By: Louaye Lamaa

Udacity Data Analyst Nanodegree

Project 4: Wrangling and Analyzing Data

Data Wrangling Report

This report is a documentation for the project 4 of the NanoDegree. This project is focused on data wrangling step of the Data Analysis process

1-Gather the data

The data of this project was gathered from three different sources:

- The first source is a flat file (CSV) given by Udacity. This file was downloaded to my computer like any other file. This file contains the archive data of the WeRateDogs twitter page. (df_archive)
- The second source is a TSV file, downloaded programmatically using the python *Requests* library. This file contains three predictions with their confidence levels of the images uploaded by the page on Twitter. (df_predictions)
- The third source is API twitter. The API was fetched using the *Tweepy* library in order to download the data about the tweets of this page. The downloaded data was stored in a text file using the JSON library. This file contains data about retweets and favorites count of each tweet. (df_tweets)

2-Assess the data

This step is divided into two parts: visual assessment and programmatic assessment. The issues found are: 13 quality issues and 5 tidiness issues.

3-Clean the data

I started by cleaning the completeness issues in the df_archive dataset, such as the missing values in some columns (columns are in the table below). I

dropped the columns that contains missing values as well as the columns that will not be used in the analysis. After that, I addressed the tidiness issues. I merged the four dog stages into one column called dog_stage and split the timestamp into date and time columns. Note that I transform the timestamp datatype to datetime before the split even though the datatype is not a tidiness issue. One tidiness issue that I found and I thought it should be cleaned (I am not sure about it because it will create tweet ids duplicates), is that some tweets contain different information such as name and stage for different dogs in one picture and each tweet stored as one record with inaccurate values for name and stage. I cleaned this issue by creating duplicates for each of those rows and then by updating the incorrect information for each dog in these tweets.

Once the dataset is tidy, I resolved the quality issues. I corrected the incorrect values in the name and dog stage by compared the values of these columns to the text of the tweets. I removed the unwanted data from the text column such as the URL. I filtered the rows that contain 'RT@' (retweets data), I removed the 'RT' from the text, and then I dropped duplicated (with the original tweets and kept the last records (oldest row) that represent the original tweets. I used this logic instead of just dropped the retweets data because of possible retweeted tweets of others users that contain dogs ratings. The most important columns for the analysis in this dataset are the rating_numerator and rating_denominator. I removed the rows with low rating numerator (under 4) of the tweets that don't have dogs, and I corrected the incorrect values by comparing it to the text column. The rating denominator should be always 10. Some denominators are copied wrong from the text column and others are the aggregate of multiple dogs ratings. I converted the datatype of these two columns to float.

In the predictions table, I dropped the columns that I will not used in the analysis and I renamed the columns I will be using to more descriptive names.

There were no issues found in the df_tweet table.

As the prediction and tweets table complete the information presented in the archive data, I merged the three datasets on the matched tweet_id (inner join) as I needed just the tweets presented in the three datasets (datasets were

gathered during different times so the oldest dataset do not contain all the tweets of the newest one.

(651 words)

The table below is just for reference.

A) Quality Issues

Table	Issue	Assessment	Dimension	Cleaning
df_archive table	Missing values and unwanted columns	Programmatic	- Completeness - Validity	Drop these columns `'in_reply_to_status_id', 'in_reply_to_user_id',' retweeted_status_id', 'retweeted_status_user_id' 'retweeted_status_timestamp' 'expanded_urls'`
	Incorrect values in the dog_stage column	Visual and programmatic	Accuracy	Change the dog stage to that written in the tweet text
	Unwanted information in the source column	Visual	Consistency	Remove this information using split function
	URL in the text column	Visual	Consistency	Remove the URL using regular expression module re
	Retweets Records and RT @ in the text column	Visual and programmatic	Consistency – Validity	- Remove the 'RT @' from the text (consistency - Remove the duplicated records resulted (validity)
	Erroneous datatypes for timestamp, rating_numerator,rating_denominator	Programmatic	Validity	Convert ratings to float using astype() and timestamp to datetime
	Incorrect values in rating_numerator	Visual and Programmatic	Validity-accuracy	-Remove records with ratings that don't belong to dogs -Update the incorrect rating compared to the tweet text
	Incorrect values in rating_denominator	Visual and programmatic	Validity-accuracy	Change the denominator to 10
	Incorrect values in name column	Visual and programmatic	Validity-accuracy	Update names by comparing the name column to the tweet text
df_predictions table	Some columns aren't needed for analysis	Programmatic	Consistency	Remove columns as not needed `-'img_num' - 'p2' -'p2_conf'`

				-'p3' 'p3_conf' 'p3_dog'
	Non descriptive columns names	Programmatic	Consistency	Rename the columns: jpg_url → image_url, p1 → dog_breed, p1_conf → breed_confidence, p1_dog → is_dog
	Underscores and no capital letters for all dog breeds	Programmatic	Consistency	Transform each word in the dog breed to capital and remove underscores
df_tweets	No quality issues was spotted	-	-	-

B) Tidiness Issues

Table	Issue	Assessment	Cleaning
df_archive	One variable (dog stage) is stored in four columns (doggo, puppo, pupper, and floofer)	Visual	Merge the four columns in one column called dog_stage, then drop the original columns
	Some rows contain data about more than one dog	Programmatic	Duplicate these rows and then update the information about each dog
	Two variables (date and time) are stored in timestamp column	Visual	Create two columns for date and time, then drop the original column
df_predictions and df_tweets	These tables are not an observational unit, as they complete the archive dataset	Programmatic	Merge the three datasets in one tidy and clean dataframe called df_clean