# ISL Project Fall 24 **TASK-3**

# Data Description

**Dataset Overview:**  
The **Apartment for Rent Classified** dataset contains **10,000 samples** and **21 features**. The dataset primarily includes a mix of **categorical** and **integer** features. It is designed for tasks like classification, regression, and clustering.

**Types of Features**:

* + **Categorical**: Includes features like:
    - *Category*: Indicates the type of rental listing.
    - *Amenities*: Lists additional facilities available in the apartment.
    - *Pets Allowed*: Indicates whether pets are permitted.
    - *Photo Availability*: Specifies whether images are included in the listing.
  + **Numerical**: Includes:
    - *Price*: Monthly rental price.
    - *Size*: Square Feet of the apartment.
    - *Bedrooms* and *Bathrooms*: Count of rooms and restrooms in the unit.
  + **Geographical Data**: Latitude and longitude for each listing, providing spatial information.

**Data Characteristics:**

* **Purpose**: The dataset was created for applications in classification, clustering, and regression tasks. For example:
  + Predicting rental prices based on apartment features.
  + Grouping similar apartment listings for better categorization.
* **Completeness**: The dataset is well-structured but may include missing values in some fields.
* **Variability**: Features like *Price* and *Size* likely exhibit high variability, offering valuable information for clustering or regression models.

Data Preprocessing:

**Handling Null Values**

The dataset contained null values in critical columns like latitude, longitude, price, bathrooms, bedrooms, and Square Feet. Addressing these missing values was essential for maintaining data integrity and ensuring accurate clustering:

1. **Latitude and Longitude**: Since these fields are critical for geospatial clustering, listings with null values in these columns were removed entirely. Retaining only complete rows ensured valid distance calculations.
2. **Bathrooms, Bedrooms, and Square Feet**: These attributes are fundamental for describing the properties. Missing values in these fields were imputed using the **mode** (most frequently occurring value). This approach was selected because mode preserves the categorical integrity of numerical features like room counts.
3. **Price**: Given its importance for clustering and variability, null values were filled with the **mean** of the column. The mean provides a reasonable central value that maintains the data's overall balance.

**Categorical Data Encoding**

The dataset contained categorical features such as amenities, pets allowed, and category type. These were converted into a numerical format using one-hot encoding, which transformed each unique category into a binary column. This approach ensured compatibility with distance-based clustering algorithms while preserving feature uniqueness.

**Scaling**

Numerical features such as price, Square Feet, latitude, and longitude were scaled to ensure uniform contribution to the distance metric. Standardization was applied to bring all features to a comparable scale, with a mean of 0 and a standard deviation of 1. This step was necessary because hierarchical clustering relies on distance calculations, and features with larger ranges could dominate others without scaling.

Exploratory data analysis:

**1. Visualization Techniques:**

* **Histograms:** Created to examine the price distribution and grouped by location to observe regional price trends.
* **Scatter Plots:** Plotted to analyze relationships between price and square footage, bedrooms, and bathrooms. 3D scatter plots were used to explore multidimensional relationships.
* **Box Plots:** Used to visualize the spread and central tendency of prices by state, while identifying potential outliers.
* **Correlation Heatmap:** Employed to quantify linear relationships among key numerical features like price, square footage, bedrooms, and bathrooms.
* **Additional Plots:** Explored specific relationships, such as price vs. bedrooms and bedrooms vs. bathrooms, using color coding for categorical variables like payment type.\

**2. Data Cleaning**

* Detected and removed an outlier in the square footage variable to improve the clarity and accuracy of visualizations.
* Also detected and removed some other outliers in bedroom and bathroom variables to improve data modeling.

**3. Findings**

**3.1 Price Distribution**

* Observed a heavily right-skewed price distribution, indicating a concentration of lower-priced apartments with fewer high-priced options. This pattern was evident in the histogram.

**3.2 Price vs. Size**

* A positive correlation exists between apartment price and size (square footage). The scatter plot highlighted this trend, with some high-square-footage outliers.

**3.3 Location Impact**

* Significant variation in apartment prices by state was observed. Some states consistently exhibited higher prices, as shown by scatter plots and grouped histograms.

**3.4 Bedrooms and Bathrooms**

* Both the number of bedrooms and bathrooms are positively correlated with price. A relationship between bedrooms and bathrooms was also evident in scatter plots with color coding by payment type.

**3.5 Outliers**

* One prominent outlier in the square footage data was removed, enhancing the clarity of subsequent analyses.

**4. Patterns Identified**

1. **Skewed Price Distribution:** Indicates affordability for most apartments with few luxury options.
2. **Feature Correlations:** Price increases with size, number of bedrooms, and bathrooms.
3. **Regional Price Differences:** States exhibit distinct pricing trends, with some being significantly costlier.
4. **Inter-feature Relationships:** A correlation exists between bedrooms and bathrooms, as well as their combined impact on price.

**5. Conclusion**

The analysis highlighted clear relationships and trends in apartment rental data, providing valuable insights into price drivers. These findings can guide further analyses or model development for predicting apartment prices.

Model Development: Hyperparameters and Selection:

**1. K-Means Clustering**

* **Hyperparameters Selected:**
  + **Number of Clusters (n\_clusters)**:
    - Initially set to **65** for clustering the original data, chosen to explore detailed segmentation of the dataset based on apartment features.
    - Reduced to **10** for clustering after Principal Component Analysis (PCA), based on the **Bend or Elbow pattern** observed in the explained variance by principal component graph, which indicated that the first 10 components captured the majority of the variance in the data.
  + **Maximum Iterations (max\_iter)**:
    - Set to **50000** for the original dataset and **500000** after PCA to ensure convergence for potentially large and complex datasets.
* **Reasoning for Selection:**
  + A higher number of clusters (65) was initially selected to capture finer distinctions in the dataset. After PCA, the dimensionality reduction and the elbow pattern justified a lower number of clusters (10) for better visualization and practical interpretability.
  + A high iteration count was chosen to guarantee convergence, especially given the potential for slower convergence with a large number of clusters or high-dimensional data.

**2. Principal Component Analysis (PCA)**

* **Hyperparameters Selected:**
  + **Number of Components (n\_components)**:
    - Initially set equal to the number of features in the dataset, enabling a full analysis of variance explained by all principal components.
    - Focused on the first 10 components for downstream clustering and visualization, as identified by the **Bend or Elbow pattern** in the explained variance by principal component graph.
* **Reasoning for Selection:**
  + Retaining all components initially ensured no information loss during the explained variance analysis. Subsequently, the top 10 components were chosen for K-Means clustering as they captured the majority of the dataset's variance, making clustering more efficient and meaningful.
  + Dimensionality reduction facilitated easier interpretation and visualization in the reduced feature space.

**3. Hierarchical Clustering**

* **Hyperparameters Selected:**
  + **Linkage Methods:**
    - Evaluated multiple linkage methods: complete, single, centroid, average, and ward. Each was plotted with varying truncation levels (p=10, p=5, p=3) to compare results.
  + **Distance Metric:**
    - Used **Euclidean distance**, a standard choice for clustering numerical data.
  + **Number of Rows (p) in Dendrograms:**
    - Truncation levels of **10**, **5**, and **3** were explored to visualize clusters at varying levels of granularity.
* **Reasoning for Selection:**
  + Multiple linkage methods were tested to determine which provided the most meaningful separation for the dataset. Ward's method was particularly valuable due to its focus on minimizing variance within clusters. The Euclidean distance metric complemented the numerical nature of the data, while varying truncation (p) allowed an analysis of cluster granularity.

**Conclusion**

The hyperparameters for both K-Means and Hierarchical Clustering were iteratively refined based on exploratory visualizations and standard clustering practices. PCA played a pivotal role in determining the optimal number of clusters by providing a clear elbow pattern, ensuring robust and interpretable clustering results.

Performance Evaluation: Model Results

# **K-Means**

In clustering analysis, traditional supervised learning metrics like accuracy, precision, and recall are not applicable because clustering is an unsupervised learning task. Instead, we evaluate the model using metrics and visualizations that assess the quality and interpretability of the clusters formed.

A graph with blue lines

Description automatically generated A graph of different colored dots

Description automatically generated

Figure 1 Figure 2

1. **Explained Variance by Principal Component**:
   * The first graph shows the explained variance ratio for each principal component. Using this graph, the **elbow point** was identified at 10 components, which guided the selection of 10 as the optimal number of clusters for K-Means clustering after performing PCA. This ensures that the clusters formed are based on a compact yet comprehensive representation of the data.
2. **Cluster Visualization**:
   * The second graph visualizes the clusters formed in a two-dimensional PCA space. Each cluster is represented by a unique color, and the centroids are marked in black.
   * Observing this visualization, the clusters exhibit reasonable separation, indicating that the K-Means algorithm effectively segmented the data into distinct groups. However, there is some overlap between certain clusters, which might suggest potential improvement areas or an inherent overlap in the dataset features.
3. **Cluster Quality**:
   * To quantitatively assess cluster quality, metrics like **silhouette score** and **inertia** (sum of squared distances of samples to their nearest cluster centroid) can be used. The elbow method applied during the PCA phase indirectly confirms the adequacy of 10 clusters by minimizing intra-cluster variance while avoiding over-segmentation.
4. **Interpretation of Results**:
   * Each cluster represents a unique group within the dataset with similar characteristics. The plotted centroids demonstrate the core of each group and can serve as a reference for identifying cluster properties.

**Key Takeaways:**

The K-Means clustering model with 10 clusters demonstrates clear separability and consistency in grouping similar data points. Visualizations and the elbow method validate the effectiveness of the chosen number of clusters. Further analysis can focus on interpreting the specific characteristics of each cluster to draw actionable insights.

# **Hierarchical Clustering**

In hierarchical clustering, the evaluation focuses on the interpretability of the dendrograms, and the meaningfulness of the clusters formed. Below is a detailed explanation of the evaluation process for hierarchical clustering in this analysis:

1. **Dendrogram Analysis**:
   * **Experimentation with different values of p**:
     + Dendrograms were plotted for three different sets of the number of principal components (**p = 10, 5, and 3**) across various linkage methods, including **complete, average, centroid, ward, and single**.
     + After careful examination and comparison, **p = 5** was found to provide the best trade-off between granularity and interpretability for most linkages. This allowed for more detailed groupings, capturing meaningful hierarchical relationships in the data while avoiding overcomplication.
2. **Comparison of Linkage Methods**:
   * **Complete, Centroid, and Average Linkage**:
     + These linkage methods provided well-formed, interpretable groupings with clear separations between clusters. The structure of the dendrograms indicated meaningful hierarchical relationships in the dataset.
   * **Ward’s Linkage**:
     + Ward’s method resulted in the most detailed groupings, producing fine-grained clusters in the dendrogram for **p = 5**. However, its increased granularity might require domain-specific interpretation to avoid over-segmentation.
   * **Single Linkage**:
     + Single linkage dendrograms appeared overly simplistic across all values of **p**, often forming long chains and failing to effectively group similar data points. This linkage method was not chosen due to its tendency to create uneven and less meaningful clusters.
3. **Cluster Formation**:
   * Based on the **dendrogram cut** at an appropriate height, the hierarchical clustering was able to generate clusters that aligned well with the inherent structure of the data. The results varied slightly depending on the linkage method, but **complete, centroid, and average linkage methods with p = 5** provided the most balanced and interpretable clusters.
4. **Comparison with K-Means**:
   * Hierarchical clustering provides a hierarchical view of the data and allows for flexible cluster selection at various thresholds, which is an advantage over K-Means. The detailed grouping and hierarchical relationships observed in the dendrograms make it complementary to the partition-based approach of K-Means.

**Key Takeaways:**

The hierarchical clustering approach was thoroughly evaluated using dendrograms across different linkage methods and dimensionality settings (p = 10, 5, 3). After analysis:

* **p = 5** was selected as the optimal setting for dimensionality reduction.
* **Complete, centroid, and average linkages** demonstrated clear and meaningful groupings, with Ward’s method providing more detailed clusters.
* Single linkage was deemed less effective due to its overly simplistic cluster formations.

This analysis highlights the strengths of hierarchical clustering in visualizing relationships between clusters and the flexibility of choosing an appropriate number of clusters by varying the dendrogram cut height. Further insights can be gained by comparing these results with K-Means clustering to validate cluster stability and interpretability.

Interpretation: why a model is performing better/worse?

**1. Comparison of Model Performance:**

* **K-Means Clustering**:
  + **Strengths**:
    - K-Means performed better in terms of forming well-separated and compact clusters, as observed in the scatter plot of clusters after PCA.
    - The explained variance plot was instrumental in determining the number of principal components to retain, and the elbow pattern in K-Means allowed us to select the optimal number of clusters (**k = 10**). This helped in segmenting the data into meaningful groups with distinct centroids.
    - The visualization of clusters with centroids plotted revealed that the model effectively grouped data points with similar characteristics. The cluster boundaries were clear, and the centroids were representative of the groups.
  + **Limitations**:
    - K-Means assumes spherical clusters and may struggle with irregular cluster shapes. It is also sensitive to initialization, though this was mitigated by using multiple initializations.
    - K-Means doesn't provide hierarchical relationships, limiting its interpretability when compared to hierarchical clustering.
* **Hierarchical Clustering**:
  + **Strengths**:
    - Hierarchical clustering excels in providing insights into the hierarchical relationships between data points through dendrograms.
    - By experimenting with different linkage methods (complete, average, centroid, ward, and single) and different values of **p** (principal components), we observed that **p = 5** offered a detailed yet interpretable grouping of data.
    - Ward’s method produced the most granular clusters, while complete and average linkages provided well-balanced groupings with meaningful separations.
    - Unlike K-Means, hierarchical clustering does not require a pre-specified number of clusters, allowing flexible exploration of cluster thresholds.
  + **Limitations**:
    - It is computationally intensive for large datasets.
    - Single linkage produced simplistic clusters with chaining effects, making it less suitable for this dataset.

**2. Significant Features:**

* Both models used the **PCA-transformed features**, which represent combinations of the original variables in the dataset. The PCA graph of explained variance demonstrated that the first few principal components explained most of the variance, suggesting that:
  + **Principal components 1 and 2 (PC1 and PC2)** are likely the most significant, as they capture the maximum variability.
  + Features contributing heavily to the first few PCs are the most significant in driving the clustering results. These might include variables that exhibit the greatest variability or differentiate groups distinctly in the original dataset.
* Hierarchical clustering’s dendrograms also hinted at relationships between data points that align with these key features, reinforcing their significance.

**3. Reasons for Specific Cluster Assignments:**

* In **K-Means**, cluster assignments are influenced by proximity to centroids. The centroids are representative of the average characteristics of the data points within a cluster, so the model groups data points based on similarity in their PCA-transformed features.
* In **Hierarchical Clustering**, clusters are formed based on linkage criteria:
  + Complete, average, and centroid linkages group points by minimizing intra-cluster distances.
  + Ward’s linkage minimizes variance within clusters, resulting in smaller, more refined groupings.
  + The observed differences in cluster structures across linkage methods and values of **p** suggest that the specific geometry and density of the data significantly affect clustering.

**4. Relationship to Dataset and Problem Context:**

* Both models reveal meaningful patterns in the data that relate to the underlying problem.
  + For example, hierarchical clustering’s flexibility in exploring different cluster thresholds (via dendrogram cuts) allowed us to see both broad and fine-grained groupings, which is critical in analyzing data hierarchies.
  + K-Means provided a practical segmentation into **10 distinct groups**, which aligns well with the problem's requirement for structured cluster analysis.
* The PCA dimensionality reduction step ensured that the models focused on the most informative features of the dataset, avoiding noise and redundancy in the original data.

**5. Overall Insights:**

* Both clustering models are complementary:
  + **K-Means** is better suited for scenarios where a fixed number of clusters is required, providing compact and well-defined groupings.
  + **Hierarchical Clustering** is more flexible for exploratory analysis, offering a hierarchical perspective and the ability to visualize relationships at multiple levels.
* The results indicate that the dataset has meaningful separations based on the key features, and these separations align well with the problem's context. The models classified or clustered data into specific groups because of clear patterns in the PCA-reduced features, highlighting the effectiveness of dimensionality reduction as a preprocessing step.

Conclusion:

The analysis of the Apartment for Rent Classified dataset focused on understanding the relationships between apartment features and rental prices. Data preprocessing involved handling missing values, encoding categorical variables, and scaling numerical features. Exploratory analysis revealed key trends such as a positive correlation between price and apartment size, as well as significant regional price differences.

For model development, K-Means clustering with 10 clusters and hierarchical clustering with varying linkage methods were used to segment the data. Both models provided valuable insights into apartment groupings, with K-Means excelling in well-separated clusters and hierarchical clustering offering flexibility in interpreting relationships.

The results suggest that the clustering models effectively captured the inherent structure of the dataset, providing a foundation for future analyses in predicting rental prices or categorizing apartment listings.

References:

<https://archive.ics.uci.edu/dataset/555/apartment+for+rent+classified>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.spatial.distance.pdist.html>

<https://docs.scipy.org/doc/scipy/reference/generated/scipy.cluster.hierarchy.dendrogram.html>