EMPLOYEE REVIEW ANALYSIS

VARSHA TRIPATHI

SPRINGBOARD DATA SCIENCE CAPSTONE PROJECT

MAY 2019

WHICH EMPLOYERS?

- AMAZON
- APPLE
- FACEBOOK
- GOOGLE
- MICROSOFT
- NETFLIX

THE APPROACH

- Create a model to determine which category of ratings affects the overall ratings of a company the most
- Find the correlation between the various employer ratings

THE CLIENT

- JOB SEEKERS
  - Use this report to determine which employers have increased their ratings over the years
  - Research potential employers without having to browse through different websites that have employee feedback
- EMPLOYERS
  - Use this report to determine the categories where they have consistently received low ratings and work upon improving them

THE DATASET

- **SOURCE:** 1 Dataset hosted by KAGGLE
- **LINK:** HTTPS://WWW.KAGGLE.COM/PETERSUNGA/GOOGLE-AMAZON-FACEBOOK-EMPLOYEE-REVIEWS
- **TOTAL RECORDS:** 67529
- **TIMELINE:** 2008-2018
- **COLUMNS:** WORK-BALANCE-STARS, CULTURE-VALUES-STARS, CAREER-OPPORTUNITIES-STARS, COMP-BENEFIT-STARS, SENIOR-MANAGEMENT-STARS, OVERALL-RATINGS

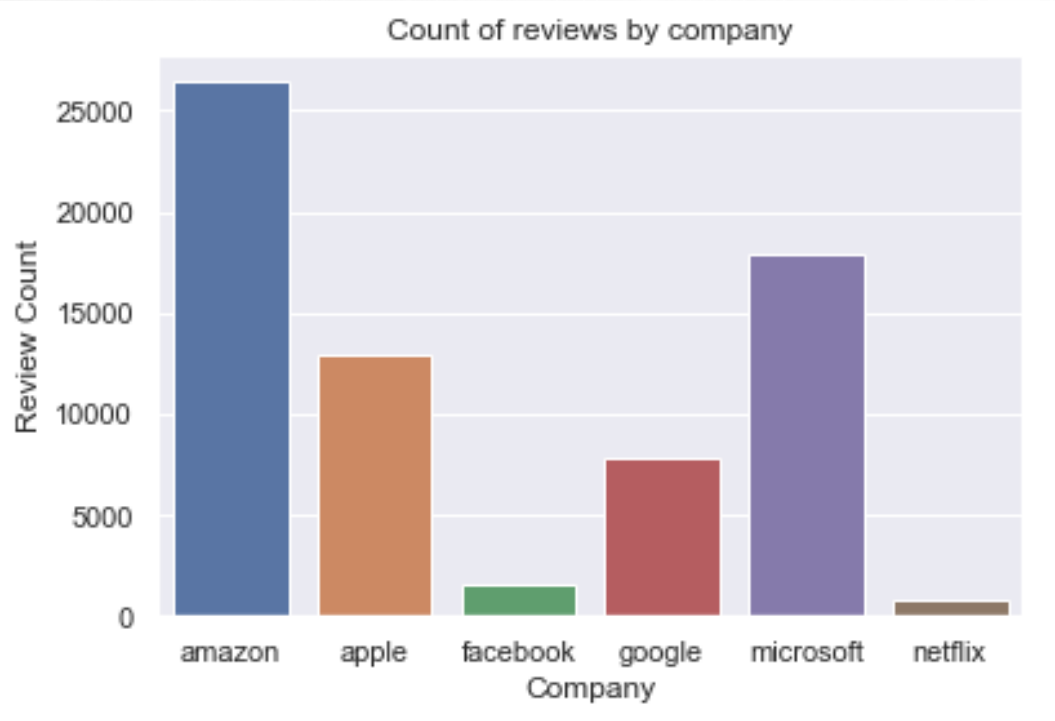
DATA WRANGLING

- Converted the “dates” column from object datatype to datetime format to reduce processing times.
- Converted the ratings columns from object datatype to float.
- Dropped the text columns – pros, cons, summary, job_title.
- Correctly labelled all the misspelled columns.
- Converted “none” values on the columns to NaN.
- Dropped “Location” and “culture_values_stars” columns as they had missing values on 20% or more rows
- Filled the missing values on other ratings columns with the means of those columns.

CLEANED DATA

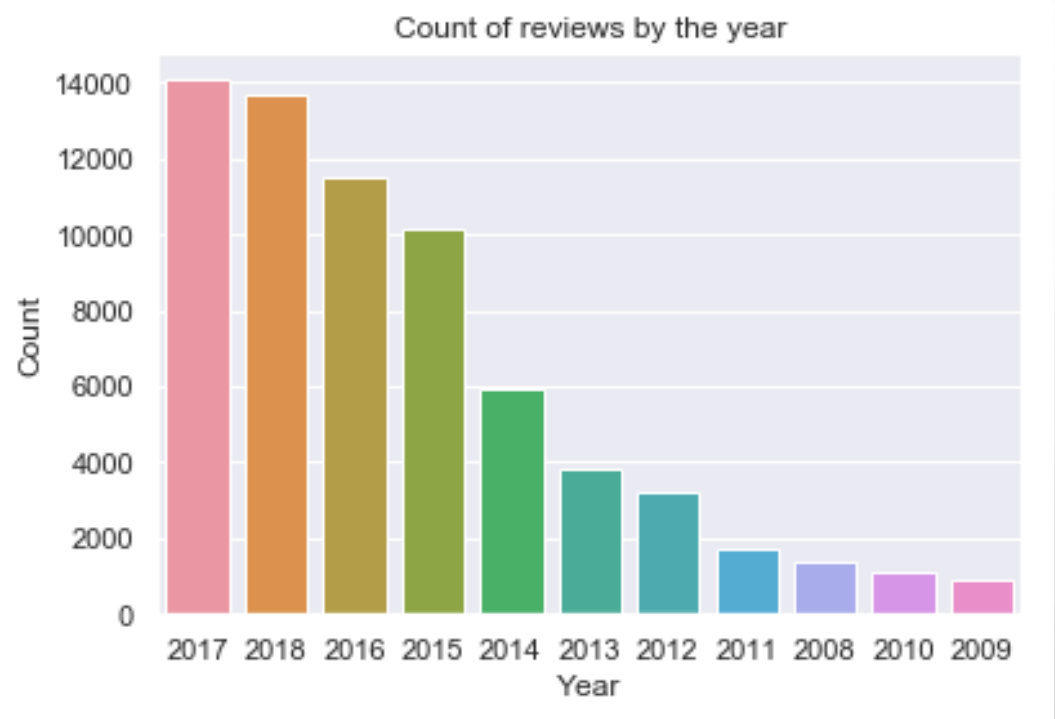
company	Name of the Company	Object
location	Branch location	Object
dates	Date of review	Datetime
job_title	Job title of employee	Object
summary	Summary	Object
pros	Company pros	Object
cons	Company cons	Object
advice_to_mgmt	Advice to senior management	Object
overall_ratings	Overall rating	Float64
work_balance_stars	Work-life balance rating	Float64
culture_values_stars	Culture and values rating	Float64
career_opportunities_stars	Career opportunities rating	Float64
comp_benefit_stars	Compensation and benefits rating	Float64
senior_management_stars	Senior management rating	Float64
helpful_count	How many found this helpful	Int64
link	Website link	Object

EDA – COUNT OF REVIEWS BY COMPANY



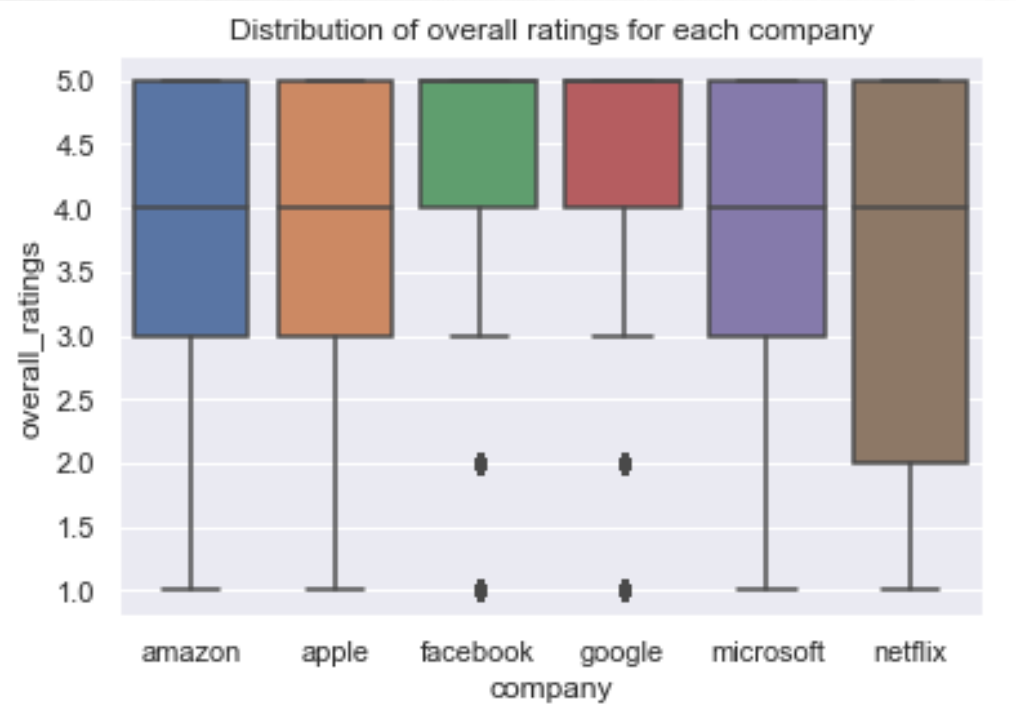
1. Amazon
2. Microsoft
3. Apple
4. Google
5. Facebook
6. Netflix

EDA – COUNTS OF REVIEWS BY THE YEAR



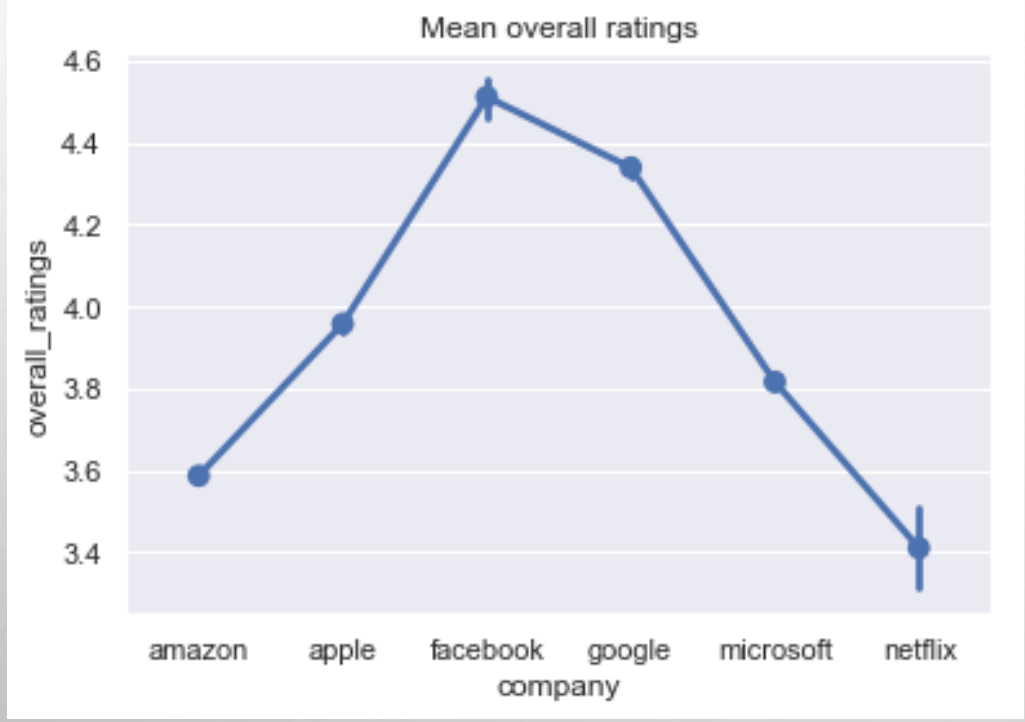
- Highest count: 2017
- Lowest count: 2009
- 2018 has less counts than 2017
- 2008 had higher counts than 2009 and 2010

EDA – OVERALL RATINGS DISTRIBUTION



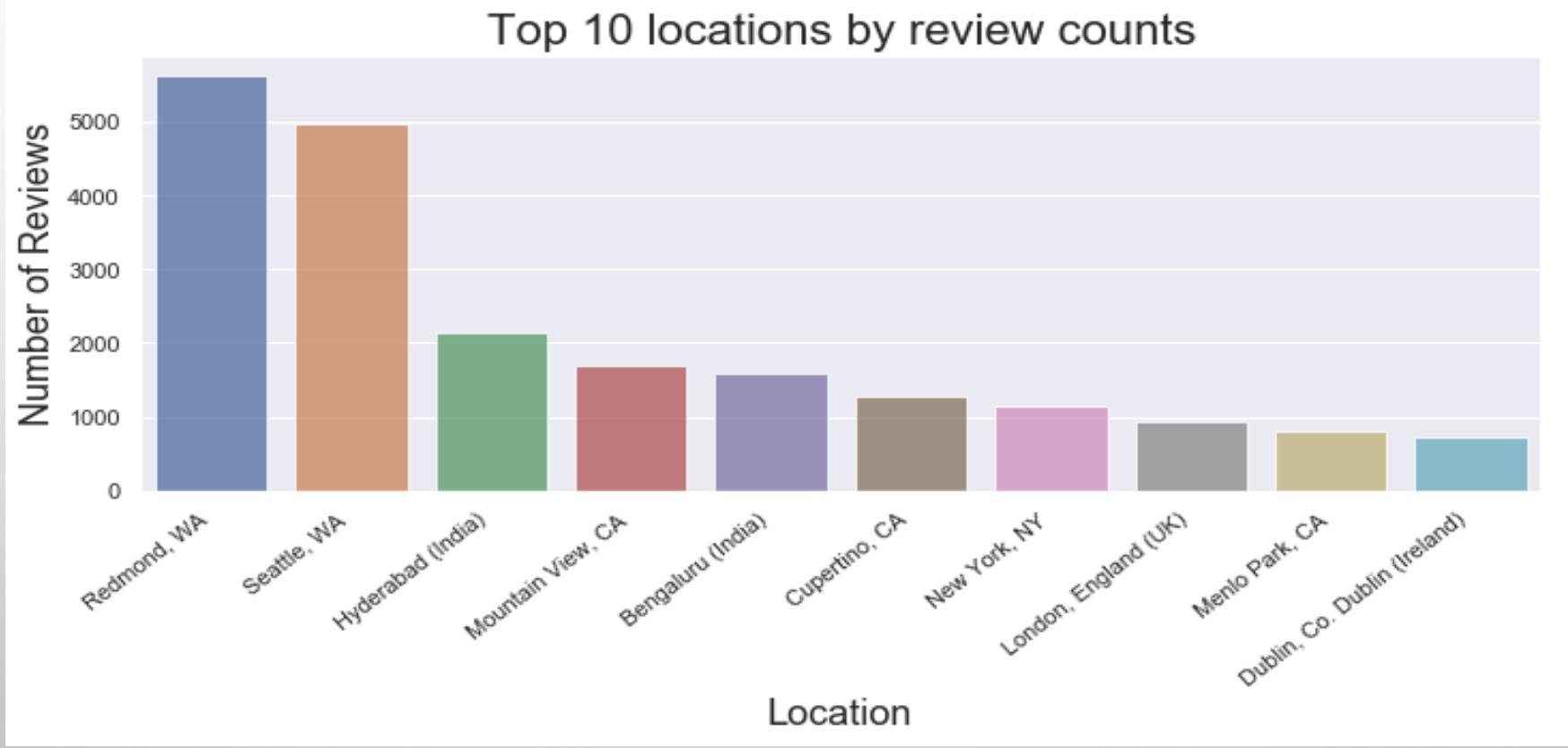
- Interquartile range is between 4 and 5 for Facebook and Google with some outliers.

EDA – MEAN OVERALL RATING

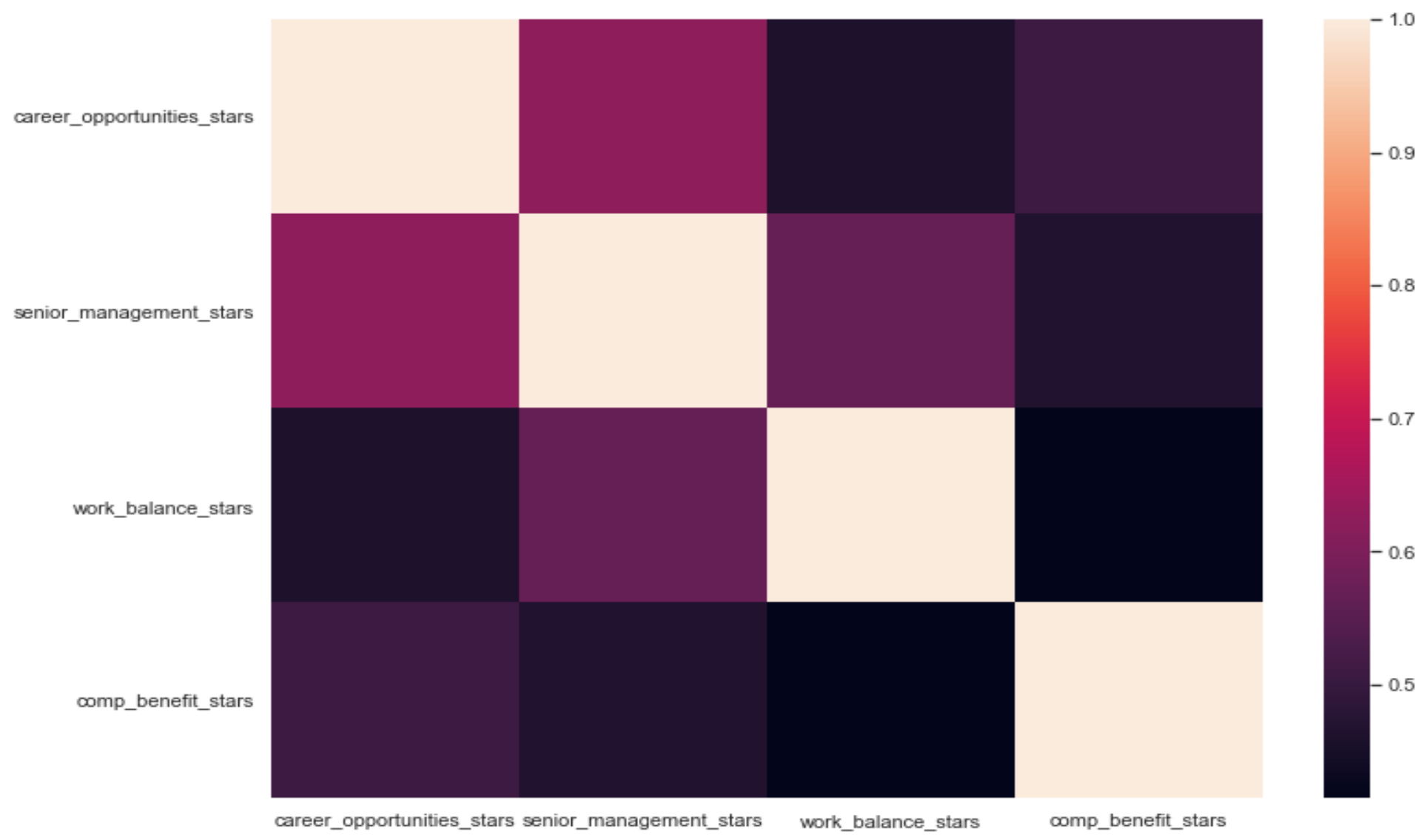


1. Facebook
2. Google
3. Apple
4. Microsoft
5. Amazon
6. Netflix

EDA – TOP 10 LOCATION COUNTS



CORRELATION MATRIX



CORRELATION WITH OVERALL_RATINGS

- SENIOR_MANAGEMENT_STARS 0.688452

  CAREER_OPPORTUNITIES_STARS 0.659956

  WORK_BALANCE_STARS 0.580291

  COMP_BENEFIT_STARS 0.512315

- Work_balance_stars and comp_benefit_stars both have a moderate positive relation with overall_ratings.
- Career_opportunities_stars and senior_management_stars both have a strong positive relation with overall_ratings.

CORRELATION BETWEEN INDEPENDENT VARIABLES

- The correlation between career_opportunities_stars and senior_management_stars is strong (correlation = 0.625).

- The correlation between work_balance_stars and comp_benefit_stars is moderate (correlation = 0.415).

MACHINE LEARNING MODELS APPLIED ON THE DATASET

- LINEAR REGRESSION
- RIDGE REGRESSION
- LASSO REGRESSION
- RANDOM FOREST REGRESSOR
- GRADIENT BOOSTING REGRESSOR

SCORE TABLE

Algorithm	r2_train	r2_test	mse_train	mse_test
LinearRegression	0.6062	0.5991	0.5265	0.5301
Lasso	0.1729	0.1730	1.1058	1.0935
Ridge	0.6038	0.5977	0.5297	0.5319
RandomForestRegressor	0.6465	0.6058	0.4727	0.5212
GradientBoostingRegressor	0.5324	0.4935	0.1023	0.1110

FUTURE WORK

- Use the text columns such as summary, pros, cons and advice to mgmt. NLP techniques can be used on their content to conduct text processing and sentimental analysis.
- Ratings by location and ratings by ex-employees vs current employees will be another good data point to analyze.
- Analyze if and why current employees give more reviews than the ex-employees
- Is there a trend between the number of reviews and dates?
- Were there a high number of reviews on a particular day?
- What is the average overall rating? Which companies have consistently continued to stay above this average?

CONCLUSION

- Overall ratings of a company are
  - Most affected by senior management ratings
  - Least affected by the compensation and benefits ratings.
- The quality of senior leadership is most important for employee satisfaction followed by career growth opportunities.
- Work-life balance and compensation/benefits matter the least for workplace happiness.

ACKNOWLEDGEMENT

- Mentor: Max Sop
- Kaggle.com dataset
- Springboard team