Summary of Clustering Models

Project of unsupervised learning

Statistical Learning Course Projects

By Wang Jian(Steven)

Course:Data science for economics (classe lm-data) (code: B79, class: LM-DATA - Data science for economics) Year 2

Matriculation number:14326A

Contents

Abstract

This report outlines a project on clustering models. The focus is on applying unsupervised learning techniques to a dataset from the Melbourne housing market[1], sourced from Kaggle. By clustering properties into distinct groups, the project aims to uncover underlying patterns and characteristics within the data, providing insights into the housing market dynamics.

Statement of the Problem/Goal and Dataset Description

The goal of this analysis is to apply clustering algorithms to the Melbourne housing market dataset to identify inherent groupings within the properties. The dataset includes features like Suburb, Type, and Price, among others. The problem involves understanding how these properties cluster based on various characteristics and determining the insights that can be drawn from these clusters.

Findings/Key Points

Distance-based clustering[2] (figure1, figure 3) yields more distinct and better-separated clusters compared to original latitude-longitude-based[3] (figure2) clustering, as evidenced by higher average silhouette scores.

The within-cluster sum of squares (WSS) indicates more compact clusters in the coordinates clustering than in the distance-based clustering.

The analysis reveals that considering properties' proximity to a central point, in combination with other features, results in more meaningful clustering than using geographical coordinates alone.

[1] https://www.kaggle.com/datasets/anthonypino/melbourne-housing-market

[2] Clustering variables: "DistanceFromCenter", "PropertyAge", "PricePerArea"

[3] Clustering variables: "Latitude", "Longitude", "PropertyAge", "PricePerArea"

Summary of Clustering Models_Wang Jian_ Matriculation number:14326A

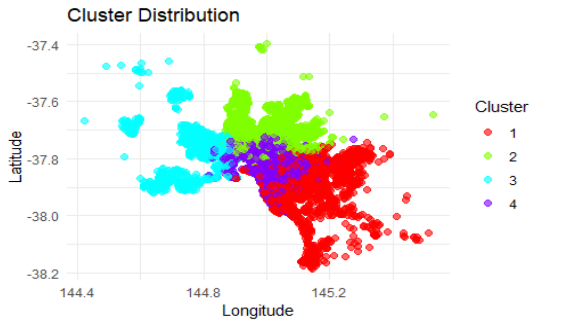

Figure 1 Clustering based on location+other variables

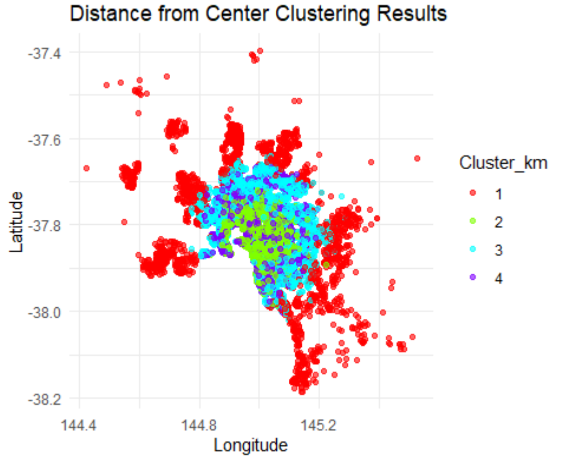

Figure 2 Clustering based on distance+other variables displayed in coordinates

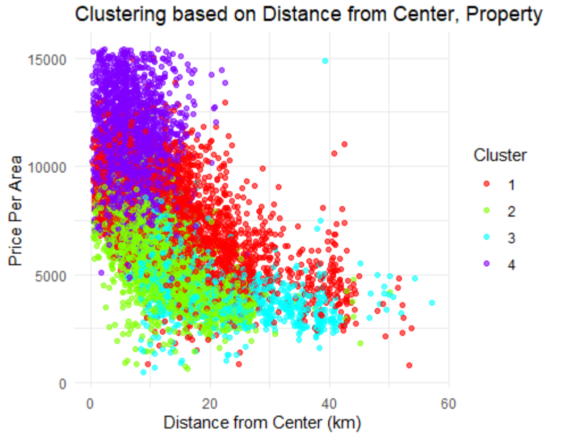

Figure 3Clustering on distance + additional variables

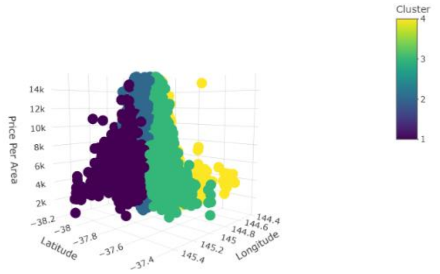

Figure 4 Clustering based on location+other variables displayed in coordinates+priceperarea

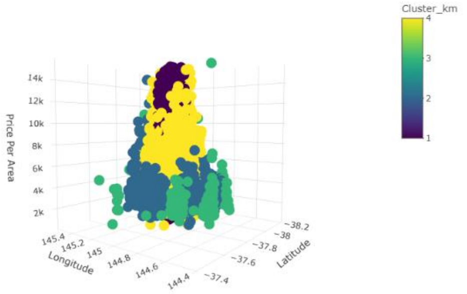

Figure 5 Clustering based on distance+other variables displayed in coordinates+priceperarea

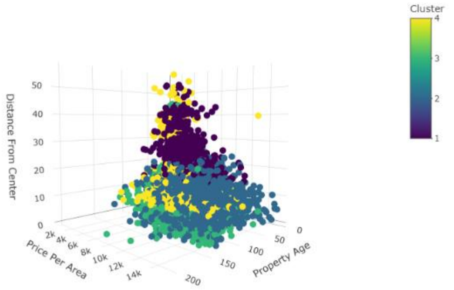

Figure 6 Clustering based on location+other variables displayed in distancefromcenter+propertyage+priceperarea

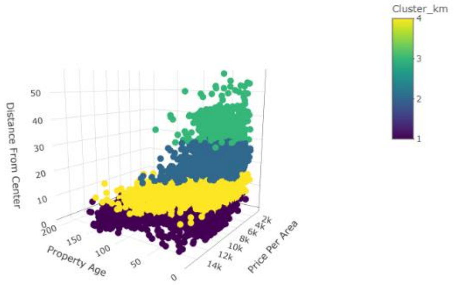

Figure 7 Clustering based on distancefromcenter +other variables displayed in distancefromcenter+propertyage+priceperarea

Analysis and Commentary

The analysis began with data cleaning and preparation, utilizing R libraries such as readr and dplyr. Critical steps included outlier elimination and feature engineering, like calculating PropertyAge and PricePerArea. Clustering was performed using the k-means algorithm, and results were visualized geographically and evaluated using metrics like average silhouette score and WSS[4].

If the primary goal is to identify well-separated groups, the distance-based clustering seems superior as indicated by the silhouette scores. However, if the focus were on creating tightly packed groups, the original method might be seen as more effective based on WSS alone. The choice between the two methods should therefore consider the specific objectives of the clustering analysis and the inherent characteristics of the dataset.

Conclusions

The project demonstrates that distance-based[5] clustering provides clearer separation between property clusters in the Melbourne housing market, as indicated by the silhouette scores. However, the choice between geographic[6] and distance-based clustering should be guided by the specific goals of the analysis. The findings underscore the importance of considering the nature of the data and the analysis objectives when choosing clustering approaches.

In the business context, either the buyer or the seller could utilize such results to further analyze the property groups in each cluster. After finding distinctive groups, further supervised learning methods could produce informative information to decide proper price. Or, when comparing several potential purchase choices, check if they are in the same cluster, to compare their different characteristics.

[4] **Original Clustering (based on Latitude and Longitude):** The average silhouette score is 0.3427781, indicating a moderate fit. This suggests that while there is some separation between the clusters, there is still room for improvement.
Distance-based Clustering: The average silhouette score is significantly higher at 0.5552604, indicating a much better fit than the original clustering. This higher score suggests that the clusters are more distinct and better separated.

[5] With other selected variables

[6] With other selected variables

Appendix

The appendix would contain brief processing steps and all R code used in the project, detailing each step of the data cleaning, preparation, clustering, and analysis process. This includes code for data manipulation, outlier detection, clustering, and result visualization.

Appendix 1, Process of clustering

Brief go-through of the steps

Step 1: Data Cleaning and Preparation

Utilize R libraries readr and dplyr for data manipulation.

Load the dataset, correct column name typos, and address parsing issues.

Columns such as Suburb, Type, and Price are included in the dataset, among others.

Step 2: Clean Data

Remove rows with missing values in critical columns (e.g., Latitude, Longitude).

Calculate new variables such as PropertyAge and PricePerArea, and remove any rows with NA or infinite values after these calculations.

Step 3: Show Data Distributions

Visualize distributions of Latitude, Longitude, Property Age, and Price Per Area using histograms to understand data distribution and variability.

Step 4: Implement Outlier Elimination Strategy

Use the Interquartile Range (IQR) method to filter out outliers from the dataset, specifically focusing on the PricePerArea and potentially other variables.

Step 5: Further Steps (Clustering, Visualization, Comparison)

Scale the data and use the elbow method to determine the optimal number of clusters, followed by k-means clustering.

Assign cluster labels to the data based on k-means results.

Step 6: Show Result in Geo Map

Visualize the geographic distribution of the clusters using ggplot2, with points colored by cluster assignment.

Step 7: Calculate the Central Point

Determine the central point (mean latitude and longitude) of the dataset.

Step 8: Calculate the Distance from the Center for Each Point

Use the Haversine formula to calculate the distance from the center for each data point.

Step 9: Cluster Using the New Feature

Perform k-means clustering again, this time using the newly calculated distance from the center along with other features.

Step 10: Visualization and Analysis

Visualize the new clustering results, analyzing them based on distance from the center, property age, and price per area.

Step 12: Visualizing Both Clustering Results

Compare the original clustering (based on latitude and longitude) with the new clustering (based on distance from the center) through visualizations.

Step 13: Calculate Metrics for Both Clustering Approaches

Calculate and compare metrics such as the average silhouette score and withinss (within-cluster sum of squares) for both clustering approaches to evaluate their effectiveness.

Step 13: Additional visualization 3D

Visualize clusters under different Axis for both cluster strategies.

Throughout these steps, after clean, process, and cluster the data, certain meaningful patterns and relationships are identified.

Evaluation and comments about the models:

The clustering project involved various steps, from data preparation to implementing different clustering strategies, and concluded with a comparison of these strategies based on specific metrics. Below is a summary of the metrics used for comparison between the original spatial feature clustering and the distance-based clustering:

Average Silhouette Score:

Original Clustering (based on Latitude and Longitude): The average silhouette score is 0.3427781, indicating a moderate fit. This suggests that while there is some separation between the clusters, there is still room for improvement.

Distance-based Clustering: The average silhouette score is significantly higher at 0.5552604, indicating a much better fit than the original clustering. This higher score suggests that the clusters are more distinct and better separated.

Within-Cluster Sum of Square (WSS):

Original Clustering: The WSS is 83.12818, which suggests that the clusters are relatively compact and well-separated, as lower WSS values are generally indicative of tighter clusters.

Distance-based Clustering: The WSS is much higher at 67887.45, which typically would suggest less compact clusters. However, this large number might also reflect the scale of the data used for clustering, particularly because this approach involves distances that could inherently have larger values and variances than those seen in typical geographic coordinates.

Comparison and Interpretation:

The silhouette scores suggest that the distance-based clustering provides a clearer delineation between clusters than the original geographic clustering. This could mean that when considering the dataset's spatial distribution, grouping properties based on their proximity to a central point (along with other features) results in more meaningful clusters than simply using their latitude and longitude.

The WSS values present a more complex picture. While the other clustering yields a low WSS, indicating tight clusters, the distance-based clustering has a much higher WSS, which typically would suggest less compact clusters. However, considering the different nature of the clustering bases (geographical coordinates vs. distances), this increase in WSS may not directly indicate worse performance but rather reflects the different scales and distributions inherent to the data used.

From a metrics point of view, if the primary goal is to identify well-separated groups, the distance-based clustering seems superior as indicated by the silhouette scores. However, if the focus were on creating tightly packed groups, the original method might be seen as more effective based on WSS alone. The choice between the two methods should therefore consider the specific objectives of the clustering analysis and the inherent characteristics of the dataset.

Appendix 2, R code as attached