## Project Deliverable 1

**Raw Dataset**

Google, Amazon and More Employee Reviews:

<https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews>

**Description**

The dataset we will be working with contains employee reviews for Google, Amazon, Apple, Facebook, Microsoft and Netflix, collected by web scraping Glassdoor. Glassdoor is one of the world’s largest job and recruiting sites, built on the foundation of increasing workplace transparency. One key feature of the site involves collecting anonymous reviews of employers from current and former employees. Reviews often contain opinions on management, company culture, company policies, and salary information. The dataset we discovered for this project contains over 67,000 reviews described across 16 variables, which are elaborated on below. It is worth noting that the data type for each variable is in its original form, which may not be the optimal type for our analysis purposes (i.e. may be changed during the data preparation phase).

**Variables**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Data Type** | **Description** |
| count | Integer | A count for each review |
| company | Factor | The name of the company |
| location | Factor | The location of where the employee worked in text, described in the format of ‘city, state’. |
| dates | Factor | The date on which the review was created. Written in the format of ‘month date, year’ |
| job-title | Factor | Describes whether the employee is a current or former employee and the job title of the employee, in the format ‘status employee, job title’ The reviewer may elect to choose to keep their title anonymous. |
| summary | Factor | A text summary of the employee’s review |
| pros | Factor | Describes in text the employee’s opinion of the positive aspects of working at the company |
| cons | Factor | Describes in text the employee’s opinion of the negative aspects of working at the company |
| advice-to-mgmt | Factor | A text of the employee’s advice to management for improving their practices |
| overall-ratings | Numerical | A numerical rating for the employee’s overall impression of working at the company. Rated on a scale from 1-5, where 1 is the lowest score and 5 is the highest score. |
| work-balance-stars | Factor | The individual employees’ rating of how good they found the work-life balance to be. Expressed as a rating from 1-5, where 1 exhibits a low work-life balance, and 5 a high (positive) work-life balance. |
| culture-values-stars | Factor | A numerical rating for the employee’s impression of the culture and values of the company. Rated on a scale from 1-5, where 1 is the lowest score and 5 is the highest score. |
| career-opportunities-stars | Factor | A numerical rating for the employee’s impression of the career opportunities at the company. Rated on a scale from 1-5, where 1 is the lowest score and 5 is the highest score. |
| comp-benefit-stars | Factor | A numerical rating for the employee’s impression of the compensation and benefits of the company. Rated on a scale from 1-5, where 1 is the lowest score and 5 is the highest score. |
| senior-management-stars | Factor | A numerical rating for the employee’s impression of the competency of their senior management at the company. Rated on a scale from 1-5, where 1 is the lowest score and 5 is the highest score. |
| helpful-count | Integer | A score of how many readers that have clicked that they found that particular review helpful |
| link | Factor | The URL to the particular review |

**Questions to Explore**

* Which company has the happiest employees?
* What are the most prominent positive aspects of working for each company?
  + Correlation between positive aspects and employee status (i.e. “current” or “former”)?
* What are the most prominent negative aspects of working for each company?
  + Correlation between negative aspects and employee status (i.e. “current” or “former”)?
* Are anonymous (hidden job title) reviews more negative than non-anonymous reviews?

**Questions to Explore cont.**

* Does the sentiment of reviews differ between locations?
  + I.e. are people more happy working at Google’s headquarters in Mountain View, than in alternative locations?
* Are we able determine how positive (negative) a review is solely based on the summary of the review?
* Are there specific types of employees (job roles) which are more satisfied than others?
* Are there core aspects on which the companies’ work cultures differ? Are company groupings (“clusters”) apparent based on these aspects?
* As time has gone by, have employees’ views on the companies become more negative, more positive, or stayed flat?

**Data Preparation**

After reading in the CSV file, we deleted the first column (which was essentially a row number for each record) and the last column (URL for review link). We replaced any “none” values in the dataset with “NA” to be able to run the mice::mice() function (and other missing-value-resolution functions) later on. We also converted all the rating-related variables [i.e. overall rating, and the “stars”-related variables (subratings) informing the overall rating] from factor to numeric data type.

Next, we cleaned up the variable names using base::names() to fix misspellings and ensure a standard format (e.g. replacing “.” with “\_” for best practice reasons). Before running the mice package, we took a closer look at the missing values in our dataset using md.pattern() to gain a better understanding of how the NA values were distributed in the data. In preparation for imputation of missing values, we created a separate data frame containing only our numeric “stars”-related variables; we named it “reviews\_onlystars.” Next, we chose to scale this limited dataframe, and named the scaled data “reviews\_onlystars\_scaled.” Finally, we ran complete(mice()) using m = 10 because that figure approximately corresponded to the percentage of missing data. After mice() was done running, we checked that we had resolved all NAs once more with md.pattern().

While working with mice(), we encountered a number of problems. The first time we used it we chose not to scale the data and noticed that running the function caused the ratings distributions to change: initially, all “stars”-related variables were constrained to a 1-5 scale, but after mice() was executed we noticed that some reviews contained ratings that were below 1 and others above 5. We thought that scaling the data before running mice() might fix the issue, but did not feel confident enough that it did so and worried that the resulting imputed values would not accurately depict the realistic distribution of the data in the missing values.

To mitigate this, we resorted to imputing missing values with column medians (which would give values within the necessary min/max range and increment size of 0.5 or 1). We also used the median because it is not possible for a user to give, for example, 3.68 stars - so using a mean of some sort would have imputed values not necessarily reflective of realistic data. We used the following code for median imputation, and found this described method to be effective:

for(i in 1:ncol(reviews\_onlystars\_clean)){reviews\_onlystars\_clean[is.na(reviews\_onlystars\_clean[,i]), i] <- median(reviews\_onlystars\_clean[,i], na.rm = TRUE)}

As the last step in the missing-value-resolution section, we merged “reviews\_exclstars” (all variables from the original dataset excluding the “stars”-related variables) with “reviews\_onlystars\_clean” (the now-complete “stars”-related variables) and called the final merged data “reviews\_merged.”

Next, we moved our attention to a prepare a few other variables within “reviews\_merged.” Using the “lubridate” package, we converted the “date” variable from factor to its correct date (mdy) type. In addition, we separated the job\_title column into two different columns called "employee\_status" and "position” and converted employee\_status to a factor variable (“current” or “former” employee). Last, we converted other factor variables containing text responses to their correct character type, for future text analysis purposes.

Finally, we wrote the cleaned data to a .csv file.