For this project, I will use a dataset from tcgcsv.com, which provides publicly accessible CSV files that can be downloaded directly or retrieved via requests. The data originates from TCGplayer, the largest online marketplace for trading card games and the industry standard for secondary market pricing. The data consists of information about pokemon TCG product listings on the platform, with observations consisting of data for one unique card per day. The dataset includes approximately 24 million rows and 31 columns, making it large enough to satisfy the project requirements. It contains both numerical variables (low_price, mid_price, high_price, market_price, hp) and categorical variables (rarity, card_type, stage, set_name, abbreviation), enabling a wide range of machine learning applications. A lot of feature engineering will be necessary to remove extraneous features and make other features more useful.

Regression Analysis

The regression task will focus on predicting market_price, the best overall measure of a card's value. Independent variables will include rarity, clean_name, release_date, hp, stage, attack details, and set_name. Feature engineering will be necessary to handle text-based features such as attacks, extracting numeric damage values and indicators for special effects. The release_date feature can be transformed into card age, which may correlate with scarcity. Categorical fields such as rarity, card_type, and set_name will require encoding before they can be used in modeling. These steps ensure the regression analysis accounts for both numerical and categorical features in a structured way.

Clustering and Classification

The classification task will focus on predicting rarity. This is a multi-class problem with categories such as Common, Uncommon, Holo, Illustration Rare, Ultra Rare, and more. Independent variables will include pricing features, hp, stage, card_type, and release_date. Feature engineering will help quantify attributes like attacks, while encoding will be needed for categorical variables such as set_name and card_type.

In addition to classification, clustering methods will be applied to identify groups of cards with similar attributes and pricing patterns. This unsupervised analysis can highlight natural groupings that may not align exactly with rarity labels, revealing hidden structure in how card attributes and values are distributed.

Association Rule Mining

Association rule mining will be applied with market_price as the independent variable and categorical fields such as rarity, card_type, set_name, stage, and release year as the dependent variable. This will identify patterns between variables which consistently alter market value. These rules are directly interpretable and provide insights into the relationships between card features and their broader distributions. In the market, there are typically patterns for certain characters from the series to have more valuable cards, so I predict that rule mining will generate some clear and interesting insights.

Applications and Relevance

Although the topic is trading cards, the dataset has strong parallels to alternative asset markets. Pokémon cards derive value from rarity, intrinsic features, and fluctuating demand, similar to other collectibles, art, or other non-traditional assets. Regression for price prediction, classification for category assignment, clustering for segmentation, and association rule mining for feature interactions are directly comparable to techniques applied in financial analytics and market research. This makes the dataset both academically suitable and practically relevant, providing insights that extend beyond the specific domain of trading cards.