Capstone Project: San Francisco Affordable Housing Predictive Modeling

By Joey Notaro

Problem Statement

- The San Francisco Mayor's Office of Housing and Community Development (MOHCD) is concerned with the pace of construction of new housing and particularly affordable housing in the city.

Problem Statement

- What traits or characteristics do large housing unit, especially large affordable housing unit, projects have in common to promote for future development?

- Goal: Clustering Model with silhouette score of at least 0.75 and distinct features in each cluster to analyze housing trends.

Problem Statement

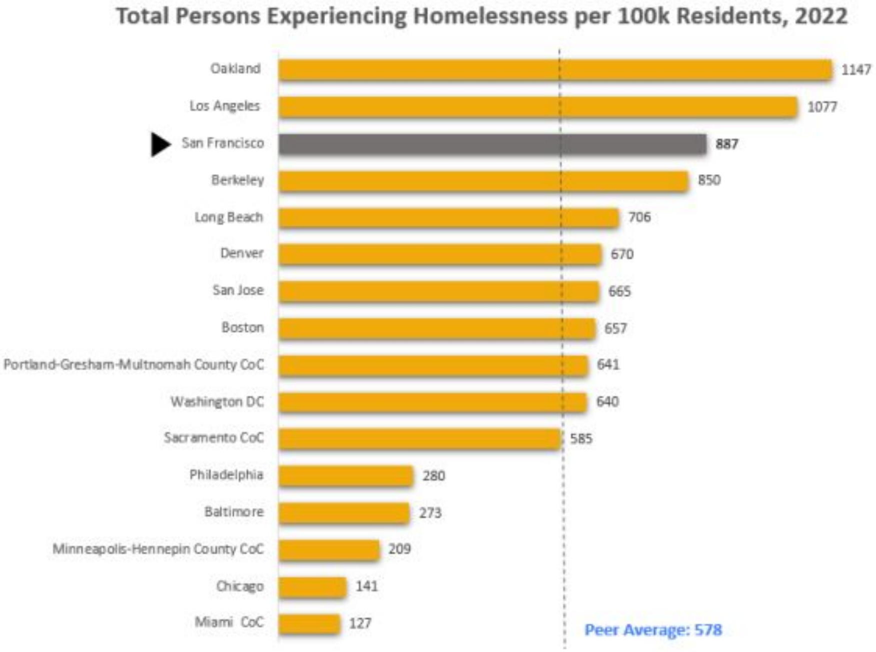
- Can we use knowledge of recent past construction numbers to predict future construction trends and make reasonable projections of whether the city is on track to keep up with affordable housing targets

- Goal: Time Series Model with root mean squared error of at least 50% less than null model.

Why Affordable Housing?

- Total unhoused population dropped 3% from 2019 to 2022.

- San Francisco remains third largest unhoused population per capita.



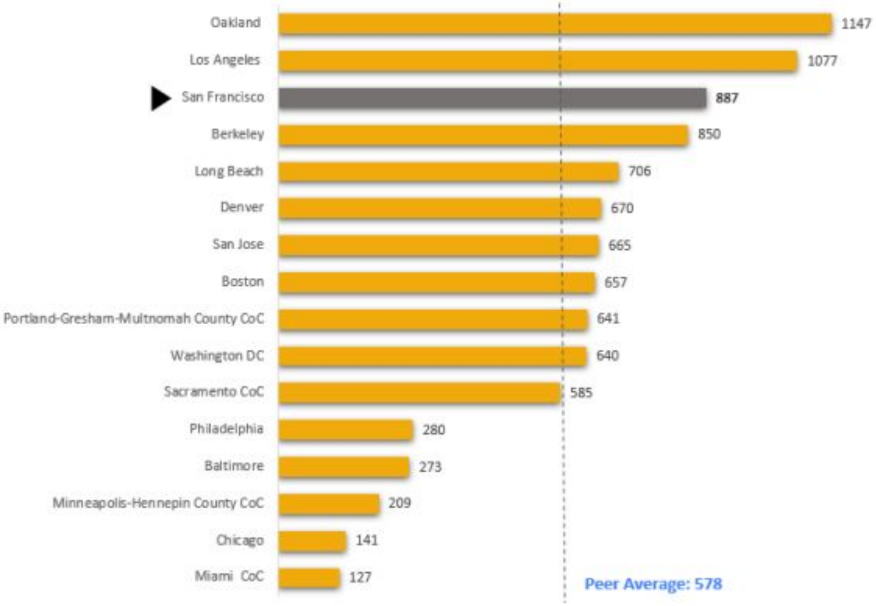
Total Persons Experiencing Homelessness per 100k Residents, 2022

Why Affordable Housing?

- Housing crisis in California and Bay Area is decades old.

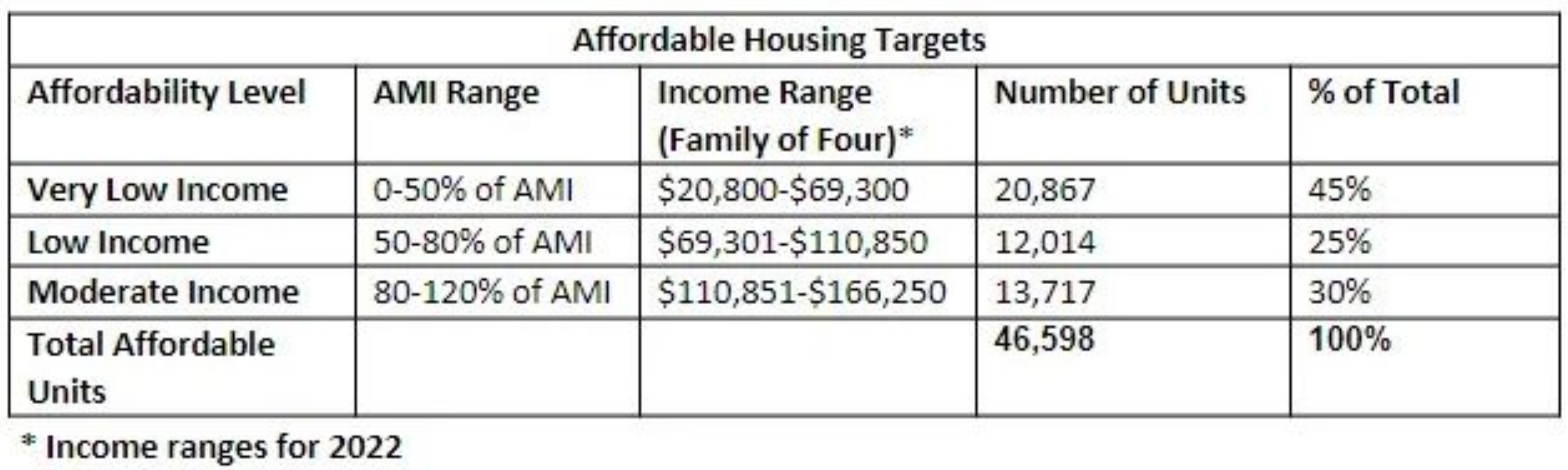
- State has specific housing plan with affordable goals. Almost every city is behind.

Total Persons Experiencing Homelessness per 100k Residents, 2022



What Does "Affordable" Mean?

- No more than 30% of a family's income.
- Based on Housing and Urban Development's Area Median Income numbers (AMI).



Affordable Housing Targets				
Affordability Level	AMI Range	Income Range (Family of Four)*	Number of Units	% of Total
Very Low Income	0-50% of AMI	$20,800-$69,300	20,867	45%
Low Income	50-80% of AMI	$69,301-$110,850	12,014	25%
Moderate Income	80-120% of AMI	$110,851-$166,250	13,717	30%
Total Affordable Units			**46,598**	**100%**

* Income ranges for 2022

Data Source

- MOHCD collects contemporary affordable housing data.

- SF Planning Department collects more robust data on all development projects in the pipeline for the whole city.

- SF Planning Department release development pipeline data quarterly.

Data Collection & Cleaning

- Development pipeline data publicly available stretching back more than 10 years.

- Despite consistent data dictionary for the **last five years**, data features and variable naming conventions changed year to year and quarter to quarter.

- Some features containing MANY nulls.

Data Architecture Finding & Recommendation

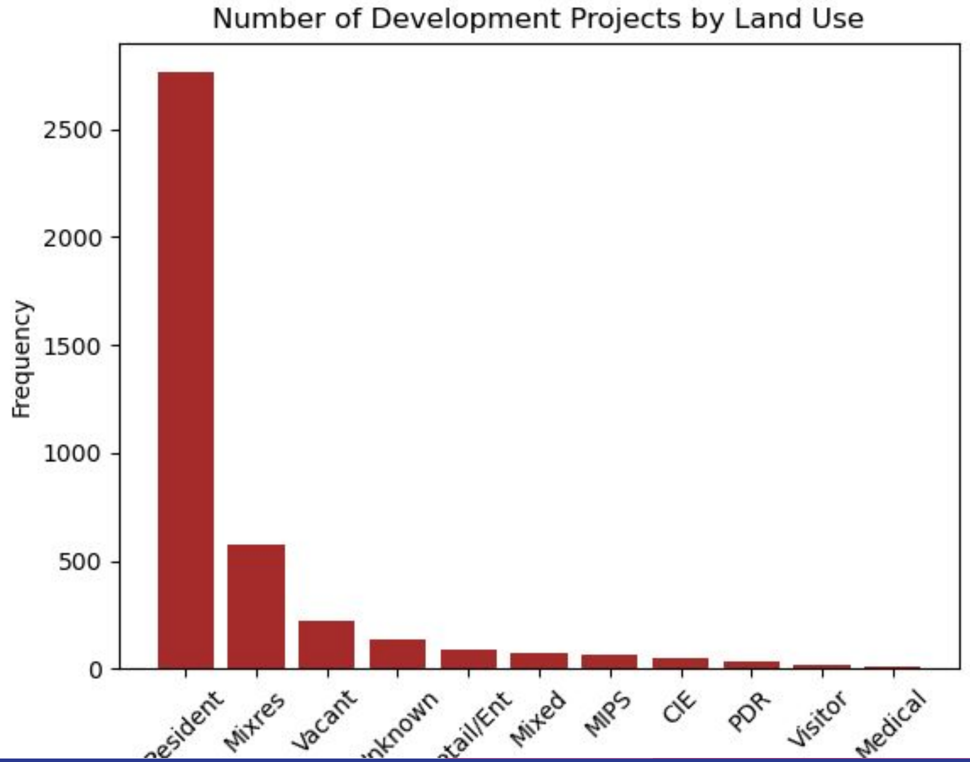
- FINDING: The city’s data sharing with the public is a great step in transparency and accountability, but its collection methods are inconsistent and incompatible with best practices.

- RECOMMENDATION: MOHCD should work with Planning Dept to urge the Mayor’s Office to hire a data engineer or data architect.

Exploratory Data Analysis

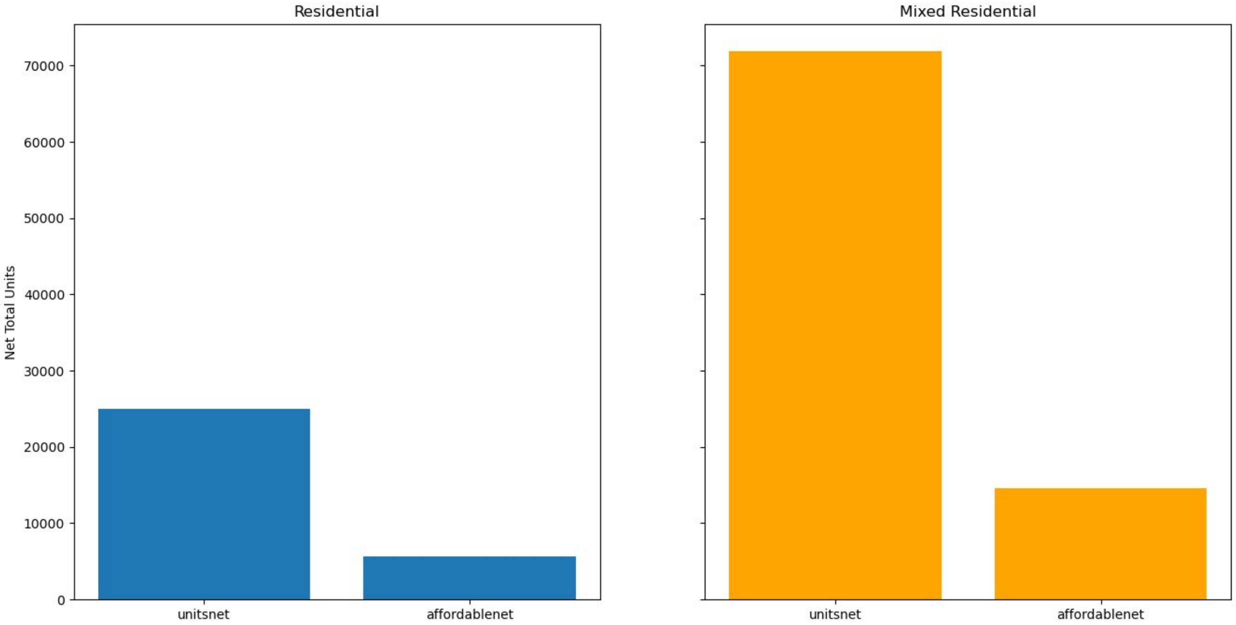
- Currently 4,106 projects in the pipeline.

- Overwhelming majority residential and mixed residential.



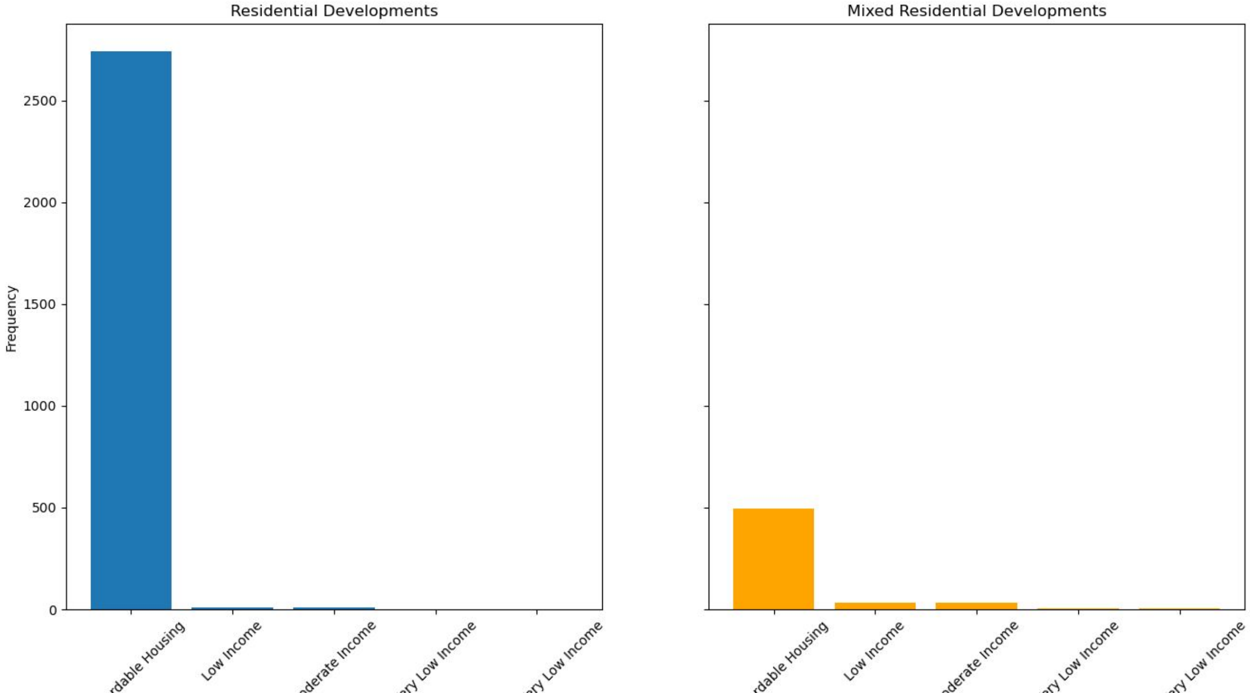
Housing Impacts

- Mixed residential projects have much larger housing impact.



Affordability Targets

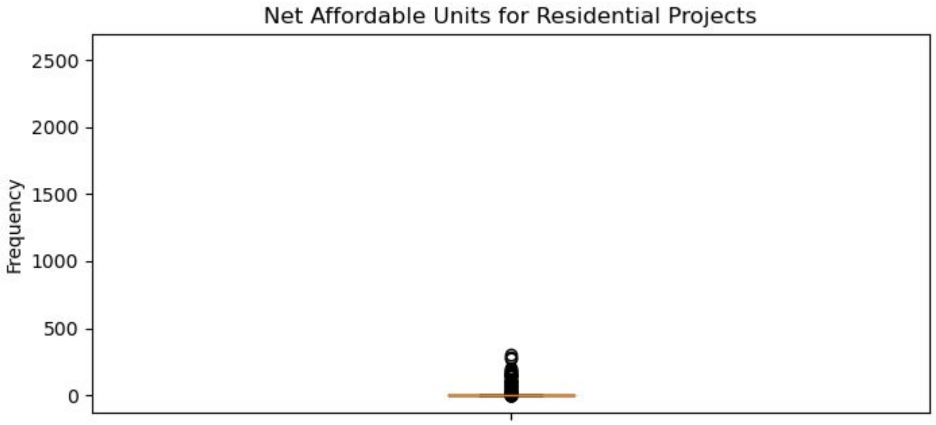
- Modest moderate income and low income goals.

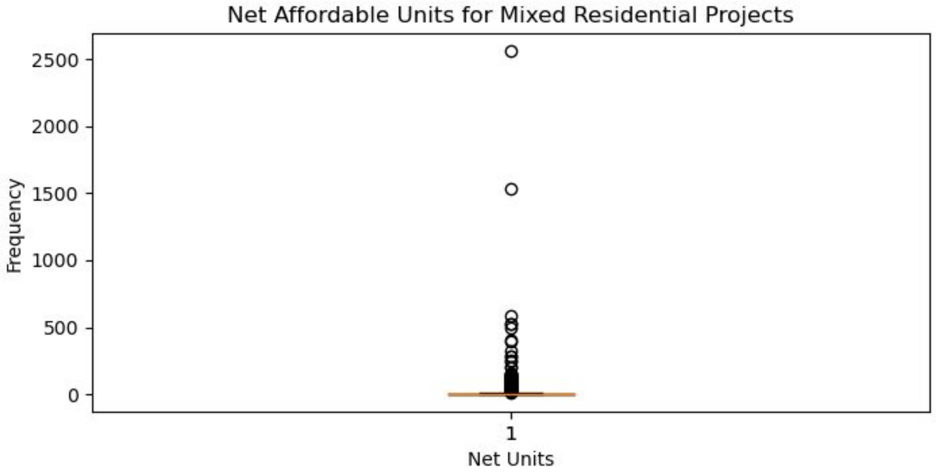


Net Units

- Middle 50% of all housing projects net 1-2 units and 0 affordable units.

- Very few outliers are responsible for almost all the net units and affordable units.





EDA Findings & Recommendations

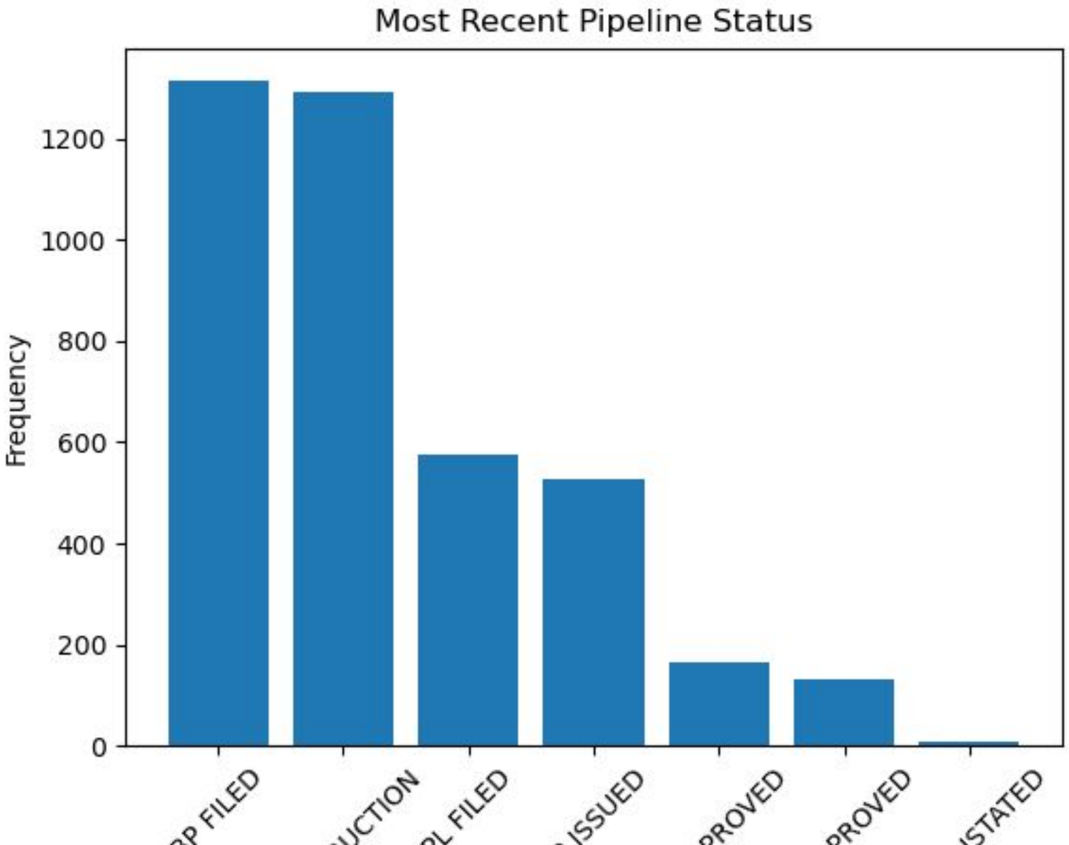
- FINDING #1: A few large outlier development projects are responsible for the overwhelming majority of net new affordable units and a larger proportion of them at in mixed residential zoned land uses as opposed to residential land uses.

- RECOMMENDATION #1: MOHCD needs to find ways to attract many more large-scale mixed residential development projects with affordable housing targets embedded in the design and construction, as these projects historically yield the most net affordable units.

Pipeline Status

- Over 1,200 projects (~30%) in construction.

- Just as many waiting Building Permit Approval.



Most Recent Pipeline Status

EDA Findings & Recommendations

- FINDING #2: With 4,106 projects in the pipeline only about 1,200 are currently under construction (~30%) with as many projects stalled at the phase of a building permit being filed with the Department of Building Inspections.

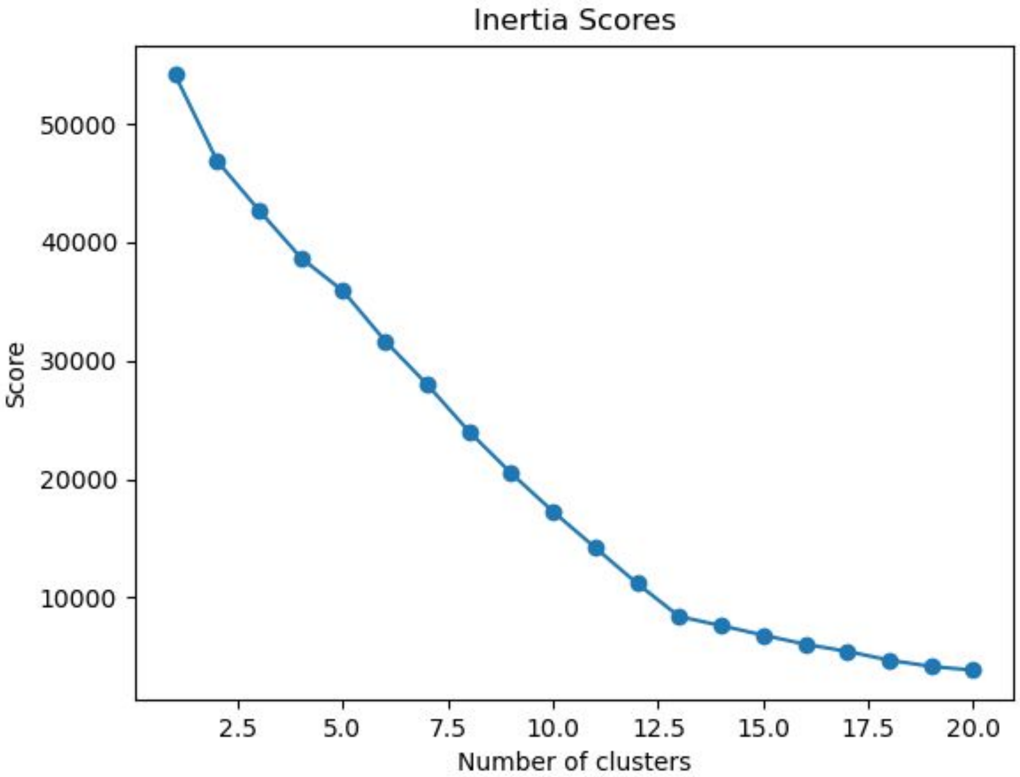
- RECOMMENDATION #2: MOHCD needs to work with the Mayor's Office to expedite development projects with clear affordable housing targets, trying to put leapfrog these projects to the head of the Building Inspections line.

Clustering Model Methodology

- After data cleaning, 51 features available for feature selection.
- Maintain most salient housing features like net units, affordable units, affordability targets, etc.
- Iterate through numeric features and categorical features separately for feature selection.
- Combine features from both iterations into final production model with goal of 0.75 silhouette score.

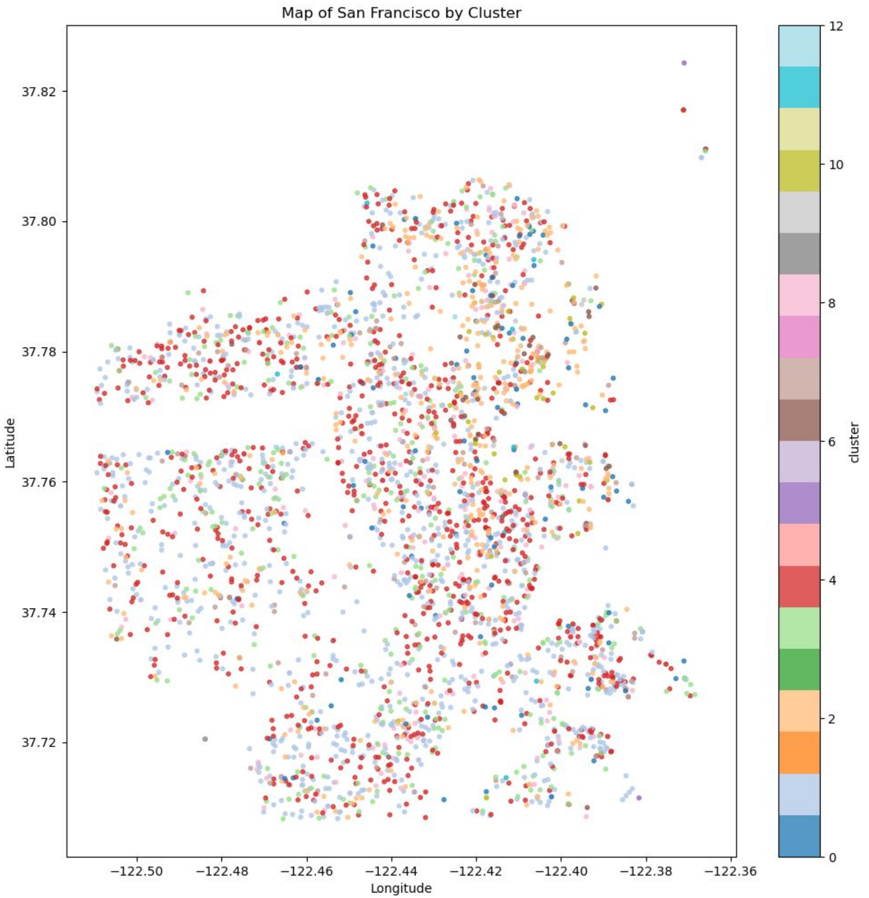
Clustering Production Model

- Six features:
  - Land Use
  - Pipeline Status
  - Net Units
  - Affordable Units
  - Affordability Targets
  - Supervisor District
- Thirteen Clusters
- Silhouette Score: ~0.59

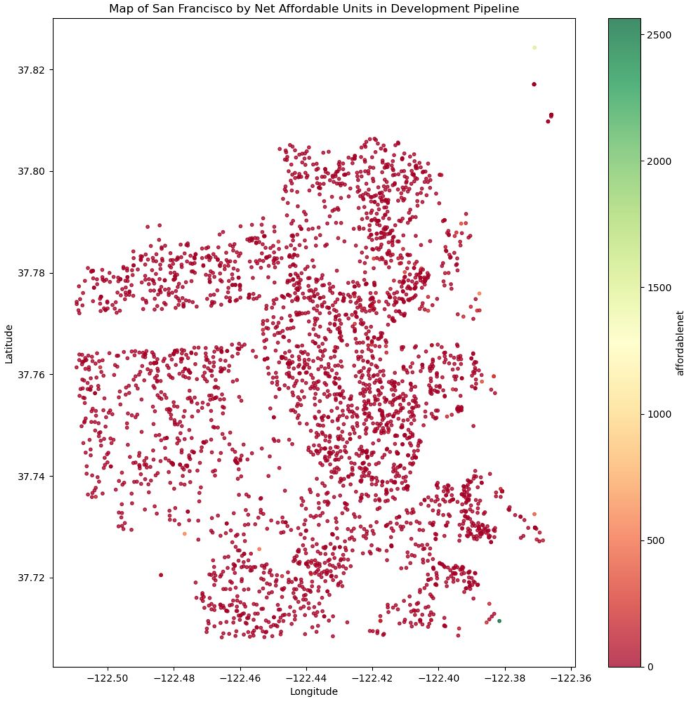


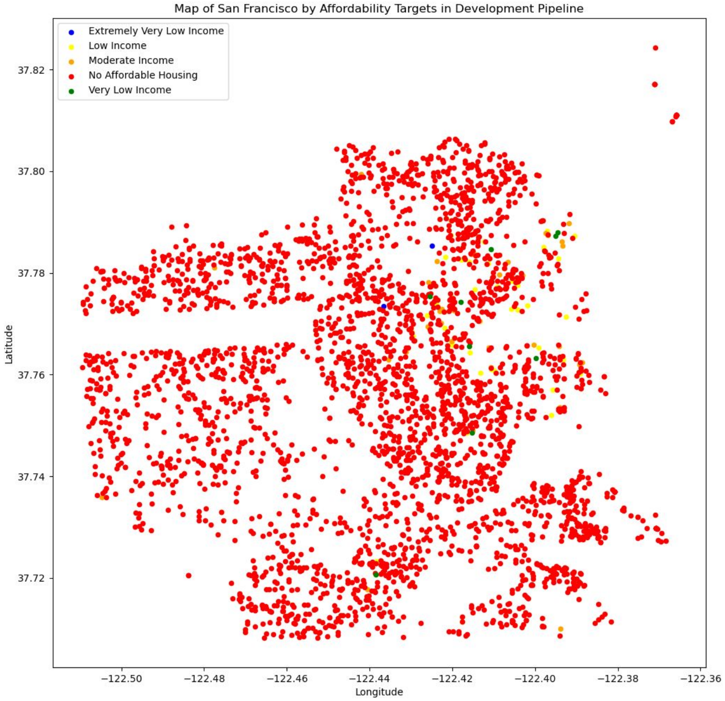
Clusters

- No discernible geographic trends.
- No visual trend between net affordable units or affordability targets and clusters.



Affordability Trends

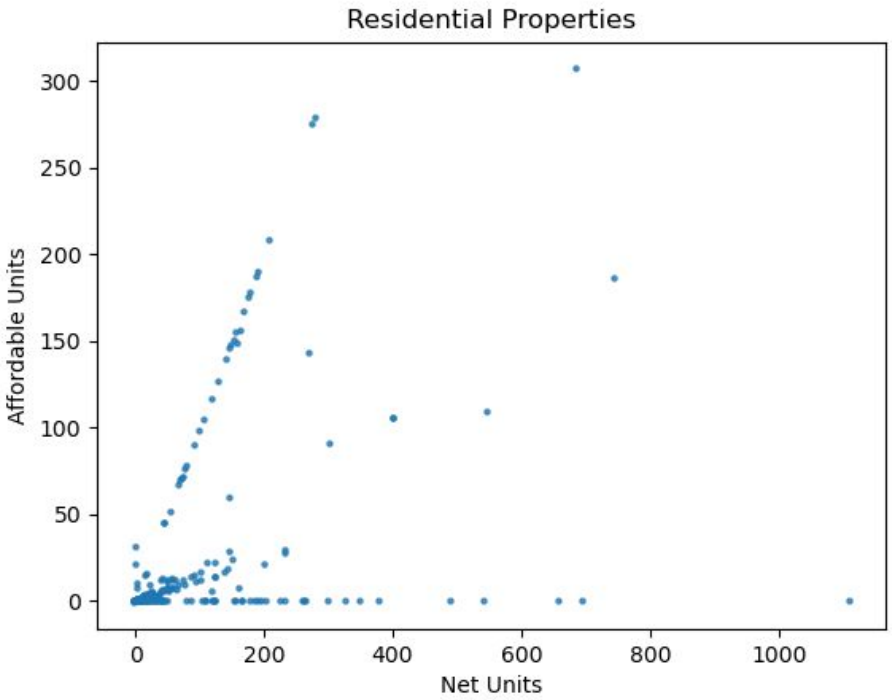




High Proportion Affordable

- Clear pattern of residential and mixed residential development projects which are 90-100% affordable.

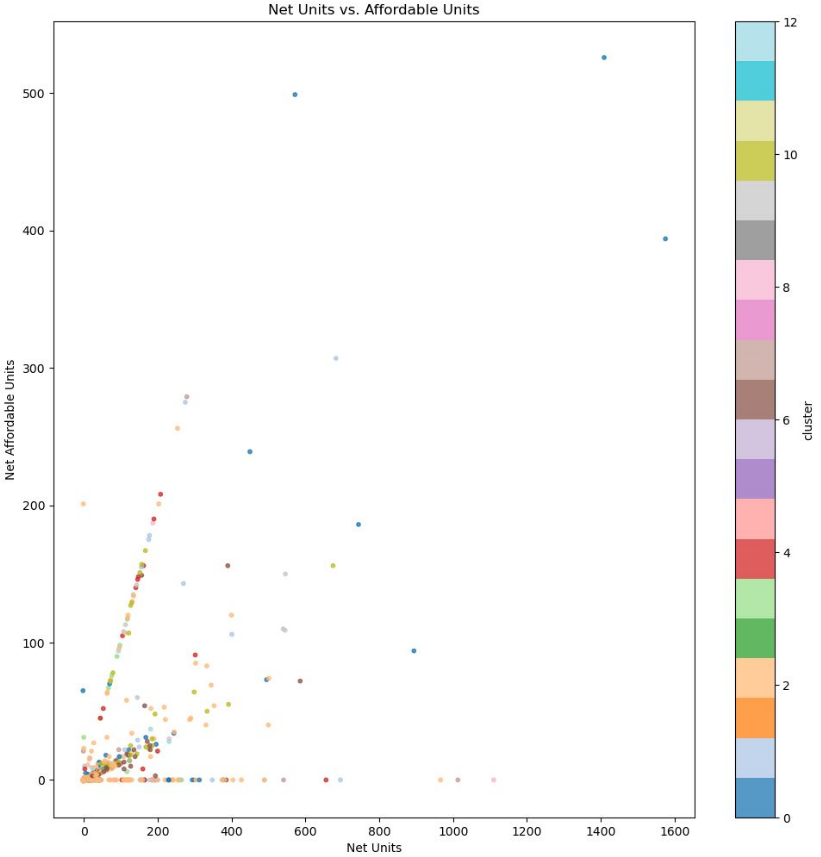
- Resemble clusters 9 and 10 on same scatter plot by cluster.



Proportion Affordable

- High proportion affordable reside mostly in clusters 9 and 10.

- Big impact outliers reside in cluster 5.



Clustering Model Finding & Recommendation

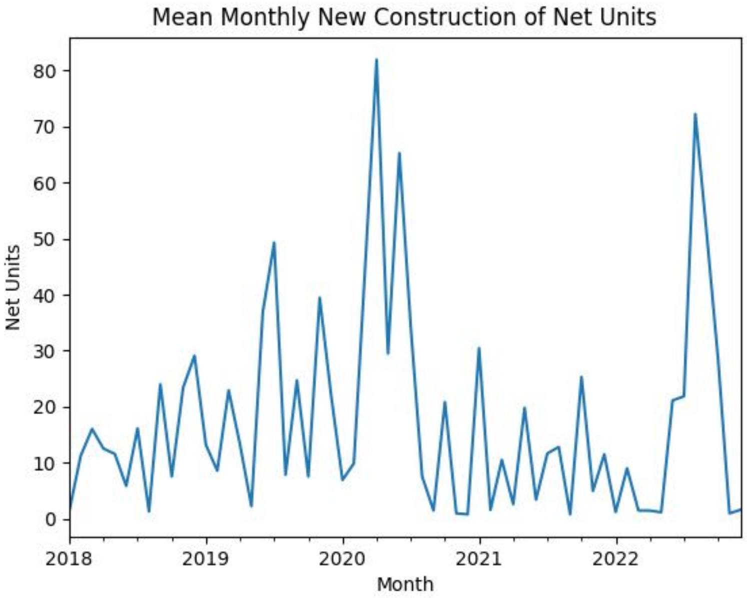
- FINDING: Our clustering model did not have a high enough silhouette score or show us clear geographic or housing distinctions in the clusters. Our EDA did uncover that the large project outliers with high net affordable units were all located in cluster 5 and that the highest proportion affordable to total net units occurred in clusters 9 and 10.

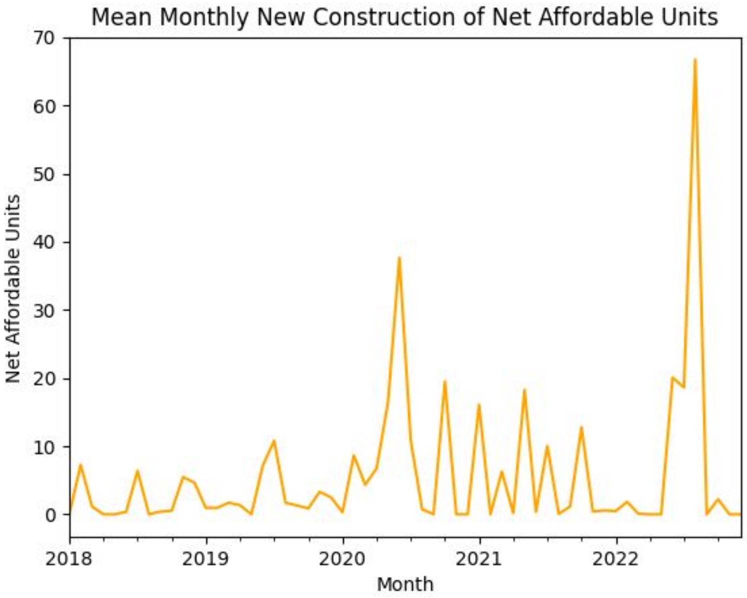
- RECOMMENDATION: Staff in the MOHCD more familiar with the specific development projects in these three clusters should gather more qualitative data about what these projects have in common.

Time Series Model Methodology

- Focus on forecasting net units and net affordable units.
- Two different time series models for each:
  - Mean monthly new construction
  - Construction in the pipeline quarterly
- Establish naive last or historical mean null model.
- Iterate through simple exponential smoothing, Holt Winters, and ARIMA models.
- Train vector autoregression for net units and net affordable units, if appropriate.

Mean Monthly New Construction



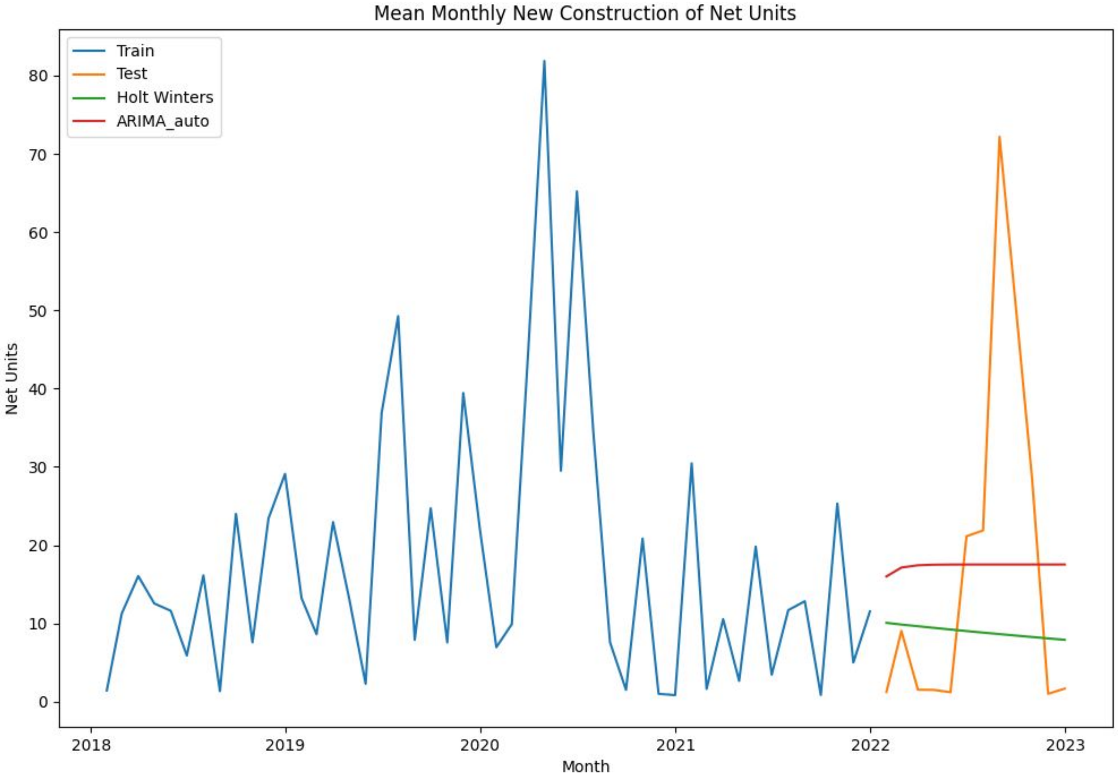


Time Series Net Units

- Null Baseline RMSE: ~22.15 net units

- ARIMA RMSE: ~22.04 net units
  - One lag
  - Zero differencing

Time Series Net Units

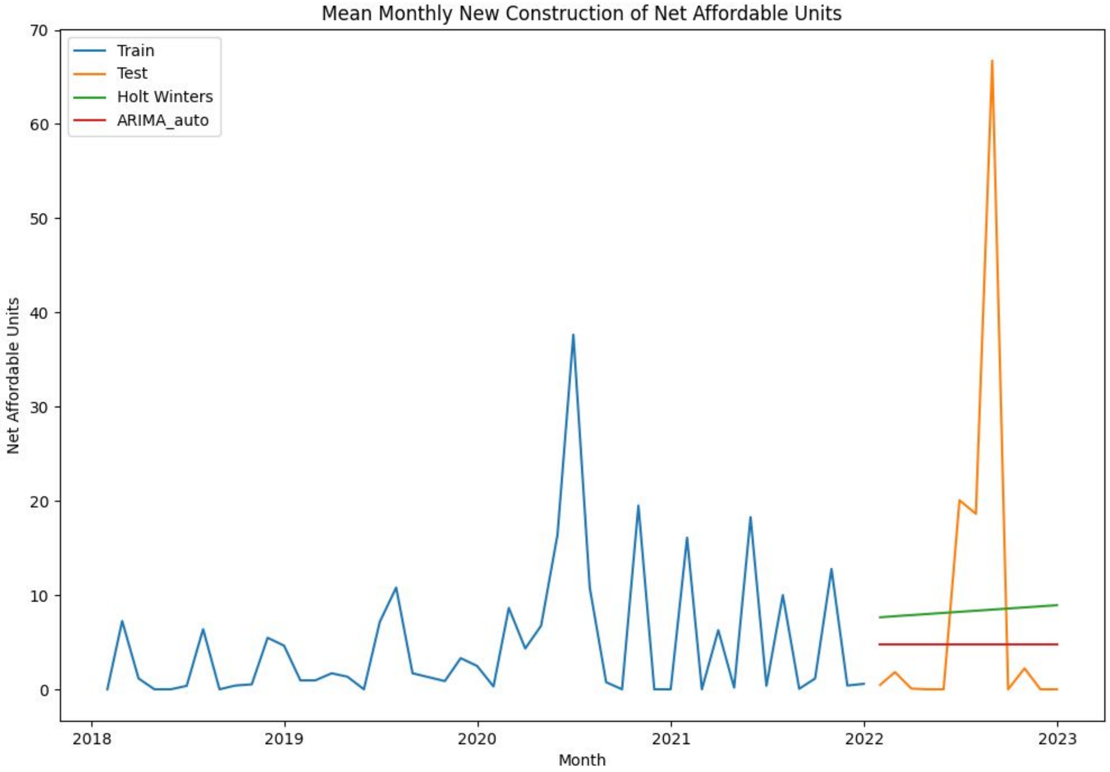


Time Series Net Affordable Units

- Null Baseline RMSE: ~19.22 net affordable units

- Holt Winters RMSE: ~18.69 net affordable units
  - More recent observation may have slightly higher impact on predicting future but barely outperforms historic mean.

Time Series Net Affordable Units



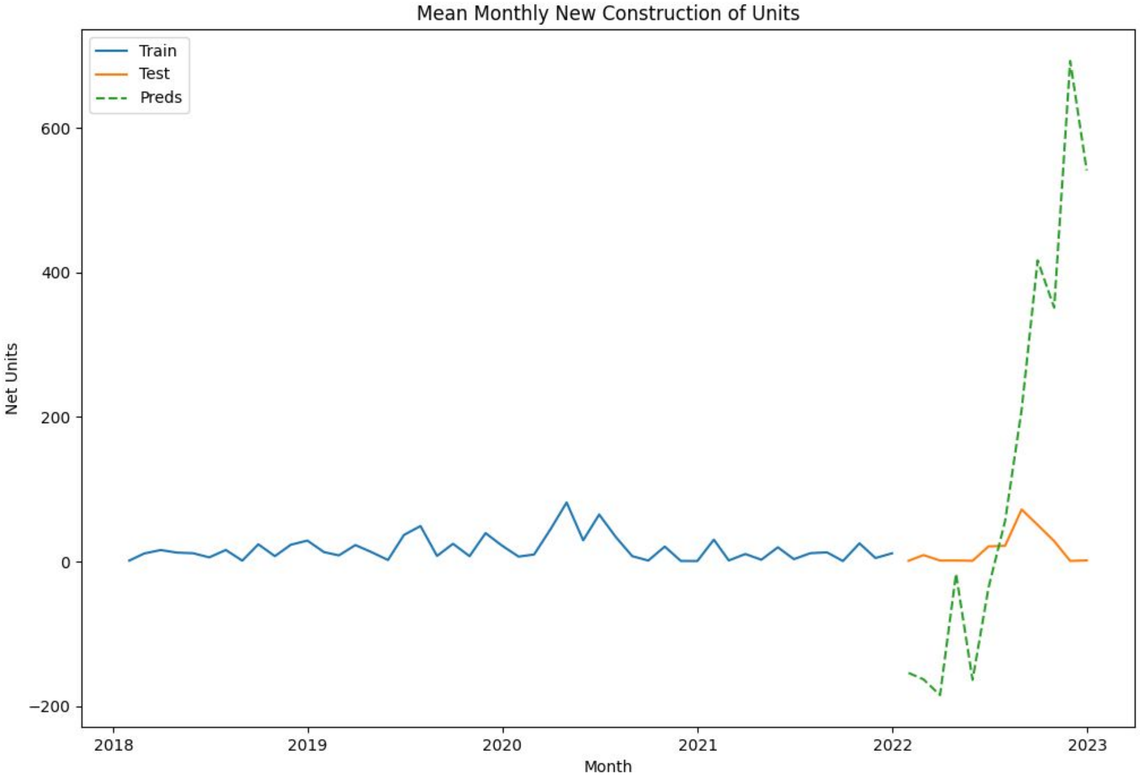
Vector Autoregression Time Series

- Augmented Dickey-Fuller Test showed stationarity for both net units and net affordable units mean monthly new construction.

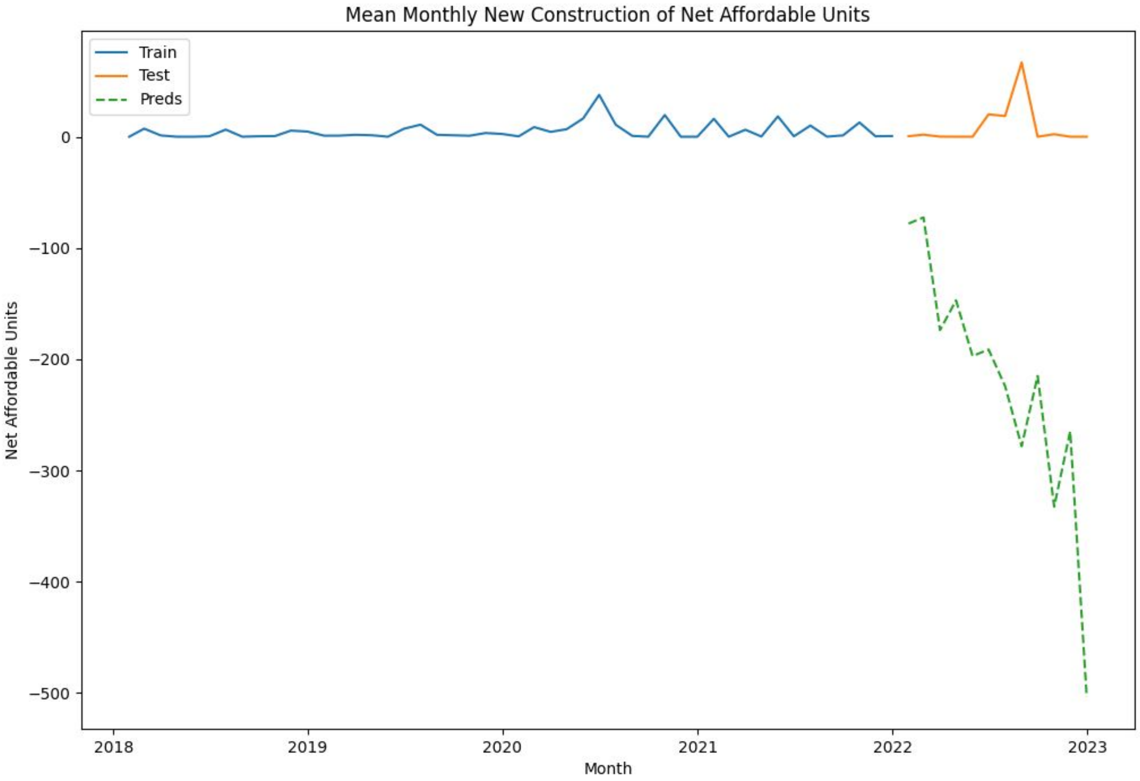
- Correlation between both variables of 0.62.

- Net Units RMSE: ~309 net units
- Net Affordable Units RMSE: ~259 net affordable units

Vector Autoregression Net Units



Vector Autoregression Net Affordable Units



Time Series Model Finding & Recommendation

- FINDING: Our multiple time series models all failed to meet the threshold of a serviceable prediction for future new construction or units in the construction pipeline. The best models performed almost as well as our null models.

- RECOMMENDATION: MOHCD cannot adequately predict future affordable housing. It is in the best interest of the Office to focus its policy efforts on increasing the number of new affordable housing projects pushed expediently through the pipeline as the historic mean for new construction remains low and well below the targets set by the city and the state.