Wrangle_report

INTRODUCTION

I wrangled and analyzed the tweet archive of WeRateDogs, better known as @dog rates on Twitter. WeRateDogs is a Twitter account that assesses people's dogs by making funny remarks about them, giving the dog a rating that is almost always more than 10, which is their rating denominator.

Brent is the name given to dogs who receive a rating of 10 or above.

IMPORTING PYTHON LIBRARIES

I began by importing the Python libraries and packages that would be needed for this research.

PANDAS
NUMPY
MATPLOTLIP
REQUEST
TWEEPY
JSON
SEABORN

GATHERING THE DATA

The data for this project came in three different formats:

1. Twitter Archive File from WeRateDogs: WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

2. Image Prediction File: The photographs in the WeRateDogs Twitter archive (the first dataset above) were processed by a neural network capable of classifying dog breeds. This file contains the picture forecasts. This TSV file was hosted on Udacity's server and was downloaded programmatically using the Python Requests library's URL.

3. Tweeter API JSON File: Using the tweet IDs from the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data with

Python's Tweepy package and saved each tweet's whole set of JSON data in a file named tweet json.txt. The JSON data for each tweet was written to its own line. Then I read the.txt file into a pandas DataFrame line by line.

ASSESSING THE DATA

The three data points have been collected and correctly analyzed. I looked for quality and tidiness concerns during the assessment.

Quality: Low-quality data is also known as filthy data. The substance of dirty data is problematic. Completeness, Validity, Accuracy, and Consistency are the Data Quality Dimensions.

Untidiness: Untidiness is sometimes referred to as "messy" data. Messy data has structural difficulties.

1. Each variable becomes a column in tidy data.
2. Each observation is arranged in a row.
3. A table is formed by each sort of observational unit.

After physically inspecting the data in DataFrames and Excel spreadsheets, as well as programmatically inspecting the three DataFrames separately, I discovered eight quality concerns and two tidiness issues in total, which I recorded in my Python notebook.

CLEANING THE DATA

During the assessment step, I cleaned up all of the flaws I documented. It's important to understand that cleaning the data does not imply changing its format. It simply implies increasing the data's quality so that it may be used. To increase its quality and tidiness, the data is cleansed.

I cleaned the data using the Define, Code, and Test cleaning steps.

Define: the concerns discovered during the evaluation are transformed into cleaning chores.

Cleaning tasks are turned into code and then executed.

To ensure that my cleaning efforts were successful, I employed codes.

Keeping the Data

After cleaning the data, I aggregated the three cleaned datasets using a common attribute, the Twitter id, and saved the master dataset as 'twitter archive master.csv'.

ANALYSIS AND VISUALIZATION

The data has been collected, evaluated, and cleaned, and it is now available for analysis. I asked the data several questions in order to gain insights from it.

According to the distribution of dog photos, the most common dog stage is 'pupper' (a little doggo, generally younger), followed by 'doggo' and 'puppo.' So, according to our study, the majority of the image forecasts were at the pupper stage. It might be because a young and immature dog is typically cuter than an adult dog, which is why most people buy, adopt, and own them.

We can observe that the pupper dog stage has the highest amount of retweet count making it the people favorite doggo stage.