Modeling Gentrification in NYC: Time Series and Neural Approaches

Tyler Mckenzie and Ben Tunney
github

I. Introduction and Literature Review

This project explores gentrification across NYC sub-borough areas from 2005 to 2019. We apply time series techniques to analyze neighborhood changes, using median household income as a primary indicator. We apply a convolutional neural network in attempts to classify gentrification based on a variety of neighborhood-specific demographic features. The time series analysis presented on this poster focuses specifically on Bushwick, and the neural network is applied to all neighborhoods.

Gentrification is a process that takes place in low-income neighborhoods of urban areas that increases housing prices, displaces existing residents, and disrupts communities. Building a machine learning framework to track changes in income level and predict gentrification could be integral to policy makers and citizens in order to implement early intervention methods and protect communities. We propose our application of time series analysis and gentrification classification model as an intervention and monitoring method for this problem.

Our first reference for defining gentrification computationally was the NYU Furman Center's *State of New York City's Housing and Neighborhoods in 2025* report. In this report the Furman center develops techniques to track changing rent in New York City and analyzes accompanying demographic characteristics of the corresponding sub-boroughs. Their definition of gentrification begins with determining which neighborhoods were low-income in 1990, "in the bottom 40 percent of neighborhoods with respect to average household income" (Austensen et al., 2016). After determining low-income status, the gentrification label is determined by tracking which neighborhoods "experienced rent growth above the median SBA rent growth between 1990 and 2010-2014" (Austensen et al., 2016). Neighborhoods that are not classified low-income in 1990 are exempt from gentrification status and labeled high-income.

The Furman Center reports disparity in many NYC neighborhood demographics including increasing rents, housing units count, and average household income. Gentrifying neighborhoods experienced a 30.4 average percent increase in mean household rent from 2000 to 2010-2014 compared to 16.1 percent in non-gentrifying neighborhoods. Housing units increased by 7.2 percent for gentrified neighborhoods from 2000-2010, and only 5.5 percent for non-gentrified neighborhoods. Gentrifying neighborhoods experienced a 6.1 percent increase in average household income from 2000 to 2010-2014 while non-gentrifying neighborhoods experienced almost an 8 percent decrease.

The U.S. Census Bureau's *Identifying Gentrification using Machine Learning* provides a valuable example of ML applications for forecasting gentrification and a strong model for training data in this area. In this paper the U.S. Census Bureau uses a combination of sources for the demographic data for the Washington D.C. area, builds a gentrification label, and applies five classification models in order to predict for the gentrification indicator.

The primary source for data is the American Housing Survey which provides socioeconomic and housing features like householder age, family type, unit size, and year built. The rest of the merged dataset features were related to household income, education, amenity, environmental risks, walking score, and education, and these data were sourced from the American Community Survey (2015, 2019) and various independent neighborhood metrics websites. This dataset serves as a comprehensive representation of D.C. neighborhoods on various axes.

This paper presents a three-pronged approach at building the gentrification indicator including household members, household income growth, and social mobility.

> "All household members in 2017 or 2019 are different than the members in 2015; household income growth (from 2015 to 2017 or from 2015 to 2019) exceeds the tract level growth rate from ACS; and household members reported moving for better jobs, homes, or neighborhoods." (Yoo, 2023)

The U.S. Census Bureau tracked gentrification for D.C. neighborhoods across these three axes using the aforementioned demographic data.

The applied machine learning models were logistic regression, k-nearest neighbors classifier, random forest, support vector machines, and gradient boosting. The U.S. Census Bureau found that the random forest classifier performed the best at 83 percent accuracy and confirms validity in exploring machine learning solutions for the gentrification forecasting problem.

Jonathan Reades's *Understanding urban gentrification through machine learning* provides a succinct definition of gentrification and justifies the place of machine learning in urban studies. The authors argue against the implication that quantitative analysis of gentrification is less valuable than analysis of "media analysis, interviews, ethnography and other forms of observational data collection" (Reades et al., 2018).

Some of the primary arguments for this application of machine learning in urban studies include scalability, explainability, and tracking. The authors emphasize the ability of big data solutions to analyze massive datasets with many features, which is important and valuable in real-world problems that have increasing data availability. Additionally, they emphasize the "availability and openness of these tools – they are not 'black boxes'" (Reades et al., 2018), ML technicians can provide valuable insights into how specific features power ML solutions and provide valuable warning frameworks for stakeholders. Finally, the authors present the optimistic potentiality of these ML models being used to predict urban changes and to influence policy decision-making.

The authors employ a random forest model to predict neighborhood change in London in 2011 and 2021 using data from 2000-2010. The target metric is a combined socioeconomic variable built from household income, occupational share, resident qualifications, and property sale value. This model presents another example of the value of using quantitative techniques for urban studies.

II. ML Methodology and Data Collection

Data Source and Feature Selection:

We sourced all data from CoreData NYC. This project focused on Sub-Borough Areas (SBAs) as our unit of analysis. Feature selection was guided by the U.S. Census Bureau's 2023 paper, Identifying Gentrification Using Machine Learning, which emphasized a wide range of **socioeconomic** and **neighborhood features**. The socioeconomic features included income diversity, racial diversity, poverty rate, educational attainment, homeownership rate, and foreclosure notices. Some additional neighborhood features variables, such as car-free commute rates, park and subway access, and public school performance metrics, were collected at the Community District level and mapped to SBAs using classification tables from the NYU Furman Center. Variables were merged using Sub-Borough Area names as unique identifiers.

Labeling Gentrification for CNN:

To create the labels "Gentrifying," "Non-Gentrifying," or "Higher-Income," we implemented a labeling approach inspired by the NYU Furman Center's State of the City 2015 report. In the Furman Center's classification, gentrifying neighborhoods are those that were low-income in 1990 and experienced rent

growth **higher than the citywide** median between 2015 and 2017. Using this approach, we used 2005 as a baseline year for household income and classified neighborhoods into three categories (using median income as a proxy for rent):

1. Higher-Income: SBAs with 2005 median household income above the citywide median
2. Gentrifying: SBAs with 2005 income below the citywide median and income growth from 2005 to 2019 above the median growth across all SBAs
3. Non-Gentrifying: SBAs with 2005 income below the citywide median and income growth from 2005 to 2019 below the median growth

We used the following code to compute the income change and label:

```
df['income_growth'] = (df['income_median_2019'] - df['income_median_2005']) / df['income_median_2005']
baseline_median_income = df['income_median_2005'].median()
median_growth = df['income_growth'].median()

def label_neighborhood(row, income_thresh, growth_thresh):
    if row['income_median_2005'] >= income_thresh:
        return "Higher-Income"
    else:
        if row['income_growth'] > growth_thresh:
            return "Gentrifying"
        else:
            return "Non-Gentrifying"

df['gentrification_label'] = df.apply(
    label_neighborhood,
    axis=1,
    args=(baseline_median_income, median_growth)
)
```

Overview of ML Models:

To analyze and predict gentrification, we implemented both time series and convolutional neural network models.

ARMA Time Series Modeling: We applied an Autoregressive Moving Average (ARMA) model to explore median income trends over time for a subset of SBAs. ARMA models combine two components:

- AR(p): An autoregressive model of order p uses previous values (lags) of the time series.
- MA(q): A moving average model of order q

The ARMA(p,q) model is defined as:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

Where:

- y□ is the value of the series at time t
- c is a constant term
- φ_i shows influence of past terms, depending on p,q values
- θ□ is the moving average parameter
- ε□ is error at time t

Convolutional Neural Network (CNN):

In addition to time series analysis, we trained a 1D convolutional neural network (CNN) to classify SBAs as gentrifying, non gentrifying or high income. We used **socioeconomic** and **neighborhood features** across multiple years (2005–2019), as previously stated. The CNN was implemented using R and consisted of:

- Randomly initialize weights and biases
- A 1D convolutional layer
- Leaky ReLU activation function
- Max pooling to reduce dimensionality
- A dense layer
- A sigmoid output unit for classification

Instead of backpropagation we updated weights according to loss changes using numerical gradient estimation, minimizing binary cross-entropy loss. This is a variation of the central difference formula, which approximates the first derivative of a function. We perturb each parameter (loss1, loss2, grad_step) by 1e-5, which helps provide the direction in which each parameter should move in order to reduce loss. Finally we use a gradient descent update (params - lr * params).

Central Difference Formula:

$$f'(x) \simeq \frac{f(x + \Delta x) - f(x - \Delta x)}{2\Delta x}$$

Modified Central Difference Formula:

grad = (loss1 - loss2) / (2 * grad_step)

Binary Cross Entropy Loss:

$$L_{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(Y_i \cdot \log \hat{Y}_i + (1 - Y_i) \cdot \log (1 - \hat{Y}_i))$$

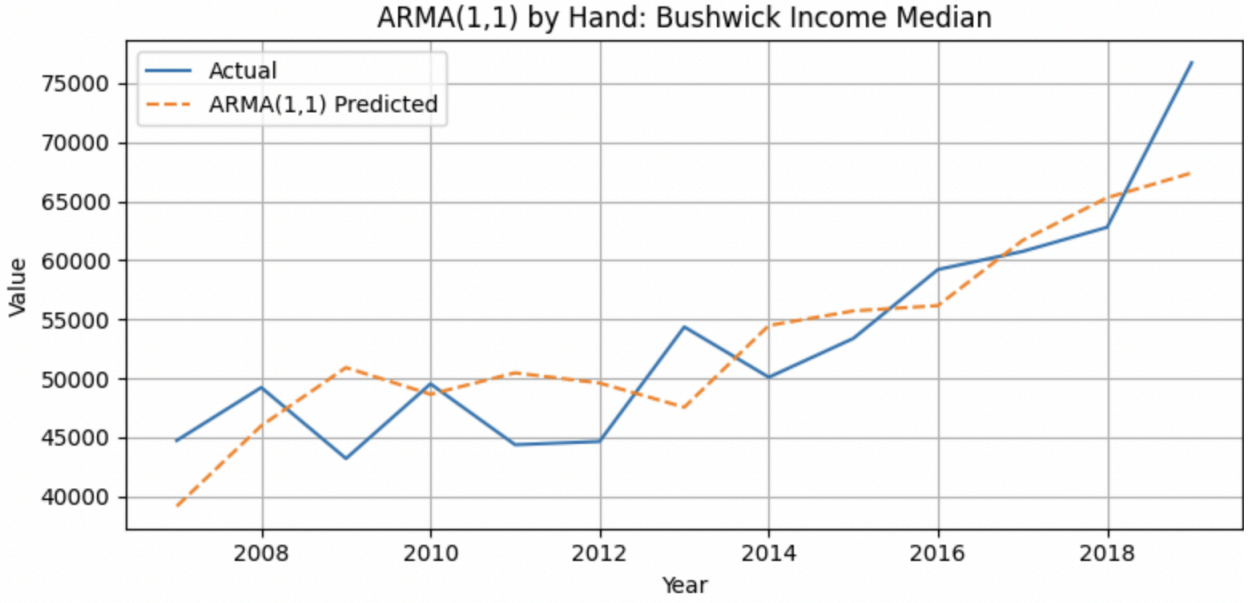
Together, these models capture key insights on how gentrification is reflected in income and housing data across neighborhoods (SBAs) in NYC.

III. ML Results and Interpretation

ARMA Time Series Modeling:

We applied ARMA time series modeling to a subset of SBAs we identified as either gentrifying or non-gentrifying in the test set of the CNN, in order to take a deeper look inside the income trends in specific SBAs, using our labels to help choose which we should analyze under what lense. In this report, we look at one gentrifying neighborhood (West Village) and one non-gentrifying neighborhood().

Model 1: Bushwick



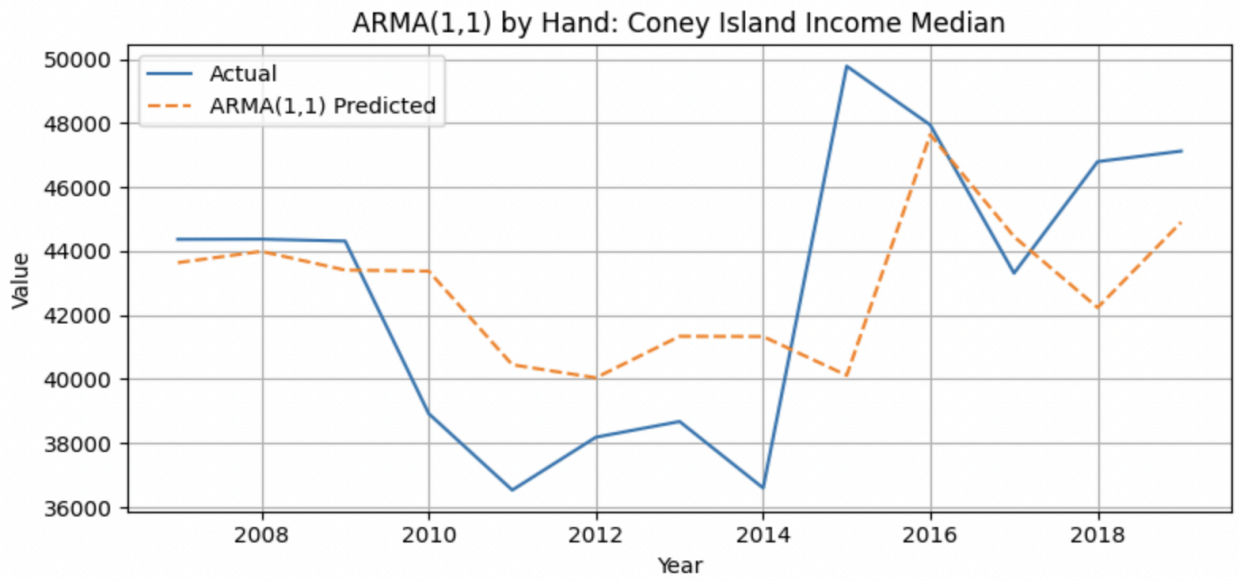
We chose the ARMA(1,1) model because it struck the best balance between capturing the underlying trends and avoiding overfitting, especially compared to AR(1) and ARMA(2,2). Both the autoregressive (AR) and moving average (MA) components were necessary. When decomposing the data, we observed a clear upward trend, which indicated the need for an AR term to incorporate past values into future predictions. Additionally, the residual plot showed remaining structure rather than random noise, suggesting that an MA term was also needed to account for the error from previous time steps.

The MAE of the model is $4,442.54, which tells us that the model's predictions deviate from the actual median income values by approximately that margin, which is not too large considering the scale of income. $5,095.02, reflecting the presence of some larger individual errors, though not to a degree that would suggest instability. This suggests the model is fairly consistent in its predictions. The Mean Absolute Percentage Error (MAPE) was 8.61%, meaning the model's predictions were off by around 8.6% on average.

As can be seen in the above figure, there is an upward trend starting in around 2012. There was a consistent low rent price, including the 2008 stock market crash. Unlike areas such as the Upper West Side, which experienced a sharp dip and a multi-year recovery following the crash, Bushwick's income trajectory surpassed pre-2008 levels more rapidly. This can be explained by Bushwick's transformation over the past decade from a historically working-class, predominantly Latinx neighborhood into a site of intense gentrification. Starting in the early 2010s, an influx of young artists, students, and professionals arrived in Bushwick, after being priced out of nearby Williamsburg. As the neighborhood attracted more real estate investment and new development, median incomes began to rise significantly (Navarro, 2016; New York Times).

This corroborated NYU Furman Centers findings, that Bushwick was one of the fastest growing neighborhoods with steep demographic shifts.

Model 2: Coney Island



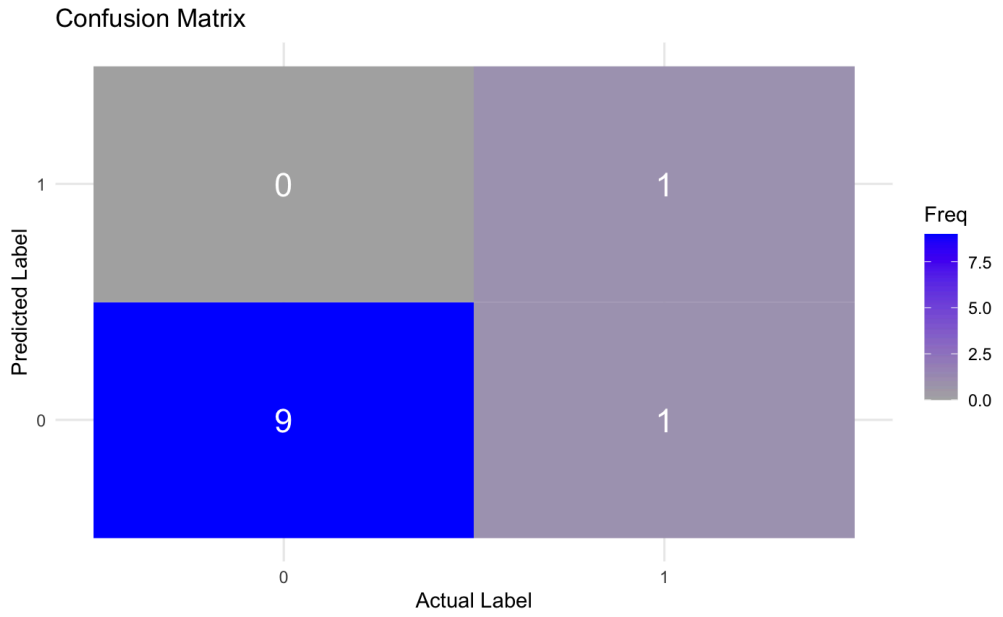
We chose ARMA(1,1) for Coney Island for similar reasons as Bushwick, it was able to loosely capture trends without overfitting to the data. The model's MSE was $2,886.69, the RMSE was $3,823.71. This shows that while most predictions were close, there were some larger errors which skewed the RMSE to be around $1000 higher. The MAPE was 6.8%, suggesting the model's predictions were relatively accurate, especially given the limited years and volatility in Coney Island's income trajectory.

As you can see in the model, the median income in Coney Island dipped post 2008 and stayed consistently low until 2014. This can be attributed to the 2008 housing crisis, and also Hurricane Sandy in 2012. Both damaged the economy and quality of life in Coney Island, as it is a coastal area. After Hurricane Sandy in 2012, Coney Island became the focus of several city- and state-led recovery and resiliency efforts, including housing redevelopment, infrastructure repair, and community investment. These interventions likely contributed to short-term income increases but did not lead to the kind of structural socioeconomic change associated with gentrification (NYC Planning, 2014; NYU Furman Center, 2016). As you can see, incomes return to around the same level as pre-2008 incomes, while in the case of Bushwick, incomes surpass 2008 levels greatly.

Convolutional Neural Network (CNN):

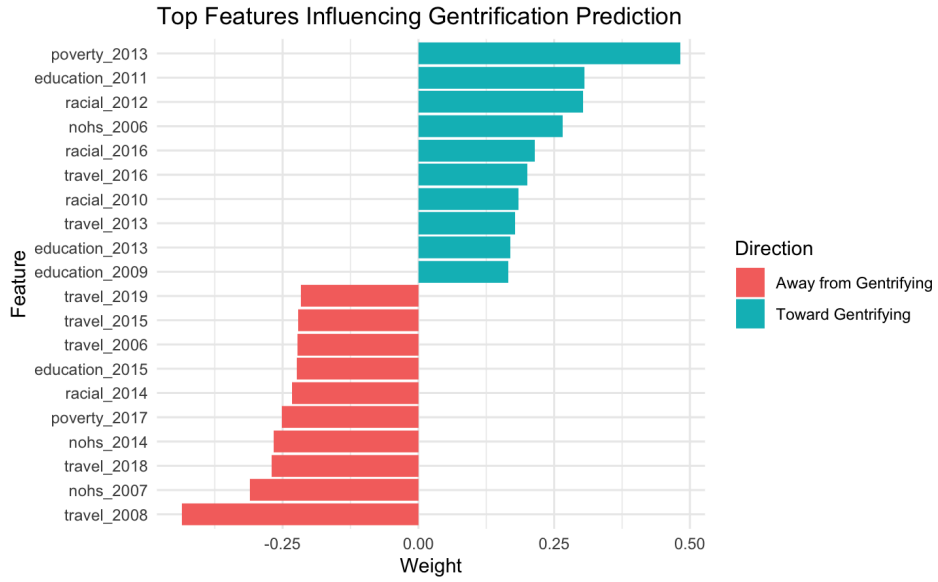
We applied the CNN to the subset of data ranging from 2005-2019, removing income-related variables as they are heavily related to the construction of the gentrification label. The gentrification label was one-hot encoded in order to have the CNN classify for 2 groups: non-gentrified (0) and gentrified (1). After experimenting with the model's learning rate at a smaller scale (25 epochs), we chose a learning rate of .0015, ran the model on the 11 neighborhood test set for 100 epochs, and achieved an accuracy of 0.9091, precision of 1, recall score of 0.5, and F1 score of .6667. The accuracy is notably higher than the other metrics. With the majority of neighborhoods in the dataset having the non-gentrifying label, any incorrect predictions of the positive gentrified label will have severe penalties on the other scores; in this case half of the true gentrified-labelled neighborhoods were incorrect and hurt this score. Still, the model has a strong accuracy and precision, potentially indicating a strong application of this neural network for this gentrification problem.

Predictions Confusion Matrix



One of the key areas of interest for our analysis was tracking which features are positive or negative contributors towards a neighborhood's gentrification-classification. By analyzing the coefficients of the 180 flattened demographic features used for classification, we can filter and analyze the strength and meaning of these coefficients. We filtered for the 10 most positive and 10 most negative feature weights and built the following plot.

Strongest Coefficients Plot



The features are flattened (all demographic features have 15 instances for each year in the dataset), so determining the exact importance of each demographic feature is slightly more difficult, but with each demographic-year as its own feature we can collect more granular insights. Poverty level in

2013 has the strongest coefficient, potentially indicating that 2013 was a turning point for gentrified/non-gentrified neighborhoods that starkly divided them in terms of poverty level, respectively. The travel measure in 2008 has the strongest negative coefficient (travel metric refers to the mean travel time to work), potentially indicating a specific cultural importance of commute time in 2008 that strongly indicated differences between gentrified and non-gentrified neighborhoods.

In terms of overall trends, we can see that education and racial demographic metrics are positive indicators for six out of the top 10 positive coefficients. The education metric refers to the percentage of fourth-graders performing at or above their grade level and the racial metric refers to the racial diversity index, or the probability that two randomly sampled residents will belong to different races. With these definitions in mind, we might conclude that improvements in elementary education, particularly in reading, and increase in racial diversity can be strong indicators that a neighborhood may become gentrified. For the negatively influencing features, travel makes up 5 out of the lowest 10 coefficients. We might conclude that higher commute times are negative contributors to the gentrification-classification.

IV. Conclusions and Future Work

In this study, we explored patterns of Gentrification in New York City analyzing median income trends over time and applying machine learning techniques to Sub-Borough Area level neighborhoods. Our approach used a CNN to identify which neighborhoods are predicted to be Gentrifying and Non-Gentrifying. Then we used ARMA time series modeling to take a deeper look into the gentrification phenomenon, in order to explain income levels shifting over time, and see what these shifts look like in a gentrifying neighborhood.

Our time series analysis showed that ARMA(1,1) models were well-suited to track income dynamics in neighborhoods undergoing change. For instance, Bushwick, a gentrifying neighborhood, had a Mean Absolute Percentage Error (MAPE) of 8.61%, while Coney Island, a non-gentrifying area, yielded a lower MAPE of 6.81%. These models captured trends without overfitting, allowing us to see gentrification trends. In Bushwick, the gentrifying neighborhood, median income spiked in the early 2010s, and is continuing to spike in 2019. This drastic increase in median income aligns with the definition of gentrification. In Coney Island, a non gentrifying neighborhood, income fluctuated, but recovered from setbacks instead of greatly surpassing previous levels.

Our implementation of the CNN proved to be a worthy solution to the neighborhood gentrification classification problem, achieving a strong accuracy of 90.91% after 100 epochs of training. After considering the parameter weights we were able to make some socially-relevant conclusions regarding specific features' impact on gentrification. Racial diversity and elementary education proved to be the most important features for a neighborhood to be given the positive gentrification classification. Additionally, high commute times are negatively associated with gentrifying neighborhoods. As referenced in Jonathan Reades's *Understanding urban gentrification through machine learning,* this coefficient analysis presents the value of quantitative analysis in urban planning and particularly in gentrification forecasting.

A limiting factor was our definition of gentrification. Gentrification is a complex, socially constructed process, which can not be defined in simple terms of median income , or rent. While median income is a factor, it is not the sole determinant of gentrification. Also, just two classes, gentrifying and non gentrifying, may not accurately represent the neighborhoods in NYC. Our binary classification system, while guided by prior research, may have oversimplified complex neighborhood dynamics. Moreover, linearity assumptions made by ARMA and independence of features assumed by CNN may not hold in reality.

Additionally, one limiting factor that may impede our confidence in model accuracy is the size of our dataset. While we feel that our set of demographic features is extensive, our number of neighborhoods was only 55, and thus the test set was rather small. The model also may be at risk of overfitting to specific trends among New York neighborhoods, and upon application to other city neighborhoods, it may

perform poorly. In the future, it would be important to collect data for many cities and many more neighborhoods.

References

Austensen, M., Yager, J., Willis, M. A., Suher, M., Stern, E., Sanders, T., Rosoff, S., Moriarty, S., Khun Jush, G., Karfunkel, B., Herrine, L., & Gould Ellen, I. (n.d.). (rep.). *State of New York City's Housing and Neighborhoods in 2015*.

Reades, J., De Souza, J., & Hubbard, P. (2018). Understanding urban gentrification through machine learning. *Urban Studies*, *56*(5), 922–942. https://doi.org/10.1177/0042098018789054

Yoo, J. (n.d.). (working paper). *Identifying Gentrification using Machine Learning*. U.S. Census Bureau.

Navarro, M. (2016, February 5). *How Gentrification Has Transformed Bushwick*. The New York Times. https://www.nytimes.com/2016/02/07/realestate/how-gentrification-has-transformed-bushwick.html

NYC Department of City Planning. (2014). *Housing New York: A Five-Borough, Ten-Year Plan.* https://www.nyc.gov/assets/hpd/downloads/pdfs/services/housing-new-york.pdf