STAT 209 Final Project

Trung Nguyen, Dat Le

1. Introduction

Our research aims to understand how various company attributes impact job satisfaction, utilizing company ratings as an indicator for employee satisfaction. This study is especially relevant to us as we are nearing graduation and are keen on identifying employers who not only excel in fostering a positive work environment but also support the professional and personal growth of their employees. By understanding the relationship between company attributes and employee satisfaction, we aim to better equip ourselves and our peers with the knowledge needed to make informed decisions about future employment opportunities.

Recent studies have illustrated the significance of factors like industry type, company size, remote work availability, and compensation structures on job satisfaction. For example, research by [1] Tansel and Gazîoğlu (2013) indicates job satisfaction levels are lower in large firms. Similarly, [2] Sharma and Gupta (2020) research highlights that the healthcare, education and travel and tourism exhibited high levels of job satisfaction, while insurance, outsourcing and logistics industries figured low on this parameter. Additionally, [3] Makridis and Schloetzer (2022) found working from home more frequently tends to confer higher satisfaction.

2. Data

We constructed our dataset by merging two separate Kaggle datasets: 'LinkedIn Job Postings' and 'Company Reviews and Ratings'. The former was collected over a span of two days, several months apart, and consists of over 33,000 job postings from LinkedIn. The latter was compiled by scraping the AmbitionBox website, which provides 10,000 rows of reviews and ratings of companies.

Since the datasets lacked common columns, we combined them using company names as the key. This method led to numerous mismatches in the combined dataset, many of which we manually corrected (further details on the data limitations are provided in the appendix). The final dataset comprises 337 rows, each representing a unique company, across 14 columns.

Within our variables, we are using:

- Rating (Continuous, 0-5): Represents the overall company rating which serves as an indicator for job satisfaction.
- Employee Count (Continuous): Number of employees, reflecting company size.
- Remote Allowed (Categorical, Yes or No): Indicates whether remote work is permissible.
- Country (Categorical, US or non-US): The location of the company.
- Average Salary (Continuous): Mean salary offered by the company.
- Average Number of Benefits (Continuous): Reflects the range of employee benefits.
- Company Age (Continuous): The time since the company was founded.

And these variables we are not going to use:

- Company id (Continuous): Unique id for every company
- Name (Categorical): Company name
- Industry (Categorical): The sector in which the company operates.
- State (Categorical): More specific location within the US.
- City (Categorical): City where the company is based.
- Company Type (Categorical): Such as public, Forbes Global 2000, or Fortune India 500.
- Number of Reviews (Continuous): Total number of reviews recorded

Although state, city, industry, and company type are aspects of a company that could influence job satisfaction, they have too many unique values. Due to our lack of time and expertise, we were unable to group these categories effectively, so we decided to exclude them from our analysis.

3. Analytic Framework

Our research focuses on determining how various elements of a company affect job satisfaction, using company ratings as an indicator for job satisfaction. The response variable (Y) in our study is the company rating, a continuous variable ranging from 0 to 5 (real numbers), which represents overall job satisfaction derived from employee feedback and other criteria.

The predictor variables (X) are those that we mentioned above, which include both continuous and categorical data:

- Number of employees (`employee_count`)
- Remote work availability (`remote_allowed`)
- Country (`country`)
- Average salary (`avg_salary`)
- Average number of benefits (`avg_benefit_count`)
- Company age (`company_age`)

Our primary supervised learning strategy will begin with an Ordinary Least Squares (OLS) regression, followed by a Lasso model. Initially, we'll use OLS regression to assess the strength of the relationship between each predictor and the outcome. Subsequently, we'll apply the Lasso model, which is effective for feature selection and regularization to prevent overfitting. By comparing the Root Mean Squared Error (RMSE) of both models, we'll determine which model performs better and select it for more detailed analysis.

We are also considering unsupervised learning methods like clustering to explore whether companies can be grouped into distinct categories based on features influencing job satisfaction. This approach can uncover patterns not immediately apparent through regression analysis alone, such as identifying clusters of companies that are similar in terms of employee satisfaction drivers but differ in other factors.

4. Results

4.1. Exploratory data analysis

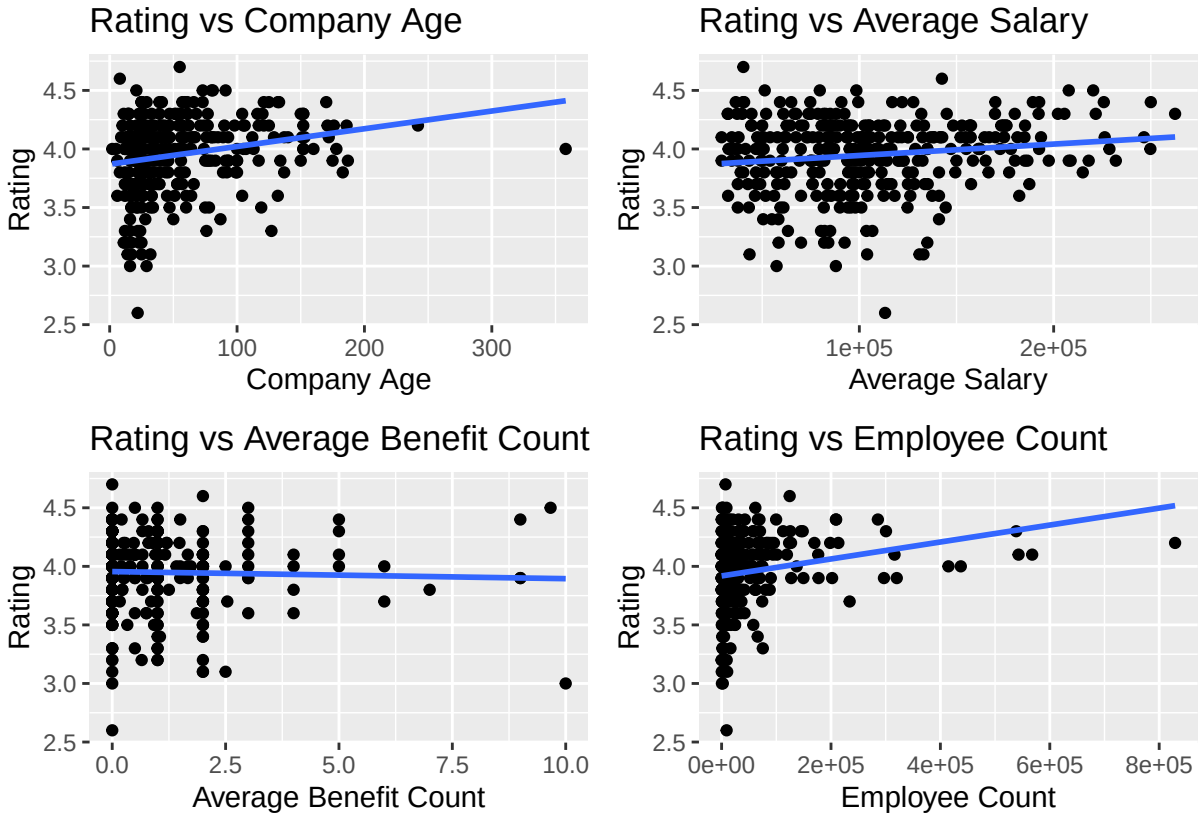
First, we will plot the distribution of each variable in our dataset. Given that some of our distributions are heavily skewed, we have chosen the median as the measure of central tendency to represent the center of each distribution. (The red line in the plot indicates the median of the distribution)

We can see that the distribution of `rating` is roughly normally distributed with a median of 4 and a standard deviation of 0.32. The distribution is slightly skewed to the left.

Look at these plots, the first thing is apparent is that the distribution of `company_age`, `employee_count`, and `avg_benefit_count` are heavily skewed to the right, which is due to the presence of a few companies with very high values in these variables. The distribution of `avg_salary` is also skewed to the right but to a lesser degree.

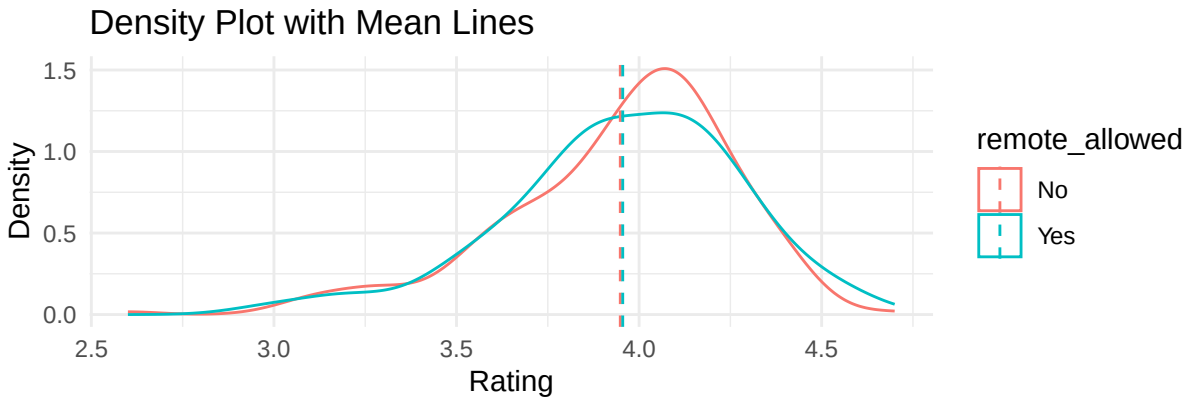
For the categorical variables, we can see for `remote_allowed`, there is a significant number of companies that do not allow remote work, with only a smaller portion that do. And for `country`, we can see the majority of companies are located in the US, as indicated by the taller bar.

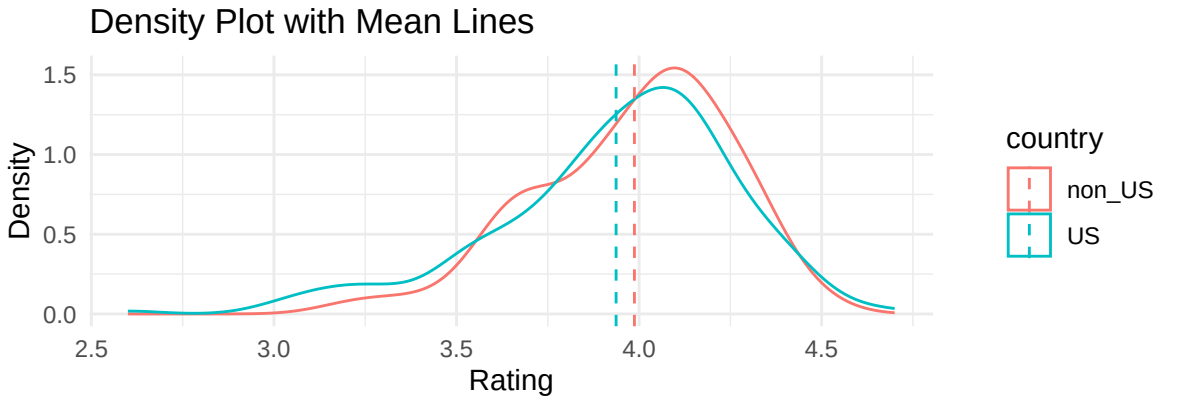
Next, let's look at the relationships between the response variables and continuous predictors.



Our analysis reveals positive correlations between the `rating` variable and several other factors: `company_age`, `avg_salary`, and `employee_count`. There's also a weak negative relationship between `rating` and `avg_benefit_count`, which appears to be influenced by an outlier. Removing this outlier might shift this relationship to a positive one, though it's likely to remain weak as the correlation line nearly levels out.

Next, let's look at the categorical variables `remote_allowed` and `country` and see if there are any relationships between them and `rating`.





The data shows no significant differences in the ratings of companies that allow remote work compared to those that do not. However, there is a noticeable variation in the ratings of companies based on their location. Specifically, companies in the US tend to have slightly lower ratings than those outside the US.

4.2. Linear & Lasso models

Let's fit our data into models to further investigate the quantitative relationships between the variables.

Since our dataset is relatively small, we decided to use 25% of the data for testing and the rest for training. We will use 5-fold cross-validation for the training data.

First let's fit a linear regression model to the data and see how it performs.

```
## # A tibble: 7 x 5
##   term                  estimate   std.error statistic   p.value
##   <chr>                    <dbl>       <dbl>     <dbl>     <dbl>
## 1 (Intercept)        3.72        0.0640         58.0   3.49e-145
## 2 company_age        0.00166     0.000420        3.95  1.03e-  4
## 3 avg_salary         0.00000126  0.000000411     3.07  2.42e-  3
## 4 avg_benefit_count -0.00565     0.0118         -0.478 6.33e-  1
## 5 employee_count     0.000000746 0.000000220     3.38  8.34e-  4
## 6 remote_allowedYes  0.00680     0.0491          0.139 8.90e-  1
## 7 countryUS         -0.0341      0.0456         -0.748 4.55e-  1
```

Setting the significance level at 0.05, we can see that `company_age`, `avg_salary`, `employee_count`, and `country` are significant predictors of `rating`, while `avg_benefit_count` and `remote_allowed` are not since their p-values are greater than 0.05.

Next, let's fit a lasso regression model to the data and see how it performs. We chose the penalty of 0.01411217 since this value yields the lowest cross-validation RMSE (see more detail in appendix)

```
## # A tibble: 7 x 3
##   term                  estimate penalty
##   <chr>                    <dbl>   <dbl>
## 1 (Intercept)        3.74         0.0141
## 2 company_age        0.00140      0.0141
## 3 avg_salary         0.000000963  0.0141
## 4 avg_benefit_count  0            0.0141
## 5 employee_count     0.000000611  0.0141
## 6 remote_allowedYes  0            0.0141
## 7 countryUS         -0.00512      0.0141
```

Lasso regression model seems to have selected `company_age`, `avg_salary`, `employee_count`, and `country` as significant predictors of `rating` as the coefficients of `avg_benefit_count` and `remote_allowed` are set to zero. This is consistent with the linear regression model.

Calculation of the 5-fold cross-validation RMSE of the linear regression and lasso regression yields

- Linear: 0.3083812
- Lasso: 0.3066943

The RMSE of the lasso regression model is slightly lower than that of the linear regression model, which means that the lasso regression model is slightly better at predicting the ratings of companies. Therefore, let's use the lasso regression model for our final analysis.

First, the RMSE of the Lasso model on our test data is 0.275391. This is smaller than the standard deviation of the rating, which is 0.32. This means that our model is able to predict the rating of companies with a reasonable degree of accuracy.

Examining the coefficients from the Lasso regression model, we can see how each variable qualitatively influences the rating of a company, controlling for other variables:

- For each 10 years increase in company age, the rating of the company increases by about 0.01
- For each $10,000 increase in average salary, the rating of the company increases by 0.01
- For each 10000 increase in employee count, the rating of the company increases by 0.006
- Companies in the US tend to have a lower rating compared to companies in other countries by about 0.004

The qualitative analysis of the Lasso regression model's coefficients indicates that the variables examined do not significantly impact company ratings in a practical sense due to the minimal changes. However, they do establish some relationships:

- The positive coefficients for `company age`, `average salary`, and `employee count` confirm their positive relationships with company ratings, as suggested by earlier visualizations. Notably, `average salary` appears to be the most significant predictor among them, aligning with the expectation that higher salaries correlate with better employee satisfaction and higher company ratings. This is also supported by Tansel and Gazîoğlu that average salary is a significant positive predictor of company ratings. However, the positive coefficient for employee count in our model contrasts with the IZA paper's findings of lower satisfaction in larger firms. This discrepancy might be due to different measures or contexts (e.g., country-specific factors).
- The coefficient for `company location in the US` is negative, suggesting that US companies have slightly lower ratings compared to those in other countries. This finding contradicts the expectation of higher ratings due to better salaries and benefits typically associated with US companies, but the effect is relatively small.
- Variables like `avg_benefit_count` and `remote_allowed` showed no significant predictive power for company ratings, contrary to expectations. The zero coefficients for these variables indicate they do not influence company ratings under the current model settings. Makridis's paper does not really address benefits directly, but it emphasizes that overall workplace environment and compensation are more critical. However, the paper highlights the complex relationship between remote work and job satisfaction. Remote work does not significantly improve satisfaction and increases the intention to leave. This suggests that merely allowing remote work is not sufficient; the quality of the remote work environment and other factors are more influential.

Overall, while the Lasso model predicts company ratings with reasonable accuracy, the strength of the relationships between the predictors and ratings is weak. This could stem from the small dataset size, data skewness, or the experimental nature of our data collection and wrangling methods. More comprehensive data might reveal stronger relationships. (More details about the limitation in the appendix)

4.3. Clustering

To further investigate to see if there are other patterns in our data, we decided to use clustering analysis. We decided to use three significant continuous variables from previous model, which are `company age`, `average salary`, and `employee count`. Using an elbow plot, we decided to split the data into 5 clusters since this is where the Total Within-Cluster Sum of Squares starts to level off. (see more details in the appendix)

The results are as follows:

```
## # A tibble: 5 x 6
##   cluster     n company_age avg_salary employee_count rating
##   <fct>   <int>       <dbl>      <dbl>          <dbl>  <dbl>
## 1 1          11        55      132500.        440903.   4.12
## 2 2         119        39.3     74181.         35359.   4.10
## 3 5          72        39.6    175758.         28719.   4.09
## 4 4          50       138.      99099.         46684.   4.08
## 5 3          85        33.2     90580.         14723.   3.53
```

We decided to characterize each cluster by the most prominent characteristic. Ordered by their ratings

- Cluster 3: Established Giants. This cluster represents very large companies with an extremely high number of employees, a mature age of about 55 years, moderately high average salaries
- Cluster 1: Low-salary Firms. This cluster is characterized by its average salary being the lowest among all clusters. Companies in this cluster have moderate age and size
- Cluster 5: High-salary Firms. This cluster is characterized by its average salary being the highest among all clusters. Companies in this cluster also have moderate age and size
- Cluster 4: Seasoned Firms. Companies here are much older on average. They offer moderate salaries and have above average employee counts compared to some other clusters
- Cluster 2: Small-scale Startups. This cluster consists of the youngest companies with the smallest employee counts. Its average salary is moderate

Small-scale Startups stand out with notably lower ratings than other clusters despite its moderate salary. Established Giants exhibit the highest ratings, and High-salary Firms and Seasoned Firms also enjoy comparably high ratings. All of which aligns the earlier linear regression results that older, larger companies offering higher salaries tend to be rated more favorably than their counterparts.

However, Low-salary Firms present an exception to this pattern. They have the lowest average salary, and both their company age and employee count are moderate compared to other clusters, yet it maintains the second highest rating. This suggests that factors other than average salary, company age, and employee count, which we have not yet considered, also significantly influence company ratings.

5. Implications

One notable implication of this analysis is that older, larger, more well-paid companies tend to have higher ratings than their counterparts. This result may not be a surprise since old and large companies tend to be more experienced firms and salary is a significant motivation of workers. Moreover, companies in the US have slightly lower ratings than non-US companies. However, we have to keep in mind that these effects are relatively small in the practical settings, some of which are negligible. Additionally, it seems that whether the company is remote or not and the amount of benefits offered by the company are not significant factors contributing to its rating compared to other variables.

Finally, our cluster analysis points out that there are other factors than average salary, company age, and employee count, which we have not yet considered, also significantly influence company ratings (some examples can be the industry of the company or whether the company is profit or non-profit)

Future research should consider more factors of the company that can contribute to its rating and use a larger dataset. This can pose a challenge since internal information of companies is often not publicly available.

Citation

[1] Tansel, A., & Gazîoğlu, Ş. (2014). Management-employee relations, firm size and job satisfaction. International journal of manpower, 35(8), 1260-1275.

[2] Sharma, S. C., & Gupta, R. (2020). Job satisfaction: difference in levels among selected industries. Int J Rec Technol Eng (IJRTE), 8(6).

[3] Makridis, C., & Schloetzer, J. D. (2022). Does working from home increase job satisfaction and retention? Evidence from the COVID-19 pandemic. Evidence from the COVID-19 Pandemic (March 6, 2022). Georgetown McDonough School of Business Research Paper, (4016657).

Appendix

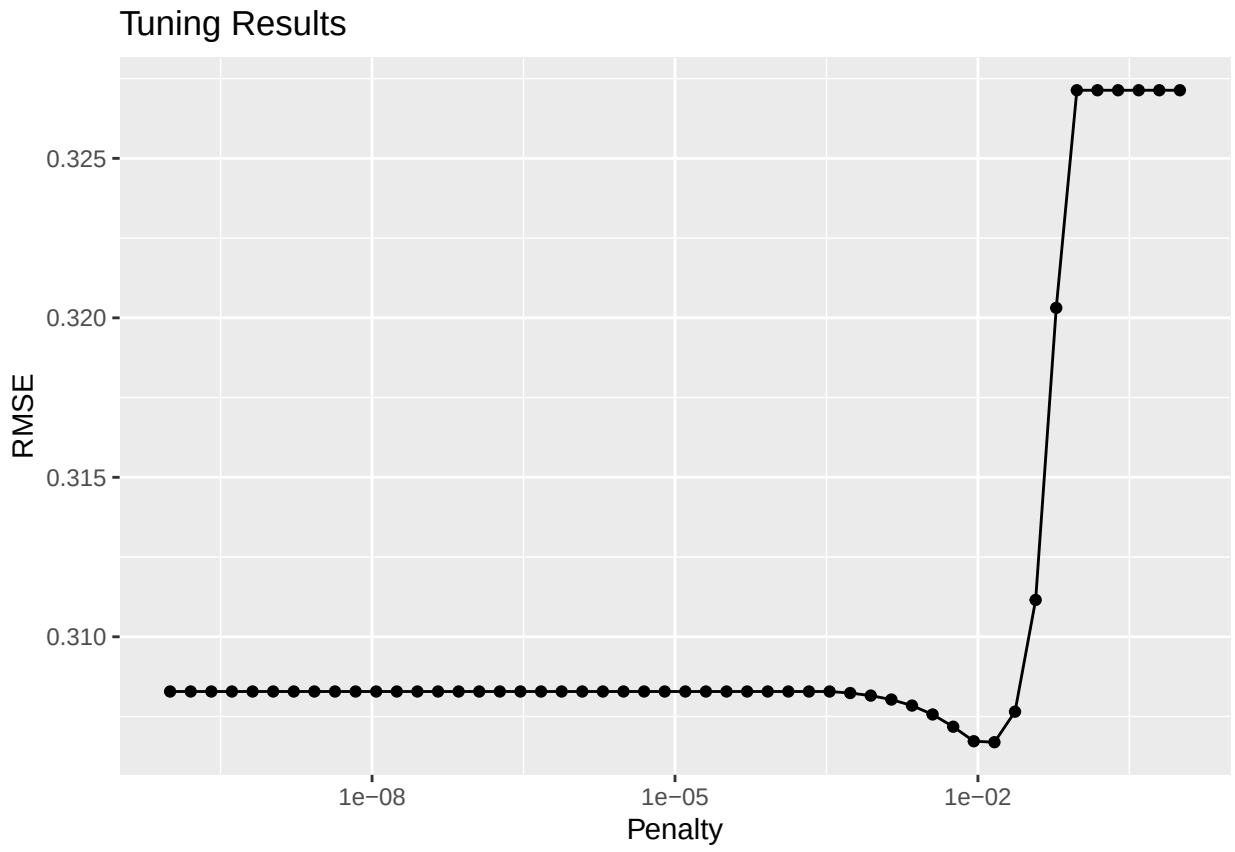
1. Data quality issues

The two dataset we chose are quite different in their nature and the type of information they provide. One dataset contains job listings from LinkedIn, while the other contains from internal employees of companies. The difference of these datasets can introduce inconsistencies and inaccuracies when merging them. Also, due to the lack of common columns for a straightforward merge, we had to rely on company names to combine the datasets. This process led to numerous mismatches that need manual correction. However, this manual correction process is prone to human error, which can introduce further inaccuracies into the dataset.

The final dataset consists of only 337 rows, each representing a unique company across 14 columns. The small size of the dataset limits the statistical power of our analyses and may not provide a comprehensive representation of the entire job market.

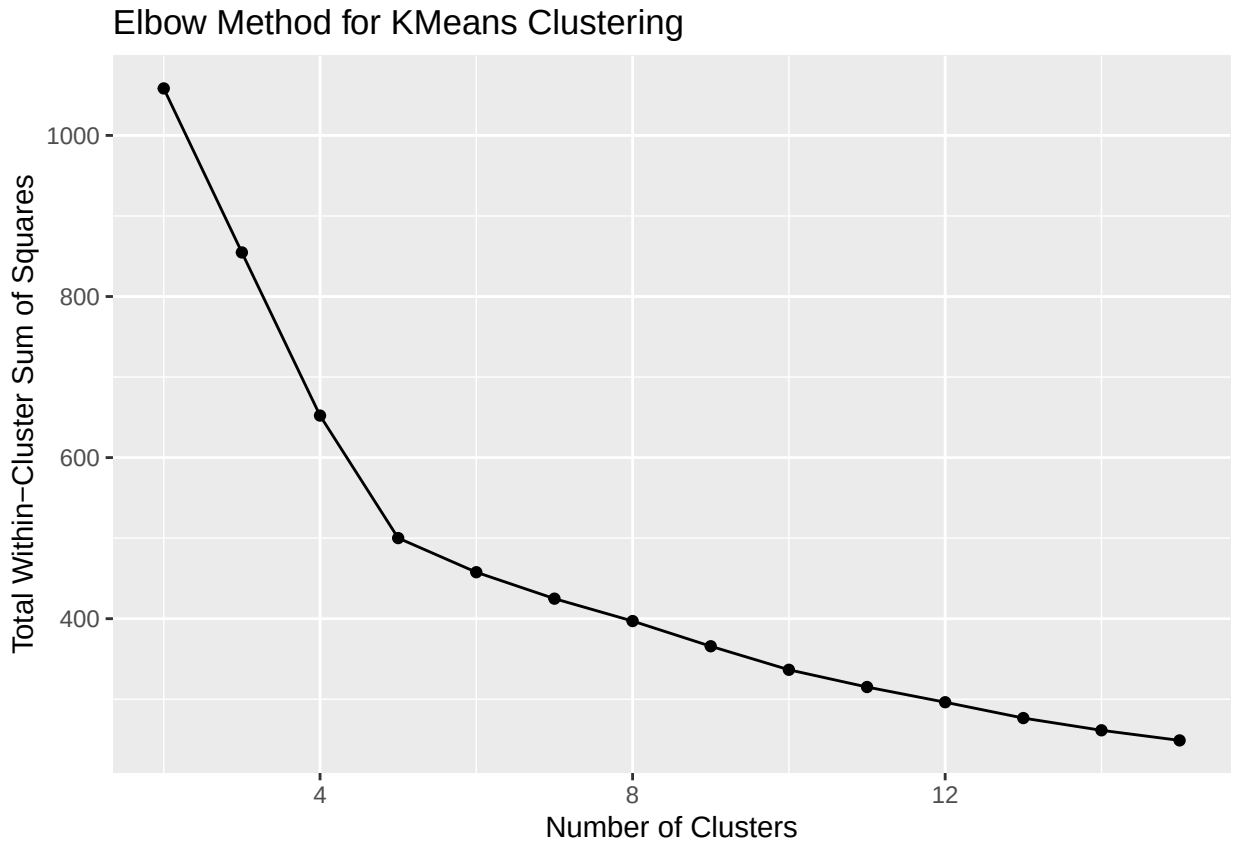
2. Choosing penalty for the Lasso model

We used the `tune_grid` function from the `tidymodels` package to find the optimal penalty for the Lasso model. We used a grid of 50 penalty values and performed 5-fold cross-validation to find the penalty that yields the lowest RMSE. The penalty value of 0.01411217 was chosen since it resulted in the lowest RMSE. This value was used to fit the final Lasso model.



3. Elbow method for KMeans clustering

We used the elbow method to determine the optimal number of clusters for the KMeans clustering analysis. The elbow method involves plotting the total within-cluster sum of squares against the number of clusters and selecting the number of clusters where the plot starts to level off. In our case, the plot started to level off at 5 clusters, so we decided to split the data into 5 clusters for further analysis.



We have adhered to the Honor Code. Trung. Dat.