Ailin Valdes
27th May 2020

Project:

Wrangle and Analyze Data

INTRODUCTION

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention for its popularity. [1]

In this project I will be wrangling, analyzing and visualizing the tweet archive of this Twitter account following the next steps: 1) Gathering, 2) Assessing and 3) Cleaning.

1) GATHERING

Three files were read to analyze the information for wrangling analysis:

A) twitter_archive_enhanced.csv: This file is provided for Udacity and downloaded manually.

B) image_predictions.tsv: This file is hosted on Udacity's servers and downloaded programmatically using the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

C) tweet_json.txt: This file is downloaded from Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive.

2) ASSESSING

Initially, the data was visually analyzed in Excel. Then a more detailed analysis was done programmatically in the Jupyter Notebook. During the programmatic analysis different instances were used to view specific portions and summaries of the data. Ten quality issues and two tidiness issues were detected in the Jupyter Notebook:

- QUALITY

  twitter_archive table:

  1) In the dataset there are retweets and replies. For our analysis the original tweets are needed.
  2) The rating numerator and denominator have wrong values.
  3) The names have invalid values.
  4) There is a URL that does not provide information in the tweet text.
  5) Null values are not detailed as such (NaN) in the different dog types.
  6) There are columns that are not used for analysis.

[1] https://en.wikipedia.org/wiki/WeRateDogs

7) Timestamp format is an object instead of datetime

image_predictions table

8) There are duplicate values in jpg_url.
9) There are various confidence columns and image estimates. The most significant one must be left for analysis.
10) There are columns that are not used for analysis.

- TIDINESS

tweet_json table

1) Tweet_id format is an object instead of datetime.
2) The three datasets must be joined in a single table

3) CLEANING

The cleaning was divided into Define, Code and Test following the steps that were explained in the chapter.

CONCLUSION

Data wrangling provides us with the correct information for our analysis. To do this, a strategic analysis must be carried out in order to determine the correct information for our objectives.

[1] https://en.wikipedia.org/wiki/WeRateDogs