|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Group Number:** | 7 | | | | | | | Text  Description automatically generated |
| **Assignment Title:** | Group Assignment 4 | | | | | | |
| **Course Code:** | RSM8413 | | | | | | |
| **Instructor Name:** | Gerhard Trippen |  |  |  |  |  |  |
|  |  | | | | | | |
|  |  | | | | | | |
|  |  | | | | | | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | | | |
|  |  |  |  |
|  | 1002183031 |  | 1005627403 |
|  | 1002897378 |  | 1007045118 |
|  | 1007554745 |  | 1006604701 |

**Executive Summary**

Canada Venture, a Canadian real estate developer, wants to decide on the optimal census tracts to develop real estate. Census tracts (CTs) are small, relatively stable geographic areas that usually have a population between 2,500 and 8,000 persons. They are located in census metropolitan areas and in census agglomerations that had a core population of 50,000 or more in the previous census.

We used a sample of 5000 records that contains information on census tracts (CT) delineated by local specialists and Statistics Canada and found five different cluster types: Renters, New Owners, Old Owners, Other Dwellers, Stagnant. It is recommended that Canada Venture customize their development strategies based on the needs and characteristics of these different clusters.

**Introduction**

One of the most important unknowns for real estate developers is how to develop new properties in different census tracts based on the characteristics and income levels of census tracts (CT). However, they do not always have the median income of new CTs. Developers wish to find out what the median income level is for different CTs that they plan to invest in.

In this report, we consult for Canada Ventures, a real estate developer, with the aim of finding the defining characteristics, based on the different available features, that can best predict the median income (per year) for census tracts.

The business objective is to use the predicted median income to ascertain the characteristics of different clusters. Hence, we can provide Canada Ventures with the framework to develop better investment decision-making practices. We aim to accomplish this by clustering different CTs and creating optimized models for each cluster.

**Part One**

Variable Summaries

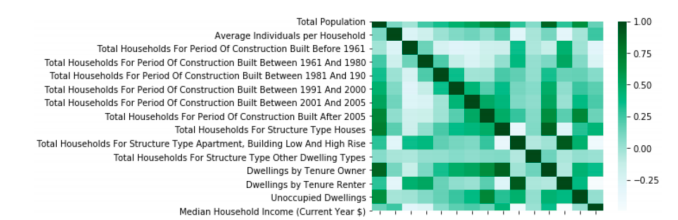
We noticed that there are a few rows with all the entries equal to 0. In order to better cluster the training data, we removed the 20 rows with missing values.

Based on the correlation matrix, we removed a few variables that are highly correlated with other variables (Total Population, Total Households, Total Households For Period Of Construction, Total Households for Tenure, Dwellings by Tenure Owner, Dwellings by Tenure Renter). And we derived two additional variables in order to capture additional information that is not provided (Total Households for Period Of Construction Built After 2005) and to capture information of some removed variables (Persons Per Household). The following table provides details for each variable in the training data and two new derived variables.

|  |  |  |
| --- | --- | --- |
| Total Population | Dropped | * Highly correlated with multiple variables (see correlation matrix graph) * Information is captured by other variable derived (Persons Per Household) |
| Total Households | Dropped | * Overlap with variables showing buildings built-in time installments (Collinearity) * Have the same value as Total Households For Period Of Construction * Information is captured by other variable derived (Persons Per Household) |
| Median Household Income | Included | Target Variable |
| Total Households For Period Of Construction | Dropped | Same as Total Households |
| Total Households For Period Of Construction Built Before 1961 | Included | Provide useful information on Occupied Private Dwellings by Period of Construction |
| Total Households For Period Of Construction Built Between 1961 And 1980 | Included | Provide useful information on Occupied Private Dwellings by Period of Construction |
| Total Households For Period Of Construction Built Between 1981 And 1900 | Included | Provide useful information on Occupied Private Dwellings by Period of Construction |
| Total Households For Period Of Construction Built Between 1991 And 2000 | Included | Provide useful information on Occupied Private Dwellings by Period of Construction |
| Total Households For Period Of Construction Built Between 2001 And 2005 | Included | Provide useful information on Occupied Private Dwellings by Period of Construction |
| Total Households For Structure Type Houses | Included | * Provide useful information on houses structure type * Highly correlated variables are removed |
| Total Households For Structure Type Apartment, Building Low And High Rise | Included | * Provide useful information on houses structure type * Highly correlated variables are removed |
| Total Households For Structure Type Other Dwelling Types | Included | * Provide useful information on houses structure type * Highly correlated variables are removed |
| Total Households for Tenure | Dropped | Since drop all of the Occupied Private Dwellings by Tenure categories |
| Dwellings by Tenure Owner | Dropped | Highly correlated with multiple variables (Total Households For Structure Type Houses) |
| Dwellings by Tenure Renter | Dropped | Highly correlated with multiple variables (Total Households For Structure Type Apartment) |
| Total Households For Period Of Construction Built After 2005 | Derived | The number for total households built after 2005 is not included in original variables |
| Persons Per Household | Derived | Provides useful information on Total Population and Total Households |

Correlation Matrix and Scatter Plots

In order to evaluate the correlations between the dependent variables and the independent variables, a correlation heatmap was developed, shown below.

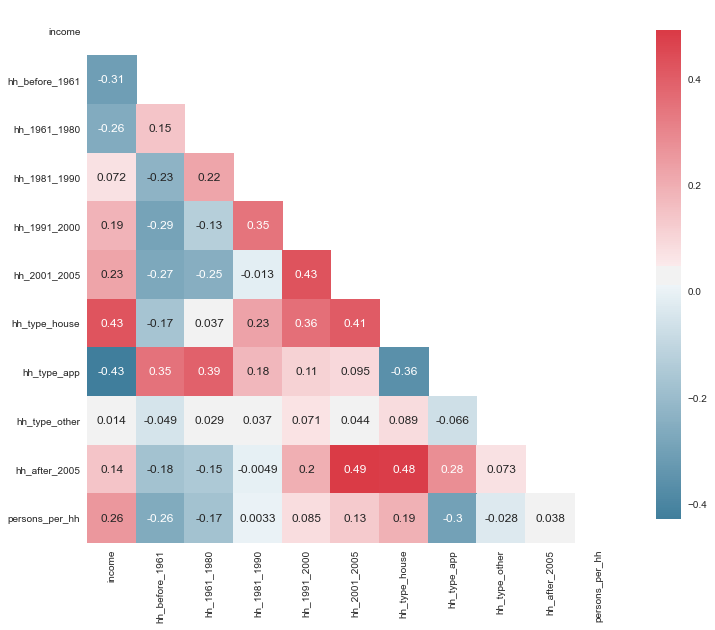


The most significant correlations are noted as:

* The households built after the year 2005 is strongly positively correlated with the total population, a scatterplot is shown for this relationship in Appendix Figure (A). The duration after 2005 exhibits the highest increase in population relative to all the other construction durations (i.e. before 1961 – 2001). This is indicative of a steep increase in population, specifically after 2005, relative to other periods.
* Apparently, the households with structure type houses exhibit a strong positive correlation with the total population as well, shown in the figure in Appendix Figure (B). This is also representative of a high demand from households after the year 2005 for houses as opposed to other household structures.
* Furthermore, from Appendix Figure (C) it can be seen that House type structures are also strongly correlated with the Dwelling by Tenure Owner, indicating that more owners rather than condominiums in buildings tend to be living in house types.
* In contrast, Appendix Figure (D) displays the strong positive correlation between Tenure Renters and Apartment type structures, indicating that more individuals that live in condominium buildings tend to be renters on average.
* Additionally, the Dwellings by Tenure Owner is positively correlated with the target variable: Median Household Income. Indicating that on average owners of households tend to have higher median household incomes, as in Appendix Figure (E).

However, based on the results of the correlation, it can be assumed that: Households that are of structure type houses, were typically built after 2005 and are currently occupied by the owners tend to have higher median incomes than average.

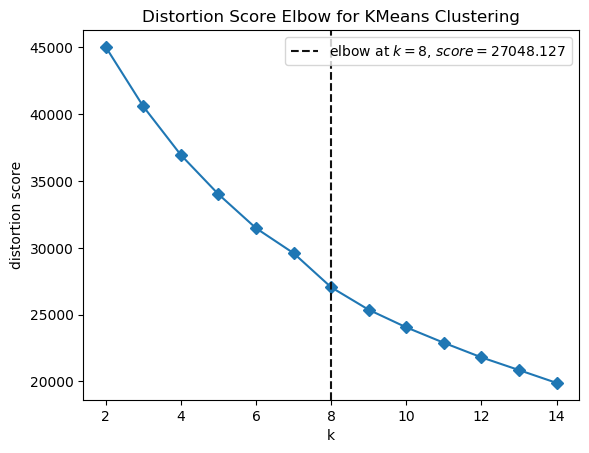
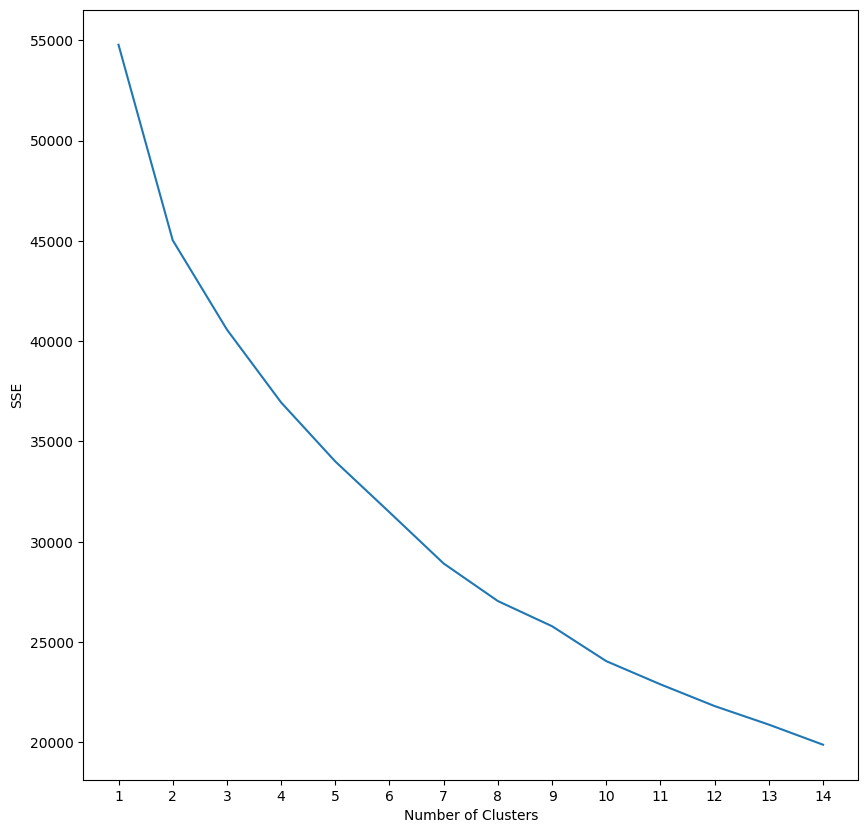
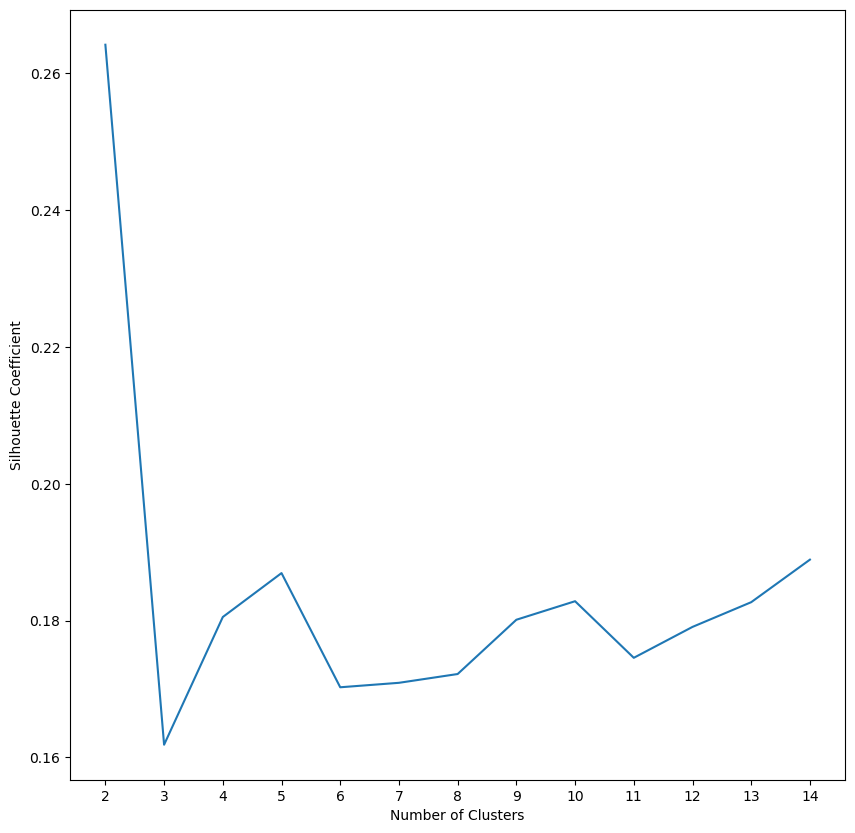
After we cleaned the original input variables, the 11 variables that we have selected and derived are not highly correlated (<0.5). The below graph shows the correlation heatmap for the final 11 variables.



Optimal K for Clustering

In order to find the optimal number of clusters we used following methods:

* Elbow method - Distortion Score
* Elbow method - Inertia (Done manually)
* Silhouette Coefficient



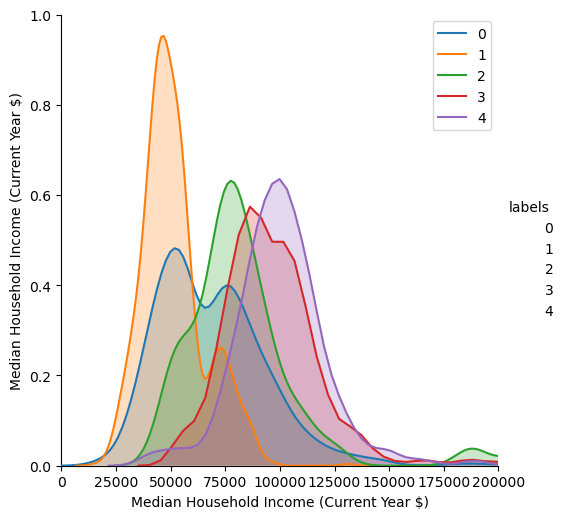
**Silhouette Coefficient Elbow Method - Inertia Elbow Method - Distortion**

It was observed that the optimal number of clusters varied based on the random\_state and other metrics used in the above methods. The above methods suggested the number of clusters range from five to eight. Based on the elbow method, better interpretability, and practicality for the business case we selected five clusters as the optimal number.

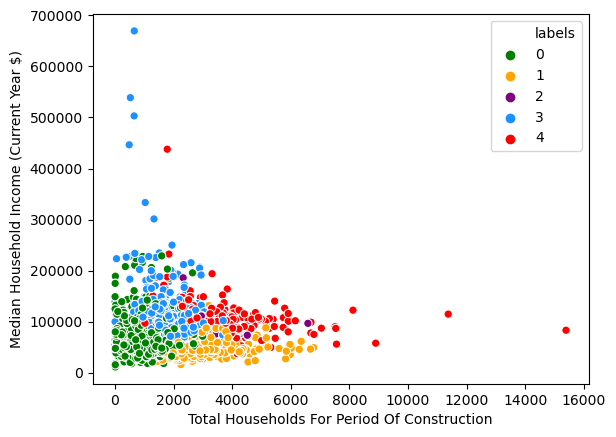
Cluster Profiles

We identified 5 clusters. Cluster 0 has 2110 records, Cluster 3 has 1407 records, Cluster 1 has 909 records, Cluster 4 has 474 records, Cluster 2 has 80 records.

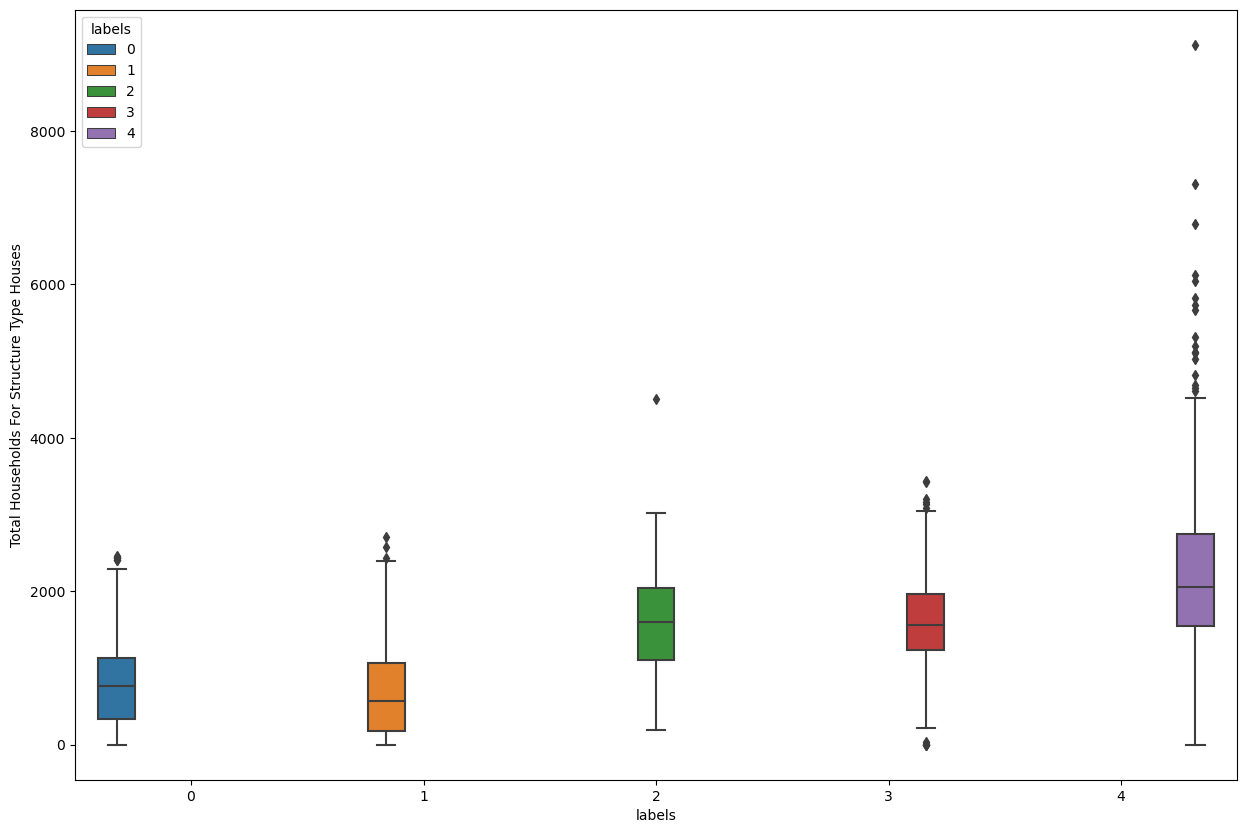
The histogram below shows the distributions of Median Household Income for the five clusters. The rank of the mean of Median Household Income for each cluster is 4,3,2,0 and 1.



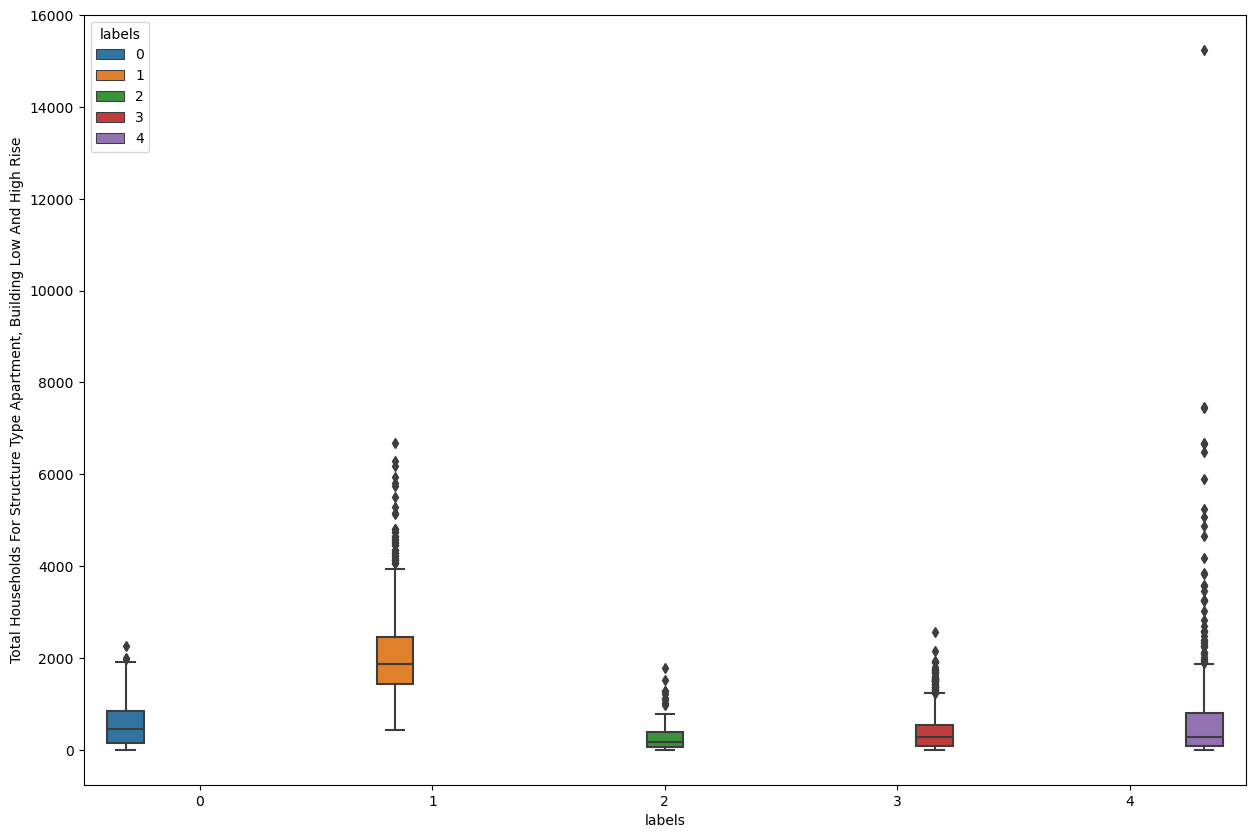
The scatterplot below shows the relationship between Median Household Income and Total Households for Period of Construction for the five clusters.



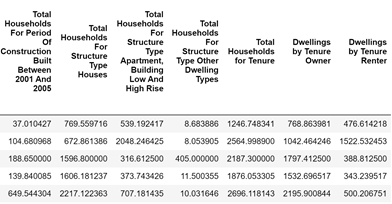
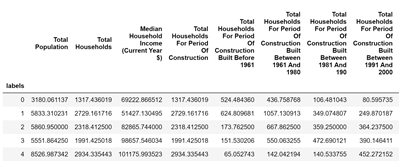
The boxplot below shows the distribution of Total Households for Structure Type Houses for the five clusters. We see that cluster 4 has the highest median, while cluster 1 has the lowest.



The boxplot below shows the distribution of Total Households for Structure Type Apartment, Building High and Low Rise) for the five clusters. We see cluster 1 has the highest median, while cluster 2 has the lowest.



The mean value for each variable is also shown below. There is also a trend we can see in the time of construction, a decreasing trend over time for clusters 0, 1, 3, and an increasing trend over time for cluster 4.



Based on all the information above, we summarize the housing characteristics of the five clusters, and determine names to represent each of them. Details are shown below.

*Cluster 0: Stagnant* - Second lowest median income, no new development, more historical, lower-middle-class

*Cluster 1: Renters Representative* - Lowest median income, mostly apartment & renter working-class resident or temporary residents

*Cluster 2: Other Dwellers Representative* -Other dwelling types dominant, all other variables fall in the middle.

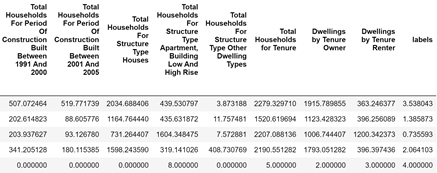
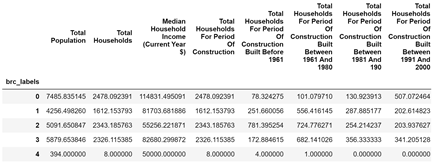
*Cluster 3: Old Owner's Representative* - High income, historically rich areas, mostly developed in the previous century, residents are largely house owners.

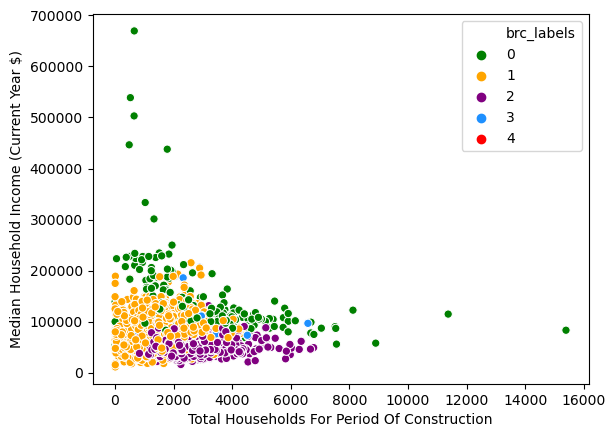
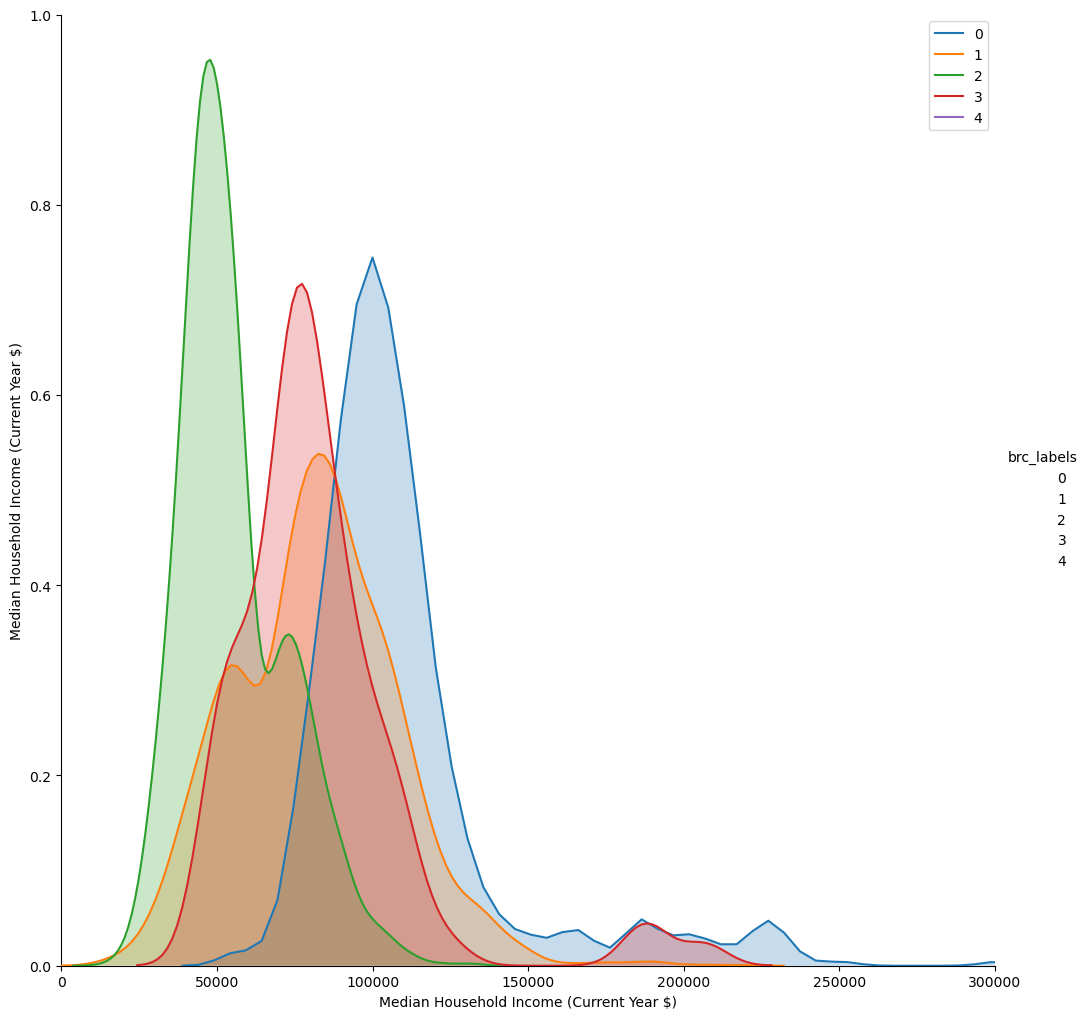
*Cluster 4: New Owners Representative* - Highest median income, recently developed areas, residents could be new upper-middle classes, most house owners.

BIRCH Clustering

We kept K=5 for BIRCH clustering. Cluster 0 has 552 records, Cluster 3 has 78 records, Cluster 1 has 2874 records, Cluster 4 has 1 record, Cluster 2 has 1475 records. The cluster distribution changes a lot, especially for cluster 4 where there is only one record. Conversely, the cluster with the most records increased (cluster 1, 2874 records).

According to the same diagrams we did before, as well as the mean statistics table (shown below; boxplots are shown in Appendix), we tried to match the clusters in BIRCH with the ones in K-means. Cluster 0 is New Owners Representative, which is very similar to K-mean cluster 4. Cluster 1 here is Old Owners Representative; it shows a high percentage of houses, a relatively high income, but the decreasing trend of construction in years is not obvious. Cluster 3 is Other Dwellers; it shows the same incredibly high Other Dwellers type housing as K-mean cluster 2, yet the statistics of other variables are a bit different. Cluster 2 is Renter Representative (cluster 1 from K-means) as it shows a low house to apartment ratio and a low median income, but not the lowest. Cluster 4 has only one data point so drawing a conclusion is harder, but it does show a low median income and decreasing trend of construction in the year, so it matches with Stagnant (cluster 0 from K-means).



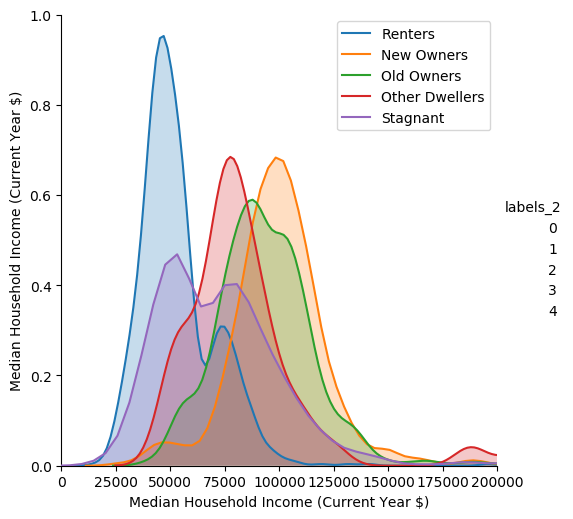


**Part Two**

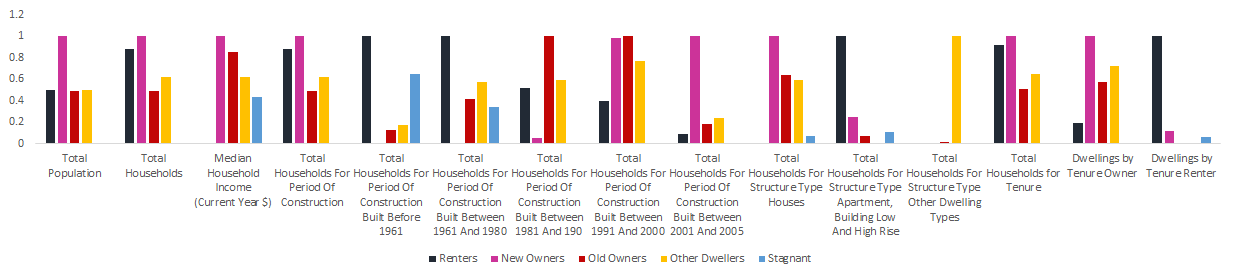
Clusters as Inputs to Downstream Models:

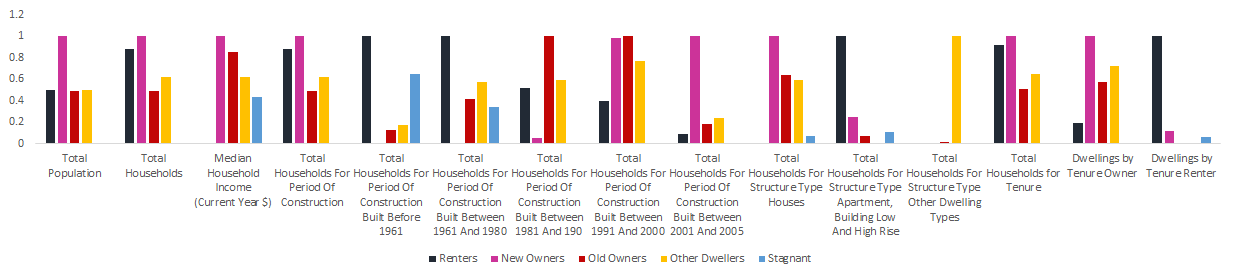
From our previous analysis, K-mean clustering with k=5 was chosen as our best clustering model. This was because BIRCH clustering did not provide the best distribution of clusters. Also, the K value was not changed to maintain interpretability. Once the records were divided into clusters, the labels were added to the original dataset as “labels\_2”. The EDA for these clusters is shown in the graphs below.

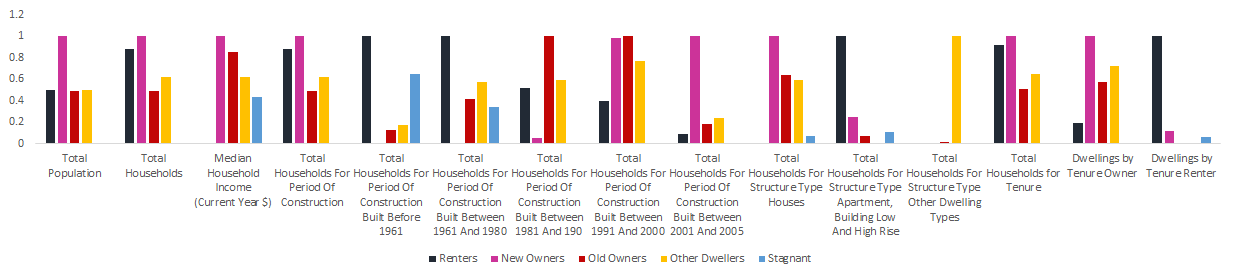
Median Household Income Distribution of New Clusters



Relative Importance of All Clusters to the Predictors of K-Means (values were Min-Max scaled)







Chosen Estimation Modeling Tools

After determining the clusters, we decided to find the optimal model for each cluster type. We first split the data into training and validation sets with a 60/40 split. Next, we used the pipeline function to standard scale and run each algorithm against every cluster. Then we used GridSearchCV for each model to determine the optimal parameters that minimize RMSE for each cluster. 10-Folds cross-validation was also used for each model.

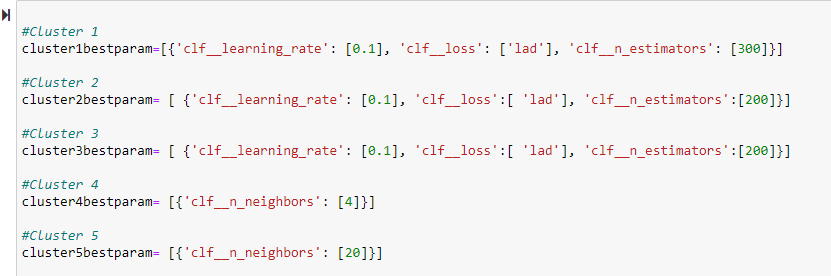
The parameters we found were different for each cluster. After some deliberation and taking into account the lack of computing power, we ended up with five algorithms: Linear Regression, KNN, SVR, Decision Tree, and Gradient Boosting Regression.

Model Comparison

The output RMSE for each model for each cluster is shown below.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Clusters** | **Linear Regression** | **Decision Tree** | **KNN** | **SVR** | **Gradient Boosting Regression** | **N** |
| **Renters** | 17404.741 | 19171.94 | 17514.557 | 19054.725 | 16628.483 | 81 |
| **New Owners** | 19737.766 | 34574.71 | 20954.298 | 22674.469 | 19351.322 | 549 |
| **Old Owners** | 18919.312 | 23474.50 | 19275.042 | 21909.535 | 17686.09 | 836 |
| **Other Dwellers** | 20162.986 | 27330.91 | 19953.755 | 24324.974 | 23330.71 | 1554 |
| **Stagnant** | 31145.849 | 39710.20 | 29233.75 | 33017.427 | 30056.512 | 1960 |

For the first three clusters, the gradient boosting regressor was the best option, and for the last two, KNN was the most optimal. The optimal parameters are shown below as well.



Predicting with Test Data

The test data was preprocessed the same as the training data. Then the cluster labels were predicted using the K-means model that was fit on the training data. Afterwards, each cluster model was fit using the training data with the optimal models and then the test data was used to predict income. The prediction results were then combined into one column with the proper index order and saved as a txt file.

**Conclusions & Recommendations**

After conducting our analysis, we recommend that Canada Ventures (CV) customize their development strategies (the type of houses built and customer acquisition) based on the clusters the CTs are in. First, we believe Canada Ventures should utilize a narrow targeting approach as opposed to a broad one. They should only focus on CTs that are worth investing in and have potentials, determined by our clustering, instead of all of them. Our recommended approaches are as follows:

* For the Old Owners Representative CTs area, challenges exist due to legacy competition. Thus, the CV should focus on renovating old buildings.
* For New Owners areas, develop primarily house structures as the area is filled with high-income investors. Competition with other developers could be a challenge, so we should focus on marketing campaigns as well to capture a higher market share.
* For Renters representative areas, apartments could be built to accommodate the low-income working-class population. Canada Ventures can also invest in property management and already developed properties. Finally, CV could also collaborate with the local government to work on projects like social housing.
* The Stagnant CTs have no development potential in terms of amenities and do not offer much room for profit so CV should not target them.
* The Other Dwellers CTs require further research. The CV should look into what buildings are considered other dwellings to see if there are any development opportunities.

**Challenges and Next Steps**

As stated above, different clusters come with their own challenges. Old Owner CTs come with legacy competition, while New Owner CTs come with fierce competition. The realities of both these clusters lead to the need for smarter investment strategies that were listed above. For Renters, the advent of property management comes with the risk of rent control limiting profit. Finally, for Other Dwellers, the lack of information about the other dwelling structure types hinders our confidence in investing in CTs within this cluster.

To improve our model and lower the risk of investment, new data on market price, competitor data, and general real estate data can be used. Additionally, data on past versions of the dataset can be used to measure things such as potential housing bubbles, gentrification, and planning & zoning changes.

**Appendix**

Scatter Plots for EDA

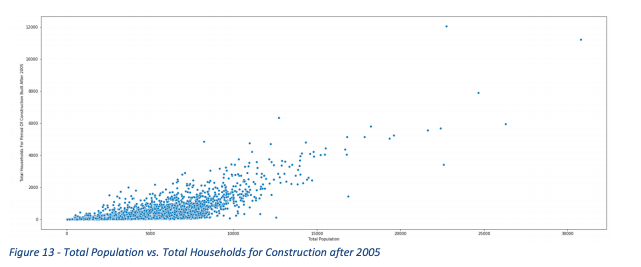


Figure A

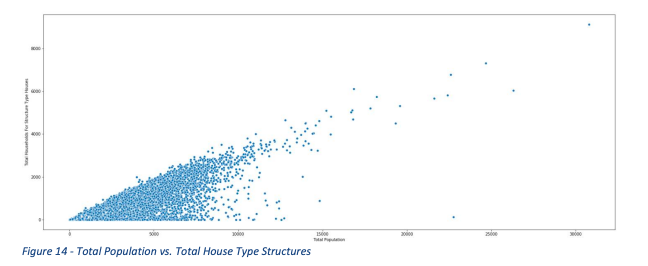


Figure B

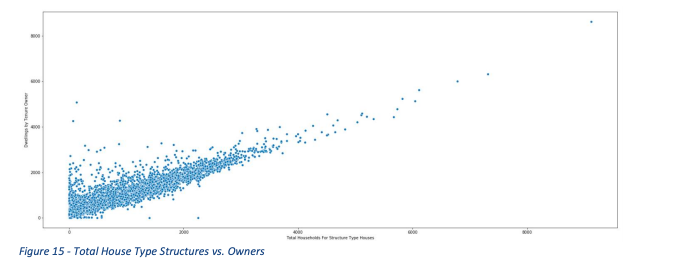


Figure C

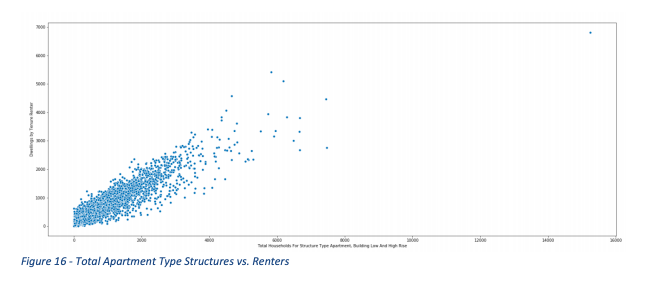


Figure D

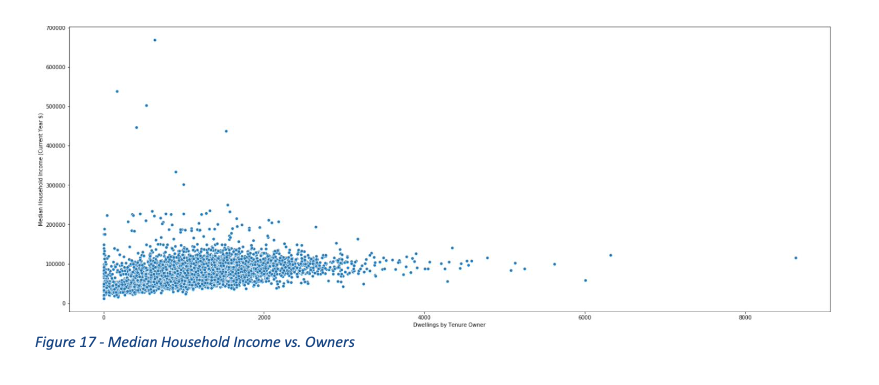


Figure E

BIRCH Clustering EDA

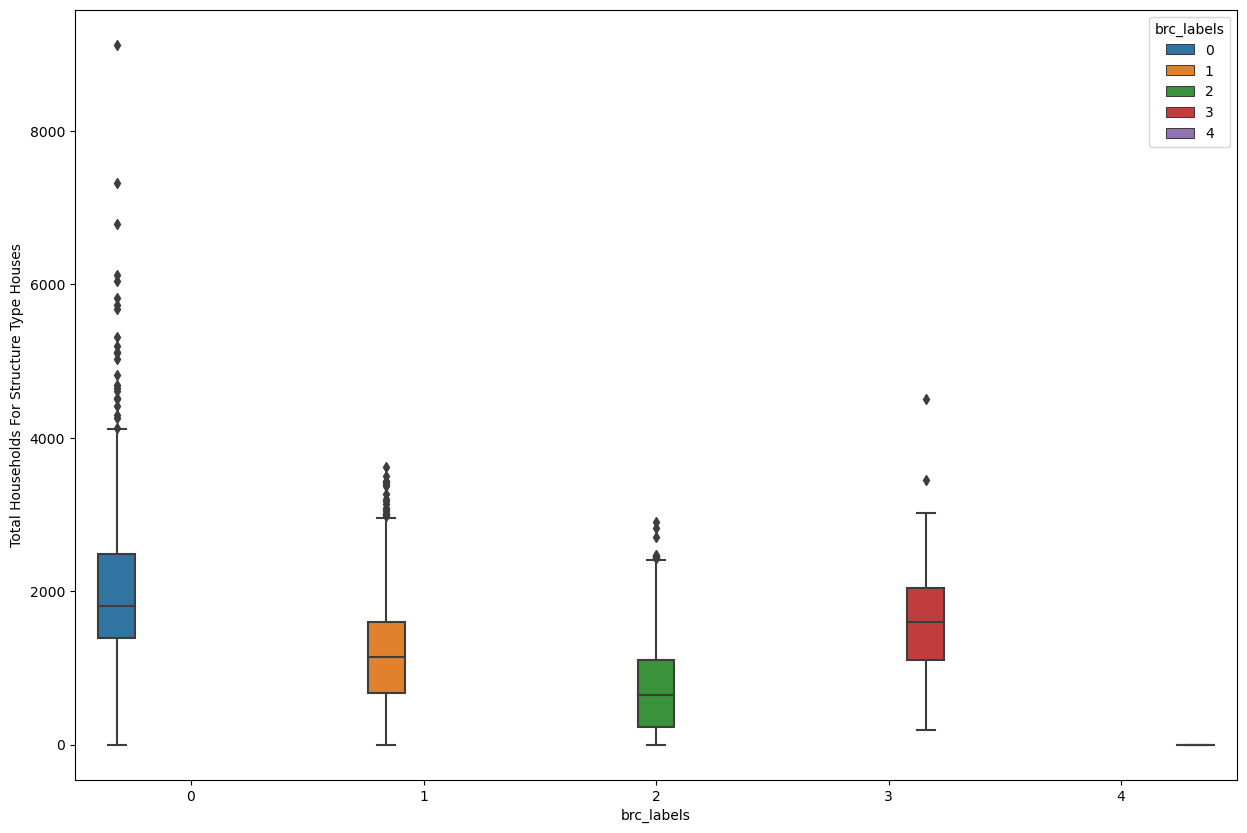


Figure A

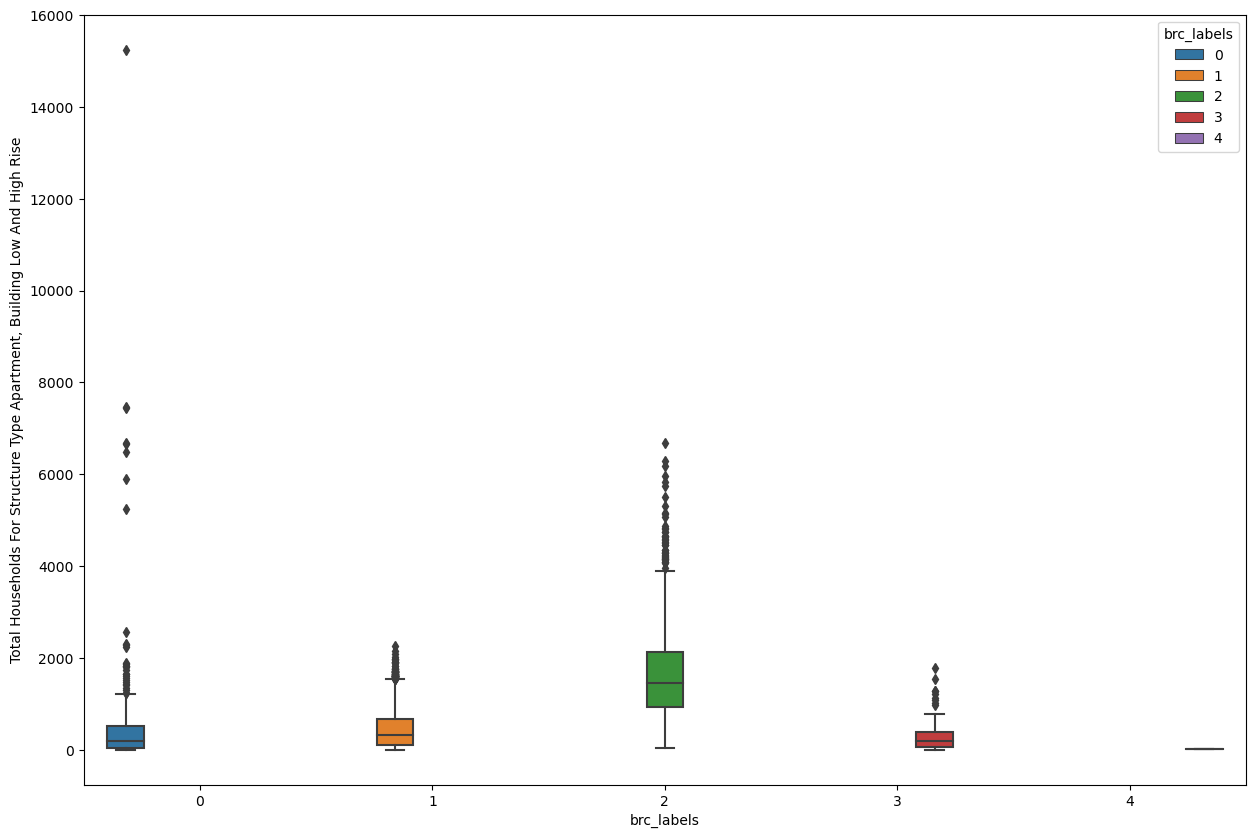


Figure B

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Final Page** | |  |
|  |  |  |  |
|  | **Grade:** |  |  |