Wrangle Report: WeRateDogs Twitter Archive

Assessing Data

You can assess data for:

- Quality: issues with content. Low quality data is also known as dirty data.
- Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as messy data. Tidy data requirements:
  1. Each variable forms a column.
  2. Each observation forms a row.
  3. Each type of observational unit forms a table.

...using two types of assessment:

- Visual assessment: scrolling through the data in your preferred software application (Google Sheets, Excel, a text editor, etc.).
- Programmatic assessment: using code to view specific portions and summaries of the data (pandas' head, tail, and info methods, for example).

https://cran.r-project.org/web/packages/tidyr/vignettes/tidy-data.html
The six core data quality dimensions are (https://strategicdb.com/2017/10/27/six-dimensions-to-data-quality):

- Completeness
- Uniqueness
- Timeliness
- Validity
- Accuracy
- Consistency

Data Quality Check Methods

Data quality should be assessing both visually and programmatically for quality and tidiness. Sometimes a quick look via Excel or a text editor is sufficient for seeing issues. Making sure datatypes are correct is another major consideration for data quality. Integers should be stored as ints, numbers with decimals as floats, datetimes stored as datetimes (not strings) and so on.

For more complex analysis, utilizing of regular expressions (RegEx) is a powerful tool to have in your toolbelt.

After analyzing the data from the three files, major issues were found and cleaned. See the next section for details.

Data Issues Summary

Data Quality Issues

1. There are 59 NULL values in expanded_urls.
2. Timestamp is stored as object instead of datetime.
3. There is one tweet with a denominator value of '0'.
4. There are 181 tweets that are retweets.
5. There are 78 tweets that are replies.
6. Duplicate data, there are 66 instances of the same photo(s) being predicted..
7. Instances of invalid dog names (‘None’ and lowercase words).
8. Dog "stage" has string value of 'None' instead of NULL for the respective columns (‘doggo’, ‘floofer’, ‘pupper’, ‘puppo’).
9. The 'source' column should be in a cleaner format.

Tidiness Issues

1. The dog 'stages' should be one column with values of 'None', 'doggo', 'floofer', 'pupper', 'puppo'.
2. Remove replies and retweets, we only want original tweets.
3. Three predictions are excessive and not useful for this project. Remove P2 and P3.
4. Combine the three dataframes into one for easier analysis.