Report Data wrangling

“WeRateDogs”

Data wrangling procedure has been completed on the datasets obtained from “WeRateDogs”. There have been 3 main stages:

1. **Data gathering**
2. **Data assessment**
3. **Data cleaning**

Data have been gathered from different resources such as a provided file (twitter-archive-enhanced.csv), a downloaded file via the internet using the provided link (programmatically) and file that was gotten from interrogating the twitter API using tweepy.

As a result 3 datasets (df, Api_df, df_image) have been collected for further assessment and cleaning.

The three saved data frames were then assessed visually inside a jupyter notebook with pandas and because the datasets were not too large, a copy of each was exported into one Excel workbook. This allowed quick scanning through the rows and use of filters to identify areas for more detailed investigation. Following this a programmatic assessment was made inside jupyter with pandas using the following functions, df.info(), df.head(), df.sample(), df.value_counts().

Data assessment lead to find 2 kind of issues: Quality and *Tidiness* issues:

Quality refers to issues related to the content of the data, sometimes called dirty data. The standard criteria of completeness, validity, accuracy, and consistency of the data were used to identify quality issues. These issues were varied and are listed in the assessment section of the “wrangle_act.ipynb” jupyter notebook.

Tidiness refers to issues related to the structure of the data, sometimes called messy data. The basis for assessment is that each variable forms a column, each observation forms a row and each type of observational unit forms a table. After assessing the three datasets, it was decided to marge them into a single data frame, reducing superfluous columns that wouldn’t be needed in any future analysis.

The following problems were identified:

Quality issues

1.unnecessary columns in df like in_reply_to_status_id ,in_reply_to_user_id..

2.Drop all rows with retweets because we only want original tweets

3.incorrect data types like timestamp and tweet_id are not in the right format.

4.dogs names aren't homogenous , some starts with capitals and others not

5.irrelevent values in columns like name (Dogs without names, but given names of “a” or “an” instead of “None.”)

6.irrelevant values in rating_numerator and rating_denominator such as denominator=0 or numerator =1776 ...

7.tweets with missing data in the expanded_urls.

8.The source column looks messy and clutters the table

Tideness issues

1.df and API_df can be gathered in one dataframe

2."doggo" "floofer" "pupper" "puppo" can be all in one column called step_dog

The final step in the wrangling process is **cleaning** the data for quality and tidiness issues. The cleaning followed the standard process of define, code and test for each of the issues that were documented before and they were tackled in a logical order. Most of the cleaning was performed using programmatic tools, such as functions or pandas built-in functions (merge, melt etc.).

It was difficult to handle the huge amount of missing or irrelevant data, and to produce in the end a one table data frame that contains clean and concise information only