Data Wrangling Project Report

Introduction

Data wrangling is a process that aims at helping a data analyst know the anomalies a dataset contains. While wrangling, one is able to identify Missing, Invalid, Inaccurate or inconsistent data in a given dataset.

Data wrangling involves three steps:

1. Gathering Data
2. Assessing Data
3. Cleaning Data

Gathering Data

For this project, I gathered data from three different data sources as discussed below

1. WeRateDogs Twitter archive.
   This dataset was provided by Udacity in the form a csv file. I downloaded the dataset and uploaded it into my notebook. I managed to load the data using panda's read_csv function and stored the dataset in tweetsdf dataframe.
2. Image Predictions dataset.
   This dataset was acquired programmatically. I downloaded the data from Udacity server using the URL that was provided. I used panda's request library and the get function to grab contents of the dataset. I stored the dataset in imagepredictions dataframe.
3. Querying Twitter's Api
   To get the JSON data of each tweet, I used tweet_id in tweetsdf and a for loop to loop through each tweet querying the Twitter's API.I then grabbed favorite_count, retweet_count, followers_count, favourites_count and created_at columns. I stored the columns in a list called listdf. The errors encountered were stored in a list called errorlist. After grabbing all the required data, I stored the dataset in a dataframe called twitterdata

Assessing

With all the datasets required at hand, the next step was to check the quality and tidiness of the datasets. I did assess the datasets programmatically and visually.I was able to identify a couple of quality and tidiness issues.

Cleaning

Cleaning data means tackling and resolving issues identified in the assessing step. Before cleaning the datasets, I made copies of the dataframes. I used the define, code, test method iteratively to clean issues I identified earlier.

Storing

At this point tweetsdfclean contained clean data. I stored the dataset in twitter_archive_master.csv file.

Analysis and Visualization

From tweetsclean dataframe that had data from twitter_archive_master.csv, I came up with two visuals to provide insights.