House Price Prediction

Aman Prakash
University at Buffalo (Graduate Student)
Buffalo, United States
amanpark@buffalo.edu

Avash Neupane
University at Buffalo (Graduate Student)
Buffalo, United States
avashneu@buffalo.edu

Bhavnish Attaluri
University at Buffalo (Graduate Student)
Buffalo,United States
battalur@buffalo.edu

Prashant Upadhyay
University at Buffalo (Graduate Student)
Buffalo,Unites States
pupadhya@buffalo.edu

Serath Chandra Nutakki
University at Buffalo (Graduate Student)
Buffalo, United States
serathch@buffalo.edu

***Abstract* — Nowadays the real estate market is using a lot of big data to standout against the price prediction and price fluctuation, and it is one of the primary fields that is using ideas of machine learning to predict the costs with high accuracy. The objective of the project is to predict the market value using features that describe the physical characteristics of a house. To get an exquisite perception of the commercialized market this predictive model system is highly conducive. The model system not only incorporates the evident factors like economy and land scarcity but also number of geographic variables that are invaluable. The project is considered as a future stride for people to make better decisions regarding house investment by analyzing and dissecting the patterns of the past data and the development in the anticipation of the market value on an unseen data. For the initial process we have checked our model on linear regression, and we got the R-squared as 0.86. Furthermore, by assimilating variable transformation we improved the accuracy to 0.88.**

Keywords—Feature, sales price, density plot

I. INTRODUCTION

Homeownership in the present real estate market has been tremendously difficult to maintain with soaring house prices becoming a hurdle. With these price hikes, it has become a guessing game for people as to how much money should they save to buy the house of their dreams. The motivation behind this project is to predict house prices based on the features of the house. The price predicted for a house with certain features will provide potential homeowners with an insight so that they plan their savings and finances to buy their dream house.

The property business has been one of the main examination regions zeroing in on present-day financial aspects, for its critical ramifications on important ventures and fields like development, speculation, and public government assistance. Purchasing and placing assets into any land undertaking will incorporate various trades between various ends. The best technique to foster a reasonable model to precisely expect the expense of land has been a badly designed point with exceptional potential for extra assessment. It is overall acknowledged by the insightful local area that precisely anticipating the extraordinary expense for explicit land is illogical since it fuses a ton of variables applying influence on the unavoidable expense.

The present examination on the property business, progressed investigation techniques, for example, AI and ML logics have been all things considered embraced in different viewpoints. Notwithstanding the way that they are utilized in surveying the expense and worth, they are applied to figure out potential expectations and challenges. The expansive gathering of AI and ML logic in the land business has normally changed this experience-driven industry with unimaginable trade opportunities to an insightful and data-driven to drive.

Dissimilar to regression logic, artificial neural network (ANN) and fuzzy logic (FL) philosophy have additionally been broadly acknowledged. In Din et al's. research on the ANN strategy for property evaluation, it performs well and produces satisfactory execution in certain regards. In any case, et likewise turns out that diverse information selections of factors would now and then create measurably various upsides of result, which demonstrates the shakiness and youthfulness of the ANN philosophy. As far as fuzzy logic, it is generally accepted to be an encouraging and conventional methodology for assessing properties. Liu et al. have developed a model generated by using fuzzy neural-net to make predictions which sums the decadent hypothesis and an incredible information base with pertinent qualities influencing the cost of properties dependent on as of late sold tasks. The result and assessment displayed that the fuzzy neural-net has a promising capacity for land value forecast given solid information with top caliber. A relative examination has additionally uncovered that various regression applications for property evaluation function admirably with given information.

Nonetheless, hypothetically, the models dependent on different regressions appear to provide more importance to measurable results. Notwithstanding the significant and expansive importance brought by this strategy, it has become important to analyze expectations on the valuation of the property. Subsequently, embracing different procedures to lead a thorough and feasible exploration to an alternate degree is crucial.

II. MOTIVATION

The vital problems faced by most of the people around the globe is the exorbitant price of housing. The real estate market contains many investors, companies and agents who strive to achieve competitive results over each other. Despite the economic hardship due to pandemic, the housing market is blooming in 2020 with the highest record of housing sales in the U.S. since 2005 where lot of factors like remote work, stable income, location and many more. With our motive to understand the contributing factors behind the pricing of the houses and understand what are the most prominent features that can help us predict the house prices. With the objective of making house search and sell/buy a hassle-free process and eliminate the need of "middle-men" that may incur more cost, we took up the research. Furthermore, to perceive and understand the influencing factors of the housing prices, we are conducting the research

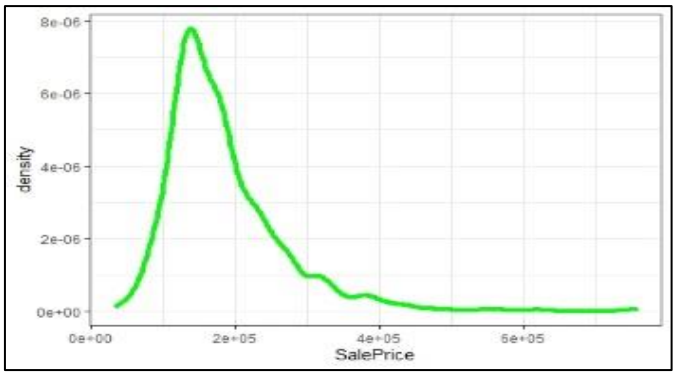
III. APPROACH

A. *Obtain Dataset*

The dataset for the project was obtained from an online data repository "Kaggle". This dataset is a modified version of 'Ames Housing Dataset' initially compiled by Dean De Cock[1] as an alternative to the popular 'Boston Housing dataset'. For the purpose the project, the dataset that consisted of 1460 observations and 81 variables was taken.

B. *Exploratory Data Analysis*

The variable 'SalePrice' is the target while the rest of the variables are the independent variables. Using the 1460 observations, it was observed that the SalePrice ranges from 34900 to 755000 with a mode of 140000. The median value is 163000 which is less than the mean of 180921 which suggests that the target variable is positively skewed. The density plot and Q-Q plot for the SalePrice shown below (Figure 1) represents the same. In order to transform the skewed target variable into normal distribution, a logarithmic transformation is taken. The density plot and Q-Q plot after taking the log transform is shown below (Figure 2).



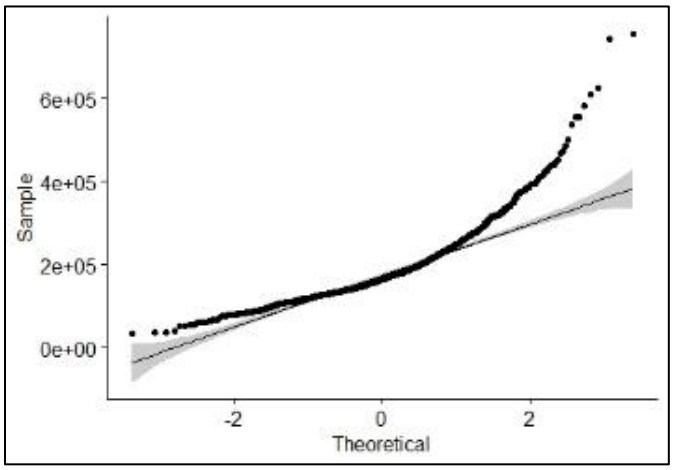
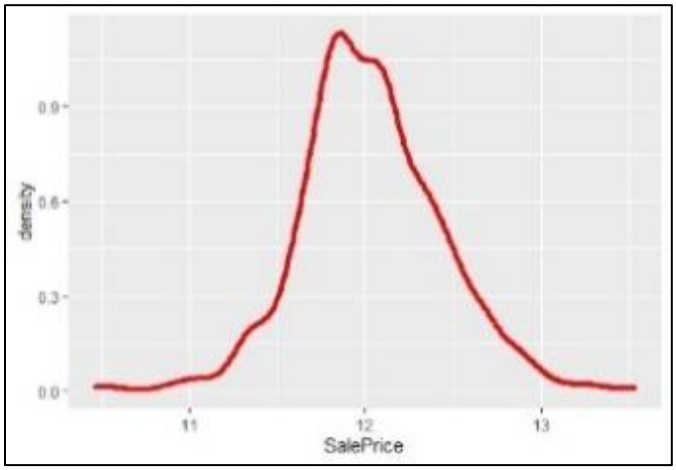


Figure 1: Density Plot and Q-Q Plot Before Transformation



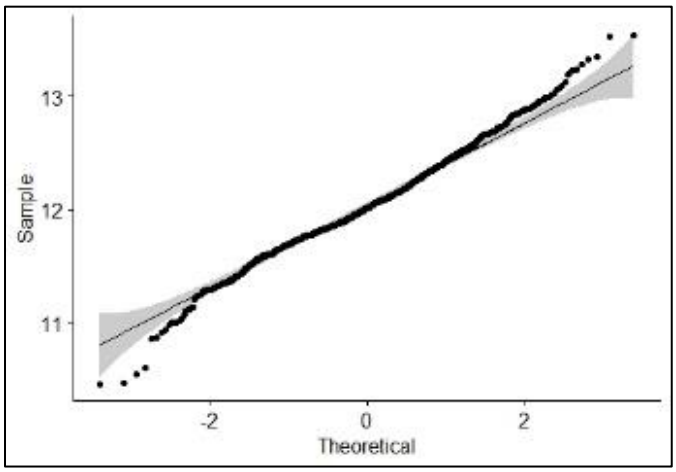


Figure 2: Density Plot and Q-Q Plot After Transformation

C. *Taking care of missing values*

The number of missing values for each of the variables were checked. It was observed that for some of the columns, there were more than 1000 missing values which accounted for more than 71% of the data. Thus, these variables were dropped off.

From the dataset with rest of the variables, the variables with more than 80 missing observations were also dropped. For the rest of the variables the observations with missing values were dropped in such a way that minimal data loss would be encountered. After taking care of the missing values, a total of 1412 observations and 69 features remained.

D. Encoding categorical variables

In the dataset, there were a wide range of categorical variables such as color, number of rooms, type of recreational facilities or the type of furniture. For regression analysis with the type of data, there is a need to create dummy variables to understand what factor of the feature has an impact on the prediction. Thus, in order to handle the categorical variables, these were encoded using one-hot encoding techniques. After performing one-hot encoding, for each of the factors of the categorical variables, new binary variables were created. With this type of encoding, it makes it easier to study to compare the features and for easier feature analysis.

E. Transformation of Independent Variables

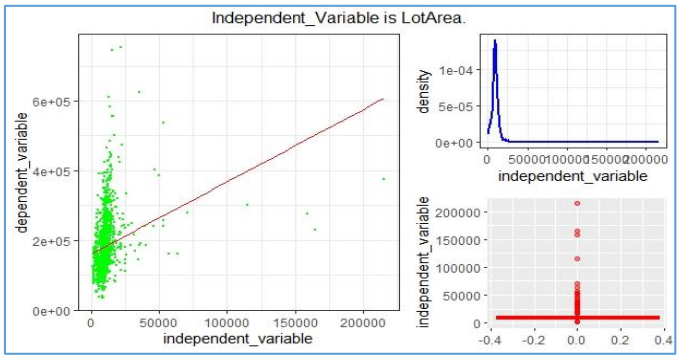


Figure 3: Before Transformation - Scatter Plot (Dependent vs Independent Variable), Density Plots, Box Plots (Here, Independent variable is LotArea)

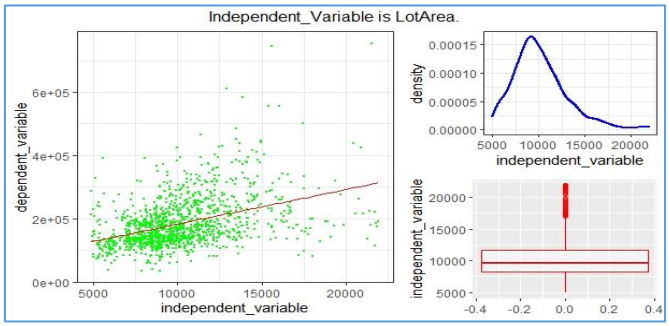


Figure 4: After Transformation - Scatter Plot (Dependent vs Independent Variable), Density Plots, Box Plots (Here, Independent variable is LotArea)

From the scatter plot, the linearity in relationship between each of the continuous random variable with 'SalePrice' was observed. The density plots for the continuous variables were plotted in order to understand the distribution of the data for each of the continuous variables. Furthermore, in order to observe the quartile values and check for outliers, a boxplot graph was also plotted.

For instance, consider the example graph shown above. Before transformation (Figure 3), we can observe a distorted linear relationship between the independent variable with the SalePrice from the scatter plot. From the density plot, we can observe skewness in data for the independent variable. As the range of the data distribution is observed to be quite large, it is prudent to transform these variables so as to encapsulate the overall range of data as far as possible. Thus, after checking these graphs for each independent variable, it was decided to take log transform of the independent variables as well.

After log transformation of variables, the graphs were plotted again to study the changes occurred (Figure 4). From the scatter plot, it was observed that a better linearity in relationship was achieved between the independent variable and the target variable. Additionally, the skewness in the data distribution was also taken care of and a more bell-shaped graph was obtained. It also helped to better grasp the outliers present in the data.

This step was not employed for development of model 1 (mentioned later).

F. Outliers treatment

Another benefit of taking the log transform of the independent variable was in de-emphasis of the outliers in the data. However, the outliers obtained post-transformation treated with minimum data loss.

G. Data Partitioning

The pre-processed data was partitioned into training set and test set in a percentage split of 80/20.

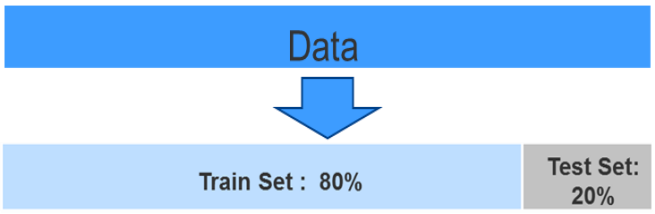


Figure 5: Data Partitioning

H. Feature Selection

As it is not feasible to use all the features in the dataset to build a predictive model, it is important to be selective of the features to be employed in the ML model. Feature selection allows us to considerably decrease the complexity of the model while making the model interpretable. A stepwise forward regression feature selection method using p-values was employed to obtain the most relevant features for the regression model.

A stepwise regression model builds regression model from a set of candidate predictor variables by entering predictors based on p values, in a stepwise manner until there is no variable left to enter any more.

While selecting the features, 20 features that were the most suitable among all other features were selected.

1. OverallQual - Rates the overall material and finish of the house
2. GrLivArea - Above grade (ground) living area square feet
3. BsmtFinSF1 - Type 1 finished square feet
4. GarageCars - Size of garage in car capacity
5. MSSubClass - Identifies the type of dwelling involved in the sale
6. YearBuilt - Original construction date
7. OverallCond - Rates the overall condition of the house
8. BedroomAbvGr - Bedrooms above grade (does NOT include basement bedrooms)
9. LotArea - Lot size in square feet
10. MasVnrArea - Masonry veneer area in square feet
11. BsmtFullBath - Basement full bathrooms
12. TotRmsAbvGrd - Total rooms above grade (does not include bathrooms)
13. ScreenPorch - Screen porch area in square feet
14. WoodDeckSF - Wood deck area in square feet
15. YearRemodAdd - Remodel date (same as construction date if no remodeling or additions)
16. TotalBsmtSF - Total square feet of basement area
17. KitchenAbvGr - Kitchens above grade
18. Fireplaces - Number of fireplaces
19. PoolArea - Pool area in square feet
20. FullBath - Full bathrooms above grade

I. Model Building

Model 1: Feature Selection + Regression Model

By employing the features selected, a multiple linear regression model was built that resulted in an adjusted R-Squared values of 0.8589.

```
Call:
lm(formula = SalePrice ~ ., data = train_transormation)

Residuals:
     Min       1Q   Median       3Q      Max
-1.93564 -0.06854  0.00151  0.07689  0.54228

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.634e+00  5.382e-01   4.895 1.10e-06 ***
OverallQual   8.502e-02  5.076e-03  16.752  < 2e-16 ***
GrLivArea     1.941e-04  1.804e-05  10.758  < 2e-16 ***
BsmtFinSF1    2.638e-05  1.310e-05   2.014 0.044176 *
GarageCars    7.316e-02  7.348e-03   9.956  < 2e-16 ***
MSSubClass   -6.886e-04  1.115e-04  -6.179 8.48e-10 ***
YearBuilt     2.933e-03  2.337e-04  12.546  < 2e-16 ***
OverallCond   4.809e-02  4.361e-03  11.028  < 2e-16 ***
BedroomAbvGr -1.751e-03  7.260e-03  -0.241 0.809427
LotArea       1.909e-06  4.276e-07   4.464 8.69e-06 ***
MasVnrArea    4.111e-06  2.535e-05   0.162 0.871180
BsmtFullBath  5.996e-02  1.022e-02   5.865 5.61e-09 ***
TotRmsAbvGrd  1.476e-02  5.321e-03   2.774 0.005605 **
ScreenPorch   3.427e-04  7.198e-05   4.761 2.12e-06 ***
WoodDeckSF    1.109e-04  3.377e-05   3.285 0.001045 **
YearRemodAdd  1.101e-03  2.853e-04   3.859 0.000119 ***
TotalBsmtSF   5.072e-05  1.535e-05   3.304 0.000978 ***
KitchenAbvGr -3.732e-02  2.417e-02  -1.544 0.122891
Fireplaces    4.898e-02  7.443e-03   6.580 6.63e-11 ***
PoolArea     -3.606e-04  9.985e-05  -3.611 0.000315 ***
FullBath      2.884e-02  1.128e-02   2.558 0.010640 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1479 on 1391 degrees of freedom
Multiple R-squared:  0.8609,    Adjusted R-squared:  0.8589
F-statistic: 430.6 on 20 and 1391 DF,  p-value: < 2.2e-16
```

Figure 6: Statistics for Model 1

Equation of the model:

log(SalesPrice) = 2.634 + (0.08502)* ***OverallQual*** + (0.001941)****GrLivArea*** + (0.00002638)* ***BsmtFinSF1*** + (0.07316)****GarageCars*** – (0.006886)****MSSubClass*** + (0.002933)****YearBuilt*** + (0.04809)****OverallCond*** - (0.001751)****BedroomAbvGr*** + (0.000001909)****LotArea*** + (0.000004111)****MasVnrArea*** + (0.05996)****BsmtFullBath*** + (0.01476)****TotRmsAbvGrd*** + (0.0003427)****ScreenPorch*** + (0.001109)****WoodDeckSF*** + (0.001101)****YearRemodAdd*** + (0.00005072)****TotalBsmtSF*** - (0.003732)****KitchenAbvGr*** + (0.04898)****Fireplaces*** - (0.0003606)****PoolArea*** + (0.02884)****FullBath***

Model 2: Variable Transformation + Feature Selection + Regression Model

As mentioned in III (E), for Model 2, the log transform of the variables was taken before fitting it in our model. With variable transformation, it was observed that a better linearity in relationship was achieved between the independent variable and the target variable. Additionally, the skewness in the data distribution was also taken care of and a more bell-shaped graph was obtained.

The transformed and selective features were used to build Model 2. The multiple linear regression model that was fitted for Model 2 resulted in the statistics mentioned below in *Figure 7*. The R-Squared value obtained was 0.8856 and Adjusted R-Squared was 0.8832.

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.85940 -0.06756  0.00120  0.07237  0.52474

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -5.738e+01  4.632e+00 -12.389  < 2e-16 ***
Id            5.918e-06  1.022e-05   0.579 0.562607
OverallQual   3.806e-01  3.086e-02  12.331  < 2e-16 ***
GrLivArea     4.378e-01  3.149e-02  13.905  < 2e-16 ***
BsmtFinSF1    1.055e-02  1.637e-03   6.448 1.78e-10 ***
GarageCars    8.877e-02  1.599e-02   5.552 3.65e-08 ***
MSSubClass    7.125e-03  8.914e-03   0.799 0.424326
YearBuilt     5.166e+00  4.844e-01  10.665  < 2e-16 ***
OverallCond   2.417e-01  2.489e-02   9.713  < 2e-16 ***
BedroomAbvGr -1.190e-01  1.979e-02  -6.013 2.58e-09 ***
LotArea       6.569e-02  1.784e-02   3.683 0.000244 ***
MasVnrArea    1.814e-03  1.943e-03   0.933 0.350838
BsmtFullBath  1.513e-01  7.489e-02   2.020 0.043692 *
TotRmsAbvGrd  1.435e-01  3.831e-02   3.746 0.000190 ***
ScreenPorch   9.350e-03  3.056e-03   3.060 0.002276 **
WoodDeckSF    6.679e-03  1.822e-03   3.665 0.000260 ***
YearRemodAdd  3.146e+00  5.965e-01   5.273 1.65e-07 ***
TotalBsmtSF   1.706e-01  1.808e-02   9.435  < 2e-16 ***
KitchenAbvGr -2.609e-01  3.830e-02  -6.812 1.69e-11 ***
Fireplaces    7.005e-02  2.428e-02   2.885 0.004002 **
PoolArea      1.235e-02  1.077e-02   1.147 0.251595
FullBath     -1.385e-03  1.935e-02  -0.072 0.942939
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1348 on 971 degrees of freedom
Multiple R-squared:  0.8856,    Adjusted R-squared:  0.8832
F-statistic:   358 on 21 and 971 DF,  p-value: < 2.2e-16
```

Figure 7: Statistics for Model 2

Model 3: Variable Transformation + Feature Selection + Clustering + Regression Model

Different combinations of features are used to classify the data for model 3 and the set of features that has produced the best set of classes has been chosen. Due to the large number of features, tuning the models has to be done on a preselected number of features. The selection of features is based on the initial model and these selected features are used for the later models. The below *Figure 8* shows the output of the classifier model. The number of clusters are chosen based on the subsequent output of the regression models, with fewer cluster there is no significant divide in the data to be able to separate the class specific features. Cluster analysis, often known as clustering, is the approach of arranging a set of items so that objects in the same group/cluster are more comparable to those in other groups.

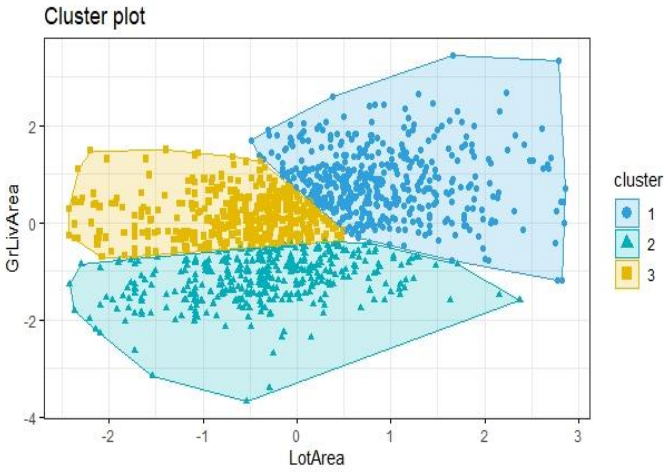


Figure 8: Cluster Plot (GrLivArea, LotArea)

With more than three classes the model 3 has underperformed that the previous models, with clusters more than 3, there are overlapping classes. This made it harder to identify the correct regression model for the datapoints in the overlapping regions. While for basic regression analysis it is common to see some data loss due to trimming of outliers with the cluster model the data can be divided into class of its own. That is higher range of data gets clustered into similar class and lower ranges of data gets clustered into similar class. This enables us to treat the data in the class as independent data. This reduces the need for trimming the outliers and maintain minimal data loss. The outliers previously – in complete data – are now in a class with data of similar range.

The below correlation plots *(Figure 9 and Figure 10)* show that indeed the features are relevant to price are varied across the classes.

Figure 9 shows the correlation plot for cluster 1, from the plot we see that overall quality of the house is highly correlated with the sale price. *Figure 10* shows the correlation plot for the cluster 2, the feature overall quality in the cluster 2 is poorly correlated with the sale price. The year in which the house is built is highly correlated with the sale price, but this feature has poor correlation with the price for the cluster 1.

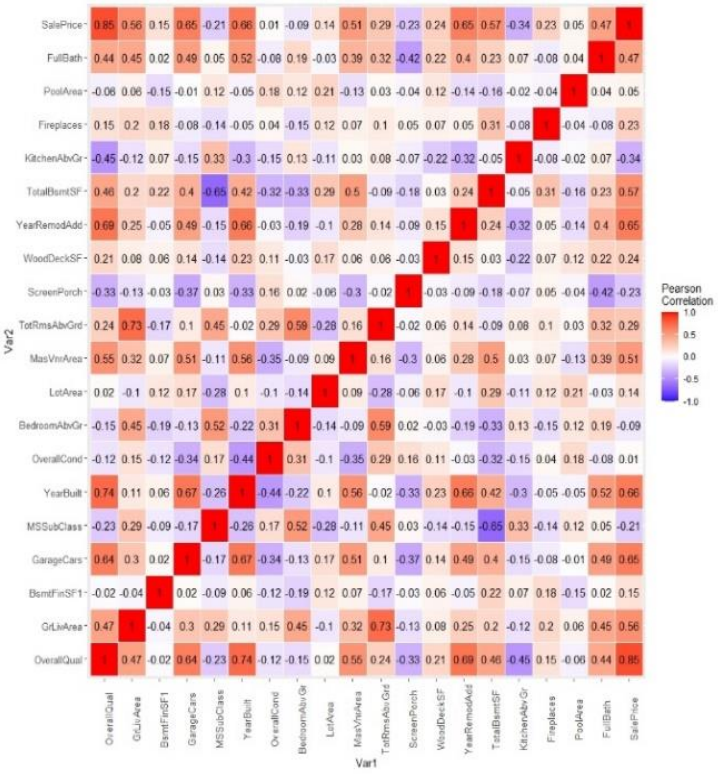


Figure 9: Correlation Plot for Cluster 1

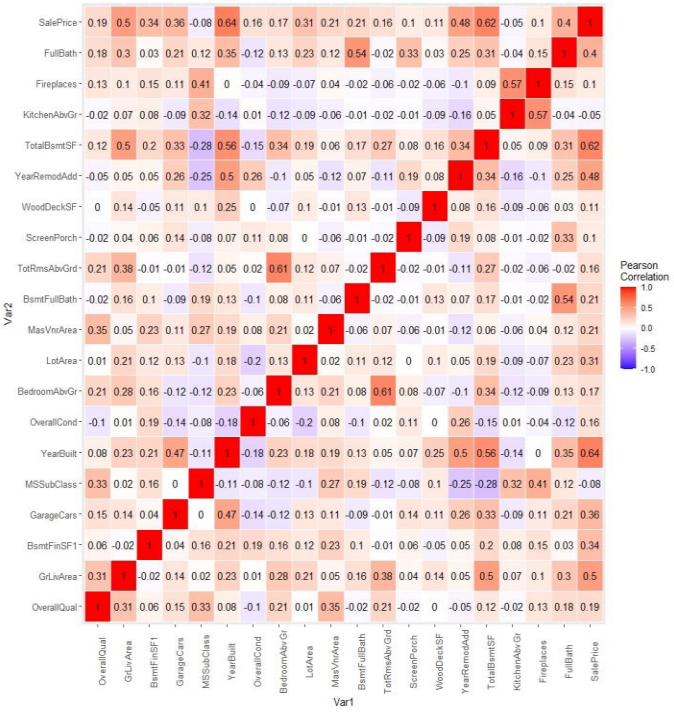


Figure 10: Correlation Plot for Cluster 2

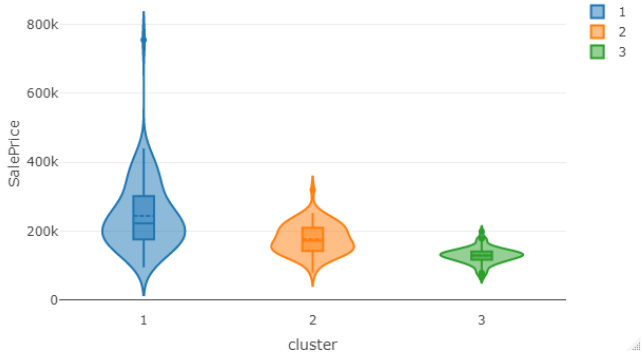


Figure 11: ViolinPlot for Actual SalePrice based on Cluster

As 'SalePrice' is the dependent that is to be predicted it cannot be used divide the data into clusters - as the variable is not readily available for the unseen data – the data has to be clustered using independent variables. So, it is important to identify how the dependent is being affected from dividing the data into clusters. The above figure 11 gives the distribution of the Actual Sale Price across the clusters, the figure clearly indicates there is a clear divide in the price range amongst the three classes. Cluster 1 has the highest range for price it also has the highest median followed by cluster 2 then cluster 3. The figure also shows an indication to the size of each cluster, with cluster 3 being the smallest and cluster 1 the biggest.

Summary of Statistics for Model 3

Cluster 1:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.80193 -0.07884  0.00731  0.08190  0.35672

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.059e+01  8.543e+00  -7.092 7.58e-12 ***
OverallQual   5.328e-01  6.101e-02   8.733  < 2e-16 ***
GrLivArea     5.342e-01  6.398e-02   8.349 1.72e-15 ***
BsmtFinSF1    9.728e-03  2.583e-03   3.767 0.000195 ***
GarageCars    1.221e-01  3.236e-02   3.773 0.000190 ***
MSSubClass    3.454e-02  2.114e-02   1.634 0.103228
YearBuilt     3.208e+00  8.655e-01   3.706 0.000245 ***
OverallCond   1.560e-01  4.905e-02   3.181 0.001603 **
BedroomAbvGr -1.039e-01  3.309e-02  -3.139 0.001843 **
LotArea       1.370e-01  4.057e-02   3.378 0.000815 ***
MasVnrArea    3.365e-04  3.015e-03   0.112 0.911183
BsmtFullBath  9.340e-02  1.223e-01   0.763 0.445700
TotRmsAbvGrd  1.010e-01  6.397e-02   1.579 0.115166
ScreenPorch   7.016e-03  4.628e-03   1.516 0.130464
WoodDeckSF    1.611e-03  2.960e-03   0.544 0.586624
YearRemodAdd  5.279e+00  1.240e+00   4.257 2.68e-05 ***
TotalBsmtSF   2.128e-01  3.615e-02   5.886 9.42e-09 ***
KitchenAbvGr -3.186e-01  6.903e-02  -4.616 5.54e-06 ***
Fireplaces    6.452e-02  3.073e-02   2.100 0.036469 *
PoolArea     -2.084e-03  1.602e-02  -0.130 0.896591
FullBath     -2.068e-02  3.297e-02  -0.627 0.530949
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.134 on 343 degrees of freedom
Multiple R-squared:  0.8579,    Adjusted R-squared:  0.8496
F-statistic: 103.6 on 20 and 343 DF,  p-value: < 2.2e-16
```

Cluster 2:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.71697 -0.06547 -0.00143  0.06450  0.42035

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -61.071238   7.250824  -8.423 1.40e-15 ***
OverallQual    0.339935   0.047957   7.088 9.25e-12 ***
GrLivArea      0.390429   0.063420   6.156 2.31e-09 ***
BsmtFinSF1     0.008843   0.002411   3.668 0.000288 ***
GarageCars     0.059284   0.026488   2.238 0.025923 *
MSSubClass     0.016024   0.014091   1.137 0.256328
YearBuilt      5.820169   0.730137   7.971 3.06e-14 ***
OverallCond    0.337978   0.036389   9.288  < 2e-16 ***
BedroomAbvGr  -0.163596   0.035294  -4.635 5.27e-06 ***
LotArea       -0.004192   0.040399  -0.104 0.917415
MasVnrArea     0.005803   0.003062   1.895 0.059043 .
BsmtFullBath   0.196224   0.093956   2.088 0.037575 *
TotRmsAbvGrd   0.252515   0.059365   4.254 2.79e-05 ***
ScreenPorch    0.014822   0.004633   3.199 0.001521 **
WoodDeckSF     0.010349   0.002795   3.703 0.000253 ***
YearRemodAdd   3.065250   0.907855   3.376 0.000828 ***
TotalBsmtSF    0.172296   0.026084   6.605 1.73e-10 ***
KitchenAbvGr  -0.238539   0.049544  -4.815 2.31e-06 ***
Fireplaces     0.057429   0.042071   1.365 0.173231
PoolArea       0.013177   0.018285   0.721 0.471674
FullBath       0.034681   0.029340   1.182 0.238107
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1155 on 309 degrees of freedom
Multiple R-squared:  0.859,     Adjusted R-squared:  0.8499
F-statistic: 94.15 on 20 and 309 DF,  p-value: < 2.2e-16
```

Cluster 3:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.83364 -0.03822  0.01481  0.06302  0.41402

Coefficients: (1 not defined because of singularities)
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  -52.970328   8.649333  -6.124 3.09e-09 ***
OverallQual    0.258959   0.054715   4.733 3.52e-06 ***
GrLivArea      0.407907   0.073677   5.536 7.13e-08 ***
BsmtFinSF1     0.008570   0.003688   2.324 0.020849 *
GarageCars     0.065163   0.024468   2.663 0.008191 **
MSSubClass     0.002763   0.014641   0.189 0.850444
YearBuilt      7.056213   0.964052   7.319 2.67e-12 ***
OverallCond    0.281539   0.044995   6.257 1.47e-09 ***
BedroomAbvGr  -0.005214   0.039835  -0.131 0.895962
LotArea        0.120193   0.037926   3.169 0.001699 **
MasVnrArea     0.003986   0.004178   0.954 0.340867
BsmtFullBath  -0.047595   0.207543  -0.229 0.818786
TotRmsAbvGrd  -0.050706   0.074738  -0.678 0.498045
ScreenPorch    0.010925   0.006771   1.614 0.107750
WoodDeckSF     0.006021   0.003522   1.710 0.088452 .
YearRemodAdd   0.725758   0.946317   0.767 0.443772
TotalBsmtSF    0.130376   0.037585   3.469 0.000605 ***
KitchenAbvGr          NA         NA      NA       NA
Fireplaces     0.065029   0.090593   0.718 0.473470
PoolArea       0.017122   0.021833   0.784 0.433574
FullBath      -0.023633   0.046128  -0.512 0.608818
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.136 on 279 degrees of freedom
Multiple R-squared:  0.7414,    Adjusted R-squared:  0.7238
F-statistic:  42.1 on 19 and 279 DF,  p-value: < 2.2e-16
```

IV. RESULTS

Feature Selection + Regression
- R-Squared: 0.8609
- Adjusted R-Squared: 0.8589

Variable Transformation + Feature Selection + Regression
- R-Squared: 0.8856
- Adjusted R-Squared: 0.8832
- RMSE: 1780.834

Clustering + Regression
(3 clusters – LotArea, Above grade living area)
- RMSE : 1602.174

The results from the 3 different models are shown in the figure above.
For Model 1 (Feature Selection + Regression), an R-Squared of 0.8609 and an adjusted R-Squared of 0.8589 was obtained, which is actually a decent value.

Nonetheless, in the pursuit of getting a better model, Model 2 (Variable Transformation + Feature Selection + Regression Model) was built which yielded in values. Both the R-Squared and the Adjusted R-Squared values increased to 0.8856 and 0.8832 respectively. The Root Mean Squared Error (RMSE) was calculated as 1780.834.

For the final model, Model 3 (with Clustering + Regression) which consisted of 3 clusters based on LotArea and Above grade living area, the regression model proved to be even better in terms of RMSE as it was determined to be 1602.174.

Figure 12: Predicted vs Actual Sale Price

The above *Figure 12* shows the Predicted Price points versus the actual over the complete range of data. The orange data points show the precited price from the final model, clustering of data using K-means model and regression on each of the cluster. The clustering is done using the features lot area and above ground living area. The idea is to find the inherent divide in data using features. The idea to classify is to identify the features that might be prevalent in one class that might not be as relevant in other classes. A good, but generic, example is a house with a pool in the house might be a relevant feature if the house is of a bigger lot size. With smaller lot size, a pool might not be relevant feature. As seen in the figure the orange points lie closer to the line - actual price versus actual price line plotted for easier understanding and error identification – that the points marked in the green. The price predicted using model 2 are plotted in green, the model uses features that are selected using feature selection and transformed features.

As the price range increases, we can see the cluster and regression model has not overestimated the price points but have underestimated in some cases. At lower price points the cluster and regression model, model 3, has produced better results than the model 2 which can be seen from the closeness of the points to the actual price line, Model 2 however has overestimated some price points in the lower price range. The error in prediction is greater at higher price points for both the models but Model 3 has some less error than model 2.

Using all the features the regression model alone was able to produce a greater adjusted R squared metric than the other models on the seen data. But the generalizability of the model has greatly reduced, all three models presented, have outperformed the model using all the features on the unseen dataset. Given the case, the model 2 has higher R squared than the model 1 and has also outperformed the model 1 on the unseen test dataset. The model 3, having three clusters of data and a multiple linear regression model for each of the cluster. The root mean squared error for model 3 was around ten percent lower than that of the model 2.

Figure 13: Predicted vs Actual Sale Price by Class for Model 3

From *Figure 3*, we can observe three different clusters of sales price. The blue dots represent the cluster of SalePrice which are mostly on the lower side. The green triangles represent the cluster of SalePrice which are on the intermediate range. The red dots represent the cluster of SalePrice which are mostly on the higher range.

V. BROADER IMPACT

With further analysis and advanced statistical techniques, we can develop model to adapt for anomalies with the unseen dataset. Not only buyers can easily access the price of their dream house, but the sellers can also reasonably perform their property valuations on their own.

The model can be further scaled up as a product (mobile app/website) where the buyers/sellers can easily list their property as well as gain the insight of house prices based on the house features for a region. With further emphasis on the scalability of the product, the system can be developed to handle housing transactions without the need of "middle-men".

VI. DISCUSSION

The real estate industry makes extensive use of big data to make price prediction and fluctuation, and it is one of the key areas that use machine learning concepts to accurately estimate costs. Thus, with the motive of understanding the influencing factors of the housing prices, we conducted the research. We started out by performing EDA on the dataset, and then pre-processed the data using data wrangling techniques such as taking care of missing values, encoding categorical variables, and variable transformation. The 3 different models were built using the selected features from the dataset.

Starting from model 1, we obtained a R-squared value which was further improved in model 2. Thus, in our case, taking the log transformation of both the independent and dependent variables produced a better model. Next, clustering was done using selective features from the dataset prior to building the regression model. A significant reduction by 10% in RMSE was obtained for the model 3 as compared to model 2.

However, we would like to further develop our model using advanced statistical techniques to account for the abnormalities that might be encountered on working with variability of unseen dataset.

All the coding was done in R.

ACKNOWLEDGMENT

We would like to express our appreciation to Professor Nazmus Sakib for his valuable and constructive suggestions during the planning and development of this project work. His willingness to give his time so generously has been very much appreciated.

REFERENCES

[1] *Dean De Cock "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project". Journal of Statistics Education Volume 19, Number 3(2011),* www.amstat.org/publications/jse/v19n3/decock.pdf

[2] *Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani "An Introduction to Statistical Learning with Applications in R Second Edition.*

[3] *Michael Patrick Allen, Understanding Regression Analysis*