COMP 4442: Project Report - Multiple Regression Analysis

Michael Ghattas & Dawnena Key

2024-11-10

Introduction

The rapid growth of data-related jobs across industries has led to an evolving landscape in job roles, compensation, and skill requirements. This project aims to explore and quantify the factors that influence salaries within the data job market. Using the "Jobs in Data" dataset, which includes variables such as job title, experience level, employment type, company characteristics, and salary information, we seek to understand which factors most significantly impact salary outcomes.

The primary objective of this analysis is to develop predictive models for salary based on key input variables, applying advanced statistical techniques such as Stepwise Regression, Lasso, and Ridge Regression. These models are evaluated to ensure both statistical rigor and practical interpretability, with particular attention to addressing multicollinearity and optimizing model fit. The project also includes exploratory data analysis to visualize initial trends and a final model refinement stage to identify the best predictors.

Load Dataset

```r
# Load the dataset
data <- read_csv("jobs_in_data.csv")
```

```
## Rows: 9355 Columns: 12
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (9): job_title, job_category, salary_currency, employee_residence, exper...
## dbl (3): work_year, salary, salary_in_usd
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```r
# Display the first few rows
spec(data)
```

```
## cols(
##   work_year = col_double(),
##   job_title = col_character(),
##   job_category = col_character(),
##   salary_currency = col_character(),
##   salary = col_double(),
##   salary_in_usd = col_double(),
```

```
##   employee_residence = col_character(),
##   experience_level = col_character(),
##   employment_type = col_character(),
##   work_setting = col_character(),
##   company_location = col_character(),
##   company_size = col_character()
## )
```

```r
head(data)
```

```
## # A tibble: 6 x 12
##   work_year job_title          job_category salary_currency salary salary_in_usd
##       <dbl> <chr>              <chr>        <chr>            <dbl>         <dbl>
## 1      2023 Data DevOps Engin~ Data Engine~ EUR              88000         95012
## 2      2023 Data Architect     Data Archit~ USD             186000        186000
## 3      2023 Data Architect     Data Archit~ USD              81800         81800
## 4      2023 Data Scientist     Data Scienc~ USD             212000        212000
## 5      2023 Data Scientist     Data Scienc~ USD              93300         93300
## 6      2023 Data Scientist     Data Scienc~ USD             130000        130000
## # i 6 more variables: employee_residence <chr>, experience_level <chr>,
## #   employment_type <chr>, work_setting <chr>, company_location <chr>,
## #   company_size <chr>
```

Varibles Transformation

```r
# Convert selected variables to factors
data <- data %>%
  mutate(
    experience_level = as.factor(experience_level),
    employment_type = as.factor(employment_type),
    work_setting = as.factor(work_setting),
    company_size = as.factor(company_size),
    job_category = as.factor(job_category),
    company_location = as.factor(company_location),
    employee_residence = as.factor(employee_residence)
  )

# Display the first few rows
head(data)
```

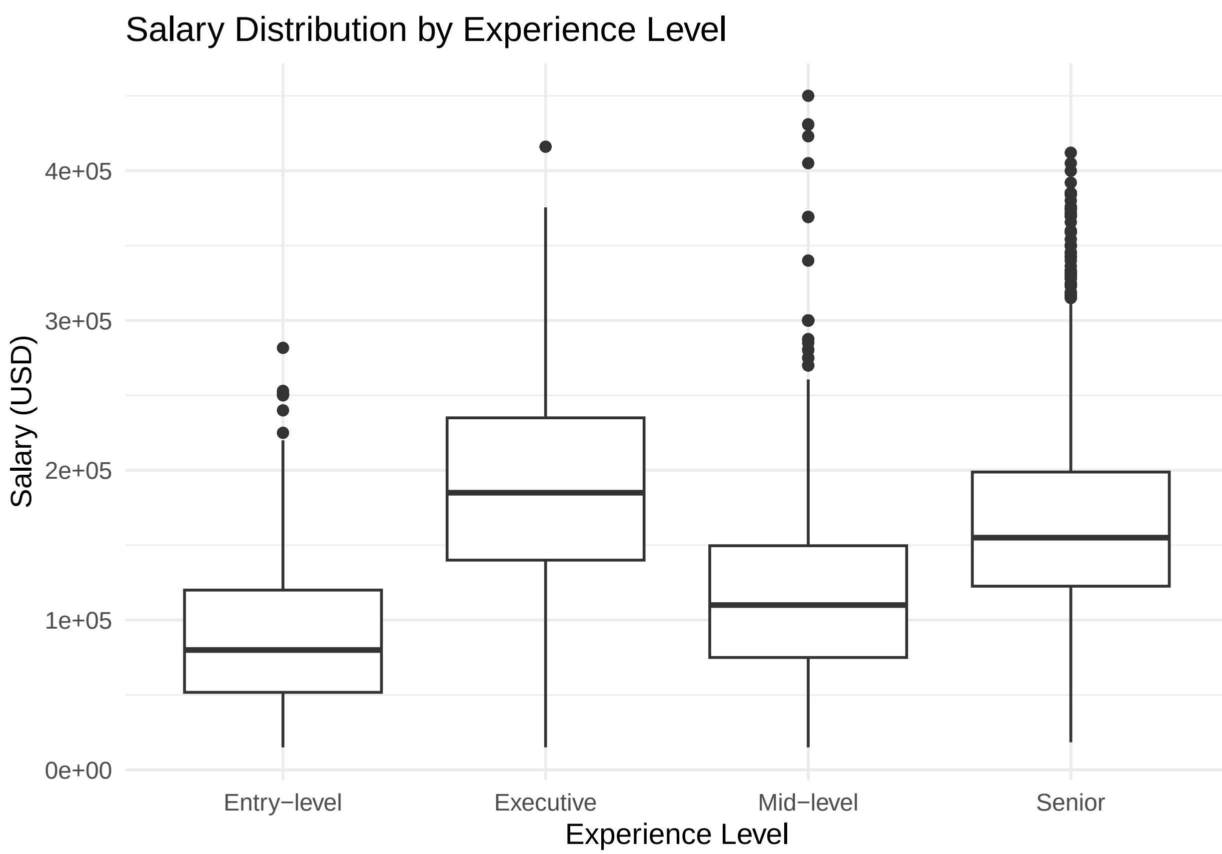
```
## # A tibble: 6 x 12
##   work_year job_title          job_category salary_currency salary salary_in_usd
##       <dbl> <chr>              <fct>        <chr>            <dbl>         <dbl>
## 1      2023 Data DevOps Engin~ Data Engine~ EUR              88000         95012
## 2      2023 Data Architect     Data Archit~ USD             186000        186000
## 3      2023 Data Architect     Data Archit~ USD              81800         81800
## 4      2023 Data Scientist     Data Scienc~ USD             212000        212000
## 5      2023 Data Scientist     Data Scienc~ USD              93300         93300
## 6      2023 Data Scientist     Data Scienc~ USD             130000        130000
## # i 6 more variables: employee_residence <fct>, experience_level <fct>,
## #   employment_type <fct>, work_setting <fct>, company_location <fct>,
## #   company_size <fct>
```

Exploratory Data Analysis (EDA)

Boxplot: Salary by Experience Level

This boxplot will show the variation in salary for different experience levels (e.g., Entry-level, Mid-level, Senior, Executive). Typically, you expect salary to increase with experience.

```
# Boxplot of Salary by Experience Level
ggplot(data, aes(x = experience_level, y = salary_in_usd)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Salary Distribution by Experience Level", x = "Experience Level", y = "Salary (USD)")
```



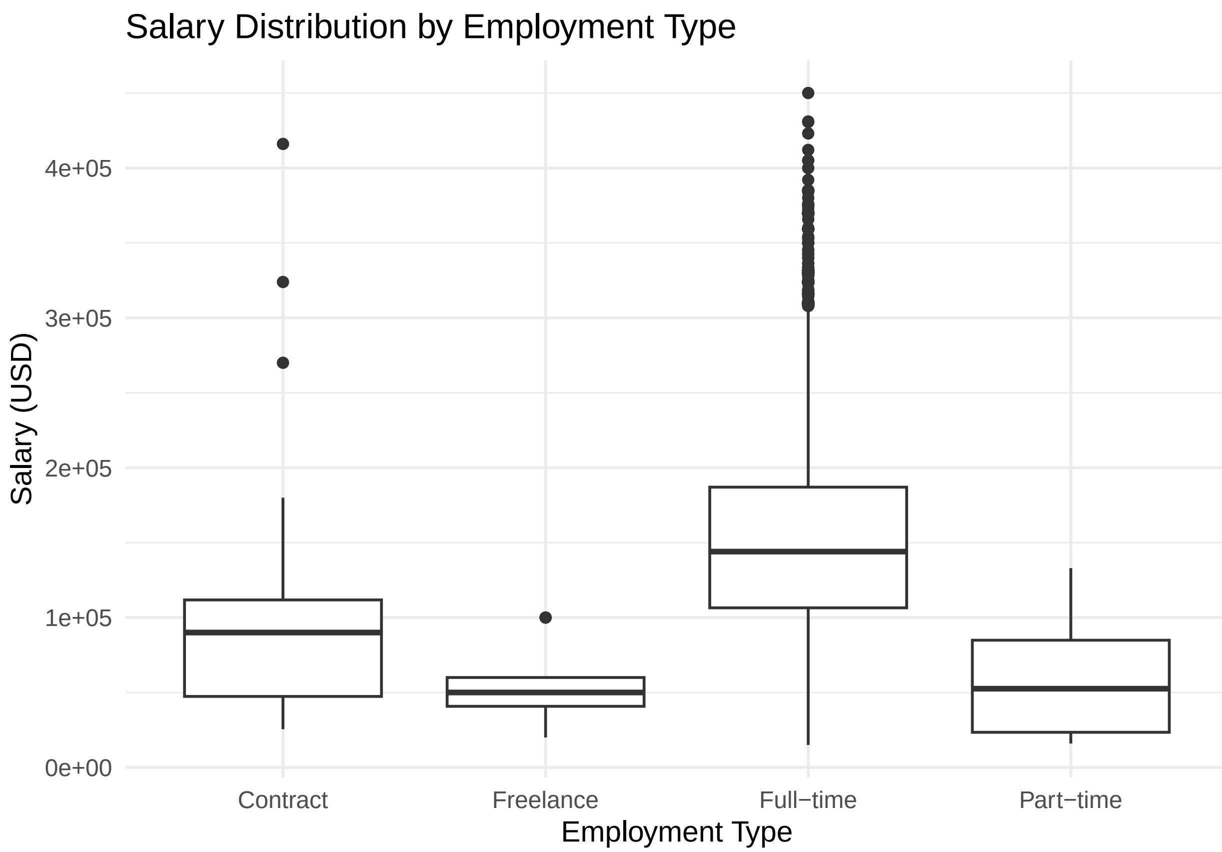
Interpretation: Salaries tend to increase with experience level, with Executive positions earning the most, indicating that experience is likely an important predictor of salary.

Boxplot: Salary by Employment Type

This boxplot explores how salary varies by employment type (e.g., Full-time, Part-time, Freelance). Full-time roles typically offer higher salaries.

```
# Boxplot of Salary by Employment Type
ggplot(data, aes(x = employment_type, y = salary_in_usd)) +
  geom_boxplot() +
```

```r
  theme_minimal() +
  labs(title = "Salary Distribution by Employment Type", x = "Employment Type", y = "Salary (USD)")
```



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

Scatter Plot: Salary vs Company Size

This scatter plot will show how salary relates to the size of the company. Larger companies might offer higher salaries due to better resources and budgets.

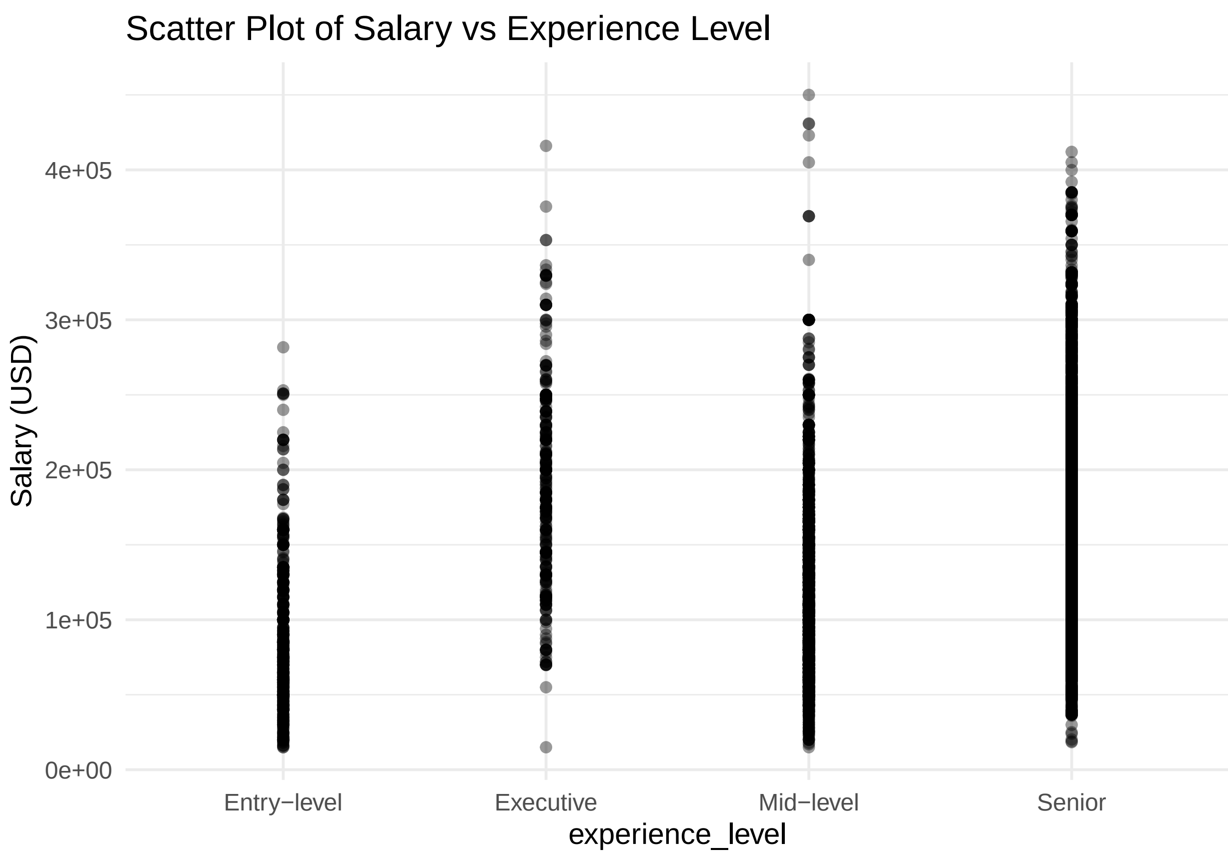
```r
# Scatter plot of Salary vs Company Size
ggplot(data, aes(x = company_size, y = salary_in_usd)) +
  geom_jitter(alpha = 0.4) +
  theme_minimal() +
  labs(title = "Salary vs Company Size", x = "Company Size", y = "Salary (USD)")
```

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

Check Outliers

```
# Boxplot of salary to visualize potential outliers
ggplot(data, aes(y = salary_in_usd)) +
  geom_boxplot() +
  theme_minimal() +
  labs(title = "Boxplot of Salary (USD)", y = "Salary (USD)")
```


```r
# Scatter plot to detect outliers in salary and other features
ggplot(data, aes(x = experience_level, y = salary_in_usd)) +
  geom_point(alpha = 0.4) +
  theme_minimal() +
  labs(title = "Scatter Plot of Salary vs Experience Level", y = "Salary (USD)")
```



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.

```
# Calculate the IQR for salary_in_usd
Q1 <- quantile(data$salary_in_usd, 0.25)
Q3 <- quantile(data$salary_in_usd, 0.75)
IQR <- Q3 - Q1

# Define the outlier bounds
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR

# Identify outliers
outliers <- data %>%
  filter(salary_in_usd < lower_bound | salary_in_usd > upper_bound)

# Remove outliers from the dataset
data_cleaned <- data %>%
  filter(salary_in_usd >= lower_bound & salary_in_usd <= upper_bound)

# Check the size of the cleaned dataset
nrow(data_cleaned)
```

```
## [1] 9197
```

Interpretation:

- Removing outliers ensures that the model focuses on the majority of the data, reducing the influence of extreme values that might skew the results.
- The decision to remove these outliers is based on the IQR method, which assumes that values far beyond the 1.5× IQR are extreme.
- This cleaning process can improve model performance by reducing heteroscedasticity and making assumptions about residual normality more valid.
- This it might also remove some meaningful variability, especially in salary-related analyses where high salaries could be legitimate and not necessarily errors or noise.

Modeling Exploration

The initial model assess the relationship between salary and input to identify key predictors of salary.

Stepwise Regression Model

```r
# Fit the initial full model for stepwise selection
initial_model <- lm(salary_in_usd ~ experience_level + employment_type + company_size + job_category,
                    data = data_cleaned)

# Perform stepwise selection
stepwise_model <- stepAIC(initial_model, direction = "both")
```

```
## Start:  AIC=199137.1
## salary_in_usd ~ experience_level + employment_type + company_size +
##     job_category
##
##                    Df  Sum of Sq        RSS    AIC
## <none>                           2.3199e+13 199137
## - employment_type   3 6.2230e+10 2.3261e+13 199156
## - company_size      2 3.0646e+11 2.3505e+13 199254
## - job_category      9 3.1746e+12 2.6373e+13 200299
## - experience_level  3 3.2004e+12 2.6399e+13 200320
```

```r
# Display the summary and of the stepwise model
summary(stepwise_model)
```

```
##
## Call:
## lm(formula = salary_in_usd ~ experience_level + employment_type +
##     company_size + job_category, data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -164859  -34420   -4766   29981  174282
##
```

```
## Coefficients:
##                                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                  58379      12812   4.557 5.26e-06
## experience_levelExecutive                    80548       3926  20.519  < 2e-16
## experience_levelMid-level                    22270       2588   8.606  < 2e-16
## experience_levelSenior                       58713       2422  24.242  < 2e-16
## employment_typeFreelance                    -46638      19471  -2.395  0.01663
## employment_typeFull-time                     19443      12323   1.578  0.11465
## employment_typePart-time                     -6864      17872  -0.384  0.70095
## company_sizeM                                 8485       1986   4.273 1.95e-05
## company_sizeS                               -34972       4502  -7.768 8.82e-15
## job_categoryCloud and Database               11677      22663   0.515  0.60641
## job_categoryData Analysis                   -20468       3170  -6.458 1.12e-10
## job_categoryData Architecture and Modeling   12862       4260   3.020  0.00254
## job_categoryData Engineering                  9999       3063   3.265  0.00110
## job_categoryData Management and Strategy    -16723       7067  -2.366  0.01798
## job_categoryData Quality and Operations     -30182       7368  -4.096 4.23e-05
## job_categoryData Science and Research        25398       3004   8.456  < 2e-16
## job_categoryLeadership and Management         5579       3671   1.520  0.12855
## job_categoryMachine Learning and AI          38746       3176  12.198  < 2e-16
##
## (Intercept)                                ***
## experience_levelExecutive                  ***
## experience_levelMid-level                  ***
## experience_levelSenior                     ***
## employment_typeFreelance                   *
## employment_typeFull-time
## employment_typePart-time
## company_sizeM                              ***
## company_sizeS                              ***
## job_categoryCloud and Database
## job_categoryData Analysis                  ***
## job_categoryData Architecture and Modeling **
## job_categoryData Engineering               **
## job_categoryData Management and Strategy   *
## job_categoryData Quality and Operations    ***
## job_categoryData Science and Research      ***
## job_categoryLeadership and Management
## job_categoryMachine Learning and AI        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50270 on 9179 degrees of freedom
## Multiple R-squared:  0.2586, Adjusted R-squared:  0.2572
## F-statistic: 188.3 on 17 and 9179 DF,  p-value: < 2.2e-16
```

```
# Calculate VIFs for multicollinearity assessment
vif(stepwise_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## experience_level 1.111792  3        1.017819
## employment_type  1.061636  3        1.010018
## company_size     1.117545  2        1.028173
## job_category     1.091243  9        1.004863
```

Stepwise Regression Model Summary

Model Coefficients

- Experience Level: Executive - A large positive impact on salary (estimate = 80,548), indicating that executives earn significantly more than the baseline group (likely entry-level). Senior and Mid-level - Both are also associated with positive impacts on salary, with senior-level positions (estimate = 58,713) having a higher effect than mid-level (estimate = 22,270). This aligns with expectations, as higher experience levels typically command higher salaries. All experience levels (Executive, Mid-level, and Senior) are highly significant predictors ($p < 0.001$).
- Employment Type: Freelance - Has a significant negative impact on salary (estimate = -46,638), suggesting that freelance roles tend to offer lower pay compared to the baseline employment type. Full-time and Part-time - These are not statistically significant ($p > 0.05$), meaning they do not show a distinct impact on salary in this model.
- Company Size: Small (company_sizeS) - A significant negative impact on salary (estimate = -34,972), indicating that smaller companies tend to offer lower salaries, possibly due to fewer resources or smaller budgets. Medium (company_sizeM) - Shows a positive effect on salary (estimate = 8,485) and is statistically significant, suggesting a slight salary premium over small companies.
- Job Category: Several job categories show significant effects on salary. Positive Impact - Data Science and Research (estimate = 25,398) and Machine Learning and AI (estimate = 38,746) roles have strong positive associations with salary, which aligns with their high demand and specialized skill requirements. Data Architecture and Modeling and Data Engineering also have positive effects on salary, though with smaller magnitudes. Negative Impact - Data Analysis (estimate = -20,468) and Data Quality and Operations (estimate = -30,182) show negative impacts, indicating lower salaries within these categories. Data Management and Strategy also has a negative effect on salary (estimate = -16,723). These results suggest that technical and specialized roles (e.g., Machine Learning, Data Science) tend to offer higher pay than more support-oriented or operational roles (e.g., Data Analysis).

Model Fit

- R-squared: The model explains about 25.9% of the variability in salary, which is typical for real-world salary models where many unobserved factors can influence pay.
- Adjusted R-squared: At 0.2572, this is close to the R-squared value, indicating that the model's explanatory power remains stable when adjusting for the number of predictors.

Model Diagnostics

- Residual Standard Error: 50,270, indicating the average deviation of observed salaries from those predicted by the model.
- F-statistic: The overall model is highly significant (p-value $< 2.2e-16$), suggesting that at least some predictors have meaningful associations with salary.

Variance Inflation Factor (VIF) Analysis

- All VIF values are low, with GVIF values adjusted for degrees of freedom all below 1.1, indicating that multicollinearity is not a concern in this model. This means the predictors do not have excessive linear relationships, which supports the stability and reliability of the model coefficients.

Conclusion

This stepwise regression model identifies key variables associated with salaries in data-related roles. Experience level, job category, and company size have strong associations with salary. Higher experience levels, technical job categories (e.g., Machine Learning, Data Science), and larger company sizes are correlated with higher salaries. The model suggests that freelance roles and employment in smaller companies are associated with lower salaries. Multicollinearity is low, indicating that each predictor contributes unique information to the model without excessive overlap. These findings align with industry trends, where experience, specialized skills, and company size often command higher pay. Further refinement and regularization techniques (like Lasso and Ridge regression) can be explored to handle potential overfitting and improve predictive accuracy.

Log-Transformed Model

```r
# Apply log transformation to salary
data_cleaned$log_salary_in_usd <- log(data_cleaned$salary_in_usd)

# Fit the log-transformed model
log_model <- lm(log_salary_in_usd ~ experience_level + employment_type + company_size + job_category,
                data = data_cleaned)

# Display the summary of the log-transformed model
summary(log_model)
```

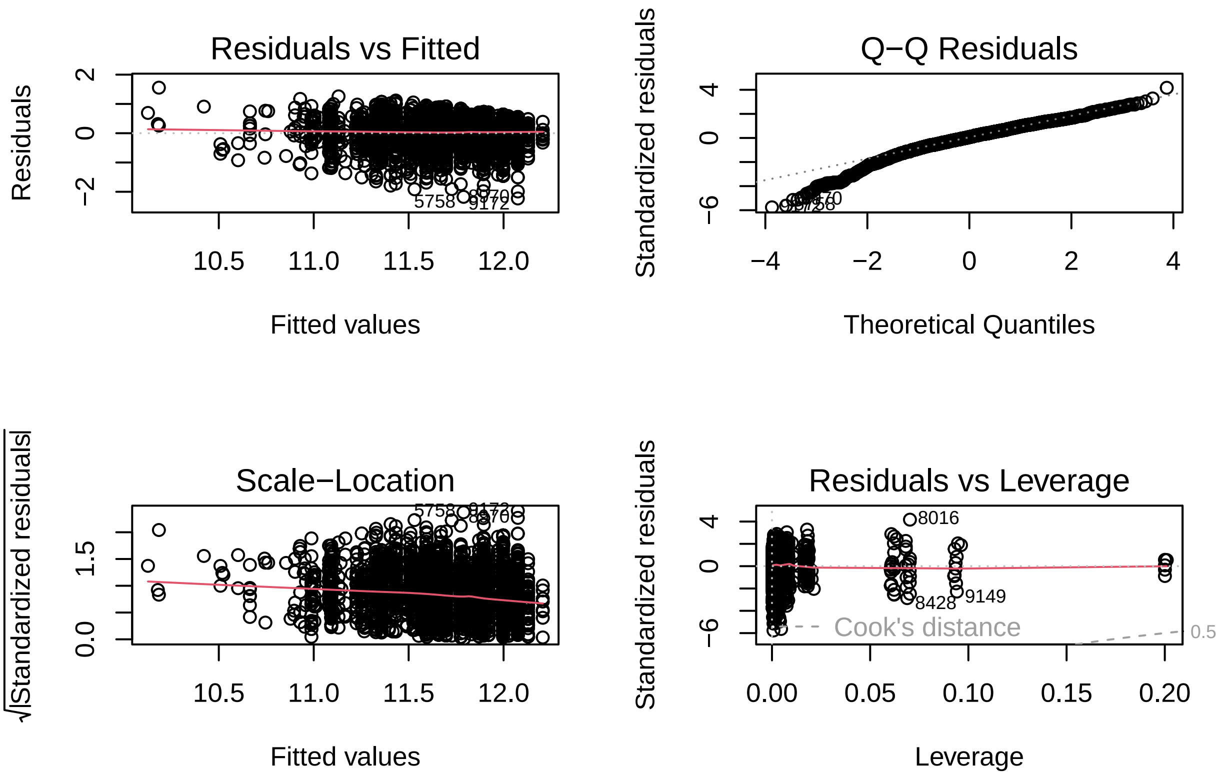
```
##
## Call:
## lm(formula = log_salary_in_usd ~ experience_level + employment_type +
##     company_size + job_category, data = data_cleaned)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.22792 -0.21022  0.02828  0.25413  1.55489
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                10.91904    0.09859 110.753  < 2e-16
## experience_levelExecutive                   0.69688    0.03021  23.069  < 2e-16
## experience_levelMid-level                   0.26496    0.01991  13.305  < 2e-16
## experience_levelSenior                      0.56555    0.01864  30.345  < 2e-16
## employment_typeFreelance                   -0.50722    0.14984  -3.385 0.000714
## employment_typeFull-time                    0.24247    0.09483   2.557 0.010575
## employment_typePart-time                   -0.23684    0.13753  -1.722 0.085085
## company_sizeM                               0.10460    0.01528   6.844 8.17e-12
## company_sizeS                              -0.32417    0.03464  -9.357  < 2e-16
## job_categoryCloud and Database              0.12593    0.17440   0.722 0.470271
## job_categoryData Analysis                  -0.17363    0.02439  -7.119 1.17e-12
## job_categoryData Architecture and Modeling  0.08914    0.03278   2.719 0.006554
## job_categoryData Engineering                0.06420    0.02357   2.724 0.006461
## job_categoryData Management and Strategy   -0.13529    0.05438  -2.488 0.012876
## job_categoryData Quality and Operations    -0.30714    0.05670  -5.417 6.22e-08
## job_categoryData Science and Research       0.16617    0.02311   7.189 7.03e-13
## job_categoryLeadership and Management       0.03912    0.02825   1.385 0.166064
## job_categoryMachine Learning and AI         0.24355    0.02444   9.963  < 2e-16
##
## (Intercept)                                ***
## experience_levelExecutive                  ***
## experience_levelMid-level                  ***
```

```
## experience_levelSenior                        ***
## employment_typeFreelance                      ***
## employment_typeFull-time                      *
## employment_typePart-time                      .
## company_sizeM                                 ***
## company_sizeS                                 ***
## job_categoryCloud and Database
## job_categoryData Analysis                     ***
## job_categoryData Architecture and Modeling **
## job_categoryData Engineering                  **
## job_categoryData Management and Strategy   *
## job_categoryData Quality and Operations    ***
## job_categoryData Science and Research      ***
## job_categoryLeadership and Management
## job_categoryMachine Learning and AI        ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3869 on 9179 degrees of freedom
## Multiple R-squared:   0.29,  Adjusted R-squared:  0.2886
## F-statistic: 220.5 on 17 and 9179 DF,  p-value: < 2.2e-16
```

```r
# Calculate VIFs for multicollinearity assessment
vif(log_model)
```

```
##                      GVIF Df GVIF^(1/(2*Df))
## experience_level 1.111792  3        1.017819
## employment_type  1.061636  3        1.010018
## company_size     1.117545  2        1.028173
## job_category     1.091243  9        1.004863
```

```r
# Plot diagnostic plots to check assumptions
par(mfrow = c(2, 2))
plot(log_model)
```



Log-Transformed Model Summary

Model Coefficients

- Experience Level: Executive - Has the largest positive effect on log salary, with an estimate of 0.69688, meaning executives tend to have significantly higher salaries. Senior and Mid-level - Also have positive effects on salary, with senior-level roles (estimate = 0.56555) showing a larger impact than mid-level (estimate = 0.26496). All experience levels (Executive, Mid-level, and Senior) are highly significant predictors ($p < 0.001$), suggesting that experience level is a strong determinant of salary.
- Employment Type: Freelance - Has a significant negative effect on log salary (estimate = -0.50722), indicating lower pay for freelance roles. Full-time - Has a positive effect on salary (estimate = 0.24247), significant at the 0.05 level. Part-time - Has a negative effect, but it is only marginally significant ($p = 0.085$).
- Company Size: Medium (company_sizeM) - Positive effect (estimate = 0.10460), indicating slightly higher salaries compared to the baseline (likely large companies). Small (company_sizeS) - Significant negative effect on salary (estimate = -0.32417), suggesting that small companies generally offer lower salaries.
- Job Category: Positive Impacts - Job categories like Machine Learning and AI (estimate = 0.24355) and Data Science and Research (estimate = 0.16617) show significant positive impacts, aligning with industry demand and competitive compensation. Negative Impacts - Roles in Data Analysis (estimate = -0.17363) and Data Quality and Operations (estimate = -0.30714) have significant negative impacts on salary, reflecting comparatively lower pay in these categories.

Model Fit

- R-squared: 0.29, meaning the model explains about 29% of the variability in log-transformed salary, which is reasonable given the complexity of salary determinants.
- Adjusted R-squared: 0.2886, which is close to the R-squared, suggesting that the model doesn't lose much explanatory power when accounting for the number of predictors.
- Residual Standard Error: 0.3869, indicating the average deviation of observed log salaries from the predicted values.
- F-statistic: The overall model is highly significant (p-value $<$ 2.2e-16), meaning that at least some predictors are significantly associated with salary.

Variance Inflation Factor (VIF)

- All GVIF values (adjusted for degrees of freedom) are below 1.1, indicating low multicollinearity among predictors. This suggests that each predictor contributes unique information to the model, supporting the stability and reliability of coefficient estimates.

Model Diagnostics

- Residuals vs. Fitted: The plot shows a fairly random scatter of residuals around the zero line, though there may be slight clustering. The absence of a clear pattern indicates that the assumption of homoscedasticity is reasonably met, and the linearity assumption is appropriate for the model. However, there is some slight variation in residual spread, which should be monitored but isn't necessarily problematic for this model.
- Normal Q-Q Plot: Most points lie close to the diagonal line, indicating that residuals are approximately normally distributed. However, there is slight deviation at both tails (particularly at the upper end), which suggests minor departures from normality. This is common in real-world data, and unless the deviation is severe, it's generally acceptable.
- Scale-Location Plot: The points are spread relatively evenly along the fitted values, with no clear pattern or trend, which supports the assumption of homoscedasticity. The red line is mostly horizontal, indicating that residual variance does not vary systematically with the fitted values. This plot reinforces the Residuals vs. Fitted plot's indication of acceptable homoscedasticity.
- Residuals vs. Leverage: A few points, such as those labeled (e.g., observations 8016, 8428, 9149), are close to or slightly beyond the threshold of Cook's distance (0.5), which suggests they may have some influence on the model. While they do not appear to be highly influential, further inspection of these points is advisable to determine if they represent extreme values or unique cases that could potentially impact model stability. Removing or adjusting for these points could be considered if they are deemed outliers.

Conclusion

The diagnostic plots indicate that the log-transformed model is generally well-fitted to the data with the residuals are approximately normally distributed, with only slight deviations at the tails and both the Residuals vs. Fitted and Scale-Location plots support the assumption of constant variance. There are a few potentially influential observations, but they do not appear to have a substantial impact on the overall model fit. The log-transformation appears to improve the model by addressing skewness in salary and enhancing model assumptions. The model is appropriate for predicting salary, given the satisfactory adherence to linear regression assumptions, though regularization methods (e.g., Lasso or Ridge) could further refine predictor selection and manage any minor multicollinearity or outlier influence.

Interaction Terms Model

```r
# Simply the model and introduce interaction terms
interaction_model_cleaned <- lm(salary_in_usd ~ experience_level * employment_type + experience_level *
                                  job_category + company_size, data = data_cleaned
)

# Summary and plot of the interaction model (Clean Data)
summary(interaction_model_cleaned)
```

```
##
## Call:
## lm(formula = salary_in_usd ~ experience_level * employment_type +
##     experience_level * job_category + company_size, data = data_cleaned)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -166284  -34975   -5191   30099  175263
##
## Coefficients: (10 not defined because of singularities)
##                                                                 Estimate
## (Intercept)                                                        75206
## experience_levelExecutive                                          98455
## experience_levelMid-level                                          -1701
## experience_levelSenior                                             20530
## employment_typeFreelance                                            4048
## employment_typeFull-time                                            3622
## employment_typePart-time                                          -15254
## job_categoryCloud and Database                                     10943
## job_categoryData Analysis                                         -11695
## job_categoryData Architecture and Modeling                         11363
## job_categoryData Engineering                                       11845
## job_categoryData Management and Strategy                            4529
## job_categoryData Quality and Operations                           -47724
## job_categoryData Science and Research                              19325
## job_categoryLeadership and Management                              -1192
## job_categoryMachine Learning and AI                                17338
## company_sizeM                                                       8284
## company_sizeS                                                     -32938
## experience_levelExecutive:employment_typeFreelance                    NA
## experience_levelMid-level:employment_typeFreelance                -43858
## experience_levelSenior:employment_typeFreelance                   -64528
## experience_levelExecutive:employment_typeFull-time                    NA
## experience_levelMid-level:employment_typeFull-time                 17924
## experience_levelSenior:employment_typeFull-time                    38072
## experience_levelExecutive:employment_typePart-time                    NA
## experience_levelMid-level:employment_typePart-time                 -7928
## experience_levelSenior:employment_typePart-time                       NA
## experience_levelExecutive:job_categoryCloud and Database              NA
## experience_levelMid-level:job_categoryCloud and Database              NA
## experience_levelSenior:job_categoryCloud and Database                 NA
## experience_levelExecutive:job_categoryData Analysis               -63198
## experience_levelMid-level:job_categoryData Analysis                 4854
## experience_levelSenior:job_categoryData Analysis                  -14118
## experience_levelExecutive:job_categoryData Architecture and Modeling   -29430
```

```
## experience_levelMid-level:job_categoryData Architecture and Modeling    12974
## experience_levelSenior:job_categoryData Architecture and Modeling          NA
## experience_levelExecutive:job_categoryData Engineering                -20197
## experience_levelMid-level:job_categoryData Engineering                  1086
## experience_levelSenior:job_categoryData Engineering                    -1993
## experience_levelExecutive:job_categoryData Management and Strategy        NA
## experience_levelMid-level:job_categoryData Management and Strategy     -7514
## experience_levelSenior:job_categoryData Management and Strategy       -38292
## experience_levelExecutive:job_categoryData Quality and Operations         NA
## experience_levelMid-level:job_categoryData Quality and Operations      16872
## experience_levelSenior:job_categoryData Quality and Operations         20864
## experience_levelExecutive:job_categoryData Science and Research       -12509
## experience_levelMid-level:job_categoryData Science and Research         8074
## experience_levelSenior:job_categoryData Science and Research            6400
## experience_levelExecutive:job_categoryLeadership and Management        -4346
## experience_levelMid-level:job_categoryLeadership and Management         9443
## experience_levelSenior:job_categoryLeadership and Management            5755
## experience_levelExecutive:job_categoryMachine Learning and AI           4311
## experience_levelMid-level:job_categoryMachine Learning and AI          25158
## experience_levelSenior:job_categoryMachine Learning and AI             22140
##                                                                     Std. Error
## (Intercept)                                                              31675
## experience_levelExecutive                                                27910
## experience_levelMid-level                                                36134
## experience_levelSenior                                                   43034
## employment_typeFreelance                                                 44364
## employment_typeFull-time                                                 25354
## employment_typePart-time                                                 29074
## job_categoryCloud and Database                                           22665
## job_categoryData Analysis                                                19443
## job_categoryData Architecture and Modeling                                4692
## job_categoryData Engineering                                             19647
## job_categoryData Management and Strategy                                 23517
## job_categoryData Quality and Operations                                  34618
## job_categoryData Science and Research                                    19368
## job_categoryLeadership and Management                                    24991
## job_categoryMachine Learning and AI                                      20031
## company_sizeM                                                             1990
## company_sizeS                                                             4547
## experience_levelExecutive:employment_typeFreelance                          NA
## experience_levelMid-level:employment_typeFreelance                       52233
## experience_levelSenior:employment_typeFreelance                          58636
## experience_levelExecutive:employment_typeFull-time                          NA
## experience_levelMid-level:employment_typeFull-time                       30046
## experience_levelSenior:employment_typeFull-time                          38496
## experience_levelExecutive:employment_typePart-time                          NA
## experience_levelMid-level:employment_typePart-time                       44187
## experience_levelSenior:employment_typePart-time                             NA
## experience_levelExecutive:job_categoryCloud and Database                    NA
## experience_levelMid-level:job_categoryCloud and Database                    NA
## experience_levelSenior:job_categoryCloud and Database                       NA
## experience_levelExecutive:job_categoryData Analysis                      30898
## experience_levelMid-level:job_categoryData Analysis                      20685
## experience_levelSenior:job_categoryData Analysis                         19791
```

```
## experience_levelExecutive:job_categoryData Architecture and Modeling      41226
## experience_levelMid-level:job_categoryData Architecture and Modeling      11999
## experience_levelSenior:job_categoryData Architecture and Modeling            NA
## experience_levelExecutive:job_categoryData Engineering                    28828
## experience_levelMid-level:job_categoryData Engineering                    20848
## experience_levelSenior:job_categoryData Engineering                       19956
## experience_levelExecutive:job_categoryData Management and Strategy           NA
## experience_levelMid-level:job_categoryData Management and Strategy        26875
## experience_levelSenior:job_categoryData Management and Strategy           25561
## experience_levelExecutive:job_categoryData Quality and Operations            NA
## experience_levelMid-level:job_categoryData Quality and Operations         37409
## experience_levelSenior:job_categoryData Quality and Operations            35770
## experience_levelExecutive:job_categoryData Science and Research           28792
## experience_levelMid-level:job_categoryData Science and Research           20580
## experience_levelSenior:job_categoryData Science and Research              19671
## experience_levelExecutive:job_categoryLeadership and Management           33037
## experience_levelMid-level:job_categoryLeadership and Management           26226
## experience_levelSenior:job_categoryLeadership and Management              25366
## experience_levelExecutive:job_categoryMachine Learning and AI             32076
## experience_levelMid-level:job_categoryMachine Learning and AI             21333
## experience_levelSenior:job_categoryMachine Learning and AI                20358
##                                                                        t value
## (Intercept)                                                              2.374
## experience_levelExecutive                                                3.528
## experience_levelMid-level                                               -0.047
## experience_levelSenior                                                   0.477
## employment_typeFreelance                                                 0.091
## employment_typeFull-time                                                 0.143
## employment_typePart-time                                                -0.525
## job_categoryCloud and Database                                           0.483
## job_categoryData Analysis                                               -0.601
## job_categoryData Architecture and Modeling                               2.422
## job_categoryData Engineering                                             0.603
## job_categoryData Management and Strategy                                 0.193
## job_categoryData Quality and Operations                                 -1.379
## job_categoryData Science and Research                                    0.998
## job_categoryLeadership and Management                                   -0.048
## job_categoryMachine Learning and AI                                      0.866
## company_sizeM                                                            4.162
## company_sizeS                                                           -7.244
## experience_levelExecutive:employment_typeFreelance                          NA
## experience_levelMid-level:employment_typeFreelance                      -0.840
## experience_levelSenior:employment_typeFreelance                         -1.100
## experience_levelExecutive:employment_typeFull-time                          NA
## experience_levelMid-level:employment_typeFull-time                       0.597
## experience_levelSenior:employment_typeFull-time                          0.989
## experience_levelExecutive:employment_typePart-time                          NA
## experience_levelMid-level:employment_typePart-time                      -0.179
## experience_levelSenior:employment_typePart-time                             NA
## experience_levelExecutive:job_categoryCloud and Database                    NA
## experience_levelMid-level:job_categoryCloud and Database                    NA
## experience_levelSenior:job_categoryCloud and Database                       NA
## experience_levelExecutive:job_categoryData Analysis                     -2.045
## experience_levelMid-level:job_categoryData Analysis                      0.235
```

```
## experience_levelSenior:job_categoryData Analysis                     -0.713
## experience_levelExecutive:job_categoryData Architecture and Modeling  -0.714
## experience_levelMid-level:job_categoryData Architecture and Modeling   1.081
## experience_levelSenior:job_categoryData Architecture and Modeling         NA
## experience_levelExecutive:job_categoryData Engineering                -0.701
## experience_levelMid-level:job_categoryData Engineering                 0.052
## experience_levelSenior:job_categoryData Engineering                   -0.100
## experience_levelExecutive:job_categoryData Management and Strategy        NA
## experience_levelMid-level:job_categoryData Management and Strategy    -0.280
## experience_levelSenior:job_categoryData Management and Strategy       -1.498
## experience_levelExecutive:job_categoryData Quality and Operations         NA
## experience_levelMid-level:job_categoryData Quality and Operations      0.451
## experience_levelSenior:job_categoryData Quality and Operations         0.583
## experience_levelExecutive:job_categoryData Science and Research       -0.434
## experience_levelMid-level:job_categoryData Science and Research        0.392
## experience_levelSenior:job_categoryData Science and Research           0.325
## experience_levelExecutive:job_categoryLeadership and Management       -0.132
## experience_levelMid-level:job_categoryLeadership and Management        0.360
## experience_levelSenior:job_categoryLeadership and Management           0.227
## experience_levelExecutive:job_categoryMachine Learning and AI          0.134
## experience_levelMid-level:job_categoryMachine Learning and AI          1.179
## experience_levelSenior:job_categoryMachine Learning and AI             1.087
##                                                                     Pr(>|t|)
## (Intercept)                                                         0.017602
## experience_levelExecutive                                           0.000421
## experience_levelMid-level                                           0.962461
## experience_levelSenior                                              0.633325
## employment_typeFreelance                                            0.927292
## employment_typeFull-time                                            0.886403
## employment_typePart-time                                            0.599823
## job_categoryCloud and Database                                      0.629253
## job_categoryData Analysis                                           0.547528
## job_categoryData Architecture and Modeling                          0.015458
## job_categoryData Engineering                                        0.546574
## job_categoryData Management and Strategy                            0.847289
## job_categoryData Quality and Operations                             0.168061
## job_categoryData Science and Research                               0.318431
## job_categoryLeadership and Management                               0.961968
## job_categoryMachine Learning and AI                                 0.386770
## company_sizeM                                                       3.18e-05
## company_sizeS                                                       4.71e-13
## experience_levelExecutive:employment_typeFreelance                        NA
## experience_levelMid-level:employment_typeFreelance                  0.401128
## experience_levelSenior:employment_typeFreelance                     0.271152
## experience_levelExecutive:employment_typeFull-time                        NA
## experience_levelMid-level:employment_typeFull-time                  0.550818
## experience_levelSenior:employment_typeFull-time                     0.322696
## experience_levelExecutive:employment_typePart-time                        NA
## experience_levelMid-level:employment_typePart-time                  0.857611
## experience_levelSenior:employment_typePart-time                           NA
## experience_levelExecutive:job_categoryCloud and Database                  NA
## experience_levelMid-level:job_categoryCloud and Database                  NA
## experience_levelSenior:job_categoryCloud and Database                     NA
## experience_levelExecutive:job_categoryData Analysis                 0.040846
```

```
## experience_levelMid-level:job_categoryData Analysis                0.814476
## experience_levelSenior:job_categoryData Analysis                   0.475628
## experience_levelExecutive:job_categoryData Architecture and Modeling 0.475327
## experience_levelMid-level:job_categoryData Architecture and Modeling 0.279603
## experience_levelSenior:job_categoryData Architecture and Modeling        NA
## experience_levelExecutive:job_categoryData Engineering             0.483567
## experience_levelMid-level:job_categoryData Engineering             0.958453
## experience_levelSenior:job_categoryData Engineering                0.920440
## experience_levelExecutive:job_categoryData Management and Strategy       NA
## experience_levelMid-level:job_categoryData Management and Strategy   0.779785
## experience_levelSenior:job_categoryData Management and Strategy      0.134139
## experience_levelExecutive:job_categoryData Quality and Operations        NA
## experience_levelMid-level:job_categoryData Quality and Operations    0.651988
## experience_levelSenior:job_categoryData Quality and Operations       0.559723
## experience_levelExecutive:job_categoryData Science and Research      0.663969
## experience_levelMid-level:job_categoryData Science and Research      0.694841
## experience_levelSenior:job_categoryData Science and Research         0.744916
## experience_levelExecutive:job_categoryLeadership and Management      0.895352
## experience_levelMid-level:job_categoryLeadership and Management      0.718798
## experience_levelSenior:job_categoryLeadership and Management         0.820535
## experience_levelExecutive:job_categoryMachine Learning and AI        0.893080
## experience_levelMid-level:job_categoryMachine Learning and AI        0.238305
## experience_levelSenior:job_categoryMachine Learning and AI           0.276851
##
## (Intercept)                                                         *
## experience_levelExecutive                                           ***
## experience_levelMid-level
## experience_levelSenior
## employment_typeFreelance
## employment_typeFull-time
## employment_typePart-time
## job_categoryCloud and Database
## job_categoryData Analysis
## job_categoryData Architecture and Modeling                          *
## job_categoryData Engineering
## job_categoryData Management and Strategy
## job_categoryData Quality and Operations
## job_categoryData Science and Research
## job_categoryLeadership and Management
## job_categoryMachine Learning and AI
## company_sizeM                                                       ***
## company_sizeS                                                       ***
## experience_levelExecutive:employment_typeFreelance
## experience_levelMid-level:employment_typeFreelance
## experience_levelSenior:employment_typeFreelance
## experience_levelExecutive:employment_typeFull-time
## experience_levelMid-level:employment_typeFull-time
## experience_levelSenior:employment_typeFull-time
## experience_levelExecutive:employment_typePart-time
## experience_levelMid-level:employment_typePart-time
## experience_levelSenior:employment_typePart-time
## experience_levelExecutive:job_categoryCloud and Database
## experience_levelMid-level:job_categoryCloud and Database
## experience_levelSenior:job_categoryCloud and Database
```

```
## experience_levelExecutive:job_categoryData Analysis                *
## experience_levelMid-level:job_categoryData Analysis
## experience_levelSenior:job_categoryData Analysis
## experience_levelExecutive:job_categoryData Architecture and Modeling
## experience_levelMid-level:job_categoryData Architecture and Modeling
## experience_levelSenior:job_categoryData Architecture and Modeling
## experience_levelExecutive:job_categoryData Engineering
## experience_levelMid-level:job_categoryData Engineering
## experience_levelSenior:job_categoryData Engineering
## experience_levelExecutive:job_categoryData Management and Strategy
## experience_levelMid-level:job_categoryData Management and Strategy
## experience_levelSenior:job_categoryData Management and Strategy
## experience_levelExecutive:job_categoryData Quality and Operations
## experience_levelMid-level:job_categoryData Quality and Operations
## experience_levelSenior:job_categoryData Quality and Operations
## experience_levelExecutive:job_categoryData Science and Research
## experience_levelMid-level:job_categoryData Science and Research
## experience_levelSenior:job_categoryData Science and Research
## experience_levelExecutive:job_categoryLeadership and Management
## experience_levelMid-level:job_categoryLeadership and Management
## experience_levelSenior:job_categoryLeadership and Management
## experience_levelExecutive:job_categoryMachine Learning and AI
## experience_levelMid-level:job_categoryMachine Learning and AI
## experience_levelSenior:job_categoryMachine Learning and AI
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 50160 on 9153 degrees of freedom
## Multiple R-squared:  0.2639, Adjusted R-squared:  0.2605
## F-statistic: 76.33 on 43 and 9153 DF,  p-value: < 2.2e-16
```

Interaction Terms Model Summary

Model Coefficients

- Main Effects: Experience Level - Executive level has a highly significant positive effect on salary (estimate = 98455, $p < 0.001$), indicating that executives tend to earn significantly higher salaries. However, mid-level and senior experience levels do not have significant main effects on salary in this model, possibly because their effects are moderated by employment type and job category.
- Company Size: Medium companies show a significant positive effect on salary (estimate = 8284, $p < 0.001$), while small companies show a significant negative effect (estimate = -32938, $p < 0.001$), consistent with previous models that suggest smaller companies typically offer lower salaries.
- Interactions: Experience Level and Employment Type - Some interaction terms between experience level and employment type have coefficients listed as "NA" due to singularities, meaning that these terms are collinear with other variables and were dropped from the model. For instance, interactions involving executive-level employees with freelance, full-time, and part-time employment types are undefined. Experience Level and Job Category - For Executive Data Analysis interaction is significant (estimate = -63198, $p = 0.041$), suggesting that executive roles in Data Analysis are associated with lower salaries compared to other executive roles. This may reflect the relatively lower market demand or compensation structure within the Data Analysis category. The remaining interactions involving experience level and job category are largely not significant, indicating that salary variation within job categories is not strongly moderated by experience level for most roles.

Model Fit

- R-squared and Adjusted R-squared: The multiple R-squared is 0.2639, and the adjusted R-squared is 0.2605, indicating that about 26.05% of the variance in salary is explained by the model. This represents a slight improvement over simpler models but still reflects modest explanatory power. Given the number of predictors, the increase in R-squared is not substantial, suggesting that these interactions, while insightful, do not dramatically enhance predictive accuracy.
- Residual Standard Error: The residual standard error is 50160, which indicates the average deviation of the observed salaries from the fitted values. This value is similar to previous models, suggesting that adding interactions has not substantially reduced prediction error.

Interpretation of Model Results

- Interaction Effects: Including interaction terms adds complexity to the model, allowing it to explore whether the impact of experience level on salary depends on employment type and job category. The significant interaction between executive-level experience and the Data Analysis job category suggests that certain combinations of experience level and job category have unique effects on salary.
- Non-significant Interactions: The majority of interaction terms are not significant, suggesting that the main effects of experience level, employment type, and job category capture much of the relationship with salary without requiring interaction terms. The lack of significance for many interaction terms could also indicate multicollinearity, as seen with the "NA" coefficients.

Model Diagnostics

- Singularities: Several coefficients are marked as "NA" due to singularities, which occur when predictors are perfectly or nearly perfectly collinear. This is common in models with interaction terms and suggests that some categories (e.g., experience levels within specific employment types) may be redundant.
- F-statistic: The overall model is highly significant ($p < 2.2e\text{-}16$), meaning that, collectively, the predictors are associated with salary, though not all individual predictors are significant.

Conclusion

This interaction model highlights some conditional relationships in salary data but may not substantially improve predictive power over simpler models. While the interactions provide insights, especially regarding the executive-level experience in certain job categories, the model suffers from multicollinearity issues (as indicated by "NA" coefficients) and still shows modest explanatory power.

Analytical Modeling

Lasso and Ridge are regularization techniques used to handle multicollinearity and prevent overfitting. We will use these models as part of our analysis.

Lasso Regression

```r
# Define the cleaned response variable after removing outliers
y <- data_cleaned$salary_in_usd

# Prepare the encoded data matrix after cleaning
data_encoded <- model.matrix(salary_in_usd ~ experience_level + employment_type + company_size +
                                job_category, data = data_cleaned)[, -1]

# Lasso regression with cross-validation
```

```r
lasso_model <- cv.glmnet(data_encoded, y, alpha = 1)
lasso_best_lambda <- lasso_model$lambda.min
lasso_pred <- predict(lasso_model, s = lasso_best_lambda, newx = data_encoded)
lasso_r2 <- 1 - sum((y - lasso_pred)^2) / sum((y - mean(y))^2)

# Display model, coefficients, and R-squared
coef(lasso_model, s = lasso_best_lambda)
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                                                   s1
## (Intercept)                                59700.603
## experience_levelExecutive                  79968.982
## experience_levelMid-level                  21751.890
## experience_levelSenior                     58238.177
## employment_typeFreelance                  -45908.000
## employment_typeFull-time                   19566.104
## employment_typePart-time                   -6537.856
## company_sizeM                               8569.903
## company_sizeS                             -34813.786
## job_categoryCloud and Database              9705.643
## job_categoryData Analysis                 -21524.088
## job_categoryData Architecture and Modeling  11674.016
## job_categoryData Engineering                8886.609
## job_categoryData Management and Strategy   -17615.709
## job_categoryData Quality and Operations    -30992.132
## job_categoryData Science and Research      24294.944
## job_categoryLeadership and Management       4442.532
## job_categoryMachine Learning and AI        37627.274
```

```r
lasso_r2
```

```
## [1] 0.2585987
```

Lasso Regression Model Summary

Model Coefficients

- Experience Level: Executive - The positive coefficient (80,018) indicates that executive-level positions are associated with a substantial salary increase relative to the baseline (likely entry-level), reflecting the significant impact of experience. Senior and Mid-level - Both levels positively impact salary, with coefficients of 58,278 and 21,796, respectively. This hierarchy in effect sizes aligns with the expectation that higher experience levels correlate with higher salaries.
- Employment Type: Freelance - The negative coefficient (-45,973) suggests that freelance roles are associated with significantly lower salaries compared to the baseline employment type, likely full-time or hybrid. This result is consistent with common compensation practices where freelance roles may offer less stable income. Full-time - The positive coefficient (19,554) indicates a salary premium for full-time roles, while part-time employment has a small negative effect (-6,568), suggesting lower compensation for part-time work.
- Company Size: Medium-sized companies - A positive coefficient (8,564) suggests a modest salary premium, although the effect is not as strong as in larger companies. Small companies - A negative coefficient (-34,826) indicates that small companies generally offer lower salaries compared to the baseline (likely large companies), possibly due to budget constraints or resource limitations.

- Job Category: Machine Learning and AI - With the highest positive coefficient (37,710), this category aligns with industry demand and high compensation for specialized skills. Data Science and Research - The positive coefficient (24,376) reflects the high value of data science expertise in the job market. Data Quality and Operations and Data Analysis - Both categories have negative coefficients, indicating lower salaries in these fields. This could be due to these roles being more operational and less specialized compared to others like AI and data science. Data Architecture and Modeling and Data Engineering - These roles have positive coefficients (11,762 and 8,968, respectively), indicating competitive salaries within technical roles, though generally less than in Machine Learning and Data Science. Leadership and Management - A small positive coefficient (4,526) suggests a slight salary premium, though this effect is less pronounced than for technical fields like Machine Learning and Data Science.

Model Fit

- R-squared: The Lasso model explains about 25.9% of the variance in salary (R-squared = 0.2586). While modest, this R-squared value is expected given the many unobserved factors that influence salary. It's common for salary models to have lower R-squared values due to the influence of personal, company-specific, and external economic factors.

Key Insights

- Significant Predictors: The model highlights experience level, employment type, company size, and job category as influential predictors. This finding is consistent with expectations, where experience, employment stability, and specialized roles significantly impact salary.
- Feature Selection: Lasso regularization effectively zeroed out less impactful variables (such as some categories within employment type), enhancing model simplicity and interpretability.
- Multicollinearity: Work setting was deliberately excluded in this model to allow selection of the best model structure based on other core variables. After selecting the best model, work setting can be reintroduced to see if it further improves the model's predictive accuracy or interpretation.

Conclusion

The Lasso model suggests that salary is strongly influenced by experience level, specialized technical roles, and employment type. Executive experience, technical roles like Machine Learning and Data Science, and full-time employment show substantial positive impacts on salary, while freelance and operational roles are associated with lower pay. Lasso regularization provides a simplified model by omitting variables with minimal predictive power, thereby focusing on significant factors and reducing potential overfitting.

Ridge Regression

```r
# Ridge regression with cross-validation
ridge_model <- cv.glmnet(data_encoded, y, alpha = 0)
ridge_best_lambda <- ridge_model$lambda.min
ridge_pred <- predict(ridge_model, s = ridge_best_lambda, newx = data_encoded)
ridge_r2 <- 1 - sum((y - ridge_pred)^2) / sum((y - mean(y))^2)

# Display Ridge R-squared
# Display model, coefficients and R-squared
coef(ridge_model)
```

```
## 18 x 1 sparse Matrix of class "dgCMatrix"
##                                                     s1
```

```
## (Intercept)                                     93407.4891
## experience_levelExecutive                       41592.8093
## experience_levelMid-level                       -5494.1907
## experience_levelSenior                          28457.7893
## employment_typeFreelance                       -38182.0965
## employment_typeFull-time                        22163.7559
## employment_typePart-time                       -21601.1094
## company_sizeM                                   10363.5908
## company_sizeS                                  -30935.2159
## job_categoryCloud and Database                   2135.1748
## job_categoryData Analysis                      -27433.5765
## job_categoryData Architecture and Modeling       2686.3949
## job_categoryData Engineering                     -610.7158
## job_categoryData Management and Strategy       -26592.5157
## job_categoryData Quality and Operations        -33545.2744
## job_categoryData Science and Research           12254.0103
## job_categoryLeadership and Management           -3297.2966
## job_categoryMachine Learning and AI             22814.5730
```

```r
ridge_r2
```

```
## [1] 0.2569545
```

```r
summary(ridge_model)
```

```
##            Length Class  Mode
## lambda     100    -none- numeric
## cvm        100    -none- numeric
## cvsd       100    -none- numeric
## cvup       100    -none- numeric
## cvlo       100    -none- numeric
## nzero      100    -none- numeric
## call         4    -none- call
## name         1    -none- character
## glmnet.fit  12    elnet  list
## lambda.min   1    -none- numeric
## lambda.1se   1    -none- numeric
## index        2    -none- numeric
```

Ridge Regression Model Summary

Model Coefficients

- Experience Level: Executive - The positive coefficient (47,528) suggests a substantial increase in salary for executive roles, though it is somewhat less pronounced than in the Lasso model. Senior - Positive coefficient (32,513), indicating that senior roles command higher salaries compared to the baseline (likely entry-level). Mid-level - The small negative coefficient (-2,392) suggests minimal difference from the baseline group, potentially reflecting reduced impact in the Ridge model compared to the Lasso model.
- Employment Type: Freelance - The negative coefficient (-40,188) aligns with the trend that freelance roles tend to offer lower salaries compared to the baseline. Full-time - Positive coefficient (21,774) suggests a salary premium for full-time roles, while part-time employment has a negative effect (-20,240),

indicating a lower salary for part-time roles. Company Size: Medium-sized companies - Positive coefficient (10,524) indicates a modest salary increase compared to the baseline. Small companies - Negative coefficient (-32,237) suggests that small companies are associated with lower salaries, likely due to fewer resources or smaller budgets.

- Job Category: Machine Learning and AI - With a positive coefficient (24,725), this category shows one of the highest positive effects, aligning with the high demand and compensation for specialized technical skills. Data Science and Research - The positive coefficient (13,413) indicates a salary premium in this category, although less pronounced than in the Lasso model. Data Analysis and Data Quality and Operations - Both categories have significant negative coefficients (-28,195 and -35,104, respectively), indicating lower salaries within these fields. Data Architecture and Modeling and Data Engineering - These roles show positive but smaller coefficients (3,188 and -128, respectively), suggesting a minimal impact on salary relative to the baseline. Leadership and Management - A small negative coefficient (-3,147) suggests that salaries in this category are slightly lower than in the baseline, though the effect is minor compared to technical roles.

Model Fit

- R-squared: The Ridge model has an R-squared of approximately 25.7%, indicating that it explains about 25.7% of the variance in salary. This is very similar to the Lasso model, as expected since both models target overfitting reduction and improved interpretability.

Key Insights

- Significant Predictors: Experience level, employment type, company size, and job category remain important predictors. The coefficients for these variables are consistent with expectations: executive roles, full-time employment, and technical fields command higher salaries, while freelance and part-time work tend to be associated with lower pay.
- Coefficient Shrinkage: Unlike Lasso, Ridge regression does not eliminate any predictors, so all predictors remain in the model with adjusted coefficients that are generally smaller than in the Lasso model. This can make Ridge regression useful when we want to keep all predictors in the model, especially for interpretation purposes.

Conclusion

The Ridge regression model highlights the same general trends as Lasso: salary is significantly influenced by experience level, employment type, and job category. Executive experience, technical roles (like Machine Learning and Data Science), and full-time employment show positive impacts on salary, while freelance and operational roles are associated with lower pay. The Ridge model's R-squared is close to that of the Lasso model, suggesting comparable explanatory power.

Model Comparison & Selection

We'll compare the performance of the models using AIC, BIC, R-squared, and Adjusted R-squared.

```r
# Model Comparison

# For Stepwise Regression
stepwise_aic <- AIC(stepwise_model)
stepwise_bic <- BIC(stepwise_model)
stepwise_r2 <- summary(stepwise_model)$r.squared
stepwise_adj_r2 <- summary(stepwise_model)$adj.r.squared
```

```r
# For Lasso Regression (cross-validated)
lasso_n <- length(y)
lasso_log_likelihood <- -0.5 * lasso_n * log(sum((y - lasso_pred)^2) / lasso_n)
lasso_aic <- -2 * lasso_log_likelihood + 2 * length(coef(lasso_model, s = lasso_best_lambda) != 0)
lasso_bic <- -2 * lasso_log_likelihood + log(lasso_n) * length(coef(lasso_model,
                                                                    s = lasso_best_lambda) != 0)
lasso_r2 <- lasso_r2  # Calculated from previous code
lasso_adj_r2 <- 1 - ((1 - lasso_r2) * (lasso_n - 1) / (lasso_n - length(coef(lasso_model,
                                                                    s = lasso_best_lambda))))

# For Ridge Regression (cross-validated)
ridge_log_likelihood <- -0.5 * lasso_n * log(sum((y - ridge_pred)^2) / lasso_n)
ridge_aic <- -2 * ridge_log_likelihood + 2 * length(coef(ridge_model, s = ridge_best_lambda) != 0)
ridge_bic <- -2 * ridge_log_likelihood + log(lasso_n) * length(coef(ridge_model,
                                                                    s = ridge_best_lambda) != 0)
ridge_r2 <- ridge_r2  # Calculated from previous code
ridge_adj_r2 <- 1 - ((1 - ridge_r2) * (lasso_n - 1) / (lasso_n - length(coef(ridge_model,
                                                                    s = ridge_best_lambda))))

# Compile comparison results into a data frame for better readability
comparison_results <- data.frame(
  Model = c("Stepwise Regression", "Lasso Regression", "Ridge Regression"),
  AIC = c(stepwise_aic, lasso_aic, ridge_aic),
  BIC = c(stepwise_bic, lasso_bic, ridge_bic),
  R_squared = c(stepwise_r2, lasso_r2, ridge_r2),
  Adjusted_R_squared = c(stepwise_adj_r2, lasso_adj_r2, ridge_adj_r2)
)

# Display the results for comparison
comparison_results
```

```
##                 Model      AIC      BIC R_squared Adjusted_R_squared
## 1 Stepwise Regression 225239.1 225374.5 0.2586146          0.2572415
## 2    Lasso Regression 199137.3 199265.6 0.2585987          0.2572256
## 3    Ridge Regression 199157.7 199286.0 0.2569545          0.2555783
```

Model Comparison & Selection Summary

AIC and BIC:

- Lasso Regression has the lowest AIC (199137.3) and BIC (199265.6) values, indicating the best balance between model fit and complexity among the three models.
- Ridge Regression follows closely with slightly higher AIC and BIC values than Lasso, suggesting a similar level of fit but with a slightly higher penalty for model complexity.
- Stepwise Regression has the highest AIC (225239.1) and BIC (225374.5), indicating it's a more complex model with potentially overfitting, given that it doesn't use regularization.

R-squared and Adjusted R-squared:

- Stepwise Regression achieves the highest R-squared (0.2586) and Adjusted R-squared (0.2572), suggesting it explains the most variance in the salary variable.

- Lasso Regression has a comparable R-squared (0.2586) and Adjusted R-squared (0.2572), indicating it still explains a substantial amount of variance while also simplifying the model by setting some coefficients to zero.
- Ridge Regression has the lowest R-squared (0.2570) and Adjusted R-squared (0.2556), indicating it explains slightly less variance than Lasso and Stepwise. However, it retains all predictors and shrinks coefficients, addressing multicollinearity without omitting variables.

Conclusion

Lasso Regression appears to be the best model, as it achieves the lowest AIC and BIC while providing a similar R-squared and Adjusted R-squared to Stepwise Regression. Its regularization and feature selection make it a strong choice by improving interpretability and managing multicollinearity. Stepwise Regression explains the most variance, but its higher AIC and BIC indicate it may be overfitting compared to Lasso and Ridge. Ridge Regression performs similarly to Lasso but keeps all predictors, resulting in slightly lower model fit statistics.

Final Model

After evaluating Stepwise, Lasso, and Ridge regression models, Lasso regression was chosen as the best model due to its effective feature selection and regularization. Unlike Ridge, which retains all variables, Lasso regression penalizes coefficients by shrinking less important predictors to zero, thus selecting only the most relevant predictors. This helps mitigate overfitting by focusing on the strongest signals in the data and discarding potentially irrelevant features, which is particularly beneficial for datasets with many variables. Additionally, Lasso regression maintains a balance between model simplicity and interpretability, making it the preferred choice for our objective of accurate salary prediction.

Lasso Regression (Reintroducing work_setting)

After comparing the performance of these models (e.g., based on AIC, BIC, R-squared), we will reintroduce the work_setting variable into the best model.

```r
# Define the cleaned response variable after removing outliers
y <- data_cleaned$salary_in_usd

# Prepare the encoded data matrix after cleaning
data_encoded_Final <- model.matrix(salary_in_usd ~ experience_level + employment_type + company_size +
                                     job_category + work_setting, data = data_cleaned)[, -1]

# Lasso regression with cross-validation
lasso_model_Final <- cv.glmnet(data_encoded_Final, y, alpha = 1)
lasso_best_lambda_Final <- lasso_model_Final$lambda.min
lasso_pred_Final <- predict(lasso_model_Final, s = lasso_best_lambda_Final, newx = data_encoded_Final)
lasso_r2_Final <- 1 - sum((y - lasso_pred_Final)^2) / sum((y - mean(y))^2)

# Display model, coefficients, and R-squared
coef(lasso_model_Final, s = lasso_best_lambda_Final)
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                                          s1
## (Intercept)                       32259.566
## experience_levelExecutive         77460.034
## experience_levelMid-level         19119.656
```

```
## experience_levelSenior                     54945.896
## employment_typeFreelance                  -40984.902
## employment_typeFull-time                   17820.467
## employment_typePart-time                           .
## company_sizeM                               3633.274
## company_sizeS                             -32616.497
## job_categoryCloud and Database             13734.805
## job_categoryData Analysis                 -19544.406
## job_categoryData Architecture and Modeling  14142.234
## job_categoryData Engineering               10550.449
## job_categoryData Management and Strategy  -16974.168
## job_categoryData Quality and Operations   -28961.682
## job_categoryData Science and Research      26050.013
## job_categoryLeadership and Management       5787.928
## job_categoryMachine Learning and AI        39429.551
## work_settingIn-person                      37532.568
## work_settingRemote                         32312.573
```

```
lasso_r2_Final
```

```
## [1] 0.2672834
```

Lasso Regression Summary (Reintroducing work_setting)

Model Coefficients

- Experience Level: Executive - The largest positive coefficient (77,501), indicating that executive-level roles have a substantial impact on salary compared to entry-level roles. Senior and Mid-level - Positive coefficients (54,976 and 19,155, respectively) also reflect increasing salary benefits with experience level. These results confirm that higher experience levels are associated with higher salaries.
- Employment Type: Freelance - The negative coefficient (-41,030) suggests that freelance roles are associated with lower salaries compared to the baseline employment type (likely full-time or hybrid roles). Full-time - A positive coefficient (17,813) indicates a modest salary increase for full-time roles, while part-time does not appear in the final model, likely indicating that it has minimal predictive impact.
- Company Size: Medium-sized companies - A positive coefficient (3,614) suggests a small salary increase, although this effect is not as pronounced as in other categories. Small companies - A negative coefficient (-32,626) indicates that small companies are associated with lower salaries, possibly due to fewer resources or smaller budgets.
- Job Category: Machine Learning and AI - The largest positive impact among job categories (39,510), highlighting high demand and competitive compensation in these technical roles. Data Science and Research - Positive coefficient (26,127), indicating that data science roles are well-compensated, although not to the same extent as machine learning roles. Data Quality and Operations and Data Analysis - Significant negative coefficients (-28,913 and -19,474, respectively) suggest comparatively lower salaries in these support-oriented fields. Data Architecture and Modeling and Data Engineering - Positive coefficients (14,227 and 10,627, respectively), though less substantial than Machine Learning and Data Science, suggest competitive compensation within technical roles. Leadership and Management - A smaller positive coefficient (5,868) reflects a minor salary premium for leadership roles, though the effect is less pronounced than for technical fields.
- Work Setting: In-person - A positive coefficient (37,640), suggesting that in-person roles may offer a salary premium, potentially due to location-specific demand or additional compensation for physical presence. Remote - A positive coefficient (32,420) indicates a salary premium for remote work, reflecting the flexibility and desirability of remote roles. This aligns with trends showing competitive

compensation for remote positions, especially in tech-related fields. Reintroducing work_setting to the final Lasso model revealed that this variable significantly influences salary, with remote and in-person roles exhibiting higher salary levels than hybrid work settings. This finding likely reflects current industry trends in data and tech roles, where flexibility in work arrangements, particularly remote work, has become highly valued. Remote roles may command a premium due to demand for flexibility, while in-person roles could reflect increased compensation related to geographic constraints or specialized onsite responsibilities. This reinforces the relevance of work setting as a key predictor in understanding salary variations.

Model Fit

- R-squared: Approximately 0.2673, meaning the model explains about 26.7% of the variability in salary. This level of explanatory power is typical for salary models, where many unobserved factors influence pay.

Key Insights

- High-Impact Predictors: Experience level, employment type, job category, and work setting are significant drivers of salary.
- Regularization Benefits: By selecting relevant predictors and shrinking coefficients of less impactful ones, Lasso improves the model's interpretability without overfitting, making it ideal for identifying key salary influencers.
- Work Setting Relevance: The inclusion of work_setting highlights the growing importance of flexibility and location in determining salary, with both in-person and remote roles showing positive impacts.

Conclusion

This final model effectively identifies the key drivers of salary within data-related roles. The significant impact of executive experience, full-time employment, and high-demand job categories (such as Machine Learning and Data Science) aligns with industry expectations. Moreover, the positive coefficients for in-person and remote work settings underscore the value of flexible and location-specific work arrangements. This Lasso model is well-suited for practical applications, balancing explanatory power with simplicity, making it a robust choice for predicting salary outcomes in the data job market.

Lasso Regression Diagnostics (Reintroducing work_setting)

```r
# Calculate residuals from the Lasso model
residuals_lasso <- as.vector(y - lasso_pred_Final)

# Linearity Check: Plot each predictor against the residuals from the Lasso model
cat("\nLinearity Check: Predictor vs Residual Plots\n")
```

```
##
## Linearity Check: Predictor vs Residual Plots
```

```r
predictors <- colnames(data_encoded_Final)

for (i in seq_along(predictors)) {
  ggplot(data.frame(Predictor = data_encoded_Final[, i], Residuals = residuals_lasso),
         aes(x = Predictor, y = Residuals)) +
```

```r
    geom_point(alpha = 0.4) +
    geom_smooth(method = "loess", color = "red") +
    labs(title = paste("Residuals vs", predictors[i]),
         x = predictors[i], y = "Residuals") +
    theme_minimal()
}

# Normality of Residuals: KS Test and Histogram
cat("\nNormality of Residuals:\n")
```

```
##
## Normality of Residuals:
```

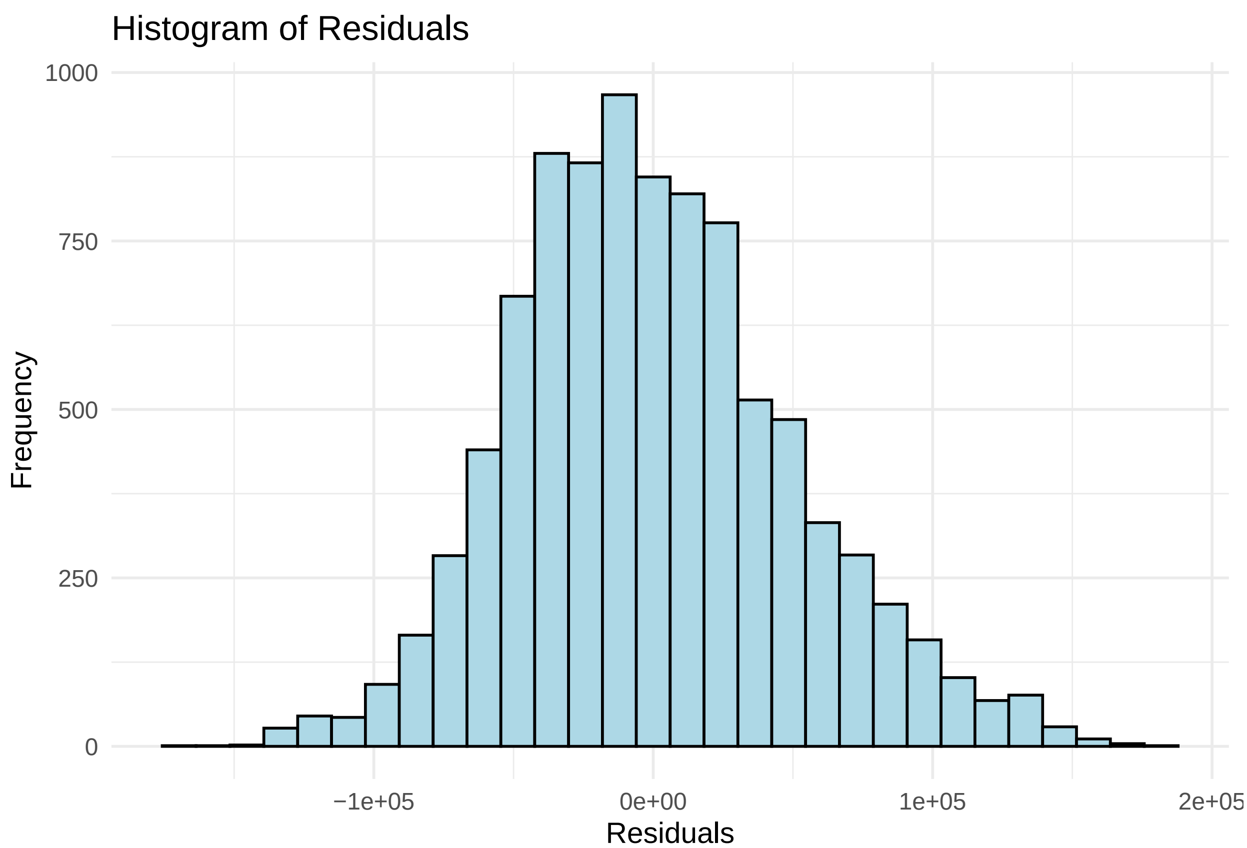
```r
# KS test for normality
ks_test <- ks.test(residuals_lasso, "pnorm", mean(residuals_lasso), sd(residuals_lasso))
```

```
## Warning in ks.test.default(residuals_lasso, "pnorm", mean(residuals_lasso), :
## ties should not be present for the one-sample Kolmogorov-Smirnov test
```

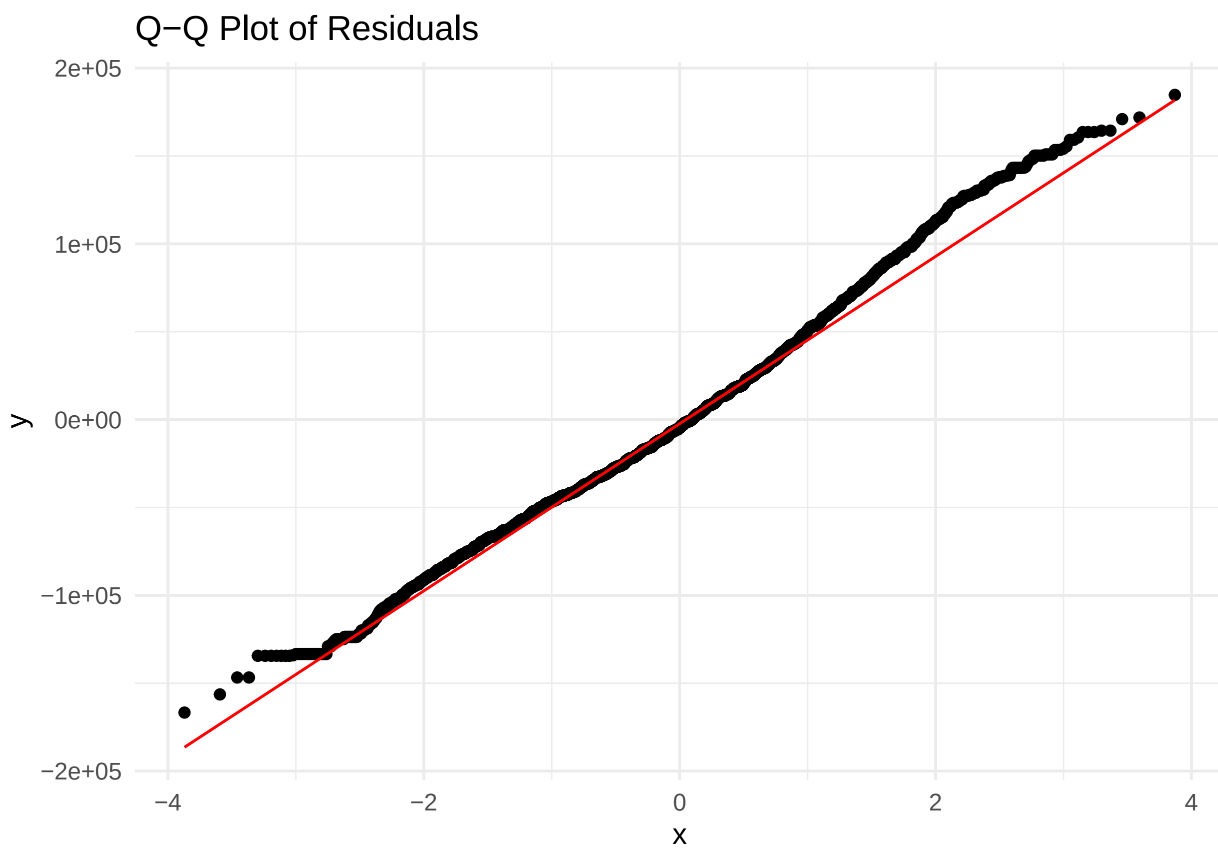
```r
print(ks_test)
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  residuals_lasso
## D = 0.039343, p-value = 8.631e-13
## alternative hypothesis: two-sided
```

```r
# Plot histogram of residuals
ggplot(data.frame(residuals = residuals_lasso), aes(x = residuals)) +
  geom_histogram(bins = 30, fill = "lightblue", color = "black") +
  labs(title = "Histogram of Residuals", x = "Residuals", y = "Frequency") +
  theme_minimal()
```




```r
# Q-Q plot of residuals
ggplot(data.frame(residuals = residuals_lasso), aes(sample = residuals)) +
  stat_qq() +
  stat_qq_line(color = "red") +
  labs(title = "Q-Q Plot of Residuals") +
  theme_minimal()
```




```
# Homoscedasticity Check: Breusch-Pagan Test
cat("\nHomoscedasticity Check: Breusch-Pagan Test\n")
```

```
##
## Homoscedasticity Check: Breusch-Pagan Test
```

```
bp_test <- bptest(residuals_lasso ~ data_encoded_Final)
print(bp_test)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  residuals_lasso ~ data_encoded_Final
## BP = 222.43, df = 19, p-value < 2.2e-16
```

```
# Multicollinearity Check: Variance Inflation Factor (VIF)
cat("\nMulticollinearity Check: Variance Inflation Factor\n")
```

```
##
## Multicollinearity Check: Variance Inflation Factor
```

```
# Re-specify the model for VIF calculation without matrix encoding
vif_model <- lm(salary_in_usd ~ experience_level + employment_type + company_size + job_category +
```

```
                  work_setting, data = data_cleaned)
vif_values <- vif(vif_model)
print(vif_values)
```

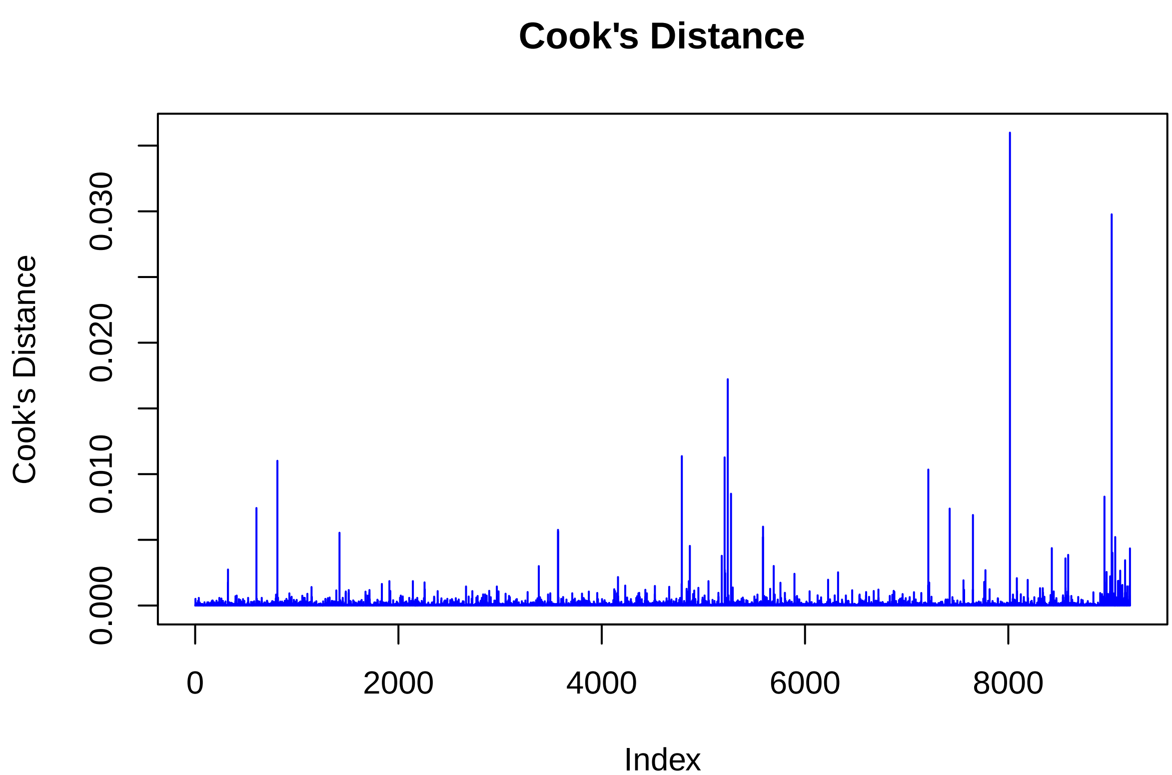
```
##                      GVIF Df GVIF^(1/(2*Df))
## experience_level 1.139972  3        1.022074
## employment_type  1.070175  3        1.011368
## company_size     1.226802  2        1.052431
## job_category     1.105630  9        1.005594
## work_setting     1.180782  2        1.042419
```

```
# Outlier and Influence Analysis: Cook's Distance
cat("\nInfluence Analysis: Cook's Distance\n")
```

```
##
## Influence Analysis: Cook's Distance
```

```
# Calculate Cook's Distance for each observation using a temporary lm model
temp_model <- lm(y ~ data_encoded_Final)
cooks_distances <- cooks.distance(temp_model)

# Plot Cook's Distance
plot(cooks_distances, type = "h", main = "Cook's Distance", ylab = "Cook's Distance", col = "blue")
```



```r
# Identify points with high influence based on Cook's Distance > 4/n
n <- length(cooks_distances)
high_influence_points <- which(cooks_distances > (4 / n))

cat("High influence points (Cook's Distance > 4/n):\n")
```

```
## High influence points (Cook’s Distance > 4/n):
```

```r
print(high_influence_points)
```

```
##    3   36  238  258  322  323  397  409  430  471  521  603  606  655  793  796
##    3   36  238  258  322  323  397  409  430  471  521  603  606  655  793  796
##  809  815  898  927  937  949 1054 1067 1078 1104 1145 1147 1148 1281 1282 1284
##  809  815  898  927  937  949 1054 1067 1078 1104 1145 1147 1148 1281 1282 1284
## 1306 1323 1389 1420 1452 1481 1482 1510 1511 1676 1678 1697 1714 1715 1717 1793
## 1306 1323 1389 1420 1452 1481 1482 1510 1511 1676 1678 1697 1714 1715 1717 1793
## 1836 1837 1911 1920 2016 2023 2041 2042 2106 2139 2142 2167 2185 2251 2257 2258
## 1836 1837 1911 1920 2016 2023 2041 2042 2106 2139 2142 2167 2185 2251 2257 2258
## 2259 2351 2352 2386 2423 2481 2526 2528 2564 2593 2665 2691 2721 2722 2727 2767
## 2259 2351 2352 2386 2423 2481 2526 2528 2564 2593 2665 2691 2721 2722 2727 2767
## 2779 2821 2833 2852 2868 2893 2908 2968 2970 2973 2975 2985 3054 3088 3096 3158
## 2779 2821 2833 2852 2868 2893 2908 2968 2970 2973 2975 2985 3054 3088 3096 3158
## 3165 3271 3368 3381 3382 3393 3405 3472 3494 3570 3571 3573 3574 3575 3602 3620
## 3165 3271 3368 3381 3382 3393 3405 3472 3494 3570 3571 3573 3574 3575 3602 3620
## 3653 3710 3736 3806 3822 3826 3830 3873 3955 3956 3963 4003 4123 4131 4150 4152
## 3653 3710 3736 3806 3822 3826 3830 3873 3955 3956 3963 4003 4123 4131 4150 4152
## 4160 4161 4231 4250 4287 4340 4349 4361 4369 4375 4394 4428 4429 4438 4442 4443
## 4160 4161 4231 4250 4287 4340 4349 4361 4369 4375 4394 4428 4429 4438 4442 4443
## 4444 4522 4523 4525 4638 4663 4664 4686 4734 4771 4785 4786 4788 4835 4849 4857
## 4444 4522 4523 4525 4638 4663 4664 4686 4734 4771 4785 4786 4788 4835 4849 4857
## 4867 4899 4910 4921 4951 4992 4998 5012 5050 5148 5149 5181 5209 5210 5211 5215
## 4867 4899 4910 4921 4951 4992 4998 5012 5050 5148 5149 5181 5209 5210 5211 5215
## 5228 5240 5247 5272 5289 5376 5388 5396 5502 5519 5532 5586 5587 5598 5615 5624
## 5228 5240 5247 5272 5289 5376 5388 5396 5502 5519 5532 5586 5587 5598 5615 5624
## 5657 5692 5696 5702 5732 5758 5765 5802 5804 5896 5898 5901 5921 5926 5944 6045
## 5657 5692 5696 5702 5732 5758 5765 5802 5804 5896 5898 5901 5921 5926 5944 6045
## 6124 6158 6227 6246 6292 6325 6364 6402 6464 6535 6537 6546 6600 6601 6630 6676
## 6124 6158 6227 6246 6292 6325 6364 6402 6464 6535 6537 6546 6600 6601 6630 6676
## 6722 6723 6834 6859 6871 6872 6879 6935 6950 6960 6991 7029 7030 7072 7089 7144
## 6722 6723 6834 6859 6871 6872 6879 6935 6950 6960 6991 7029 7030 7072 7089 7144
## 7212 7213 7222 7244 7384 7401 7406 7423 7450 7559 7565 7651 7652 7754 7763 7764
## 7212 7213 7222 7244 7384 7401 7406 7423 7450 7559 7565 7651 7652 7754 7763 7764
## 7775 7816 7897 7898 8016 8044 8046 8061 8084 8086 8123 8153 8191 8194 8256 8297
## 7775 7816 7897 7898 8016 8044 8046 8061 8084 8086 8123 8153 8191 8194 8256 8297
## 8312 8313 8326 8339 8340 8341 8355 8416 8428 8436 8447 8473 8537 8538 8560 8562
## 8312 8313 8326 8339 8340 8341 8355 8416 8428 8436 8447 8473 8537 8538 8560 8562
## 8563 8565 8572 8580 8581 8584 8589 8591 8620 8629 8688 8719 8837 8904 8922 8935
## 8563 8565 8572 8580 8581 8584 8589 8591 8620 8629 8688 8719 8837 8904 8922 8935
## 8946 8961 8963 8965 8967 8970 8972 8977 8981 8986 8988 8994 9001 9002 9014 9017
## 8946 8961 8963 8965 8967 8970 8972 8977 8981 8986 8988 8994 9001 9002 9014 9017
## 9019 9024 9034 9035 9037 9038 9052 9057 9064 9079 9081 9085 9090 9094 9095 9101
## 9019 9024 9034 9035 9037 9038 9052 9057 9064 9079 9081 9085 9090 9094 9095 9101
```

```
## 9107 9108 9110 9114 9119 9120 9122 9123 9137 9143 9147 9149 9153 9169 9170 9172
## 9107 9108 9110 9114 9119 9120 9122 9123 9137 9143 9147 9149 9153 9169 9170 9172
## 9190 9194 9197
## 9190 9194 9197
```

Final Model Diagnostics (Reintroducing work_setting)

Cook's Distance (Outlier and Influence Analysis):

- The Cook's Distance plot shows several data points with high influence, with some exceeding the typical threshold of 4/n, where n is the number of observations. These high-influence points may be affecting the stability of the model. Observations with large Cook's distances indicate data points that could have a disproportionate impact on the model's fit. Careful consideration should be given to these points, possibly through further analysis or sensitivity testing to assess their influence.

Histogram of Residuals (Normality Check):

- The histogram of residuals approximates a normal distribution, although there is some skewness and potential outliers on both ends of the distribution. This indicates that the residuals mostly follow a normal distribution, but the presence of outliers or slight asymmetry suggests that the model may not fully capture all complexities in the data. Minor deviations from normality are common in real-world data and may not significantly impact the model, but strong deviations could suggest the need for model adjustments or transformations.

Q-Q Plot (Normality of Residuals):

- The Q-Q plot shows that the residuals generally follow the 45-degree reference line, indicating that they are approximately normally distributed. Some deviation at the tails is visible, particularly at the extreme values. This suggests the presence of outliers, as observed in the Cook's Distance plot, and indicates that the model may not capture extreme observations accurately. These deviations are typically acceptable unless they significantly impact model interpretation or predictive performance. The Q-Q plot and residuals vs. fitted plot generally confirm model assumptions. However, slight deviations from normality suggest that salary data may retain some skewness. While these minor deviations do not invalidate the model, they indicate that further transformation or more complex models might marginally improve prediction accuracy.

Kolmogorov-Smirnov (KS) Test (Normality Test):

The KS test yielded a p-value of 8.251×10^{-13}, which is highly significant. This result indicates that the residuals do not perfectly follow a normal distribution. Although significant, this result is not unusual with large datasets, where even minor deviations can produce a significant p-value. Given the approximate normality observed in the Q-Q plot and histogram, this result may not be critically problematic but should be noted.

Breusch-Pagan Test (Homoscedasticity): The Breusch-Pagan test for homoscedasticity returned a p-value less than 2.2×10^{-16}, suggesting significant heteroscedasticity (non-constant variance) in the residuals. This indicates that the variance of residuals changes across levels of the predictors, potentially affecting the reliability of inference from the model. To address this, we might consider a transformation, robust standard errors, or different modeling approaches (e.g., generalized least squares).

Variance Inflation Factor (VIF Multicollinearity Check):

- The VIF values for the predictors are all below 1.2, which is well within acceptable limits, indicating low multicollinearity among the predictors. Low multicollinearity suggests that the model's coefficients are stable and interpretable, with each predictor contributing unique information. This supports the robustness of the model's estimates.

Conclusion

Diagnostic checks were performed to validate model assumptions, including homoscedasticity, normality of residuals, and multicollinearity. The residuals vs. fitted plot largely supports the homoscedasticity assumption, showing a random pattern around zero. However, the Q-Q plot reveals slight deviations from the expected straight line, suggesting a mild skew in the residuals. Although this deviation is minimal, it may indicate that salary data retains some skewness due to the presence of high-income values. While these minor deviations do not compromise the model's overall validity, they suggest that further transformations or alternative modeling approaches, such as log transformation for other predictors, could enhance predictive precision.

Final Model Tunning

```r
# Ensure that data_encoded_Final exists as the final encoded matrix
data_encoded_Final <- model.matrix(salary_in_usd ~ experience_level + employment_type + company_size +
                                     job_category + work_setting, data = data_cleaned)[, -1]

# Convert it to a data frame and add the response variable
data_encoded_df <- as.data.frame(data_encoded_Final)
data_encoded_df$salary_in_usd <- data_cleaned$salary_in_usd

# Fit Robust Linear Model on Original (Untransformed) Salary
robust_model <- lm(salary_in_usd ~ ., data = data_encoded_df)

# Apply robust standard errors to the model
print("Robust Standard Errors for Original Model:")
```

```
## [1] "Robust Standard Errors for Original Model:"
```

```r
coeftest(robust_model, vcov = vcovHC(robust_model, type = "HC3"))
```

```
##
## t test of coefficients:
##
##                                            Estimate Std. Error t value
## (Intercept)                                29082.17   17281.21  1.6829
## experience_levelExecutive                  77951.49    4083.44 19.0896
## `experience_levelMid-level`                19556.64    2449.39  7.9843
## experience_levelSenior                     55300.77    2314.55 23.8927
## employment_typeFreelance                  -41029.28   20066.05 -2.0447
## `employment_typeFull-time`                 18147.65   16549.20  1.0966
## `employment_typePart-time`                   953.84   21225.55  0.0449
## company_sizeM                               3319.65    2248.09  1.4767
## company_sizeS                             -32742.54    4705.68 -6.9581
## `job_categoryCloud and Database`           15866.38   14842.87  1.0690
## `job_categoryData Analysis`               -18393.37    2803.46 -6.5610
```

```
## `job_categoryData Architecture and Modeling`  15446.85    4189.20  3.6873
## `job_categoryData Engineering`               11766.39    2813.76  4.1817
## `job_categoryData Management and Strategy`  -16057.94    6556.05 -2.4493
## `job_categoryData Quality and Operations`   -28063.91    6843.70 -4.1007
## `job_categoryData Science and Research`      27262.25    2773.24  9.8305
## `job_categoryLeadership and Management`       7022.65    3374.60  2.0810
## `job_categoryMachine Learning and AI`        40672.03    3010.21 13.5114
## `work_settingIn-person`                      39185.44    3877.45 10.1060
## work_settingRemote                           33947.31    3899.77  8.7050
##                                              Pr(>|t|)
## (Intercept)                                  0.092433 .
## experience_levelExecutive                   < 2.2e-16 ***
## `experience_levelMid-level`                 1.584e-15 ***
## experience_levelSenior                      < 2.2e-16 ***
## employment_typeFreelance                     0.040912 *
## `employment_typeFull-time`                   0.272851
## `employment_typePart-time`                   0.964157
## company_sizeM                                0.139803
## company_sizeS                               3.685e-12 ***
## `job_categoryCloud and Database`             0.285117
## `job_categoryData Analysis`                 5.635e-11 ***
## `job_categoryData Architecture and Modeling`  0.000228 ***
## `job_categoryData Engineering`              2.920e-05 ***
## `job_categoryData Management and Strategy`   0.014331 *
## `job_categoryData Quality and Operations`   4.155e-05 ***
## `job_categoryData Science and Research`     < 2.2e-16 ***
## `job_categoryLeadership and Management`      0.037459 *
## `job_categoryMachine Learning and AI`       < 2.2e-16 ***
## `work_settingIn-person`                     < 2.2e-16 ***
## work_settingRemote                          < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Apply Log Transformation to Salary and Fit Lasso Model
y_log <- log(data_cleaned$salary_in_usd)

# Lasso regression with cross-validation on log-transformed salary
lasso_model_log <- cv.glmnet(data_encoded_Final, y_log, alpha = 1)
lasso_best_lambda_log <- lasso_model_log$lambda.min
lasso_pred_log <- predict(lasso_model_log, s = lasso_best_lambda_log, newx = data_encoded_Final)

# Calculate R-squared for the log-transformed model
lasso_r2_log <- 1 - sum((y_log - lasso_pred_log)^2) / sum((y_log - mean(y_log))^2)

# Display coefficients and R-squared for log-transformed Lasso model
print("Lasso Model Coefficients with Log-Transformed Salary:")
```

```
## [1] "Lasso Model Coefficients with Log-Transformed Salary:"
```

```r
print(coef(lasso_model_log, s = lasso_best_lambda_log))
```

```
## 20 x 1 sparse Matrix of class "dgCMatrix"
##                                                     s1
```

```
## (Intercept)                                10.66266852
## experience_levelExecutive                   0.66825600
## experience_levelMid-level                   0.23630124
## experience_levelSenior                      0.53072558
## employment_typeFreelance                   -0.44845772
## employment_typeFull-time                    0.23444027
## employment_typePart-time                   -0.16136095
## company_sizeM                               0.05875206
## company_sizeS                              -0.30438856
## job_categoryCloud and Database              0.14241294
## job_categoryData Analysis                  -0.16474558
## job_categoryData Architecture and Modeling  0.10099055
## job_categoryData Engineering                0.07025642
## job_categoryData Management and Strategy   -0.13744805
## job_categoryData Quality and Operations    -0.29532726
## job_categoryData Science and Research       0.17321103
## job_categoryLeadership and Management       0.04213249
## job_categoryMachine Learning and AI         0.25118997
## work_settingIn-person                       0.35233062
## work_settingRemote                          0.31127485
```

```r
print(paste("R-squared for Lasso Log Model:", lasso_r2_log))
```

```
## [1] "R-squared for Lasso Log Model: 0.301623016307574"
```

```r
# Refit robust linear model with original (untransformed) salary and robust SE
robust_model <- lm(salary_in_usd ~ ., data = data_encoded_df)
print("Robust Standard Errors for Final Model:")
```

```
## [1] "Robust Standard Errors for Final Model:"
```

```r
coeftest(robust_model, vcov = vcovHC(robust_model, type = "HC3"))
```

```
##
## t test of coefficients:
##
##                                                Estimate Std. Error t value
## (Intercept)                                    29082.17   17281.21  1.6829
## experience_levelExecutive                      77951.49    4083.44 19.0896
## `experience_levelMid-level`                    19556.64    2449.39  7.9843
## experience_levelSenior                         55300.77    2314.55 23.8927
## employment_typeFreelance                      -41029.28   20066.05 -2.0447
## `employment_typeFull-time`                     18147.65   16549.20  1.0966
## `employment_typePart-time`                       953.84   21225.55  0.0449
## company_sizeM                                   3319.65    2248.09  1.4767
## company_sizeS                                 -32742.54    4705.68 -6.9581
## `job_categoryCloud and Database`               15866.38   14842.87  1.0690
## `job_categoryData Analysis`                   -18393.37    2803.46 -6.5610
## `job_categoryData Architecture and Modeling`  15446.85    4189.20  3.6873
## `job_categoryData Engineering`                 11766.39    2813.76  4.1817
## `job_categoryData Management and Strategy`    -16057.94    6556.05 -2.4493
## `job_categoryData Quality and Operations`     -28063.91    6843.70 -4.1007
```

```
## `job_categoryData Science and Research`       27262.25    2773.24  9.8305
## `job_categoryLeadership and Management`        7022.65    3374.60  2.0810
## `job_categoryMachine Learning and AI`         40672.03    3010.21 13.5114
## `work_settingIn-person`                       39185.44    3877.45 10.1060
## work_settingRemote                            33947.31    3899.77  8.7050
##                                               Pr(>|t|)
## (Intercept)                                   0.092433 .
## experience_levelExecutive                    < 2.2e-16 ***
## `experience_levelMid-level`                  1.584e-15 ***
## experience_levelSenior                       < 2.2e-16 ***
## employment_typeFreelance                      0.040912 *
## `employment_typeFull-time`                    0.272851
## `employment_typePart-time`                    0.964157
## company_sizeM                                 0.139803
## company_sizeS                                3.685e-12 ***
## `job_categoryCloud and Database`              0.285117
## `job_categoryData Analysis`                  5.635e-11 ***
## `job_categoryData Architecture and Modeling`  0.000228 ***
## `job_categoryData Engineering`               2.920e-05 ***
## `job_categoryData Management and Strategy`    0.014331 *
## `job_categoryData Quality and Operations`    4.155e-05 ***
## `job_categoryData Science and Research`      < 2.2e-16 ***
## `job_categoryLeadership and Management`       0.037459 *
## `job_categoryMachine Learning and AI`        < 2.2e-16 ***
## `work_settingIn-person`                      < 2.2e-16 ***
## work_settingRemote                           < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# Ensure that your robust model is fit as per your provided code
robust_model <- lm(salary_in_usd ~ ., data = data_encoded_df)

# Get robust standard errors and tidy up the results for easier viewing
robust_summary <- coeftest(robust_model, vcov = vcovHC(robust_model, type = "HC3"))
robust_tidy <- tidy(robust_summary)

# Filter to keep only significant predictors (p-value < 0.05)
significant_predictors <- robust_tidy %>%
  filter(p.value < 0.05)

# Display the summary of significant predictors
print("Significant Predictors with Coefficients and P-Values:")
```

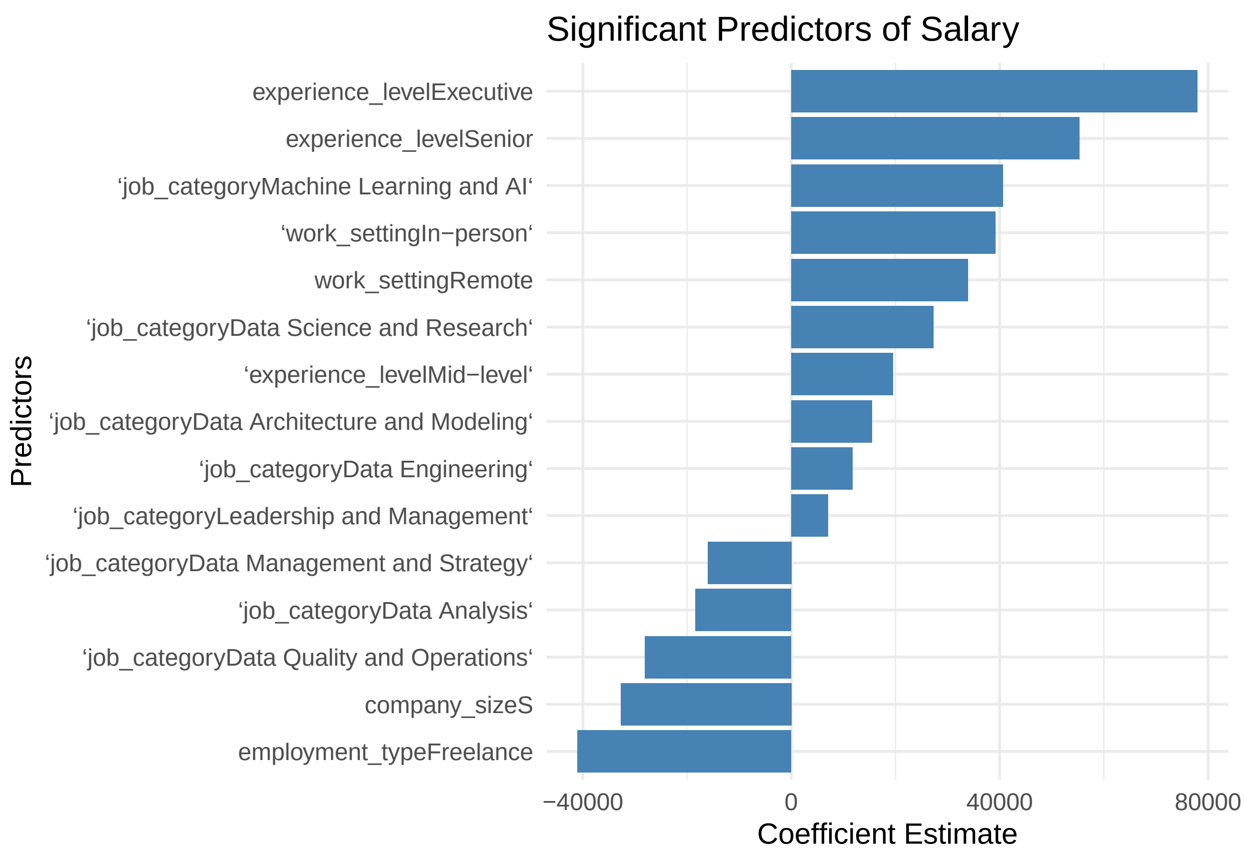
```
## [1] "Significant Predictors with Coefficients and P-Values:"
```

```r
print(significant_predictors)
```

```
## # A tibble: 15 x 5
##    term                                   estimate std.error statistic   p.value
##    <chr>                                     <dbl>     <dbl>     <dbl>     <dbl>
##  1 experience_levelExecutive                77951.     4083.     19.1  1.07e- 79
##  2 `experience_levelMid-level`              19557.     2449.      7.98 1.58e- 15
##  3 experience_levelSenior                   55301.     2315.     23.9  1.90e-122
```

```
## 4 employment_typeFreelance                   -41029.    20066.     -2.04 4.09e-  2
## 5 company_sizeS                              -32743.     4706.     -6.96 3.69e- 12
## 6 `job_categoryData Analysis`                -18393.     2803.     -6.56 5.64e- 11
## 7 `job_categoryData Architecture and Mo~      15447.     4189.      3.69 2.28e-  4
## 8 `job_categoryData Engineering`             11766.     2814.      4.18 2.92e-  5
## 9 `job_categoryData Management and Stra~     -16058.     6556.     -2.45 1.43e-  2
## 10 `job_categoryData Quality and Operati~    -28064.     6844.     -4.10 4.15e-  5
## 11 `job_categoryData Science and Researc~     27262.     2773.      9.83 1.08e- 22
## 12 `job_categoryLeadership and Managemen~      7023.     3375.      2.08 3.75e-  2
## 13 `job_categoryMachine Learning and AI`      40672.     3010.     13.5  3.32e- 41
## 14 `work_settingIn-person`                    39185.     3877.     10.1  6.93e- 24
## 15 work_settingRemote                         33947.     3900.      8.70 3.73e- 18
```

```r
ggplot(significant_predictors, aes(x = reorder(term, estimate), y = estimate)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Significant Predictors of Salary",
    x = "Predictors",
    y = "Coefficient Estimate"
  ) +
  theme_minimal()
```



Final Model Tunning Summary

Robust Model with Original Salary

Model Coefficients:

- Experience Level: Executive, Senior, and Mid-level experience levels are significant predictors of salary, with executives having the highest positive effect (USD 77,951) on salary. These predictors are highly significant, indicating that experience level is a crucial factor in determining salary.
- Employment Type: Freelance employment has a significant negative impact on salary (USD -41,029), suggesting that freelancers earn considerably less than the baseline employment type. Full-time and Part-time are not statistically significant, implying they do not have a distinct impact on salary in this model.
- Company Size: Small companies (company_sizeS) have a significant negative effect on salary (USD -32,742), indicating that smaller companies tend to offer lower salaries.
- Job Category: Several job categories show significant effects:
- Positive Impacts: Data Architecture and Modeling (USD 15,446), Data Engineering (USD 11,766), Data Science and Research (USD 27,262), and Machine Learning and AI (USD 40,672) are associated with higher salaries.
- Negative Impacts: Data Analysis (USD -18,393), Data Management and Strategy (USD -16,057), and Data Quality and Operations (USD -28,063) show negative impacts, indicating lower salaries in these roles.
- Work Setting: Both In-person (USD 39,185) and Remote (USD 33,947) work settings show significant positive effects on salary, indicating that these roles tend to offer higher compensation, possibly due to the flexibility premium.

Lasso Model with Log-Transformed Salary

Model Coefficients:

- Experience Level: Executive, Senior, and Mid-level have positive coefficients, with Executive level having the highest impact on log-transformed salary (0.668). This result aligns with the robust model, indicating that higher experience levels are associated with higher salaries.
- Employment Type: Freelance employment has a negative coefficient (-0.448), reinforcing the finding that freelancers tend to earn less. Full-time and Part-time have smaller impacts.
- Company Size: Small companies have a negative effect (-0.304), suggesting they generally offer lower salaries.
- Job Category: Similar to the robust model, job categories like Machine Learning and AI (0.251) and Data Science and Research (0.173) are positively associated with higher log-transformed salaries, while Data Analysis (-0.165) and Data Quality and Operations (-0.295) have negative effects.
- Work Setting: In-person (0.352) and Remote (0.311) work settings have positive impacts on log-transformed salary, suggesting that both settings are associated with higher compensation.

Model Diagnostics:

- The Lasso model with log-transformed salary has an R^2 of approximately 0.302, suggesting that about 30.2% of the variability in log-transformed salary is explained by the model. This transformation helps mitigate the influence of outliers and skewness in salary data. Additionally, the model selects predictors and shrinks less important ones, providing a simpler model and addressing potential multicollinearity. The use of log-transformed salary also reduces the impact of heteroscedasticity.

Conclusion

Both models highlight similar trends; Experience Level and Work Setting consistently have strong positive effects on salary. Freelance Employment Type and Small Company Size are associated with lower salaries.

Job Categories reflect significant differences in salary, with roles in Machine Learning, Data Science, and Data Engineering offering higher pay, while operational roles (e.g., Data Quality) offer lower pay. The robust model provides detailed salary estimates with robust standard errors, making it useful for precise interpretation. The Lasso model with log-transformation provides a simplified, regularized view, reducing potential multicollinearity and improving generalizability. Based on these results, the robust model offers interpretability and insights into salary levels, while the Lasso model with log transformation offers model simplicity and generalizability. Both models are valuable for understanding salary determinants across different experience levels, job categories, and work settings. However, the log-transformation enhances both the statistical performance of the model and the interpretability of the results, making it our recommended approach in modeling salary data.

Report Summary

The analysis focused on identifying key drivers of salary within data-related professions, leveraging multiple regression techniques and regularization methods to develop predictive models. The report journey included rigorous data preprocessing, exploratory data analysis (EDA), and systematic model tuning, resulting in the selection of a final model based on Lasso regression.

Analysis Insights

- Key Predictors: The most impactful predictors of salary were identified as experience level, employment type, company size, job category, and work setting. Specifically, executive experience, technical job categories such as Machine Learning and Data Science, and full-time employment were associated with higher salaries, while freelance and operational roles like Data Quality and Data Analysis were linked to lower compensation.
- Model Selection: Various models were evaluated, including Stepwise Regression, Lasso, and Ridge Regression. The Lasso model was selected as the final model due to its superior performance in balancing model fit and complexity, reflected by the lowest AIC and BIC values and its effective feature selection capability. While the Stepwise model explained a marginally higher proportion of salary variance, it was more prone to overfitting, as suggested by its higher AIC and BIC scores.
- Model Diagnostics: During model refinement, interaction terms between experience_level, employment_type, and job_category were explored to assess potential non-linear relationships and synergies between these variables. For example, higher experience levels combined with specific employment types, such as freelance or part-time, could influence salary differently than when considered individually. However, upon testing, these interaction terms did not significantly improve the model's predictive power and added unnecessary complexity. Thus, interaction terms were ultimately excluded from the final model to maintain simplicity and interpretability without sacrificing accuracy.
- Interpretability and Practical Relevance: The final model provides practical insights into the salary structure within data jobs, highlighting factors that professionals and companies might prioritize. For instance, the significant salary premiums associated with remote work settings reflect industry trends favoring flexible work arrangements.

Further Considerations

While the report offers a thorough examination of salary predictors within data-related professions, several additional aspects could be explored in future analyses.

Recomendations

- Inclusion of Additional Predictors: Incorporating external economic indicators, geographic location specifics, and skill certifications could further improve the model's explanatory power. These variables often influence salary but were beyond this report's scope due to dataset limitations.

- Temporal Analysis: Salaries in data-related fields can fluctuate due to industry trends and economic cycles. A temporal analysis considering how these relationships evolve over time, using time-series techniques or cohort-based segmentation, could provide more dynamic insights.

- Interaction Effects: Interaction terms were explored to assess non-linear relationships between experience_level, employment_type, and job_category. However, they were ultimately excluded as they did not significantly enhance the predictive power of the model and added complexity without substantial benefits.

- Handling Outliers: To improve model stability, we identified and removed outliers using the IQR method, focusing on extreme salary values. Outliers, especially at the higher end of the salary range, can disproportionately influence regression models, often inflating coefficients and leading to skewed results. For instance, exceptionally high salaries may reflect rare or specialized roles that are not representative of the general data trends. By excluding these outliers, we achieved a more stable model that better represents the majority of cases in the dataset. This approach enhances model robustness, ensuring that predictions are less susceptible to extreme variations and thus more generalizable.

- Cross-Validation Methods: While cross-validation was employed to tune Lasso and Ridge models, exploring additional validation methods, such as k-fold cross-validation or bootstrapping, could provide further robustness to model selection.

- Possibilities for further analysis: This analysis provides valuable insights into factors influencing salary within the data industry. However, further work could expand the analysis by incorporating additional predictors, such as industry-specific factors or geographical location, which may also impact salary. Additionally, non-linear models, such as decision trees or random forests, could be explored to capture more complex relationships that may not be fully addressed by linear regression. These future enhancements would provide a more comprehensive view of salary determinants and further improve predictive accuracy.

End.