# A Data Analytics Model for Extended Real Estate Comparative Market Analysis

***Abstract*:** As part of real estate sales, a comparative market analysis is an analysis that is provided to most of the prospects and customers. Currently, this market analysis typically includes only nearby properties and is rather superficial. In the case of an investor, the list of properties in the analysis should include properties beyond the nearby area. In this study, data analytics tools for clustering, classification, and recommendation models were explored with the aim of producing a larger and more detailed list of properties to resolve these issues.

## Introduction

In the real estate market, buyers and sellers usually receive a Comparative Market Analysis (CMA) from their agents and this includes only nearby properties that are manually selected in most cases [7]. Investors who want to know similar information from different cities in a state currently must depend on multiple real estate agents to get this information which it also performed manually by the various real estate agents. To provide better insight to the investor, it would be valuable if there was a data analytics solution that could find the comparable properties in many cities in a state and also predict the future prices and/or rental estimates of the found comparable properties. In addition to this, a person or an agent has to go through many factors that are involved in the short listed properties to give a suggested properties list to the prospect. Currently online and offline solutions that are aiding the process of filtering real estate listing as per the user selected or provided criteria are handled mainly by online service providers and none of them are effectively using a data analytics approach or machine learning algorithms to tackle the situation.

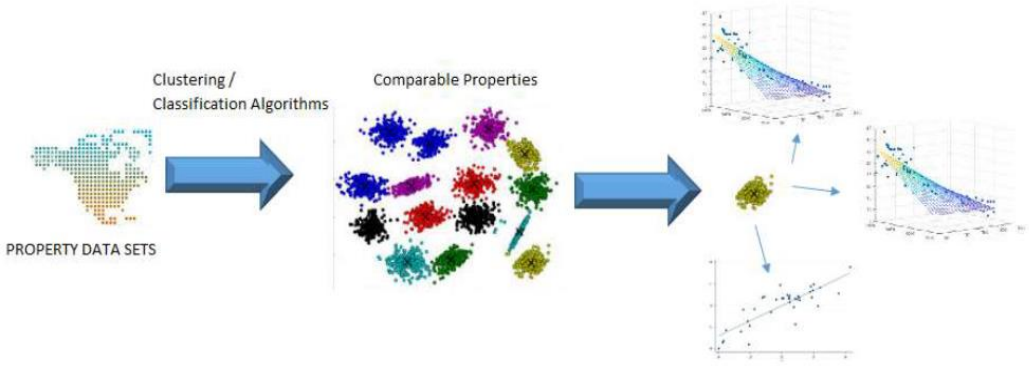



Figure 1 A Graphical Representation of Extended CMA.

A software solution or an algorithm that can help in automatically filtering the results intelligently considering all the facts can be very helpful. A data analytics model can be a solution for the above mentioned problem.

### Background

Whenever a client receives comparative market analysis from an agent, it will be prepared from comparable properties from the nearby areas. It is an important factor from the agents' point of view when preparing CMA (comparative market analysis) to include nearby properties. Most of the agents has access only to one or two real estate associations or they will be subscribed to only minimal amount of multiple listing providers. In addition to this, price of the real estate is considered

largely depend on the sales and price of the nearby properties [7, 8, 9]. But what happens when an investor consider purchasing a property that can be anywhere in a large area that can span areas such as north east or south west etc. This kind of search is not limiting to one agent. Then the users have to depend on multiple agents to get the property choices available to them according to their criteria in which all of these agents provide the nearby properties in their respective geographic area. Another way to get this is to depend on a multi-state or nationwide realtor company such as Century 21 or Coldwell Banker in which they assign many agents to come up with a list of properties from different places. However this also requires co-ordination between agents and arranging and rearranging multiple reports provided by multiple agents.

In some cases software may be available to give an automatic comparison or nearby properties but in some case it will be manually handpicked by agents. If the investor tries another way and searches himself online websites such as Zillow.com, Realtor.com, Trulia.com etc., still their searches mainly done by inputting a property and the site then will provide a list of properties nearby to compare. Here also it can be a tedious task to come up with properties that are matching or comparable to the selected choices of the investor [10, 11, 12]. The search itself can be a huge task. Then they have to spend additional time to consider various factors of the properties given in the list and make a choice. Instead of these time consuming and cumbersome methods of selecting a property for an investor, it will be an easy process if a data analytics approach using a machine learning or deep learning model were to provide or suggest a list of properties in any given geographical area. The list of properties can be input to the model to train it and after clustering or classifying according to correlating features, it can be submitted to a recommendation model which will analyze various factors and provide best options to the investor according to a recommendation score that the model is producing.

**Extended Comparative Market Analysis**

While comparative market analysis takes nearby properties to come up with a list of property that are close enough to present to client, Extended Comparative market analysis's main target audience is investors. While residential buyers mainly concentrate on a specific geographic area called nearby area, investor's property search area doesn't need to be limited to nearby area. Investors can buy property anywhere in the nation wherever a close match can be found. This close match can be in the sense of price, return on investment, rent, easiness to rent, easy to sell etc. etc. depending on the investor. In other words, comparative market analysis produces a list of properties that are nearby while Extended Comparative Market analysis produces a list of properties that are nearby and beyond nearby area depends on the user's choice of geographical area. This geographic area can extended to anywhere in the nation where the investors area of business. Note that both comparative and extended comparative market analysis make use of sold price [7, 8, 9]. However, there are other factors that depends on the pricing and value of real estate ownership when it goes beyond nearby area. Extended Comparative Market Analysis includes properties in nearby areas as well as beyond nearby areas. The beyond nearby areas can be anywhere where comparable economic, financial, marketing conditions can come a close match.

**Nearby area**

Nearby area is a concept that uses in traditional comparative market analysis as well as all real estate similar property searches. A nearby area generally is a geographic area that is close to the user's interested property location. However, this nearness can vary depends on users requirements. It can be a small area of a development site, or a large are that includes one or two zip codes. In addition a nearby area can even include a city or a specific school zone. The nearby area can be dictated by many factors. It can be a user defined area, it can be a small development, and it can be an area around the property that are enough similar properties can be found. It can be an area coming under a zip code or multiple zip code. It can be include a city where the property resides. However, In the case of Extended Comparative market analysis, the above mentioned geographic area concept in ordinary comparative market analysis is ignored and the selection of the properties can be anywhere depends on the data involved.

**Technologies and Platforms**

Extended Comparative Market Analysis is mainly using Clustering, Classification, and Regression Analysis machine and deep learning algorithms to come up with its final list of properties. The software technologies such as Tesnorflow, Keras, Python libraries, R, Hadoop, and Spark are used in various steps of the study. IBM Bluemix, AWS, Google Coud/Colab, Anaconda,

Weka etc. were also used as part of the study in different steps. Resilient Distributed Datasets (RDD), Numpy Arrays, Pandas DataFrames and comma separated value files were used to handle the data during the processing of the software solution.

## Literature Review

There are lot of documents available for house price prediction. However almost all of them are not giving priority on a classification or clustering projects. A few documents explore the correlation among features that are usually considered along with house prices such as square feet, number of rooms, number of beds, price etc. Some of the literature review done are given here. Sifei Lu et al. investigate hybrid regression techniques for house price predication. A practical and composite data pre-processing and creative feature engineering method is examined in this paper. [1]. Debanjan Banerje et al. trying to predict the house price direction using machine learning techniques in their research. This research work applies various feature selection techniques such as variance influence factor, information value, principle component analysis and data transformation techniques such as outlier and missing value treatment as well as box-cox transformation techniques. In this study, the performance of the machine learning technique is measured by accuracy, precision, specificity, and sensitivity [2].Parasich Andrey Viktorovich et al. describing regression methods of machine learning they used in their work. In this work they used classic machine learning algorithms and their own original methods to predict the sale price. [3].S. Zhou, L. Cao and Y. Li has done a study on real estate investment environment using support vector machine. Natural environment, economic environment, policy environment, social culture environment features were considered for this study. [4].Huawang Shi, in his paper titled "Determination of Real Estate Price Based on Principal Component Analysis and Artificial Neural Networks" uses principal components analysis method of multi-dimensional statistical analysis as well as artificial neural networks to determine the price of real estate. [5]. HuXiaolong and Zhong Ming studied neural network's ability to resolve the problems associated with prediction of real estate price. [6].

## Data Collection and Preparation

In various steps of the study, different sub sets of the same data set have been used. The raw dataset with forty five features has been subjected to various data preparation and visualization steps. The features with better correlation have been selected for the data set for the main study. The comma separated value file of the raw data set allowed to manipulate the file using Excel. However, after initial data preparation steps, the source code of the project has included many steps that addresses the noise issues of the data set. Outlier values are removed from the data set after visualizations. Many of the rows contained spaces in the original raw data set have been removed. Null values also removed from the data set to further achieve a better cleanness of the data. attomdata.com [13] is the main source of the data used in this study

### Data Visualization

Data visualization is done using Python code in this study. Matplotlib, Seaborn, and Pandas framework are mainly used to create graphs. Univariate analysis has been done by creating one dimensional histogram using Pandas framework. One dimensional histogram (Figure 1) and density plots are created to visualize continuous numeric attributes.

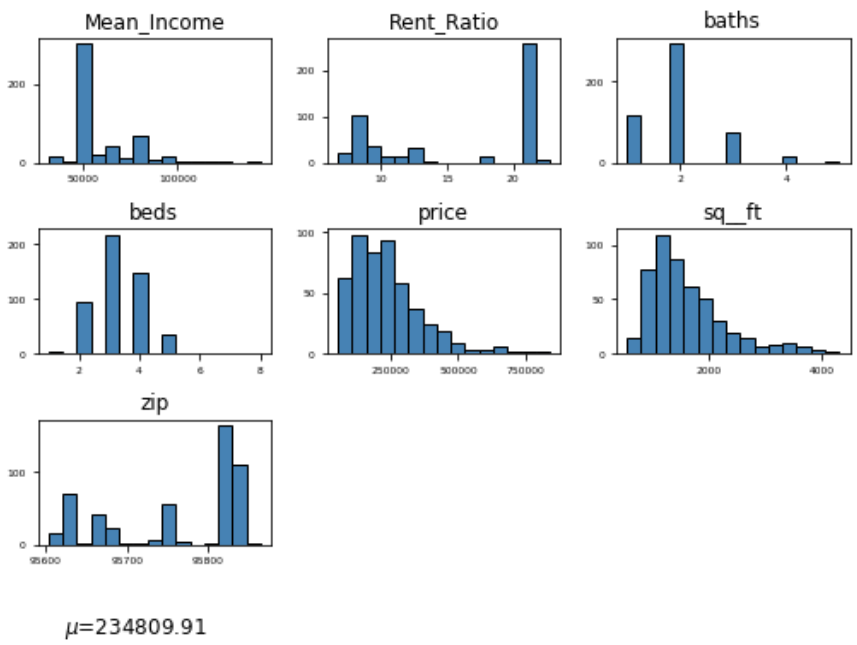



Figure 1 One Dimensional Graphs

Bar plots were created to visualize discrete categorical data attributes. To do multivariate analysis pairwise correlation matrix and heat map has been created. A multidimensional scatter plot with zip, price, and square feet was created to visualize data. In addition to this, size and color was used to represent number of baths and number of beds in the same graph.

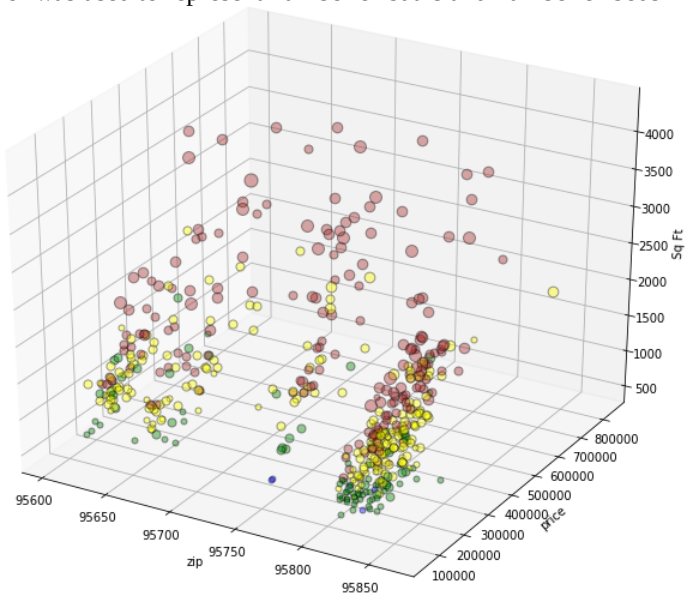



Figure 2. Graph 5D.

This made it as 5d graph (Figure 2). Pair-wise plots (Figure 3) against each attributes visualized the data effectively when it used along with correlation heat map [8].

**Exploratory Data Analysis**

The housing data set used has more than forty columns. However, after the data preparation, the columns selected for this study are street, city, zip, state, beds, baths, square feet, price, type, Mean Income, Inventory, rent ratio, and various other ratios and synthetic features that are derived during the execution of the code. Various one dimensional and multi-dimensional graphs including a five dimensional graph was created and studied.

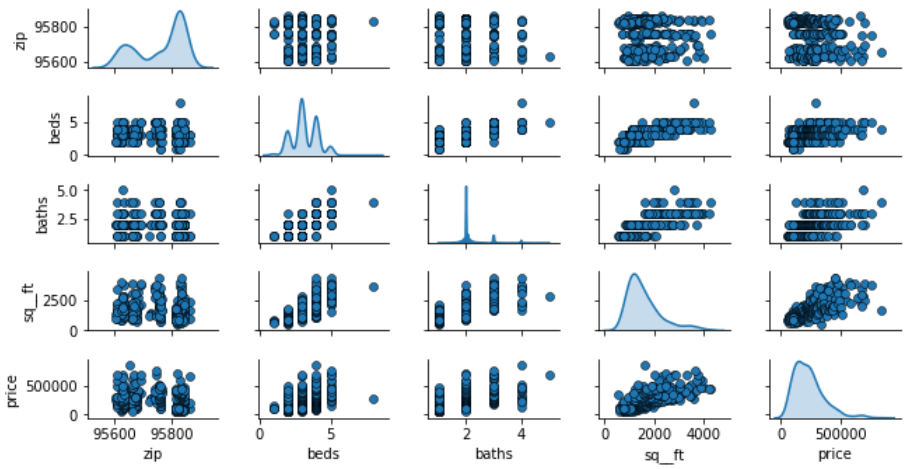



Figure 3 Pair-wise Plots

Histogram square feet vs price (Figure 4) and a density plot for square feet vs price (Figure 5) shows a normal skew with minor issues with some zips. A scatter plot was created using zip, price, square feet data, beds, and bath. The color and size was used to show beds and baths in this graph. The graph has clearly shown the distribution of properties.

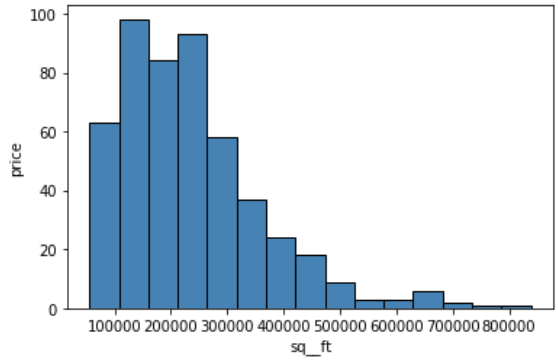



Figure 2 Histogram square feet vs price

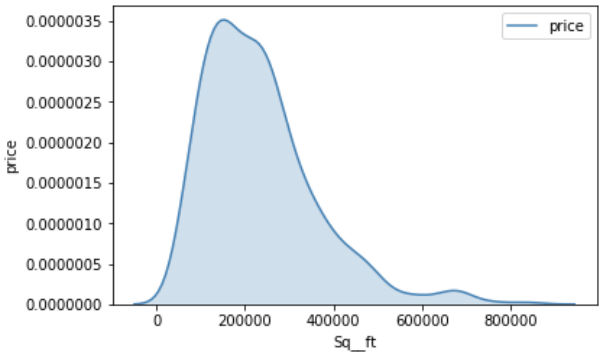



Figure 5 density plot for square feet vs price

As zip is not a major factor in the model, no data were removed because of the minor abnormality in the bell curve.

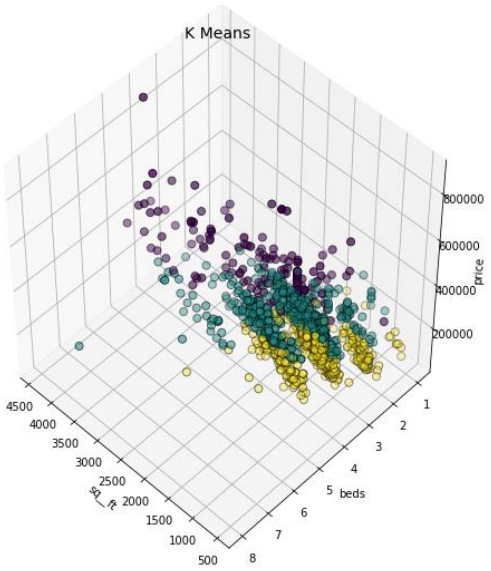



Figure 6. A Sample K-Means Clustering Graph Created During the Study.

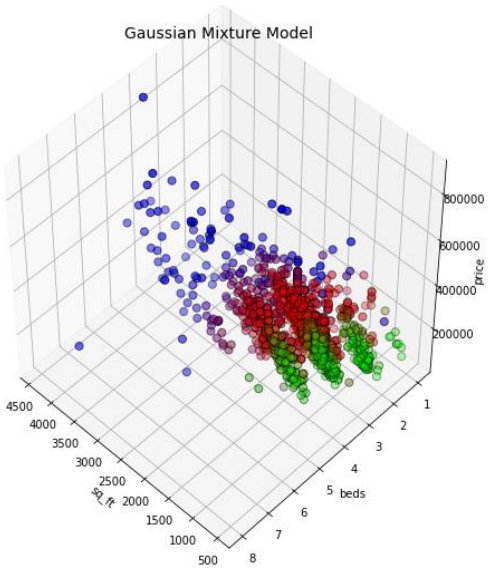



Figure 7. A Sample Gaussian Mixture Clustering Graph Created During the Study.

**Feature Engineering**

Relationship between available features were studied in the data set. A correlation heat map (Figure 8) was used to find the correlation between features. Street, Latitude, and Longitude correlation was poor and hence not used in almost all of the algorithms. The percentage of correlation between square feet and price was expected to be significant.

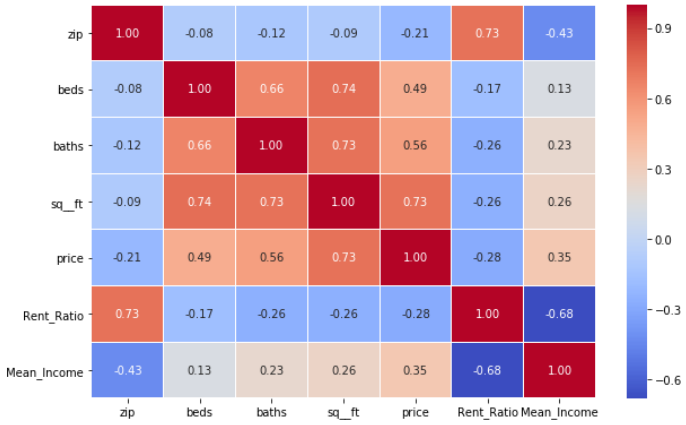



Figure 8 Correlation Heat map

However in the available data the correlation between square feet and price was only up to 31 percentage. At the same time, there is a better correlation percentage between other features such as beds, baths, and square feet. Therefor beds, baths, and square feet were also used in a significant way in the algorithms used.

## Methodology

Machine learning algorithms are implemented to the task of filtering the property lists according to various criteria in different levels of processing and finally to come up with a small list of matching properties

Extended Comparative market analysis methodology involve many steps of grouping or clustering based on various types of features. As the model has to come up with a list of properties that are beyond the nearby area, geographical features are ignored when the clustering or grouping takes place and filter properties. However, the algorithm has to find the matching economies or locations based on the economic and financial features and its relationship to price. Once this step is done, the algorithm gives importance to real property features and find its matches. After the real property features then comes the amenity and other features. In the category of other features, anything specific to geographical identifiers are ignored to avoid the grouping based on geographical boundaries.

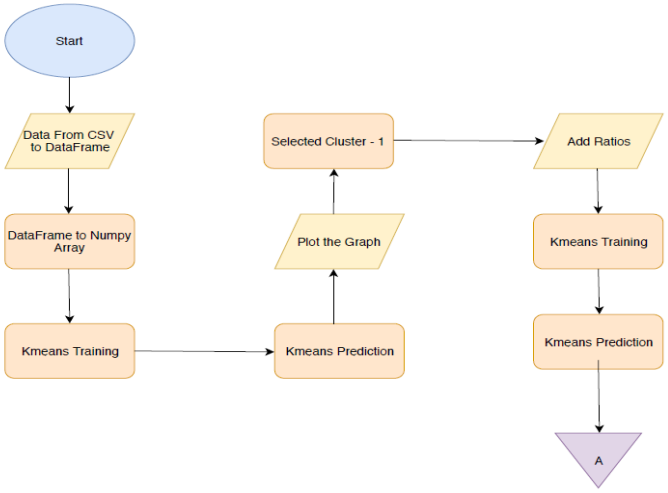



Figure 9. Flow Chart 1.

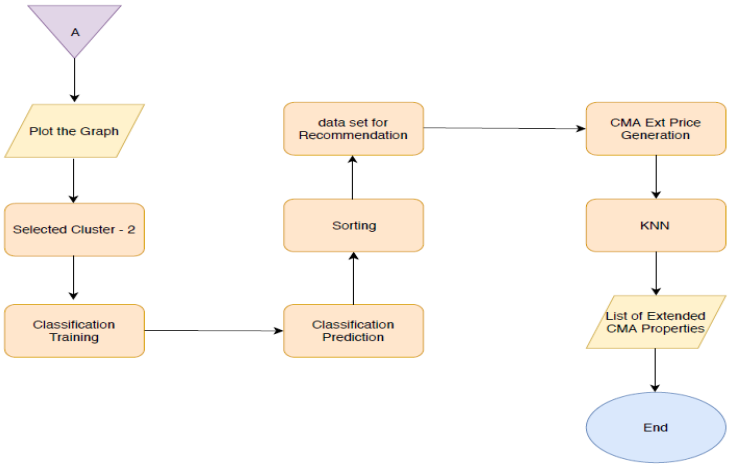



Figure 10. Flow Chart 2.

**Clustering**

The clustering algorithm used in the model is K-Means clustering which can show us possible cluster (or K) there are in the dataset. The algorithm then iteratively moves the k-centers and selects the data points that are closest to that centroid in the cluster [14].

1. Initialize **cluster centroids** $\mu_1, \mu_2, \ldots, \mu_k \in \mathbb{R}^n$ randomly.
2. Repeat until convergence: {

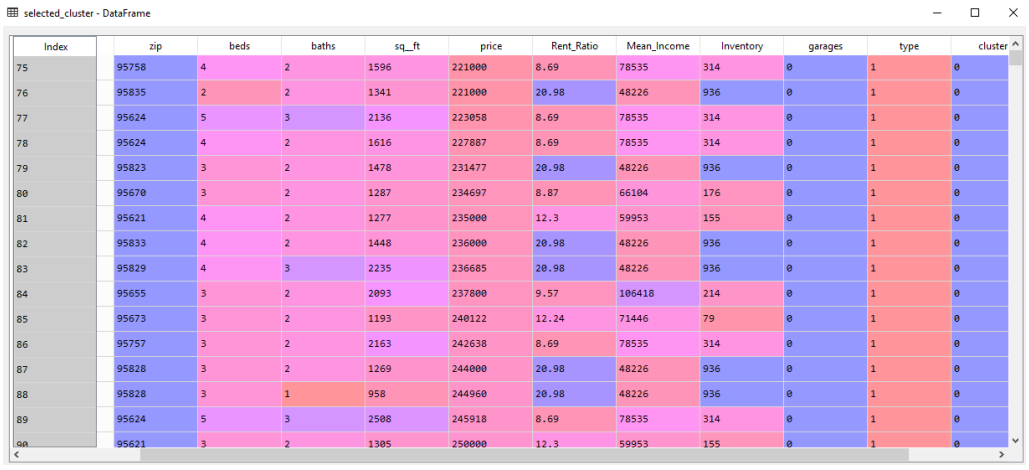
   For every $i$, set

   $$c^{(i)} := \arg\min_j ||x^{(i)} - \mu_j||^2.$$

   For each $j$, set

   $$\mu_j := \frac{\sum_{i=1}^{m} 1\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^{m} 1\{c^{(i)} = j\}}.$$

   }



selected_cluster - DataFrame

| Index | zip | beds | baths | sq__ft | price | Rent_Ratio | Mean_Income | Inventory | garages | type | cluster |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 75 | 95758 | 4 | 2 | 1596 | 221000 | 8.69 | 78535 | 314 | 0 | 1 | 0 |
| 76 | 95835 | 2 | 2 | 1341 | 221000 | 20.98 | 48226 | 936 | 0 | 1 | 0 |
| 77 | 95624 | 5 | 3 | 2136 | 223058 | 8.69 | 78535 | 314 | 0 | 1 | 0 |
| 78 | 95624 | 4 | 2 | 1616 | 227887 | 8.69 | 78535 | 314 | 0 | 1 | 0 |
| 79 | 95823 | 3 | 2 | 1478 | 231477 | 20.98 | 48226 | 936 | 0 | 1 | 0 |
| 80 | 95670 | 3 | 2 | 1287 | 234697 | 8.87 | 66104 | 176 | 0 | 1 | 0 |
| 81 | 95621 | 4 | 2 | 1277 | 235000 | 12.3 | 59953 | 155 | 0 | 1 | 0 |
| 82 | 95833 | 4 | 2 | 1448 | 236000 | 20.98 | 48226 | 936 | 0 | 1 | 0 |
| 83 | 95829 | 4 | 3 | 2235 | 236685 | 20.98 | 48226 | 936 | 0 | 1 | 0 |
| 84 | 95655 | 3 | 2 | 2093 | 237800 | 9.57 | 106418 | 214 | 0 | 1 | 0 |
| 85 | 95673 | 3 | 2 | 1193 | 240122 | 12.24 | 71446 | 79 | 0 | 1 | 0 |
| 86 | 95757 | 3 | 2 | 2163 | 242638 | 8.69 | 78535 | 314 | 0 | 1 | 0 |
| 87 | 95828 | 3 | 2 | 1269 | 244000 | 20.98 | 48226 | 936 | 0 | 1 | 0 |
| 88 | 95828 | 3 | 1 | 958 | 244960 | 20.98 | 48226 | 936 | 0 | 1 | 0 |
| 89 | 95624 | 5 | 3 | 2508 | 245918 | 8.69 | 78535 | 314 | 0 | 1 | 0 |
| 90 | 95621 | 3 | 2 | 1305 | 250000 | 12.3 | 59953 | 155 | 0 | 1 | 0 |

Figure 3 Pandas DataFrame with First Cluster Column

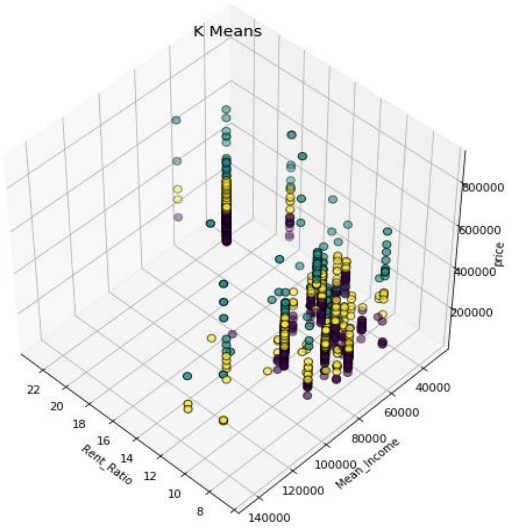



Figure 12 Cluster Graph Created During Frist Cluster Formation

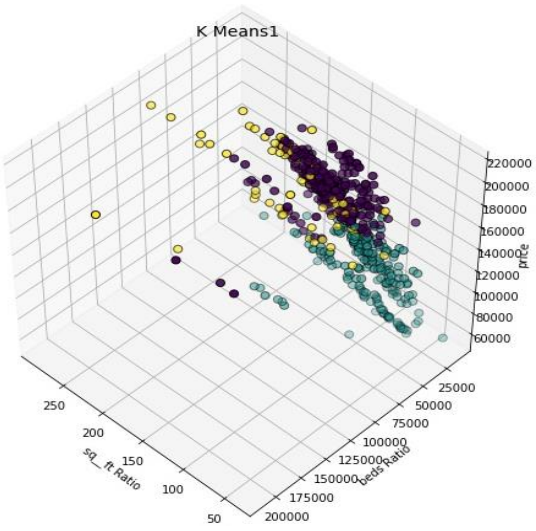



Figure 13 Cluster Graph During the Second Cluster Formation

The graph shown here in Figure 36 is created during the execution of the Extended CMA code and the optimal value of k is 3 and thereby 3 is the number of clusters used as value of k in the code [17].

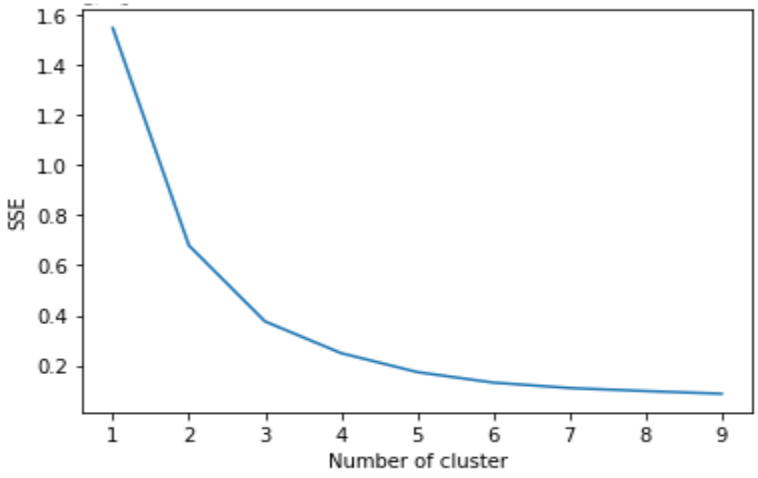



Figure 14 Scikit K-means clustering Performance Measure

**Classification**

A neural network for binary classification was used in the study. There are three input variables and one output variable. According to the heatmap created using the correlation for the dataset and the seaborn pair plot showed the relationship among the variables and were fit for the use. Keras sequential model is used to build the neural network. After standardizing the input feature, divided the data into training and testing data set.

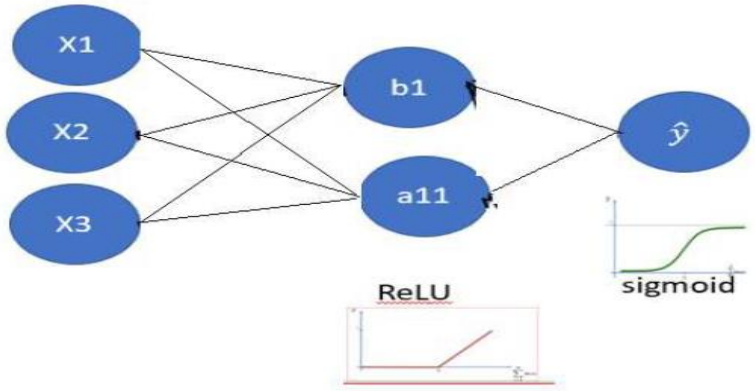



Figure 15 Neural Network Layer

A graphical representation is shown here that explains the neural network layer (Figure 38). As we have 3 input features and one target variable, there is 1 Hidden layer and the hidden layer has 3 nodes. ReLu is the activation function for hidden layer. As this is a binary classification problem we will use sigmoid as the activation function. Dense layer implements Output = activation (dot (input, kernel) + bias). Kernel is the weight matrix. Kernel initialization defines the way to set the initial random weights of Keras layers .As this is a binary classification problem, used binary_crossentropy to calculate the loss function between the actual output and the predicted output. To optimize the neural network Adam is used. Adam stands for Adaptive moment estimation. Adam is a combination of RMSProp + Momentum. Momentum takes the past gradients into account in order to smooth out the gradient descent.

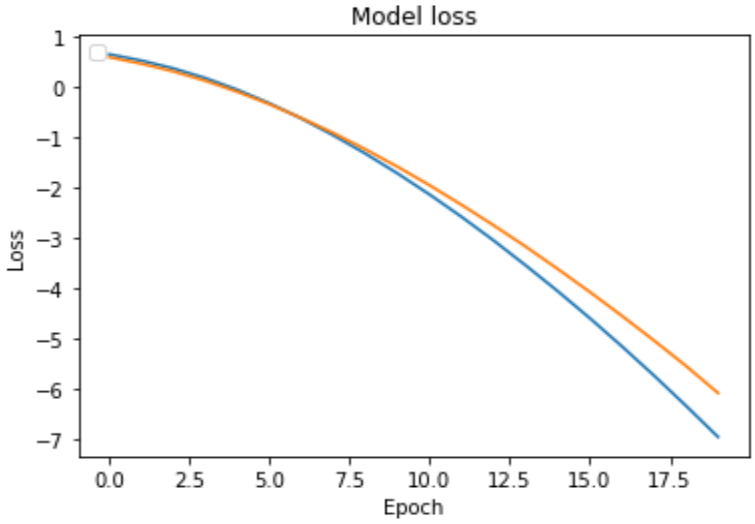



Figure 16 Loss Vs Epoch Graph for Training and Testing Data

The model used a batch size of 10. This implies that we use 10 samples per gradient update. Iterates over 20 epochs to train the model. An epoch is an iteration over the entire data set. A graph is drawn to show the model loss (Figure 16). Number of epoch is shown on the x axis and loss is shown on the y axis. Both training and testing loss is show as two separate lines in the graph [15, 16].

**Recommendation**

The final algorithm used in the model for recommendation is K-nearest neighbors (KNN). K-nearest neighbors is a supervised algorithms that is easy to implement for the purpose of recommendation systems.

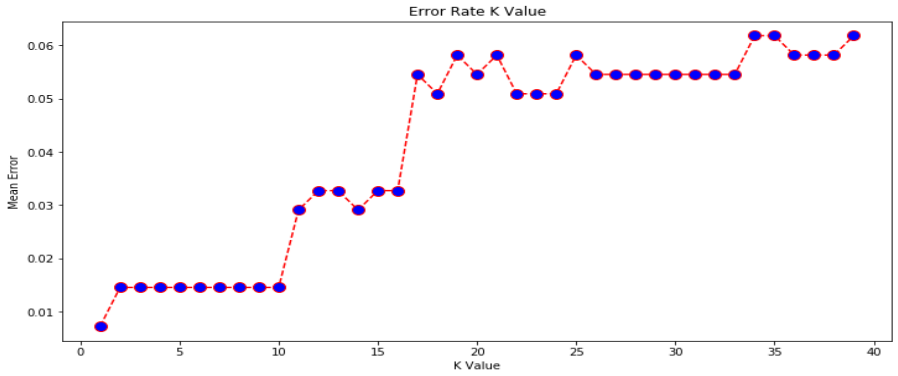



Figure 17 Mean Error Vs K value graph

Sklearn NearestNeighbors library is used in the model to do the K-nearest neighbors algorithm to recommend the properties.

## Results

The model successfully produced a list of comparable properties beyond the nearby area. The model effectively picked similar economies and compared and matched various features such as real estate features, financial features, Amenity features, different ratios to produce the final results. The order and the purpose of clustering, classification and recommendation algorithms used in the solution was effective which is verified by the results produced.

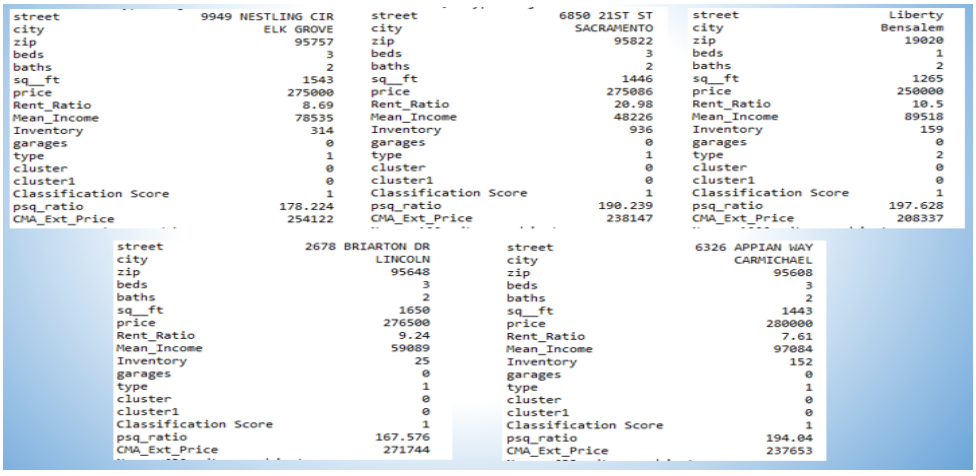



Figure 18 Final Display of Recommended Properties

A clustering model alone was not enough to produce the output and hence methodology included a classification and finally a recommendation model to come up with the final result. Each model has used different features and relationship to optimize the result. When the clustering models used economic and financial feature in the beginning step, and basic feature ratios in the second step, classification used basic real estate features. At the same time recommendation model mainly made use of square feet, beds, bath, and price to come up with list of properties that are comparable to the user selected property.

## Conclusions

The Extended CMA algorithm generates a set of properties beyond nearby areas. An application developed using the Extended CMA algorithm can address the need of investors beyond a small geographical area. An application developed using Extended CMA algorithm can eliminate the dependency to multiple realtors to find properties in a large area. An application developed using Extended CMA algorithm can automate real estate search in terms of geographical area, number of properties etc. The algorithm considers a wide range of features and synthetic features. Extended CMA can provide a far greater number of properties to the user than the regular CMA and existing real estate portal searches.

## References


[1] Lu, S., Li, Z., Qin, Z., Yang, X., & Goh, R. S. (2017). A hybrid regression technique for house prices prediction. *2017 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*. doi:10.1109/ieem.2017.8289904USA: Abbrev. of Publisher, year, ch. *x*, sec. *x*, pp. *xxx–xxx.*

[2] Banerjee, D., & Dutta, S. (2017). Predicting the housing price direction using machine learning techniques. *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)*. doi:10.1109/icpcsi.2017.8392275.

[3] Viktorovich, P. A., Aleksandrovich, P. V., Leopoldovich, K. I., & Vasilevna, P. I. (2018). Predicting Sales Prices of the Houses Using Regression Methods of Machine Learning. *2018 3rd Russian-Pacific Conference on Computer Technology and Applications (RPC)*. doi:10.1109/rpc.2018.8482191

[4] S. Zhou, L. Cao and Y. Li, "Evaluation model for Real Estate Investment Environment Based on SVM," *2008 International Workshop on Modelling, Simulation and Optimization*, Hong Kong, 2008, pp. 192-195.

[5] Shi, H. (2009). Determination of Real Estate Price Based on Principal Component Analysis and Artificial Neural Networks. 2009 *Second International Conference on Intelligent Computation Technology and Automation***:** 314-317.

[6] Hu Xiaolong and Zhong Ming, "Applied research on real estate price prediction by the neural network," *2010 The 2nd Conference on Environmental Science and Information Application Technology*, Wuhan, 2010, pp. 384-386.

[7] A. Barone, “Comparative Market Analysis,” Investopedia, 18-Nov-2019. [Online]. Available: https://www.investopedia.com/terms/c/comparative-market-analysis.asp. [Accessed: 20-Nov-2019].

[8] J. Kimmons, “The Real Estate CMA,” The Balance Small Business, 08-May-2019. [Online]. Available: https://www.thebalancesmb.com/comparative-market-analysis-in-real-estate-2866366. [Accessed: 20-Nov-2019].

[9] K. Treece , “Comparative Market Analysis: Ultimate Guide to a CMA in Real Estate,” Fit Small Business, 22-Feb-2019. [Online]. Available: https://fitsmallbusiness.com/comparative-market-analysis/. [Accessed: 18-Nov-2019].

[10] Zillow. [Online]. Available: http://www.zillow.com/. [Accessed: 22-Nov-2019].

[11] “Find Real Estate, Homes for Sale, Apartments & Houses for Rent: realtor.com®,” Find Real Estate, Homes for Sale, Apartments & Houses for Rent | realtor.com®. [Online]. Available: https://www.realtor.com/. [Accessed: 22-Nov-2019].

[12] “Discover a place you'll love to live,” Trulia Real Estate Search. [Online]. Available: https://www.trulia.com/. [Accessed: 22-Nov-2019].

[13] “Property Characteristics,” ATTOM Data Solutions. [Online]. Available: https://www.attomdata.com/data/property-data/property-characteristics/. [Accessed: 18-Nov-2019].

[14] S.S. Nazrul, “Clustering Based Unsupervised Learning,” Medium, 15-May-2018. [Online]. Available: https://towardsdatascience.com/clustering-based-unsupervised-learning-8d705298ae51. [Accessed: 02-Dec-2019].

[15] 15SimoneSimone 2, Matias ValdenegroMatias Valdenegro 39.8k55 gold badges7575 silver badges9595 bronze badges, Rahul VermaRahul Verma 55322 silver badges1414 bronze badges, and Ashok Kumar JayaramanAshok Kumar Jayaraman 1, “Keras - Plot training, validation and test set accuracy,” Stack Overflow, 01-Mar-1967. [Online]. Available: https://stackoverflow.com/questions/41908379/keras-plot-training-validation-and-test-set-accuracy. [Accessed: 03-Dec-2019].

[16] Visualization - Keras Documentation. [Online]. Available: https://keras.io/visualization/. [Accessed: 03-Dec-2019].

[17] O. Prakash, “Scikit K-means clustering performance measure,” Stack Overflow, 01-Jun-1967. [Online]. Available: https://stackoverflow.com/questions/43784903/scikit-k-means-clustering-performance-measure. [Accessed: 02-Dec-2019].