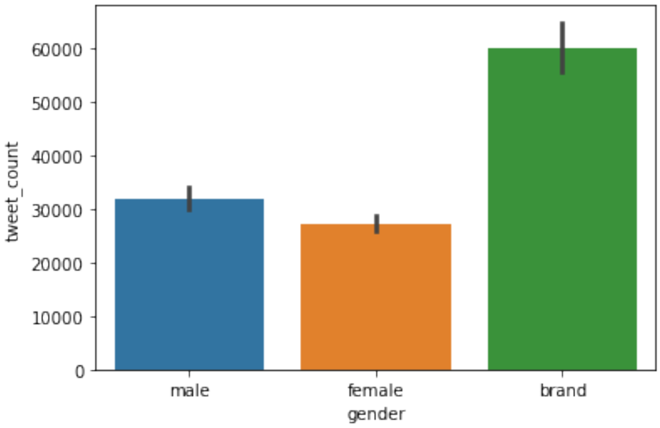
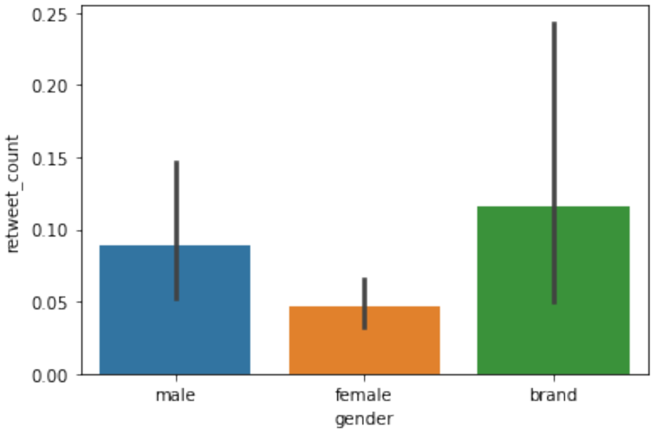
SUMMARY

Start by importing library modules . Create data frame . Start cleaning data by dropping unnecessary columns . Later clean the text by dropping null values from gender . Merge the ‘text’ and ‘description’ to combine all sorts of text and then finding out the common words. Start by cleaning junk words and letters other than the English vocab words are filtered out. Later remove stop words from ‘text_description’ . Start filtering out 'text_description' and printing most commonly used words by eliminating stopwords. Later started solving problems like

1.Represent the number of tweets based on gender? ,
Ans:



2.Represent the number of retweets based on gender? ,
Ans:



3.What is the number of females present ? ,
Ans: 35.6% females are present

4.Which year has the maximum tweets? ,
Ans: 2014

5.At what hour minimum tweets are done? .
Ans: 1 am

Later performed 3 ML algorithms and finding out which algorithm best suits the dataset with respect to the accuracy of the algorithm .

Observations:
Accuracy of Logistic Regression is 48.80%
Accuracy of Random forest is 48.48%
Accuracy of Naive Bayes is 61.86%

Naive bayes approach is better in this Gender prediction as it uses nltk (natural language toolkit) which uses the description as the predicting variable to detect the gender