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 2008 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. They were converted from object datatype to float.

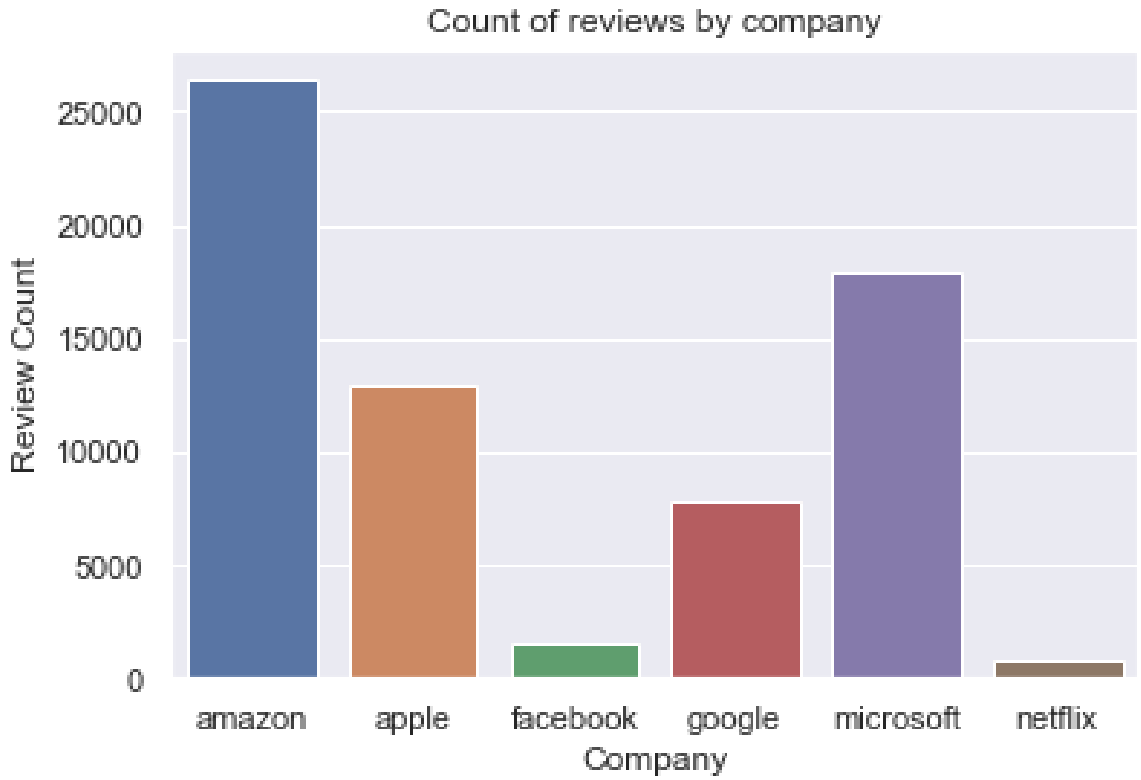
The “dates” column was converted from object datatype to datetime format to reduce processing times.

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.

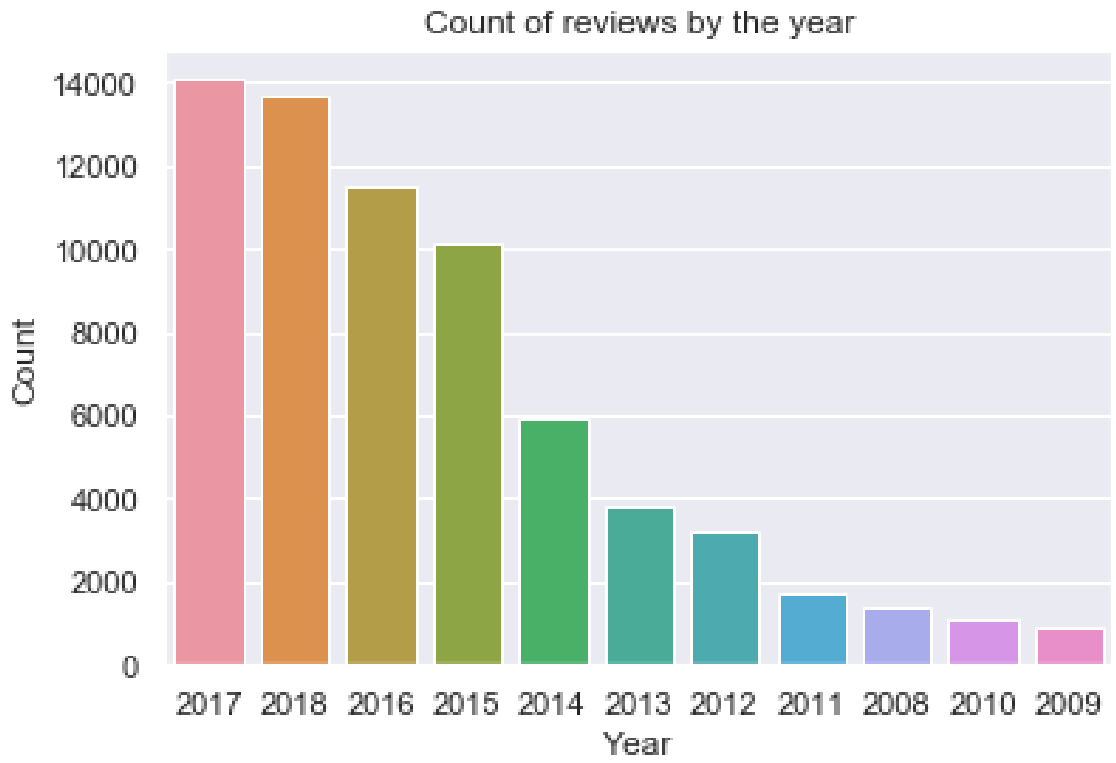
Let's first review the counts by each company.



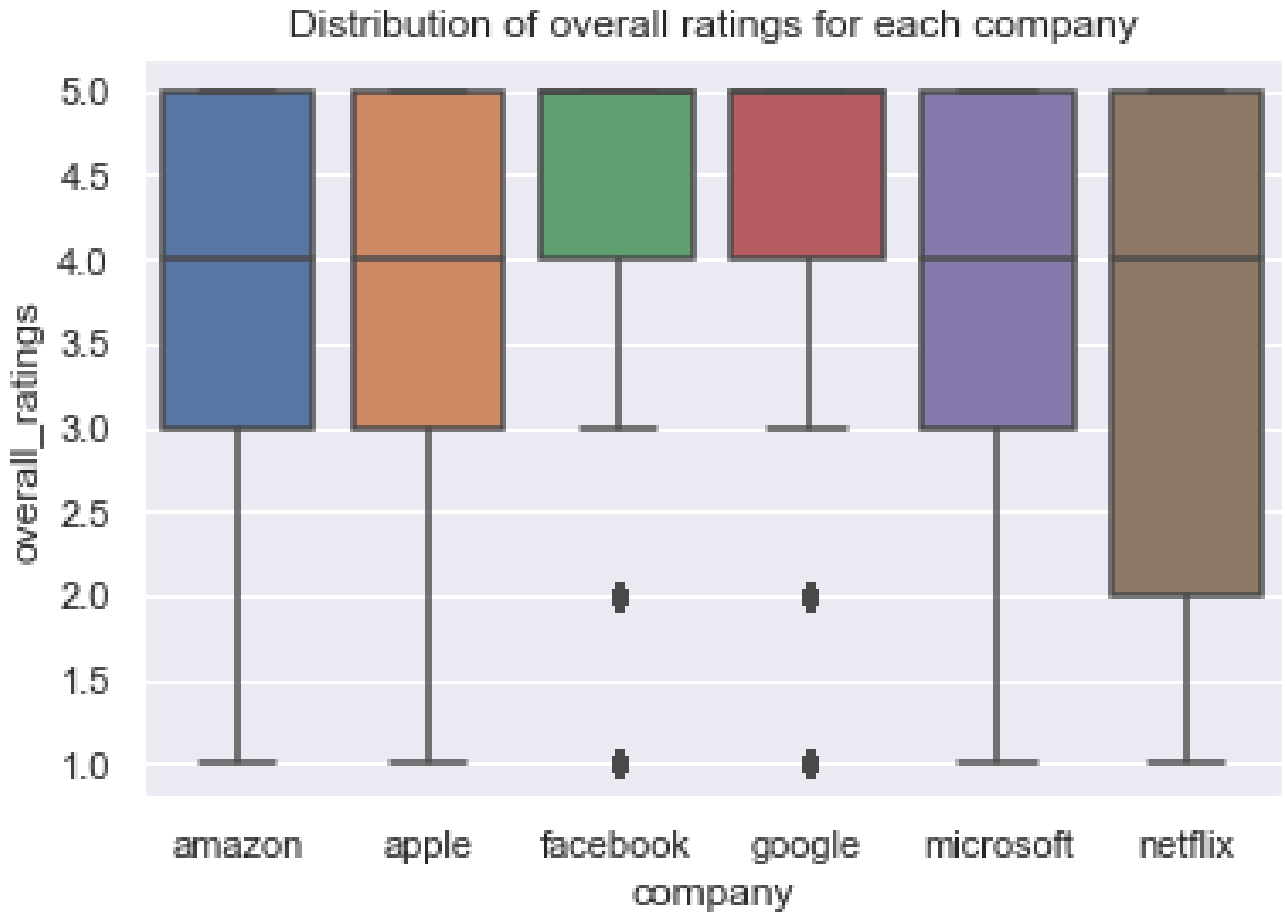
Apple and Microsoft have more than 10K reviews. Facebook and Netflix have less than 2000 reviews. Google has little less than 8K reviews. Amazon tops the charts with more than 25K

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.

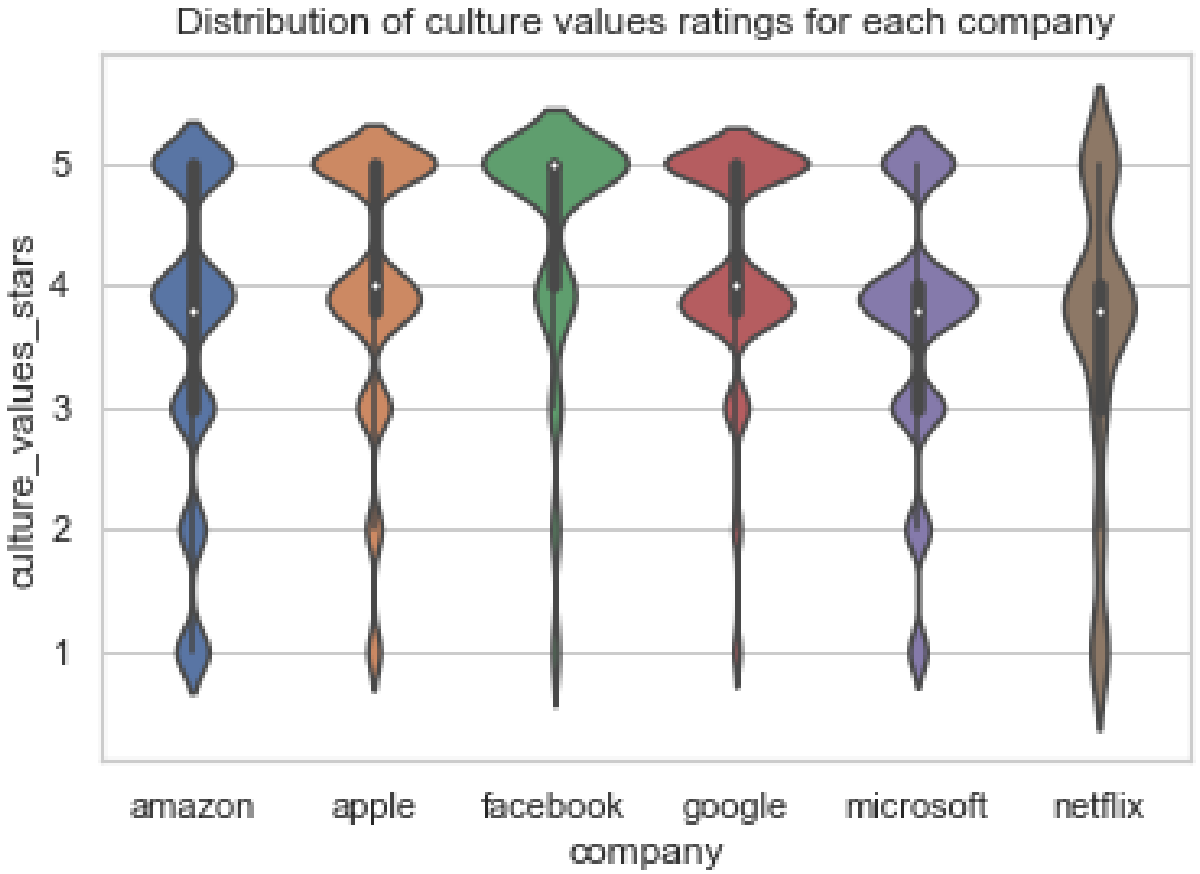


The counts are highest in recent years (2017-18) vs the earlier years (2009-10). One reason for this could be Glassdoor has become more popular over the years and considered more reliable for by employees for posting reviews. Although it must be noted that 2008 had more reviews than 2009 and 2010.



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.

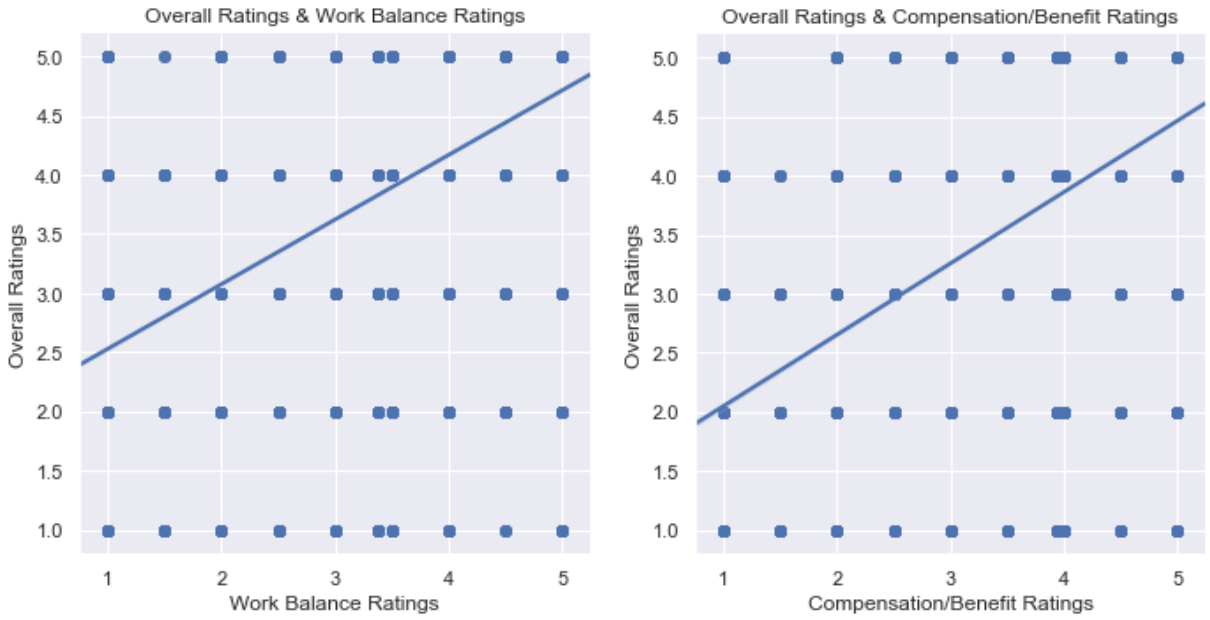
Let’s look at a violin plot to find the distribution of culture and values ratings for each company. In a box plot, all the plot components correspond to actual data points whereas a violin plot features a kernel density estimation of the underlying distribution.



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 lmplots below shows the relationship between overall ratings and work_balance_stars, comp_benefit_stars, career_opportunities_stars and senior_management_stars.

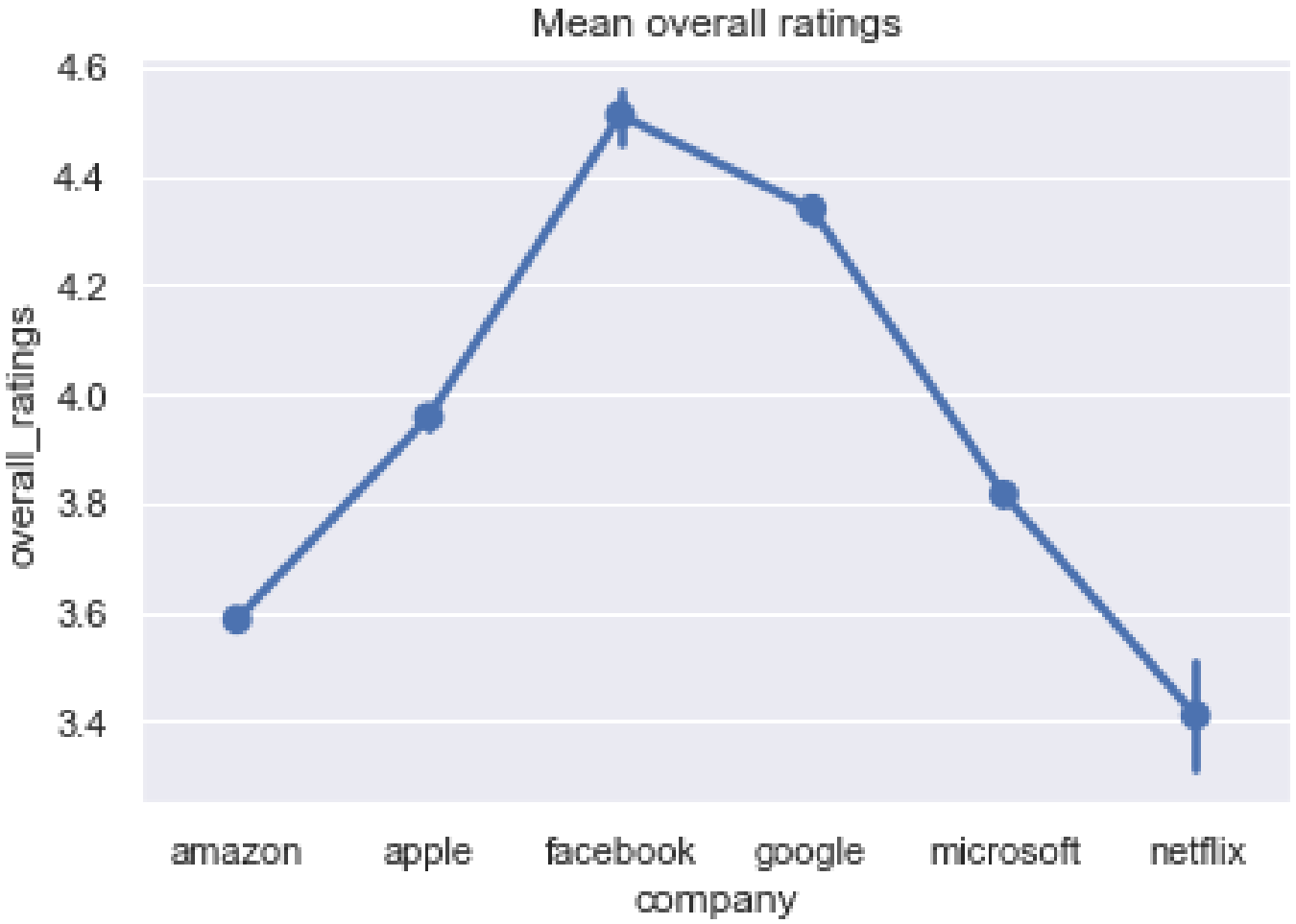
Looking at them, it is easy to say that all these variables have a positive relationship with overall ratings.




Overall Ratings & Career Opportunities Ratings
Overall Ratings
Career Opportunities Ratings
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1
2
3
4
5

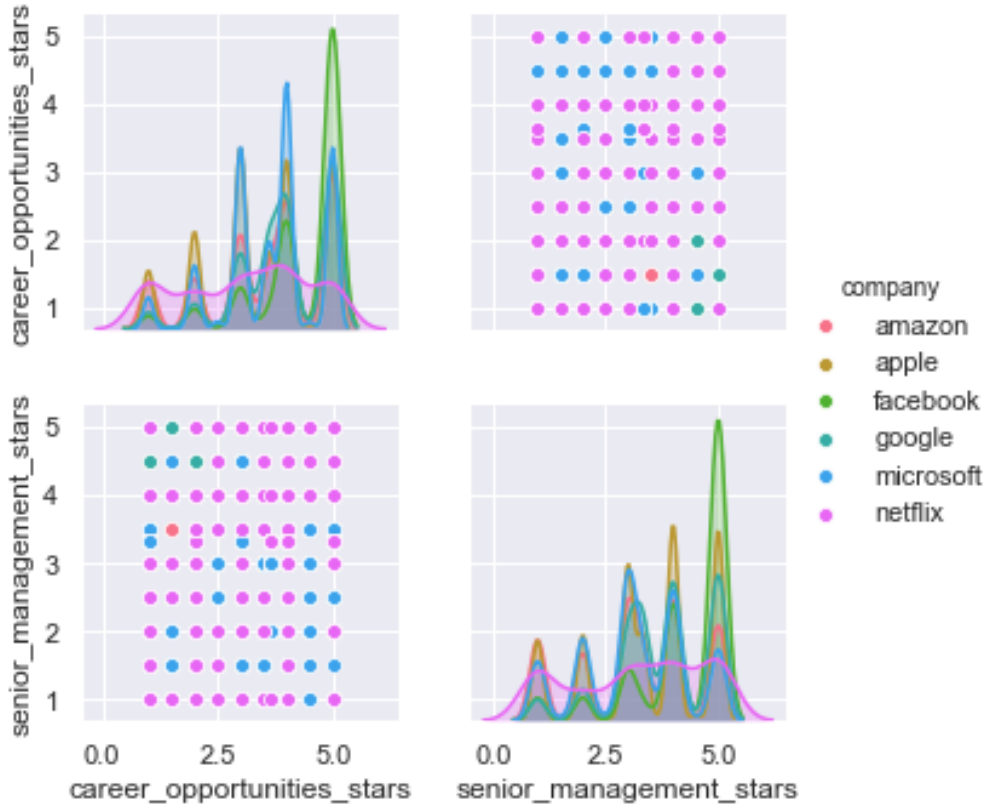

Overall Ratings & Senior Management Ratings
Overall Ratings
Senior Management Ratings
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
1
2
3
4
5

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



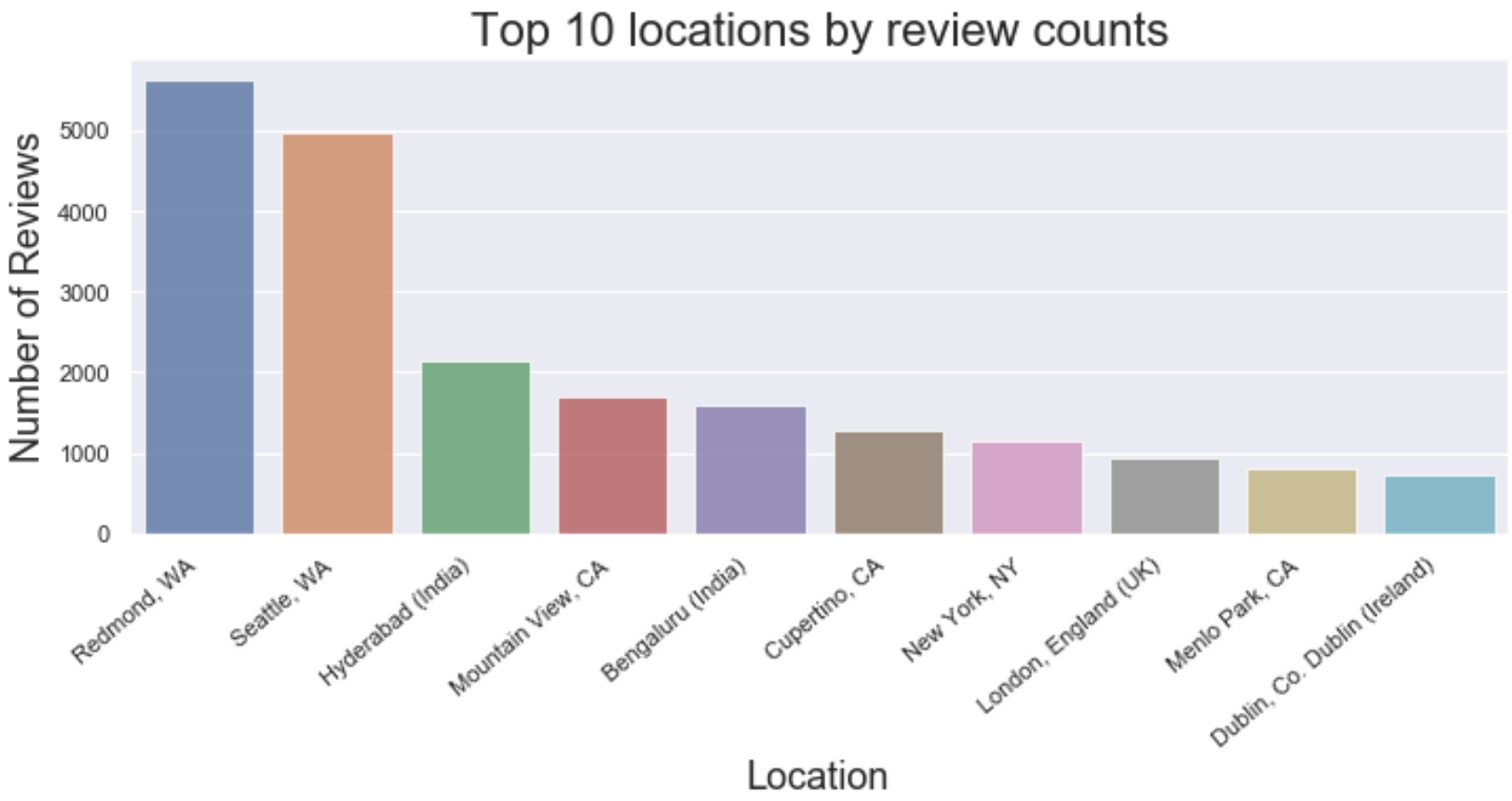
It is quite evident that Facebook received the highest average overall ratings whereas Netflix got the lowest.

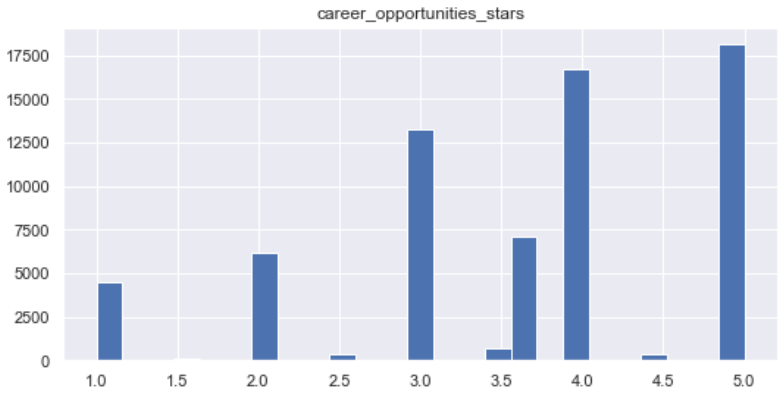
Finally, we have a pairplot which will show pairwise relationships between two of our variables. A pairplot allows us to see both the distribution of single variables and relationships between two variables. Pair plots are a great method to identify trends for follow-up analysis.

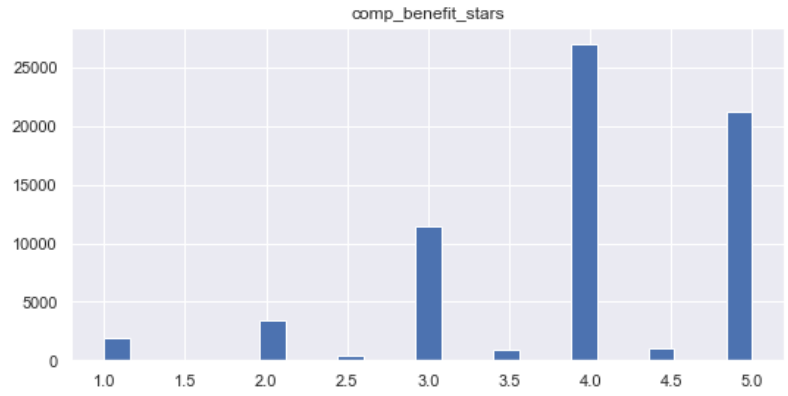


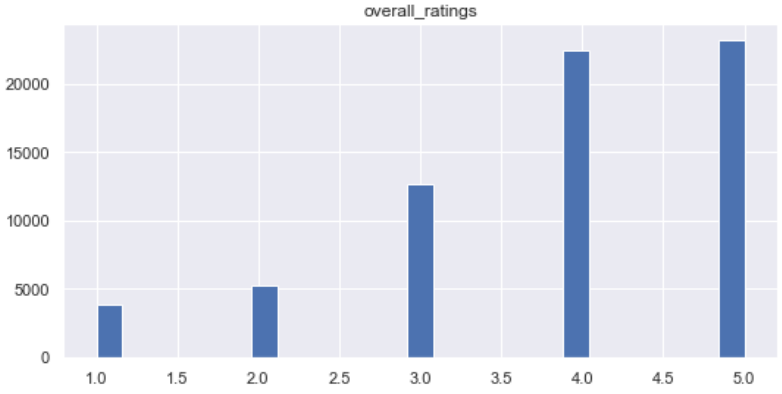
The histogram on the diagonal shows the distribution of a single variable. The scatter plots on the upper and lower triangles show the relationship between the 2 variables which in this case are senior management stars and career opportunities stars.

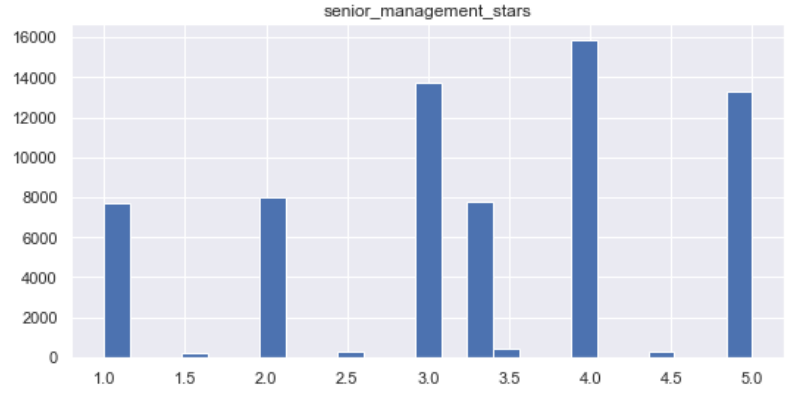
By plotting the Top 10 locations by review counts, we find that most reviews came from Redmond, WA and Seattle, WA. Amazon and Apple are both headquartered in these cities and we have seen in the graph above that these 2 companies have received the most reviews from their former and current employees.

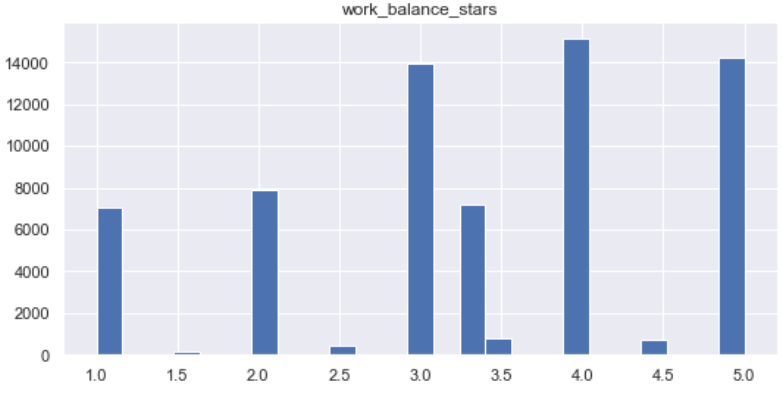



career_opportunities_stars
17500
15000
12500
10000
7500
5000
2500
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


comp_benefit_stars
25000
20000
15000
10000
5000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


overall_ratings
20000
15000
10000
5000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


senior_management_stars
16000
14000
12000
10000
8000
6000
4000
2000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0


work_balance_stars
14000
12000
10000
8000
6000
4000
2000
0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0

Exploratory Data Analysis:

Now it's time to make some inferences from our data.

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. Which variables are the biggest predictor for the target variable (overall_ratings)?
6. Trend of ratings using date and location
7. More than 35% of the rows have missing location and about 20% have missing values in culture_values_stars. Therefore, we will not be using these two columns for any of our analysis. The other ratings columns have less than 10% rows with missing values. We computed the mean and median of these columns and found them to be about the same. So the missing values on them were filled with the mean of that column.

What is the main question for this project?

I would say the most important question is: What are the various factors that influence the overall_ratings review an employee gives to their ex- or current employer?

Are there variables that are particularly significant in terms of explaining the answer to your project question?

In the dataset, we have many variables that could be significant in how an employee rates his employer. Right now, we will just focus on the variables that have a numerical value for the rating. These variables are: work_balance_stars, career_opportunities_stars, comp_benefit_stars and senior_management_stars. All the other features have been removed from the dataframe.

Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?

We ran a correlation matrix between the target variable (overall_ratings) and the four numerical ratings columns. From that we found that the variables work_balance_stars and comp_benefit_stars both have a moderate positive relation with overall_ratings. On the other hand, career_opportunities_stars and senior_management_stars both have a strong positive relation with overall_ratings. We plotted scatter plots earlier which confimed that our four predictor variables have a linear and positive relationship with overall_ratings.

Typically, employees are most happy when they receive good career growth opportunities within their organization. And such opportunities are usually provided by the senior management. So let's

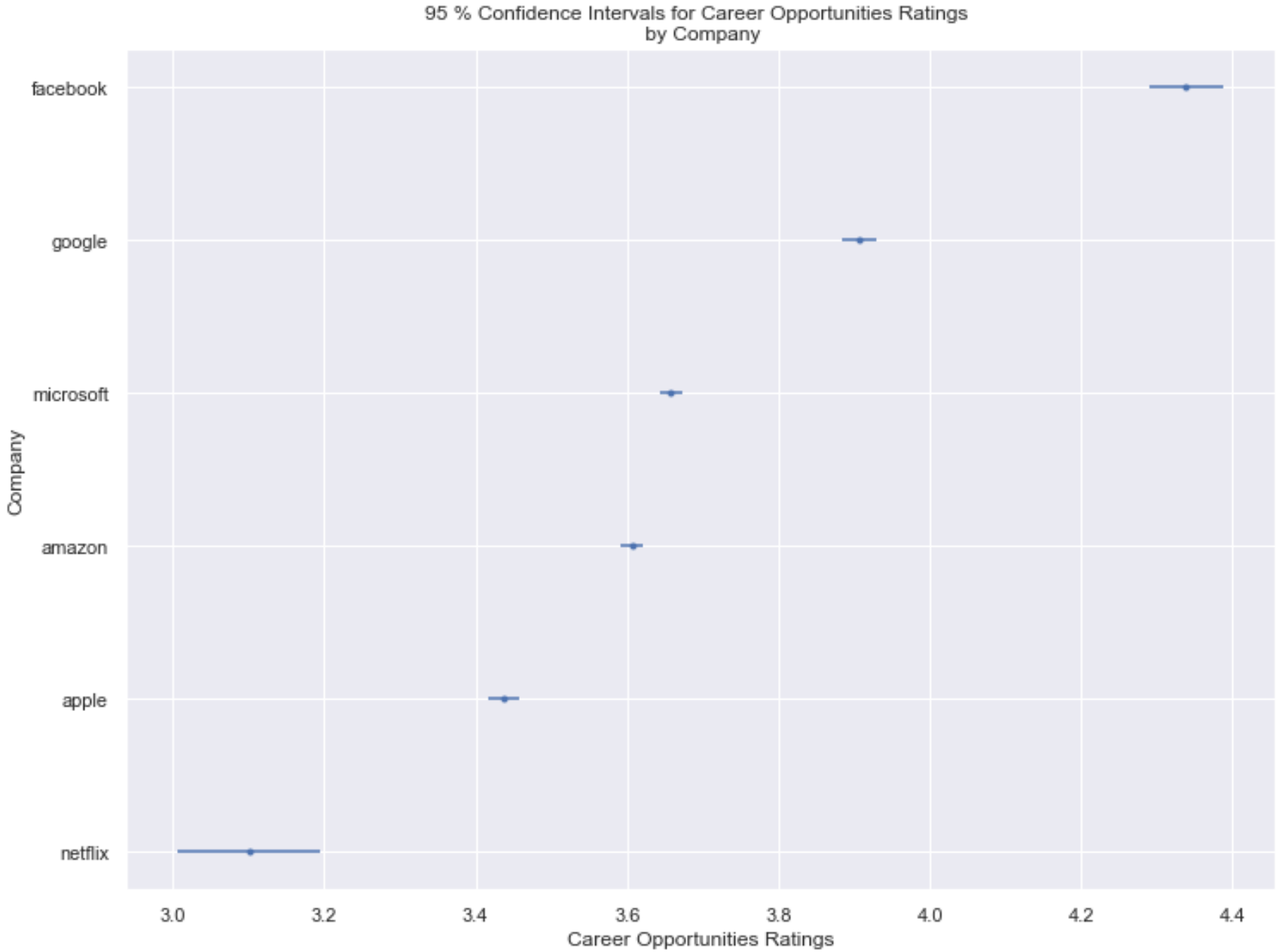
check the correlation between career_opportunities_stars and senior_management_stars and see if we can prove our theory.
The correlation between career_opportunities_stars and senior_management_stars is strong (correlation coefficient = 0.62). The correlation between work_balance_stars and comp_benefit_stars is moderate (correlation coefficient = 0.41).

What are the most appropriate tests to use to analyse these relationships?

The Pearson correlation test assumes that the variables are normally distributed whereas the Spearman Correlation does not assume that the datasets are normally distributed. We ran both tests on our variables.

Null hypotheses: the variables are independent
Alternative hypotheses: the variables are not independent but correlated.
The Pearson correlation between overall_ratings and the predictor variables reported a p-value of 0.0 so we reject the null hypotheses and accept the alternative hypotheses that there is a correlation.

Another test that was ran is the chi-square test.
This test also gave us a p-value of 0.0 which is less than 0.05. Hence we reject the null hypotheses and deduce that the variables are correlated.


95 % Confidence Intervals for Career Opportunities Ratings
by Company
facebook
google
microsoft
amazon
apple
netflix
Company
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
Career Opportunities Ratings

Conclusion:

All the ratings variables are correlated and show as pattern. Further analysis should be done using regression models and F-tests. I think it will be interesting to also analyze the summary, pros, cons and advice_to_mgmt columns. They contain textual data and I would like to use NLP techniques to conduct text processing and sentiment analysis on them.