Rotman

OUTLINE

Executive Summary

The following report investigates the use of clustering and prediction models to estimate the median annual income of communities based on available demographic data for Statistics Canada census tracts. With prediction models deployed on different established clusters, concentrations of low-income and high-income areas can be identified in order to aid in urban planning design and governmental policy initiatives to reduce economic segregation and inequality.

The census tract records were clustered into 3 groups based on demographic commonalities by our k-means model. These clusters were identified as Downtown, Suburb, and Midtown. Cluster 1 Downtown had the highest proportion of apartments as dwellings and the highest percentage of renters as residents. Cluster 2 Suburb had the highest proportion of houses and the highest proportion of owners. Cluster 3 Midtown had an even mix of houses and apartments as well as owners and renters.

Segmentation models exhibited the lowest test error rates when compared to the proposed global models. Therefore, KNN, Random Forest, and Regression Trees were respectively fitted on Cluster 1, Cluster 2 and Cluster 3 as the best models to predict median household income per year for the provided census tracts.

Upon investigation, the percentage of tenure type was observed as the premier feature in predicting median household income. This can be intuitively explained as purchasing property requires a large sum of upfront cash. This can be a reasonable explanation for the observations of high median household incomes in areas with high percentages of owners as residents.

I. Introduction

The emergence of agriculture has diverted humans from traditional nomadic practices into farming societies. The Neolithic Revolution and the transition in lifestyle away from hunting and gathering have allowed humans to establish permanent settlements. This newfound ability to grow and maintain a source of food has allowed for the capability to sustain a larger population of humans than ever before. This in the combination with permanent settlements has led to the development of cities and dense living population centers. Continuing into the modern era, globalization has led humans into a period of wealth inequality among the populations. This economic disparity has many significant consequences including the manifestation of socioeconomic segregation of metropolitan regions and neighbourhoods. The rise of residential sorting by income class and the increased concentration in poverty among communities is a byproduct of increased wealth disparity. Spatial segregation by income class can exacerbate social and economic inequalities, which underpin health inequities. Social and ethnic heterogeneity in communities is the desired outcome of urban planning. Careful design of urban infrastructures ensures that all residents have more equitable access to livelihood opportunities, community and recreational facilities, social services, and education.

The rise of homogeneity among communities, and its associated features, can be studied to predict the income of those regions. The identification and analysis of demographic characteristics can be used to predict the median annual household income for census tracts. As defined by Statistics Canada, Census Tracts (CTs) are small and relatively stable geographic areas that usually have a population between 2,500 to 7,500 persons, with a preferred average of 5,000. At the time of creation by StatCan, the delineation of CTs should be as homogeneous as possible in socioeconomic characteristics such as economic status and social living conditions. Census tracts are located in census metropolitan areas (CMAs) and in census agglomerations (CAs) that had a core population of 50,000 or more in the previous census.

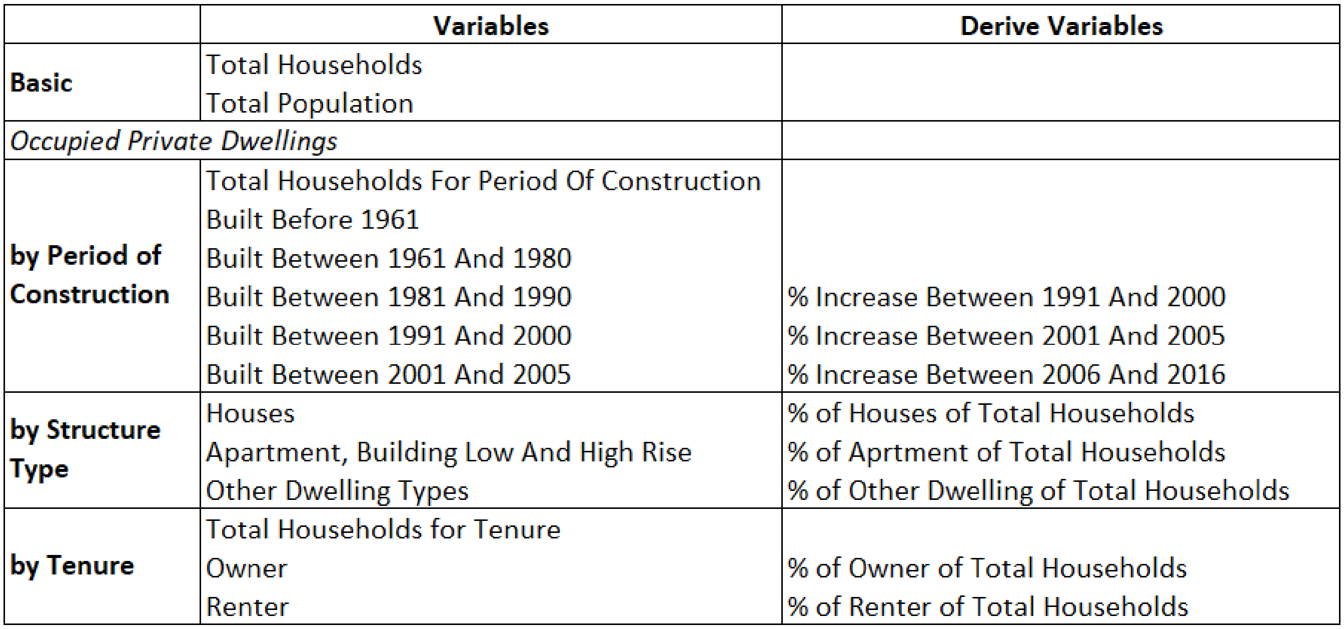
Utilizing demographic characteristics of publicly accessible CT records provided by Statistics Canada, we have developed a model to predict missing median income data for 721 records. This investigation aims to utilize predicted median income for the purposes of urban planning and better-designed public policy to increase equitable access to the benefits of urban life by identifying and reducing spatial economic segregation.

II. Data Preprocessing

The census Canada training dataset contains fourteen attributes (Table 1) and one output variable *Median Household Income* with 5000 records. Three additional columns were derived from the *occupied private dwellings by the period of construction* variables to assess the growth rate in household construction for each census tract. Only the period from 1991 to 2016 was included in this analysis since the data prior to 1991 is outdated and is not helpful for this analysis. In order to compare census tracts of different sizes more effectively, the *occupied private dwellings by structure type* and *occupied private dwellings by tenure type* variables were converted to the percentage of total households for each census tract. In addition, 14 outlier records were dropped as their value for the *% Owner of Total Household* and *% Renter of Total Household* were greater than 1. The duplicate variables for *Total Households* that were eliminated are *Total Households For Period Of Construction* and *Total Households for Tenure*.

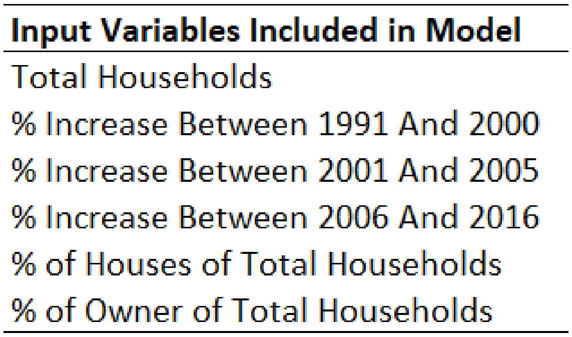
Moreover, correlation analysis was performed on all remaining variables to further select the list of significant input variables to include in the model. The correlation matrix graph and scatter plot in Figure 1 and 2 display that *Total Households* and *Total Population* are highly correlated, with a correlation of 0.8975. Additionally, as shown in Figure 3 and 4, there is a negative linear relationship between *% Houses of Total Household* and *% Apartment of Total Household* as well as *% Owner of Total Household* and *% Renter of Total Household* (-0.97 and -0.91, respectively). As listed in Table 2, there are six input variables with 4,986 records in the training dataset and six input variables with 721 records in the test dataset were retained after highly correlated and duplicate variables were removed.

Table 1: Original Variables and Derive Variables



	Variables	**Derive Variables**
Basic	Total Households Total Population	
Occupied Private Dwellings		
by Period of Construction	Total Households For Period Of Construction	
	Built Before 1961	
	Built Between 1961 And 1980	
	Built Between 1981 And 1990	% Increase Between 1991 And 2000
	Built Between 1991 And 2000	% Increase Between 2001 And 2005
	Built Between 2001 And 2005	% Increase Between 2006 And 2016
by Structure Type	Houses	% of Houses of Total Households
	Apartment, Building Low And High Rise	% of Aprtment of Total Households
	Other Dwelling Types	% of Other Dwelling of Total Households
by Tenure	Total Households for Tenure	
	Owner	% of Owner of Total Households
	Renter	% of Renter of Total Households

Table 2: Input Variables Included in Model



Input Variables Included in Model
Total Households
% Increase Between 1991 And 2000
% Increase Between 2001 And 2005
% Increase Between 2006 And 2016
% of Houses of Total Households
% of Owner of Total Households

III. Clustering Model

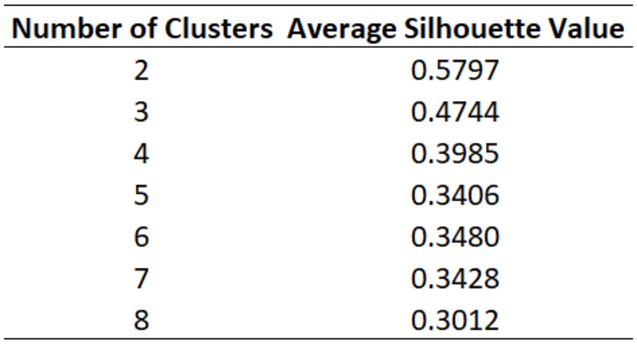
Elbow Method

To determine the optimal number of clusters to use in the clustering model. Two methodologies have been employed in this analysis. The Elbow method calculate the inertia which is the distance between each record and its centroid. The inertia is monotonically decreasing with increasing numbers of clusters and the optimal point of cluster is where an elbow occurs. The plot in Figure 5 exhibits a sharp decline up until k reaches 3 and thereafter decreases slowly. It seems like k=3 is the best choice. However, the elbow method only takes into account cluster cohesion. Thus, a silhouette analysis that gauges both cluster separation and cluster cohesion was preferable.

Silhouette Method

The Silhouette approach evaluates how well the cluster solution fits. The average silhouette value is from -1 to +1. A positive value indicates that the assignment is good, with higher values being preferable to lower values. In this analysis, a silhouette analysis was performed on clusters with a number from 2 to 8. As shown in Table 3, the values of 2 and 3 for the number of clusters have the highest average silhouette values (0.5797 and 0.4777, respectively). The red dashed line in the silhouette plot (Figure 6 and 7) indicates that the silhouette coefficient value for each cluster is higher than the average silhouette value. It is also important to consider the thickness of the silhouette plot representing each cluster. The thickness of the bottom cluster is substantially greater than the top cluster in the plot with two clusters (Figure 6). Conversely, the thickness for the plot with three clusters (Figure 7) is more comparable in size, therefore the best number of clusters is 3.

Table 3: Average Silhouette Value



Number of Clusters	Average Silhouette Value
2	0.5797
3	0.4744
4	0.3985
5	0.3406
6	0.3480
7	0.3428
8	0.3012

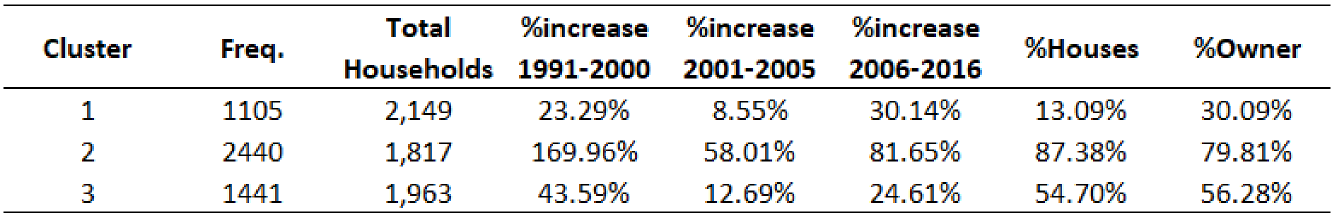
K-Means

The variable *total household* has been scaled using the Min-Max method since the majority of the other input variables have percentage values between 0 and 1. The census data are divided into three clusters by the K-Means clustering model. Out of a total of 4,986 census tract records, cluster 2 contains the most records (2,440), while clusters 1 and 3 have medium-sized records (1,105 and 1441, respectively). According to the aggregate mean variables by cluster (Table 4), cluster 2 has seen a significant increase in household construction during the past 25 years (from 58.01% to 169.96%), compared to clusters 1 and 3, which have experienced a moderate growth rate.

Furthermore, as shown by Figure 8 and Figure 9, cluster 1 has a lowest percentage of houses and percentage of owners, which indicates that the majority of dwellings in cluster 1 are apartments, low- or high-rise buildings and most households are renters. In contrast, cluster 2 has the largest percentage of houses and owners (87.38% and 79.81% on average). In cluster 3, there are roughly half and half of houses and apartments, as well as owners and renters. Nevertheless, the histogram of total households by clusters (Figure 10) shows that there is only little difference between three clusters, indicating that this attribute cannot be utilized to distinguish between the clusters from one another. The algorithm is relying heavily on *the percentage of house* and *percentage of owner* variables.

The differences in target variable distribution among the clusters are shown by the histogram of median household income (Figure 11). It demonstrates that cluster 1, which comprises a high percentage of apartments and renters, is likely to have a lower median income than cluster 2, which contains a higher percentage of houses and owners, while cluster 3 is in the middle.

Table 4: Aggregate Mean Statistic by Cluster



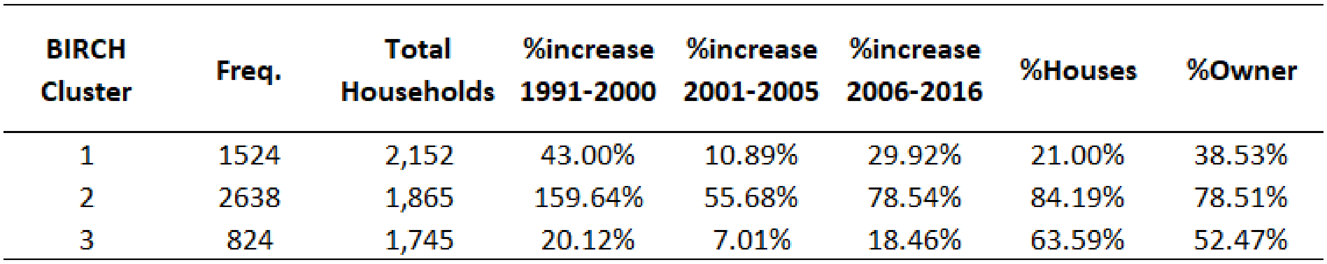
Cluster	Freq.	Total Households	%increase 1991-2000	%increase 2001-2005	%increase 2006-2016	%Houses	%Owner
1	1105	2,149	23.29%	8.55%	30.14%	13.09%	30.09%
2	2440	1,817	169.96%	58.01%	81.65%	87.38%	79.81%
3	1441	1,963	43.59%	12.69%	24.61%	54.70%	56.28%

BIRCH

The Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering model was compared to the K-Means model using the same number of clusters. The default branching factor and threshold of 0.1 are the parameters used in the BIRCH model. The BIRCH model produced similar results and treatment of the data. The size of clusters differs slightly from the K-Mean model. There are only 824 entries present in cluster 3 and more records in cluster 1 and cluster 2 (1524 and 2638, respectively).

According to aggregate statistics (Table 5), cluster 2 continues to have the highest growth rate in household construction, as well as the highest percentage of houses and apartments, while cluster 1 has a relatively low percentage of houses and owners and a moderate growth rate that is higher than cluster 3. The cluster 3 had on average 52% of owners and 62% of houses. However, Figure 12 and Figure 13 demonstrate that there is greater overlap across three clusters than K-Means on the important variables *%Houses* and *%Owner*, suggesting that K-Means more accurately distinguishes each cluster. Therefore, K-Means was selected as the clustering model to be employed in the segmentation modelling.

Table 5: Aggregate Mean Statistic by Cluster



BIRCH Cluster	Freq.	Total Households	%increase 1991-2000	%increase 2001-2005	%increase 2006-2016	%Houses	%Owner
1	1524	2,152	43.00%	10.89%	29.92%	21.00%	38.53%
2	2638	1,865	159.64%	55.68%	78.54%	84.19%	78.51%
3	824	1,745	20.12%	7.01%	18.46%	63.59%	52.47%

Clusters Profile

Based on the above analysis, the three clusters could be classified as follows:

Cluster 1: Downtown

Downtown, the inner city or the city center are terms used interchangeably that feature the highest urban density of these three classifications. Dwellings in these areas consist of apartments, low, and high-rise buildings.

Cluster 2: Suburbs

The suburbs are defined as an area that is primarily residential that is located on the periphery of a city. These communities are not as densely populated as an inner city but not as sparsely populated as a rural area. In Canada, the typical dwellings in these communities are houses rather than dense apartment units.

Cluster 3: Inner Suburb/Midtown

Closer to the inner city than the suburbs are the inner suburbs. The inner suburbs have a lower urban density than that of the inner city, but higher than that of a suburb. They tend to feature mixed-use developments, with dwellings typically comprising an even distribution of houses and apartments.

IV. Segmentation Modeling

Three candidate models were proposed as the predictive segmentation models, including KNN, Regression Tree, and Random Forest. After the preprocessed data was partitioned into 3 clusters using the optimal k-means model, three candidate prediction models were applied to these 3 clusters separately. The data was subsequently split into training and test sets, by the ratio of 75:25, in order to compare the test errors. Based on the comparison of mean absolute errors (MAEs), the best prediction model was selected for each cluster.

The total and segmented training set size and validation set size after partitioning and cluster were shown in Table 6.

Table 6: Training and Validation Set

	Training Set Size	Validation Set Size
Cluster 1	828	277
Cluster 2	1835	605
Cluster 3	1076	365
Total	3739	1247

KNN

K-Nearest Neighbor algorithm uses feature similarity to predict new data points. K-fold Cross-Validation with 5 folds was used to select the best number of neighbours. A KNN model was fit on the entire training dataset. The optimal number of neighbours selected for the global model by K-fold cross-validation was 26. The 3 other KNN models with 15, 24, and 28 neighbours were optimally fitted on Cluster 1, 2 and 3, respectively. These worked as the segmentation models for the three clusters.

The models were then used to make predictions on corresponding validation sets. Their training and test Mean Absolute Errors are listed in Table 7.

Table 7: Training and Test MAE of KNN Model

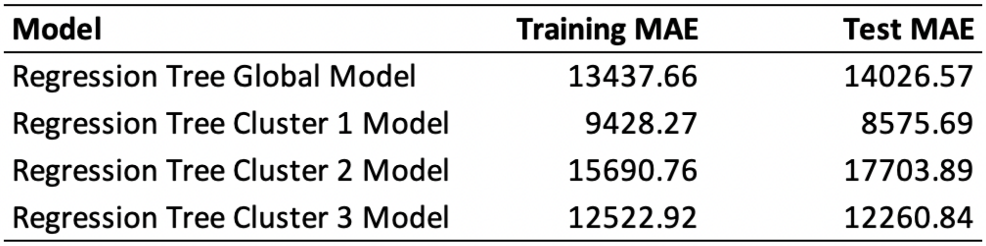
Model	Training MAE	Test MAE
KNN Global Model	13469.06	14484.45
KNN Cluster 1 Model	8645.57	7727.25
KNN Cluster 2 Model	15856.72	17351.2
KNN Cluster 3 Model	12955.44	14362.79

The average training MAE for 3 segmentation models was 12,485.91. This is lower than the training MAE for KNN global model (13,469.06). Similarly, the average test MAE for 3 segmentation models was 13,147.08 which was also lower than the test MAE for KNN global model (14,484.45). Thus, the global KNN model performed worse than the average of segmentation KNN models.

Regression Tree

The Regression Tree is one of the CART (Classification and Regression Tree) models, it recursively partitions data into binary groups according to feature thresholds, until it comes to a stop and produces leaf nodes. These leaf nodes are the predictions. The regression tree model was fit first on the entire training set. The 3 clusters were then fitted with the regression tree that was restricted to a minimum of 50 samples per leaf. Their training and test Mean Absolute Errors are listed in Table 8.

Table 8: Training and Test MAE of Regression Tree Model



Model	Training MAE	Test MAE
Regression Tree Global Model	13437.66	14026.57
Regression Tree Cluster 1 Model	9428.27	8575.69
Regression Tree Cluster 2 Model	15690.76	17703.89
Regression Tree Cluster 3 Model	12522.92	12260.84

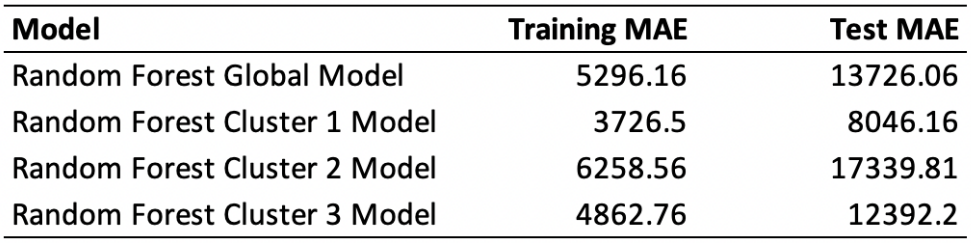
The average training MAE if applying Regression Tree model on 3 segments was 12,547.32. This was lower than the training MAE for the Regression Tree global model (13,437.66). Similarly, the average test MAE for 3 segmentation models is 12,846.81. This was also lower than the test MAE for the Regression Tree global model (14,026.57). It indicated the global regression tree model performed worse than the average of segmentation models.

Random Forest

Random forest is an ensemble learning model for classification and regression, by constructing a multitude of classification or regression trees, and for this case regression is used. Every individual tree will generate a prediction result, and the final result is the average prediction of the individual trees.

The random forest model was fit first on the entire training set. The 3 clusters were then fitted with the random forest tree that was restricted to 128 decision trees. Their training and test Mean Absolute Errors are listed in Table 9.

Table 9: Training and Test MAE of Random Forest Model



Model	Training MAE	Test MAE
Random Forest Global Model	5296.16	13726.06
Random Forest Cluster 1 Model	3726.5	8046.16
Random Forest Cluster 2 Model	6258.56	17339.81
Random Forest Cluster 3 Model	4862.76	12392.2

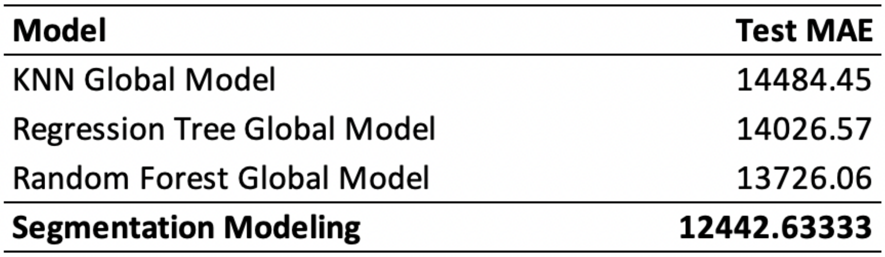
The average training MAE if applied to the Random Forest model on 3 segments was 4,979.27. This was lower than the training MAE for Random Forest global model (5,296.16). Similarly, the average test MAE for 3 segmentation models is 12,592.72. This was also lower than the test MAE for Random Forest global model (13,726.06). Similar to the two proposed models above, the global random forest model also performed worse than segmentation models.

Model Comparison & Choice

The best model for each cluster was selected based on the test MAE results. For cluster 1, the model with the best performance was KNN model, with lowest test MAE of 77,27.25. For cluster 2, the model with the best performance was Random Forest model, with lowest test MAE of 17,339.81. For cluster 3, the model with the best performance was the Regression Tree model, with lowest test MAE of 12,260.84. The visualization of test MAE comparison was shown in Figure 14.

Using the chosen segmentation models can lower the average test MAE to 12,442.63 when compared to using any one of the three global models directly, which has corresponding test MAEs of 13,726.06, 14,026.57 and 14,484.45, as shown in Table 10. As a result, segmentation modeling offers lower test MAE and better performance than proposed global models.

Table 10: Test MAE of 3 Global Models v.s. Average Test MAE of 3 Selected Models



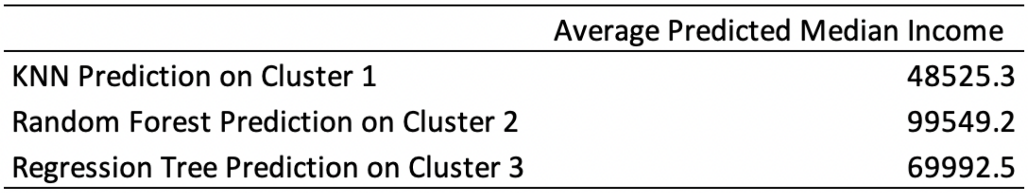
Model	Test MAE
KNN Global Model	14484.45
Regression Tree Global Model	14026.57
Random Forest Global Model	13726.06
Segmentation Modeling	**12442.63333**

Feature Importance

A feature importance analysis was run on cluster 2 and cluster 3 using random forest and regression tree models. For the KNN algorithm, feature importance is not defined. According to Figure 15 and 16, the graphs indicated that on both clusters, the proportion of owners was always the most important variable in predicting median household income, meaning that for Census Tract areas, higher the proportion of owners, higher the median household income would be.

Predictions on Test Data

Table 11: Predicted Median Income on Test Data



	Average Predicted Median Income
KNN Prediction on Cluster 1	48525.3
Random Forest Prediction on Cluster 2	99549.2
Regression Tree Prediction on Cluster 3	69992.5

According to the average of predicted median income for each cluster shown in Table 11 above, cluster 2 (suburb) has the highest average predicted median income around 9,9549.2, while cluster 1 (downtown) has the lowest average predicted median income around 4,8525.3.

V. Conclusion

In conclusion to what have been done in this study, an optimal K-means model was firstly used to cluster the census data into Downtown, Suburb, and Midtown, mainly based on the percentage of building structure (i.e., house or apartment), the percentage of tenure type (i.e., owner or renter), as well as the growth rate of household construction. Then, KNN, Random Forest, and Regression Tree were selected as best segmentation models to predict median household income respectively for Cluster 1 (Downtown), Cluster 2 (Suburb), and Cluster 3 (Midtown). In addition, it was found that the proportion of owners was the most important feature in predicting median household income of Census geographic units.

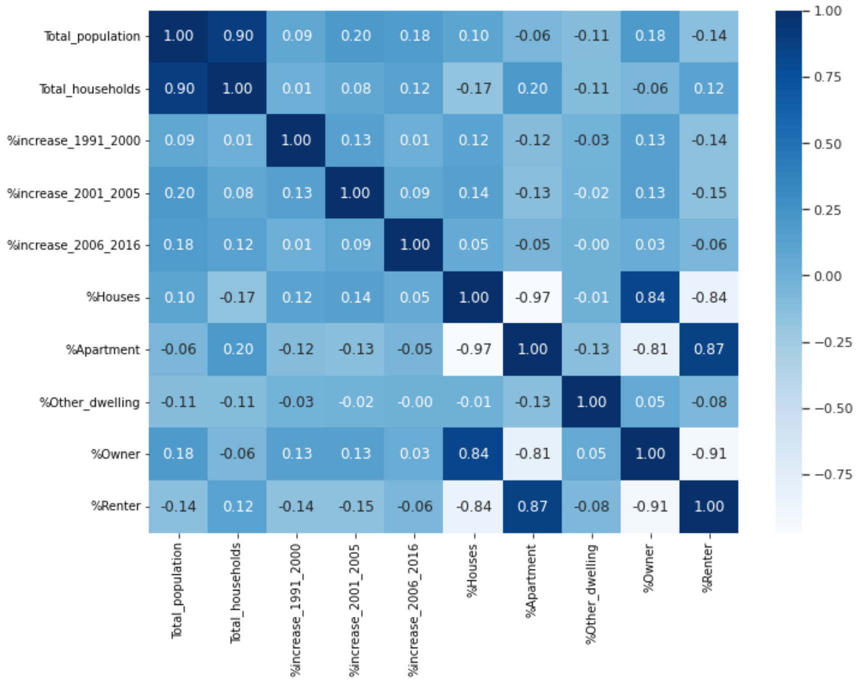
However, there exist limitations and thus some future steps to take regarding further study on median household income. First of all, there are only a limited number of input variables available for this study. More research can be done in the future on other possible factors associated with household income. Second, other useful models can be investigated to provide more choices for model comparison and improve model accuracy. Furthermore, data of more census tracts can possibly be explored or collected to improve the prediction accuracy.

VI. Recommendations

Residential clustering by income will always be an inevitable element of urban life. However, early identification and careful planning of regions with lower median income can greatly benefit the welfare of all citizens. Distinguishing the separate concentrations of low and high wealth in communities allow for the introduction of policy and zoning regulations between these areas to encourage equitable access to facilities, jobs, and services. The construction of schools and community centers between neighbourhoods with great wealth inequalities can aid in the social inclusion of families with diverse income levels. The ability to recognize areas of higher affluence enables affordable housing developments to be targeted to these areas to encourage income heterogeneity in these communities. Further, developing and building public infrastructure to support public transit in low-income areas can improve the socioeconomic status of those citizens and health outcomes.

Appendix

Figure 1: Correlation Matrix Heatmap



	Total_population	Total_households	%increase_1991_2000	%increase_2001_2005	%increase_2006_2016	%Houses	%Apartment	%Other_dwelling	%Owner	%Renter
Total_population	1.00	0.90	0.09	0.20	0.18	0.10	-0.06	-0.11	0.18	-0.14
Total_households	0.90	1.00	0.01	0.08	0.12	-0.17	0.20	-0.11	-0.06	0.12
%increase_1991_2000	0.09	0.01	1.00	0.13	0.01	0.12	-0.12	-0.03	0.13	-0.14
%increase_2001_2005	0.20	0.08	0.13	1.00	0.09	0.14	-0.13	-0.02	0.13	-0.15
%increase_2006_2016	0.18	0.12	0.01	0.09	1.00	0.05	-0.05	-0.00	0.03	-0.06
%Houses	0.10	-0.17	0.12	0.14	0.05	1.00	-0.97	-0.01	0.84	-0.84
%Apartment	-0.06	0.20	-0.12	-0.13	-0.05	-0.97	1.00	-0.13	-0.81	0.87
%Other_dwelling	-0.11	-0.11	-0.03	-0.02	-0.00	-0.01	-0.13	1.00	0.05	-0.08
%Owner	0.18	-0.06	0.13	0.13	0.03	0.84	-0.81	0.05	1.00	-0.91
%Renter	-0.14	0.12	-0.14	-0.15	-0.06	-0.84	0.87	-0.08	-0.91	1.00

Figure 2: Scatter Plot of Total Population ve Total Households

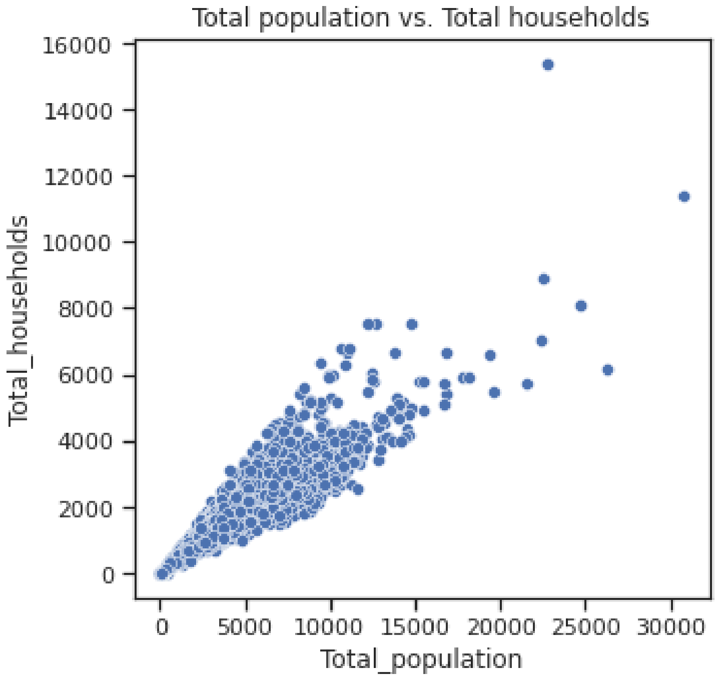


Figure 3: Scatter Plot of % Owner vs %Renter

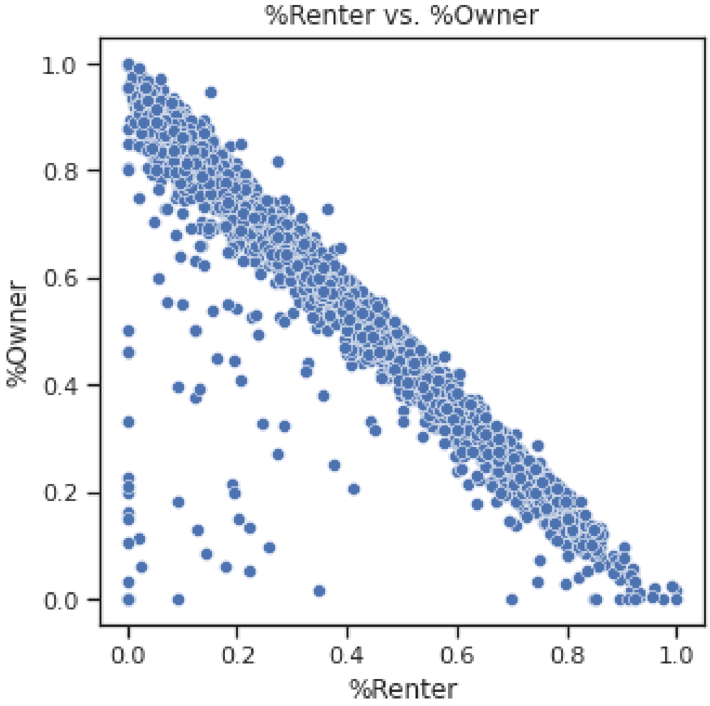


Figure 4: Scatter Plot of % Houses vs % Apartment

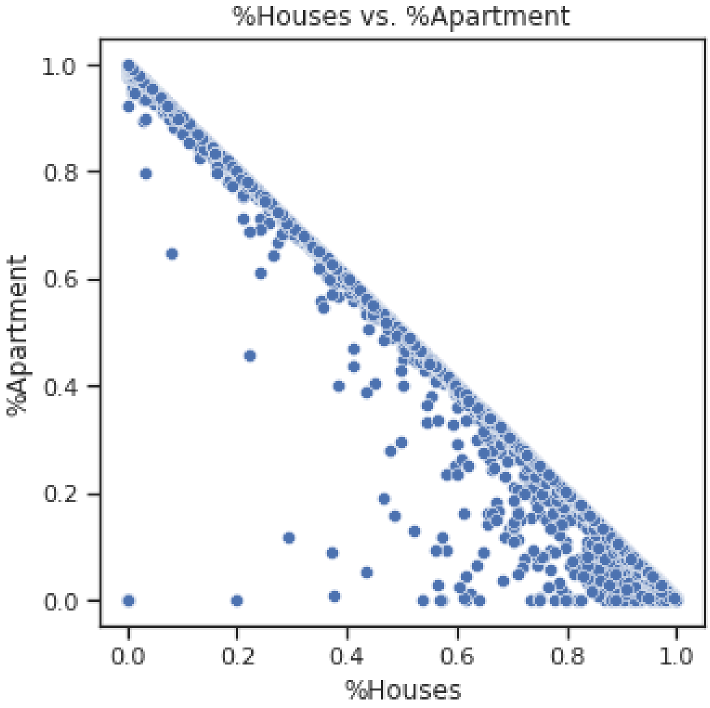


Figure 5: Elbow Plot

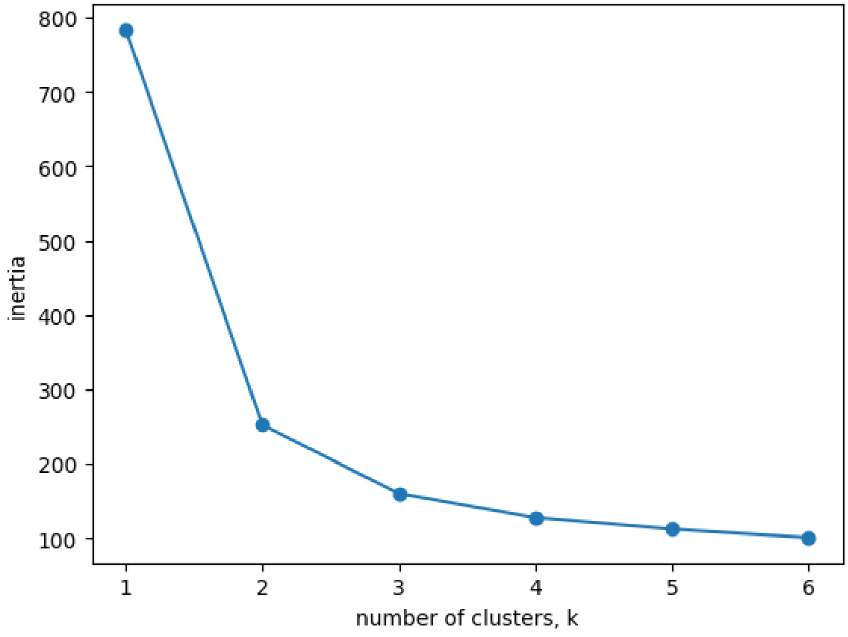


Figure 6: Silhouette Plot with n_clusters = 2

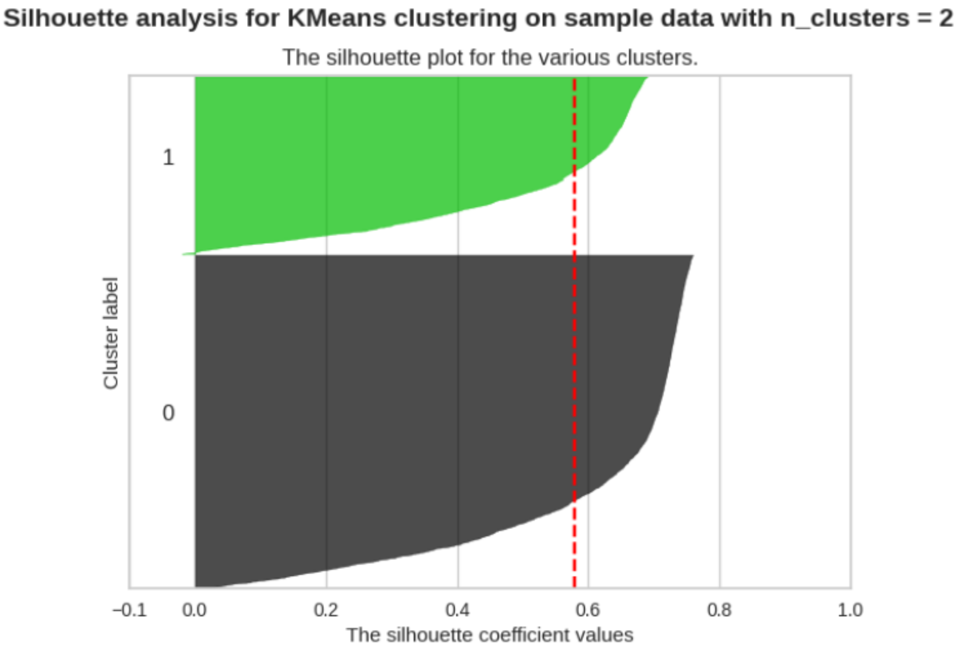


Figure 7: Silhouette Plot with n_clusters = 3

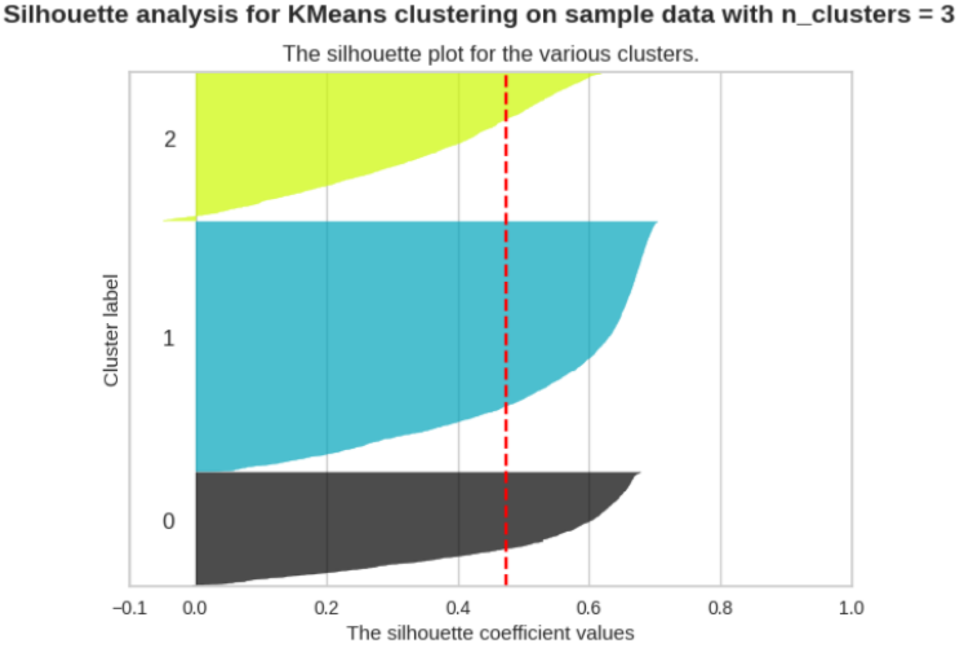


Figure 8: Boxplot of % of Houses of Total Households by Cluster

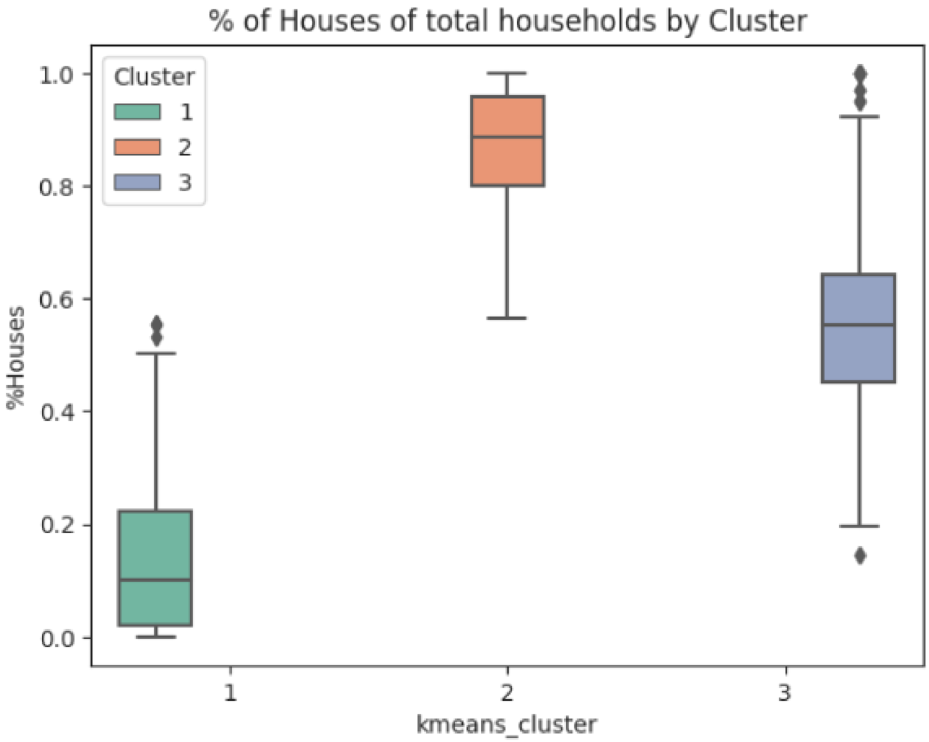


Figure 9: Boxplot of % of Owner of Total Households by Cluster

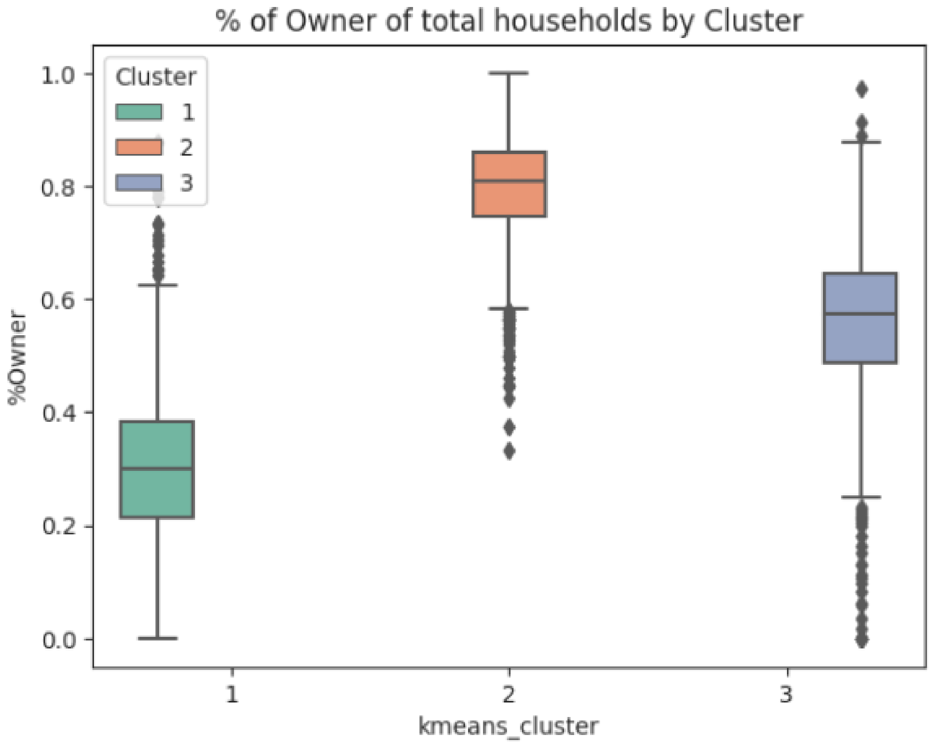


Figure 10: Histogram of Total Households by Cluster

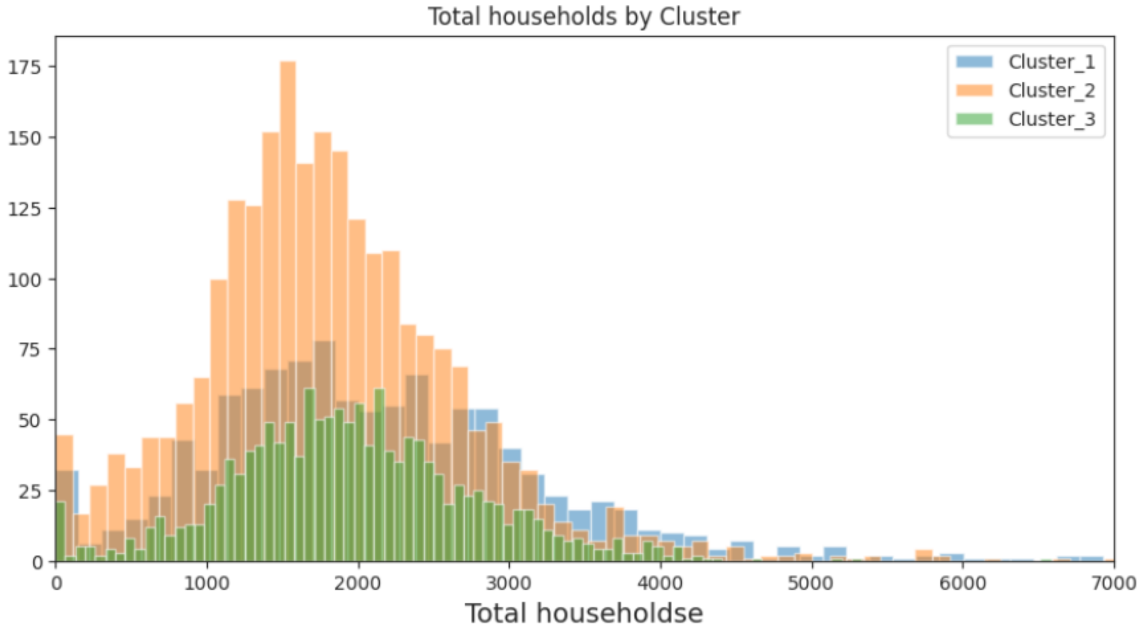


Figure 11: Histogram of Median Income by Cluster

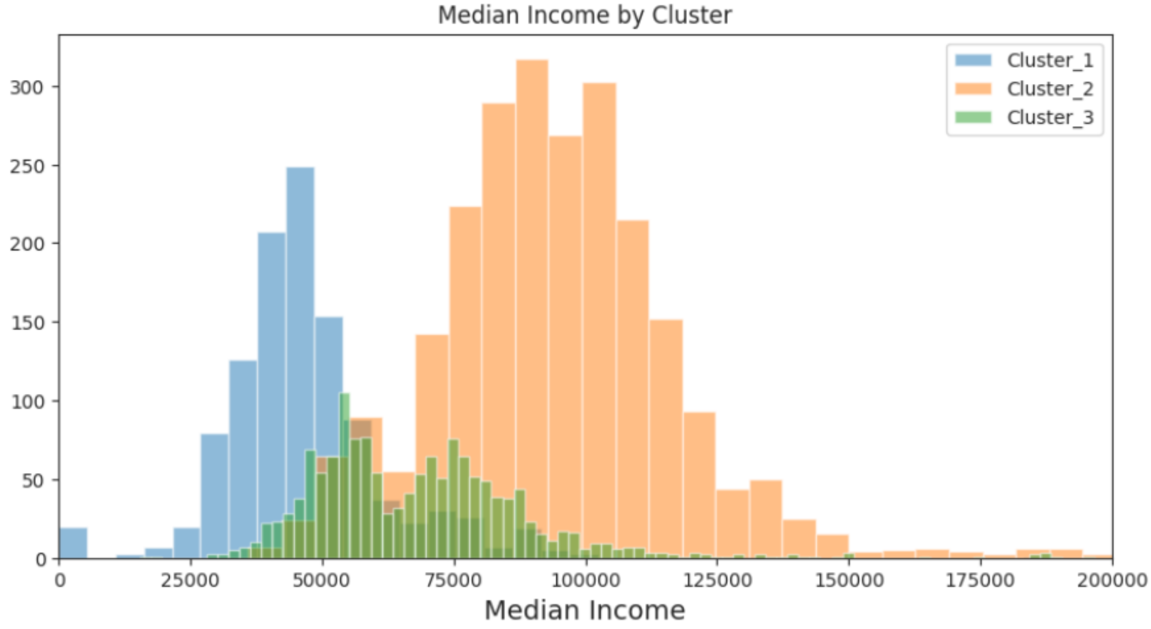


Figure 12: BIRCH: Histogram of % Houses by Cluster

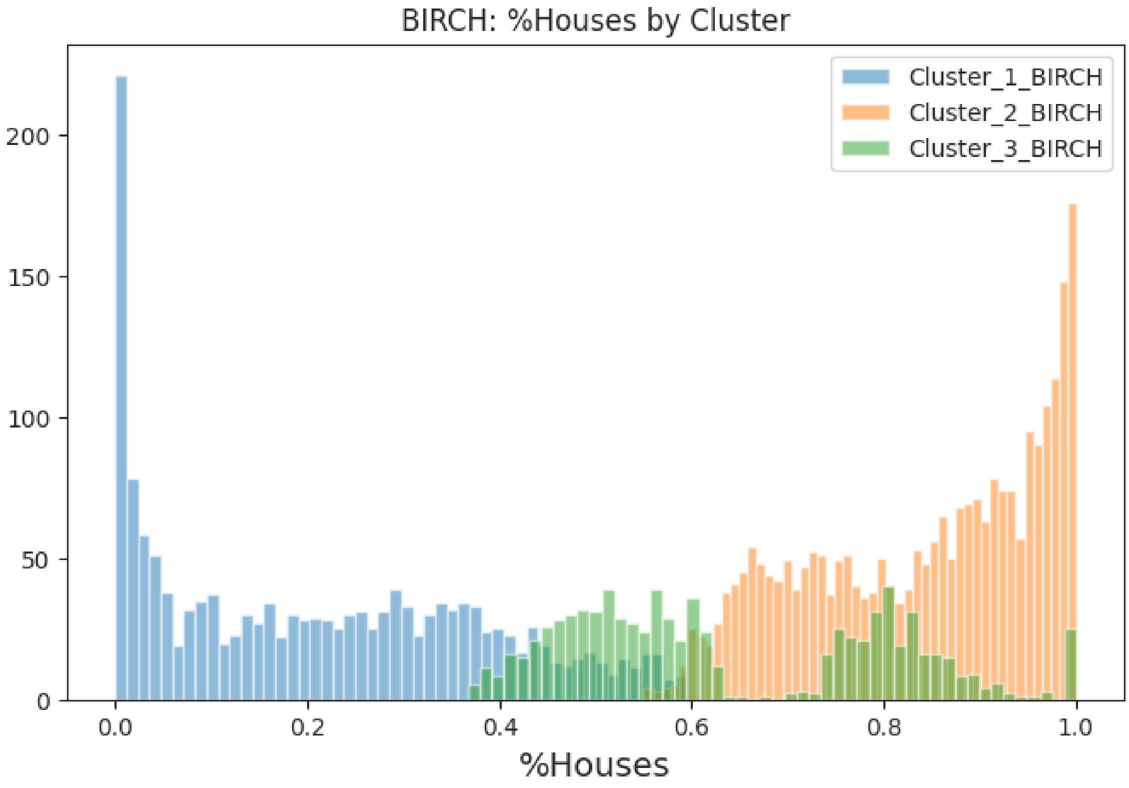


Figure 13: BIRCH:Histogram of % Owner by Cluster

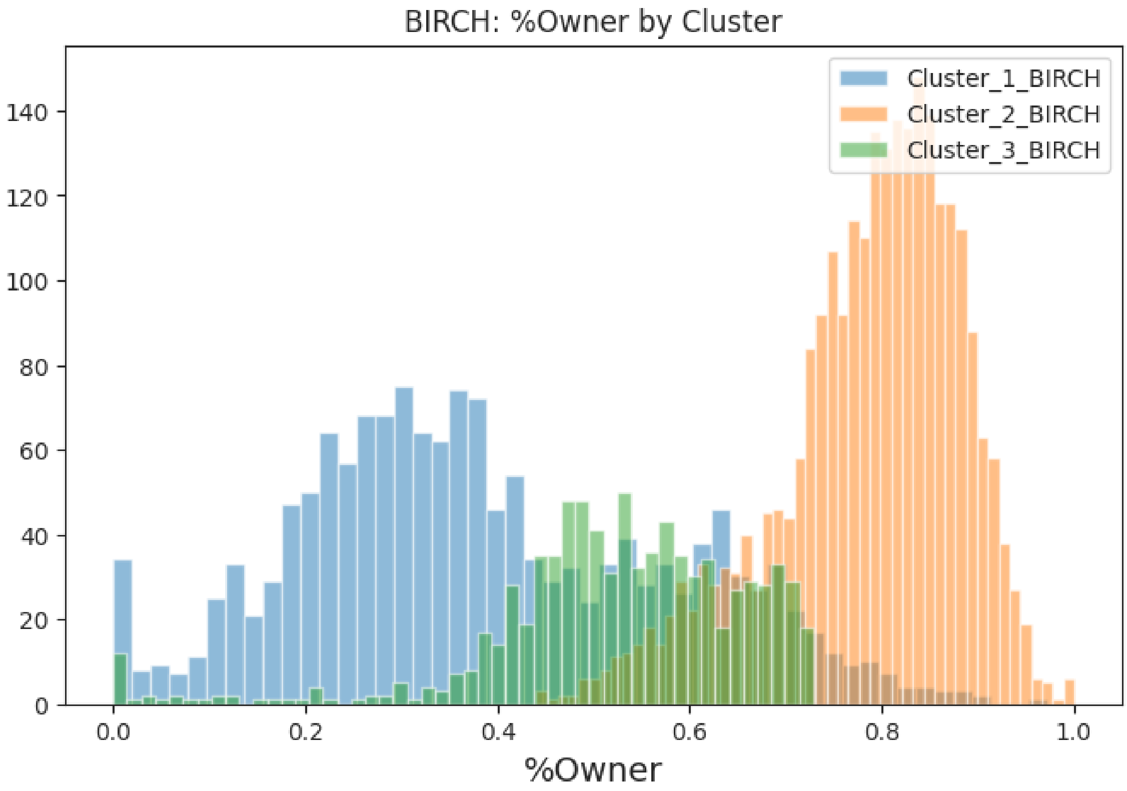


Figure 14: Test MAE Comparison

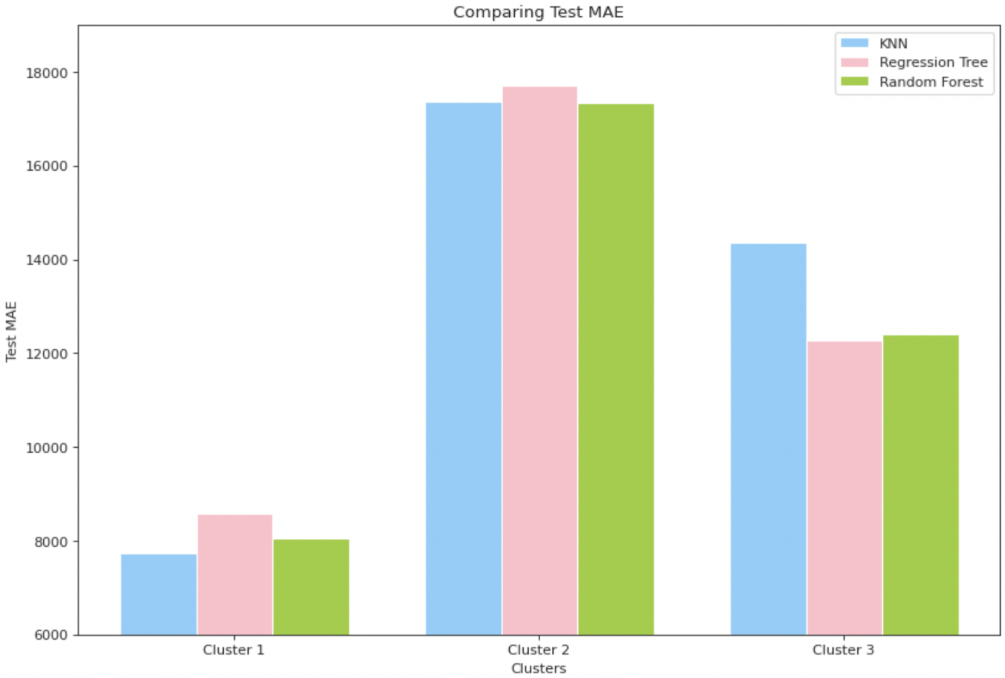
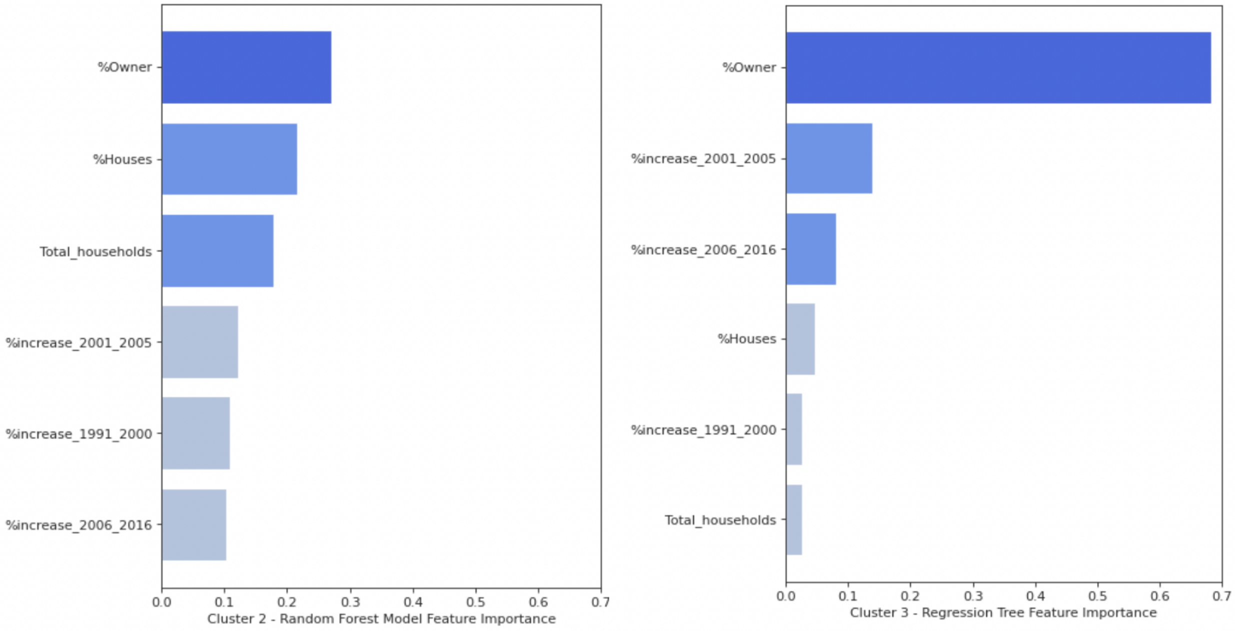


Figure 15 and 16: Feature Importance of Cluster 2 and Cluster 3


%Owner
%Houses
Total_households
%increase_2001_2005
%increase_1991_2000
%increase_2006_2016
Cluster 2 - Random Forest Model Feature Importance
%Owner
%increase_2001_2005
%increase_2006_2016
%Houses
%increase_1991_2000
Total_households
Cluster 3 - Regression Tree Feature Importance