Preliminary Report
(The Perceptrons)

***Parameters (independent variables) by Zip Code*:**

- Race in 2015, 2019 (ACS)
- Age in 2015, 2019 (ACS)
- Educational attainment in 2015, 2019 (ACS)
- Household income in 2015, 2019 (ACS)
- Home values in 2015, 2019 (zillow) - excluded for preliminary model
- Rent values in 2015, 2019 (zillow) - excluded for preliminary model

Outcome variable:

- Percentage change in median home prices between Jan 2015 to Dec 2019
- proxy for gentrification for baseline model

Summary statistics:

% Change Between 2015 and 2019

	Home Price	Male	White	Age	College Degree
count	54	54	54	54	54
mean	11.038262	0.003445	0.120296	0.033108	0.133410
std	3.585594	0.035975	0.353610	0.060753	0.145550
min	-41.298622	-0.078431	-0.247059	-0.071823	-0.176471
25%	5.893871	-0.020306	-0.025183	0.000663	0.039044
50%	11.188944	0.000000	0.005618	0.019554	0.103914
75%	17.899045	0.020408	0.116667	0.038244	0.190616
max	43.954938	0.120000	2.000000	0.316602	0.666667

We have explored the datasets from American Community Survey and Zillow Housing Research for the years of 2015 and 2019 on the following parameters:

- Race (% White)
- Sex (% Male)
- Median Age
- Education (% High school graduate or higher)
- Education (% Bachelor's degree or higher)
- Median Home value
- Median Rent value - excluded for preliminary model

The main aim of our baseline regression model was to explore the correlation between our feature variables and neighborhood change. In our baseline model, we calculated the percentage changes in the variables on race, age, education, household income between the years 2015 and 2019 across zip codes in Chicago and these were used as features (x) to classify Chicago zip codes by the type of neighborhood change ("Significantly Declining", "Declining", "Stable", "Growing", "Significantly Growing"). For our predicted label (y), we used the percentage change in median home prices between Jan 2015 to Dec 2019 as a proxy for neighborhood change across zip codes in Chicago.

We coded the multi-class "neighborhood change" variable as follows:

- "Significantly Declining": At least 50% below Chicago median
- "Declining": At least 25% above Chicago median
- "Stable": Within 25% above or below Chicago median
- "Growing": At least 25% above Chicago median
- "Significantly Growing": At least 25% below Chicago median

Testing and Training:

We split our data into training and testing data with a 75:25 split, so that there are 40 zip codes in the training data and 14 zip codes in the testing data. Our initial model regressed change in home price in a zip code on change in male population, change in population age, change in the population that was white, and change in the population that had a college education. We obtained an initial set of weights:

$\beta_0 = 10.03$
$\beta_{male} = 13.82$
$\beta_{white} = -3.937$
$\beta_{age} = 1.56$
$\beta_{college} = -29.707$

We used the weights to predict home values for the training and test data and used mean squared error as our loss function. We obtained the following MSE for training and testing:

$MSE_{traing} = 102.03$
$MSE_{test} = 510.45$

Future Plans

Regarding adjustments, we will most likely shift the bounds of our categorical output variable, given that none of the neighborhoods in the sample grew or declined by more than 44%, which means that none could be classified as “Significantly Declining” or “Significantly Growing”.

Regarding updates, we will most likely be bringing in new feature variables to improve the predictive power of our model, which is currently underperforming. This should include exploring two other housing-related datasets on building permits and AirBnb reviews in Chicago to understand which model/feature set is more predictive of neighborhood change. This may involve parsing and potentially working with unstructured data, which would be a larger undertaking. In addition to this, we may want to add a second demographic component to our prediction variable so that it’s a better proxy for gentrification, since changing demographics are an important component of gentrification. We will need to look into how to do this, which may include engineering a new feature by combining demographic change with home value change.

Regarding the model for the main approach, we are tentatively planning on using a multi-class logistic regression and sentiment analysis for the unstructured Airbnb data.