Wrangling report

April 27, 2018

1 Introduction

In this project, we had to conduct data wrangling on twitter data of 'WeRateDogs'. This is a twitter account which basically rates dog images with humorous content. The rating denominator is usually 10, however, the numerators are usually greater than 10. This aspect was not cleaned as it is part of the humor and popularity of WeRateDogs.

The tasks for this project were: - Data wrangling comprising of following steps: - Gathering data - Assessing data - Cleaning data - Storing, analyzing, and visualizing the wrangled data - Creation of report on the analyses and visualizations of the wrangled data

1.1 Data wrangling

1.1.1 1. Gather

Gathering data for this project was focused around three files. We were provided with two files, twitter archive file containing data of 2356 tweets and second file, consisted of the image prediction data. For the third file, we had to query the twitter API for each tweet's JSON data. This was carried out using Python's Tweepy library and the output of this was stored in a text file. Now, these are the three files on which next steps, assessing and cleaning was to be carried out.

1.1.2 2. Assess

Assessment on the gathered data had to be done visually and programmatically. Visual assessment of these three files helped finding the three tidiness issues and a few of the quality issues. Programmatic assessment helped filtering the issues.

Quality Quality issues were detected on the basis of completeness, validity, accuracy and consistency. Ten issues were detected based on these parameters.

Tidiness Tidiness revolves around structural issues. Here, we could find 3 issues regarding to the structure.

During programmatic assessment, I checked for data types, null entries, duplicate values, value counts etc. As for structural issues, I have combined the data in three files and stored in to a single file. Also, some columns were condensed into a single column and few columns were removed.

1.1.3 3. Clean

This is where, all the issues found in step 2 had to be fixed. Process for cleaning is define, code and test. Each issue was fixed with these three steps.

1.2 Analysing and visualizing

The master data file was then analysed to find basic results like which is the most favourited tweet, which is the most common dog type, what is the most common dog name? etc. Visualization helped in summarizing these trends over the years. The inferences are all summarized in a pdf file.

2 Conclusion

Data wrangling helped finding insights into the data which normally would have been misleading. This was a tough yet an interesting exercise. The issues which were solved here are only some of the issues in the dataset. There are many issues that can help refining the data further. I hope to carry those out in the future.