

15/7/2024

act report

Wrangle_act



Data Wrangling and Analysis Report

Introduction

This report presents the insights and visualizations derived from the wrangled Twitter data. The data includes tweets from the WeRateDogs Twitter account, which were cleaned and analyzed to uncover interesting patterns and trends.

Data Cleaning Process

1. **Removing Retweets and Replies:** Retweets and replies were excluded to focus on original tweets.
2. **Filtering Tweets with Images:** Only tweets with images were retained for analysis.
3. **Converting Timestamps:** The 'timestamp' column was converted to datetime format.
4. **Correcting Dog Names:** Erroneous dog names were corrected by ensuring they are alphabetic.
5. **Combining Dog Stages:** Dog stages (doggo, floofer, pupper, puppo) were combined into a single column.
6. **Merging Datasets:** The cleaned Twitter archive, image predictions, and tweet JSON data were merged into a master DataFrame.
7. **Handling Missing Values:** Rows with missing values in critical columns were removed.
8. **Combining Rating Columns:** The rating numerator and denominator were combined into a single 'rating' column.
9. **Dropping Unnecessary Columns:** Columns not needed for analysis were dropped.

Insights and Visualizations

10. **Distribution of Dog Ratings:** Distribution of Dog Ratings The histogram shows the distribution of dog ratings, with most ratings clustering around the higher end, indicating generally positive ratings.
11. **Most Common Dog Names:** Top 10 Most Common Dog Names The bar plot highlights the top 10 most common dog names, with "Charlie" and "Cooper" being the most frequent.
12. **Relationship Between Retweets and Favorites:** Relationship Between Retweets and Favorites The scatter plot illustrates a positive correlation between retweet counts and favorite counts, suggesting that popular tweets tend to receive both retweets and likes.

Conclusion

The data wrangling process was essential in transforming raw Twitter data into a clean and analyzable format. The insights gained from the analysis provide a deeper understanding of the WeRateDogs Twitter account's content and audience engagement.

"Data is a precious thing and will last longer than the systems themselves."

Tim Berners-Lee