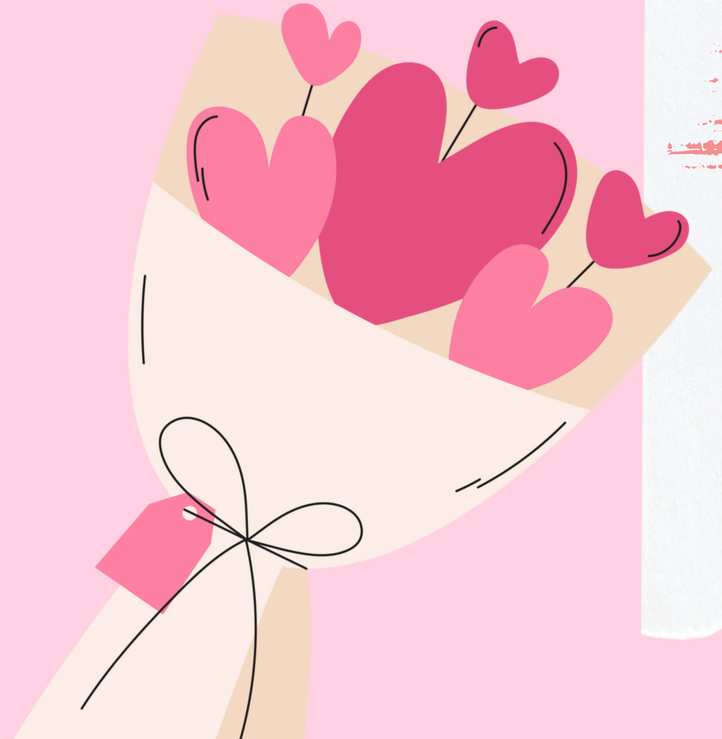
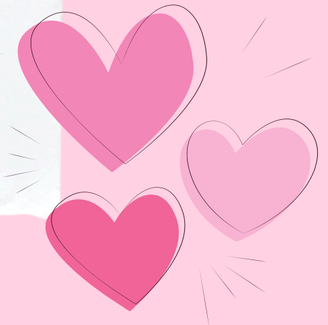
How Couples Meet and Stay Together





How Couples Meet and Stay Together (HCMST):

A study is a comprehensive, nationally representative survey examining the dynamics of how American adults find and maintain romantic relationships. Initiated in 2009 by researchers Michael J. Rosenfeld, Reuben J. Thomas, and Maja Falcon, the study has provided valuable insights into the evolving landscape of romantic partnerships in the United States.

How Couples Meet and Stay Together (HCMST 2009):
It comprises 388 variables, encompassing both categorical and numerical data.

Categorical Variables:

- Relationship Status
- Sexual Orientation

Numerical Variables:

- Age
- Relationship Duration

Introduction to K-Means Clustering:

K-Means is an unsupervised machine learning algorithm used to partition data into K distinct clusters based on feature similarity.

How K-Means Works:

Initialization: Select K initial centroids randomly.

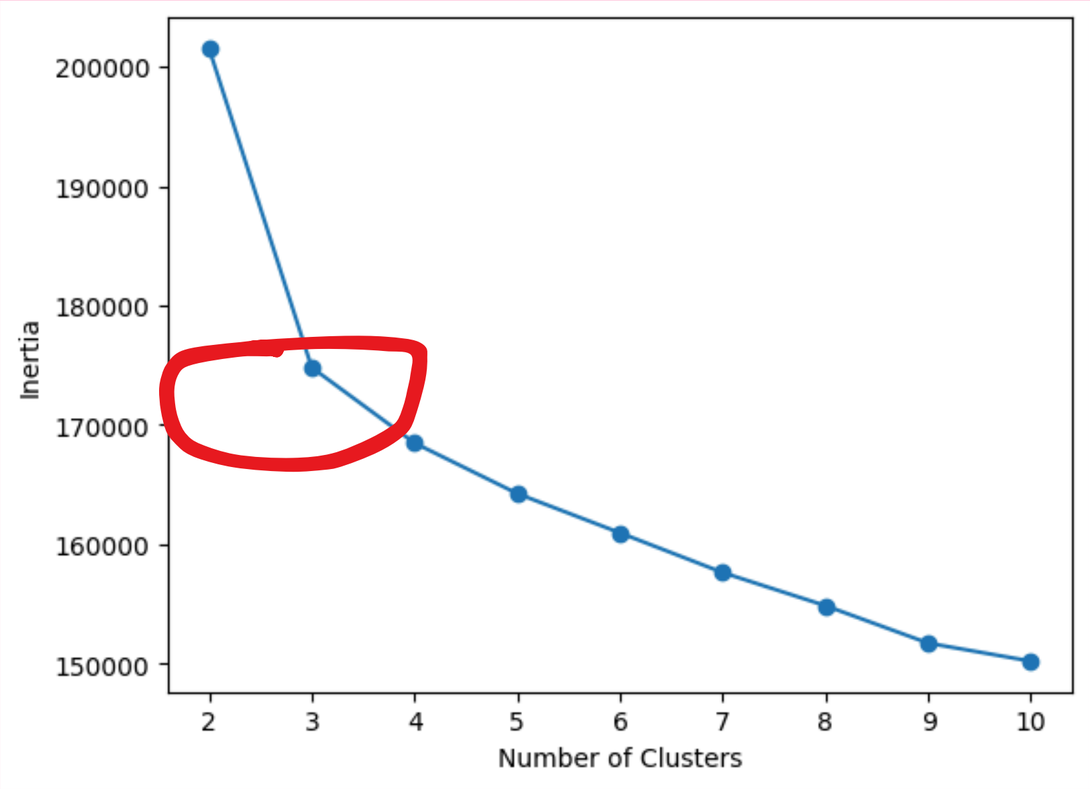
Assignment Step: Assign each data point to the nearest centroid, forming K clusters.

Update Step: Recalculate the centroids as the mean of all points in each cluster.

Iteration: Repeat the assignment and update steps until centroids stabilize (i.e., no change in clusters).

Elbow Method:

Plot the sum of squared distances from each point to its assigned centroid for different K values. The 'elbow point' indicates an optimal K.



Cluster Age Statistics:

Cluster 0: Mean Age: 62.94, Median Age: 62
Cluster 1: Mean Age: 52.51, Median Age: 52
Cluster 2: Mean Age: 50.61, Median Age: 53
Cluster 3: Mean Age: 33.86, Median Age: 33

Interpretation:

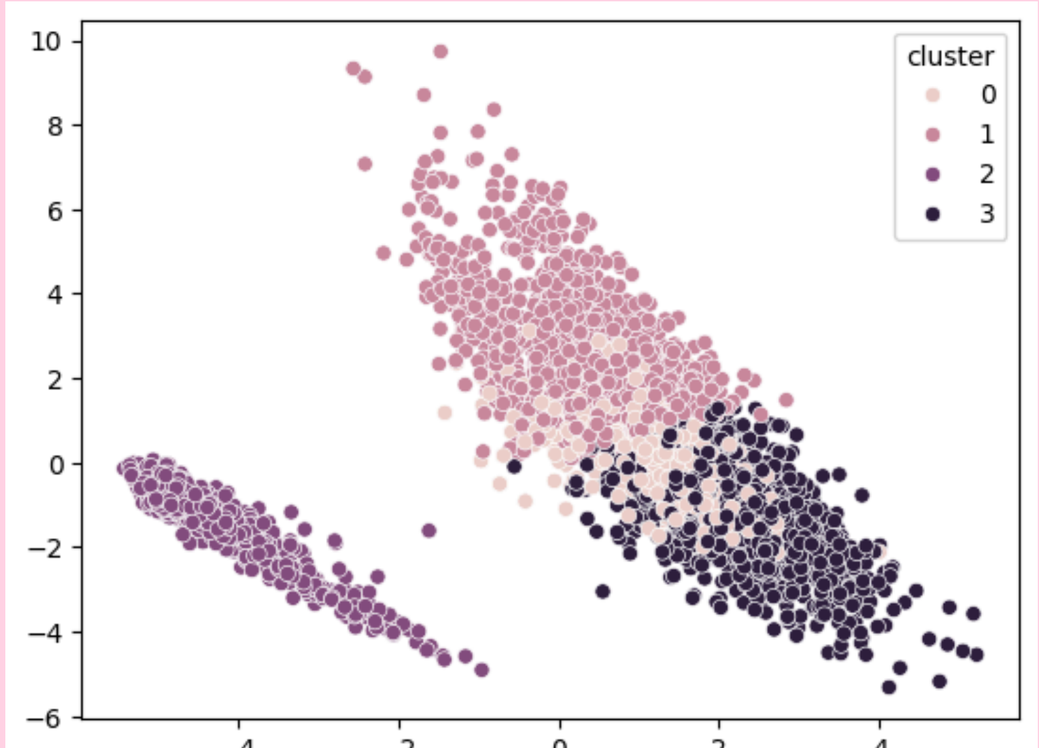
The distinct mean and median ages across clusters suggest that age is a primary determinant in the clustering algorithm.

Clusters are likely formed based on age-related groupings, reflecting different age demographics.

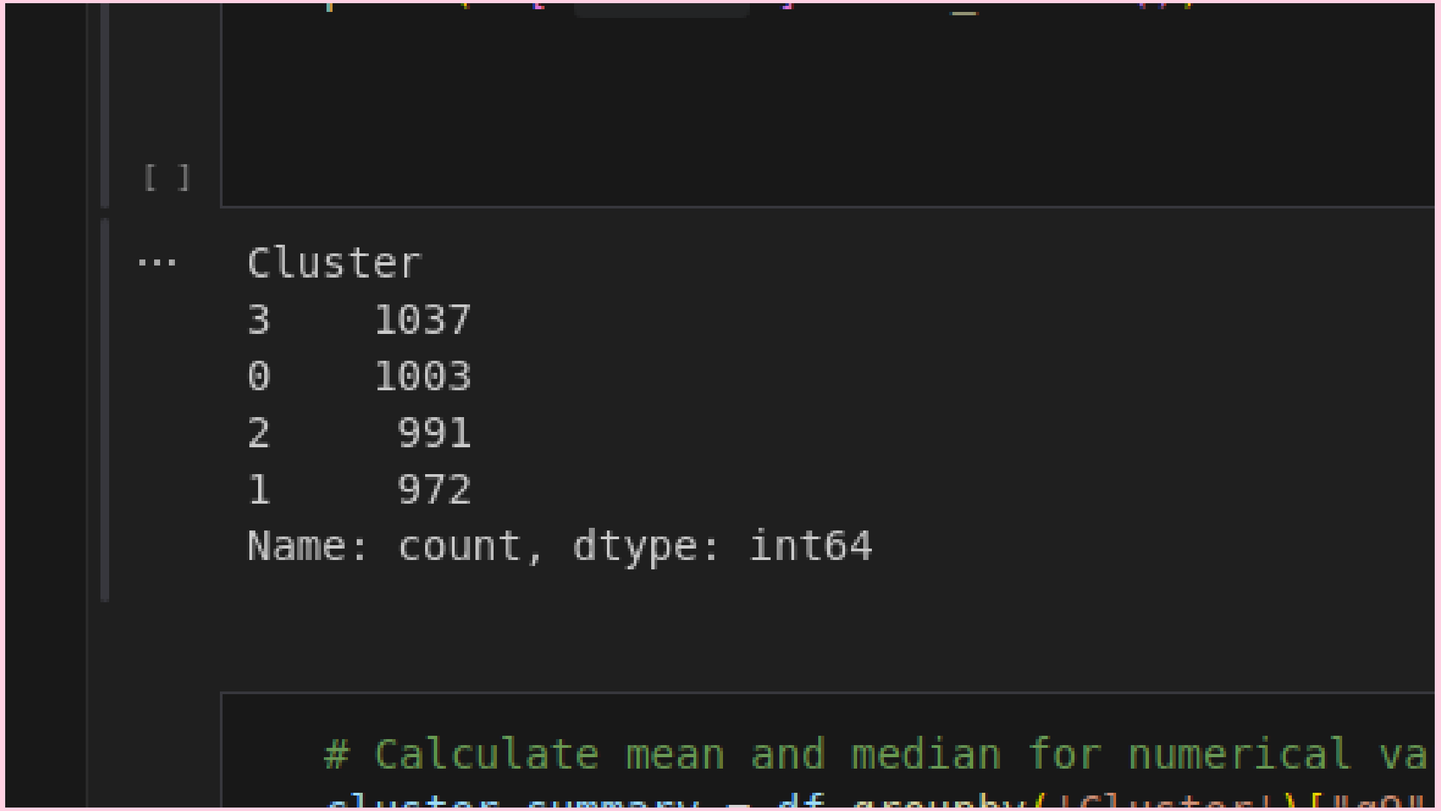
```
df['cluster'].value_counts()
```

```
cluster
3    1080
2     993
1     970
0     959
Name: count, dtype: int64
```

```
# Calculate mean and median for numerical va
cluster_summary = df.groupby('cluster')["q9"
```



KPrototypes



Findings

Data Complexity

Mixed data types: Numerical, categorical, and textual data.

Missing values and inconsistencies in the dataset.

Presence of outliers that could skew the results.

Impact on Models

KMeans: Assumes numerical data, problematic with categorical or mixed types.

KPrototypes: Works with mixed data, but can be sensitive to noise and missing values.

Recommendations

Data Cleaning Steps for KMeans and KPrototypes

- **1. Handle Missing Data**
  - For KMeans, use imputation techniques (mean, median, or KNN).
  - For categorical data, impute using the mode or prediction models.
- **2. Normalize Numerical Data**
  - Standardize numerical variables for KMeans to prevent dominance of one feature over others.
  - Consider scaling for KPrototypes when necessary.
- **3. Encode Categorical Data**
  - Use one-hot encoding or label encoding for categorical variables in KMeans.
  - For KPrototypes, encode categorical variables appropriately (e.g., using the ‘Hamming’ distance).
- **4. Remove Outliers**
  - Detect outliers using Z-scores or IQR (Interquartile Range) and remove or treat them.