Wrangle Report

INTRODUCTION

This project involves assessing and cleaning data derived from Twitter database as an example of how real-life data exists.
There are several steps required to be carried out in the project which are:

Steps Taken

1. Data Gathering
2. Assessing the Data gathered
3. Cleaning data
4. Storing the Data,
5. Analyzing the Data, and Visualization to derive insight.
6. Reporting

Key Metrics/Action Points

- To derive original ratings (no retweets) that have images from the dataset.
- To assess and clean the dataset with at least eight (8) quality issues and two (2) tidiness issues taken care of.
- To create copies of datasets and merge all into a master dataset, then saved into an external csv file.
- To generate at least three insights and one visualization from the dataset.
- To report wrangling and visualization activities carried out.

1. Data Gathering

Data was gathered from three sources: The We rate twitter archive provided by Udacity, The tweet image predictions downloaded programmatically from Udacity and the Api Data from Twitter which was imported through a JSON file provided by Udacity.

2. Assessing the Data

The Data were assessing visibly and programmatically for quality and tidiness issues which were to be worked on in the wrangling aspect of the project. The assessments carried out are outlined in the table below:

Twitter Archive	
The columns related to replies are not applicable for original tweets	Quality
The numerator and denominator columns have invalid values.	Quality
The numerator and denominator columns have values that were not accurately extracted.	Quality
timestamp should be datetime instead of object.	Quality
in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp columns are not needed.	Quality
Incorrect or missing names in name column	Quality
Values of "None" in the name column.	Quality
There are 4 columns for dog stages (doggo, floofer, pupper, puppo) with the same measurement.	Tidiness
The source column has html	Tidiness
Image Prediction	
About 281 IDs are missing comparing the data against the archive data.	Quality
The p1, p2, p3 data have columns instead of spaces between them	Quality
p1, p2 and p3 should be categoral datatype	Quality
p1_conf, p2_conf and p3_conf columns should be merged	Tidiness
p1_dog, p2_dog and p3_dog columns should be merged	Tidiness
Api Data	
Number of missing IDs is about 2when compared to the archive dataset.	Quality
The id column should be renamed to allow merging of the datasets	Tidiness

3. Cleaning

- Drop columns: (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp)
- Cleaning the Numerator and Denominator
- Adjust numerator values that were not correctly extracted
- Adjust timestamp datatype
- Replacing all names with errors and the value of None in the column
- Merge the dog stages columns
- Clean the Source column, remove the html
- converting p1, p2, p3 to categorical datatypes
- remove false vlaues in the prediction ~and confidence column
- Renaming the id column

4. Storing the Data

The three datasets were merged and stored as a Master Dataset then imported back for Analysis and Visualization

5. Analysis and Visualization

The Data was analysed and the insights were conveyed with graphics as seen the Act Report.

6. Conclusion/Reporting

a) Most of the users were using iPhones which depicts that most of the dog owners can afford the expenses of taking care of their pets.
b) Most people named their dogs cooper.
c) The Saluki breed has the highest score which reflects the likability of the dog breed compared to others.
d) The Pupper stage is the most common dog stage which could also indicates why most of the pictures got a likes.
e) The golden retriever is the most common dog breed in the data which could also be because of the breeds' companionship trait
f) The Data also shows that there is a positive correlation between the retweets and likes, which shows that as the retweets increased, the likes also increased.