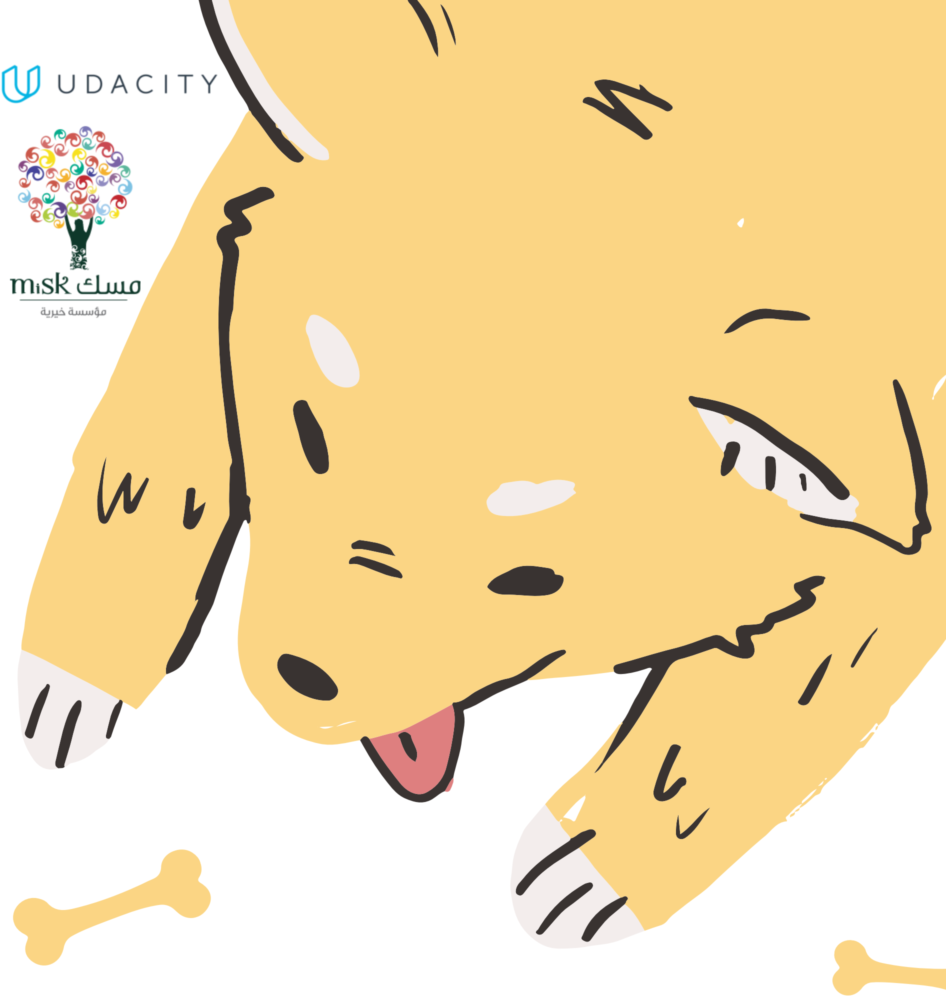

WRANGLE REPORT

MALAK ALMOHAMMADI

Project Goal

wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

Data Gathering

Data Gathered from 3 resources as :

df1: twitter_archive_enhanced.csv

df2: image_predictions.tsv

df3: Using the tweet IDs in the WeRateDogs Twitter archive, and query the Twitter API for each tweet's JSON data.

Assess Data

Quality

Issue1: df1 -> Unnecessary column(retweeted_status_id, retweeted_status_user_id).

Issue2: df1 -> Wrong data type (in_reply_to_status_id,in_reply_to_user_id) most be string not float.

Issue3: df1 -> Wrong data type (timestamp) most be datetime not string.

Issue4: df1 -> Some rows has wong rating

Issue5: df2 -> wrong column names (1,2,3,...)

Issue6: df2 -> Wrond data type (p1_conf,p2_conf,p3_conf) most be float not string

Issue7: df2 -> Wrong data type (p1_dog,p2_dog,p3_dog) most be boolean not string

Issue8: df3 -> All values is null in (contributors,coordinates,geo) and (place) have only one value

Issue9: df3 -> Sources not readable.

Assess Data

Tidiness

- Issue1: df1 -> The dog stage stored in many columns.
- Issue2: df3 -> df2 and df3 should be in df1.
- Issue3: df1 -> Create a column name (rating)equal to rating numerator/rating denominator. Drop rating numerator and rating denominator.

Clean Data

Quality

- Create a copy of all data frames as (df1_c,df2_c,
- df3_c)
- Issue1: df1 -> Unnecessary column

  Define Issue1: Drop Unnecessary column
- Issue2: df1 -> Wrong data type (in_reply_to_status_id,
- in_reply_to_user_id) most be string not float

  Define Issue2: Convert (in_reply_to_status_id,
- in_reply_to_status_id) from string not float

Clean Data

- Issue3: df1 -> Wrong data type (timestamp) most be DateTime not string
  Define Issue3: Convert (timestamp) from string to DateTime
- Issue4: df1 -> Some rows have wong rating
  Define Issue4: Drop wong rating
- Issue5: df2 -> wrong column names (1,2,3,...)
  Define Issue5: Change column names from (1,2,3,...) to right names
- Issue6: df2 -> Wrond data type (p1_conf,p2_conf,p3_conf) most be float not string
  Define Issue6: Convert (p1_conf,p2_conf,p3_conf) from string to float
- Issue7: df2 -> Wrong data type (p1_dog,p2_dog,p3_dog) most be boolean not string
  Define Issue7: Convert (p1_dog,p2_dog,p3_dog) from string to bool

Clean Data

- Issue8: df3 -> All values is null in (contributors,coordinates,geo) and (place) have only one value
  Define Issue8: Drop Column (contributors,coordinates,geo,place)
- Issue9: df3 -> Sources not readable.
  Define Issue9: Extract sources categories from Source column

Tidiness

- Issue1: df1 -> The dog stage stored in many columns
  Based on what I learned in this course the faster and effective way to handle the Pandas data frame problems is using Numpy package functions. From our data, the dog stage represent in 4 columns, each row has one stage and the others is none, So using a where() function is very useful here.
  Define Issue1: Create a column (DogeStage). Ectract not null value from each column. Drop (doggo,floofer,pupper,puppo)

Clean Data

- Issue2: df1 -> There's two column for (rating) rating_numerator, rating_denominator.
  Define Issue2: df1 -> Create a column name (rating)equal to rating numerator/rating denominator. Drop rating numerator and rating denominator.
- Issue3: df3 -> df2,df3 should be in df1.
  Define Issue3: Merge df1 with df2 and df3 (left join)