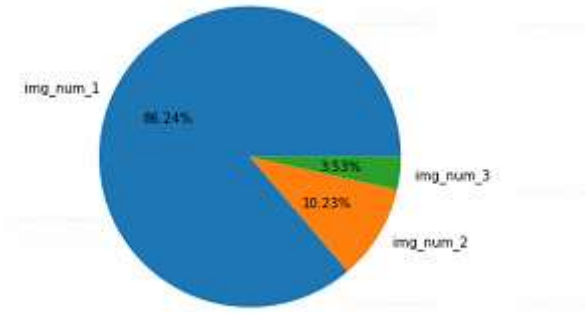
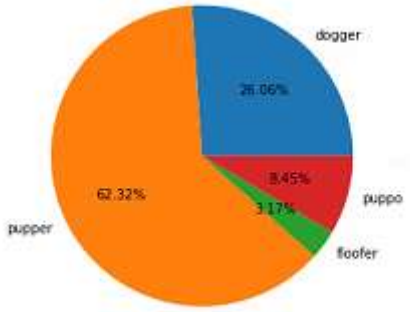
PROJECT2 : WRANGLING AND ANALYZE DATA

ANALYZE AND VISUALIZATION

1. Proportions of tweet per prediction model
   - Write function pie_plot(data, labels) which plot a pie graph
   - Create a list of number of tweet_id per image number and call pie_plot() function



2. Proportions of tweet per dog stage
   - Create a list which contain number of tweet for each dog stage
   - Call pie_plot function

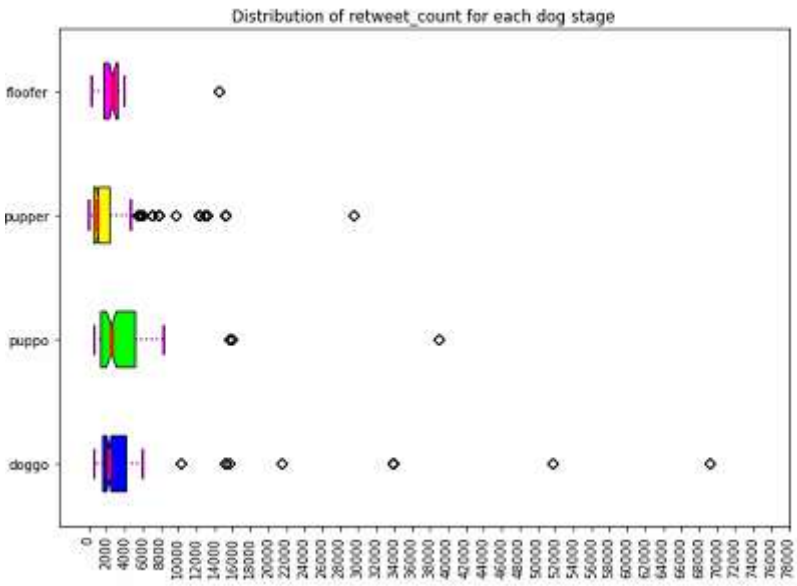


3. Description of retweet_count and favorite_count per dog stage
   - Extract rows from twitter_archive_master table for each dog stage
   - Extract rows from twitter_archive_master table for tweet which have multiple_stages to create multipleStage dataframe
   - Write function named AppendValue (multipleStage_tweet, dog_stage_num, initialArr, prop) to append property values for tweet_id with multiple dog stage
     - Dog_stag_ num is number of dog stage in an array of four elements where each element equalled to 0 or 1(first element

represents doggo, second element represents puppo, third element represents pupper, and fourth element represents floofer)

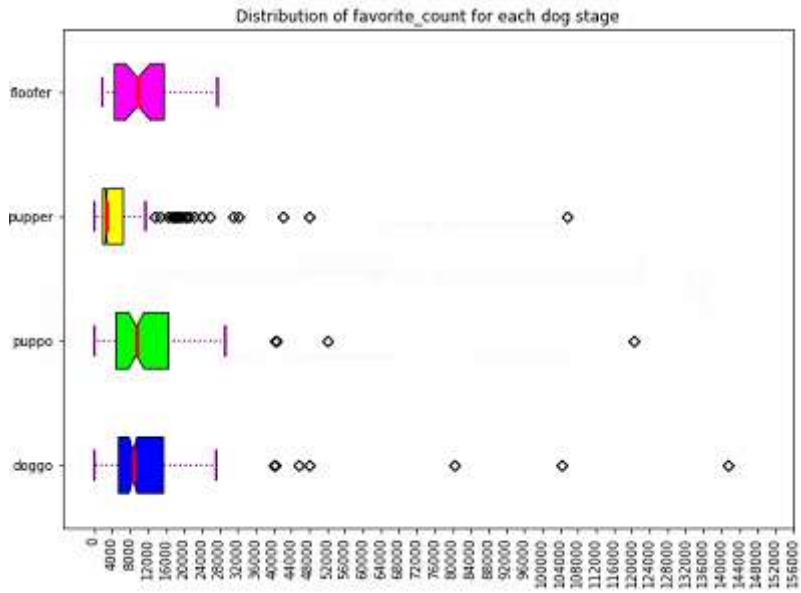
- initialArr is array to append new value
- prop is property to find value to append

- write function named SumBoxPlot(param, stop, step) to plot box-plots for distributions of retweet_count and favourite_count
  - param is retweet_count or favorite_count
  - stop is end of x-labels
  - step is difference between two consecutive labels
- call SumBoxPlots('retweet_count', 80000, 2000) and interpreting result :
  - floofer dog stage has a larger average of retweet_count(about 4000)
  - for floofer dog stage, retweet_count max is 4000 (retweet_count average is 4000)
  - for pupper dog stage, retweet_count max is 5000 (retweet_count average is 1000)
  - for puppo dog stage, retweet_count max is 9000 (retweet_count average is 3000)
  - for doggo dog stage, retweet_count max is 6000 (retweet_count average is 3000)

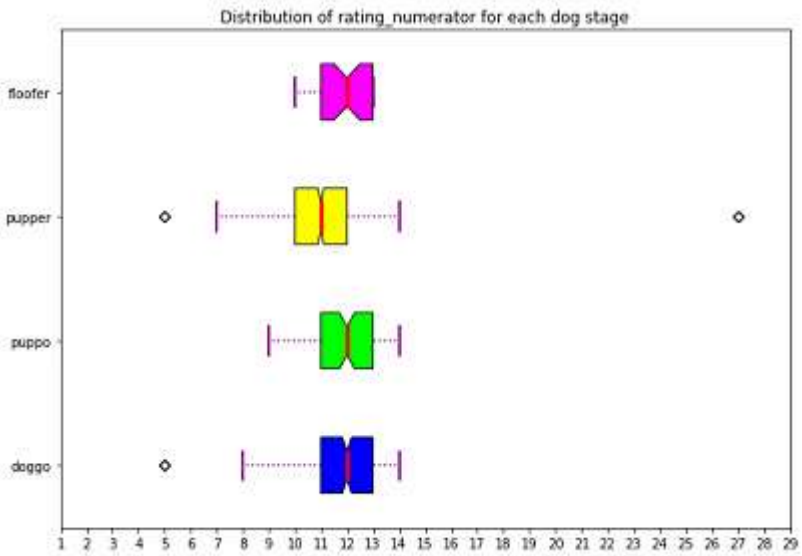


- SumBoxPlots('favorite_count', 80000, 2000) and interpreting result :
  - floofer and puppo dog stages has a larger average of favorite_count(about 10000)
  - for floofer dog stage, favorite_count max is 26000 (favorite_count average is 10000)

- for pupper dog stage, favorite_count max is 12000 (favorite_count average is 2000)
- for puppo dog stage, favorite_count max is 28000 (favorite_count average is 10000)
- for doggo dog stage, favorite_count max is 28000 (favorite_count average is 8000)



4. Description of rating_numerator and rating_denominator per dog stage
   - Reuse AppenValue function
   - Write function named SumBoxPlots2(param) to plot box plots for distributions of rating numerator and rating Denominator
     - Param is rating_numerator or rating_denomination
   - Call SumBoxPlots2('rating_numerator') and interpreting result graph :
     - for floofer dog stage, rating numerator min is 10 and rating numerator max is 13(rating numerator average is 12)
     - for pupper dog stage, rating numerator min is 7 and rating numerator max is 14(rating numerator average is 11)
     - for doggo dog stage, rating numerator min is 8 and rating numerator max is 14(rating numerator average is 12)
     - for puppo dog stage, rating numerator min is 9 and rating numerator max is 14(rating numerator average is 13)



- Call SumBoxPlots2('rating_denominator') and interpreting result graph :
  - rating_denominator is always equal to 10.0