# REAL ESTATE INVESTMENT

**Group No**.: Group 21

**Student Names**:

Akash Parekh

Honglin Wu

# I. Introduction & Purpose

## 1.1 Background

Real estate properties are attractive option for safe investment. It is an investment which does not decline in value rapidly. However, the latest recession shows that real estate prices cannot keep rising all the time. In different areas, fluctuations in prices are different. These risks need to be constantly monitored and managed. The company, who plan to invest on real estate, has to evaluate real estate property very critically, and consider it in long-term.

In order to make an appropriate decision to reach the maximum profit, the investment company should find a proper tool to analyze data. All the problems to be solved depend on the correct analyses of real estate data.

During the past few decades, the pattern recognition and machine-learning communities have greatly expanded their areas of application and the kind information to be extracted. The database community joined the endeavor in early 1990 and a new multi-disciplinary field began, which we now call data mining. It is the process of extracting valid, previously unknown, comprehensible and actionable information from large database and using it to make decisions, the combination of modern artificial intelligence and statistics. The methods of data mining include neural network, heredity arithmetic, decision-tree arithmetic, rough set arithmetic, fuzzy sets theory and so on. The main task of different data mining activity can be divided into four kinds: relevancy analysis, clustering analysis, time-sequence model and prediction, deviation analysis. Data mining can help us understand the running mechanism of the object, discover the future trend and make relative prediction. All the mined information will be very useful for final decision.

The typical uses of data mining technology in real estate investment are showed in table 1:

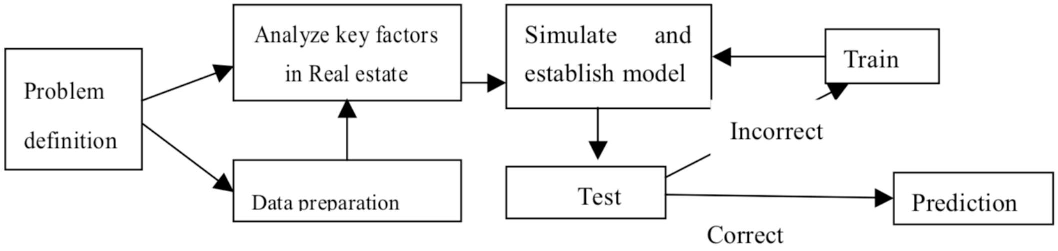


Table 1

## 1.2 Purpose

There is a company who wants to choose a most profitable location to do the real estate investment, so it uses Data Mining method to do a project to make prediction and comparison. This project will use R to find appropriate location to analysis current house prices, thereby forecasting the best location for investment. After the evaluation, the company can make the final decision.

## II. Data Preparation and Preprocessing

**Data resource:** the data was obtained from Zillow.com and https://www.kaggle.com/goldenoakresearch/us-household-income-stats-geo-locations#Kaggle_deffs.pdf

https://www.zillow.com/research/data/

The data represents the median property values for various regions and locations, cities show the change of house prices according to the locations.

**Variable definitions:**

1. id

Description: The Id of the location of which you are analyzing

Example: 0101909

2. State_Code

Description: The state code reported by the U.S. Census Bureau for the specified geographic location.

Example: 01

3. State_Name

Description: The state name reported by the U.S. Census Bureau for the specified geographic location.

Example: Alabama

4. State_ab

Description: The abbreviated state name reported by the U.S. Census Bureau for the specified geographic location.

Example: AL

5. County

Description: The county name reported by the U.S. Census Bureau for the specified geographic location.

Example: Calhoun County

6. City

Description: The city name reported by the U.S. Census Bureau for the specified geographic location.

Example: Alexandria

7. Place

Description: The place name reported by the U.S. Census Bureau for the specified geographic location.

Example: Alexandria

8. Type

Description: The place Type reported by the U.S. Census Bureau for the specified geographic location.

Example: CPD

9. Primary

Description: Defines whether the location is a geographic location or a track and block group.

Example: Place

10.Zip Code

Description: The zip code reported by the U.S. Census Bureau of the closest geographic location with a zip code.

Example: 36250

11.Area Code

Description: The zip code reported by the U.S. Census Bureau of the closest geographic location with a zip code.

Example: 256

12.ALand

Description: The Square area of land at the geographic or track location.

Example: 28834974

13.AWater

Description: The Square area of water at the geographic or track location.

Example: 33099

14.Lat

Description: The mean household income of the specified geographic location.

Example: 33.760819

15.Lon

Description: The standard deviation of the household income for the specified geographic location.

Example: -86.872008

16.Mean

Description: The mean household income of the specified geographic location.

Example: 93216

17.Median

Description: The median household income of the specified geographic location.

Example: 92686

18.Stdev

Description: The standard deviation of the household income for the specified geographic location.

Example: Albertville

19.Households

Description: The number of households used in the statistical calculations

20. Property median values

Description: The median house value used in calculations.

21. Rental median values

Description: The median rent value used in calculations.

The variable chosen has been careful analyzed and stated. The real estate investment is a broad topic of investment and the variables mentioned above plays a major role of significance in the investment decision. The technique used for finding the solution relies on the above stated variables.

## III. Data Mining Techniques and Implementation

**Solution Statement**: The solution design is given below for the above model.

- Variable Elimination & Selection
- Data Normalization
- Separate Dataset
- Set Module –>Using KNN
- Verification

**Assumption:** The factors which affect the real estate investment in the USA is divided in terms of two areas:

- MIDWEST MARKET
- COASTAL MARKET

The investment decision is always based on the coastal markets since the rate of appreciation of value of property is highest on the coastal than mid-west where the rental value is higher. Since the real estate company needs better value on investment so the selection will be based on the coastal markets. This leads to elimination of large number of mid-west states from the data examined.

As discussed about the stated variables, we need to use the above variables to find an appropriate solution design for finalizing the investment for maximizing the profit.
The solution design consists of data selection that is variable selection as performed above and to perform of data normalization. The key attributes of the data are variable

median house value, median household income and median rent value which have been converted to variable Monthly budget and the region name.

After selection and normalization we separate the data sets into two parts for training and validation. 30% of the data set is used for training and remaining is used for validation of the data. The validation technique is used to evaluate the performance of the model. The technique used for the implementation of the model is KNN.

The KNN technique is used, **KNN** is a **non-parametric, lazy** learning algorithm. Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point. The concept of non-parametric is not making any underlying assumptions of the data. The technique does not have explicit training phase but it needs it for testing of the data.

The main pros of the KNN method are:

- No assumptions about data—useful, for example, for nonlinear data
- Simple algorithm—to explain and understand/interpret
- High accuracy (relatively)—it is high but not competitive in comparison to better supervised learning models
- Versatile—useful for classification or regression

The main cons of KNN method are:-

- High memory requirement
- Stores all (or almost all) of the training data
- Prediction stage might be slow (with big N)
- Sensitive to irrelevant features and the scale of the data

KNN method does not learn any model and it makes prediction of the data between input sample and training data. After application of the method, the accuracy and mistakes are computed for the model. These are the data mining techniques and implementation of the model using KNN method.

## IV. Results

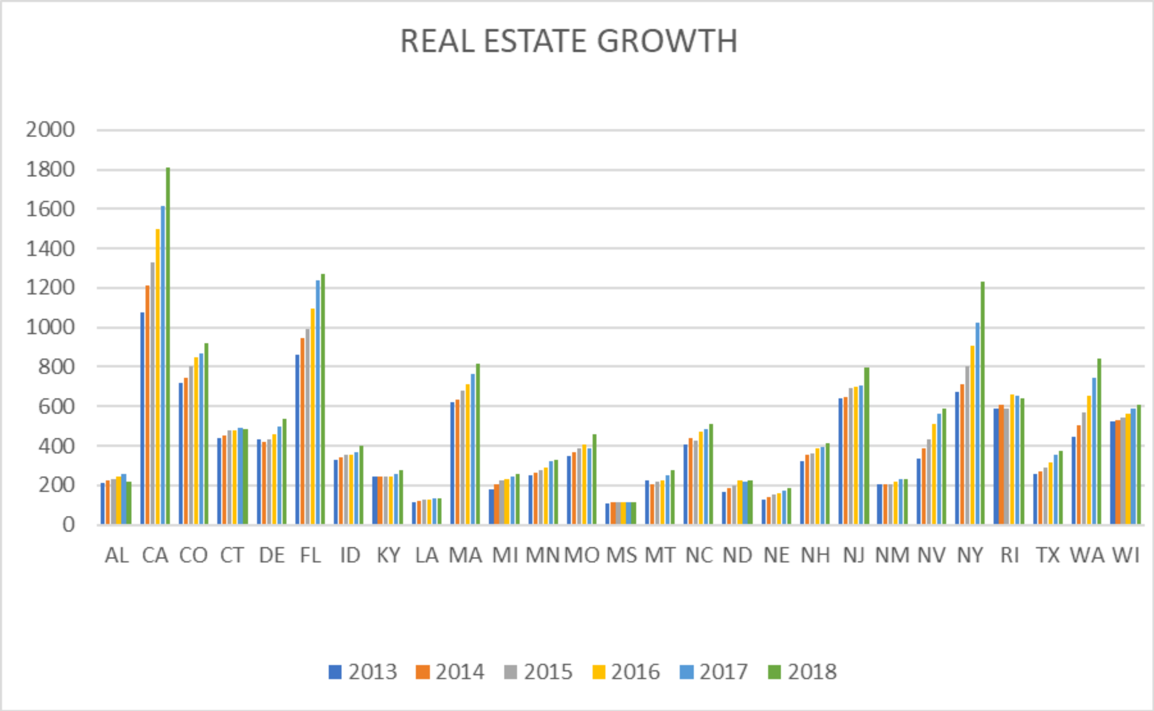
In terms of real estate investment, we cannot use regression models and time series model. The data variable selected does not have the indication of the assumptions. The regression model and time series used to the collection of past data where longer the duration of the data more are the chances of inaccuracy. Since the real estate is only type of investment model which works on the concept of cycles, federal interest rate, economy growth rate, rise in employment rates etc. All these factors with large variable change every year does not lead to accurate assumptions of the investment. In case of investing every $1 movement of economy can create huge difference in margin of profits.

### RESULT BASED ON DATA VISUALIZATION

As spoken above factors which affect the real estate investment in the USA is divided in terms of two areas:-

- MIDWEST MARKET
- COASTAL MARKET

The investment decision is always based on the coastal markets since the rate of appreciation of value of property is highest on the coastal than mid-west where the rental value is higher. Since the real estate company needs better value on investment so the selection will be based on the coastal markets. This leads to elimination of large number of mid-west states from the data examined. The following bar chart gives the general examination of the above data



As you can see the appreciation is highest in the coastal regions of the states over the years. We are concluding for investment in the coastal markets. According to data visualization we predict the selection of **coastal area of New York** due to present conditions of the real estate cycle being in the expansion phase and the highest appreciation % is being obtained in the state of New York near the coastal regions.

**RESULT BASED ON KNN**

Using the above obtained model and conditions of KNN. The value of k to be computed is very critical as lower the k more sensitive and more noise influence and higher the k can always have varied influence on the results. Using the above model, we obtain 4 locations for investment which can generate better rate of return on investment. These locations are:

**1.) New York**

**2.) California**

**3.) Washington**

4.) **New Jersey**.

All these locations belong near the coastal regions. The proportion of investment is in ascending order.

## V. Conclusion

The model adopted and the approach of KNN technique whose con and pros were adequately shown. The model errors and accuracy were computed. This investment idea was very general using this method. The regression model and time series not used because of larger duration of data leading to inaccuracy. The real estate cycle has movement of 16 years and depending on the phase the data can be misleading. We can reduce the data time by predicting the real estate cycle movement and make decision and forecast on basis of the above data for better accuracy and prediction using the above methods. The model assumptions of coastal and mid-west market are true with the fact it has been working from last 60 years based on previous data of real estate cycle in the US. The time series and regression can predict the price but the uncertainty due to large number of missing variables like the federal interest rate changes, economy rate, type of house , area , location , employment numbers, household income changes proper data available, and there are many factors which affect the decision hence to generalize and not depend on underlying and large number of data variables and variation we preferred to use the KNN technique to just obtain the ideal location without the stress of the underlying data and then other factors can be applied for the location in the state and finding the area and region ideal for growth and investment. Since real estate decision are based on factor, we can say the decisions obtained by the model can be inaccurate but in majority according to the trend the location obtained are bound to apply for better investment for profitability. The next scenario after searching the locations will be using predictive analytics by considering big data of the missing macro and micro economic factors for better predictions. Also, the main consideration is the amount of investment the company is willing to make and in what proportions in states. The model is perfect for making base decisions and then going for deeper analysis for better profitability.
We prefer to go deeper in the analysis and use large variations and sensitivity analysis to obtain better results.

## VII. References

https://www.kaggle.com/rossrco/passnyc-socio-economic-needs-index/notebook

Shmueli, G., Bruce, P. C., Yahav, I., Patel, N. R., & Lichtendahl, K. C. (2018). *Data mining for business analytics concepts, techniques, and applications in R*. Hoboken, NJ, USA: Wiley.

## Appendix: R Code for use case study

```
#Read Data
mydata=read.csv("City_MedianValuePerSqft_AllHomes.csv", head= “TRUE”)
#Preprocessing
 #Feature Elimination
mydata= mydata[, c(-1,-3,-4,-5)]
head(mydata)
 #data normalization
mydata=scale(mydata[,-1],center = TRUE, scale = TRUE)
head(mydata)
#separate data
set.seed(3)
ind= sample(2, nrow(mydata), replace = TRUE, prob= c(0.7,0.3))
train-data= mydata[ind==1,]
head(train-data)
summary(train-data)
test-data= mydata[ind==2,]
head(test-data)
summary(test-data)

#knn
trainControl <- trainControl(method="repeatedcv", number=10, repeats=3)
metric <- "Accuracy"
set.seed(1)
grid <- expand.grid(.k=seq(1,20,by=1))
```

```
fit.knn <- train(RegionName~., data=train-data, method="knn", metric=metric,
tuneGrid=grid,
             preProc=c("BoxCox"), trControl=trainControl, na.action=na.omit)
print(fit.knn)#k=50
plot(fit.knn)
#knn prediction
pred=predict(fit.knn,test-data)
confusionMatrix(pred,test-dat$RegionName)
```