Kaggle Data Sets

Kaggle is a data science community allowing users to share data and enter competitions to solve data science challenges. Analyzing data set usage by download counts and other meta-data provides insights on the most popular areas of study.

Audience

Data Sets from Kaggle are used by Data Scientists & Data Analysts to solve data science challenges. This poster provides such users with insights into the most popular areas of study to aid in the selection of new challenges.

Data

The Kaggle data analyzed has 11 rows and 18518 columns providing metadata about the data sets on kaggle as of June 2016. Aggregations of the number of data sets associated with each keyword were performed for the keyword data only, removing any records without keywords associated.

Questions

What are the 10 most downloaded data sets & what keywords do they use? What keywords are used the most across all data sets?

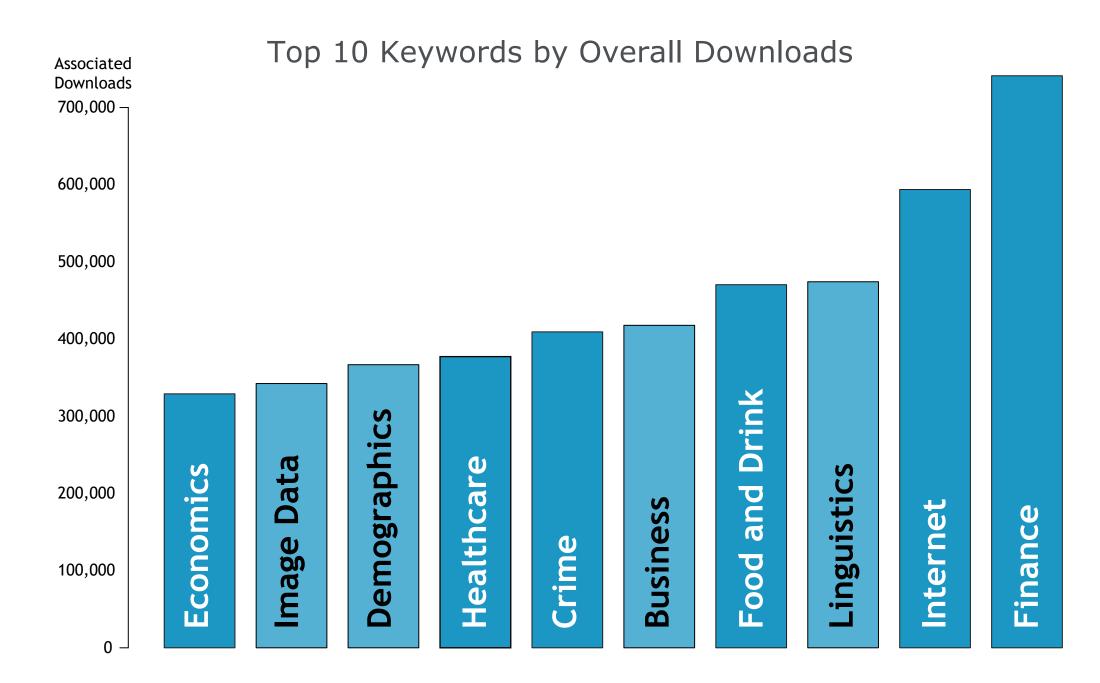
Which keywords are associated with the most overall downloads?

Is there a correlation between downloads and data set size, kernels, likes or views?

Keywords

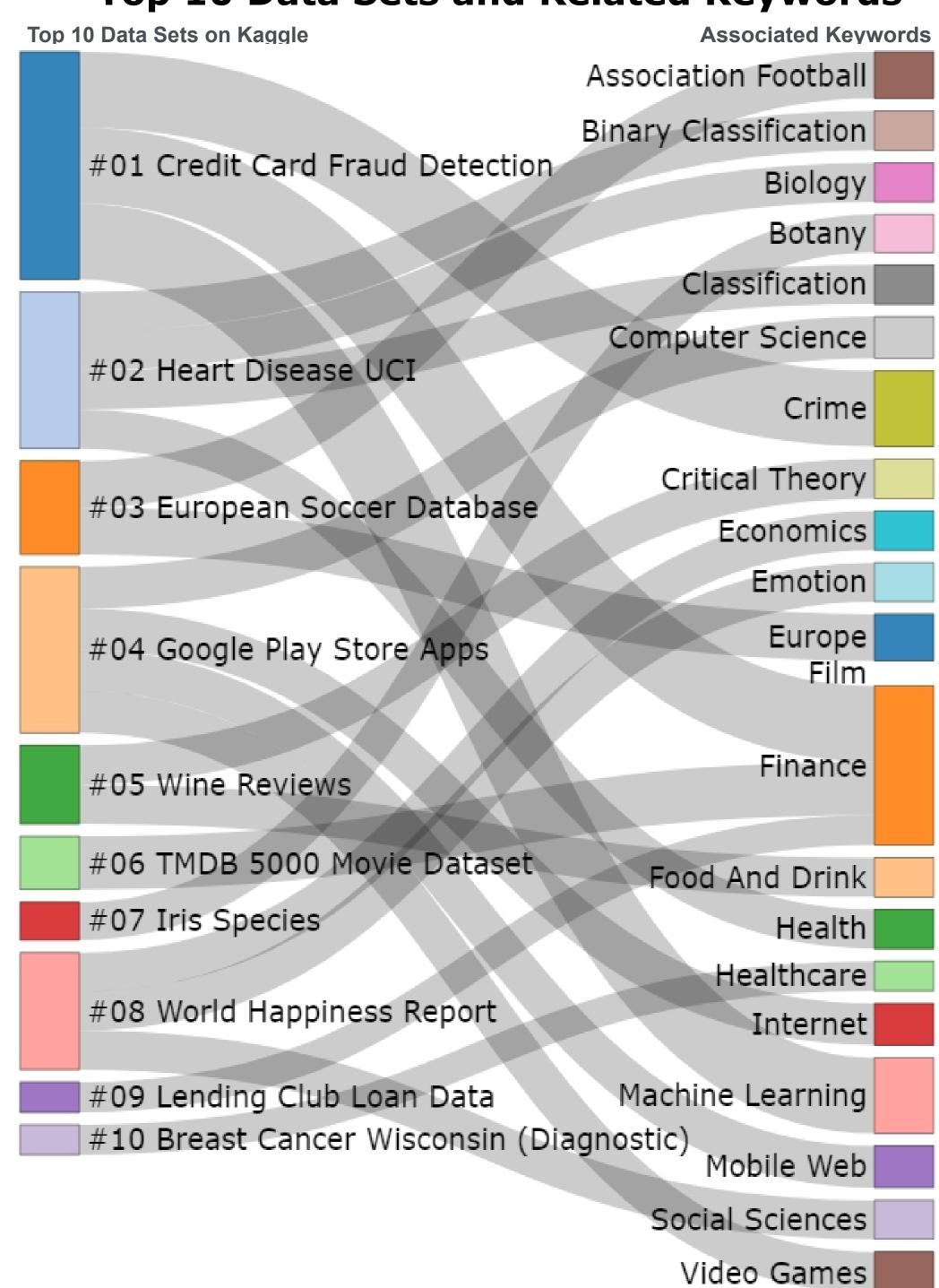


Socrata is the most frequently used keyword, associated with 923 data sets. The next most used keyword is **Internet** associated with only 424.



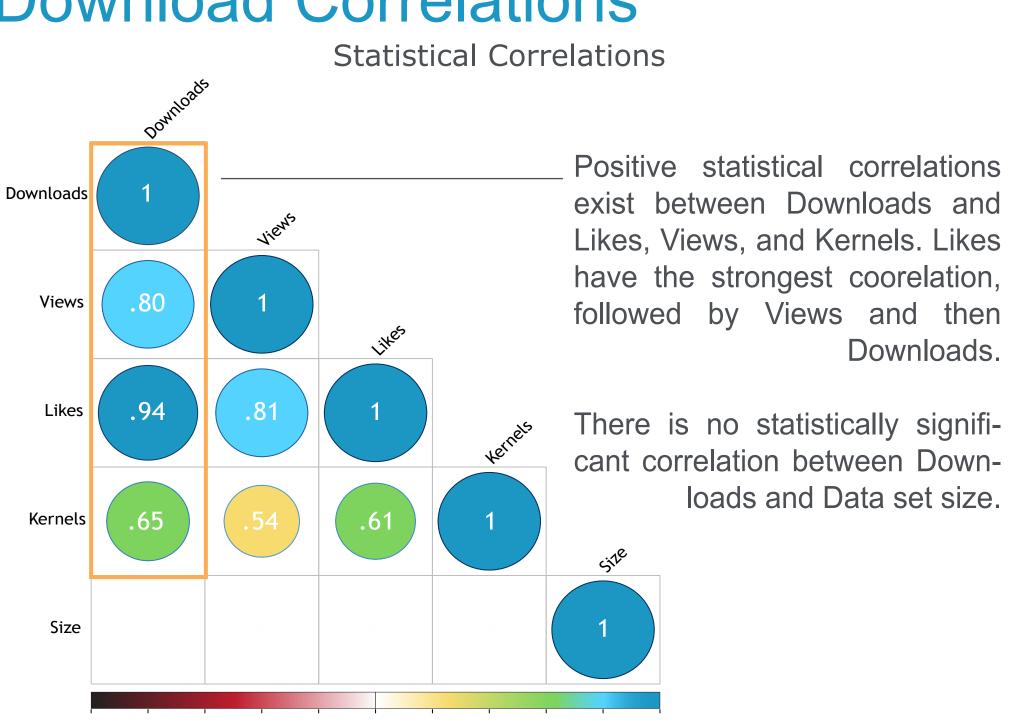
Finance, Internet, Food and Drink, Crime, Healthcare & Economics are top 10 keywords also associated with the top 10 data sets and Finance is the keyword associated with the most total downloads across all data sets.

Top 10 Data Sets and Related Keywords

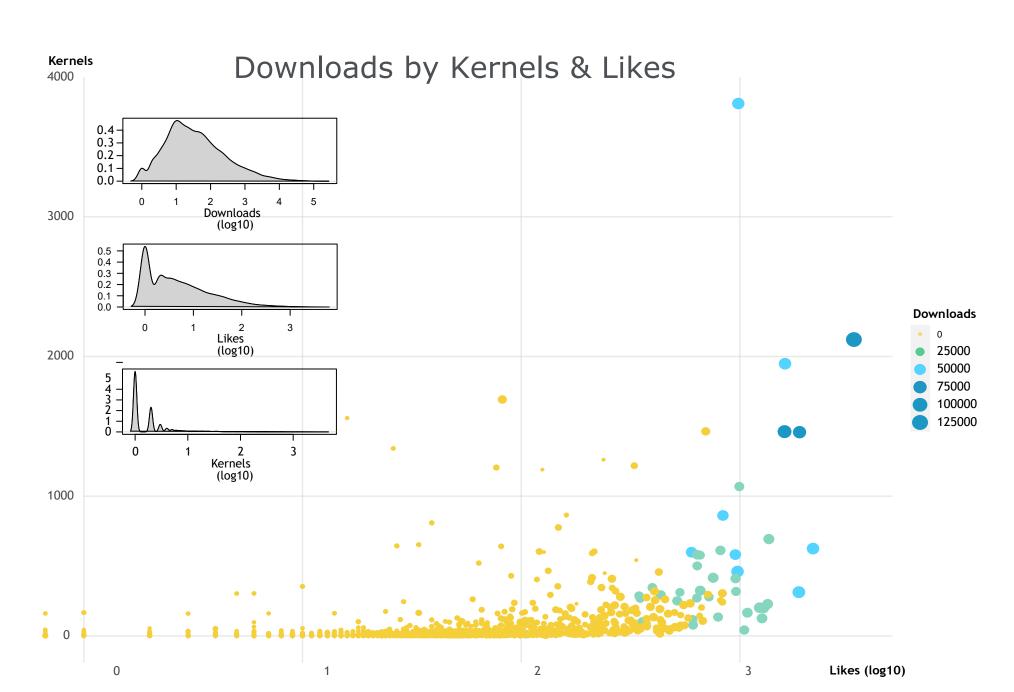


The most downloaded data set is **Credit Card Fraud Detection** which uses the keywords **Finance**, **Crime**, and **Machine Learning**.

Download Correlations



Author: Andrea Bradshaw
Class: IST 719 Information Visualization



Visualizing the statistical correlations the shows the relationship where as Likes increase so do Kernels and Downloads.