Data Wrangling Report

Introduction

This Wrangle and Analyze Project is part of Udacity's Data Analyst Program. In this project the dats wrangling, which consists of:

- Gathering data, download resourses from diferent references
- Assessing data, for quality and tidyness issues
- Cleaning data, that identified in previous step
- Storing, analyzing, and visualizig wrangled data
- Reporting on data analyses and visualizations

The most tools, libraries and programming language used in this project are:

- Python
- Numpy Library
- Pandas Library
- Requests Library
- Tweepy Library
- Json Library
- Matplotlib Notebook
- Twitter's APi

Gathering

Data was gathered from 3 different sources:

1. The 'twitter-archive-enhanced' was provided by Udacity. This file includes a huge variety variables foe each tweet as tweet_id, timestamp, text, name, rating, ect.
2. The 'image_prediction file was downloaded using the Requests library.
3. Additional data, including favorite count and retweet count were gathered using the Twitter API.

Assessing Data

After gathering data were used the following methods:

- .head()
- .sample()
- .info()
- .value_counts()

Quality issue:

- Tweet_id was the incirrect data type
- The columns conteined the 'None' insted of 'NaN'
- Timestamp was the incorrect datatype
- Missing values from images datased(2075 rows insted of 2356)
- Sources are not readable

Tidiness:

- Dogstage was in 4 columns(doggo, floofer, pupper, puppo), there is no reason to keep them
- Merge all dataframes togethr as they all contained the proper information

Cleaning Data

This part of the data wrangling was performed in three stages: Define, Code and Test. First, copies of DataFrames were created before cleaning. Then used methods for investigation: merge(), reduce(), drop(), replace(), head(), etc

Storing

The final datset was stored as 'twitter_archive_final' in a csv file. At this step, the data was successfully wrangled and therefore ready for analysis and visualization

Analysis and Visualization

Visualizations are provided in act_report.pdf