Data Visualization Report

Introduction:

This project work contains datasets of twitter user @dog_rates, also known as 'WeRateDogs'. This is a twitter handle that rates people's dogs with a humorous comment about the dog. @dog_rates datasets used in this data analysis contains tweet data of over 5000+ which were obtained from three sources.

Figure 1: A golden retriever named Stuart (Source: Udacity Data analysis course)

The aim of this data analysis project is to gather these data from its sources, evaluate and assess these datasets visually and programmatically for its quality and tidiness, then clean and visualize content of the data in a written report.

Insight 1: Distribution of WeRateDogs Tweet timeline

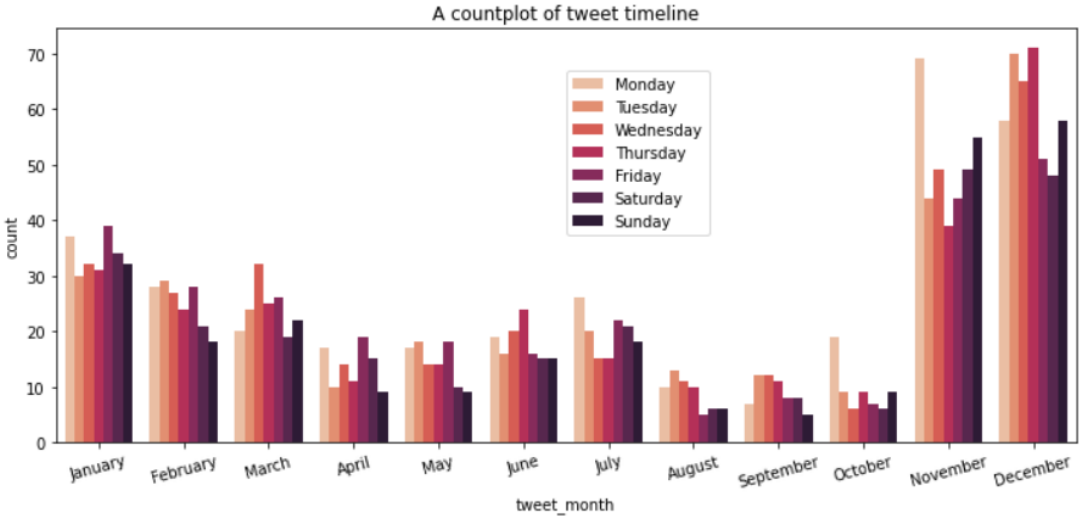

Figure 2: A count plot of WeRateDogs tweet timeline

Observation: Based on the visualization plots illustrated above, Tweets on dog rates are usually issued on Mondays for weekdays and more commonly in December of a year. It is also observed that the latter and early part of the year shows high dog rates tweets. Generally high mean dog ratings are usually given in the month of July.

Insight 2: Impact of display_text_range on tweet ratings

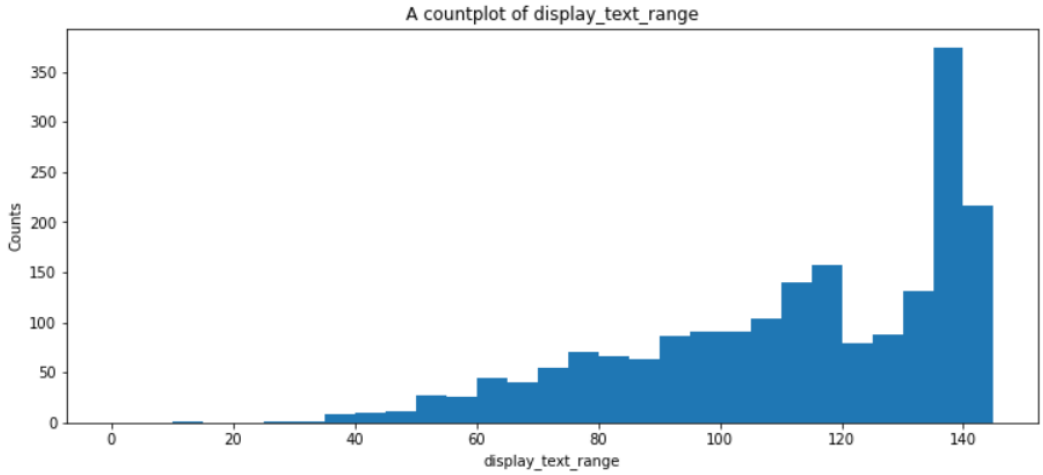

Figure 3: A count plot of display_text_range

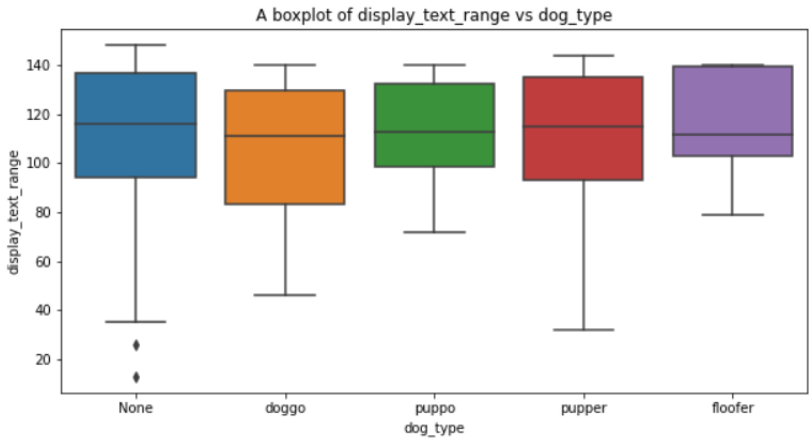

Figure 4: A boxplot of display_text_range vs dog_type

Observation: Results shows that 75% of display_text_range is about 137 characters. Furthermore, it is not quite clear how mean dog ratings are affected by length of display_text_range. However, is a very minor tendencies that the longer the display text chat the less the mean dog rating. The heatmap plot show that many of the distant_text_range characters fall between 130-137 for a peak mean rating of 1.13-1.20. Based on the box plot above, the median population of the dog types are almost the same, however, display_text_range with pupper dog types have a little bit higher number of characters than other dog types apart from 'None'.

Insight 3: Impact of dog_type on tweet ratings

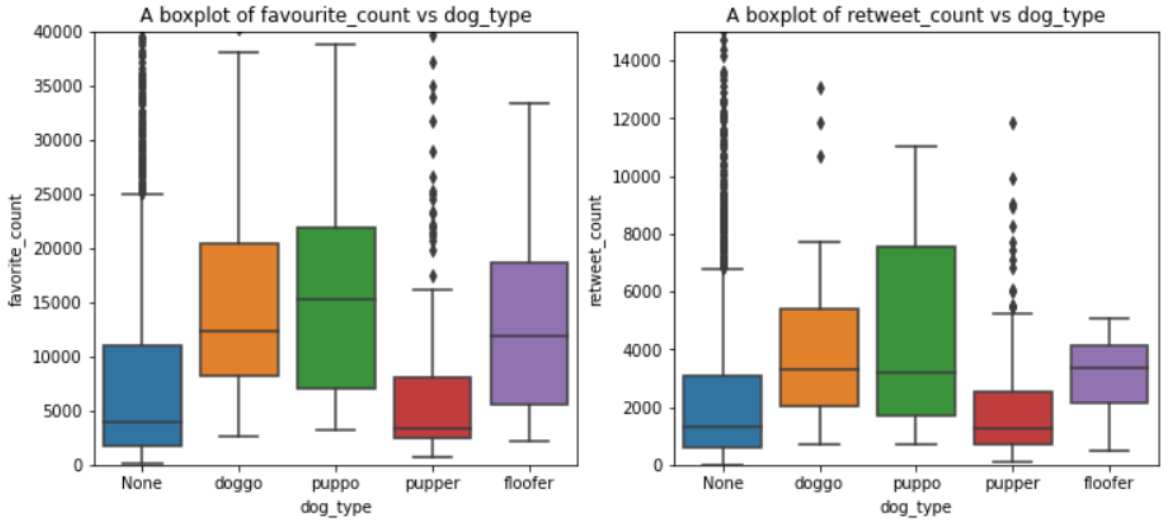

Figure 5: A boxplot of favourite_count vs dog_type and retweet_counts vs dog_type

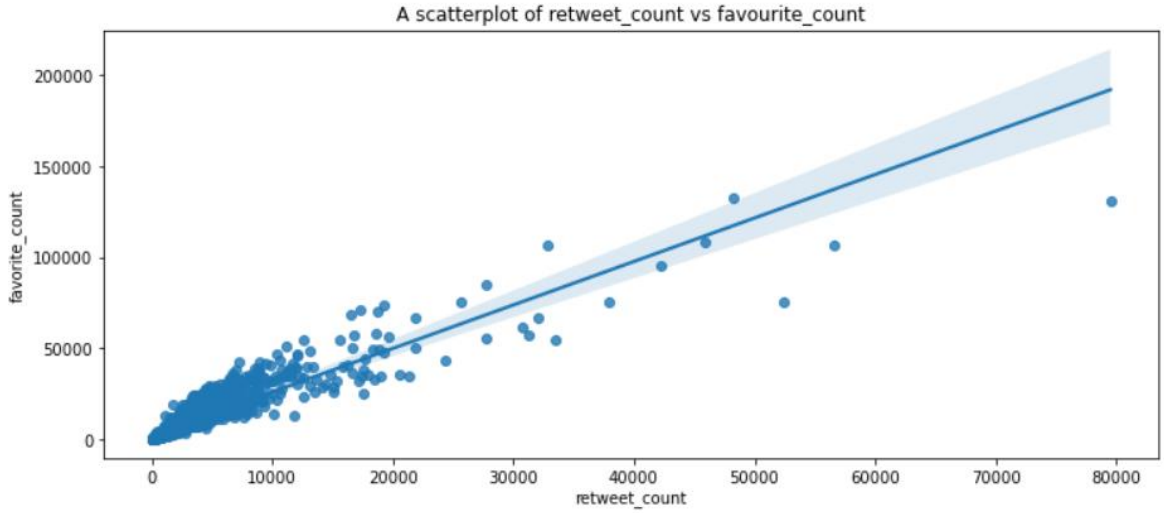

Figure 6: A scatterplot of retweet_counts vs favourite_count

Observation: Results from the plot above shows that generally puppo dog type have a higher favourite count than other dog types while tweets with doggo and puppo have the highest retweet ratings. From the above favourite_count vs retweet count plot indicates a positive linear correlation distribution.