REPORT: DATA WRANGLING EFFORTS

Introduction:

The purpose of this report is to provide an overview of the data wrangling efforts undertaken to gather, assess, and clean the WeRateDogs Twitter data. The objective was to transform the initial raw data into a cleaned and structured format that could be used to generate meaningful insights and visualizations.

Data Sources:

The primary data source for this project was the Enhanced Twitter Archive, which contained basic tweet data for WeRateDogs. However, it lacked comprehensive information. To enhance the dataset, two additional sources were utilized. The first was the Twitter API, accessed through the tweepy library, which provided data on retweet counts and favorite counts for each tweet. The second source was the Image Predictions file, which included image predictions for the tweets, such as the breed of the dog depicted in the image.

Data Wrangling Process:

1. Gathering:

The Enhanced Twitter Archive was obtained by directly downloading the provided csv file and loading it into a dataframe named 'twitter_archive_df'. The Image Predictions file was downloaded using the requests.get() method and saved as a TSV file. It was then loaded into a dataframe named 'image_predictions_df'. Unfortunately, access to the Twitter API was not available, so a file provided by Udacity, containing retweet counts and favorite counts, was used to create a dataframe named 'api_data_df'.

2. Assessing:

During the assessment phase, both visual and programmatic techniques were employed to identify data quality and tidiness issues. A comprehensive assessment was conducted on all three data sources, including the archive, image predictions, and API data.

3. Cleaning:

A total of seven quality issues and two tidiness issues were identified during the assessment. These issues were addressed using various cleaning techniques, such as dropping unnecessary columns, renaming columns, correcting data types, and resolving inconsistencies. The cleaning process resulted in a clean and well-structured dataset.

Insights and Visualizations:

To gain insights from the cleaned dataset, various data visualization libraries, including Seaborn and Matplotlib, were utilized. Key visualizations, such as scatter plots, bar charts, and regression lines, were created to showcase the relationships between variables and uncover interesting patterns and trends. These visualizations allowed for a better understanding of factors such as favorite counts, top dog names, and dog stages.

Conclusion:

The data wrangling efforts involved gathering, assessing, and cleaning the WeRateDogs Twitter data to create a high-quality dataset. The cleaning process addressed quality and tidiness issues, resulting in a clean and structured dataset suitable for analysis. Visualizations were then generated to present meaningful insights and patterns within the data. The resulting dataset and

visualizations can be used to draw conclusions and make informed decisions regarding dog ratings, audience engagement, and popular dog breeds.