Navigating Corporate Labyrinth: A 10,000 Company Rating Analysis

Bridget Sheng, Class of 2024
Data Science Major Capstone

Background and Research Questions

Understanding the corporate landscape is important for students to make informed decisions for life after college. Based on the profiles of 10,000 companies, **this capstone project seeks to understand the relevant factors that influence corporate success**, aiming to provide peers with a deep understanding of what drives company ratings in today's environment.

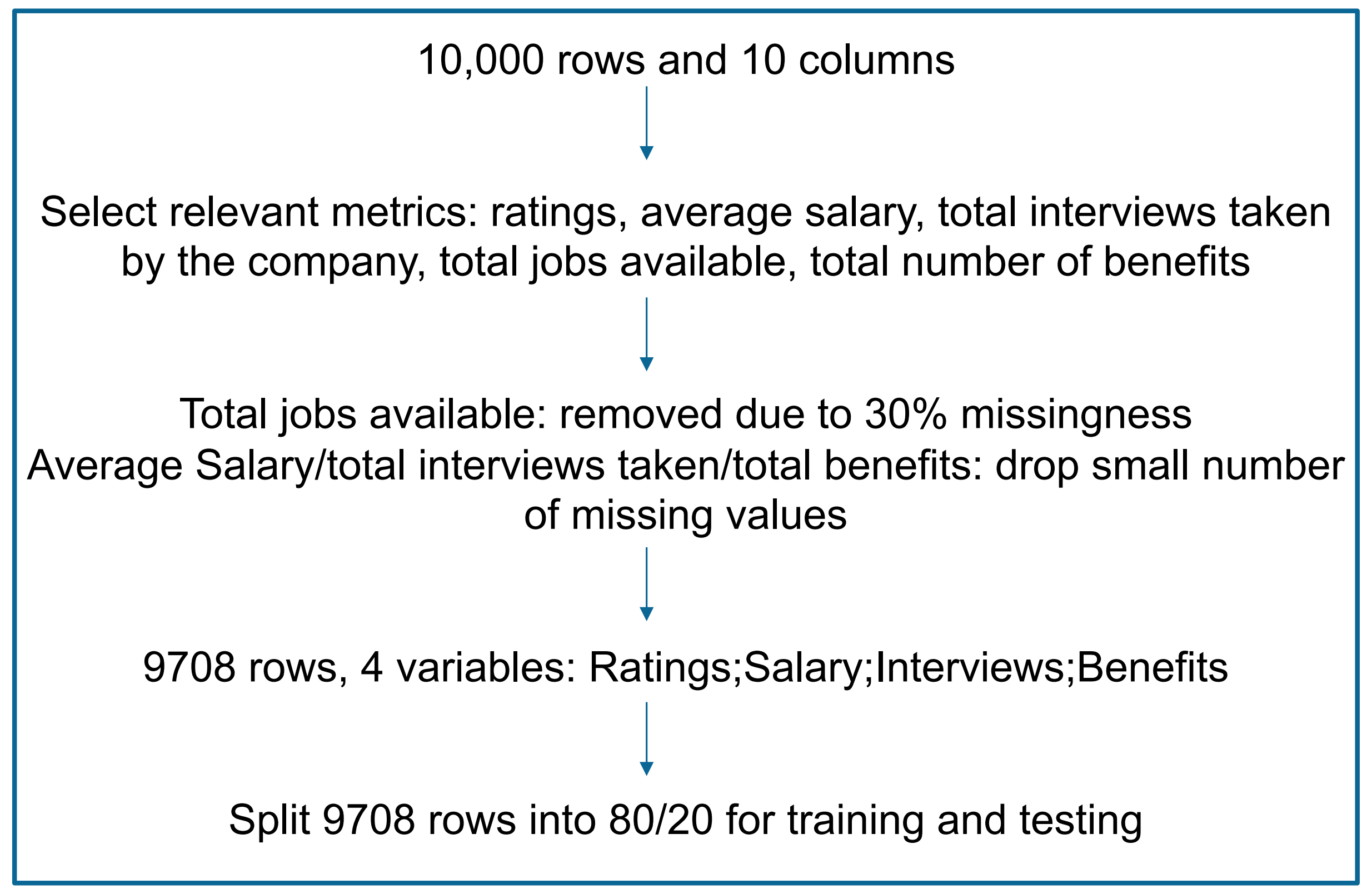
This study uses the top 10,000 company or startup profiles from Ambition Box [1]. By investigating three variables, this project intends to answer the following research questions:

1. How are average salary, total number of interviews conducted by the company, and total number of benefits related to a company's rating?
2. Among average salary, number of interviews taken, and total benefits, what variables are most relevant to predict a company's ratings?

Data Processing and Methodology

Overview and Cleaning:

The dataset was extracted from Ambition Box, a website that provides company reviews. It includes character variables (company name, description, etc.) and numeric variables (ratings, total review, average salary, etc.). The dataset was processed in the following steps:


10,000 rows and 10 columns

Select relevant metrics: ratings, average salary, total interviews taken by the company, total jobs available, total number of benefits

Total jobs available: removed due to 30% missingness
Average Salary/total interviews taken/total benefits: drop small number of missing values

9708 rows, 4 variables: Ratings;Salary;Interviews;Benefits

Split 9708 rows into 80/20 for training and testing

Methods:

- Based on initial visualization analysis, the original data may not meet assumptions for linear regression. Hence, log transformations were performed to normalize the data and reduce outliers for regression models
- Simple and first-order linear regression models were used to analyze the dataset

Initial Data Visualization

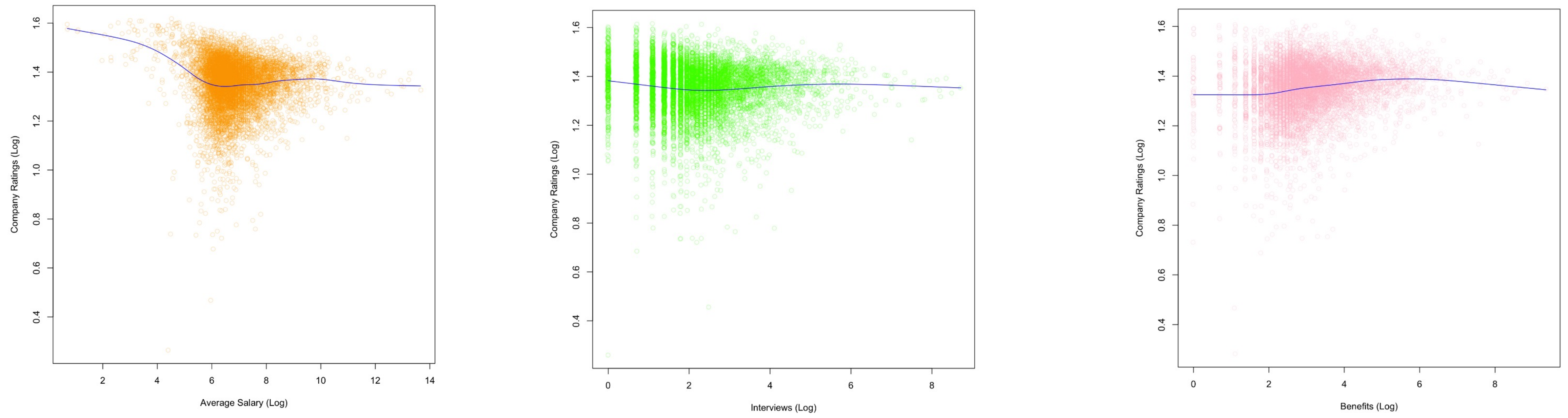


Figure I. Relationship between each predictor (average salary, total interviews taken, total benefits) and company rating on a log scale

Data Modeling: Simple Linear Regression

- Simple linear regression models for each variable were run to identify their associations with company ratings.
- All three models generated statistically significant results with small p-values.
- The model using *Benefits Log* as a predictor had significant positive coefficient and highest R^2 of 0.0269 and adjusted R^2 of 0.0267 among the three models.
- *Benefits Log* was chosen as the base predictor variable to perform first-order models for better fits.

Model	Intercept	Coefficient	P-Value	R^2	Adjusted R^2	BIC
Salary Log	1.383	-0.00443	6.38e-05	0.00206	0.00193	-13094
Interviews Log	1.358	-0.00249	0.0133	0.000789	0.000661	-13084
Benefits Log	1.305	0.0154	<2e-16	0.0269	0.0267	-13289

Data Modeling: First Order Linear Regression

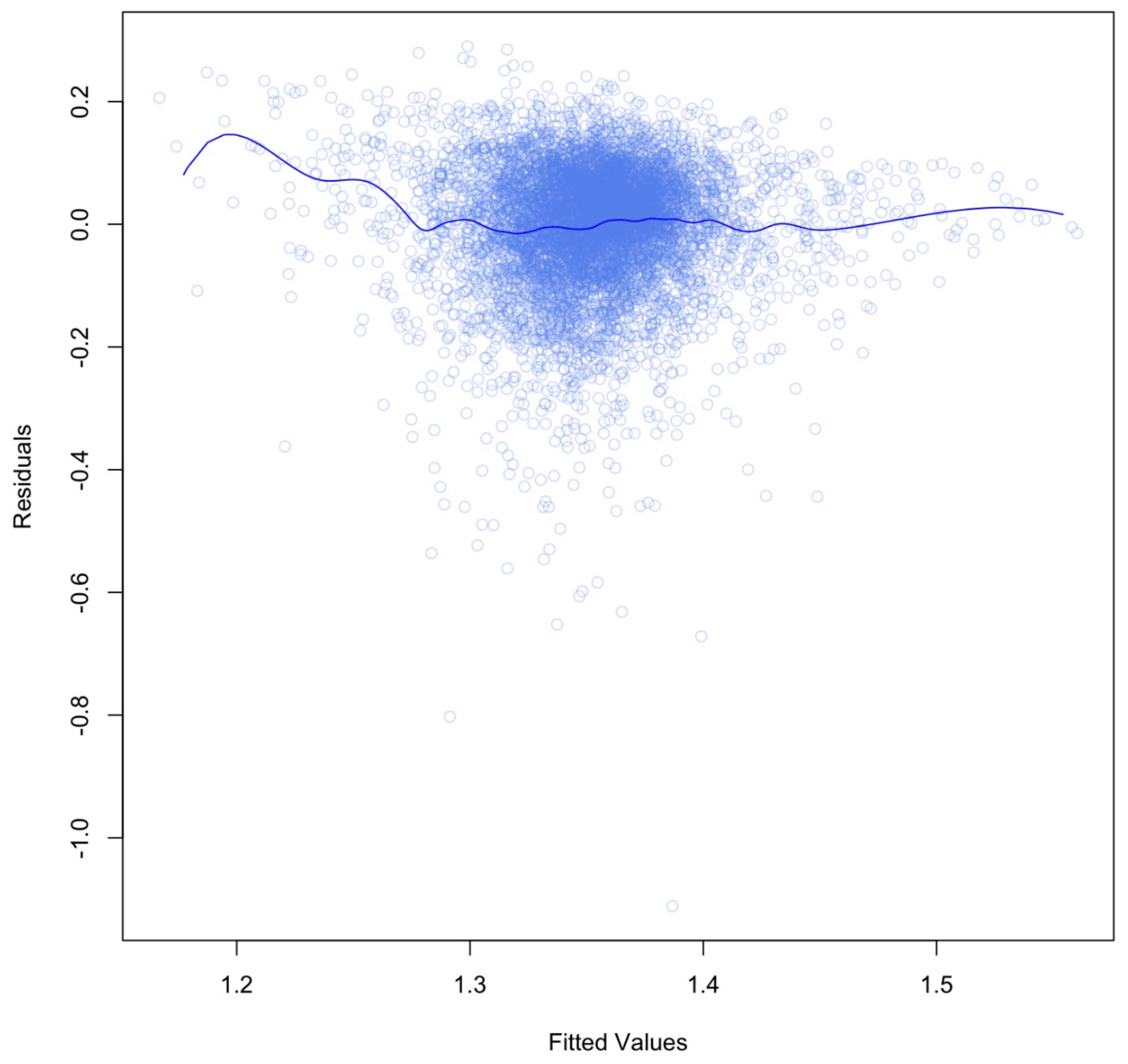


Figure II. Residuals versus Fitted values for the best fit model (benefits+interviews+salary). All the variables are log-transformed.

•*Interviews* and *Salary* were added to the first order model. Each was combined with *Benefits* to test model fit alone, and all three predictors were combined together.
•R^2 and adjusted R^2 increase for first-order models and BIC decreases, indicating better fits with more predictors

Model	Coefficient / P-Value	R^2	Adjusted R^2	BIC
Benefits +Salary	(Intercept) 1.489 <2e-16 BenefitsLog 0.0447 <2e-16 SalaryLog -0.0400 <2e-16	0.0965	0.0963	-13857
Benefits +Interviews	(Intercept) 1.300 <2e-16 BenefitsLog 0.0363 <2e-16 InterviewsLog -0.0273 <2e-16	0.0721	0.0719	-13650
Benefits +Interviews +Salary	(Intercept) 1.450 < 2e-16 BenefitsLog 0.0478 < 2e-16 SalaryLog -0.0321 < 2e-16 InterviewsLog -0.0116 8.35e-12	0.102	0.102	-13895

Results

- Simple linear models show that average salary and total number of interviews conducted have a slightly negative association with company ratings, while total number of benefits have a positive relationship.
- The model with all three predictors appears to be the best one with the highest adjusted R^2 of 0.102 and lowest BIC of -13895
- Applying the best fit model for the testing data, a mean absolute error (MAE) of 0.277 and a mean squared error (MSE) of 0.129 were obtained.

Discussions and Limitations

Unexpected negative coefficients under simple linear regression: Higher salaries and more interviews correlate with lower company ratings, possibly due to demanding job environments and competitive selection processes. Conversely, a greater number of benefits improves ratings, which highlights the value placed on welfare.

First-order Model: The model with all three predictor variables presents the best fit, indicating that a company's rating is a multifaceted issue.

Testing data: The model's predictions closely match actual ratings, with low errors (MSE of 0.129 & MAE of 0.277 for a scale from 1 to 5). It performs well on test data and is potentially suitable for real-world applications.

Limitation 1) The analysis focuses on three variables. Incorporating additional ones, such as aspects of the company that are highly rated for, could offer for a more comprehensive understanding for the corporate landscape. **2)** Despite using log transformation, the residual and fitted value pattern (Figure II) indicate potential model assumption violations. This suggests further data transformation or alternative modeling practices might be needed.

Data Ethics & Conclusion

Information about individual companies' jobs and salaries could raise concerns regarding the transparency of data collection. Additionally, biases within the rating processes should also be considered.

In conclusion, company ratings are affected by multiple factors: benefits increase ratings, while higher salaries and more interviews conducted decrease them. Peers should thoroughly assess companies before considering opportunities

Reference:
[1] Kaggle Dataset: https://www.kaggle.com/datasets/vedantkhapekar/top-10000-companies-dataset/data