Udacity Data Analyst Nanodegree

Wrangle Report: WeRateDogs

Data Wrangling : Gather, Access, Clean

(Source: https://twitter.com/dog_rates)

Introduction

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that

rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

[*Ref: Project Overview section under concepts in Wrangle and Analyze Data*]

Project Details:

Tasks in this project are as follows:

- Data wrangling, which consists of:
    1. Gathering data
    2. Assessing data
    3. Cleaning data
- Storing, analyzing, and visualizing the wrangled data
- Reporting on
    1. Data wrangling efforts
    2. Data analyses and visualizations

Data Gathering:

In this project of analysing WeRateDogs (@dog_rates) Twitter handle, we are goona gather data in both manual apporach and programatic approach, by using the data given by Udacity Team we are done with the manual process. Further, we are gonna use Twitter API services and perform requests to get the data.

1. **twitter_archive_enhanced.csv** : Provided by Udacity Team
2. **image-predictions.tsv**: Provided by Udacity Team | Can be retrived with an Enpoint request without Authorization
    - [URL of the file: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv)]
3. Gathering Tweets, retweets and count etc,. Data from twitter API using Tweepy library with personal API credentials

Data Assessing:

Assessing is also identifying structural (tidiness) issues that make analysis difficult.
In this specific segment on Data Assessing, we'll be going through two major Data Issues.

1. **Data Quality Issues**
2. **Data Tidiness Issues**

Data Quality Issues

issues with content. Low quality data is also known as dirty data. Dirty data = low quality data = content issues

From the Lesson 3 of Data Wranglins section. Thr four important data quality factors are :
1. Completeness : Which concludes, weather the data consists of Missing data.
2. Validity : All about the structured data communication that converys any meaning
3. Accuracy : Deals about the inaccurate data, in which there are chances that dirty data can show up as valid data.

3. Consistency : Deals with data Standrdization

Let's go through the data quality issues with the three various data sources that we gathered in the **Data Gathering** section.

archive_data:

1. Completeness:
    - Missing data found in following features:
        - in_reply_to_status_id
        - in_reply_to_user_id
        - retweeted_status_id
        - retweeted_status_user_id
        - retweeted_status_timestamp
        - expanded_urls
2. Validity:
    - dog names: few dogs names conisits of
        - 'None' as a name, or 'a', or 'an.'
    - This data consists of Duplicate data, which is a result of having retweets.
3. Accuracy:
    - retweeted_status_timestamp
    - timespamp
      The above mentioned features are in type Object.
4. Consistency:
    - The well known column 'rating_denominator' supossed to be standard 10, but there are multitude of various values.

    - Score feature consists of HTML tags

image_data:

- Validity:
    - p1
    - p2
    - p3
- columns have invalid data
- Consistency:
    - p1
    - p2
    - p3
- columns aren't consistent when it comes to capitalization
    - in p1, p2 and p3 columns there is an underscore for multi-word dog breeds

twitter_counts_df:

- Completeness:
    - Missing Data Available

Messy data = untidy data = structural issues

Tidiness Issues

Three requirements for tidiness:

1. Each variable forms a column
2. Each observation forms a row
3. Each type of observational unit forms a table

archive_data:

- The last four columns all relate to the same variable (dogoo, floofer, pupper, puppo)

image_data:

- this data set is part of the same observational unit as the data in the archive - one table with all basic information about the dog ratings

twitter_counts_df:

- this data set is also part of the same observational unit - one table with all basic information about the dog ratings

Data Cleaning:

Cleaning is the third step in the data wrangling process:

Types of cleaning:
1. Manual (not recommended unless the issues are single occurrences)

2. Programmatic

The **programmatic** data cleaning process:

- ***Define***: convert the data into defined cleaning tasks. These definitions also serve as an instruction list so others (or yourself in the future) can look at your work and reproduce it.
- ***Code***: convert those definitions to code and run that code.
- ***Test***: test your dataset, visually or with code, to make sure your cleaning operations worked.

ref: [Cleaning summary in Cleaning Data lesson]

Define

1. Merging the cleaned data frames of images, archive and twitter_count_df and corret the dog types.
2. Framing a single specific column for different dog types:
    - Doggo, Floofer, Pupper, Puppo.
    - and remove columns which are not required
        - in_reply_to_status_id
        - in_reply_to_user_id
        - retweeted_status_id
        - retweeted_status_user_id
        - retweeted_status_timestamp
3. Deleting retweets.
4. Removing / Deleting the columns/feature which are not required
5. Changing or converting the tweet_id from an integet type to a string type.
6. Changing the timestamp into actual datetime format.
7. Naming issues cocrrection.
8. Standardization over dog ratings.