DATA WRANGLING PROCESS

WRANGLING DATA CONSISTS OF THREE MAIN STEPS:

- GATHERING DATA.
- ASSESSING DATA.
- CLEAN DATA.

WE REPEAT STEPS EVEN AFTER CLEANING (WE SEE HIDDEN FINDINGS) THE WE BACK AGAING AND AGAIN UNTIL WE GET THE BEST RESULT.

GATHERING DATA

DATA WERE GATHERED FROM 3 DIFFERENT SOURCES:

- FIRST DATA SET (THE WERATEDOGS TWITTER ARCHIVE) - DOWNLOADED MANUALLY, WE CONVERT IT DIRECTLY INTO DATA FRAME USING PANDAS LIBRARY; THROUGH READ_CSV METHOD.
- SECOND DATA SET (THE TWEET IMAGE PREDICTIONS) -DOWNLOADED PROGRAMMATICALLY USING PYTHON REQUESTS LIBRARY, THE WE STORE IT IN OUR DEVICE USING OS LIBRARY THE IT IS EASY TO CONVERT IT INTO DATA FRAME USING PANDAS AGAIN.
- THIRD DATA SET (TWEETS JSON DATA) - I DOWNLOADED USING TWEETER API (TWEEPY) AS DATA WERE SCRAPED FROM TWITTER FROM SOURCES PROVIDED IN JSON FILE, THE SCRAPED DATA WILL BE IN TXT FILE (JSON FORMAT) WE WILL LOAD THIS JSON TXT INTO PYTHON (DICTIONARY) AS CONVENTIN!
  THEN WE LOAD THIS DICTIONAY INTO A PANDAS DATA FRAME.

ASSESING DATA

WE ASSESS DATA BY SEEING THE DATA WITH EYES AND BY APPLYING SOME CODING.

AFTER ASSESSING STEP, WE CAN LIST BOTH QUALITY AND TIDINESS ISSUES.

1. **QUALITY:**
   - **ARCHIVED DATASET**
     1. WRONG DATA TYPES IN (TIMESTAMP, RETWEETED_STATUS_TIMESTAMP)
     2. NULL VALUES IN (IN_REPLY_TO_STATUS_ID, IN_REPLY_TO_USER_ID, RETWEETED_STATUS_ID, RETWEETED_STATUS_USER_ID, RETWEETED_STATUS_TIMESTAMP, EXPANDED_URLS) - WE CAN EXCLUDE SOME COLUMNS AND CONSIDER THEM AS EXTRANEOUS COLUMNS.
     3. HTML TAGS IN SOURCE.
     4. SOME RECORDS HAVE DENOMINATOR NOT EQUAL TO 10(NORMALIZATION IS NEEDED).
     5. INVALID DOG NAMES.
     2. 4. SOME RECORDS ARE RETWEETS AND REPLIES.
     6. DOGGO, PUPPO, PUPPER AND FLOOFER COLUMNS HAVE A LOT OF 'NONE' VALUES.
   - **PREDICTION DATASET**
     1. WRONG DATA TYPES IN (TWEET_ID)
     2. P1, P2, P3 HAS NO REAL DOG NAMES SOMETIMES
     3. COLUMNS NAMES IN IMAGE PREDICTION
3. **TIDENESS:**
   - ALL THREE DATASET SHOULD BE JOINED TOGETHER.
   - DOGGO, PUPPO, PUPPER AND FLOOFER CAN BE JOINED AS EACH DOG HAS ONE TRUE VALUE ASSIGNED TO ONLY ONE OF THEM
   - ALL RETWEETS AND REPLIES SHOULD BE DELETED.

CLEANING DATA

CLEANING DATA THIS PROCESS CAME AFTER CONCLUSIONS ON ASSESSING DATA TRYING TO FIX QUALITY AND TIDINESS ISSUES LIKE:

- CORRECT WRONG DATA TYPES.
- DELETE EXTRANEOUS COLUMNS ON DATA.
- FIX NULL PROBLEM …ETC

IT CONTAINS THREE MAIN PARTS:

- DEFINE.
- CODE.
- TEST.

BEFORE CLEANING, IT IS BETTER TO WORK ON A COPY OF DATA.

AFTER GETTING A CLEANDE DATA WE SHOULD EXPORT IT IN CASE OF WE USE IT AGAIN.

WE CANNOT DO ANALYSIS OR VISUALIZATION WITHOUT DOING WRANGLING PROCCES.