REPORT ON

DATA WRANGLING PROJECT

BY

SANGOTADE IDOWU

DATA WRANGLING STEP

1. Gathering

2. Accessing

3. Cleaning

Project Details

Our tasks in this project are as follows:

Data wrangling, which consists of:

- • Gathering data .
- • Assessing data
- • Cleaning data
- • Storing, analyzing, and visualizing our wrangled data
- • Reporting on o our data wrangling efforts and on our data analyses and visualizations

Gathering Data for this Project

1. The WeRateDogs Twitter archive file was given us . So, I downloaded this file manually by clicking the following link as provided : twitter_archive_enhanced.csv

2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is presented in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL:

https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-pre dictions/image-predictions.tsv

3. Each tweet's retweet count and favorite ("like") count at minimum. Using the second option of not creating a developer, the API Code for accessing the data was read and understood before pasting in my jupyter notebook. The resulting test file stored in JSON was then used to create dataframe from relevant columns extracted from the file for my twitter_info table

Assessing Data for this Project

These are the following Quality and Tidiness Issues what found .

Quality Issue

Twitter Archive

Quality

S/N	ISSUESS	CLEANING
1	Remove retweets	Delete retweet by filtering the Null values of retweeted_status_user_id column
2	Drop Columns not needed for future analysis	Drop columns with pandas drop function
3	Erroneous datatypes in these columns (tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, source, retweeted_status_user_id, retweeted_status_timestamp, doggo, floofer, pupper, and puppo)	• Convert tweet_id to str from twitter_archive • Convert timestamp to datetime • Convert source to category datatype
4	Source column is in HTML-formatted string, not a normal string.	Extract HTML values from source column
5	Error in dog names (e.g a,an,actually) are not a dog's name.	Change error name in dog name to None.
6	Text column includes a text and a short link.	Remove http links in tweets.

7	Fix the rows that was not extracted properly for rating_numerator data	Spot those records and confirm changes made.

IMAGE PREDICTION

S/N	ISSUES	CLEANING
1	Erroneous datatype (tweet_id)	Convert tweet_id to str

API TABLE

S/N	ISSUES	CLEANING
1	Erroneous datatype (tweet_id)	Convert tweet_id to str

TIDINESS

S/N	ISSUES	CLEANING
1	doggo, floofer, pupper and puppo columns in twitter_archive table should be merged into one column named "dog_stage"	Merge columns into one column named 'dog_stage'
2	Image predictions table should be added to twitter archive table	Merge table with twitter archive table.
3	twitter api table columns to be added to twitter archive table	Merge table with twitter archive table.