Final Report

Elaine Dai

2024-04-30

1. Introduction

In the rapidly evolving job market, the role of a data science related job has emerged as a pivotal position within companies. As a student approaching the completion of data science degree, I am preparing to step into the professional world, and get interested in understanding the landscape of data science related positions. Therefore, this project aims to dig into the characteristics of data analyst related job postings, with a specific focus on uncovering the markers that signal high-salary opportunities. By digging into this, we can gain insights into the attributes and qualifications that employers value most for.

This project seeks to answer the question: "What characteristics of data science related job postings are indicative of high-salary positions?". The dataset utilized in this analysis was acquired from Kaggle, specifically from the "LinkedIn Job Postings" dataset by user Arsh Kon here. This dataset comprises a wide range of United States job postings on LinkedIn in 2023, including various fields such as job title, company, location, salary, and more.

Some columns that are deemed relevant to the topic are:

- job_id: Job ID as defined by LinkedIn
- company_id: Company ID as Defined by LinkedIn
- title: Job Title
- description: Job Description
- max_salary: Maximum Salary
- med_salary: Median Salary
- min_salary: Minimum Salary
- pay_period: Pay Period
- formatted_work_type: Work Type
- location: Job Location
- formatted_experience_level: Experience Level Required

2. Methods

First I loaded libraries and the data set.

2.1 Data Cleaning and Wrangling

Upon acquiring the dataset, I examined the structure of the data: the types of each column and any missing values. The target attribute is the salary, so this preliminary examination is focused mainly on salary. The examination revealed that salary information was provided either in a range format (minimum and maximum salary) or a median value (med_salary), and not all entries had a valid salary. To address this, a new column, salary, was created by averaging the min_salary and max_salary values for each entry where med_salary was missing. 2087(the average number of work hours in a year) is multiplied to the hourly based salary, in order to convert salary information from an hourly to a yearly basis for those entries listed with hourly rates. Another modification involved extracting the state information from the column location, for future analysis.

The further cleaning entailed filtering the dataset to retain only those job postings that are directly related to data roles. This was achieved by keeping rows where the job title included the term "data". After examining the summary statistics for salary, 4 outliers with unreasonably low annual salary have been removed. And the final stage is removing all job postings with na value and selecting specific columns that were deemed relevant for the analysis.

2.2 EDA

Summary Statistics: salary I extracted the summary statistics for salary and formulated the table using the kable. To further explore the categorical variables within the dataset, I implemented a visualization function using the ggplot2 library. The function was designed to produce two types of visual representations for each categorical variable: a bar plot and a box plot.

Exploratory Graphs: word type, state, experience level The function first groups the data by the categorical variable provided, summarises the data to get counts of each category within the variable, and then creates a barplot using ggplot2 with categories on the y-axis (due to coord_flip()) and their respective counts on the x-axis. Then a boxplot is created with the categorical variable on the y-axis (again, due to coord_flip()) and salary on the x-axis. The boxes are colored based on the categorical variable to differentiate between categories visually.

Text Mining: description Finally, an additional analysis was incorporated by utilizing text mining techniques to dissect job descriptions. The tidytext package was used to tokenize the text, allowing for the identification and visualization of the most frequent tokens. Common stopwords were removed, including universally frequent English words and additional terms like "years", "will", "work", "job" and "role".etc which are expected to be recurrent in job descriptions but offer little analytical value. Additionally, any tokens containing numbers were filtered out to focus purely on textual data. The text data was then tokenized into bigrams and trigrams to facilitate a granular analysis of phrase patterns within the job descriptions. Word/phrase count bar plots and a word cloud were created for visualization.

2.3 Model Building and Preparation

Dataset Preparation For preparing the data specifically for modeling purposes, I concentrated on our main interest, full-time job postings only. I also modified the "state" column to classify the states with too few observations into "other" to avoid splitting error. Initially, a unique identifier (doc_id) was assigned to each job posting to facilitate individual tracking through subsequent steps.

A significant portion of this phase involved processing the text data within job descriptions using TF-IDF (Term Frequency-Inverse Document Frequency). TF-IDF is a statistical measure used to evaluate the importance of a word to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus.

This method was chosen because it effectively highlights the most relevant words in job descriptions, which are likely indicative of the skills and responsibilities valued in higher-paying roles. Words typical across many job descriptions or irrelevant (such as common English stopwords and numbers) were filtered out to refine the analysis.

Following the computation of TF-IDF scores, these were aggregated for each document, then pivoted to create a wide-format dataset where each row represents a job posting and each column represents a word's TF-IDF score. This data was then merged back with the original dataset, preserving the essential variables such as salary, state, and experience level, and ensuring that all categorical variables were appropriately encoded as factors for analytical consistency.

The final dataset was split into training and test sets, with 70% of the data allocated for training to build the predictive models and 30% reserved for testing to evaluate model performance.

Modeling

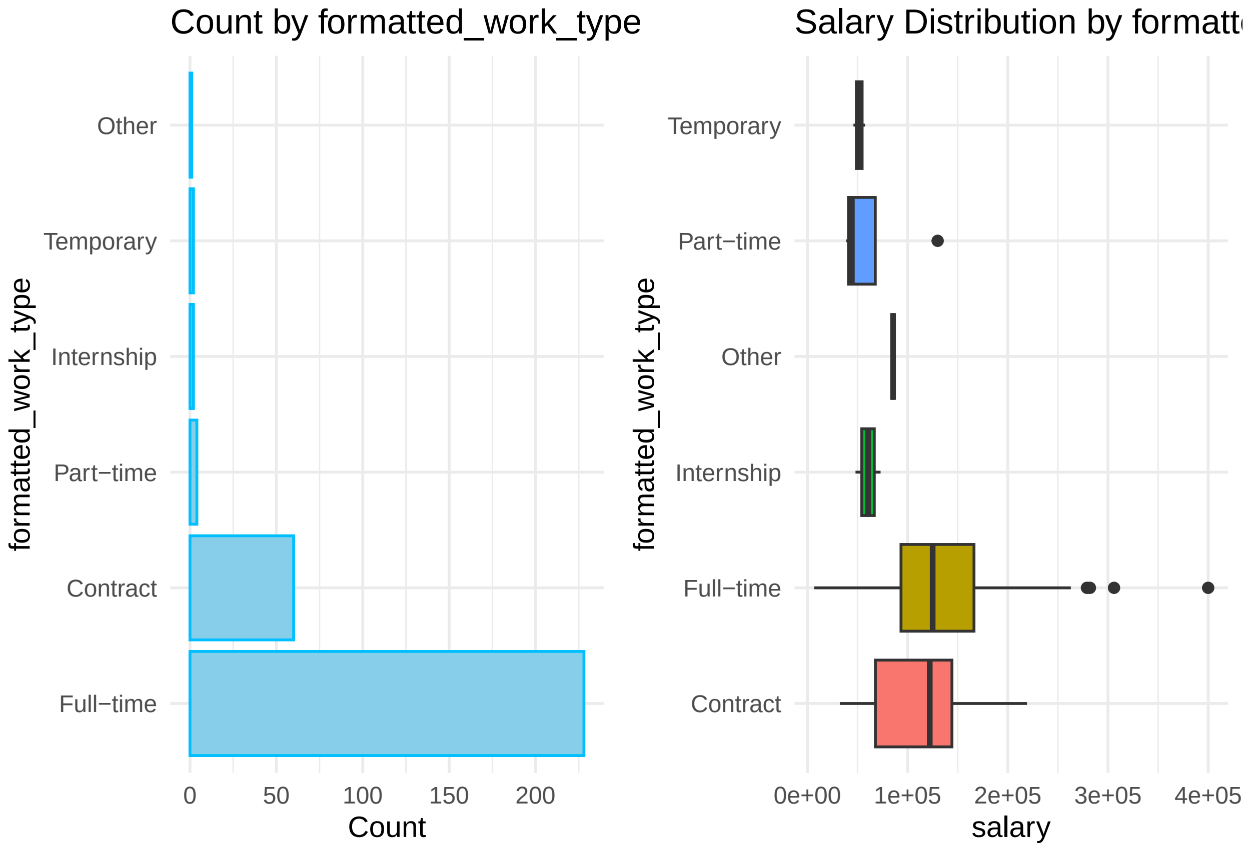
3. Results

3.1 EDA Results

After cleaning and wrangling, the cleaned dataset contains 297 observations and 6 variables with no missing values. The summary statistics for the numeric variable salary and the exploratory plots are given below.

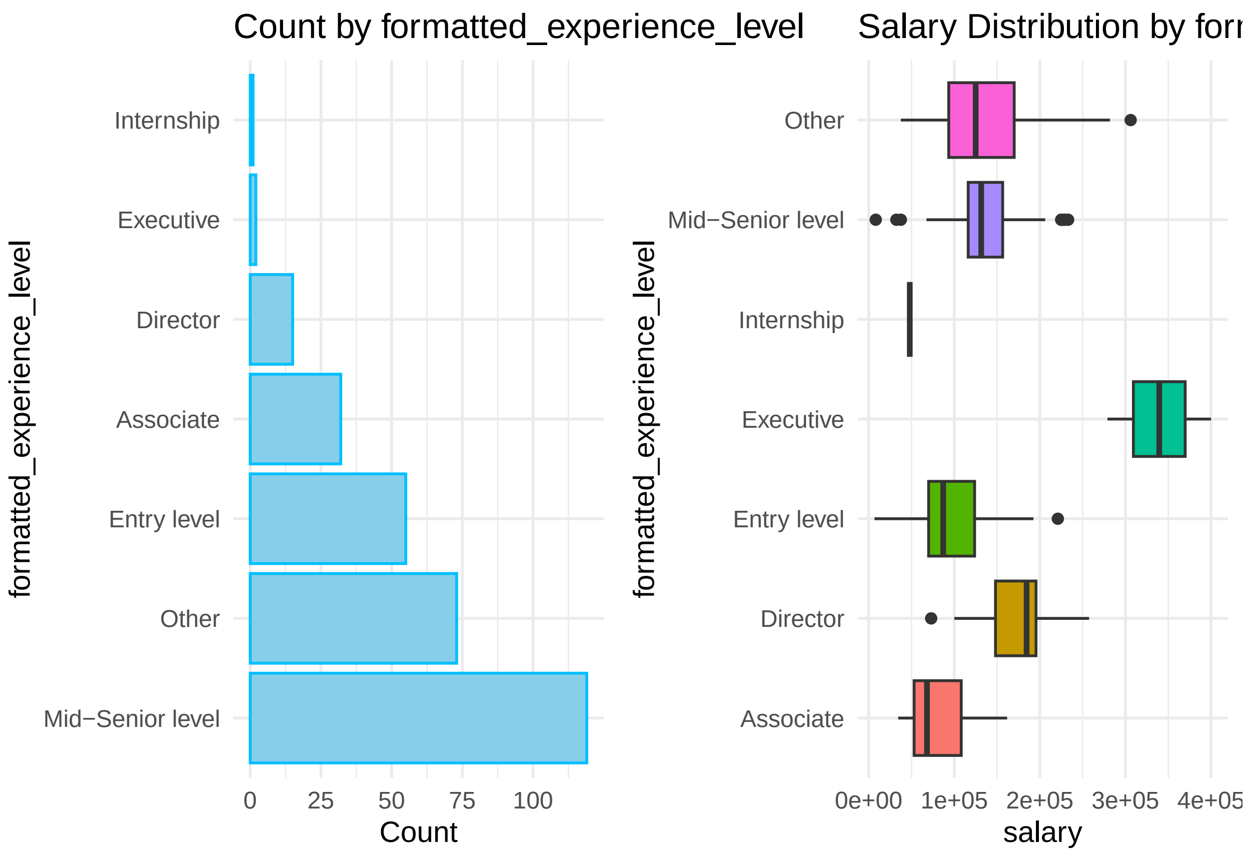
Statistic	Value
mean	124343.88
median	123133.00
sd	54841.42
min	6769.50
max	400000.00

In the work type category, full-time positions dominate the job market. Internships, part-time and temporary jobs are very less frequent. Salary-wise, full-time positions also lead with higher pay. Contract roles offer lower median salaries than full-time positions but are still well above temporary, part-time, and internship categories, which present the lowest pay. This is a predictable trend.

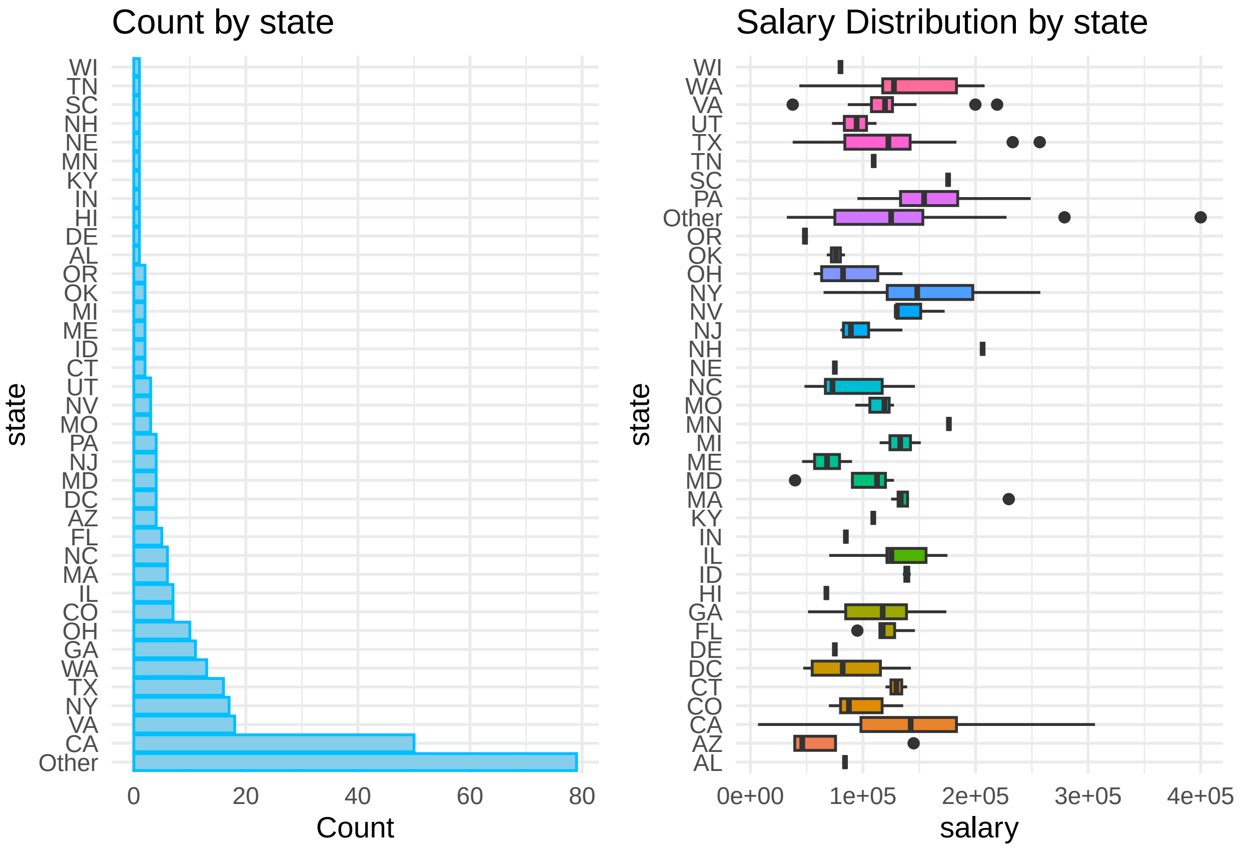


Regarding experience levels, mid-senior jobs are the most abundant, entry-level positions and associate follow in frequency. However, the high-ranking executive and director positions are rare, which reflects their specialized and leadership-focused nature.

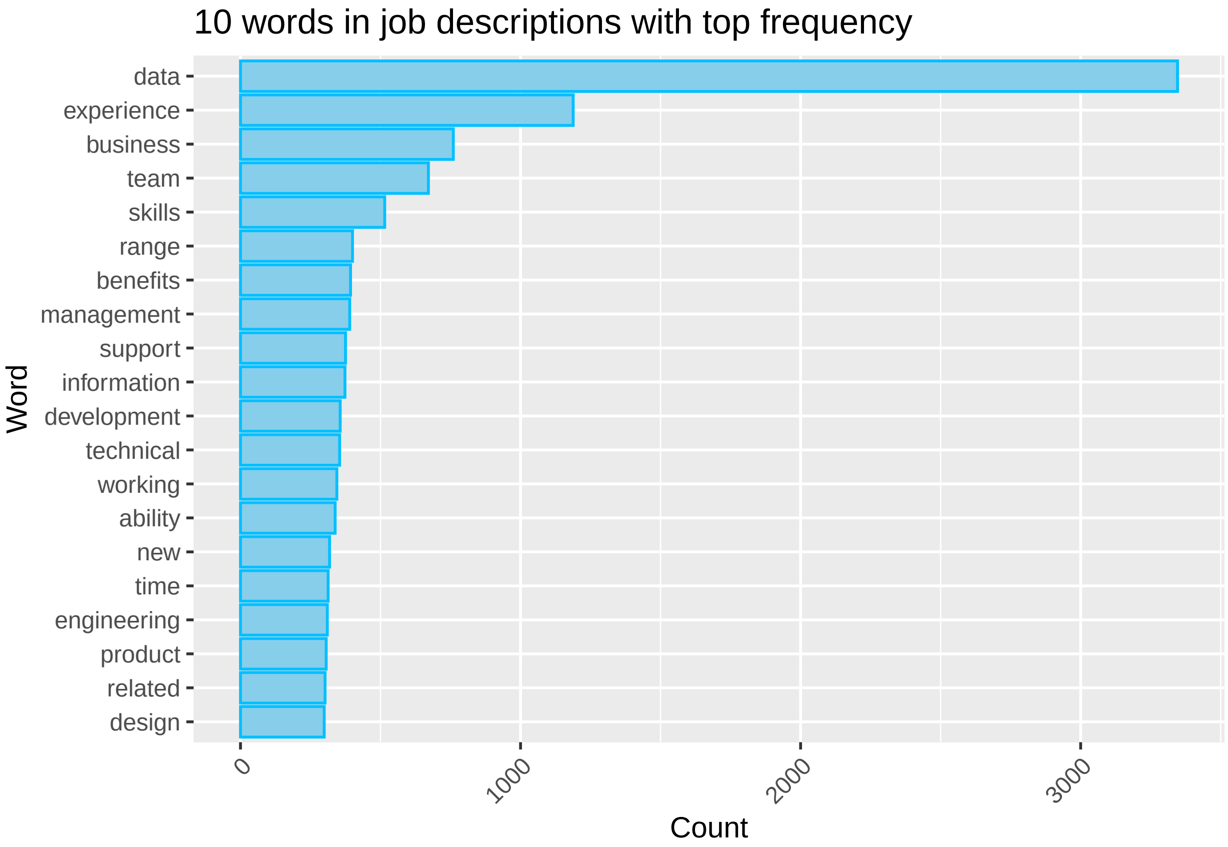
Salaries increase notably with experience. Top-tier roles like executives enjoy the highest salaries, and understandably, those at the entry-level earn the least. This gradient in pay is expected, aligning with the increased responsibilities and expertise required at higher levels.

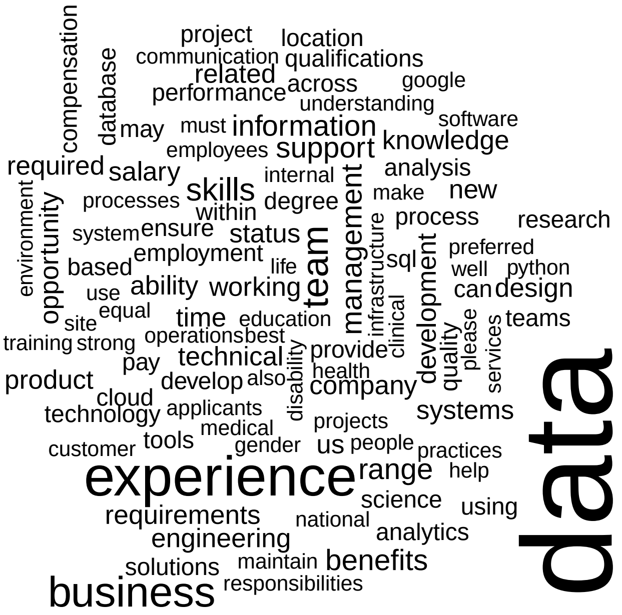


A state-wise look at job counts places California at the front, suggesting a bustling job market. New York and Texas also show significant job availability. Salaries by state reveal disparities that could be influenced by living costs or industry concentration, with places like California and New York showing higher median salaries. Other states exhibit a broad range of salaries, pointing to diverse economic landscapes and job sectors within each state.

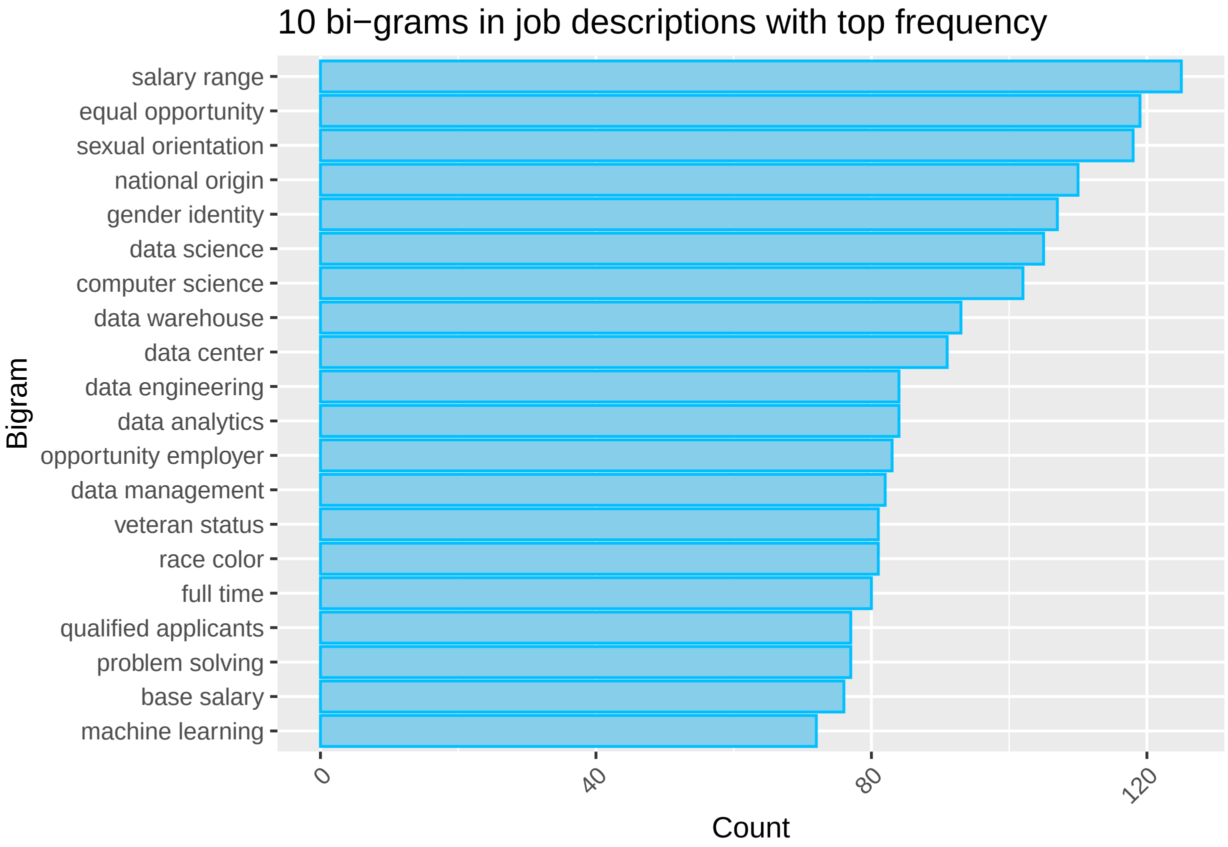


This bar graph displays the most frequently occurring individual words within job descriptions. The words with highest frequencies includes “data”, “experience”, “business”, “team”, “skills”, highlighting the general importance of these aspects in the professional environment.

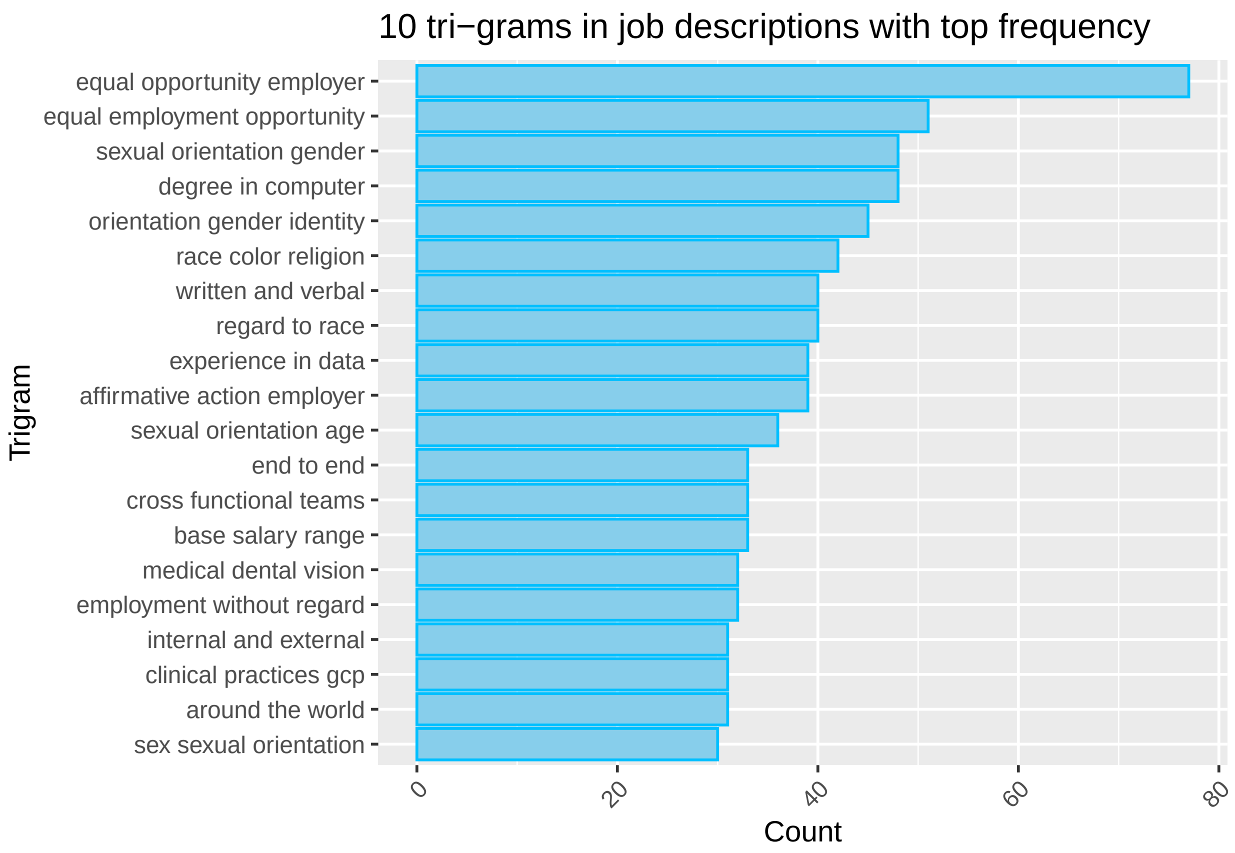

10 words in job descriptions with top frequency
Word
data
experience
business
team
skills
range
benefits
management
support
information
development
technical
working
ability
new
time
engineering
product
related
design
Count
0
1000
2000
3000



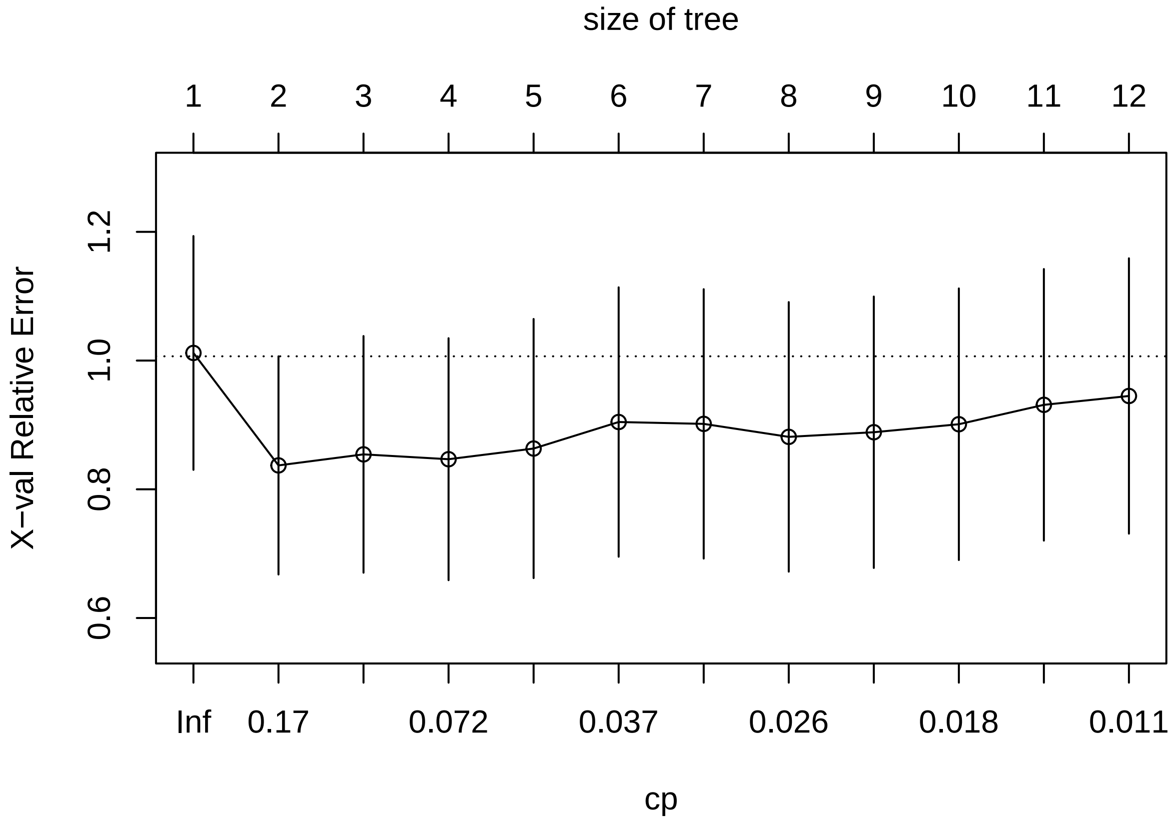
The bigram (two-word phrase) frequency graph sheds light on common pairings such as "machine learning", "base salary", "computer science" and "problem solving". These reflect specific skills, compensation expectations, and competencies valued in the job market. The relatively even distribution suggests no overwhelming focus on a particular phrase but instead a variety of important attributes and benefits. Noticeably gender and equality concepts are brought up frequently, indicating the significant emphasis on diversity and inclusion within the job market.

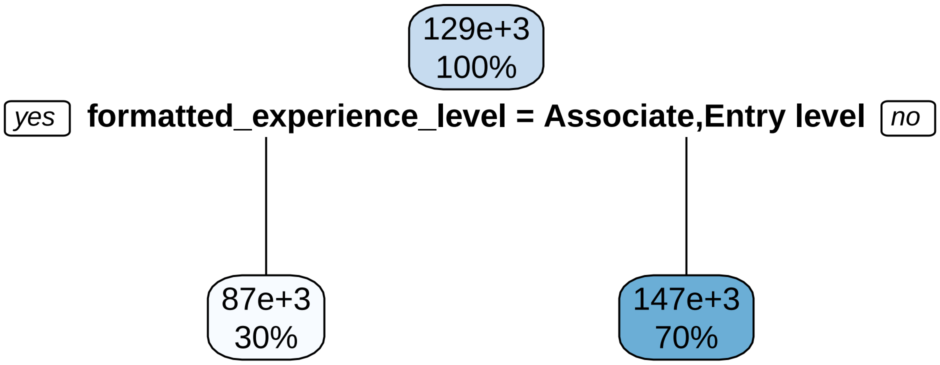


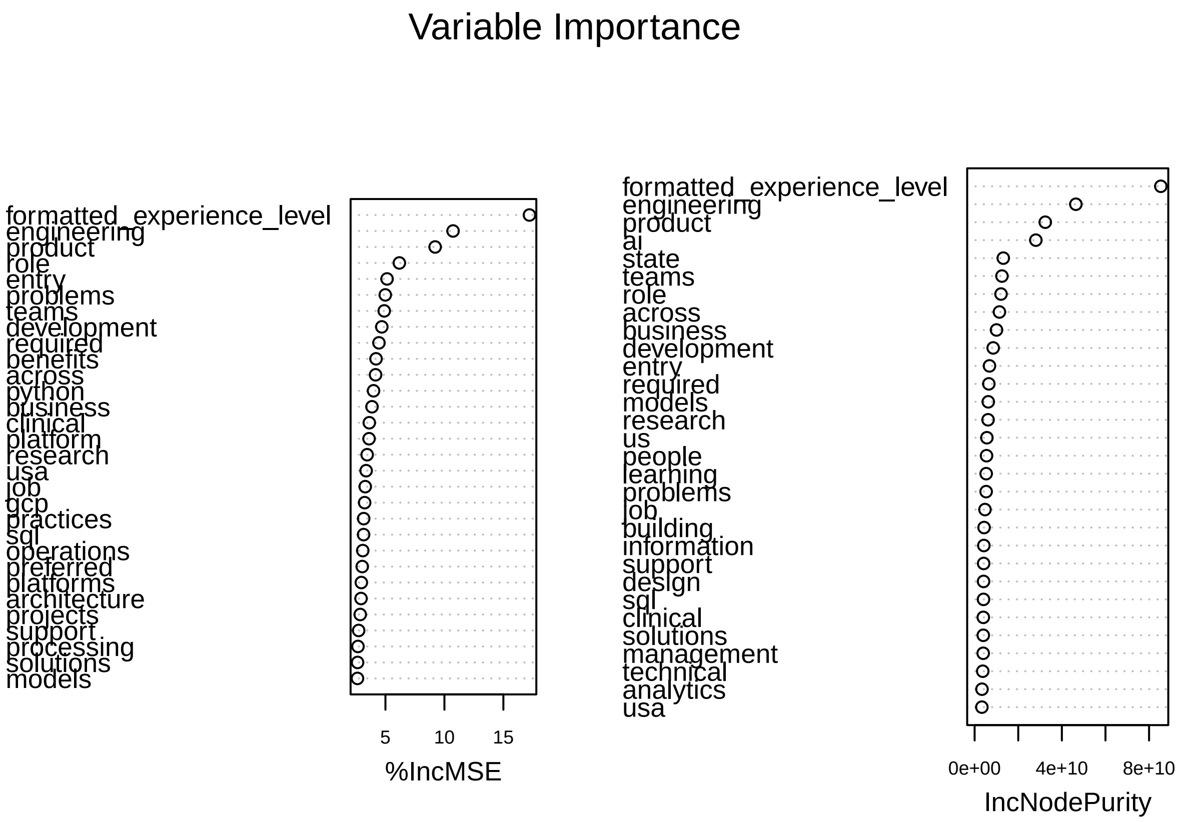
The trigram (three-word phrase) bar graph reveals frequent phrases with a more contextual understanding, such as ‘equal opportunity employer’, which denotes a commitment to workplace equality. The prevalence of phrases like ‘race color religion’, and ‘gender identity’ points to a focus on diversity and inclusion in hiring practices.


10 tri−grams in job descriptions with top frequency
Trigram
equal opportunity employer
equal employment opportunity
sexual orientation gender
degree in computer
orientation gender identity
race color religion
written and verbal
regard to race
experience in data
affirmative action employer
sexual orientation age
end to end
cross functional teams
base salary range
medical dental vision
employment without regard
internal and external
clinical practices gcp
around the world
sex sexual orientation
0
20
40
60
80
Count

3.2 Model Results




129e+3
100%
yes
formatted_experience_level = Associate,Entry level
no
87e+3
30%
147e+3
70%



```r
calculate_mse <- function(model, data) {
  predictions <- predict(model, newdata = data)
  mean((data$Salary - predictions)^2)
  # data.frame(s = data$Salary, p = predictions)
}
```

```r
calculate_mse(tree_pruned, test)
```

```r
mse_lm <- calculate_mse(lm_model, test)
mse_tree_pruned <- calculate_mse(tree_pruned, test)
mse_bagging <- calculate_mse(bagging_model, test)
mse_rf <- calculate_mse(rf_model, test)
mse_boosting <- calculate_mse(boosting_model, test, n.trees = 1000)
mse_xgboost <- calculate_mse(xgboost_model, test)

mse_comparison <- data.frame(
  Method = c("Linear Regression", "Regression Tree", "Bagging", "Random Forest", "Boosting", "XGBoost")
  MSE = c(mse_lm, mse_tree_pruned, mse_bagging, mse_rf, mse_boosting, mse_xgboost)
)

mse_comparison
```

```r
tf_idf_numeric <- select(filtered_tf_idf_wide, -doc_id)

# Apply PCA
```

```r
pca_result <- prcomp(tf_idf_numeric, center = TRUE, scale. = TRUE)

# View summary of PCA results
summary(pca_result)
```

```r
loadings <- pca_result$rotation
biplot(pca_result)
```

```r
k <- 36  # for example, change this based on your scree plot and summary
tf_idf_pca <- pca_result$x[, 1:k]

# Convert to data frame, if you need to use it further in data processing
tf_idf_pca_df <- as.data.frame(tf_idf_pca)
model_data_with_pca <- cbind(full_time_data, tf_idf_pca_df) |> select(-description, -doc_id)
```

```r
lm_pca <- lm(salary ~ ., data = train)
```

Summary

This project aimed to discover which aspects of data science job postings correlate with higher salaries. The exploration of work type, location, and experience level revealed distinct variations in average salaries across different groups. Full-time positions, certain states, and advanced experience levels generally command higher pay. Natural Language Processing (NLP) applied to job descriptions also helped by identifying frequent keywords, bigrams, and trigrams. Dominant terms like 'data', 'experience', and 'business', along with phrases emphasizing 'machine learning' and 'problem-solving', were prominent. Additionally, gender, race and equality concepts were recurrent terms, signifying an ethical emphasis on hiring.