WRANGLE REPORT

PROJECT: WRANGLE AND ANALYZE DATA

SUBMITTED BY: BILAL ARSHAD CHAUDHRY

INTRODUCTION

This report is part of the seventh project of Udacity's Data Analyst Nanodegree. The purpose of this project was to demonstrate key principles and methodologies related to the data wrangling and cleaning process, which is known to take up a major portion of the Data Analyst's time.

In this project we dealt the dataset provided by the Twitter account @dog_rates, which is also known as WeRateDogs. In short, the tweets in this account contain pictures of people's dogs, a rating on a scale from 1 – 10 and humorous comments about the dogs in question. The project was executed using Jupyter Notebooks and contains the following sections:

1. Gathering Data
2. Assessing Data
3. Observations and Course of Action
4. Cleaning Data
5. Generating Insights

GATHERING DATA

The data used in this project was gathered from three different sources, which in turn made up three data sets. These were:

1. **Twitter Data:** We were provided the master archive called "twitter_archive_enhanced.csv" by Udacity.
2. **Image Prediction Data:** This file contains a classification of dog breeds based on the output results of a Convolutional Neural Network, which analyzed the images of dogs in each tweet.
3. **Twitter API and JSON Data:** This data was gathered by querying the Twitter API for each tweet's JSON data, using Python's "Tweepy" library. The data was then stored in a file valled "tweet_json.txt". The file was then read, line by line, into a dedicate pandas dataframe. It must be noted that since I wasn't able to gain access to a Twitter developer account, I proceeded to manually download the file from Udacity's project resources section.

Gathering, by itself, can prove to be pretty challenging. This is especially true for scraping website APIs for JSON data. However, even in the face of incomplete documentation and support on various websites on how to access their APIs, I took comfort in the fact that there was a number of highly detailed blogs and YouTube tutorials which clarified the process for a beginner like myself.

ASSESSING DATA

Following the gathering phase, where the datasets were loaded into their respective data frames, I then proceeded to assess the data by visual and programmatic means.

1. **Visual Assessment:** To conduct visual assessments of the three data frames, the data frames were accessed and assessed, line-by-line, in the Jupyter Notebook itself. Some casual assessments were also carried out using Microsoft Excel, since these were reasonably small and manageable data sets.
2. **Programmatic Assessment:** Pandas provides users with numerous methods to assess data programmatically in an efficient and speedy manner. For the purposes of this project, methods like .info(), .duplicated(), .isnull(), value_counts(), contains() with regex statements, .query(), .isupper(), .islower() etc. were used on multiple occassions. Finally, the .sample() method was also used to generate random samples of the data for further assessments.

OBSERVATIONS AND COURSE OF ACTION

Once all visual and programmatic assessments were carried out, I then proceeded to note observations pertaining to tidiness and quality. As a summary, 18 data quality issues were highlighted in the data frames in addition to 2 tidiness issues.

DATA CLEANING

Once all observations had been made and a course of action was finalized, I then proceeded to clean the data. But, before proceeding any further, copies of the data frames, which would be worked upon, were then created in accordance to data wrangling best practices.

This process was broken down into individual steps to maximize readability for future use. Each step could then be broken down into three parts i.e. define, code and test. Special care was given to ensuring that only original tweets were used, instead of retweets. Data types were also corrected for various columns in accordance with the information they contained. I also chose to standardize the ratings columns, which were split into numerators and denominators, by choosing entries with denominators fixed at 10, while numerators with decimals contained value errors which were also corrected.

Some parts of the cleaning exercise were more challenging than others, for e.g. using functions, for loops and nested if-elif-else statements to replace incorrect names and to consolidate the image prediction results into single columns. Once all of this, in addition to other smaller corrections, was carried out, I consolidated the three cleaned data frames into one master data frame. This was then exported as a .csv file called “twitter_archive_master.csv”.

GENERATING INSIGHTS

Once the data had been cleaned to my liking, I then proceeded to generate statistical and graphical insights which focused on the statistical summaries of the various variables involved, the relationship between retweets and favorites and their correlation coefficients and finally, an analysis of the popularity of the dog breeds and the retweets and favorite counts they generated. The plots generated during this phase of the project were then saved as .png files to be used for additional reporting.