Udacity
Data Analyst Nanodegree Program
Data Wrangling
Wrangle and Analyze Data

Wrangle Report

Wrangling WeRateDogs

Done by:

Zyad alatar

CONTENT

❖ Introduction:

In this project I will put the fundamentals learned in videos in order to complete the project. In this project I will be wrangling & analyzing WeRateDogs data set by first gathering data then assessing the data and finally cleaning data.

❖ Gathering:

Gathering phase is divided into 3 steps:

1) Manually downloading (twitter_archive_enhanced.csv) file
2) Programmatically downloading (image_predictions.tsv) file using Python's Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3) I will use Python's Tweepy library and store JSON object data of each tweet in a file called tweet_json.txt

❖ Assessing:

In this phase I have done both visual and programmatic assessment in order to get the data in the best quality and tidiness and I came with this results:

Quality problems :

1)Retweet will be dropped ,we only care about tweets
2)unused columns will be dropped
3)changing the data type of rating_numerator into float
3.1)changing the data type of rating_denominator into float
4)changing the data type of timestap into datetime then split them
5)changing the data type of doggo , floofer , pupper , puppo into category
6)changing the data type of tweet_id from integer to string
7)name attribute has None as missing value instead of Nan
8)source column is not clear enough ,needed to be clear more

Tidiness problems :

1)There are a number of attributes that needed to merged into a column because splitting them is making the analysis harder & to have a higher cohesiveness for the dataset. The attributes are: doggo , floofer , pupper and puppo

2)creating one dataset that contains all three dataset

❖ Cleaning

In cleaning part, I fixed quality &tidiness issues arise above using the great fundamentals that Python has. This part was divided into define, code and test for each quality and tidiness issue found. Define part includes defining the problem and how to handle it. Code part includes the code that solve that problem. Test part includes ensuring that the problems have been solve.

❖ conclusion:

In conclusion, this project was amazing the I have learned a lot and I will be happy to practice what I learned in my career.