WERATEDOGS WRANGLED DATA REPORT BY AKIN YEKE EMMANUEL TOLANI

❖ INTRODUCTION

According to Wikipedia, WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRate Dogs asks people to send photos of their dogs, and then tweets selected photos rating and a humorous comment. Dogs are rated on a scale of one to ten, but are invariably given ratings in excess of the maximum, such as "13/10". Popular posts are re-posted on Instagram and Facebook.

❖ PROJECT TITLE

Wrangle And Analyze Weratedogs Twitter Data

Wrangled Report

The goal of this project is to wrangle and and analyze WeRateDogs Twitter data in order to create interesting and trustworthy analyses and visualizations.
The following steps were duly followed in order to achieve the main objective.

Step 1: Gathering data

The WeRateDogs Twitter dataset consist of 3 major files. As instructed, I downloaded and uploaded the Twitter Enhanced Archive file (twitter_archive_enhanced.csv) which contains at least 5000 tweets, the tweet image prediction file (image_predictions.tsv) and the tweet_json.txt file that was provided by my instructor on Udacity. It is however imperative to note that I couldn't query the each tweet's retweet count and favorite ("like") count from via the twitter api channel. I therefore used the json file provided as an alternative.

Step 2: Assessing data

After previewing my dataset using the required codes, I checked the tweet archive enhanced column variable definitions in order to understand what the dataset contains. This however gave me a basic understanding of the type of data that's expected to be in each column. Upon inspection, assessment(programmatic and visual) the following Qualities and Tidiness issues was discovered

Quality Issues

1. Missing tweet_id in for so many column (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, etc)

2. Wrong data type for tweet_id and date column, in twitter archive enhanced dataframe and tweet_id in tweet Image dataframe

3. Erroneous data present at the dog names column (a, an)(twitter archive dataframe), some starting with small letter

4. Erroneous data present at the dog names column (multiple dog stage exist for one row of data)

5. Erroneous column names for some data

6. Unnecessary html url tags in the source column

7. The types of dogs found in the columns p1, p2, and p3 has some uppercase and lowercase letters

8. The text column contains unwanted elements

Tidyness Issues

1. The 4 Columns that depicts the dog stages (doggo, floofer, pupper and puppo) ought to be in a column in order to simplify analysis

2. The rating numerator and denominator column title too long making the colum bogus

3. The 3 dataset has similar column which can be merged to improve data analysis

4. Remove unwanted columns in the dataframe

Step 3: Cleaning data

In order to clean the data, I ensured that I created a new copy of the dataset I was going to work with. While cleaning the data, I used the define-code-test framework and I also carried out proper documentation using comments and visualization tools

Step 4: Storing data

As instructed, after successfully cleaning my dataset, I saved the cleaned dataset with the name 'twitter_archive_master.csv' on my computer.

Step 5: Analyzing, and visualizing data

As instructed, I carried out 6 insighful analysis and visualizations on my wrangle dataset. The insight includes

1. The Most used twitter source for werate dogs post

2. Timeline of the number of tweets posted by WeRateDogs

3. WeRateDogs Dog Tweet and Rating Analysis(analysis such as Top 10 Most Common Dog Names, Dog Average retweet and favorite tweet counts, Retweeting and Favoriting trend over time and the Dog rating statists that's greater than 10)

Step 6: Reporting

I have created a minimum of 600 words written
report called wrangle_report.pdf. This report shows my wrangling effort and the procedures that I used in archiving my project goals. I also created and saved saved another report called wrangle_act.pfg this file contains and communicates all the insights, displays the visualization(s) produced from my

wrangled data. The references to the document has also been in the codes contained in the file.