**Title: Exploration Of User Preferences Through Clustering Analysis**

Subtitle - A Study on Recreational Activities and Cultural Interests

Author details here

**Abstract:**

This project explores a dataset obtained from the BuddyMove Data Set, accessible through the UCI Machine Learning Repository, focusing on users' preferences across multiple categories such as Sports, Religious, Nature, Theatre, Shopping, and Picnic. Employing unsupervised learning techniques, specifically clustering analysis and dimensionality reduction, we aim to uncover inherent structures and patterns within the dataset.

Initially, we conduct exploratory data analysis using basic techniques like box plots, histograms, and swarm plots to gain insights into the distribution of preferences across categories. Subsequently, we employ the K-means clustering algorithm to segment users into distinct clusters based solely on their preferences. A detailed analysis of each cluster's characteristics reveals unique patterns and tendencies among user groups.

To enhance understanding, Principal Component Analysis (PCA) is utilized to reduce the dataset's dimensionality while preserving its variance. Despite challenges, we successfully implement PCA, resulting in a visually compelling 2-D plot highlighting distinct clusters. Additionally, hierarchical clustering is explored as an alternative method, suggesting superior interpretability.

Overall, this project underscores the importance of employing diverse analytical techniques to extract meaningful insights from datasets, shedding light on user preferences and behavior across various categories.

**Introduction:**

The field of data science often involves unraveling patterns and extracting insights from datasets that encapsulate diverse aspects of human behavior and preferences. In this project, we embark on an exploration journey focusing on a dataset obtained from the BuddyMove Data Set, which offers a glimpse into users' recreational activities and cultural interests. The dataset is accessible through the UCI Machine Learning Repository, providing researchers and enthusiasts with a rich resource for analysis and experimentation.

Our primary objective is to delve into this dataset and uncover underlying structures and patterns inherent in users' preferences across multiple categories. Given the absence of labeled data and the unsupervised nature of the task, we turn to clustering analysis as a promising approach to segment users into distinct groups based solely on their preferences. This method allows us to discern meaningful clusters within the data, facilitating a deeper understanding of user behavior and preferences.

To kickstart our analysis, we begin with an initial exploration of the dataset using basic data analysis techniques such as box plots, histograms, and swarm plots. These techniques offer valuable insights into the distribution of preferences within each category, laying the foundation for further investigation.

Subsequently, we employ the K-means clustering algorithm to partition users into distinct clusters based on their preferences across various categories such as Sports, Religious, Nature, Theatre, Shopping, and Picnic. Post-clustering, we conduct a detailed analysis of each cluster's characteristics by computing the mean values of preferences within them. This analysis unveils unique patterns and tendencies among user groups, offering valuable insights into their preferences and behavior.

In our quest for deeper understanding, we leverage Principal Component Analysis (PCA) to reduce the dataset's dimensionality while preserving its variance. Despite encountering initial challenges, we successfully implement PCA, resulting in a visually compelling 2-D plot highlighting four distinct clusters. This visualization aids in a clearer comprehension of underlying structures and relationships within the data.

Additionally, we explore hierarchical clustering as an alternative method to K-means clustering. Through thorough discussions and analyses, hierarchical clustering emerges as a potentially more interpretable solution. By visualizing the dendrogram and evaluating the hierarchical structure of clusters, we conclude that grouping the data into three clusters would likely yield superior separation and facilitate easier interpretation.

In summary, this project underscores the significance of employing diverse analytical techniques in the absence of labeled data. Through clustering and dimensionality reduction methods, we extract meaningful insights from the dataset, shedding light on user preferences and behavior across various categories.

**Methodology**

**1 Data Acquisition and preparation:**

**Data Description:**

1. **Data Source:** The dataset utilized in this project is the BuddyMove Data Set, accessible through the UCI Machine Learning Repository. This dataset captures user preferences across a spectrum of categories, including Sports, Religious, Nature, Theatre, Shopping, and Picnic.
2. **Data Content:**
   * Each entry in the dataset represents a user's ratings or preferences across the specified categories.
   * Structurally, the dataset comprises rows representing individual users and columns corresponding to each category.
   * Ratings within each cell denote the degree of preference or engagement a user has for a particular category.

**Data Cleaning and Preprocessing:**

1. **Handling Missing Values:**
   * Any missing or null values within the dataset are identified and subsequently addressed through imputation or removal. This ensures the dataset's integrity and consistency.
2. **Outlier Detection:**
   * Anomalies or outliers in the dataset are identified using statistical methods or domain-specific knowledge. Appropriate actions are taken to handle these outliers to prevent them from skewing the analysis.
3. **Normalization and Standardization:**
   * To ensure uniformity and comparability across features, data normalization or standardization techniques may be applied. This process brings all features to a similar scale, aiding in the effectiveness of clustering algorithms.

This comprehensive data description outlines the origin and structure of the dataset, as well as the steps taken to prepare it for subsequent analysis.

**2 Data Exploration:**

**2.1 Column wise exploration**

1. **Descriptive analysis:**

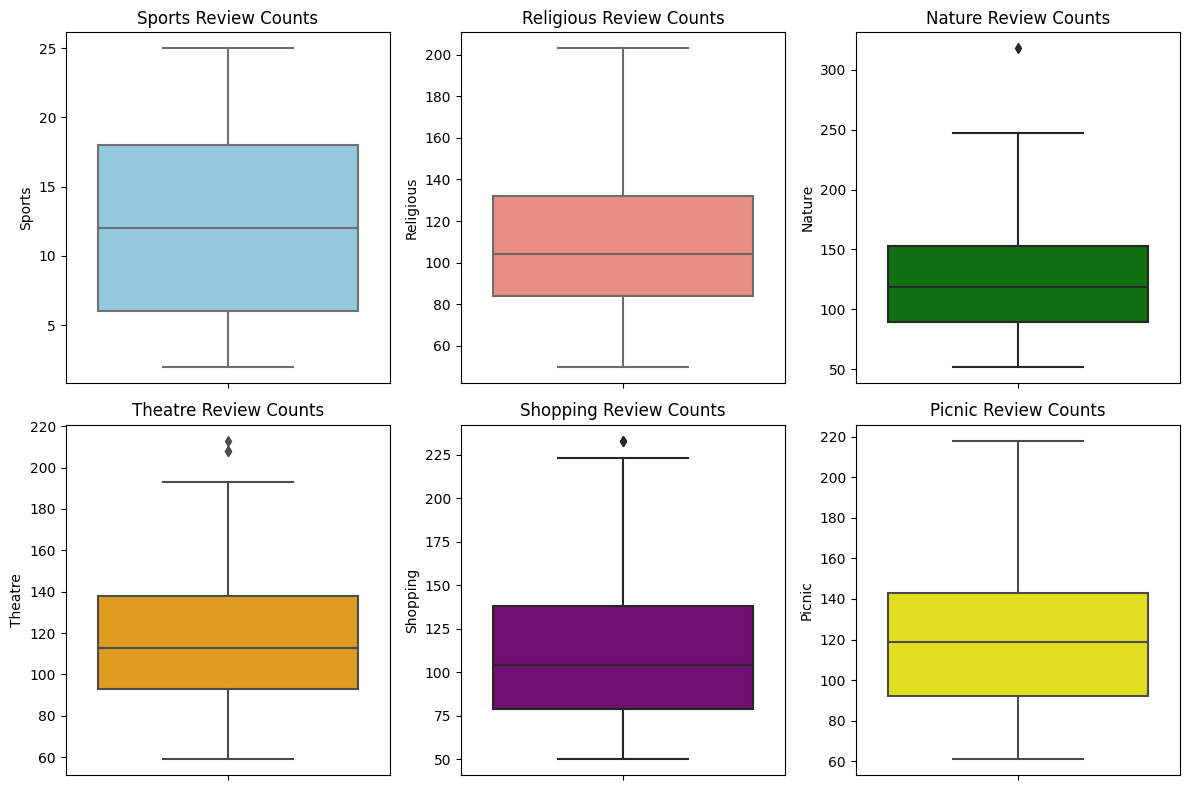
The descriptive statistics provide valuable insights into the distribution of preferences across different categories.

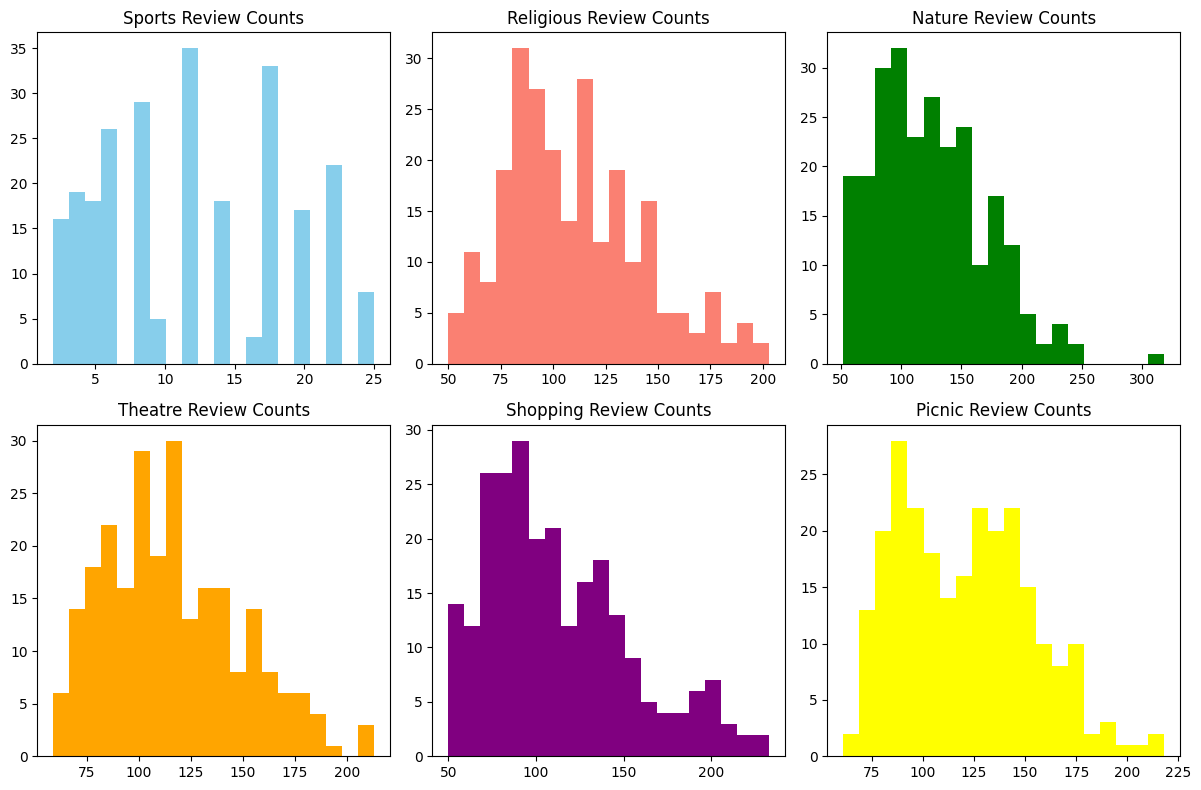
* **Sports:** The mean number of sports-related preferences is approximately 12, with a standard deviation of 6.6. This suggests that preferences for sports vary widely among users, with some showing a strong interest and others less so.
* **Religious:** On average, users have around 110 religious-related preferences, with a notable standard deviation of 32. This indicates a considerable diversity in religious preferences among the user base.
* **Nature:** The mean number of preferences related to nature is approximately 125, with a standard deviation of 45. This suggests that while many users have a strong affinity for nature-related activities, there is significant variability in the extent of this preference.
* **Theatre:** Users have an average of 116 theatre-related preferences, with a standard deviation of 32. This indicates a moderate level of interest in theatre activities among the user population.
* **Shopping**: On average, users express around 113 shopping-related preferences, with a standard deviation of 42. This highlights a wide range of shopping preferences among users, with some showing a strong inclination towards shopping activities.
* **Picnic:** The mean number of picnic-related preferences is approximately 120, with a standard deviation of 33. This suggests that while picnicking is a popular activity among users, there is variability in the intensity of this preference.

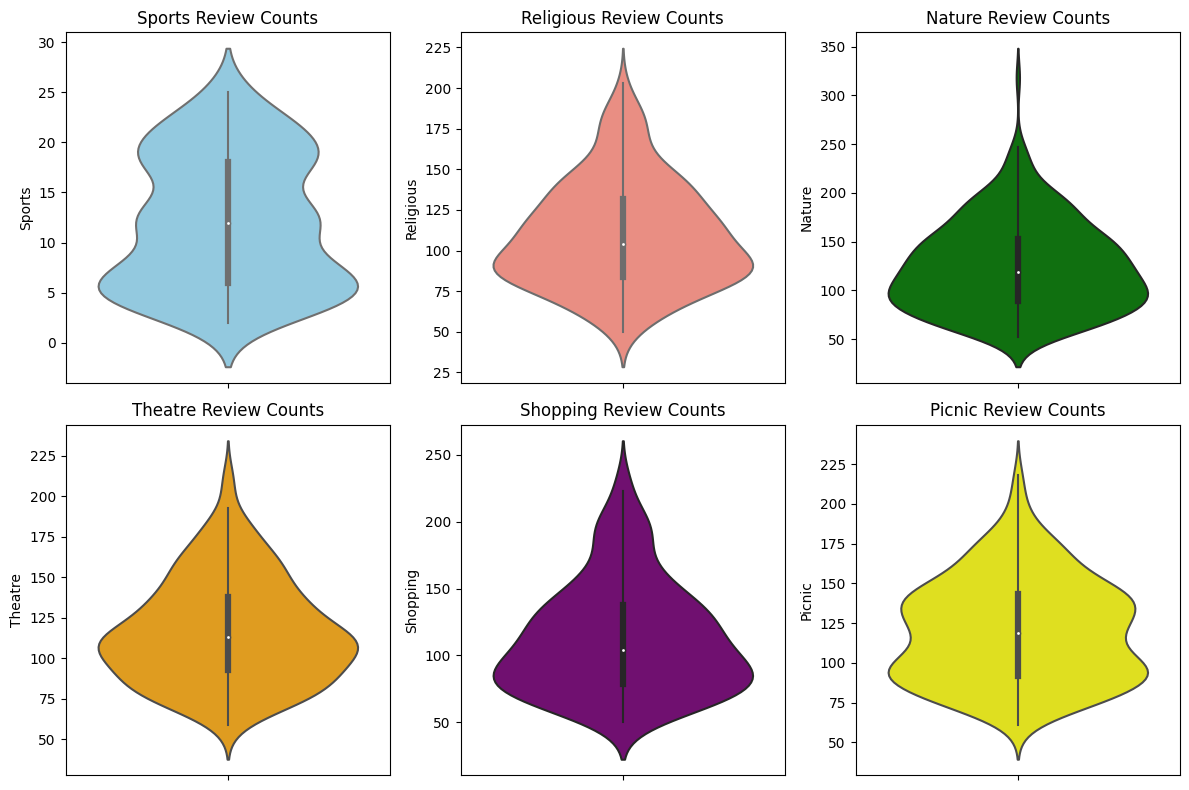
Overall, these descriptive statistics provide a comprehensive overview of user preferences across various categories, highlighting both common trends and individual differences within the dataset.

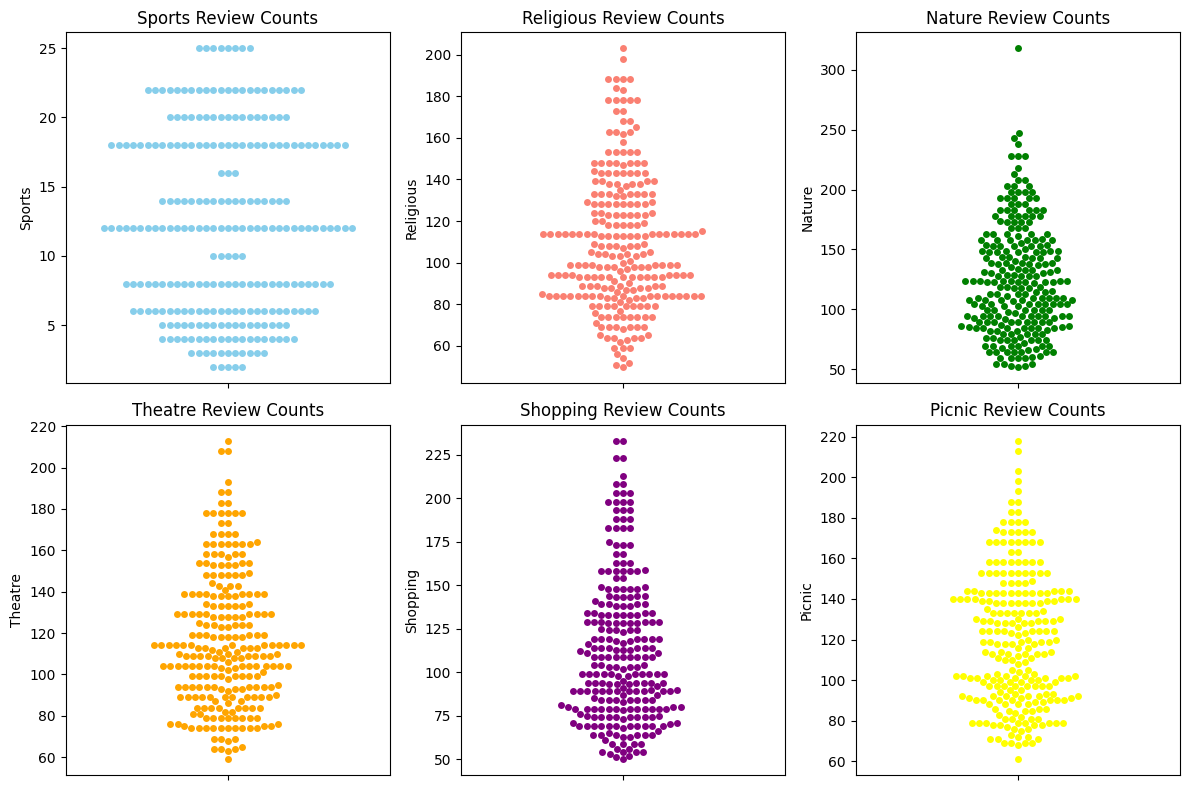
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | count | mean | std | min | 25% | 50% | 75% | max |
| Sports | 249 | 11.99 | 6.62 | 2 | 6 | 12 | 18 | 25 |
| Religious | 249 | 109.78 | 32.45 | 50 | 84 | 104 | 132 | 203 |
| Nature | 249 | 124.52 | 45.64 | 52 | 89 | 119 | 153 | 318 |
| Theatre | 249 | 116.38 | 32.13 | 59 | 93 | 113 | 138 | 213 |
| Shopping | 249 | 112.64 | 41.56 | 50 | 79 | 104 | 138 | 233 |
| Picnic | 249 | 120.4 | 32.63 | 61 | 92 | 119 | 143 | 218 |

**Unveiling Review Trends: A Look at Four Data Visualization Techniques**









**Unveiling Review Trends: Analysis of these 4 data visualization**

This report explores the distribution of review counts for various categories using four data visualization techniques: box plots, histograms, swarm plots, and violin plots. Each plot offers a unique perspective on the data, revealing insights into the number of reviews received by different categories and the presence of outliers.

**Box Plots: Unveiling the Median and Spread**

Box plots provide a concise summary of the distribution by displaying the median (center line), quartiles (box edges), and outliers (data points beyond the whiskers). Analyzing the review data through box plots reveals the following:

* **Median Review Count:** Across most categories, the median number of reviews falls between 125 and 160.
* **Spread of Reviews:** Sports and nature reviews exhibit the widest spread, indicating a higher variability in the number of reviews received. Religious, theatre, and shopping reviews have a tighter cluster around the median. Picnic reviews show the least spread, suggesting a more consistent number of reviews.
* **Outliers:** Some categories like sports and nature have outliers, with a few reviews receiving significantly more reviews than the majority.

**Histograms: Unveiling the Frequency Distribution**

Histograms depict the frequency distribution of review counts using bars. Each bar represents the number of reviews within a specific range. This visualization highlights the following:

* **Concentration Around Specific Ranges:** For all categories, the most frequent number of reviews falls within a range of 100 to 175. However, each category has its own "sweet spot" within this range. Sports reviews tend towards the lower end (10-15 reviews) before surging between 15-75. Religious reviews peak between 125-150, while theatre and shopping reviews concentrate around 150-175 and 100-150 respectively. Picnic reviews show the most consistent distribution around 100-125.

**Swarm Plots: Unveiling Individual Reviews**

Swarm plots showcase each review as a single dot, revealing the relationship between the number of reviews and the rating given. This plot offers insights into:

* **Number of Reviews:** The swarm plot confirms the observations of box plots and histograms, with most reviews clustered between 100-150. However, it also allows for visualization of outliers with very high review counts.
* **Correlation Between Number of Reviews and Ratings:** Due to the scaling of the rating axis, the swarm plot doesn't definitively show a strong correlation between review count and rating. However, it allows for visual exploration of potential trends in future analyses.

**Violin Plots: Combining the Best of Both Worlds**

Violin plots combine the features of box plots and histograms, displaying the median, interquartile range, and overall distribution of review counts. This visualization provides the following insights:

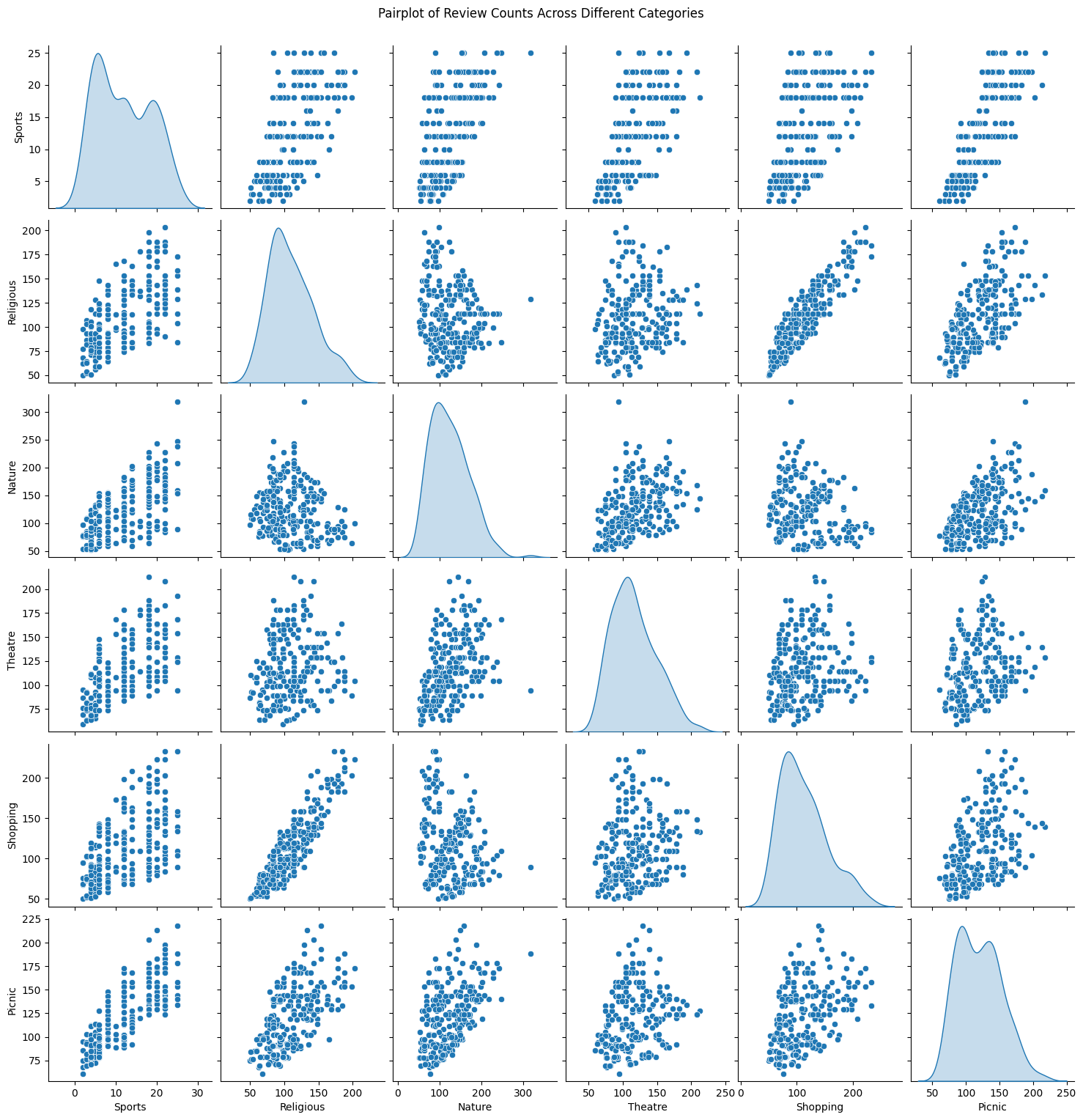
* **Overall Distribution:** Like box plots, violin plots confirm the clustering of review counts around 150, with wider tails for sports and nature categories and narrower tails for shopping and picnic reviews.
* **Outlier Identification:** Similar to box plots, violin plots visually identify outliers, allowing for further investigation into these specific reviews.

**2.2 Relational Analysis of Features**

1. **Pairplot Analysis:**

Pairplot analysis involves visualizing the relationship between each pair of features in the dataset. This technique provides a comprehensive overview of the correlations and distributions between variables.

Pairplot visually represents the pairwise relationships in a dataset, displaying scatterplots for numerical variables and histograms for univariate distributions along the diagonal.



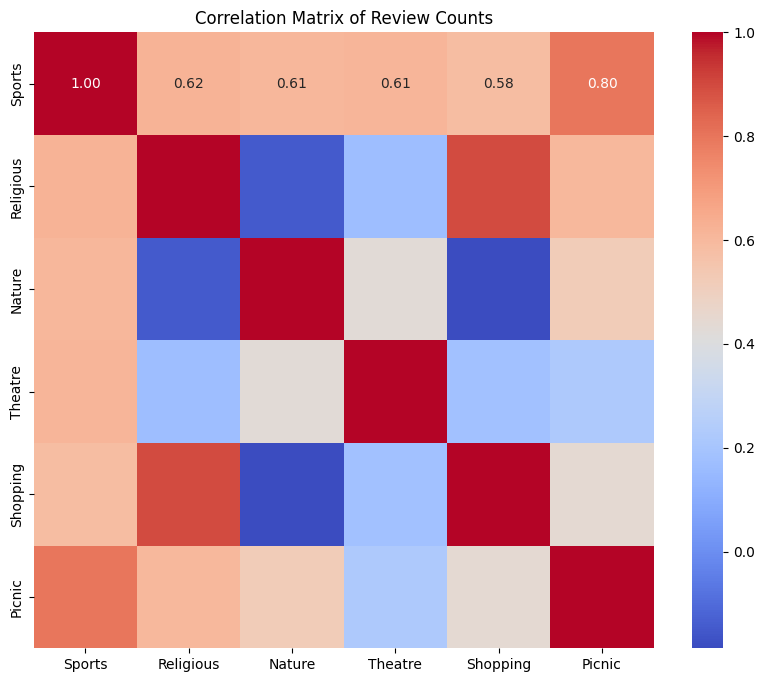
The above image is a pair plot, which is a collection of scatter plots created to show the relationship between pairs of variables in a dataset. This specific pair plot shows the relationship between the number of reviews for places in 6 different categories: Sports, Religious, Nature, Theatre, Shopping, and Picnic.

Here’s a summary of the relationship between each feature:

* **Religious vs Shopping:** There appears to be a positive correlation between the number of reviews for religious places and the number of reviews for shopping places. This means that places that have a lot of reviews for religious places also tend to have a lot of reviews for shopping places.
* **Religious vs Nature:** The data shows a weak positive correlation between the number of reviews for religious places and places listed under nature.
* **Nature vs Shopping:** There is no clear relationship between the number of reviews for places listed under nature and shopping places.
* **Nature vs Picnic:** There is a weak positive correlation between the number of reviews for places listed under nature and picnic places.
* **Shopping vs Picnic:** There is no clear relationship between the number of reviews for shopping places and picnic places.

1. **Correlation matrix analysis:**

Correlation matrix summarizes data by showing relationships between variables. Each cell represents a correlation coefficient, indicating how strong and directional the link is between two features.



**Positive Correlation:** when one feature increase other also increases

**Strong relation:** Strong positive correlation means two things increase (or decrease) together in a very clear way.

**Weak Positive**

 Nature & Sports

 Nature & Theatre

 Theatre & Shopping

**Moderate Positive**

 Sports & Religious

 Sports & Shopping

 Religious & Picnic

 Nature & Picnic

**Strong Positive**

 Sports & Picnic

 Shopping & Religious

**Weak Relation:**  Weak correlation means two things increase (or decrease) together, but it's not a strong trend.

**Moderate Relation:** Moderate correlation means two things increase (or decrease) together, but it's not as strong as a clear upward trend.

**No Relation**

 Theatre & Picnic

 Theatre & Religious

 Theatre & Picnic

**Weak Negative**

 Religious & Nature

 Shopping & Nature

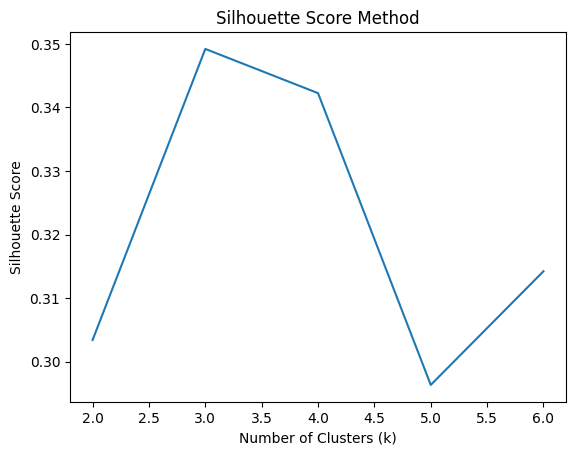
**3 Data Modelling:**

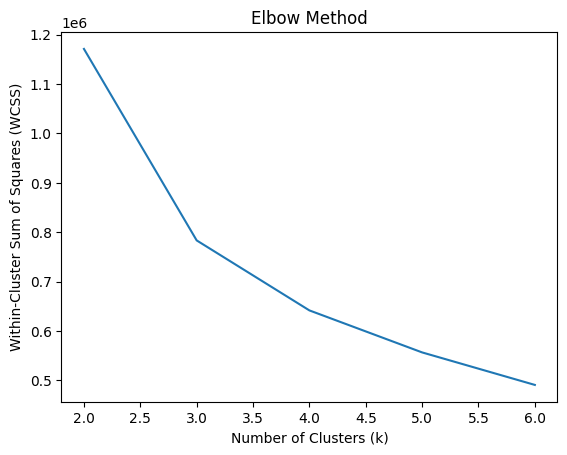
**Clustering Algorithm Selection**

Given the nature of our dataset with unlabelled data, we opted for clustering algorithms for data modeling. Our primary focus was on two clustering algorithms: K-means and hierarchical clustering.

**3.1 K-means Clustering:**

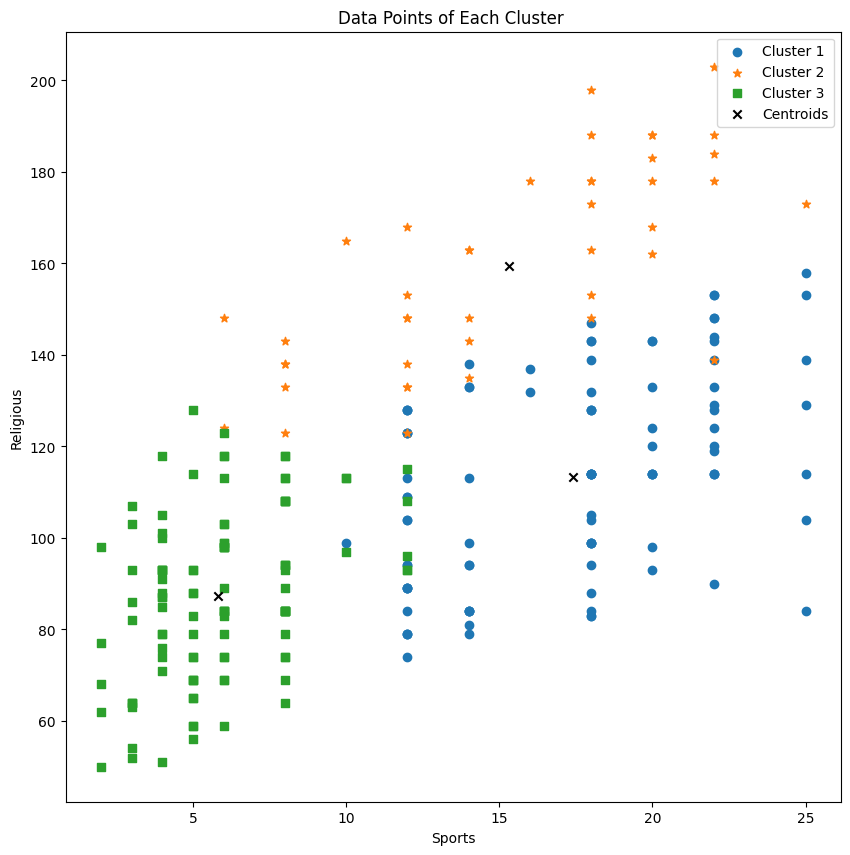
We performed K-means clustering with a range of values for 𝑘, from 8 down to 2. The selection of the optimal *k* value was critical for achieving meaningful clusters. To determine the best *k* value, we employed techniques such as the elbow method and silhouette score. After evaluating the performance at each *k*, we found that *k*=3 yielded the most cohesive and interpretable clusters. This decision was based on the balance between intra-cluster cohesion and inter-cluster separation.





**K-means Clustering Analysis**

Following the determination of the optimal number of clusters, we applied the K-means algorithm to partition users based on their preferences across various categories. Post-clustering, we conducted a comprehensive analysis of each cluster's characteristics by computing the mean values of preferences within them. This analysis unveiled distinct patterns and tendencies among user groups, offering valuable insights into their preferences and behavior.





**Cluster Analysis:**

**Cluster 0:**

* This cluster exhibits moderate preferences across all categories.
* Users in this cluster tend to have average scores for Sports, Religious, Nature, Theatre, Shopping, and Picnic.
* No specific feature stands out as having a significant influence in this cluster.

**Cluster 1:**

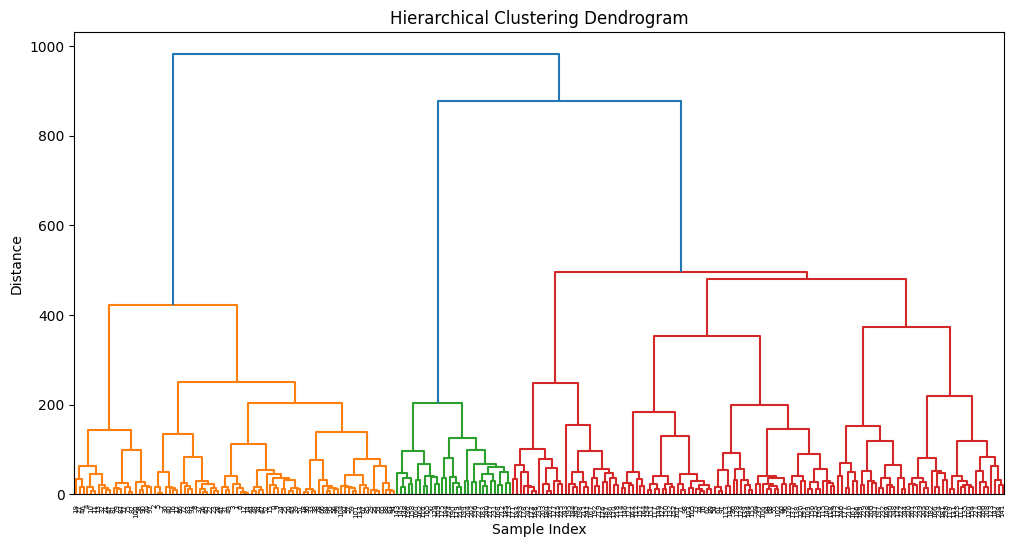
* This cluster represents users with strong preferences for Religious and Shopping.
* Users in this cluster have high scores for Religious and Shopping, with relatively lower scores for Nature and Theatre.
* Religious and Shopping are the dominant features influencing this cluster.

**Cluster 2:**

* This cluster represents users with low to moderate preferences across all categories.
* Users in this cluster have the lowest scores overall, indicating a general lack of interest or preference.
* No specific feature stands out as having a significant influence in this cluster.

**3.2 Hierarchical Clustering:**

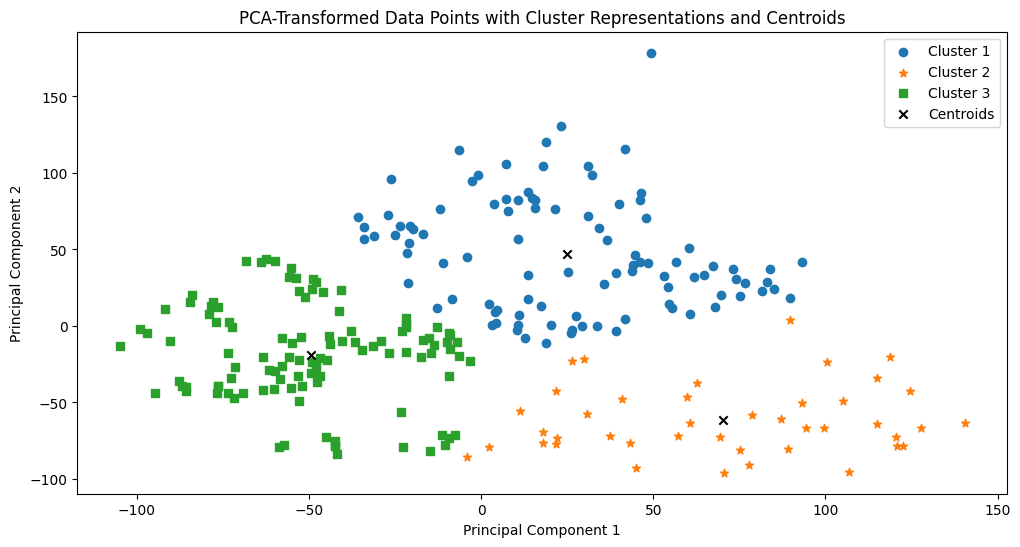
In addition to K-means clustering, we explored hierarchical clustering as an alternative approach. By plotting the dendrogram, we gained insights into the hierarchical structure of the data and identified the optimal number of clusters. The dendrogram revealed that partitioning the data into three clusters would provide the most meaningful segmentation, ensuring clear separation and interpretability.



So from here also it is clearly visible that we should take 3 clusters (Orange, Green and Red part)

**3.3 Principal Component Analysis (PCA)**

To gain deeper insights and visualize the data in a lower-dimensional space, we employed Principal Component Analysis (PCA). Despite initial challenges, we successfully implemented PCA, resulting in a visually compelling 2-D plot highlighting four distinct clusters. This visualization aided in a clearer comprehension of underlying structures and relationships within the data.



So here we have clearly visible 3 clusters Star, Square and Disc with all the centroids of all 3 clusters

**Results:**

The results of our analysis reveal insightful patterns and trends within the dataset, shedding light on user preferences across various categories. Through a combination of data exploration, clustering algorithms, and dimensionality reduction techniques, we uncovered distinct clusters and gained deeper insights into user behavior.

**Exploratory Data Analysis:**

Our initial exploration of the dataset provided valuable insights into the distribution of preferences across different categories. Descriptive statistics such as mean, standard deviation, and quartiles offered a comprehensive overview of user preferences, highlighting both common trends and individual differences within the dataset.

**Clustering Analysis:**

**K-means Clustering:**

* We performed K-means clustering with a range of values for 𝑘*k*, ultimately selecting 𝑘=3*k*=3 as the optimal number of clusters.
* Cluster analysis revealed three distinct groups of users:
  + Cluster 0: Users with moderate preferences across all categories.
  + Cluster 1: Users with strong preferences for Religious and Shopping.
  + Cluster 2: Users with low to moderate preferences across all categories.
* Religious and Shopping emerged as dominant features influencing Cluster 1, while no specific feature stood out in Clusters 0 and 2.

**Hierarchical Clustering:**

* Hierarchical clustering, visualized through a dendrogram, suggested partitioning the data into three clusters for optimal segmentation.
* This hierarchical structure provided clarity and interpretability, confirming the decision to adopt three clusters.

**Dimensionality Reduction:**

**Principal Component Analysis (PCA):**

* Despite initial challenges, PCA was successfully implemented to reduce the dataset's dimensionality while preserving its variance.
* The resulting 2-D plot highlighted four distinct clusters, offering a clearer comprehension of underlying structures and relationships within the data.

**Insights:**

1. **User Segmentation:** Clustering analysis revealed distinct user segments based on their preferences. These segments provide valuable insights for targeted marketing strategies and personalized recommendations.
2. **Dominant Features:** Religious and Shopping emerged as influential factors driving user preferences in certain clusters, indicating potential areas for further investigation and marketing campaigns.
3. **Optimal Cluster Size:** Both K-means and hierarchical clustering methods converged on three clusters, suggesting a consistent segmentation approach for this dataset.
4. **Dimensionality Reduction:** PCA facilitated visualization and interpretation by reducing the dataset's dimensionality, enabling clearer insights into cluster structures.

**Conclusion:**

Our analysis demonstrates the effectiveness of clustering algorithms and dimensionality reduction techniques in uncovering meaningful patterns within unlabelled datasets. By leveraging these methods, we gained valuable insights into user preferences and behavior, laying the foundation for informed decision-making and targeted strategies in various domains such as marketing, product development, and user engagement initiatives.

Top of Form