Rotman

Canada Ventures-Property Classification

Team 7

February 5, 2021
Prepared by
Humza Butt, Alan Liu, Cynthia Zhuo, Ifeyinwa Kofo-Alada, Priya Varshini G, Josh Samadi

Rotman School of Management
UNIVERSITY OF TORONTO

Agenda

Canada Ventures

Rotman

- **Executive Summary**
- **Business Problem Setting**
- **Problem Statement**
  - Managerial Question
  - Analytical Question
- **Key Data Sources Overview**
- **EDA**
- **Modeling Approach**
  - Cluster Model
  - Segmentation Modelling
- **Key Findings**
- **Recommendations**
- **Next Steps**

Executive Summary

Rotman

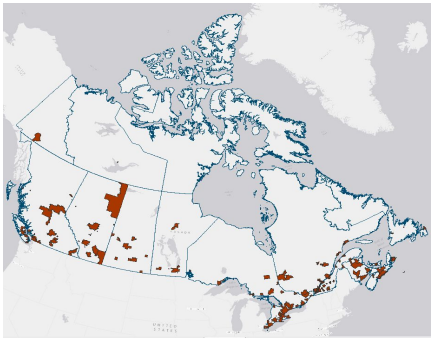
Business Case

Business Problem

- Canada Ventures wants to decide the optimal CTs to develop properties for.
- CTs are small, stable geographic areas that usually have a population between 2,500 and 8,000 persons.
- Assumes that specific demographics prefer a specific type of property.
- “If you build for the people, you build for success”

Data Used

- Sample of 5000 records that contains information on census tracts (CT) delineated by local specialists and Statistics Canada.



Main Findings

- We found 5 different cluster types: Renters, New Owners, Old Owners, Other Dwellers, Stagnant.
- The regression models that gave optimal median income predictions varied with the cluster we ran the model on.

Recommendation

Cluster	Recommendation	Challenges
New Owners	Luxury Houses, Marketing	Fierce Market Competition
Old Owners	House Renovations	Legacy Competition
Renters	Apartments Property Management	Rent Control

Business Problem

Rotman

Current Situation

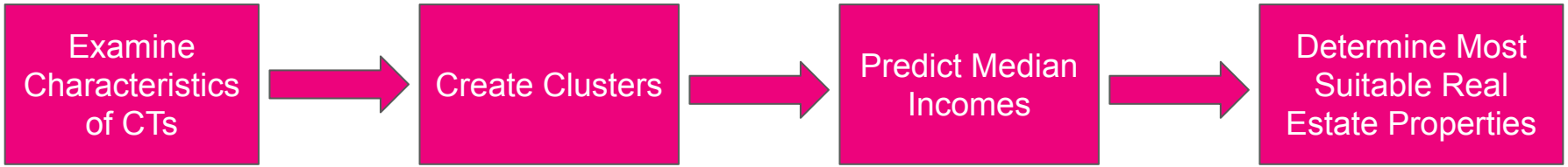
Business Overview

- CV develops properties on census tracts.
- A data-driven company that makes decisions based on demographic data.
- Work under the assumption that specific demographics prefer specific properties.
- “If you build for the people, you build for success”

Problem Statement

- The real estate developers want to develop new properties in different census tracts based on characteristics of different demographics. However they do not have the median income of the CTs.
- Developers want to find out what the median income level is for different CTs that they plan to invest in.

Investment Process

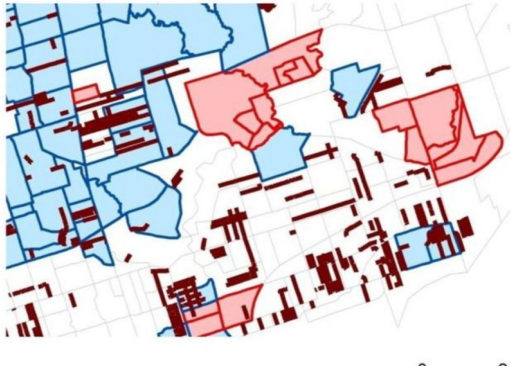


Managerial and Analytical Objectives

Rotman

Problem Statement

- **Managerial question**
  - How are we going to expand our property development to different CTs and what the best type of property to develop to fit the local characteristics and income level?
- **Analytical question**
  - Are there similarities between different census tracts that can be categorized?
  - What are the defining characteristics for different clusters based on the input variables?
  - What is the level of median income for each CTs?
  - For each cluster, what is the optimal algorithm and parameters that will minimize RMSE given input predictors for median income?



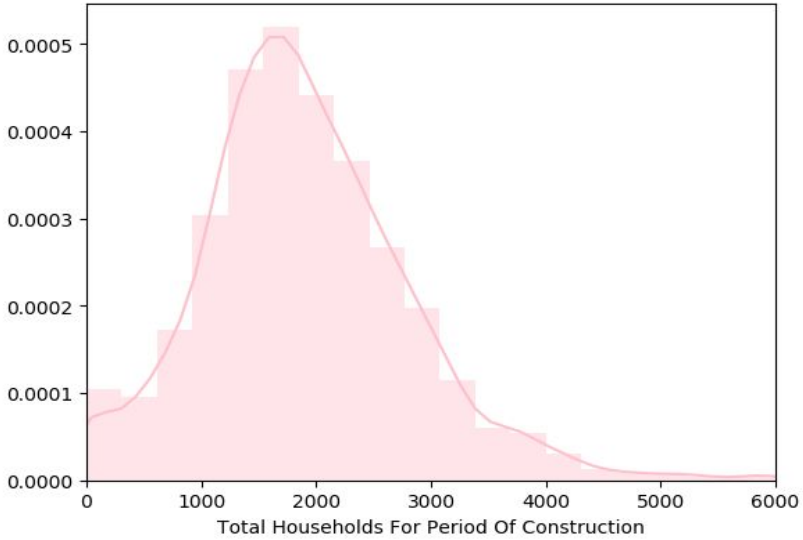
Key Data Sources

Rotman

Handling Data

Problems in Data

- Missing Data when median income is equal to 0.
  - Remove only 20 rows with this issue



Data Sources Used

Data	Number of Records	Number of Variables	Used
CensusCanada2016Training.csv	5000	14	Yes
CensusCanada2016Test.csv	721	13	Yes

Description of Data

- Census Data from Statistics Canada on housing characteristics of various Census Tracts.
- Total households are the same as Total Households For Period Of Construction
  - Both overlap with variables showing buildings built-in time installments (Collinearity)
  - Created Total Households Built after 2005
- Other highly correlated variables

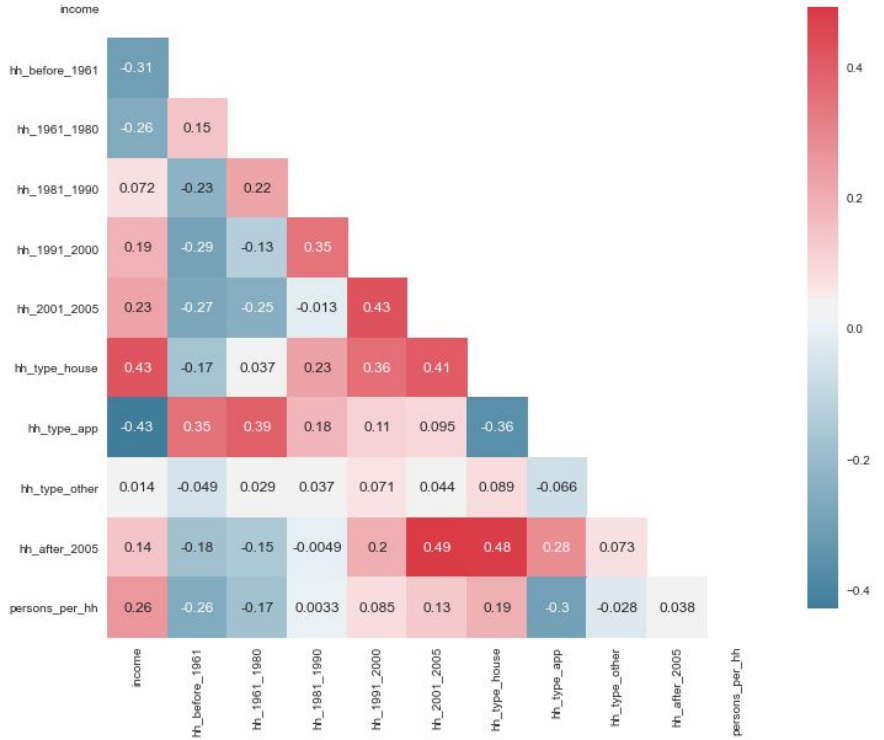
Key Data Sources

Rotman

Feature Selection

Variable	Handling	Reason
Total Households	Dropped	Info captured by other variables
Total Population	Dropped	Info captured by other variables
Total Households For Period Of Construction	Dropped	Same as Total Households
Total Households For Tenure	Dropped	Same as Total Households
Dwellings By Tenure Renter	Dropped	High correlation with hh_type_app
Dwellings by Tenure Owner	Dropped	High correlation with hh_type_house
Total Households For Period Of Construction Built After 2005	Created	Provides info in Total Households not captured by existing variables
Persons Per Household	Created	Provides info in Total Population

New Correlation Matrix



Rotman

Exploratory Data Analysis

Correlated Variable Analysis

- Total Population vs Total Household construction after 2005
- Total Population vs Total Household structure type houses
- Dwelling by Tenure Owner vs Total Household structure type houses
- Dwelling by Rentor vs Total Household structure type apartment buildings (High & Low Rise)
- Median Household Income vs Dwellings by Tenure Owner
- Households that were Constructed after 2005, of House type, and are Occupied by Owners will likely have higher Median Incomes.

Modeling Approach

Rotman

Clustering & Segmentation

Model Overview: Clustering Model

- Unsupervised learning algorithm where there is no target variable.
- K=5 from elbow method criterion.
- Both BIRCH and K-means were used to create labels and compared to determine which is the best clustering model.
- K-means was also conducted without the median income to compare future clusters.

Model Overview: Segmentation Model

- The optimal model for each cluster is created from K-means.
- RMSE was utilized as the evaluation metric.
- Models used for each segment include Linear Regression, KNN Regressor, Decision Tree Regressor, SVR, and Gradient Boosting Regressor.
- 10-Fold Cross-Validation was utilized with parameter tuning.

Cluster Variables and Output Variable

Input Variables

- 'Hh_before_1961',
- 'Hh_1961_1980'
- ,'Hh_1981_1990'
- 'Hh_1991_2000'
- 'Hh_2001_2005'
- ,'Hh_type_house,
- 'Hh_type_app'
- ,'hh_type_other'
- 'Hh_after_2005'
- 'Persons_per_hh’

Output Variables

- Birch Cluster Label
- K-means Cluster Label

Model Predictors and Response Variable

Predictors

- 'Hh_before_1961',
- 'Hh_1961_1980'
- ,'Hh_1981_1990'
- 'Hh_1991_2000'
- 'Hh_2001_2005'
- ,'Hh_type_house,
- 'Hh_type_app'
- ,'hh_type_other'
- 'Hh_after_2005'
- 'Persons_per_hh,

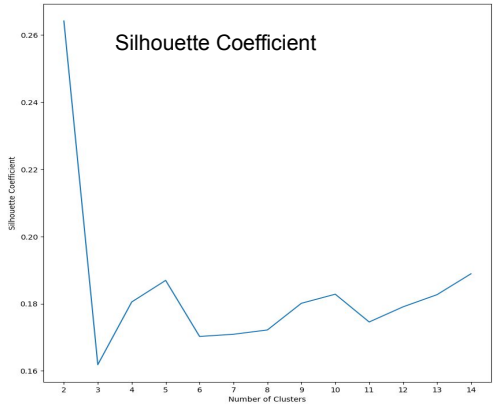
Response Variable

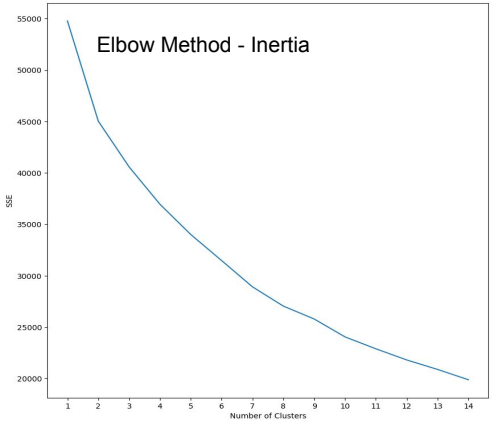
- Median Household Income

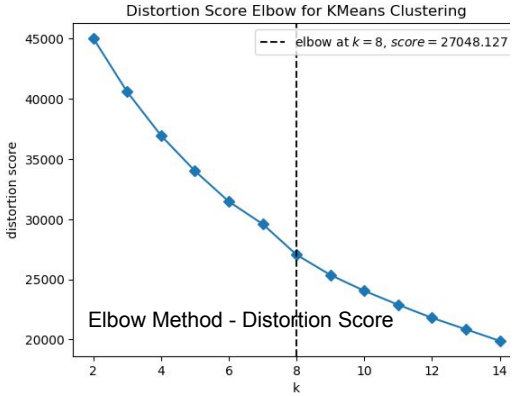
Modeling Approach

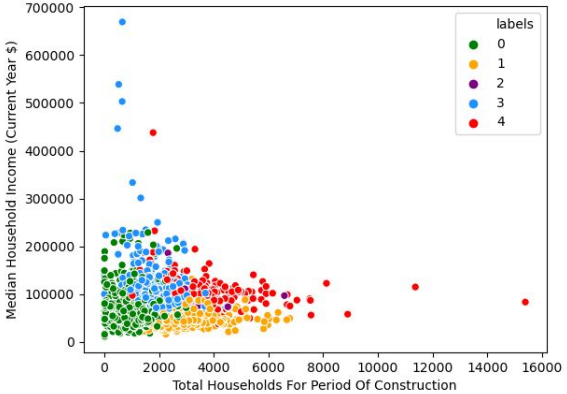
Rotman

Selecting K



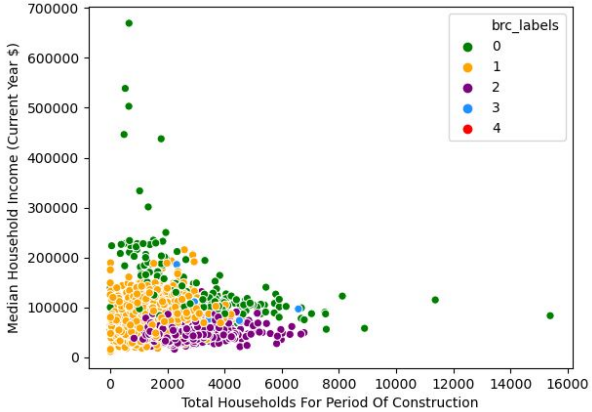






Optimal K Criterion

- The optimal k changed based on random_state and metric used, and range between 5 to 8.
- Silhouette Coefficient implies 2 as the optimal K. But in a business case this would be too basic of an answer.
- Selected 5 based on elbow method, silhouette coefficient value, and interpretability for the business.

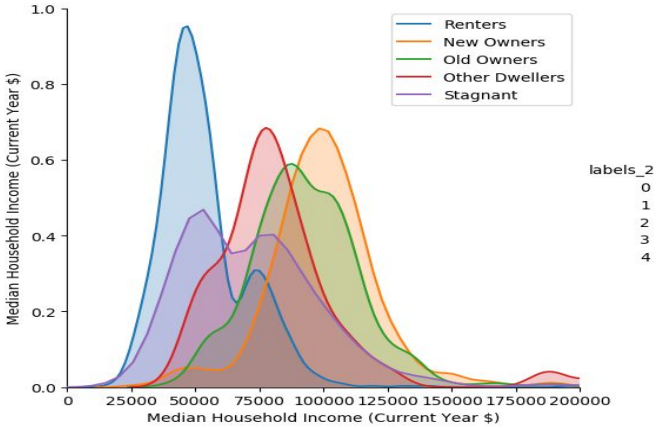


Key Findings

Rotman

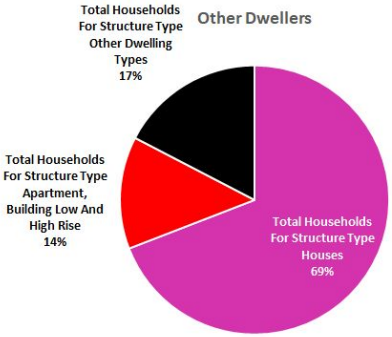
Cluster Profiles

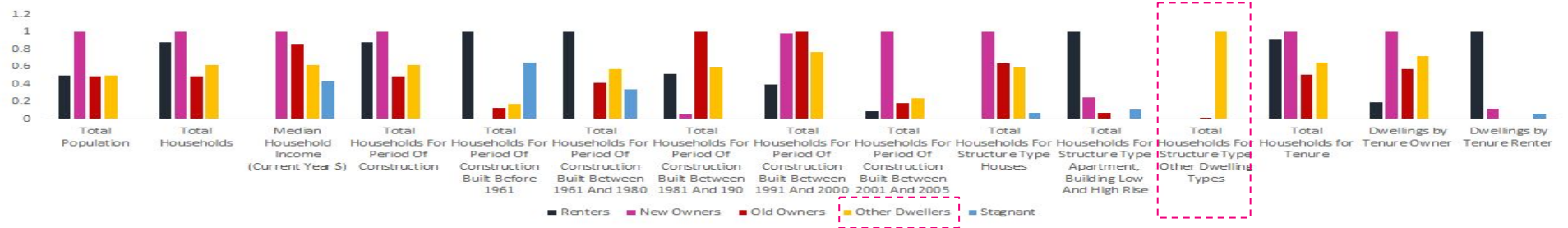
Median Income Distributions



Cluster Labels

1. **Renters Representative:** Lowest median income, mostly apartment & renter class resident or temporary residents
2. **New Owners Representative**: Highest median income, newly developed areas, residents could be new upper-middle classes house (mostly house owners)
3. **Old Owners Representative:** High income, historically rich area, mostly developed in the last century, and residents are largely house owner
4. **Other Dwellers Representative:** Other dwelling types dominant, all other variables fall in the middle
5. **Stagnant:** Second least median income, no new development, and more historical lower-middle class





Key Findings

Rotman

Segmentation Model

Optimal Models RMSE Overview

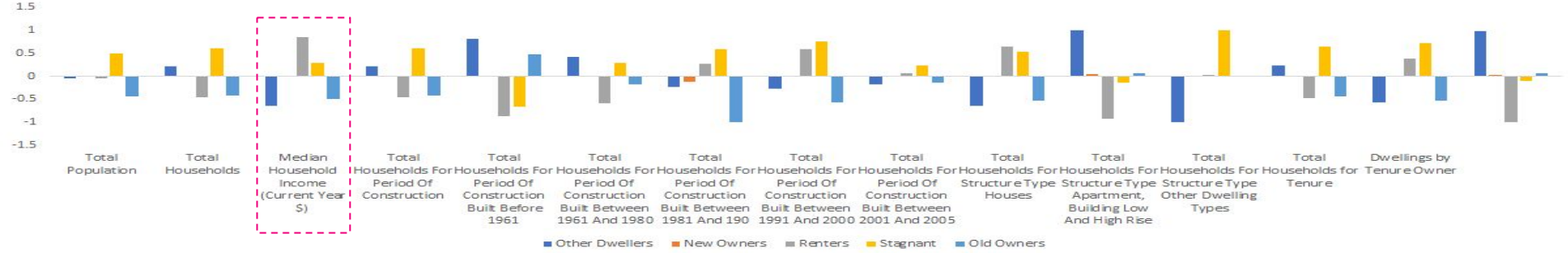
Clusters	Linear Regression	Decision Tree	KNN	SVR	Gradient Boosting Regression	N
Renters	17404.741	19171.94	17514.557	19054.725	16628.483	81
New Owners	19737.766	34574.71	20954.298	22674.469	19351.322	549
Old Owners	18919.312	23474.50	19275.042	21909.535	17686.09	836
Other Dwellers	20162.986	27330.91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

Model Performance Variable

Median Household Income (Current Year $)

Model Interpretation

- Each Cluster was recalculated without median income before determining the best model.
- The Weighted RMSE score is 23104.95.
- A regression model to predict median income can helpfully classify new CTs when we do not have that information.
    - Allows us to complete the classification of different clusters to make better-informed investments.
- Below is a graph that demonstrates the absolute change in the mean value for the different clusters when the median income is not used.



Recommendations

Rotman

Canada Ventures

Overall Findings

- From our cluster analysis, we found that different clusters constituted different housing characteristics.
    - Some clusters that were renting heavy also had a low median income, and other clusters that were ownership heavy also have a high median income.
    - Understanding the type of structures present in CTs implies the behavior of the residents.
- Different clusters were optimized for different regression models when predicting median income.
- The household population is a large indicator of construction activity.

Tactical Recommendations

- Canada Ventures should customize their development strategies (the type of houses build and customer acquisition) based on the clusters the CTs are in.
- Only focus on CTs that are worth investing in and have potentials, determined by our clustering.

Cluster	Recommendation	Challenges
New Owners	Luxury Houses, Marketing	Fierce Market Competition
Old Owners	House Renovations	Legacy Competition
Renters	Apartments, Property Management	Rent Control

Next Steps

Rotman

Future Considerations

- Additional data we could use:

Variable	Reasoning
Real Estate data	Such as market price, lands under development, and competitors' information (the presence of competing real estate developers in our targeted CTs could eat into our expected profits).
Further Demographic data	This would help to create more informative clusters. E.g. No of children in the household could be an influencing factor on the housing type our target customers in a CT prefer.

- Further Research into cluster 4 (Other Dwellers). What buildings are considered other dwellings and any development opportunities?
- Time series of the census tracts data to track for things such as gentrification, housing bubbles, and changes of government planning & zoning policy.
- Use cluster labels as an input variable and rerun our model on a combined data frame.

Thank You

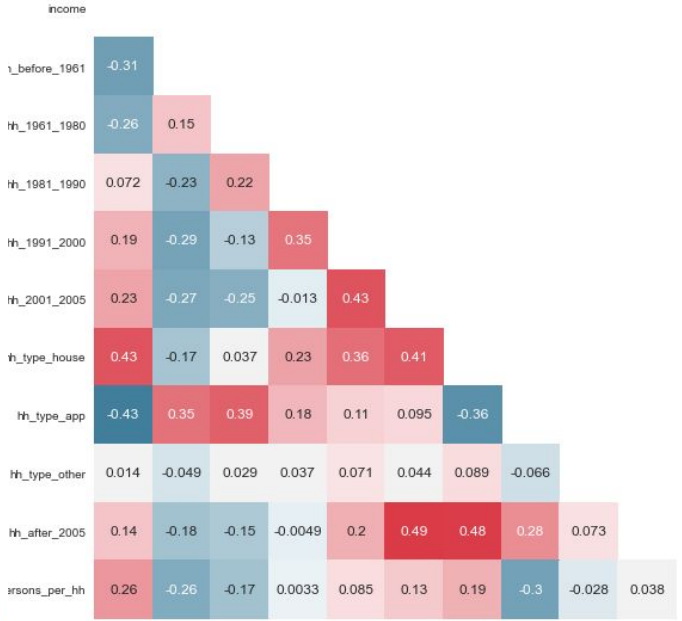
Q&A

Appendix:

Appendix

Rotman

EDA: Correlation Before And After Feature Selection



Appendix

Rotman

Exploratory Data Analysis: Scatter Plots (1/2)

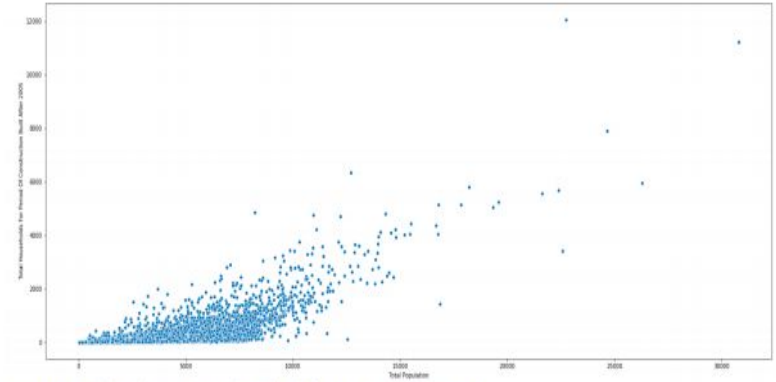


Figure 13 - Total Population vs. Total Households for Construction after 2005

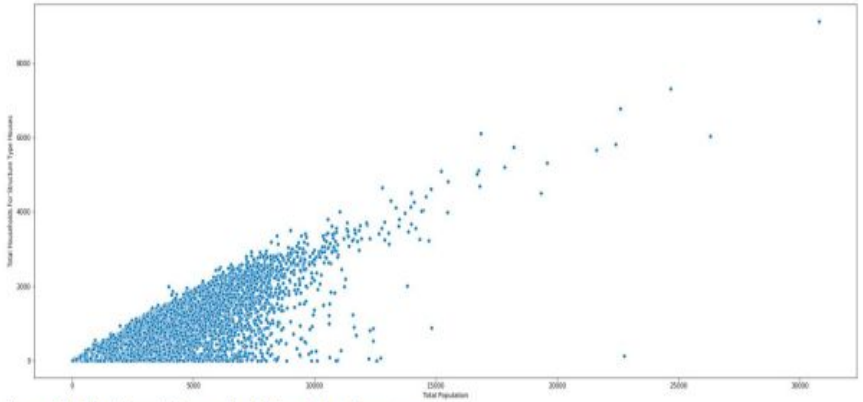


Figure 14 - Total Population vs. Total House Type Structures

Firstly, the households build after the year 2005 is strongly positively correlated with the total population.

Secondly, the households with structure type houses exhibits a strong positive correlation with the total population

Rotman

Exploratory Data Analysis: Scatter Plots (2/2)

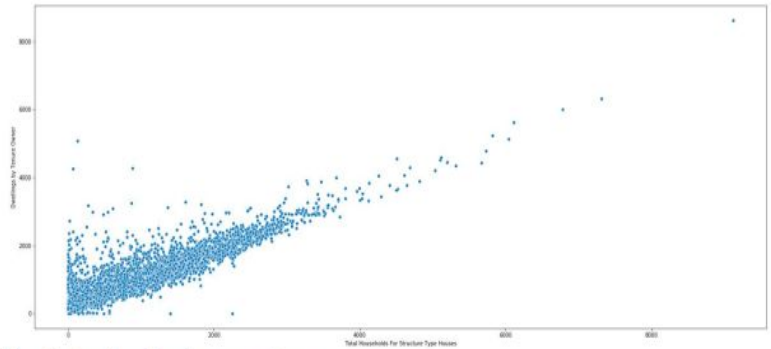


Figure 15 - Total House Type Structures vs. Owners

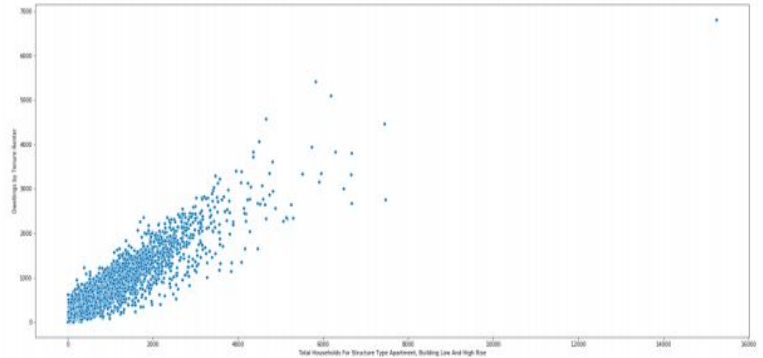


Figure 16 - Total Apartment Type Structures vs. Renters

It can be seen that House type structures are also strongly correlated with the Dwelling by Tenure Owner, indicating that the more owners tend to be living in house types rather than condominiums in buildings.

Strong positive correlation between Tenure Renters and Apartment type structures, indicating that more individuals on average that live in condominium buildings tend to be renters.

Appendix

Rotman

Segmentation Model

Performing model optimizations...
Cluster: 1
Estimator: Linear Regression
Best params: {'clf__copy_X': True, 'clf__fit_intercept': True}
Test set RMSE score for best params: 17404.741
Estimator: Decision Tree
Best params: {'clf__criterion': 'mse'}
Test set RMSE score for best params: 19171.941
Estimator: KNN
Best params: {'clf__n_neighbors': 17}
Test set RMSE score for best params: 17514.557
Estimator: SVR
Best params: {'clf__C': 10, 'clf__kernel': 'linear'}
Test set RMSE score for best params: 19054.725
Estimator: Gradient Boosting Regression
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'lad', 'clf__n_estimators': 300}
Test set RMSE score for best params: 16628.483
Regressor with best test set accuracy: Gradient Boosting Regression

Cluster: 2
Estimator: Linear Regression
Best params: {'clf__copy_X': True, 'clf__fit_intercept': True}
Test set RMSE score for best params: 19737.766
Estimator: Decision Tree
Best params: {'clf__criterion': 'mse'}
Test set RMSE score for best params: 34574.710
Estimator: KNN
Best params: {'clf__n_neighbors': 20}
Test set RMSE score for best params: 20954.298
Estimator: SVR
Best params: {'clf__C': 10, 'clf__kernel': 'linear'}
Test set RMSE score for best params: 22674.469
Estimator: Gradient Boosting Regression
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'lad', 'clf__n_estimators': 200}
Test set RMSE score for best params: 19351.322
Regressor with best test set accuracy: Gradient Boosting Regression

Appendix

Rotman

Segmentation Model

Cluster: 3
Estimator: Linear Regression
Best params: {'clf__copy_X': True, 'clf__fit_intercept': True}
Test set RMSE score for best params: 18919.312
Estimator: Decision Tree
Best params: {'clf__criterion': 'mse'}
Test set RMSE score for best params: 23474.508
Estimator: KNN
Best params: {'clf__n_neighbors': 10}
Test set RMSE score for best params: 19275.042
Estimator: SVR
Best params: {'clf__C': 10, 'clf__kernel': 'linear'}
Test set RMSE score for best params: 21909.535
Estimator: Gradient Boosting Regression
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'lad', 'clf__n_estimators': 200}
Test set RMSE score for best params: 17686.090
Regressor with best test set accuracy: Gradient Boosting Regression

Cluster: 4
Estimator: Linear Regression
Best params: {'clf__copy_X': True, 'clf__fit_intercept': True}
Test set RMSE score for best params: 20162.986
Estimator: Decision Tree
Best params: {'clf__criterion': 'mse'}
Test set RMSE score for best params: 27330.917
Estimator: KNN
Best params: {'clf__n_neighbors': 4}
Test set RMSE score for best params: 19953.755
Estimator: SVR
Best params: {'clf__C': 10, 'clf__kernel': 'linear'}
Test set RMSE score for best params: 24324.974
Estimator: Gradient Boosting Regression
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'lad', 'clf__n_estimators': 300}
Test set RMSE score for best params: 23330.710
Regressor with best test set accuracy: KNN

Appendix

Rotman

Segmentation Model

Cluster: 5
Estimator: Linear Regression
Best params: {'clf__copy_X': True, 'clf__fit_intercept': True}
Test set RMSE score for best params: 31145.849
Estimator: Decision Tree
Best params: {'clf__criterion': 'mse'}
Test set RMSE score for best params: 39710.209
Estimator: KNN
Best params: {'clf__n_neighbors': 20}
Test set RMSE score for best params: 29233.750
Estimator: SVR
Best params: {'clf__C': 10, 'clf__kernel': 'linear'}
Test set RMSE score for best params: 33017.427
Estimator: Gradient Boosting Regression
Best params: {'clf__learning_rate': 0.1, 'clf__loss': 'lad', 'clf__n_estimators': 200}
Test set RMSE score for best params: 30056.512
Regressor with best test set accuracy: KNN