5. Create Chart for “Undervalued” vs “Overvalued” for High Regressivity neighborhoods based on that qualified year’s median VCS sale ratio
 - a. Overvalued if indiv SR- VCS median >0, undervalued if indiv SR-VCS med < 0, Appropriate if =0

Quality Model

DV:

- Overvalued vs not Overvalued

IV:

- Quality Grade
- Year Built
- Assessed Value

Create Final Modeling Dataset (in R)

- *Create a sale quarter + year categorical variable*
- *Create a location category or use a zoning one?--- Planning_Jurisdictionproxy*
- *Create a “active gentrification” proxy – Percent decrease in black home owners*
- *Trim to required modeling fields:*
 - *Sale QR+YR (categorical)*
 - *Heated square feet (continuous)*
 - *Acerage (continuous)*
 - *Year built (continuous)*
 - *Planning Jurisdiction (categorical)*
 - *Gentrification Proxy- 25% or greater reduction in black home own (binary)*
 - *Others:*
 - *Effective year built (continuous) == check for collinearity with year built- pick one*
 - *% Black population (continuous)--- run with and without*
 - *Sale price (put in groups?) (categorical?)*
 - *OUTCOME VARIABLE: Sales Ratio (continuous- but need a transformation?)*

8. Ethics and Confidentiality

- **Ethical Considerations:** Address any ethical considerations related to data analysis.
- **Confidentiality:** Detail measures to ensure participant confidentiality.

9. Timeline

- **Schedule:** Provide a timeline for completing the statistical analysis.

2024

By August 9 (for Aug 15 report)

Initial Model Findings

Initial Neighborhood Equity Comparisons

August

Iterate on and Evaluate Model

Final Neighborhood Equity Boundaries Analysis

September

Interpret and write up model findings

Individual property boundaries analyses

Create initial materials to share with neighborhoods, get feedback from stakeholders

October

Add-in results from characteristic-based group Analysis

Share results with community

November

Draft methodology for Durham and Orange counties

Write final report and documentation for project

December

Final report to NNIP

10. References

- **Literature:** List references to relevant statistical methods and previous studies.