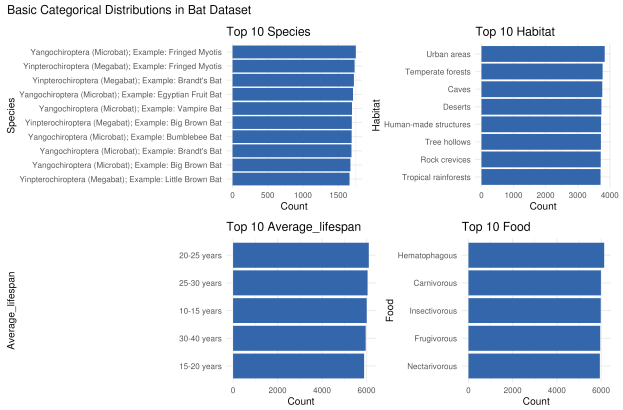
BIOS611Final Project

The **bats genetic adaptation dataset**(https://www.kaggle.com/datasets/jockeroika/bats-data/data ) contains around 30000 genomic and phenotypic records of bats sampled across diverse geographic regions, with variables describing morphology, diet, ecological traits, and genetic variation. After cleaning and inspection, the dataset provides a structured foundation for analyzing how bat species adapt to environmental pressures, particularly through the lens of genetic divergence, ecological niche differences, and phenotypic patterns.

The goal of this project is to use exploratory data analysis and machine learning techniques to uncover signals of genetic adaptation, identify key traits associated with ecological specialization, and visualize how phenotypic and environmental features contribute to adaptive divergence across bat species.

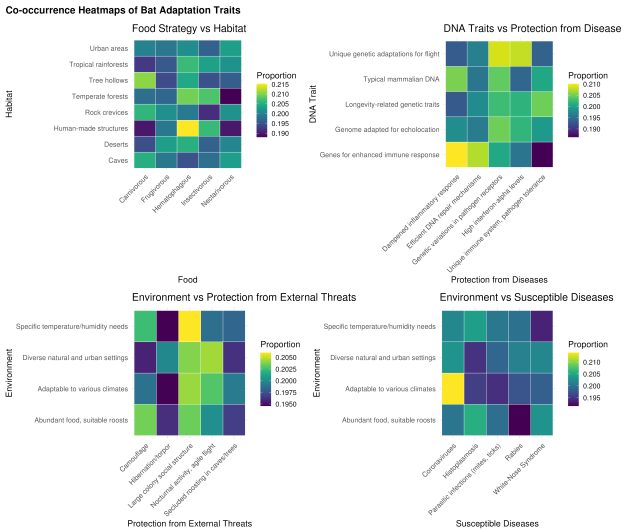
Exploratory Data Analysis

I cleaned and standardized the dataset, extracted four key categorical variables (species, habitat, lifespan group, and diet), calculated their top 10 most frequent categories, and visualized them using horizontal bar plots.



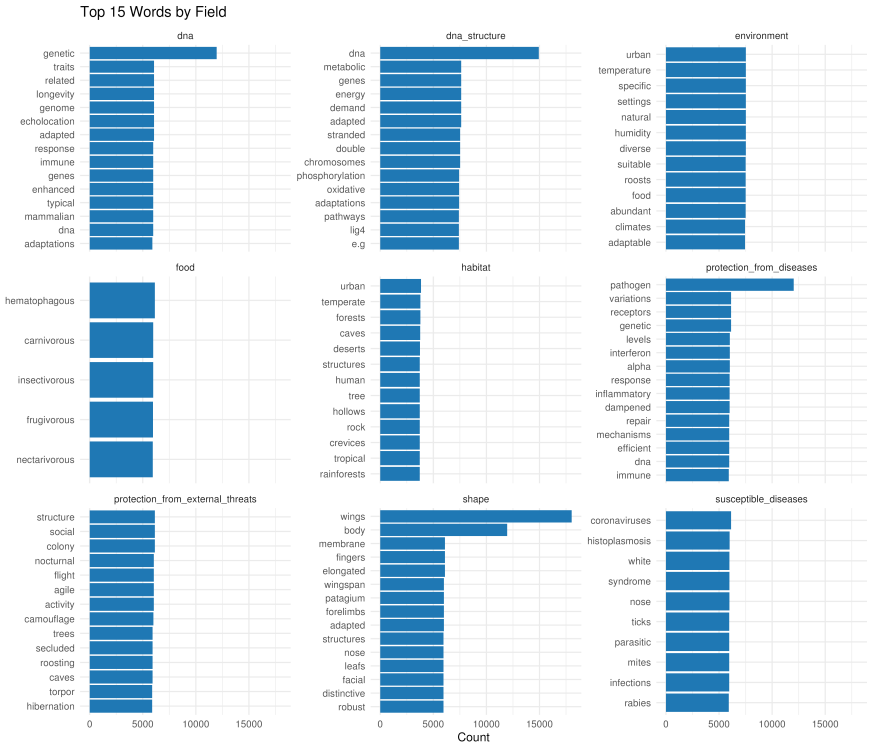

The distributions show that the dataset is dominated by a few recurring species, common habitats such as urban areas and forests, typical lifespans between 10–30 years, and a balanced mix of dietary types. Overall, the dataset contains broad ecological and biological diversity across bats, indicating that it is suitable for studying patterns of adaptation across species, environments, and diets.

I calculated co-occurrence proportions between pairs of categorical traits: food strategy, habitat, DNA traits, disease protection, environment, and disease susceptibility, and then visualized each pair as a heatmap using a consistent color scale to highlight higher or lower co-occurrence patterns.




Across the heatmaps, no extreme hotspots dominate, but consistent moderate co-occurrence patterns suggest that bat ecological traits tend to distribute broadly rather than forming tight, exclusive clusters. Certain habitats and food strategies align slightly more strongly, and DNA traits show modest association with disease-resistance categories, hinting at adaptive links between genetics and survival pressures. Environmental conditions appear evenly distributed across both protective traits and disease susceptibilities, implying that multiple ecological factors jointly shape adaptation rather than any single dominant driver. Overall, these co-occurrence patterns support the idea that bat adaptation is multidimensional, with many small contributions across traits rather than a few strong, isolated relationships.

Since all variables though categorical but are described by text, tokenized text data (unnest_tokens + stopword removal), computed word frequencies for each descriptive field (e.g., dna, habitat, shape) and then counted occurrences of each token within each field, selected the top 15 most frequent words per field and visualized them with faceted bar charts.

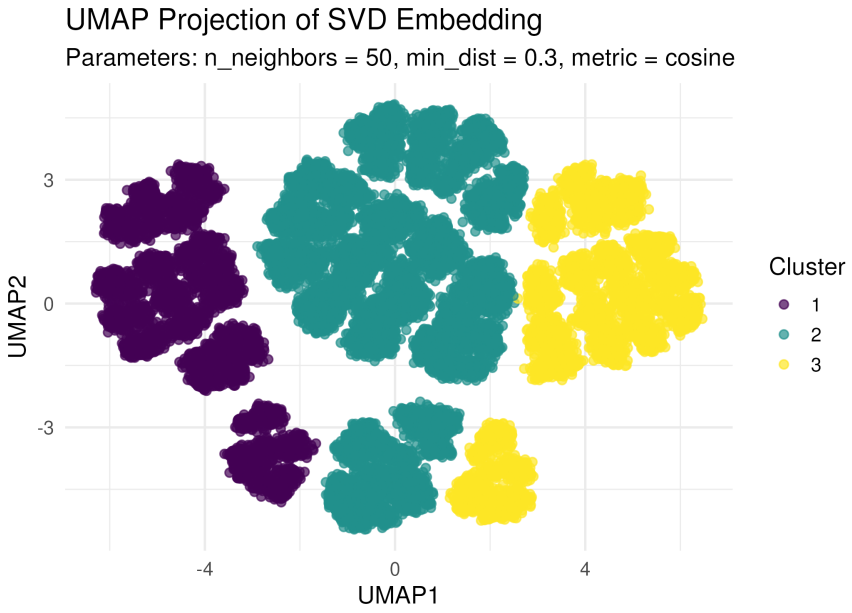


The top-word profiles highlight the characteristic language associated with each trait: DNA-related fields emphasize genetic mechanisms and adaptations, habitat descriptions focus on environmental settings like forests, caves, and urban areas, while feeding and susceptibility fields highlight biological roles such as "hematophagous," "carnivorous," or "coronaviruses." Structural traits consistently reference morphology like wings, membranes, and body features, and protection fields emphasize immune, behavioral, or environmental defenses.

Clustering

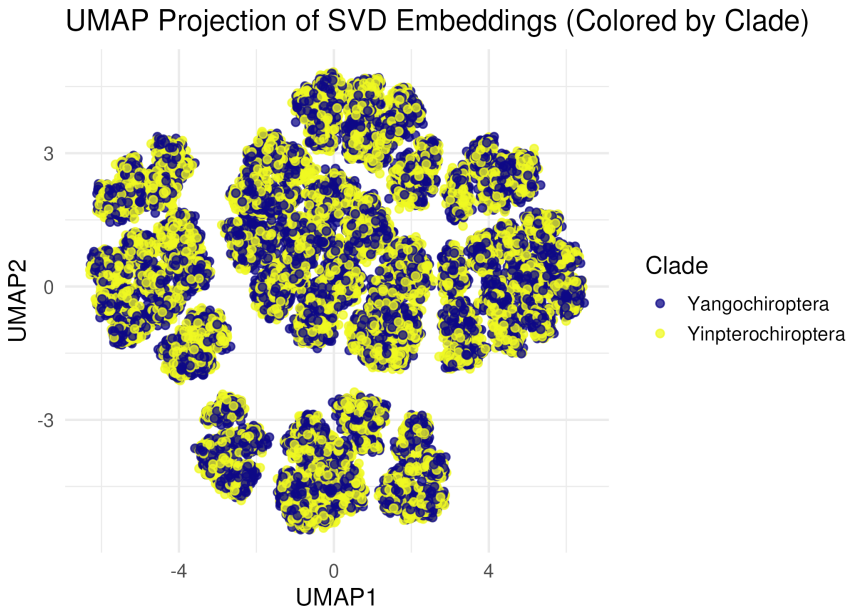
**Method 1: Treat all descriptive variables as text**

I embedded all bat-description text into a document-level **TF–IDF** matrix, reduced dimensionality using **truncated SVD** (30 components), and then **applied UMAP** with n_neighbors = 50, min_dist = 0.3, and metric = "cosine". Finally, **ran k-means clustering** (centers = 3) on the SVD embeddings and visualized the resulting cluster labels on the 2D UMAP projection.




The UMAP projection shows three clearly separable clusters, indicating that the text-based SVD embeddings capture meaningful structure within the bat descriptions. Cluster 1 (purple), Cluster 2 (teal), and Cluster 3 (yellow) each form dense, well-defined groups, suggesting distinct underlying themes or trait combinations present in the textual fields. The strong separation implies that even without explicit biological labels, the combined text information contains enough signal to partition bats into coherent adaptation profiles.

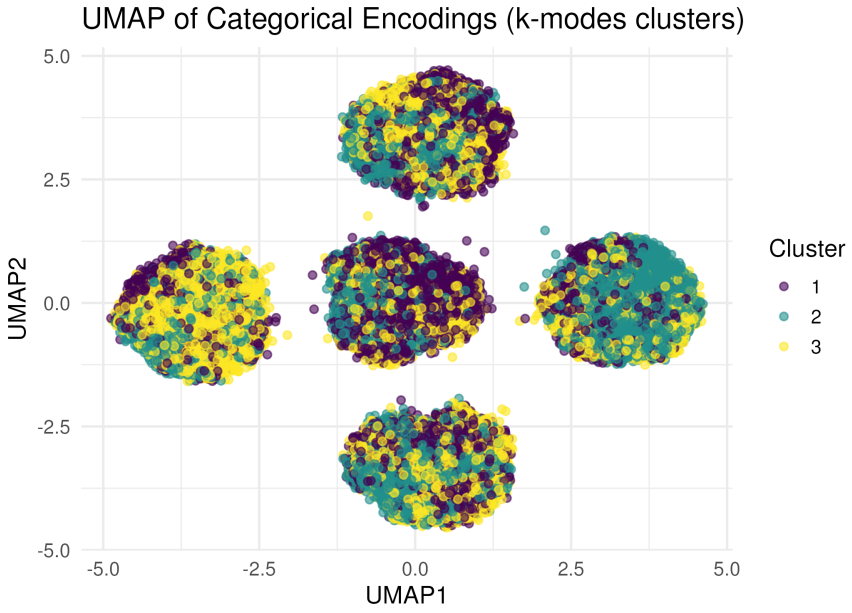
**Drawback.** Since the input text is generated from repeated categorical descriptions rather than natural sentences, the resulting TF–IDF and SVD embeddings capture discrete label patterns rather than smooth semantic structure. As a result, the UMAP projection appear fragmented and shattered, reflecting artificial repetition in the data rather than true continuous variation in biological traits.




The UMAP projection shows extensive overlap between Yangochiroptera and Yinpterochiroptera, indicating that the text-based embeddings do not strongly separate bats by clade. This suggests that the descriptive fields used to construct the TF–IDF/SVD embedding such as habitat, food, environment, and disease traits, do not systematically differ between the two phylogenetic lineages in this dataset. Combined with the earlier observation that UMAP clusters are partly shaped by repeated categorical descriptions rather than natural language structure, the mixed distribution of clades reinforces that the embedding captures ecological and descriptive similarities, not evolutionary history.

**Method 2: k-modes Clustering on Categorical Traits**

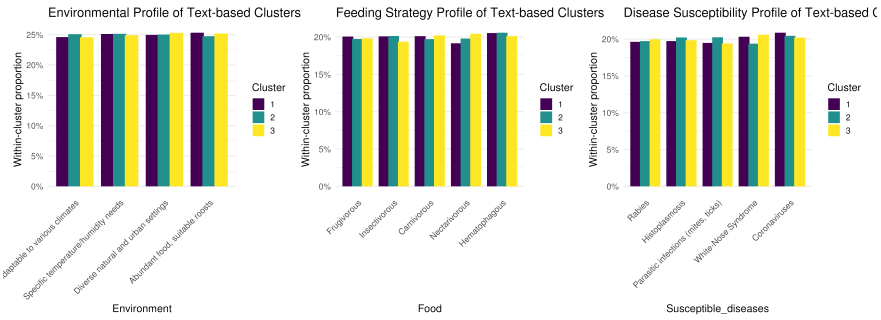
Converted all descriptive columns (habitat, food, environment, diseases, DNA traits, etc.) into categorical factors and applied **k-modes clustering** ($k = 3$), which is designed for purely categorical data. To visualize cluster structure, use one-hot encoded the categorical variables and projected them into 2D using UMAP with a Manhattan metric and then plotted each sample according to its UMAP coordinates, coloring points by the assigned k-modes cluster.




The UMAP projection reveals several compact, island-like clusters, but the k-modes labels are evenly mixed within each island rather than forming clean cluster-specific regions. This suggests that while UMAP separates groups based on combinations of categorical traits, the global k-modes solution does not align neatly with these local structures, likely because the categorical traits do not form strong, globally consistent partitions. Compared with the text-based SVD embedding, the categorical-only approach produces more discrete UMAP islands, highlighting that categorical repetition creates sharp boundaries but may not yield interpretable or biologically meaningful clusters on its own.

Cluster Trait Profiling

Took the k-means clusters derived from the **text-based SVD embeddings** and profiled them against key categorical traits: environment, feeding strategy, and disease susceptibility. For each cluster, computed the **within-cluster proportion** of each category and visualized the results using grouped bar charts .




Across all three profiles, the clusters show very similar proportional distributions, with only minimal variation across environments, diets, and disease susceptibilities. This indicates that the text-based embedding clusters do not strongly correspond to these categorical adaptive traits in the dataset. The high uniformity suggests that the textual descriptions, being repetitive categorical labels rather than rich narratives, do not encode enough distinct biological information to drive differentiated cluster profiles. In short, while the embedding produced visually separable groups, these groups do not map cleanly onto ecological or physiological categories, limiting their interpretability for biological adaptation analysis.

Discussion

Overall, the exploratory analyses show that while the bat dataset is rich in ecological and categorical de-

scriptors, the textual fields, being repetitions of structured labels rather than natural language, limit the ability of TF–IDF, SVD, and UMAP to recover biologically meaningful latent structure. Both text-based and categorical-based embeddings produced visually distinct clusters, but these clusters did not map cleanly onto ecological traits, feeding strategies, disease susceptibilities, or evolutionary clades. This suggests that additional biological signal is needed to uncover deeper adaptation patterns.

**Future work** could incorporate several enhancements: **Use natural-language descriptions** (if available) or augment the dataset with external ecological or genomic metadata to create richer text embeddings. **Model categorical traits explicitly** using probabilistic graphical models or latent class analysis, which may better capture multi-trait interactions. **Integrate continuous traits** (e.g., body size, wingspan) with categorical fields to obtain smoother embedding spaces. These extensions would provide a more robust framework for identifying meaningful adaptation signatures in bats and overcoming the limitations inherent in the current categorical-text representation.