Unsupervised Learning Clusters: Gainesville Housing

Scott Fischer, Feb 2020

Generating New Insights

Unsupervised clustering leverages machine learning's ability to discover patterns and insights not readily supplied or inherently obvious. When applied to real estate, the impact expands beyond just the individual and creates a mosaic of an entire region.

Marketplace

What are the archetypes being sold / developed?

City Trends

With shifting population comes new infrastructure from both the city and commercial developers.

Socioeconomic

How are the area residents fairing with transaction trends?

The Data Used

Location: *Gainesville, FL*

Data Origin: *Local MLS*

Study Focus: *Residential Sales*

Time Frame: *2017 - 2019*

Data Cleaning

- Removing skewed data
- Correcting input errors
- Aggregating disparate values

Modeling the Data

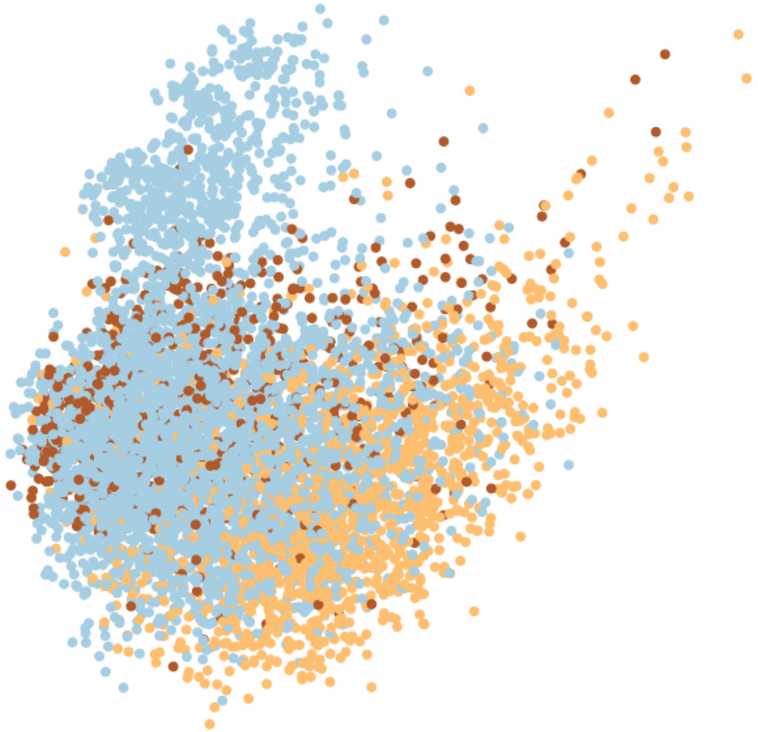
Distance Metric

Gower

Since there was a heavy amount of mixed categorical and continuous data, I decided to use the Gower method of calculating distances.

DBSCAN

Hierarchical



Gaussian Mixture

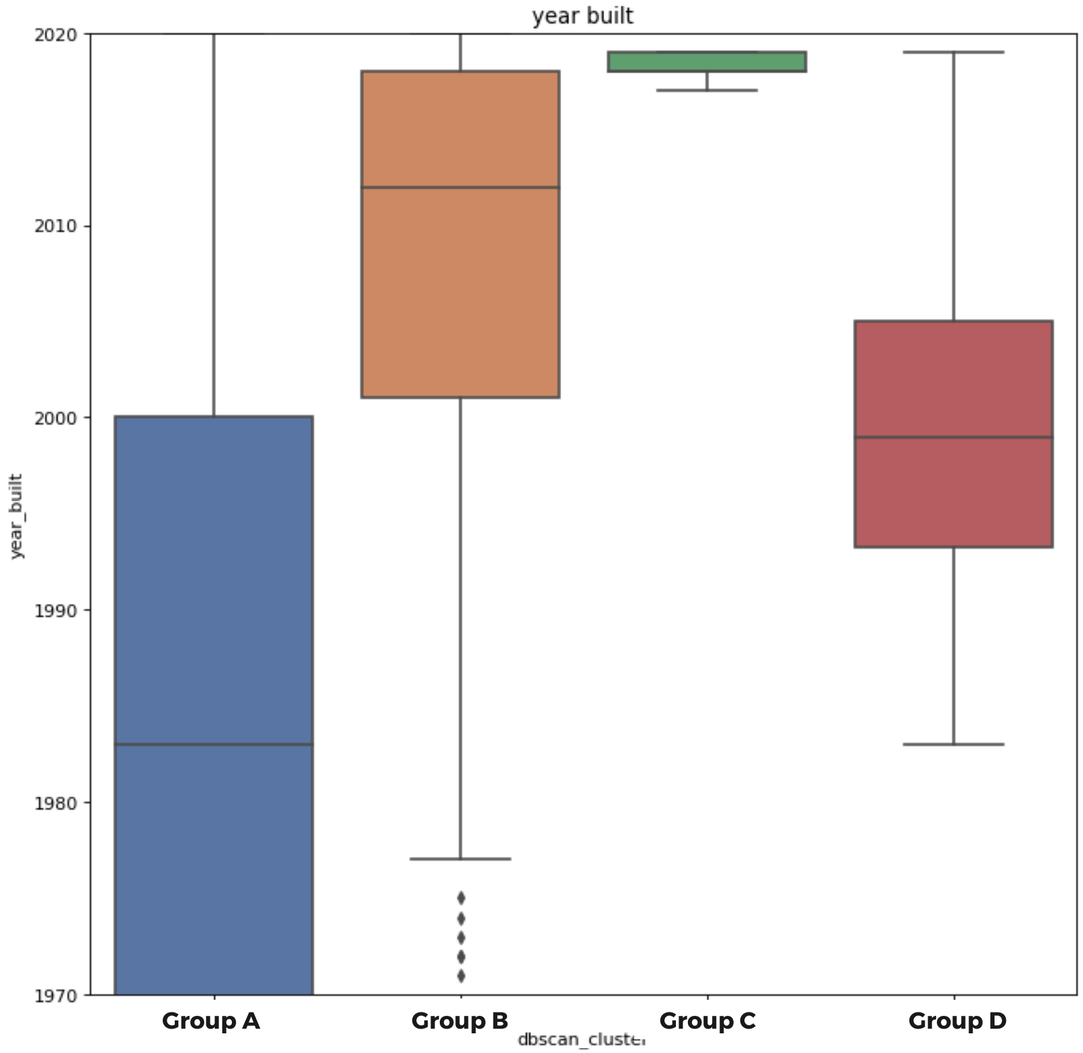
Best Performing Model?

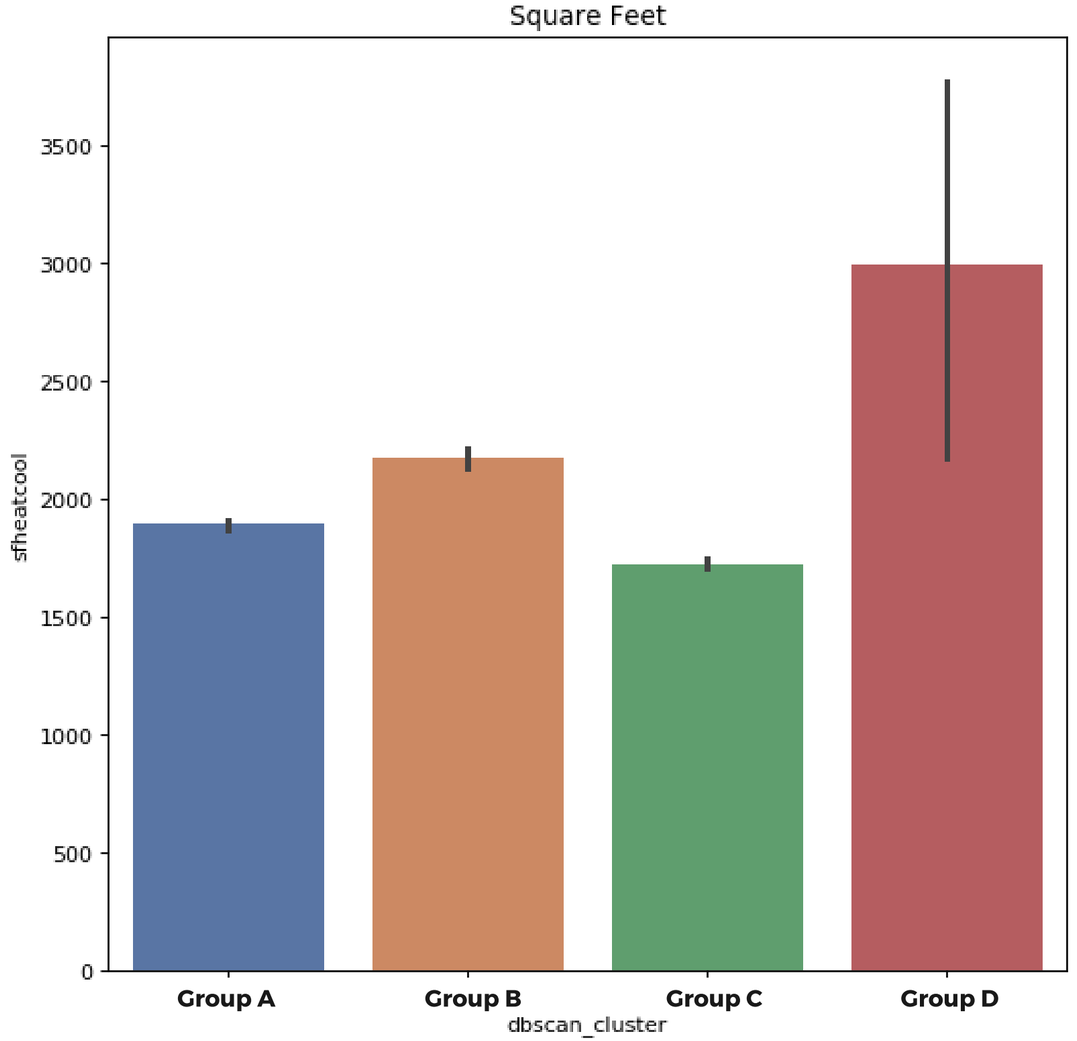
DBSCAN had the most reliable and interpretable results of the three I used.

The Cluster Groups: An Overview

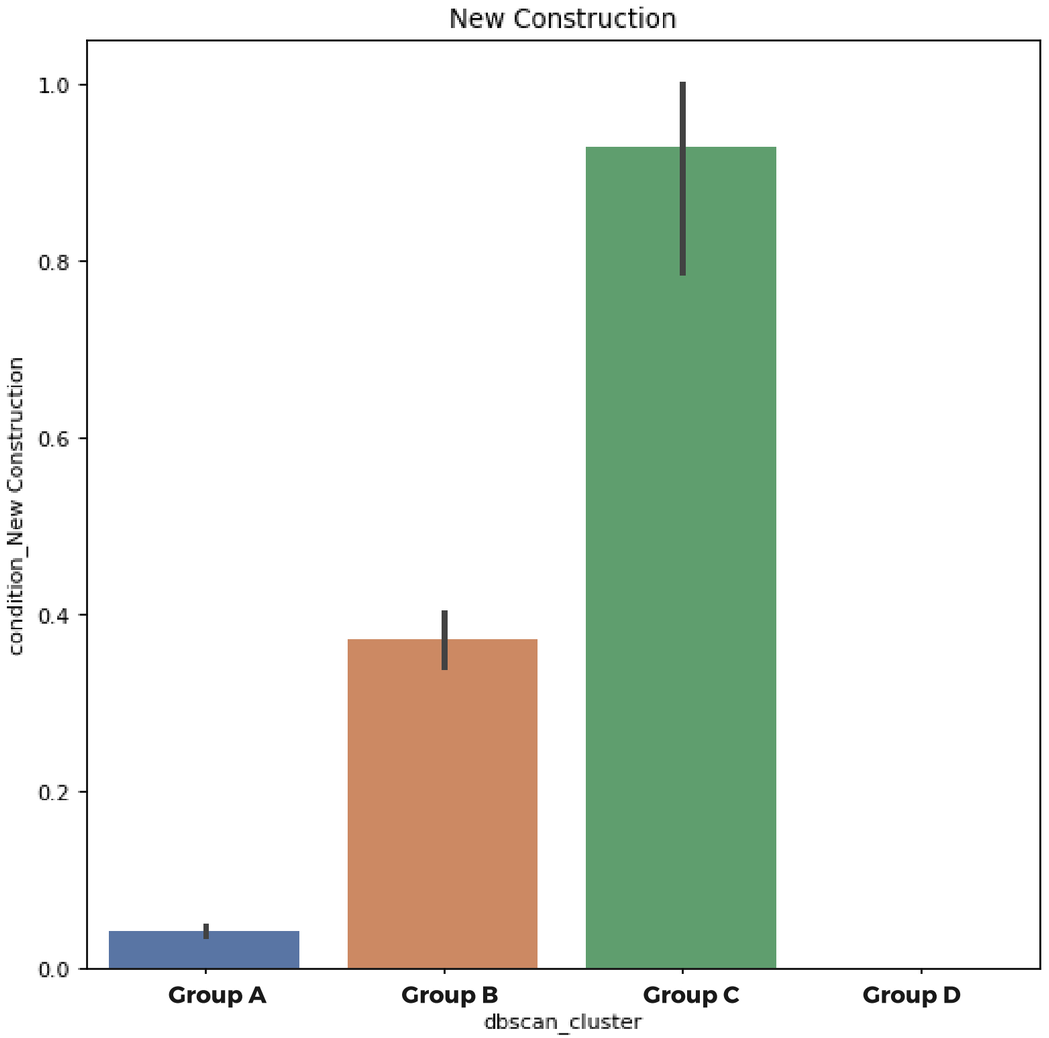
Key Averages	Group A	Group B	Group C	Group D
Days on Market	94 days	140 days	202 days	110 days
Price per Sqft	$128	$155	$170	$110
Sqft heated/cooled	1892 sqft	2170 sqft	1723 sqft	2993 sqft
Acres	.56 acres	.28 acres	.12 acres	.70 acres
Year Built	1983	2008	2018	1999
Amount of homes in cluster	5288	1052	14	8

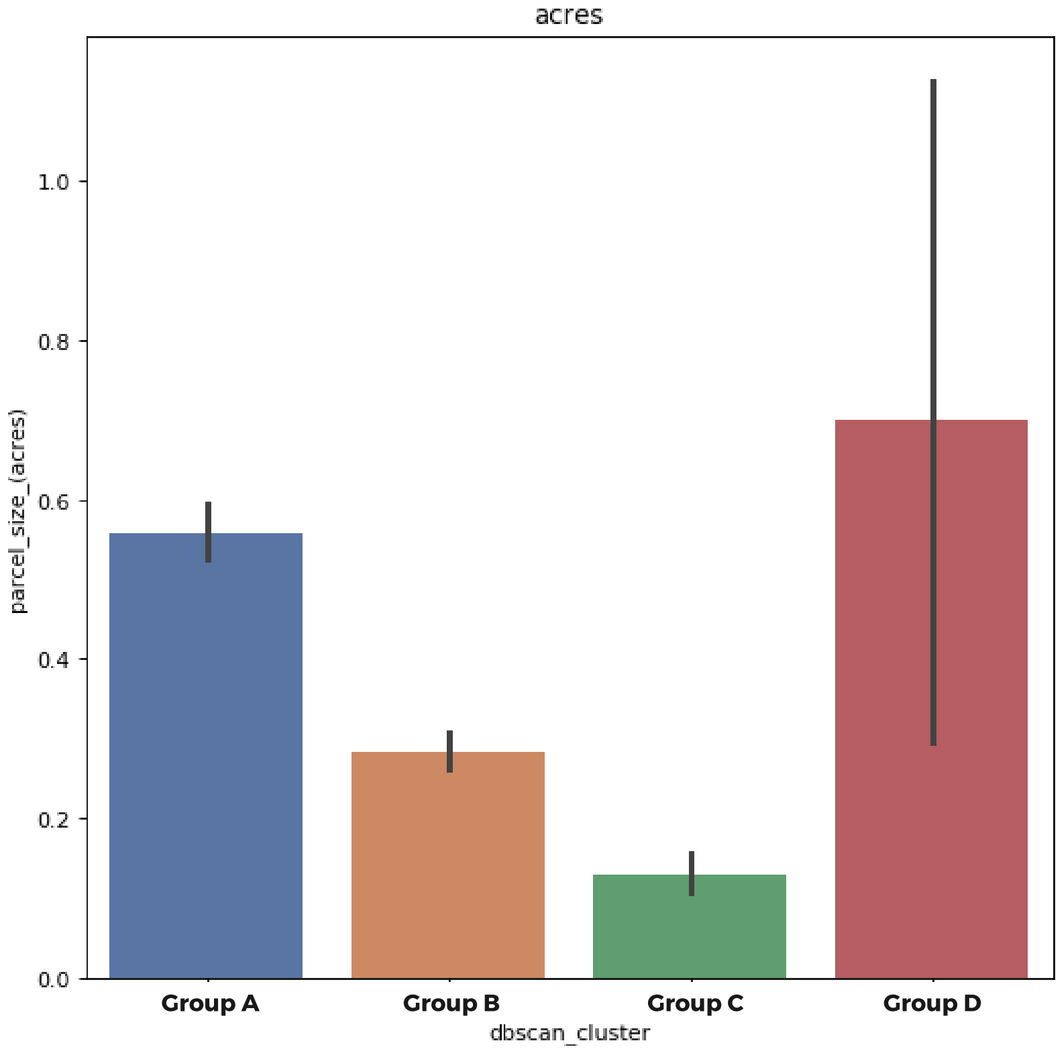
The Cluster Groups: Comparisons





The Cluster Groups: Comparisons, Cont.





Identifying The Clusters

There are some fairly clear delineations between the different clusters, and I've decided on some archetypes to apply.

Group A
"The Average"

Group B
"The Gentrifier"

Group C
"The Newest"

Group D
"The Luxury"

"The Average"

The vast majority of houses in Gainesville fall within this category.

On average they:

- Are the oldest houses
- Have the lowest cost per sqft
- Fastest to sell
- Second biggest lot size

$242,176
Avg Sales Price

"The Gentrifier"

These houses are typically newer but with some older houses in the subset. This is reflective of the recent trend of either ground up developing or renovating in traditionally underprivileged neighborhoods in the city center.

On average they have:

- Second highest cost per sqft
- Second longest time on market
- 40% are new constructions
- Given average sqft, second most expensive

$336,350
Avg Sales Price

"The Newest"

The vast majority of these homes were purchased as a new construction. They perfectly share a number of features, which likely means they are all part of a development or from the same builder.

On average they have:

- Highest cost per sqft
- 93% are new constructions
- ALL have slab frame foundations
- ALL have concrete siding

$301,410
Avg Sales Price

"The Luxury"

Not brand new, but the most expensive property cluster on the market. These "McMansions" usually have more yard space and a lot of interior space.

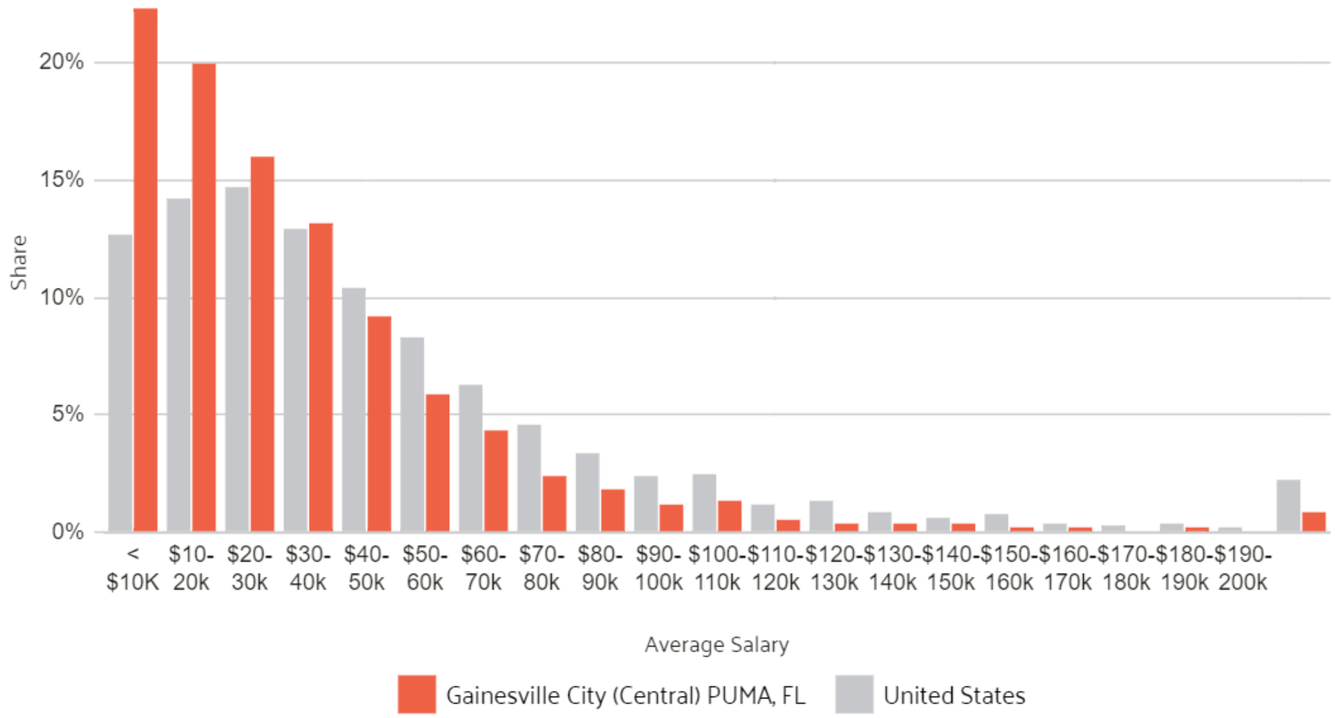
On average they have:

- Second lowest cost per sqft
- Highest amount of acreage
- Largest sqft
- ALL are classified as having multiple layout styles

$430,992
Avg Sales Price

Contextual Economic Overview

Wage Distribution



$1,006
Median Gross Rent

58.3%
Housing units Renter occupied

$36,047
Median Earnings

$159,800
Median Home Value

Groups

Group A
"The Average"

5288

$242,176
Avg Sales Price

Group B
"The Gentrifier"

1052

$336,350
Avg Sales Price

Group C
"The Newest"

14

$301,410
Avg Sales Price

Group D
"The Luxury"

8

$430,992
Avg Sales Price

Concluding Insights

There is a disturbing trend for sales from the past two years not matching the market's inherent need for housing. It is important to note that Gainesville is a University town, with 1/4 of it's population as students. However, demographic studies show that minorities disproportionately make up the lowest income brackets. It's also important to note that most of the census data is reflective of local residents which most students don't identify as.

Marketplace

Houses are selling for values that are above what their tax assessed value. With most housing occupied by renters, this means more people are investing in housing as a passive income source. I expect rents to increase with increasing housing costs.

City Trends

The market is getting more expensive and wages aren't improving. Most houses are bought and rented out with minimal improvements. This reflects the affordable housing crisis rampant. Gainesville has a low cost of living, but also an incredibly low average wage.

Socioeconomic

Gentrification has become a trend in town, especially in inner city areas as downtown develops and the land becomes more valuable. Historically black neighborhoods, such as Porters, have undergone drastic changes and developments after the city invested in a $35 million park under a mile away.

Improving the Study

Due to time constraints, I had to limit the number of clusters I could study from DBSCAN. Some versions of my clustering produced over 20 different groups and I would have liked to explore further segmentation of the average group and to hopefully lesson the data imbalance.

Other quality of life improvements:

- Incorporating latitude / longitude information and graphing clusters
- Examining a historic housing price graph - how does this compare to 2007/2008?
- Include demographic information and visual general distribution in relation to graphed clusters
- Gather historic sales data on each of the parcels and include a % profit based on how much the seller spent.