DATA WRANGLING REPORT

Aisulu Raganina

Udacity Project: Wrangle and analyze Data

Introduction

The project shows proficiency in data wrangling and analyses and includes several steps:

1. Gathering;
2. Assessing;
3. Cleaning;
4. Storing.

1.Gathering

The data was gathered from 3 sources and stored in separate files:

1. *df_tw_archive* was read as CSV file from twitter_archive_enhanced.csv;
2. *image_predictions* was downloaded programmatically;
3. *tweet_info* JSON data was downloaded by querying the Twitter API using the Tweepy library.

2.Assessing

Each dataframe was assessed visually using several pandas functions: head(), tail(), info(), unique(), value_counts(), count(), describe(). The possible duplicates and missing values were checked. The main objectives were to look for the quality and tidiness issues and to understand the data.

The quality and tidiness issues were inspected for further cleaning:

DF	Quality Issues	Tidiness Issues
df_tw_archive	Incorrect data type for timestamp and retweeted_status_timestamp	The columns doggo, floofer, pupper, puppo should be in one column „Stage“
	Invalid values: 'O' to 'O'Malley'; 'Al' to 'Al Cabone'; 'my' to 'Zoey'	
	We want to have only original ratings (no retweets) that have images	
	Column „Name“ includes invalid names "a", "an", "the" etc.	

	There are 'None' missing values in „Name“	
	The columns rating_denominator and rating_numerator have invalid values:	
image_predictions	According to the p1-dog, p2-dog and p3-dog together, there are 324 rows with no dog on the picture	
	The p1,p2 and p3 include Uppercase and Lowercase for the first letter	
tweet_info	The column „id“ should be changed to „tweet_id“ for joining DFs	Join tweet_info with df_tw_archive and image_predictions
	Duplicates were found	

During the cleaning of the data, additional assessment was necessary because after resolving the tidiness issue with column „Stage“ for doggo, floofer, pupper, puppo, the invalid values were identified: doggopupper- 10; doggofloofer-1; doggopuppo-1.

3.Cleaning

For cleaning purposes, the copies of original dataframes were created:

1. df_tw_archive_clean = df_tw_archive.copy()
2. image_predictions_clean = image_predictions.copy()
3. tweet_info_clean = tweet_info.copy()

DF	**Cleaning**
df_tw_archive_clean	Changed to datetime by using pd.to_datetime
	Replaced 'O' to 'O'Malley'; 'Al' to 'Al Cabone'; 'my' to 'Zoey'
	Removed 'retweeted_status_id' which were not null and removed columns 'retweeted_status_id', 'retweeted_status_user_id', 'retweeted_status_timestamp'
	As most of the invalid values for names are starting from the lowercase letter and the text includes the real names, the names were retrieved and replaced to the column „Name“ from „Text“ column with the information after the key words: 'named' and 'name is' .
	Additionally, 'None' missing values were changed to NaN in column „Name“

	The denominator should be 10. In the column „Text“ we can find the correct rating based on 10. Therefore, the denominator and numerator were changed where the text gives the correct rating
image_predictions_clean	324 rows were dropped using drop_duplicates()
	The first letter of dog breeds‘ names replaced with uppercase in columns p1,p2 and p3
	The new column „Stage“ was created in order to store the variables doggo, floofer, pupper, puppo
	The invalid values for doggopupper, doggofloofer and doggopuppo were additionally assessed. There were some NaN in column „Name“ for invalid values doggopupper. Hence, it was impossible to detect the correct dog stage. Thus, the invalid values of doggopupper where column „Name“ included NaN were changed to np.nan Other incorrect values had the correct dog stage in the text and were renamed correspondingly.
tweet_info_clean	The column „id“ was renamed to „tweet_id“ using function rename()
	The duplicates were dropped using drop_duplicates()
	All three tables were joined using pd.merge()

4.Storing

Finally, the new cleaned and merged dataframe was stored as 'twitter_archive_master.csv'.