# Wrangle Report on Twitter Account WeRateDogs

The goal of analyzing the Twitter account WeRateDogs was to generate exciting insights and visuals regarding the performance of the account's tweets. WeRateDogs is a Twitter account that portrays the dogs of dog owners in a humorous way. What makes WeRateDogs special is the rating system, with a denominator of 10 and a nominator of at least 10.

The first step of the analysis was to collect the data. Data from three different data sources were used for a comprehensive analysis. First, a WeRateDogs Twitter Archive was accessed, which was available as a csv file and included over 2350 tweets. Second, the results of a neural network for automatic dog breed recognition were downloaded programmatically. Here, the data set amounted to circa 2075 tweets. Third, other relevant information such as the number of favorites and retweets counts were requested using the Twitter API and saved as a JSON file. The Twitter ID from the CSV file served as the basis of the query. All three data sources were saved as a separate data frame for further analysis.

The second step was to assess the data from the three data frames both visually and programmatically to identify quality issues and tidiness issues. For all three data frames, a total of 18 quality problems and 3 tidiness problems were listed. Quality issues included missing data, duplicate entries, incorrect data types, incorrect data, or incorrect formatting. Tidiness problems included multiple information being represented in one column or the same information being spread across multiple columns.

The third step consisted of cleaning the data. For this purpose, copies of the three existing data frames were first created. Then, first, columns and rows with missing or duplicate data were removed. Second, the tidiness issues were targeted and each information was assigned to a column. In the third step, the quality problems were addressed.

In the fourth step, the data frames that were available as CSV files or requested via the Twitter API were merged into a new data frame based on the Twitter ID and saved as a CSV file.This new data frame was then merged with the programmatically downloaded image data frames, also based on the Twitter ID, and again saved as a separate CSV file.

The reason for the split was the analysis performed in the last step. The first merged data frame (CSV and API), which has more entries than the second merged data frame (image predictions), was used to determine meaningful average values of favorites, retweets, and ratings and to display them visually as bar charts. The second merged data frame (image predictions) has fewer entries, as only those tweets in which the neural network detected a dog were considered. These results were used to determine meaningful data regarding the most recognized dog breeds and display them as a chart.