**Abstract/Introduction**

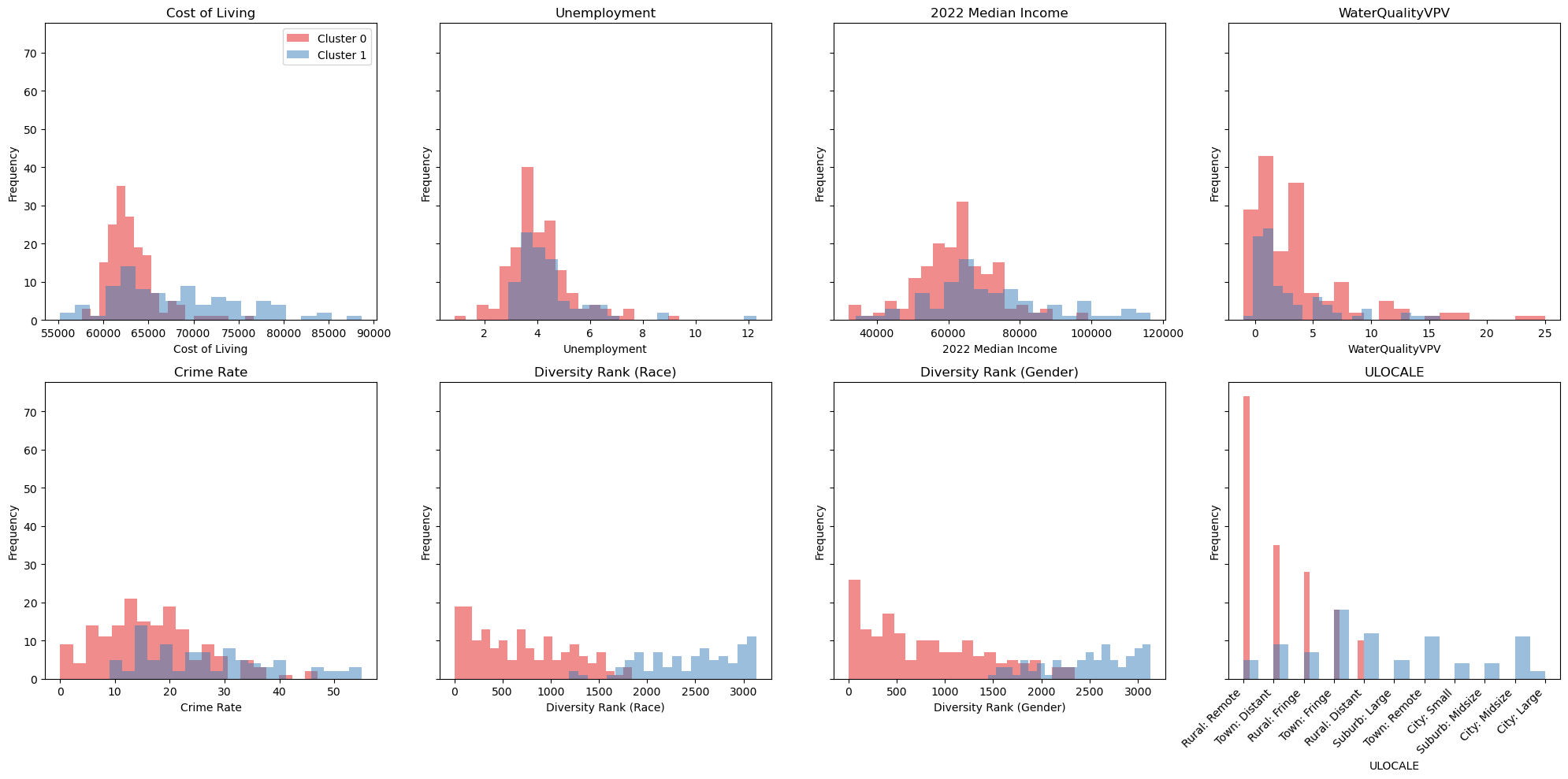
The state of Texas is the second largest in the US, covering almost 270,000 square miles of land, with 254 distinct counties (World Atlas, 2024). Each county has varied walkability, air quality, diversity, and economic prosperity, raising the question: which are the top five counties in the state of Texas most desirable to live in as a recent college graduate? Furthermore, which counties share similar attributes, and are there natural groupings of these counties? We inferred that counties would naturally group together based on similar attributes, as aspects such as unemployment rate and diversity can be influenced by the urban or rural location of the county. In determining livability, we used data from every county in Texas that included their corresponding cost of living, unemployment rate, crime rate, median income, water quality, diversity rank, and locality (urban, rural, etc.) to run a Principle Component Analysis (PCA) and a subsequent K-means clustering analysis. In running the PCA, we found that the first principle component was most influenced by diversity features and locality while the second principle component was most influenced by unemployment rate and crime rate. In running the K-means clustering, we found that dividing the data into two clusters yielded the highest silhouette score (0.261) and that Dimmit County was scored as the most livable county for college graduates.

**Methods**

We first extracted data from the All-in-One Quality of Life Analysis dataset from Gigasheet that included every county in the United States and the corresponding factors contributing to quality of life (Gigasheet, 2025). We then filtered the dataset to capture Texas counties and identified our features (listed in the introduction), cleaning the rows to include proper syntax and encoding the ordinal feature (locality) to ensure numeric values are represented in preparation for analysis (Open AI, 2023). Our assumption was that college graduates prefer more walkable/urban areas, so we encoded higher values to the urban counties and ranked them based on that criteria. With our new features, we inverted some of the values (cost of living, unemployment, diversity) to correspond a higher score to a higher livability. Our last step of data cleaning was scaling the features, making sure that each feature had the same weight of influence on the model. In terms of analyzing the data, we ran a PCA as the first step in enhancing visualization of our eight features on a two dimensional graph. We visualized the PCA as a scatterplot of each county, inspecting the two components to see which features contributed the most/least to the analysis. With our PCA analysis, we tested several K-means clustering models with varying numbers of clusters to find the highest silhouette score. This ended up being just two clusters. After plotting our final K-means clustering, we wanted to make the data more interpretable, so we created another visualization using Plotly that labeled each county within each cluster and its “score” within each dimension (Open AI, 2023).

**Results**

After fitting and plotting our PCA (Figure 1), we found that our diversity features (gender and race) and locality contributed most to the first principle component, with values around 0.50 (diversity) and 0.15 (locality), respectively (Table 1, Figure 2a). In the second principle component, unemployment (0.56) and crime rate (0.44) contributed more significantly (Table 1, Figure 2b). Of the eight K-means models, the model with two clusters performed the best, outputting a silhouette score of 0.261 (Figure 3 & 4). The remaining seven models did not score higher than 0.221, and did not score less than 0.177 (Figure 3). We found that the primary difference between the clusters were their rank means of the diversity features (Figure 5).

****

In cluster one, race garnered an average mean of 2446.80, while cluster zero had an average mean of 671.81 (Table 2). Similarly, gender in cluster one generated an average mean of 2485.23, while cluster zero had an average mean of 796.58 (Table 2). Overall, cluster zero had an average lower cost of living ($63,070), unemployment rate (4.07), and crime rate (16.52) compared to cluster one, as well as a higher water quality amongst the counties (3.44) (Table 2). On the other hand, cluster one had a higher median income, garnering an average of $72,401 compared to an average of $62,423 in cluster zero counties respectively (Table 2). Based on the livability score determined by the PCA-transformed values by amount of variance explained by each of the components, our analysis designated Dimmit County, Brooks County, Jim Hogg County, Kenedy County and Starr County as the best counties for recent college graduates to live (Table 3).

**Discussion**

Our results show that Texas counties can be grouped by certain quality-of-life features, but the groupings themselves may not be strong. With the silhouette score of our final K-means model being 0.261 (considered fairly weak), many clusters may not belong to one cluster over another. Despite this, we could effectively observe general trends, specifically regarding metrics such as diversity ranks, crime and unemployment rate. In our final two-cluster model, cluster zero counties, which scored higher in livability for recent college graduates, generally offered lower costs of living, lower crime, better water quality and more rural environments. On the other hand, cluster one counties had higher diversity and median income but also higher unemployment and crime. This may suggest that safety and affordability may outweigh urban living and potential salary for new college graduates, but should be interpreted with caution. With our results, we were surprised that urban counties, such as Harris (home to Houston) and Dallas county did not rank higher. For future work, it is important to consider more quality-of-life metrics, such as public transportation access, housing availability, average age and job data in order to get a more well-rounded model. It may also be beneficial for future research to add weights to features based off of actual surveys completed by recent college graduates in terms of what they look for in a post-grad location. In addition, we also initially wanted to include more states to observe if livability was consistent or different across states. A larger dataset may have led to more natural groupings, improving our silhouette scores.

**Appendix**

Table 1: Depiction of the contribution of each feature to the eight principal components

|  | **PC1** | **PC2** | **PC3** | **PC4** | **PC5** | **PC6** | **PC7** | **PC8** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Diversity Rank (Gender) | 0.52 | -0.29 | -0.02 | -0.05 | 0.15 | 0.36 | 0.11 | -0.70 |
| Diversity Rank (Race) | 0.51 | -0.31 | -0.02 | -0.07 | 0.16 | 0.33 | -0.02 | 0.71 |
| Locality | 0.15 | 0.13 | -0.42 | 0.89 | 0.03 | 0.00 | 0.04 | 0.02 |
| Unemployment | 0.09 | 0.56 | -0.18 | -0.16 | -0.54 | 0.57 | -0.07 | 0.02 |
| Water Quality | 0.08 | -0.01 | 0.86 | 0.40 | -0.27 | 0.12 | -0.02 | 0.01 |
| Crime Rate | -0.22 | 0.45 | 0.19 | 0.03 | 0.75 | 0.38 | 0.04 | 0.01 |
| 2022 Median Income | -0.42 | -0.42 | -0.09 | 0.12 | -0.01 | 0.38 | -0.69 | -0.06 |
| Cost of Living | -0.46 | -0.35 | -0.06 | 0.07 | -0.13 | 0.36 | 0.71 | 0.06 |

Table 2: Calculation of the average mean values of each feature in the two-cluster model. Locality is non-numeric, so therefore is calculated for mode.

| Cluster | *Cost of Living ($)* | *Unemployment*  *(%)* | *2022 Median Income ($)* | *Water Quality (VPV)* | *Crime Rate*  *(% per 100,00 people)* | *Diversity Rank (Race)* | *Diversity Rank (Gender)* | *Locality*  *(*Mode*)* |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 63,070.43 | 4.0746 | 62,423.11 | 3.44 | 16.52 | 671.83 | 796.58 | Rural |
| **1** | 68,029.7 | 4.4264 | 72,400.94 | 2.91 | 26.78 | 2446.80 | 2485.23 | Town |

Table 3: The top five counties to live as a recent college graduate in Texas according to our model. Livability score based on PCA-transformed values by the amount of variance explained by each of the components.

| **County** | **Livability Score** |
| --- | --- |
| 1. Dimmit | 1.650 |
| 1. Brooks | 1.454 |
| 1. Jim Hogg | 1.397 |
| 1. Kenedy | 1.381 |
| 1. Starr | 1.299 |

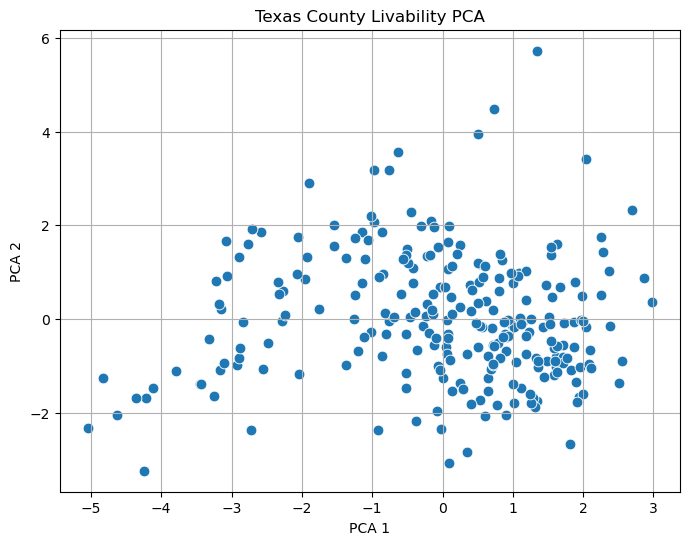
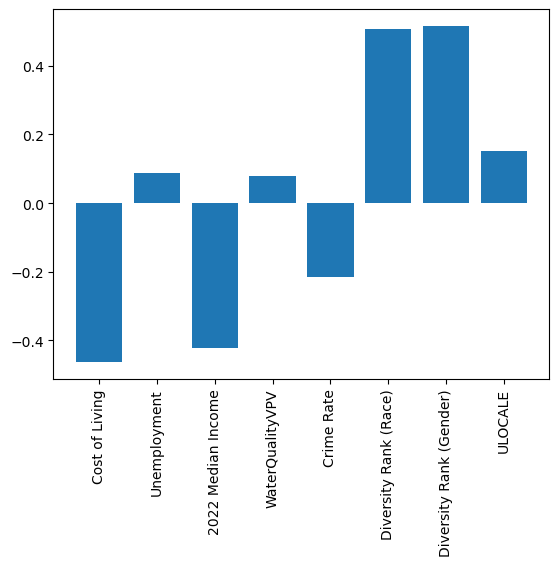


Figure 1: Principal Component Analysis (PCA) of Texas Counties based on Quality-of-Life features (Diversity Ranks, Water Quality, Crime Rate, Cost of Living, 2022 Median Income, Unemployment, Locality).



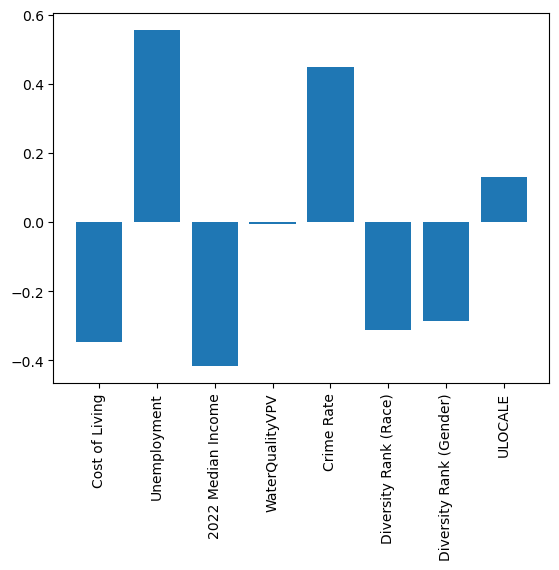


Figure 2: Two bar graphs depicting contribution of each feature into PCA1 (a.) and PCA (b.).

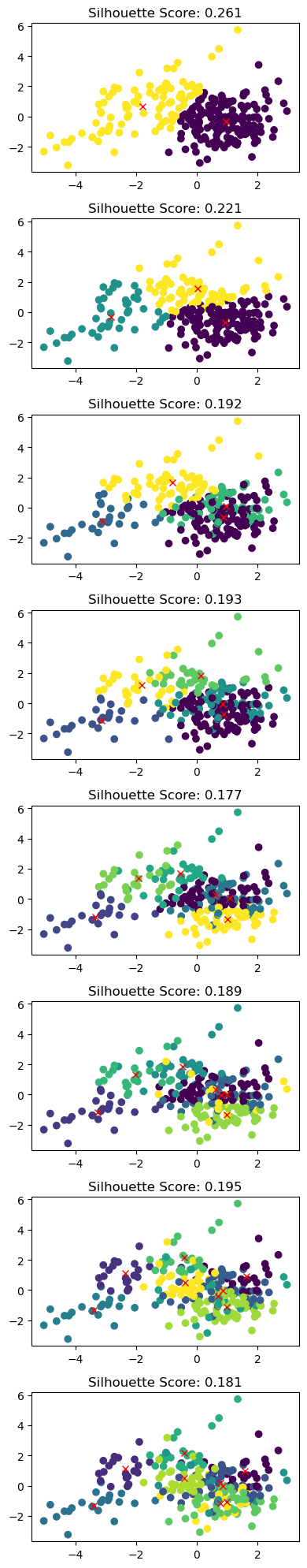


Figure 3: K-Means Clustering results across varying cluster counts (k = 2 to 9) with corresponding silhouette scores.

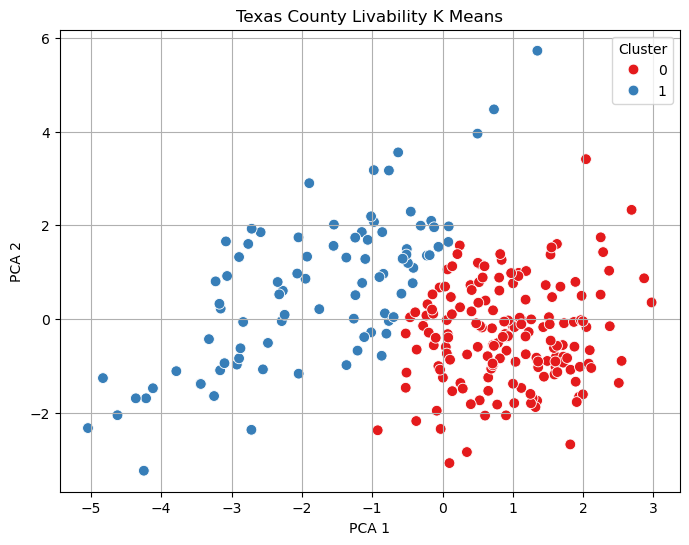
****

Figure 4: Final K-means clustering (k = 2) of Texas counties. Cluster zero defined by diversity ranks and locality, while cluster one defined by unemployment and crime rate.

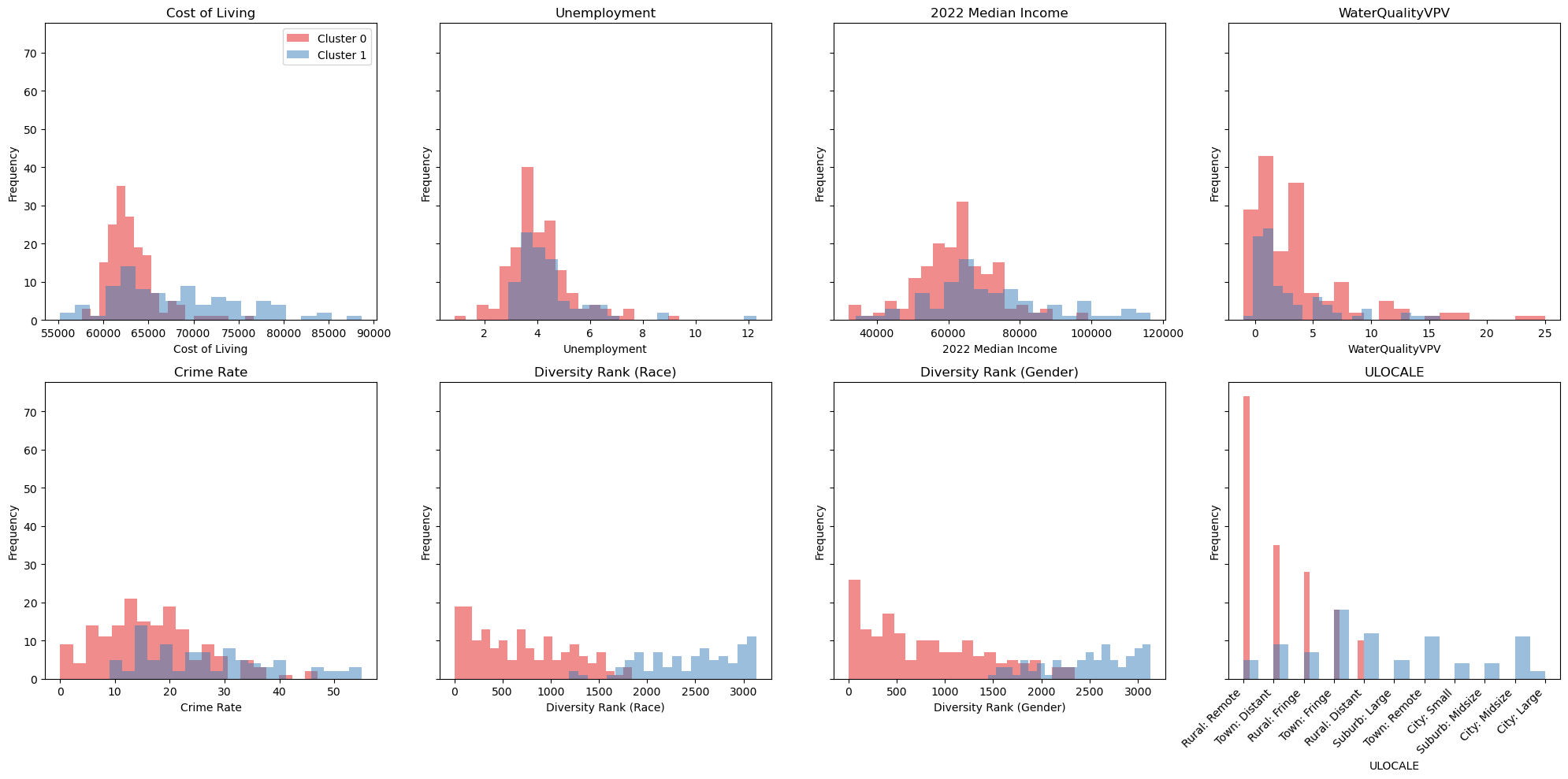
****

Figure 5: Distribution of Quality-of-Life features by Cluster.

**References:**

*All-in-One Quality of Life Analysis: City, ZIP, County, and FIPS Data | Spreadsheet Download |*

*Gigasheet*. (n.d.). Retrieved May 1, 2025, from <https://www.gigasheet.com/sample-data/cityzipcountyfips---quality-of-life-us>

OpenAI. (2023). *ChatGPT* (Mar 14 version) [Large language model].

*Texas Maps & Facts*. (2024, January 18). WorldAtlas.

<https://www.worldatlas.com/maps/united-states/texas>