Adrian Marinovich
Springboard Data Science Career Track
Capstone Project #1 - Data Wrangling
September 24, 2018

Project: Detection of smiles in images of faces

What kind of cleaning steps did you perform?

The primary data wrangling tasks were not cleaning, but rather were focused on writing scripts to fill the AWS S3 buckets with data, establishing a pipeline from the Jupyter notebook to the S3 buckets using Boto, and then using the lists of smile and non-smile image file names to select the corresponding images from the larger LFWcrop dataset of images.

How did you deal with missing values, if any?

Two file names in the smile list did not have matching image files, and were identified as extraneous .txt file names ('listt.txt', and 'SMILE_list.txt') which did not interfere with the correct matching of image file names.'

Were there outliers, and how did you handle them?

Images corresponding to the smile and non-smile lists were reviewed to check for validity of the smile/non-smile labels, and found to be consistent with the given labels. There also appears to be a diversity of age, sex, ethnicity, head position, facial hair, and presence of eyeglasses in both the smile and non-smile sets, with the exception that children appear to be absent from both sets. Given that children's photos would not be used without careful consideration and protections, this is not seen as a flaw or outlier issue in the dataset, for the purposes of an initial analysis as undertaken here.

Submit a link to the document.

https://github.com/adriatic13/springboard/blob/master/dsct_capstone1/Adrian_Marinovich_Cap1_smiles_data_wrangling.ipynb