Reporting: wragle_report

Introduction

In order to record the project's data wrangling , this wrangle report is a component of the Wrangle and Analyze Data project .The tweet history of Twitter user@dog rates, better known as WeRateDogs, served as the project's data source.A Twitter account called WeRateDogs awards ratings and hilarious remarks about people's dogs.The three processes of data wrangling are documented in the wrangle report: gathering data, assessing data, and cleaning data.

- ## Gathering Data

I must acquire information for my project from several sources and in a variety of forms.

1. The archive of WeRateDogs tweets. The project provides the file, which is available for direct download from the Udacity website
2. The predicted tweet images. The file is stored on servers owned by Udacity. I programmatically downloaded this file using Python's Requests module.
3. Retweet count and favorite count data from another file can be obtained in place of that from the Twitter archive. Since I don't have a Twitter account, I decided to obtain the tweet JSON file programmatically using the Requests package.

- ## Assessing data

After collecting the data, I evaluated it programatically and visually to find any problems with data quality and tidiness. Tidiness pertains to data structure, whereas quality relates to content. Organize your data by dividing each variable into a column and each observation into a row, a table is formed by each sort of observational unit.I used code in Jupyter Notebook to display particular data subsets and summaries , for example , describe , duplicated, query, value_counts, head and info methods. In order to address the problems later, during the cleaning process, I have taken notes of the observations I made when evaluating the data.

- ## Quality Issues

1 - Remove retweets and replies

2- change datatype of timestamp to datetime

3- Remove columns that not important

4- There are more sources but it not important

5- Some rating_numerator are not correct

6- Change data type for tweet_id to str

7- p1,p2 and p3 some of the start with lowercase others with uppercase

8- There are some photos are missing for IDs

- **Tidiness Issues**

1- All dataframe are related we should marge them

2- There are 4 columns for dog stages: puppo, pupper, floofer, doggo , should be in one column

- **Cleaning data**

As I assessed the issues, I cleaned each one. Despite the fact that the entire dataset has a lot of problems, fixing them all would take a lot of time. I therefore concentrated only on those relevant to my analysis. The previous issues were cleaned as appropriate resulting in a high quality and tidy master pandas DataFrame