U.S. Multi-Family Rental Market

Breakout Room 2:
Jonah Gerstel, Michael McGuigan,
Lena Chretien, Ben Burkey

Table of Contents

Business Approach

About the Business:

- **Markerr** provides insights for real estate investors about the jobs, people, and financial trends on a location basis.

Our Objective:

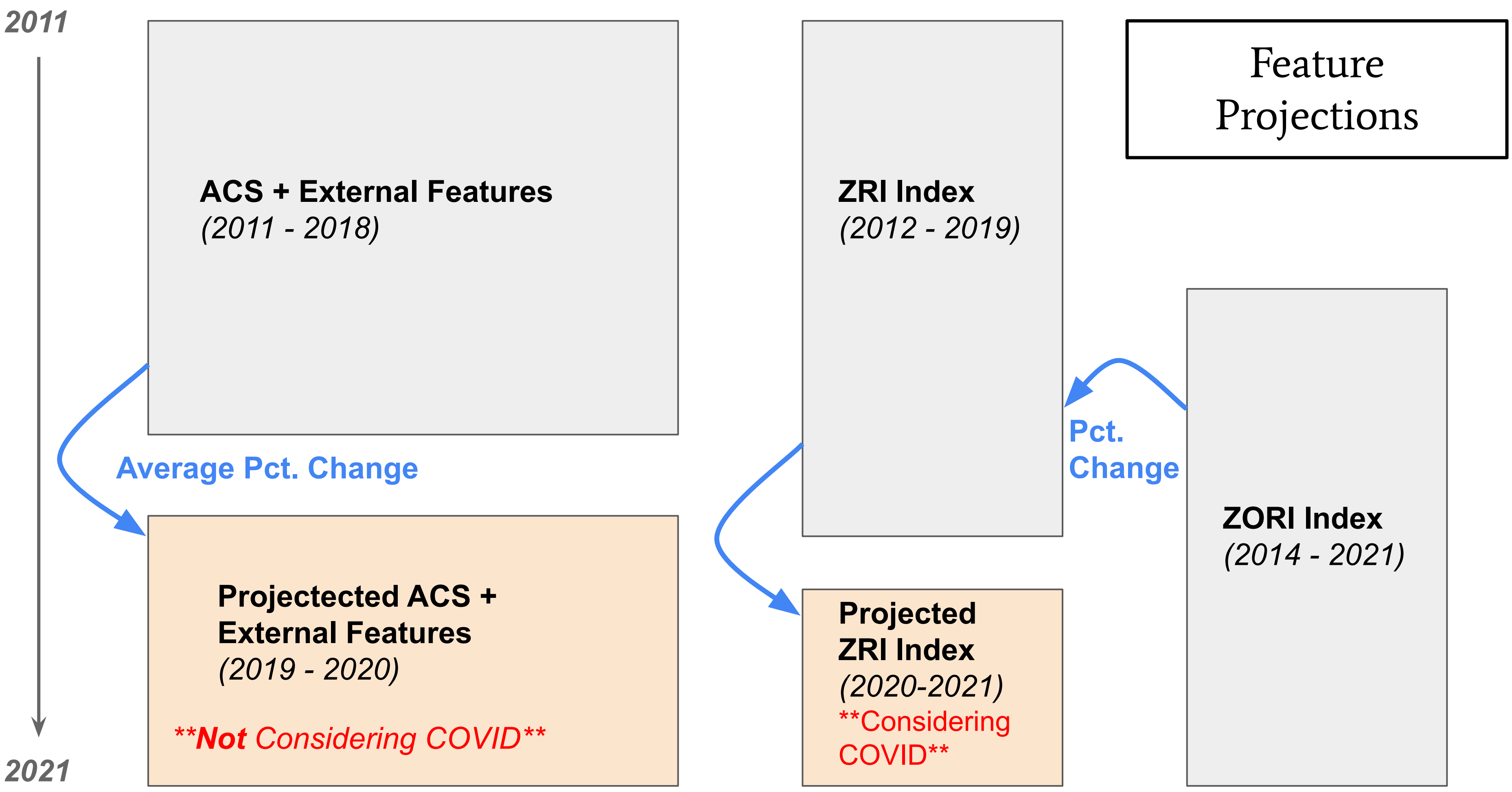
- Gain insight into rent price indices by comparing pre-Covid predictions to Covid projections for multi-family homes in large cities.

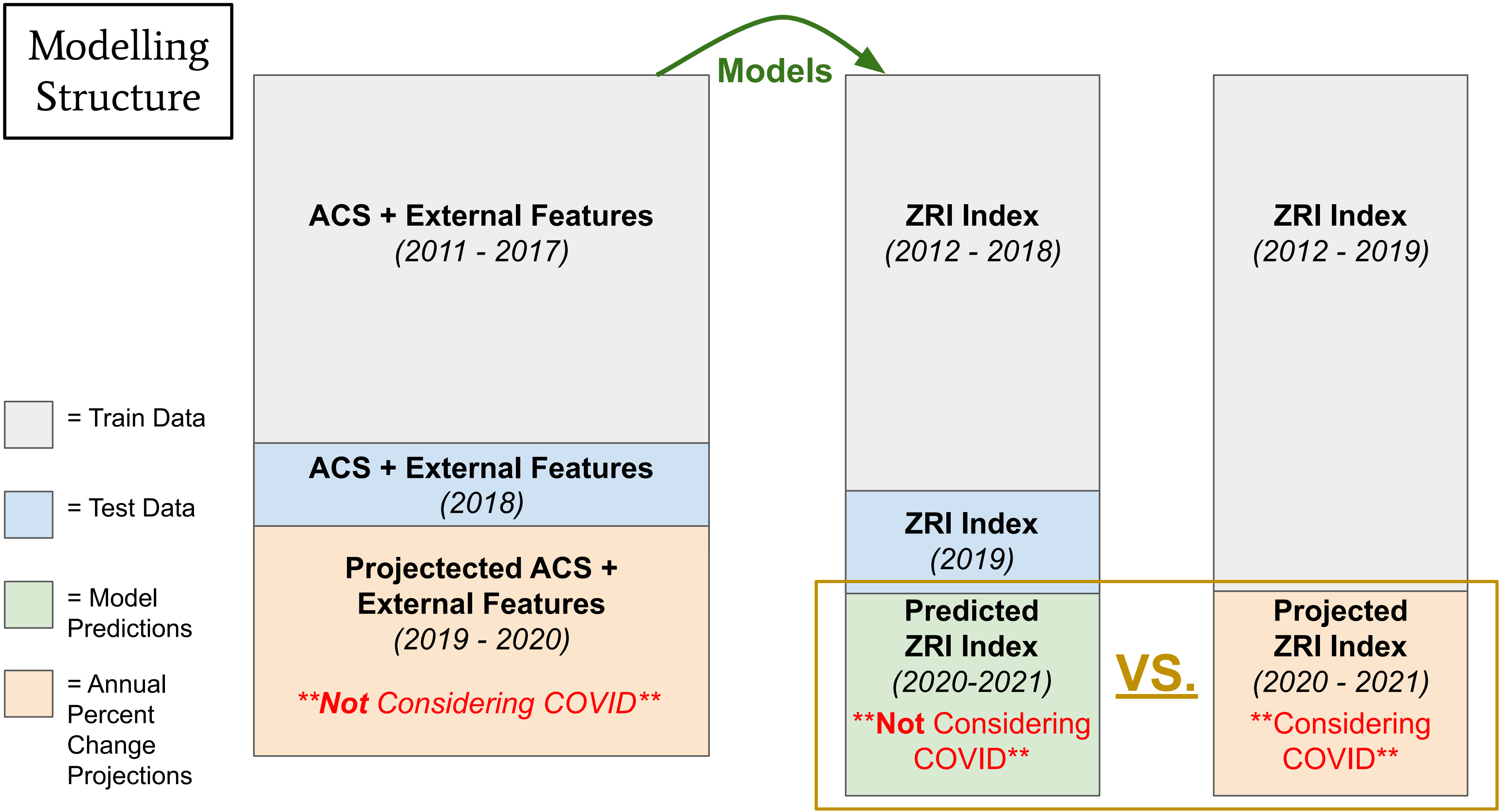
Data Engineering

Data Collected:

- American Community Survey (ACS) data between 2011-2018
- Gross Domestic Product (GDP) between 2011-2018
- Effective Federal Funds Rate (EFFR) between 2011-2018
- Number of Businesses per ZIP code between 2011-2018
- Consumer Price Index (CPI) data for ‘shelter’ between 2011-2018
- Zillow Observed Rent Index (ZORI) data for all homes between 2019-2021

- **Target:** Zillow Rent Index (ZRI) data for Multi-family homes between 2012-2019


2011
ACS + External Features
(2011 - 2018)
ZRI Index
(2012 - 2019)
Feature Projections
Pct. Change
Average Pct. Change
ZORI Index
(2014 - 2021)
Projectected ACS + External Features
(2019 - 2020)
Not Considering COVID
Projected ZRI Index
(2020-2021)
Considering COVID
2021


Modelling Structure
= Train Data
= Test Data
= Model Predictions
= Annual Percent Change Projections
ACS + External Features (2011 - 2017)
ACS + External Features (2018)
Projectected ACS + External Features (2019 - 2020)
Not Considering COVID
Models
ZRI Index (2012 - 2018)
ZRI Index (2019)
Predicted ZRI Index (2020-2021)
Not Considering COVID
VS.
ZRI Index (2012 - 2019)
Projected ZRI Index (2020 - 2021)
Considering COVID

Feature Engineering

- Reduced multicollinearity by creating ratios for features based on population count. (Divided by total population)
  - Employed Ratio = Employed Population /Total Population
  - Percentage Male/Female

- Dropped features we considered overly specific
  - Example: Males between the age 45-64 with an Associates Degree was one feature
  - This data was captured in more general features such as percent of males and percent with associates degrees

- Filtered all data for zip codes in cities with populations greater than 100,000 people

PCA (Principal Component Analysis)

- Many features did not contribute to the variability of the data (noise)
- To preserve 95% of the variability, 35 features were retained.
- Two most important features were:
    - 'total_pop',
    - 'Income_per_capita'
- They alone explain 37% and 15% of the variability, respectively

- Data added by us is also important:
    - CPI (inflation)
    - GDP (Gross domestic Product)

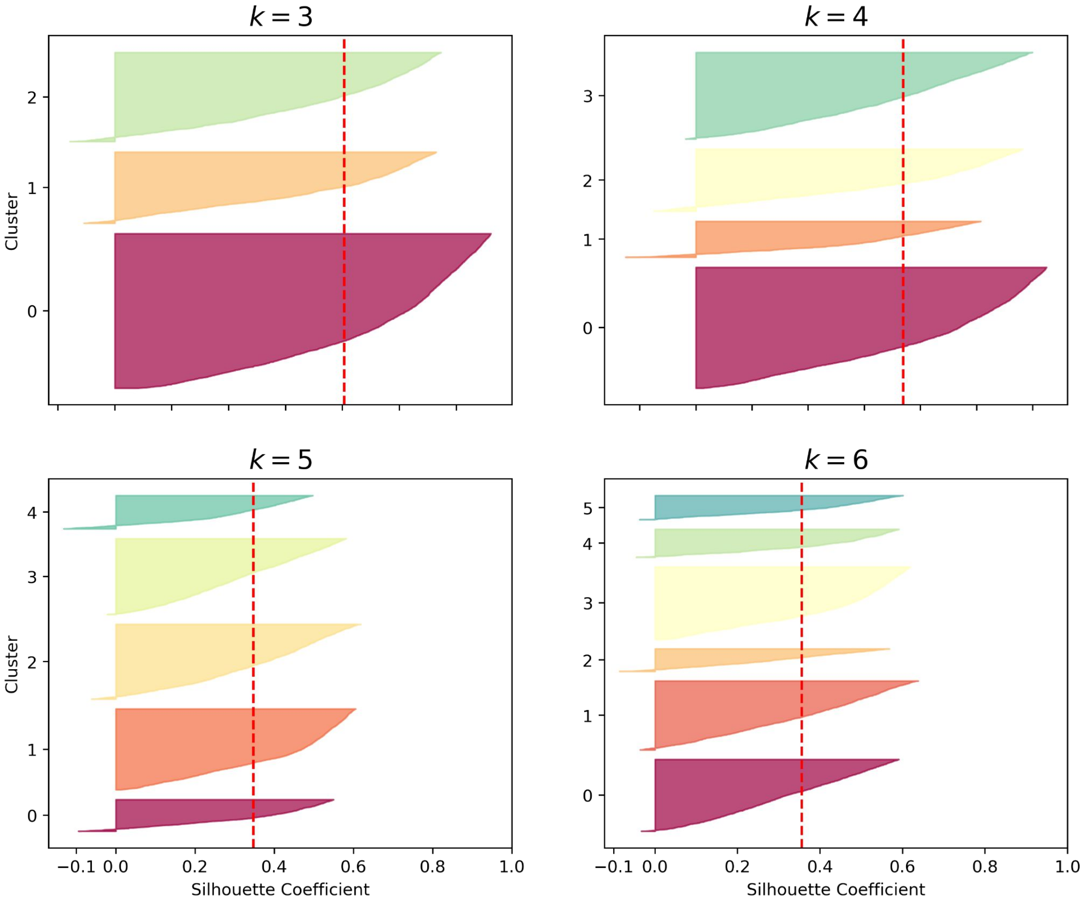
- All models were run **WITHOUT** and **WITH** PCA (results shown later)

Clustering

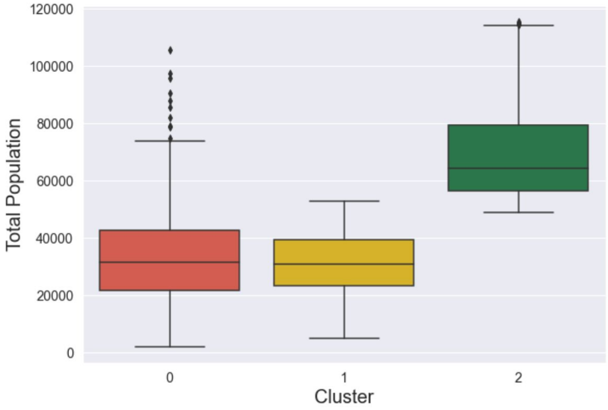
- KMeans:
  - Distance of a point to cluster center.
  - Silhouette shows distance of point to cluster
- Features used for clusters:
  - Rent, income, and population
- Chose 3 clusters

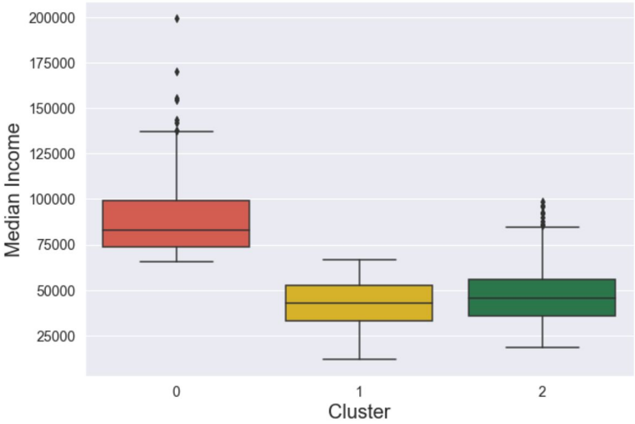
Future work:
Try other clustering methods

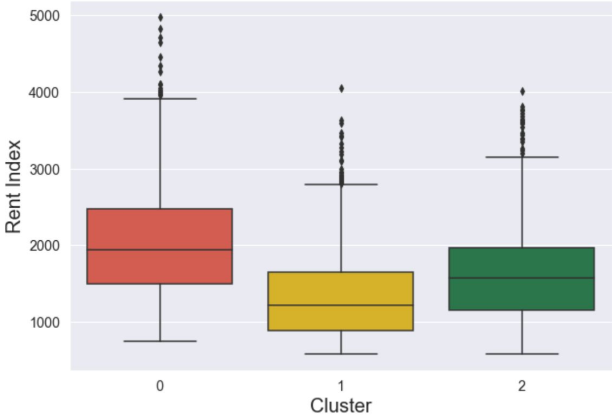
- DBScan:
  - Uses density of a region as clusters
- Spectral clustering:
  - Reduces dimensionality



Box Plots by Cluster





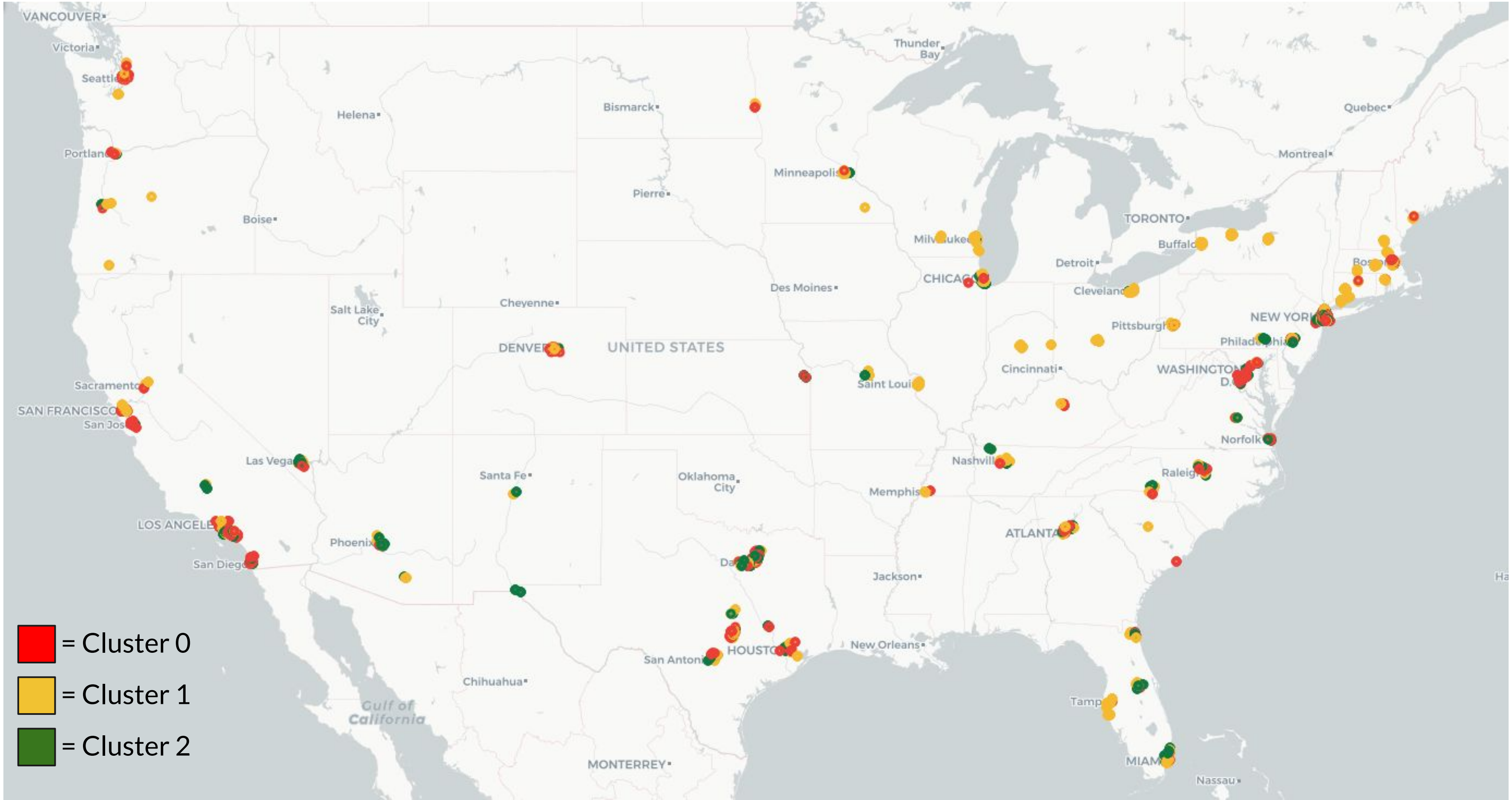


<u>Cluster 0</u>
- Lower Total Pop
- High Median Income
- High Rent Index
- Ex. Tribeca

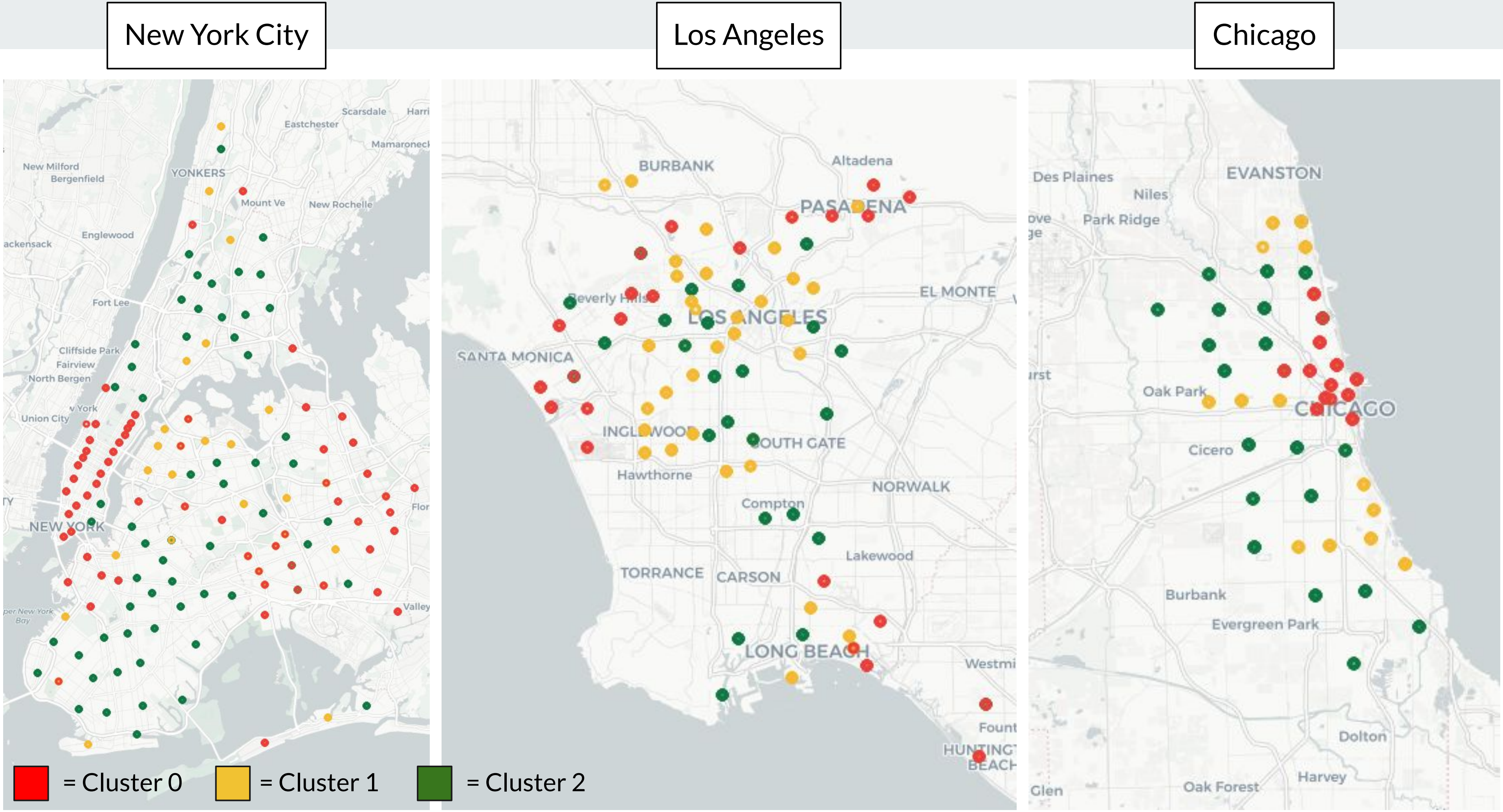
<u>Cluster 1</u>
- Lower Total Pop
- Lower Median Income
- Lower Rent Index
- Ex. Astoria, Queens

<u>Cluster 2</u>
- Highest Total Pop
- Lower Median Income
- Medium Rent Index
- Ex. Lower East Side

Cluster Map

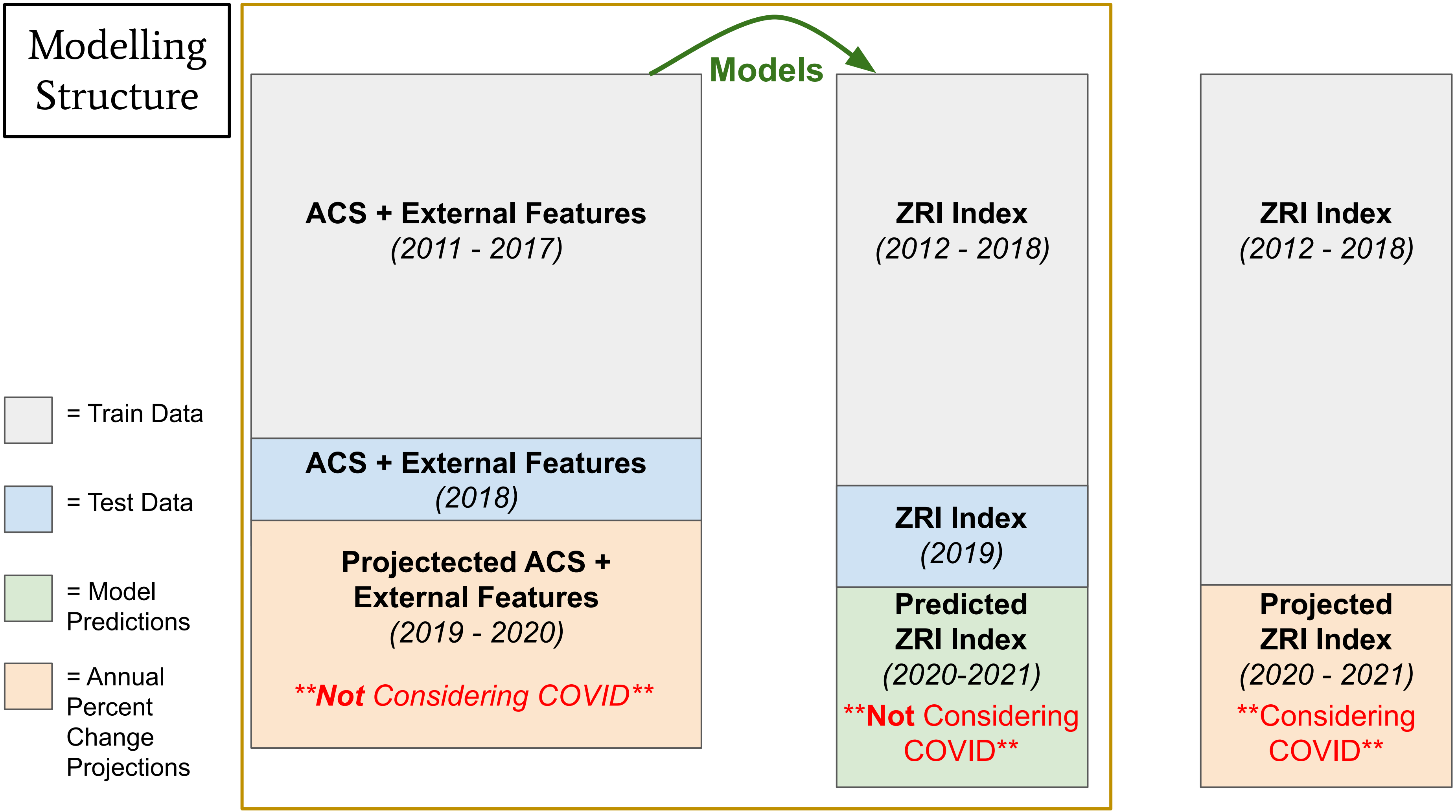


VANCOUVER
Victoria
Seattle
Portland
Helena
Bismarck
Thunder Bay
Quebec
Montreal
Minneapolis
Pierre
Boise
TORONTO
Buffalo
Milwaukee
Detroit
Boston
CHICAGO
Des Moines
Cleveland
Cheyenne
Salt Lake City
NEW YORK
Pittsburgh
DENVER
UNITED STATES
Philadelphia
Cincinnati
WASHINGTON D.C.
Sacramento
Saint Louis
SAN FRANCISCO
San Jose
Norfolk
Nashville
Las Vegas
Santa Fe
Oklahoma City
Raleigh
Memphis
LOS ANGELES
ATLANTA
Phoenix
San Diego
Dallas
Jackson
New Orleans
San Antonio
HOUSTON
Chihuahua
Tampa
Gulf of California
MIAMI
MONTERREY
Nassau
= Cluster 0
= Cluster 1
= Cluster 2


New York City
Los Angeles
Chicago
Scarsdale
Eastchester
Mamaroneck
New Milford
Bergenfield
YONKERS
Mount Ve
New Rochelle
Englewood
Fort Lee
Cliffside Park
Fairview
North Bergen
Union City
NEW YORK
BURBANK
Altadena
PASADENA
Beverly Hills
EL MONTE
LOS ANGELES
SANTA MONICA
INGLEWOOD
SOUTH GATE
Hawthorne
NORWALK
Compton
Lakewood
TORRANCE
CARSON
LONG BEACH
Des Plaines
Niles
EVANSTON
Park Ridge
Oak Park
CHICAGO
Cicero
Burbank
Evergreen Park
Dolton
Harvey
Oak Forest
= Cluster 0
= Cluster 1
= Cluster 2

Machine Learning Models

1. Lasso Regression
2. Random Forest
3. Gradient Boosting


Modelling Structure
= Train Data
= Test Data
= Model Predictions
= Annual Percent Change Projections
Models
ACS + External Features (2011 - 2017)
ACS + External Features (2018)
Projectected ACS + External Features (2019 - 2020)
Not Considering COVID
ZRI Index (2012 - 2018)
ZRI Index (2019)
Predicted ZRI Index (2020-2021)
Not Considering COVID
ZRI Index (2012 - 2018)
Projected ZRI Index (2020 - 2021)
Considering COVID

Lasso Regression

Lasso Regression Results								
	All	All (PCA)	Cluster 1	Cluster 1 (PCA)	Cluster 2	Cluster 2 (PCA)	Cluster 3	Cluster 3 (PCA)
Training RMSE	76.62	76.58	90.98	92.08	67.52	69.61	68.33	72.05
Test RMSE	83.51	86.46	139.81	79.46	64.86	73.38	76.21	89.34
Training R2	0.988	0.988	0.986	0.985	0.985	0.984	0.987	0.986
Test R2	0.987	0.986	0.966	0.987	0.989	0.985	0.986	0.982

Why Lasso?

- Automatic feature selection
- Easily interpretable (linear regression)
- Avoids overfitting

Results:

- Reduced number of features to below 10
- PCA did not improve results outside of Cluster 1
- Clustering improved results of linear regression

Random Forest

Random Forest Results								
	All	All (PCA)	Cluster 1	Cluster 1 (PCA)	Cluster 2	Cluster 2 (PCA)	Cluster 3	Cluster 3 (PCA)
Training RMSE	50.46	56.81	61.09	66.83	52.07	45.83	43.79	41.27
Test RMSE	75.36	73.57	79.84	78.55	71.92	71.22	85.37	83.82
Training R2	0.979	0.980	0.959	0.960	0.976	0.977	0.965	0.967
Test R2	0.989	0.990	0.986	0.986	0.987	0.986	0.984	0.985

Why Random Forest?

- Not affected by multicollinearity
- High performance and accuracy
- No feature scaling required

Results:

- Tendency to overfit the training data
- PCA reduced overfitting on all features & cluster 1
- PCA improved results on all grouping
- Clustering did not improve overall results
- Best performing model overall

Gradient Boosting

Gradient Boosting Results								
	All	All (PCA)	Cluster 1	Cluster 1 (PCA)	Cluster 2	Cluster 2 (PCA)	Cluster 3	Cluster 3 (PCA)
Training RMSE	62.67	63.60	50.65	57.52	52.47	55.02	50.08	44.34
Test RMSE	76.32	75.96	79.89	79.23	72.38	72.57	85.52	84.24
Training R2	0.984	0.984	0.968	0.970	0.981	0.982	0.967	0.975
Test R2	0.989	0.990	0.986	0.987	0.986	0.986	0.984	0.985

Why Gradient Boosting?

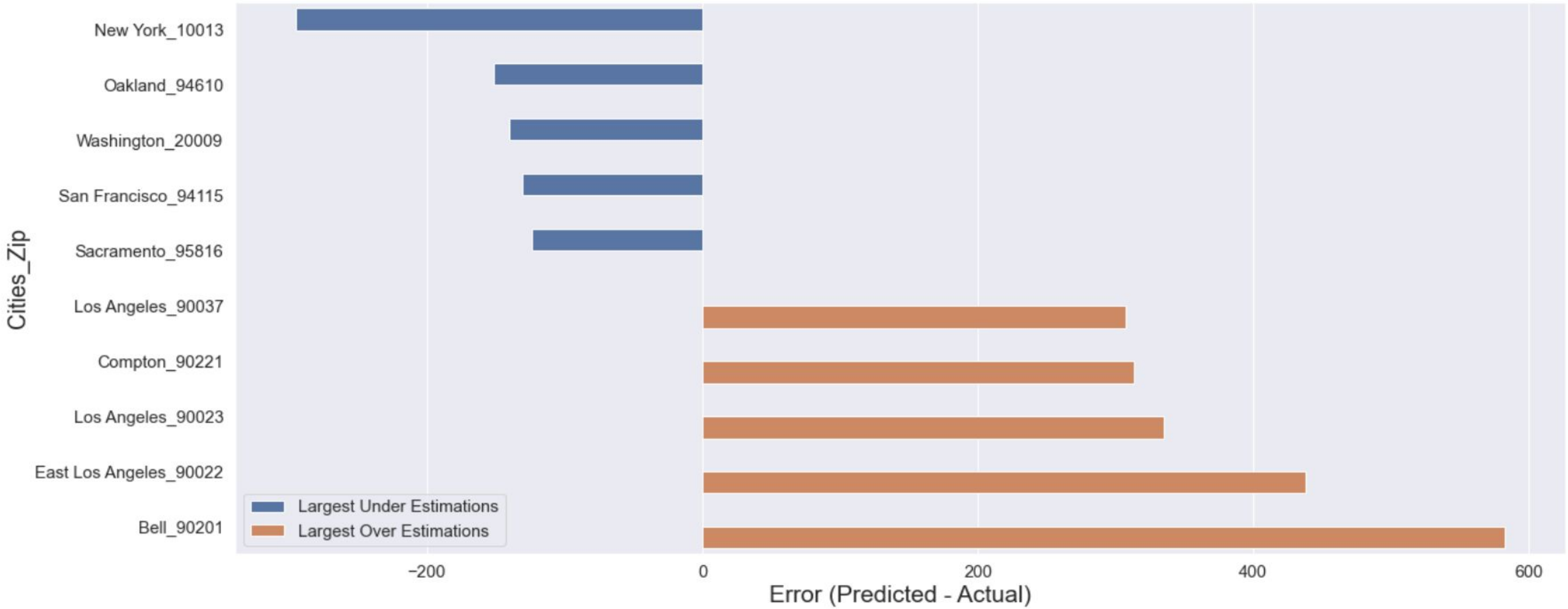
- Not affected by multicollinearity
- High performance and accuracy
- No feature scaling required

Results:

- Tendency to overfit the training data
- PCA reduced overfitting on all features, cluster 1, and cluster 2
- PCA improved results on all grouping
- Clustering did not improve overall results

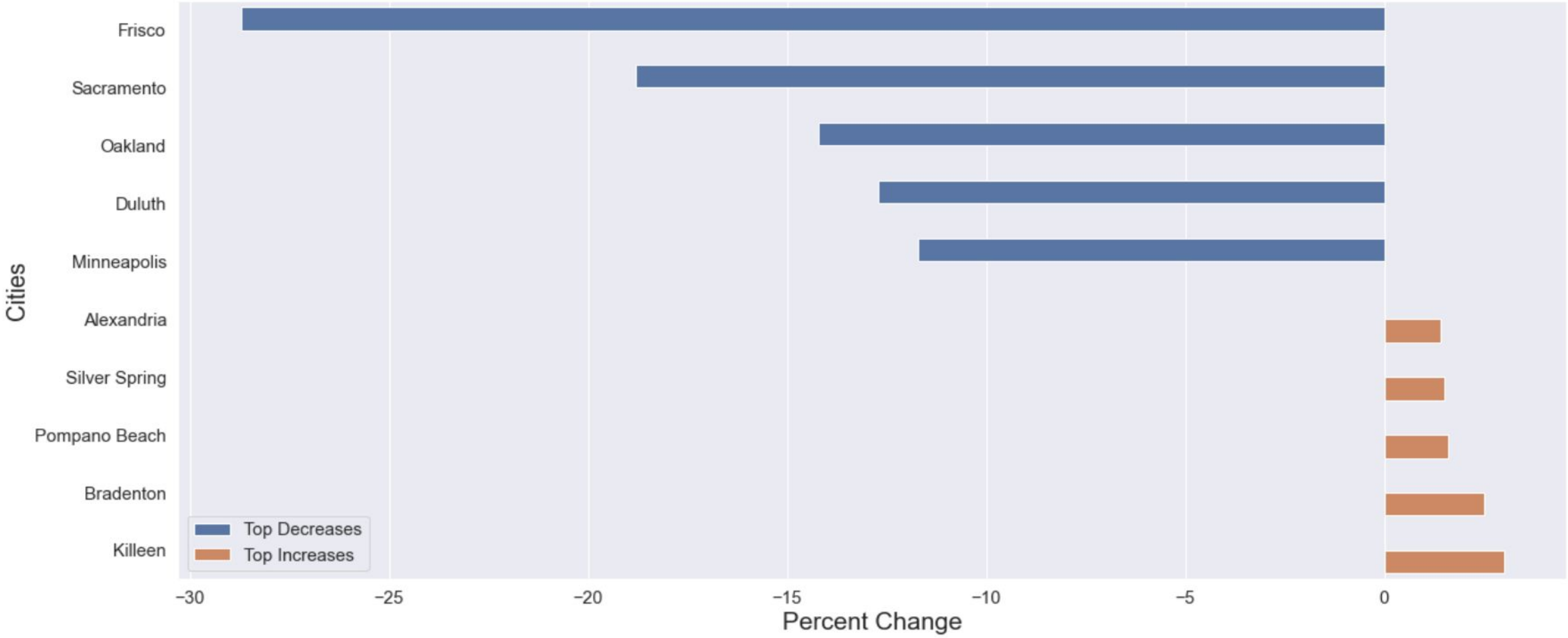
Top 5 Cities with Largest Over and Under Estimations from Random Forest Model (Test Data - 2018)

RMSE of $73.57 (on average, Random Forest Model over estimated by $18.67)



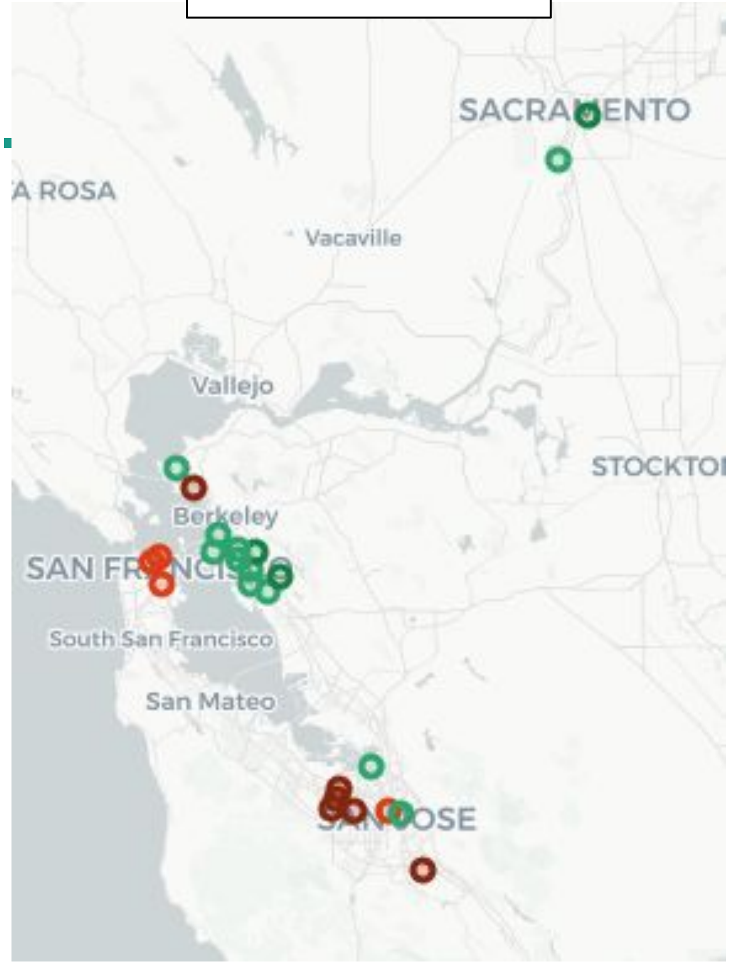
Top 5 Cities with Largest Decreases and Increases versus Pre-Covid Projections

Average Decrease of 5.5% across all ZIP codes analyzed

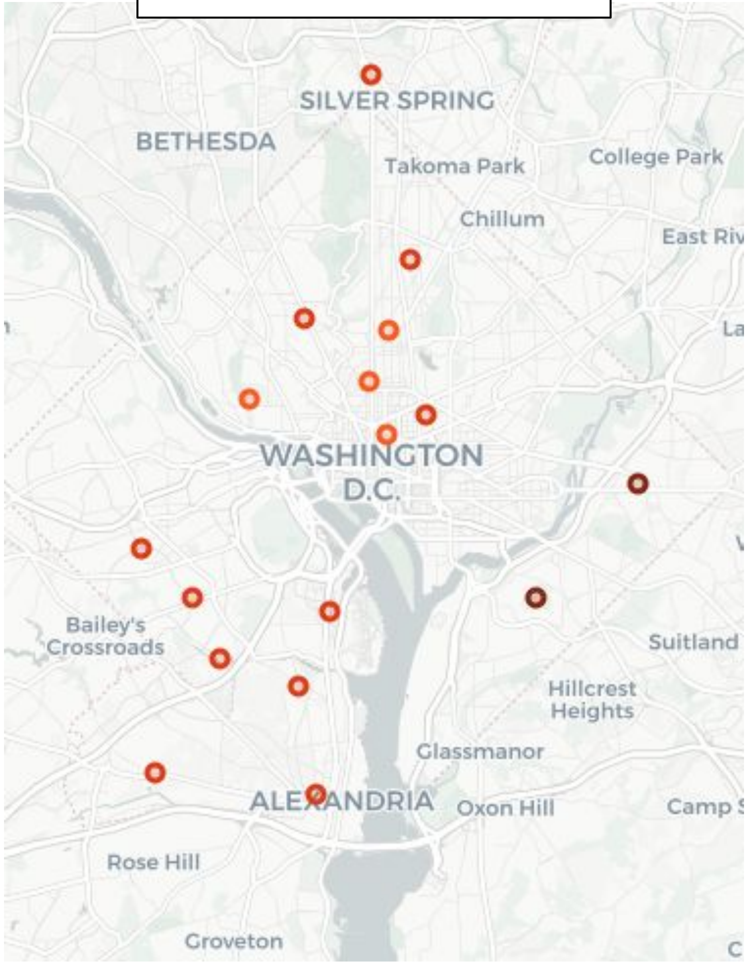


Percent change = (ZRI(nonCovid) - ZRI(Covid) - Model Error) / (ZRI(nonCovid))

SFO / Sacramento



Silver Spring / Alexandria



Conclusion / Next Steps

Conclusion:

- On average, urban rental markets have seen a decreased rent when compared to pre-Covid predictions (Average Percent Change = -5.5%)
- Rental market in the north east has returned to pre-Covid predicted levels, more than the west coast
- PCA improved results on tree based models, but not linear models
- Clustering improved results on linear models, but not tree based models

Next Steps:

- Test different clustering methods and groupings
- Include more data for more accurate predictions (ex. Include housing prices)
- Incorporate Covid data directly (ex. Instead of using ZORI to predict ZRI)
- Look at single family homes in suburban areas