SocioEconAdopt: SocioEcon-Aware Dog Adoption Recommender

Milestone 1: Project Proposal

Group Members:

Name	Email	Github Usernames
Lufei Chen	lufeic22@seas.upenn.edu	sc102299
Sophie Shi	xueshi1@seas.upenn.edu	sophieshixue
Xilin Chen	chen0810@seas.upenn.edu	cxl0810
Jier Yin	jieryin@seas.upenn.edu	jieryin

Application Overview:

Build a data-driven web application that recommends dog breeds for adoption based on city and community characteristics, combining both demographic (Census) and dog adoption data. The system will help users find breeds suited to their household and neighborhood profile—factoring in coat, size, child-friendliness, and socioeconomic indicators such as household income, family size, and housing type—while also offering an interactive map that visualizes breed and demographic patterns across the U.S.

Dog Recommendation System
Suggest dog breeds that best fit a user's environment and household context. Recommendations use both the adoption dataset (breed, coat, size, temperament) and local Census variables.

City-Based Breed Explorer
Enable users to search for popular dog breeds in their city or county, and examine how adoption trends relate to demographics.

Demographic and Breed Map Visualization
Interactive U.S. map displaying: The most common dog breeds by county; Demographic and socioeconomic indicators (income, family size, urbanicity); Correlations between local conditions and adoption characteristics; Users can hover over a city or county to see breed frequency, demographic statistics, and adoption insights.

Description of the dataset/website idea

1. **Adoptable Dogs in the US**
   (https://www.kaggle.com/datasets/thedevastator/adoptable-dogs-in-the-us)

   This dataset, sourced from Kaggle, contains detailed information about over 58,000 adoptable dogs across the United States, with 36 features describing each dog's characteristics including each dog's breed, size, age, coat type, and shelter location (city, state, and ZIP code).

   I. Size statistics: 67.29 MB CSV file, 36 attributes, 58180 rows

II. Summary statistics:

Numeric attributes such as ID and index have consistent distributions (mean ≈ 4.4×10^7, std ≈ 3.8×10^6).

For categorical attributes, preliminary analysis shows that: 51.4% of dogs are medium-sized, 27.1% are large, and 19.9% are small; 76.0% have short coats; 64.7% are house-trained; 72.3% are vaccinated (shots_current = True).

These results provide a clear view of the overall distribution of dog characteristics and support future correlation analysis with climate data.

2. **US Census Demographic Data** (https://www.kaggle.com/datasets/muonneutrino/us-census-demographic-data)

   This dataset, sourced from the U.S. Census Bureau (via Kaggle), provides demographic and socioeconomic data for 74,001 census tracts across the United States, with 37 attributes describing population, income, poverty, and employment characteristics. Each row represents one census tract, which typically covers a population of 1,200–8,000 residents.

   I. Size statistics: 14.4 MB CSV file, 37 attributes, 74,001 rows

   II. Summary statistics:

   Key numeric attributes show substantial variation across tracts. The average total population per tract is 4,325 (std = 2,129), with a mean household income of $57,226 (std = $28,663) and an average poverty rate of 17.0% (std = 13.2%). The mean per-capita income is $28,491, while unemployment averages 9.0%. These statistics highlight large regional disparities in socioeconomic conditions across U.S. communities.

   This dataset will be joined with the adoptable dogs dataset through the state identifier, enabling analysis of how income levels, poverty, and employment relate to regional patterns of dog adoption and breed distribution.

Queries:

1. Find the top 5 dog breeds most adoptable in certain counties based on different socio-economic status by joining the two dataset on the same county.
2. Calculate an index for each breed in a given city using attributes of adoptable dogs, and recommend breeds that best match users based on their socioeconomic status and needs.
3. Show adopted dog breeds, average age, and size based on the county/counties that the user is interested in.
4. Find the most common dog size in areas with the top 10% income nationwide by Subquery, ranking, and filtering. Use a subquery to select the top decile of tracts by income, then join with adoption data to find dominant size types within those regions.
5. Compute the number of dogs adopted per 1 000 households in each county and rank counties by this income-normalized adoption rate.
6. Analyze whether larger average family sizes correspond to adoption of larger dog breeds by computing correlation between avg_family_size (from Census) and avg_dog_size_index (from Dogs).