Wrangle Report

Introduction

The wrangle act analysis focuses on the process of data wrangling which is composed of 3 steps:
- Gathering
- Assessing
- Cleaning
- EDA (Exploratory data analysis)

After wrangling the data in a quite thorough manner, a quick analysis will be performed.

The data selected in this view is from Twitter. The data that we have comes from multiple sources including a manually downloaded file, a file from an online server and data from Twitter's API. A lot of processing must be done in order to wrangle this data. In the final steps, we will have a look at tweets from the famous profile @WeRateDogs and derive trends from them.

Gathering Data

In this part, we gathered data from 3 different sources in 3 different formats:
- **A downloadable CSV file source:** Twitter enhanced archive data 'twitter-archive-enhanced.csv'
- **An online server source TSV file** : Image prediction data based on tweets from the archive 'tweet_image_pred.tsv'
- **An API JSON source data to load into a txt file** : Using API to get more data based on tweets from the archive (our main focus will be to gather retweet count and favorite count) 'tweet_json.txt'

These datasets were read in a dataframe format using Pandas Library

Assessing Data

In this second step, eight (8) quality issues and two (2) tidiness issues were detected and documented. It has been done by looking at the data programmatically using useful pandas features and also visually by looking at the dataframe and random samples of it.
The **quality issues** were composed of different cases that can be separated this way:

- Missing data – *and also some unnecessary data*
- Datatype issues
- Inconsistency issues
- Duplicates issues
- Invalid data
- Inaccurate data

The **tidiness issues** were represented by pivoted columns that are better analyzed in one category and by the presence of several datasets when all could be grouped in one.

Cleaning Data

The cleaning process follows along to correct each data issue detected in the assessment. There are 3 steps for each issue:

- **Define** how the issue will be corrected
- **Code** to correct the issue
- **Test** if the issue was corrected

Finally the end result **clean dataset** has been **stored** in a master csv file 'twitter_archive_master.csv'

Exploratory Data Analysis

- Number of tweets overtime? How is the WeRateDogs general account activity?
- Which is the most retweeted post vs most favorite post?
- What makes a good dog rating? - *correlation between rating, retweet and favorite? Rating by top dog category?*
- What is the best name for a dog?
- How good is the prediction of the dog breed? *Check confidence level range and test image*