Exploring the relationship between cat popularity and breed characteristics using public data sources

Name: Tianyu Wang

Course: DSCI 510

Date: *2025/12/15*

Abstract

This study investigated whether measurable cat breed characteristics are related to breed popularity. Collect data from Wikipedia through web scraping, Wikimedia page browsing API, and TheCatAPI. After cleaning and integrating the dataset, exploratory analysis, correlation analysis, and clustering were conducted. The results indicate a weak correlation between the popularity of breeds and personal characteristics such as intelligence or grooming requirements, suggesting that the public interest is influenced by broader social and contextual factors. This project showcases a Python project that combines heterogeneous public data sources to solve research problems.

1. Introduction

Cat breeds vary greatly in appearance and behavioral characteristics, but some breeds have always received public attention. Understanding whether the popularity of a breed is related to its inherent characteristics can provide insights into human preferences and information seeking behavior. This project addresses the following research question: Is the breed characteristics of cats related to the popularity of breeds measured by public interest? To answer this question, the project integrated network scanning data and API based data to analyze the relationship between popularity and breed characteristics.

2. Data and Methods

2.1 Data Sources

Three data sources were used. Firstly, Wikipedia was crawled to obtain a list of cat breeds and basic attributes such as origin, body size, and fur type. Secondly, the Wikimedia Pageviews API is used to measure the popularity of the breed, summarizing the page views of recent months as a proxy for public interest. Thirdly, TheCatAPI provides standardized digital features, including intelligence, energy level, affection level, grooming requirements, lifespan, and weight。

2.2 Data Cleaning and Integration

The breed names have been standardized to achieve reliable merging across sources. Footnotes and citation marks on Wikipedia have been removed, and range-based variables are converted to numerical values using midpoint estimation. Breeds with missing key attributes are excluded. The final dataset contains approximately 50-60 breeds with over 15 features.

2.3 Analytical Approach

Exploratory analysis is used to determine the most popular breeds. Pearson correlation analysis examined the relationship between popularity and breed characteristics. Apply K-means clustering to standardized behavioral and physical features to identify populations of breeds with similar characteristics.

3. Results

3.1 Breed Popularity

According to the ranking of breeds based on page views on Wikipedia, well-known breeds such as Maine Coon, Persian, and ragdoll are the most popular. An extreme outlier of "Cyprus" shows abnormally high page views due to ambiguity between breed names and country names on Wikipedia, indicating the limitations of popularity metrics based on page views.

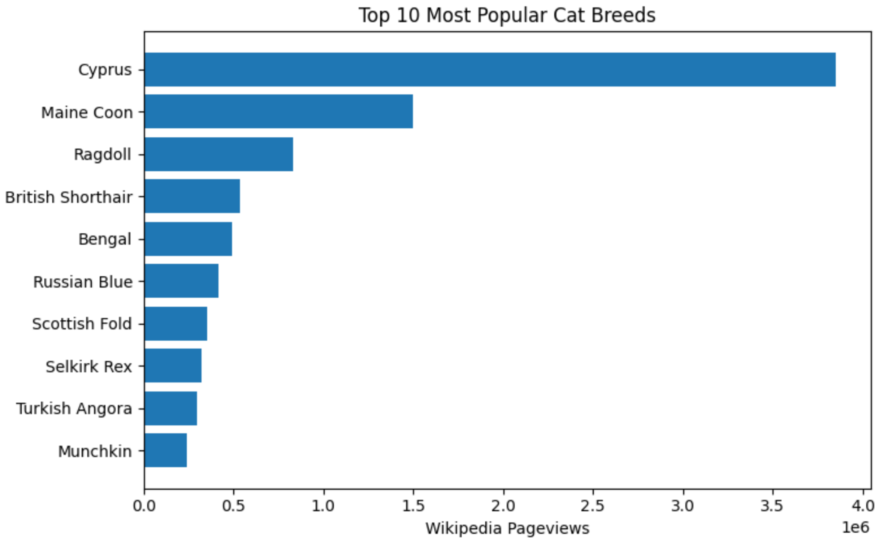

Figure.1

3.2 Trait–Popularity Relationships

Scatter plot and correlation analysis indicate that the correlation between popularity and individual characteristics is weak. The correlation between intelligence and affection level and page views is small, while grooming requirements show a slightly positive but weak relationship. These findings indicate that no single breed characteristic can strongly predict popularity.

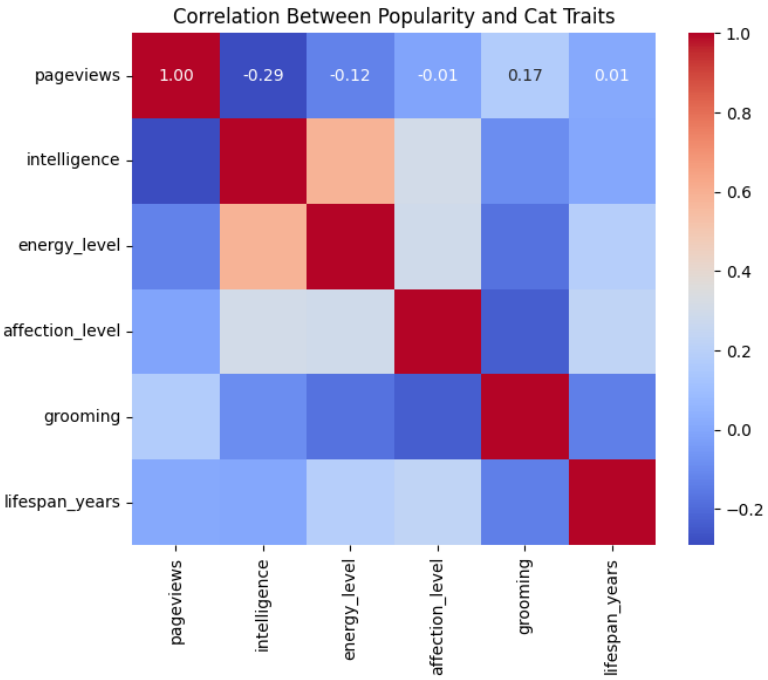

Figure. 2

3.3 Clustering Analysis

K-means clustering identified three different breeds based on behavioral and physical characteristics. These clusters reflect the balance between energy levels and grooming requirements, highlighting groups that require moderate maintenance such as high-energy and low-energy breeds, as well as groups with the least grooming needs. These clusters provide another perspective on breed diversity beyond popularity rankings.

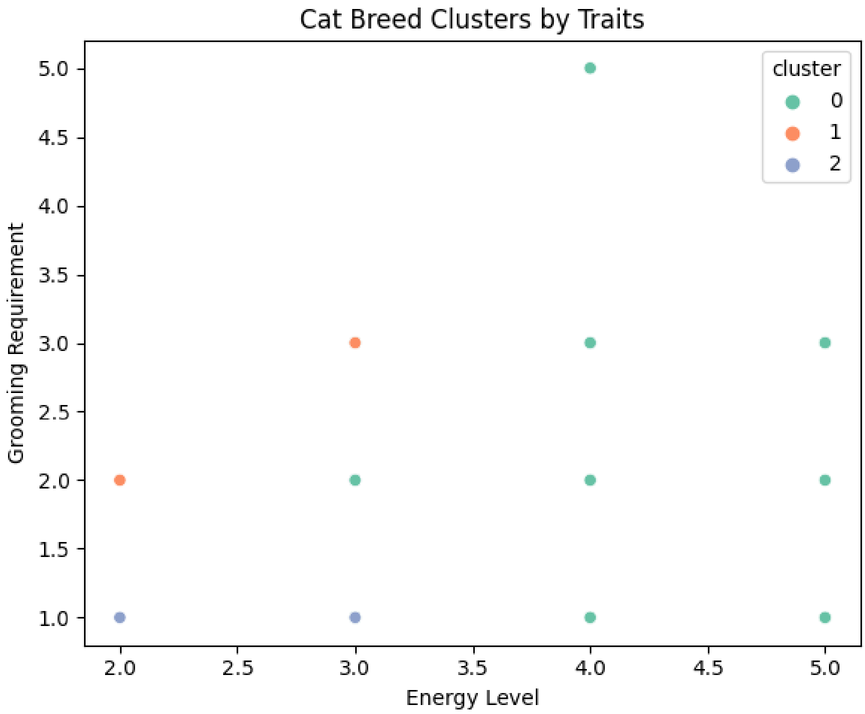

Figure.3

4. Discussion

The results indicate that the popularity of cat breeds is not strongly driven by individual measurable characteristics. On the contrary, popularity seems to be influenced by broader background factors, including historical recognition, media exposure, and naming ambiguity. The presence of outliers such as "Cyprus" highlights the importance of validating popularity proxies when integrating heterogeneous data sources. Overall, the research findings indicate that behavior or physical characteristics alone cannot fully explain public interest in cat breeds.

5. Limitations and Future Work

This study relies on the page views of Wikipedia as a proxy for popularity, which reflects attention rather than actual ownership or adoption rates. In addition, unclear naming may increase the page views of certain breeds, and the sample size is limited to the breeds available in all data sources. Future work may incorporate other popular indicators, such as Google Trends or the use of statistical data, as well as temporal or geographic analysis of product popularity.

6. Conclusion

This study relies on the page views of Wikipedia as a proxy for popularity, which reflects attention rather than actual ownership or adoption rates. In addition, unclear naming may increase the page views of certain breeds, and the sample size is limited to the breeds available in all data sources. Future work may incorporate other popular indicators, such as the use of Google Trends or statistical data, as well as temporal or geographic analysis of product popularity.