WRANGLE REPORT

BY

WILLIAMS MIEBI GBENEOWEI

OUTLINE:

INTRODUCTION

DATA GATHERING PROCESS

DATA ASSESSMENT

DATA CLEANING

STORING DATA

ANALYZING AND VISUALIZING DATA

INTRODUCTION:

This report aims to give a summary explanation of the wrangling and analysis of the data gotten from the twitter user @dog_rates also known as we rate dogs. It's a part of the data analysis process for the data gotten from the tweets of the twitter we rate dogs. The data analysis is meant to create some insights about dogs that were tweeted about.

The report documents the data gathering, assessment, cleaning, storing and visualization.

DATA GATHERING PROCESS:

In this project, data was gathered from several sources and were of different formats.

1. The WeRateDogs Twitter archive. The file was provided by the project and was downloaded directly from Udacity website.

2. The tweet image predictions. The file is hosted on Udacity's servers. I downloaded this file programmatically by using the Requests library in Python.

3. Get retweets count and favorite count information missing from the Twitter archive from another file. I chose to download the tweet JSON file programmatically by using the Requests library because I encountered a challenge with twitter that I was unable to overcome.

DATA ASSESSMENT:

After downloading the data, the data sets were assessed programmatically and visually for data quality issues and data tidiness issues.

Programmatic assessment involved the use of several functions such as *head(), tail(), value_counts(),isnull() , query, describe, info()* etc. to get a proper understanding and feel of the data sets.

I assessed the data visually in the jupyter notebook by altering the display settings. Each data set was evaluated for quality and tidiness issues and documented.

The following observations were made:

Quality issues

Twitter archive (TA) data frame

1. There are ridiculously very high values such as 204,1776,960,666,143,182,144,88,84,165,60,50,44,26,24,80,75,420, in the numerator column and 2333 for the denominator column which makes no sense

2. The timestamp column is in string format should be in date time and +0000 will have to be removed

3. columns such as text, source, retweets and replies ,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp do not appear to be useful to the analysis since the analysis should involve only original tweets ,hence should be dropped.

4. in_reply_to_status_id, in_reply_to_user_id,retweeted_status_id,retweeted_status_user_id,retweeted_status_timestamp have a staggering amount of nulls.

5. Name column is full of variables that are not dog names

Images data frame

1. Some entries are not actually dog breeds, we are having entries such as banana, paper towel, bagel, orange, spatula, etc.

2. Presence of duplicate image URLs

3. There are 31 occurrences in the data frame that do not fall into the most confident prediction for dog breed analysis.

4. Grouping the p1, p2, p3 columns into one since they are all dog breeds, it makes more sense to have them in a column.

5. The p1_conf, p2_conf and p3_conf levels should all be in a column.

6. The capitalization in the p- columns are inconsistent.

DATA CLEANING:

This part of the project was the most difficult part because most of the datasets were very dirty. All the observations made in the assessment part had to be corrected using code. The most challenging part was when I had to manually look out for wrong entrances in the dog breed column to make sure everything entered was a confirmed dog breed. After the new data frame was created it was assessed again before storing.

STORING DATA:

The newly created data frame was stored as csv file.

ANALYSIS AND VISUALIZATIONS:

The newly created data frame was loaded into the workbook and used to derive a few insights about the data.