DATA ANALYST NANODEGREE
PROJECT: WRANGLE AND ANALYZE DATA

WRANGLE REPORT

Berk TEZKOSAR

Istanbul, TURKEY

2020

DATA GATHERING

We will obtain data from three different sources:

- Manually downloaded "Twitter_Archive_Enhanced.csv" file
- Programmatically downloaded "image-predictions.tsv" file
- Scraped data using Twitter API.

CRUCIAL CRITERIA

To select the appropriate data, we are supposed to use the following criteria:

- Retweets shouldn't contain details
- Just tweets which have pictures

In addition, reply tweets that can also contain improved / downgraded dog ratings are also included in the database. This implies that in some cases, there are two observations / scores for the same dog. As a result, I preferred to include only original ratings and thus established an additional criterion:

- The dataset should not include responses

DATA CLEANING

Multiple quality and tidiness issues were identified for the three tables.

Missing data issues are addressed first**.** ***Tidiness*** issues were addressed second and remaining ***quality*** issues were addressed in the third. Details of the issues identified and solutions are found in the following table.

Archive Table	Quality	**1**	Retweets are included in the dataset	Used for loop and .str.contains() to re-identify if text contains each column header.
		2	Replies are included in the dataset	Used for loop and .str.contains() to re-identify if text contains each column header.
		3	Erroneous datatypes in multiple columns	tweet_id changed to str, dog_type changed to categorical type, timestamp changed to datetime
		4	Missing names are identified in the text and name columns	Created a function to identify all pet names and add them to the name column
		5	Some entries in the name column are not names	Identified but not handled
		6	Some posts do not have images	Rows deleted
		7	The "text" column includes both text and short link	Created a function to remove links and applied to the "text" column"
		8	Values in the rating_numerator column are incorrect	Created a function that identifies the value before the last '/' in the "text" column"
		9	Values in the rating_denominarator column are incorrect	Created a function that identifies the value after the last '/' in the "text" column"
	Tidiness	**10**	Multiple columns containes the same type of data ("doggo" to puppo")	Created "dog_type" column and filled with dog types, .fillna() and dropped unneeded columns
		11	Unneeded columns ("in_reply_status_to","in_reply_user_id", "retweet_status_id", "retweet_status_user_id", "retweet_status_timestamp")	Dropped unneeded columns
Predictions	Quality	**12**	Erroneous datatypes in multiple columns	tweet_id changed to str, prediction_order changed to categorical
		13	Reduced entries shows that some entries in the archive does not have images	Removed any tweets without an image
	Tidiness	**14**	Multiple columns have same kind of data in "p1" to "p3"	Column names changed and merged into a single column using pd.wide_to_long funciton
API Data	Quality	**15**	Erroneous datatypes in multiple columns	tweet_id changed to str
		16	Retweets and favorites data is missing for some tweets	Used for loop and .str.contains() to re-identify if text contains each column header.
	Tidiness	**17**	Two different tables occurred, tweet information is in the other table	Used a left join to merge api_data with archive_data table on"tweet_id" column