WeRateDogs Data Wrangling Report

Wrangle & Analyze Data

This project was completed to fulfill the course requirements of Udacity's Data Analyst Nanodegree certification.

Overview

The project uses data from the WeRateDogs Twitter account. The data was gathered, assessed, cleaned and analyzed to provide accurate insights into Twitter account that rates people's dogs with funny comments about the dog and account follower behavior.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Software Needed

Need to be able to work in a Jupyter Notebook.

The following packages (libraries) need to be installed:

- pandas
- NumPy
- requests
- tweepy
- matplotlib
- json

Project Details

The tasks in this project are as follows:

Data wrangling, that consists of.

- ❖ Gathering data.

- ❖ Assessing data.
- ❖ Cleaning data.
- ❖ Storing, analyzing, and visualizing your wrangled data
- ❖ Reporting on
    a. Data wrangling efforts.
    b. Data analyses and visualizations.

Gathering Data

Data was gathered from 3 different sources:

1. From WeRateDogs twitter archive given by Udacity in csv format.
2. Image predictions file downloaded programmatically using Requests library and the URL provided by Udacity in tsv format.
3. Data retrieved by querying Twitter's APIs to get tweet’s Json data using Tweepy library.

Assessing Data

After gathering the data and storing them in DataFrames, the following step was assessing the data for quality and tidiness. Data were assessed visually and programmatically.

Data quality for:

1. Completeness.
2. Validity.
3. Accuracy.
4. Consistency.

Programmatically:

- ✓ Detecting an issue.
- ✓ Documenting issue.

After gathering the data and storing it in dataframs, the data assessed for quality visualy and programmatically.

Quality

Issues with content. Low quality data or dirty data.

Identified quality issues:

- twitter archive
    - Missing Values :
        - in_reply_to_status_id, in_reply_to_user_id : have 2278 Null Values.
        - retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp: have 2175 Null Values.
- tweet_id of integer datatype.
- in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id, retweeted_status_user_id are float datatype.
- timestamp , retweeted_status_timestamp are object datatype (str).
- Source column is written in html contain (tags).
- Columns doggo,floofer,pupper, puppo has None values instead of Null.
- Column name: there are some inaccurate values.
- image_predictions
    - Have missed data (only 2075 observations).
    - tweet_id are of integer datatype.
    - some images are not related to dogs
- tweet_json

- Column id is of integer datatype.
- Has missed values.

Tidiness

Issues with structure, untidy data, also known as messy data.

Identified tidiness issues are:

- No need for Dog stage is in 4 columns (doggo, floofer, pupper, puppo).
- Columns p1, p1_dog, p1_conf , p2, p2_dog, p2_conf , p3, p3_dog, p3_conf should be renamed full name instead of abbreviations in image_predictions dataset.
- Name of tweet_json (column id) differ from other datasets.
- Need to Merge Dataframes.

Cleaning

It is the process of fixing and resoling issues identified in the assessment process.

The (define, code, and test) steps were performed in the cleaning process. A copy of Dataframes were created before cleaning. Then the steps of cleaning were applied iteratively on all issues.

Storing

The Dataframes Merged in one Datafram then stored in a csv file called " twitter_archive_master.csv".

At this point, the data was successfully wrangled and therefore it was ready for analysis and visualization in separate file called 'act_report.pdf '.