Real estate data analysisusing principal component analysis and 'R'

Article *in* International Journal of Pure and Applied Mathematics · January 2018

CITATIONS
0

READS
660

5 authors, including:

Sailaja Kumar
Bangalore University
11 PUBLICATIONS **24** CITATIONS
SEE PROFILE

Suresh Kumar T.V.
Bharat Institute of Engineering and Technology
33 PUBLICATIONS **125** CITATIONS
SEE PROFILE

Real Estate Data Analysis using Principal Component Analysis and 'R'

K Sailaja Kumar, Kameshwari Soundarya, Harshitha R, D Evangelin Geetha and T V Suresh Kumar
sailajakumar.k@gmail.com, kameshwari.soundarya@gmail.com, harshitha3072@gmail.com, devangelin@gmail.com,
Department of Computer Applications, M S Ramaiah Institute of Technology, Bangalore, India.

***Abstract*—The primary focus of real estate data analysis is on analyzing the past data and then predicting the real estimates in further time. This analysis is used to figure out the right price for the property in view of buyer and seller. In this paper a methodology is proposed to estimate the real estate housing price based on various predictor variables using regression models. The models are then implemented using 'R' statistical tool. It is used to make data easy to explore and visualize. Principal component analysis (PCA) is used to emphasize the variations and bring out strong patterns in the real estate dataset. The accuracy of these three regression models M1, M2 and M3 are 0.3907, 0.3467 and 0.5825 respectively. From these accuracy values it is observed that the model M3 is giving better results in comparison to M1 and M2. Moreover M3 is constructed using the results obtained from the PCA technique employed in M2. PCA is identified as the powerful data science technique for predictor variable reduction in particular for real estate dataset attributes reduction analysis. Hence this technique is adopted for estimating the real estate property price accurately. Further the model can be extendable to other real estate data sets in future.**

I. INTRODUCTION

THE economic decline had a significant impact on the real estate community. One of the challenges with the real estate data analysis was to figure out the right price for the property in view of buyer and seller. The analysis of the real estate dataset and then identifying the most influential parameters that effect the house price is a very challenging task due to the high dimension of the dataset. Therefore in this paper, we present the best model to predict the house price with good accuracy and will help the customers to buy a house considering the influential factors.

This paper is focused specifically on Washington City's real estate data, with 159 rows and 330 columns for instance, the total-floors-in-building, price-per-sqft, number-of-beds, number-of-baths and many more. This data is prepared for generating the regression models after handling the missing values. Using principal component analysis (PCA) the significant predictor variables (columns in the real estate dataset) are identified which are useful in predicting the house price. PCA is commonly used to handle large datasets associated with the social sciences, market research, and other communities. It uses orthogonal transformation to convert a set of correlated predictor variables (columns in the real estate dataset) into a set of linearly uncorrelated variables called *'principal components' or 'predictor variables'*, from the real estate dataset. The goal of PCA is to explain the amount of house price variation considering the fewer number of these principal components.

In this paper, three regression models M1, M2 and M3 are constructed to identify the prevailing predictor variables used to estimate the housing price in view of owner of the house and the real estate agent. These models were developed based on regression analysis, principal component analysis (PCA) and with the selected original raw variables using R for the real estate price estimates respectively.

First regression model M1 is constructed from the training dataset considering the 101 numeric predictor variables. The target variable is house price. The performance of this model is evaluated using the test dataset and the accuracy of the model as 0.3907. This shows 39% price variation is due to these numeric predictor variables.

The second regression model M2 is based on PCA. Using this technique principal components Dim1 to Dim6 from the numeric predictor variables are extracted. These extracted principal components do not have any multicollinearity between them as they are orthogonal or perfectly uncorrelated. Using these components regression model is developed. The performance of this model is evaluated using the test dataset. The accuracy of this model is 0.3467. Approximately 35% of the variation in the house price is due to these principal components.

The third model M3 is built from model M2 using the selected numeric predictor variables based on their significance w.r.t principal components Dim1, Dim2 and Dim3. The performance of this model is evaluated using the test dataset. The accuracy of this model is 0.5825. Here 58% price variation is there due to these selected significant predictor variables.

There is a marginal improvement in the accuracy from first model to the third model. The second model explains only 35% of the price variation due to the complexity involved in the PCA technique. The model M3 is simple and gives more accurate results in comparison to other models M1 and M2. More over this model is based on the PCA results obtained in model M2. Thus PCA is a highly intuitive and powerful data science technique that can be used to construct predictive models to estimate the real estate property price.

'R' is a software useful for statistical computing and visualization. It has become the most popular and versatile language for data science. It is an interpreted language and comes with a command line interpreter - available for Linux,

Windows and Mac machines - but there are IDEs like RStudio or JGR to support development. RStudio is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting. The above mentioned models are implemented using R.

The paper is further organized as follows. In Section II literature on real estate data analysis is presented. Section III provides the methodology to collect and examine the real estate data and developing models to identify the influential factors for property price prediction. In Section IV the methodology illustration is provided. Experimental study that is carried out focusing the analysis of real estate dataset of Washington City, and results obtained using the proposed methodology is discussed in Section V. The conclusion and future work are presented in Section VI.

II. LITERATURE

Prediction of house prices based on NJOP houses in Malang city with regression analysis and particle swarm optimization (PSO) is discussed in [1]. PSO is used for selection of affect variables and regression analysis is used to determine the optimal coefficient in prediction. The result from this research proved combination regression and PSO is suitable and get the minimum prediction error.

Modeling and forecasting of land price in Chennai Metropolitan Area (CMA) in the state of Tamilnadu, India using multiple regression and neural network techniques is presented in [2]. Thirteen locations spread over CMA are selected at random as study areas. Both multiple regression and neural network models are validated and used to forecast the land price in CMA. Both the models are found to be well fit for the trend of land price; however the model using neural network shows better accuracy.

Regression models, using various features to have lower Residual Sum of Squares error are described in [3].

The predictive performance of artificial neural networks and multiple regression analysis for single family housing sales is compared in [5]. Multiple comparisons are made between the two data models in which the data sample size, the functional specification and the temporal prediction are varied. The authors have observed that ANN (Artificial Neural Networks) results better than MRA (Multiple Regression Analysis) when a moderate to large data sample size is used [4].

The functioning [5] involves a website which accepts customer's specifications and then combines the application of multiple linear regression algorithm of data mining. This application will help customers to invest in an estate without approaching an agent. It also decreases the risk involved in the transaction.

Prominent factors affecting property value in Indian housing market are described in [6]. Total nineteen variables were identified from literature as well as discussion with experts. Data obtained from the survey was processed using PCA which helped to decrease the number of variables into seven most important factors affecting value of real property in Indian housing market. The seven most important factors that affect Indian real property value included living conditions of residents are identified by the authors as; characteristics of housing; regional influence; utilities; age of property; economic, political and social influence; area and legal aspect.

In this paper, a methodology is proposed to estimate the real estate housing price considering various predictor variables using regression models. The models are then implemented using 'R' statistical tool.

III. METHODOLOGY

The proposed methodology construct the regression models to predict the housing price. The steps to be followed are:

Step1. Load the real estate data into R

Step2. Identify and handle the missing values

Step3. Identify numeric and categorical predictor Variables

Step4. Identify training dataset and test dataset from the sample dataset for constructing the three regression models M1, M2 and M3.

Step5. Model M1:
- Construct regression model with all the numeric predictor variables from the training dataset
- Evaluate the performance of the model using the test dataset

Step6. Model M2:
- Construct regression model with the orthogonal principal components derived from the original predictor variable using PCA
- Evaluate the performance of the model using the test dataset

Step7. Model M3:
- Construct regression model from the selected significant predictor variables obtained from the significant principal components in Model M2
- Evaluate the performance of the model using the test dataset

Step8. Compare the three models based on the accuracy levels and select the model with high accuracy.

Step9. Use the model selected in step 9 for predicting the housing price.

IV. ILLUSTRATING THE METHODOLOGY

Methodology presented in Section III is illustrated as follows.

A. Load the real estate data into R

Consider sample real estate dataset as .csv file with 932 rows representing the values under 7 columns (Variables) [7]. The screenshot is shown in Figure 1. The columns are designated as Distance-to-BT, Distance-to-Mall, Distance-to-Hospital, Carpet-Area, Built-up-Area, Parking-Status and House-Price. Load this dataset into R.

Distance_to_Taxi	Distance_to_Market	Distance_to_Hospital	Carpet_Area	Builtup_Area	Parking_Status	City_Category	Rainfall	House_Price
9796	5250	10703	1659	1961	Open	CAT B	530	6649000
8294	8186	12694	1461	1752	Not Provided	CAT B	210	3982000
11001	14399	16991	1340	1609	Not Provided	CAT A	720	5401000
8301	11188	12289	1451	1748	Covered	CAT B	620	5373000
10510	12629	13921	1770	2111	Not Provided	CAT B	450	4662000
6665	5142	9972	1442	1733	Open	CAT B	760	4526000
13153	11869	17811	1542	1858	No Parking	CAT A	1030	7224000
5882	9948	13315	1261	1507	Open	CAT C	1020	3772000
7495	11589	13370	1090	1321	Not Provided	CAT B	680	4631000
8233	7067	11400	1030	1235	Open	CAT C	1130	4415000
4278	10646	8243	1187	1439	Covered	CAT A	1090	7128000
8066	11149	12936	1751	2098	No Parking	CAT B	720	5762000
7693	9130	14684	1746	2064	Open	CAT B	1050	6047000
5236	10853	13054	1615	1931	Covered	CAT B	1160	5913000
6027	6707	10176	1469	1756	Open	CAT B	770	6636000
9648	14789	12812	1644	1950	Covered	CAT A	790	7887000

Fig. 1. Sample real estate dataset screen shot

B. Identify numeric and categorical predictor variables

Sample dataset columns are classified into numeric and categorical predictor variables based on their properties. Out of 7 columns, 5 columns are identified as numeric variables (Distance-to-BT, Distance-to-Mall, Distance-to-Hospital, Carpet-Area and Built-up-Area), 1 is identified as categorical (Parking-Status) and house-price is identified as target variable.

C. Identify and handle the missing values

Missing value indicates that no data value is stored for the variable in the current observation. Missing data will have significant effect on the conclusions drawn from the data. Hence handling missing data is important for accurate predictions. In this dataset missing numeric variables are replaced with mean value and categorical variables are replaced with mode value.

D. Identify training data and test data for regression Modeling

Sample dataset is divided into training and test datasets. 70% of the sample dataset is considered as training dataset and 30% as test dataset. The training dataset is used to build the three regression models and the test dataset is used to evaluate the performance of the model.

E. Construct the regression model M1

Regression model M1 is built using the numeric and categorical predictor variables and house price as target variable. The results of the regression analysis summary is shown in Figure 2.

```
Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                3389911.00  574492.82   5.901 5.97e-09 ***
Dist_to_BT                      10.67      42.76   0.250 0.802962
Dist_to_Mall                    19.90      32.98   0.603 0.546491
Dist_to_Hospital               113.14      47.77   2.369 0.018159 *
Carpet_Area                  -2171.76    5462.14  -0.398 0.691060
Builtup_Area                  2454.20    4558.84   0.538 0.590538
Parking_statusNo Parking    -841197.09  220319.11  -3.818 0.000148 ***
Parking_statusNot Provided  -497942.08  199155.62  -2.500 0.012668 *
Parking_statusOpen          -397319.59  181527.52  -2.189 0.028988 *
---
```

Fig. 2. Regression Analysis Summary of Model M1 from R

From Figure 2 results it is observed all the variables are part of this model including the categorical variables where each category is represented as a separate variable. The last column in the table indicates the level of significance or importance of these variables in the model. In this model, carpet-area and built-up-area of the house are not shown as important. But these variables are having high correlation with the house price. Hence to address this issue, second model is built using PCA.

The performance of the model is evaluated using the test dataset. The correlation between the estimated and the observed values of the house prices is calculated. The correlation value explains the level of accuracy of the model. The square of the correlation is the R-square value or the predictive power of the model. The accuracy of this model is 0.01226.

The next model is constructed to handle multicollinearity using PCA. PCA modifies a set of numeric variables into uncorrelated principal components. A new model M2 is constructed with the orthogonal principal components derived from the original variables.

F. Construct the regression model M2

Five orthogonal principal components Dim1 to Dim6 are derived from the original numeric variables. These orthogonal components are shown in the PCA graph presented in Figure 3

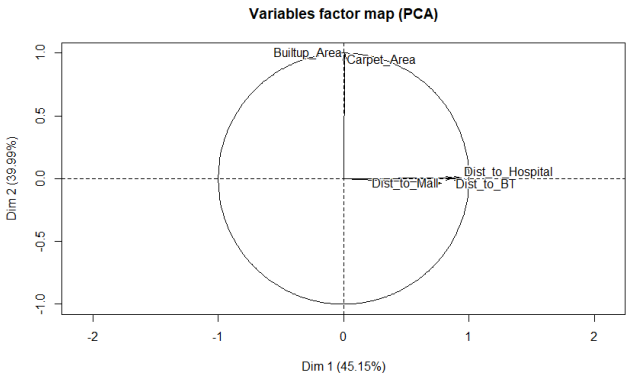

Fig. 3. PCA Graph from R

From Figure 3 it is observed that Distance-to-BT, Distance-to-Mall, and Distance-to-Hospital have formed principal component variable Dim1 which explains 45.15% information in data. Where in Carpet-Area and Builtup-Area explains the remaining 39.99% of information through Dim2 principal component variable. The eigenvalue is used in the principal component analysis to calculate the % variance as shown Figure 4. From Figure 4 it is observed that Comp1 and Comp2 explains 45.15% and 39.99 % of variance which is high among all.

```
       eigenvalue percentage of variance cumulative percentage of variance
comp 1 2.257514941             45.15029882                          45.15030
comp 2 1.999723923             39.99447846                          85.14478
comp 3 0.567327242             11.34654485                          96.49132
comp 4 0.174319980              3.48639960                          99.97772
comp 5 0.001113914              0.02227828                         100.00000
```

Fig. 4. Eigen value and % of variance explained by 5 components, sample screenshot from R

Regression model M2 is built using these principal components and categorical variable Parking-Status. The results of the regression analysis summary is shown in Figure 5.

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                6325900     147691  42.832  < 2e-16 ***
Parking_statusNo Parking    -835510     219374  -3.809 0.000154 ***
Parking_statusNot Provided  -501925     198856  -2.524 0.011849 *
Parking_statusOpen          -397342     180930  -2.196 0.028453 *
Dim.1                        224088      44531   5.032 6.36e-07 ***
Dim.2                        135554      48077   2.820 0.004963 **
```

Fig. 5. Regression Analysis Summary of Model M2 from R

From Figure 5 it is observed there are no insignificant numeric predictor variables in the model. The performance of the model is evaluated using the test dataset. The accuracy of this model is 0.01401. There is a slight improvement in the accuracy from model M1.

A correlation matrix constructed with the predictor variables and the principal components as shown in Figure 6. From the correlation matrix it is observed that 88% of the Distance-to-Hospital is loaded on Dim1 and 100% of both Carpet-Area and Built-up-Area is loaded on Dim2.

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
Dist_to_Hospital	88	0	2	10	0
Dist_to_BT	77	0	17	6	0
Dist_to_Mall	61	0	38	1	0
Carpet_Area	0	100	0	0	0
Builtup_Area	0	100	0	0	0

Fig. 6. Correlation Matrix of predictor variables and the principal components, sample screenshot from R

The model accuracy levels are satisfactory but the complexity involved in the operationalization of PCA model is high. Hence we can build a model by selecting the most likely predictor variables under the principal components Dim1 and Dim2 from the correlation matrix shown in Figure 6.

G. Construct the regression model M3

From Figure 6 it is observed that the dominant variables in Dim1 and Dim2 are Distance-to-Hospital and Carpet-Area. Hence, regression model M3 is built with these variables. The R-square value for this model is 0.01006. This value is acceptable since the operationalization of this model will be easy.

H. Compare the three models based on the accuracy levels and select the model with high accuracy.

From the above three models accuracy results it is observed that there is a marginal improvement in the accuracy from the first model to the second model. The model is simple and gives acceptable results in comparison to other models M1 and M2. More over this model is based on the PCA results obtained in model M2. Thus can be used to construct predictive models to estimate the real estate property price.

V. EXPERIMENTAL STUDY, RESULTS AND DISCUSSION

To address the expectations, needs and requirements of buyer and seller of real estate, an experimental study is carried out on Washington City's real estate data, with 159 rows and 330 columns. The dataset is available in .CSV format. The sample snapshot of the data is presented in Figure 7.

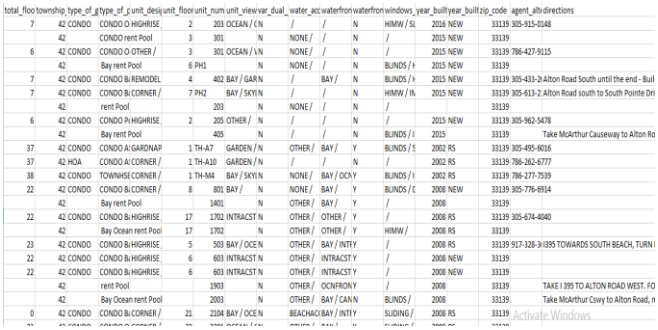

Fig. 7. Washington City's real estate sample dataset

The system configuration used for the experimental study is described below

Intel® Pentium® iV
RAM: 3.99GB, 64 bit operating system
Software tools: R Studio

A. Data Preparation

As the first step load the Washington City's real estate dataset into R to build the model. Missing values under each attribute list are identified and replaced with appropriate values. Numeric and categorical predictor variables are classified in the dataset. 101 numeric predictor variables are identified out of 330, remaining variables are considered as categorical. The target variable or response variable is identified as house price. In the next step, the dataset is divided into training set and test set. The training set is used to build the model and the test dataset is used to evaluate the performance of the model. These datasets are created randomly by making 70% of data as the training dataset and the remaining 30% as the test dataset.

B. M1 Model Building

A regression model is constructed using 101 numeric variables as predictor variables from the test dataset and house-price as the target variable. The summary of regression model obtained using R is presented in Figure 8. From Figure 8, it is observed that some of the predictor variables used in the analysis are tagged with *, **, *** based on their correlation with the target variable.

The performance of the model is evaluated using the test dataset. The correlation between the estimated and the observed values of the house prices is calculated. The correlation value explains the level of accuracy of the model. The square of the correlation is the R-square value or the predictive power of the model. The accuracy of this model is 0.3907. This explains 39.07% of the variation in the house price is due to these predictor variables. It is observed that in this model most significant predictor variables are tagged as unimportant. In fact, these variables have high correlation with the target variable. This problem is addressed in the next model using PCA.

```
Call:
lm(formula = list_price ~ ., data = PCA_data)

Residuals:
    Min      1Q  Median      3Q     Max
-711256 -189836       0  129616 1095417

Coefficients: (33 not defined because of singularities)
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -7.103e+10  5.538e+10  -1.282  0.20687
number_of_beds                  -3.832e+05  2.366e+05  -1.620  0.11300
number_of_fbaths                -5.121e+05  3.121e+05  -1.641  0.10848
number_of_hbaths                -4.199e+05  2.075e+05  -2.024  0.04954 *
number_of_ceiling_fans           1.510e+05  2.803e+05   0.539  0.59302
number_of_garage_spaces          1.920e+05  1.871e+05   1.026  0.31089
number_of_interior_levels       -1.145e+06  1.026e+06  -1.116  0.27108
number_of_times_leased_per_year -3.438e+03  3.298e+03  -1.042  0.30331
i_number                         1.945e+04  1.088e+04   1.787  0.08135 .
area                            -5.948e+04  4.119e+04  -1.444  0.15633
dade_assessed_dollars_soh_value  2.971e-01  1.924e-01   1.545  0.13015
maintenance_charge_per_month     7.424e+01  2.039e+02   0.364  0.71769
minimum_number_of_days_for_lease -1.946e+03  1.463e+03  -1.330  0.19095
rec_lease_per_month                     NA         NA      NA       NA
section                         -5.217e+04  9.438e+04  -0.553  0.58340
sqft_liv_area                    8.086e+03  7.060e+03   1.145  0.25874
subdivision_number               1.227e+03  1.496e+03   0.820  0.41685
tax_amount                       2.319e+01  1.222e+01   1.897  0.06487 .
total_number_of_units_in_buildin 1.129e+03  1.866e+03   0.605  0.54842
total_number_of_units_in_complex 8.241e+02  2.070e+03   0.398  0.69267
total_floors_in_building        -2.801e+04  1.029e+04  -2.721  0.00952 **
township_range                   1.321e+04  1.631e+04   0.810  0.42284
unit_floor_location             -2.033e+04  1.754e+04  -1.159  0.25316
price_per_sqft                   1.290e+03  5.557e+02   2.322  0.02529 *
mls_latitude                     8.720e+08  4.741e+08   1.839  0.07312 .
mls_longitude                   -6.058e+08  5.482e+08  -1.105  0.27561
number_of_stories               -3.731e+01  1.253e+04  -0.003  0.99764
parcel_number                    6.005e+02  2.173e+02   2.763  0.00853 **
parcel_number.1                         NA         NA      NA       NA
furn_annual_rent                 6.155e+01  5.599e+01   1.099  0.27801

flp_PAGE1                       -1.310e+02  1.702e+02  -0.770  0.44585
flp_QCD2                        -5.179e+04  9.393e+04  -0.551  0.58442
flp_SALES_YR2                           NA         NA      NA       NA
flp_SALES_MM2                           NA         NA      NA       NA
flp_BOOK2                               NA         NA      NA       NA
flp_PAGE2                               NA         NA      NA       NA
flp_MAIL_ZIP                     6.718e+00  5.888e+00   1.141  0.26047
flp_MKAR                         8.484e+04  6.825e+04   1.243  0.22094
flp_NBRHDCD                     -4.863e+03  2.911e+03  -1.671  0.10241
flp_TAXAUTHCD                           NA         NA      NA       NA
flp_SEC                         -7.598e+04  1.666e+05  -0.456  0.65066
flp_PH_ZIP                              NA         NA      NA       NA
flp_ASS_DIF_TRNS                -3.271e+05  6.450e+05  -0.507  0.61474
flp_ASS_DIF_SPLT                        NA         NA      NA       NA
flp_ASS_DIF_VL                          NA         NA      NA       NA
flp_ASS_DIF_CNTY                        NA         NA      NA       NA
flp_ASS_DIF_YR                          NA         NA      NA       NA
flp_EXMPT01                             NA         NA      NA       NA
flp_EXMPT02                             NA         NA      NA       NA
flp_SEQNO                       -4.508e+01  1.108e+02  -0.407  0.68623
flp_FCF1                                NA         NA      NA       NA
flp_FCF2                        -1.954e+01  3.594e+02  -0.054  0.95690
flp_OWNER_LATITUDE               8.630e+02  1.714e+04   0.050  0.96010
flp_OWNER_LONGITUDE             -5.819e+03  1.618e+04  -0.360  0.72093
flp_STREET_LEVEL                 1.975e+05  1.165e+05   1.695  0.09760 .
flp_GIS_AREA                            NA         NA      NA       NA
flp_LATITUDE                            NA         NA      NA       NA
flp_LONGITUDE                           NA         NA      NA       NA
latitude                                NA         NA      NA       NA
longitude                               NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 486100 on 41 degrees of freedom
Multiple R-squared:  0.9911,    Adjusted R-squared:  0.9765
F-statistic: 67.53 on 68 and 41 DF,  p-value: < 2.2e-16
```

Fig. 8. Snapshot of regression analysis summary obtained from R for Model M1

C. *M2 Model Building*

The orthogonal principal components Dim1 to Dim6 are extracted from the 101 numeric predictor variables using PCA. These predictor variables do not have any multicollinearity between them as they are perfectly uncorrelated. PCA graph is shown in Figure 9. From the graph it is observed that principal component variable Dim1 explains 15.23% of data and Dim2 explains 13.33% of data. Using these principal components regression model is developed.

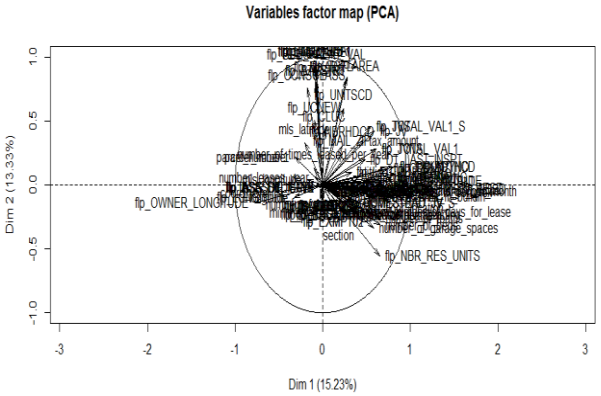

Fig. 9. PCA graph, screenshot from R

Eigenvalues and Eigenvectors are obtained during the PCA. A sample of 36 components comp1 to comp36 is shown in Figure 4. Eigenvalues are used in principal component analysis to explain the percentage of variance in the house-price due to each component. Eigenvectors represents the vector locations of these components. A correlation matrix is constructed using the numeric predictor variables and the principal components as shown in Figure 10. From the correlation matrix, we can address the issues of multicollinearity in the numeric predictor variables.

The performance of this model is evaluated using the test dataset. The accuracy of this model is 0.3467. This explains approximately 35% of the variation in the house price is due to these principal components which is lesser compared to model M1. But the complexity involved in the operationalization of PCA model is high. Hence we can build a model by selecting the most likely predictor variables under the principal components Dim1 to Dim3 from the correlation matrix shown in Figure 11.

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	1.538011e+01	1.522783e+01	15.22783
comp 2	1.346435e+01	1.333104e+01	28.55887
comp 3	1.019045e+01	1.008955e+01	38.64842
comp 4	7.576380e+00	7.501366e+00	46.14978
comp 5	5.780303e+00	5.723073e+00	51.87286
comp 6	5.173841e+00	5.122615e+00	56.99547
comp 7	4.758243e+00	4.711131e+00	61.70660
comp 8	4.117738e+00	4.076968e+00	65.78357
comp 9	3.761004e+00	3.723767e+00	69.50734
comp 10	3.077823e+00	3.047349e+00	72.55469
comp 11	2.766833e+00	2.739438e+00	75.29412
comp 12	1.640761e+00	1.624516e+00	76.91864
comp 13	1.528869e+00	1.513732e+00	78.43237
comp 14	1.516549e+00	1.501534e+00	79.93391
comp 15	1.400798e+00	1.386929e+00	81.32083
comp 16	1.350491e+00	1.337120e+00	82.65795
comp 17	1.251018e+00	1.238631e+00	83.89659
comp 18	1.161504e+00	1.150004e+00	85.04659
comp 19	1.078459e+00	1.067781e+00	86.11437
comp 20	1.037453e+00	1.027181e+00	87.14155
comp 21	9.959640e-01	9.861030e-01	88.12765
comp 22	9.390201e-01	9.297229e-01	89.05738
comp 23	8.603545e-01	8.518362e-01	89.90921
comp 24	8.170181e-01	8.089288e-01	90.71814
comp 25	7.800026e-01	7.722798e-01	91.49042
comp 26	7.284232e-01	7.212111e-01	92.21163
comp 27	6.766534e-01	6.699539e-01	92.88159
comp 28	6.294176e-01	6.231857e-01	93.50477
comp 29	5.912740e-01	5.854198e-01	94.09019
comp 30	5.181384e-01	5.130083e-01	94.60320
comp 31	5.016121e-01	4.966456e-01	95.09985
comp 32	4.715354e-01	4.668667e-01	95.56671
comp 33	4.101034e-01	4.060430e-01	95.97276
comp 34	3.738423e-01	3.701409e-01	96.34290
comp 35	3.643113e-01	3.607043e-01	96.70360
comp 36	3.530485e-01	3.495530e-01	97.05315

Fig. 10. Eigen value and % of variance explained by 36 components, sample screenshot from R

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
flp_SEQNO	68	1	27	0	0
flp_COUNTYNO	67	2	28	0	0
flp_TAXAUTHCD	67	2	28	0	0
flp_PH_ZIP	67	2	28	0	0
flp_YR_BUILT	63	0	27	0	0
flp_ACT_YR_BLT	63	0	27	0	0
flp_IMPQUAL	61	0	29	0	0
flp_NOBULDNG	61	0	29	0	0
flp_OWNER_LONGITUDE	60	2	27	0	0
flp_GRPNO	55	2	17	0	0
sqft_liv_area	45	5	32	0	0
number_of_fbaths	43	7	19	0	0
number_of_beds	42	10	18	0	1
flp_NBR_RES_UNITS	42	31	20	0	0
flp_JVNS	38	8	35	12	0
flp_TOTAL_VAL1	37	8	35	13	0
flp_JV	35	18	35	5	0
flp_FCF2	35	1	7	10	1
number_of_garage_spaces	34	11	7	1	0
flp_OWNER_LATITUDE	32	0	24	0	0
flp_AV_NON_HMSTD_RESD	31	0	26	28	0
flp_JVS	30	21	31	11	0
flp_TOTAL_VAL1_S	30	21	31	11	0
flp_JV_NON_HMSTD_RESD	30	0	25	26	1
maintenance_charge_per_month	29	0	35	3	2
coord_level	29	0	5	1	0
price_per_sqft	25	0	35	12	3
flp_DT_LAST_INSPT	23	4	20	5	0
tax_amount	21	12	34	0	0
unit_floor_location	18	3	12	14	1
township_range	17	0	6	0	0
i_number	16	5	0	2	0
area	16	0	0	0	0
total_floors_in_building	15	5	9	26	5
price_vs_market	15	0	9	27	9

Fig. 11. Correlation Matrix of predictor variables and the principal components, sample screenshot from R

A. *M3 Model Building*

The third model M3 is built using the predictor variables selected from the correlation matrix of Figure 11. These variable section is based on their significance w.r.t principal components Dim1, Dim2 and Dim3 of M2 model correlation matrix. Hence 19 significant predictor variable from 3 principal components are selected. Regression model is constructed by using these predictor variables. The summary of the regression model obtained using R is shown in Figure 12. The accuracy of this model is 0.5825. This 58% house price variation is due to the selected significant predictor variables.

```
Call:
lm(formula = list_price ~ ., data = cdata[train, c(Target, new_predictors)])

Residuals:
     Min       1Q   Median       3Q      Max
-8610203  -750236   215736   877499 11326215

Coefficients: (8 not defined because of singularities)
                     Estimate  Std. Error t value Pr(>|t|)
(Intercept)        1111600.000  981452.912   1.133 0.260143
flp_LAND_VAL1          -14.860       9.312  -1.596 0.113749
flp_NOUNITS                 NA          NA      NA       NA
flp_BLDG_SQFT               NA          NA      NA       NA
flp_SPEC_FEAT_VAL       440.788    1173.239   0.376 0.707951
acf3                     11.456     746.520   0.015 0.987788
flp_FCF1                    NA          NA      NA       NA
flp_GIS_AREA                NA          NA      NA       NA
flp_SEQNO               -20.316      57.990  -0.350 0.726839
flp_COUNTYNO        1152274.191 1080716.945   1.066 0.288947
flp_TAXAUTHCD               NA          NA      NA       NA
flp_PH_ZIP                  NA          NA      NA       NA
flp_YR_BUILT         563603.047 1113419.872   0.506 0.613859
flp_ACT_YR_BLT      -577748.184 1111438.654  -0.520 0.604361
flp_IMPQUAL                 NA          NA      NA       NA
flp_NOBULDNG                NA          NA      NA       NA
flp_OWNER_LONGITUDE  -16800.385   50239.373  -0.334 0.738787
flp_GRPNO            217785.729  447071.577   0.487 0.627248
acf2                   7100.478    1963.691   3.616 0.000475 ***
flp_TOTLAREA          -4926.669    2134.531  -2.308 0.023097 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2195000 on 98 degrees of freedom
Multiple R-squared:  0.5689,    Adjusted R-squared:  0.5205
F-statistic: 11.76 on 11 and 98 DF,  p-value: 0.0000000000001037
```

Fig. 12. Regression Analysis results of Model M3

From the above three models accuracy results it is observed that there is a marginal improvement in the accuracy from the first model to the third model. The second model explains only 35% of the price variation. This is due to the complexity involved in the usage of PCA. The model M3 is simple and gives more accurate results in comparison to other models M1 and M2. More over this model is based on the PCA results obtained in model M2. Thus PCA is a highly intuitive and powerful data science technique that can be used to construct predictive models to estimate the real estate property price.

VI. Conclusion

In this paper, three regression models are developed and validated using R to estimate the real estate housing price based on various predictor variables. The accuracy of the models M1, M2 and M3 are 0.3907, 0.3467 and 0.5825 respectively. From these accuracy values it is observed that the model M3 is giving better results in comparison to M1 and M2. Moreover M3 is a simple model constructed using the results obtained from the PCA technique employed in M2, which is identified as the powerful data science technique for predictor variable reduction in particular for real estate dataset attributes reduction analysis. This technique is also useful for constructing the models to estimate the real estate property price accurately. In this paper only numeric attributes are considered for PCA to create models for analyzing and predicting the real estate housing price. In future there is a scope to extend this data model to support categorical attributes as well. Further the model can be extendable to other real estate data sets in future.

References

[1] Adyan Nur Alfiyatin, Ruth Ema Febrita, Hilman Taufiq, Wayan Firdaus Mahmudy: "Modeling House Price Prediction using Regression Analysis and Particle Swarm Optimization - Case Study: Malang, East Java, Indonesia, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 8, No. 10, 2017, pp. 323-326.

[2] V.Sampathkumara, M.Helen Santhib, J.Vanjinathan: "Forecasting the land price using statistical and neural network software", 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015), Procedia Computer Science, Science Direct, 2015, pp.112 – 121.

[3] R Manjula, Shubham Jain, Sharad Srivastava and Pranav Rajiv Kher: "Real estate value prediction using multivariate regression models", 14th ICSET-2017, IOP Conf. Series: Materials Science and Engineering, 2017, pp. doi:10.1088/1757-899X/263/4/042098.

[4] Nguyen Nghiep and Cripps Al "Predicting Housing Value: A Comparison of Multiple Regression Analysis and Artificial Neural Networks" Journal of Real Estate Research: 2001

[5] Nihar Bhagat, Ankit Mohokar, Shreyash Mane: "House Price Forecasting using Data Mining", International Journal of Computer Applications (0975 – 8887) Volume 152 – No.2, October 2016, pp. 23-26.

[6] Sayali Sandbhor, N.B. Chaphalkar: "Determining attributes of Indian real property valuation using principal component analysis", Journal of Engineering Technology (ISSN: 0747-9964) Volume 6, Issue 2, July. 2017, PP. 483-495.

[7] http://ucanalytics.com/blogs/category/pricing-case-study-example