Module 3: Clustering

9/24/2025

40/40 Points

Attempt 1

Review Feedback
9/22/2025

Attempt 1 Score:
40/40

View Feedback

Anonymous Grading: **no**

Unlimited Attempts Allowed

Details

Goal of the assignment:

In this assignment, you will analyze the clustering results of the housing dataset we looked at in class. Your goal is to **generate and plot distributions of key features in the different clusters and interpret the characteristics of each cluster,** including the `ocean_proximity` feature that can be found in the original housing dataset.

Steps:

1. Create a new Colab notebook. Name it **your-name-clustering** (replacing your-name with *your actual name*, of course!!) Using the code in the `03-Clustering.ipynb` notebook, perform k-Means clustering (with k=4) on the `housing_minus_coords` dataset, which is the z-score standardized housing dataset without the geographical coordinates. You can simply reuse the code in the notebook for this.
2. Now that you have the cluster labels (stored in the kmeans.labels_ object), load the original `housing_data` as done in the first cell of the `02-Toolbox-Part3.ipynb` notebook.
3. Next, paste the clustering labels into the original housing_data dataframe (hint: I'm doing something similar when adding the cluster labels to the `long_lat_cl` object in `03-Clustering.ipynb` ...)
4. **For each of the four clusters,** create boxplots or histograms or kernel density plots (whatever you prefer) for the following features: housing_median_age, total_rooms, population, households, median_income, median_house_value, and a barplot (since it's a discrete and not a numeric feature) for the `ocean_proximity` feature.
5. By comparing the distribution of each feature in each cluster, see if you can find some way to interpret the clusters. For example, how do the features differ between clusters? What can you infer about each cluster based on the `ocean_proximity` and other features? (e.g., are certain clusters closer to the coast? Are they associated with higher or lower incomes? Do the clusters contain vacation properties?)
6. Write your interpretation of the clusters directly in the notebook as a text cell.

Deliverables:
A Colab notebook containing all the code and plots, and a cell with your interpretation of the clusters. Please upload that to Canvas. That's all you need to upload.

Use of LLMs:

You are allowed (and even encouraged) to ask Gemini (available directly in Colab) or other LLMs for help if you don't know how to do something. The only thing that is not directly provided in the notebooks I shared is how to plot distributions for each cluster, so that might be something you need to do a little bit of research on.

View Rubric

Select Grader

Dario Ghersi (Teacher)

Rubric for clustering assignment

Criteria	Ratings					Pts
Data Loading, Preprocessing, and K-Means Clustering	**15 pts** **Excellent** Correctly loads dataset. K-Means clustering is implemented using all appropriate features with correct parameters. Code runs without errors.	**10 pts** **Good** Minor issues with loading the data or clustering, but the approach is correct overall.	**7 pts** **Satisfactory** Loads data and applies K-Means but with substantial issues (e.g., missing features or incorrect parameters).	**4 pts** **Needs improvement** Dataset is not fully loaded or preprocessed. K-Means is applied incorrectly or not as required.	**0 pts** **Incomplete** No dataset loaded or K-Means clustering attempted.	12 / 15 pts
Feature Distributions for Clusters	[✓ **10 pts** **Excellent** All required plots are generated and correctly visualize feature distributions for each cluster. Clear, well-labeled plots.]	**7 pts** **Good** Plots are mostly correct, but there are minor labeling or clarity issues.	**5 pts** **Satisfactory** Visualizations are missing or unclear for some features.	**3 pts** **Needs improvement** Incomplete or incorrect plots, lacking labels or clarity.	**0 pts** **Incomplete** No plots provided.	10 / 10 pts
Interpretation of Clusters	[✓ **10 pts** **Excellent** Provides detailed, insightful interpretation of the clusters. Makes strong connections between ocean_proximity and other features with real-world implications.]	**7 pts** **Good** Good interpretation with mostly clear connections between features and clusters.	**5 pts** **Satisfactory** Satisfactory interpretation with some insights, but lacks depth or has unclear connections with features.	**3 pts** **Needs improvement** Limited or superficial interpretation, with unclear or incorrect conclusions about the clusters.	**0 pts** **Incomplete** No interpretation provided.	10 / 10 pts
Code Quality and Presentation	[✓ **5 pts** **Excellent** Code is clean, well-commented, and easy to follow. Notebook is well-organized.]	**3 pts** **Good** Code is mostly clean with minor issues. The notebook is organized but could be more clear.	**2 pts** **Satisfactory** Code is functional but messy or lacks adequate comments. Notebook organization is lacking.	**1 pts** **Needs improvement** Code has significant issues, lacks comments, and the notebook is difficult to follow.	**0 pts** **Incomplete** Code is disorganized or missing, and the notebook is hard to follow.	5 / 5 pts

Total Points: 37