APPLIED MACHINE LEARNING FOR THE ANALYSIS OF THE VIRGINIA HOUSING MARKET

A PREPRINT

Siddharth Nanda
Department of Computer Science
University of Virginia
sn9dq@virginia.edu

Dale Wilson
Department of Computer Science
University of Virginia
dsw6ru@virginia.edu

April 27, 2019

Abstract: In recent years, rising income and wealth inequality have become a topic of national debate prompting increasing questions about sustainable and affordable housing for the many people whose stagnant wages have not kept up with rising costs. Recent political events and increased participation in socioeconomic movements protesting the growing income disparity have had public policy officials and leaders in the tech industry unite to look for a solution. In our project, we attempt to investigate the underlying market factors that influence the pricing of houses in Virginia as well as discern critical differences in housing markets nationwide using several Machine Learning techniques.

Introduction: The search for affordable housing has remained at the forefront of issues faced by Virginians and the general American populace. This is evident after examining the effects of the latest American housing crisis as well as the birth and explosive growth of companies like Airbnb which seek to disrupt the state of the housing industry. Recent events such as the decision to move Amazon's second headquarters to Northern Virginia also bring into question the future stability of affordable housing in Virginia. The rippling effects of tax breaks and rising housing costs for zip-codes situated around corporate megaliths have been studied in areas like Silicon Valley which have been historically known for bleeding-edge technology, the effects have yet to be fully understood for an area such as Virginia. For an area so intertwined with government work due to proximity to D.C., will Virginia be able to sustain such movements? Can we identify zip-codes in Virginia that might experience crippling blows or booms to housing value?

Previous research, such as the study conducted by Park and Bae, have leveraged regression on a more granular scale with respect to housing (Park, Bae 2015). Park and Bae looked into forecasting housing prices for Fairfax County Virginia to determine how housing prices in Fairfax reflect trends in the US housing market as a whole. Our investigation differs from these as we set out to determine what factors influence housing price by zip code as opposed to within a given county of Virginia. We are interested in predicting changes in housing value for zip codes, then determining if these values can be explained by multi-county or state-wide events.

Other researchers sought to track segmentation of the housing market into sub-markets by using neural networks to evaluate a combination of socioeconomic, physical and location relevant data. Kauko, Hooimeijer, and Hakfoort applied such techniques to housing data in the city of Helsinki, but our investigation seeks to apply a clustering methodology to zip codes within an entire state (Kauko et al., 2010). Both studies seek to uncover dimensions of housing sub-markets through evaluating patterns in our respective datasets. We opted to use K-means to determine if interesting aspects of Virginia's housing market may arise, such as differences between zip codes located in northern and southern Virginia, or areas around cities like Charlottesville, Richmond, and Lynchburg as opposed to differences within one city.

In this project, we analyzed aggregated housing market and economic data (collected by Zillow's Economic Research Team) with Machine Learning techniques (specifically Regression and Clustering) in order to draw conclusions about selling prices of housing in Virginia, the similarity of the Virginia housing market and the national market, and possible reflections of recent socioeconomic movements in Virginia in housing data. This is a direct application problem, as it allows us to present conclusions that would be immediately of use to individuals across socioeconomic groups as they hope to build themselves and their families a better future as well as to legislators and students who could use the conclusions to direct policy and apply similar methodologies to other time series problems.

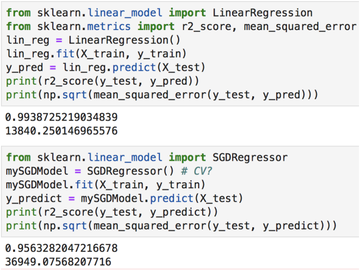
Method: In this experiment, we utilized several variants of Linear Regression in order to predict the Zillow Home Value Index (ZHVI) of Virginia Homes and used K-Means clustering to expose the underlying structure of the Virginia Housing Market. Data was initially acquired from the Zillow Research Data web-page and subsequently preprocessed using the pandas and numpy packages. Feature and dataset exploration was conducted in order to find the number of features, feature names, feature types, and possible correlations between features. The dataset contained 14 features: 'Date', 'RegionID', 'RegionName', 'State', 'Metro', 'County', 'City', 'SizeRank', 'Zhvi', 'MoM', 'QoQ', 'YoY', '5Year', '10Year', 'PeakMonth', 'PeakQuarter', 'PeakZHVI', 'PctFallFromPeak', and 'LastTimeAtCurrZHVI'. The 'State', 'Metro', 'County', 'City', 'Date', 'PeakMonth', 'PeakQuarter', and 'LastTimeAtCurrZHVI' features were categorical, while the remaining features were numerical. For both tasks, all rows of Virginia data with missing values were dropped leaving 363 rows of complete Virginia data for analysis. Exploration of statistics, histogram of features, and a scatter matrix provided us with some initial context for patterns in the data that could be exposed through clustering. Data was split into train and test sets and due to the unreliability in encoding some features in the test set due to the uniqueness of some values, the 'County', 'City', 'PeakMonth', 'Metro', 'PeakQuarter', 'LastTimeAtCurrZHVI' columns were dropped for the Regression models. Data was then transformed using Column Transformer, Pipeline, OneHotEncoder, and StandardScaler before the construction of the models. Generalized Linear Models were imported from scikit-learn as was the K-Means clustering algorithm for use in data analysis. Several Linear Regression models were constructed using solely Virginia housing data for the purpose of predicting the ZHVI label and a second set of Linear Regression models were constructed using data from the entire country for testing on Virginia Data in order to assess the ability of the model to predict Virginia housing prices based on national trends. K-Means clustering was executed on Virginia housing data and several graphics were constructed in order to analyze the traits by which the homes in each cluster were grouped. The specifics of the Linear Regression and K-Means experiments are detailed in the following section.

Experiments: In our experiments, we utilized various Linear Models from sklearn in order to find the best performing regressor for the ZHVI of Virginia Homes. We trained and tested six models and employed cross-validation for the Lasso (L-1 regularized), Ridge (L-2 regularized), and Elastic Net (L-1 and L-2 regularized) models. From best to worst in terms of Root Mean Square Error of ZHVI (our performance metric for regression), the models constructed were: the Lasso regressor, the Ridge regressor, the Linear regressor, the Linear regressor trained on national data, the Elastic Net regressor, and the SGD regressor. Linear models were found to fit the data extremely well, with the worst R squared value (SGD) being slightly over 0.95 and the best RMSE (Lasso) at only 13840. The Linear regressor trained on the national data also performed extremely well yielding an RMSE of only 15814 when tested on Virginia data. The parameters and RMSE of the Lasso regression indicate that L-1 normalization suits the model and feature selection/sparsity are critical factors in the performance of the regressors. The results of this portion of our experiment are viewable in Figures 1-5 of the subsequent section.

For the clustering portion of our experiments, we utilized the K-Means algorithm from sklearn in order to perform our analysis. After including got previously dropped features from the regression portion of our experiment, three clusters were generated using K-Means. These clusters were then graphed over 5 and 10 year changes in ZHVI vs ZHVI as well as Quarter-over-Quarter, Month-over-Month, and Year-over-Year changes in ZHVI vs ZHVI (as seen in Figure 6). We also graphed the clusters over plots of ZHVI vs ZHVI, Peak ZHVI vs ZHVI, and Percentage Fall from Peak ZHVI vs ZHVI (also seen in Figure 6). We decided to use these plots as these are a useful tool for showing the deviation in growth that homes in different regions have seen since the end of the last large recession (which greatly impacted the housing market) and easily demonstrate how this discrepancy is both a large problem to a population facing rising income inequality.

Results:

Figure 1: R Squared and Root Mean Square Error of Linear and Stochastic Gradient Descent Regressors for the ZHVI of Virginia Homes



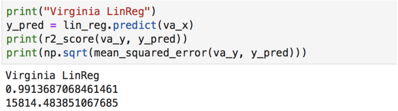
```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score, mean_squared_error
lin_reg = LinearRegression()
lin_reg.fit(X_train, y_train)
y_pred = lin_reg.predict(X_test)
print(r2_score(y_test, y_pred))
print(np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
0.9938725219034839
13840.250146965576
```

```
from sklearn.linear_model import SGDRegressor
mySGDModel = SGDRegressor() # CV?
mySGDModel.fit(X_train, y_train)
y_predict = mySGDModel.predict(X_test)
print(r2_score(y_test, y_predict))
print(np.sqrt(mean_squared_error(y_test, y_predict)))
```

```
0.9563282047216678
36949.07568207716
```

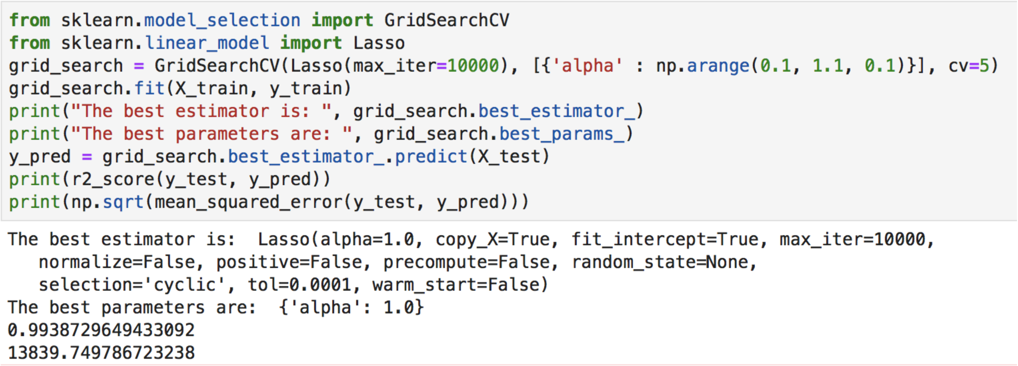
Figure 2: R Squared and Root Mean Square Error of Linear Regression Model Trained on National Data for the ZHVI of Virginia Homes



```
print("Virginia LinReg")
y_pred = lin_reg.predict(va_x)
print(r2_score(va_y, y_pred))
print(np.sqrt(mean_squared_error(va_y, y_pred)))
```

```
Virginia LinReg
0.9913687068461461
15814.483851067685
```

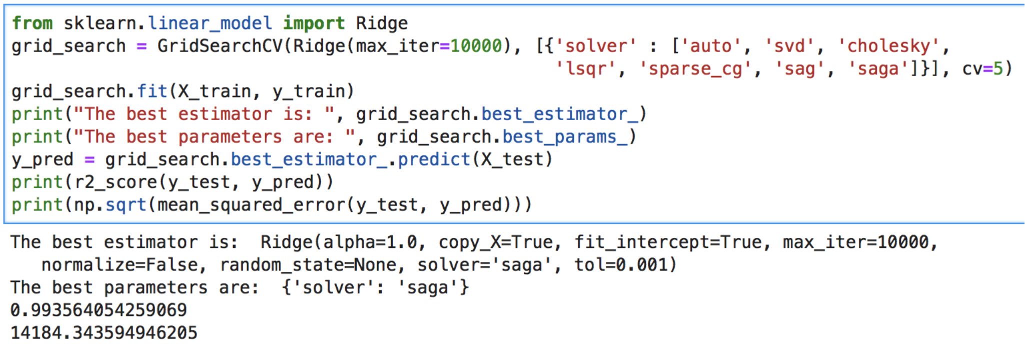
Figure 3: R Squared and Root Mean Square Error of Optimal Lasso Regressor for the ZHVI of Virginia Homes



```
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Lasso
grid_search = GridSearchCV(Lasso(max_iter=10000), [{'alpha' : np.arange(0.1, 1.1, 0.1)}], cv=5)
grid_search.fit(X_train, y_train)
print("The best estimator is: ", grid_search.best_estimator_)
print("The best parameters are: ", grid_search.best_params_)
y_pred = grid_search.best_estimator_.predict(X_test)
print(r2_score(y_test, y_pred))
print(np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
The best estimator is:  Lasso(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=10000,
   normalize=False, positive=False, precompute=False, random_state=None,
   selection='cyclic', tol=0.0001, warm_start=False)
The best parameters are:  {'alpha': 1.0}
0.9938729649433092
13839.749786723238
```

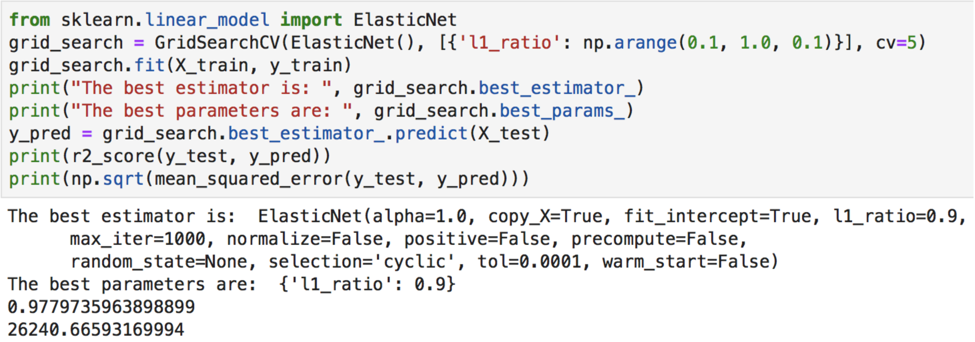
Figure 4: R Squared and Root Mean Square Error of Optimal Ridge Regressor for the ZHVI of Virginia Homes



```
from sklearn.linear_model import Ridge
grid_search = GridSearchCV(Ridge(max_iter=10000), [{'solver' : ['auto', 'svd', 'cholesky',
                                                    'lsqr', 'sparse_cg', 'sag', 'saga']}], cv=5)
grid_search.fit(X_train, y_train)
print("The best estimator is: ", grid_search.best_estimator_)
print("The best parameters are: ", grid_search.best_params_)
y_pred = grid_search.best_estimator_.predict(X_test)
print(r2_score(y_test, y_pred))
print(np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
The best estimator is:  Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=10000,
   normalize=False, random_state=None, solver='saga', tol=0.001)
The best parameters are:  {'solver': 'saga'}
0.993564054259069
14184.343594946205
```

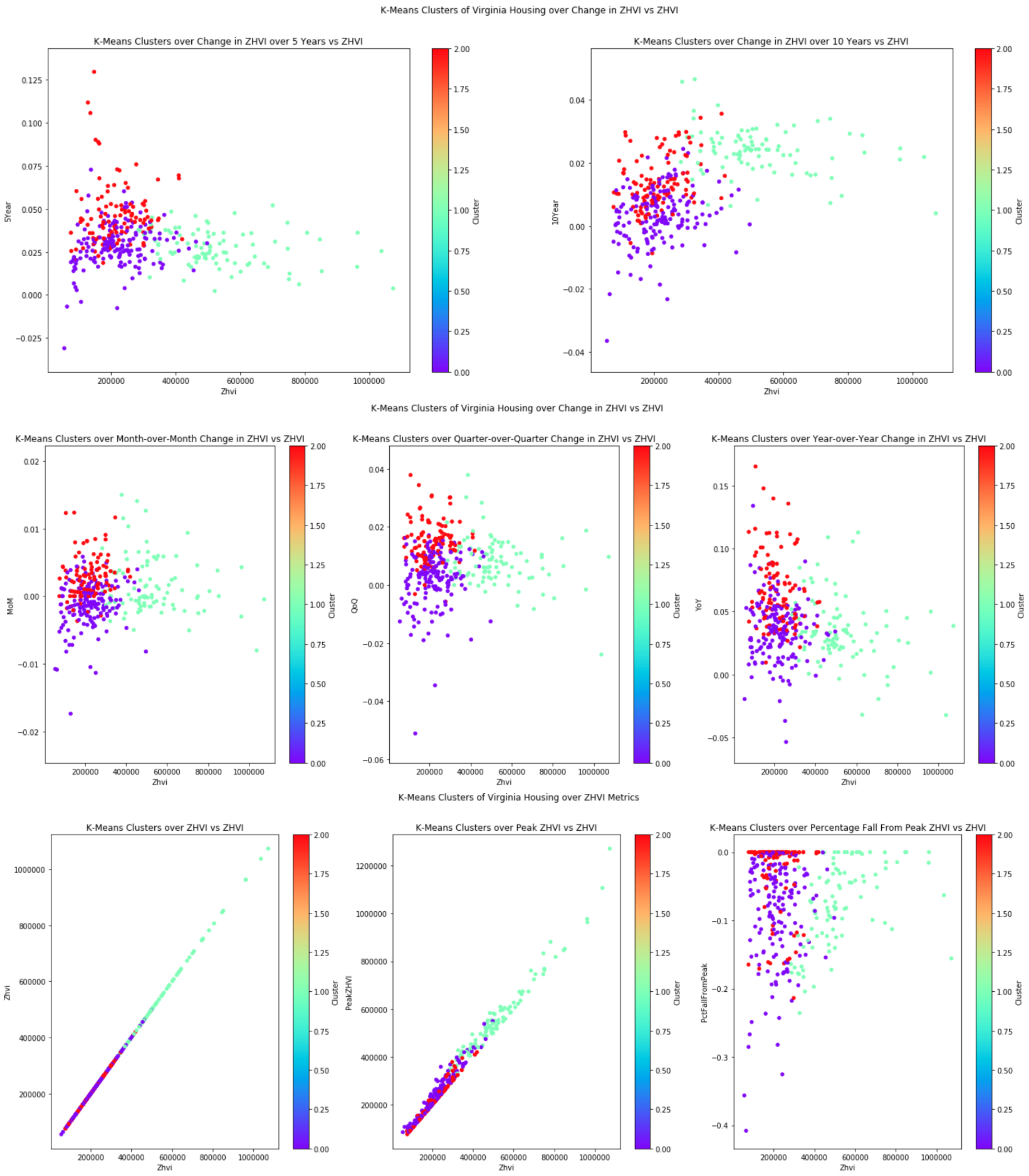
Figure 5: R Squared and Root Mean Square Error of Optimal Elastic Net Regressor for the ZHVI of Virginia Homes



```
from sklearn.linear_model import ElasticNet
grid_search = GridSearchCV(ElasticNet(), [{'l1_ratio': np.arange(0.1, 1.0, 0.1)}], cv=5)
grid_search.fit(X_train, y_train)
print("The best estimator is: ", grid_search.best_estimator_)
print("The best parameters are: ", grid_search.best_params_)
y_pred = grid_search.best_estimator_.predict(X_test)
print(r2_score(y_test, y_pred))
print(np.sqrt(mean_squared_error(y_test, y_pred)))
```

```
The best estimator is:  ElasticNet(alpha=1.0, copy_X=True, fit_intercept=True, l1_ratio=0.9,
      max_iter=1000, normalize=False, positive=False, precompute=False,
      random_state=None, selection='cyclic', tol=0.0001, warm_start=False)
The best parameters are:  {'l1_ratio': 0.9}
0.9779735963898899
26240.66593169994
```

Figure 6: Visualization of Results of K-Means Clustering Across Relevant Features



Conclusion: The Zillow housing data successfully facilitates the creation of varied models to extract underlying patterns both in the national housing market as well as the housing market across Virginia. A variety of factors could help explain the linear nature of housing market data seen throughout our investigation. The dramatic drop in interest rates as well as the Federal Reserve's use of policies such as Quantitative Easing since the last recession could explain the existence of an increasing linear relationship for predicting housing prices. An influx of new jobs due to expanding corporate infrastructure, such as movements from Amazon, Parsons, and Google to hotbeds like Northern Virginia can also explain an increase in housing value over time due to increased demand for local properties. In addition, the use of L-1 norm based regularization greatly improved the performance of the model which suggests that the scope of causes for this effect can be narrowed.

The effectiveness of our national regressor when applied to Virginian data points suggests differences between the national housing market and state housing markets may be fewer than originally thought. This could mean differences in housing values across the nation are influenced by features similar to those influencing housing values within any given state. This brings into question the extent to which housing value is affected by more local sociopolitical events as opposed to nation-wide changes and signals that while certain Virginia specific factors are certainly critical to the rising price of homes in the state: wider trends must be carefully considered in order to provide a reasonable solution to this problem.

We were able to create a regressor that accurately predicts the value of properties based on percent changes in price over time for many time intervals. Extremely high R-squared values for most all of the linear and optimized linear regressions further validate the existence of strong linear relationships within the dataset. Our linear regression's acceptable mean squared error of 13,383 dollars is excellent when considering housing values varied widely from 99,200 dollars to 1,036,700 dollars and the standard deviation of prices was close to 174,000 dollars. This could specifically be used to help explain and expose the effects of socioeconomic events on housing markets in Virginia and predict how housing values might change in light of future socioeconomic events: greatly strengthening the case that actions must be taken to ensure affordable housing for those who will need it in the future.

Clustering seemed to segment the housing market into three distinguishable sets: higher housing value zip codes with historically increasing housing values (green points), lower housing value zip codes with historically increasing housing values (red points), and lower housing value zip codes with historically decreasing housing values (purple points). It is important note how many data points located in counties like Fairfax, Loudoun, and Prince William were clustered to the green centroid while data points located around Richmond, Charlottesville, and Virginia Beach were clustered to the purple centroid (the former of which are booming economically, while the latter have faced some struggles). The red centroid appears to have less definite geographical markers, zip-codes for this cluster were well distributed throughout Virginia. Further analysis within these specific zip-codes could yield meaningful insights as to what factors can result in housing value increasing and decreasing over time on a larger geographical scale. In the future, this work could be used to drive further affordable housing research, inform policy initiatives to help counteract migration-based wealth disparity and gentrification, and advise investors interested in entering real-estate and housing markets.

References:

1. Jonshon, Michael (2018). A Quick Exploration of Virginia Time Series.
2. Kauko, T., Hooimeijer, P., Hakfoort, J. (2002). Capturing Housing Market Segmentation: An Alternative Approach based on Neural Network Modelling. Housing Studies, 17(6), 875-894.
3. Park, B., Bae, J. K. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Expert Systems with Applications, 42(6), 2928-2934.