WRANGLE REPORT Udacity DAND: Wrangle and Analyze Data Project :

Data wrangling steps:

1) GATHERING DATA:

Data was gathered from 3 different sources:

1) The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.

2) Additional data, including favorite count and retweet count, were gathered using the Twitter API.

3) The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers. Using machine learning techniques, the breed of dog was predicted based on the picture.

2) ASSESSING DATA:

After the data was gathered, assessment was performed using the following methods:

- head()
- sample()
- info()
- value_counts()

1> Tidiness issues that were cleaned:

1. Combining all dataframes together as they all contained information about the same tweets .
2. Combining 4 variables about dog type into 1 column "dog_stage"

2> Quality issues that were cleaned:

1) Data contained retweets
2) Tweet id was the incorrect data type
3) Timestamp was the incorrect datatype
4) Name contained the string “None” instead of a NaN
5) Name contained various inaccuracies which were regular lowercase words
6) The name O’Malley was incorrectly extracted as “O”
7) Rating numerators which contained decimals were incorrected exported
8) Ratings are unstandardized
9) Undesired columns present

3) CLEANING DATA

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- merge()
- reduce()
- .extract()
- .drop()
- .isnan
- .astype()
- .to_datetime()
- .islower()
- .replace()
- .rename()
- set_option()
- .loc[]
- .value_counts()
- .info()
- .head()
- Loops
- Regular expressions