Wrangle Report

Introduction

This project is a requirement for Udacity Data Analyst Nano Degree Program, Data Wrangling and Analysis. In this project, we are gathering, analyzing and visualization data from various data sources.

Project Stages:

This project involves working with tweet data from WeRateDogs twitter account. The account tweets about wonderful dogs and rates them. They have a unique rating system where they rate each dog more than the score. The entire analytics exercise involves 6 stages –

- Gathering
- Assessing
- Cleaning
- Storing
- Insights
- Visualization

- Data Sets
    - In this project, we are gathering three different data sets.
        - twitter_archive_enhanced.csv –
            - This data contains information like tweet_id, timestamp, source, dog stages etc.
        - image_predictions.tsv –
            - This data is provided via URL by Udacity
            - It is being pulled by utilizing th e request package
        - twitter_data –
            - Gathered programmatically via API directly from twitter
            - This data contains information like retweet counts, favorite counts, followers etc.
- Gathering
    - Each data set was gathered in three different methods
        - 1. twitter_archive_enhanced - is manually downloaded from Udacity
        - 2. Image_predictions.tsv - is provided via URL by Udacity
            - We are pulling this data by utilizing requests package and then storing the request response into a data frame
        - 3. twitter_data - programmatically extracted data by utilizing tweepy package, twitter api and personal credentials
            - We stored the tweet_id column in a variable
            - Created a loop to test the api pulling of data for one tweet id
            - We then created a loop for the entire tweet_id list to pull those tweets in json format.
            - Twitter has 15 min intervals condition so we had to apply wait on rate limit logic due to which the loop would stop every time it hit 15 min interval
            - This way we were able to pull the tweet data in json format and then we read it by utilizing pandas' read_json function
        - These 3 steps completed our data collection for 3 different methods
- Assessment
    - In this stage, we are assessing the data sets obtained to identity quality as well as tidiness issues followed by insights and visualization requirements. The assessment involved visual as well as programmatic assessment.
        - Visual Assessment
            - In this assessment, we were able to identify the columns that were needed top be dropped to improve quality of data
            - The tweet id in the twitter_data was incorrectly named, and it needed to be fixed to merge all three data properly

        - Programmatic Assessment - This assessment type helped us with the below issues and insights -
            - Shape of data
            - Columns and their data types for each data set
            - We were able to identify that we can pull follower information which was stored in the one of the columns to twitter_data
            - Dog stages needs to be converted into one column
            - fixing of headers
- Cleaning
    - The cleaning stage was divided between two stages
        - Quality Issues
            - 9 Quality issues were identified
            - Dataset - twitter_data
                - Remove unwanted columns
                - has follower data in json format in one of the columns, we need to extract that information to create a separate column
                - We need to create bins to categorize and analyze followership of the account
                - The twitter_data has tweet id named as id we need to rename it to tweet_id to merge the data
            - Dataset - Twitter Archive Data - we rate dog
                - Remove unwanted columns
                - The timestamp column has erroneous datatype
            - Dataset - Image Predictions
                - Drop unwanted columns
                - Fix column headers for prediction column to make it more relevant
                - Improve consistency for dog breeds
        - Tidiness Issues
            - 2 Tidiness issues were identified
            - This project involves three datasets we will combine them to create one dataset
            - Combining the 4 columns (dog stages to create one column)
- Storage
    - We created a master data frame by combining the three datasets and stored it as a csv
    - This data set was further analyzed to generate insights and visualizations