Wrangle Report

INTRODUCTION

The current, project is about the data quality of three Twitter data sets that share the same key identifier tweet id.

The first data set Twitter archive, which contain a lot of unnecessary column columns which can be dropped, also some columns contain missing value (more than 90%), and some columns were supposed to be one single column.

The second data set contains information on tweet images URL of dogs, this contains some six columns that were supposed to be combined into two columns.

The third data set contains a lot of unnecessary columns that can be dropped from the data to remain with the necessary information that is used for further analysis.

Wrangling stages:

In the first step, data were gathered data from Twitter and assessed for quality and tidiness.

Python and its libraries are used as tools for data manipulation, visualization, and analysis.

- Some data (archive Twitter data) were downloaded manually,
- Some data were downloaded programmatically (image prediction data), using this link: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/imagepredictions.tsv
- the data in JSON format were supposed to be opened by Twitter API, but we did not get the Twitter developer account, then it was opened manually by pandas.

In the assessing step, different python functions were used to identify data quality issues, only archive data had missing values

In Archive data columns such as doggo, floofer, pupper, and puppo should be combined to the dog stage. Also, the in text column there was RT before some values which should be removed.

In image prediction, columns such as p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_conf, and p3_dog were supposed to be merged based on dog condition

In three data sets, the tweet id was in the wrong format. All issues were fixed to improve the quality of data for further analysis.

Storing

After succeeding with the quality control of three data frames together were merged into one only data frame which was saved as twitter_archive_master.csv.

Conclusion

Quality control of the data was done to improve the quality of data insight and findings using python and its libraries.