wrangle_report

July 8, 2022

0.1 Reporting: Wrangling Efforts

0.2 Introduction

This is a report on the Data Wrangling efforts for my second project in the ALX-T Data Analyst Nanodegree Program on Udacity. The aim of the project is to gather & analyse the data of a Twitter account named @dog_rates. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. To analyze the data and generate insights, we must first go through a data Wrangling Process to ensure we're generating a trustworthy analyses and visualizations. #### The Data Wrangling Process is in stages, these include: 1. Data Gathering 2. Data Assessment 3. Data Cleaning 4. Data Storing

This report will be structured as sub-headings under the different Data Wrangling stages

0.2.1 Data Gathering

The data utilized for the analysis were gathered from three different sources; these include: 1. A CSV (comma-separated values) file that was provided by Udacity for WeRateDogs Twitter archive, this was downloaded from the Udacity Website. * After Downloading this file to the local device; it was read into a pandas DataFrame named `twitter_archive` using pandas.read_csv. 2. A TSV file that was downloaded programatically from the Udacity Server, this contains predictions about the image posted by WeRateDogs. It was downloaded programatically using Requests library. This was read in a Dataframe named `predictions` 3. Additional data were retrieved through on Twitter using the Twitter API library Tweepy. Each tweet's JSON data were written to its own line and stored in a text file named `tweet_json.txt`. This text file was then read line by line into a pandas DataFrame named `twitter_additions`

0.2.2 Assessing Data

After gathering the three datasets & reading into the Pandas Dataframe, The datasets were assessed visually and programmatically for quality and tidiness issues. The steps for this included: 1. **Visual Assessment**: To make the Visual assessment easier the datasets that were downloaded programmatically were read into a CSV file, therefore visual assessment could be done on the three datasets using a spreadsheet application (Excel) for easier assessment. > During these steps, rows that were suspected to have issues were further investigated by loading the `expanded_url` for that particular Tweet to investigate the issue that was identified. One of those was the rating_numerator for 420; the score was inflated therefore further investigation was done to load the Url for the Tweet, this showed that the rating was not a Dog's rating but rather a human picture (named Snoop Dogg). 2. **Programmatic assessment**: Further assessment was done on the Datasets using code to view specific portions and summaries of the data. Pandas & python libraries was used here. > All the issues that were identified were recorded and will be Cleaned during the Data Cleaning Stage of this Data Wrangling process.

0.2.3 Cleaning Data

The Quality and Tidiness issues that were identified and documented during the assessment stage were cleaned properly here, during the cleaning stage other issues were identified therefore the Assessment & Cleaning were repeated multiple times. > The Define-Code-Test framework was used while cleaning, this will make it easier for other analyst to go over the workdone to understand fully the cleaning steps.

0.2.4 Storing Data

The wrangling steps were completed and the three Datasets were merged together. This merged dataset was saved to another dataframe named `twitter_archive_master.csv`. This merged dataframe was used for the analysis.