Statistical analysis on Ames House Prices

Unsupervised Learning Algorithms

Clustering

Professor:

Silvia Salini

Presenter:

Shojaat Joodi Bigdilo

Student number: 14088A

June 26, 2023

List of contents:

1. Introduction

The following analysis takes place on the **Ames House Price Estimation dataset** was compiled by Dean De Cock for use in data science education The goal of this analysis is to explore the possibility of exploiting the data by employing k-Means Clustering algorithm which is one of **Unsupervised learning** techniques. The **goal** is to cluster houses with different features in different homogenous groups. This can help us to determine which cluster belongs a new house with its characteristics.

2. Data Collection

The dataset analyzed in this study has been uploaded under the name ”House Prices - Advanced Regression Techniques” on the Kaggle1 platform.

In this dataset, there are 1460 observations with 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa. Among explanatory variables, there are 37 integer variables, such as Id, MSSubClass, LotFrontage, and 43 factor variables, such as MSZoning, Street, LotShape. Descriptive analysis and quantitative analysis will be used. Target variable (SalePrice) has a continuous value.

3. Data Preprocessing

Data preprocessing is almost the same as preprocessing in the Supervised learning part. Therefore, this part skipped in order to avoid duplication.

4. Clustering[2]

Clustering is a widely used technique in exploratory data analysis that helps us gain insights into the underlying structure of the data. Its goal is to identify subgroups, or clusters, within the data where data points within the same cluster are highly similar, while those in different clusters are significantly different. Essentially, we aim to find homogeneous subgroups in the data, maximizing the similarity between data points using measures like Euclidean distance or correlation-based distance. The choice of similarity measure depends on the specific application.

Unlike supervised learning, clustering is categorized as an unsupervised learning method. This is because we do not have ground truth labels to evaluate the clustering algorithm's performance. Instead, our goal is to explore the data's structure by assigning data points to distinct subgroups or clusters.

[1] https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

[2] https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

4.1. K-Means Clustering Algorithm

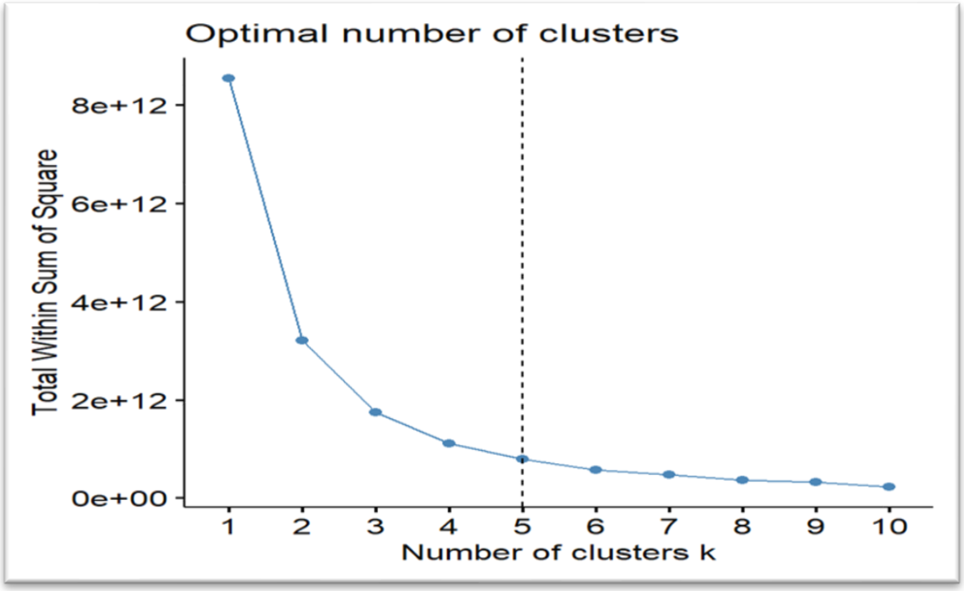
The K-means algorithm is an iterative algorithm designed to divide a dataset into K distinct and non-overlapping subgroups, or clusters, where each data point belongs to only one cluster. Its objective is to maximize the similarity among data points within the same cluster while ensuring that the clusters themselves are as dissimilar as possible. It achieves this by assigning data points to clusters in a way that minimizes the sum of squared distances between the data points and the centroid of each cluster. By reducing the variation within clusters, the algorithm aims to create more homogeneous groups of data points.

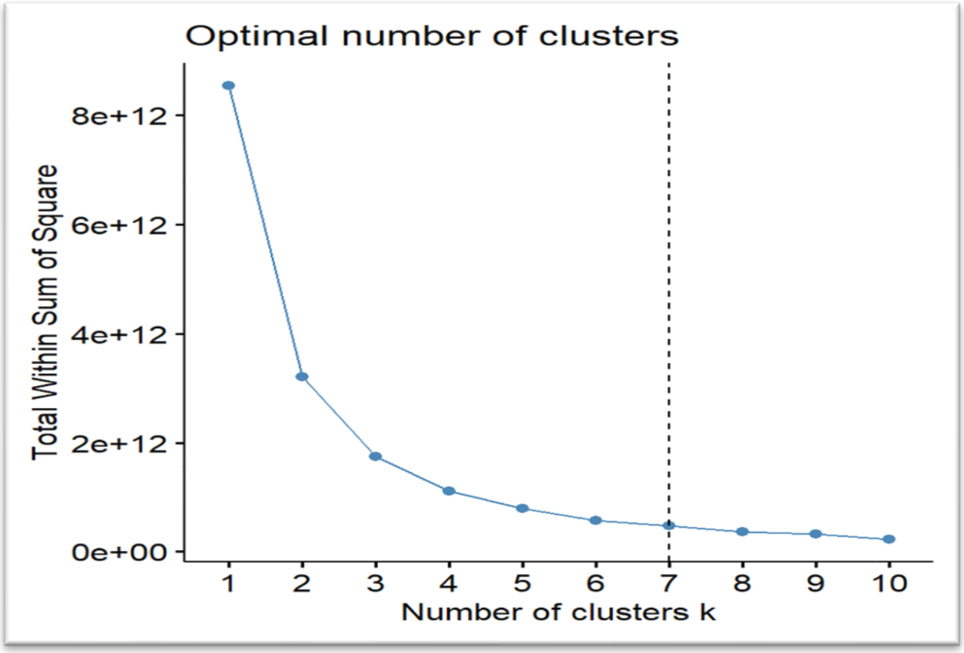
The steps involved in the K-means algorithm are as follows:

1. Specify the desired number of clusters, K.
2. Initialize the centroids by shuffling the dataset and randomly selecting K data points as centroids, without replacement.
3. Iterate until the centroids no longer change, indicating stable assignments of data points to clusters.
4. Compute the sum of squared distances between data points and all centroids.
5. Assign each data point to the cluster with the closest centroid.
6. Update the centroids of the clusters by calculating the average of all data points belonging to each cluster.

Result of K-Means Algorithm for Ames House Price Dataset:

4.1.1. Finding optimal number of clusters





Description:

In clustering analysis, the first step is known as "Finding optimal number of clusters" involves determining the most suitable number of clusters for a clustering algorithm. This step entails creating a plot that visualizes the relationship between the number of clusters and a specific evaluation metric.

The goal of finding the optimal number of clusters is to strike a balance between having enough clusters to capture the inherent structure of the data and avoiding an excessive number of clusters that could lead to overfitting or reduced interpretability.

To achieve this, a common approach[3] is to plot an evaluation metric, such as the within-cluster sum of squares (WSS), against different numbers of clusters. The aim is to create clusters in such a way that the overall variation within each cluster, also known as the within-cluster sum of squares (WSS), is minimized. This evaluation metric serves as an indicator of the clustering solution's quality.

The plot typically displays the evaluation metric on the y-axis and the number of clusters on the x-axis. By examining the plot, one can search for an "elbow point" or a point where the evaluation metric no longer shows significant improvement as the number of clusters increases. This elbow point or a notable change in the plot's slope suggests a promising choice for the optimal number of clusters. For this analysis the number of clusters based on the result of two above plots is

[3] https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

considered 5 cluster, since these plots show after cluster number 5, the evaluation metric does not show significant improvement.

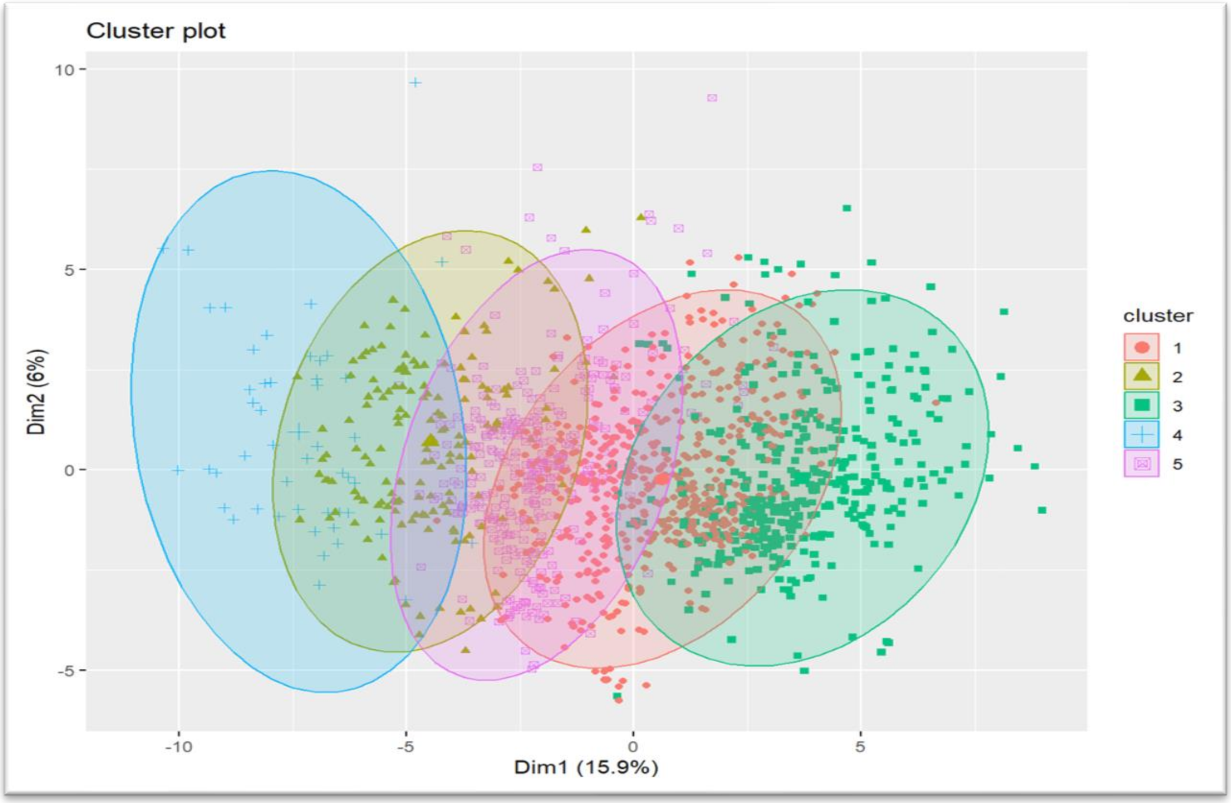
4.1.2. Plotting of K-means clustering

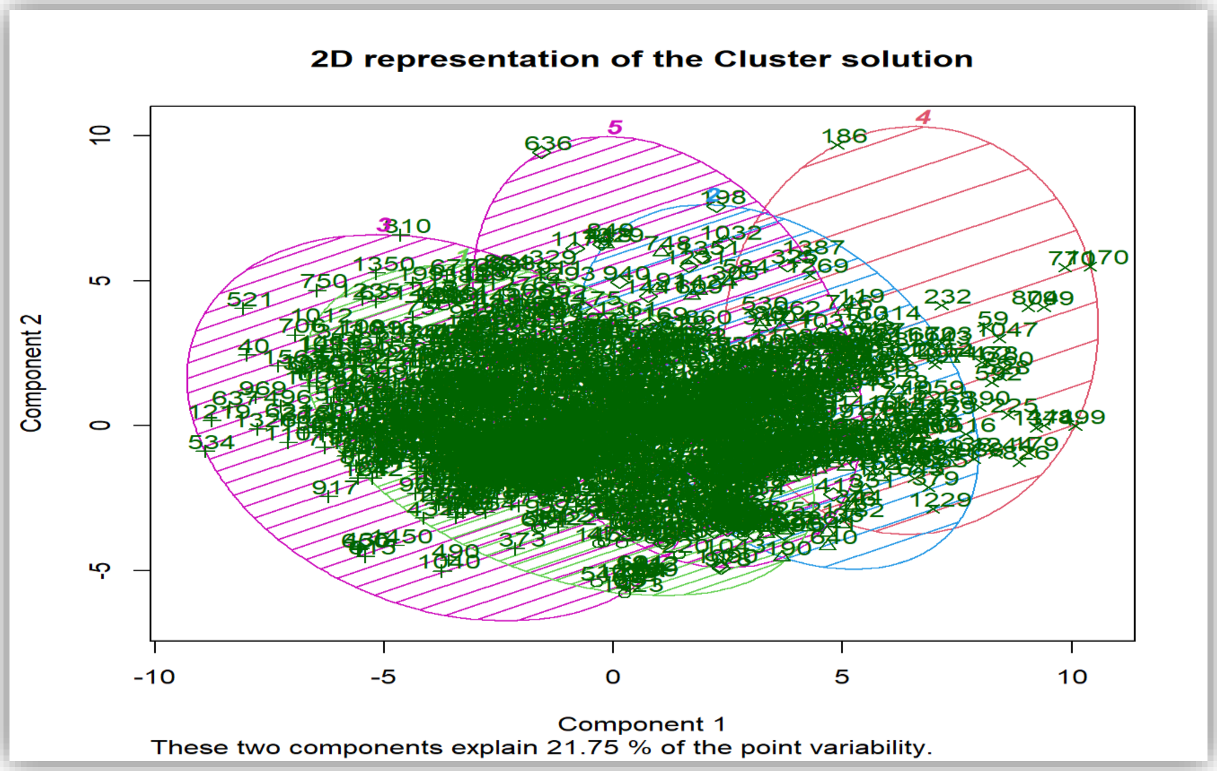
4.1.2.1. Main Plot

the "Main Plot" refers to the visualization of the clusters created by the K-means algorithm. It is the primary plot that displays the data points and their assigned clusters.

In the following Plots, we will obtain a visualization of the clusters in the Ames house Price dataset, where the data points are represented as **points** on the plot, and **ellipses** indicate the shape and boundaries of the clusters.

Plot 1: Main Plot





First Plot:

The values "dim1: 15.9%" and "dim2: 6%" represent the proportions of variance explained by each principal component on the respective axes of the plot.

When performing dimensionality reduction techniques, such as principal component analysis (PCA), the data is transformed into a lower-dimensional space. In this case, the data is projected onto a two-dimensional plot for visualization purposes. The two axes in the plot, "dim1" and "dim2," represent the first and second principal components, respectively.

The **percentages** associated with each dimension indicate the **amount of variance** in the data explained by that specific principal component. For example, "dim1: 15.9%" means that the first principal component explains 15.9% of the total variance in the dataset, while "dim2: 6%" indicates that the second principal component explains 6% of the total variance.

These values provide insights into the importance or contribution of each principal component in representing the data in the reduced dimensional space. Higher percentages suggest that the corresponding principal component **captures more information** and plays a more significant role in the clustering results depicted in the plot.

Second Plot:

the statement "These two components explain 21.75% of points variability" refers to the amount of variance in the data that is accounted for by the two components used for the plot. The statement indicates that the first two principal components, or the two dimensions of the plot, capture 21.75% of the total variability present in the dataset. This means that these two dimensions are able to explain and represent a significant portion of the variation in the data.

Summary of plots:

These two plots almost show the same result, however, the underlying data captured by these **two components** is limited and does not encompass a significant amount of information. This implies that these two dimensions do not capture a substantial portion of the data's variability or the distinct characteristics of the clusters. Therefore, additional dimensions or features might be required to obtain a more comprehensive understanding of the clusters and their underlying patterns.

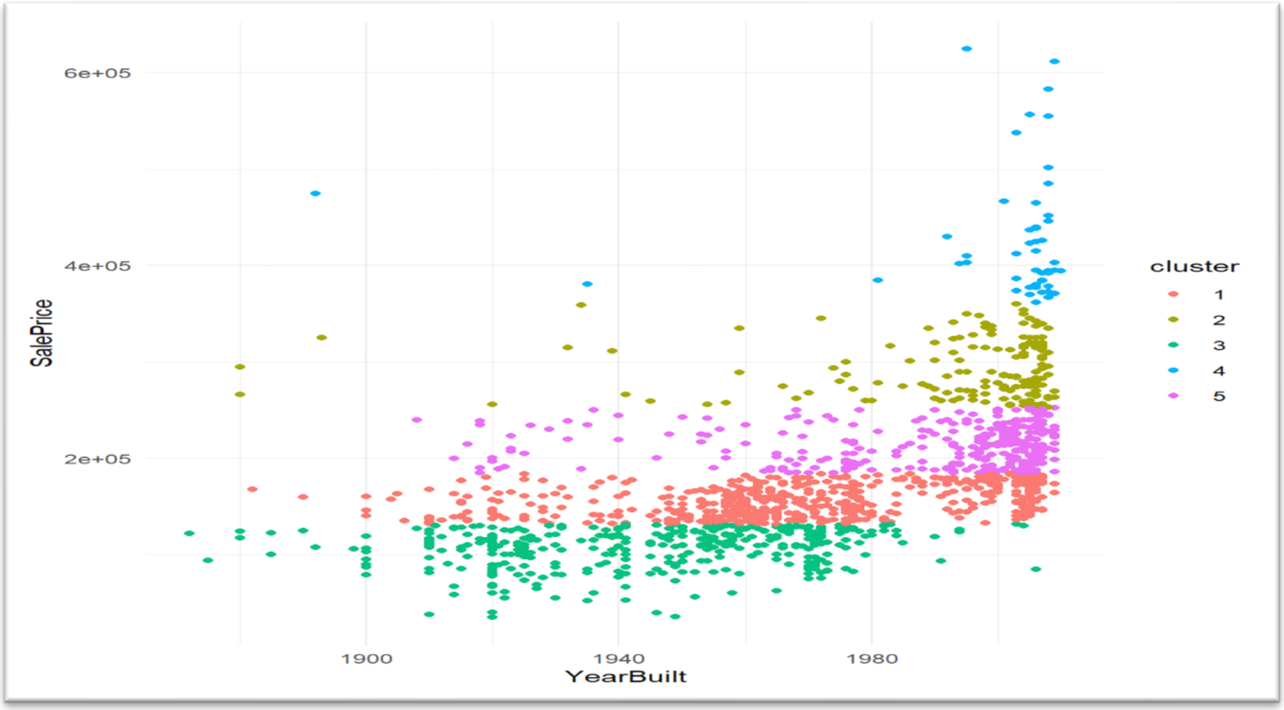
The number of data points assigned to each cluster are 536, 159, 384, 48, 324 respectively. Which shows cluster 4 (Blue) has minimum number of data points, while cluster number 1 (Red) contains maximum number of datapoints, 536.

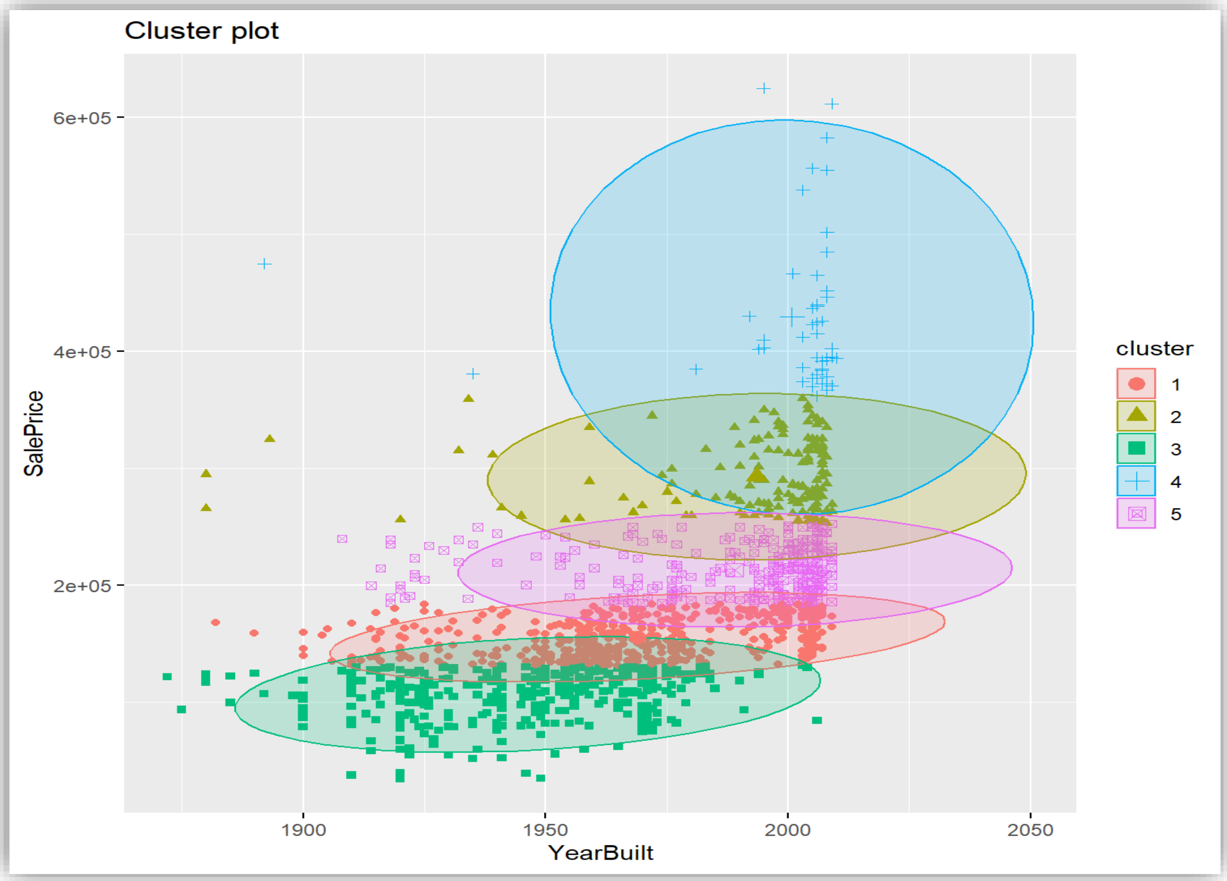
4.1.2.2. Visualizing the clusters based on some important features

By examining the **clusters**, we can analyze the distribution of data points across the clusters and gain insights into the grouping patterns. It allows us to identify which data points belong to each cluster, which is crucial for further analysis and interpretation.

1. **Ploting YearBuilt vs. SalePrice**

In this plot ‘YearBuilt’ is the original construction date of house.


6e+05
4e+05
2e+05
SalePrice
1900
1940
1980
YearBuilt
cluster
1
2
3
4
5


Cluster plot
6e+05
4e+05
2e+05
SalePrice
1900
1950
2000
2050
YearBuilt
cluster
1
2
3
4
5

Description:

The scatter plot of YearBuilt against SalePrice, color-coded by the cluster, reveals interesting insights into the relationship between the year of construction and the corresponding sale price of houses. The clusters are differentiated by color, with each point representing an individual house.

Observing the plot, we can identify distinct patterns within each cluster. Cluster 2 (dark yellow) and 4 (blue) predominantly consist of newer houses, built more recently. These houses tend to have higher sale prices, suggesting that newer constructions are generally associated with higher values.

Conversely, cluster 3 (green) and 1 (Red) primarily contain houses built before 1980, indicating older constructions. The majority of these houses show lower sale prices compared to other clusters. Cluster 3, in particular, consists mostly of houses built before 1980, contributing to its lower overall sale prices.

However, within cluster 4, we observe a few houses built before 1980 that have notably high sale prices. This anomaly can potentially be explained by other influential factors such as the number of bedrooms, lot area, or additional features that contribute to their higher values.

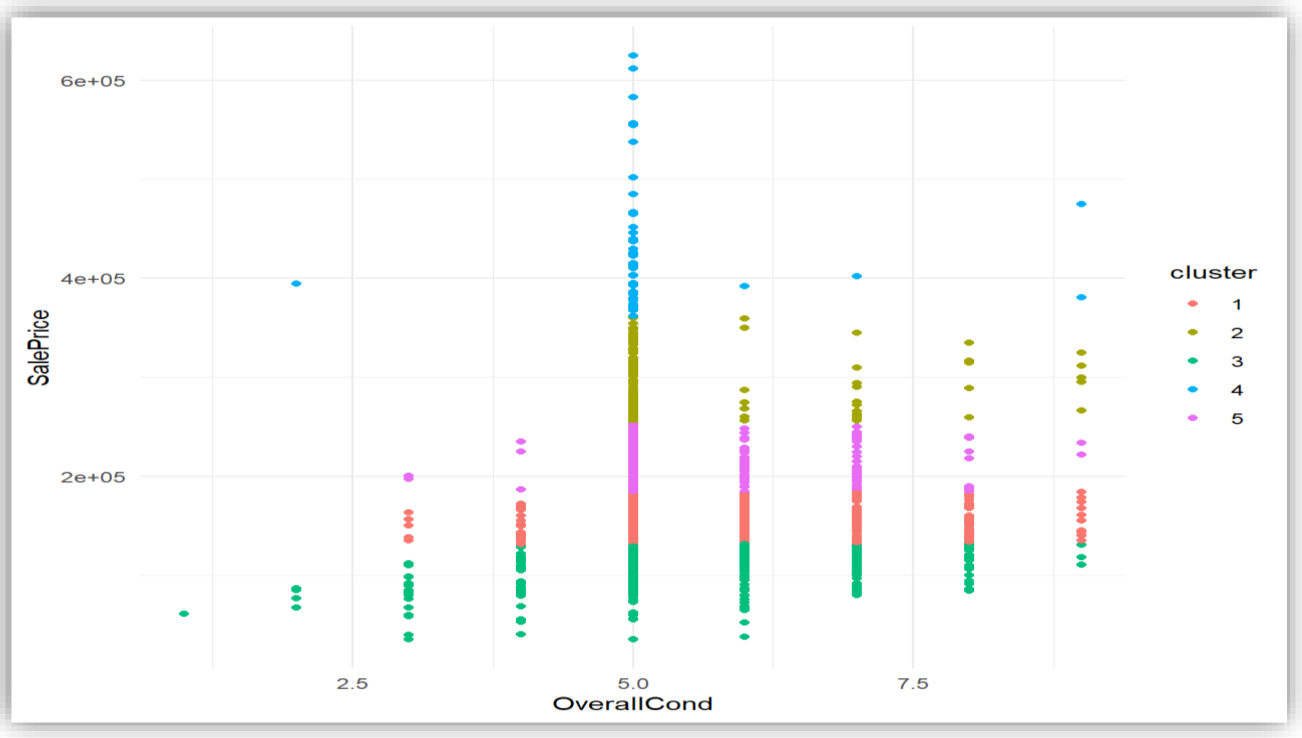
In the case of cluster 5 (Pink), characterized by intermediate sale prices, approximately 80% of the observations correspond to houses built after 1990. It suggests that newer constructions in this cluster tend to have moderate sale prices, potentially influenced by factors such as the number of bedrooms, garage area, size of the houses, and other relevant parameters.

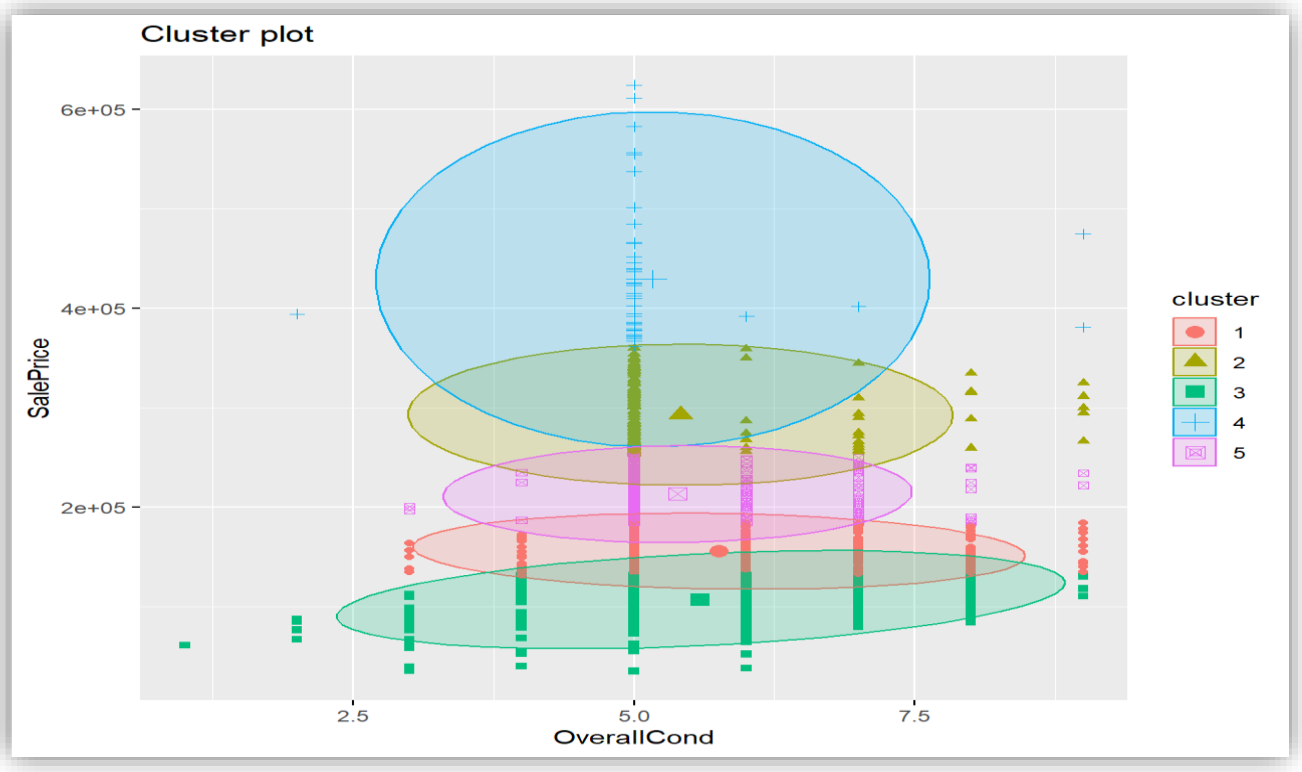
Overall, this plot provides valuable insights into how the year of construction (**YearBuilt**) relates to the **sale prices** of houses within different clusters. It highlights the influence of construction year on the pricing dynamics, with newer houses generally commanding higher values and older constructions often associated with lower sale prices.

2. Plotting 'OverallCond' and 'SalePrice'

In this plot 'OverallCond' is the Rate of **overall condition** of the house, in the following range:

10	Very Excellent	9	Excellent	8	Very Good
7	Good	6	Above Average	5	Average
4	Below Average	3	Fair		
2	Poor	1	Very Poor		


6e+05
4e+05
2e+05
SalePrice
2.5
5.0
7.5
OverallCond
cluster
1
2
3
4
5


Cluster plot
6e+05
4e+05
2e+05
SalePrice
2.5
5.0
7.5
OverallCond
cluster
1
2
3
4
5

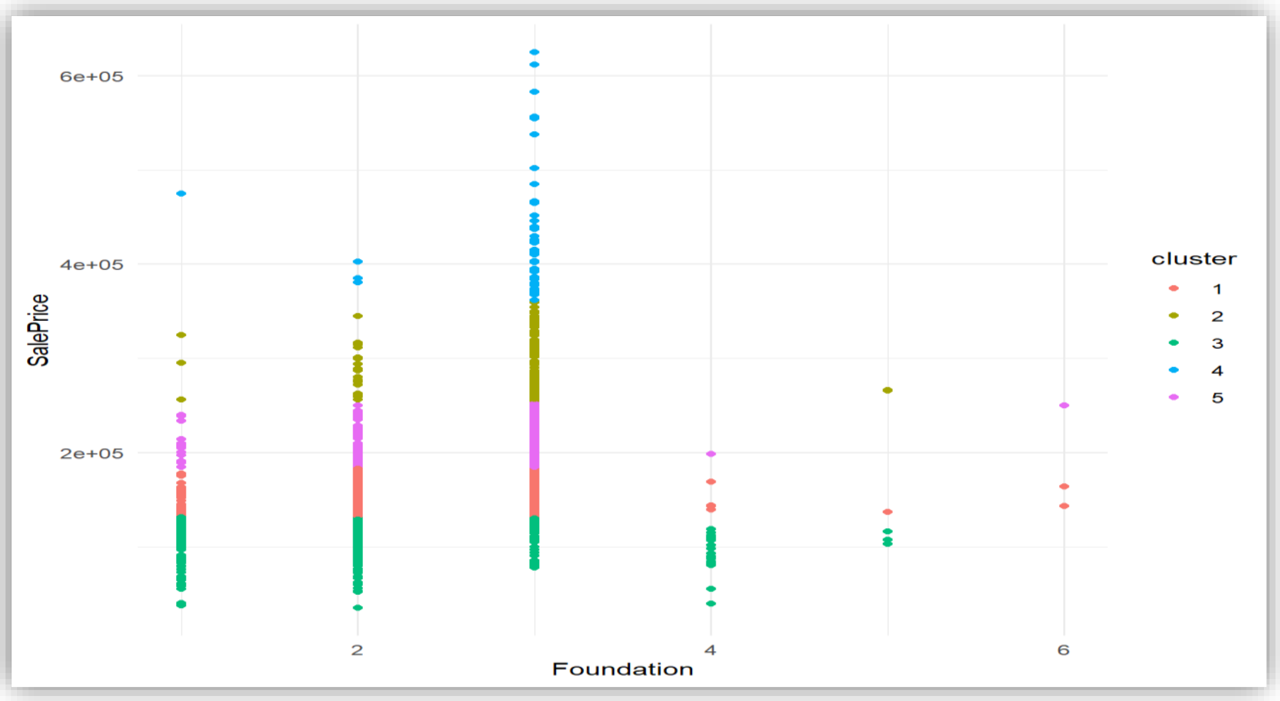
Description of plots: Almost all house that have high amount of SalePrice, they have good Overall condition (more than equal 5) too. But for cluster 3 (Blue) and 1 (Red), which both have Low amount of SalePrice, in every condition level exist some considerable number of observations. But still majority of them have good Overall condition (more than equal 5). However, in cluster number 1, 3, 5 exist some houses with high level of Overall condition but their Saleprice is low because of other features like neighborhood or house size.

In the case of cluster number 2 (Yellow), with middle SalePrice, there are not exist any observation with low Overall condition (less than 5), this shows that if the Overall Condition of house is good, customer can not expect to buy these kinds of houses with low Price, unless their other important features like location is not good.

3. Plotting 'Foundation' and 'SalePrice'

In this plot 'Foundation' is the Type of foundation of the house, consider the following types with their encoded numbers:

BrkTil	Brick & Tile	(1)
CBlock	Cinder Block	(2)
PConc	Poured Contrete	(3)
Slab	Slab	(4)
Stone	Stone	(5)
Wood	Wood	(6)

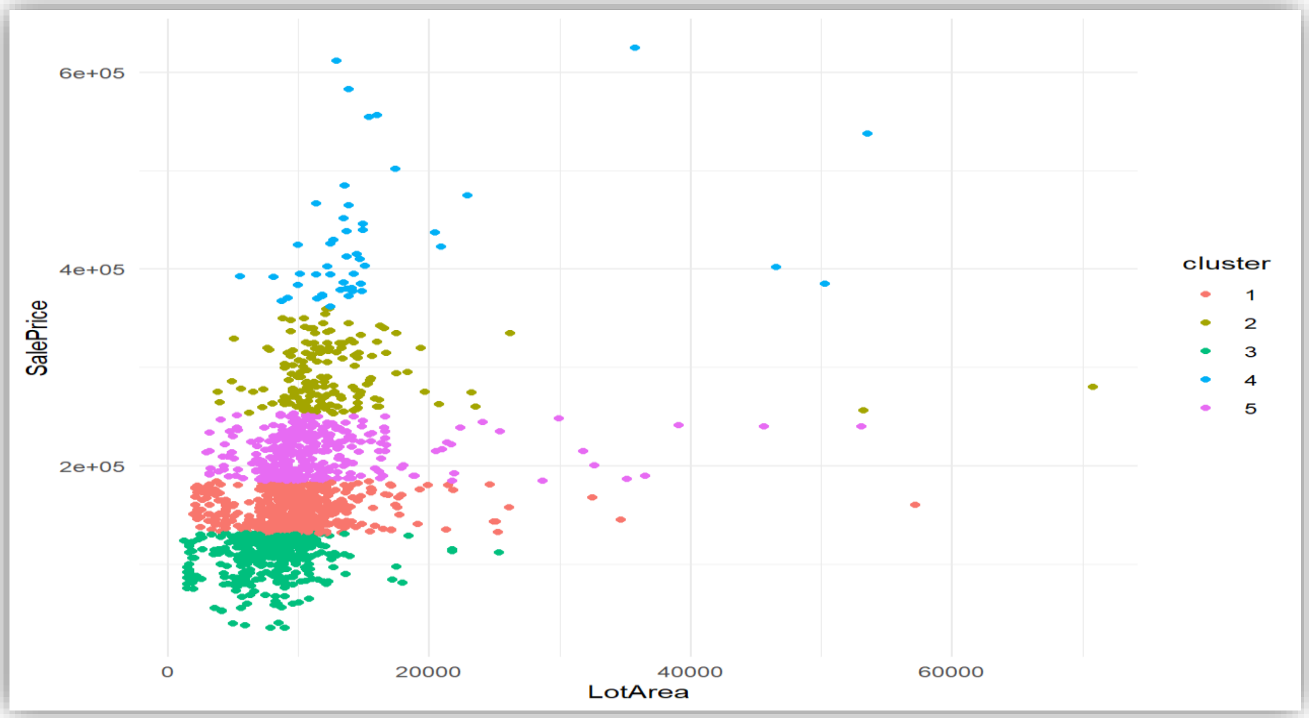


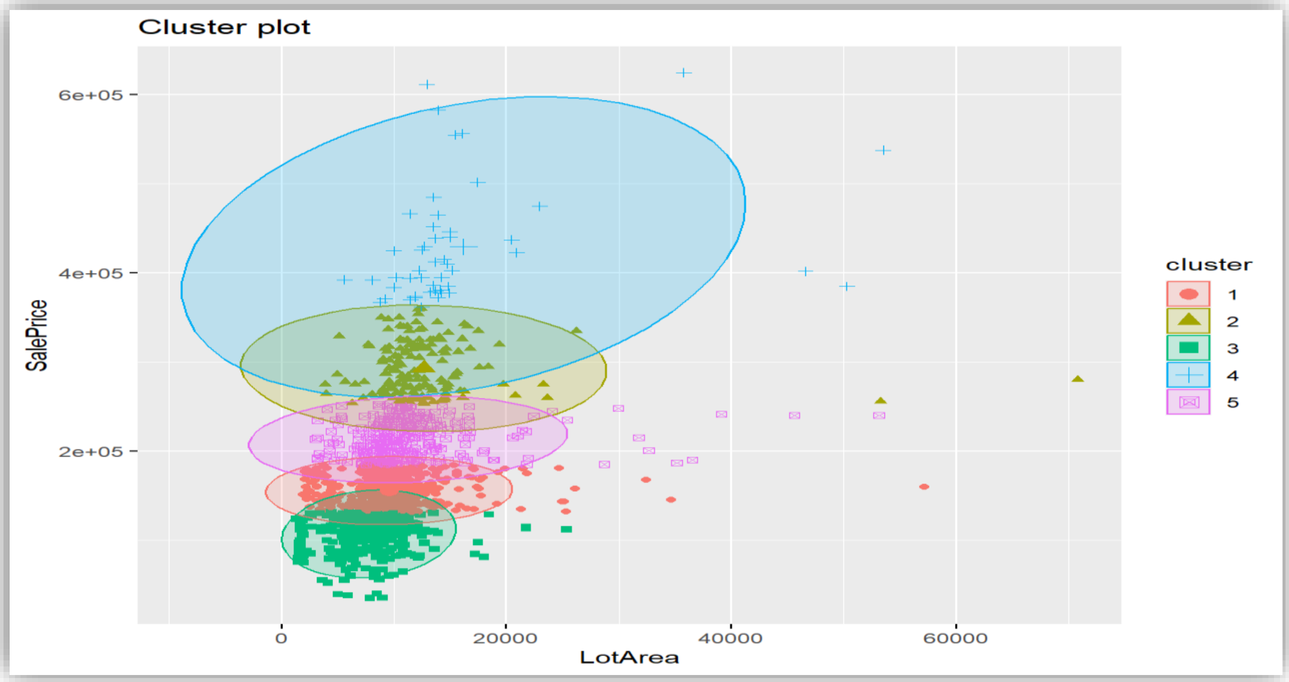
Description of Plot:

Almost all houses with high amount of SalePrice belong to cluster 4 (Blue) that their foundation type are mostly **Poured Contrete**, even houses inside the cluster 2 (dark yellow), with middle price, most of their foundation is Poured Contrete. While houses that are built by Stone, Wood, and Slab almost all are belong to cluster number 3 and 1 which their Saleprice is low.

4. Plotting 'LotArea' and 'SalePrice'

'**LotArea'** is Lot size in square feet.





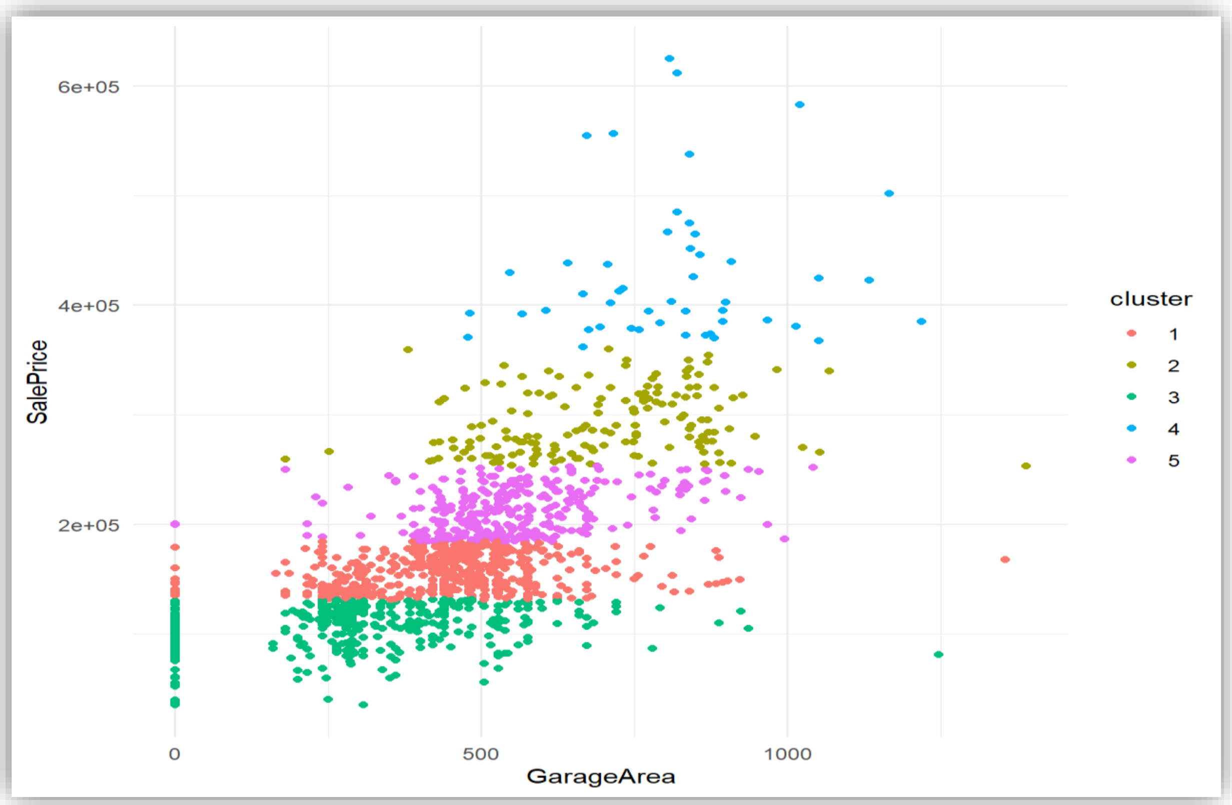
Description of plots:

As second plot shows, the radius size of cluster is so big for cluster number 4 while for cluster number 3 (Green) it is small. This shows that data are spread with high variation inside cluster 4 (blue), it means that data points have high distance from each other.

Clusters show that as the **lot area** increases, the SalePrice increases a bit, but does not affect the SalePrice Considerably.

5. Plotting ‘GarageArea’ and ‘SalePrice’

GarageArea is Size of garage in square feet.



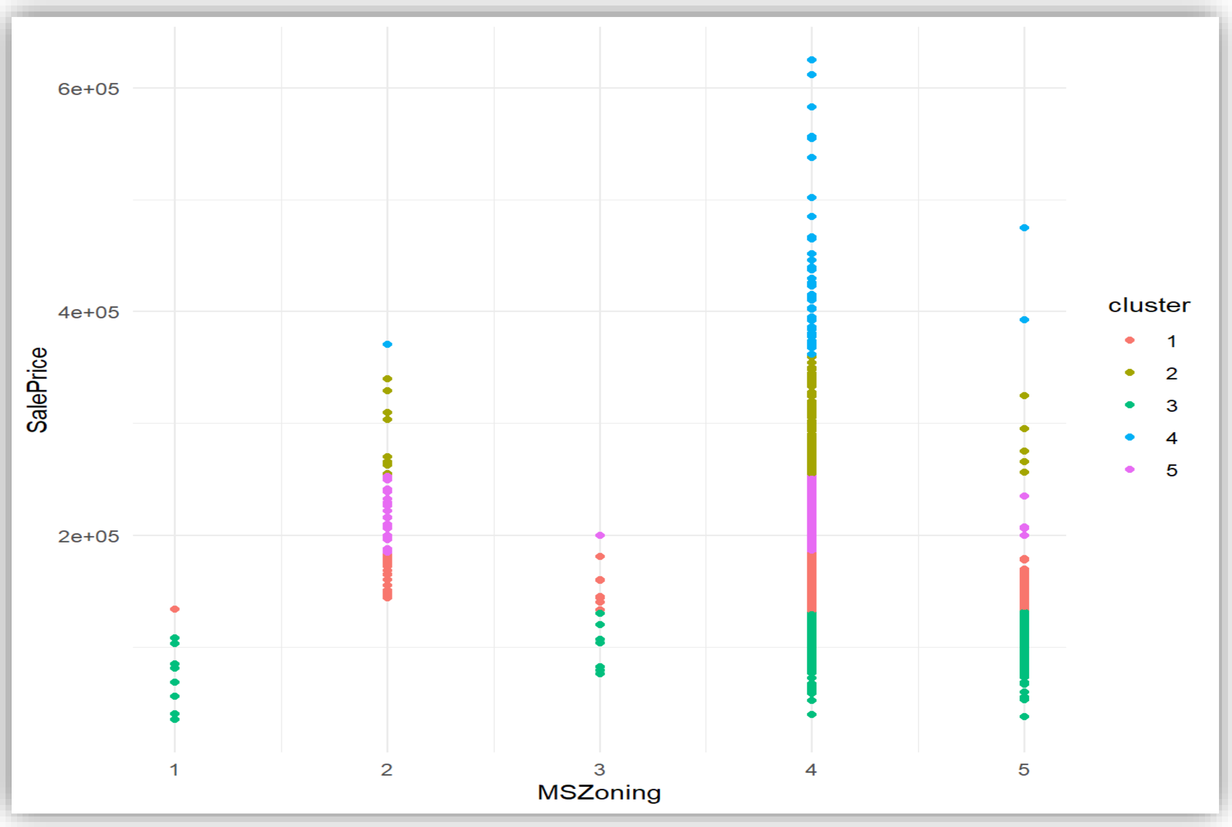
Description of plot:

This plot Shows that, as Garage Area increase, the SalePrice of houses tend to increase. It is considerable to mention that all of the houses without garage (zero Garage Area) belong to the Cluster 1 and 3 which they have low amount of SalePrice. The Cluster 4 (blue) which has high amount of price, all contain big GarageArea, more than 500 square feet. This trend is almost same for cluster 2 (dark yellow).

6. Plotting ‘MSZoning’ and ‘SalePrice’

MSZoning: Identifies the general **Zoning** classification of the sale.

C : Commercial
FV : Floating Village Residential
RH : Residential High Density
RL : Residential Low Density
RM : Residential Medium Density

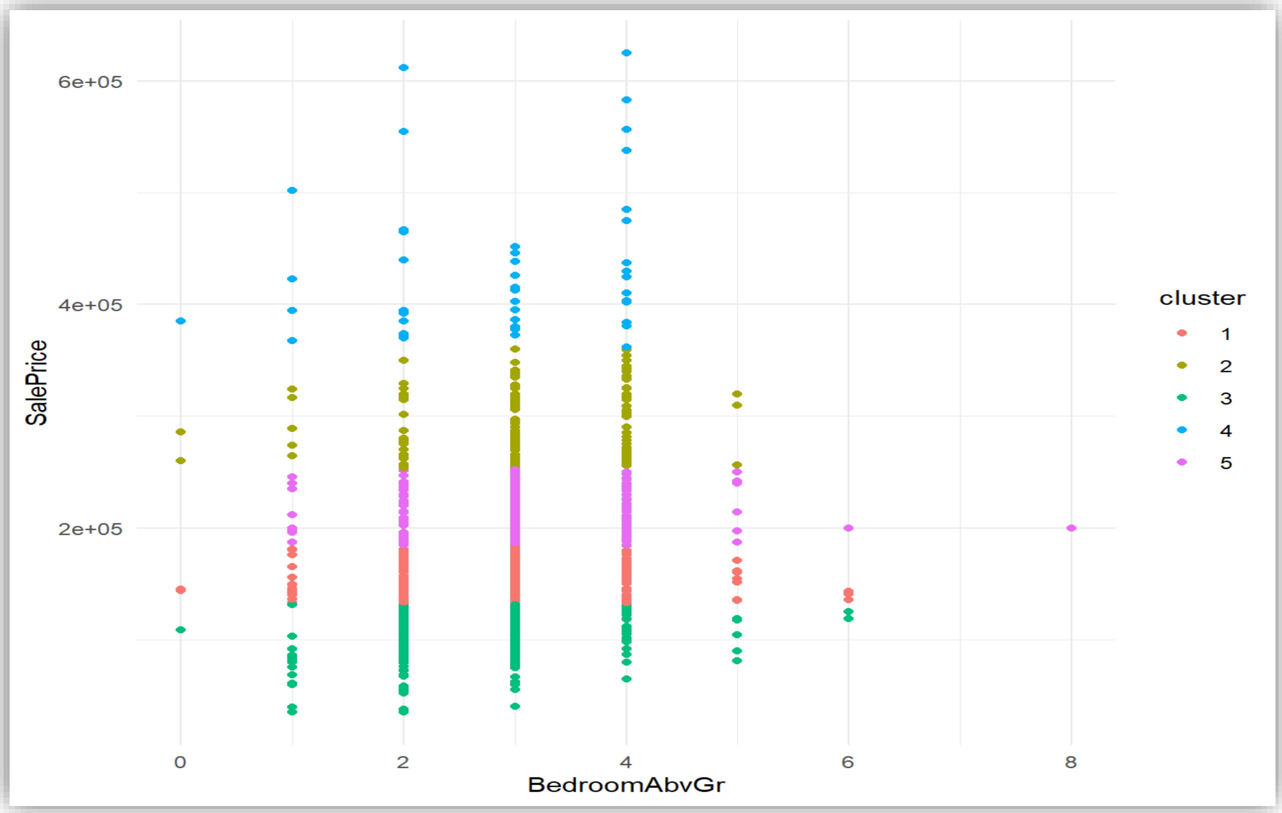


Description of plot:

Almost all houses with high amount of SalePrice belong to cluster 4 (Blue) which their location are in residential area with Low Density. Even most of the houses are inside cluster 2 (dark yellow), with middle price, most of them located in Low Density area. All of the houses in Residential area with High Density belong to cluster 3 (green) and 1 (red) which they have low amount of Saleprice. It is interesting all hoses except one in the commercial area are belong to cluster 3 (green), which has so low Saleprice.

7. Plotting 'BedroomAbvG' and 'SalePrice'

BedroomAbvGr is Bedrooms above grade (does NOT include basement bedrooms), with a range between 0 to 8.

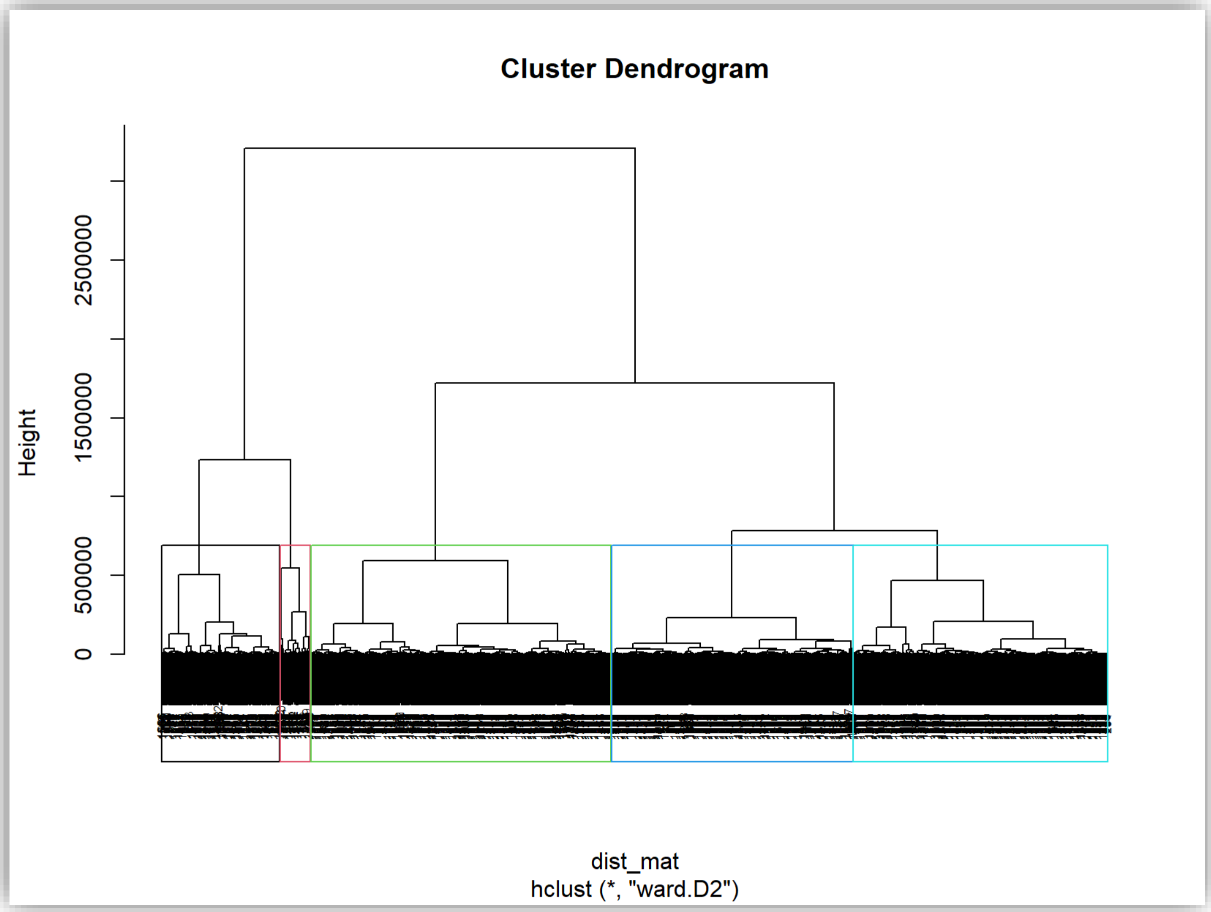


Description of plot:

Since this plot shows that for houses with 2 to 4 bedrooms, all clusters have enough number of houses in different SalePrice levels. On the other hand, exist some houses with 5 and 6 bedrooms, but they are belonging to clusters with low Saleprice. As a result, the number of the bedrooms does not affect the Saleprice significantly, so other features are mostly affecting the house price.

5. Hierarchical Clustering

The following plot is the Dendrogram of hierarchical clustering for this analysis with 5 clusters.



Summary:

The analysis of the Ames house price dataset using k-means clustering revealed interesting patterns and insights into the factors influencing house prices. The plots and descriptions showed that newer houses generally have higher sale prices, while older constructions tend to have lower prices. However, anomalies were observed, with some older houses in certain clusters having unexpectedly high prices, potentially due to other influential features. The analysis also highlighted the importance of overall condition, with houses in good condition commanding higher prices. Additionally, factors such as foundation type, zoning, density, and the number of bedrooms were found to have varying effects on sale prices across different clusters.

Overall, this analysis provides valuable information for understanding the dynamics of the housing market in Ames. The findings can assist buyers, sellers, and real estate professionals in making informed decisions by considering the influential factors identified through the clustering analysis. By understanding how features such as construction year, overall condition, foundation type, zoning, density, and the number of bedrooms affect house prices, stakeholders can gain insights into market trends and make more accurate pricing assessments. These insights contribute to a comprehensive understanding of the Ames housing market and can aid in maximizing returns and making well-informed real estate decisions.

Reference:

https://posit.co/download/rstudio-desktop/

https://www.r-project.org/other-docs.html

https://ggplot2.tidyverse.org/

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/overview

https://towardsdatascience.com/k-means-clustering-algorithm-applications-evaluation-methods-and-drawbacks-aa03e644b48a

https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/