Final Project:
Principal Component Analysis, Exploratory Factor Analysis and k-Means Clustering

Serra Uzun, MSDS_411 FALL 2020
11/19/2020

Introduction

Unsupervised learning methods give a first and thorough look into large datasets and provide useful insights that can be directly practiced or used as a part of further modeling. The following report will be conducting multiple unsupervised models on the Ames Iowa Housing data to determine if and how predictor variables impact SalePrice, which causes a visible grouping of observations due to high correlation, similarities, and underlying patterns.

The Dataset

The Ames Iowa Housing dataset contains the information on residential properties listed in the housing market in Iowa. The raw dataset consists of 82 variables and 2930 observations. With 81 predictor variables, the dataset provides almost every aspect of the homes listed in Iowa, such as their indoor and outdoor design features, area take-offs, HVAC systems, lot properties, etc. The dataset's response variable is SalePrice, which is the sale price of the home in dollars.

SID (num): Sale ID
PID (num): Post ID
SubClass (num): The building class
Zoning (chr): The general zoning classification
LotFrontage (num): Linear feet of street connected to property
LotArea (num): Lot size in square feet
Street (chr): Type of road access
Alley (chr): Type of alley access
LotShape (chr): General shape of property
LandContour (chr): Flatness of the property
Utilities (chr): Type of utilities available
LotConfig (chr): Lot configuration
LandSlope (chr): Slope of property
Neighborhood (chr): Physical locations within Ames city limits
Condition1 (chr): Proximity to main road or railroad
Condition2 (chr): Proximity to main road or railroad (if a second is present)
BldgType (chr): Type of dwelling
HouseStyle (chr): Style of dwelling
OverallQual (num): Overall material and finish quality
OverallCond (num): Overall condition rating
YearBuilt (num): Original construction date
YearRemodel (num): Remodel date
RoofStyle (chr): Type of roof
RoofMat (chr): Roof material
Exterior1 (chr): Exterior covering on house
Exterior2 (chr): Exterior covering on house (if more than one material)
MasVnrType (chr): Masonry veneer type
MasVnrArea (num): Masonry veneer area in square feet
ExterQual (chr): Exterior material quality

ExterCond (chr): Present condition of the material on the exterior
Foundation (chr): Type of foundation
BsmtQual (chr): Height of the basement
BsmtCond (chr): General condition of the basement
BsmtExposure (chr): Walkout or garden level basement walls
BsmtFinType1 (chr): Quality of basement finished area
BsmtFinSF1 (num): Type 1 finished square feet
BsmtFinType2 (chr): Quality of second finished area (if present)
BsmtFinSF2 (num): Type 2 finished square feet
BsmtUnfSF (num): Unfinished square feet of basement area
TotalBsmtSF (num): Total square feet of basement area
Heating (chr): Type of heating
HeatingQC (chr): Heating quality and condition
CentralAir (chr): Central air conditioning
Electrical (chr): Electrical system
FirstFlrSF (num): First Floor square feet
SecondFlrSF (num): Second floor square feet
LowQualFinSF (num): Low quality finished square feet (all floors)
GrLivArea (num): Above grade (ground) living area square feet
BsmtFullBath (num): Basement full bathrooms
BsmtHalfBath (num): Basement half bathrooms
FullBath (num): Full bathrooms above grade
HalfBath (num): Half baths above grade
BedroomAbvGr (num): Number of bedrooms above basement level
KitchenAbvGr (num): Number of kitchens
KitchenQual (chr): Kitchen quality
TotRmsAbvGrd (num): Total rooms above grade (does not include bathrooms)
Functional (chr): Home functionality rating
Fireplaces (num): Number of fireplaces
FireplaceQu (chr): Fireplace quality
GarageType (chr): Garage location
GarageYrBlt (num): Year garage was built
GarageFinish (chr): Interior finish of the garage
GarageCars (num): Size of garage in car capacity
GarageArea (num): Size of garage in square feet
GarageQual (chr): Garage quality
GarageCond (chr): Garage condition
PavedDrive (chr): Paved driveway
WoodDeckSF (num): Wood deck area in square feet
OpenPorchSF (num): Open porch area in square feet
EnclosedPorch (num): Enclosed porch area in square feet
ThreeSsnPorch (num): Three season porch area in square feet
ScreenPorch (num): Screen porch area in square feet
PoolArea (num): Pool area in square feet
PoolQC (num): Pool quality
Fence (chr): Fence quality
MiscFeature (chr): Miscellaneous feature not covered in other categories
MiscVal (num): Value of miscellaneous feature
MoSold (num): Month Sold
YrSold (num): Year Sold
SaleType (chr): Type of sale
SaleCondition (chr): Condition of sale
SalePrice (num): the property's sale price in dollars.

Research Question

Our analysis aims to determine if the size of a single-family house built in the past 40 years (between 1970 and 2010) in Iowa results in any segmentation of our response variable, SalePrice. So, if we were to state our question specifically, it would be:

"Does the size of a single-family suburban house that was built in the past 40 years in Iowa has an effect on Sale Price that causes obvious segmentation?"

We will start by creating several drop conditions to eliminate unnecessary data and gather the eligible observations and variables per our research question. Following this step, we will conduct a data preparation where we will have a first look at the missing values, outliers, and abnormalities within the dataset before the exploratory analysis stage.

Sample Definition and Data Preparation

Before the exploratory data analysis, we will be implementing the conditions to the dataset to establish a sample group that contains only the observations and variables that are eligible per our research question. In addition to narrowing the dataset to single-family homes built post-1970, we will remove all the listings of the houses that do not come with utilities, have had a normal sale, no basement, and no garage. 505 observations are not single-family homes, 423 observations of not normal home sales, 1054 observations of homes built pre 1970, 1 observation without a basement, and 14 observations without a garage space. When we drop the observations that do not meet our criteria, we are left with 933 observations.

The variables relevant to our research question, which indicates that we will be looking into the effect of the 'size' of the home in the listing, we will be only keeping the area and room count related variables. Per these criteria, below is the list of variables we will be using in the remainder of our analysis.

1. LotArea
2. TotalBsmtSF
3. TotAbvGrdSF
    - (FirstFlrSF+SecondFlrSF)
4. GrLivArea
5. FullBath
6. HalfBath
7. BedroomAbvGr
8. BsmtFullBath
9. BsmtHalfBath
10. TotRmsAbvGrd
11. GarageArea
12. WoodDeckSF
13. OpenPorchSF
14. YearBuilt
15. SalePrice

We will be continuing our analysis with the 14 predictor and one response variables as listed above. The FirstFlrSF and SecondFlr SF were added together to make a new variable called TotAbvGrdSF (Total Above Grade SF)

Exploratory Data Analysis

We start our EDA with 933 observations, 14 predictors, and one response variable. The predictor variables are informally separated into two groups as area-related variables, all continuous, and room count related variables, all discrete. In order to see the general distribution of all types of individual variables, we conduct a histogram plot, presented in Figure 1 below.

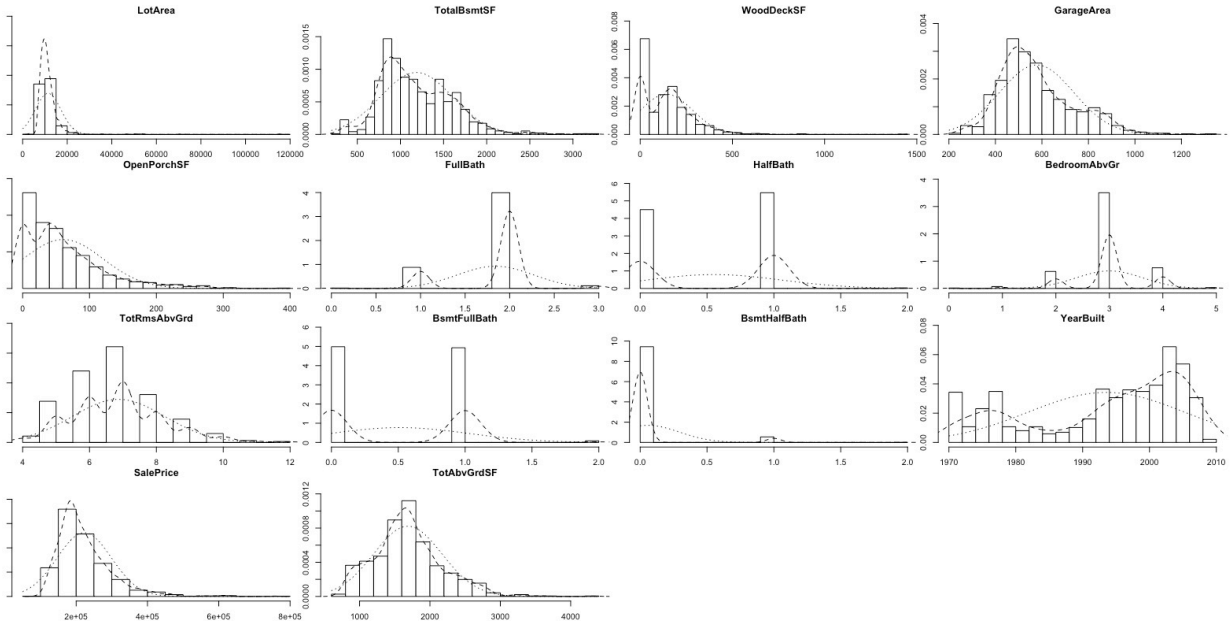


Figure 1: Histogram Plot of All Selected Variables

Upon reviewing the histogram plot in Figure 1, we see a general left skewness with histograms of continuous variables, which we mentioned before, area/square footage related. The left skew suggests possible outliers towards the higher end of the spectrum. On the other hand, if we observe the YearBuilt histogram, we can see that our sample dataset's house listings are mostly of single-family houses built in recent years, specifically post 1990. Finally, if we look at the SalePrice histogram plot, we can see that generally, single-family homes in Iowa are priced between $100,000 and $300,000. Next, we will explore the outliers through boxplots..

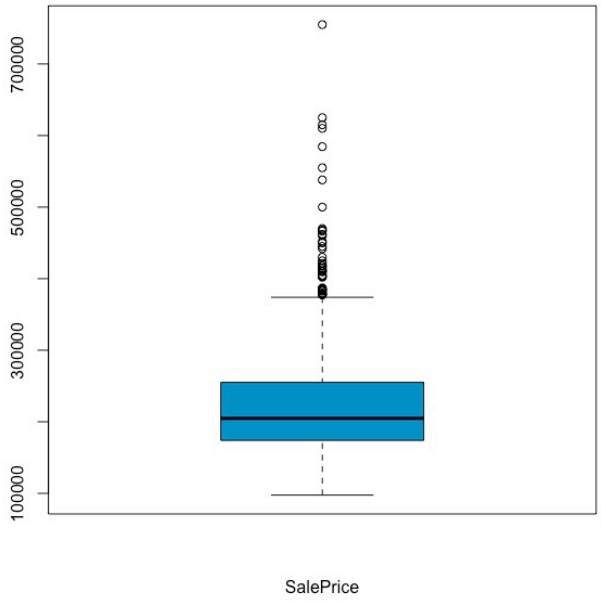


Figure 2: Boxplot for SalePrice

The boxplot of the response variable, Sale Price, shows some considerable outliers. As we also saw from the histogram plot, the distribution of the SalePrice is accumulated mainly between $100,000 and $300,000, so having several Sale Prices above $500,000 can be considered concerning. Before removing any of the outliers, we will boxplot the continuous variables to see if multiple variables need cleaning.

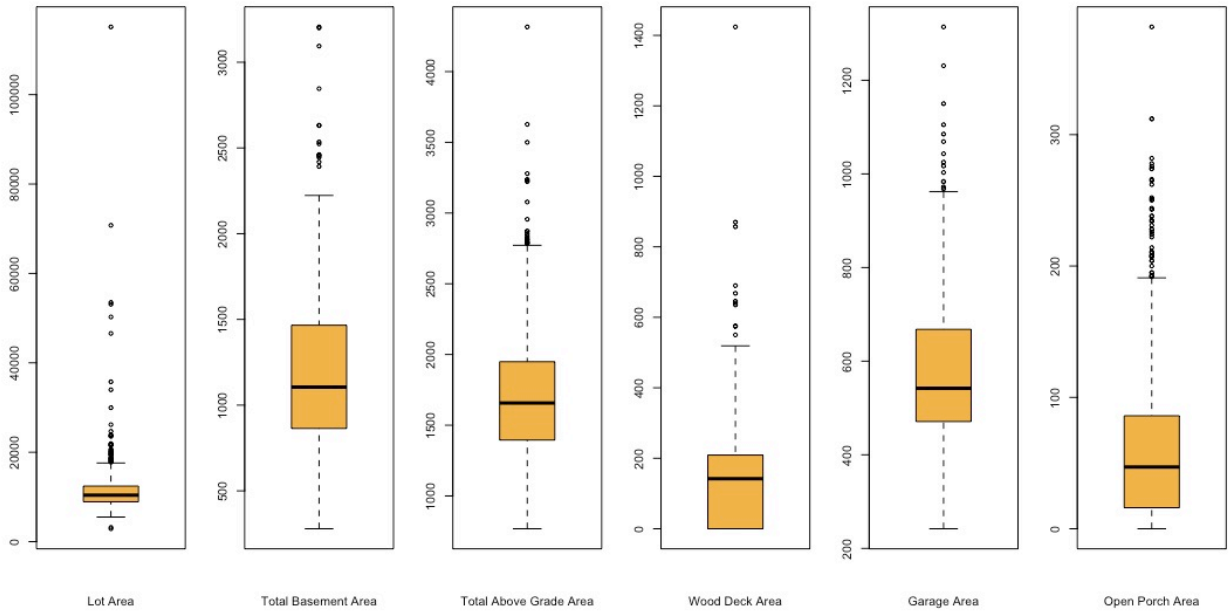


Figure 3: Boxplot for Area Related Continuous Variables

The boxplots presented in Figure 3 show outliers for several predictor variables: Lot Area, Wood Deck Area, Open Porch Area, and Total Above Grade Area. As the observations that are extreme outliers of both variables listed above and the response variable can skew and negatively impact our results, we will keep the observations that fall between Q1 and Q3, plus IQR of 1.5.

Exhibited in Figure 4 and Figure 5 below, the boxplots show a more uniform distribution of observations around and beyond the Q3. Including with our response variable's boxplot, Sale Price, there are no longer extreme outliers that are too distant and off from the observations seen under the same variable. With this data cleaning, the skewness of the observations observed in Figure 1 area corrected, and the dataset is prepared for analysis. Through outlier removal, a total of 132 observations were taken out of our working dataset, which leaves us with 801 observations moving forward.

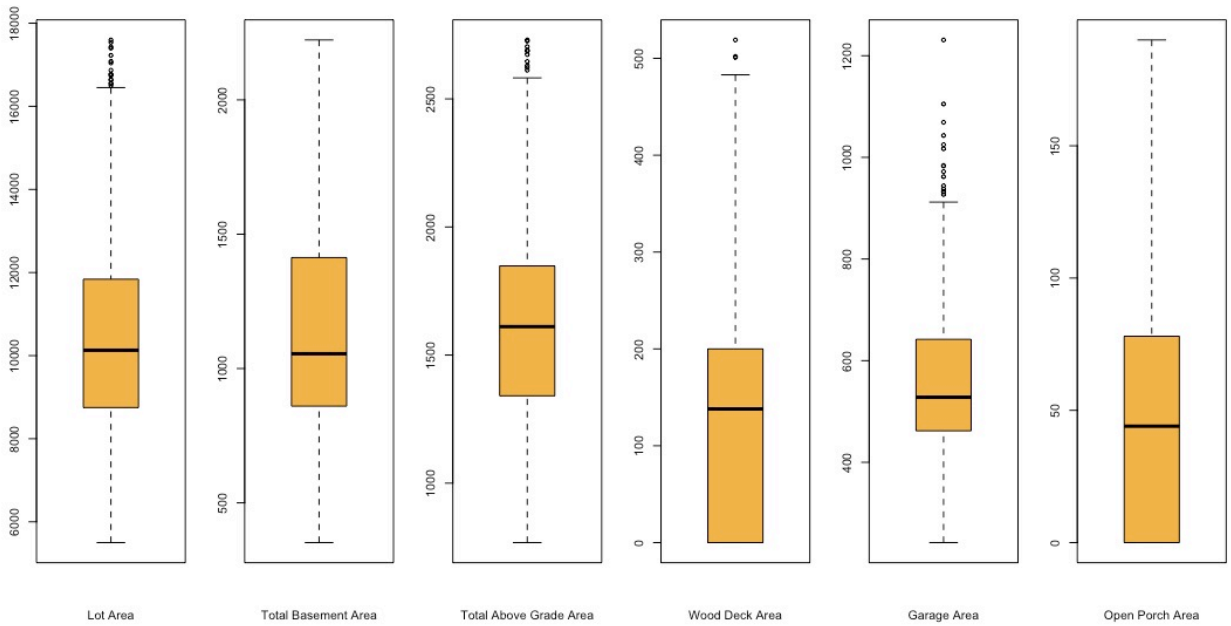


Figure 4: Boxplot for Area Related Continuous Variables without Outliers

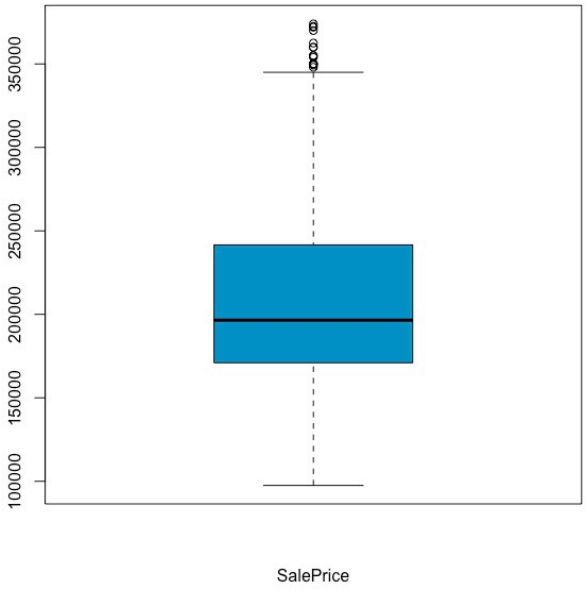


Figure 5: Boxplot for SalePrice without Outliers

As a part of our EDA, we will also conduct a correlation plot with all predictor variables and response variable. The correlation plot is conducted to visualize high or love correlations between variables and a possible sign of multicollinearity. Our threshold in identifying high and low correlations are +/- 0.5. The correlation plot is presented in Figure 6 below.

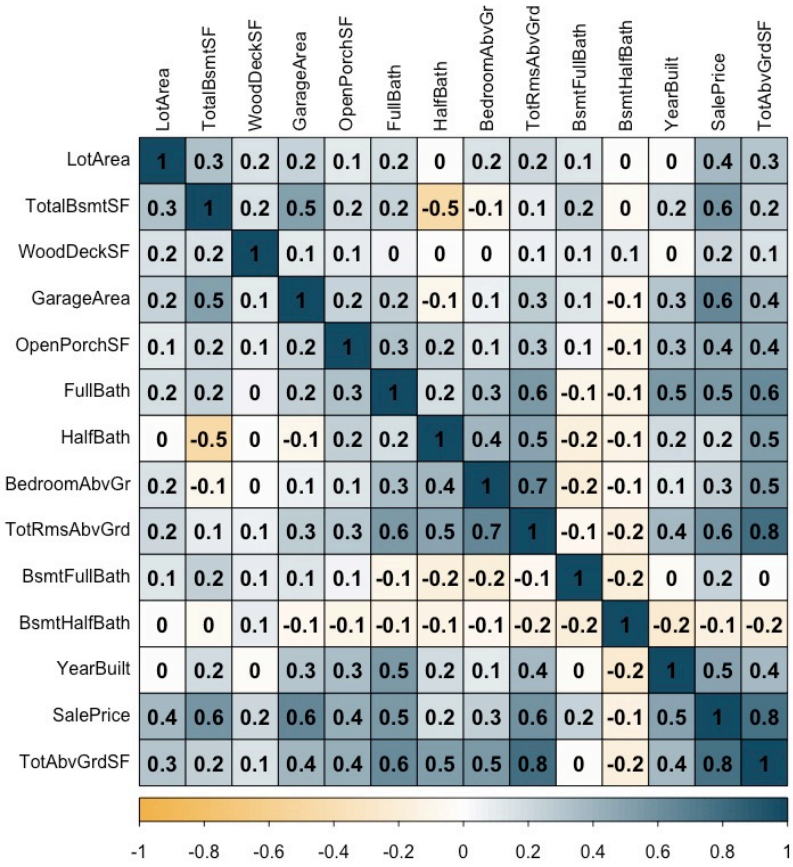


Figure 6: Correlation Plot

The correlation plot presents several strong and noticeable correlations between the response variable and predictor variables, as well as between predictor variables alone. If we first look at the final row of the correlation plot where we have the response variable, we can see that TotAbvGrdSF is the predictor variable that has the highest correlation with SalePrice with 0.8, followed by TotalBsmtSF, GarageArea and TotRmsAbvGrd, all have 0.6 positive correlation with the response variable. As three out of four variables that are most correlated with response variable are area related, we can determine from our correlation plot that the square footage of a house has more impact on the SalePrice than both size of the outdoor areas and count of the rooms. The home's size has a more noticeable impact on SalePrice than how many rooms it has or its lot size.

Other high negative and correlations observed in Figure 6 are between TotAbvGrdSF and TotRmsAbvGrd with a positive correlation of 0.8, and HalfBath and TotalBsmtSF with a negative correlation of -0.5. The positive correlation observed between the TotAbvGrdSF and TotRmsAbvGrd is expected as the more rooms a house has, the more area it will have. The negative correlation between HalfBath and TotalBsmt is interesting and indicates that the larger the Total Basement SF the house has, the less likely that it has a HalfBath.

As the final step of our EDA, we will look at some scatter plots between the response variable, SalePrice, and variables that hold the information on indoor and outdoor areas, Total Above Grade Area, and Lot Area. We know from our correlation plot that TotAbvGrdSF has a high correlation with the response variables while LotArea correlates with SalePrice of 0.4, below our threshold.

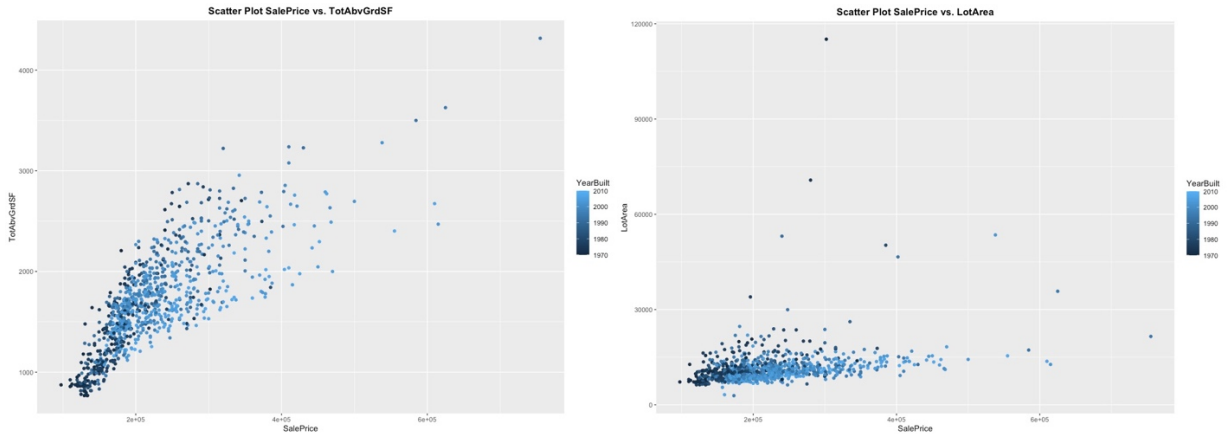


Figure 7: Scatter Plots for SalePrice vs. GrLivArea (left) and SalePrice vs. LotArea (right)

The scattering of the observations presented in each scatter plot is parallel with the correlation plot results. The observations were colored per the YearBuilt of the home listing, which shows few critical insights. Firstly, from the plot on the right, we can see that even though LotArea has not increased in years, the SalePrice has continually increased. The newer homes are presented to be more expensive; the LotArea of the houses has remained close to what it was 30-40 years ago.

Additionally, if we look at the plot on the left, we can see that as the TotAbvGrdSF has increased, the SalePrice has also increased over the years. The scatter plot also suggests that newer homes are relatively larger and more expensive; the houses that have similar TotAbvGrdSF with newer homes tend to cost less. So, while we did not see a change in LotArea against the SalePrice in the past 40 years, we have seen an increase in TotAbvGrdSF against SalePrice between 1970 and 2010.

Principal Component Analysis

As the first step of the PCA, we start with a scree plot, and the total variance explained. While the scree plot shows us the eigenvalue for each component, the total variance explained plot shows us the cumulative explained variance. Per the scree plot by the third component, we reach an eigenvalue around 0.1. Per this scree plot presented on the left in Figure 8, the elbowing occurs when we reach the third component, so the optimal number of components is likely to be three.

On the other hand, from the total variance explained plot on the right side of Figure 8, we can see that we reach our minimum variance threshold with seven components, which is 80%. This suggests that with the first seven components, we can explain over 80% of the variance in the Ames Dataset.

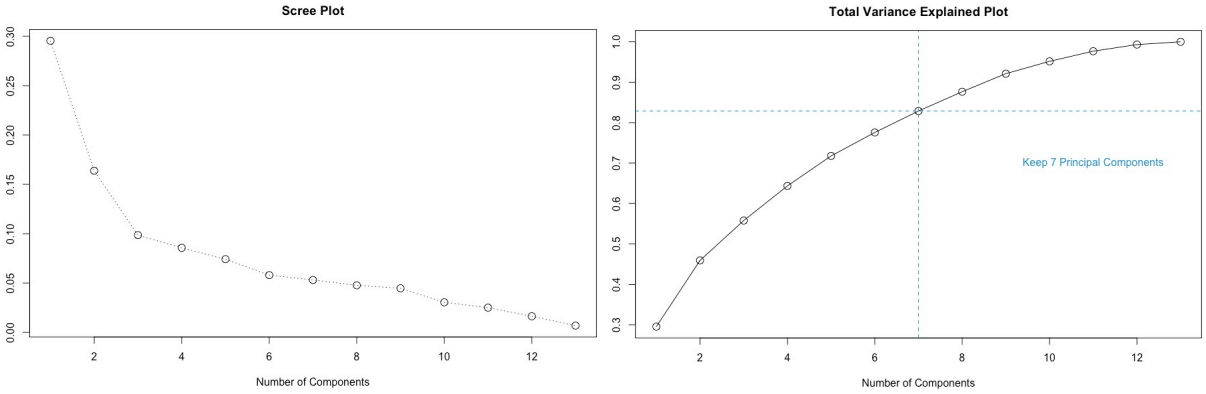
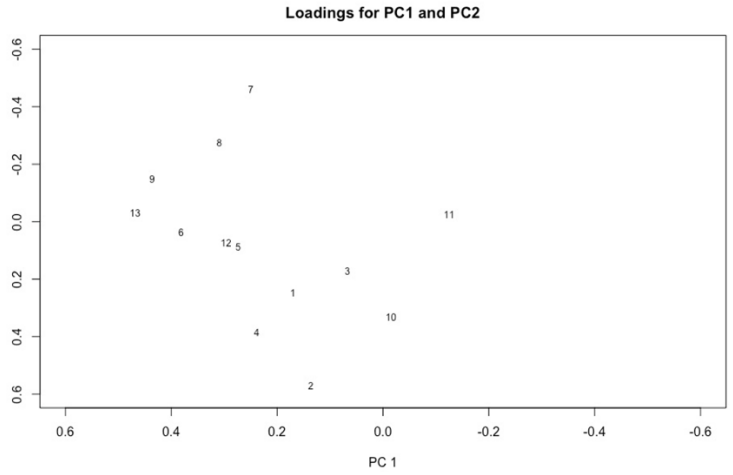


Figure 8: Scree and Total Variance Plots

Next, we will take principal component 1 (PC1) and principal component 2 (PC2). They are the two components with eigenvalues over 0.1 and can explain about 50-60% of the variance in the dataset plot their loadings against each other. The PC1 and PC2 loadings plotted against each other are presented in Figure 9 below.



		Comp.1	Comp.2
1	LotArea	0.1701	0.2467
2	TotalBsmtSF	0.1364	0.5704
3	WoodDeckSF	0.0672	0.1723
4	GarageArea	0.2382	0.3832
5	OpenPorchSF	0.2734	0.0878
6	FullBath	0.3813	0.0366
7	HalfBath	0.2496	-0.4616
8	BedroomAbvGr	0.3093	-0.2747
9	TotRmsAbvGrd	0.4362	-0.1489
10	BsmtFullBath	-0.0155	0.3327
11	BsmtHalfBath	-0.1252	-0.0251
12	YearBuilt	0.2964	0.0723
13	TotAbvGrdSF	0.4680	-0.0319

Figure 9: PC1 and PC2 Loadings Scatterplot

Per the distribution of variables within the Loadings plot, we can see that area related variables such as Lot Area, Total Basement SF, Wood Deck SF, and Garage Area are the variables that are most loaded in PC2, whereas PC1 is loaded more with variables that are related to room count such as Total Rooms Above Grade, Full Baths, etc. The Total Above Grade Area variable, which is the predictor variable that we observed to be most positively correlated with the response variable, is heavily loaded in PC1 but almost none in PC2. Finally, the Basement Half Bath variable is the least significant of all variables when PC1 and PC2 are concerned.

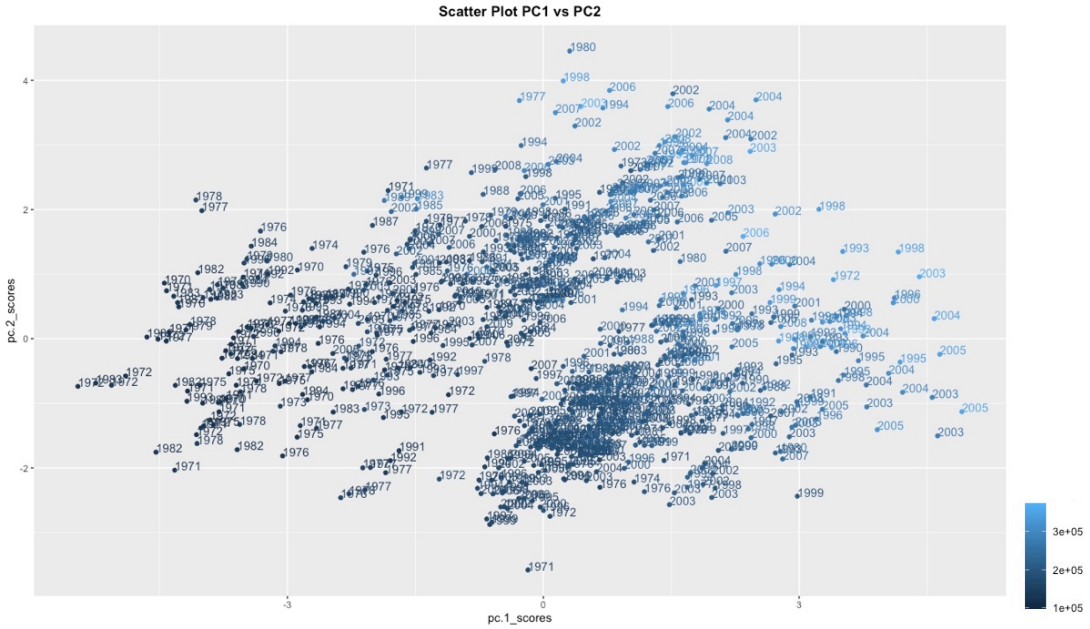


Figure 10: PC1 and PC2 Scores Scatterplot

Lastly, the above scatter plot presented in Figure 10 is the PC1 and PC2 scores plotted against each other. The SalePrice and labels on them color the scores show the YearBuilt. Through this plot, we can see some gradual distribution amongst scores, with the older homes having negative scores in both components, and the scores increase as the home has a YearBuilt that is more recent. The homes that are built between 1970 and 1985 have low sale prices and have relatively negative scores. Almost all homes built between 1985 and 1995 have a score of about zero in PC1 and relatively low scores in PC2 while having a medium sale price. Finally, the most recently built homes have higher prices and have higher scores in PC1 and PC2.

Exploratory Factor Analysis

The exploratory factor analysis is the method chosen as a part of our analysis to uncover any underlying or hidden strong correlations within the dataset. The exploratory factor analysis is the method selected as a part of our analysis to uncover any underlying or hidden strong correlations within the dataset. As the initial step of the EFA, we plot the scree plot for both the PCA conducted in the previous section and EFA together. Once we review the plot, we see that the FA Actual Data and FA Simulated Data lines intersect at factor number 5. Per our parallel analysis with 801 observations and 100 iterations, our model and plot suggest that the number of factors to be 5 and the number of components to be 3. This outcome is taken into account; we will be conducting the EFA with five factors, varimax rotation, and the most likelihood factoring method as our next step.

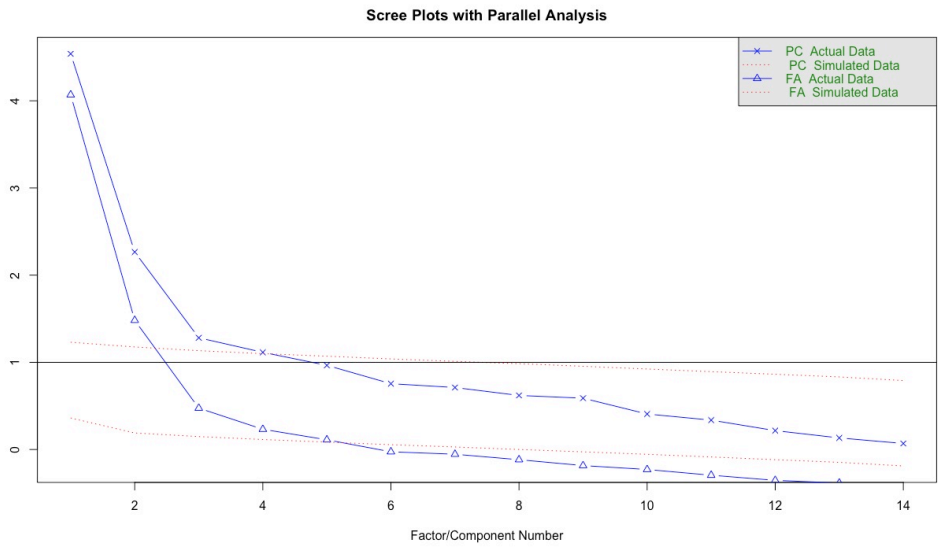


Figure 11: Scree Plots with Parallel Analysis

The 5-factor varimax rotated factor analysis shows that factor 3 (ML3) is the most loaded factor with 2.11 SS Loadings, followed by factor 4 (ML4), and factor 2 (ML2) with 2.03 and 1.91 SS loadings, respectively. When we further look into the factors presented in Table 1 below, we see that factor 3 is heavily loaded with SalePrice, Total Basement SF, and Garage Area. Next, when we look at factor 4, we see that both Bedroom Above Grade and Total Rooms Above Grade are highly loaded, together with Total Above Grade Area. In factor 2, YearBuilt is weighted above all other variables with 0.96 and, together with SalePrice, are heavy loaders in ML2. Lastly, if we observe the final two factors with least SS Loadings, which are factor 1 and 5, we see that factor 1 has a relatively well distributed and insignificant spread of loadings amongst variables, whereas factor 5 has FullBath variable at 0.58 loadings, significantly more than any other variable.

	ML3	ML4	ML2	ML1	ML5
LotArea	0.41	0.17	-0.02	0	0.09
TotalBsmtSF	0.67	-0.04	0.16	-0.45	0.17
WoodDeckSF	0.26	0.01	-0.03	0.03	0.04
GarageArea	0.6	0.11	0.33	-0.12	-0.01
OpenPorchSF	0.24	0.08	0.33	0.19	0.23
FullBath	0.12	0.39	0.45	-0.06	0.58
HalfBath	-0.15	0.34	0.21	0.9	0.04
BedroomAbvGr	-0.03	0.81	0.05	0.12	0.04
TotRmsAbvGrd	0.19	0.75	0.33	0.22	0.19
BsmtFullBath	0.4	-0.18	0.02	-0.04	-0.15
BsmtHalfBath	-0.03	-0.07	-0.23	-0.07	-0.01
YearBuilt	-0.02	0.06	0.96	-0.05	0.15
SalePrice	0.75	0.29	0.49	0.08	0.21
TotAbvGrdSF	0.45	0.6	0.32	0.36	0.39
SS loadings	2.11	2.03	1.91	1.27	0.71
Proportion Var	0.15	0.14	0.14	0.09	0.05
Cumulative Var	0.15	0.3	0.43	0.52	0.57
Proportion Explained	0.26	0.25	0.24	0.16	0.09
Cumulative Proportion	0.26	0.52	0.75	0.91	1

Table 1: Factor Loadings

The below correlation plot shows each variable as well as each factor presented above. We choose to save factor scores of each factor and plot it in a correlation plot together with all the variables to visualize and understand the underlying relationships of the variables within our dataset.

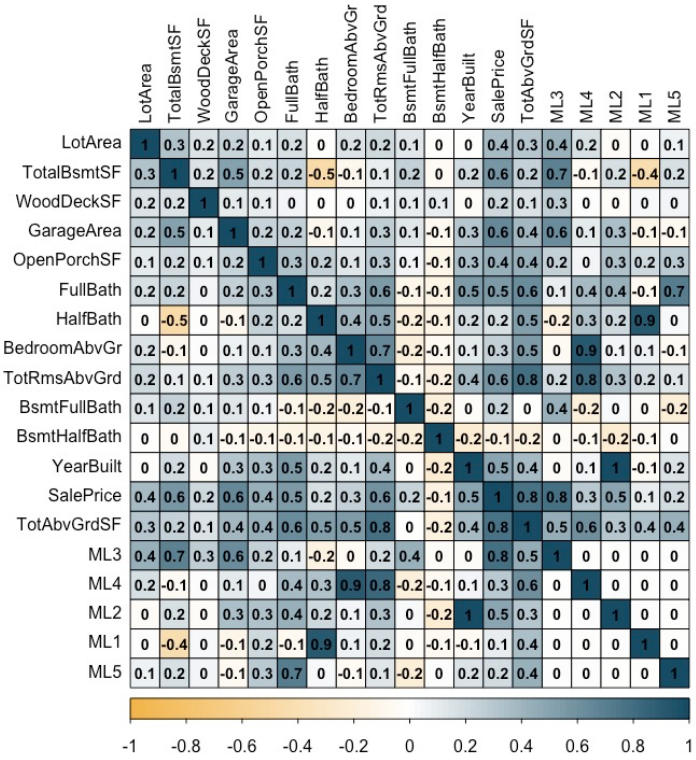


Figure 12: Correlation Plot with Factors

Through the correlation plot, we are able to visualize the loadings in the factor of each variable through strong positive or negative correlations. For example, ML4 is heavily correlated with Bedroom Above Grade and Total Rooms Above Grade, which we already determined through the above. Yet, we can now clearly see that with 0.9 and 0.8 correlations, we can remove one of them from our dataset. Another observation we make through the correlation plot that we haven't already discussed is the low correlation of Lot Area, Wood Deck SF, Open Porch SF, Basement Full Bath, and Basement Half Bath with all factors. These variables do not have a correlation over +/- 0.5, which is our threshold, which suggests that these variables do not significantly impact the variance within the dataset.

As we wrap up the EFA, we can conclude that Bedroom Above Grade and Total Rooms Above Grade are positively correlated with an individual factor, and we will not be losing information from our dataset if we are to keep one and remove the other. In addition, we see that some variables, such as Year Built, Half Bath, Full Bath, Total Basement SF, Garage SF are predictor variables that are significant and should remain in our dataset. Moving forward, we will not be including the TotRmsAbvGrd variables in our analysis per EFA outcome.

k-Means Clustering

k-Means Clustering is a widespread method to determine grouping within the dataset due to common similarities between observations and variables. We start the k-means clustering analysis with the 'optimal number of clusters' plot. Presented in Figure 13 below, we have the Between Sum of Squares Percentage (right) and Within the Sum of Squares (left) plotted against the number of clusters. By looking at the y-axis and the BSS and WSS, we see that we can get relatively ineffective and insignificant clustering through our model. The elbowing in the plot occurs around 5 clusters, yet the BSS is around 50% as we reach that point, which is quite lower than what we would have hoped to see.

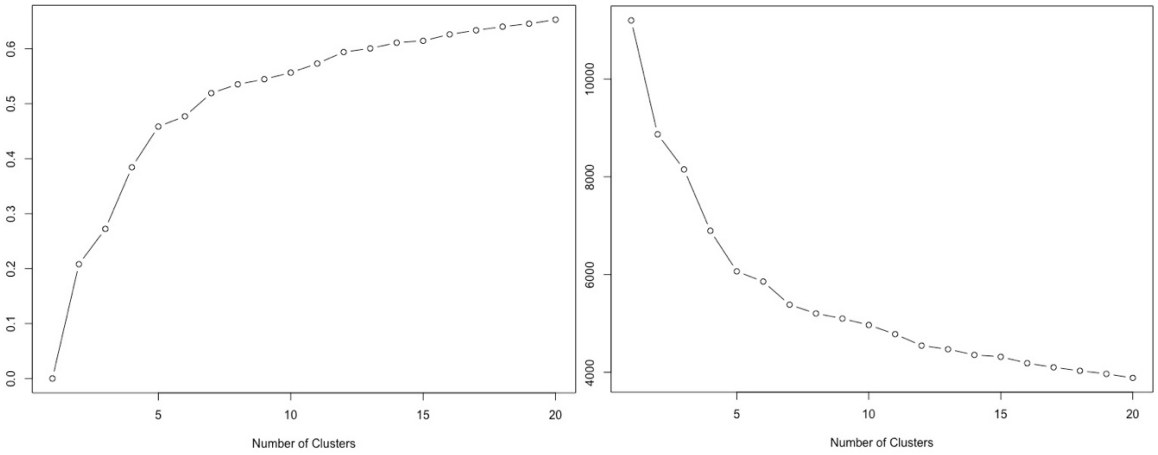


Figure 13: Optimal Number of Clusters

Even though we are aware that it will not give us a high BSS and a low WSS, as the elbowing occurs around 5 clusters in the above plots, we will be using k=5 for our analysis.

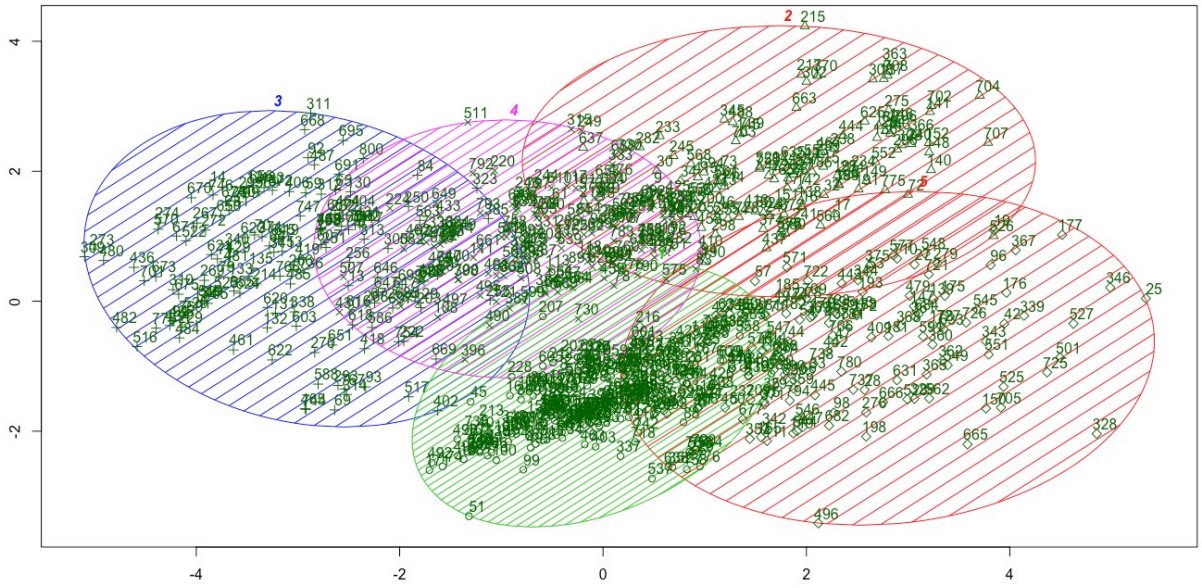


Figure 14: Cluster Plot with k=5

	LotArea	TotalBsmtSF	WoodDeckSF	GarageArea	OpenPorchSF	FullBath	HalfBath
1	-0.327	-0.876	-0.251	-0.493	0.133	0.391	0.921
2	0.653	1.503	0.397	1.181	0.544	0.358	-0.891
3	-0.440	-0.571	-0.095	-0.533	-0.873	-1.966	-0.500
4	-0.052	0.621	-0.154	-0.258	-0.310	0.456	-1.051
5	0.525	0.097	0.373	0.709	0.588	0.527	0.926

	BedroomAbvGr	BsmtFullBath	BsmtHalfBath	YearBuilt	SalePrice	TotAbvGrdSF
1	0.170	-0.444	-0.159	0.373	-0.281	0.203
2	-0.482	0.378	-0.235	0.710	1.118	0.188
3	-0.593	0.195	0.294	-1.201	-1.178	-1.358
4	-0.309	0.002	0.254	-0.181	-0.313	-0.459
5	1.075	0.234	-0.140	0.235	1.128	1.432

Table 1: Cluster Means

Firstly, when the cluster plot in Figure 14 is reviewed, we can see that cluster #1 is the densest and the most crowded cluster with 236 data points. We observe several significant overlaps and no distinct clusters that indicate high similarities between data points. The Between Sum of Squares of the 5-cluster model is approximately 43%, meaning our model is not a great fit, and the dispersion between data points is not significant. Yet if we look into the cluster means for each variable, we see that the cluster means to tell a similar story with what we have uncovered so far through PCA and EFA.

Within cluster #2, we see that Total Basement SF and Garage Area, followed by Sale Price, Lot Area, and Open Porch SF has high means, suggesting that through these characteristics of the listing, the data points show similarities in cluster #2. Another cluster to review is cluster #5, where the mean for Sale Price is 1.128. In addition to the high mean of Sale Price, the mean for Above Grade SF and Bedrooms Above Grade are noticeably high as well.

In addition to cluster #2 and #5, the SalePrice mean is relatively high in cluster #3, which is where we see that YearBuilt and Total Above Grade SF also has high negative cluster means. As the means of these variables are high and relatively parallel and close to each other in cluster #3, we can say that cluster #3 suggests a similarity between Year Built and Total Above Grade SF and SalePrice.

While our analysis doesn't show a clear and distinct segmentation or clustering of SalePrice, the observations mentioned above indicate (a) per cluster #2 as Total Basement SF, Garage SF, Open Porch and Lot Area increases, Sale Price is also very likely to increase, and (b) Bedrooms Above Grade and Total Above Grade SF also show significant similarities with SalePrice in the context of cluster #5.

Conclusion

Exploratory Data Analysis, Principal Component Analysis, Exploratory Factor Analysis, and k-Means Clustering analysis all shown us valuable insights on the underlying correlations, similarities, and patterns. As we have taken on the sample dataset that we initially set through our criteria, through these analyses, we observed that:

1. SalePrice of a single-family home with a basement and garage in Iowa ranges typically between $100,000 and $300,000.
2. Square footage/area related variables have a larger impact on SalePrice than room count related variables.
3. YearBuilt is not the variable with the highest correlation with SalePrice, yet it is a noticeable factor in changing SalePrice observed dataset.
4. Total Basement SF and Half Bath count have a considerable negative correlation, and the presence of one can indicate the absence of the other in cases of missing values.

Per our studies, we can state that size of indoor spaces has a more noticeable effect on SalePrice than the room counts, which suggests that counts are secondary variables in predicting SalePrice in the presence of SF related information. Variables such as Year Built and Total Above Grade SF, in addition to Lot Area, Garage Area, and Number of Bedroom Above Grade, are good candidates for a linear regression model following this analysis and the top predictors of Sale Price of a single-family home in Iowa.