Capstone Project 1: Data Wrangling

The data has been acquired as a csv file from <https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews>. It was created by web scraping over 67K employee reviews for employers such as Amazon, Apple, Google, Facebook, Microsoft and Netflix. The reviews were collected over a time span of 10 years (2009-2018).

The columns in this dataset are:

company Name of company

location Company location

dates Date of review

job\_title Job title of reviewer

summary Review summary

pros Advantages

cons Drawbacks

advice\_to\_mgmt Advice to senior mgmt

overall\_ratings Overall rating for the company

work\_balance\_stars Work balance rating

culture\_values\_stars Culture and values rating

career\_opportunities\_stars Career opportunities rating

comp\_benefit\_stars Compensation and benefits rating

senior\_management\_stars Senior management rating

helpful\_count User count who found this review helpful

link Website link for the review

The first step was to read the csv file into a pandas DataFrame. The dataset consists of 67529 observations with 17 columns.

Some of the columns were poorly labeled so the first step was to rename them.

Some columns had “none” as value. So it didn’t qualify as missing value, even though it was null. The next step was to replace all “none” values with np.NaN so they could be identified as a null value.

The ratings columns (overall\_ratings, work\_balance\_stars, culture\_values\_stars, career\_opportunities\_stars, comp\_benefit\_stars, senior\_management\_stars) are the most important variables in this project. In order to fill the missing values on them, they were first converted from object datatype to float type. And then, the null values were replaced with mean of the column.

Finally, the dataframe was sorted by company and null values in location column were filled using the ffill and bfill methods. At the end of this process, the only variables that continue to have null values are summary and advice\_to\_mgmt but that can be ignored for our analysis.