Data Wrangling Report

By Maria George

1-Gathering The Data :

- Download the file **manually** by clicking the following link: **twitter_archive_enhanced.csv** then read into a dataframe with a name archive_df using the pandas library
- Second dataset “Image_predication.tsv” I download it programtically using the request library following URL “ https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv” then reading the file then read it into a data frame called Image_predication_df
- Third Dataset I can’t do twitter developer account as many issues appear so I download the file in the class room “tweet_json.txt” read it line by line form a dataframe called api_df with three coumns tweet_id , retweet_count , favorite_count

- So the output from this stage is three dataframes archive_df , Image_predication_df , api_df

2- Data Assessment

□ in this step , we investigate our imported dataset both visually and programmatically for quality and tidness issue

□ the visual assessment done on spreadsheet application like excel and then programmatic assessment is conducted on Jupiter notebook

□ Missing data were addressed first then messy structured were addressed to facilitate the tackling of the rest of quality issues that fall in bucket of validity , accuracy , inconsistency class od data quality aspects

□ Some of data cleaning efforts were guided by the scope of the project that mandated the exclusion of retweets and replies and tweets featuring no images

Table	#	Issue	Solution
archive		Quality ISSue	
	1	Data type(consistency issue) All timestamp is object type	Type conversion to datetime type
	2	There are retweets and replies in dataset	Remove those tweets by slicing and comparing with image predication dataset
	3	Error in names like a and an	Their relevant retweet were reinvestigated and the correct names were extracted if existed
	4	Missing entries in expanded_urls	Dropped as those don't feature image
	5	Incorrect and weired values of the rating numerator wich has a maximum of 1776 , the same holds as for the rating denominator with illogical maximum of 170	Absurdly high values (there were two) were deleted others were closely investigated the correct values were extracted programmatically and manually
	6	Source value have uncleaned data with a full <a> tag we need only the source	Optimize the source content by 'Twitter for iphone', 'Twitter Web Client', and 'TweetDeck
	7	Some names begin with capital and other with small	Captilize all
Image_predication_df	8	The column name p1,p1_conf ,p1_gog ,p2 ….. are non descriptive names	Change the column name for better readability
twitter_archive_master	9	All tweets id are integer (rather than change it in each df before Merging them together)	Type conversion to String

		Tidness Issue	
Archive	1	Values are column names [doggo,floofer,pupper,puppo]	Combined in one column called stage as it is a dog stage
Api_df	2	This isn’t considered an observational unit to have it’s own table	Merged to the archive data table