wrangle_report

February 13, 2023

1 Wrangle and Analyze Data

1.1 Reporting: wragle_report

Prepared by: Nouf AlGhamdi

1.1.1 Table of content:

1. Project Overview
2. Gathering Data
3. Assessing Data
4. Cleaning Data
5. Storing Data

6. Visualizing Data
7. Project Challenges and Limitations

1.1.2 1. Project Overview

Wrangling data, also known as data cleansing, data remediation, or data munging, involves transforming raw data into easily usable formats. Methods vary depending on the data being utilized and the goal being pursued.

The dataset that we have wrangled (and analyzed and visualized) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs along with a humorous commentary about the dog. The WeRateDogs account has over 4 million followers and has been covered by international media outlets.

1.1.3 2. Gathering Data

This project uses three different datasets, which were obtained in the following manner: #### 1. The WeRateDogs Twitter archive: To use this dataset we Downloaded the file twitter_archive_enhanced.csv manually by clicking on the provided link Once it is downloaded, we upload it and read the data into a pandas DataFrame **twitter_archive** . #### 2. The tweet image predictions This file (image_predictions.tsv) is present in each tweet according to a neural network. It is hosted on Udacity's servers so we downloaded programmatically using the Requests library and the provided URL, and stored it into **image_predictions** DataFrame #### 3. Additional data from the Twitter API We also gathered each tweet's retweet count and favorite ("like") count and additional data we find interesting like the folower count, and the friends count. Using the tweet IDs in the WeRateDogs Twitter archive, and have query the Twitter API for each tweet's JSON data using Python's Tweepy library and storeed each tweet's entire set of JSON data in a file called tweet_json.txt file. We store this additional data into the DataFrame **tweet_detailed_data**

1.1.4 3. Assessing Data

After gathering all the three pieces of data, we assessed them visually and programmatically for quality and tidiness issues. we have Detected and documented **ten (10) quality issues** and **four (4) tidiness issues**, The issues identified during the assessment phase were documented into the **Assessing Data** section of a Jupyter notebook and categorized into Quality Issues and Tidiness Issues

1.1.5 4.Cleaning Data

In this section, we cleaned all of the issues we documented while assessing. also as an anssioal step we have a copy of the original data before the cleaning. And best approach to clean data in this three steps: * Define — express in words how you intend to resolve the problem. * Code — turn your definitions into executable code. * Test — test your data to confirm that your code was properly implemented.

Once the data had been cleaned, individual pieces were merged according to the rules of tidy data into a high-quality and tidy DataFrame **merged_dataset**

1.1.6 5. Storing Data

In this section we have stored the cleaned DataFrame in a CSV file named **twitter_archive_master.csv**.

1.1.7 6. Visualizing Data

The final cleaned dataset was used to derive insights and make visualizations. These latter will be presented in the act report file.

1.1.8 7. Project Challenges and Limitations

I found working with Twitter API to be the most challenging aspect of this project. Even though I completed all the steps required to make the project work, I was not granted access to some of the main functions that were necessary to gather the information needed for the project, however the **tweet_json.txt** file was provided to make this task easier.

```
In [ ]:
```