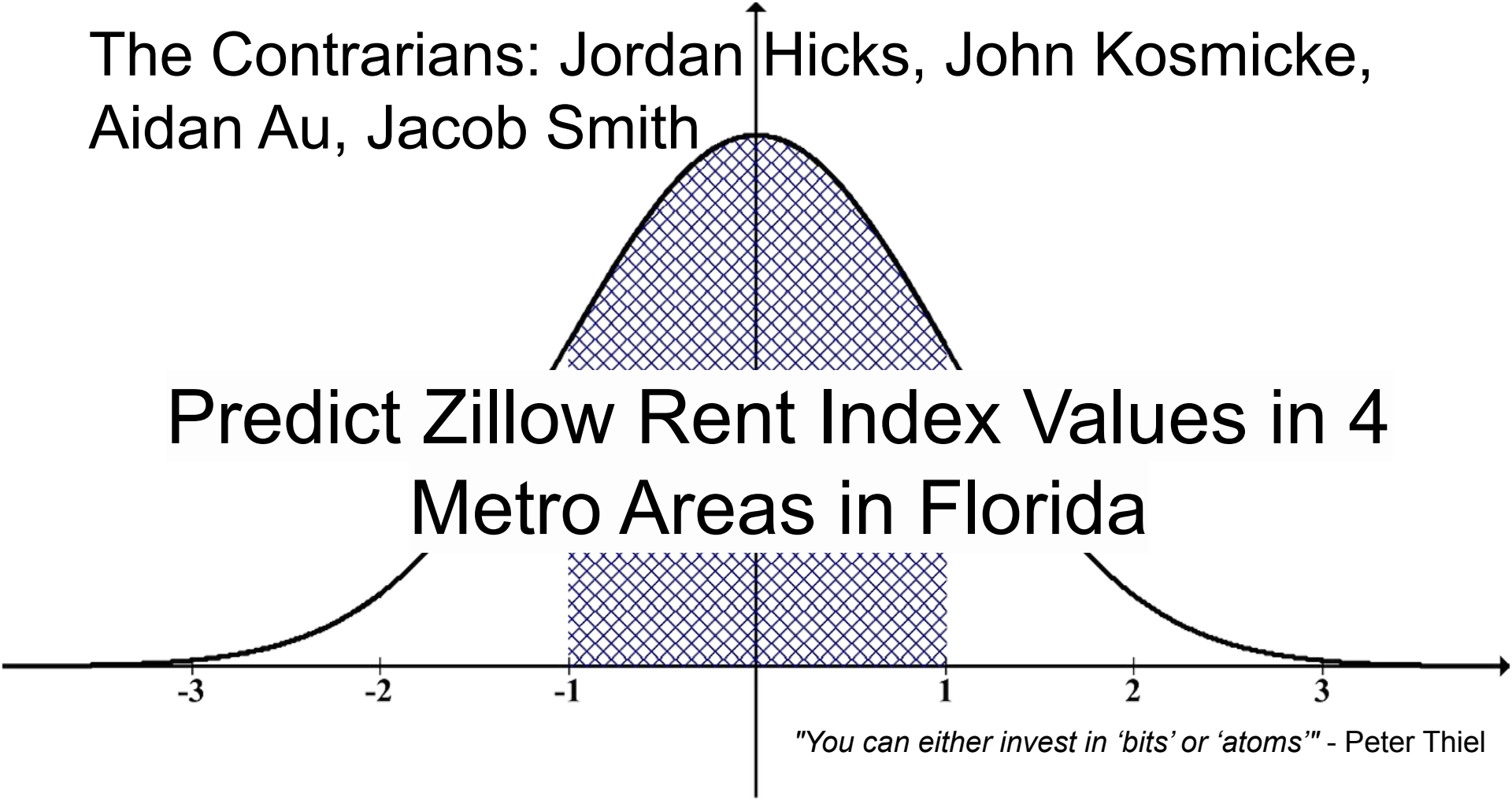
The Contrarians: Jordan Hicks, John Kosmicke, Aidan Au, Jacob Smith

# Predict Zillow Rent Index Values in 4 Metro Areas in Florida




*"You can either invest in 'bits' or 'atoms'"* - Peter Thiel

# Agenda

- **Research Questions and Motivation**
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

# Research Questions/Motivations

- What models can we use to to predict the average monthly rent prices for all metro areas in Florida with a lower average error?

- What features are important in predicting the average monthly Zillow Rent Index in each of the 4 metro areas?

# Agenda

- Research Questions and Motivation
- **Data Used**
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

# Data Used

- (**ACS**) American Community Survey (5 year data)
- **FRED** State Total Housing Price Index
- **BLS** Florida Unemployment

# Agenda

- Research Questions and Motivation
- Data Used
- **Why Florida**
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

# Why Florida, the Sunshine State?

- Many people have their second homes in Florida.
  Great for **investment properties** as well as primary residence.
- **Steadily rising rents** year over year
- **Population** have been **growing** even BEFORE Covid, **influx of new residents** from bigger cities and states
- The Mayor of Miami welcomes Silicon Valley and tech firms to move in, which can bring **higher wages**
- Yearly-round warm weather, popular travel destinations
- No personal state income tax (This is not legal or tax advice)
- Some say, “if you marry California and Texas, you’d get Florida”

# 'S QUARTERLY JOBS AND INCOME REPORT SHOWS NATIONAL GAINS WITH ORLANDO LEADING THE WAY

December 10, 2021

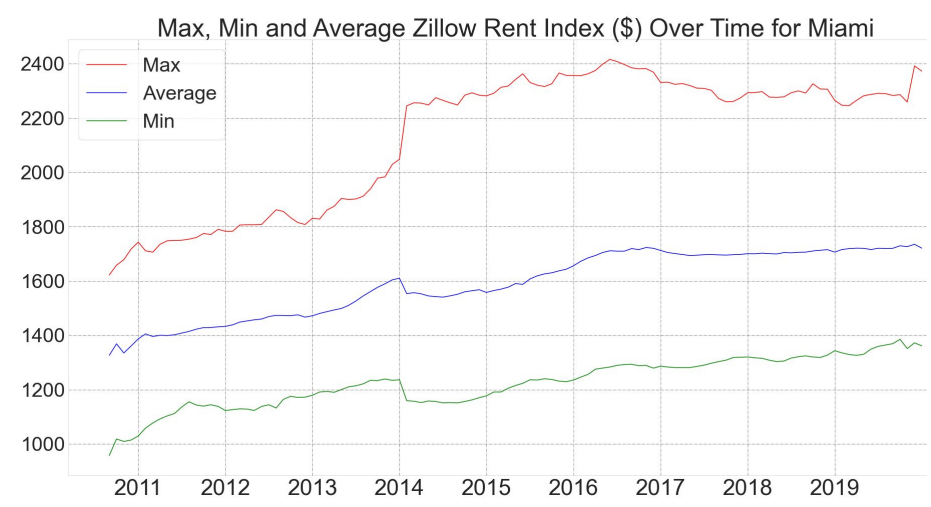
# THESE SOUTHEAST FLORIDA NEIGHBORHOODS ARE THRIVING THANKS TO A GROWING BASE OF HIGH-INCOME RESIDENTS

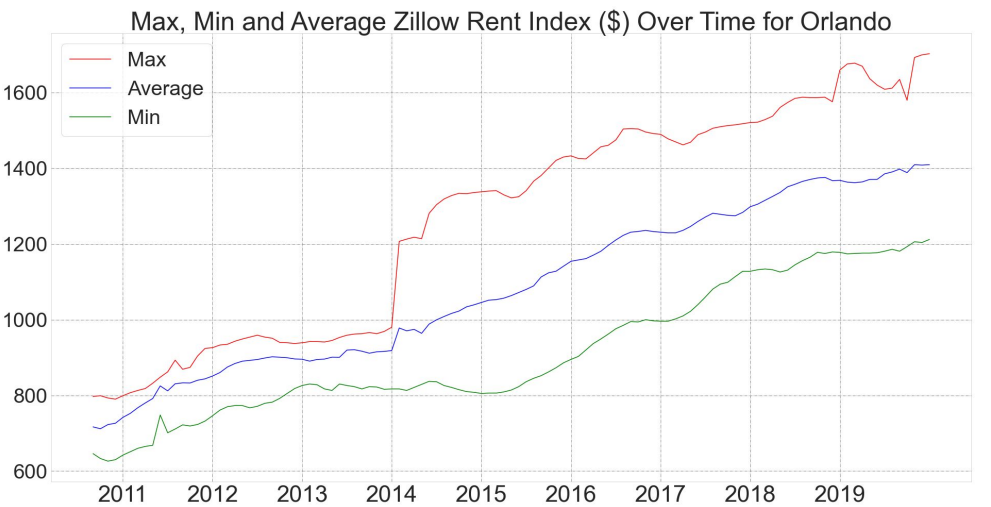
October 18, 2021

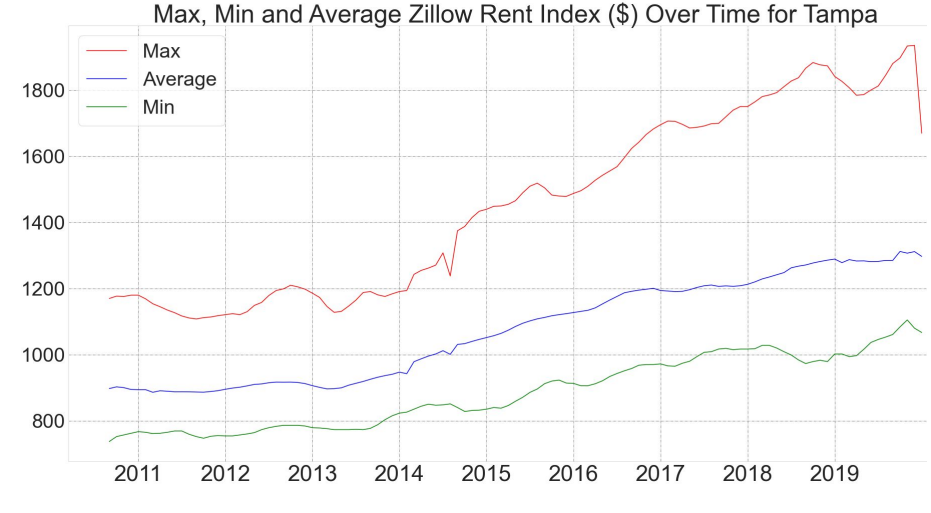
# Agenda

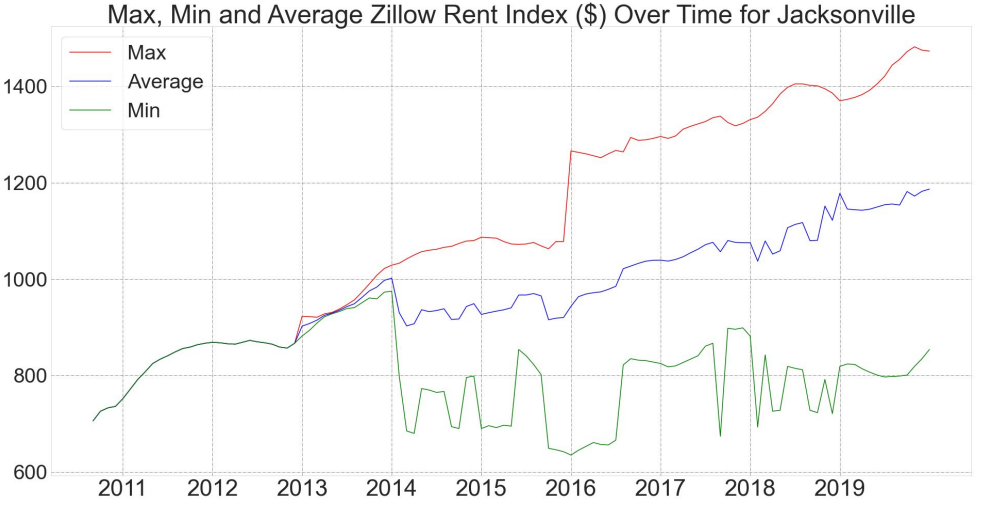
- Research Questions and Motivation
- Data Used
- Why Florida
- **Data Visualization**
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

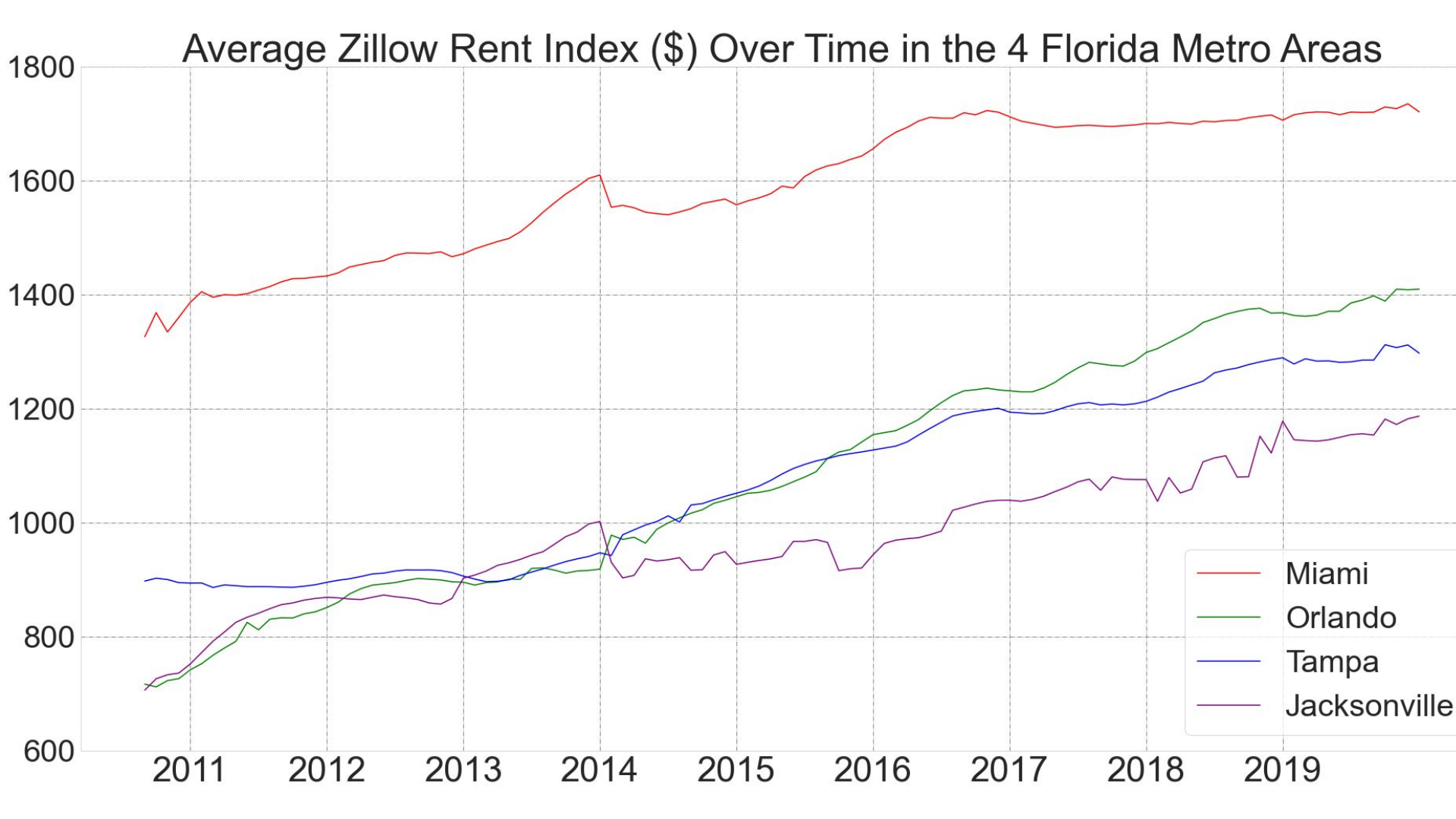
Visualization:


Max, Min and Average Zillow Rent Index ($) Over Time for Miami
Max
Average
Min
2400
2200
2000
1800
1600
1400
1200
1000
2011
2012
2013
2014
2015
2016
2017
2018
2019


Max, Min and Average Zillow Rent Index ($) Over Time for Orlando
Max
Average
Min
1600
1400
1200
1000
800
600
2011
2012
2013
2014
2015
2016
2017
2018
2019


Max, Min and Average Zillow Rent Index ($) Over Time for Tampa
Max
Average
Min
1800
1600
1400
1200
1000
800
2011
2012
2013
2014
2015
2016
2017
2018
2019


Max, Min and Average Zillow Rent Index ($) Over Time for Jacksonville
Max
Average
Min
1400
1200
1000
800
600
2011
2012
2013
2014
2015
2016
2017
2018
2019


Average Zillow Rent Index ($) Over Time in the 4 Florida Metro Areas
1800
1600
1400
1200
1000
800
600
2011
2012
2013
2014
2015
2016
2017
2018
2019
Miami
Orlando
Tampa
Jacksonville

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- **Data Cleaning**
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

# How We Cleaned the Data

- Removed ACS features with Nan values
  - ACS features with Nan values were 50%+ Nan

# Feature Transformation

- Standard Scaler (important for linear models)

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- **Clustering**
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

# Clustering

Clustered based off economic conditions in the area.

Variables Used:

- Local unemployment population
- Local poverty rates
- Local median income
- Local median rent
- Local million-dollar housing units
- Local average percent income spent on rent

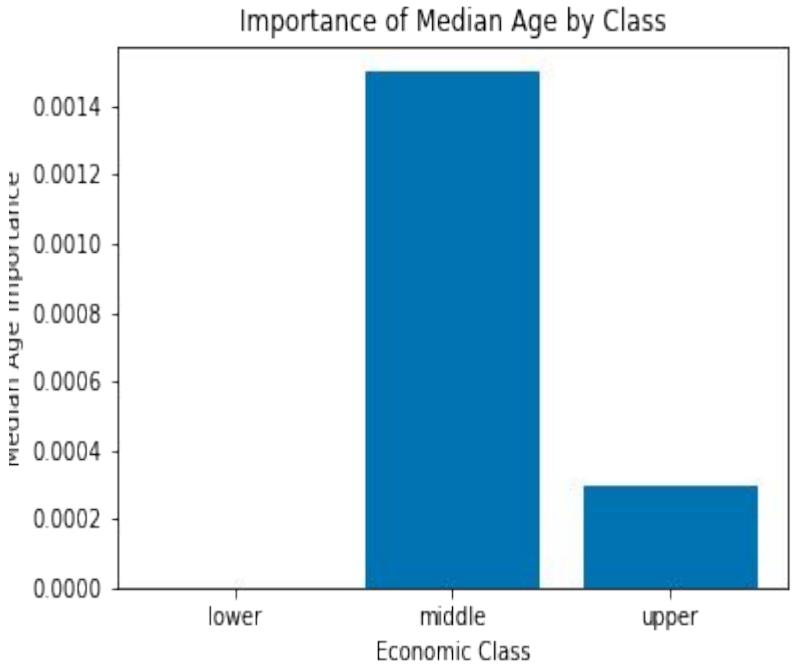
# Clustering: Feature Importances

- Some features are more important for certain economic conditions

- Clustering can help identify those

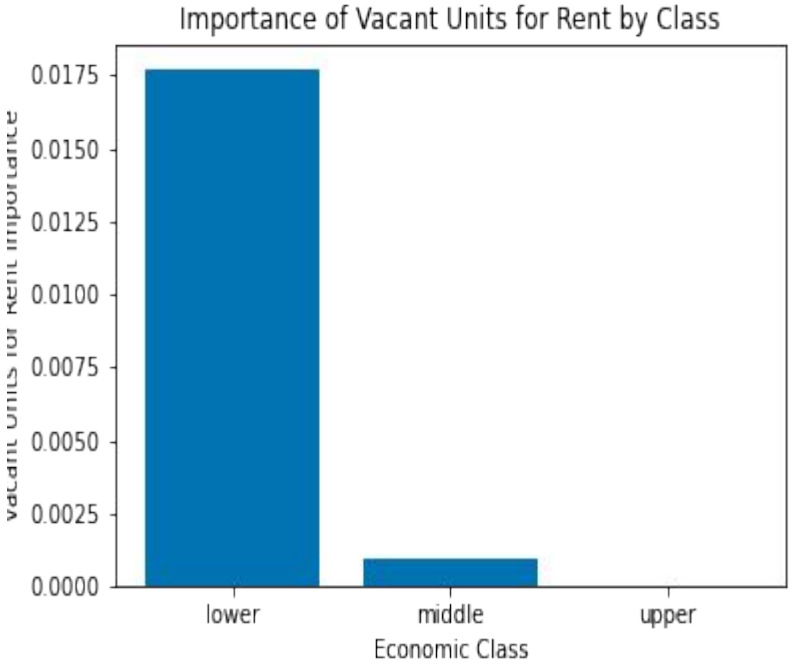
# Feature Importance: Median Age

- Some importance for middle class
- Low importance otherwise



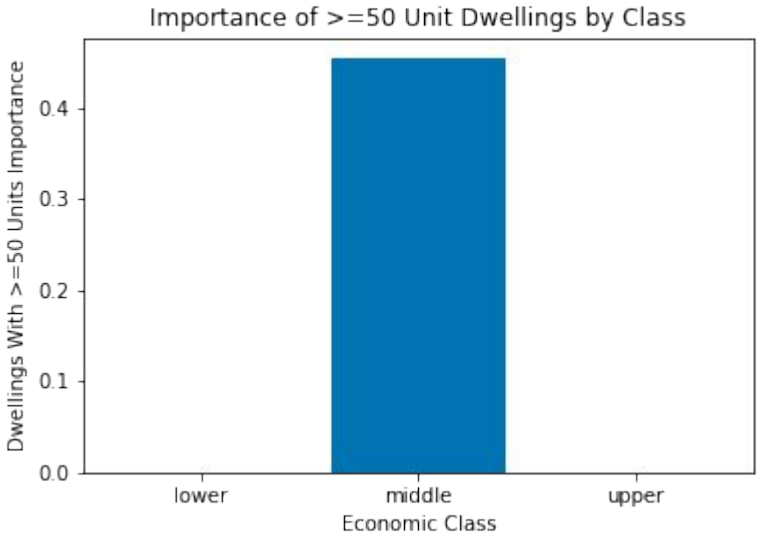
# Feature Importance: Vacant Units for Rent

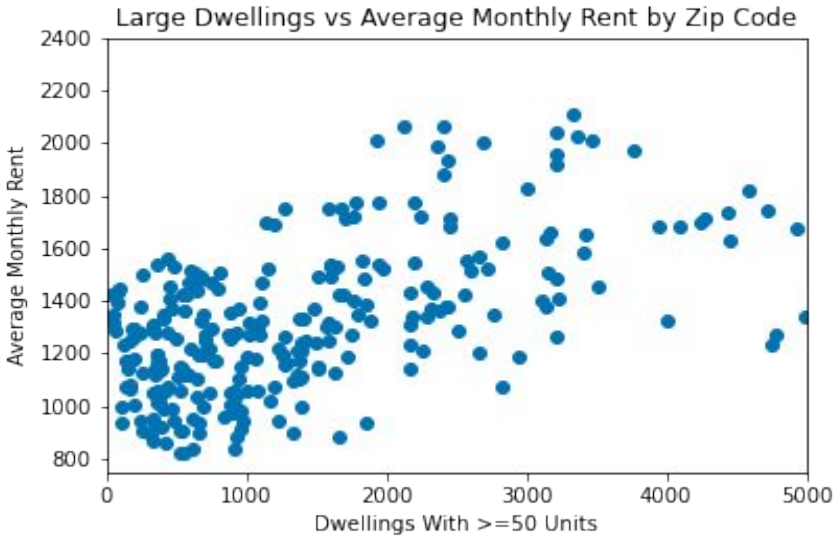
- Some importance for lower class
- Low importance otherwise



# Feature Importance: Large Dwellings

- A large dwelling here is defined as a building with at least 50 housing units.
- Whether all of these are occupied is not considered



# Clustering Algorithms


# K Means vs Agglomerative

K Means

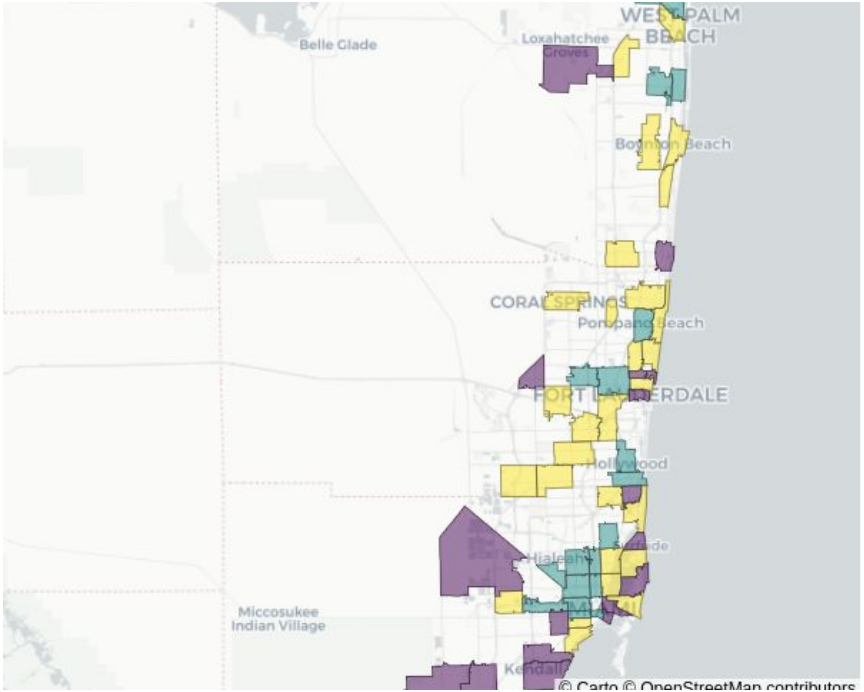
- Centroidal
- Minimizes Distance Between Clusters
- Initialized K Means++

Agglomerative

- Hierarchical
- Minimizes Variance Among Clusters (Ward's Method)

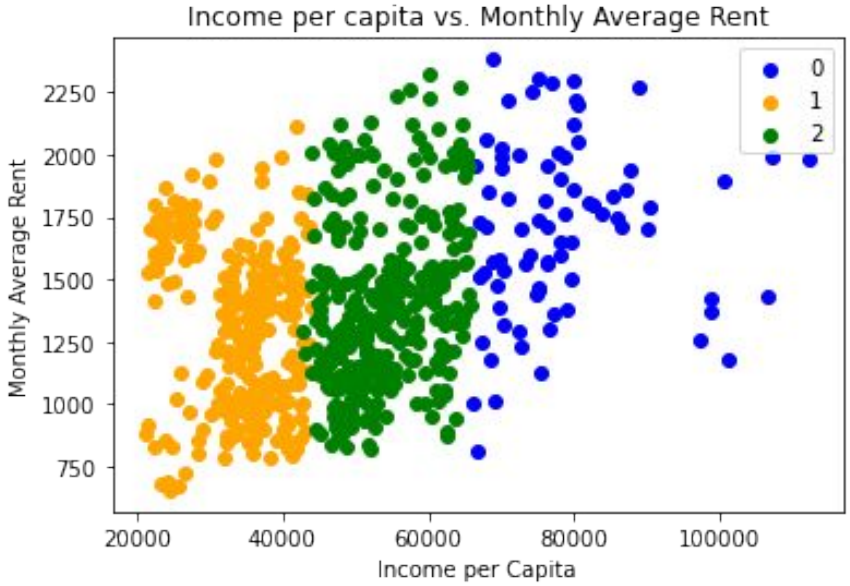
- "Dependent Variable" is the cluster mean of median YOY percentage rent increase
- All clusters between 2% and 5%
- The scatterplots represent a projection of each cluster onto average rent and annual income

# Subset Clustering (prior variables - K Means)

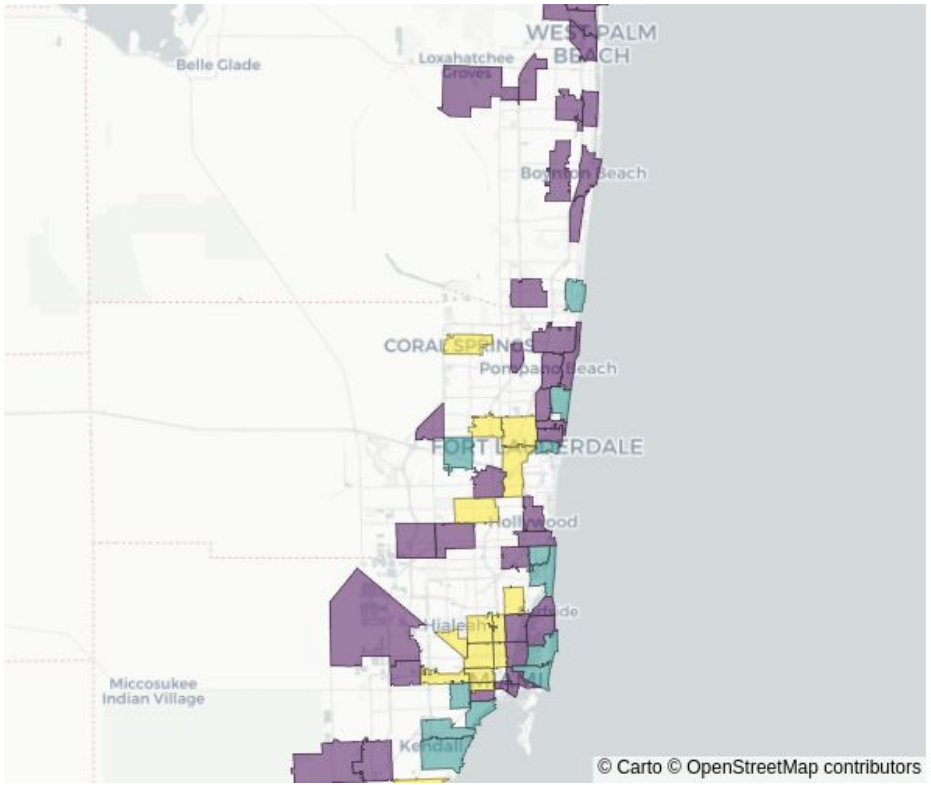




| Median YOY Rent Change By Cluster, KMeans | |
|---|---|
| **0** | 0.036005 |
| **1** | 0.040510 |
| **2** | 0.041471 |



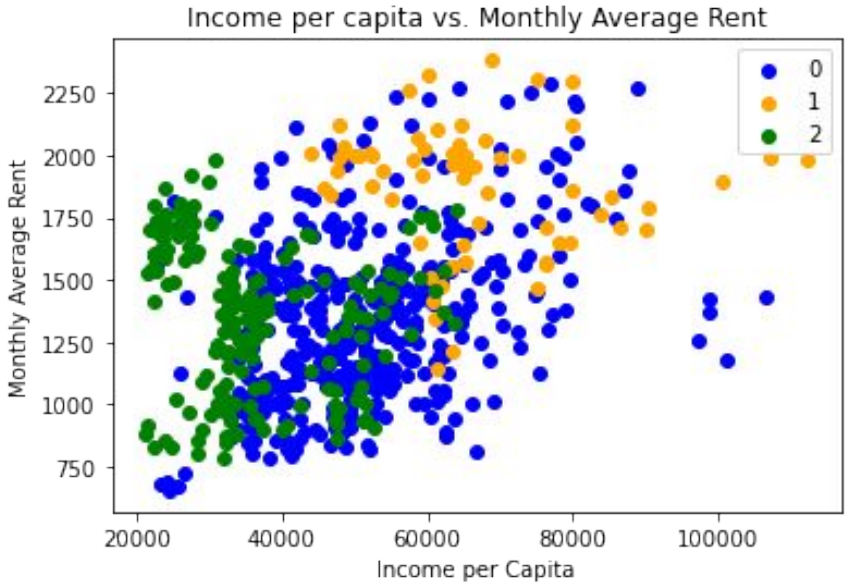
# Subset Clustering (prior variables - Agglomerative)



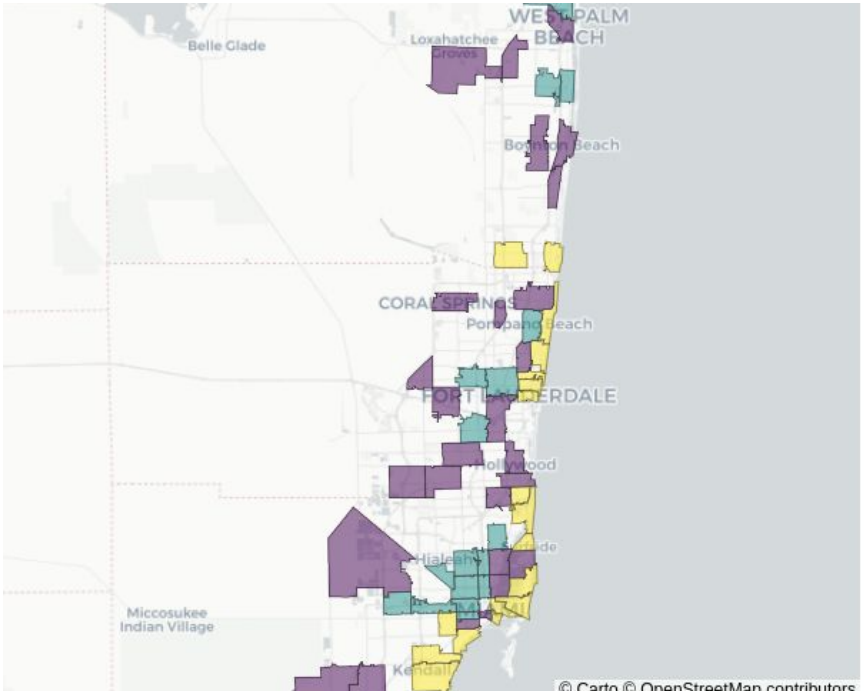


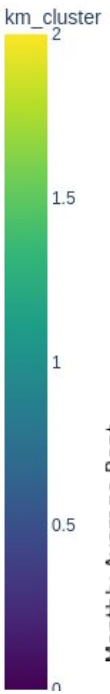
Median YOY Rent Change By Cluster, Agglomerative

| | |
|---|---|
| **0** | 0.043422 |
| **1** | 0.028761 |
| **2** | 0.038067 |

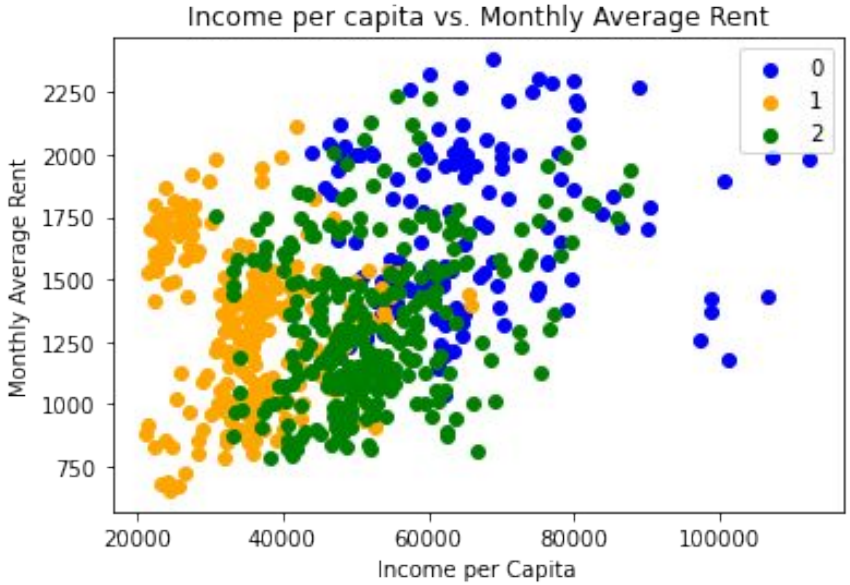


# Subset Clustering (new variables, K Means)



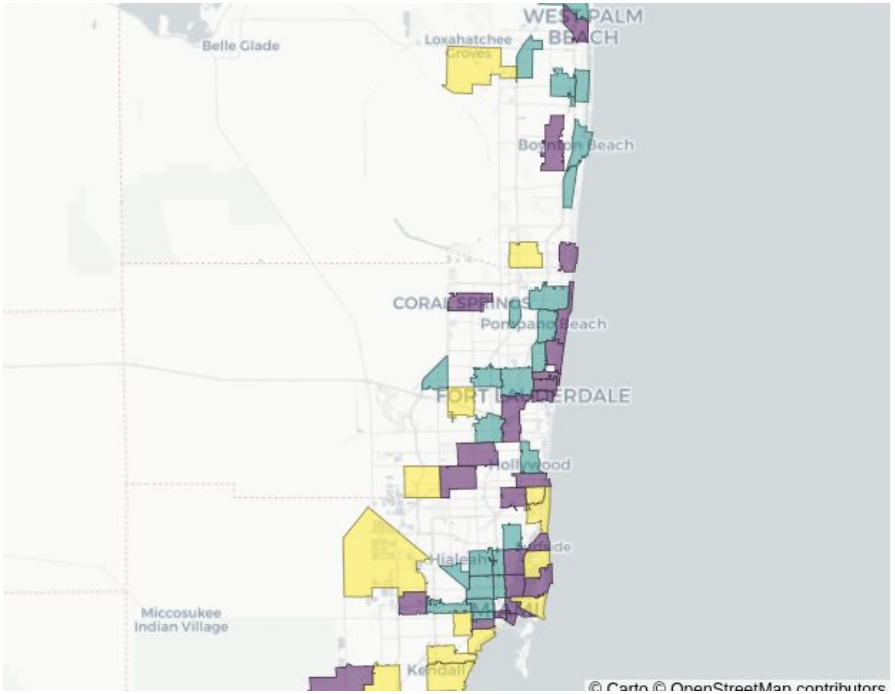




| Median YOY Rent Change By Cluster, KMeans | |
|---|---|
| **0** | 0.036438 |
| **1** | 0.040498 |
| **2** | 0.042260 |




- Gini_index
- Graduate_professional_degree
- income_per_capita

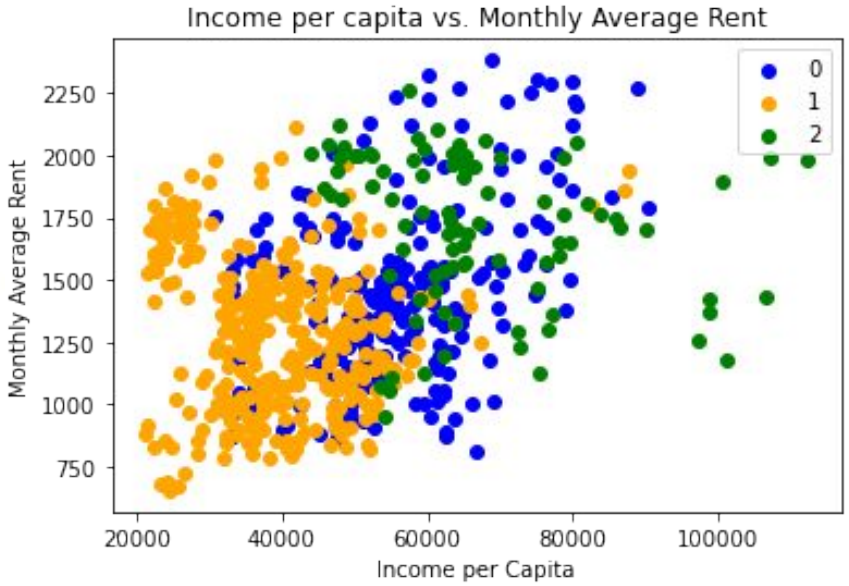
# Subset Clustering (new variables - Agglomerative)





Median YOY Rent Change By Cluster, Agglomerative

| | |
|---|---|
| **0** | 0.043410 |
| **1** | 0.042049 |
| **2** | 0.027068 |




- Gini_index
- Graduate_professional_degree
- income_per_capita

## Subset 3

- not_us_citizen_pop
- median_age
- Median_year_structure built

| | Median YOY Rent Change By Cluster, KMeans |
|---|---|
| 0 | 0.035157 |
| 1 | 0.044831 |
| 2 | 0.028834 |

| | Median YOY Rent Change By Cluster, Agglomerative |
|---|---|
| 0 | 0.043953 |
| 1 | 0.032485 |
| 2 | 0.030442 |

## Subset 4

- total_pop
- unemployed_pop
- high_school_diploma

| | Median YOY Rent Change By Cluster, KMeans |
|---|---|
| 0 | 0.037521 |
| 1 | 0.041084 |
| 2 | 0.044045 |

| | Median YOY Rent Change By Cluster, Agglomerative |
|---|---|
| 0 | 0.040125 |
| 1 | 0.041266 |
| 2 | 0.039659 |

## Subset 5

- mobile_homes
- masters_degree
- no_car

| | Median YOY Rent Change By Cluster, KMeans |
|---|---|
| 0 | 0.034445 |
| 1 | 0.042031 |
| 2 | 0.047845 |

| | Median YOY Rent Change By Cluster, Agglomerative |
|---|---|
| 0 | 0.047845 |
| 1 | 0.033962 |
| 2 | 0.041427 |

## Subset 6

- male_45_64_bachelors_degree
- male_45_64_graduate_degree
- male_45_64_high_school
- male_45_64_some_college

| | Median YOY Rent Change By Cluster, KMeans |
|---|---|
| 0 | 0.038863 |
| 1 | 0.042877 |
| 2 | 0.037594 |

| | Median YOY Rent Change By Cluster, Agglomerative |
|---|---|
| 0 | 0.044496 |
| 1 | 0.041617 |
| 2 | 0.027575 |

## Subset 7

- FLSTHPI_Yearly_Avg
- children
- employed_pop

Median YOY Rent Change By Cluster, KMeans

| | |
|---|---|
| 0 | 0.038370 |
| 1 | 0.041989 |
| 2 | 0.041271 |

Median YOY Rent Change By Cluster, Agglomerative

| | |
|---|---|
| 0 | 0.040720 |
| 1 | 0.038202 |
| 2 | 0.045297 |

## Subset 8

- FLSTHPI_Yearly_Avg
- female_25_to_29
- female_30_to_34
- female_35_to_39

Median YOY Rent Change By Cluster, KMeans

| | |
|---|---|
| 0 | 0.041785 |
| 1 | 0.034797 |
| 2 | 0.046023 |

Median YOY Rent Change By Cluster, Agglomerative

| | |
|---|---|
| 0 | 0.039921 |
| 1 | 0.039125 |
| 2 | 0.046082 |

## Subset 9

- FLSTHPI_Yearly_Avg
- income_50000_59999
- income_60000_74999
- income_75000_99999

Median YOY Rent Change By Cluster, KMeans

| | |
|---|---|
| 0 | 0.040067 |
| 1 | 0.038365 |
| 2 | 0.042988 |

Median YOY Rent Change By Cluster, Agglomerative

| | |
|---|---|
| 0 | 0.040248 |
| 1 | 0.037804 |
| 2 | 0.044888 |

## Subset 10

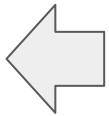
- married_households
- children
- pop_25_64
- vacant_housing_units

Median YOY Rent Change By Cluster, KMeans

| | |
|---|---|
| 0 | 0.037689 |
| 1 | 0.040967 |
| 2 | 0.043154 |

Median YOY Rent Change By Cluster, Agglomerative

| | |
|---|---|
| 0 | 0.041652 |
| 1 | 0.040970 |
| 2 | 0.025457 |



# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- **Feature Selection**
- Target Transformation
- Modeling
- Results
- Conclusion
- Future Work

# Feature Selection

- To reduce multicollinearity
- Stepwise Selection (Forward Selection)
- Removed racial features to minimize potentially discriminatory biases
- Leaving these in could potentially make housing prices be affected by racial demographics
- Removed features that can cause data leakage
  (e.g. median rent, percentage of income spent on rent)

- Removed duplicated columns

- No *historical* rent prices were used as a feature

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection
- **Target Transformation**
- Modeling
- Results
- Conclusion
- Future Work

# Target Transformation

- Average monthly Zillow Rent Index for the corresponding year

(All observations are at the zip-code level; 145 zip codes from all 4 metro areas - Miami: 75, Orlando: 27, Tampa: 32, Jacksonville: 11)

| Zip Code | Month | ZRI | Average Monthly ZRI in 2018 |
|---|---|---|---|
| Zip Code 1 | 2018-Jan | 1432 | 1532.41 |
| Zip Code 1 | 2018-Feb | 1540 | 1532.41 |
| Zip Code 1 | 2018-Mar | 1630 | 1532.41 |
| ⋮ | | | |
| Zip Code 1 | 2018-Nov | 1574 | 1532.41 |
| Zip Code 1 | 2018-Dec | 1645 | 1532.41 |

# Target Transformation

- New ACS data released annually; for simplicity to predict an annualized target
- Same amount of monthly rent in a 12-month lease
- Pros: less noise, months with a higher index are averaged out by those with a lower index
- For each zip code, Instead of having 12 rows of monthly data in a year, there’s only 1 row of annualized data
- A simpler model with a little bias learn less from the noise, avoid overfitting
- Cons: less granular, no “true” monthly index prediction

# Target, Training and Test Sets

- Target: Average Monthly Zillow Index in the respective year
- Train Set: American Community Survey (ACS) data from 2013 to 2017
- Test Set: American Community Survey (ACS) data from 2018

Trained the model with the ACS data from 2013-2017.

Tested the model with the ACS data from 2018 to predict the Average Monthly Zillow Index in 2018

Note: **No** ACS data from 2018 were used when training the models

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- **Modeling**
- Results
- Conclusion
- Future Work

# Models Used

- Model 1: Multiple Linear Regression
- Model 2: Lasso Regression
    - Hypertuned with Grid Search, 3-fold Cross Validation

- Model 3: Random Forest
  (For getting Feature Importances only but not prediction)

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- **Results**
- Conclusion
- Future Work

# Results

| Miami | Multiple Linear Regression | Lasso Regression |
|---|---|---|
| RMSE | 122.12 | 121.91 |
| Average Monthly Index in 2018 | 1705.16 | 1705.16 |
| % of Avg. Monthly Index in 2018 | 7.16% | 7.15% |

| Orlando | Multiple Linear Regression | Lasso Regression |
|---|---|---|
| RMSE | 107.2 | 86.32 |
| Average Monthly Index in 2018 | 1350.46 | 1350.46 |
| % of Avg. Mthly Index in 2018 | 7.94% | 6.39% |

| Tampa | Multiple Linear Regression | Lasso Regression |
|---|---|---|
| RMSE | 86.96 | 90.98 |
| Average Monthly Index in 2018 | 1258.61 | 1258.61 |
| % of Avg. Mthly Index in 2018 | 6.91% | 7.23% |

| Jacksonville | Multiple Linear Regression | Lasso Regression |
|---|---|---|
| RMSE | 67.82 | 37.86 |
| Average Monthly Index in 2018 | 1065.16 | 1065.16 |
| % of Avg. Mthly Index in 2018 | 6.37% | 3.55% |

# Feature Importances for Miami

| | Random Forest | Lasso | Signs of Coefficients |
|---|---|---|---|
| 1 | owner_occupied_housing_units_median_value | pop_25_years_over | (+) |
| 2 | income_200000_or_more | children | (-) |
| 3 | commute_less_10_mins | mortgaged_housing_units | (-) |
| 4 | male_male_households | housing_units | (-) |
| 5 | not_us_citizen_pop | owner_occupied_housing_units_median_value | (+) |
| 6 | gini_index | children_in_single_female_hh | (+) |
| 7 | dwellings_50_or_more_units | FL_Unemployment | (-) |
| 8 | dwellings_5_to_9_units | owner_occupied_housing_units | (+) |
| 9 | children_in_single_female_hh | income_125000_149999 | (+) |
| 10 | male_35_to_39 | male_5_to_9 | (+) |

# Feature Importances for Orlando

| | Random Forest | Lasso | Signs of Coefficients |
|---|---|---|---|
| 1 | owner_occupied_housing_units_median_value | dwellings_50_or_more_units | (+) |
| 2 | FL_Unemployment | FLSTHPI_Yearly_Avg | (+) |
| 3 | FLSTHPI_Yearly_Avg | commute_less_10_mins | (+) |
| 4 | unemployed_pop | male_male_households | (+) |
| 5 | male_85_and_over | commuters_by_bus | (+) |
| 6 | income_15000_19999 | children | (-) |
| 7 | two_parents_mother_in_labor_force_families_with_young_children | employed_arts_entertainment_recreation_accommodation_food | (+) |
| 8 | children_in_single_female_hh | dwellings_20_to_49_units | (-) |
| 9 | male_male_households | male_85_and_over | (-) |
| 10 | employed_arts_entertainment_recreation_accommodation_food | median_year_structure_built | (-) |

# Feature Importances for Tampa

| | Random Forest | Lasso | Signs of Coefficients |
|---|---|---|---|
| 1 | unemployed_pop | dwellings_50_or_more_units | (+) |
| 2 | income_40000_44999 | pop_25_years_over | (+) |
| 3 | income_30000_34999 | not_us_citizen_pop | (+) |
| 4 | income_200000_or_more | children_in_single_female_hh | (+) |
| 5 | FL_Unemployment | owner_occupied_housing_units_ median_value | (+) |
| 6 | FLSTHPI_Yearly_Avg | some_college_and_associates_ degree | (-) |
| 7 | Owner_occupied_housing_units_ median_value | FL_Unemployment | (-) |
| 8 | male_45_64_grade_9_12 | commuters_by_bus | (-) |
| 9 | income_15000_19999 | income_125000_149999 | (-) |
| 10 | dwellings_50_or_more_units | unemployed_pop | (-) |

# Feature Importances for Jacksonville

| | Random Forest | Lasso | Signs of Coefficients |
|---|---|---|---|
| 1 | households_public_asst_or_food_stamps | Owner_occupied_housing_units_ median_value | (+) |
| 2 | owner_occupied_housing_units_median_value | employed_arts_entertainment_ recreation_accommodation_food | (-) |
| 3 | income_200000_or_more | vacant_housing_units_for_rent | (-) |
| 4 | commuters_by_bus | female_62_to_64 | (-) |
| 5 | median_year_structure_built | children_in_single_female_hh | (+) |
| 6 | walked_to_work | employed_transportation_ warehousing_utilities | (-) |
| 7 | income_125000_149999 | commute_35_44_mins | (+) |
| 8 | FL_Unemployment | median_year_structure_built | (+) |
| 9 | income_30000_34999 | some_college_and_associates_degree | (-) |
| 10 | FLSTHPI_Yearly_Avg | in_grades_5_to_8 | (+) |

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- **Conclusion**
- Future Work

# Conclusion

- Lasso in general achieved a higher accuracy than Multiple Linear Regression with a smaller average error/RMSE, with proper hyperparameter tuning

- Within different economic classes, variables can have smaller or lager effects on the price, especially the number of dwellings with 50 or more housing units

- Also, using algorithms that do not directly predict the rent level, we can also offer important guidance to property developers and other stakeholders as to the growth profile for smaller areas within a larger metro area

# Agenda

- Research Questions and Motivation
- Data Used
- Why Florida
- Data Visualization
- Data Cleaning
- Clustering
- Feature Selection and Target Transformation
- Modeling
- Results
- Conclusion
- **Future Work**

# Next Steps/Future Work

- Clean the data more. Remove outliers
- Real Estate often has very localized trends, talking to domain experts in each area should provide more insight.
- Incorporate time series analysis for future predictions
- Use more counties.  Zip codes get more scarce than you appreciate as you drill down
- Normalize the data for KMeans clustering
- Analyze columns with large amounts of nans to investigate better imputation methods

You've Got Questions.
We've Got You Covered