UDACITY

2021

Data Analyst Nanodegree Program

Project 3: Wrangel and Analyzed Data

REFAH M. ALDOSSARY

Wrangel Report

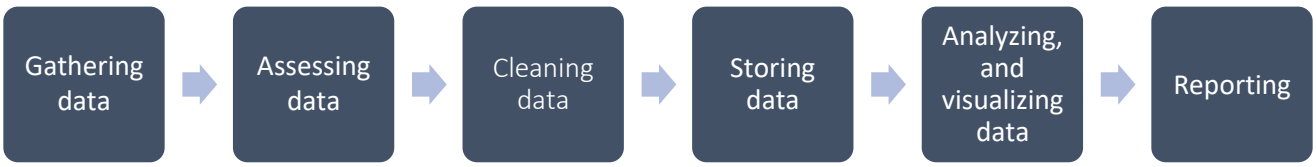
Project Objective:

The project main objective is to test student skills in gathering data from a variety of sources and in a variety of formats, assessing its quality and tidiness, then cleaning it using programming languages like python.

Project Summary:

The project is about wrangling WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. Tweet archive of Twitter user @dog_rates was used, also known as WeRateDogs, which is a Twitter account that rates people's dogs with a humorous comment about the dog. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Project Steps Overview:



Data Wrangling Process:

- **Data Gathering:**

1. Enhanced Twitter Archive

The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. Data includes text, dog name, dog "stage" (i.e. doggo, floofer, pupper, and puppo), tweet_id, and other additional data. The provided csv file was imported using python as shown below:

```
twitter_archive = pd.read_csv("twitter-archive-enhanced.csv")
twitter_archive.head(2)
```

2. Image Predictions

This dataset contains data about image predictions and confidence levels. The file was programmatically downloaded using Requests library as the following:

```
url = "https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv"
response = req.get(url)
image_predictions = pd.read_csv(io.StringIO(response.text), sep='\t')
image_predictions.to_csv('image-predictions.tsv',sep='\t',index=False)
image_predictions.head(2)
```

3. Tweet Json

This json file contains additional data such as favorite counts, retweet counts, and date of creation. The file was downloaded as the following:

```
#Downloading JSON file (tweet_json.txt):
full_tweet_data = pd.read_json('tweet-json.txt', lines=True)
full_tweet_data.head(2)
```

- **Data Assessment:**

The three datasets were assessed programmatically using pandas' functions such as info(), describe(), value_counts(), and shape. The goal is to find out some quality and tidiness issues in order to clean them later and have a comprehensive dataset that can be analyzed and visualized. This process results in the following issues:

Quality
twitter_archive: timestamp data type (convert to datetime), tweet_id data type (convert to string), Remove retweets and replies data, and Remove records where name = none or a

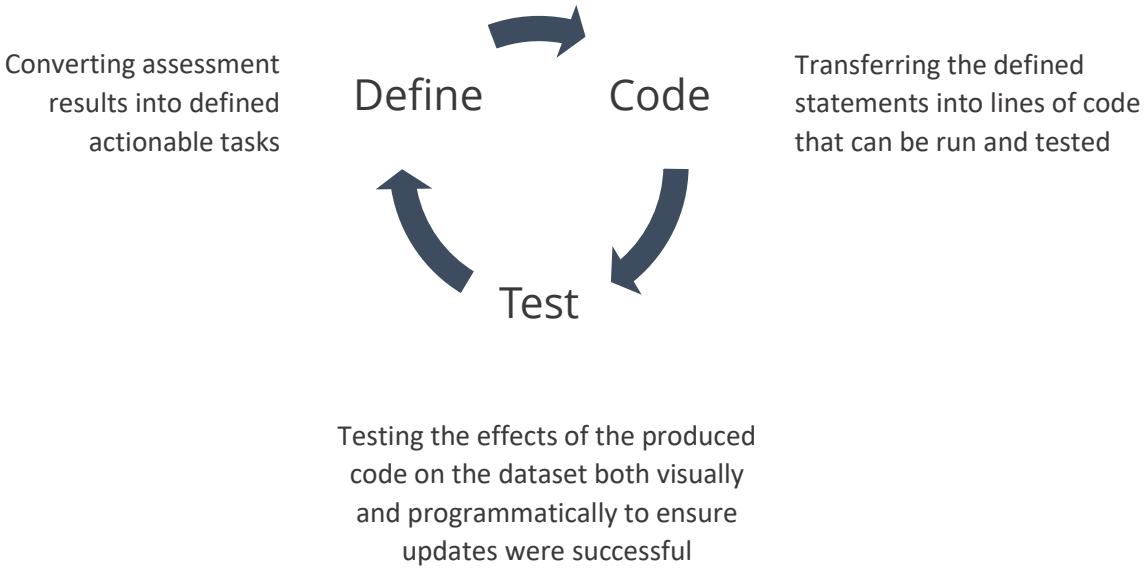
image_predictions: Update data type (tweet_id should be string)

full_tweet_data: Update column name (id should be renamed as tweet_id), id data type (convert to string), source data type (convert to category), and extract the exact source of tweet

Tidiness
1. Merging the 3 datasets for better analysis 2. Adding new categorical variable to combine doggo, floofer, pupper, puppo in a single column called "stages_of_dogs" (in twitter_archive) 3. Select only needed columns

- **Data Cleansing:**

After assessing and identifying the above issues, I started data cleansing process following these steps for each identified issue:



- **Data Storing:**

At this point, I have all data cleansed and ready to be stored in csv format to be further analyzed and visualized for better insights. Data was stored as the following:

```
final_tweets_data.to_csv('twitter_archive_master.csv')
```

- **Data Analysis & Visualization:**

After storing the cleansed data, I started drawing insights based on some visualizations. More details about this phase are provided in the act_report.pdf.