WRANGLE REPORT

Udacity DAND: Wrangle and Analyze Data Project

By: Himanshu Tripathi

INTRODUCTION

This Wrangle and Analyze Data Project is part of Udacity's Data Analyst Nanodegree Term 2. The project involves wrangling of data from various sources associated with tweets from the Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs rate's pictures of people's dogs in a humorous manner, most often giving ratings higher than 10/10. After scraping together the data, quality and tidiness issues were assessed and then cleaned. Finally, two visualizations were created and insights can be found in the act_report.pdf document.

GATHERING DATA

Data was gathered from 3 different sources:

1) The enhanced twitter archive file was provided and downloaded manually. This file includes various variables for each tweet including tweet id, timestamp, text, rating numerator and denominator, name, etc.

2) Additional data, including favorite count and retweet count, were gathered using the Twitter API.

3) The tweet image predictions file was downloaded programmatically using the Requests library from Udacity's servers. Using machine learning techniques, the breed of dog was predicted based on the picture.

ASSESSING DATA

After the data was gathered, assessment was performed using the following methods:

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

Tidiness issues that were cleaned:

- • Combining all dataframes together as they all contained information about the same tweets
- • Combining 4 variables about dog type into 1 column "dog_stage"

Quality issues that were cleaned:

- • Data contained retweets
- • Tweet id was the incorrect data type
- • Timestamp was the incorrect datatype
- • Name contained the string "None" instead of a NaN
- • Name contained various inaccuracies which were regular lowercase words
- • The name O'Malley was incorrectly extracted as "O"
- • Rating numerators which contained decimals were incorrected exported
- • Ratings are unstandardized
- • Undesired columns present

CLEANING DATA

The issues found during the assessment process were cleaned and tested using the following methods and techniques:

- • merge()
- • reduce()
- • .extract()
- • .drop()
- • .isnan
- • .astype()
- • .to_datetime()
- • .islower()
- • .replace()
- • .rename()
- • set_option()
- • .loc[]
- • .value_counts()
- • .info()
- • .head()
- • Loops
- • Regular expressions

CONCLUSION

Rarely does all the data you want for a project come from 1 source and is already tidy. This project emphasized that you will need to using Python and its various libraries to scrape data from various sources in various formats, and clean various quality and tidiness issues, before any data analysis can be performed.