REPORT
WERATEDOGS

Data Source:

The data was gathered from a given CSV, a website, and Twitter's API. Tweepy was used to access the API and gather the JSON data for the tweets. The JSON data was stored in a text file, then loaded what is required into a pandas data frame.

Process:

- The quality and tidiness issues were identified.
- A copy was created for each dataset before cleaning.
- The issues were fixed data was cleaned.
- The cleaned data was saved.
- The dataset was explored using data visualization.

Assessment:

The following Tidiness issues were found:

- Erroneous datatype of tweed_id.
- Erroneous datatype of stage column.
- Presence of ratings and links in text.
- Irrelevant names starting with small letters.
- Erroneous datatype of timestamp.
- Erroneous datatype of retweet_counts and favorite counts.
- Presence of html anchor tabs in expanded_url column.

The tidiness issues were fixed by the following:

- The respective erroneous datatype of variables were converted to the appropriate ones.
- The ratings and links in text were removed.
- The names starting with small letters were removed.
- The html anchor tabs were removed from expanded_url column.

The following quality issues were found:

- Retweets doesn't fit into the data analysis and needs to be removed.
- Lack of seperate column dog stage that contains observations from duggo,floofer,pupper and puppo .
- The ratings are wrong and needs to be normalized.
- Seperate columns for dog prediction and probability of confidence is required that contains observations from columns 'p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog'.
- Data which has images should only be considered for analysis.
- Retweet count and favourite count must be added.

The quality issues were fixed by the following:

- The retweet columns were removed.
- Separate column dog stage that contains observations from duggo,floofer,pupper and puppo is added.
- Separate columns named predicted breed and p_conf are created that contains observations from columns 'p1','p1_conf','p1_dog','p2','p2_conf','p2_dog','p3','p3_conf','p3_dog'.
- Retweet count and favourite count were added from json data file.

Data visualization:

- A bar plot was plotted to find the number of dogs that belongs to each stage.

Inference:

Pupper have the highest count in the data and Doggo & Floofer the least.

- A scatter plot was plotted to find the correlation betweet favorite count and retweet count.

Inference:

retweet_count and favorite_count are positively correlated.

- A boxplot was plotted between rating and dog stage.

Inference:

The overall rating is almost in the same range for all stages with pupper in lower end and doggo,floofer and puppo in higher end.

- Rating vs timestamp was plotted to view the change in rating density over time.

Inference:

One could observe that mostly, irrespective of the time the rating given is 12.