TULKINOV SHAVKATJON 李克涵
Student ID: 502023145004

EXPLORING TRENDS AND INSIGHTS IN DATA SCIENCE JOB POSTINGS: A GLASSDOOR ANALYSIS

#1. Introduction: The field of data science has experienced exponential growth in recent years, with an increasing demand for skilled professionals who possess a combination of analytical, statistical, and programming expertise. As organizations strive to harness the power of data to drive informed decision-making, the role of data scientists has become essential. Glassdoor, a popular job search and review website, serves as a valuable platform for employers to post job openings and for job seekers to explore potential opportunities. This dataset, "Data Science Job Postings on Glassdoor" obtained from Kaggle, provides a comprehensive collection of job postings in the data science domain. In this project, we aim to delve into this dataset to uncover trends, insights, and patterns pertaining to data science job postings, enabling us to gain a deeper understanding of the current landscape of the field.

##Reason for choosing particular dataset. I chose to focus on data science job postings on Glassdoor for several reasons. Firstly, Glassdoor is a popular and widely-used platform that provides valuable insights into job opportunities and company information. By analyzing data science job postings on Glassdoor, I can gain insights into the current demand for data scientists, the skills and qualifications sought by employers, and the overall trends in the data science job market. Additionally, Glassdoor allows employees and job seekers to share their experiences and provide reviews about companies. This can give me a better understanding of the work culture, employee satisfaction, and other factors that may influence my decision-making process. Furthermore, Glassdoor provides a rich dataset that includes various attributes such as job titles, company sizes, salaries, and reviews. By analyzing this data, I can uncover patterns and correlations that can help me make informed decisions about my career path, such as identifying the most common job titles, exploring the relationship between salaries and company sizes, or understanding the distribution of job types and seniority levels.

#2. Problem Statement: The exponential growth of the data science field has led to an increased number of job postings across various industries. However, navigating through the vast array of job opportunities can be overwhelming for job seekers, and employers need insights into the competitive landscape to attract top talent. As a student specializing in information resource management and data analytics, my goal is to establish a successful career as a data scientist. Therefore, the problem we aim to address in this project is to analyze the dataset of data science job postings on Glassdoor and derive valuable insights to help job seekers (such as myself and my classmates) informed decisions and assist employers in understanding the dynamics of the job market.

#3. Research Questions:

1. What are the most common job titles in the data science field?
2. How do the salary ranges vary for different job titles and industries in the data science field?
3. Which industries and locations have the highest number of job postings in the data science field?
4. Is there a correlation between company ratings and job postings, indicating whether highly-rated companies offer better salaries?
5. What are the most in-demand skills for data science jobs based on the skill analysis?
6. How does company size and revenue relate to the number of job opportunities in the data science field?
7. What is the distribution of job types and seniority levels in the data science field, providing insights into the composition of data science positions?

By addressing these research questions, we aim to provide valuable insights and actionable recommendations for both job seekers and employers in the data science field.

#4. Data Analysis and Visualization Plan.

In order to address research questions above, these plans/steps will be taken:

- **Distribution of Job Titles:** Visualize the distribution of different job titles in the data science field to identify the most common positions.
- **Salary Analysis:** Explore the salary estimation column to analyze the salary ranges for different job titles and industries. To create visualizations such as box plots, histograms, or bar charts to compare salaries across various factors.
- **Industry and Location Analysis:** Analyze the industries and locations of the companies posting data science jobs. Create bar charts to identify the industries and locations with the highest number of job postings.
- **Company Ratings and Job Postings:** Investigate the relationship between company ratings and job postings. Analyze whether highly-rated companies tend to offer better salaries or have more job opportunities.
- **Skill Analysis:** Explore the skills required for data science jobs, such as Python, Excel, Hadoop, Spark, AWS, Tableau, and Big Data. Visualize the frequency of these skills using bar charts to identify the most in-demand skills.
- **Company Size and Revenue Analysis:** Analyze the relationship between company size, revenue, and job postings. Visualize the distribution of company sizes and revenues in relation to job opportunities.
- **Job Type and Seniority:** Examine the job type (e.g., data scientist, data analyst) and seniority (senior or not) to understand the composition of data science positions. Create visualizations such as pie charts or stacked bar charts to represent the distribution of job types and seniority levels.

#5. Data Analysis and Visualization

This section details the data analysis and visualization of the aforementioned dataset. This analysis will help in answering the aforementioned research questions and visualization will provide visual detail of the results of the analysis. ##Inspection of dataset. The first step in the analysis process is the loading of the required libraries for the whole analysis, such as: ‘readr’, ‘dplyr’.

```
#Step_1# Get libraries for the whole analysis and visualization

# Load libraries

library(readr)

library(dplyr)

library(ggplot2)

library(RColorBrewer)

library(scales)

library(lubridate)

library(reshape2)
```

These libraries loaded collectively offer a comprehensive set of tools for data import, manipulation, visualization, and formatting, enabling efficient and effective analysis and visualization workflows for the whole project. In the next step, csv file has been imported into RStudio:

```
#Step_2# Load dataset
DS_Jobs <- read_csv("DS_Jobs.csv")

## Rows: 660 Columns: 27

── Column specification ────────────────────────────────────────────────

Delimiter: ","

chr (15): Job Title, Salary Estimate Thousands, Job Description, Company Name, Location, Headquarters...

dbl (12): Rating, min_salary, max_salary, avg_salary, same_state, python, excel, hadoop, spark, aws, ...

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.
```

In the above steps, the dataset is read using the '**read_csv**' function from the '**readr**' library. Finally, the data set has been called and inspected:

```
#Step_3# Inspect the dataset
DS_Jobs

# A tibble: 660 × 27

   Job Ti…¹ Salar…² Job D…³ Rating Compa…⁴ Locat…⁵ Headq…⁶ Emplo…⁷ Type …⁸ Indus…⁹ Sector Reven…˟ min_s…˟

   <chr>    <chr>   <chr>    <dbl> <chr>   <chr>   <chr>   <chr>   <chr>   <chr>   <chr>  <chr>     <dbl>

 1 Sr Data… 137-171 "Descr…    3.1 Health… New Yo… New Yo… 1001-5… Nonpro… Insura… Insur… Unknow…     137

 2 Data Sc… 137-171 "Secur…    4.2 ManTech Chanti… Herndo… 5001-1… Compan… Resear… Busin… $1000-…     137

 3 Data Sc… 137-171 "Overv…    3.8 Analys… Boston… Boston… 1001-5… Privat… Consul… Busin… $100-$…     137

 4 Data Sc… 137-171 "JOB D…    3.5 INFICON Newton… Bad Ra… 501-10… Compan… Electr… Manuf… $100-$…     137

 5 Data Sc… 137-171 "Data …    2.9 Affini… New Yo… New Yo… 51-200  Compan… Advert… Busin… Unknow…     137

 6 Data Sc… 137-171 "About…    4.2 HG Ins… Santa … Santa … 51-200  Compan… Comput… Infor… Unknow…     137

 7 Data Sc… 137-171 "Posti…    3.9 Novart… Cambri… Basel,… 10000+  Compan… Biotec… Biote… $10000+     137

 8 Data Sc… 137-171 "Intro…    3.5 iRobot  Bedfor… Bedfor… 1001-5… Compan… Consum… Retail $1000-…     137

 9 Staff D… 137-171 "Intui…    4.4 Intuit… San Di… Mounta… 5001-1… Compan… Comput… Infor… $2000-…     137

10 Data Sc… 137-171 "Ready…    3.6 XSELL … Chicag… Chicag… 51-200  Compan… Enterp… Infor… Unknow…     137

# … with 650 more rows, 14 more variables: max_salary <dbl>, avg_salary <dbl>, job_state <chr>,

#   same_state <dbl>, company_age <chr>, python <dbl>, excel <dbl>, hadoop <dbl>, spark <dbl>,

#   aws <dbl>, tableau <dbl>, big_data <dbl>, job_simp <chr>, seniority <chr>, and abbreviated variable

#   names ¹`Job Title`, ²`Salary Estimate Thousands`, ³`Job Description`, ⁴`Company Name`, ⁵Location,

#   ⁶Headquarters, ⁷`Employee Size`, ⁸`Type of ownership`, ⁹Industry, ˟`Revenue Millions`, ˟min_salary

# ℹ Use `print(n = ...)` to see more rows, and `colnames()` to see all variable names
```

Now our Data is ready for Analysis. We will analyze and visualize in the following sections.

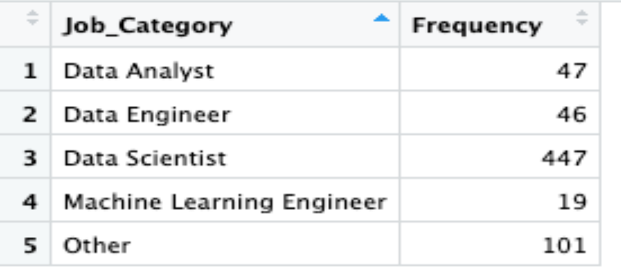
ANALYSIS 1. Count the frequencies of job titles in order to answer research question 1: ***What are the most common job titles in the data science field?***

Visualize the distribution of different job categories in the data science field to identify the most common positions and their frequencies. We classified using the **grepl()** function to check if each job title matches specific patterns (e.g., "Data Scientist", "Data Engineer", etc.). Based on the matches, we assign corresponding job categories using the **case_when()** function. The remaining job titles that don't match any specific pattern are assigned to the "Other" category. After classifying the job titles, we use the **count()** function to count the occurrences of each job title and store the results in the **job_title_counts** variable. Then it has been previewed:

```
#Step_1# Define classification rules
job_title_categories <- DS_Jobs %>%

  mutate(Job_Category = case_when(

    grepl("Data Scientist", `Job Title`, ignore.case = TRUE) ~ "Data Scientist",

    grepl("Data Engineer", `Job Title`, ignore.case = TRUE) ~ "Data Engineer",

    grepl("Machine Learning Engineer", `Job Title`, ignore.case = TRUE) ~ "Machine Learning
 Engineer",

    grepl("Data Analyst", `Job Title`, ignore.case = TRUE) ~ "Data Analyst",

     TRUE ~ "Other"))

#2 Count the frequency of each job category

job_category_counts <- table(job_title_categories$Job_Category)

#3 Create a data frame with job categories and their frequencies

job_category_df <- data.frame(Job_Category = names(job_category_counts),

                              Frequency = as.vector(job_category_counts))

#4 Preview the classified job titles
job_category_df

#

               Job_Category Frequency

1              Data Analyst        47

2             Data Engineer        46

3            Data Scientist       447

4 Machine Learning Engineer        19

5                     Other       101
```

Understanding the distribution of job titles allows job seekers to tailor their job search and focus on the most in-demand roles. It helps them align their skills and qualifications with the prevalent job titles, increasing their chances of finding relevant job openings. Here is the analysis result shown below in Table 1:

Table 1. Most common job categories.



	Job_Category	Frequency
1	Data Analyst	47
2	Data Engineer	46
3	Data Scientist	447
4	Machine Learning Engineer	19
5	Other	101

Now let’s create a bar chart to visualize the distribution of job categories in the data science field. For visualization, we use the following:

```
#Step_2# Create a bar chart with improved aesthetics and theme
ggplot(job_category_df, aes(x = Job_Category, y = Frequency, fill = Job_Category)) +

  geom_bar(stat = "identity", color = "black", width = 0.6) +

  labs(x = "Job Category", y = "Frequency", title = "Distribution of Job Categories") +

  scale_fill_manual(values = c("#FF6F61", "#6B5B95", "#88B04B", "#F7CAC9", "#A6CEE3"), guid
e = FALSE) +

  theme_minimal() +

  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 11, color = "black"),

        axis.text.y = element_text(size = 10, color = "black"),

        axis.title.x = element_text(size = 12, face = "bold"),

        axis.title.y = element_text(size = 12, face = "bold"),

        plot.title = element_text(size = 14, face = "bold", hjust = 0.5),

        panel.grid.major = element_line(color = "lightgray", linetype = "dashed"),

        panel.grid.minor = element_blank(),

        panel.border = element_blank(),

        panel.background = element_blank())
```

The x-axis represents the different job categories, including 'Data Analyst', 'Data Engineer', 'Data Scientist', 'Machine Learning Engineer', and 'Other'. The y-axis represents the frequency or count of job titles falling under each category. The bars in the plot correspond to each job category, and their heights indicate the frequency of job titles in that category. The colors of the bars represent different job categories, providing a visual distinction between them (see Figure 1).

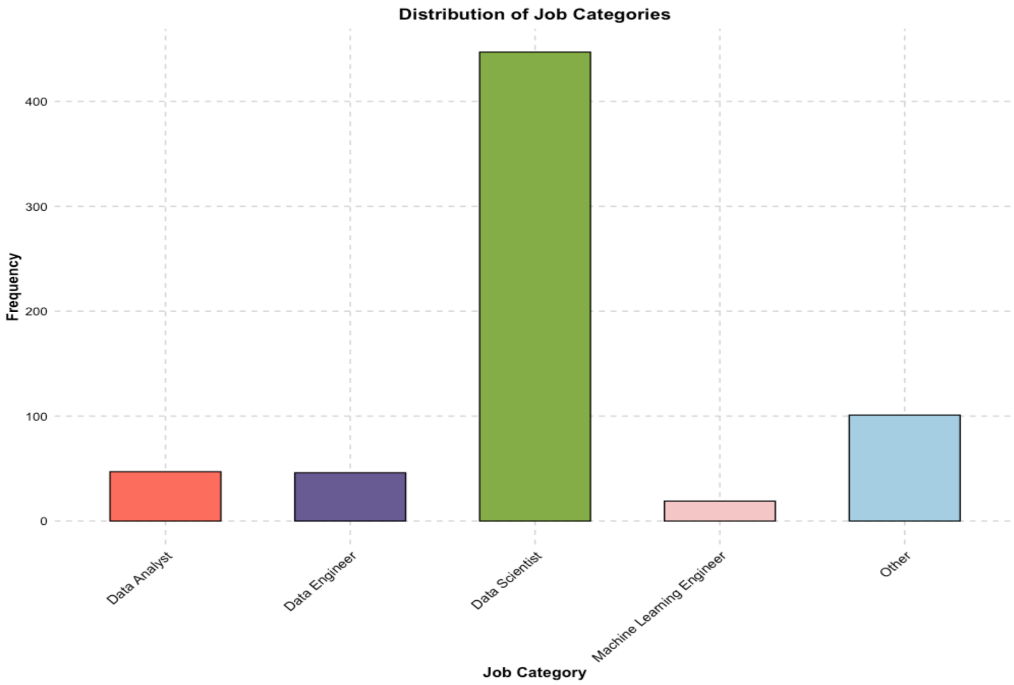


Figure 1. Visualization of the most common job categories.

Result: The figure 1 shows the number of job titles falling under each category, including 'Data Analyst', 'Data Engineer', 'Data Scientist', 'Machine Learning Engineer', and 'Other'. From the figure 1, we can observe that the most common job category is 'Data Scientist', with a frequency of 447 job titles. It is followed by 'Data Analyst' with 47 job titles, 'Data Engineer' with 46 job titles, 'Other' with 101 job titles, and 'Machine Learning Engineer' with 19 job titles.

The findings from the figure 1 provide valuable insights into the job market and the demand for different roles in the data science field. **Data Scientist:** The high frequency of 'Data Scientist' job titles suggests a strong demand for professionals with skills in data analysis, statistical modeling, and machine learning. This indicates the significance of data scientists in various industries and organizations that rely on data-driven decision-making. **Data Analyst:** The presence of a considerable number of 'Data Analyst' job titles highlights the need for professionals who can extract insights from data and provide meaningful interpretations. Data analysts play a crucial role in analyzing and visualizing data to support business strategies and decision-making processes. **Data Engineer:** The relatively high number of 'Data Engineer' job titles signifies the importance of professionals who can design and build data infrastructure. Data engineers are responsible for managing and organizing data, ensuring its quality, and enabling efficient data processing. **Machine Learning Engineer:** Although the frequency of 'Machine Learning Engineer' job titles is relatively lower than the other categories, it indicates a growing demand for professionals skilled in machine learning algorithms and model development. Machine learning engineers are involved in building and deploying machine learning models for various applications. **Other:** The category of 'Other' includes job titles that may not fit directly into the specified categories but are still relevant to the data science field. These job titles could represent specialized roles, emerging job titles, or positions that require a combination of skills from multiple categories.

Figure 1 provides a snapshot of the job market in the data science field, highlighting the demand for different roles. Professionals aspiring to pursue a career in data science can use this information to identify popular job categories and align their skills and expertise accordingly. Employers and organizations can benefit from these findings by understanding the job market trends and tailoring their hiring strategies to meet the industry's demands. It is important to note that the analysis is based on the provided data, and the actual job market dynamics may vary based on different factors such as location, industry, and time period. Continuous monitoring of job trends and staying updated with evolving technologies and skills can further enhance career opportunities in the data science field.

ANALYSIS 2. Salary Analysis by Job Title to answer research question 2: **How do the salary ranges vary for different job titles and industries in the data science field?**

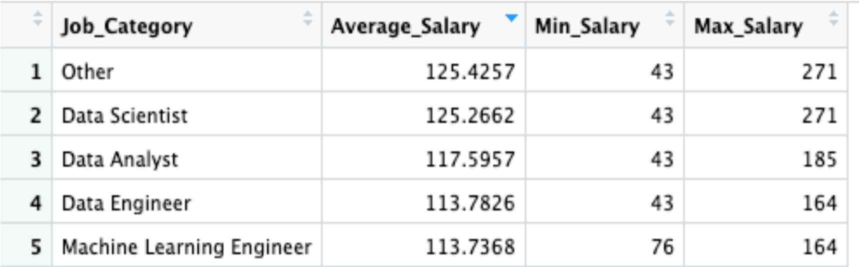
In this step, we analyze and visualize the salary ranges for different job categories and industries in the classified dataset. Salary analysis is performed by grouping the data by job category and industry using the '**group_by**' function from '**dplyr**'. The minimum, maximum, and average salary values are calculated using the '**min**', '**max**', and '**mean**' functions, respectively.

```
#Step_1# Group the data by job category and calculate average, minimum, and maximu
m salaries

salary_analysis <- job_title_categories %>%

  group_by(Job_Category) %>%

  summarise(Average_Salary = mean(avg_salary),

            Min_Salary = min(avg_salary),

            Max_Salary = max(avg_salary))
```

```
#2 Sort the data by average salary in descending order
salary_analysis <- salary_analysis %>%
  arrange(desc(Average_Salary))
#3 Print the salary analysis
print(salary_analysis)
##
# A tibble: 5 × 4
  Job_Category              Average_Salary Min_Salary Max_Salary
  <chr>                              <dbl>      <dbl>      <dbl>
1 Other                               125.         43        271
2 Data Scientist                      125.         43        271
3 Data Analyst                        118.         43        185
4 Data Engineer                       114.         43        164
5 Machine Learning Engineer           114.         76        164
```

The salary analysis by job categories enables job seekers to make informed decisions regarding their career choices and negotiate fair compensation. Employers can utilize this data to stay competitive in talent acquisition and retention. The results shown in Table 2:

Table 2. Salary by job categories.



	Job_Category	Average_Salary	Min_Salary	Max_Salary
1	Other	125.4257	43	271
2	Data Scientist	125.2662	43	271
3	Data Analyst	117.5957	43	185
4	Data Engineer	113.7826	43	164
5	Machine Learning Engineer	113.7368	76	164

The analysis of salaries by job categories provides insights into the salary ranges and average salaries for different positions in the data science field. The Table 2 presents information on minimum salary, maximum salary, and average salary for each job categories.

For better visualization of salary analysis by job categories, a bar plot is created using '**ggplot2**':

```
#Step_2# Visualize Salary Analysis by Job Categories
ggplot(salary_analysis, aes(x = Job_Category, y = Average_Salary)) +
  geom_bar(stat = "identity", fill = "skyblue", color = "black") +
  geom_errorbar(aes(ymin = Min_Salary, ymax = Max_Salary), width = 0.2, color = "black") +
  labs(x = "Job Category", y = "Salary (USD)", title = "Salary Analysis by Job Title") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 11, color = "black"),
        axis.text.y = element_text(size = 10, color = "black"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),

        plot.title = element_text(size = 14, face = "bold", hjust = 0.5),

        panel.grid.major.x = element_line(color = "lightgray", linetype = "dashed"),

        panel.grid.minor = element_blank(),

        panel.border = element_blank(),

        panel.background = element_blank())
```

Here, the Figure 2 below, represents a comparison of average salaries for different job categories in the field of Data Science.

The x-axis represents the job categories, including "Other," "Data Scientist," "Data Analyst," "Data Engineer," and "Machine Learning Engineer." The y-axis represents the salary in USD. Each bar in the plot represents the average salary for a particular job category. The height of the bar indicates the average salary value. The bars are filled with a sky-blue color. Additionally, error bars are shown on each bar, representing the range of salaries within each job category. The lower and upper ends of the error bars correspond to the minimum and maximum salaries, respectively:

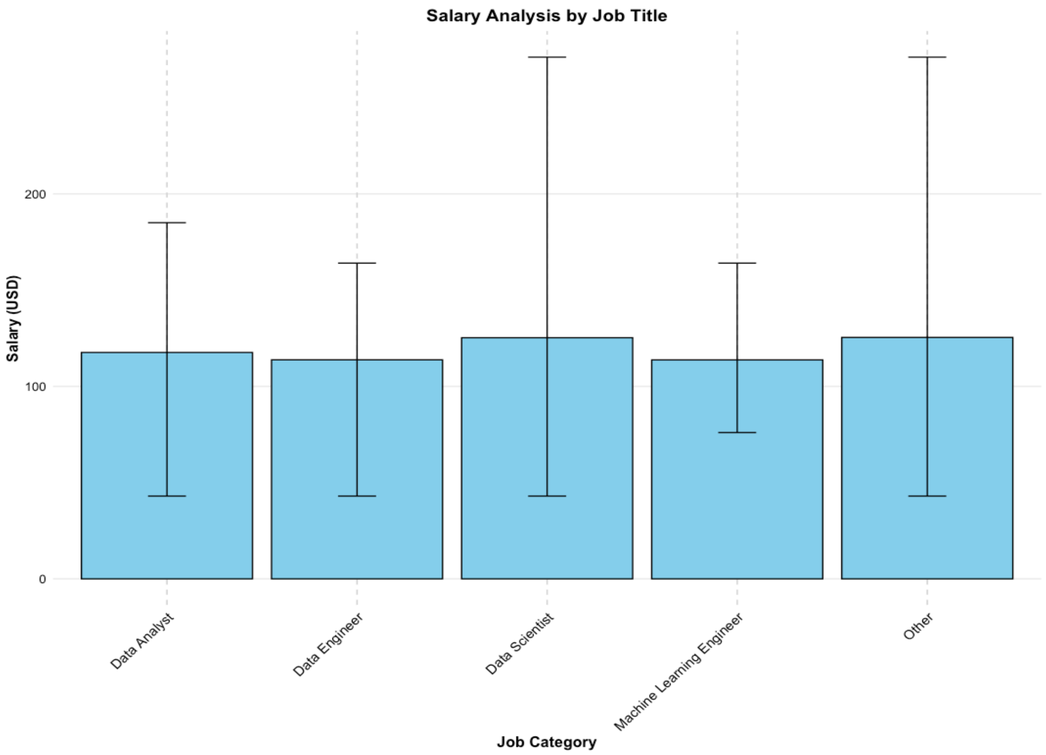


Figure 2. Comparison of average salaries for different job titles.

Result: From Figure 2, we can observe that the job category "Data Scientist" and "Other" have the highest average salary, as indicated by the taller bars. This suggests that professionals in these categories tend to receive higher salaries compared to other job categories in the data science field. The job categories "Data Analyst," "Data Engineer," and "Machine Learning Engineer" have slightly lower average salaries compared to "Data Scientist" and "Other." The error bars provide a visual representation of the salary range within each job category. For example, the "Data Scientist" category has a wider salary range compared to other categories, as indicated by the longer error bars. Overall, the plot highlights the variation in salaries across different job categories in the data science field. It provides a quick comparison of average salaries and allows for a visual understanding of the salary ranges within each category.

The benefit of these results for job seekers is that they provide valuable insights into the salary ranges within the Data Science field. Job seekers can use this information to make informed decisions about their career choices and negotiate better compensation packages. By identifying job titles with higher average salaries, job seekers can target those positions to potentially earn a higher income. Conversely, they can also use this data to avoid job titles with lower average salaries if their primary goal is to maximize their earning potential.

Employers can also benefit from this analysis as it provides them with a benchmark for salary ranges in the data science field. Employers can use this information to align their salary offerings with industry standards and ensure they remain competitive in the job market. This knowledge can empower them during salary negotiations, as they can reference the average salaries associated with specific job titles to support their requests for competitive compensation.

ANALYSIS 3. Industry and Location Analysis for Data Science Job postings to answer research question 3: **Which industries and locations have the highest number of job postings in the data science field?**

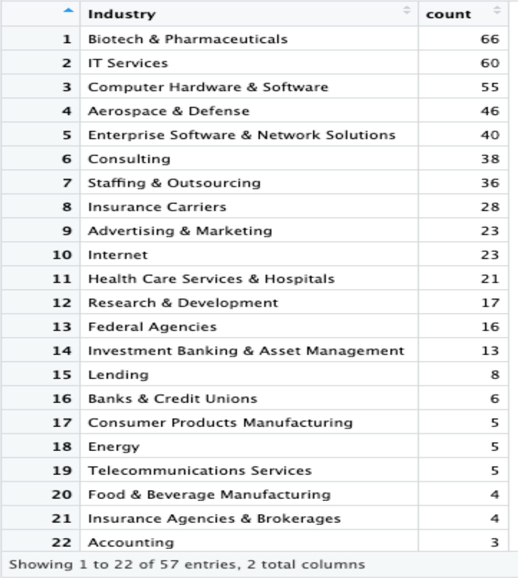
In this step, an industry analysis is conducted on the dataset. The goal is to analyze the distribution of job postings across different industries in the field of Data Science. The code below calculates the count of job postings for each industry also removes unknown industries by filtering and arranges them in descending order:

```
#Step_1# Industry Analysis

industry_counts <- DS_Jobs %>%

  filter(Industry != "Unknown") %>%

  group_by(Industry) %>%

  summarize(count = n()) %>%

  arrange(desc(count))

#2 Preview

industry_counts

##

# A tibble: 57 × 2

   Industry                                 count

   <chr>                                    <int>

 1 Biotech & Pharmaceuticals                   66

 2 IT Services                                 60

 3 Computer Hardware & Software                55

 4 Aerospace & Defense                         46

 5 Enterprise Software & Network Solutions     40

 6 Consulting                                  38

 7 Staffing & Outsourcing                      36

 8 Insurance Carriers                          28

 9 Advertising & Marketing                     23
```

```
10 Internet                              23

# … with 47 more rows
```

The results of the analysis are presented in the Table 3. The table has two columns: "Industry" and "count." The "Industry" column lists the names of the industries, while the "count" column shows the number of job postings associated with each industry. The industry analysis step and its results provide an overview of the distribution of job postings across different industries in the field of Data Science, helping job seekers and professionals make informed decisions about their career choices and job search strategies.

Table 3. The top industries with the highest number of job postings



	Industry	count
1	Biotech & Pharmaceuticals	66
2	IT Services	60
3	Computer Hardware & Software	55
4	Aerospace & Defense	46
5	Enterprise Software & Network Solutions	40
6	Consulting	38
7	Staffing & Outsourcing	36
8	Insurance Carriers	28
9	Advertising & Marketing	23
10	Internet	23
11	Health Care Services & Hospitals	21
12	Research & Development	17
13	Federal Agencies	16
14	Investment Banking & Asset Management	13
15	Lending	8
16	Banks & Credit Unions	6
17	Consumer Products Manufacturing	5
18	Energy	5
19	Telecommunications Services	5
20	Food & Beverage Manufacturing	4
21	Insurance Agencies & Brokerages	4
22	Accounting	3

Showing 1 to 22 of 57 entries, 2 total columns

The Table 3 reveals the top industries with the highest number of job postings in the dataset. The industry with the highest count is "Biotech & Pharmaceuticals" with 66 job postings, followed by "IT Services" with 60 job postings, and so on.

The Figure 3 displays the distribution of industries based on their count, offering valuable insights into the popularity of various sectors within the dataset:

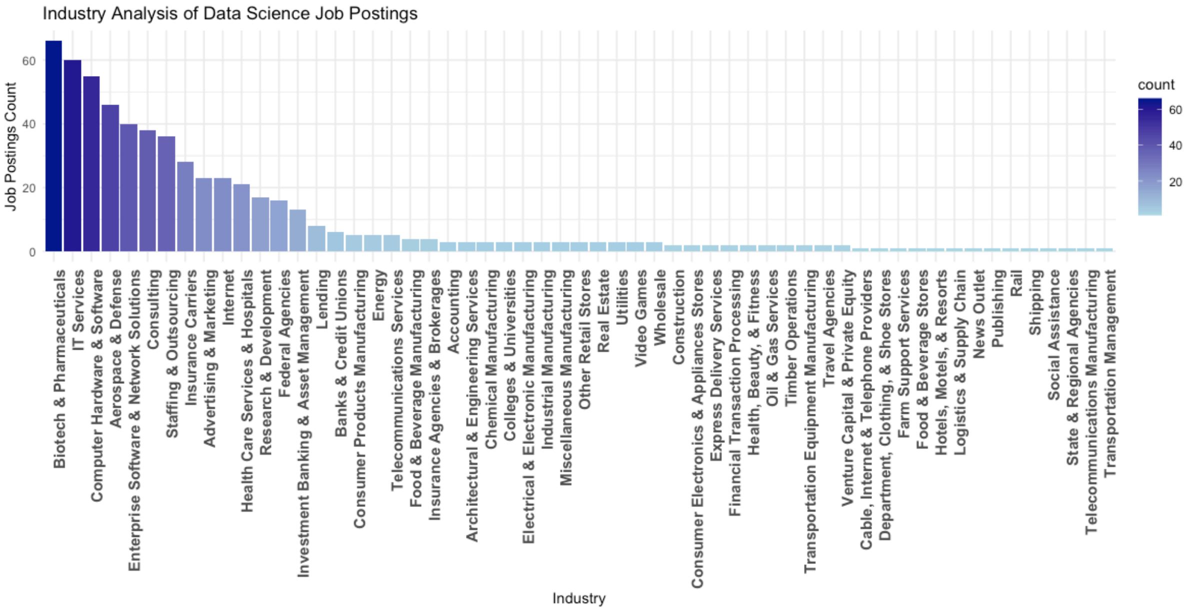


Figure 3. Distribution of industries based on their Job posting count

Result: As shown in Figure 3, among the highest job posting industries, the top five are "Biotech & Pharmaceuticals," "IT Services," "Computer Hardware & Software," "Aerospace & Defense," and "Enterprise Software & Network Solutions." The highest industry, "Biotech & Pharmaceuticals," with 66 entries, suggests a strong demand for professionals in this sector. This could be attributed to advancements in medical research, pharmaceutical development, and the increasing importance of healthcare solutions in society. Job seekers with backgrounds in life sciences, research, or pharmaceuticals may find abundant opportunities within this industry. The IT Services industry follows closely with 60 entries, indicating a significant need for skilled professionals in the technology sector. This demand could be driven by the rapid digital transformation of businesses, increasing reliance on technology infrastructure, and the emergence of new software and IT solutions. Job seekers with expertise in software development, data analysis, cybersecurity, and IT consulting may find numerous prospects within this industry. With 55 entries, the "Computer Hardware & Software" industry demonstrates a substantial presence in the dataset. This can be attributed to the ongoing advancements in computer technology, the rise of software-driven solutions, and the demand for hardware components. Job seekers specializing in computer engineering, software development, or hardware design may find promising opportunities within this field. The "Aerospace & Defense" industry, with 46 entries, highlights the significance of the aerospace sector in terms of job listings. This industry encompasses aircraft manufacturing, defense systems, and space exploration. The demand for skilled professionals in this sector may be driven by government contracts, defense investments, and the pursuit of technological advancements. Job seekers with backgrounds in engineering, aviation, or defense-related disciplines may find rewarding careers within this industry.

On the other hand, several industries have relatively low representation in the dataset. For instance, industries like "Cable, Internet & Telephone Providers," "Department, Clothing, & Shoe Stores," "Farm Support Services," "Food & Beverage Stores," and "Hotels, Motels, & Resorts" each have only one entry. This suggests a comparatively lower number of job opportunities in these sectors within the dataset. Job seekers targeting these industries may face more limited options and should consider exploring related sectors or expanding their job search to increase their chances of success.

Understanding the distribution of industries in the dataset can be advantageous for job seekers. It allows them to assess the demand for specific industries, make informed decisions about their career paths, and tailor their job search strategies accordingly. By focusing on industries with higher counts, such as biotech & pharmaceuticals, IT services, computer hardware & software, aerospace & defense, and others, job seekers can potentially tap into sectors that offer a greater number of job opportunities. Additionally, they can consider the reasons behind the prevalence of certain industries, such as technological advancements, market demand, government investments, or societal needs, to align their skills and experiences with the current job market trends.

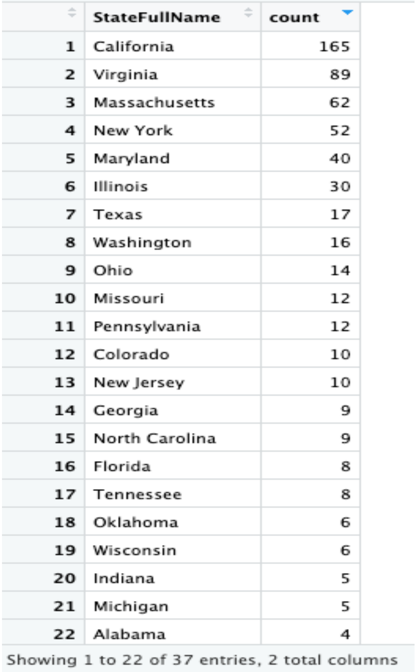
Part 2 of the Analysis 3: Location Analysis.

In this step, a location analysis is conducted. The location analysis of data science job postings involved several steps to summarize and examine the distribution of jobs across different states. By extracting state information from the **Location** column, we obtained the counts of job postings for each state. Duplicate state names were merged, and the counts were aggregated to provide a comprehensive view of job distribution by state. After removing any rows with missing or incomplete data, we obtained a clean dataset with valid state names and their corresponding counts.

```
#Step_1# Define the state abbreviation to full name mapping

state_mapping <- data.frame(StateAbbreviation = c("AL", "AK", "AZ", "AR", "CA", "CO", "CT",
 "DE", "FL", "GA", "HI", "ID", "IL", "IN", "IA", "KS", "KY", "LA", "ME", "MD", "MA", "MI",
"MN", "MS", "MO", "MT", "NE", "NV", "NH", "NJ", "NM", "NY", "NC", "ND", "OH", "OK", "OR", "
PA", "RI", "SC", "SD", "TN", "TX", "UT", "VT", "VA", "WA", "WV", "WI", "WY"),

                            StateFullName = c("Alabama", "Alaska", "Arizona", "Arkansas", "
California", "Colorado", "Connecticut", "Delaware", "Florida", "Georgia", "Hawaii", "Idaho
", "Illinois", "Indiana", "Iowa", "Kansas", "Kentucky", "Louisiana", "Maine", "Maryland", "
Massachusetts", "Michigan", "Minnesota", "Mississippi", "Missouri", "Montana", "Nebraska",
"Nevada", "New Hampshire", "New Jersey", "New Mexico", "New York", "North Carolina", "North
 Dakota", "Ohio", "Oklahoma", "Oregon", "Pennsylvania", "Rhode Island", "South Carolina", "
South Dakota", "Tennessee", "Texas", "Utah", "Vermont", "Virginia", "Washington", "West Vir
ginia", "Wisconsin", "Wyoming"))

#Step_2# Location Analysis

location_counts <- DS_Jobs %>%

  group_by(Location) %>%

  summarize(count = n()) %>%

  arrange(desc(count))

#3 Extract state information from Location

location_counts <- location_counts %>%

  mutate(StateAbbreviation = str_extract(Location, "\\b[A-Z]{2}\\b"))

#4 Join with the state mapping table to add full state names

location_counts <- left_join(location_counts, state_mapping, by = "StateAbbreviation")

#5 Remove the StateAbbreviation column

location_counts <- location_counts %>%

  select(-StateAbbreviation)

#6 Merge duplicate state names and sum the counts

location_counts <- location_counts %>%

  group_by(StateFullName) %>%

  summarize(count = sum(count))

#7 Remove rows with NA values

location_counts <- na.omit(location_counts)

#8 Preview the classified locations with merged state names and removed NA values

print(location_counts)

##

# A tibble: 37 × 2

   StateFullName count

   <chr>         <int>

 1 Alabama           4

 2 Arizona           4

 3 California      165
 4 Colorado          10

 5 Connecticut        4

 6 Delaware           1

 7 Florida            8

 8 Georgia            9

 9 Illinois          30

10 Indiana            5

# … with 27 more rows
```

This analysis helps us identify the states with the highest demand for data science professionals and offers valuable insights into the geographic landscape of the data science job market as shown in table 4:

Table 4. Number of job listings associated with each location (by State).



	StateFullName	count
1	California	165
2	Virginia	89
3	Massachusetts	62
4	New York	52
5	Maryland	40
6	Illinois	30
7	Texas	17
8	Washington	16
9	Ohio	14
10	Missouri	12
11	Pennsylvania	12
12	Colorado	10
13	New Jersey	10
14	Georgia	9
15	North Carolina	9
16	Florida	8
17	Tennessee	8
18	Oklahoma	6
19	Wisconsin	6
20	Indiana	5
21	Michigan	5
22	Alabama	4

Showing 1 to 22 of 37 entries, 2 total columns

By examining the Table 4, we can gain insights into the distribution of job listings across different locations.

But let's create a graph for better visualization:

```
#Step_3# Bar chart for Visualization of Location Analysis

ggplot(location_counts, aes(x = reorder(StateFullName, -count), y = count, fill = count)) +

  geom_bar(stat = "identity", width = 0.6) +

  scale_fill_gradient(low = "#4C72B0", high = "#FFA500") +  # Adjust color gradient

  labs(x = "State", y = "Job Postings Count") +

  ggtitle("Location Analysis of Data Science Job Postings") +

  theme_minimal() +

  theme(axis.text.x = element_text(angle = 90, hjust = 1, size = rel(1.3)),

        axis.text.y = element_text(size = 12),
        plot.title = element_text(size = 16, face = "bold"),

        plot.margin = unit(c(1, 1, 1, 3), "lines"))
```

Here is the output:

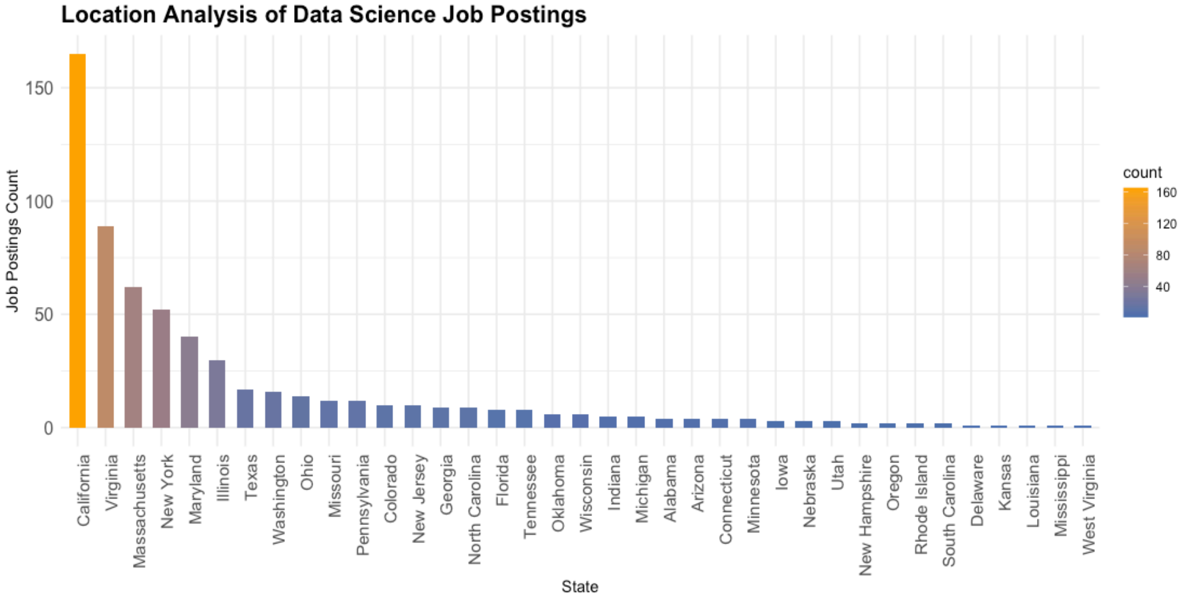


Figure 4. The overview of the job postings counts in different locations (by State).

Result: The figure 4 illustrates the distribution of data science job postings across various states. California stands out as the top state with the highest number of job postings, followed by Virginia, Massachusetts, New York, and Maryland. These states reveal a strong demand for data science professionals, presenting plenty opportunities for job seekers in the field. The figure 4 emphasizes the significance of location when considering career prospects in data science, as certain states offer a greater concentration of job opportunities. While California dominates in terms of job postings, other states such as Virginia, Massachusetts, New York, and Maryland also show a considerable number of opportunities. These states are emerging data science centers, indicating the growth and expansion of the field beyond traditional technology hotspots. Job seekers who are open to exploring opportunities outside of the well-established data science locations can benefit from these emerging hubs, which offer a balance between job availability and potentially lower competition.

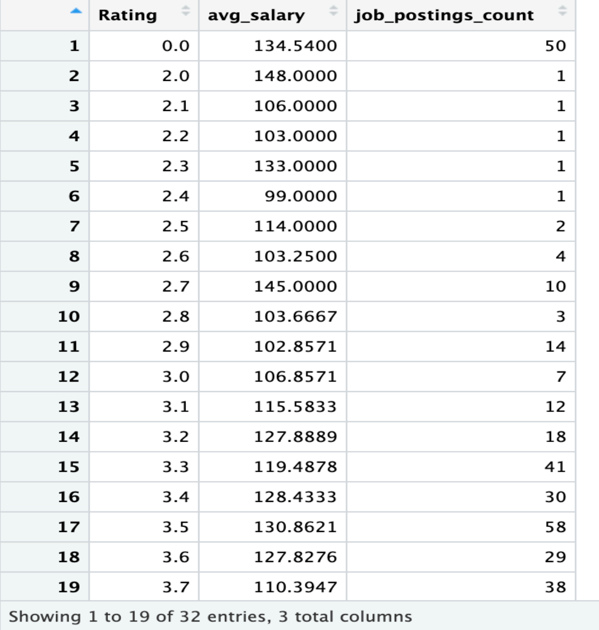
ANALYSIS 4. Company Ratings vs Average Salary and Job Postings Analysis to find out **if there is a correlation between company ratings and job postings, indicating whether highly-rated companies offer better salaries or have more job opportunities.**

The analysis focuses on examining the relationship between company ratings and job postings. We start by grouping the dataset by the "Rating" column. Then, calculate two metrics for each rating group: the average salary ("avg_salary") and the count of job postings ("job_postings_count"). The 'mean()' function is used to calculate the average salary, and the 'n()' function counts the number of job postings within each rating group. The results are stored in a new table called "rating_analysis".

```
#Step_1# Company Ratings Analysis

rating_analysis <- DS_Jobs %>%

  group_by(Rating) %>%

  summarize(avg_salary = mean(avg_salary),
               job_postings_count = n())
#2 Preview
rating_analysis
##
# A tibble: 32 × 3
   Rating avg_salary job_postings_count
    <dbl>      <dbl>              <int>
 1    0         135.                 50
 2    2         148                   1
 3    2.1       106                   1
 4    2.2       103                   1
 5    2.3       133                   1
 6    2.4        99                   1
 7    2.5       114                   2
 8    2.6       103.                  4
 9    2.7       145                  10
10    2.8       104.                  3
# … with 22 more rows
```

The output of this step shown in table 5, consists of three columns: “Rating”, “avg_salary”, and “job_postings_count”. The “Rating” column represents the company rating, while the “avg_salary” column displays the average salary associated with each rating. The “job_postings_count” column indicates the number of job postings for each rating.

Table 5. Number of job listings associated with each location



	Rating	avg_salary	job_postings_count
1	0.0	134.5400	50
2	2.0	148.0000	1
3	2.1	106.0000	1
4	2.2	103.0000	1
5	2.3	133.0000	1
6	2.4	99.0000	1
7	2.5	114.0000	2
8	2.6	103.2500	4
9	2.7	145.0000	10
10	2.8	103.6667	3
11	2.9	102.8571	14
12	3.0	106.8571	7
13	3.1	115.5833	12
14	3.2	127.8889	18
15	3.3	119.4878	41
16	3.4	128.4333	30
17	3.5	130.8621	58
18	3.6	127.8276	29
19	3.7	110.3947	38

Showing 1 to 19 of 32 entries, 3 total columns

Result: Table 5 provides valuable insights into the relationship between company ratings and two important factors: average salary and job postings count. Analyzing company ratings and job postings count provides valuable insights for job seekers. It helps identify popular companies with a high number of job postings, indicating active hiring and more job opportunities. The analysis also reveals job market demand and industries with a strong need for professionals. Highly-rated companies tend to have more job postings, suggesting positive work environments and growth prospects. Job seekers can target reputable companies with abundant job postings and align their skills accordingly. The analysis highlights competition levels, potential growth areas, and market trends, empowering job seekers to make informed decisions and navigate the job market effectively.

The next step, we create a bar graph to visualize company ratings and job postings count. As there are more levels in the **‘Rating’** variable, we need to generate a larger palette with 32 colors to accommodate all the levels in the **‘Rating’** variable. Then each rating level will be assigned a different color from the expanded color palette. Afterwards the ‘**Rating’** variable is converted to a factor using the ‘**factor()’** function. This ensures that it is treated as a discrete variable in the plot:

```
 #Step_2# Visualize Company Ratings and Job Postings

   #2 Generate a larger color palette with 32 colors

colors <- c("red", "blue", "green", "yellow", "purple", "orange", "pink", "brown", "gray",
"cyan", "magenta",

            "darkgreen", "darkblue", "darkred", "darkorange", "darkgray", "darkcyan", "dark
magenta", "lightgreen",

            "lightblue", "yellow", "lightpink", "brown", "lightgray", "darkcyan", "darkmage
nta",

            "steelblue", "goldenrod", "olivedrab", "tomato", "slategray", "deepskyblue", "o
rchid")

      ##3 Convert Rating to factor

rating_analysis$Rating <- factor(rating_analysis$Rating)

    #4 Create Bar Graph

ggplot(rating_analysis, aes(x = Rating, y = job_postings_count, fill = Rating)) +

  geom_bar(stat = "identity") +

  scale_fill_manual(values = colors) +  # Set the fill colors using the larger color palett
e

  labs(x = "Company Rating", y = "Job Postings Count") +

  ggtitle("Company Ratings and Job Postings Count") +

  theme_minimal()
```

The bar graph visualizes the relationship between company ratings and job postings count is generated correctly. Each bar represents a specific company rating, and the height of the bar corresponds to the number of job postings for that rating as shown below in Figure 5:

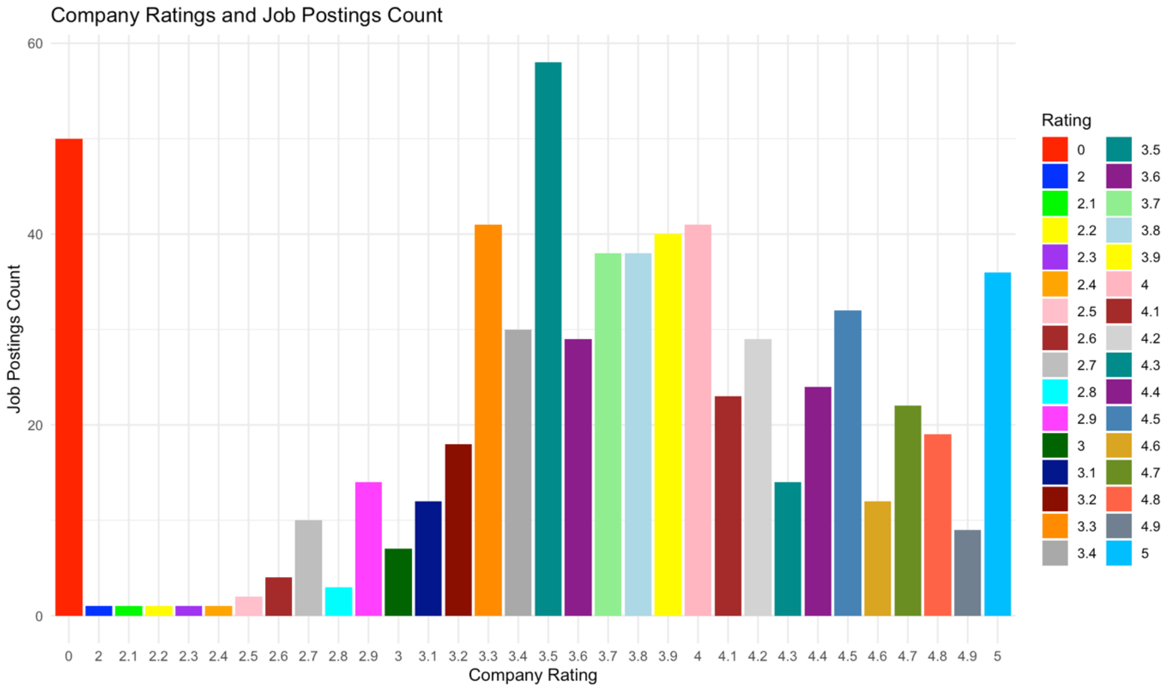


Figure 5. Visualization of Company Ratings and Job Postings Count

Result: The Figure 5 shows that there is variation in the number of job postings across different company ratings. Higher-rated companies tend to have a higher number of job postings compared to lower-rated companies. However, there are a few exceptions where lower-rated companies have a relatively high number of job postings as well. One possible reason for this trend is that highly-rated companies often have a good reputation and are known for providing better job opportunities. Job seekers tend to be attracted to companies with higher ratings due to factors such as better work culture, competitive salaries, growth opportunities, and employee benefits. As a result, these companies receive a larger number of job applications and hence have a higher number of job postings. On the other hand, it is also possible that some lower-rated companies are actively hiring or expanding their workforce, which leads to a higher number of job postings despite their lower ratings. These companies may be undergoing a restructuring phase, targeting specific skill sets, or offering competitive compensation packages to attract potential candidates.

For job seekers, this analysis and graph provide valuable insights. They can use the information to identify companies with higher job posting counts, which could indicate more employment opportunities. Additionally, it allows job seekers to gauge the overall market demand for their skills and target companies that align with their career goals.

Part 2 of the Analysis 4: Average salaries across different company ratings

Let’s move to step 3, to provide a visual understanding of how company ratings and average salaries are related and any potential patterns or trends that may exist. It involves visualizing the relationship between company ratings and average salaries using a scatter plot with a trend line:

```
#Step_3# Visualize Company Ratings and Average Salaries

ggplot(rating_analysis, aes(x = Rating, y = avg_salary)) +

  geom_point(size = 3, color = "purple") +

  geom_smooth(method = "lm", se = FALSE, color = "green") +

  labs(x = "Company Rating", y = "Average Salary") +
  ggtitle("Company Ratings and Average Salary") +

  theme_minimal()
```

By executing this code, a scatter plot will be generated, showing the relationship between company ratings and average salaries. The scatter points will be colored purple, and a trend line will be overlaid in green, representing the general direction of the relationship between the variables. But I wanted to make it more visible and better to read, so then, these additional customizations have been added to enhance the visualization:

```
#Step_4# Customized for Better Visualization

ggplot(rating_analysis, aes(x = Rating, y = avg_salary, color = "Data Points")) +

  geom_point(size = 3, alpha = 0.8) +

  geom_smooth(aes(color = "Regression Line"), method = "lm", se = FALSE, linetype = "dashed") +

  labs(x = "Company Rating", y = "Average Salary") +

  ggtitle("Company Ratings and Average Salary") +

  theme_minimal() +

  theme(

    plot.title = element_text(size = 16, face = "bold", hjust = 0.5),

    axis.title = element_text(size = 12, face = "bold"),

    axis.text = element_text(size = 10),

    axis.line = element_line(color = "black"),

    legend.position = "bottom",

    legend.title = element_blank(),

    panel.grid.major = element_line(color = "gray", linetype = "dashed"),

    panel.grid.minor = element_blank(),

    panel.background = element_rect(fill = "lightgray")) +

  scale_color_manual(values = c("orange", "darkblue"), labels = c("Data Points", "Regression Line"))
```

After including these customizations, the scatter plot points will be slightly transparent and colored orange, representing "Data Points," while the trend line will be dashed and colored dark blue, representing the "Regression Line." The legend will be positioned at the bottom of the plot and will display the custom colors and labels. Additionally, other elements such as the plot title, axis titles, and grid lines are modified to improve readability and aesthetics as shown in Figure 6:

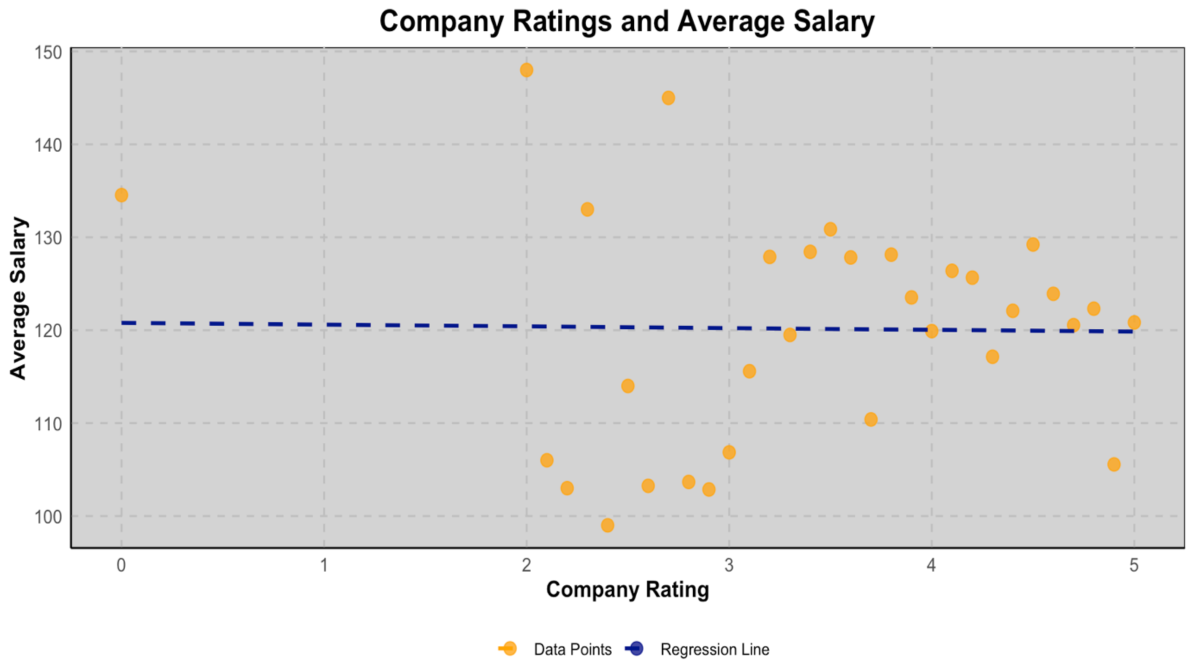


Figure 6. The relationship between company ratings and average salaries

Result: One key observation from the Figure 6 is the variation in average salaries across different company ratings. It visualizes a moderately positive trend, where higher company ratings tend to be associated with higher average salaries. As the company rating increases, there is a tendency for the average salary to also increase. However, there are some deviations from the trend line, indicating that other factors may influence salary levels. Higher-rated companies tend to have higher average salaries, while lower-rated companies generally offer lower average salaries. This pattern suggests a positive correlation between company ratings and salaries, indicating that highly-rated companies may be more likely to provide better compensation packages. Job seekers who prioritize competitive salaries may find it advantageous to target companies with higher ratings.

The positive relationship between company ratings and average salaries suggests that highly-rated companies are more likely to offer better compensation packages. This could be attributed to several reasons: *Reputation and Market Position:* Highly-rated companies often have a strong reputation in the industry and are well-established. Their positive image and market position may enable them to attract top talent and offer competitive salaries to retain skilled professionals. *Performance and Success:* Companies with higher ratings may have a track record of financial success, growth, and profitability. This success can provide them with the resources to invest in higher salaries as a means of attracting and retaining talented employees. *Employee Satisfaction and Engagement:* Companies that prioritize employee satisfaction and engagement are more likely to receive higher ratings. Such organizations may invest in employee benefits, including competitive salaries, to foster a positive work environment and motivate their workforce.

Benefits for Job Seekers: The results of this analysis have several benefits for job seekers: *Salary Expectations:* Job seekers can use this information to gain insights into salary expectations based on the company's rating. It can help them assess whether a company's compensation aligns with their financial goals and expectations. *Targeting Highly-Rated Companies:* Job seekers interested in higher salaries may consider targeting companies with higher ratings. These organizations are more likely to offer competitive compensation packages, indicating potential long-term financial benefits. *Negotiation and Decision Making:* Armed with knowledge about the relationship between ratings and salaries, job seekers can use this information to negotiate better offers or make more informed decisions when evaluating job opportunities.

However, it's important for job seekers to consider other factors beyond just company ratings and salaries. Cultural fit, career development, work-life balance, and personal growth opportunities are also essential aspects to evaluate when making career decisions.

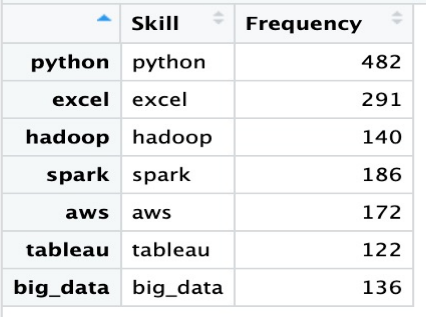
ANALYSIS 5. Skill Analysis for Data Science Jobs to find out: **What are the most in-demand skills for data science jobs based on the skill analysis?**

The analysis focuses on the most in-demand skills for data science jobs based on the skill analysis. The steps for analysis have been run using these codes below:

```
  #Step_1# Extract skill columns

skill_columns <- c("python", "excel", "hadoop", "spark", "aws", "tableau", "big_data")

  #2 Skill Analysis

skill_counts <- DS_Jobs %>%

  summarise(across(all_of(skill_columns), sum))

   #3 Convert skill counts to a data frame

skill_df <- data.frame(Skill = names(skill_counts), Frequency = unlist(skill_counts))

  #4 Sort skills by frequency in descending order

skill_df <- skill_df[order(skill_df$Frequency, decreasing = TRUE), ]

#5 Preview

skill_df

##

Skill Frequency

python      python       482

excel        excel       291

hadoop      hadoop       140

spark        spark       186

aws            aws       172

tableau    tableau       122

big_data big_data        136
```

By executing these steps, provide a summary of the most in-demand skills for data science jobs, along with their respective frequencies as shown in table 6 below:

Table 6. Extracted skill frequencies for Skill Analysis



	Skill	Frequency
python	python	482
excel	excel	291
hadoop	hadoop	140
spark	spark	186
aws	aws	172
tableau	tableau	122
big_data	big_data	136

Then it is better to create bar chart for better visualization and to gain more knowledge from the analysis. The following step, we executed these codes for creating visualization:

```
#Step_2# Bar chart for Skill Analysis

ggplot(skill_df, aes(x = Skill, y = Frequency, fill = Skill)) +

  geom_bar(stat = "identity") +

  scale_fill_manual(values = c(aws = "darkblue", big_data = "red", excel = "purple", hadoop
 = "green", python = "brown", spark = "orange", tableau = "blue")) +

  labs(x = "Skill", y = "Frequency") +

  ggtitle("Skill Analysis for Data Science Jobs") +

  theme_minimal() +

  theme(

    axis.text.x = element_text(angle = 90, hjust = 1, size = rel(1.5), face = "bold"),

    axis.text.y = element_text(face = "bold"),

    axis.title = element_text(face = "bold"),

    plot.title = element_text(face = "bold", size = 14, hjust = 0.5),

    axis.line = element_line(color = "black"),

    axis.ticks = element_line(color = "black"),

legend.position = "none")
```

The created bar chart visualizing the skill analysis results provides valuable insights into the most in-demand skills for data science jobs. The Figure 7 shows the frequency of each skill in different colors and allows us to interpret the results more effectively.

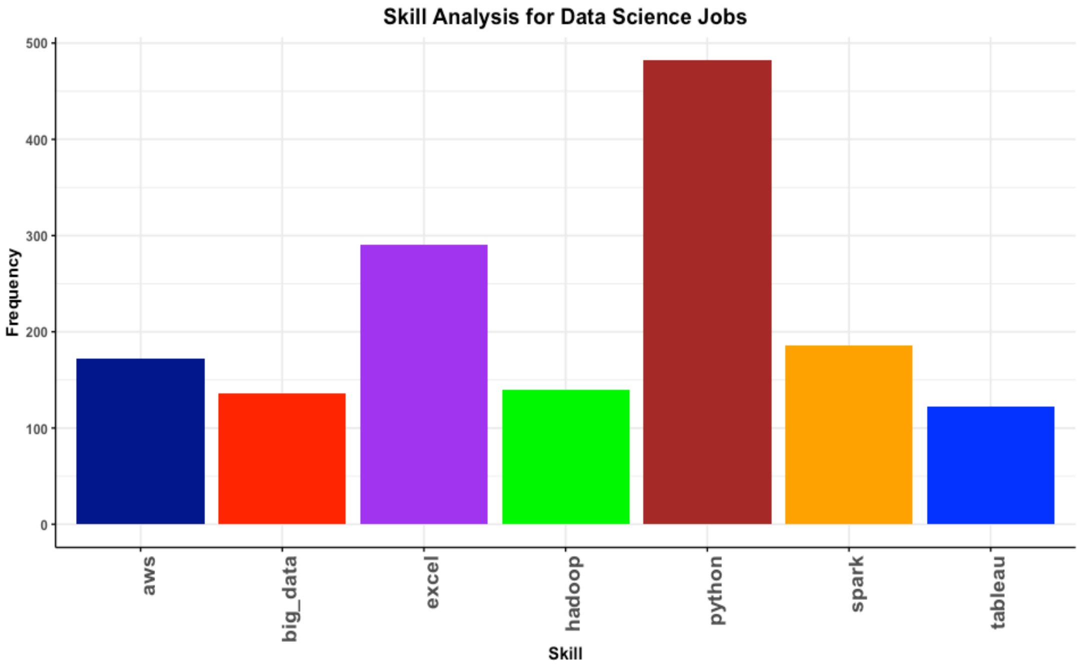


Figure 7. Overview of the high-demand skills for Data Science Jobs.

Result: From the Figure 7, we can observe that *Python* is the most in-demand skill for data science, with a frequency of 482. This reaffirms the widespread use of Python in the field of data science due to its flexibility, extensive libraries for data analysis and machine learning, and its user-friendly syntax. *Excel*, with a frequency of 291, is the second most sought-after skill. Although Excel may not be a dedicated data science tool, its presence in the graph indicates its importance in data manipulation, basic analytics, and reporting tasks. Proficiency in Excel is valuable for handling and organizing data effectively. *Hadoop, Spark, and AWS*, with frequencies of 140, 186, and 172 respectively, highlight the significance of big data processing and cloud computing skills in data science roles. Companies dealing with large volumes of data require professionals who can work with distributed computing frameworks like Hadoop and Spark, as well as utilize cloud platforms like AWS for data storage, processing, and analysis. *Tableau*, appearing 122 times, emphasizes the importance of data visualization skills in the field of data science. Effective data visualization enables data scientists to communicate insights and findings to stakeholders in a visually appealing and meaningful manner. The frequency of *big data*, with 136 occurrences, underscores the growing need for professionals who can handle and analyze large and complex datasets. As data continues to grow in size and complexity, individuals with skills and expertise in managing big data will remain in high demand.

Generally, this analysis provides job seekers with valuable information about the skills that are currently sought after in the data science job market. By acquiring or improving these skills, job seekers can enhance their employability and stand out among other candidates. It is advisable for individuals interested in data science careers to focus on developing proficiency in Python, Excel, Hadoop, Spark, AWS, Tableau, and big data techniques to increase their chances of securing desired positions.

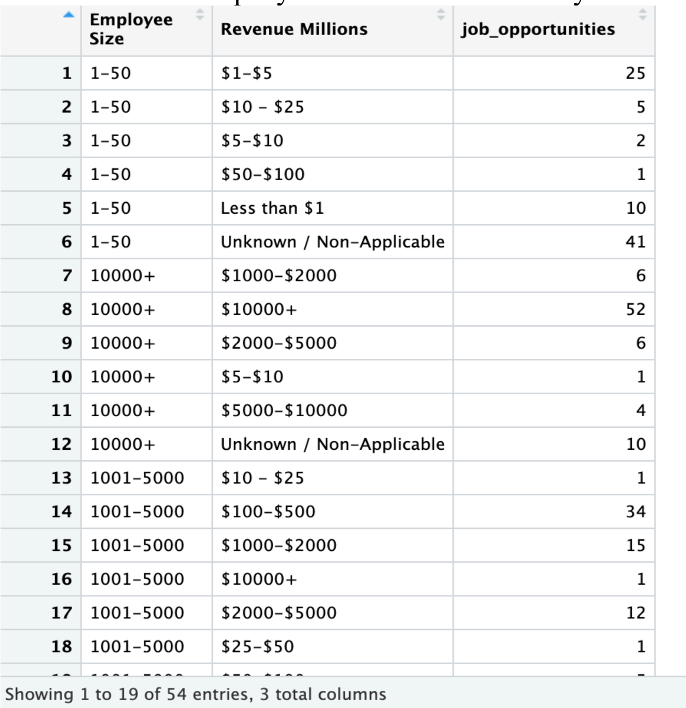
ANALYSIS 6. Distribution of Company Size (Employees) and revenues in relation to Job opportunities**: How does company size and revenue relate to the number of job opportunities in the data science field?**

The analysis focuses on exploring the relationship between company size (measured by the number of employees) and revenue with the number of job opportunities in the data science field. The analysis provides insights into how these factors are related and their impact on the availability of job opportunities. Here is the applied codes step by step:

```
  #Step1# Company Size and Revenue Analysis

size_revenue_analysis <- DS_Jobs %>%

  group_by(`Employee Size`, `Revenue Millions`) %>%

  summarize(job_opportunities = n())

#2 Preview

size_revenue_analysis

##

# A tibble: 54 × 3

# Groups:   Employee Size [8]

   `Employee Size` `Revenue Millions`       job_opportunities

   <chr>           <chr>                                <int>

 1 1-50            $1-$5                                   25

 2 1-50            $10 - $25                                5
 3 1-50           $5-$10                          2
 4 1-50           $50-$100                        1
 5 1-50           Less than $1                   10
 6 1-50           Unknown / Non-Applicable       41
 7 10000+         $1000-$2000                     6
 8 10000+         $10000+                        52
 9 10000+         $2000-$5000                     6
10 10000+         $5-$10                          1
# … with 44 more rows
```

The analysis summarized and the results of the analysis are presented in a table 7, showing the different combinations of company size and revenue categories along with the corresponding count of job opportunities.

Table 7. Company Size and Revenue Analysis



	Employee Size	Revenue Millions	job_opportunities
1	1-50	$1-$5	25
2	1-50	$10 - $25	5
3	1-50	$5-$10	2
4	1-50	$50-$100	1
5	1-50	Less than $1	10
6	1-50	Unknown / Non-Applicable	41
7	10000+	$1000-$2000	6
8	10000+	$10000+	52
9	10000+	$2000-$5000	6
10	10000+	$5-$10	1
11	10000+	$5000-$10000	4
12	10000+	Unknown / Non-Applicable	10
13	1001-5000	$10 - $25	1
14	1001-5000	$100-$500	34
15	1001-5000	$1000-$2000	15
16	1001-5000	$10000+	1
17	1001-5000	$2000-$5000	12
18	1001-5000	$25-$50	1

Showing 1 to 19 of 54 entries, 3 total columns

The resulting table provides insights into this relationship. Here are some observations: The "Employee Size" column represents different ranges of company sizes based on the number of employees. Examples include "1-50" (indicating companies with 1 to 50 employees) and "10000+" (representing companies with 10,000 or more employees). The "Revenue Millions" column categorizes the revenue of companies in million-dollar ranges. For instance, "$1-$5" signifies companies with revenues between 1 million and 5 million dollars. The "job_opportunities" column indicates the number of job opportunities available for each combination of employee size and revenue range. This metric helps us understand the level of job opportunities in relation to company size and revenue.

```
#Step_2# Bar chart for Company Size and Revenue Analysis

ggplot(size_revenue_analysis, aes(x = `Employee Size`, y = `Revenue Millions`, fill = job_o
pportunities, tooltip = job_opportunities)) +

  geom_tile(color = "white", size = 0.5) +

  labs(x = "", y = "Revenue", fill = "Job Opportunities") +
  ggtitle("Distribution of Company Sizes and Revenues in Relation to Job Opportunities") +

  scale_fill_gradient(low = "lightblue", high = "darkblue") +

  theme_minimal() +

  theme(axis.text.x = element_text(angle = 45, hjust = 1), legend.position = "right")

         #3 More visible

ggplot(size_revenue_analysis, aes(x = `Employee Size`, y = `Revenue Millions`, fill = job_opportunities)) +

  geom_tile(color = "white", size = 0.5) +

  labs(x = "Company Size", y = "Revenue (Millions)", fill = "Job Opportunities",

       title = "Distribution of Job Opportunities by Company Size and Revenue") +

  ggtitle("Distribution of Company Sizes and Revenues in Relation to Job Opportunities") +

  scale_fill_gradient(low = "#88CCEE", high = "#004488", name = "Job Opportunities") +

  theme_minimal() +

  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 12, face = "bold"),

        axis.text.y = element_text(size = 12),

        axis.title = element_text(size = 14, face = "bold"),

        plot.title = element_text(size = 16, face = "bold", hjust = 0.5),

        legend.position = "right",

        legend.title = element_text(size = 12),

        legend.text = element_text(size = 10),

        panel.grid = element_blank(),

        panel.border = element_rect(color = "black", fill = NA, linewidth = 0.5))
```

The visualization created represents the distribution of job opportunities based on company size and revenue in the data science field. The analysis provides insights into how company size and revenue relate to the number of job opportunities available. The x-axis of the Figure 8 represents different ranges of company sizes. The y-axis represents revenue ranges in million dollars. The number of job opportunities is displayed in the tooltip, providing specific information about the job count for each company size and revenue combination. The fill color of the tiles represents the number of job opportunities in each category. Darker shades of blue indicate a higher number of job opportunities, while lighter shades represent a lower number. By examining the size and color of the tiles, we can quickly identify the revenue and company size categories with the most job opportunities.

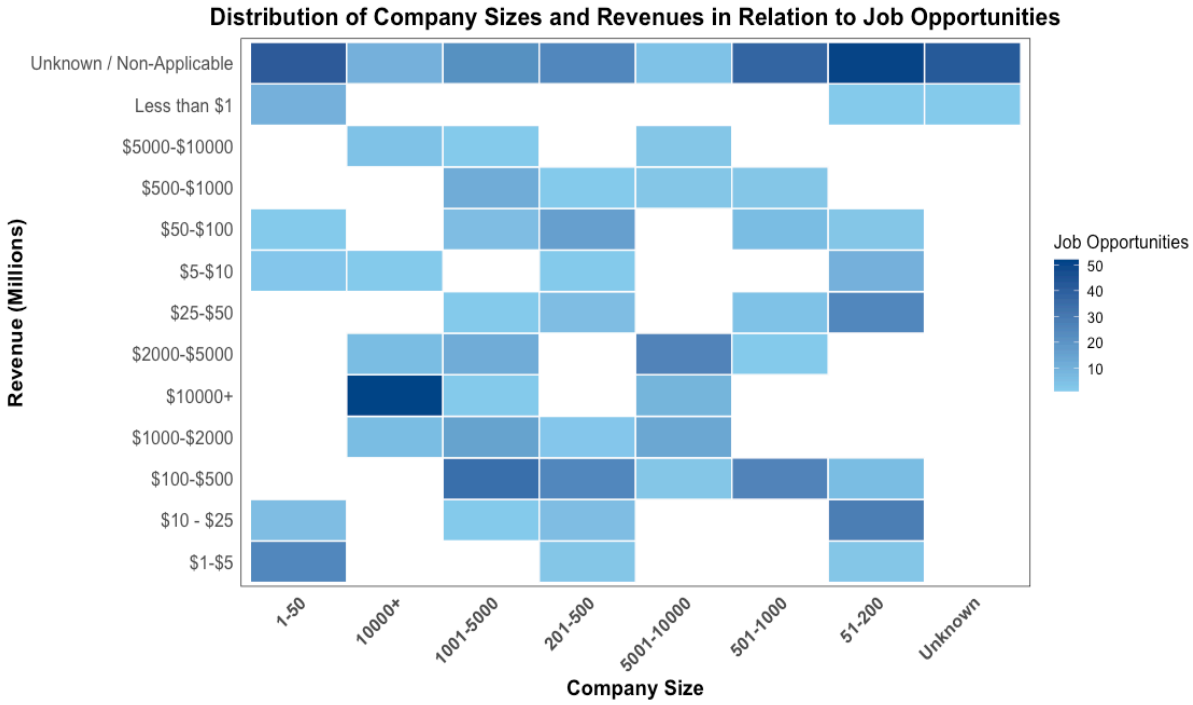


Figure 8. Distribution of Company Size and revenues in relation to Job opportunities

Result: As shown in Figure 8, Smaller companies, particularly those in the "1-50" employee range, offer a substantial number of job opportunities. This suggests that job seekers who enjoy a dynamic and agile work environment may find sufficient prospects in startups and smaller organizations. Companies with higher revenues, such as those in the "$1000-$2000" and "$10000+" categories, provide a significant number of job opportunities. These organizations typically have more resources to invest in data science teams and advanced technologies, making them attractive for those seeking stability and career growth. Startups display a notable presence within the smaller company size range, highlighting their potential for high growth and disruptive innovation. Job seekers with an entrepreneurial mindset and a passion for cutting-edge projects might find exciting opportunities in this sector. Different revenue ranges correspond to different job specialization needs. For instance, the "$2000-$5000" revenue range may indicate a demand for data scientists with specialized domain expertise, such as healthcare analytics or financial modeling. Mid-sized companies, including those in the "201-500" employee range, offer a diverse range of job opportunities. They provide a balance between stability and growth potential, making them an attractive option for job seekers seeking a collaborative environment with a mix of projects.

In conclusion, the Figure 8 suggests that job opportunities in data science are available across various company sizes and revenue ranges. Conducting further research and understanding the specific industry dynamics can help individuals align their interests with the right company size and revenue range to enhance their chances of success in the data science job market.

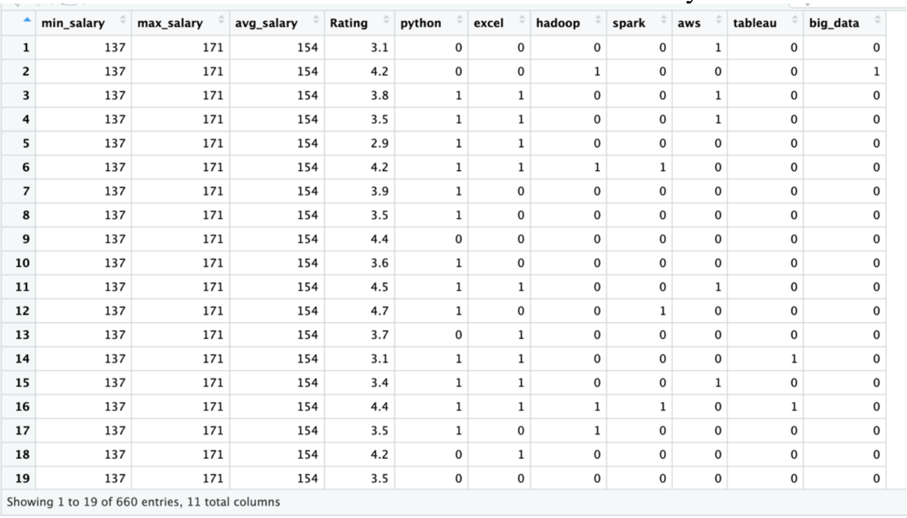
ANALYSIS 7: Identify how skills and ratings correlate with salary indicators.

The analysis focuses on performing correlation analysis on this dataset, we can investigate the relationships between different variables. Correlation measures the strength and direction of the linear relationship between two variables. In this case, we can explore how skills and ratings correlate with salary indicators.

```
 #Step_1# Select relevant columns for correlation analysis

correlation_data <- DS_Jobs %>%
  select(min_salary, max_salary, avg_salary, Rating, python, excel, hadoop, spark, aws, tab
leau, big_data)

#2 Preview

correlation_data

##

# A tibble: 660 × 11

   min_salary max_salary avg_sa…¹ Rating python excel hadoop spark   aws tableau big_d…²

        <dbl>      <dbl>    <dbl>  <dbl>  <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>   <dbl>

 1        137        171      154    3.1      0     0      0     0     1       0       0

 2        137        171      154    4.2      0     0      1     0     0       0       1

 3        137        171      154    3.8      1     1      0     0     1       0       0

 4        137        171      154    3.5      1     1      0     0     1       0       0

 5        137        171      154    2.9      1     1      0     0     0       0       0

 6        137        171      154    4.2      1     1      1     1     0       0       0

 7        137        171      154    3.9      1     0      0     0     0       0       0

 8        137        171      154    3.5      1     0      0     0     0       0       0

 9        137        171      154    4.4      0     0      0     0     0       0       0

10        137        171      154    3.6      1     0      0     0     0       0       0

# … with 650 more rows, and abbreviated variable names ¹avg_salary, ²big_data
```

Table 8 has been generated for correlation with selected data.

Table 8. Selected data for Correlation Analysis



	min_salary	max_salary	avg_salary	Rating	python	excel	hadoop	spark	aws	tableau	big_data
1	137	171	154	3.1	0	0	0	0	1	0	0
2	137	171	154	4.2	0	0	1	0	0	0	1
3	137	171	154	3.8	1	1	0	0	1	0	0
4	137	171	154	3.5	1	1	0	0	1	0	0
5	137	171	154	2.9	1	1	0	0	0	0	0
6	137	171	154	4.2	1	1	1	1	0	0	0
7	137	171	154	3.9	1	0	0	0	0	0	0
8	137	171	154	3.5	1	0	0	0	0	0	0
9	137	171	154	4.4	0	0	0	0	0	0	0
10	137	171	154	3.6	1	0	0	0	0	0	0
11	137	171	154	4.5	1	1	0	0	1	0	0
12	137	171	154	4.7	1	0	0	1	0	0	0
13	137	171	154	3.7	0	1	0	0	0	0	0
14	137	171	154	3.1	1	1	0	0	0	1	0
15	137	171	154	3.4	1	1	0	0	1	0	0
16	137	171	154	4.4	1	1	1	1	0	1	0
17	137	171	154	3.5	1	0	1	0	0	0	0
18	137	171	154	4.2	0	1	0	0	0	0	0
19	137	171	154	3.5	0	0	0	0	0	0	0

Showing 1 to 19 of 660 entries, 11 total columns

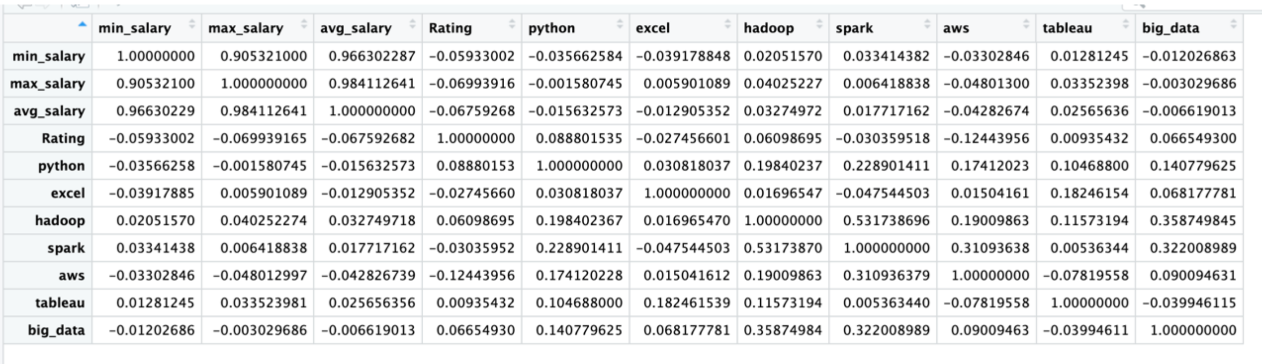
Then we calculate the correlation matrix for the selected dataset. The **'correlation_matrix'** is calculated using the **'cor()'** function on the selected dataset. This function computes the correlation coefficients between all pairs of variables in the dataset.

The resulting correlation matrix is then printed, displaying the correlation coefficients between each pair of variables. The matrix is a square table where the rows and columns represent the variables, and each cell contains the correlation coefficient between the corresponding variables.

```
#Step_2# Calculate correlation matrix

correlation_matrix <- cor(correlation_data)

#2 Print correlation matrix

print(correlation_matrix)

##

min_salary   max_salary   avg_salary      Rating       python        excel      hadoop        spark          aws     ta
bleau

min_salary  1.00000000  0.905321000  0.966302287 -0.05933002 -0.035662584 -0.039178848 0.02051570  0.033414382 -0.033
02846  0.01281245

max_salary  0.90532100  1.000000000  0.984112641 -0.06993916 -0.001580745  0.005901089 0.04025227  0.006418838 -0.048
01300  0.03352398

avg_salary  0.96630229  0.984112641  1.000000000 -0.06759268 -0.015632573 -0.012905352 0.03274972  0.017717162 -0.042
82674  0.02565636

Rating     -0.05933002 -0.069939165 -0.067592682  1.00000000  0.088801535 -0.027456601 0.06098695 -0.030359518 -0.124
43956  0.00935432

python     -0.03566258 -0.001580745 -0.015632573  0.08880153  1.000000000  0.030818037 0.19840237  0.228901411  0.174
12023  0.10468800

excel      -0.03917885  0.005901089 -0.012905352 -0.02745660  0.030818037  1.000000000 0.01696547 -0.047544503  0.015
04161  0.18246154

hadoop      0.02051570  0.040252274  0.032749718  0.06098695  0.198402367  0.016965470 1.00000000  0.531738696  0.190
09863  0.11573194

spark       0.03341438  0.006418838  0.017717162 -0.03035952  0.228901411 -0.047544503 0.53173870  1.000000000  0.310
93638  0.00536344

aws        -0.03302846 -0.048012997 -0.042826739 -0.12443956  0.174120228  0.015041612 0.19009863  0.310936379  1.000
00000 -0.07819558

tableau     0.01281245  0.033523981  0.025656356  0.00935432  0.104688000  0.182461539 0.11573194  0.005363440 -0.078
19558  1.00000000

big_data   -0.01202686 -0.003029686 -0.006619013  0.06654930  0.140779625  0.068177781 0.35874984  0.322008989  0.090
09463 -0.03994611

               big_data

min_salary -0.012026863

max_salary -0.003029686

avg_salary -0.006619013

Rating      0.066549300

python      0.140779625

excel       0.068177781

hadoop      0.358749845

spark       0.322008989

aws         0.090094631

tableau    -0.039946115

big_data    1.000000000
```

As shown in table 9, the correlation matrix provides insights into the relationships between different variables in the dataset. The values in the matrix range from -1 to 1, where -1 represents a strong negative correlation, 1 represents a strong positive correlation, and 0 represents no correlation.

Table 9. Calculated Correlation Matrix



	min_salary	max_salary	avg_salary	Rating	python	excel	hadoop	spark	aws	tableau	big_data
min_salary	1.00000000	0.905321000	0.966302287	-0.05933002	-0.035662584	-0.039178848	0.02051570	0.033414382	-0.03302846	0.01281245	-0.012026863
max_salary	0.90532100	1.000000000	0.984112641	-0.06993916	-0.001580745	0.005901089	0.04025227	0.006418838	-0.04801300	0.03352398	-0.003029686
avg_salary	0.96630229	0.984112641	1.000000000	-0.06759268	-0.015632573	-0.012905352	0.03274972	0.017717162	-0.04282674	0.02565636	-0.006619013
Rating	-0.05933002	-0.069939165	-0.067592682	1.00000000	0.088801535	-0.027456601	0.06098695	-0.030359518	-0.12443956	0.00935432	0.066549300
python	-0.03566258	-0.001580745	-0.015632573	0.08880153	1.000000000	0.030818037	0.19840237	0.228901411	0.17412023	0.10468800	0.140779625
excel	-0.03917885	0.005901089	-0.012905352	-0.02745660	0.030818037	1.000000000	0.01696547	-0.047544503	0.01504161	0.18246154	0.068177781
hadoop	0.02051570	0.040252274	0.032749718	0.06098695	0.198402367	0.016965470	1.00000000	0.531738696	0.19009863	0.11573194	0.358749845
spark	0.03341438	0.006418838	0.017717162	-0.03035952	0.228901411	-0.047544503	0.53173870	1.000000000	0.31093638	0.00536344	0.322008989
aws	-0.03302846	-0.048012997	-0.042826739	-0.12443956	0.174120228	0.015041612	0.19009863	0.310936379	1.00000000	-0.07819558	0.090094631
tableau	0.01281245	0.033523981	0.025656356	0.00935432	0.104688000	0.182461539	0.11573194	0.005363440	-0.07819558	1.00000000	-0.039946115
big_data	-0.01202686	-0.003029686	-0.006619013	0.06654930	0.140779625	0.068177781	0.35874984	0.322008989	0.09009463	-0.03994611	1.000000000

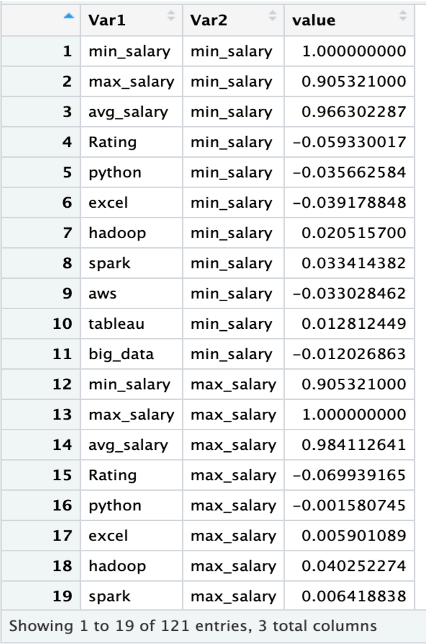
In the next step, the **'melt()'** function is applied to the **'correlation_matrix'** to reshape it into a data frame format. The **'melt()'** function is commonly used to transform wide-format data into long-format data, making it easier to work with and visualize.

```
#4 Reshape the correlation matrix into a data frame
correlation_df <- melt(correlation_matrix)
```

The resulting '**correlation_df'** data frame will have the following structure:

- The data frame will consist of three columns: '**variable1'**, '**variable2'**, and '**value'**.
- The '**variable1'** and '**variable2'** columns will represent the variables in the original correlation matrix, indicating the pair of variables for which the correlation coefficient is calculated.
- The '**value'** column will contain the correlation coefficient values from the original correlation matrix.

Table 10. Reshaped the correlation matrix into a data frame



	Var1	Var2	value
1	min_salary	min_salary	1.000000000
2	max_salary	min_salary	0.905321000
3	avg_salary	min_salary	0.966302287
4	Rating	min_salary	-0.059330017
5	python	min_salary	-0.035662584
6	excel	min_salary	-0.039178848
7	hadoop	min_salary	0.020515700
8	spark	min_salary	0.033414382
9	aws	min_salary	-0.033028462
10	tableau	min_salary	0.012812449
11	big_data	min_salary	-0.012026863
12	min_salary	max_salary	0.905321000
13	max_salary	max_salary	1.000000000
14	avg_salary	max_salary	0.984112641
15	Rating	max_salary	-0.069939165
16	python	max_salary	-0.001580745
17	excel	max_salary	0.005901089
18	hadoop	max_salary	0.040252274
19	spark	max_salary	0.006418838

Showing 1 to 19 of 121 entries, 3 total columns

Finally, in the next step, we can generate a heatmap visualization of the correlation matrix using the **'ggplot2'** package. The **'scale_fill_gradient2()'** function is used to define the color gradient for the fill based on the correlation coefficient values. The gradient ranges from light blue to dark blue, with a midpoint at 0 (white color). Also, additional theme settings are specified using the **'theme()'** function to customize the appearance of the plot, such as bold axis labels, title size, and legend text size:

```
#Step_3# Create the Heatmap

ggplot(correlation_df, aes(x = Var2, y = Var1, fill = value)) +

geom_tile(color = "white") +

  scale_fill_gradient2(low = "lightblue", mid = "white", high = "darkblue", midpoint = 0) +

  labs(title = "Correlation Heatmap",

       x = "Variables",

       y = "Variables",

       fill = "Correlation") +

  theme_minimal() +

  theme(axis.text.x = element_text(angle = 45, hjust = 1, face = "bold"),  # Make x-axis labels bold

        axis.text.y = element_text(face = "bold"),  # Make y-axis labels bold

        plot.title = element_text(size = 16, face = "bold"),

        axis.title = element_text(size = 14, face = "bold"),

        legend.title = element_text(size = 12, face = "bold"),

        legend.text = element_text(size = 10, face = "bold"))
```

The resulting plot will be a heatmap showing the correlations between variables, with positive correlations indicated by darker blue shades and negative correlations indicated by lighter blue shades. The diagonal elements of the heatmap will represent the self-correlations of variables (which will be 1.0 by definition), see Figure 9 below:

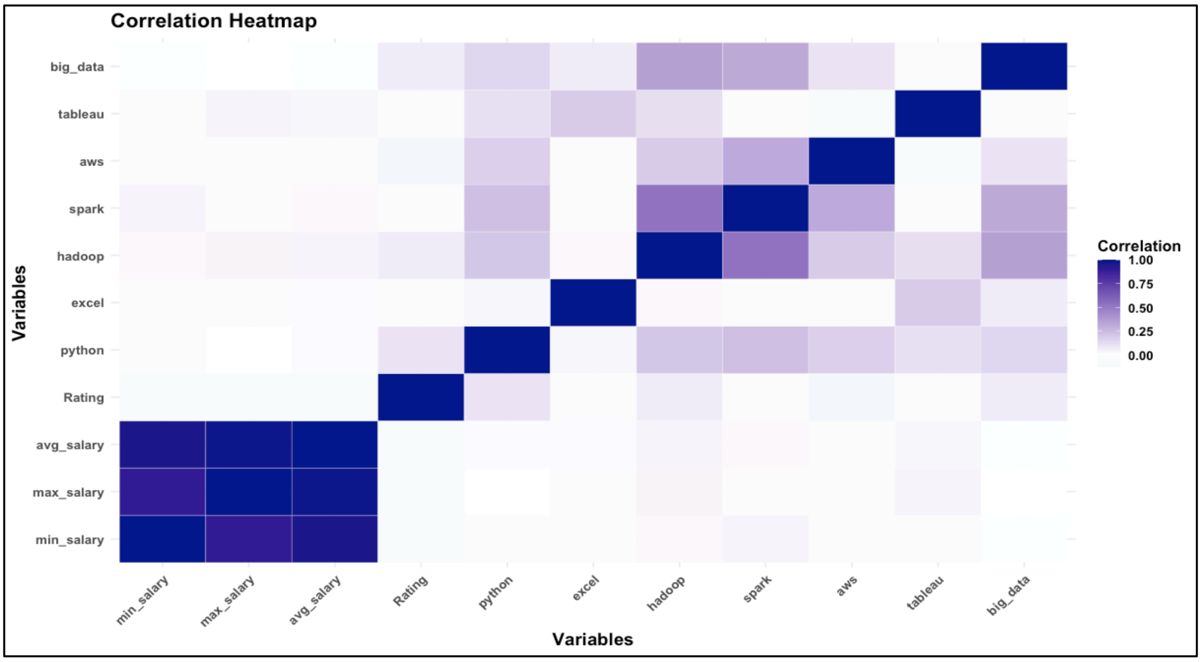


Figure 9. Correlations between variables.

Result: Each cell in the figure 9 represents the correlation coefficient between two variables, with color indicating the strength and direction of the correlation. From the figure 9, The correlation heatmap, we can gain the following knowledge:

1. Salary Correlations: The salary variables (**min_salary**, **max_salary**, **avg_salary**) exhibit strong positive correlations with each other. This is evident from the dark blue tiles along the diagonal, indicating that higher salaries tend to be positively correlated across the different measures.
2. Technical Skills: The heatmap reveals interesting correlations between technical skills. For example, there is a moderate positive correlation between **python** and **excel**, indicating that individuals with proficiency in Python also tend to have proficiency in Excel. Similarly, there are moderate positive correlations between **python** and other skills like **hadoop**, **spark**, and **aws**, suggesting that these skills are often used together.
3. Rating: The **Rating** variable, representing the rating of a job, shows weak negative correlations with the salary variables. This suggests that higher-rated jobs tend to have slightly lower salaries, although the correlation is not very strong.
4. Big Data: The **big_data** variable shows a moderate positive correlation with several skills like **hadoop**, **spark**, and **aws**. This indicates that proficiency in big data technologies is often associated with skills in these specific areas.

To conclude, the correlation heatmap in figure 9 highlights strong correlations between salary measures, associations between technical skills, and the relationship between job rating and salary. However, it's important to note that correlation does not imply causation, and further analysis is required to understand the underlying factors driving these relationships.

ANALYSIS 8. Job Type Seniority Distribution: **What is the distribution of job types and seniority levels in the data science field, providing insights into the composition of data science positions?**

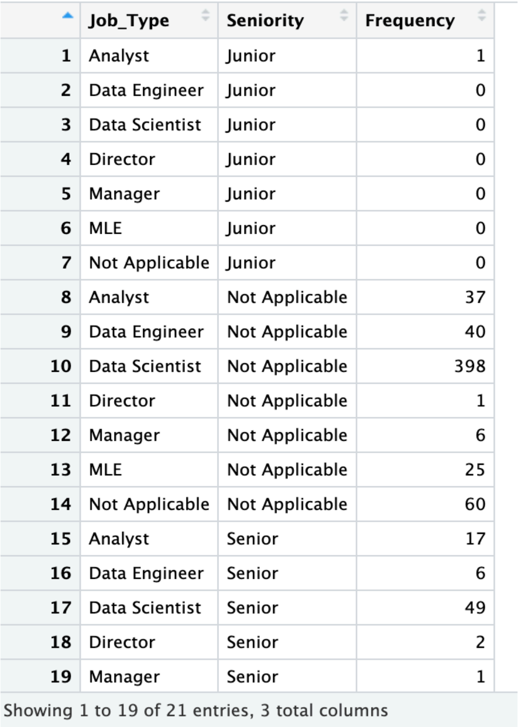
The analysis focuses on the distribution of job types and seniority levels in the data science field, allowing for further analysis and insights into the distribution of job roles and seniority within the data. To achieve the results, the following codes has been applied:

```
#Step_1# Calculate the frequency of job types and seniority levels

job_type_seniority_counts <- table(DS_Jobs$job_simp, DS_Jobs$seniority)

#2 Create a data frame with job types, seniority levels, and their frequencies

job_type_seniority_df <- as.data.frame(job_type_seniority_counts)

#3 Preview

job_type_seniority_df

##

  Var1           Var2 Freq

1         Analyst         Junior     1

2    Data Engineer         Junior     0

3   Data Scientist         Junior     0

4        Director         Junior     0

5         Manager         Junior     0

6             MLE         Junior     0

7   Not Applicable         Junior     0

8         Analyst Not Applicable    37

9    Data Engineer Not Applicable    40
```

```
10 Data Scientist Not Applicable  398

11       Director Not Applicable    1

12        Manager Not Applicable    6

13            MLE Not Applicable   25

14 Not Applicable Not Applicable   60

15        Analyst         Senior   17

16  Data Engineer         Senior    6

17 Data Scientist         Senior   49

18       Director         Senior    2

19        Manager         Senior    1

20            MLE         Senior    9

21 Not Applicable         Senior    8
       #4 Rename the columns for better readability

colnames(job_type_seniority_df) <- c("Job_Type", "Seniority", "Frequency")
```

Then the generated table columns renamed for better readability and the resulting data frame, '**job_type_seniority_df**', displays the job types, seniority levels, and their respective frequencies as shown in table 11 below:

Table 11. Job seniority counts



	Job_Type	Seniority	Frequency
1	Analyst	Junior	1
2	Data Engineer	Junior	0
3	Data Scientist	Junior	0
4	Director	Junior	0
5	Manager	Junior	0
6	MLE	Junior	0
7	Not Applicable	Junior	0
8	Analyst	Not Applicable	37
9	Data Engineer	Not Applicable	40
10	Data Scientist	Not Applicable	398
11	Director	Not Applicable	1
12	Manager	Not Applicable	6
13	MLE	Not Applicable	25
14	Not Applicable	Not Applicable	60
15	Analyst	Senior	17
16	Data Engineer	Senior	6
17	Data Scientist	Senior	49
18	Director	Senior	2
19	Manager	Senior	1

Showing 1 to 19 of 21 entries, 3 total columns

Then we move to step 2, to observe more through creating a stacked bar chart. The stacked bar chart visualizes the distribution of job types and seniority levels in the data science field. Each bar represents a job type, and the bars are stacked to represent the frequencies of different seniority levels within each job type.

```
#Step_2# Create a stacked bar chart for job type and seniority distribution
```

```
ggplot(job_type_seniority_df, aes(x = Job_Type, y = Frequency, fill = Seniority)) +
  geom_bar(stat = "identity", color = "black", width = 0.6) +
  labs(x = "Job Type", y = "Frequency", fill = "Seniority") +
  ggtitle("Job Type and Seniority Distribution") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1, size = 11, color = "black"),
        axis.text.y = element_text(size = 10, color = "black"),
        axis.title.x = element_text(size = 12, face = "bold"),
        axis.title.y = element_text(size = 12, face = "bold"),
        plot.title = element_text(size = 14, face = "bold", hjust = 0.5),
        legend.title = element_text(size = 10, face = "bold"),
        legend.text = element_text(size = 10),
        legend.position = "right",
        panel.grid.major = element_line(color = "lightgray", linetype = "dashed"),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank())
```

After executing the codes above, a stacked bar chat has been created in figure 10. The figure 10 includes job types such as "Analyst," "Data Engineer," "Data Scientist," "Director," "Manager," "MLE," and "Not Applicable". The seniority levels are represented by different colors in the figure 10, including "Junior," "Not Applicable," and "Senior". The height of each stack represents the frequency of a specific seniority level within a job type. The total height of each bar represents the total frequency of that particular job type.

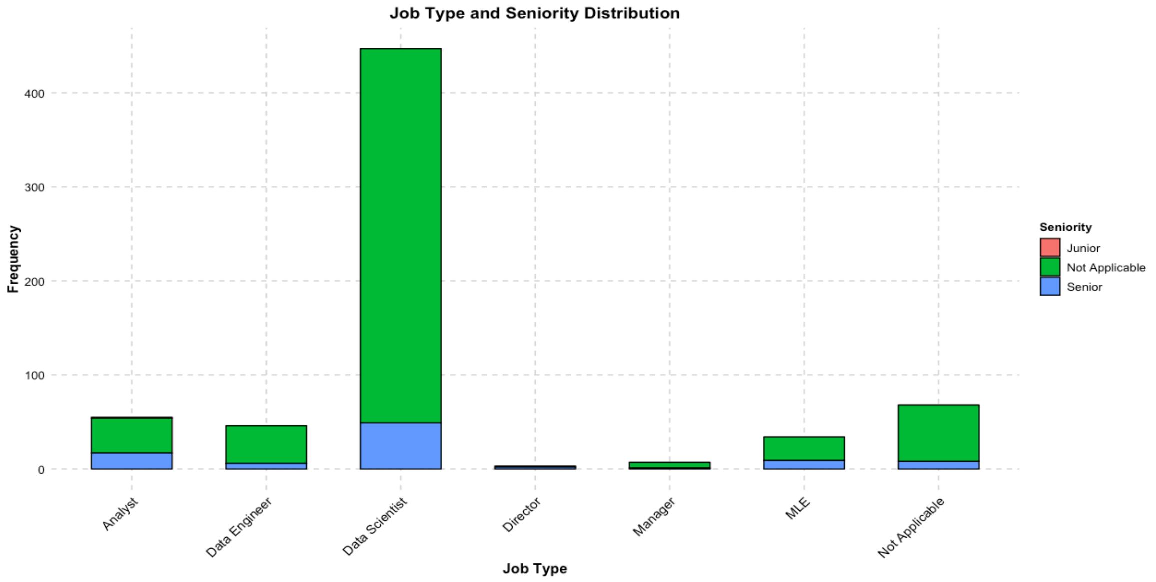


Figure 10. Job Type Seniority Distribution

Result: The distribution of job types and seniority levels in the data science field, as shown in the Figure 10, provides valuable insights into the composition of data science positions and can be related to various aspects. The figure 10 reveals the relative frequencies of different job types in the data science field. The "Data Scientist" job type has the highest frequency of senior positions, followed by "Analyst," "MLE," and "Data Engineer." Other job types such as "Manager" and "Director" have a relatively lower frequency of senior positions. Job seekers can gain an understanding of the demand and popularity of various roles. For example, the high frequency of "Data Scientist" positions indicates a strong demand for professionals in this role. Conversely, job types with lower frequencies may suggest a relatively smaller market or specific requirements for those roles. The distribution of seniority levels highlights the opportunities available at different career stages. Job seekers can assess the prevalence of junior, senior, and managerial positions in the data science field. This information helps in aligning career goals and expectations with the current market trends. For instance, if a job seeker aims to secure a senior position, they can identify the job types and seniority levels that offer more opportunities in that regard. The "Not Applicable" seniority level is visible across multiple job types, indicating that many positions do not specify a particular seniority level.

Conclusion

Throughout this project, we have explored and analyzed a dataset containing information about data science job postings. We began by loading the clean dataset and performing an initial exploration to understand its structure and variables. We examined the first few rows, checked for missing values, and summarized the dataset's key statistics.

Next, we dived into the exploratory data analysis phase. We investigated various aspects of the data science job market, including the distribution of job titles, the distribution of job seniority levels, and the distribution of job locations. Through visualizations such as bar charts, pie charts, and heatmaps, we gained insights into the frequencies and proportions of different variables, allowing us to identify popular job titles, prevalent seniority levels, and prominent job locations.

Moving forward, we conducted a correlation analysis to explore relationships between different variables in the dataset. By selecting relevant columns and calculating the correlation matrix, we obtained a numeric representation of the relationships between job salary, company ratings, and various technical skills such as Python, Excel, Hadoop, Spark, AWS, Tableau, and Big Data. We visualized the correlation matrix using a heatmap, which provided a clear overview of the associations between these variables.

Lastly, we examined the distribution of job types and seniority levels within the data science field. By calculating the frequencies and creating a stacked bar chart, we gained insights into the composition of data science positions, identified prevalent job types, and explored the availability of junior, senior, and managerial roles. We discussed the implications of these findings for job seekers, such as understanding market demand, identifying career progression paths, and uncovering niche opportunities.

Acknowledgement: I would like to express my heartfelt appreciation to **Professor Kang Lele** for his invaluable teachings and guidance throughout the course/semester. His expertise in data analytics and use of R language, combined with practical examples and hands-on practice, have equipped us with essential skills and knowledge. I am grateful for his dedication and support, which have made this learning experience and this project both enjoyable and enriching. Thank you, **Professor Kang Lele**, for your exceptional contribution to our growth as data analysts.