ACT REPORT

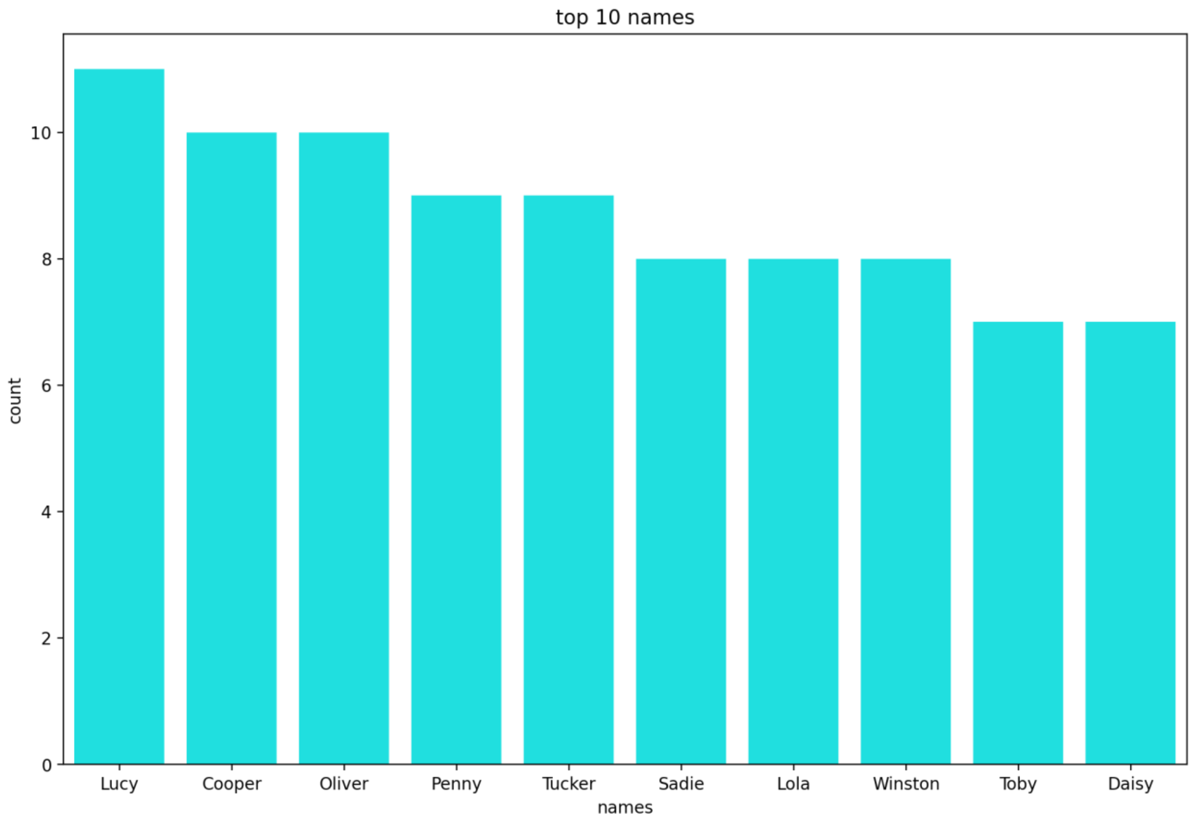
This is a report on the insights that were found in the data which was gathered from different sources. Some of the files were downloaded by web scarping and some obtained by downloading the data provided by Udacity. Some tidiness and quality issues where detected, Hence I proceeded in cleaning the data to ensure the insights are accurate and of good quality. Here are the measures that were taking to clean the data:

- The doggo, floofer, Puppo, and pupper columns were cleaned by replacing the None values with empty cells, after which it was merged into one column and the individual columns dropped from the data frame.
- The duplicated jpg URLs were removed from the df_image_predictions after which it was moved to the df_archive-clean data frame since this was done the expanded URL was dropped from this data frame.
- The timestamp and dog stages columns were changed to date-time and categorical datatypes respectively.
- All images that are not dogs were removed from the image predictions data frame.
- The p1,p2, and p3 columns were all changed to lowercase letters so as to have the same casings and values.
- Dropped columns are not needed from the data frame.
- Cleaned the text column by removing the ratings and URLs presents.

The following insights were made after merging the data frames:

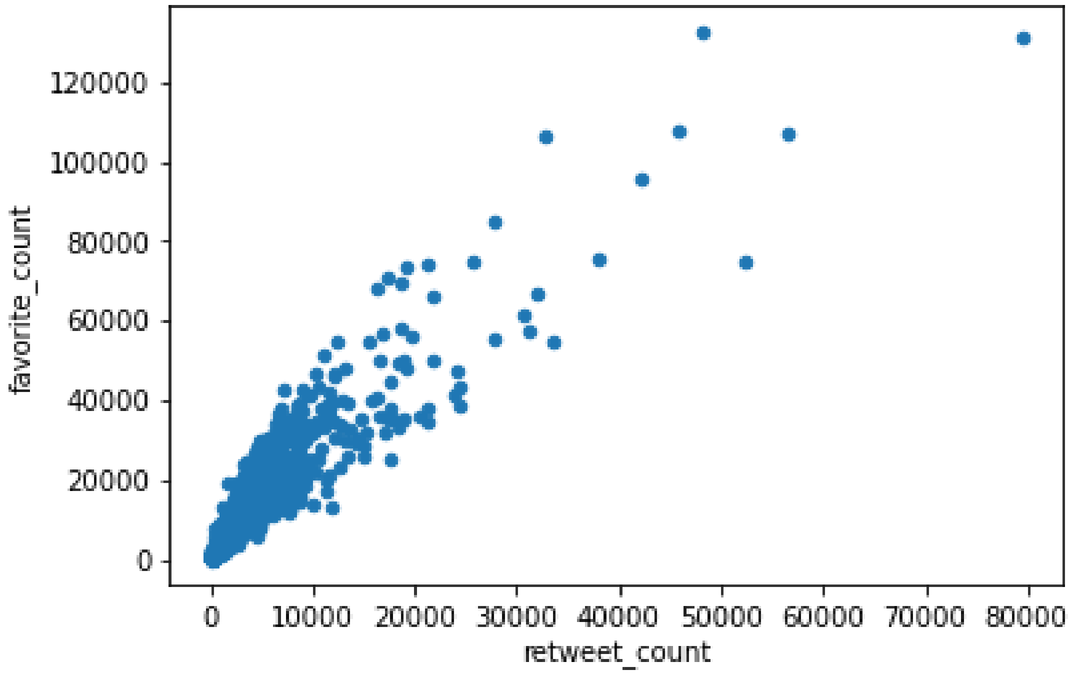
1. To determine the top 10 dog names in the data frame:

	Names	count
0	Lucy	11
1	Cooper	10
2	Oliver	10
3	Penny	9
4	Tucker	9
5	Sadie	8
6	Lola	8
7	Winston	8
8	Toby	7
9	Daisy	7



2. Determine the outlier present in the rating_numerator: Observed that some of the ratings are abnormal from the rating style of the We_rate_dogs.

3. Determine the correlation between retweet_count and favourite_count: With a correlation of 0.91, this indicates a strong positive correlation.


favorite_count
retweet_count