Wrangle report

Introduction

With this re report I'll walk through the steps of data wrangling I did in the project .

Step 1: Gathering data

I gathered data from many files in this step

the first was Twitter archive data, which was a csv file, followed by (image predictions.tsv), which i downloaded programmatically, and finally, tweet-json file.

Step 2: Assessing data

This step was split into two sections.

the first was a visual assessment, and the second was a programmatic assessment

and the results are :

Quality Issues

Twitter_archive_data

- There are unneeded columns
- The timestamp , the rating_numerator , tweet_id columns has incorrect datatypes
- The rating_numerator and rating_denominator columns has invalid values
- Incorrect names in name column
- Column expanded_urls has null values

Image_predictions_data

- There are unneeded columns
- tweet_id has incorrect datatype
- The columns p1 , p2 , p3 values some are capitalized and some are not
- The columns p1, p2, p3, p1_conf, p2_conf, p3_conf, p1_dog, p2_dog and p3_dog dont have efficient names

Tweet_json_data

- id column has incorrect datatype
- The 'id' column sholud be named tweet_id

Tidness

- In Twitter_archive_data, the doggo, floofer, pupper and puppo columns shold be in one column called 'dog_stage'
- The dataframes should be merged in one dataframe.

Step 3: Cleaning data

All of the data assessment results were cleaned and tested using a variety of techniques, including (drop , rename , fillna , merge..) after the dataframes were cleaned, they were merged and saved in a csv file called twitter archive master.csv/