# ***Data wrangle***

## *1-Data gathering*

***we have 3 data frame :***

**1-twitter_archive** :

This data frame is downloaded manually from audacity resource.

**2-image_predictions:**

This data frame is downloaded  programmatically using request library.

**3-api_df:**

-Since I prefer not to create twitter account for personal reasons I choose the second approach.

-In this approach we download the json-tweet.txt file from udacity resource.

-since this file is large we will read it line by line and each line will represent an observation in the data frame that we want to create .

-we convert each line from string into a dictionary and from the dictionary we can easily get the data of each column we will use in the dataframe and append each observation's data into dictionary and append it  into the list that we will lastly convert it to dataframe.

**Output :** twitter_archive,image_predictions,api_df

## *2-assessment*

***Visual assessment:***

-I just try to open the csv file of twittwer_archive and notice the none values of both dog'classifer and name .

- i opened the tsv file and noticed the floating value of confidence columns.

-i opened the api_txt file and it was totally hard to be read

**Programmatic assessment**

I use pandas functions like head,info ,shape,describe and value_counts.

**Quality:**

**`twitter_archive` table**

- tweet_id string not int (consistency issues):
- timestamp datetime not object (consistency issues):
- it is only tweets not retweets (completeness)
- it is only tweets not reply(completeness)
- retweets and reply columns are useless now (consistency issues):
- it is tweet without image(completeness)
- in expanded_urls columns remove missing values(completeness)
- in dog classifier columns change none into ‘’ (consistency issues)
-in name column change none into ‘’(consistency issues)

**`image_predictions` table**

- tweet_id string not int(consistency issues):
- name of prediction and confidence columns is not readable (consistency issues)
- confidence columns refer to accuracy of prediction so it should be *100 then convert into int to be statistically (consistency issues)
- tweet_id has to be the same of twitter_arcive after dropping(completeness)

**`api_df` table**
- tweet_id string not int(consistency issues):
- tweet_id has to be the same of twitter_arcive after dropping(completeness)

**Tideness**

**twitter_archive**
1- instead of (doggo, floofer, pupper, puppo) only use one column for dog_bread.
2- since Each type of observational unit forms a table , so we don't need to have api and archive separable since we deal with the same type of observational unit.

## 3-cleaning

- cleaning has 3 steps: define , code and test.
- in cleaning we usually start with drop missing value and completeness issues
- then we deal with tiredness issues.
- then dealing with quality issues.

| **Quality issues** | **solution** |
|---|---|
| Tweets we deal with are only tweets without retweets.(completeness) | Use retweet columns that contain values not null as an indication of retweets and drop these rows form all data frames. |
| Tweets we deal with are only tweets without reply.(completeness) | Use reply columns that contain values not null as an indication of reply and drop these rows form all data frames. |
| Retweets and replies columns are useless now. | We will drop these columns. |
| Expand_url column has nan values so we .(completeness) | We will drop all nan value rows from all data frames. |
| We deal with tweets with images .(completeness) | We will drop all rows not existed in image_perdiction data frame. |
| Twitter _id is int value and we don't use it in any arithmetic operation .(consistency) | We will convert it into string in all data frames. |
| Datastamp column in twitter_archive is in object form . (consistency) | We will convert it into datetime . |
| Dog classifier columns in twitter_archive data frame contain "None".(consistency) | We will convert None into "" |
| Dog classifier columns in twitter_archive data frame contain "None".(consistency) | We will convert None into "" |
| Prediction and confidence columns in image_prediction have unreadable names.(consistency) | We will change these names into readable format. |
| Confidence columns in image_prediction are in float format not easy in statistical | Since confidence refers to the accuracy of the prediction so we will use it in a |

| | |
|---|---|
| use. | percentage format multiple by 100 and convert it into integer . |
| Rating columns in twitter_archive have invalid values. | Some of the values need to be modified manually and others need to be dropped and others refer to not a dog image . |

| **Tidiness issues** | **solution** |
|---|---|
| In dog classifier columns the column header contain a value | We will use only one column as dog_bread. |
| we don't need to have api and archive separable since we deal with the same type of observational unit. | We will merge the two data frames since they deal with the same observational unit. |

## Store

After the cleaning step we will merge twitter_archive and image_predition in the one dataframe to be used in analysis step
Then we will store this data frame in csv file called  twitter_archive_master.csv