Wrangle Report

Introduction

The aim of this project is handling dirty dataframes from different sources with tidiness issues and use knowledge both from the course and learn one of the most important skills for data analyst (know how to search for the right answer) and for me personally I'm always looking for the simplest code structure to write to get the results I need

Our goal for the project was getting data from different sources regarding a Twitter account @dog_rates, it's a twitter account that rates other people dogs in humorous way, and wrangle all the data, then store it, then analyze the finding and visualize it and finally write a report for the stakeholders.

Steps

1. Data Wrangling
    I. Gathering Data
    II. Assessing Data (Visually & Programmatically)
    III. Cleaning Data
2. Storing Data
3. Analyzing Data
4. Visualizing Data
5. Reporting The Data

I. Gathering Data

Data required to complete the project had different sources not all was provided directly to me

- twitter-archive-enhanced.csv file it was provided directly
- image-predictions.tsv file a link was provided to download from the internet for all students
- json_tweets.csv file which was gathered and stored by myself by using Twitter API tweepy

II. Assessing Data

At first all three data files were inspected visually to understand all data fields and what exactly they are representing and find if there are any missing or odd data fields, then all three dataframes were assessed programmatically for any duplicated values, missing values, inconsistency, wrong data types of columns, outliers, and columns that aren't usable for analysis.
Then all quality and tidiness issues were written down and reported for next steps.

III. Cleaning Data

A copy of the three dataframes were made prior to any cleaning steps to make a checkpoint for our code if any step went wrong

Then each issue reported from the previous step were transformed into code and tested for cleaning the data and our end result is one dataframe with values from all 3 cleaned tables (twitter_archive_master.csv) that is ready for next step (analysis and visualization)