wrangle_report

May 26, 2020

1 Data Wrangling Report

Shakhawat Hassan

1.0.1 1) Gather Data

- import tweepy
- from tweepy import OAuthHandler
- import json
- from timeit import default_timer as timer
- import pandas as pd
- import numpy as np
- import seaborn as sns
- import urllib.request
- import matplotlib.pyplot as plt
- %matplotlib inline

After importing the libries, I did the following - df1 is 'twitter-archive-enhanced.csv' - df2 is image-predictions.tsv - df3 is 'tweet-json.txt' - every Twitter API for each tweet in the Twitter archive and save JSON in a text file and these are hidden to comply with Twitter's API terms and conditions > consumer_key = 'HIDDEN' consumer_secret = 'HIDDEN' access_token = 'HIDDEN' access_secret = 'HIDDEN' >auth = OAuthHandler(consumer_key, consumer_secret) auth.set_access_token(access_token, access_secret) api = tweepy.API(auth)

```
In [ ]:
```

1.1 2) Assess Data

After gathering the data, I assess the datasets by

- .info
- .describe
- .head()
- all_columns = pd.Series(list(df1) + list (df2)+ list(df3))
- all_columns[all_columns.duplicated()]
- df['names'].value_counts()

1.2 3) Clean Data

After gathering and assessing the data. I clean the datasets by

Quality - Drop unnecessary columns in df1 - Drop unnecessary columns in df2 - Extract the columns in df3 which are needed ('id', 'retweet_count', 'favorite_count') - Change 'id' name to 'tweet_id' in df3 - Change 'tweet_id' data type to string in df1 - Change 'tweet_id' data type to string in df2 - Change 'id' data type to string in df3 - Remove "_" and capitalize the first letter for p1, p2, and p3 in df2 - Change string to datetime for timestamp in df1 - Capitalize first letter in 'names' in df1 - Replace the names 'DoggoPupper" to "Doggo", 'DoggoPuppo', 'Puppo', 'DoggoFloofer', 'Floofer' Tidiness - Merge all three dogs' stages into one single column and then drop the empty rows - Merge all three columns into a single column by 'tweeter_id'