Reporting: wragle_report

Create a 300-600 word written report called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

DATA WRANGLING REPORT

by Onu Chinonso

WeRateDogs™ @dog_rates · Apr 16
Say hello to Alice. I'm told she enjoys car rides and smells good. 12/10 would give her everything she could ever want

WeRateDogs™ @dog_rates · Apr 16
A photographer took pictures before and after he told his bunny he's a good boy. Here are the results. 13/10

Data wrangling is the process of gathering your data, assessing its quality and structure, and cleaning it before you do things like analysis, visualization, or build predictive models using machine learning.

I will be wrangling (and analyzing and visualizing) three(3)datasets,They are tweet archive from Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.the three datasets are as listed below

1. A downloaded (twitter_archive_enhanced.csv) from the WeRateDogs Twitter archive data

2. downloaded (image_predictions.tsv) from Udacity Servers

3. A tweet_json.txt file generated from twitter api

The Outlined steps below is what I used in wrangling my datasets

Step 1: Gathering data

Step 2: Assessing data

Step 3: Cleaning data

Step 4: Storing data

Step 5: Analyzing, and visualizing data

Step 6: Reporting

Report Goal:

The goal of this Report is to effectively wrangle data related related to dog ratings.

Report Details:

The tasks of this project are as follows:

Step 1 : Gathering Datasets

The data used for this project consisted of three different datasets that were obtained as following:

twitter_archive_enhanced.csv : This dataset was provided in the project workspace . I downloaded it by clicking on the link, I then uploaded it in my Project workspace and read it into pandas dataframe.i.e df=pd.read_csv('twitter-archive-enhanced.csv')

Tweet image prediction file: I imported the Python requests, numpy and os libraries. With the get() function of the requests library, I got the data through its url and saved it in a response variable. using the open function, I wrote it into a "tav" format then loaded it to pandas dataframe,

Tweet_Json text: I created a twitter developer account and sent to request to Twitter, It has not been granted yet so I used the Tweet_Json text file provided to us by Udacity to work with. I uploaded it to my Udacity Workspace, With the Python with open function again and a for loop, I read the tweet_json.txt line by line and loaded each line as json file. I saved each tweet_id, retweet_count, favorite_count, followers_count and friends_count which I later converted to a dataframe named tweet_json

With the Python with open function again and a for loop, I read the tweet_json.txt line by line and loaded each line as json file. I saved each tweet_id, retweet_count, favorite_count, followers_count and friends_count which I later converted to a dataframe named tweet_json.

Step 2: Assessing Data

I assessed the data using the following technique:

Visually: I read the three different dataframes individually in a jupiter notebook and also visually assessed the csv files in Excel spreadsheet.

Programmatically: I did various programmatic assessment with various python and pandas methods and functions such as .info(),.columns , .shape , .describe() , .duplicated() , .isnull().sum() , .sample(8) ,

Step 3: Cleaning Data

Before performing the cleaning , I made a copy of each databsets with the copy method.

Data Cleaning three processes namely: Define, Code and Test. which processes were followed in cleaning the data.

The Iaauea observed are listed Below

Quality issues

1. tweet_json.txt : ID Column has similar values with Tweet_id in the other dataframe, should be rename as tweet_id

2. twitter-archive-enhanced.csv : We observed that retweeted_status_id, retweeted_status_user_id, and retweeted_status_timestamp have (not null and not empty value) which affects the quality of data; we need to remove it.

3. twitter-archive-enhanced.csv table : Wrong Datatype TimeStamp

4. twitter-archive-enhanced.csv table : Wrong Datatype Text column

5. twitter-archive-enhanced.csv table : missing values for in_reply_to_status_id

6. tweet_json.txt: friends_count has one value, needs to be dropped.

7. twitter-archive-enhanced.csv table : missing values for in_reply_to_user_id ,retweeted_status_user_id ,expanded_urls, retweeted_status_timestamp.

8. twitter-archive-enhanced.csv table : Erroneous datatypes for name column

9. image-predictions.tsv table: jpg_url column is not needed, we need to drop it

Tidiness issues

1. Columns (doggo, floofer, pupper, puppo) are categorical data for dog and should be in a column

2. tweets_id column is spread across the three Datasets, we will merge the datasets

The Define Solution are as listed

1 renaming the id column

2 We will remove all rows that have values (not blank or non-null)

3 change the Datatype of Timestamp to Datetime

4 Change Datatype from Object to String.

5 I will drop the in_reply_to_status_id column it will not be need in our analysis

6 I will drop the friends_count column in tweet-json.txt.

7 Drop the columns - those coumns will not be needed for derivng insight.

8 Erroneous datatypes for (name) column

9 I will drop the jpg_url column

Tidiness

1 combine the four columns and form a new column called Stage_attribute

2 merge the three Datasets

Step 4: Storing the Data

After gathering, assessing and cleaning the data, I saved the merged data in a csv file named twitter_archive_master.csv.

Conclusion

This project has show me how to use the Wramgling process in analyzing my Dataset.