Visual Data Science

Model

Clara Pichler[1]

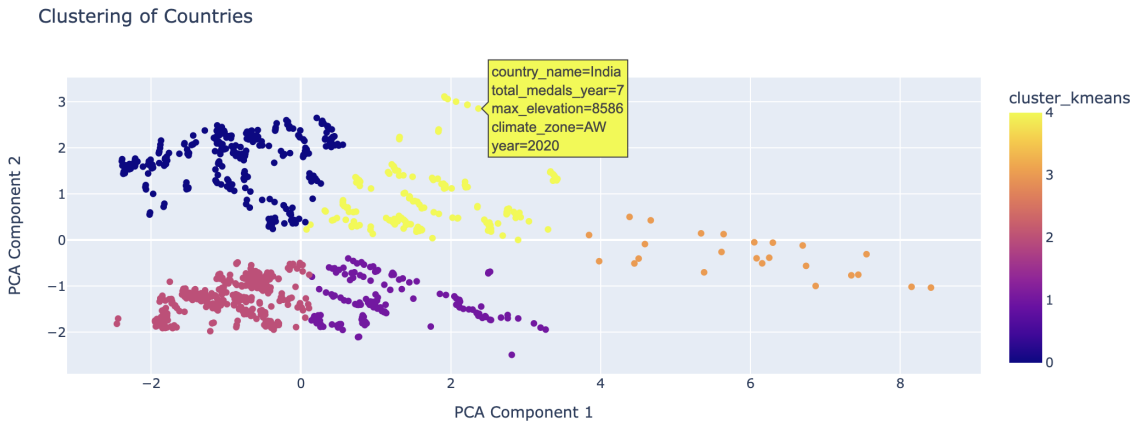
In this project, we aimed to explore patterns among countries participating in the Olympics by clustering them based on their medal counts and related features.

1 Model Description

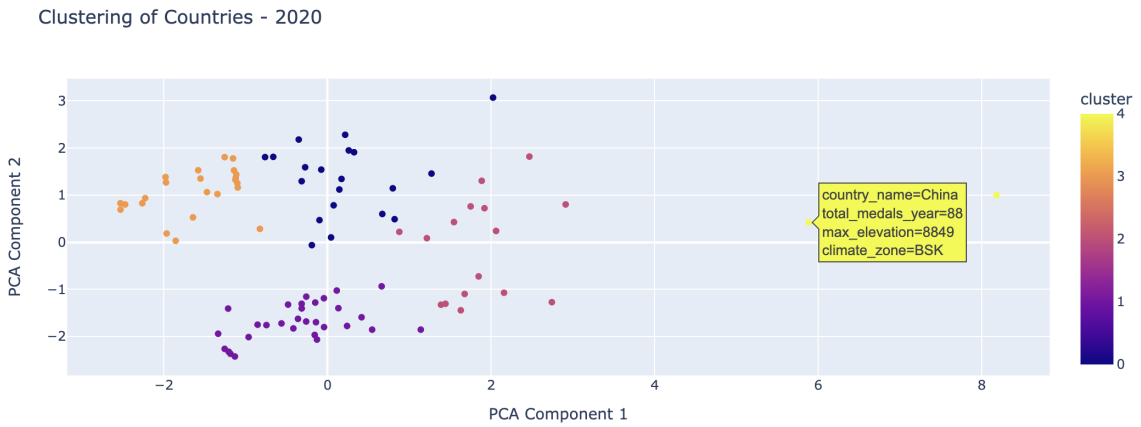
In this project, we aimed to explore patterns among countries participating in the Olympics by clustering them based on their medal counts and related features. Our modeling approach began with a bit of further data preprocessing. We considered not only the whole data set but also for each year. Non-essential columns such as `discipline_title`, `event_title`, and `event_gender` were removed to focus on attributes directly related to country performance. We addressed missing values by excluding incomplete entries and used one-hot encoding to transform categorical variables like `climate_zone` into a suitable numerical format for analysis.

Next, we applied standardization using `StandardScaler` to normalize the numerical features such as GDP, max elevation, and average temperature. This step was crucial to ensure that all features contributed equally to the clustering process, preventing features with larger scales from dominating the results. We then employed *Principal Component Analysis* (PCA) to reduce the dimensionality of the dataset. PCA helped in simplifying the dataset into two principal components, making it easier to visualize and interpret while retaining the most significant variance. With the data prepared, we utilized *K-means clustering* to group countries based on their similarities. The number of clusters was chosen based on empirical observation and validation techniques like the silhouette score.

To present the clustering results, we used Plotly Express to create interactive scatter plots. These plots allowed us to visualize the relationships between countries in the reduced PCA space, with each point representing a country (or a combination of country and year when considering the whole data set) and clusters differentiated by color. Hover features provided detailed insights into each country's performance, including medal counts, GDP, and other relevant attributes.



[1] 11917694 - e11917694@student.tuwien.ac.at



2 Increasing Trust through Data Visualizations

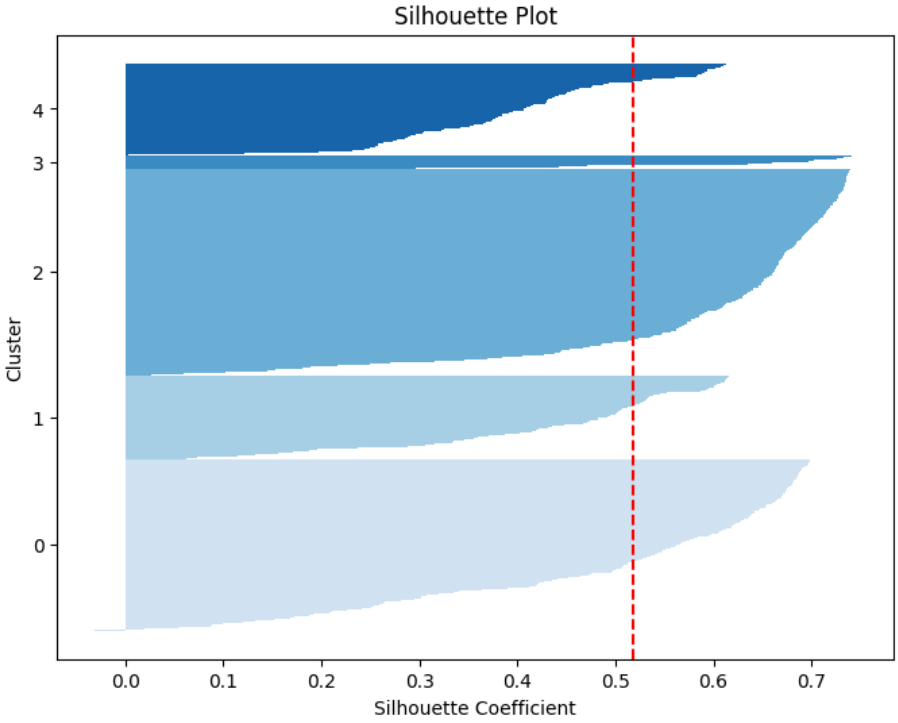
Data visualization plays a pivotal role in building trust among our customers or colleagues by making complex data more accessible, understandable, and actionable. The following visualizations exemplify how we can enhance transparency and foster confidence in the modeling process.

Scatter Plot of PCA Components: The PCA scatter plot visually represents the clustering of countries based on their principal component scores. By color-coding the clusters and allowing for hover-over details such as the country name, total medals, maximum elevation, and climate zone, this plot makes the data more interactive and informative. Customers can easily explore and validate the clustering results, fostering trust in the model's ability to group countries meaningfully. Two examples of those scatter plots can be seen in the section above.

The **Silhouette Plot** provides a visual summary of the clustering performance, indicating how well each data point fits within its assigned cluster compared to other clusters. By displaying the silhouette coefficients for each cluster, this plot enables customers to assess the quality of the clustering process. The red dashed line representing the average silhouette score further aids in evaluating the overall performance of the model, thereby reinforcing confidence in the cluster validity.

Grouping countries by clusters and displaying their total medals (or other features like the GDP or maximum elevation) in **Boxplots** provides insights into the distribution and variability within each cluster. This visualization highlights which clusters have the highest or most consistent performance in terms of total medals, making it easier for customers to interpret the results in the context of their expectations and domain knowledge.

In conclusion, using a variety of visualizations allows us to present data and model results in a more engaging and comprehensive manner. By offering multiple perspectives and emphasizing interpretability, these visual tools help build a narrative around the data that is transparent and trustworthy. This approach ensures that customers can not only see the results but also understand and trust the processes that led to those results.


Silhouette Plot
Cluster
4
3
2
1
0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Silhouette Coefficient