Project Title: Cat Breed Popularity and Trait Analysis Using Web Scraping and APIs
Team Members: Tianyu Wang

1. What problem are you trying to solve?

This project aims to explore the relationship between the characteristics of a cat breed and its popularity. Although different breeds of cats vary greatly in their physiological and behavioral characteristics, public popularity does not always correspond to these characteristics. By studying the relationship between breed traits and their presentation or frequency of appearance online, this project attempts to answer the question: "How do breed traits affect the popularity of a cat breed?" Its goal is to provide a data-driven perspective on understanding the patterns and correlations between cat breed traits and their perceived popularity.

2. How will you collect data and from where?

Data will primarily come from two sources: Wikipedia and TheCatAPI. Wikipedia's list of cat breeds and individual breed pages contain structured tables with data such as country of origin, lifespan, weight, and coat type; these pages will be crawled using Python's requests and BeautifulSoup libraries. TheCatAPI will provide additional quantitative trait data, such as friendliness, energy level, grooming difficulty, intelligence, and friendliness towards children/pets; this data will be obtained through API calls.

3. What analysis will you do?

The analysis will begin by cleaning and merging data scraped from Wikipedia and feature data obtained from TheCatAPI. Specific measures include standardizing lifespan and weight ranges, unifying breed names, handling missing values, and normalizing feature scales. The project will conduct descriptive analyses of features such as coat length distribution, average lifespan, weight range, and the number of breeds in each country. Furthermore, it will explore relationships between features, such as the correlation between weight and lifespan, grooming needs and coat type, and friendliness and energy levels. The project will also use frequency statistics to estimate breed popularity; by studying the number of photos uploaded for each breed in TheCatAPI, the exposure of different breeds can be analyzed.

4. What visualizations will you create?

Visualizations will include bar charts, scatter plots, and heatmaps. Additionally, cluster plots can be created to group dogs based on their overall physiological and behavioral characteristics. All visualizations will be implemented using Matplotlib or Seaborn to ensure clarity and understandability.