WeRateDogs Project

Author: Eduardo Kaneko

Date: 10.04.2019

Process

Purpose: Gather and clean data from multiple sources for analysis

Language: Python

Skills:

- gathering data: programmatic file download, downloading API data
- assessing data: documenting cleanliness and tidiness issues
- cleaning data: cleaning for completeness, tidiness, validity, accuracy, consistency
- storing data: CSVs
- analyzing data
- visualizing data
- written report

Libraries: matplotlib.pyplot, missigno, seaborn, pandas, numpy, resquests, warnings, tweetpy, json, re, os.
Software: Jupyter Notebook. </p>

Data Gathering

Data Gathering or Data Colletion is the process of gathering information on targeted variables and different sources. It is the first step of data wrangling proccess. Before gathering, we do not have data to work with. So, for this project, the data gathering was gather data from the csv given by udacity and by the twitter's API.

Data Assessment

In a perfect world, data would always be complete, accurate, current, pertinent, and unambiguous. In the real world, data is generally flawed on some or all of these dimensions. Data assessment in practice has tended to focus on look to tidiness and quality. Evaluating my data was the second step of wrangling data. In evaluating, I've act as a research detective, inspecting my data set for two things: data quality problems (ie, content problems) and lack of organization (ie, structural problems). For this specific dataset, I have found a lot of problems, assessing programatic and visually.

Ref: Technopedia

Data Cleaning

Nowadays, everyone knows that we are producing a lot of data, but, most of this data, is not clean. For this project, it was not different. Clearing my data was the third step in wrangling data. This was where I correct the quality and storage problems I identified in the evaluation step, and it is done more efficiently using code. In this step, I've had clear all the problems I identified in data assessment using Python and Pandas.

Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to **analyzing data sets** to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task. Exploratory data analysis was promoted by John Tukey to encourage statisticians to explore the data, and possibly formulate hypotheses that could lead to new data collection and experiments. EDA is different from initial data analysis (IDA), which focuses more narrowly on checking assumptions required for model fitting and hypothesis testing, and handling missing values and making transformations of variables as needed. EDA encompasses IDA.

Ref: Wikipedia