WRANGLE REPORT

Project: Wrangle WeRateDogs Twitter data

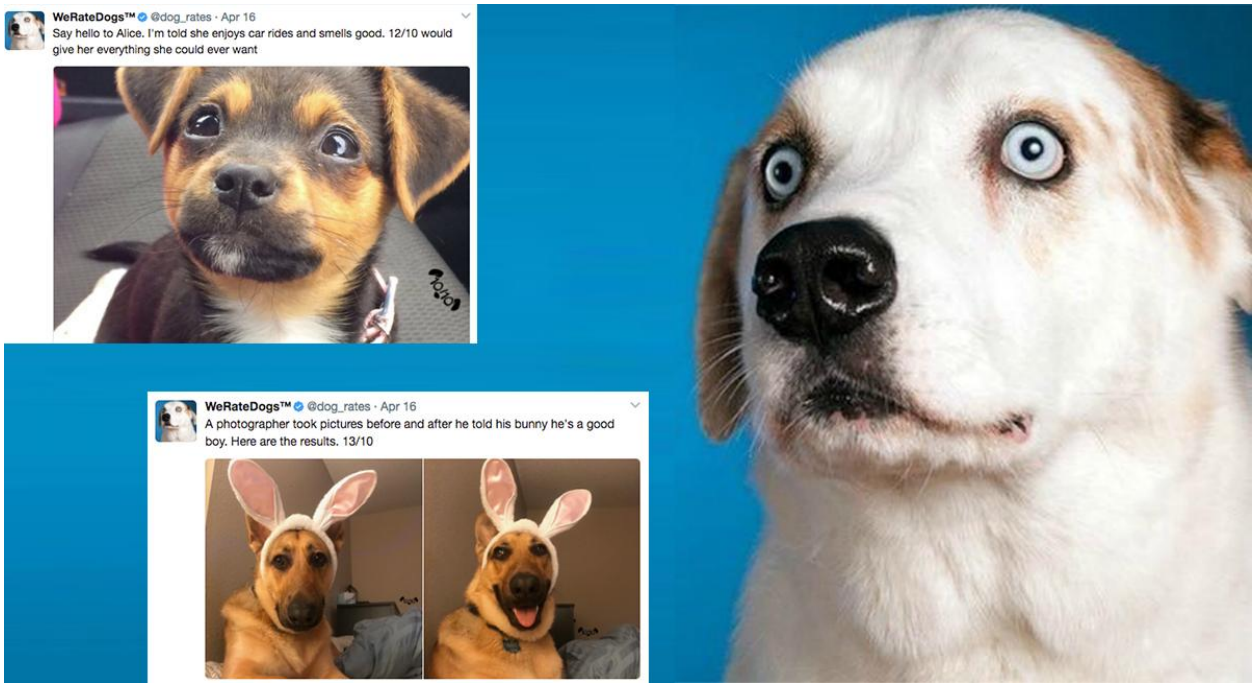

Image via Boston Magazine

Buthaina Alotaibi

06.05.2021

Data Analyst Nanodegree - Udacity

1.

INTRODUCTION

Today's resources of data are very diverse and millions of data generated in mints from social media platforms. It's important to have the knowledge of how to gather these data and maximize its usefulness by preprocessing it to make it efficient for the analysis.

In this project, multiple dataset formats have been dealt with in order to create interesting and trustworthy analysis and visualizations.

GATHERING

For this project, three different files with different formats have been gathered.

1- The WeRateDogs Twitter archive. This file has been downloaded manually.

2- The tweet image prediction. It's a flat file that has been downloaded programmatically using Requests library.

3- Additional file that contains additional data for the WeRateDogs Twitter archive. Has been downloaded using Tweepy library and Twitter API. For this file we have extracted only the needed columns which are: 'id','favorite_count' and 'retweet_count'.

ASSESSING

To assess the quality and tidiness issues two different methods have been used, visual assessing and programmatic assessing.

Quality Issues:

The quality issue	File	Column	Dimension
1. None instead of NaN	twitter-archive-enhanced.csv	name,doggo,floofer, pupper,puppo	Validity

2.Wrong name entry	twitter-archive-enhanced.csv	name	Validity
3.Retweet and replies doesnt have real ratings	twitter-archive-enhanced.csv		Validity
4.Wrong column type for tweet_id	All files	tweet_id	Accuracy
5.Some predicted names starts with small letter and others with capital letter	image_predictions.tsv	p1, p2, p3	Consistency
6.Underscore instead of space	image_predictions.tsv	p1, p2, p3	Consistency
7.Wrong rating denominator	twitter-archive-enhanced.csv	rating_denominator	Consistency
8.Missing data	twitter-archive-enhanced.csv	name	Completeness
9.Unused columns			
10.Timestamp column has two different variables			

Tidiness Issues:

1. The columns (doggo, floofer,pupper,puppo) related to the same variable
2. The two tables (image_predictions.tsv, tweet-json.txt) related to the same observational units in 'twitter-archive-enhanced.csv' table

CLEANING

ISSUE	DEFINE	CODE
Quality		
1.	None changed into NaN in the columns (name, doggo, floofer, pupper, puppo)	.replace('None', np.NaN, inplace=True)
2.	Names starts with lowercase dropped	master_df['name'].str[0].str.islower()
3.	Created dataframe with values in 'in_replay_to_status_id' and 'retweeted_status_id' then dropped	master_df.drop(remove_replies.index) master_df.drop(remove_retweet.index)
4.	tweet_id column change to object instead of inger	.astype('object')
5.	Prediction capitalized	.str.capitalize()
6.	Underscores replaced with space	.replace('_', ' ')
7.	Created dataframe with wrong rating_denominator then dropped it from master.df	master_df.drop(wrong_den.index)
8.	Rows with missing names dropped	master_df.dropna
9.	Unused columns dropped	master_df.drop()
10.	'timestamp' splitted into 'Time' and 'Date' columns	pd.to_datetime
Tidiness		
1.	New column 'stage' created to aggregate	IN [56]
2.	The three files merged	.merge

REFERENCES

1. https://stackoverflow.com/questions/35595710/splitting-timestamp-column-into-separate-date-and-time-columns
2. https://stackoverflow.com/questions/33098383/merge-multiple-column-values-into-one-column-in-python-pandas