Quantifying Gentrification Technical Report

Amy Boncelet
Jacobs Technion-Cornell Institute
Cornell Tech
New York City, United States of America
ajb347@cornell.edu

Matthew Shen
Jacobs Technion-Cornell Institute
Cornell Tech
New York City, United States of America
mds377@cornell.edu

Thomas Wallace
Jacobs Technion-Cornell Institute
Cornell Tech
New York City, United States of America
tw526@cornell.edu

***Abstract*—The following document is a technical write-up on how to quantify gentrification. This project was done as part of INFO 5430: Urban Data at Cornell Tech.**

***Index Terms*—Urban Technology, Urban Data, Data Science, Information Systems**

I. Introduction

The objective of this project was to evaluate New York City demographic characteristics determine which neighborhoods have gentrified over the past decade.

II. Analysis

A. How we defined gentrification

In order to define gentrification we looked at six different metrics. These metrics can be seen in Table I.

TABLE I
Gentrification Metrics

Increase Gentrification	Decrease Gentrification
% White Population*	Median Age
Median Income	% Foreign Born(Non-US Citizen)*
% Bachelor's Degree*	
% Median Rent	

Per Capita Calculation Required*

B. Project Scope

Data was collected on the different metrics found in Table I through the US Census Bureau. Data between 2011 and 2019 was analyzed as part of this project. These dates were selected because they sat between the decennial censuses in 2010 and 2020. This timline also allowed for the data to not be influenced by the 2020 COVID-19 pandemic. ZCTA5 values were used as the baseline geo-location. A more granular baseline geo-location would have been census tracts; however, our project looks at long term changes in areas and census tracts change from year to year. Meanwhile, ZCTA5 values remain relatively constant. New York City was selected as the test bed for our analysis.

C. Methodology

1) Year Organization: The initial data that we received from the United States Census Bureau was collected from four different surveys. These different surveys can be seen in Table II. Each of the surveys also came in separate files for each year. As a result, we had a total of 36 different data sets. These data sets were then organized into each individual year. This meant that we had a consolidated data set with all the survey data for each year.

TABLE II
US Census Bureau Surveys

Survey Code	Survey Description
DP05	ACS Demographic & Housing Estimates
S1901	Income in the Past 12 Months
DP02	Selected Social Characteristics
DP04	Selected Housing Characteristics

2) Column Metadata: Unfortunately the column metadata changed for different variables at different points in our analysis. As a result, we had to individually go through each data set and determine how the column names changed with each year. The relevant columns used to subset can be seen in Table A. Once all the columns were renamed we subset the data to only include the ZCTA5, the data in Table I, and total population.

3) Cleaning Data Types: We then found that some of the data were not float64. As a result we had to either convert the data to a float64 or remove the data. We also found that the ZCTA5 data was in the format of "ZCTA5 00000". To mitigate this we striped the "ZCTA5 " and then changed the five digit code to an integer. Another problem we found in the data was that some ZCTA5 values did not have a population. This was likley due to some new buildings which have independent ZCTA5 values but do not have any residents. As a result, we decided to remove these values. The final minor data cleaning measure was to subset the ZCTA5s to only include ones in NYC.

4) Interpolating Missing Data: Once each data set was cleaned and subset we noticed that some data was sporadically missing. In order to fill in this data we interpolated with the nearby ZCTA5 values. This was done because we can make the assumption that ZCTA5 values near each other would have similar statistical values.

5) Per Captia Calculations: The white population, bachelor's degrees, and foreign born(non-US citizen) data was given to us as raw numbers for each ZCTA5. We needed to ensure to make sure that an increase in the white population, bachelor's degrees, and foreign born(non-US citizen) is not due to an increase in population but a change in the percentage of white people, bachelor's degrees, or foreign born(non-US citizen).

6) Calculate change over time: The next step was the look at how our different metrics changed over time. To do this we merged the cleaned data sets form 2011 and 2019 on the ZCTA5s. This would allow us to remove any ZCTA5s that did not exist in both data sets. We then found the difference between 2011 and 2019. Next, made the metrics in Table I that decrease gentrification into negatives. This ensure that an increase in those values results in a lower gentrification score.

7) Normalize data: We finally normalized the data with Equation 1.

$$\frac{x - \mu}{\sigma} \tag{1}$$

In Equation 1 x is the value we are changing, μ is the mean of the series that data set is in, and σ is the standard deviation of the series that data set.

8) Creating the Gentrification Metric: Now that our data is normalized we can create our gentrification metric by adding all the values in Table I for each ZCTA5.

III. Results

The gentrification metric was then plotted. This visualization can be seen in Appendix A.

APPENDIX
COLUMN NAMES OVER TIME

TABLE III
BACHELOR'S DEGREE

Year	Code
2011	DP02_0064E
2012	
2013	
2014	
2015	
2016	
2017	
2018	
2019	DP02_0065E

TABLE IV
FOREIGN BORN, NOT A US CITIZEN

Year	Code
2011	DP02_0095E
2012	
2013	
2014	
2015	
2016	
2017	
2018	
2019	DP02_0096E

TABLE V
MEDIAN INCOME

Year	Code
2011	S1901_C01_012E
2012	
2013	
2014	
2015	
2016	
2017	
2018	
2019	

TABLE VI
WHITE NON-HISPANIC

Year	Code
2011	DP05_0072E
2012	
2013	
2014	
2015	
2016	
2017	DP05_0077E
2018	
2019	

TABLE VII
MEDIAN AGE

Year	Code
2011	DP05_0017E
2012	
2013	
2014	
2015	
2016	
2017	DP05_0018E
2018	
2019	

TABLE VIII
RENT - GROSS RENT MEDIAN PRICE

Year	Code
2011	DP04_0132E
2012	
2013	
2014	
2015	DP04_0134E
2016	
2017	
2018	
2019	

APPENDIX
GENTRIFICATION MAP