COMP 4442 PROJECT REPORT: MULTIPLE REGRESSION ANALYSIS

Presented by Michael Ghattas & Dawnena Key
Date: November 13, 2024

Project Overview

Objective:

Understand and predict salary determinants in data jobs using advanced statistical models (Stepwise, Lasso, Ridge regression).

Data:

Variables include job title, experience level, employment type, company size, and salary (USD).

Key Variables and Dataset Preparation

work_year <dbl>	job_title <chr>	job_category <chr>	salary_currency <chr>	salary <dbl>
2023	Data DevOps Engineer	Data Engineering	EUR	88000
2023	Data Architect	Data Architecture and Modeling	USD	186000
2023	Data Architect	Data Architecture and Modeling	USD	81800
2023	Data Scientist	Data Science and Research	USD	212000
2023	Data Scientist	Data Science and Research	USD	93300
2023	Data Scientist	Data Science and Research	USD	130000

Variables

Converted categorical factors (e.g., experience level, employment type)

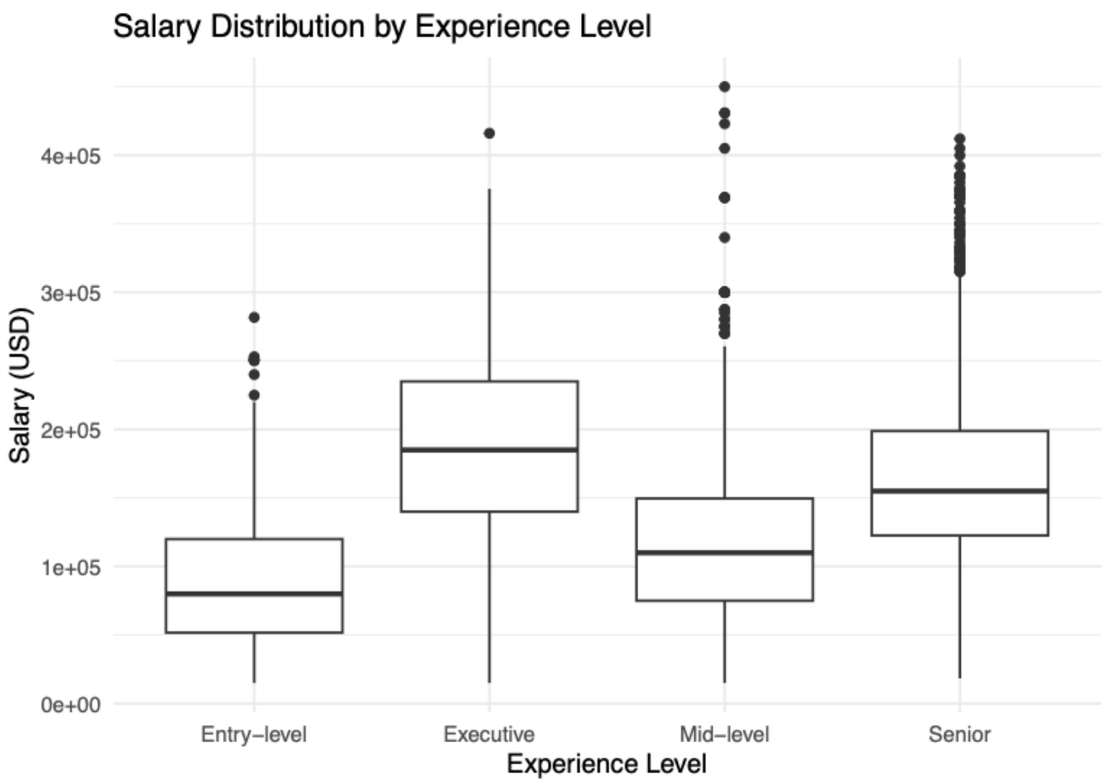
Data Profile

- Rows: 9355
- Columns: 12
  - Char-Variables: 9
  - Dbl-Variables: 3

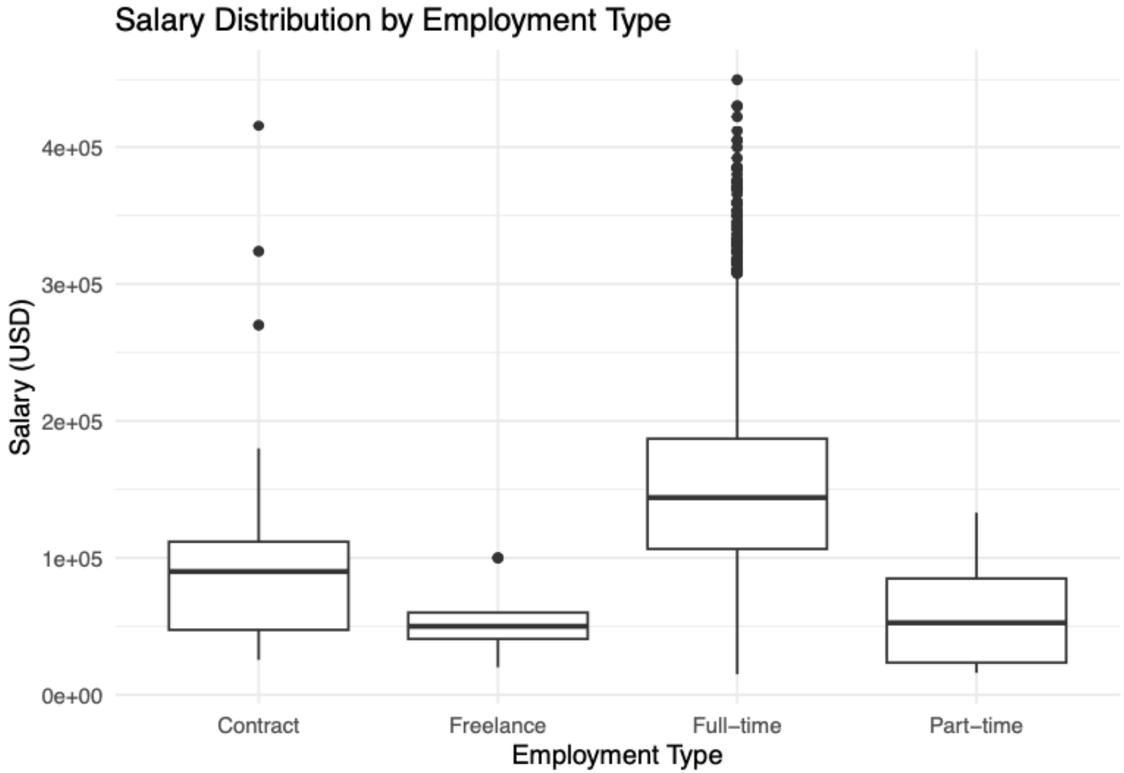
Data Cleaning

Converted categorical variables and removed outliers (IQR method) for accurate modeling.

Exploratory Data Analysis



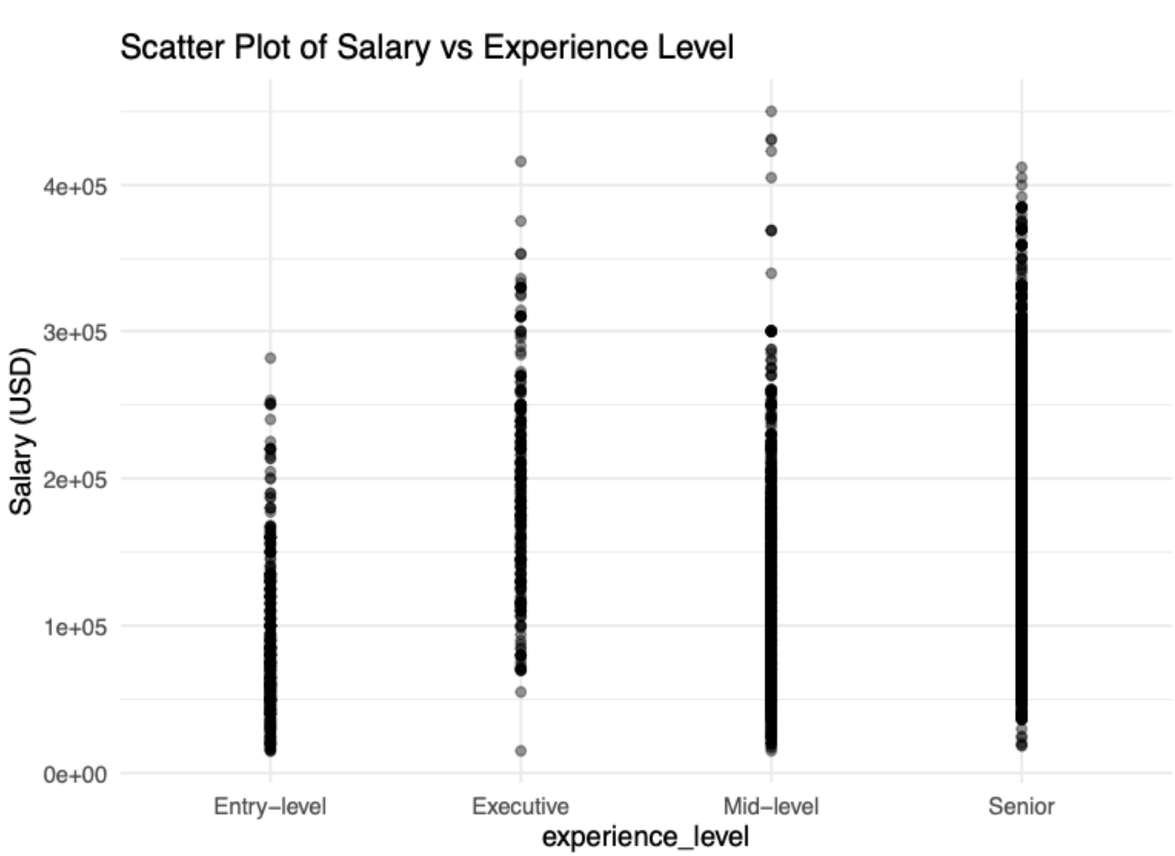
Interpretation: Salaries tend to increase with experience level, with Executive positions earning the most, indicating that experience is likely an important predictor of salary.



Interpretation: Full-time jobs are generally associated with higher salaries compared to part-time or freelance positions, suggesting that employment type affects salary.

Exploratory Data Analysis

Interpretation: The data suggests that medium sized companies (size M) may offer higher salaries compared to large and small companies.



Interpretation:
- **The boxplot provides a visual summary of the salary distribution.**
- **The majority of salaries are concentrated below 200,000 USD, with a median salary around 100,000 USD.**
- **A significant number of outliers are present in the upper salary range, extending well beyond 300,000 USD.**

Modeling Techniques

Overview of the different modeling approaches:

- **Stepwise Regression:** AIC-based selection of key predictors
- **Lasso Regression:** Regularization to reduce overfitting, select critical predictors.
- **Ridge Regression:** Minimizes multicollinearity, keeps all predictors.

Model Comparison and Selection:

State the purpose of comparing models to find the best approach

01 Lasso Regression

selected for its balance of interpretability and predictive accuracy. Emphasized roles like Machine Learning, with an Adjusted R^2 of 0.2572.

02 Step-wise Regression

Highlighted experience level and job category as significant predictors.

03 Ridge Regression

Showed similar predictors but with lower interpretability.

Model Comparison & Selection Summary

Model <chr>	AIC <dbl>	BIC <dbl>	R_squared <dbl>	Adjusted_R_squared <dbl>
Stepwise Regression	225239	225375	0.259	0.257
Lasso Regression	199137	199266	0.259	0.257
Ridge Regression	199158	199286	0.257	0.256

3 rows

Model Diagnostics Context

Cook's Distance: Some data points exceed the typical 4/n threshold, indicating high influence on the model. These points should be carefully reviewed through sensitivity testing.

Residuals Histogram: Residuals approximate normality, with slight skewness and potential outliers, suggesting the model captures most data patterns but might benefit from further adjustments.

Q-Q Plot: Residuals follow the normality line with minor deviations at the tails, indicating slight outliers that don't significantly compromise the model but hint at possible improvements through transformations.

Kolmogorov-Smirnov Test: Significant p-value (8.251×10^{-13}) suggests residuals deviate from perfect normality; however, given the histogram and Q-Q plot, this may not critically impact the model.

Breusch-Pagan Test: Significant p-value ($< 2.2 \times 10^{-16}$) indicates heteroscedasticity, suggesting variable residual variance. Consider robust errors or transformations to improve reliability.

VIF (Multicollinearity): All VIF values below 1.2, indicating low multicollinearity, which supports the stability and interpretability of the model's coefficients.

Conclusion: Diagnostic checks confirm model validity with minor deviations in residual normality and heteroscedasticity. Further transformations, like log transformation, could enhance predictive accuracy by addressing these small inconsistencies.

Final Model (Log-Transformed)

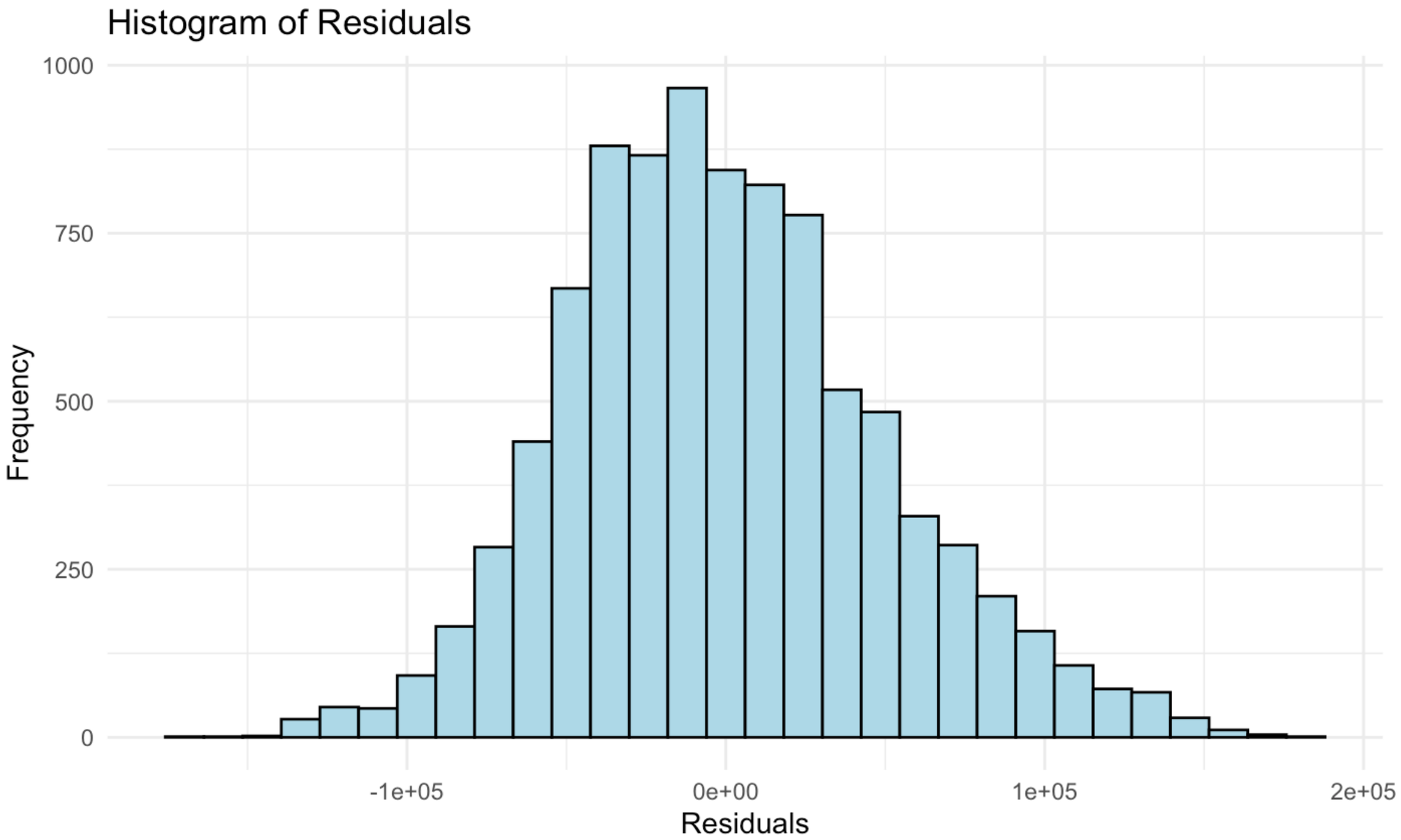
term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
experience_levelExecutive	77951	4083	19.09	1.07e-79
`experience_levelMid-level`	19557	2449	7.98	1.58e-15
experience_levelSenior	55301	2315	23.89	1.90e-122
employment_typeFreelance	-41029	20066	-2.04	4.09e-02
company_sizeS	-32743	4706	-6.96	3.69e-12
`job_categoryData Analysis`	-18393	2803	-6.56	5.64e-11
`job_categoryData Architecture and Modeling`	15447	4189	3.69	2.28e-04
`job_categoryData Engineering`	11766	2814	4.18	2.92e-05
`job_categoryData Management and Strategy`	-16058	6556	-2.45	1.43e-02
`job_categoryData Quality and Operations`	-28064	6844	-4.10	4.15e-05

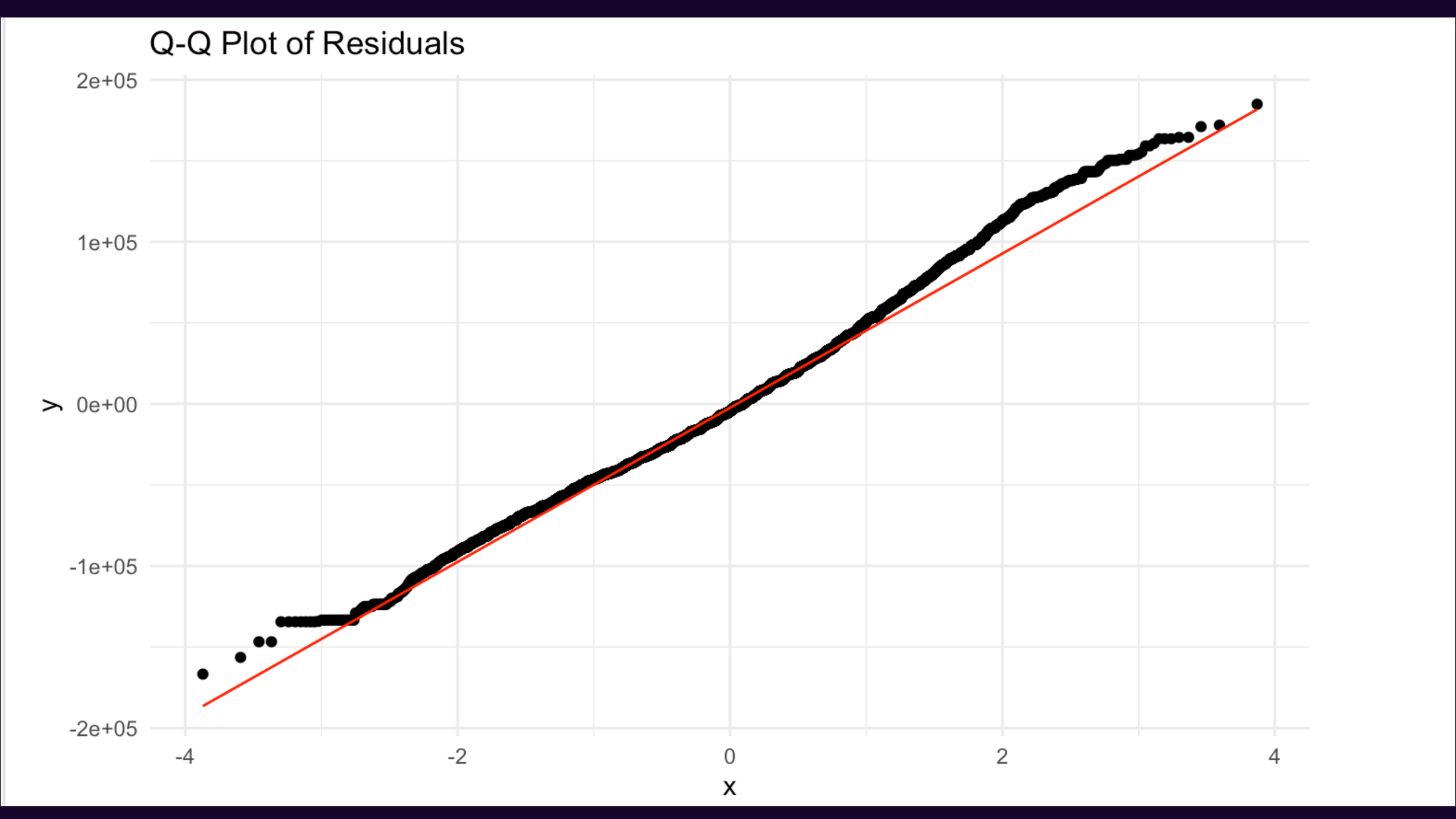
R^2 of Lasso Model: Approximately 0.302, indicating 30.2% of the variability in log-transformed salary is explained.

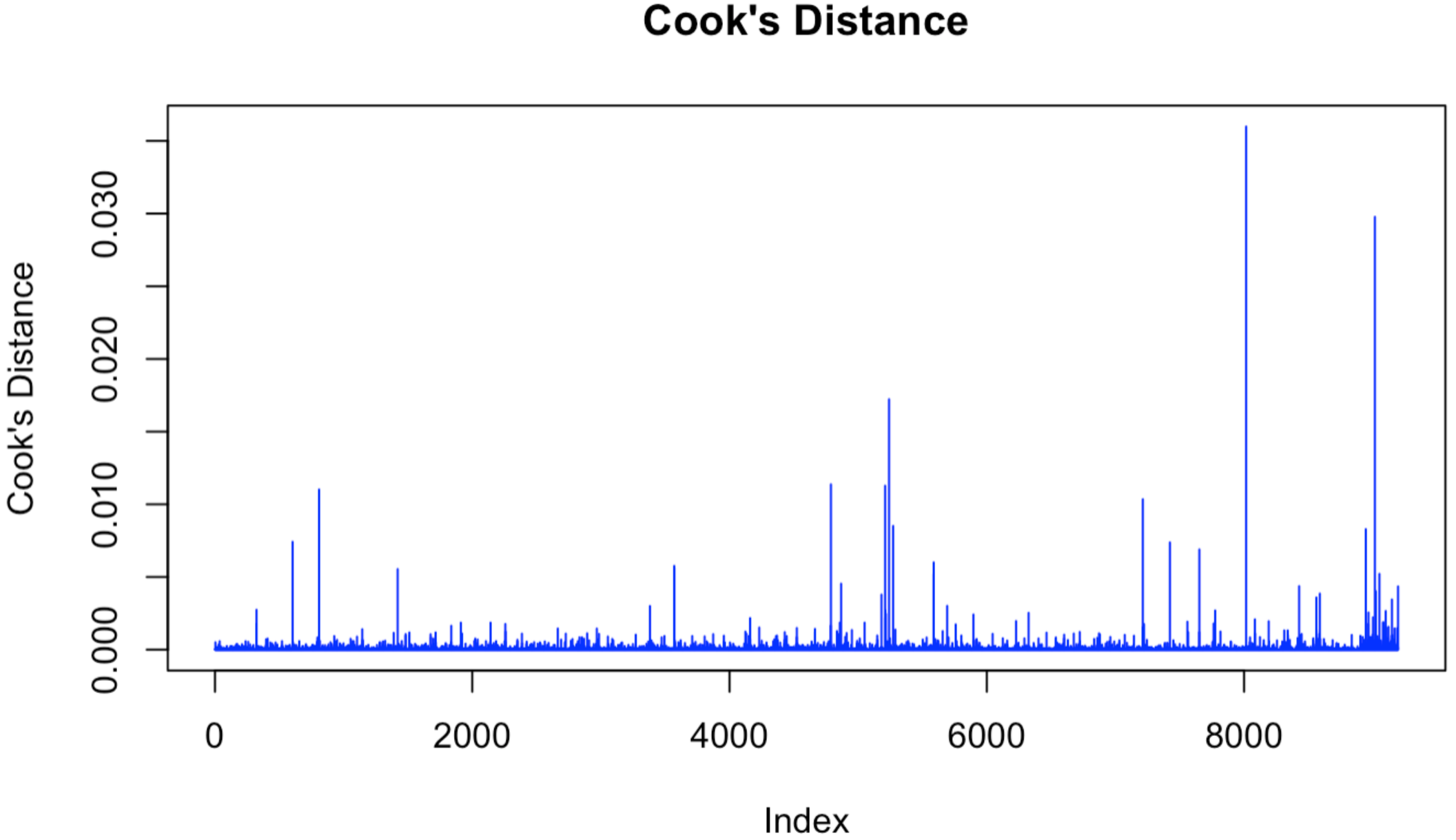
Log Transformation: Helps reduce the impact of outliers, skewness, and heteroscedasticity in salary data.

Predictor Selection: Lasso model effectively selects key predictors, shrinking less important ones for simplicity.

Multicollinearity: Addressed by the Lasso model's regularization, enhancing interpretability and stability.


Histogram of Residuals
1000
750
500
250
0
Frequency
-1e+05
0e+00
1e+05
2e+05
Residuals


Q-Q Plot of Residuals
2e+05
1e+05
0e+00
-1e+05
-2e+05
y
-4
-2
0
2
4
x


Cook's Distance
Cook's Distance
0.000
0.010
0.020
0.030
0
2000
4000
6000
8000
Index


Significant Predictors of Salary
experience_levelExecutive
experience_levelSenior
`job_categoryMachine Learning and AI`
`work_settingIn-person`
work_settingRemote
`job_categoryData Science and Research`
`experience_levelMid-level`
`job_categoryData Architecture and Modeling`
`job_categoryData Engineering`
`job_categoryLeadership and Management`
`job_categoryData Management and Strategy`
`job_categoryData Analysis`
`job_categoryData Quality and Operations`
company_sizeS
employment_typeFreelance
Predictors
-40000
0
40000
80000
Coefficient Estimate

Future Considerations

- **Additional Predictors:** Including economic indicators, geographic details, and skill certifications could enhance the model's accuracy but were omitted due to dataset limitations.
- **Temporal Analysis:** Using time-series or cohort segmentation could provide insights into salary trends over time, accounting for industry cycles and economic shifts.
- **Interaction** Effects: Interaction terms for experience level, employment type, and job category were tested but excluded, as they added complexity without improving predictive accuracy.
- **Outlier Handling:** Outliers, especially high salaries, were removed using the IQR method to prevent skewed results and improve model stability, focusing the model on representative data trends.
- **Cross-Validation:** While cross-validation was used for tuning, additional methods like k-fold or bootstrapping could further enhance model reliability.
- **Further Analysis:** Expanding with non-linear models (e.g., decision trees, random forests) and additional predictors like industry and location could capture complex relationships and improve predictive power.

Q & A

Thank You!