wrangle_report_re_submission

January 8, 2019

1 Wrangle Report

Data Wrangling leads the foundation step for any kind of Data Analysis. In this Project, we first read the files and tried to list down the Quality issues which are being listed down and along with the fixes below

1.0.1 Data Quality issues

1. Missing values in expanded_urls column >*The missing values are issues of Completeness of a dataset, thus missing values should be dropped, drop.na function is being used*
2. Columns timestamp and retweeted_timestamp are in strings >*The timestamp and retweeted_timestamp are in strings which needs to converted to datetime format and pandas date-time function is used to fix this issue*
3. vine.co in source column is not a valid website >*This is an typo error needs to corrected and str.replace fuction is used*
4. " characters in the source column >*These characters are not necesarry, might be generated due to HTML Codes. String substring is used to get rid of this*
5. tweet_id, retweeted_status_id, retweeted_status_user_id, in_reply_to_status_id and in_reply_to_user_id in floats >*id variables can't be in float or integer format. Thus integers are first converted to floats and then floats are converted to strings using .astype*
6. expanded_urls have comma separated repetative values >*We need to get rid of this as this is no use. str.split function is used*
7. The columns rating_numerator and rating_denominator is in integer format >*ratings can be in float so we changed the format from integer to float ussing .astype*
8. 'a', 'an', 'by', 'the', 'nothing' are not valid dog names >*We see if the first character is lower, the name is not valid. So we check the first character case, if it is lower we drop those records*

1.0.2 Tidiness

1. The timestamp column contains date and time integrated into 1 >*.date() and .time() function is used to separate timestamp column into date and time*
2. Having 4 different columns like 'puppo', 'doggo', 'floofer', 'pupper' increases the size of the data >*Created a new column based on the variety of the dogs by collating all the informations from columns like puppo, doggo, floofer and pupper*