Data Preprocessing

We Rate Dogs Tweets

Wrangling Act

Introduction

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.

In this project, we use WeRateDogs twitter data to create interesting insights and visualizations.

Gathering

In this step, we gathered all three pieces of data as described below:

The WeRateDogs Twitter archive

This file was directly accessible as its download link was provided. I downloaded it and then uploaded it to the workspace. Then I loaded the file to dataframe.

The tweet image predictions

Link to this file was provided. We needed to download it to the workspace programmatically. Then I loaded the file in the Jupyter Notebook as a Dataframe.

Data from the Twitter API

This data was a bit more complex to gather. We first needed to create our own twitter account. Then, we were supposed to get a developer account to get necessary tokens to authenticate with Twitter API. Then, I put those details in the tweepy twitter APL module of Python and downloaded a file containing likes and retweets of the tweets by WeRateDogs twitter account. Then we loaded it as DataFrame.

Assessing Data

After gathering all three pieces of data, assess them visually and programmatically for quality and tidiness issues. We assessed data using following two ways :

- **Visual assessment:** each piece of gathered data was displayed in the Jupyter Notebook for visual assessment purposes.
- **Programmatic assessment:** pandas' functions and/or methods were used to assess the data.

I have documented all the issues that I found in the Jupyter notebook itself.

Cleaning Data

In this step, I cleaned all the issues that I found while assessing. While cleaning, I also found some additional issues which I cleaned along the way.

Storing Data

After the data is cleaned, we need to store the cleaned data for future references. We saved the clean master dataset as twitter_archive_master.csv in the workspace.

Analyzing and Visualizing Data

In this step, we tried to analyze the clean data to find some interesting insights and visualizations.

Reporting

In this step, we will make a report of the insights and visualizations that we found in the above step. This report is also included in the reporting step.