**Capstone Project 1**

**Employee Reviews for Amazon, Apple, Facebook, Google, Microsoft and Netflix**

**Introduction:**

In today’s job market, job seekers like to research their potential employers before applying to open positions or accepting an employment offer. They use reviews and feedback from ex- or current employees of that organization to determine whether or not that workplace would be a good fit for them. There are several websites that offer reviews of companies based on user submitted feedback. However, it would be tedious for users to go through each and every review to make a sound judgement. Using the datasets on these websites, a model could be created to visualize the categorical and overall ratings of employers, so that it can be easily used.

Job seekers will be able to use this report and analysis to determine which employers have gone up in their ratings and which have gone down in the last 10 years. The dataset includes ratings on various categories such as work-life balance, compensation/benefits, career opportunities, culture values, senior management as well as an overall rating. Job seekers can use this project to steer their research on potential employers, without having to browse through all feedbacks on the various company review websites.

This report can also be used by the employers featured on it to determine the categories where they have consistently received low ratings and work upon improving them.

**The Employee Reviews Dataset:**

The dataset used for this project was created by web scraping over 67K employee reviews for these six companies. It will be acquired as a csv file from <https://www.kaggle.com/petersunga/google-amazon-facebook-employee-reviews>. The dataset spans between 2009 and 2018. It contains 67529 rows and 17 columns. This is a real world dataset and likely to have a lot of null and missing values. Those will be handled by data wrangling methods.

Some of the interesting columns are:

|  |  |
| --- | --- |
| company | Name of the company |
| location | Branch location |
| dates | Date the review was posted |
| job-title | Job title of the employee |
| overall-ratings | Overall rating (1-5) |
| work-balance-stars | Work life balance rating (1-5) |
| culture-values-stars | Culture and values rating (1-5) |
| carrer-opportunities-stars | Career opportunities rating (1-5) |
| comp-benefit-stars | Compensation benefits rating (1-5) |
| senior-mangemnet-stars | Senior management rating (1-5) |

**Data Wrangling:**

The employee review dataset is a real world dataset. Upon inspection, it was found that some of the critical columns had an incorrect value of none which would neither qualify as a null value nor a numeric value. Some of the columns were poorly labelled and/or misspelled. The date when the review was posted and all the ratings columns were stored as object datatype in the dataset.

The first step was to correctly label all the columns. Using unique values, it was found that “none” was found in all the columns as a value. Since this value of “none” would serve no purpose, they were collectively converted to np.NaN. This translated all the null values to be technically null.

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. And then, the null values were replaced with mean of the column.

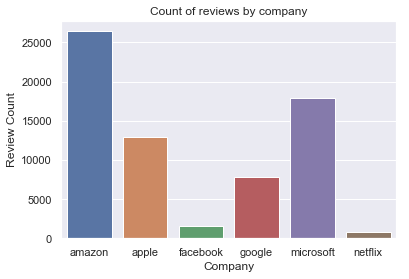
Since location is not a numerical column, it’s null values were filled using ffill and bfill methods after sorting the dataset by company. The “dates” column was converted from object datatype to datetime format to reduce processing times.

For the initial exploration, some of the columns (such as pros, cons, summary, job\_title, helpful\_count) are irrelevant, so they were dropped. Before doing that, a copy of the cleaned dataframe was created for future use.

**Data Story:**

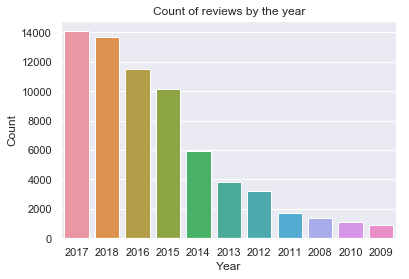
Using the nice cleaned dataset, we will begin our exploratory data analysis. The idea is to find patterns, anomalies, or relationships to inform our subsequent analysis.

Let’s first review the counts by each company.

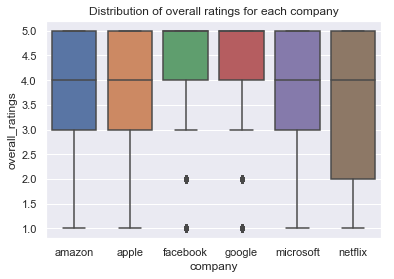


Apple and Microsoft have more than 10K reviews. Facebook and Netflix have less than 2000 reviews. Google has little less than 8K reviews. Amazon tops the charts with more than 25K reviews. The difference between number of Amazon and Netflix reviews is quite high so it is hard to determine if our analysis would be fair and accurate.

Next, let’s see the counts by year.

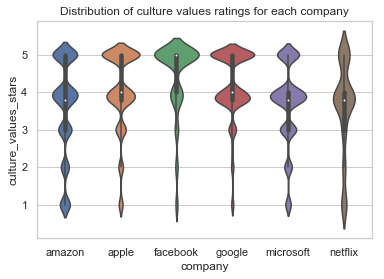


The counts are highest in recent years (2017-18) vs the earlier years (2009-10). One reason for this could be Glassdoor has become more popular over the years and considered more reliable for by employees for posting reviews.



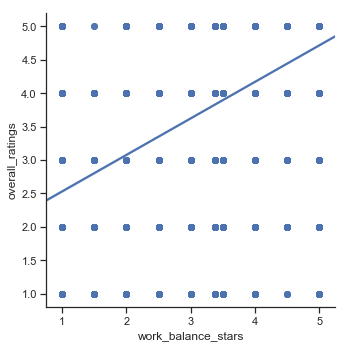
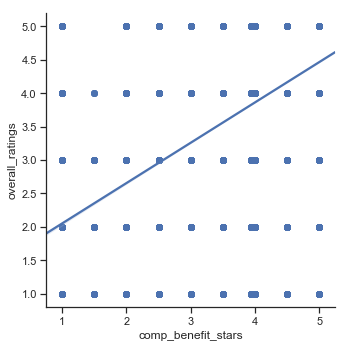
The above boxplot shows the distribution of overall ratings for each company. Interestingly, the interquartile range for Facebook and Google is from 4.0 to 5.0 with some outliers. The boxplot is skewed towards the higher values for all companies except Netflix.

Let’s look at a violin plot to find the distribution of culture and values ratings for each company. In a box plot, all the plot components correspond to actual data points whereas a violin plot features a kernel density estimation of the underlying distribution.



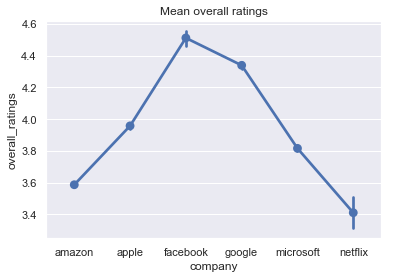
In this violin plot, the median is 5 for Facebook and close to 4 for all other companies. The interquartile range is lowest for Microsoft and Netflix.

Let’s draw an lmplot which is a 2D scatterplot with an optional overlaid regression line. The lmplot on the below left shows the relationship between work balance stars and overall ratings. The one on the right shows the relationship between compensation benefit stars and overall ratings.

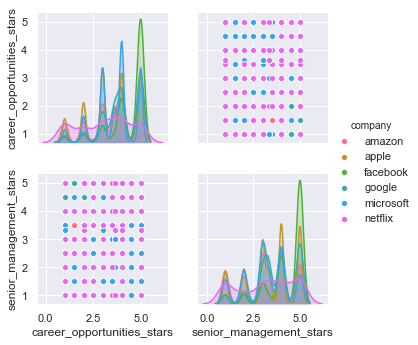
Looking at them, it is easy to conclude that the work balance stars as well as compensation benefit stars both have a positive relationship with overall ratings.

Next we have a point plot to determine the mean overall ratings for each company.



It is quite evident that Facebook received the highest average overall ratings whereas Netflix got the lowest.

Finally, we have a pairplot which will show pairwise relationships between two of our variables. A pairplot allows us to see both the distribution of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis.



The histogram on the diagonal shows the distribution of a single variable. The scatter plots on the upper and lower triangles show the relationship between the 2 variables which in this case are senior management stars and career opportunities stars.

**Conclusion:**

The dataset is rich and interesting and can be used to answer a lot of other questions such as:

1. Do current employees give more reviews than the ex-employees?
2. Is there a trend between the number of reviews and dates?
3. Were there a high number of reviews on a particular day?
4. What is the average overall rating? Which companies have consistently continued to stay above this average?
5. How do the ratings vary for each company by individual locations?
6. Trend of ratings using date and location