wrangle_report

September 1, 2022

1 WeRateDogs DAND Project

1.1 A Data-Wrangling project by Matthew Olushola Odebunmi)

1.2 Introduction

WeRateDogs by Matthew Olushola Odebunmi Data wrangling project is a coursework in udacity Data Analysis Nanodegree program.

Objective: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations

The WeRateDogs Data wrangling consist of these following processes: 1. Gathering Data 2. Assessing Data 3. Cleaning Data 4. Store Data 5. Analyzing and Visualization

1. Gathering Data: The gathering of data is obtain from three different sources. The first dataset, which is the largest dataset was downloaded manually as a CSV file named 'twitter_archive_enhanced.csv'. The second dataset This file is called 'image_predictions.tsv' is present in each tweet according to a neural network. It is downloaded programmatically using the *Requests library* . The file is in tsv format which needed to be open csv format. The last data set stored as 'tweet_json.txt' is query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

2. Assessing Data: Assessing the 3 data set is done visually and programmatically in jupyter note book for quality and tidiness issues. The aim is to detect and document at least eight (8) quality issues and two (2) tidiness issues in the "Accessing Data" section in the wrangle_act.ipynb Jupyter Notebook. I deploy two types of assessment: **Visual assessment**: each piece of gathered data is displayed in the Jupyter Notebook for visual assessment purposes. Once displayed, data can additionally be assessed in an external application (e.g. Excel, text editor). **Programmatic assessment**: pandas' functions and methods are used to assess the data. such as tweet_df.info(), tweet_df.head(), tweet_df['timestamp'][0], etc. A

3. Cleaning Data: Cleaning all the issues and document them.Before i perform the cleaning, make a copy of the original data. as 'tweet_df_copy', 'image_pred_df_copy' and ’tweet_json_df_copy’. I use the define-code-test framework and clearly document it. i obsered during Cleaning includes merging individual pieces of data according to the rules of tidy data. The result was a high-quality and tidy master pandas DataFrame (or DataFrames, if appropriate).

4. Storing Data: The tidy master dataset ’twitter_archive_master_df’, was stored as a CSV file named "twitter_archive_master.csv".

5. Analyzing and Visualization: The tidy master dataset was anlyse and visualize to produce insight of the wrangling dataframe