act_report

May 15, 2020

0.1 Quality Issues

0.1.1 df:

1)Missing data in the following columns: in_reply_to_status_id, in_reply_to_user_id, 2)retweeted
3)This dataset includes retweets, which means there is duplicated data [Solved]
4)Timestamp and retweeted_status_timestamp is an object [Solved]
5)The source column still has the HTML tags [Solved]
6)Dogs name have 'None', or 'a', or 'an.' and some more lower case words as names [Solved]
7)Multiple dog stages occurs such as 'doggo puppo', 'doggo pupper', 'doggo floofer' [Solved]

0.1.2 image_pred_df:

1)dog breeds are not consistently in p1,p2,p3 columns [Solved]

0.1.3 df_tweet_json:

1)Missing data [Solved]
2)tweet_id is an object [Solved]

0.2 Tidiness Issues

0.2.1 df:

1)The variable for the dog's stage (dogoo, floofer, pupper, puppo) is spread in different column

0.2.2 image_pred_df:

1)This data set is part of the same observational unit as the data in the archive_df [Solved]

0.2.3 tweet_json_df:

1)This data set is also part of the same observational unit as the data in the archive_df [Solve

0.3 Findings of the analysis

1)The pred_breed column is created based on the the confidence level of minimum 20% and 'p1_dog'
2)Based on dog types: doggo, floofer, pupper, puppo, 'doggo, puppo', 'doggo, pupper', 'doggo, fl
3)tweet_id is set as object type as it is not going to use for calculation.