This project is divided into four stages:

Data Cleaning: Initially, I cleaned the data by creating 'avg\_weight' and 'avg\_height' columns using the minimum and maximum values provided for them in Excel. Then, I subsetted the dataset, retaining only the necessary columns. Further, I converted the character variables to factors, removed missing values, and merged other CSV files obtained by scraping data from the AKC website. However, this process resulted in missing values, approximately 24% of the breeds missing ranks. Consequently, I removed these entries, considering that it would not significantly impact our study. Notably, the popularity ranking is based on registration statistics from the AKC. It's important to note that the AKC's breed rankings are determined by registration statistics, not by factors like entertainment value, intelligence, or appearance.

Data Exploration and Feature Engineering: In this stage, I utilized various visualizations to gain insights into the data. Histograms were employed to understand the distribution of weight and height of breeds, based on which I decided on cut-off points to create a new column named 'size' categorizing breeds into small, medium, and large. Additionally, I created another variable named "score" by assigning weights to various attributes in the dataset, drawing from personal expertise (a potential source of bias). These newly created columns were added to the CSV file. Outliers were identified using z-scores.

Further Exploration: I delved deeper into the data using additional visualizations. For instance, I created a data frame with only three variables and visualized the data to gain further insights. I experimented with scatterplots to understand the correlations/relationships between certain variables. Notably, a strong correlation between weight and lifespan was observed. I also explored line graphs and bar plots to glean additional insights.

K-means Clustering: In this stage, I applied the k-means algorithm to cluster the dogs into four categories. Since I had predefined categories in mind, I opted for k-means clustering. The dogs were clustered into categories labeled as 'hotdogs', 'overlooked treasures', 'rightly ignored', and 'overrated'. I then used ggplot to visualize this clustering.