Data Wrangling

Ayano Yamamoto

3rd April 2021

Table of Contents

Report: Best Companies to Work for in IT

Best Reviews Overall
Facebook
(☆4.51)

Best Reviewed Locations
Menlo Park, CA
(☆4.55)

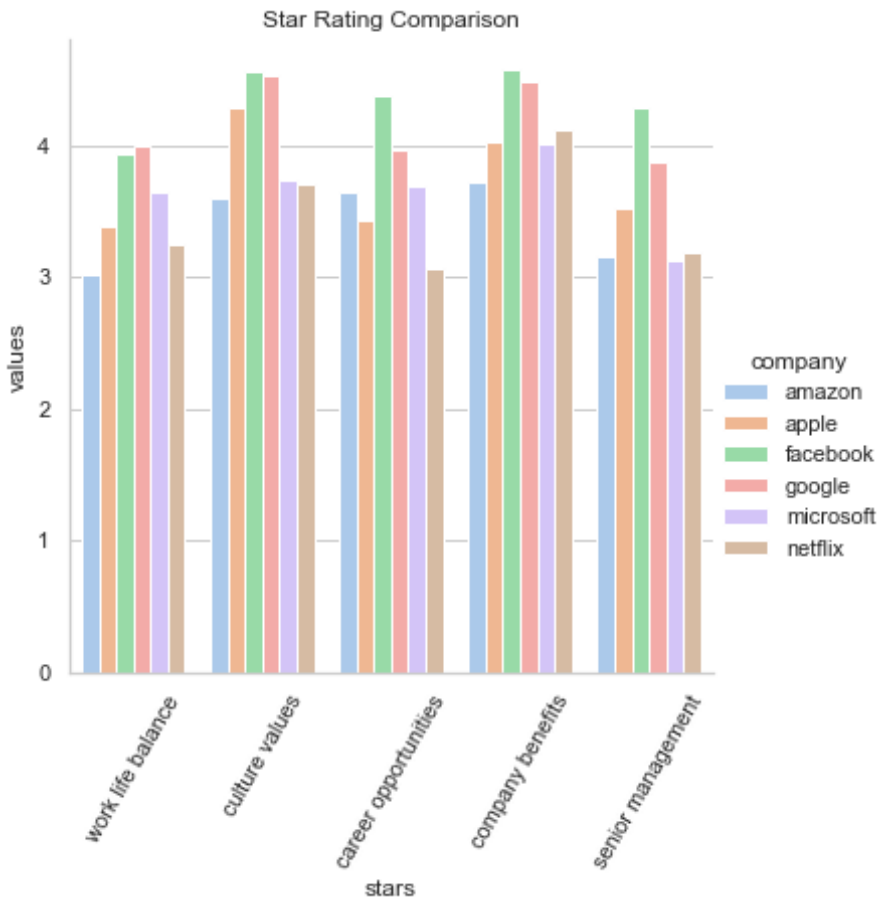
Best Reviews in 2018
Google
(☆4.38)

Highest Revenue in 2021
Amazon
(469.822 billion USD)

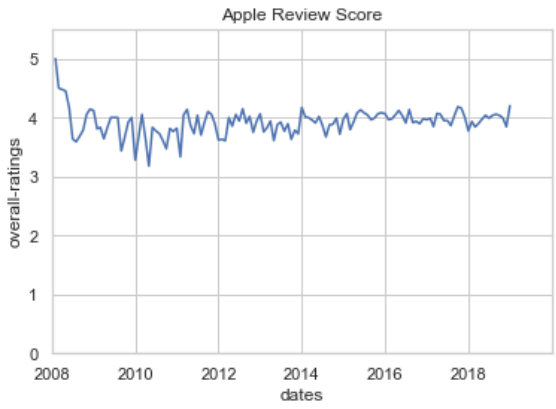
Highest Number of Employees in 2021
Amazon
(1,608,000)

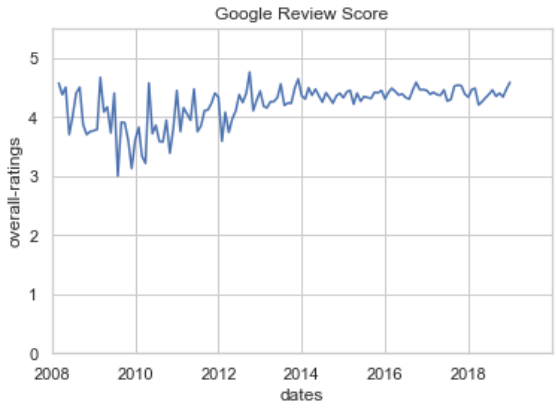
Highest % of Female Employees in 2020
Netflix
(47.1%)

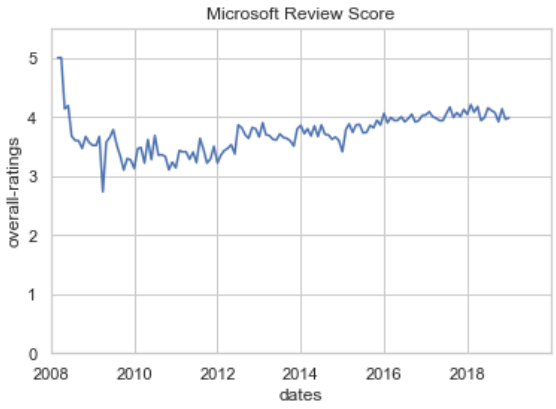
Comparison of Companies for each Star Ratings

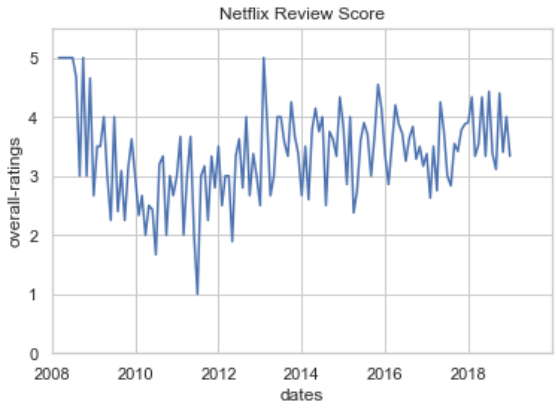


Trend Overtime of Overall Ratings for Each Company









Summary of Analysis Performed

Importing and cleaning the data

- 'employee_reviews.csv' was imported using encoding='ISO-8859-1.' All the columns were imported except for the index and 'link' since they are not relevant for this analysis.

- Inspection of the data showed that there are 969 duplicates. They have been removed during cleaning, reducing the number of rows from 67,529 to 66,560.

- Most null values were represented as 'none.' They have been replaced with np.nan for the analysis.

- 'dates' column was converted into date format using .to_datetime.

- Star rating columns ('work-balance-stars', 'culture-values-stars', 'carrer-opportunities-stars', 'comp-benefit-stars', and 'senior-mangemnet-stars') were converted into numerical format using .to_numeric.

- After converting star rating columns to numerical format, the number of missing values in star rating columns were as below.

Column Names	Values Missing	% Values Missing
work-balance-stars	7,052	10.59%
culture-values-stars	13,386	20.11%
carrer-opportunities-stars	6,998	10.51%
comp-benefit-stars	7,050	10.59%
senior-mangemnet-stars	7,659	11.51%

Since the % values missing were all low enough (less than 30%), the NaN values have been replaced with median values for each company per star rating. Median values were chosen to avoid influences from occasional very low reviews that are outliers.

Company with best reviews overall

The column 'overall-ratings' had no missing values from the original dataset. After removing duplicates, company with the best reviews overall has been calculated by grouping a dataframe by 'company' and calculating the mean 'overall-ratings' for each company. The result was sorted descending, and the top row was selected.

Location with best reviews

From the original dataset, there were 2,044 unique values for 'location,' of which 1,075 of them had only one review. This would interfere when calculating mean 'overall-ratings' for each location, since the top-rated locations will be those with one 5-star ratings.

To avoid this, a dataframe 'locationsize' was created to show the number of reviews per location. This dataframe was filtered to only show those that have above mean number of reviews.

Mean 'overall-ratings' for each location were calculated only for locations that exists in 'locationsize' dataframe. The result was sorted descending, and the top row was selected.

Trend overtime (review score) for specific companies

After the 'dates' column was converted into date format, dataframes were created for each companies grouping the reviews by months and calculating the mean 'overall-ratings' for each month.

Line plot was used to visualise the trend overtime for each company, with x-limits and y-limits standardised over all six graphs.

Company with current best reviews

Since the latest reviews in the dataset were from 2018, I have defined the "current reviews" as reviews from 2018. A dataframe was created to filter for reviews with 'dates' larger than pd.to_datetime('2017-12-31') and the mean 'overall-ratings' for each company were calculated. The result was sorted descending, and the top row was selected.

Comparison of companies in terms of each star ratings

In the data cleaning section, the star ratings were converted to numerical format and their missing values were replaced with median values for each company per star rating.

They were then grouped by 'company' and the mean values were calculated per star rating. This dataframe 'meanstardf' was converted to a long form and visualised in a bar plot grouped by star ratings.

Company with highest revenue & highest number of employees in 2021

Revenue and number of employees are external information brought in from Wikipedia through read_html. The original sources are:

- https://en.wikipedia.org/wiki/List_of_largest_technology_companies_by_revenue
- https://en.wikipedia.org/wiki/Amazon_(company)
- https://en.wikipedia.org/wiki/Netflix

Tables from these three links were standardised for column names and concatenated into one dataframe. This dataframe 'companydata' was then merged into the main dataframe using keys from the main dataframe.

Company with highest percentage of female employees in 2020

Percentage of female employees is an external information brough in through read_csv. The csv was created from each of the company's diversity reports:

- Our workforce data (aboutamazon.com)
- Apple shares new data about diversity in the company - 9to5Mac
- 2020 Report – Diversity (fb.com)
- Workforce Representation - Google Diversity Equity & Inclusion
- Microsoft's 2020 Diversity & Inclusion report: A commitment to accelerate progress amidst global change - The Official Microsoft Blog
- About Netflix - Inclusion Takes Root at Netflix: Our First Report

Dataframe 'genderdf' was created and merged into the main dataframe using keys from the main dataframe.