WRANGLE **REPORT**

By

Doaa Salahedin Ahmed

Project: Wrangle & Analyze Data

INTRODUCTION

This report is prepared with the purpose of illustrating a personal journey of wrangling and analyzing a dataet from the Twitter account **@dog_rates** which also referred to as **WeRateDogs**

My data analysis process went through the following steps:

1- Data gathering
2- Data assessment
3- Data cleaning

I. DATA GATHERING

- This project is concerned with the analysis of a specific number of observations from a given dataset named “twitter_archive_enhanced.csv”.
- However, throughout the data gathering step, this dataset was supplemented with additional data gathered by 2 different ways:

PROGRAMMATICALLY:

The image predictions dataset named “image_predictions.tsv”, was downloaded programmatically from Udacity severs using Requests library and a hyperlink.

These predictions supplied the main dataset with the phenomenon “dog breeds”.

API QUERY:

The third data source was obtained directly from twitter by querying Twitter API for the ID of each observation in the main dataset.

This process was accomplished using the Tweepy library, and the information was stored in the work directory as JSON file named “tweet_json.txt”.

Querying Twitter API supplied the main data set with additional 2 direct variables affecting our phenomena, which are “favorite count, retweet count”.

II. DATA ASSESSMENT

- This part started with appropriate definition of the important variables within each table, the 3 dataframes were assessed according to:

VISUAL ASSESSMENT:

Through careful exploration of 200 observations on each dataframe. In the way to do that inside jupyter notebook, the minimum displayed number of rows was set to 200 and the dot info method.

PROGRAMMATIC ASSESSMENT:

several methods were used for instance, dot info, dot value_counts, dot describe, dot sample, dot duplicated… etc.

- The addressed assessment issues were noted down in points in terms of quality and tidiness at the end of this section.

III. DATA CLEANING

- The cleaning part started out with creating copies of the 3 original tables to revert to the originals and create more copies upon the incidence of any error.
- The following step was merging the 3 copies into 1 master table to manipulate data within.
- Each cleaning step followed a sequenced structure of: Define, Code, Test.
- The cleaning process has gone from programmatic codes at the beginning to facilitate manual cleaning at the end. Therefore, the duplicates were dropped first as well as retweets and replies, scattered data between columns were aggregated to end up with single column for dog type, a single column for dog breed, and a single column for confidence level.
- All raw columns were dropped except for some few columns which would act as a guide through the manual and the visual cleaning afterwards.
- Cleaning the ratings was a bit sophisticated, so, the process started out with correcting outliers, then the tricky part was to remove those interfering data belonging to non-dogs.
- Some keywords in addition to the degree of confidence were explored to form a conditional state to recognize non-dogs' observations and removing them.

CONCLUSION

Throughout the data analysis process, gathering data from multiple sources alongside careful assessment and detailed cleaning, have altogether proven being the cornerstone for drawing meaningful analytical insights and visuals.