Wrangle & Analyze WeRateDogs Data

Wrangling Report

In this project, we went through Data Wrangling. We first started by Gathering the data then Assessing it, and finally cleaning it.

Dataset:

The tweet history of Twitter user @dog rates, commonly known as WeRateDogs, is the dataset I will be manipulating. WeRateDogs is a Twitter account that rates users' dogs and adds a lighthearted comments.

- **Data Gathering**

In this step, we gathered data from three different sources: Enhanced Twitter Archive, Twitter API, and a TSV file (download programmatically)

- **Assessing Data**

This step includes assessing data visually and programmatically from which we concluded the following Quality and Tidiness issues.

❖ **Quality issues**

Enhanced Twitter Archive Data

1. The dataset is incomplete. It contains 2075 samples, not 5000.
2. "<a href=" exists in the source column
3. Incorrect Datatypes:(retweeted_status_timestamp, timestamp..)
4. Missing Data in in_reply_to_status_id and in_reply_to_user_id
5. "None" instead of NaN in Doggo, floofer, pupper, puppo
6. In Sample 193, name is quite (incorrect).
7. Delete unnecessary columns for Analysis(retweeted_status_timestamp...)
8. In sample 2311, the name is ‘a’ instead of Octaviath.

Image Predictions Data

1. The dataset is incomplete. It contains around 2000 samples, not 5000.
2. None descriptive columns' names.

Twitter API Data

1. The dataset is incomplete. It contains around 2000 samples, not 5000.

❖ **Tidiness issues**

Enhanced Twitter Archive Data

1. timestamp contains two pieces of information Date and time. So we can separate these two into two columns.
2. Doggo, floofer, pupper, puppo should be in one column.

- **Cleaning Data**

In this step, we cleaned some of the issues we found in the data.

➔ Issue #1: Delete unnecessary columns for Analysis(retweeted_status_timestamp...) + Drop Duplicated rows(Retweets)

➔ Issue #2: Doggo,floofer,pupper,puppo should be in one column.

➔ Issue #3: Correct timestamp type then Separate the information

➔ Issue #4: "<a href=" exists in the source column (Enhanced Twitter Archive)

➔ Issue #5: Image Predictions dataset columns' names

As a final step, after cleaning the datasets we merge all three of them and export them into a CSV file called Twitter_archive_master.csv.