Determinants of Salary Packages in Indian Companies:

A Multiple Regression Analysis

Group Members

Anmol Sai Ramchandran Ramanan (AR2867) | Devang Chaturvedi (DC1872) | Kanishk Deshwal (KD1088) | Shreyas Pantangi (SP2976)

1. Introduction

Compensation benchmarking has gained relevance in India's expanding corporate ecosystem, where salary competitiveness varies significantly across industries, cities, and organizational categories. This study analyzes salary determinants across **10,000+ Indian companies** using multiple regression modeling conducted in **R**.

The primary goals were:

1. Determine which company attributes significantly explains salary levels.
2. Build a statistically valid and interpretable regression model.
3. Diagnose and correct potential modeling violations (skewness, outliers, heteroscedasticity).

---

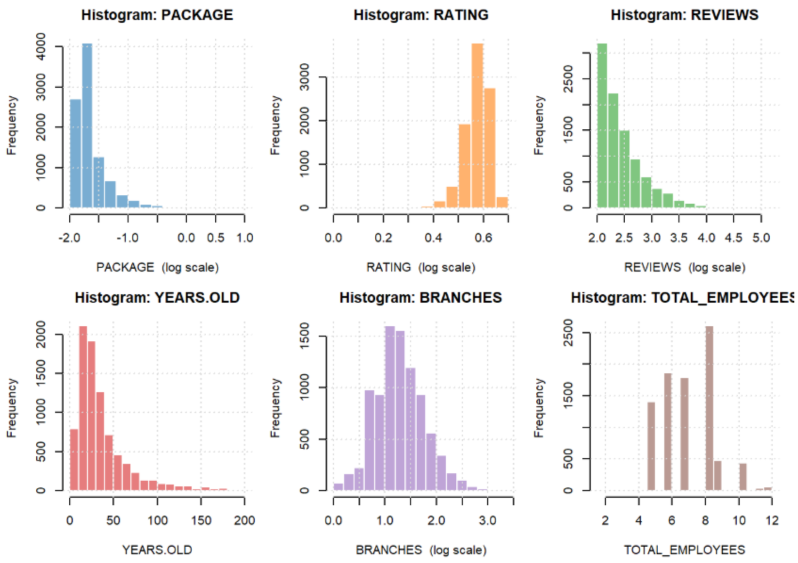
2. Dataset Overview & Preprocessing

The dataset was obtained from Kaggle: 'https://www.kaggle.com/datasets/ashura369/indian-companies-complete-data-2025-10000'

Variables Scraped

- Salary package (dependent variable)
- Company rating
- Reviews volume
- Years of establishment
- Branch count
- Total employee count
- Headquarters location
- Industry category

Due to heavy right-skewness (start-up vs corporate spread), numeric features were cleaned, normalized, and log-transformed wherever necessary.



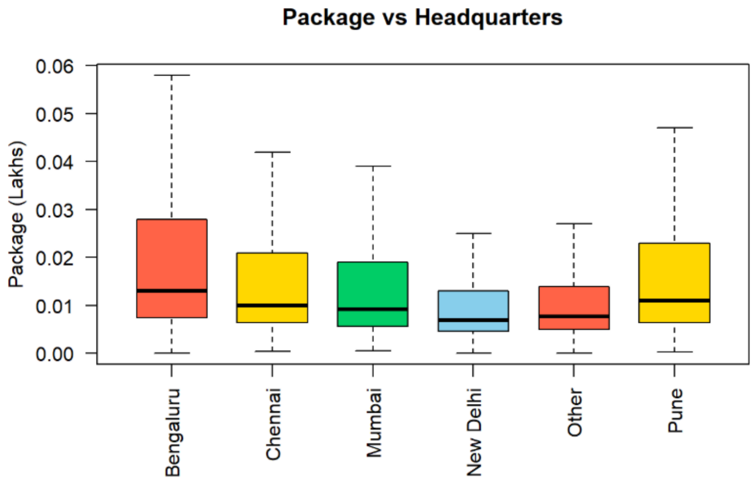
Key Cleaning Actions

- Removal of commas, currency symbols, suffixes (K, L)
- Conversion of categorical features into dummy variables
- Log transformation based on Box-Cox (λ ≈ 0)
- Consolidation of rare industry and headquarter categories under “Other”

---

3. Exploratory Insights

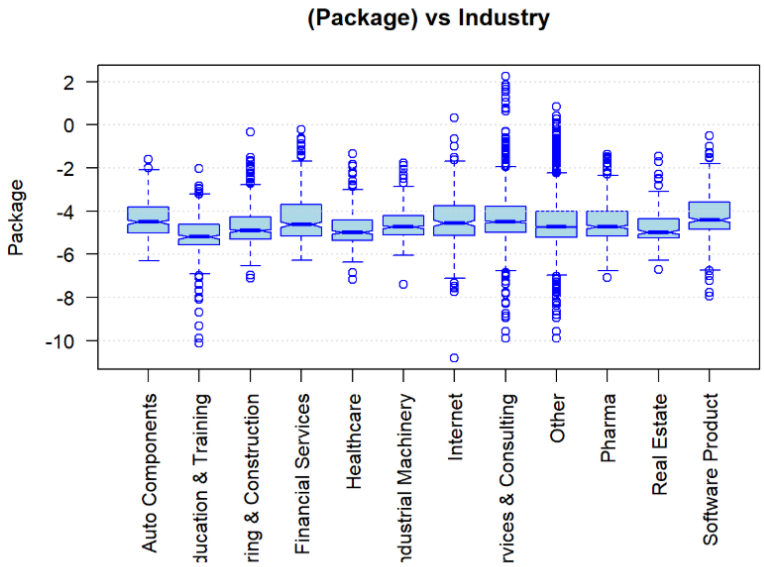
Salary by City



- **Bengaluru consistently leads** salary levels with wide upper spread.

- **Mumbai & Chennai** follow with moderate but stable pay scales.
- **New Delhi & Tier-2 clusters** show lower median ranges.

Salary by Industry



- **Software Product & Financial Services** are top salary contributors.
- **Education, Construction, Healthcare** are structurally lower paying.
- **Internet & Consulting** shows high dispersion reflecting start-up vs MNC bandwidths.

---

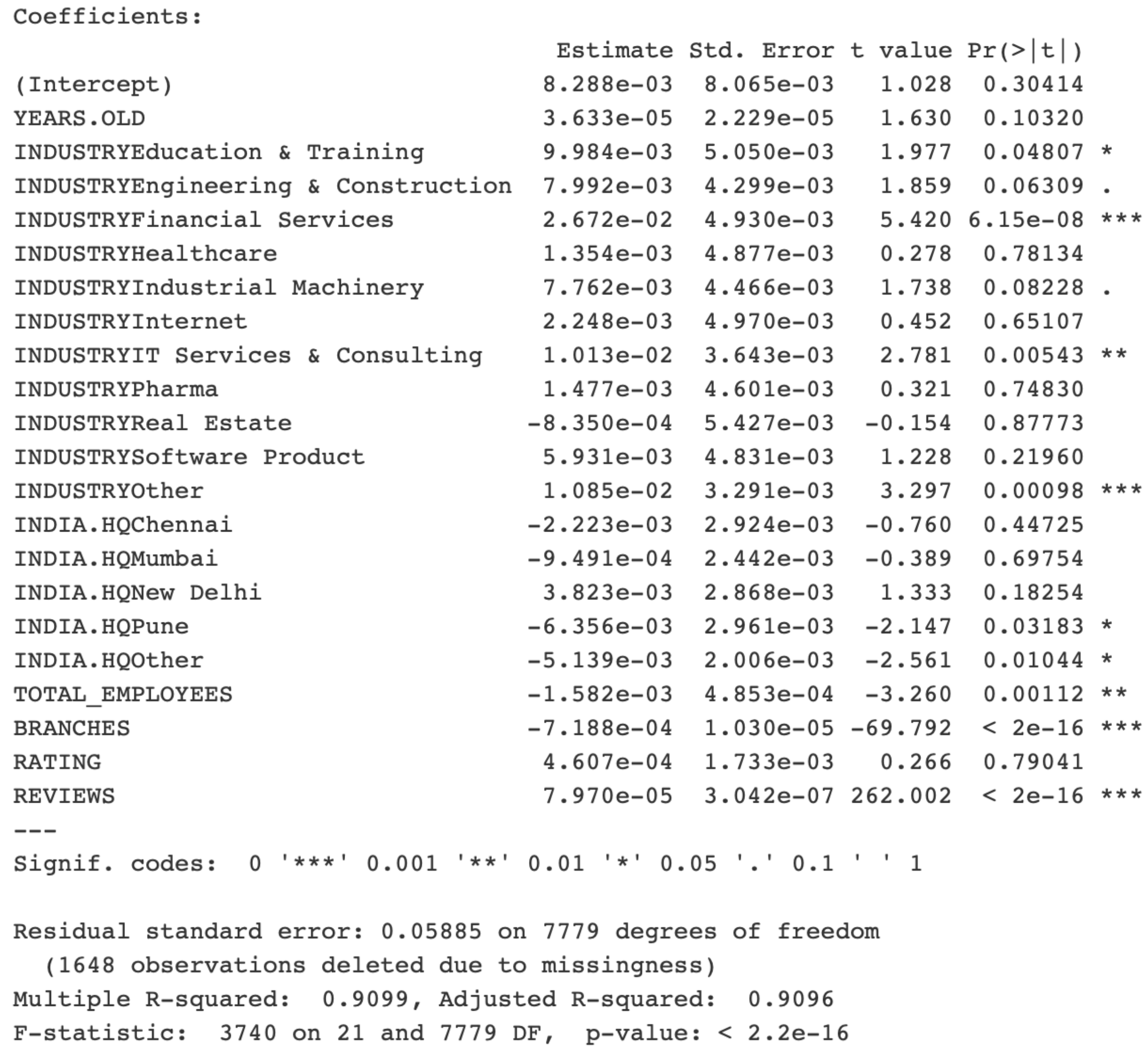
4. Hypothesis Testing Summary

Hypothesis	Result
Is the mean rating 3.5?	Rejected → Mean rating significantly **higher**
Bengaluru vs non-Bengaluru packages	Rejected → Bengaluru offers higher packages
Salary differs by industry	Rejected → Strong categorical effect
Industry independent of location	Rejected → Industries cluster by city
Company age affects salary	Failed to reject → negligible impact

5. Initial Regression Model (Before Diagnostics)

The initial model included **all predictors** without transformation:

$$\text{Package} = \beta_0 + \beta_1.\text{Industry} + \beta_2.\text{HQ} + \beta_3.\text{Years Old} + \beta_4.\text{Branches} + \beta_5.\text{Employees} + \beta_6.\text{Rating} + \beta_7.\text{Reviews}$$



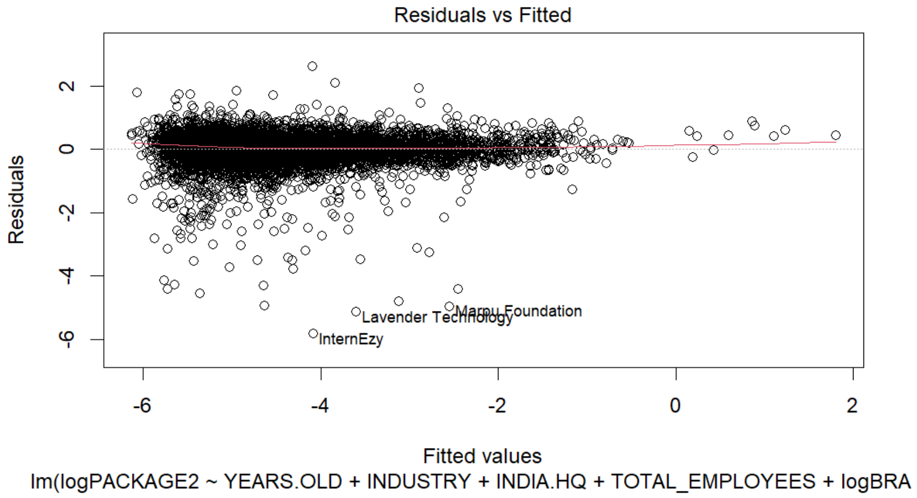
```
Coefficients:
                                      Estimate Std. Error t value Pr(>|t|)
(Intercept)                          8.288e-03  8.065e-03   1.028  0.30414
YEARS.OLD                            3.633e-05  2.229e-05   1.630  0.10320
INDUSTRYEducation & Training         9.984e-03  5.050e-03   1.977  0.04807 *
INDUSTRYEngineering & Construction   7.992e-03  4.299e-03   1.859  0.06309 .
INDUSTRYFinancial Services           2.672e-02  4.930e-03   5.420 6.15e-08 ***
INDUSTRYHealthcare                   1.354e-03  4.877e-03   0.278  0.78134
INDUSTRYIndustrial Machinery         7.762e-03  4.466e-03   1.738  0.08228 .
INDUSTRYInternet                     2.248e-03  4.970e-03   0.452  0.65107
INDUSTRYIT Services & Consulting     1.013e-02  3.643e-03   2.781  0.00543 **
INDUSTRYPharma                       1.477e-03  4.601e-03   0.321  0.74830
INDUSTRYReal Estate                 -8.350e-04  5.427e-03  -0.154  0.87773
INDUSTRYSoftware Product             5.931e-03  4.831e-03   1.228  0.21960
INDUSTRYOther                        1.085e-02  3.291e-03   3.297  0.00098 ***
INDIA.HQChennai                     -2.223e-03  2.924e-03  -0.760  0.44725
INDIA.HQMumbai                      -9.491e-04  2.442e-03  -0.389  0.69754
INDIA.HQNew Delhi                    3.823e-03  2.868e-03   1.333  0.18254
INDIA.HQPune                        -6.356e-03  2.961e-03  -2.147  0.03183 *
INDIA.HQOther                       -5.139e-03  2.006e-03  -2.561  0.01044 *
TOTAL_EMPLOYEES                     -1.582e-03  4.853e-04  -3.260  0.00112 **
BRANCHES                            -7.188e-04  1.030e-05 -69.792  < 2e-16 ***
RATING                               4.607e-04  1.733e-03   0.266  0.79041
REVIEWS                              7.970e-05  3.042e-07 262.002  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05885 on 7779 degrees of freedom
  (1648 observations deleted due to missingness)
Multiple R-squared:  0.9099, Adjusted R-squared:  0.9096
F-statistic:  3740 on 21 and 7779 DF,  p-value: < 2.2e-16
```

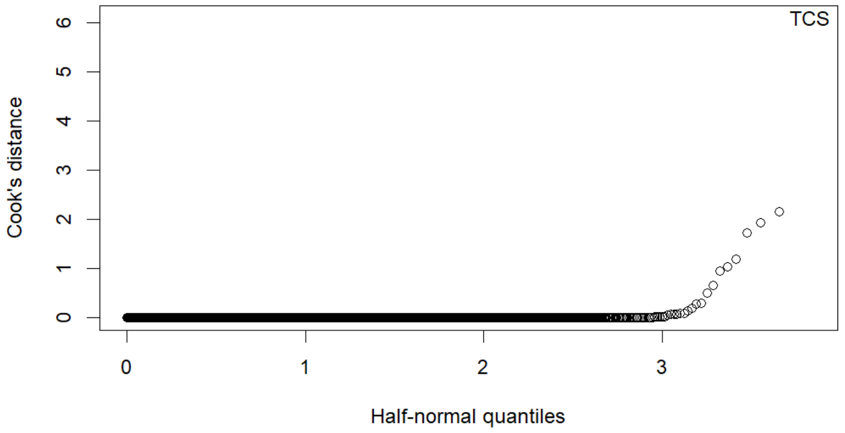
Interpretation

- Moderate explanatory strength (≈91%).
- Employee Count, Reviews, and Branches were significant; **Age of Company and Rating were not**. Overall model was significant.

6. Diagnostics, Outliers & Model Repair

Test	Result	Action
Residuals vs Fitted Values	Non-Constant Variance	Log Transform
Jackknife Residuals	Many Outliers	Examine influence not elimination
Cook’s Distance	Only TCS Influential	Remove non-influential extreme outliers
BIC Stepwise	Age and Employees removable	Dropped for parsimony





Outlier Handling Note

Numerous high-profile firms (Infosys, Deloitte, HDFC, Accenture, BYJU'S) appeared as outliers, but only TCS was truly influential.

After applying a log transformation, some points that appeared highly influential became less influential, suggesting that their apparent influence was mainly due to scale/skewness rather than true structural outlying-ness.

---

7. Stabilizing estimates

Multicollinearity can reduce the effective rank of the design matrix and thus making parameter estimates unstable. To ensure unique optimal solution, multicollinearity was evaluated using model diagnostics.

Method	Result	Interpretation
Checking Correlation matrix	Max Corr = 0.59 (Total Employees and Reviews). Only weak correlation exists among parameters.	No sign of Multicollinearity
Calculating Condition Number	K = 920, K>>30	May suggest strong multicollinearity
Calculating VIF	Average VIF = 1.9 Max. VIF = 6.1	No sign of Multicollinearity

Correlation matrix and low Variance Inflation Factor (VIF) suggest Multicollinearity is not problematic. The high condition number is due to design-matrix structure (dummy coding), not harmful collinearity in the regression coefficients. Therefore, no action is required to stabilize parameter estimates.

---

8. Model Selection

In practice, a less complex model is always preferred over a complex one on the condition that the predictive power of the model is not compromised. So, a hybrid search based on BIC (strong penalty for model size) was used to select a smaller model.

Lasso was not considered as model selection method as we want explicit inclusion or exclusion of predictors rather than shrinkage. In addition, Lasso introduces asymptotic Bias as opposed of criterion-based model selection.

Final Model:

$$\log(\text{Package}) = \beta_0 + \beta_1.\text{Industry} + \beta_2.\text{HQ} + \beta_3.\text{Branches} + \ \beta_4.\text{Rating} + \beta_5.\text{Reviews}$$

As a result, Years Old and Total Employees were dropped from the model.

Metric	Value
R-squared	0.8026
Adjusted R-squared	0.8021
F-statistic	1660
p-value	< 2.2e-16

Why Final Model Is Not Overfitted

Despite removal of non-influential outliers, R² and Adjusted R² remain nearly identical:

$$R^2 = 0.8026 \text{ vs Adj } R^2 = 0.8021$$

A large gap would imply overfitting; a tight convergence indicates:

- retained predictors add true explanatory value
- model complexity did not artificially inflate fit
- stability improved for prediction without distortion

---

9. Key Determinant Summary

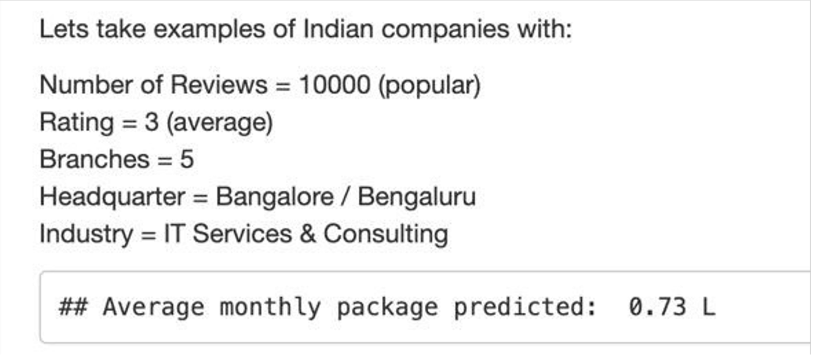
Driver	Outcome
Industry sector	Primary determinant of salary level
City (HQ location)	Bengaluru wage premium dominates
Reputation indicators (Rating, Reviews)	Strong predictors of compensation competitiveness
Branches	Expansion maturity correlates with higher salary
Age & Employee Count	Statistically negligible

Strategic Insight

Compensation is not governed by company maturity or workforce volume but by **market reputation, locational premium, and sector economics**.

---

10. Illustration of Predictions

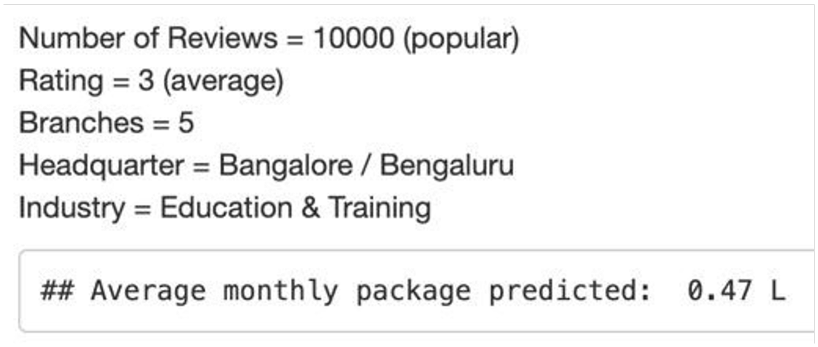


Lets take examples of Indian companies with:

Number of Reviews = 10000 (popular)
Rating = 3 (average)
Branches = 5
Headquarter = Bangalore / Bengaluru
Industry = IT Services & Consulting

```
## Average monthly package predicted:  0.73 L
```

- Bengaluru IT Services firm with strong brand score ≈ **₹73,000/month** predicted



Number of Reviews = 10000 (popular)
Rating = 3 (average)
Branches = 5
Headquarter = Bangalore / Bengaluru
Industry = Education & Training

```
## Average monthly package predicted:  0.47 L
```

- Education & Training firms under identical quantitative attributes → **significantly lower salary**

---

11. Conclusion

After comprehensive regression diagnostics, this study shows that:

- Salary determination in India is structurally linked to industry and headquarters city.
- Reputation (public rating and review intensity) prevails as a powerful compensation signal.
- After dealing with skewness, heteroscedasticity, and non-influential outliers, the final model explains 80% of the data. The model is stable and is also not overfitted.
- Transformation and variable removal did not distort the model; rather they enhanced predictive capability.

12. Contribution of Team Members

- **Data Cleaning:**
  Conducted collectively by **all team members**, ensuring preprocessing, handling missing values, fixing inconsistencies, and preparing the dataset for analysis.

- **Exploratory Data Analysis (EDA):**
  Performed by **Devang Chaturvedi**, including generation of descriptive statistics, visualizations, and identification of initial trends to guide further modeling.
- **Hypothesis Formation:**
  Developed by **Anmol Sai Ramchandran Ramanan**, who structured the theoretical assumptions and analytical framework for testing relationships within the dataset.
- **Model Diagnostics:**
  Completed by **Shreyas Pantangi**, involving variance stabilization, detection of outliers, leverage points, and Influential points.
- **Multicollinearity Assessment and Prediction Modeling:**
  Executed by **Kanishk Deshwal**, including correlation checks, condition number, VIF analysis, Model Selection and construction/evaluation of prediction models.

13. Appendix

- **Data:**
  - Indian companies complete data 2025.csv
  - Indian_companies_processed.csv
  - Indian_companies_transformed.csv
- **Code:**
  - DataCleaning.Rmd
  - DataCleaning.pdf
  - regressionPLOTS.Rmd
  - regressionPLOTS.pdf
  - Hypothesis_Testing.Rmd
  - Hypothesis_Testing.pdf
  - Diagnostics.Rmd
  - Diagnostics.pdf
  - Model Selection and Prediction.Rmd
  - Model Selection and Prediction.pdf